Meaning Refinement to Improve Cross-lingual Information Retrieval

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von: M.Sc. Farag Ahmed
geb. am 02. März. 1972 in Libyen

Gutachter:
Prof. Dr. Andreas Nürnberger
Prof. Dr. Kamel Smaïli
Prof. Dr. Maciej Piasecki

Magdeburg, den 17.01.2012

Farag Ahmed

Meaning Refinement to Improve Cross-lingual Information Retrieval

Abstract

Cross-lingual information retrieval (CLIR) adds a way to efficiently transfer information across languages. However, to achieve this goal, the limitations imposed by the language barriers, such as problems with multiple word meanings, is a serious issue. Therefore, to support a user, to get information across languages, the user's information need (e.g., a specific query) has to be translated. This translation is not a trivial task, especially for some morphologically complex languages such as Arabic. Arabic is a morphologically complex language, in that it provides flexibility in word formation (inflection), making it possible to derive hundreds of words from only one root. Furthermore, due to the lack of coverage of existing dictionaries, compounds that appear frequently in languages such as German, Dutch etc., cause low performance in cross-lingual retrieval. Therefore, in order to improve the performance of cross-lingual systems, these compounds need to be decompounded before translation. After possible translations (senses) are obtained, one of the main problems that impacts the performance of cross-lingual retrieval systems is how to disambiguate translations and - since this usually cannot be done completely automatically - how to smoothly integrate a user in this disambiguation process.

In this thesis, firstly, fundamental approaches such as stemming, spelling correction, decompounding and cross-lingual retrieval approaches and issues are studied in detail. Furthermore, state-of-the art cross-lingual interactive tools are reviewed and discussed. The spotlight of the work, presented in this thesis, builds on exploiting word correspondence across languages for Word Sense Disambiguation (WSD) in a query-based translation scenario. Furthermore, it builds on exploiting parallel linguistic resources for overcoming the user's lack of knowledge in the target language. We designed a cross-lingual interactive tool in order to investigate the feasibility and the validity of utilizing translations for cross-lingual retrieval. To ensure that a user has a certain confidence in selecting a translation, which he/she possibly cannot even read or understand, the designed tool provides sufficient information about translation alternatives and their meaning so that the user has a certain degree of confidence in the translation. This is achieved by automatically translating the user query and then providing possibilities to interactively select relevant terms obtained from corpora. The selected relevant terms can be used to improve the translation (and thus improve the cross-lingual retrieval process), if needed. A human judgment experiment was designed to obtain an evaluation of the functionality of the tool. The result of the user study was used as a reference point to improve the tool's functionality, which has been employed in a revised design.

Zusammenfassung

Sprachübergreifende Suche ermöglicht eine effiziente Informationenübertragung über Sprachgrenzen hinweg. Dazu müssen jedoch verschiedene, durch Sprachbarrieren hervorgerufene Hürden überwunden werden wie beispielsweise das Problem der Wortmehrdeutigkeiten. Um den Nutzer dabei zu unterstützen, Informationen über verschiedene Sprachen hinweg zu erhalten, muss die Anfrage zunächst übersetzt werden. Diese Übersetzung ist keine triviale Aufgabe, insbesondere für morphologisch komplexe Sprachen wie Arabisch. Arabisch ist eine morphologisch komplexe Sprache, da Flexibilität in der Wortbildung (Flexion) erlaubt ist und so Hunderte von Wörtern aus einem einzigen Wortstamm abgeleitet werden können. Weiterhin stellen zusammengesetzte Wörter, die häufig in Sprachen wie Deutsch oder Niederländisch auftreten, ein Problem dar, da sie unzureichend von existierenden Wörterbüchern abgedeckt werden. Solche Wörter müssen daher vor dem Übersetzen aufgespaltet werden. Nachdem eine Anfrage-Übersetzung (Bedeutung) durchgeführt wurde, besteht ein wesentliches performanzkritisches Problem darin, Mehrdeutigkeiten gefundener Übersetzungen aufzulösen (disambiguieren) und - da dies nicht vollständig automatisch erfolgen kann - den Benutzer dabei nahtlos in den Begriffsklärungsprozess zu integrieren.

In dieser Arbeit werden zunächst grundlegende Techniken wie Wortstammbildung und Rechtschreibkorrektur sowie Ansätze und Probleme sprachübergreifender Suche im Detail untersucht. Darüber hinaus wird der Stand der Technik interaktiver Werkzeuge zur sprachübergreifenden Suche diskutiert. Der Kern der Arbeit beschreibt, wie Wort-Korrespondenzen über verschiedene Sprachen hinweg zur Auflösung von Mehrdeutigkeiten in einem anfragebasierten Übersetzungsszenario genutzt werden können und wie sich mit Hilfe paralleler linguistischer Ressourcen fehlendes Wissen des Benutzers über die Zielsprache kompensieren lässt. Im Rahmen der Arbeit wurde ein sprachübergreifendes interaktives Werkzeug entwickelt um die Machbarkeit und Wirksamkeit der Verwendung von Übersetzungen für sprachübergreifende Suche zu untersuchen. Das entwickelte System bietet interaktiv kontextuelle Informationen zu alternativen Übersetzungen und deren Bedeutungen, wodurch sich beim Benutzer ein gewisses Vertrauen in die Auswahl einer Übersetzung aufbauen lässt, welche sie oder er möglicherweise nicht einmal lesen oder verstehen kann. Dies wird erreicht, indem zunächst die Nutzeranfrage automatisch übersetzt wird und anschließend die Möglichkeit besteht, interaktiv relevante Worte auszuwählen, welche aus Corpora gewonnen wurden. Die ausgewählten Worte können bei Bedarf zur Verbesserung der Übersetzung (und damit zur Verbesserung des gesamten sprachübergreifenden Suchprozesses) genutzt werden. Um eine Bewertung der Funktionalität des Werkzeugs zu erhalten wurde eine Nutzerstudie durchgeführt. Die aus der Studie gewonnenen Erkenntnisse bildeten einen Bezugspunkt für die Verbesserung des Funktionalität des Werkzeugs und führten zu einem überarbeiteten Design.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Prof. Dr. Andreas Nürnberger, for his continuous support, guidance and valuable advice in researching and implementing all approaches and tools proposed in this thesis. My appreciation also goes to Prof. Dr. Kamel Smaïli (Nancy university in France) and Prof. Dr. Maciej Piasecki (Wroclaw University of Technology) for being my thesis second examiners. I would like to express my special thank to Evelyn Wilson for the language proofreading. I would also like to thank my parents and all members of my family for their continuous encouragement and moral support.

Contents

List of Figures

List of Tables

Chapter 1

Thesis Introduction

1.1 Introduction

The increase of multilingual information on the World Wide Web has led to the necessity to develop methods and applications to make use of this multilingual information. However, language barriers are a serious issue to world communication and to economic and cultural exchange. In order to allow users, to overlap across languages, cross-lingual information retrieval (CLIR) can be used.

Cross-lingual information retrieval provides means to retrieve information written in one language while using a query expressed in a different language. However, the main research obstacle that prevents cross-lingual retrieval from performing well is the lexical ambiguity of source and target languages. In every language, there are words which have multiple meanings, which will lead to the fact that the user query can have several possible translations. In order for cross-lingual information retrieval to perform the cross-lingual search task to a good extent, this lexical ambiguity needs to be tackled or at least alleviated. In addition to the classical information retrieval tasks, cross-lingual retrieval requires that the query (or the documents) be translated from one language into another. Query translation is widely used for cross-lingual tasks, as query translation requires fewer computational resources compared to translating a large set of retrieved documents (Carbonell et al., 1997). Furthermore, users who are able to understand more than one language might not be able to effectively express their need in those languages. Those users with cross-lingual system support can cover more multilingual resources with a single query expressed in a language they are fluent in. Furthermore, cross-lingual based on query translation, can also be useful for users who can read a single language. Using query translation can narrow the examined documents, by the user, in the target language. This can reduce time and effort in comparison to translating all documents in the data set and then retrieving the relevant documents out of them. In some cases, cross-lingual retrieval can be useful for the monolingual user. For example, an industrial expert is looking for a specific pump, in a specific country and he/she would like to know if this pump is produced there. Using the cross-lingual system, the query will be translated and relevant documents will be provided. Based on examining these documents, the user might find images for the pump which meet the expectations about his/her information need. Furthermore, the user might then select one or two documents to automatically translate. Therefore, using query translation and then retrieval can be more beneficial than document translation and then retrieval (Oard, 1997b).

Despite many advantages of query translation, query translation suffers from translation ambiguity as queries are often short and do not provide rich context for disambiguation (Hull and Grefenstette, 1996; Gabrilovich et al., 2009). An alternative to translating the user query, using the cross-lingual system, is to use Machine Translation (MT). However, although it seems that cross-lingual retrieval systems and Machine Translation (MT) systems are related, the way both systems translate the given text is different. Their commonality is that both systems must produce the same given text in different languages. Machine translation systems put a lot of effort into producing syntactically correct sentences and should be read like naturally produced text, while cross-lingual retrieval systems are based on individual word translations without focusing on producing a syntactically correct translation. One clear drawback, that machine translation systems are not suitable for the cross-lingual retrieval task is that the user queries are often short and formed, usually without any proper syntactic structure (Hull and Grefenstette, 1996). Furthermore, machine translation systems provide no possibilities for the user to be involved in refining the translation in the hope of improving the retrieval performance. Therefore, the performance of current machine translation systems is low for cross-lingual retrieval (Pirkola, 1998). In the early seventies, experiments for retrieving information across languages were first initiated by Salton (1973). Currently, cross-lingual retrieval issues are addressed in several evaluation forums, such as TREC[1], CLEF[2], SemEval[3] and NTCIR[4], while each of them covers different languages: TREC includes Spanish, Chinese, German, French, Italian, and Arabic; CLEF includes French, German, Italian, Swedish, Spanish, Dutch, Finnish, and Russian; SemEval includes Dutch, French, German, Spanish and Italian and NTCIR includes Japanese, Chinese and Korean. Finding

[1] http://trec.nist.gov/trec_eval/

[2] http://clef-campaign.org/

[3] http://semeval2.fbk.eu/semeval2.php/

[4] http://research.nii.ac.jp/ntcir/index-en.html/

the most effective way to bridge the language barrier between queries and documents is the central challenge in cross-lingual retrieval (Yang and Ma, 2002).

In this thesis, besides the improvements and the implementations of statistical approaches to disambiguate the user query, a novel approach is proposed to support the user in having more confidence in the automatic translation, which they can not read or understand. The core idea is to provide possibilities to interactively select relevant terms from contextual information, in a language the user is familiar with, in order to improve the translation and thus improve the cross-lingual information retrieval process. The contextual information is displayed to the user in a language he/she is familiar with. This information is needed in order to give the user a confidence in the translation he/she can not understand and in some extreme cases can not even read.

In the following, a brief overview of the main research topics that are covered in this thesis and how they are related to each other, are given. Furthermore, an overview of the different thesis chapters is given.

1.2 General Overview of the Main Research Topics

In Figure 1.1, an abstract view of the research topics in this thesis is presented based on the building blocks of an interactive cross-lingual retrieval system. This structure will be used as a reference system throughout this thesis. The cross-lingual process starts by sending a natural language user query. This query is first pre-processed for misspelling words. Then, in order to have the appropriate translation, first, the word stem has to be identified. This step is important, especially for high morphological languages such as Arabic, since not all word form variations can be found in existing dictionaries. Second, due to the lack of coverage of existing dictionaries, compounds that appear frequently in languages such as German, Dutch etc., cause low performance in cross-lingual retrieval. Therefore, in order to improve the performance of cross-lingual systems, these compounds need to be decompounded before translation. The processed query is then translated and ranked translations are displayed to the user. Based on this translation, contextual information that describes each translation, in the user's own language, is obtained and displayed to the user. Along with this contextual information, relevant interactive terms are also displayed, which can be used to improve the translation. A post-processing step is needed for some languages such as Arabic in order to include all word form variations to improve the retrieval performance. Once the user confirms one of the translations, this translation can be submitted to the user's favorite search engine so the relevant documents will be obtained and displayed to the user.

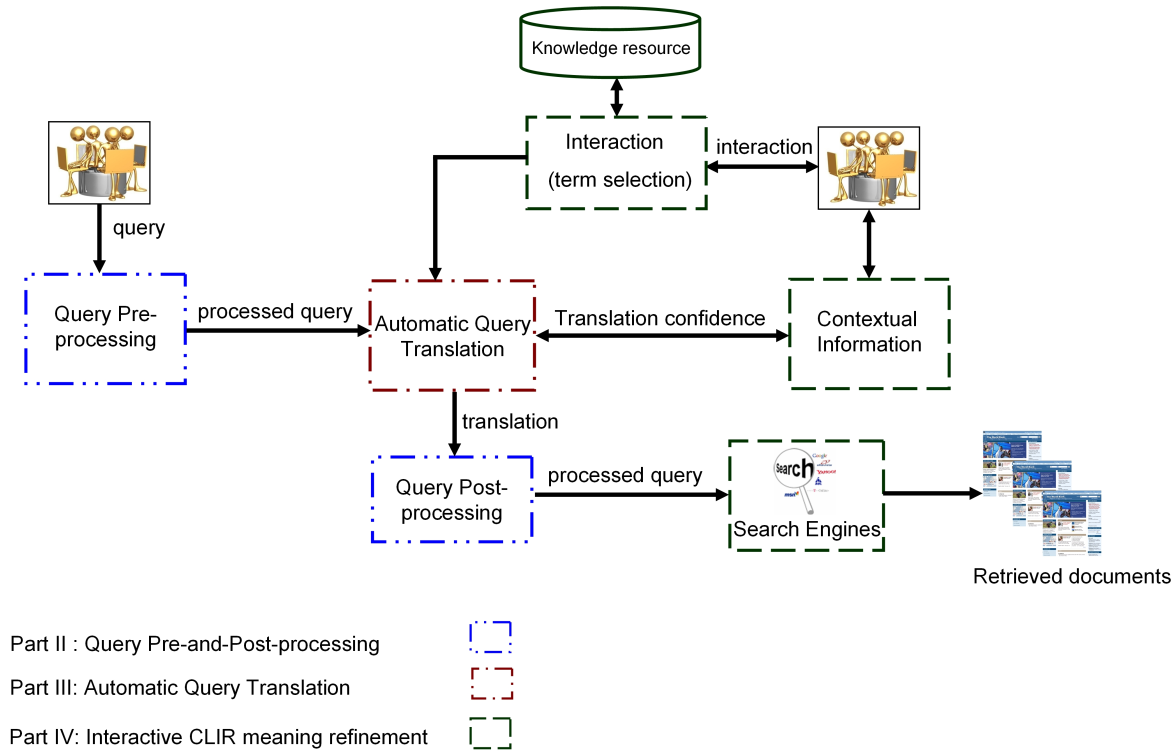


Figure 1.1: An abstract view of the main research topics.

1.3 Thesis Layout and Brief Overview of Chapters

The thesis is organized as follows:

- Part I (Fundamentals and Related Work on Cross-lingual Information Retrieval): A detailed review of the state-of-the art cross-lingual retrieval approaches and their limitations is discussed. Furthermore, state-of-the art cross-lingual retrieval tools, which consider the user as integral part of the retrieval process is researched and a summary of their limitations and advantages are discussed.

  – Chapter 2 (Fundamentals): Chapter 2 gives an overview of different cross-lingual information retrieval approaches. Furthermore, gives an overview of different cross-lingual information retrieval research issues - with focus on the Arabic and German languages - that impedes the development of cross-lingual retrieval systems with good performance. These issues are explained in detail with helpful figures and examples. In addition, in this chapter, different state-of-the art approaches to overcome these issues are discussed.

  – Chapter 3 (Related Work on Interactive Cross-lingual Retrieval Tools): Chapter 3 describes state-of-the art cross-lingual retrieval tools. The chapter begins with a discussion of each cross-lingual tool, how the tool performs the retrieval task, what the task of the user is, how the translation and the disambiguation process is performed. The chapter concludes with a discussion of the limitations of the state-of-the art cross-lingual tools.

- Part II (Query Pre-and-Post Processing): A pre-post-processing approaches such as spelling correction, decompounding, word form variations detection etc., which has to be done before and after translation is reviewed and discussed.

  - Chapter 4 (Pre-processing: Spelling Correction): Chapter 4 describes the approaches developed to deal with spelling errors in the user query. This chapter describes in detail the MultiSpell approach which is a language-independent spell-checker that is based on an enhancement of the n-gram model. At the end of the chapter an evaluation is described in detail. The proposed MultiSpell approach has been compared with the state-of-the art approaches.

  - Chapter 5 (Post-processing: Word Inflection): Chapter 5 describes the approaches developed to deal with word inflection issue (Arabic). This chapter describes, in detail, a conflation approach, based on dealing with the special properties of the Arabic language in order to improve the retrieval performance. This chapter ends with a description of a language independent system (araSearch). araSearch supports a user with an extension of his/her query, by automatically including all word forms to the submitted query. As a result, the user does not need to be concerned with including all word forms of the submitted query. At the end of the chapter an evaluation is described in details. The proposed approaches in this chapter have been compared with the state-of-the art approaches.

- Part III (Query Translation and Disambiguation): The proposed approaches to tackle the ambiguity in the user query are discussed. Furthermore, detailed evaluations, to evaluate the performance of the proposed approaches, are presented.

  - Chapter 6 (Algorithms for Query Translation and Disambiguation): Chapter 6 begins with the description of the automatic translation approach followed by a general overview of how the disambiguation process is performed. The first disambiguation method is based on Naïve Bayesian Classifier (NB) and parallel corpora, where different classifiers from different subsets of features and combinations of them are built. The second method is based on Mutual Information (MI) and monolingual corpora where we present the data sparseness issue which is tackled through the enhancement of the Mutual Information approach.

  - Chapter 7 (Disambiguation Algorithms Evaluation): Chapter 7 presents an evaluation of the proposed disambiguation algorithms which contains: translation accuracy evaluation based on parallel corpora and Naïve Bayesian Classifier (NB) and translation accuracy based on monolingual corpora and the Mutual Information approach. For Naïve Bayesian Classifier approach, we used Arabic/ English parallel corpora as source of the statistical co-occurrences data. Based on the performed experiments, results could show that our algorithm achieved promising results when the inflectional form issue for Arabic words is considered. For the Mutual Information approach, we used monolingual corpora as source of the statistical co-occurrences data. Based on the experiments that we performed, using monolingual corpora and the web, results showed that our algorithm achieved promising results especially when using web as source of statistical data.

- Part IV (Interactive Meaning Refinement): Describes how all developed approaches are integrated to form the cross-lingual tool proposed in this thesis. Furthermore, describes how the user feedback can be used with the support of the tool to refine the translation and thus refine the cross-lingual process. A detailed user study and a disambiguation algorithm evaluation are presented and discussed.

  - Chapter 8 (Interactive Meaning Refinement): Chapter 8 describes how query pre-post-processing and (query translation and disambiguation) are integrated in the proposed interactive cross-lingual retrieval approach. The chapter begins with a short description about the initial work (first prototype) which we performed as initial step. Furthermore, the identified issues and shortcomings in the state-of-the art cross-lingual tools which we tackled in the proposed cross-lingual tool in this thesis are described and discussed in detail. In addition, we conducted a broad user study to consider more points of interest in evaluating the proposed approach and identify more issues in the first prototype which is tackled in the revised prototype. Furthermore, we described the different interface components and how they are integrated in order to perform the cross-lingual task (i.e., how we tackle the state-of-the art cross-lingual tools and the initial interface issues and shortcomings), from submitting the query till getting the relevant documents.

- Chapter 9 (Prototype Evaluation): In chapter 9, the second prototype has been used to evaluate the performance of the disambiguation algorithm for English/German language pair. Furthermore, we performed an evaluation to check whether the support provided by our cross-lingual tool is significant to guide the user in improving the translation and thus improve the performance of the cross-lingual retrieval system.

- Part V (Concluding Remarks and Future Work Perspectives): Describes concluding remarks about all parts in this thesis. Furthermore, future work perspectives is presented and discussed.

  - Chapter 10 (Concluding Remarks and Future Work Perspectives): Chapter 10 gives a brief summary about the thesis and future work perspectives. The approaches to tackle the problems of cross-lingual retrieval, which have been proposed in this thesis, are limited to web applications dealing particularly with vagueness in the user query. In this chapter a discussion about the limitations of the approaches proposed in this thesis in covering other different domains is presented. Furthermore, in this chapter, hints in how to deal with these issues are discussed and proposed.

- Part VI: (Appendix): Appendix contains the evaluation tables that describe the results achieved, in detail. Furthermore, a description of a preliminary Arabic WordNet is presented and discussed.

  - In Appendix A, the spelling correction evaluation tables show the detailed evaluation for the spelling correction task for the proposed approach Multi-Spell, comparing it to some state-of-the art approaches such as Aspell, TST, spell checker integrated into Microsoft Word and Google. In the Appendix B, the conflation approach evaluation tables show a detailed conflation task evaluation for the proposed conflation approach with respect to other state-of-the art conflation techniques e.g., pure n-grams, edit distance etc.

  - In Appendix C, disambiguation evaluation results based on Naïve Bayesian Classifier and Mutual Information approaches are presented.

– Appendix D describes the construction of a preliminary Arabic WordNet. A brief overview of the current development of the Arabic WordNet is presented followed by a brief overview of the Arabic morphological analyzers. An approach in supporting lexicographers in creating Arabic WordNet SynSets is presented. This creation is done query-oriented, where an Arabic word is searched and secondly annotated with English SynSets. Parallel corpora are then used to create glosses for every newly created Arabic SynSet. A user interface, including the functionalities described in our approach, is presented and discussed

Part I

Fundamentals and Related Work on Cross-lingual Information Retrieval

Chapter 2

Fundamentals

In this chapter, we give an overview of different cross-lingual information retrieval approaches. Furthermore, we give an overview of different cross-lingual information retrieval research issues - with focus on the Arabic and German languages - that impedes the development of Cross-lingual retrieval systems with good performance. These issues are explained in detail with helpful figures and examples. In addition, different state-of-the art approaches to overcome these issues are reviewed discussed.

2.1 Cross-lingual Retrieval Approaches

Cross-lingual information retrieval approaches can be classified into two main approaches, the knowledge-based approach and the corpora-based approach (Oard, 1997a) (see Figure 2.1). The knowledge-based approach, represents approaches that exploit explicit representations of translation knowledge, such as bilingual dictionaries (Dictionary-based), e.g., (Ballesteros and Croft, 1996; Oard and Diekema, 1998; Oard et al., 2008) or (Ontology-based) e.g., (Cheng et al., 2006). The corpora-based approach, on the other hand, represents systems that automatically extract useful translation knowledge from comparable or parallel corpora using statistical/probabilistic models, e.g., (Brown, 1998; Nie et al., 1999; Chan and Ng, 2007).

In the following, we describe these approaches in detail.

2.1.1 Knowledge-based Approach

The knowledge-based approach can provide very useful information, to improve the performance of word sense disambiguation applications and thus, improve cross-lingual retrieval performance. While in English, and some major European languages, the "lexical

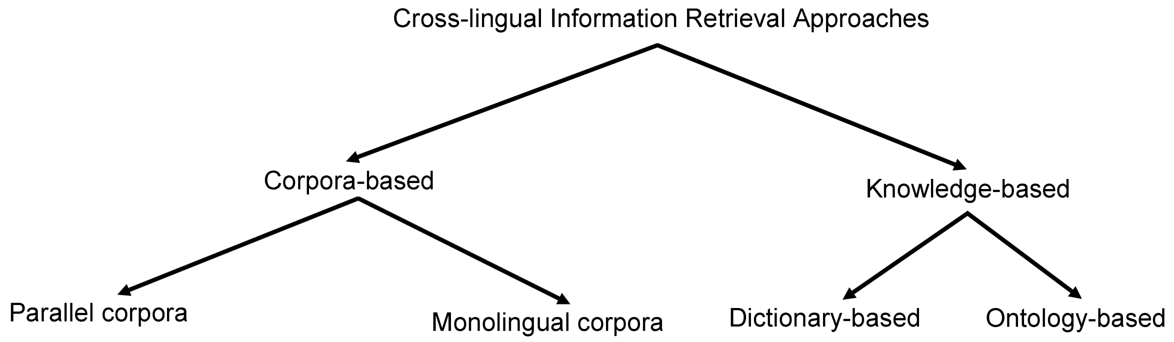


Figure 2.1: Cross-lingual retrieval approaches (Oard, 1997a).

bottleneck" problem likely softened, e.g., for English WordNet (Miller, 1995) and for (Dutch, Italian, Spanish, German, French, Czech and Estonian) EuroWordNet (Vossen, 1998), there are no available wide-range lexical resources for other languages such as Arabic. For European languages, for example, De Luca et al. (2006) proposed the MultiLexExplorer tool to support multilingual users in performing their web search. The MultiLexExplorer allows users to explore combinations of query term translations by visualizing EuroWordNet relationships together with search results and search statistics obtained from web search engines. Brown (1998) proposed an approach to construct a thesaurus based on translating the word in the original query then counting its co-occurrences information and storing it with the corresponding word in the target language.

In the following, we focus on studying the research issues that arise on using the dictionary-based approach.

For a dictionary-based approach, one can use a general-purpose dictionary or a special dictionary for a special task, e.g., a medical terminology dictionary for translation (Abusalah et al., 2005). The fundamental idea of using the dictionary-based approach is to search the dictionary, in order to extract a list of possible translations, in the target language, for each query term. However, the performance of the dictionary-based approach is very limited, due to many research issues e.g., translation ambiguity, out-of-vocabulary words (OOV), special properties for some languages hinder the correct match in the dictionary and the lack of context in the dictionary that should help to select the correct translation. In the following, we give an overview of the research issues involved with the dictionary-based approach.

Based on (Gearailt et al., 2005) four process stages for the dictionary-based query translation have been identified:

- Pre-Translation Query Modification: This involves that the source query is reformulated. (e.g., any addition, deletion or weighting of the query terms before translation).

- Dictionary Lookup: This involves the lookup mechanism for the alternative translations in the dictionary.
- Translation alternatives Selection and Term-Weighting: This involves the selection process of the best translation out of the translation alternatives for each query term. In addition, this stage also involves a Term-Weighting process, where the alternative translations can be weighted based on their co-occurrences.
- Post-Translation Query Modification: This involves the possibility of adding and deleting any translation alternatives carried out after all translation stages have been performed.

Finding correct translations for cross-lingual retrieval task in machine readable dictionary raises a number of issues (see Figure 2.2):

- It is possible that one word might have multiple translations (meanings) in the target language and thus it is very difficult to determine the correct meaning that should be chosen for the translation (see Section 2.2.3.1).
- The out of vocabulary words (OOV) issues. Dictionary does not contain all words, e.g., compound words, technical terms, proper names or spelling variants. For some language pairs, that use almost the same alphabets, this issue presents no great challenge. However, this issue is more complicated for language pairs that employ totally different alphabets and sound systems such as Arabic and English or Arabic and Japanese (see Section 2.2.1.3).
- For a high morphological inflectional language, such as Arabic, it is not possible that the dictionary can include all word forms, instead including just the root forms. Therefore, using the dictionary approach will necessitate a pre-processing step by using conflation approaches such as n-gram or stemming to identify the morphological root for the given query term (see Section 2.2.1.2).
- Lack of context in the dictionary, which is very essential to disambiguate the ambiguous query terms (see Section 2.2.3.1).

In the following, we outlined some dictionary-based approaches. As an example, we focused on Arabic cross-lingual retrieval (Ahmed, 2010).

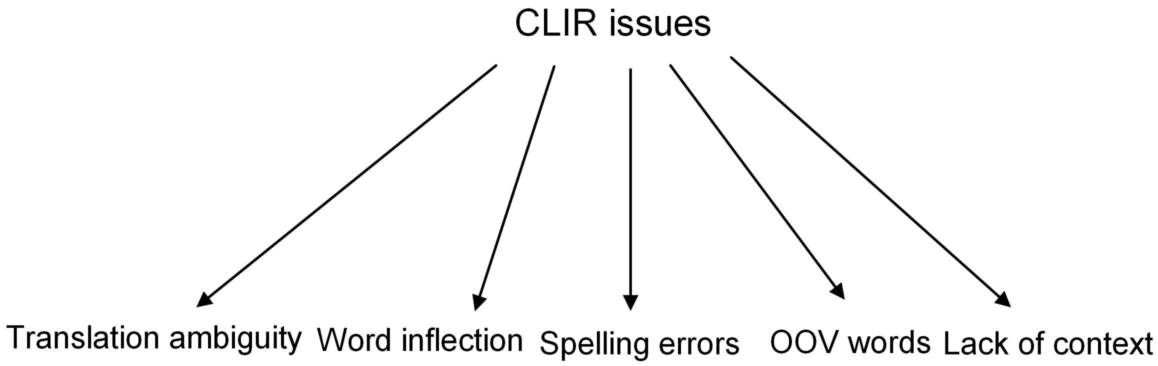


Figure 2.2: Overview of the main CLIR issues.

Specificities of Arabic

For Arabic cross-lingual retrieval several studies have been done so far. Aljlayl et al. (2002) evaluated the effectiveness of a machine translation-based Arabic-English cross-lingual retrieval by using the ALKAFI translation system and two standard TREC collections and topics. They pointed out that the experimental results indicate that the less source terms that are needed to form a context, the better the retrieval accuracy and efficiency is. Aljlayl and Frieder (2001) investigated the effectiveness of machine translation and MRD (Machine Readable Dictionary) approaches to Arabic-English cross-lingual retrieval. They studied three methods of query translation using an Arabic English bilingual dictionary: the Every-Match (EM), the First-Match (FM), and the Two-Phase (TP) methods. In the EM method they include all translations found in the dictionary for the query term. Using this method, the translation ambiguity will be higher and will result in poor effectiveness. In the FM method, they consider only the first translation provided by the bilingual dictionary. They claim that usually the translations provided by dictionaries are presented in an ordered way based on its common use and thus the more common translation is listed first. In the TP method, they select only the translation that returns the original query term when being re-translated. Based on their experimental results, they point out that the TP approach outperforms EM and FM approaches. Although translation in cross-lingual retrieval and machine translation seems to have the same concerns, it should be noted that machine translation and cross-lingual retrieval tackle quite different problems: Machine translation focuses more on providing sentences with correct syntactic information, while cross-lingual retrieval focuses more on providing translations without considering any syntactic information. Furthermore, cross-lingual retrieval systems, in some cases, allow for more than one translation for each of the query terms (translation relevant), while machine translation focuses on providing a unique translation for each query term, and in cross-lingual retrieval users are often involved in the translation refinement process, while in machine translation the user plays no role in the translation process.

Levow et al. (2005) pointed out that limitations of the dictionary-based approach can be softened by the corpora-based approach (hybrid approach) e.g., using monolingual corpora to overcome translation ambiguity as a result of using the dictionary-based approach. They mentioned that other ideas from the dictionary-based cross-lingual retrieval might find productive applications with corpus-based or interactive techniques; for example, using corpora and user feedback to enhance the translation dictionary. Using a statistical/ probabilistic model, based on corpora, a dictionary translation can be automatically improved because related cross-lingual word-pairs appear in similar context, in such a collection.

In this thesis, we proposed a hybrid cross-lingual approach that combines the dictionary-based approach and the corpora-based approach. We used the dictionary-based approach to extract all possible translations for the given user query and the corpora-based approach was used to tackle the translation ambiguity issue. To improve the proposed approach, a user feedback was used to refine the translation.

2.1.2 Corpora-based Approach

In parallel corpora the same text is written in different languages. A statistical approach to find statistical associations between words in two languages, using parallel corpora, has been studied, e.g., in (Yang et al., 1998). Resolving translation ambiguity, based on text corpora of source and target languages, was studied and evaluated, e.g., Spanish and English (Cabezas and Resnik, 2005). Statistical techniques applied in this corpora can be used to produce bilingual term equivalence by comparing which words co-occur in the sentence over the whole corpora. Corpora-based approaches, uses translations extracted from bilingual corpora to perform the query translation (Yang et al., 1998). Corpora based approaches provide an alternative solution for overcoming the lexical acquisition bottleneck by gathering information directly from textual data e.g., bilingual corpora. Due to the expense of manual acquisition of lexical and disambiguation information, where all necessary information for disambiguation has to be manually provided, supervised approaches suffer from major limitations in their reliance on predefined knowledge source, which affects their ability to handle large vocabulary in a wide variety of contexts. Resolving translation ambiguity, based on text corpora of source and target languages, was studied and evaluated, e.g., for English and Japanese (Doi and Muraki, 1992), French and English (Vickrey et al., 2005), Spanish and English (Cabezas and Resnik, 2005), Arabic and English (Ahmed and Nürnberger, 2008a,b), Portuguese and English (Specia et al., 2007) and Chinese and English (Chan and Ng, 2007). For Arabic cross-lingual retrieval using corpora approach, we presented in (Ahmed and Nürnberger, 2008a,b) a word sense disambiguation method applied in automatic translation of a query from Arabic to English. The developed machine learning approach is based on statistical models that can learn from parallel corpora by analysing the relations between the items included in these corpora in order to use them for selecting the most suitable translation of the query term.

In order to resolve the translation ambiguity inherent in bilingual dictionaries, the hybrid approach can be used. The hybrid approach uses bilingual dictionaries to extract the possible translation for each query term and uses corpora to find the cohesion score between all possible translation candidates. Unlike the corpora-based translation approach, which relies only on the use of bilingual corpora to translate the user query, a target language corpora (monolingual corpora) can be used to resolve the translation ambiguity inherent in bilingual dictionaries (see (Ballesteros and Croft, 1998; Chen et al., 1999; Ahmed et al., 2009a)). The core idea of using the target language corpora for disambiguation is to retrieve the translation candidates for each query term from bilingual dictionaries, then construct the translation combination between those candidates. The approach selects the translation combination that frequently co-occurs in the target language corpora. Parallel corpora can be used alone for cross-lingual retrieval but it is also applicable to the hybrid approach (Davis and Ogden, 1997; Ahmed and Nürnberger, 2008d). The idea behind this is that initially, possible translation candidates, using a dictionary, will be derived. Thereafter, source and translated query are used to retrieve the source and target documents from the parallel corpora, respectively. Finally, only translation that retrieves documents aligned to the documents retrieved by the source query, is selected.

In the following, we discuss some of the problems that cross-lingual retrieval approaches are currently facing in more detail. Problems found with cross-lingual retrieval approaches, hindering better performance, are translation ambiguity, word inflection, translating word compounds, phrases, proper names, spelling errors, spelling variants and special terms (Hedlund et al., 2004) (Ahmed, 2010). In the following, we discuss the most important issues - with giving special attention to Arabic and German languages - that impeded the development of cross-lingual retrieval systems with a good performance.

2.2 Cross-lingual Retrieval Issues

2.2.1 Pre-Processing Task

In this section, we describe the pre-processing step which has to be done before the cross-lingual retrieval system can perform its task. There is an urgent need to correct the user's misspelled query terms. Misspelled query terms in the user query results in poor cross-lingual retrieval. Furthermore, the user query needs to be pre-processed. This pre-processing step is useful to transform a word to its basic form. The stemming of the user query terms is very important because the dictionary does not include all word forms, instead just the root form. The use of stemming leads to a clear benefit with respect to the cross-lingual retrieval task. The user does not need to pay any attention to word form inflection issues, as different forms of his/her query terms are automatically conflated into the basic form. Furthermore, stemming provides many other benefits, such as improved retrieval performance and storage saving. As an example, we focused on Arabic which is a highly morphological language. For stemming Arabic words, we used the araMorph package based on the Buckwalter Arabic morphological analyzer (Buckwalter, 2002).

In the following, we start with describing different spelling correction approaches. Next, for Arabic, we describe the word inflection issue, followed by a detailed description of approaches which are used to solve, or at least to alleviate, some of the problems raised by a high inflectional morphology. For the German language, we describe in detail the problem of compound words and how it affects the performance of cross-lingual retrieval. Different approaches for decompounding are reviewed and discussed.

2.2.1.1 Spelling Correction Issue

The problem of devising algorithms and techniques for automatically correcting words is very essential for improving the retrieval performance. Research in this field began as early as the 1960s on computer techniques for automatic spelling correction and automatic text recognition, and it has continued up to the present, there are good reasons for the continuing research efforts in this area in order to improve quality and performance and to broaden the spectrum of possible applications (Kukich, 1992). For example, even though systems programs (language processors, operating systems, etc.,) have become increasingly powerful and sophisticated, they do not assist the user - with a very few exceptions - in correcting many of the obvious spelling errors in the source input. There are two types of word errors, the *real-word error* and the *non-word error*. Real-word errors are misspelled words that have a meaning and can be found in a dictionary. Non-word errors are words that have no meaning and are thus not included in a dictionary. We concentrate on the correction of the non-word error with the proposed algorithm. Damerau (1964) found that 80% of misspelled words that are non-word errors are the result of a single insertion, deletion, substitution or transposition of letters. Therefore, it seems reasonable to base correction algorithms on measures that consider these simple operations. However, also approaches based on pure n-gram statistics - which account for these operations only implicitly - have proven to provide good performance (Kukich, 1992; Hodge and Austin, 2003). Algorithmic techniques for detecting and correcting spelling errors in text have a long and robust history in computer science (Kukich, 1992). Many approaches have been applied since people started to deal with this problem. Different techniques like edit distance (Wagner and Fischer, 1974), rule-based techniques (Yannakoudakis and Fawthrop, 1983), n-grams (ming Zhan et al., 1998), probabilistic techniques (K.W. and W.A., 1991), neural nets (Hodge and Austin, 2003), similarity key techniques (Pollock and Zamora, 1983, 1984) and noisy channel model (Brill and Moore, 2000; Toutanova and Moore, 2002) have been proposed. All of these are based on the idea to calculate the similarity between the misspelled word and the words contained in a dictionary. In the following, we describe shortly one of the most popular approaches (Aspell) and one recently proposed approach for the Portuguese language (TST) (Martins and Silva, 2004) that we used for comparison. GNU Aspell, usually called just Aspell, is a standard spelling checker software for the GNU software system. There are Dictionaries for about 70 languages available. GNU Aspell is a Free and Open Source software[1]. In contrast to Ispell, which suggests words with small edit-distance, Aspell in addition compares soundslike equivalents (computed for English words using the metaphone algorithm (Deorowicz and Ciura, 2005)) up to a given edit distance. The Ternary Search Trees (Martins and Silva, 2004) approach (TST) is a dictionary data structure working with string-keys. It can find, remove and add these keys quickly and also easily search the tree for partial matches. Additionally near-match functions can be implemented. These give the possibility to suggest alternatives for misspelled words.

2.2.1.2 Word Inflection Issue

In word inflection items are added to the base form of a word to express grammatical meanings such as tense, mood, voice, aspect, person, number, gender and case (Alvarez et al., 2011). Word inflection causes a real problem for translations as well as for cross-

[1] http://aspell.sourceforge.net/

lingual retrieval systems whereas languages exhibiting a rich inflectional morphology face a challenge for machine translation systems. In the following we give a brief description of the Arabic language, clarifying some of its properties followed by a brief discussion of approaches that try to overcome the word inflection issues with respect to the Arabic language.

Specificities of Arabic

Arabic is a Semitic language that is based on the Arabic alphabet containing 28 letters. Its basic feature is that most of its words are built up from, and can be analyzed down to common roots. The exceptions to this rule are common nouns and particles. Arabic is a highly inflectional language with 85% of words derived from triliteral roots. Nouns and verbs are derived from a closed set of around 10,000 roots (Al-Fedaghi and Al-Anzi, 1989). Arabic has three genders, feminine, masculine, and neuter; and three numbers, singular, dual, and plural. The specific characteristics of Arabic morphology make the Arabic language particularly difficult for developing natural language processing methods for information retrieval. One of the main problems in retrieving Arabic language text is the variation in word forms. For example, the Arabic word كاتبkātb (author) is built up from the root كتبktb (write). Conjunctions and prepositions are also attached as prefixes to nouns and verbs, hindering the retrieval of morphological variants of words (Moukdad, 2004). In Table 2.1 some word form variations for the word "student" is presented in order to clarify this issue. Arabic is different from English and other Indo-European languages with respect to a number of important aspects: words are written from right to left; it is mainly a consonantal language in its written forms, i.e., it excludes vowels; its two main parts of speech are the verb and the noun in that word order, and these consist, for the main part, of triliteral roots (three consonants forming the basis of noun forms that are derived from them); it is a morphologically complex language, in that it provides flexibility in word formation: as briefly mentioned above, complex rules govern the creation of morphological variations, making it possible to form hundreds of words from one root (Moukdad and Large, 2001). Furthermore, the letter shapes are changeable in form, depending on the location of the letter at the beginning, middle or at the end of the word.

Stemmer Approaches

In information retrieval systems stemming is used to reduce variant word forms to common roots and thereby improve the ability of the system to match query and document

Feminine	Masculine	English
طالبةṭālbh	طالبṭālb	student
الطالبةālṭālbh	الطالبālṭālb	the student
طالبتانṭālbtān	طالبانṭālbān	(two) students(dual)
بطالبةbṭālbh	بطالبbṭālb	by student
بالطالبةbālṭālbh	بالطالبbālṭālb	by the student
وطالبةwṭālbh	وطالبwṭālb	and student
والطالبةwālṭālbh	والطالبwālṭālb	and the student
للطالبةlṭālbh	للطالبlṭālb	to the, for a student
طالبتهṭālbtha	طالبهṭālbh	his student
طالبتهاṭālbthā’	طالبهاṭālbhā	her student
طالباتهṭālbāth	طلبتهṭlbth	his students
طالباتهاṭālbāthā’	طلبتهاṭlbthā	her students
...	...	...

Table 2.1: Word form variations for طالبṭālb (Student).

vocabulary (Xu and Croft, 1998). Although stemming has been studied mainly for English, stemming approaches have also been developed for several other languages such as Malay (Tai et al., 2000), Latin (Greengrass et al., 1996), Indonesian (Berlian et al., 2001), Swedish (Carlberger et al., 2001), Dutch (Kraaij and Pohlmann, 1996), German (Monz and de Rijke, 2002), French (Moulinier et al., 2001), Slovene (Popovic and Willett, 1992), Turkish (Ekmekcioglu et al., 1996) and Arabic (Khoja and Garside, 1999; Larkey et al., 2007). There are three main types of approaches for stemming, dictionary-based, rule-based, and statistical-based (mainly n-gram based) approaches (Gelbukh et al., 2004). *Dictionary based approaches* provide very good results at the cost of high development efforts for the dictionary. The dictionary contains all known words with their inflection forms. The main weakness for this approach is the missing words in the dictionary which would not be recognized by the system for stemming. Another weakness is the inability of this method to stem inert names and foreign words. Also the need to process a large dictionary during runtime can result in high requirements for storage space and processing time. The closest Arabic equivalent for this kind of stemmer is the *root-based stemmer* for Arabic (Khoja and Garside, 1999) which is based on extracting the root of a given Arabic surface word by striping off all attached prefix and/or suffix then attempt to extract the root of it. Several morphological analyzers were developed based on this concept (Khoja and Garside, 1999; Buckwalter, 2002). The weaknesses of this stemmer are: it does nothing when it comes across some words which have no root. Furthermore, the construction of the corresponding dictionaries or rules is a tedious and labor consuming task due to the result of the morphology complexity of Arabic language. Another problem is that only some small linguistic resources are available for Arabic language. The weaknesses of this stemmer is that the construction of the corresponding dictionaries or rules is a tedious and labor consuming task due to the result of the morphology complexity of Arabic language. Another problem is that only some small linguistic resources are available for Arabic language. The second type are the *rule-based approaches*. They are based on set of predefined conditions rules. The most well known stemmer of this type is Porter stemmer (Porter, 1980). The main weakness for this stemmer is that building the rules for the arbitrary language is time consuming. Furthermore, there is a need for experts with linguistic knowledge in that particular language. The Arabic equivalent for this is the *Light stemmer* (Larkey et al., 2007). Unlike English, both prefixes and suffixes need to be removed for effective stemming. It is based on striping off prefix and suffix from the word, it use predefined list of prefix and suffix, it is simply striping off prefix and/or suffix without any further processing in the rest of the stemmed word (Roeck and Al-Fares, 2000; Larkey et al., 2007). The weakness of this stemmer is that the striping off prefixes or suffix in Arabic is a not an easy task. Removing them can lead to unexpected results, as many words start with one letter or more which can mistakenly assumed to be prefix or suffix.

2.2.1.3 Out of Vocabulary Words (OOV)

In cross-lingual retrieval systems the translation of out of vocabulary words that are not part of a standard dictionary such as (compound words, technical terms, named entities and acronyms) is a very important point for an effective cross-lingual retrieval system (Pirkola et al., 2003). For some language pairs, that use almost the same alphabets, this issue presents no great challenge. However, this issue is more complicated for language pairs that employ totally different alphabets and sound systems such as Arabic and English or Arabic and Japanese. Bilingual dictionaries usually avoid including OOV words like named entities, numbers, technical terms and acronyms. Davis and Ogden (1998) and Al-Fedaghi and Al-Anzi (1989) find around 50% of OOV words to be named entities. If no translation exists for these words, they have to be "converted". The process of converting a word from one orthography into another is called transliteration. Unfortunately, people usually follow no standard transliteration rules when converting foreign words into Arabic. For example, Table 2.2 shows 15 different spellings for the name Condoleezza; four of them were found in the same news web site ("CNN-Arabic")

[2].

S/N	Transliteration	Occurrence in web	Comments
1	كونداليزاkwndālyzā	3.000.000	CNN
2	كوندوليزاkwndwlyzā	197.000	CNN
3	كوندليزاkwndlyzā	51.100	CNN
4	كونداليساkwndālysā	26.300	
5	كوندوليساkwndwlysā	26.200	CNN
6	كاندوليزاkāndwlyzā	12.700	
7	كنداليزاkndālyzā	2.310	
8	كانداليزاkāndālyzā	1.530	
9	كونداليزةkwndālyzh	491	
10	كندليساkndlysā	344	
11	كونداليزهkwndālyzh	195	
12	كنداليساkndālysā	144	
13	كانداليساkāndālysā	9	
14	كونداليسةkwndālysh	9	
15	كوندليسيkwndlysy	4	

Table 2.2: Multiples spellings for the name "Condoleezza" (Ahmed and Nürnberger, 2011).

Arbabi et al. (1994) developed an algorithm at IBM using automatic transliteration of Arabic personal names into the Roman alphabet. Their approach was based on using a hybrid neural network and knowledge-based system approach. In (Stalls and Knight, 1998) an algorithm based on probabilistic models for Translating Names and Technical Terms from Arabic to English translation is proposed. This work was based on (Knight and Graehl, 1997) that describe a back transliteration system for Japanese. Al-onaizan and Knight (2002a) presented a transliteration algorithm based on sound and spelling mappings using nite state machines. Larkey et al. (2003) conducted experiments for Arabic/English cross-lingual retrieval using TREC2001 and TREC2002 to evaluate the effectiveness of the translation of proper names in information retrieval using different sources of name translation for Arabic. N-gram based approaches were widely proposed to deal with this issue. Aqeel et al. (2006) addressed the name search for Arabic transliterated names using n-gram and soundex techniques to improve precision and re-

[2] http://arabic.cnn.com/, Retrieved on 01/03/2010, www.Google.com

call of name matching against well-known techniques. Furthermore they investigated the performance of n-grams of varying length. They used in their test approximately 7,939 Arabic first names translated to English. From their experiments they pointed out that using the n-gram techniques improves precision and recall of Arabic name matching search. de Gispert and Mariño (2006) studied the performance of n-gram-based statistical machine translation (including OOV words) in two independent tasks: English-Spanish European Parliament Proceedings large-vocabulary task and Arabic-English Basic Travel Expressions small-data task. They pointed out that the result obtained outperform all previous techniques. Using bilingual and monolingual resources were also used to deal with this issue. Al-onaizan and Knight (2002b) presented a Name Entity translation algorithm for translating Arabic name entities to English without using any dictionary. They compared their results with results obtained from human translators and commercial systems. They claim that the translations obtained by their algorithm showed significant improvement over the commercial system and in some cases it outperforms the human translator. In the context of Name Entity (NE) recognition, Samy et al. (2005) used parallel corpora of 1200 sentence pairs in Spanish and Arabic with a Name Entity tagger for Spanish. For their experiments, they randomly selected 300 sentences from the Spanish corpus with their equivalent Arabic sentences. For each sentence pair the output of the NE tagger was compared to the manually annotated gold standard set. They reported that using their approach they gained higher recall and precision than state-of-the-art approaches. Although new words and word combinations can be generated readily in natural languages, the dictionaries' lack of full coverage leads to low cross-lingual retrieval performance. An example is the compound word problems which are a real issue for some languages such as German, in respect to information retrieval or cross-lingual retrieval. In the following, we focused on describing this problem with respect to the German language.

Specificities of German - Compound Word Issue

A compound word is a word that is a result of joining two or more words together. Compound words can result in having out-of-vocabulary (OOV) problems in cross-lingual information retrieval. In order to improve cross-lingual information retrieval effectiveness, these compound words need decompounding before translation. Compounds appear more frequently in some languages, such as German, while they appear less in other languages, such as English. It is possible to find two-word compounds in English such as "airmail", "airplane", "birthplace", "backbone", "cowboy", "football", "hammerhead" etc. However, it is very rare to find English compounds for three or more words. In other languages, such as German, the matter is different where compounds of two or more words are not uncommon. As an example, we consider these German compounds ("kinderwagen" "stroller"), ("liebesgedicht" "love poem"), ("Straßenreinigungsgebühr", "street cleaning fee"), ("Einkommensteuer", "income tax"), ("Suchmaschinentechnologien", "search engine technology"), ("Geschwindigkeitsüberschreitung", "exceeding the speed limit"), ("Geschwindigkeitsanzeigetafel", "Speed display board"), ("Lehrgangsteilnahmebestätigung", "training course participation confirmation"), ("Donaudampfschifffahrtsgesellschaftskapitän", "Danube steamship company captain"), ("Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz", "beef labelling regulation & delegation of supervision law") etc. In order to improve the performance of cross-lingual systems, these compounds need to be decompounded before translations. Decompounding is the process of splitting compounds into their constituent parts. For high methodological languages such as German, Dutch or Finish, decompounding has been found to improve the effectiveness of information retrieval because it can tackle the vocabulary mismatch problems (Chen and Gey, 2004). Due to the productive nature of languages, quite often many words can be combined into new compounds. When the search for a query in languages which have a high frequency of compounds, such as German or Dutch, cross-lingual retrieval performance is lower than for other language pairs (Piroi, 2010). This issue is due to the presence of compound words in the query or in the collection of documents, which will result in a higher rate of OOV compound terms. These OOVs, in most cases, can't be translated and will result in poor cross-lingual retrieval performance. Therefore, for such languages, the search or translation for cross-lingual information retrieval shouldn't only be performed based on full compounds but also in their component words.

In the following, we describe two algorithms for German compound splitting that represent two different approaches, a dictionary-based approach and a rule-based approach.

Dictionary-based Approach

Chen and Gey (2004) used a dictionary-based decompounding on the CLEF 2001 and 2002 test collection. The dictionary-based decomposition of a word checks whether prefix strings of a compound are valid words. This is done by searching for them in a dictionary. Most decompounding approaches for German information retrieval consider the most frequent rules for word formation. An example would be using the so called letter "s" connection that appears between component words and represents one of the most frequent patterns in German compound word formation. For example, the word ("Inhaltsverzeichnis", "table of contents") consists of two constituents, ("Inhalt" and "verzeichnis") that are connected by the connector "s".

The algorithm works as follows:

- A German dictionary which contains non-compound words, in various forms, is built.

- A compound German word is decompounded based on the created dictionary in the first step. For example, the German based dictionary contains ("ball", "fuss", "fussball", "meisterschaft") and others, the German compound word ("fussballmeisterschaft" "European Football Cup") is decompounded into several compound words based on the German based dictionary. So, based on this step, we have these two compounds "fuss ball europa meisterschaft" and "fussball europa meisterschaft".

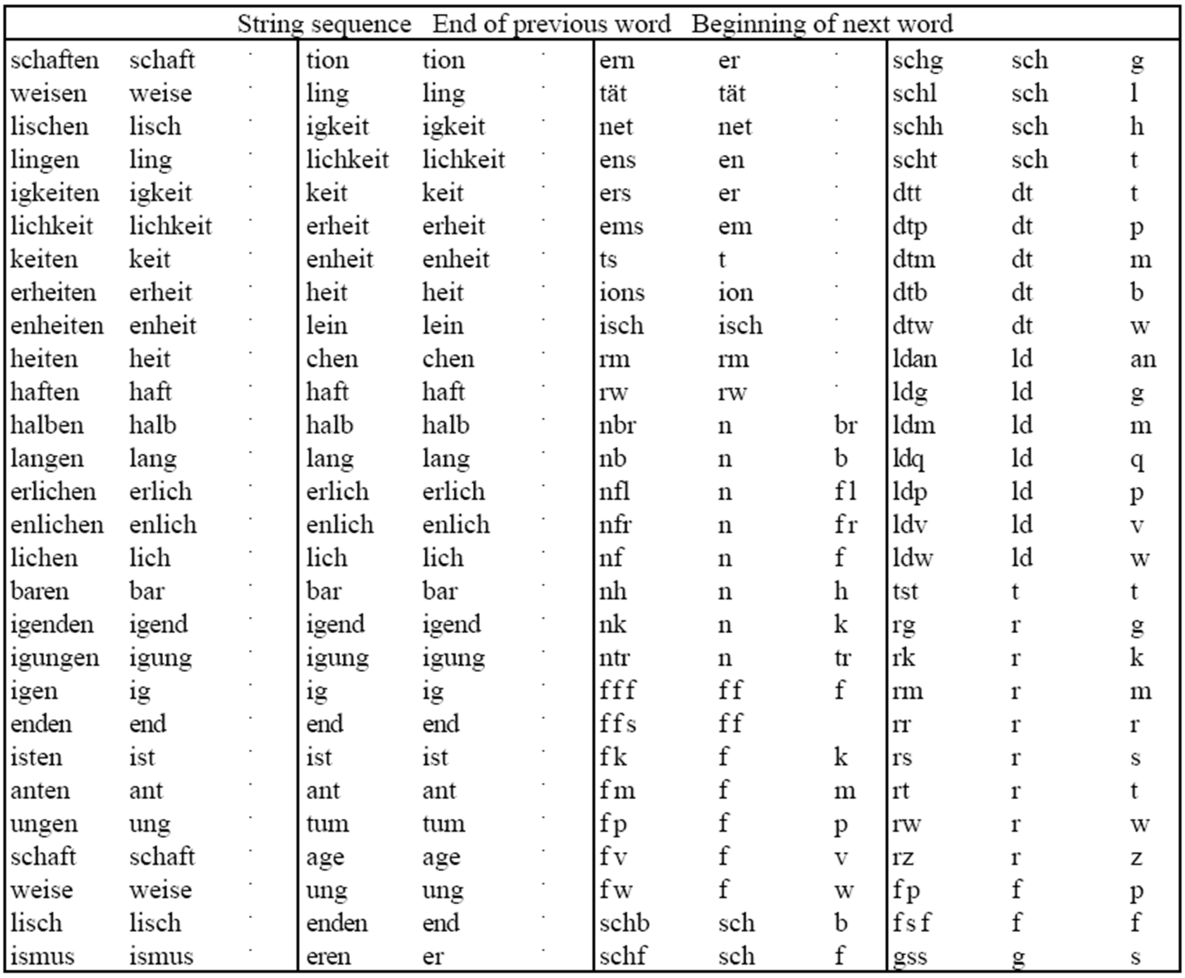
- The decomposition with the smallest number of component words is chosen. In the previous example, the decomposition "fussball europa meisterschaft" will be selected as the decompounding for the German compound "fussballmeisterschaft".

If there is more than one decomposition share with the same number of component words, the one with the highest probability of decomposition will be chosen. The probability is estimated by the product of the relative frequencies of the component words in the training collection.

Rule-based Approach

Savoy (2002) proposed a German decompounding approach based on a set of pre-defined patterns. The approach is based on breaking any words having an initial length greater than or equal to eight characters, taking into account that decomposition will not take place before any initial sequence (word might begin with a serious of vowels that must be followed by at least one consonant). In order to perform the decompounding process, a set of decompounding patterns for German is defined.

For clarification, we take the following example, the German compound ("Betreuungsstelle", "care center"). This word is more than eight characters long. In order to start splitting the compound, the algorithm seeks occurrences of one of the patterns. For this example, the patterns ("String sequence: "gss", End of previous word: "g", Begining of next word: "s") refer that when we find the character string "gss" the algorithm can cut the compound term, so the first word ends with "g" and the second word begins with "s" (see Figure 2.3). This will lead to the forming of the words "Betreuung" (care) and "Stelle" (center, place) out of the compound word ("Betreuungsstelle", "care center"). Given that the term "Stelle" is less than eight characters long, the algorithm will not attempt to decompound this term.



String sequence	End of previous word	Beginning of next word
schaften	schaft	·
weisen	weise	·
lischen	lisch	·
lingen	ling	·
igkeiten	igkeit	·
lichkeit	lichkeit	·
keiten	keit	·
erheiten	erheit	·
enheiten	enheit	·
heiten	heit	·
haften	haft	·
halben	halb	·
langen	lang	·
erlichen	erlich	·
enlichen	enlich	·
lichen	lich	·
baren	bar	·
igenden	igend	·
igungen	igung	·
igen	ig	·
enden	end	·
isten	ist	·
anten	ant	·
ungen	ung	·
schaft	schaft	·
weise	weise	·
lisch	lisch	·
ismus	ismus	·
tion	tion	·
ling	ling	·
igkeit	igkeit	·
lichkeit	lichkeit	·
keit	keit	·
erheit	erheit	·
enheit	enheit	·
heit	heit	·
lein	lein	·
chen	chen	·
haft	haft	·
halb	halb	·
lang	lang	·
erlich	erlich	·
enlich	enlich	·
lich	lich	·
bar	bar	·
igend	igend	·
igung	igung	·
ig	ig	·
end	end	·
ist	ist	·
ant	ant	·
tum	tum	·
age	age	·
ung	ung	·
enden	end	·
eren	er	·
ern	er	·
tät	tät	·
net	net	·
ens	en	·
ers	er	·
ems	em	·
ts	t	·
ions	ion	·
isch	isch	·
rm	rm	·
rw	rw	·
nbr	n	br
nb	n	b
nfl	n	fl
nfr	n	fr
nf	n	f
nh	n	h
nk	n	k
ntr	n	tr
fff	ff	f
ffs	ff	
fk	f	k
fm	f	m
fp	f	p
fv	f	v
fw	f	w
schb	sch	b
schf	sch	f
schg	sch	g
schl	sch	l
schh	sch	h
scht	sch	t
dtt	dt	t
dtp	dt	p
dtm	dt	m
dtb	dt	b
dtw	dt	w
ldan	ld	an
ldg	ld	g
ldm	ld	m
ldq	ld	q
ldp	ld	p
ldv	ld	v
ldw	ld	w
tst	t	t
rg	r	g
rk	r	k
rm	r	m
rr	r	r
rs	r	s
rt	r	t
rw	r	w
rz	r	z
fp	f	p
fsf	f	f
gss	g	s

Figure 2.3: Decompounding patterns for German (Savoy, 2002).

In this thesis, the goal is not to evaluate any approaches for decompounding instead to implement one of the reported successful approaches particularly "dictionary-based decompounding" proposed by (Chen and Gey, 2004). The implemented "dictionary-based decompounding" is used to improve the performance of the cross-lingual tool proposed in this thesis when no translation for a compound word is found in the dictionary.

2.2.2 Post-Processing Task

2.2.2.1 Word Form Variations Issue (Arabic)

The characteristics of highly inflectional languages result very often in a poor information retrieval performance. As a result, current search engines suffer from serious performance with the direct query-term-to-text-word matching for these languages, thus search engines need to be able to distinguish different variants of the same word. Detecting all word form variations in the query, which, processed by search engines, is considered essential for achieving good retrieval results and the alternative is the loss of vast amounts of information. In the following, we focus on one of the conflation approach which is n-gram.

n-gram Approaches

The main idea of n-gram based approaches, which groups together words that contain identical character substrings of length n called n-grams (Adamson and Boreham, 1974), is that the character structure of the word can be used to find semantically similar words and word variants. N-gram approaches differ from stemmers in terms of not requiring language knowledge, predefined rules or a vocabulary database. Furthermore; n-gram approaches take into account the misspelled and the transliterated words[3].

Computing Similarity Scores based on n-grams

The idea of using n-grams in language processing was discussed first by Shannon (1951). After this initial work the idea of using n-grams has been applied to many problems such as word prediction, spelling correction, speech recognition, translated word correction and string searching. One main advantage of the n-gram method is that it is language independent. In a spelling correction task an n-gram is a sequence of n letters in a word or a string. The n-gram model can be used to compute the similarity between two strings by counting the number of similar n-grams they share. The more similar n-grams between two strings exist, the more similar they are. Based on this idea the similarity coefficient can be derived. The similarity coefficient δ is defined by the following equation:

$$\delta = \frac{|\alpha \bigcap \beta|}{|\alpha \bigcup \beta|} \tag{2.1}$$

[3] Transliteration is the process of converting one orthography from one language into another.

where α and β are the n-gram sets for two words a and b to be compared. $\alpha \bigcap \beta$ denotes the number of similar n-grams in α and β, and $\alpha \bigcup \beta$ denotes the number of unique n-grams in the union of α and β.

Specificities of Arabic

Over the last years there were several studies have been performed to explore the use of n-grams for processing Arabic text. Mayfield et al. (2002) have found that n-grams work well in many languages; furthermore, they investigated the use of character n-grams for Arabic retrieval in TREC-2001 and found that n-grams of length 4 were most effective. Darwish and Oard examined multiple tokenization strategies for retrieval of scanned Arabic documents, they found out that n-grams of size $n = 3$ or $n = 4$ are well suited to Arabic document retrieval (Darwish and Oard, 2002). Mustafa (2004) assessed the overall performance of two n-gram techniques that he called conventional and hybrid. In his results Mustafa pointed out that the hybrid approach outperforms the conventional approach. Classifying Arabic text using n-gram frequencies also have been fruitful (Khreisat, 2006). Abu-Salem (2004) found out that all of the proposed n-gram methods outperform the word, stem, and root index methods. Ghaoui et al. (2005) investigated a new morphological class based language model. They used the morphological rules to derive the different words in a class from their stem. Furthermore a linear interpolation between the n-gram model and the morphological model has been evaluated. In their experiments they pointed out that the morphological class-based model yields lower performance compared to a classical trigram. However, all of the previous studies rely on studying the use of n-gram on the Arabic text based on the following aspects: The effectiveness of n-gram size and assessing the performance of existing n-gram approaches. None of the prior studies attempt to modify the pure n-gram model so that it reduce the ambiguity i.e., obtains a higher precision and recall.

Due to the mentioned insufficiencies of the existing approaches, we proposed in (Ahmed and Nürnberger, 2009) a "revised" n-gram algorithm that makes it possible to handle one-character infixes, prefixes, and suffixes, which are frequent in Arabic. The proposed method obtained superior results on a large newspaper corpus. For detailed investigation about the use of n-gram on Arabic text, we refer the reader to (Meftouh et al., 2010).

2.2.3 Automatic Query Translation Task

2.2.3.1 Translation Disambiguation Issue

In natural language there are many words that have multiple meanings and therefore the meaning of such equivocal or ambiguous words may vary significantly according to the context in which they occur. This problem is even more complicated when those words are translated from one language into another due to the ambiguities in both languages. Therefore, there is a need to disambiguate the ambiguous words that occur during the translations. Word Translations Disambiguation WTD, or more general Word Sense Disambiguation (WSD) is the process of determining the correct sense of an ambiguous word given the context in which the ambiguous word occurs. We can define the WSD problem as the association of an occurrence of an ambiguous word with one of its proper senses.

Specificities of Arabic

Arabic poses a real translation challenge for many reasons; Arabic sentences are usually long and punctuation has no or little affect on interpretation of the text. Contextual analysis is important in Arabic in order to understand the exact meaning of some words. Characters are sometimes stretched for justified text, i.e., a word will be spread over a bigger space than usual, which prevent a (character based) exact match for the same word. Furthermore, in Arabic synonyms are very common, for example, "year" has three synonyms in Arabic عامʿām , حولḥwl , سنةsnh that are all widely used in everyday communication.

Another real issue for the Arabic language is the absence of diacritics (sometimes called voweling). Diacritics can be defined as symbols over and under letters, which are used to indicate the proper pronunciations, hence also define the meaning of a word and therefore have important disambiguating properties. The absence of diacritics in Arabic texts poses a real challenge for Arabic natural language processing as well as for translation, leading to high ambiguity. Though the use of diacritics is extremely important for readability and understanding, diacritics is very rarely used in real life situations. Diacritics don't appear in most printed media in Arabic regions nor on Arabic internet web sites. They are visible in religious texts such as the Quran, which is fully diacritized in order to prevent misinterpretation. Furthermore, the diacritics are present in children's books in school for learning purposes. For native speakers, the absence of diacritics is not an issue. They can easily understand the exact meaning of the word from the context, but for inexperienced learners as well as in computer usage, the absence of the diacritics is a real issue. When the texts are unvocalized, it is possible that several words have the same form but different meaning. For example, the Arabic word يعد yʿd can have the meanings "promise", "prepare", "count", "return", "bring back" in English or the Arabic word علم ʿlm can have the meanings "flag", "science", "he knew", "it was known", "he taught", "he was taught" (see (Ahmed and Nürnberger, 2008b, 2009)).

The task of disambiguation therefore involves two processes: First, identifying all senses of the ambiguous word considered. Second, assigning the appropriate sense each time this word occurs. The first step can be tackled, e.g., using a list of senses for each of the ambiguous words existing in everyday dictionaries. The second step can be done by analysing the context in which the ambiguous word occurs, or by using an external knowledge source, such as lexical resources as well as a hand-devised source, which provides data (e.g., grammar rules) useful to assigning the appropriate sense to the ambiguous word. In WSD it is very important to consider the source of the disambiguation information (e.g., a hand-devised source may provide a better quality than a source derived by statistical processing - see Appendix D where we proposed an automatic method in supporting lexicographers in creating Arabic WordNet SynSets), the way of constructing the rules using this information and the criteria of selecting the proper sense for the ambiguous word, using these rules. WSD is considered an important research problem and is assumed to be helpful for many applications such as machine translation (MT) and information retrieval. Approaches for WSD can be classified into two main categories: knowledge-based approaches and corpora based approaches In the following, we describe in detail the state-of-the-art for these two categories.

Knowledge-based Approaches

The Knowledge-based approach for WSD exploits lexical knowledge stored in machine-readable dictionaries e.g., LDOCE (Longman English Dictionary Online) (Cowie et al., 1992; Wilks et al., 1993), thesauri e.g., Roget's International Thesaurus or ontology's (Yarowsky, 1992) or with ontologies e.g., WordNet (Sussna, 1993; Voorhees, 1993; Resnik, 1995). In (Davis, 1996) a dictionary based query translation was proposed. For disambiguation, the system uses a Part of Speech tagger to tag query terms with parts of speech information. Based on this information, the system selects the relevant terms from the dictionary, which have the same part of speech. A similarity measure is then used to compare the source language query terms and the equivalent translated terms of the aligned sentences in the parallel corpora. Only the sentences whose ranking is most similar to the source language terms will be selected. For the Arabic language, Ali et al. (2009) proposed a dictionary graph based on the WSD approach. The Authors presented a hybrid semantic-statistical method, based on computing word relatedness and a statistical measure of association to get the relationship between ambiguous words. Recently Mihalcea (2007) classified the Knowledge-based approaches for WSD into four main types: The Lesk algorithm, Semantic similarity, Selectional preferences and Heuristic methods.

Lesk Algorithm and its Variants

Lesk (1986) proposed one of the earliest dictionary based approaches to WSD. He proposed a method for counting the overlap between the words in the target context and the dictionary definitions of the senses. Patwardhan et al. (2003a) generalizes the Adapted Lesk Algorithm of Banerjee and Pedersen (2002) to a method of word sense disambiguation based on semantic relatedness. Recently Gaona et al. (2009) proposed a new sense number weight measure based on web count info obtained by a search engine. They evaluated their adapted Lesk algorithm using SemCor[4] and Senseval 2[5] test data. They pointed out that their adapted Lesk algorithm, using SemCor data, always gives an answer. On the Senseval 2 data, their adapted Lesk algorithm outperformed some other Lesk based methods. The Lesk algorithm has been applied to other languages other than English e.g., Japanese. Baldwin et al. (2008) showed that definition expansion via ontology produced a significant performance gain.

Semantic Similarity

In Semantic similarity, words in the same context are supposed to be related to each other in meaning. Thus an appropriate sense of an ambiguous word can be selected based on those meanings that are found within the smallest size n of semantic distance window (Rada et al., 1989; Mihalcea, 2007). The Semantic similarity measure is categorized into

[4] http://www.cse.unt.edu/ rada/downloads.html

[5] http://www.senseval.org/

two main categories based on the size of the context window used. Local Context relies only on information (e.g., syntactic relations) concerning the words within the context window of size n. Words that are not within the window will not be considered.

Different from local context, Global Context considers contextual information that is not within the small size of the window e.g., using Lexical chains. Lexical chains are well structured in meaning, in that related words are semantically connected. It is performed by creating a set of chains that represent different threads of relatedness throughout the text (Galley and McKeown, 2003; Nelken and Shieber, 2007).

Selectional Preferences

Selectional preferences between predicating words (verbs and adjectives) and nouns are types of linguistic information which have previously been combined with statistical methods to perform word sense disambiguation (Resnik, 1997; McCarthy and Carroll, 2003). It captures information about the possible relationships between word categories, and represents commonsense knowledge about classes of concepts. Despite the fact that selectional preferences are intuitive and understandable, WSD systems that are using selectional preference have achieved limited success (Ye, 2004). One interpretation of this deficit is that it is difficult to apply selectional preferences, in practice, to solve the problem of WSD (Mihalcea, 2007).

Heuristic Methods

A direct way to discover word meanings, in a given context, is to rely on heuristics obtained from linguistic properties in a given large text. There are three Heuristic methods, first, most-frequent-sense heuristic (it relies on the availability of the word frequency data - among all possible senses that a word may have, it is true to a great extent, that one sense occurs more often than the other senses in a given context) (McCarthy et al., 2004; Preiss et al., 2009), second, one sense-per-discourse heuristic (it appears to be extremely usual to find only one sense of a polysemous word in the same discourse) (Gale et al., 1992b), third, one-sense-per-collocation heuristic, it relies on the co-occurrence of two words in some defined relationship e.g., part-of-speech, syntactic (word tends to preserve its meaning when used in the same collection - neighboring words in a window of size n in the context of the ambiguous word provides very useful information to select the proper sense) (Yarowsky, 1993).

Corpora-based Approach

In the last few years amount of parallel corpora available in electronic format have been increased, which helps to extend the coverage of the existing system or train new system. For example, in (Brown et al., 1991; Gale et al., 1992c) the parallel aligned Hansard Corpus of Canadian Parliamentary debates was used for WSD, and in (Dagan and Itai, 1994) a monolingual corpora of Hebrew and German. The use of a bilingual corpus to disambiguate words was proposed in e.g., (Ide, 1999). Several methods for word sense disambiguation based on corpora using a supervised learning technique have been proposed. This include approaches based on Naïve Bayesian (Gale et al., 1992a), Decision List (Yarowsky, 1994), Nearest Neighbor (Ng and Lee, 1996), Transformation Based Learning (Mangu and Brill, 1997), Winnow (Golding and Roth, 1999), Boosting (Escudero and Rigau, 2000), and Naïve Bayesian Ensemble (Pedersen, 2000). For all of these methods, the ones using Naïve Bayesian Ensemble are reported to obtain the best performance for word sense disambiguation tasks with respect to the data sets used (Pedersen, 2000). Furthermore, the significant performance of the Naïve Bayesian classifier for the word sense disambiguation task has been reported by many researchers. For example, Bas et al. (2008) performed an accuracy comparison ,over 13 Polish words, between three word sense disambiguation approaches, Naïve Bayesian, Decision Table Classifier and k-Nearest Neighbours. Bas et al. (2008) found out that the Naïve Bayesian approach outperformed Table classifier and k Nearest Neighbours approaches.

The idea behind all these approaches is that it is almost always possible to determine the sense of the ambiguous word by considering its context, and thus all methods attempt to build a classifier, using features that represent the context of the ambiguous word. In addition to supervised approaches for word sense disambiguation, unsupervised approaches and combinations of them have been also proposed for the same purpose. For examples, Schütze (1998) proposed an automatic word sense discrimination which divides the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense or not, which then used for full word sense disambiguation task. Examples of unsupervised approaches were proposed in (Litkowski, 2000; Lin, 2000; Indrajit Bhattacharya, 2004). Nigam et al. (2000) proposed an unsupervised learning method using the Expectation-Maximization (EM) algorithm for text classification problems which was later improved (Shinnou and Sasaki, 2003) in order to apply it to WSD tasks. Agirre et al. (2000) combine both supervised and unsupervised lexical knowledge methods for word sense disambiguation. In (Yarowsky, 1995) and (Towell and Voorhees, 1998) approaches using rule-learning and neural networks were proposed respectively. All of the previous studies are based on the assumption that the mapping between words and word senses is widely different from one language to another. Unlike machine translation dictionaries, parallel corpora usually provide high quality translation equivalents that have been produced by experienced translators. However, in order to increase the efficiently of exploiting existing parallel corpora aligned at sentence level, explicit word-level alignments should be added where possible between sentence pairs in the training corpora. For word alignment two approaches have been proposed: statistical-based approaches, e.g., (Gale and Church, 1991; Dagan et al., 1993; Chang and Chert, 1994) and lexicon-based approaches, e.g., (Ker and Chang, 1997). Several application for word alignment in natural language processing have been studied, e.g., (Och and Ney, 2000; Yarowsky and Wicentowski, 2000). One important application for word alignment methods are the automatic extraction of bilingual lexica and terminology from corpora (Smadja et al., 1996; Melamed, 2000) and statistical machine translation systems, e.g., (Berger et al., 1994; Wu, 1996; Wang and Waibel, 1998; Niessen et al., 1998). For a more detailed overview of word alignment approaches in nature language processing see (Och and Ney, 2003a). In the past few years, the Mutual Information approach has been used to resolve translation ambiguities. For example, Mutual Information, has been used based on the target language corpora (monolingual corpora) as source of the statistical co-occurrence data e.g., (Jang et al., 1999; Qu et al., 2002; Fernandez-Amoros et al., 2010) or based on parallel corpora as source of the statistical co-occurrence data e.g., (Sari and Adriani, 2008). Furthermore, mutual information has been used to improve phrase-based machine translation e.g., in (Latiri et al., 2011). An integration of WSD in translation tasks for several languages was studied and improved by many researches. e.g., for English and Japanese (Doi and Muraki, 1992), French and English (Vickrey et al., 2005), Spanish and English (Cabezas and Resnik, 2005), Arabic and English (Ahmed and Nürnberger, 2008a,b), Portuguese and English (Specia et al., 2007) and Chinese and English (Chan and Ng, 2007).

2.3 Conclusion

Cross-lingual information retrieval provides the possibility of retrieving information across languages, without having knowledge in the target language. Two strategies to achieve this goal are to translate the query or the documents. Studies have indicated that query translation is the most used strategy in cross-lingual retrieval, due to its low computational expense. Furthermore, users who are able to understand more than one language might not be able to effectively express their need in that language. Those users, with cross-lingual system support, can cover more multilingual resources with a single query, expressed in a language they are fluent in. Despite query translation strategy advantages, there are serious limitations, such as short user queries, which provide little context, leading to a high ambiguity in translations. In order to explore approaches to tackle such issues, in this chapter, we carefully reviewed, in detail, cross-lingual retrieval approaches and issues. Furthermore, approaches to tackle cross-lingual retrieval issues has been reviewed and discussed. For specificities, we have been focused on some issues related to Arabic and German cross-lingual retrieval. For example, for the Arabic language, we have been focusing on the pre-processing step, on spelling correction and stemming. Based on the pre-processing step, a user does not need to pay any attention to word form inflection issues and thus he does not need to issue his query in the basic form. For German, besides the spelling correction, the pre-processing step was needed to tackle word compound problems, which are a real issue for some languages such as German, in respect to information retrieval or cross-lingual retrieval. This is due to the fact that compound words can result in having out-of-vocabulary (OOV) problems in cross-lingual information retrieval. In order to improve cross-lingual information retrieval effectiveness, these compound words need decompounding before translation. Once the pre-processing approaches were studied, we reviewed the word sense disambiguation approach that is used to tackle translation ambiguity issues. For example, the knowledge-based approach and the corpora-based approach was reviewed and discussed. Once the translation disambiguation approaches were reviewed, we reviewed post-processing approaches e.g., n-gram. For example, the n-gram approach can be used to detect all word form variations in the user query in order to improve the retrieval performance.

Chapter 3

Related Work on Interactive Cross-Lingual Information Retrieval Tools

3.1 Introduction

One of the main problems that impact the performance of cross-lingual information retrieval systems is how the users express their information need in form of a query. The ideal situation, in performing the cross-lingual task, is that the query term is properly formulated, and information related to it is also found in the cross-lingual system knowledge resources (e.g., bilingual dictionary, monolingual corpora, parallel corpora, etc.,) (see Chapter 2). When the query term is poorly formulated, it limits the possibilities of finding information related to it in the cross-lingual system knowledge resources and thus will limit the possibility of translating it properly. Traditional cross-lingual retrieval systems are not fully effective when the user need is not expressed appropriately. Traditional cross-lingual retrieval systems do not include any interaction scenario where the user can (with the support of the system) refine his/her need and thus improve the cross-lingual performance.

In the past, most research has been focused on the retrieval effectiveness of cross-lingual systems through information retrieval test collection approaches (Braschler et al., 2000), whereas few researchers have been focused on the user interface requirements with respect to the multilingual retrieval task (Ogden and Davis, 2000). Despite the clear effort which has been directed toward retrieval functionality and effectiveness, only little attention was paid to developing multilingual interaction tools, where users are really considered as an integral part of the retrieval process. One potential interpretation of this problem is that users of cross-lingual retrieval might not have sufficient knowledge of the target languages and therefore they are usually not involved in multilingual processes (Petrelli et al., 2004).

3.2 Cross-lingual Tools Categorization

In this thesis, we selected cross-lingual tools for review on the basis of general criteria. The studied cross-lingual tools should be web-based and perform similar or related tasks as the proposed cross-lingual tool in this thesis. Moreover, at least some of the reviewed cross-lingual tools should be developed, specifically to deal with some natural language processing issues e.g., high inflectional morphology issues for Arabic. In addition, at least the majority of the selected cross-lingual tools should consider the user as a main integral part of the retrieval process.

Since the translation is the most important part of any cross-lingual retrieval process, based on the previously mentioned criteria, we further classified the selected cross-lingual tools into two categories, depending on the best match of their used features. These two categories are cross-lingual tools that provide automatic query translation (automatic disambiguation) and cross-lingual tools that provide a user based query translation (user-based disambiguation):

- Cross-lingual tools that integrate automatic translations are the Maryland Interactive Retrieval Advanced Cross-lingual Engine *MIRACLE* (Oard et al., 2008), the cross-lingual interactive system *WORDS* (Lopez-Ostenero et al., 2002), the cross-lingual information system *LIC2M* (Semmar et al., 2005) and the cross-lingual patent retrieval system (we named it *Patent CLIR*) (Bian and Teng, 2008).

- Cross-lingual tools that use user-based translation are the German Research Center for Artificial Intelligence's (DFKI) *MULINEX* system (Capstick et al., 2000), the New Mexico State University *Keizai* system (Ogden and Davis, 2000), a Multilingual Information Retrieval Tool *UCLIR* (Abdelali et al., 2003) and *MultiLex-Explorer* (De Luca et al., 2006).

In Section 3.3, a detailed overview of the main properties for each reviewed cross-lingual tool is presented and discussed.

In the following, we describe the cross-lingual tools in detail starting with each classified category.

3.2.1 Automatic Translation CLIR Tools

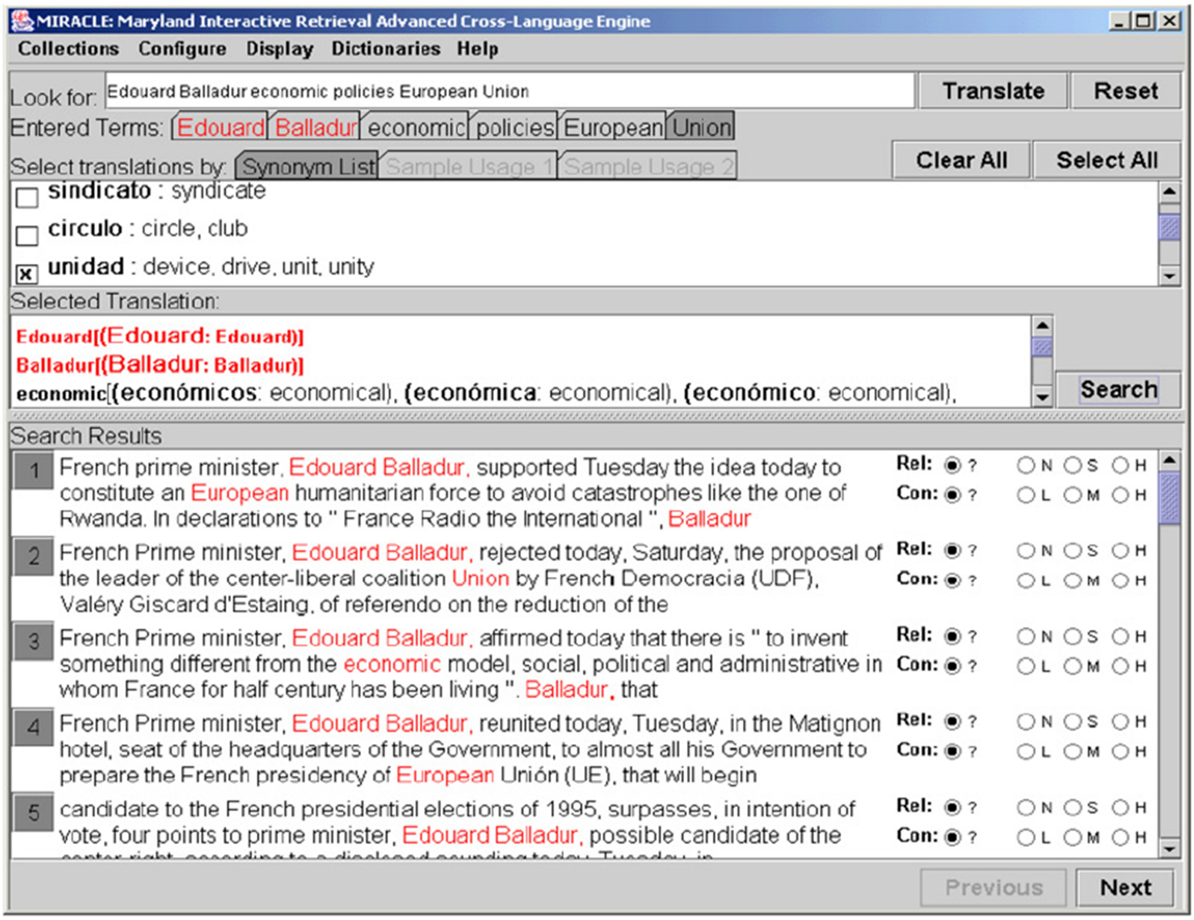
3.2.1.1 MIRACLE

In order to support the interactive cross-lingual retrieval, the system uses the *user-assisted query translation* (Oard et al., 2008). The user assisted-query translation feature supports the user to select the correct translation. However, there might be a case when the user might delete a correct translation. The system reacts, in that the searcher can see the effect of the choice and have possibilities to learn better control of the system. This is done by providing the following features, the meaning of the translation (loan word or proper name), using back translation, a list of possible synonyms are provided. Translation examples of usage are obtained from translated or topically-related text.

In MIRACLE, there are two types of query translations, fully automatic query translation (using machine translation) and user-assisted query translation. In fully automatic translation the user can be involved only once. After the system translates the query and retrieves the search results, the user can refine the query if he/she is not satisfied after examining the search results. In the user-assisted query translation, four possible refinement steps give the user an opportunity to be involved in the translation process. First, based on evidence about the meanings of the proposed translations by the system, the user has an opportunity to deselect some of the proposed translations before the search can be performed. Second the user can reform the query based on evidence about the meanings of the proposed translations. Third, the user can reform the query based on examining the search results. Fourth, in case the search result doesn't satisfy the user's needs, the user has a possibility to deselect/reselect the translations.

In other words, the user submits his/her query; the system provides her/him with translation alternatives. Before the search can be performed, the user has an opportunity to deselect some of the proposed translations. The user has an opportunity to refine his/her query based on evidence about the meanings of the proposed translations by the system. After the search is performed, the system provides the user with the search results (see Figure 3.1). If the user is satisfied with the search result then there will be no further actions by the system. In contrast, based on examining the search result, the user has two opportunities: refine his/her query and perform a new search or deselect/reselect a translation out of the translation alternatives proposed by the system. The interaction between the system and the user, gives the user possibilities to see the effect of his/her decision (selection, deselection of the translation or query refinement) in that the user can cycle the search till it satisfies his/her needs. A very important aspect in MIRACLE, is that the system provides the user with immediate feedback in response to any action, which gives the user an important opportunity to refine his/her search. The rapid adaption to new languages was taken into account in the design of the MIRACLE system. The query language is always English, in MIRACLE. However, language resources that are available for English can be leveraged, regardless of the document language. Currently, MIRACLE works with a simple bilingual term list. However, it is designed to readily leverage additional resources when they are available.

Although MIRACLE overcomes some of the limitations of the previously mentioned cross-lingual retrieval interaction tools, it also has some limitations. For example, despite the use of automatic translation in MIRACLE, the user has no influence on refining the translation before the search can be conducted e.g., providing contextual information that describes the translation in the user's own language, in that the user can have a certain degree of confidence in the translation. In addition, to the previously mentioned limitation, in MIRACLE, single word translations are used, which forces the user to spend a lot of effort checking each single translation alternative with their meanings before he/she can select/deselect translations.


MIRACLE: Maryland Interactive Retrieval Advanced Cross-Language Engine
Collections Configure Display Dictionaries Help
Translate
Reset
Synonym List
Sample Usage 1
Sample Usage 2
Clear All
Select All
Search
Previous
Next

Look for: Edouard Balladur economic policies European Union

Entered Terms: Edouard Balladur economic policies European Union

Select translations by: Synonym List

☐ sindicato : syndicate

☐ circulo : circle, club

☒ unidad : device, drive, unit, unity

Selected Translation:

Edouard[(Edouard: Edouard)]
Balladur[(Balladur: Balladur)]
economic[(económicos: economical), (económica: economical), (económico: economical),

Search Results

1 French prime minister, Edouard Balladur, supported Tuesday the idea today to constitute an European humanitarian force to avoid catastrophes like the one of Rwanda. In declarations to " France Radio the International ", Balladur — Rel: ● ? ○ N ○ S ○ H; Con: ● ? ○ L ○ M ○ H

2 French Prime minister, Edouard Balladur, rejected today, Saturday, the proposal of the leader of the center-liberal coalition Union by French Democracia (UDF), Valéry Giscard d'Estaing, of referendo on the reduction of the — Rel: ● ? ○ N ○ S ○ H; Con: ● ? ○ L ○ M ○ H

3 French Prime minister, Edouard Balladur, affirmed today that there is " to invent something different from the economic model, social, political and administrative in whom France for half century has been living ". Balladur, that — Rel: ● ? ○ N ○ S ○ H; Con: ● ? ○ L ○ M ○ H

4 French Prime minister, Edouard Balladur, reunited today, Tuesday, in the Matignon hotel, seat of the headquarters of the Government, to almost all his Government to prepare the French presidency of European Unión (UE), that will begin — Rel: ● ? ○ N ○ S ○ H; Con: ● ? ○ L ○ M ○ H

5 candidate to the French presidential elections of 1995, surpasses, in intention of vote, four points to prime minister, Edouard Balladur, possible candidate of the — Rel: ● ? ○ N ○ S ○ H; Con: ● ? ○ L ○ M ○ H

Figure 3.1: MIRACLE query assistance.

3.2.1.2 WORDS

Lopez-Ostenero et al. (2002) proposed a cross-lingual interactive approach which provides the user with the possibility of formulating and refining a query. It includes a reference system (WORDS) that supports the user in selecting proper translations for each of the query terms. Furthermore, it includes the possibility of assisting the user in formulating his/her query, by providing him/her with a set of relevant phrases. The user has the possibility of selecting promising phrases, in the presented documents, in order to improve the search. The reference system (WORDS) includes a user query translation assistance and refinement.

As Figure 3.2 shows, the WORDS translates each query term (in Spanish) by providing all possible translations for each term in English. In order to give the user confidence in the translation, WORDS uses back translation (from English to Spanish). This allows the user to deselect any translation before the search can be performed. Once the user selects the suitable translation the search can be performed. In this case, English documents will be retrieved based on the English translation. Once the documents are retrieved (in English) the system provides the user with a summary of each retrieved document, in the user's own language. This summary includes translation of all noun phrases in the document, using the Systran machine translation system[1] and the document title is automatically translated. Based on this information, the user can mark the document as relevant, irrelevant or unsure (see Figure 3.2). In addition, the user has the possibility of taking no action and leaving the document unmarked.

In order to refine the query to improve the search, the user can check the retrieved document translation (in Spanish) and point to any query term in the document. The system then points to the English query terms (one of the possible translations for the Spanish query term). The user then has the possibility to select or deselect any English term. This allows the user to keep only the appropriate translations for the Spanish query term. In addition, the user also has the possibility of selecting any term in the translation he/she thinks can improve the search e.g., any term in the context of the translation. The selected term will then be added to the query as an extra term. Furthermore, the system provides the possibility of phrase-based searching, where the system first extracts noun phrases from a dataset (iCLEF topic[2]); filter phrases with appropriate translations, which will be displayed to the user for selection.

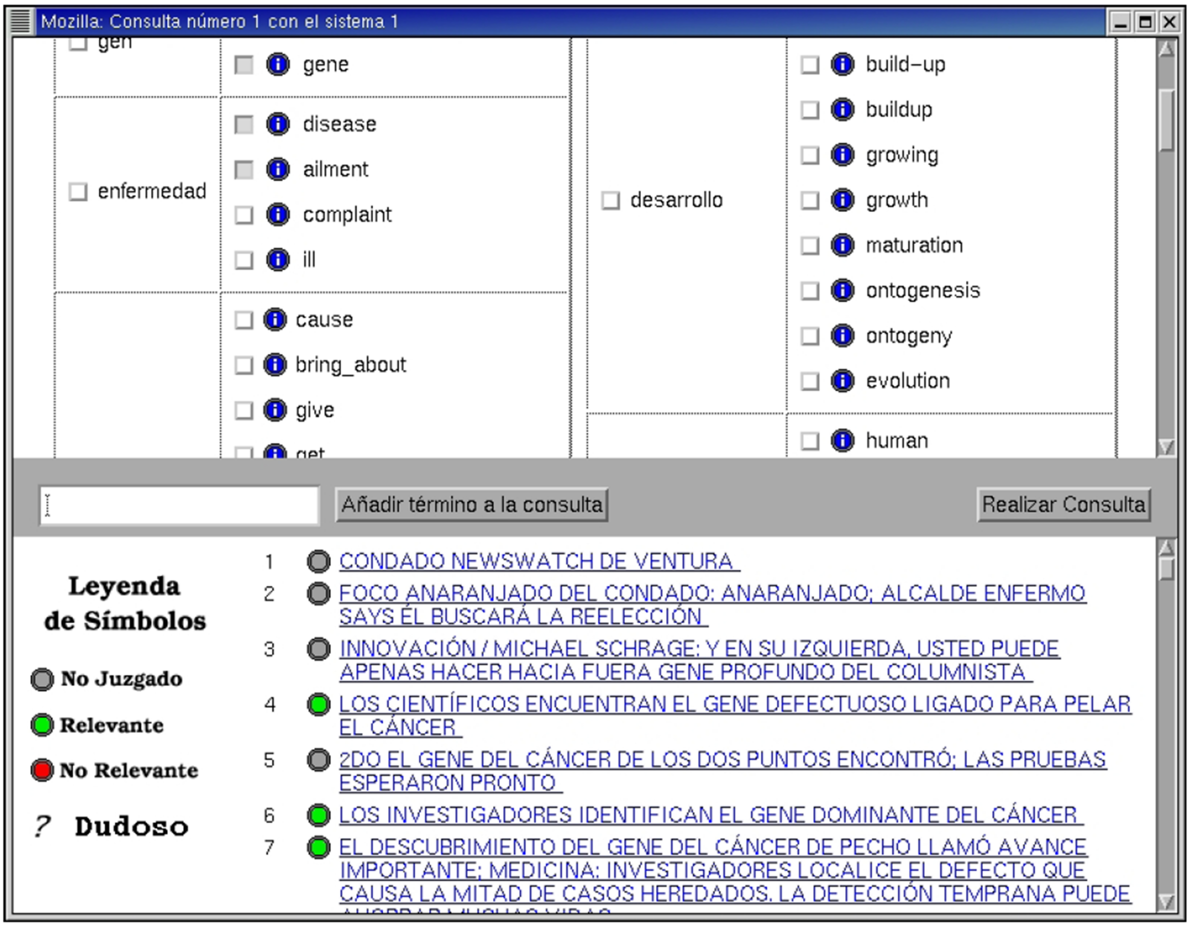
Once the user selects any phrases, those selected phrases are automatically translated

[1] http://www.systran.de/

[2] http://nlp.uned.es/iCLEF/2002/guidelines.htm#dataprov

so the user then can perform a monolingual search in his/her language. The translation of the phrases selected by the user can be used to refine the query, in the form of term suggestions, if needed. This is done as follows: the user clicks on a noun-phrase in a document, the systems automatically translates the selected noun-phrase and uses it to enlarge the original query before the monolingual search is performed. The user can then check the re-ranked document and see the affect of the query refinement on his/her search.

Despite lots of support for the user to perform the cross-lingual task, the user has to check all translation alternatives with their definitions in order to disambiguate translations. Furthermore, the query refinement depends on the automatic translation, which can't be accurate in all cases i.e., inaccurate translation leads to low retrieval performance.



Leyenda de Símbolos

○ No Juzgado

🟢 Relevante

🔴 No Relevante

? Dudoso

1 ○ CONDADO NEWSWATCH DE VENTURA

2 ○ FOCO ANARANJADO DEL CONDADO: ANARANJADO; ALCALDE ENFERMO SAYS ÉL BUSCARÁ LA REELECCIÓN

3 ○ INNOVACIÓN / MICHAEL SCHRAGE: Y EN SU IZQUIERDA, USTED PUEDE APENAS HACER HACIA FUERA GENE PROFUNDO DEL COLUMNISTA

4 🟢 LOS CIENTÍFICOS ENCUENTRAN EL GENE DEFECTUOSO LIGADO PARA PELAR EL CÁNCER

5 ○ 2DO EL GENE DEL CÁNCER DE LOS DOS PUNTOS ENCONTRÓ; LAS PRUEBAS ESPERARON PRONTO

6 🟢 LOS INVESTIGADORES IDENTIFICAN EL GENE DOMINANTE DEL CÁNCER

7 🟢 EL DESCUBRIMIENTO DEL GENE DEL CÁNCER DE PECHO LLAMÓ AVANCE IMPORTANTE; MEDICINA: INVESTIGADORES LOCALICE EL DEFECTO QUE CAUSA LA MITAD DE CASOS HEREDADOS. LA DETECCIÓN TEMPRANA PUEDE

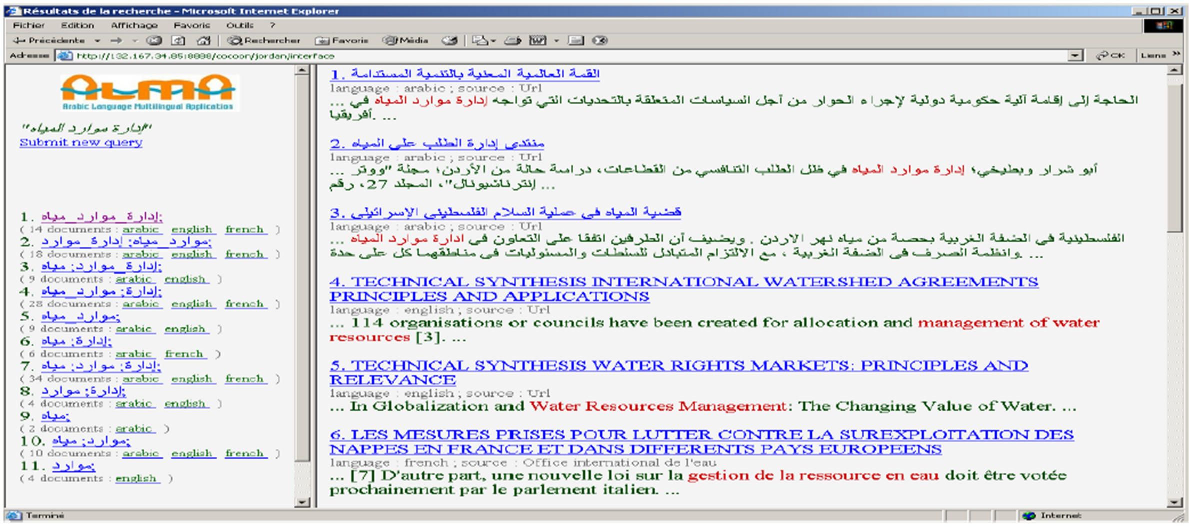
Figure 3.2: System assisted translation and judged retrieved documents.

3.2.1.3 LIC2M

Semmar et al. (2005) proposed a cross-lingual information system, based on rich linguistic analysis of documents and queries (LIC2M). LIC2M supports Arabic, English and French languages. The LIC2M cross-lingual system consists of six models:

- A linguistic analyzer which is responsible for processing the query and the documents that will be indexed. To perform its task, a linguistic analyzer uses linguistic resources. For each language, a proper linguistic resource is provided.
    - A full form dictionary: in this dictionary each word is assigned with its part-of-speech tags and its linguistic feature e.g., gender, number, etc.
    - A monolingual reformulation dictionary: used to expand the query e.g., adding synonyms, hyponyms, etc.
    - Bilingual dictionary: used for translations between languages.
    - A set of rules: used for tokenization purposes.
    - A parser: used to parse sentences, extracting compounds etc.
    - Name entity recognition: used to identify named entities etc.
- A statistical analyzer, which is responsible for providing statistical information about the documents that will be indexed. It is used to compare the similarity between documents and queries. In order to improve the retrieval process, a weight is assigned to each word in the database according to its discrimination power.
- A reformulator, which is responsible for expanding the user query. It is needed when significant results are not obtained using the previous models e.g., linguistic analysis etc. It expands the query with related terms e.g., synonym, hyponyms, etc.
- A comparator, which is responsible for computing semantic similarity between the indexed documents and the query.
- An indexer, which is responsible for storing the documents in a database.
- A search engine, which is responsible for searching the index and retrieving the relevant documents.

Figure 3.3 shows the LIC2M interface, where the user submits a query in his/her native language. The system processes the query, and expands it, if needed. The query is then submitted to the integrated search engine, which is responsible for retrieving the relevant documents from the local collection. An integrated translation engine is responsible for translating the retrieved documents. This translation engine is used by the system via its web API.


Résultats de la recherche - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://132.167.34.85:8888/cocoon/jordan/interface
OK
Liens
ALMA
Arabic Language Multilingual Application
Terminé
Internet

"إدارة موارد المياه"
Submit new query

1. إدارة_موارد_مياه;
(14 documents : arabic english french)
2. موارد_مياه; إدارة_موارد;
(18 documents : arabic english french)
3. إدارة_موارد; مياه;
(9 documents : arabic english)
4. إدارة; موارد_مياه;
(28 documents : arabic english french)
5. موارد_مياه;
(9 documents : arabic english)
6. إدارة; مياه;
(6 documents : arabic french)
7. إدارة; موارد; مياه;
(34 documents : arabic english french)
8. إدارة; موارد;
(4 documents : arabic english)
9. مياه;
(2 documents : arabic)
10. موارد; مياه;
(10 documents : arabic english french)
11. موارد;
(4 documents : english)

1. القمة العالمية المعنية بالتنمية المستدامة
language : arabic ; source : Url
... الحاجة إلى إقامة آلية حكومية دولية لإجراء الحوار من آجل السياسات المتعلقة بالتحديات التي تواجه إدارة موارد المياه في
أفريقيا. ...

2. منتدى إدارة الطلب على المياه
language : arabic ; source : Url
... أبو شرار وبطيخي؛ إدارة موارد المياه في ظل الطلب التنافسي من القطاعات، دراسة حالة من الأردن، مجلة "ووتر
إنترناشيونال"، المجلد 27، رقم ...

3. قضية المياه في عملية السلام الفلسطيني الإسرائيلي
language : arabic ; source : Url
... الفلسطينية في الضفة الغربية بحصة من مياه نهر الاردن . ويضيف آن الطرفين اتفقا على التعاون في ادارة موارد المياه
وانظمة الصرف في الضفة الغربية ، مع الالتزام المتبادل للسلطات والمسئوليات في مناطقهما كل على حدة ...

4. TECHNICAL SYNTHESIS INTERNATIONAL WATERSHED AGREEMENTS PRINCIPLES AND APPLICATIONS
language : english ; source : Url
... 114 organisations or councils have been created for allocation and management of water resources [3]. ...

5. TECHNICAL SYNTHESIS WATER RIGHTS MARKETS: PRINCIPLES AND RELEVANCE
language : english ; source : Url
... In Globalization and Water Resources Management: The Changing Value of Water. ...

6. LES MESURES PRISES POUR LUTTER CONTRE LA SUREXPLOITATION DES NAPPES EN FRANCE ET DANS DIFFERENTS PAYS EUROPEENS
language : french ; source : Office international de l'eau
... [7] D'autre part, une nouvelle loi sur la gestion de la ressource en eau doit être votée prochainement par le parlement italien. ...

Figure 3.3: Search results user interface.

Although in the LIC2M system, the query is expanded with extra terms extracted from the target languages using a bilingual dictionary, it wasn't mentioned how the system deals with the translation ambiguity i.e., not all translations are relevant to the user query. Using POS tagging for disambiguation would not be enough as it is very difficult to extract any syntactic information from the user query i.e., user search engine queries are usually between 2.4 and 2.7 in length (Gabrilovich et al., 2009). Furthermore, users have no possibility to interact with the system and refine the retrieved document translations which are obtained using a web translator.

3.2.1.4 Patent CLIR

Bian and Teng (2008) proposed a cross-lingual patent retrieval and classification system that makes use of the various free web translators to translate the user query (see Figure 3.4). The system was designed for Japanese/English cross-lingual patent retrieval. The proposed system provides monolingual and cross-lingual functionalities. The input to the system is the query or the selection of the topic file. The user then can use one of the different web translators to translate the query. The proposed system gives the user a possibility to modify the translation. The different system modules are described in the following:

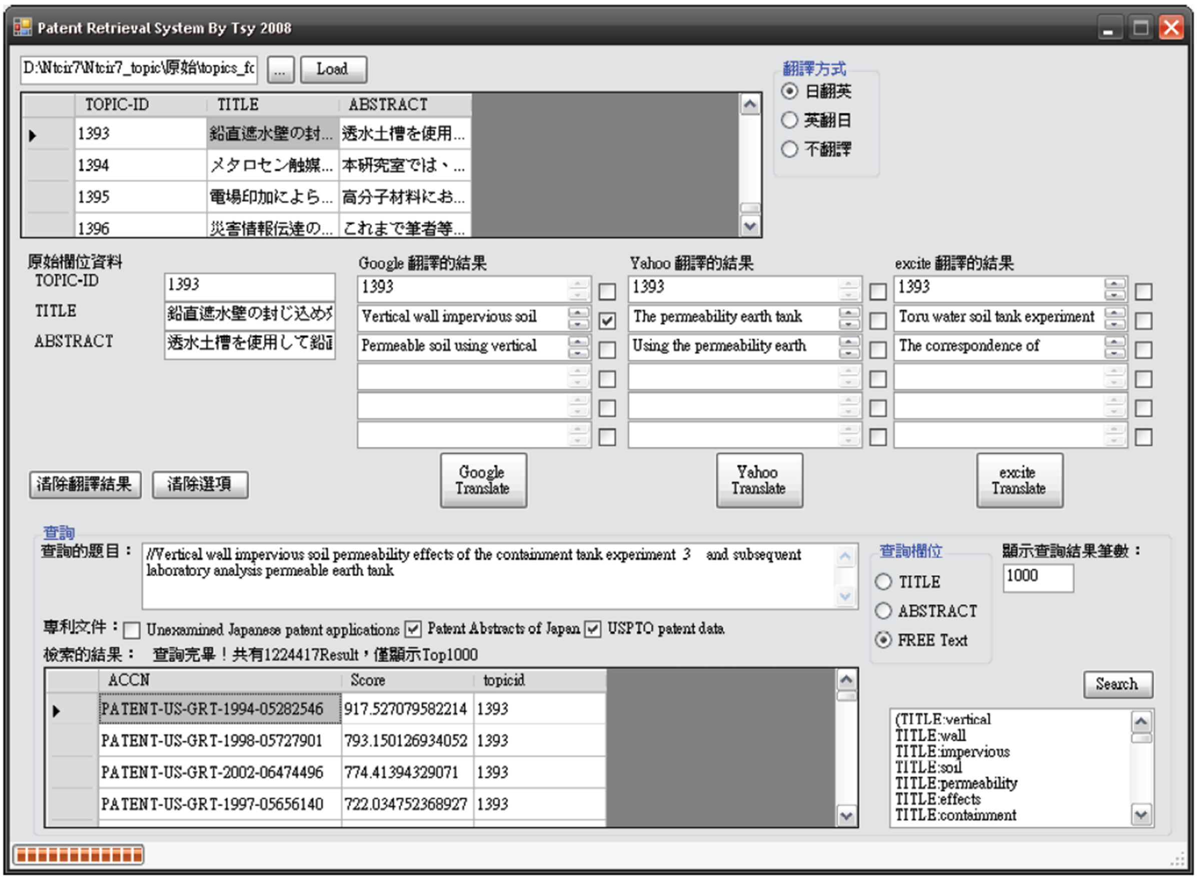
- Indexing module: in the indexing module, the multi-lingual patent document sets are processed and indexed. The system uses two types of indexing methods, a word-based method to index the English text collection and the bigram-based method to index the Japanese text collections.

- Translation module: in the translation module the query is translated from the source language to the target language. The query is sent via the system to the selected online translator system by the user. The obtained translation is then obtained and displayed to the user. Since the user can use different translators at the same time, it is possible that the user can review and modify the translation based on the results from different translators.

- Classification module: in the classification module, the retrieved patent documents are processed in order to classify them based on the International Patent Classification (IPC)[3]. This process is performed as follows:
  - The documents are retrieved based on the topic of the input patent (query).
  - The first top ranked 3000 patent documents and their IPC code from the patent data collection are retrieved.
  - The score of the IPC code is computed. This is done by computing the similarity score between the query and the retrieved documents.
  - The IPC codes, in step 3, are sorted by their score.

Despite the possibility of refining the web translators translation integrated in the tool (selecting or removing translations from 3 different web translators integrated in the tool), users with low knowledge in the target language will have no possibility to select suitable translations from different translators i.e., no information in the user's own language to describe the translation so the user can interact with it effectively.

[3] http://www.wipo.int/classifications/ipc/en/



D:\Ntcir7\Ntcir7_topic\原始\topics_fc

翻譯方式
● 日翻英
○ 英翻日
○ 不翻譯

TOPIC-ID	TITLE	ABSTRACT
1393	鉛直遮水壁の封…	透水土槽を使用…
1394	メタロセン触媒…	本研究室では、…
1395	電場印加によら…	高分子材料にお…
1396	災害情報伝達の…	これまで筆者等…

原始欄位資料
TOPIC-ID 1393
TITLE 鉛直遮水壁の封じ込め
ABSTRACT 透水土槽を使用して鉛

Google 翻譯的結果		Yahoo 翻譯的結果		excite 翻譯的結果	
1393	□	1393	□	1393	□
Vertical wall impervious soil	☑	The permeability earth tank	□	Toru water soil tank experiment	□
Permeable soil using vertical	□	Using the permeability earth	□	The correspondence of	□
	□		□		□
	□		□		□
	□		□		□

查詢
查詢的題目： //Vertical wall impervious soil permeability effects of the containment tank experiment 3 and subsequent laboratory analysis permeable earth tank

專利文件：□ Unexamined Japanese patent applications ☑ Patent Abstracts of Japan ☑ USPTO patent data

檢索的結果： 查詢完畢！共有1224417Result，僅顯示Top1000

ACCN	Score	topicid
PATENT-US-GRT-1994-05282546	917.527079582214	1393
PATENT-US-GRT-1998-05727901	793.150126934052	1393
PATENT-US-GRT-2002-06474496	774.41394329071	1393
PATENT-US-GRT-1997-05656140	722.034752368927	1393

查詢欄位
○ TITLE
○ ABSTRACT
● FREE Text

顯示查詢結果筆數： 1000

(TITLE:vertical
TITLE:wall
TITLE:impervious
TITLE:soil
TITLE:permeability
TITLE:effects
TITLE:containment

Figure 3.4: The cross-lingual patent information retrieval interface.

3.2.2 User-based Translation CLIR Tools

3.2.2.1 Mulinex

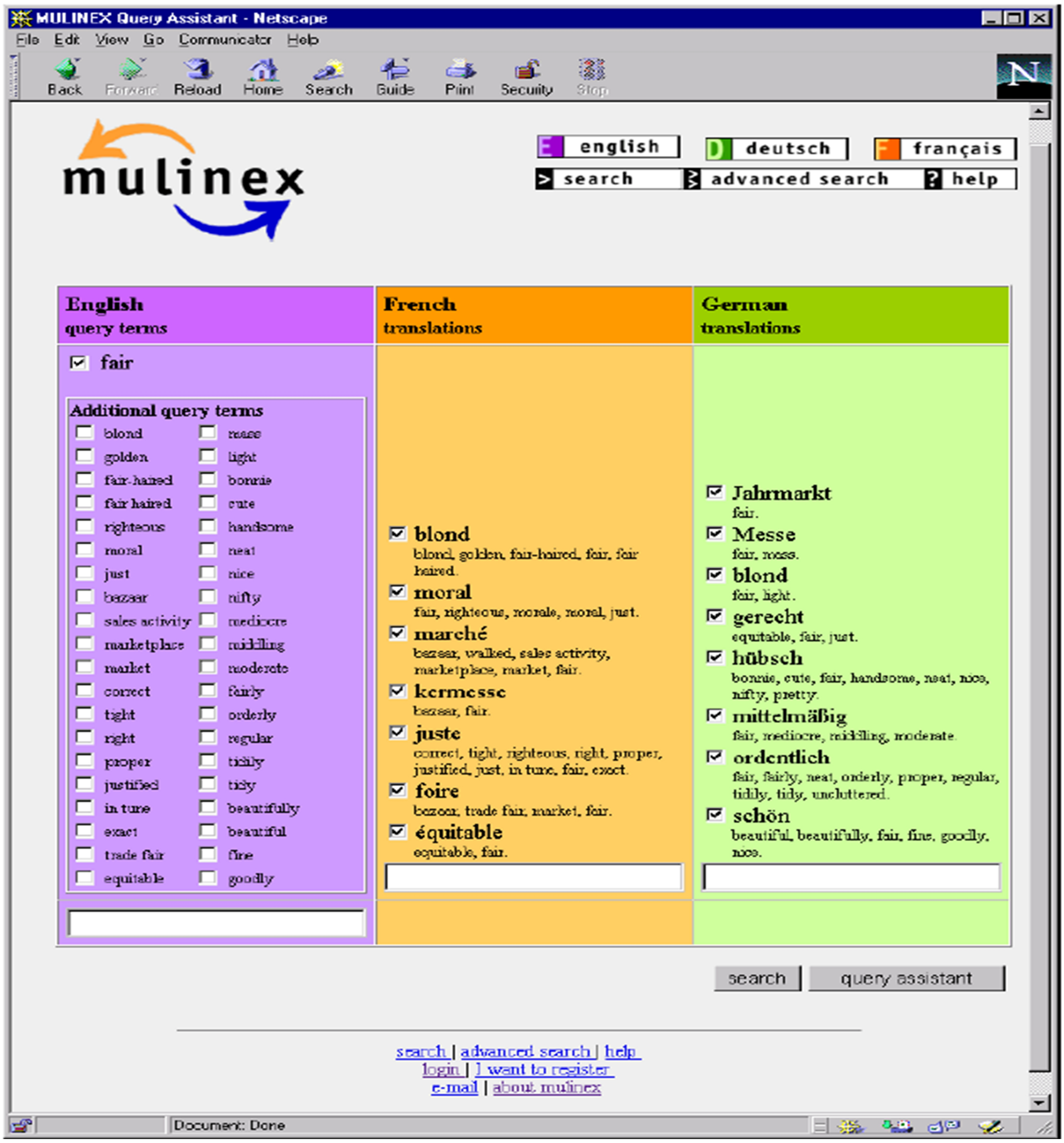
Mulinex supports cross-lingual search by giving the users possibilities to formulate, expand and disambiguate queries. Furthermore, the users are able to filter the search results and read the retrieved documents by using only their native language (Capstick et al., 2000). Mulinex performs the multilingual functionality based on a dictionary-based query translation. Besides the cross-lingual functionality, where the query is submitted in one language and the retrieved documents are presented in another language, Mulinex provides the automatic translation of documents and their summaries. In Mulinex, three languages are supported, French, German, and English. In Mulinex, the cross-lingual retrieval process is fully supported by the translation of the queries, documents and their summaries. Hereby, users do not need to have any knowledge about the target language. Mulinex provides a lot of functionality to support the retrieving of the documents in multilingual collections. Examples of these functionalities are translation of the user's query, interactive disambiguation of the query translation, interactive query expansion, on-demand translation of summaries and search results, etc.

The Mulinex interface is available in three languages English, German, and French. Since the search engine queries are usually between 2.4 and 2.7 in length (Gabrilovich et al., 2009) which typically does not provide enough context for automatic disambiguation, Mulinex using *"query assistant"* provides an opportunity for interactive query translation disambiguation. This task is performed by the "query assistant" by performing the back translation. The translated query terms are translated back into the original query language. However, this approach has some clear limitations. When no synonyms can be found in the dictionary, the technique is not helpful; and significant homonymy in the target language can result in confusing back translations (Oard et al., 2008). In Mulinex, the back translation concept is used for expanding the original query with potentially relevant terms. The query term translation is translated back to the original query language; the result of this step is having a list of possible translation in the query's original language. The user, in this case, can select some of these translation alternatives, in order to expand the user query. For example, the user submits the query,"fair", in English. The system provides the user with alternative translations in French and German.

For French, the system provides the following translations: ("blond", "moral","marché", "kermesse", "juste", "foire" and "équitable"). For German the system provides the following eight translations: ("Jahrmarket", "Messe", "blond", "gerecht", "hübsch", "mittelmäßig", "ordentlich" and "schön") (see Figure 3.5). In order to expand the query, the system translates back the translated user query terms. The result of this step is having a translation alternatives in the user's original query language. For example, the back translation alternatives for the French translation "marché" are ("bazaar", "walked", "sales activities", "marketplace", "market" and "fair") and the back translation alternative for the French translation "foire" are ("bazaar", "trade fair", "market" and "fair"). Based on the translation alternatives provided by the system, the translation "sales activities" and "trade fair" can be selected by the user as relevant expanded terms to the original query "fair".

3.2.2.2 Keizai

The goal of the Keizai project is to provide a Web-based cross-lingual text retrieval system that accepts the query in English and searches Japanese and Korean web data (Ogden and Davis, 2000). Furthermore, the system displays English summaries of the top ranking retrieved documents. In Keizai the query terms are translated into Japanese


MULINEX Query Assistant - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
mulinex
english deutsch français
search advanced search help
search query assistant
search | advanced search | help
login | I want to register
e-mail | about mulinex
Document: Done

English query terms

☑ fair

Additional query terms

☐ blond ☐ mass
☐ golden ☐ light
☐ fair-haired ☐ bonnie
☐ fair haired ☐ cute
☐ righteous ☐ handsome
☐ moral ☐ neat
☐ just ☐ nice
☐ bazaar ☐ nifty
☐ sales activity ☐ mediocre
☐ marketplace ☐ middling
☐ market ☐ moderate
☐ correct ☐ fairly
☐ tight ☐ orderly
☐ right ☐ regular
☐ proper ☐ tidily
☐ justified ☐ tidy
☐ in tune ☐ beautifully
☐ exact ☐ beautiful
☐ trade fair ☐ fine
☐ equitable ☐ goodly

French translations

☑ **blond**
blond, golden, fair-haired, fair, fair haired.

☑ **moral**
fair, righteous, morale, moral, just.

☑ **marché**
bazaar, walked, sales activity, marketplace, market, fair.

☑ **kermesse**
bazaar, fair.

☑ **juste**
correct, tight, righteous, right, proper, justified, just, in tune, fair, exact.

☑ **foire**
bazaar, trade fair, market, fair.

☑ **équitable**
equitable, fair.

German translations

☑ **Jahrmarkt**
fair.

☑ **Messe**
fair, mass.

☑ **blond**
fair, light.

☑ **gerecht**
equitable, fair, just.

☑ **hübsch**
bonnie, cute, fair, handsome, neat, nice, nifty, pretty.

☑ **mittelmäßig**
fair, mediocre, middling, moderate.

☑ **ordentlich**
fair, fairly, neat, orderly, proper, regular, tidily, tidy, uncluttered.

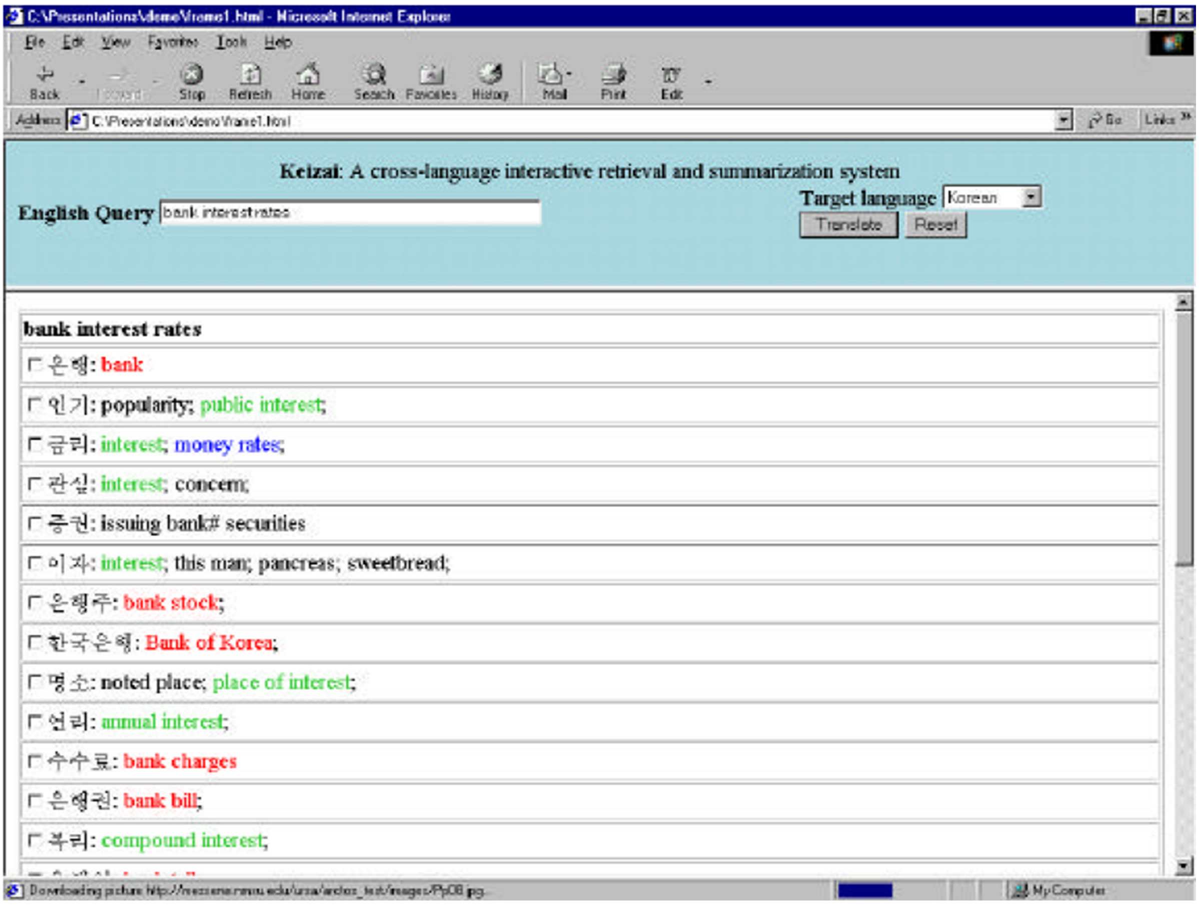
☑ **schön**
beautiful, beautifully, fair, fine, goodly, nice.

Figure 3.5: MULINEX query assistance.

or Korean languages along with their English definitions and thus this feature allows the user to disambiguate the translations (see Figure 3.6).

Based on the English definitions of the translated query terms, the user who does not understand the Japanese or Korean language can select the appropriate translation, out of several possible translations. Once the user selects those translations whose definitions are consistent with the information needed, the search can be performed. Only documents that are relevant to the selected translations will be retrieved. For each retrieved document in Japanese or Korean, an English summary along with a target document language summary will be displayed in the Keizai interface.

Keizai investigates the effectiveness of representing the retrieved documents together with small images, which they call *"Document Thumbnail Visualizations"*. Using this document representation, the retrieved documents are retained with a familiar shape and format and thus the user can see how the query terms are distributed in the retrieved documents. Using this technique the authors investigated the potential advantage of the representation of the documents as one image within the context of different interactive text retrieval tasks. In Keizai, the authors could show that the visualization improved recall and efficiency.


C:\Presentations\demo\frame1.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Edit
Address
Go
Links
Translate
Reset
Downloading picture http://messene.nmsu.edu/ursa/archos_test/images/Pp08.jpg
My Computer

Address: C:\Presentations\demo\frame1.html

Keizai: A cross-language interactive retrieval and summarization system

English Query bank interest rates

Target language Korean

bank interest rates

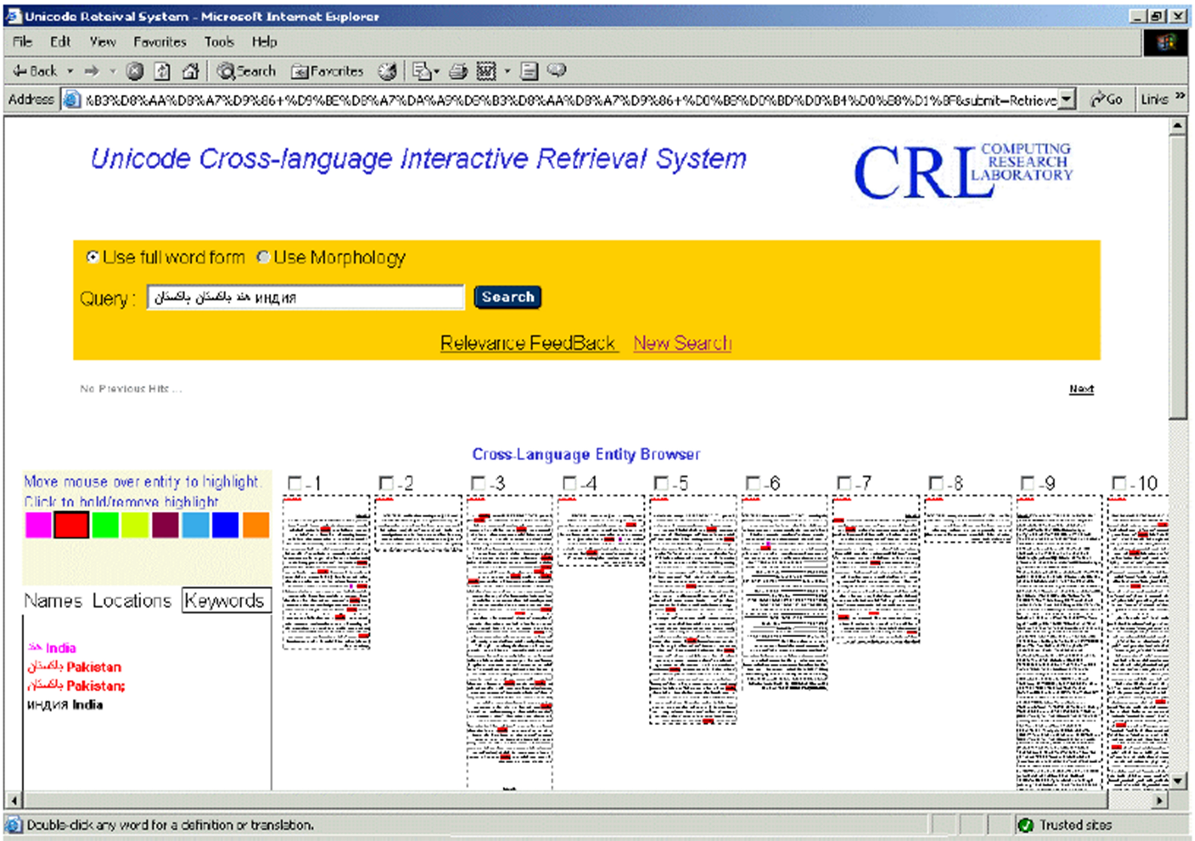
- □ 은행: bank
- □ 인기: popularity; public interest;
- □ 금리: interest; money rates;
- □ 관심: interest; concern;
- □ 증권: issuing bank# securities
- □ 이자: interest; this man; pancreas; sweetbread;
- □ 은행주: bank stock;
- □ 한국은행: Bank of Korea;
- □ 명소: noted place; place of interest;
- □ 연리: annual interest;
- □ 수수료: bank charges
- □ 은행권: bank bill;
- □ 복리: compound interest;

Figure 3.6: Keizai query term selection.

3.2.2.3 UCLIR

In UCLIR, the Arabic language was included. The system performs its task in any of the following three different modes: the first mode, using a multilingual query (query can consist of terms of different languages), the second mode using an English query without user involvement in the multilingual query formulation, the third mode using an English query with user involvement in the formulation of the multilingual queries (Abdelali et al., 2003). The first system mode: Multilingual query, in this mode the system accepts a query which consists of terms of various languages. The system will retrieve the relevant documents regardless of the query term language. The documents in the entire multilingual collection, those relevant to one of the query terms, will be retrieved. The second system mode: English query: non-interactive approach, this mode is based on the use of a set of bilingual dictionaries for translating an English query into the different target languages. First, for the English query term a set of possible translations will be obtained from the bilingual dictionaries. Second, the set of possible translations will be compared with an index word list (obtained from the system's entire multilingual resource); the translations which are not in the index word list will be eliminated from the query. The filtered query then can be used to retrieve the relevant documents from the system's entire resource and these retrieved documents are then displayed to the user in the system interface. The third system mode: English query: interactive approach, in this system mode, the user is involved in the selection of appropriate translations. The same as in the second mode, a set of possible translations will be obtained from the bilingual dictionaries and compared with the index word list; the translations which are not in the index word list will be eliminated from the query. The rest will be kept and presented to the user in the system interface along with their English translation beside other information e.g., part of speech. At the end, the user selects the appropriate translation out of the filtered translation list. The selected multilingual terms then can be used to form the multilingual query which is then submitted to retrieve the relevant documents from the system's entire multilingual resource. After the retrieval process is performed, the relevant retrieved documents can be then translated into English. To perform the document translation, two approaches are used. The first approach is word-level translation, where the user can click on the selected word and this word will be translated using the dictionary and displayed as a pop-up view to the user with its lexical information. The second approach is a document-level translation, where the whole retrieved document, using a translation system, is translated into English.

Similar to Keizai, UCLIR uses "Document Thumbnail Visualizations" (see Figure 3.7). The retrieved documents are retained with familiar shape and format which make it possible for the user to see how the query terms are distributed in the retrieved documents. Although the system in the second mode automates the process of the appropriate translation selection by comparing a set of possible translations with an index word list (the translations which are not in the index word list will be eliminated from the query). However, this can include an irrelevant translation to the user query since it is possible that not all translations can be relevant to the original query term.



Unicode Retreival System - Microsoft Internet Explorer

Address: %B3%D8%AA%D8%A7%D9%86+%D9%BE%D8%A7%DA%A9%D8%B3%D8%AA%D8%A7%D9%86+%D0%B8%D0%BD%D0%B4%D0%B8%D1%8F&submit=Retrieve

Unicode Cross-language Interactive Retrieval System

● Use full word form ○ Use Morphology

Query : هند پاکستان پاکستان индия

No Previous Hits ...

Cross-Language Entity Browser

Move mouse over entity to highlight.
Click to hold/remove highlight

Names Locations Keywords

هند India
پاکستان Pakistan
پاکستان Pakistan;
индия India

Double-click any word for a definition or translation.

Figure 3.7: UCLIR document thumbnail visualizations.

3.2.2.4 MultiLexExplorer

The goal of the MultiLexExplorer tool is to support multilingual users in performing their web search. Furthermore, the MultiLexExplorer supports the user in disambiguating word meanings by providing the user with information about the distribution of words in the web (De Luca et al., 2006). The tool allows users to explore combinations of query term translations by visualizing EuroWordNet[4] relations together with search results and search statistics obtained from web search engines. Based on the EuroWordNet, the tool supports the user with the following functionality:

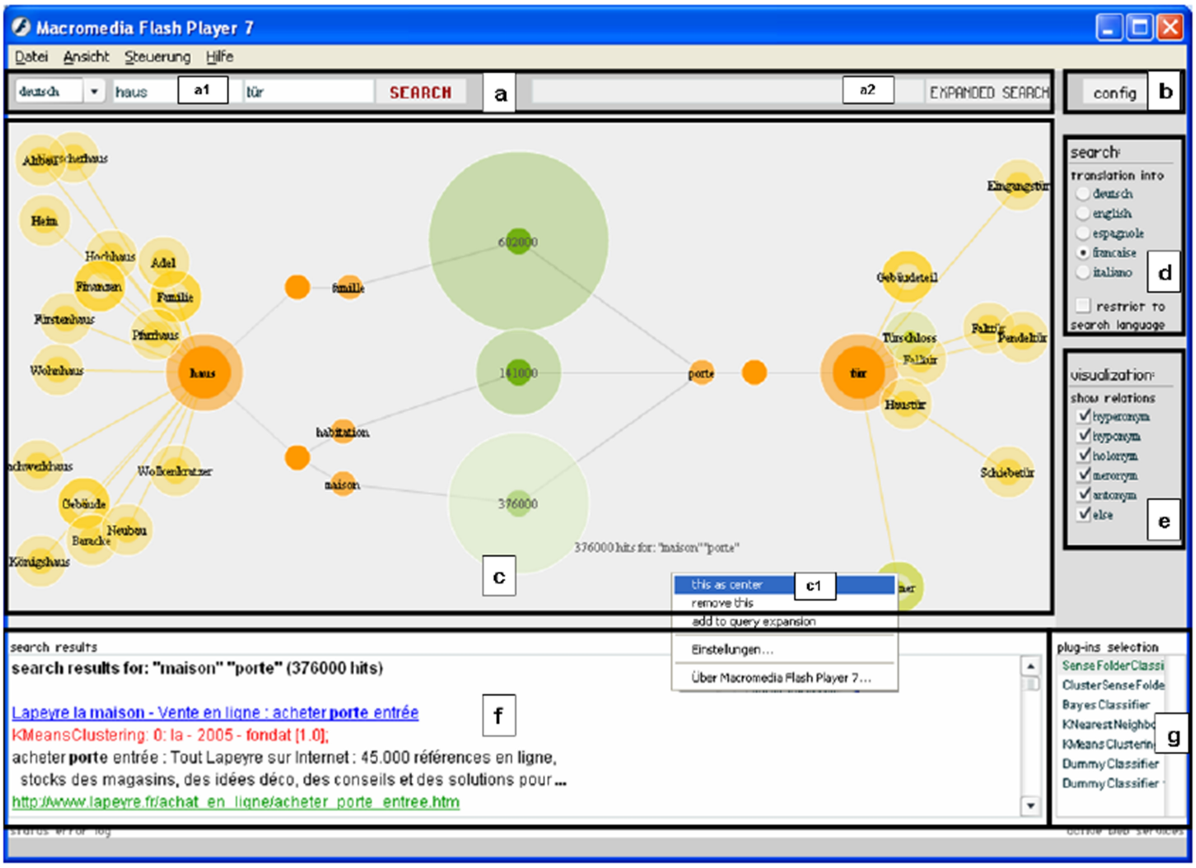
- exploring the context of a given word in the general hierarchy,
- searching in different languages, e.g., by translating word senses using the interlingual index of EuroWordNet,

[4] http://www.illc.uva.nl/EuroWordNet/

- disambiguating word sense for combinations of words,
- provide the user with the possibility to interact with the system i.e., changing the search word and the number of retrieved documents,
- expanding the original query with extra relevant terms, and in
- automatically categorizing the retrieved web documents.

As Figure 3.8 shows, the different parts of the user interface are labelled. In Figure 3.8, the user expresses his/her needs (Label a1). In addition, in Figure 3.8, the user can select the source language he/she would like to use with the help of the language resource to explore the context of the query (Label a1). The user has the possibility to interact with the tool in modifying the query context by selecting different linguistic relations i.e., Hypernym or Hyponym (Label e). In order to conduct a cross-lingual search, the user can select the target language (Label d). The tool then automatically provides translations of all possible source language senses in the target language. This translation is performed, based on the interlingual entries of EuroWordNet. After the translation is performed, the tool retrieves the number of relevant documents. The number of documents is then presented to the user in a visualization manner (circle visualization, which shows the distribution of document hits of the translations). The larger the number of retrieved documents is, the bigger the circle is (Label c). The tool automatically searches for all combinations between all senses including synonyms. With a mouse click, the user can display the relevant documents to the selected translation on the tool interface (Label f), based on the displayed "circles". The user also has the possibility to change the search context (Label c1). For example, with a right mouse click, the user can select a new word (given by the linguistic relation) and replace it with the originally searched word.

For example, the original query was (haus tür), with a right mouse click the user can select a new term (gebäude). In this case, the tool reacts by automatically repeating the same process which was done for the original query. This will involve translation, disambiguation and the visualization of the searched terms. Furthermore, another important aspect in the MultiLexExplorer, is that the user is given the possibility of removing any term/terms that are not of interest (Label c). In addition, the user can select any desired term/terms as expansion term/terms to the original query. These expanded term/terms will be presented along with the original query terms in (Label a2). As shown in Figure 3.8, the tool provides the user with different categorization techniques to categorize the huge search results for better navigation (Label g).



search results for: "maison" "porte" (376000 hits)

Lapeyre la maison - Vente en ligne : acheter **porte** entrée

KMeansClustering: 0: la - 2005 - fondat [1.0];

acheter **porte** entrée : Tout Lapeyre sur Internet : 45.000 références en ligne, stocks des magasins, des idées déco, des conseils et des solutions pour ...

http://www.lapeyre.fr/achat_en_ligne/acheter_porte_entree.htm

Figure 3.8: MultiexEXplorer interface.

In MultiLexExplorer, very useful aspect was taken into account. The information is expressed in a visually attractive manner, which makes the user's task easier. For example, in the retrieved document hits, the user does not need to check numbers, instead he/she just checks the "circle" (the bigger the circle, the greater the retrieved document hits are) that expresses the retrieved document hits.

3.3 Conclusion and Discussion

We studied in detail the state-of-the art cross-lingual retrieval interaction tools that can be used to support the user to perform his/her cross-lingual search. A comparison is made between the state-of-the art cross-lingual tools and the proposed cross-lingual tool in this thesis. The proposed tool aims to compensate for any potential deficits in the state-of-the art cross-lingual tools. More details, in how this is tackled, are discussed in Chapter 8.

Table 3.1 shows an abstract view of some of the important features that are needed to support the user in the cross-lingual task. Table 3.1 contains "Translation supported by" to clarify which translation approach is used by each tool, "Translation confidence" to clarify whether a cross-lingual tool gives the user confidence in the translation, "Translation improvement" to clarify whether a cross-lingual tool provides the user with the possibility of improving the translation, "User support" to clarify if a cross-lingual tool gives support to the user in all stages of the cross-lingual process e.g., will the user be notified about any tool failure, is the information displayed in a visual way etc., and "New language adaptations" clarify if it is possible to adapt a cross-lingual tool to handle more languages.

All of the previously mentioned tools consider the user as an integral part of the retrieval process, in that the user can plays an essential role in improving the search. One notices that there are some insufficiencies in supporting the user when he/she wants to retrieve documents written in a language which differs from the language he/she speaks. A possible reason for this deficit is that the user is requested to perform the translation disambiguation process.

For example, using Keizai, MULINEX or UCLIR, the user is requested to check all translation alternatives for each query term with the dictionary definition, in order to select the correct translation. However, the disambiguation process needs full concentration from the user, in that the user has to scroll up all translation alternatives in order for her/him to select relevant expanded terms. This can be very laborious especially for query terms that have abundant possible translations e.g., based on the given example in MULINEX, the user has to very large number of back translation alternatives in order to select the appropriate translations out of them.

In addition, the previously mentioned tools rely on the use of a bilingual dictionary or WordNet for translation as well as for disambiguation. However, bilingual dictionaries or WordNet in which the definitions of source language are available for each translation for the target languages are very rare and very laborious. Despite the good visual and functional design of MultiLexExplorer, it relies on the use of EuroWordNet, which only employs a limited number of languages. Furthermore, no automatic translation is integrated into the tool instead the user has to check many word sense combinations. We believe this review cross-lingual tools in this thesis, represents the most comprehensive review of cross-lingual tools in supporting the user seeking information in languages they are not familiar with.

Tools	**Properties**				
	Translation supported by	**Translation confidence**	**Translation improvement**	**User support**	**New language adaptations**
Keizai	Dictionary-based	no	no	partially [1]	not specified
UCLIR	Dictionary-based	no	no	partially[1]	not specified
MULINEX	Dictionary-based	yes[2]	no	no	not specified
Multi Searcher[3]	Hybrid approach[4] & Spelling correction	yes[5]	yes[6]	yes[7]	yes
MultiLexExplorer	Ontology-based	no	yes	yes[8]	no[9]
MIRCULE	Dictionary&MT[10] based	no	yes[11]	yes	yes
WORDS	Dictionary&MT[10] based	yes[2]	yes[12]	yes	not specified
LIC2M	Dictionary&MT[10] based	no	no	no	not specified
Patent CLIR	MT-based[10]	no	yes[13]	no	not specified

Table 3.1: Overview of some main properties of cross-lingual tools.

1. Support after the translation is performed "the retrieved documents together with small images are represented, which are called Document Thumbnail Visualizations.

2. Using back translation however, when no synonyms can be found in the dictionary, the technique is not helpful; and significant homonymy in the target language can result in confusing back translations.

3. Tool proposed in this thesis.

4. The automatic translation (Dictionary and Corpora based approaches) done by the tool itself and provides a possibility to refine the translation.

5. Provide the user with contextual information that describes the translation in a language the user is familiar with.

6. Done through the current translation process by suggesting interactive terms related to the user query.

7. Tool provides intensive user support e.g., user notification if no translation from dictionary was found, no available statistical co-occurrence data available to perform the automatic translation or a classified representation of the contextual information terms are represented e.g., query terms are in bold black and its synonym are in grey etc.

8. Information is expressed in a visually attractive manner, which makes the user's task easier and expands the original query with extra relevant terms.

9. Relies on using EuroWordnet which employs only a limited number of languages.

10. Using machine translation give no possibility to refine the translation.

11. Done by initiating new translation process after examining the search result.

12. Depend on adding new translation from the one proposed by the system. However, one term to one term translation employs high ambiguity.

13. Users can modify translations provided by three web translators. However, users with low knowledge in the target language will have no possibility to select suitable translations.

Part II

Query Pre-and-Post Processing

Chapter 4

Pre-processing: Spelling Correction

4.1 Introduction

Before the cross-lingual retrieval system can perform its task, the user query need to be pre-processed. This pre-processing step is useful to correct any spelling errors and to transform the a word to its basic form. The stemming of the user query terms is very important because the dictionary does not include all word forms instead just the root form. For stemming (e.g., for Arabic) we used the araMorph package based Buckwalter Arabic morphological analyzer (Buckwalter, 2002). However, if the target language is high inflectional language, in order to improve the performance of the retrieval process, all translation form variations need to be detected and included in the query (see Section 2.2.1.2 and Chapter 5).

People are using Internet search engines to retrieve information from the web. However the user misspelled query terms can lead to poor search results. Based on search logs investigation, Cucerzan and Brill (2004) claimed that around 10%-15% queries were misspelled. Before the cross-lingual retrieval system can perform its task there is an urgent need to correct the user misspelled query terms. In this thesis, we address this problem by developing a language-independent spell-checker that is based on an enhancement of the n-gram model. The spell checker is able to detect the correction suggestions by assigning weights to possible list of correction candidates based on n-gram statistics and lexical resources. We compared the results of our algorithm with state-of-the-art approaches and show that we provide very useful corrections, reaching better results than the other methods.

The algorithm we propose in the following is a language-independent spell-checker that is based on an enhancement of the n-gram model (Ahmed et al., 2007, 2009b). It is able to detect the correction suggestions by assigning weights to possible list of correction candidates based on n-gram statistics and lexical resources in order to detect the non-word errors and to derive correction candidates.

4.2 Revised n-gram Based Approach (MultiSpell)

Yannakoudakis and Fawthrop (1983) found that in most cases the first letter in the misspelled word is almost always correct and also the misspelled and real word will be either the same length or the length differ just by one. For some examples we like to refer the reader to the list of commonly misspelled words in English[1]. Furthermore, the pure n-gram based approach to compute the similarity coefficient as described above (see Eq. (4.1)) does not consider the order of the n-grams in the target word (Khaltar et al., 2006). This increases the probability that the matching score between two strings will be higher even though they do not share the same concept. Therefore, we revised the computation of the similarity between words to take these two aspects into account.

For simplicity, we describe our algorithm for $n = 2$ (bigrams). However, the approach can be applied for trigrams and n-grams with $n > 3$ as well. We define bigrams of words by their respective position in the word $w_{i,i+(n-1)}$ where i defines the position of the first letter and $i + (n - 1)$ the position of the last letter of the considered n-gram. Thus, the last possible position of an n-gram, in a word, is defined by $j = |w| - n + 1$ where $|w|$ defines the length of the word. In order to deal with the first and second aspect mentioned above, we define a window of n-grams of the target candidate words that should be compared, i.e., while in Eq. (4.1) all n-grams are compared with each other, we only compare n-grams that are in close proximity to the position of the n-gram in the word to be compared when computing the similarity score.

An example is given in Figure 4.1, where $\acute{w}$ defines the misspelled word and w a correction candidate. Here, the n-gram $\acute{w}$ of $\acute{w}_{4,5}$ will only be compared to the n-grams $w_{3,4}$, $w_{4,5}$ and $w_{5,6}$ of the correction candidate w, i.e., even if the n-gram $\acute{w}_{4,5}$ is similar to $w_{2,3}$ this would not count towards the similarity score of the words $\acute{w}$ and w.

Overall, the computation of the similarity score S for a given n-gram size n and a given odd-numbered window size m can be defined as follows. Assuming that u is the longer word (if v is longer than u then u and v can be simply exchanged):

[1] http://simple.wikipedia.org/wiki/Commonly_misspelled_words

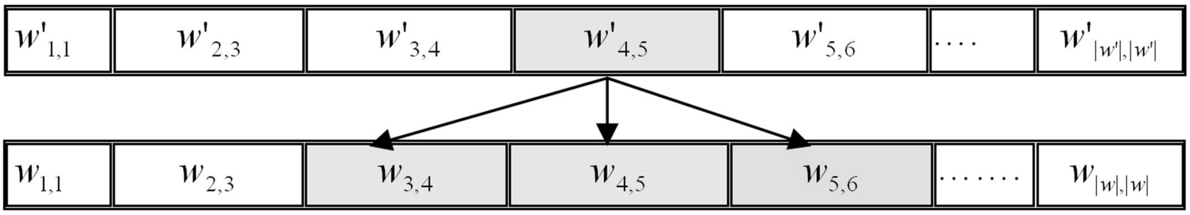


Figure 4.1: Bigram comparison for misspelled word $\acute{w}$ and a correction candidate w using a comparison window of size 3. Remark that the first and last n-gram represent the first and the last letters only and are therefore always of size one.

$$S_{n,m}(u,v) = \frac{g(u_{1,1}, v_{1,1}) + g(u_{|u|,|u|}, v_{|v|,|v|}) + \sum_{i=2}^{|u|-n+1} \sum_{j=\frac{m-1}{2}}^{\frac{m-1}{2}} g(u_{i,i+(n-1)}, v_{i+j,i+j+(n-1)})}{N} \tag{4.1}$$

$$\text{where } g(a,b) = \begin{cases} 1 & if \;\; a = b \\ 0 & \text{otherwise.} \end{cases} \text{ and } u_{i,j} = \begin{cases} substring(u,i,j) & if \;\; i \leq j \\ ,,, & \text{otherwise.} \end{cases}$$

Here, u and v are the words to be compared, the nested sum counts the number of n-grams in v that are similar to n-grams in a window the size of m around the same position in word v. N is computed similarly as in Eq. (4.1).

In Figure 4.2 the specific cases that have to be considered when computing the similarity score S are summarized.

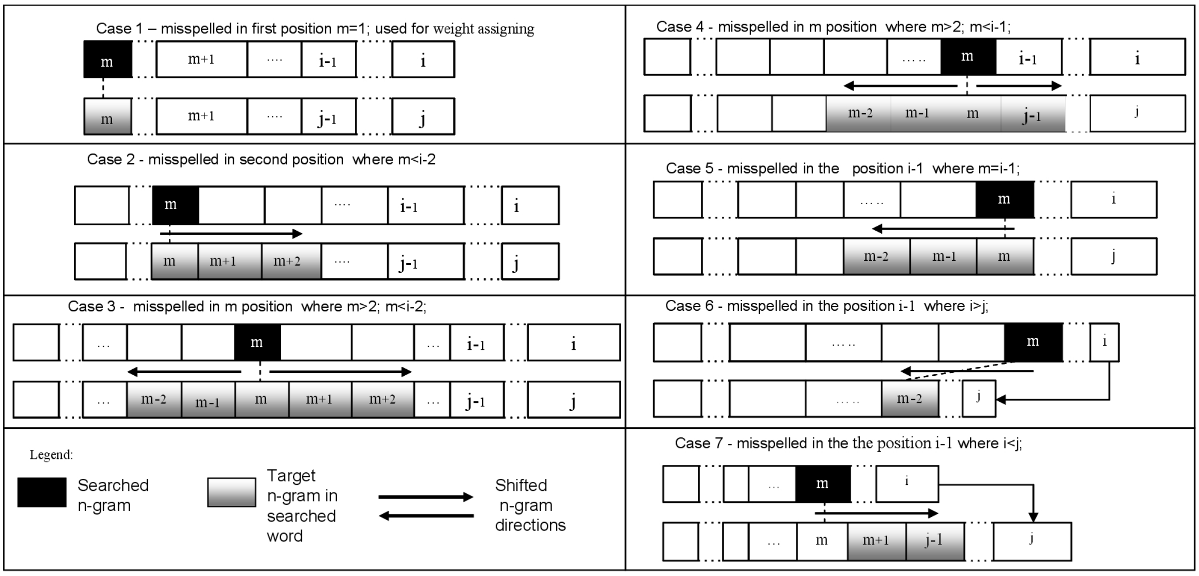

Case 1 – misspelled in first position m=1; used for weight assigning
m
m+1
....
i-1
i
m
m+1
....
j-1
j
Case 2 - misspelled in second position where m<i-2
m
....
i-1
i
m
m+1
m+2
....
j-1
j
Case 3 - misspelled in m position where m>2; m<i-2;
...
m
...
i-1
i
...
m-2
m-1
m
m+1
m+2
...
j-1
j
Legend:
Searched n-gram
Target n-gram in searched word
Shifted n-gram directions
Case 4 - misspelled in m position where m>2; m<i-1;
.....
m
i-1
i
m-2
m-1
m
j-1
j
Case 5 - misspelled in the position i-1 where m=i-1;
.....
m
i
m-2
m-1
m
j
Case 6 - misspelled in the position i-1 where i>j;
.....
m
i
.....
m-2
j
Case 7 - misspelled in the the position i-1 where i<j;
...
m
i
...
m
m+1
j-1
j

Figure 4.2: Comparing n-grams based on the MultiSpell algorithm (Ahmed et al., 2009b).

4.2.1 The MultiSpell Algorithm

The first stage of the MultiSpell algorithm is to compare the keywords given from the user with the correct words contained in the dictionary (Ahmed et al., 2007, 2009b). The list of words we used in the evaluation were extracted from MultiWordNet[2]. First of all, we check based on the used dictionary (here, based on the words extracted from MultiWordNet) if the word is misspelled. If this is the case, the algorithm builds n-grams for the misspelled word. Then we select correction candidates from the dictionary. In order to keep the number of correction candidates as small as possible we select only words as candidates that are two charters shorter or longer than the misspelled word. This is motivated by the work of Turba (1982), who has shown that most misspelled words differ in length only by one character from the correct word. For the selected words the n-grams are constructed and the similarity score is computed according to Eq. (4.1). The correction candidates can then be simply sorted by the obtained similarity score and the word with the highest score is proposed as best correction candidate.

In Section 4.3, we show results of the spelling correction experiments done for the English and Portuguese language. The first evaluation was done on a list of English common misspelled words [3]. Afterwards, we compared the results of our spell checker MultiSpell with the results of the TST approach (in one experiment, for the Portuguese language) and of the Aspell approach (in two experiments, for the Portuguese and the English language), showing that the proposed approach achieved always the best results (Ahmed et al., 2007, 2009b).

4.3 Evaluations for Different Languages in a Spelling Correction Task: Pre-Processing

The goal of the evaluation is to evaluate the query pre-processing approaches that we need as pre-processing step before the user query translation. For the spelling correction task, the evaluation was done on the whole list of commonly misspelled English words found in Wikipedia[4]. Afterwards, we compared the results of our spell checker MultiSpell with the results of the ternary search trees (TST) approach (in one experiment, for the Portuguese language) and of the Aspell approach (in two experiments, for the Portuguese and the English language), showing that the proposed approach always achieved the best

[2] www.multiwordnet.fbk.eu/english/home.php

[3] http://simple.wikipedia.org/wiki/Commonly_misspelled_words

[4] http://simple.wikipedia.org/wiki/Commonly_misspelled_words

results. For the first evaluation, we used the whole list of commonly misspelled words in English consisting of 3,975 words as published in Wikipedia. This list of common spelling mistakes is represented by a table consisting of two columns. The first one shows the misspelledword, the second the correct spelling. For the evaluations, we only considered the correction words that were ranked as best correction word, i.e., even if the second word would have been the correct candidate, this was counted as a wrong correction (Ahmed et al., 2007, 2009b).

4.3.1 Evaluations between Bigram and Trigram for English

For the first evaluation, we used the whole list of commonly misspelled words in English consisting of 3,975 words as published in Wikipedia. We first used all misspelled words of the list, using the bigram case and just the first candidate correction. Multi-Spell corrected 3,334 misspelled words (84%) and failed for 641 misspelled words (16%) although it provided similar corrections in many cases. For example the word "advice" was suggested instead of "advised" for the misspelledword "adviced". Another example is the provided correction "algebraically" instead of "algebraic" for the misspelled word "algebraical" (see Table A.1 in the Appendix). These suggestions were classified as wrong in our approach, even though they belong to the same word sense. Second, we used trigrams, this showed lower performance and efficiency. MultiSpell corrected 2,900 words (73%) and failed for 1,075 (27%) as shown in Table 4.1 and Figure 4.3.

	bigram	**trigram**
correct	3334 (84)%	2900 (73%)
wrong	641 (16%)	1075 (27%)

Table 4.1: Comparison between bigram and trigram in whole English data set (3,975 words).

4.3.2 Evaluation of English Spelling Correction

For this evaluation, we randomly selected a set of only 120 misspelled words obtained from Wikipedia and not the whole list. All error types and starting letters of the words were taken into account. We compared MultiSpell with Aspell, MicrosoftWord, and Google. Since Aspell provides a list of candidate corrections we took just the first candidate from the list assuming that the first candidate is the most likely one proposed

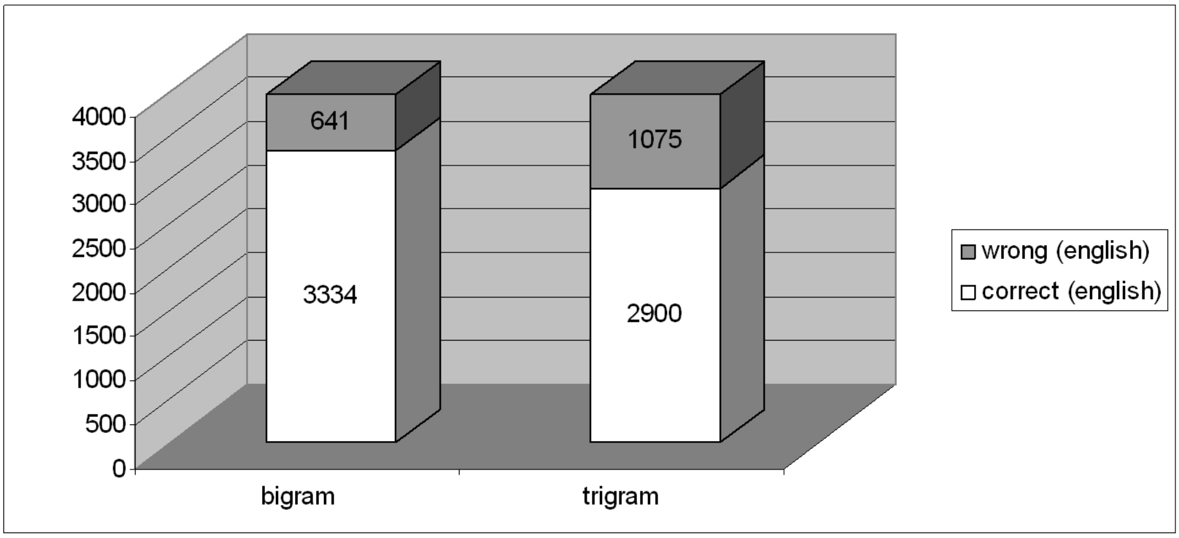


Figure 4.3: Comparison between bigram and trigram in whole English data set (3,975 words).

by the algorithm. MicrosoftWord and Google provided only one correction candidate. Table 4.2, Figure 4.4 and Table A.1 (in the Appendix) show that MultiSpell finds the correct spelling for 110 words (91.7%). In comparison, Google could correct 106 (88.3%) words, while Aspell and MicrosoftWord could correct 105 words (87.5%). MultiSpell detected 6 of 16 of the multiple correction words (which have more than one possible correction), but it doesn't fail to provide at least one correct suggestion. Aspell detected just two of the multiple corrections and it failed just one time to provide a suggestion for one of the multiple corrections.

	MultiSpell	**Aspell**	**MicrosoftWord**	**Google**
correct	110 (91.7%)	105 (87.5%)	105 (87.5%)	106 (88.3%)
wrong	10 (8.3%)	15 (12.5%)	15 (12.5%)%	14 (11.7%)

Table 4.2: Comparison of MultiSpell, Aspell, MicrosoftWord, and Google for English.

4.3.3 Evaluation of Portuguese Spelling Correction

The last evaluation was done for the Portuguese language. Martins and Silva (2004) implemented an algorithm using ternary search trees. The authors show experiments in correcting a list of some Portuguese words and comparing their results with Aspell. Here we compared MultiSpell on the whole list (120 Portuguese words) available from their experiments explained in (Martins and Silva, 2004), applying our algorithm and

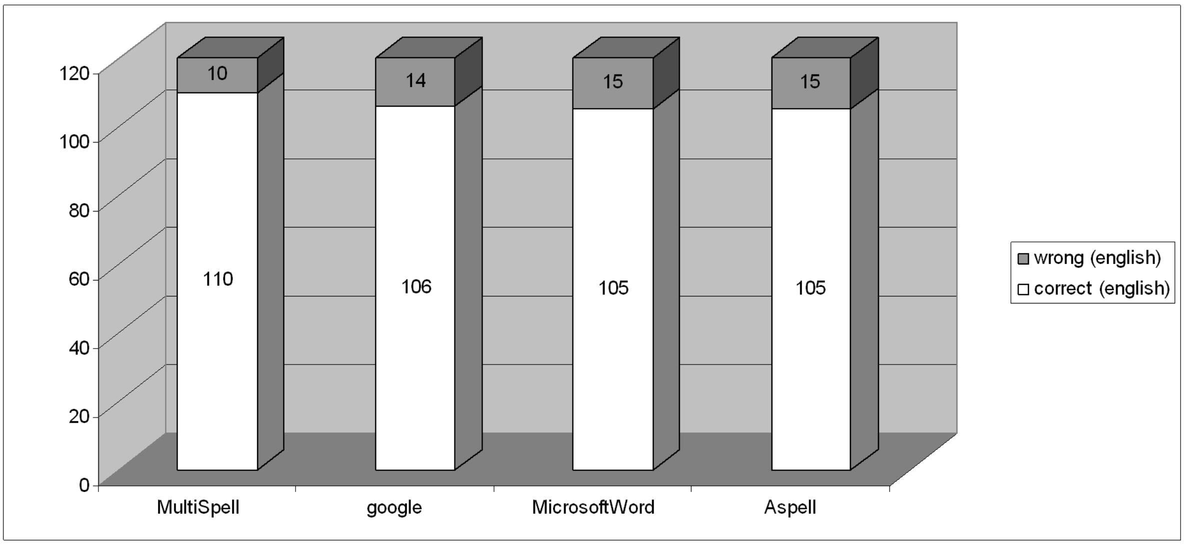


Figure 4.4: Comparison of MultiSpell, Aspell, MicrosoftWord, and Google for English.

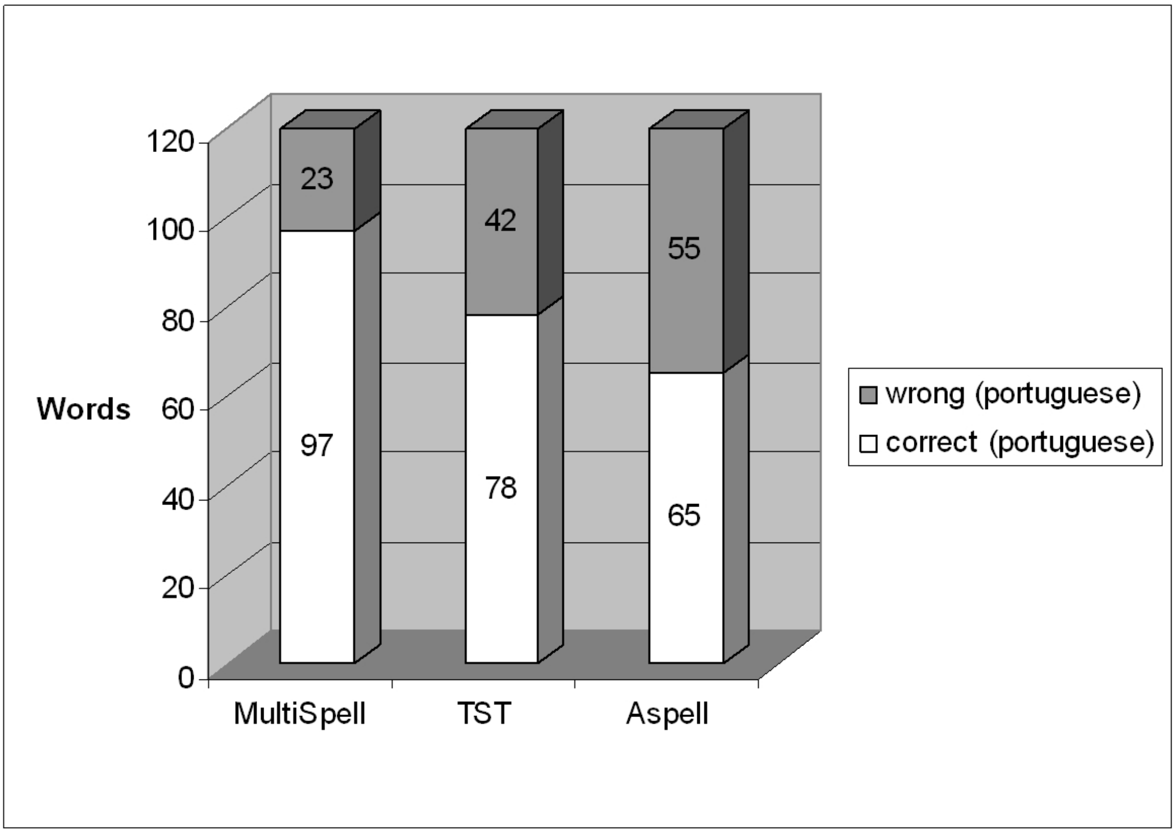


Figure 4.5: Comparison of MultiSpell, Aspell, and TST for the Portuguese language.

comparing it with the Aspell and TST algorithm. Given that MultiWordNet does not provide any Portuguese word senses, we used the dictionary made available from Martins and Silva (2004). Our algorithm succeeded in correcting 97 misspelled words (80.8%), TST succeeded in correcting 78 misspelled words (65%), and Aspell succeeded in correcting 65 misspelled words (54%) as shown in Table 4.3, Figure 4.5 and Table A.2 (in the Appendix).

	MultiSpell	**TST**	**Aspell**
correct	97 (80.8%)	78 (65%)	65 (54%)
wrong	23 (19.2%)	42 (35%)	55 (46%)

Table 4.3: Comparison of MultiSpell, Aspell, and TST for the Portuguese language.

4.4 Conclusion

In this chapter, we proposed a language-independent spell-checker that is based on an enhancement of a pure n-gram based model. Furthermore, we presented evaluations on English and Portuguese benchmark data sets of misspelled words. The obtained results outperformed other state-of-the-art methods.

Chapter 5

Post-processing: Word Inflection

5.1 Introduction

As described in Chapter 2.2.1.2 problem need to be tackled before the query can be retrieved is the variations in word form. The characteristics of highly inflectional languages result very often in a poor information retrieval performance. As a result, current search engines suffer from serious performance with the direct query-term-to-text-word matching for these languages, thus search engines need to be able to distinguish different variants of the same word. Detecting all word form variations in the query, which, processed by search engines, is considered essential for achieving good retrieval results and the alternative is the loss of vast amounts of information.

In the following we describe (as an example for Arabic language case) in detail the post-processing step (word inflection conflation), that need to be performed after the query is translated. For the evaluations, we implemented the n-gram model (using n=2,3) and their enhancement and edit distance conflations approaches (see Chapter 4.3 and Chapter 5.6).

5.2 Conflation Approach based on Revised n-gram

Arabic nouns and verbs are heavily prefixed and suffixed as described in the first section. As a result of that, it is possible to have words with different lengths that share same principal concept. Therefore, there is a need to conflate all words that refer to the same concept. Conflation is a general term for all processes of merging together nonidentical words which refer to the same principal concept i.e., to merge words which belong to same meaning class. The primary goal of conflation is to allow matching of different variants of the same word (Ahmed and Nürnberger, 2007, 2009).

Based on our previous work (Ahmed et al., 2007, 2009b) (see Chapter 4) where we applied a revised n-gram approach (Multispell) for spelling error corrections, we propose here a modified version for the conflation task. For example, there is no need for the conflation task to include the finding of Yannakoudakis and Fawthrop (1983) that refer to the fact that in most cases the first letter in the misspelled word is almost always correct and also the misspelled and real word will be either the same length or the length differ just by one. Therfore, the first part of the Eq. (4.1) can be removed (see Eq. 5.1).

$$_{n,m}(u,v) = \frac{\sum_{i=1}^{|u|-n+1} \sum_{j=\frac{m-1}{2}}^{\frac{m-1}{2}} g(u_{i,i+(n-1)}, v_{i+j,i+j+(n-1)})}{N} \tag{5.1}$$

For example, based on the fact that the average of the Arabic prefix length is 3, the compared n-grams window size can be defined. Figure 5.1 show the comparision between two words "استمرارestmrār " (Continued) and "الاستمراريةālāstmrāriyh " (the Continuation) whos differ in prefix and suffix.

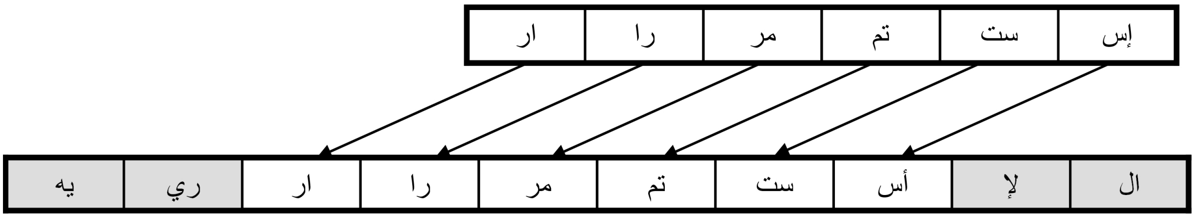


Figure 5.1: Bigram similarity measure between 2 words with different lengths.

In order to clarify how the comparision between two words is done an example is given in Figure 5.2, where $\acute{w}$ defines the given word "متسلسلةmotasalselh " (Serialized) and w a target candidate "سلسلةselslh " (Series), in case we don't find the n-gram $\acute{w}_{3,4}$ of $\acute{w}$ in the proper location the algorithm will shift the search to the right side in specific locations, so the n-gram $\acute{w}_{3,4}$ will be compared first with the n-grams $w_{3,4}$, then $w_{2,3}$ or $w_{1,2}$ of the target candidate w, in case w greater than $\acute{w}$ then the search will shift to left side.

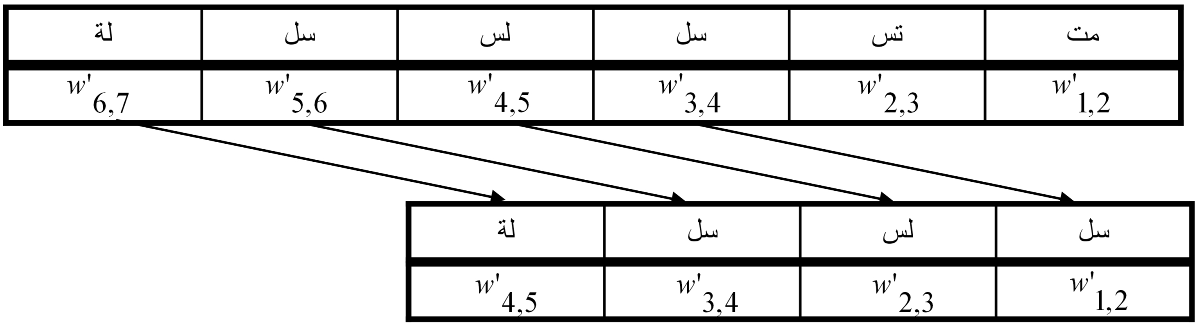


Figure 5.2: Words with different word lengths that belong to same meaning class.

As it is shown in Figure 5.3, the revised n-gram approach improve the accuracy of the string matching since it take into account the order of the n-grams. Using the pure n-gram approach the similarity measure between the Arabic word "التحالفاتaltḥālfāt " (the Alliances) and "الفاتحalfātḥ " (the Conqueror or the Light) is 85.72% although the two words have different meaning (see Figure 5.3 left). In other hand using the revised

n-gram approach where the order of the n-grams are taken into account, the similarity measure between the giving words is 28.57% (see Figure 5.3 right).

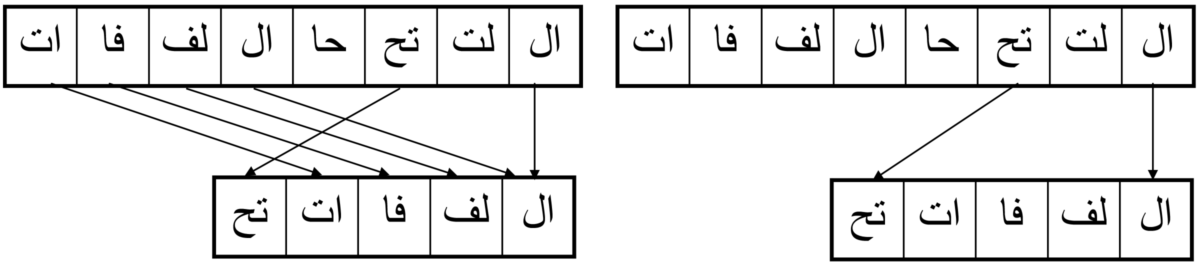


Figure 5.3: Pure bigram (left) and revised bigram (right).

5.3 Conflation Process Improvement (Web Statistics Approach)

In order to detect and eliminate conflation terms that are created by the n-gram approach, but that are most likely not relevant for the query ("noisy terms"), we propose here an approach based on Mutual Information (MI) scores computed based on web statistical co-occurrences data (Ahmed and Nürnberger, 2011). The n-gram based approach assumes two strings are alike based only on a string similarity comparison: the more n-grams existing between two strings, the more similar they are. However, there are many words that have a very similar text pattern but a quite different meaning. Therefore, we improved our n-gram approach by eliminating such noisy terms that could have been generated. This is done by computing the cohesion score between all revised n-gram generated expanded terms using the mutual information measure. The term/terms that have a lower MI score than the MI score mean for all expanded terms can be considered as noisy term/terms and thus will be eliminated.

5.3.1 Mutual Information (MI)

Given a query, the set of possible expanded terms using the revised n-gram will be generated; the coherence between the expanded terms is computed based on mutual information (MI). Given a query term $q_i = \{t_1, t_2, ..., t_n\}$ and a set of its revised n-gram model generated expanded terms $\{ext_{i,1}, ext_{i,2}, ..., ext_{i,m_i}\}$, where m_i defines the number of extended terms for t_i and $1 \leq i \leq n$. Given the set of $\frac{n(n-1)}{S}$ combinations, where S is the size of each combinations set, then the set of combinations between all expanded terms is defined as $Com_i = \{\{ext_{i,j}, ext_{i,k}\} | 1 \leq j < n, j < k \leq n\}$. The mutual information of each combination set can be computed based on the following equation:

$$MI(q_{t_1}, q_{t_2}) = log_2 \frac{p(q_{t_i}, q_{t_j})}{p(q_{t_i})p(q_{t_j})} \tag{5.2}$$

where $p(q_{t_i}, q_{t_j})$ being the joint probability of both expanded terms in the combination sets to occur in web. The probability is estimated by the relative frequency of the expanded terms in a given corpus, here the web, i.e., it is estimated by how many times q_{t_i}, q_{t_j} occur together in a (web) document.

Expanded Terms Combinations	MI Score
(وصحيفةwṣḥyfh ,للصّحيفةllṣḥyfh) "and Newspaper, for the Newspaper"	28.651
(للصّحيفةllṣḥyfh ,لصحيفةlṣḥyfh) "for the Newspaper, for a Newspaper"	28.075
(بصحيفةbṣḥyfh ,لصحيفةlṣḥyfh) "by Newspaper, for a Newspaper"	27.054
(وصحيفةwṣḥyfh ,لصحيفةlṣḥyfh) "and Newspaper, for a Newspaper"	27.047
(بصحيفةbṣḥyfh ,وصحيفةwṣḥyfh) "by Newspaper,and Newspaper"	26.486
(بصحيفةbṣḥyfh ,للصّحيفةllṣḥyfh) "by Newspaper, for the Newspaper"	25.186
(للصّحيفةllṣḥyfh ,نحيفةnḥyfh) "for the Newspaper, slim"	23.793
(بصحيفةbṣḥyfh ,نحيفةnḥyfh) "by Newspaper, slim"	23.790
(وصحيفةwṣḥyfh ,نحيفةnḥyfh) "and Newspaper, slim"	23.165
(نحيفةnḥyfh ,لصحيفةlṣḥyfh) "slim, for a Newspaper"	21.314
The MI score mean	25.456

Table 5.1: Expanded term combinations and their MI scores.

Expanded Terms	MI average Score
(للصّحيفةllṣḥyfh) "for the Newspaper"	26.421
(وصحيفةwṣḥyfh) "and Newspaper"	26.337
(لصحيفةlṣḥyfh) "for a Newspaper"	25.872
(بصحيفةbṣḥyfh) "by Newspaper"	25.629
(نحيفةnḥyfh) "slim"	23.015

Table 5.2: Expanded terms and their average MI scores.

5.3.2 A Walk Through Example

To illustrate the improvement of the revised n-gram algorithm using the statistical co-occurrences data obtained from web, let us consider the following example. The user query صحيفةṣḥyfh (Newspaper), the system using the revised n-gram model with similarity threshold of 60% expanded the user query with the following terms: (بصحيفةbṣḥyfh "by Newspaper", وصحيفةwṣḥyfh "and Newspaper" , للصّحيفةllṣḥyfh "for the Newspaper" , نحيفةnḥyfh ("slim" Feminine) and لصحيفةlṣḥyfh "for a Newspaper"). The algorithm starts by generating all possible combinations between the expanded terms where $Com_i = \{\{ext_{i,j}, ext_{i,k}\}|1 \leq j < 5, j < k \leq 5\}$. After generating all possible combinations between the expansion terms, the mutual information score for each expansion term combination will be calculated based on Eq. (5.2). Table 5.1 illustrates possible expanded term combinations and their mutual information score. As shown in Table 5.1, one of the expanded term combinations included the expanded term نحيفةnḥyfh "slim". It has the lowest mutual scores (23.793, 23.790,23,165 and 21.314). As shown in Table 5.2, the same expanded term has the lowest MI average score (23.015), which is below the MI score mean (25.456) that we defined as threshold based on prior experiments, and thus will be classified by the proposed approach as a noisy term and will be eliminated. In contrast, all other expanded terms have an average mutual score, which is above the MI score mean and thus should be correct expanded terms for the user's query. In Section 5.6, we show results of the experiments done in the conflation task. In our experiments we compared our approach with the Edit distance, pure n-gram approach for bigrams and trigrams.

5.4 araSearch: A Meta-Searcher Enhanced by Query Post-Processing

Based on the encouraging results that we achieved in previous work e.g., (Ahmed et al., 2007; Ahmed and Nürnberger, 2007; Ahmed et al., 2009b), we developed a user adaptive interface called *araSearch* (Ahmed and Nürnberger, 2008c). araSearch is a metasearcher that serves as an interface to standard search engines. We currently provide, for example, access to Google using the Google Web Services API. araSearch is based on an n-gram based similarity feature that is able to account for textual variation in Arabic. araSearch works as a "guide" in the sense that it helps users to issue their queries. araSearch offers an intuitive visual overview of the user extended query in order to allow the user to verify the query terms and select the desired ones. araSearch was designed to be a language-independent system that is able to handle other languages besides Arabic. Only minor modifications are needed in order to handle other languages. There is no need to adapt any one of the module codes. The only change required is to import a new lexicon for the target language.

In order to start using the system, the user must first access the araSearch Web site. This site was developed using jsp and java servlets and is based on the Tomcat server, which runs the programs responding to the user requests and returns the dynamic results to the user's browser. Figure 5.4 illustrates the general overview of the interaction between the user, the system, and a search engine, in this case Google. If the user sends his/her query to araSearch, it extends the query and forwards the extended query to Google, fetching the results and then displaying them to the user on the araSearch interface.

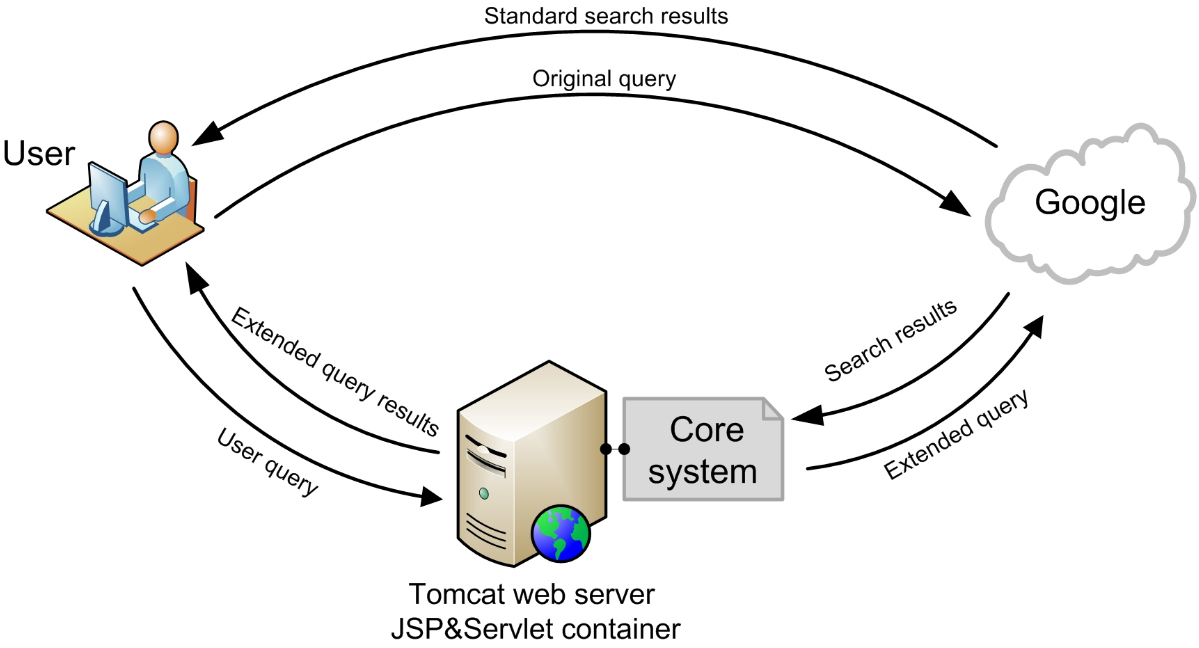


Figure 5.4: General overview of the interaction between the user, araSearch, and Google.

5.4.1 araSearch System Architecture

The following section outline the architecture of the araSearch Framework. Using the *Natural Language Query Interface (NLQI)*, the user types his/her query. Stop words will be eliminated from the query before passing it to the next module *Spelling Correction Module (SCM).* The spelling correction process is followed by the *Query Processing Module (QPM).* Using the query processing module, the system receives the query and transforms it by selecting suitable terms from lexical data, and then the transformed query is passed to the *Result Presentation Module (RPM)*, which is responsible for the graphical representation of the reformulated query results that have been received from the search engine to the user. Figure 5.5 illustrates the general view of the araSearch architecture.

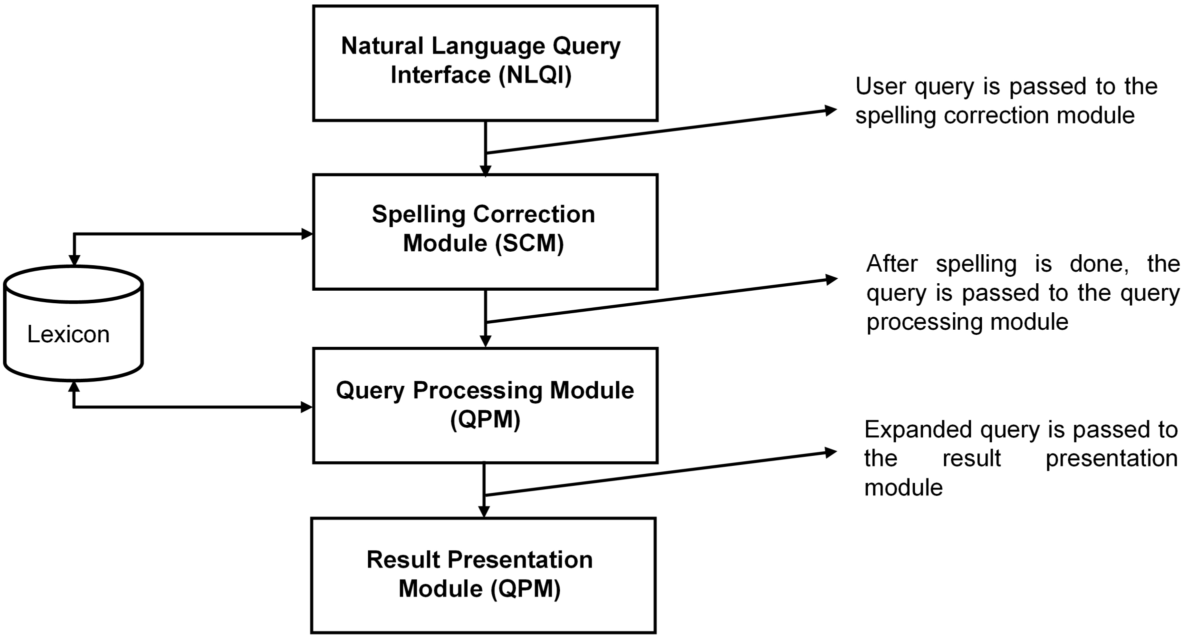

Natural Language Query Interface (NLQI)
User query is passed to the spelling correction module
Spelling Correction Module (SCM)
Lexicon
After spelling is done, the query is passed to the query processing module
Query Processing Module (QPM)
Expanded query is passed to the result presentation module
Result Presentation Module (QPM)

Figure 5.5: General overview of araSearch architecture.

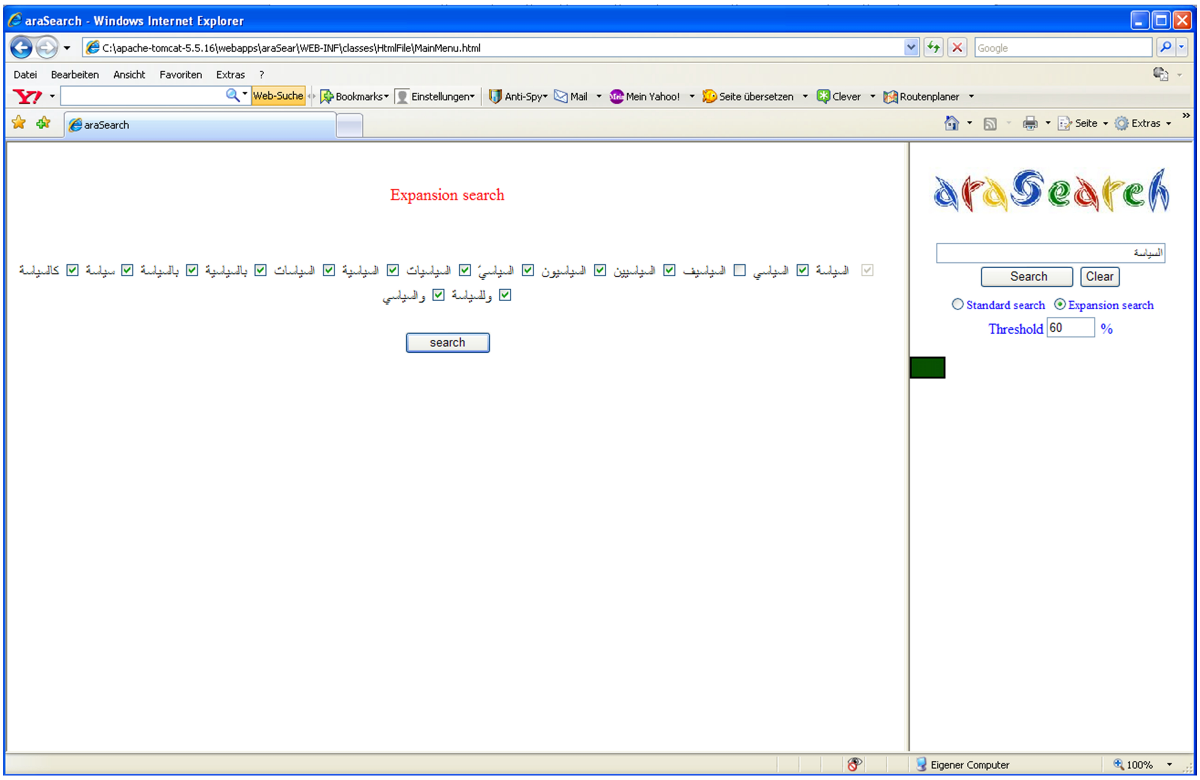
5.4.2 araSearch Modules Tasks

The Natural Language Query Interface Module (NLQI) allows the user to type the query in the natural language, which is then submitted to the next modules. The Natural Query Interface (NLQI) is an intermediate level of access to the system between all modules. The natural language query interface presents the reformulated query that is processed by QPM in a visual manner to the user which allows the user to adapt his/her needs by adjusting the threshold or by a simple mouse click, deselecting unsuitable additional query terms. During the system run time, the system improves the retrieval performance through the Spelling Correction Module (SCM), which is responsible for identifying spelling errors in query terms. The MultiSpell approach was used from the SCM to do this task. Multispell was developed as a language independent spell checker that is based on an enhancement of the n-gram model. The spell checker is able to detect the correction suggestions by assigning weights to a possible list of correction candidates based on n-gram statistics and lexical resources. For a more conclusive overview of MultiSpell see Chapter 4. The Query Processing Module (QPM) is the core of the system where all word form variations are constructed. It is responsible for converting the query terms to an extended representation. When the user submits the query, the QPM executes the query and starts the extended procedure. As the n-gram constructions and similarity coefficient calculations explained in Chapter 4.2, the algorithm starts to construct the n-grams for the query term, then computes the similarity between the query term n-grams array with each of the word dictionary n-grams array. The similarity coefficient is then calculated and compared with the threshold that the user submitted with the query. In case the similarity coefficient is greater than the threshold, the dictionary word will be suggested as a possible variant of the user's query. Figure 5.6 shows an example of the extended query-terms interface. The user submits the query (politics) with a threshold of 60%, and the system suggests possible additional query terms to the user's query. With a simple mouse click, users have the ability to deselect any one of the additional terms that don't satisfy their need. In order to display the reformulated query results in the araSearch system, Result Presentation Module (RPM) was implemented. It is responsible for the graphical representation of the reformulated query results that have been received from the search engine to the user. Figure 5.7 and 5.8 illustrates the result of the query with direct search and with the query-reformulated search. As shown in Figure 5.7, 2,320,000 relevant documents were retrieved while in Figure 5.8, 4,150,000 documents were retrieved.

In the following, we discuss the edit distance string similarity technique that we used for evaluation.

5.5 The Levenshtein Distance Techniques

The Levenshtein distance, also known as the edit distance, is a technique that is used to measure the similarity between two strings (Levenshtein, 1966). Wagner and Fischer (1974) describe an algorithm to calculate the edit distance that makes use of a technique called dynamic programming. The algorithm dynamically reuses already computed values of the edit distance so that the required number of computations can be decreased;


araSearch - Windows Internet Explorer
C:\apache-tomcat-5.5.16\webapps\araSear\WEB-INF\classes\HtmlFile\MainMenu.html
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
Web-Suche Bookmarks Einstellungen Anti-Spy Mail Mein Yahoo! Seite übersetzen Clever Routenplaner
araSearch
Seite Extras
araSearch
Search
Clear
Eigener Computer
100%

Expansion search

☑ السياسة ☑ السياسي ☐ السياسيف ☑ السياسيين ☑ السياسيون ☑ السياسيّ ☑ السياسيات ☑ السياسية ☑ السياسات ☑ بالسياسية ☑ بالسياسة ☑ سياسة ☑ كالسياسة
☑ وللسياسة ☑ والسياسي

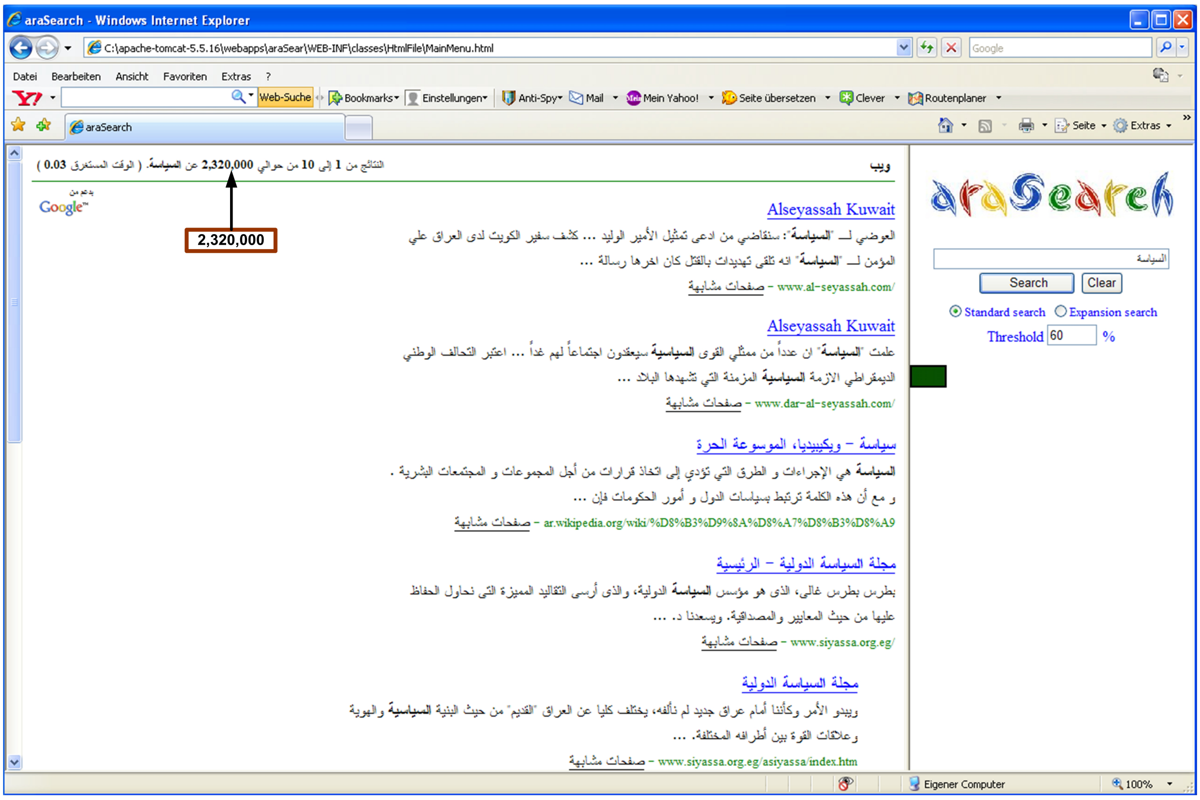
search

السياسة

○ Standard search ● Expansion search

Threshold 60 %

Figure 5.6: Relevant extended query terms.


araSearch - Windows Internet Explorer
C:\apache-tomcat-5.5.16\webapps\araSear\WEB-INF\classes\HtmlFile\MainMenu.html
Google
Datei Bearbeiten Ansicht Favoriten Extras ?
Web-Suche Bookmarks Einstellungen Anti-Spy Mail Mein Yahoo! Seite übersetzen Clever Routenplaner
araSearch
Seite Extras
بدعم من
Google
2,320,000
araSearch
Search
Clear
Eigener Computer
100%

ويب

النتائج من **1** إلى **10** من حوالي **2,320,000** عن **السياسة**. (الوقت المستغرق **0.03**)

Alseyassah Kuwait
العوضي لـــ "**السياسة**": سنقاضي من ادعى تمثيل الأمير الوليد ... كشف سفير الكويت لدى العراق علي
المؤمن لـــ "**السياسة**" انه تلقى تهديدات بالقتل كان اخرها رسالة ...
www.al-seyassah.com/ - صفحات مشابهة

Alseyassah Kuwait
علمت "**السياسة**" ان عدداً من ممثلي القوى **السياسية** سيعقدون اجتماعاً لهم غداً ... اعتبر التحالف الوطني
الديمقراطي الازمة **السياسية** المزمنة التي تشهدها البلاد ...
www.dar-al-seyassah.com/ - صفحات مشابهة

سياسة - ويكيبيديا، الموسوعة الحرة
السياسة هي الإجراءات و الطرق التي تؤدي إلى اتخاذ قرارات من أجل المجموعات و المجتمعات البشرية .
و مع أن هذه الكلمة ترتبط بسياسات الدول و أمور الحكومات فإن ...
ar.wikipedia.org/wiki/%D8%B3%D9%8A%D8%A7%D8%B3%D8%A9 - صفحات مشابهة

مجلة السياسة الدولية - الرئيسية
بطرس بطرس غالى، الذى هو مؤسس **السياسة** الدولية، والذى أرسى التقاليد المميزة التى نحاول الحفاظ
عليها من حيث المعايير والمصداقية. ويسعدنا د. ...
www.siyassa.org.eg/ - صفحات مشابهة

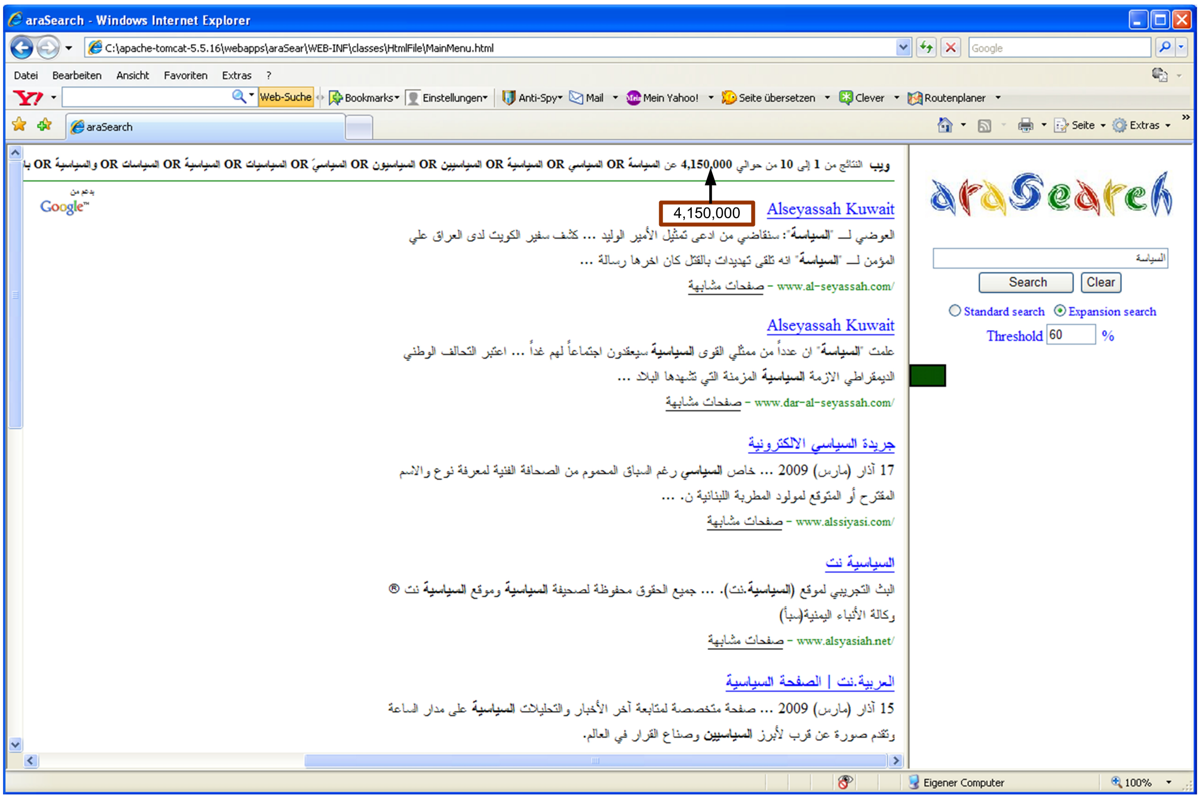
مجلة السياسة الدولية
ويبدو الأمر وكأننا أمام عراق جديد لم نألفه، يختلف كليا عن العراق "القديم" من حيث البنية **السياسية** والهوية
وعلاقات القوة بين أطرافه المختلفة. ...
www.siyassa.org.eg/asiyassa/index.htm - صفحات مشابهة

السياسة

● Standard search ○ Expansion search

Threshold 60 %

Figure 5.7: Documents retrieved with standard search.



ويب النتائج من 1 إلى 10 من حوالي 4,150,000 عن السياسة OR السياسي OR السياسية OR السياسيين OR السياسيون OR السياسيَ OR السياسيات OR السياسية OR السياسات OR والسياسية با

4,150,000

Alseyassah Kuwait
العوضي لـــ "السياسة": سنقاضي من ادعى تمثيل الأمير الوليد ... كشف سفير الكويت لدى العراق علي المؤمن لـــ "السياسة" انه تلقى تهديدات بالقتل كان اخرها رسالة ...
www.al-seyassah.com/ - صفحات مشابهة

Alseyassah Kuwait
علمت "السياسة" ان عدداً من ممثلي القوى السياسية سيعقدون اجتماعاً لهم غداً ... اعتبر التحالف الوطني الديمقراطي الازمة السياسية المزمنة التي تشهدها البلاد ...
www.dar-al-seyassah.com/ - صفحات مشابهة

جريدة السياسي الالكترونية
17 آذار (مارس) 2009 ... خاص السياسي رغم السباق المحموم من الصحافة الفنية لمعرفة نوع والاسم المقترح أو المتوقع لمولود المطربة اللبنانية ن. ...
www.alssiyasi.com/ - صفحات مشابهة

السياسية نت
البث التجريبي لموقع (السياسية.نت). ... جميع الحقوق محفوظة لصحيفة السياسية وموقع السياسية نت ® وكالة الأنباء اليمنية(سبأ)
www.alsyasiah.net/ - صفحات مشابهة

العربية.نت | الصفحة السياسية
15 آذار (مارس) 2009 ... صفحة متخصصة لمتابعة آخر الأخبار والتحليلات السياسية على مدار الساعة وتقدم صورة عن قرب لأبرز السياسيين وصناع القرار في العالم.

السياسة

Search Clear

○ Standard search ● Expansion search

Threshold 60 %

Figure 5.8: Documents retrieved after query reformulation.

thus the performance (speed) of the algorithm is improved.

The Levenshtein distance is defined as the minimal number of edit operations (insertions, substitutions, and deletions) that are necessary to transform one string into another. In other words, the two considered strings are aligned, using these transformations. More formally, given two strings s_1 and s_2, an alignment A of these strings is a sequence $(a_1 \rightarrow b_1), (a_2 \rightarrow b_2), \cdots, (a_n \rightarrow b_n)$ of edit operations where $s_1 = a_1, \cdots, a_n$ and $s_2 = b_1, \cdots, b_n$. To each edit operation a weight function δ is assigned. For each $a = b$ the weight function $\delta(a \rightarrow b) = 0$ and if $a \neq b$ the weight function $\delta(a \rightarrow b) = 1$. For example, letting I denote the insert operation, D denote the delete operation, R the substitute (or replace) operation, and M the nonoperation of "match," only one operation is needed to transform the first string مساعدmsāʿd (helper) to the second string تساعدtsāʿd (She helps). The alignment operations of the two string is represented in Figure 5.9. To calculate the number of operations needed to transform the first string into the other, we have to add up the costs of all edit operations applied. $\lambda \rightarrow$ تt denotes the operation that has to be carried out. In this case the operation is a substitution, and has a cost of $\delta(\lambda \rightarrow$ ت$t)$ = 1 The cost for all other edit operations where $a = b$ is $\delta(a \rightarrow b) = 0$

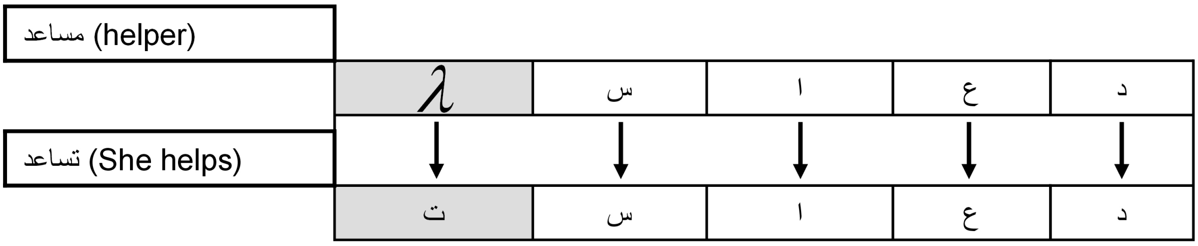


Figure 5.9: Operation needed to convert مساعدmsāʿd (helper) to تساعدtsāʿd (She helps).

5.6 Evaluation of Conflation Approaches: Post-Processing

The goal of the evaluation is to evaluate the query Post-processing approaches which we need to use as Post-processing step after the user query translation. In our experiments we compared our approach with the Edit distance, pure *n*-gram approach for bigrams and trigrams (Ahmed and Nürnberger, 2007, 2009, 2011). The reason for not taking a larger value for *n* is the problem of eliminating short words.

For example, when trying to retrieve the query "يقرyqer " (Acknowledges) using trigrams, the relevant result "قرqr " (Acknowledged) will be eliminated because no *n*-grams can be constructed for it as it is less than 3 characters long. The targets words must be at least one character longer than the size of *n* in order to have the chance to be retrieved. For this reason, we used n=2 in the proposed approach to enable retrieval of short words, as well as other words lengths Furthermore, we used the revised *n*-gram model to avoid ambiguity as described in Chapter 4.2.

5.6.1 Data Selection

To collect test data for our evaluations, we crawled the Web for articles published on one popular Arabic newsWeb site (CNN-Arabic[1]) in the period from January 2002 to March 2007. We obtained 5792 Arabic documents, all of which are abstracts of articles on

[1] http://arabic.cnn.com/

politics, sports, art, economy, and information science (size 60 MB). More than 1,400,000 Arabic words were extracted with 101,210 unique words. These articles are supposed to be correctly written and have both a large and rich vocabulary and therefore offer more investigation points in terms of the number of word variations. The approaches were evaluated against 500 queries that were formulated randomly, ensuring that the length of the query terms vary and short as well as long query terms are included. In order to construct the random queries, the algorithm requires the availability of a lexicon of terms that were extracted from the test data.

Techniques	**Precision %**
Revised bigram	91.3
Pure bigram	79.4
Revised trigram	98.7
Pure trigram	95.7
Edit distance	87.3

Table 5.3: Average precision for all approaches.

5.6.2 Comparison of Conflation Approaches

In the first experiment, based on the giving data set for a similarity threshold of 60%, we calculated the average precision for conflation approaches based on the revised and pure n-gram model (using n=2,3) and edit distance. As shown in Table 5.3, the results are quite similar. The reason for this is that only 6.5% out of 500 query words had a length of less than 3 characters, which is the length that affects the ambiguity. The revised bigrams and trigrams showed improvement over edit distance and the pure bigrams and trigrams due to the reduction in ambiguity. In order to provide a more detailed analysis, we also calculated the average precision for the pure trigram and the revised bigram for the similarity thresholds of 60, 65, 70, 75, 80, 85, 90, and 95%. Tables 5.4, 5.5 and 5.6 show the comparison of retrieved, relevant, irrelevant, and average precision between the revised bigram, pure bigram, and pure trigram approaches. The revised bigram achieved clear improvement over the pure trigram and pure bigram. The reason is that is that the revised bigram approach takes into account all word lengths, which will increase the retrieved performance. On the other hand, it takes into account the order of the n-gram, which will decrease the pure n-gram ambiguity results. This results in decreasing the number of irrelevant documents retrieved (Ahmed and Nürnberger, 2007, 2009).

Revised bigram				
Threshold	**Retrieved**	**Relevant**	**Irrelevant**	**Precision**
60 %	5992	5472	520	91.3 %
65 %	4367	4196	171	96.1 %
70 %	2960	2882	78	97.3 %
75 %	2464	2393	71	97.1 %
80 %	1817	1803	14	99.2 %
85 %	694	694	0	100 %
90 %	518	518	0	100 %
95 %	518	518	0	100 %
Average Precision				**97.6 %**

Table 5.4: Average precision of revised bigram model for different threshold on 500 words queries.

Pure bigram				
Threshold	**Retrieved**	**Relevant**	**Irrelevant**	**Precision**
60 %	6890	5472	1418	79.4 %
65 %	5200	4196	1004	80.6 %
70 %	3560	2882	678	80.9 %
75 %	2722	2393	329	87.9 %
80 %	2010	1803	207	89.7 %
85 %	744	694	50	93.2 %
90 %	552	518	34	93.8%
95 %	537	518	19	96.4%
Average Precision				**87.7 %**

Table 5.5: Average precision of revised bigram model for different threshold on 500 words queries.

The trigram approach retrieved better results in terms of the ratio of relevant documents retrieved to the (total) documents retrieved. The revised bigram approach achieved better results in terms of how many relevant documents were retrieved compared to the total number of documents retrieved (relevant and irrelevant). For example, when a threshold of 60% is selected, the revised bigram retrieved 5472 relevant documents and 520 irrelevant ones, while the pure trigram retrieved 4253 relevant documents

Pure trigram				
Threshold	**Retrieved**	**Relevant**	**Irrelevant**	**Precision**
60 %	4442	4253	189	95.7 %
65 %	3086	2969	117	96.2 %
70 %	2075	2045	30	98.5 %
75 %	1872	1843	29	98.4 %
80 %	1015	1007	8	99.2 %
85 %	549	549	0	100 %
90 %	549	549	0	100 %
95 %	549	549	0	100 %
Average Precision				**98.5 %**

Table 5.6: Average precision of pure trigram model for different thresholds on 500 words queries.

and 189 irrelevant ones. Compared with pure bigram the revised bigram decreases the number of irrelevant documents retrieved, and in so doing, gains a higher precision. The pure trigram approach retrieved fewer irrelevant documents at the expense of the total number of relevant documents retrieved, while the revised bigram retrieved fewer irrelevant documents compared to the total number of relevant documents retrieved. Figure 5.10 compares the three approaches with respect to (a) average precision, (b) total documents retrieved, (c) relevant documents retrieved, and (d) irrelevant documents retrieved. It is important to notice, when interpreting Figure 5.10 (c), one needs to consider the significant difference between the relevant documents retrieved from each method for different thresholds. Figure 5.10 (a) shows that the revised bigram gains higher precision compared to pure bigram. As shown in Tables 5.7 the performance of the revised n-gram approach is better than that of the pure n-gram approach in terms of the total number of relevant documents retrieved. Table B.1 in the Appendix provides a typical example, where the revised bigram model retrieved 33 relevant documents, while the pure trigram model retrieved 25 relevant documents. Figure 5.10 (a) illustrates that although with a threshold of 85% both approaches have maximum precision, the revised bigram approach performs better than the pure trigram in terms of the number of relevant documents retrieved. Although both pure and revised bigram have the same number of relevant documents retrieved, the revised bigram approach performs better than the pure bigram in terms of the number of irrelevant documents retrieved. Figure 5.10 (a) shows that the revised bigram approach gained clearly higher precision compared with the pure bigram. In the second experiment we estimated the average recall and F-measure for a sample of 30 queries out of 500. The query terms were selected in the same way as described above. Figure 5.12 illustrates that the revised bigram approach gained a higher average recall than the pure trigram approach, since it took into account different word lengths and similarity enhancement. As shown in Table 5.7 the revised bigram approach gained a higher F-measure of up to 85% compared to the pure trigram, pure bigram, and edit distance approaches. These results showthat the revised n-gram has gained an overall higher degree of retrieval performance than the pure n-gram and edit distance approaches. Table B.2, B.3, B.4 and B.5 in the Appendix shows a detailed example (three queries) how we perform the conflation process using bigram.

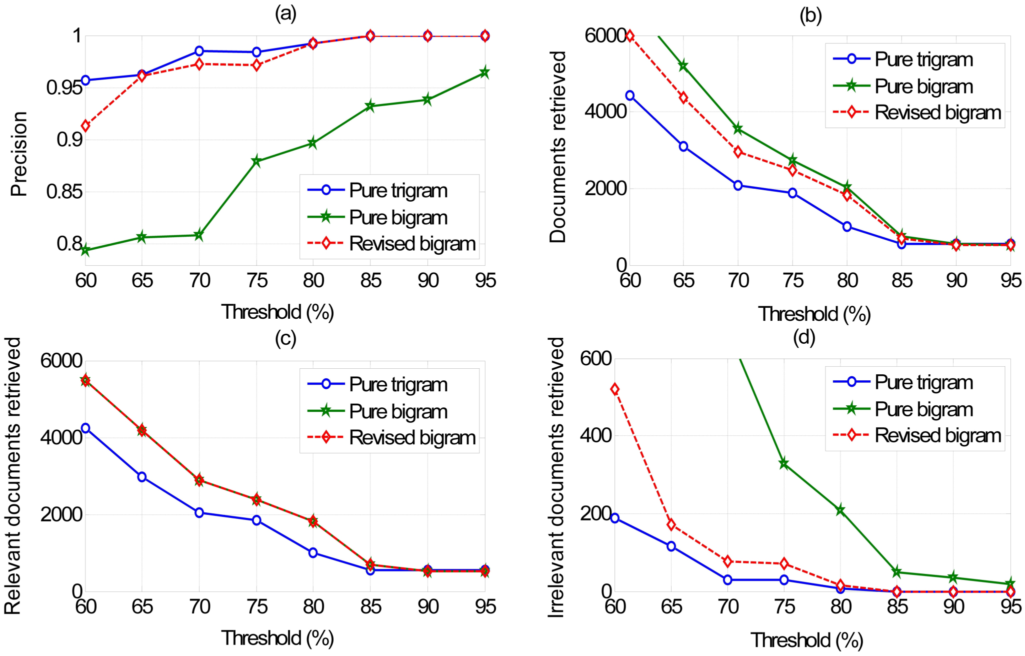


Figure 5.10: (a) Average precision. (b) Total documents retrieved. (c) Relevant documents retrieved. (d) Irrelevant documents retrieved.

5.6.3 Conflation Process Improvement (Web Statistics Approach) Evaluation

In the first evaluation, we conducted the same precision experiment in Section 5.6 to evaluate if the web statistics approach improves the precision of the revised bigram approach. As table 5.8 shows, we calculated again the average precision (based on the randomly selected 500 queries) for the pure trigram, edit distance, revised bigram and (revised bigram + MI) for the similarity thresholds of 60, 65, 70, 75, 80, 85, 90, and

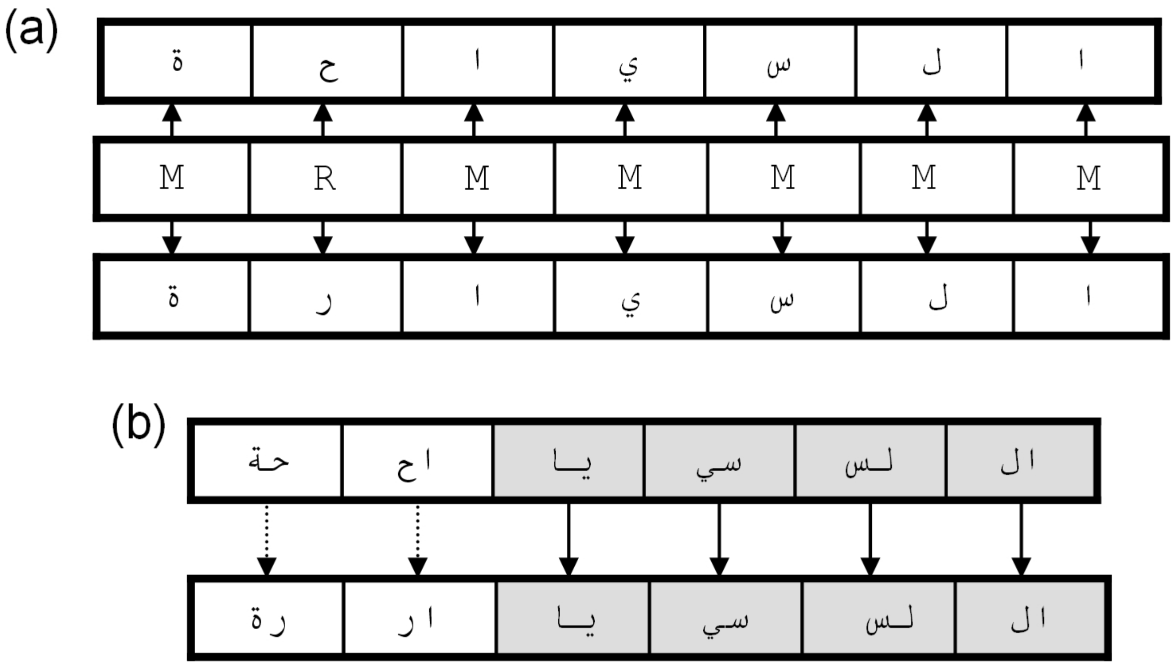


Figure 5.11: (a) One operation is needed to transform the first word into the second. I denotes the insert operation, R the substitute (or replace) operation, D the delete operation, and M the nonoperation of (or) "match". (b) Using the n-gram approach (with n=2) the similarity score is 66.66%.

Ret.	**Rel.**	**Irr.**	**Miss. Rel.**	**Precision**	**Recall**	**F-Measure**
Pure trigram						
366	360	6	374	98 %	49 %	65 %
Pure bigram						
629	539	90	195	86 %	73 %	80 %
Edit distance						
400	358	42	376	89 %	49 %	64 %
Revised bigram						
596	554	42	180	93 %	76 %	84 %
Revised-bigram + MI						
571	554	17	180	97 %	76 %	86 %

Table 5.7: Average recall, precision, and F-measure for the four approaches for a sample of 30 queries out of 500 (Ret.=Retrieved, Rel.=Relevant,Irr.=Irrelevant, Miss. Rel.=Missing Relevant).

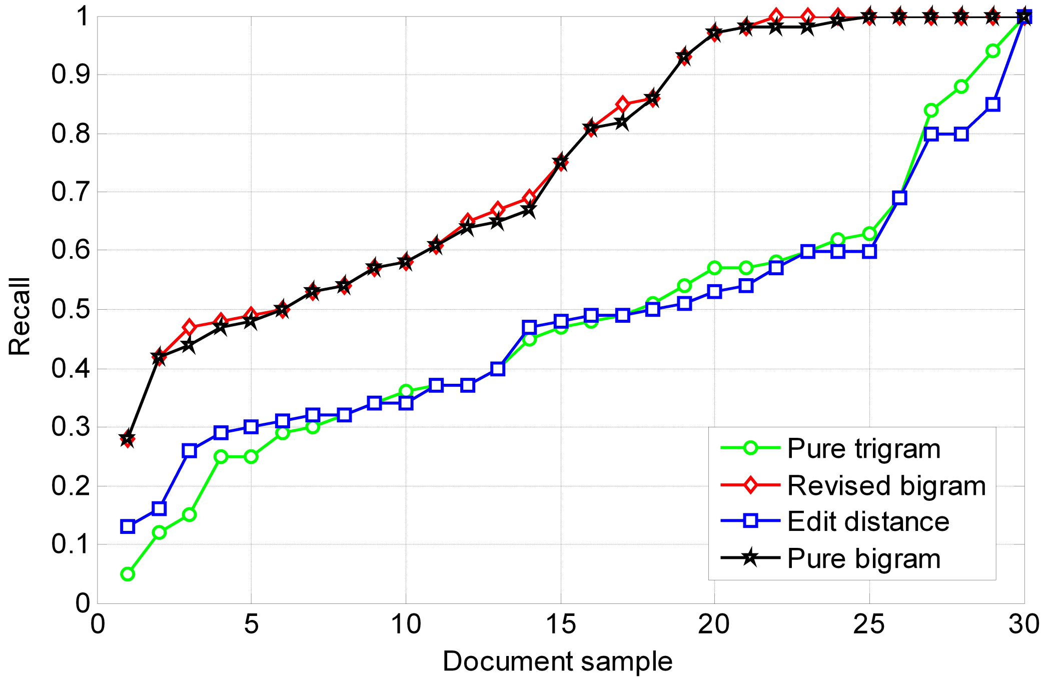


Figure 5.12: Average recall for revised bigram, pure bigram, edit distance, and pure trigram approaches (sorted by recall value).

95% (Table 5.8 shows the precision average). The trigram approaches (pure and revised) achieved higher precision than the revised bigram approach but in the same time it achieved lower recall than the revised bigram as it will be shown next in this section. The revised bigram precision was improved by 3.3% using mutual information approach based on statistical data obtained from web. In the second evaluation, we estimated the average recall and F-measure for a sample of 30 queries out of 500 (based on the experiment conducted in Section 5.6). We were interested to evaluate if the conflation approaches improvement based on web statistics data improves the precision of the revised bigram approach. We performed the web experiments using the mutual information approach to improve the precision of revised bigram approach. This was done by eliminating the bigram generated noisy expanded terms as discussed in Section 5.3. Table 5.9 and Figure 5.13 shows that the mutual information approach using statistical co-occurrence data obtained from the web succeeded in eliminating 25 irrelevant expanded terms generated by the revised bigram approach. The failed cases were counted when the algorithm failed to eliminate the noisy terms or when the algorithm eliminate a corrected expanded term/terms along with the noisy one.

For example, we consider the query افريقياāfryqyā "Africa", the algorithm succeeded

Techniques	**Precision %**
Revised bigram	91.3
Revised-bigram + MI	94.6
Pure bigram	79.4
Revised trigram	98.7
Pure trigram	95.7
Edit distance	87.3

Table 5.8: Average precision for all approaches.

in eliminating the noisy term فريقيfryqy "my team" or "two teams" but at the same time, it eliminated a relevant term بافريقياbāfryqyā "by Africa". One interpretation for this lack, is that the word فريقيfryqy "my team" or "two teams" with average MI scores (27.999) frequently appeared in the context of African sport and thus it increases the MI score mean (28.437) in that the average MI scores for the relevant word بافريقياbāfryqyā "by Africa" (27.708) is below the MI score mean.

	Pure-trigram	**Pure-bigram**	**Edit distance**	**Revised-bigram**	**Revised-bigram +** MI
Retrieved	366	629	400	596	571
Relevant	360	539	358	554	554
Irrelevant	6	90	42	42	17
Miss Relevant	6	195	376	180	180
Precision	0.98	0.86	0.89	0.93	0.97
Recall	0.49	0.73	0.49	0.76	0.76
F-Measure	0.65	0.80	0.64	0.84	0.86

Table 5.9: Average recall, precision, and F-measure for the five approaches for a sample of 30 queries out of 500.

5.7 Conclusion

We presented a language-independent conflation approach, i.e., an approach that does not depend on any predefined rules or previous knowledge of linguistic information about

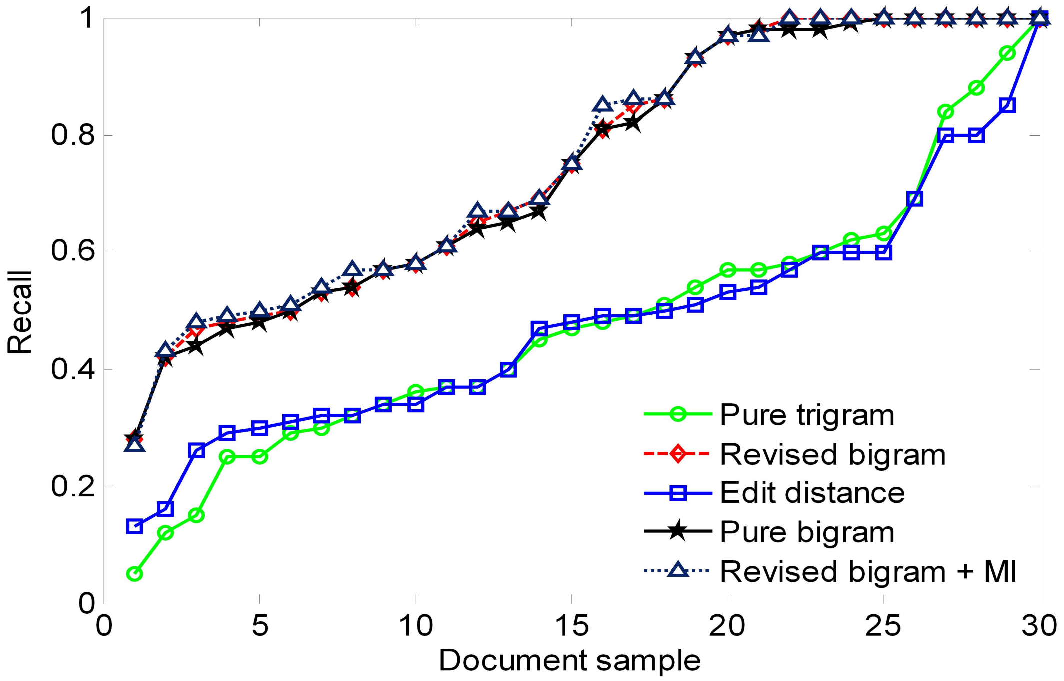


Figure 5.13: Average recall for pure trigram, edit distance, pure bigram, revised bigram and (revised bigram $+MI$) approaches (sorted by recall value).

the target language. Furthermore, we evaluated our approach on the Arabic language, which is one of most inflected languages in the world. The experimental results indicate that the selection of the n-gram size affects the retrieval performance, i.e., the number of relevant and irrelevant documents retrieved. Using a large n-gram size leads to the result that most of the documents retrieved are relevant but at the expense of missing many relevant documents, since the selection of a large n will eliminate short words from consideration. On the other hand, selecting a small value for n leads to the result that, though many relevant documents are retrieved, many irrelevant documents are retrieved at the same time due to the ambiguity that results from the small size of the n-grams. Therefore, we proposed a revised approach to compare the similarity of words based on n-grams that take the order of the n-grams into account. Based on the experimental results we show that the revised bigram approach provided better results compared to pure trigrams as well as n-grams with $n > 3$. Furthermore, we demonstrated that the enhancement of the n-gram model provided very good results in terms of conflation for heavily inflected languages such as Arabic. In addition, the proposed algorithm was evaluated based on 500 randomly selected queries. The quantitative and qualitative experimental results show that our algorithm achieved better results than pure n-gram approaches. Consequently, the proposed algorithm helps to achieve a higher degree of accuracy overall, in the conflation task. In order to deal with n-gram noisy expanded terms, a mutual information approach applied to statistical co-occurrences data obtained from web was developed, in that the terms that have less cohesion score with other will be assumed as noisy terms and thus will be eliminated. The eliminations of the n-gram noisy generated terms improved the precision of the revised n-gram with 3.3%. The failed cases by the algorithm can be interrelated by the lack of the training data or by the very generic term usage where terms can appear in different contexts.

In addition, an adaptive user interface called araSearch is proposed. araSearch is used to help the user to extend a query in order to improve the search by adding relevant word-form variations. araSearch serves as an interface to the standard search engines; it is based on an n-gram-based similarity feature that is able to account for textual variation with special attention to the Arabic language. araSearch offers a simple but intuitive visual overview of the user-extended query in order to allow the user to verify the query terms by selecting those that are suitable.

Part III

Query Translation and Disambiguation

Chapter 6

Algorithms for Query Translation and Disambiguation

The proposed cross-lingual tool in this thesis has been developed to overcome one of the main deficits in the state-of-the art cross-lingual tools, mainly where disambiguation is performed by the user (see Chapter 3.3). Usually user-based disambiguation does not encourage the user to use the cross-lingual system system and can result in frustration and loss of time.

The automatic translation, which is one of the important components in our cross-lingual tool, works independently, without any user effort. Usually, however, in order to refine the achieved automatic translations provided by the system, the user can be integrated in this process (see Chapter 8). We would like to emphasize, that in our proposed cross-lingual tool, the user task is reduced to a great extent, while in the state-of-the-art cross-lingual tools, the user is requested to check all possible query term translation alternatives with their dictionary definition, in order for him/her to disambiguate (Ahmed et al., 2011). This way of disambiguation results in the user losing time and being frustrated especially for query terms with abundant translations.

In our proposed cross-lingual tool, this task is softened to a great extent. The user query is automatically translated and thereafter a user takes over, only to refine and improve the automatic translation. The integrated automatic translation component is responsible for obtaining all query term translation alternatives, generating the translation combinations, and then the final step is to disambiguate and select the appropriate translation. This selection is based on the disambiguation score provided by a statistical approach integrated in the proposed cross-lingual tool e.g., Mutual Information or Naïve Bayesian classifier approaches.

6.1 Automatic Translation

In order to disambiguate the user query, in this thesis, first approach we use a word sense disambiguation method applied in automatic translation of a query from source to target language. The developed machine learning (Naïve Bayesian Classifier) approach is based on statistical models that can learn from parallel corpora by analysing the relations between the items included in these corpora in order to use them for selecting the most suitable translation of the query term. In order to resolve the translation ambiguity inherent in bilingual dictionaries, this hybrid approach can be used (Ahmed and Nürnberger, 2008a,b,d) (see Section 6.1.1).

Since obtaining a parallel corpora is not easy, in a second approach in order to disambiguate the user query, we use mutual information applied in monolingual corpora to calculate the cohesion scores for possible translation-candidate pairs to resolve the translation ambiguity (Ahmed and Nürnberger, 2010; Ahmed et al., 2009a; Ahmed, 2010; Ahmed et al., 2011). However, this approach is affected by the sparseness of translation combinations in the underlying corpora. One poorly distributed term can affect the whole cohesion scores obtained from the corpus and therefore in some cases only few - and thus unreliable - statistical co-occurrence data is available or in the worst case none at all. In order to obtain robust disambiguation methods, this data sparseness issue is researched and tackled in this thesis (see Section 6.1.2). The automatic translation method consists of two main steps (e.g., Arabic as source language): First, using an Arabic analyzer, the query terms are analyzed and the senses (possible translations) of the ambiguous query terms are identified. Second, the most likely correct senses of the ambiguous query terms are selected based on co-occurrence statistics.

6.1.1 Approach based on Naïve Bayesian Classifier (NB)

The proposed approach is based on exploiting parallel texts, in order to find the correct sense for the translated user query term (Ahmed and Nürnberger, 2008b,d). The minimum query length that the proposed approach accepts is two and the maximum query length is unlimited. Given the user query, the system begins by translating the query terms using the araMorph package[1]. In case the system suggests more than one translation (senses inventory) for each of the query terms, the system then starts the disambiguation process to select the correct sense for the translated query terms. The disambiguation process starts by exploiting the parallel corpus, in which the Arabic

[1] http://www.nongnu.org/aramorph/

version of the translation sentences matches fragments in the user query. A matched fragment must contain at least one word in the user query besides the ambiguous one. The words could be represented in the surface form or in one of its variant forms. Therefore, and to detect all word form variants in the translation sentences in the training corpus, special similarity score measures are applied (see Part II).

Bridging the Inflectional Morphology Gap

As motiviated in Chapter 2.2.1.2 languages exhibiting a rich inflectional morphology face a challenge for machine translation systems, as it is not possible to include all word form variants in the dictionaries. Inflected forms of words for those languages contain information that is not relevant for translation. The inflectional morphology difference between high inflectional language and poor inflectional language presents a number of issues for the translation system as well as for disambiguation algorithms. This inflection gap causes a matching challenge when translating between rich inflectional morphology and relatively poor inflectional morphology language. It is possible to have the word in one form in the source language, while having the same word in a few forms in the target language. This causes several issues for word translation disambiguation, e.g., where more unknown words forms exist in the training data and will not be recognized as being relevant to the searched words. Therefore, it is possible to have lower matching score for those words even though they have a high occurrence of them in the training data. To motivate the problem more clearly, we consider, for simplicity, the Arabic word دينdyn (religion or debt). As described in Chapter 2.2.3.1. The absence of the diacritics from the Arabic printed media or the Internet web sites causes high ambiguity. The Arabic word دينdyn has two translations in English (religion or debt). We calculate the occurrences of this word in the training corpus for both senses. This is done by searching for this word in the corpora and based on its context; we map it to the appropriate sense. As it is shown in Table 6.1 the word دينdyn was found in basic form for the sense (religion) 49 times and for the sense (Debt) only 10 times. As Table 6.2 shows, when we consider the inflectional form for the word دينdyn (religion or debt) we see that the occurrence of the inflectional form for the word دينdyn with the sense (religion) is 1146 and with the sense (debt) is 240. Table C.1 in the Appendix shows sentence examples from the training corpus where the ambiguous word دينdyn appears in basic or inflectional form with both senses. Detecting all word forms variants of the user query terms in the corpus is very essential when computing the score of the synonym sets, as it is shown in Table 6.2. More than 1386 sentences will be considered by the WSD algorithm

The ambiguous word	**Senses**	Co-occurrence/basic form
دينdyn	Religion	49
دينdyn	Debt	10
Total		**59**

Table 6.1: The occurrence of the ambiguous word دينdyn in the basic form for both senses.

The ambiguous word	**Senses**	Co-occurrence/Inflectional form
الدينāldyn	The Religion	75
والدينwāldyn	And the Religion	22
الآديانālʾādyān	The Religions	45
والآديانwālʾādyān	And the Religion	7
الدينيةāldynyh	The Religious	63
والدينيةwāldynyh	And the Religion	28
Total		**240**
الدينāldyn	The debt	860
والدينwāldyn	And the debt	22
الديونāldywn	The debts	255
والديونwāldywn	And the debts	9
Total		**1146**

Table 6.2: The occurrence of the ambiguous word دينdyn in the inflectional form for both senses.

to disambiguate the ambiguous word دينdyn .

The Naïve Bayesian Algorithm was first used for general classification problems. For WSD problems it had been used for the first time in (Gale et al., 1992c). The approach is based on the assumption that all features representing the problem are conditionally independent giving the value of classification variables. For a word sense disambiguation tasks, giving a word w , candidate classification variables $S = \{s_1, s_2, ..., s_n\}$, which represent the senses of the ambiguous word, and the feature $F = \{f_1, f_2, ..., f_n\}$ which describe the context in which an ambiguous word occurs, the Naïve Bayesian finds the proper sense s_i for the ambiguous word W by selecting the sense that maximizes the conditional probability of occurring in the given the context. In other words, NB constructs rules that achieve high discrimination level between occurrences of different word-senses by a probabilistic estimation. The Naïve Bayesian estimation for the proper sense can be defined as follows:

$$P(S_i \mid f_1, f_2, ..., f_n) = P(S_i) \prod_{j=o}^{m} P(f_i \mid S_i) \tag{6.1}$$

The sense s_i of an ambiguous word w_{amb} in the source language is defined by a synonym set (one or more of its translations) in the target language. The features for WSD, that are useful for identifying the correct sense of the ambiguous words, can be terms such as words or collocations of words. Features are extracted from the parallel corpus in the context of the ambiguous word. The conditional probabilities of the features $F = \{f_1, f_2, ..., f_n\}$ with observation of sense s_i, $P(f_i \mid S_i)$ and the probability of sense s_i, $P(s_i)$ are computed with $P(f_i \mid S_i) = \frac{C(f_i,S_i)}{C(s_i)}$ and $P(C_i) = \frac{C(s_i)}{N}$. $C(f_i, S_i)$ denoting the number of times feature f_i and sense s_i have been seen together in the training set. $C(s_i)$ denoting the number of occurrences of s_i in the training set. N is the total number of occurrences of the ambiguous word w_{amb} in the training dataset.

Feature Selection

The selection of an effective representation of the context (features) plays an essential role in WSD. The proposed approach is based on building different classifiers from different subsets of features and combinations of them. Those features are obtained from the user query terms (not counting the ambiguous terms), topic context and word inflectional form in the topic context and combinations of them. In our algorithm, query terms are represented as sets of features on which the learning algorithm is trained. Topic context is represented by a bag of surrounding words in a large context of the ambiguous word:

$$F = \{w_{w_{amb-k}}, ..., w_{w_{amb-2}}, w_{w_{amb-1}}, w_{amb}, w_{w_{amb+1}}, w_{w_{amb+2}}, ..., w_{w_{amb+k}}, q_1, q_2, ..., q_n\}$$

where k is the context size, w_{amb} is the ambiguous word and amb its position.

The ambiguous word and the words in the context can be replaced by their inflectional forms. These forms and their contexts can be used as additional features. Thus, we obtain $\acute{F}$ which contains in addition to the ambiguous word w_{amb} and its context the inflectional forms w_{inf} of the given sense and their context. As it is shown in Table 6.2 detecting all word form variants of the user query terms in the corpus will make 1386 sentences considered by the WSD algorithm to disambiguate the ambiguous word دينdyn . In addition, we count for each context word the number of occurrences of this word and all its inflectional forms, i.e.

$$\acute{F} = F \bigcup_{i=o}^{l} \{w_{w_{inf_i-k}}, ..., w_{w_{inf_i-2}}, w_{w_{inf_i-1}}, w_i nf, w_{w_{inf_i+1}}, w_{w_{inf_i+2}}, ..., w_{w_{inf_i+k}}\}$$

General Overview of the Query Translation Process

As Figure 6.1 shows, the system starts by processing the user query. The input is a natural language query Q. The query is then parsed into several words $q_1, q_2, q_3, ..., q_n$. Each word is then further processed independent of the other words. Since the dictionary does not contains all word forms of the translated word, only the root form, for each q_m in our query, we find its morphological root using the araMorph. After finding the morphological root of each term in the query, the query term is translated. In case the query term has more than one translation, the model provides a list of translations (sense inventory) for each of the ambiguous query terms. Based on the obtained sense inventory for the ambiguous query term, the disambiguation process can be initiated. The algorithm starts by computing the scores of the individual synonym sets. This is done by exploiting the parallel corpora in which the Arabic version of the translated sentences matches words or fragments of the user query, while matched words of the query must map to at least two words that are nearby in the corpus sentence. These words could be represented in the surface form or in one of its inflectional forms. In order to detect all word form variants in the translation sentences in the training corpora, special similarity score measures are applied. Since the Arabic version of the translation sentences in the bilingual corpora matches fragments in the user query, the score of the individual synonym sets can be computed based on the features that represent the context of the ambiguous word. As additional features, the words in the topic context can be replaced by their inflectional form. After we have determined the features, the score of each of the sense sets can be computed. The sense which matches the highest number of features will be considered as the correct sense of the ambiguous query term and then it is assumed to be the best sense that describes the meaning of the ambiguous query term in the context.

Illustrative Example

To consider how the algorithm performs the disambiguation steps, for simplicity we consider the following query with size 3 however the algorithm work for unlimited query

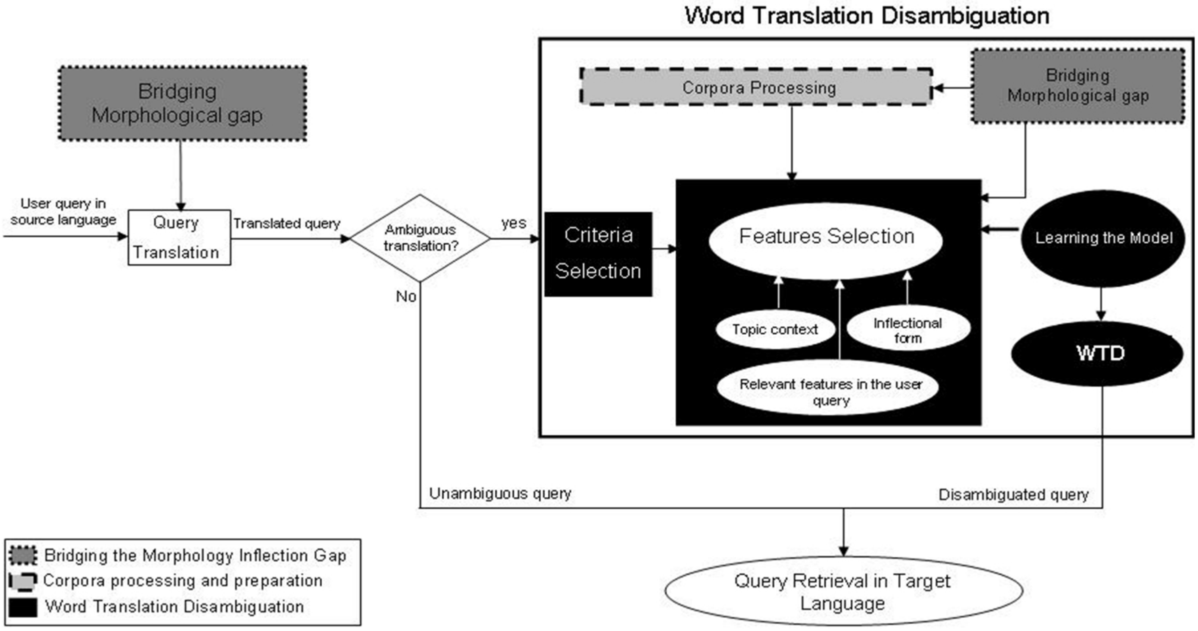

Word Translation Disambiguation
Bridging Morphological gap
User query in source language
Query Translation
Translated query
Ambiguous translation?
yes
No
Criteria Selection
Corpora Processing
Bridging Morphological gap
Features Selection
Topic context
Inflectional form
Relevant features in the user query
Learning the Model
WTD
Unambiguous query
Disambiguated query
Query Retrieval in Target Language
Bridging the Morphology Inflection Gap
Corpora processing and preparation
Word Translation Disambiguation

Figure 6.1: General overview of the query translation process (Ahmed and Nürnberger, 2008d).

size: رسم جمركى للسلع rsem ğmrkā lelslˤ (tax customs commodities):

- The natural language query Q is parsed into several words $q_1, q_2, q_3, ..., q_n$.

- For each q_m in the query, we find its morphological root, since the dictionary does not contain all word forms, the algorithm before translation will find the single form of each of the given query terms . For example, the Arabic word قرارهمqrārhm (their decision) which is not exist in the dictionary because it is not in the root form will be processed and converted to the basic form which is قرارqrār (decision).

- Translation of the query terms and creation of the sense inventory for each of the query term is done. Table 6.3 shows the sense inventory for each of the ambiguous query terms.

- The disambiguation process is initiated. The algorithm starts by computing the scores of the individual synonym sets (translation combinations):

- Number of times feature f_i and sense s_i which have been seen together in the training set is computed.
- Number of occurrences of s_i in the training set is computed.
- The total number N of occurrences of the ambiguous word w_{amb} in the training dataset is computed.
- The disambiguation score is computed and the sense which matches the highest number of features is considered as the correct sense of the ambiguous query term.

Table C.2 in the Appendix shows that there are 135 possible translations set for the giving query رسم جمركى للسلع rsem ǧmrkā lelslʿ . Furthermore, Table C.2 shows the disambiguation scores of the individual synonym sets for each ambiguous query terms with other query terms with 4934 occurrences of the ambiguous word w_{amb} in the training dataset.

Query Terms	**Sense inventory (Possible English Translations)**
رسمrsem	[fee, tax, drawing, sketch, illustration, prescribe, trace, sketch, indicate, appoint]
جمركىǧmrkā	[customs, tariff, customs, control]
للسلعlelslʿ	[crack, rift, commodities, commercial, goods]

Table 6.3: Sense inventory for each of the ambiguous query terms

6.1.2 Approach based on Mutual Information (MI)

Giving a source of data, Mutual Information (MI) is a measure to calculate the correlation between terms in specific space (corpora or web). MI approach has been frequently used in word sense disambiguation task e.g., (Fernandez-Amoros et al., 2010). The automatic translation process (Ahmed and Nürnberger, 2010) starts by translating each query term independently. This is done by obtaining a set of possible translations of each of the query terms from the dictionary. Based on the translation sets of each term, sets of all possible combinations between terms in the translation sets are generated. Using co-occurrence data extracted from monolingual corpora[2], the translations are then ranked based on a

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T07

cohesion score computed using Mutual Information: Given a query $q = \{q_1, q_1, ..., q_n\}$, and its translation set $S_{qk} = \{q_k, t_i\}$, where $1 \leq k \leq n, 1 \leq i \leq m_k$ and m_k is the number of translations for query term k. The MI score of each translation combination can be computed as follows:

$$MI(q_{t_i}, q_{t_j}) = log_2 \frac{p(q_{t_i}, q_{t_j})}{p(q_{t_i})p(q_{t_j})} \tag{6.2}$$

The probability $p(q_{t_i}, q_{t_j})$ is estimated by counting how many time each two terms, in the translation combination, appear together in corpora (see Table 6.6), e.g., how many time the term $p(q_{t_1})$ and the term $p(q_{t_2})$ co-occur together in the corpora.

The probabilities $p(q_{t_i})$ and $p(q_{t_j})$ are estimated by counting the number of individual occurrences of each possible translated query term in the corpora.

Illustrative Example

Given a user query (منظمة الصحة العالميةmnẓmh ālṣḥh ālʿālmīh , "World Health Organization") in the source language, the algorithm retrieves a set of possible translation for each query term $S_{qk} = \{q_k, t_i\}$ for each query term q_m from a dictionary (see Table 6.4). For example, we are considering the first query term (منظمةmnẓmh , "organization"), that has six translations (organization, organized, orderly, arranged, organizer, sponsor). The set of translations is thus defined with $k = 1$ and $1 \leq i \leq 6$ as $S_{q1} = \{q_{1,t_1}, q_{1,t_2}, q_{1,t_3}, q_{1,t_4}, q_{1,t_5}, q_{1,t_6}\}$. The translation sets for all query terms are retrieved from the bilingual dictionary. After the translation sets are retrieved, the next step is to generate the translation combinations between the translations for each of the query terms. The total number of combinations can be computed by simply multiplying the sizes of all translation sets. For the previous example we thus obtain total number of combinations $6 \cdot 3 \cdot 5 = 90$ as listed in Table C.3 in the Appendix.

Finally, the MI score will be calculated for all possible combinations of the translation-candidate pairs (translation sets). The translation combination that maximizes the MI score will be selected as the best translation for the user query (three translations will be selected). Before we present an evaluation of this approach in Chapter 7, we first

Query Terms	Sense inventory (Possible English Translations)
منظمةmnẓmh	[organization, organized, orderly, arranged, organizer, sponsor]
الصحةālṣḥh	[health, truth, correctness]
العالميةālʿālmīh	[universality, internationalism, international, world, wide]

Table 6.4: Sense inventory for each of the ambiguous query terms

discuss one of its main drawbacks, the data sparseness issue (Ahmed et al., 2011), in the following.

Revised MI to Overcome Data Sparseness Issue

In order to clarify the data sparseness issue, let us consider the following example. When translating the Arabic query "ضريبة مبيعات الادويةḍrībh mbīʿāt ālādwīh " (medications tax sales), there might be no enough statistical co-occurrences data obtained from the corpora and thus the algorithm will fail to translate this query. However, the revised algorithm can exploit the corpora and check out which term has no cohesion score with other terms and thus this term can be detected and eliminated. In this case, the term that affects the cohesion score is "الادويةālādwīh " (the medications, the remedies) and eliminating this term will allow to obtain sufficient statistical co-occurrence data. The rest of the terms are "ضريبة مبيعاتḍrībh mbīʿāt " (tax sales) have very high cohesion score due to the fact that these terms are widely available in the corpora.

For the translation of the noise term, as explained above, the first ranked translation will be taken from the dictionary. Looking at the translation provided by the araMorph package that we use for translation, the translation is ranked as follows (the medications, the remedies), so the algorithm will select the (the medications) as translation for " الادويةālādwīh ". The noise term detection process will be performed only if the proposed disambiguation algorithm failed to provide the translation due to the lack of statistical co-occurrences data for the query terms as a whole.

In the following, we describe, in detail, how the elimination process is performed by the algorithm. For simplicity, let's consider the previous given example "ضريبة مبيعات الادويةḍrībh mbīʿāt ālādwīh " (medications tax sales). The elimination process is done as follows:

- The algorithm generates all possible translation combinations: Given the user query $Q = \{t_1, t_2, ..., t_n\}$ "ضريبة مبيعات الادويةḍrībh mbīʿāt ālādwīh " (medications tax sales), the set of possible translation combinations $\{Tcom_1, Tcom_2, ..., Tcom_n\}$, where n defines the number of possible translation combinations for the user query Q. In our example, $n = 8$, so 8 translation combinations are generated (See Table 6.5).

S/N	Translation Combinations
1	tax AND sold AND remedies
2	tax AND sold AND medications
3	tax AND sales AND remedies
4	tax AND sales AND medications
5	levy AND sold AND remedies
6	levy AND sold AND medications
7	levy AND sales AND remedies
8	levy AND sales AND medications

Table 6.5: Translation combination for "ضريبة مبيعات الادويةḍrībh mbīʿāt ālādwīh ".

- The algorithm constructs possible term combinations between the generated translation combinations: Given a translation combination $Tcom_i = \{t_1, t_2, ..., t_n\}$, we compute its possible term combinations as follows: Given the set of $\frac{n(n-1)}{2}$ combinations, n is the number of terms in the given translation combination. The set of term combinations between all translation combination terms is defined as $Com_i = \{\{Tcom_{i,j}, Tcom_{i,k}\} | 1 \leq j < n, j < k \leq n\}$ Let's consider the translation combination (tax AND sales AND medications) number 4. Here $i = 4$, $n = 3$ and $S = 2$. After generating all possible combinations between the translation combination terms, the mutual information score for each term combination will be calculated based on Eq 6.2.

- The algorithm computes the MI score for each individual term combination, and then the MI score mean will be calculated. The term that has the lowest MI score, which is below the MI core mean, will be considered as a noise term and thus the term combination that includes this term will be eliminated. As shown in Table 6.6, term combinations with the term (medications) always have the lowest MI score (1.38629 and 3.98898).

S/N	Term combinations	MI Score
1	tax AND sales	8.86319
2	tax AND medications	1.38629
3	sales AND medications	3.98898
	The MI score mean	4.746

Table 6.6: Term combinations and their MI Scores.

- The algorithm calculates the average MI score individually for all terms in the constructed term combinations and compares them with the MI score mean. As shown in Table 6.7 ,the term "الادوية ālādwīh " (medications) has the lowest MI average score (2.687), which is below the MI score mean (4.746), and thus will be classified as a noise term and will be eliminated. In contrast, all other terms have an average mutual score, which is above the MI score mean and thus have significant statistical co-occurrence data needed for translation.

S/N	Term	MI average Score
1	sales	6.426
2	tax	5.124
3	medications	2.687

Table 6.7: Terms and their average MI Scores.

- Using the dictionary method, possible translations with contextual information for the noise term will be suggested. Ultimately, if the user agrees with the translation of the noise term based on the contextual information, the translated noise term will be included in the translation, otherwise the translated noise term will be cancelled.

6.2 Conclusion

We proposed two approaches for word translation disambiguation, one based on Naïve Bayesian Classifier and the other based on Mutual Information approach. For Naïve Bayesian Classifier approach, we used a bilingual parallel corpus together with sense definitions by translations into another language. The disambiguation for each sense of the polysemous word is done by defining a sense of each of the ambiguous words. In order to train the algorithm, a set of features was defined. The algorithm then selects the sense that maximizes the score. For the Mutual Information approach, we used monolingual corpora as source of the statistical co-occurrences data. In order to deal with the data sparseness issue we proposed a revised mutual information approach. The revised algorithm dealt with data sparseness issue by counting the cohesion between all terms in the user query and eliminating the term or terms that have a cohesion score close to zero.

Chapter 7

Evaluation of Disambiguation Algorithms

In the following, we show an evaluation of the query translation and disambiguation algorithms: accuracy evaluation based on parallel corpora and Naïve Bayesian Classifier (NB) (see Section 7.1), accuracy evaluation based on monolingual corpora and Mutual Information approach (see Section 7.2).

7.1 Accuracy Evaluation Based on Naïve Bayesian Classifier (NB)

We evaluated our approach through an experiment using the Arabic/English parallel corpus aligned at sentence level (Ahmed and Nürnberger, 2008b,d). We selected 30 Arabic sentences from the corpus as queries to test the approach. These sentences have various lengths starting from two words. These queries had to contain at least one ambiguous word, which has multiple English translations.

In order to enrich the evaluation set, these ambiguous words had to have higher frequencies compared with other words in the training data, ensuring that these words will appear in different contexts in the training data. Furthermore, ambiguous words with high frequency sense were preferred. The senses (multiple translations) of the ambiguous words were obtained from the dictionary. The number of senses per test word ranged from two to nine, and the average was four. For each test word, training data were required by the algorithm to select the proper sense. The results of the algorithm were compared with the manually selected sense. For our evaluation, we built different classifiers from different subsets of features and combinations of them. The first classifier based on features that were obtained from the user query terms and topic context, which was represented by a bag of words in the context of the ambiguous word. The second classifier was based on the topic context and its inflectional form. In order to evaluate the performance of the different classifiers, we used two measurements: applicability and precision (Dagan and Itai, 1994; Kang, 2003; Fakhrahmad et al., 2011). The applicability is the proportion of the ambiguous words that the algorithm could disambiguate. The precision is the proportion of the corrected disambiguated senses for the ambiguous word. The performance of our approach is summarized in Table 7.1. The sense, which is proposed by the algorithm was compared to the manually selected sense. As it is expected the approach is better in the case of long query terms which provide more rich features and worse in short queries, especially the one consisting of two words.

classifiers	**Applicability**	**Precision**
Query term + Topic context	52 %%	65 %
Query term+ feature Inflectional form	82 %%	93 %

Table 7.1: The overall performance using applicability and precisions.

We consider that the reason for the poor result for the short queries is that, when the query consists of few words it is possible that the features which are extracted from the query terms can appear in the context of different senses.

7.2 Accuracy Evaluation Based on Mutual Information Approach

We conducted two experiments in order to evaluate the proposed approach. In the first experiment co-occurrence data was used, which was obtained from the monolingual corpus (English Gigaword Corpus)[1] and the second was based on co-occurrence data, which was obtained from the web using a particular search engine (here, Yahoo) (Ahmed et al., 2009a; Ahmed and Nürnberger, 2010; Ahmed, 2010; Ahmed et al., 2011). The English Gigaword Corpus is a comprehensive archive of newswire text data that has been acquired over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. We used the third edition of the English Gigaword Corpus. The dictionary included in the araMorph package was used to define the senses of each query word. In order to evaluate the disambiguation algorithm, we selected randomly from the parallel corpora, 20 Arabic queries. These queries included at least one ambiguous word which has multiple English translations. In order to enrich the evaluation set, these ambiguous words have higher frequencies comparing with other words in the training data ensuring that, these words appear in different contexts in the training data. The number of senses per test word ranged from 1 to 14, and the average was 4.3. The number of query translation combinations ranged from 4 to 200 with the average being 29.1. In order to evaluate the performance of the algorithm, we used two measurements: applicability and precision (Dagan and Itai, 1994; Kang, 2003; Fakhrahmad et al., 2011). The applicability is the proportion of the ambiguous words that the algorithm could disambiguate. The precision is the proportion of the corrected disambiguated senses for the ambiguous word. Table 7.2 shows, the applicability and precision of the proposed algorithm, using monolingual corpora, over the 20 test queries. The applicability and precision were 75% and 70%, respectively. The algorithm was unable to disambiguate 25% of the queries due to insufficient statistical co-occurrence data obtained from the monolingual corpus. However, dealing with the sparseness data issue in the revised algorithm, this error rate was reduced by 5%. This error rate 20% was due to the lack of some statistical co-occurrences even after the elimination of the noise terms. In addition to this the ranked translation in the dictionary was not correct for all cases.

For example, consider the Arabic query "عجز سداد الدين ʿǧz sdād āldyn " (The deficient debt payment). Based on the cohesion score calculated for all possible combinations of the query terms, the term "عجز ʿǧz " has the lowest cohesion score compared to the rest of the terms and thus it is considered to be a noise term and will be eliminated. The

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T07

Co-occurrence data source	Applicability	Precision
Monolingual corpora	75%	70%
WEB	90%	80%

Table 7.2: Tool overall performance using monolingual corpora and the web.

rest of the terms "سداد الدينsdād āldyn " have a high enough cohesion score and thus the tool is able to translate them. As it is explained in Chapter 6.1.2, the translation of the eliminated term "عجزʿğz " will be selected based on the first ranked translation in the dictionary. The dictionary provided the following translations for the eliminated term " عجزʿğz ": ("rear", "part", "deficit", "insolvency", "incapable", "impotent", "incapacitate", "immobilize", "grow", "old", "weakness" and "inability"). The correct translation of the term "عجزʿğz " would be (deficit), which is ranked in position number three, in the dictionary. The applicability and precision of the proposed algorithm, using the web, averaged over the 20 test queries, were 90% and 80%, respectively. Due to very generic sense, the algorithm was unable to disambiguate 10% of the test queries. For example, consider the Arabic query "رسم جمركي علي اللّوحات الفنيةrasem ğmrky ʿlī āll̃wḥāt ālfnyh " (Customs tax of Paintings). The Arabic word " رسمrasem " has the following translations in English, ("drawing", "sketch", "illustration", "fee", "tax", "trace", "indicate", "appoint" and "prescribe"). What made this query very difficult to disambiguate is that the word " رسمrasem " can be found frequently in the context of (Customs) or in the context of (Paintings), which both exist in the query. These results show that the performance varies according to the query topics. Using monolingual data, our algorithm is better in the case of topic-specific senses and worse in the case of generic senses. Although the corpora used by the algorithm is rich corpora, which covers a broad range of different topics with a significant number of co-occurrence data, this corpora failed to provide co-occurrence data for 20% (this error rate also due to very generic sense cases) of the test queries e.g., the previously mentioned example: "رسم جمركي على اللّوحات الفنيةrasem ğmrky ʿlā āll̃wḥāt ālfnyh " (Customs tax of Paintings). In contrast, the algorithm using the co-occurrence data, obtained from the web, could disambiguate 18 queries and failed only to provide co-occurrence data for two queries. This is clearly due to the fact that the web provides significant co-occurrence data compared to other resources. However, obtaining statistical co-occurrences data from web is not trivial task from the implementational point of view. At least for long queries it would be almost impossible for a cross-lingual retrieval system to ensure that querying the web

Query Terms	**Sense inventory (Possible English Translations)**
سادادsādād	[payment, appropriateness, obstruction, embolism, plug, stopper]
الدينāldyn	[implacable, mortal, religious, debt, religion]

Table 7.3: Sense inventory for each of the ambiguous query terms.

and obtaining the statistical co-occurrence data enables real time performance for an interactive system. Table 7.3, shows the possible English translations for each of the original query terms " رسم جمركي على اللّوحات الفنيةrasem ğmrky ‹lā āll̃wḥāt ālfnyh ". For the first query term, 6 possible English translations were identified. For the second query term 5 English translations were identified. The total number of translation combinations is 30. Table 7.4 and Table 7.5 show an example for only the first 10th translations combinations of the c-occurrence data obtained from monolingual corpora and the web, respectively. One can notice the huge difference between the abundance of the co-occurrence data obtained from the web compared with co-occurrence data obtained from corpora. For example, using the monolingual corpora, the highest cohesion co-occurrence was 1460 for the translation combination (payment dept) and 0 was for 10 translation combinations. In contrast, the highest cohesion co-occurrences, using the web, for the same translation combination (payment dept) was 176000000, while the lowest cohesion co-occurrences, using the web, was 10500.

S/N	**Translation Combinations**	**Occurrence**	**MI Score**
1	payment AND debt	1460	7,28611
2	plug AND debt	151	5,01727
3	payment AND religious	41	3,71355
4	plug AND religious	36	3,58350
5	payment AND religion	31	3,43369
6	obstruction AND debt	20	2,99572
7	appropriateness AND debt	8	2,07944
8	plug AND religion	6	1,79175
9	obstruction AND religion	4	1,38629
10	embolism AND religious	4	1,38629
-	-	-	-

Table 7.4: Example of the co-occurrence data obtained from the monolingual corpora.

S/N	Translation Combinations	Occurrence	MI Score
1	payment AND debt	176.000.000	19,08977
2	appropriateness AND debt	612.000	17,52598
3	appropriateness AND religious	639.000	17,13441
4	obstruction AND debt	676.000	17,12512
5	obstruction AND mortal	197.000	17,09611
6	payment AND religious	34.400.000	17,02261
7	obstruction and religious	772.000	16,82318
8	appropriateness AND religion	663.000	16,66818
9	appropriateness AND mortal	3.750.000	16,54273
10	payment AND religion	30.000.000	16,38265
-	-	-	-

Table 7.5: Example of the co-occurrence data obtained from the web.

7.3 Conclusion

For Naïve Bayesian Classifier based on the experiments that we performed, using Arabic/ English parallel corpus, results could show that our algorithm achieved certain promising results when the inflectional form for Arabic words is considered. The applicability and precision using 30 polysemous words were 52% and 65% for the first classifier and 82% and 93% for the second classifier, respectively. For the Mutual Information approach, we used monolingual corpora as source of the statistical co-occurrences data. Based on the experiments that we performed, using monolingual corpora and the web, results showed that our algorithm achieved certain promising results. The applicability and precision for 20 test queries, using monolingual corpora, were 75% and 70%. Furthermore, in this evaluation, the revised algorithm that dealt with data sparseness issue by counting the cohesion between all terms in the user query and eliminating the term or terms that have a cohesion score close to zero with other terms was tested. The revised algorithm reduces the error rate from 25% to 20%. To enrich the source of the statistical co-occurrence data needed to enhance the algorithm for better selection of the correct translation, the web was used as a rich source of this statistical co-occurrence data. The applicability and precision for the 20 test queries, using the web, were 90% and 80%.

Part IV

Interactive Meaning Refinement

Chapter 8

Interactive Meaning Refinement

8.1 Introduction

In the past few years, the interest in interactive cross-lingual retrieval systems has increased significantly. Logical explanations for this phenomenon are that cross-lingual retrieval is a very difficult task to perform by the cross-lingual retrieval system itself. The difficulty lies in dealing with natural lexical ambiguity of the source and target language which is not a trivial task that the cross-lingual retrieval system can resolve fully automatic. In every language, there are words which have multiple senses, which results in the user query having several possible translations. Furthermore, difficulties occur in cross-lingual information retrieval, due to the fact that users in some cases are looking for documents written in languages they can not understand and in some extreme cases they can not even read. This may lead to the result that the users can not recognize the desired documents even if they have received them. Therefore, there is a need for users and the cross-lingual retrieval system to overcome the shortcomings for each other. The cross-lingual retrieval system provides users with helpful information in the user's native language and based on this information, the user can provide the cross-lingual retrieval system with useful feedback that would likely help to improve the translation and thus improve the cross-lingual retrieval quality. Therefore, the accuracy of the cross-lingual retrieval system depends to a strong extent on the interaction between the user and the system (Ahmed and Nürnberger, 2010; Ahmed et al., 2011; Ahmed and Nürnberger, 2012).

Based on the cross-lingual tool literature review (see Chapter 3), we identified several issues and shortcomings, which we have tackled in the cross-lingual tool proposed in this thesis. We proposed a smooth design that is on the one hand supported by significant back-end components and on the other hand gives the user some control over the query translation. The proposed cross-lingual tool, in this thesis, considers the user as an integral part of the cross-lingual process, in that the user can interact with the cross-lingual tool in a way that allows him/her improve the translation and thus improve the cross-lingual retrieval process. To achieve this crucial goal, the user needs valuable information from the system. For example, how the request is made to improve the translation, when the user has no knowledge about the target language.

In the following, we outline the identified state-of-the art cross-lingual tools shortcomings and the proposed solutions to tackle them.

8.2 Tackled State-of-the art Cross-lingual Tools Shortcomings

In the following, we clarify points of interest that we have focused on, to analyze the ability of the state-of-the art cross-lingual tools to support the user in a cross-lingual search. Furthermore, based on this analysis, we discuss which solutions we proposed to tackle the identified issues and shortcomings in the state-of-the art cross-lingual tools. The main points of interest are translation confidence, automatic translation, translation improvement, user support and new language adaptations.

- Translation confidence: An important point, which has been studied in depth in the state-of-the art cross-lingual tools analysis, is the translation confidence. How we expect the user to rely on the translation provided by the cross-lingual tool when he/she is not able to understand or even read this translation. Based on the analysis of eight cross-lingual tools, we found out that only two cross-lingual tools provide a possibility of giving the user some confidence in the translation. However, both cross-lingual tools used back translation, where the translation is translated back to the source language. If there is overlap between the query and the back translation then one might have some confidence in the translation. However, this approach suffers from a clear drawbacks, when no synonyms can be found in the dictionary, the technique is not helpful; and significant homonymy in the target language can result in confusing back translations (Oard et al., 2008). Some state-of-the art cross lingual tools used the dictionary definitions to give the user some confidence in the translation. However, bilingual dictionaries, in which the definitions of source language are available for each translation for the target languages, are very rare and very laborious. Some times, in the existing of translation definition, it does not resolve the problem clearly because this definition is displayed for each translation term independent from other translated terms.

  - In the proposed cross-lingual tool, in order to tackle such clear shortcomings, we used parallel corpora that provided us with significant numbers of definitions (context), which we can use to describe the translation in a language the user is familiar with. We call this type of context "contextual information". This information is extracted from the parallel corpora and describes the complete translation (all terms in each possible translation at once) in a language the user is familiar with, in that he/she can have confidence in the translation. These parallel corpora in most cases can be freely obtained in the internet. One important aspect in this resource is that it is continuously growing for different language pairs e.g., Europarl parallel corpora[1] available for 21 European language pairs or the United Nations corpora[2] available for 6 language pairs. Furthermore, parallel corpora are a significant source of contextual information for different types of terms i.e., OOV words such as proper names, technical terms and acronyms. In the state-of-the art cross-lingual tools, a translation definition is usually short and is displayed to the user as raw text without any further classification i.e., which meaning can each term in the definition represent to the user. In the cross-lingual proposed tool, in this thesis, the contextual information is not delivered to the user as raw text; instead a classified representation for each term in the contextual information is generated. For example, an interesting point for the user, to rely on the translation, is to see the query terms in the contextual information. These terms are highlighted in bold black and are displayed with their context in different sentences. In order to improve this feature, synonyms for the given query terms in the contextual information are highlighted with light grey (selecting these highlighting mechanisms can avoid issues for people who are color blind).

- Automatic translation: Most of the studied state-of-the-art cross-lingual tools provide no possibility for automatic translation and thus for automatic translation disambiguation. They are based on individual term translations, where the user is requested to perform the disambiguation process. This disambiguation process by the user is done based on the translations definition, which in some cases is

[1] http://www.statmt.org/europarl/
[2] http://www.un.org/

displayed with the translation. Despite the lack of this translation definition in the dictionary, where in some cases it is very short or does not exist at all, the user needs to make a huge effort and check each translation alternative along with its definition, in order to disambiguate. This task can take significant time, especially for query terms that have an abundance of possible translations. We want to emphasize that some of the state-of-the art cross-lingual tools that use the automatic translation, don't use their implemented approaches for automatic translation, instead a free machine translation is used. Using machine translation will give no possibility for the user to interact with the translation so he/she can improve it. Furthermore, one clear drawback, that machine translation systems are not suitable for the cross-lingual task (machine translation expect syntactically written sentences) is that the user queries are often short (Gabrilovich et al., 2009) and formed, usually without any proper syntactic structure (Hull and Grefenstette, 1996).

  - In the proposed cross-lingual tool, we alleviate the user task to perform the disambiguation, where the user needs to check all translation alternatives with their dictionary definitions (this can lead to frustration and lack of desire to use the tool), by researching and implementing an automatic translation component in the interface. In the proposed cross-lingual interactive tool, we gave the user a possibility to interact with the automatic translation by selecting relevant terms suggested by the tool to see the affect on improving the automatic translation on his/her cross-lingual search. The integrated automatic translation in the proposed cross-lingual tool is based on statistical methods that we enhanced to deal with the translation ambiguity e.g., Naïve Bayesian Classifier (NB) or Mutual Information (MI). In order to give the user flexible possibilities to interact with the tool, five automatically ranked translations are provided. Using user-selected interactive terms; the automatic translation algorithm will re-rank the translation, based on the user interaction.

- Translation improvement: This was one of the important aspects that we carefully studied in the state-of-the art cross-lingual tools. We wanted to check whether the state-of-the art cross-lingual tools really consider the user as an integral part and whether the state-of-the art cross-lingual tools provides the user with significant information to perform the cross-lingual task. We found out only two cross-lingual tools out of the eight studied cross-lingual tools provided some kind of translation improvement. However, this support was deficient in various aspects. For example, some tool provides a translation improvement possibility by providing the user with the retrieved documents relevant to his/her information need. The user can initiate a new translation process, based on the examined retrieved documents (based on the search result the user can use different query terms). There is no possibility of improving the translation during the translation process, which leads the user to lose time and be frustrated. Another tool provides a translation improvement possibility by using EuroWordnet[3] relations. However, EuroWordNet employs only a limited number of languages.

  - In the proposed cross-lingual tool, in order to give the user wide possibilities to interact with the cross-lingual tool proposed in this thesis, the cross-lingual tool provides the user with five ranked translations along with their contextual information. Furthermore, a list of possible interactive related terms, to the user query, is extracted from a corpora and presented to the user. Using this term/terms the user can interact with the system and has impact on refining the translation. The user can immediately see his/her interactive term/terms selection impact on the automatic translation, as well as in the cross-lingual search results. The selected interactive term/terms are only used for re-ranking purposes and they will not be added to the query as new term/terms.

- User support: Two of the state-of-the art cross-lingual tools provide partial support for the user. They provide support after the translation is performed, the retrieved documents together with small images are represented, which are called Document Thumbnail Visualizations. However, examining the retrieved documents has no clear impact on supporting the user in performing the cross-lingual search. Another four cross-lingual tools provide more support to the user in interacting or alleviating his/her task in using the system e.g., expressing the information in a visually attractive manner, which makes the user's task easier and expands the original query with extra relevant terms. Based on our state-of-the art cross-lingual tools analysis, we found that the state-of-the art cross-lingual tools suffer from clear shortcomings in supporting the user during the cross-lingual retrieval process. For example, the state-of-the art cross-lingual tools lack of clear error notification e.g., when there is no translation available from the dictionary for some term/terms or when the algorithm failed to provide a translation for the given user query (provide reason of failure so that the user can have some idea as to what the problem could

[3] http://www.illc.uva.nl/EuroWordNet/

be). Another deficit is the lack of cross-lingual process stages e.g., when the user is not happy with the current interaction step and would like to take a step backward.

  - In the proposed cross-lingual tool, a significantly wide range of user support was taken into account, when designing the tool. The proposed cross-lingual tool includes a significant error notification mechanism in that it provides a description and an automatic recovery for each possible failure. For example, when the user submits three query terms and one of them has no possible translation in the dictionary, the cross-lingual tool will notify the user that there is no possible translation found for this term. Furthermore, based on the rest of the query terms, the cross lingual tool automatically suggests the relevant terms to the user's query so that he/she can replace the term that has no translation in the dictionary, if needed. Another significant notification, when the cross-lingual tool fails to provide the user with any automatic translation, is that the tool notifies the user that there is no significant statistical data obtained from the automatic translation algorithm knowledge source, so the user can reformulate his/her information need.

- New language adaptations: A very important feature to consider when designing cross-lingual tools is the ability of the cross-lingual tool to handle more languages. One of the researched state-of-the art cross-lingual tools provides this possibility. However, it was not described how and to what extent.

  - The proposed cross-lingual tool, in this thesis, has been designed to accommodate new languages, when the language resources are available. All algorithms integrated in the proposed cross-lingual tool, such as spelling correction algorithm (n-gram based approach), word sense disambiguation algorithms (Naïve Bayesian or Mutual Information) and the contextual information provider is language independent. In order to include new languages, a bilingual dictionary and parallel corpora are needed. No need to adapt any algorithm integrated in the proposed cross-lingual tool for any new language. An exception for this is that pre-processing algorithms might be needed. For example, when we first included Arabic, an Arabic analyzer was needed to tackle the high morphological issue or when including German, a decomposing algorithm was needed.

To summarize, this chapter aims to answer these main research questions:

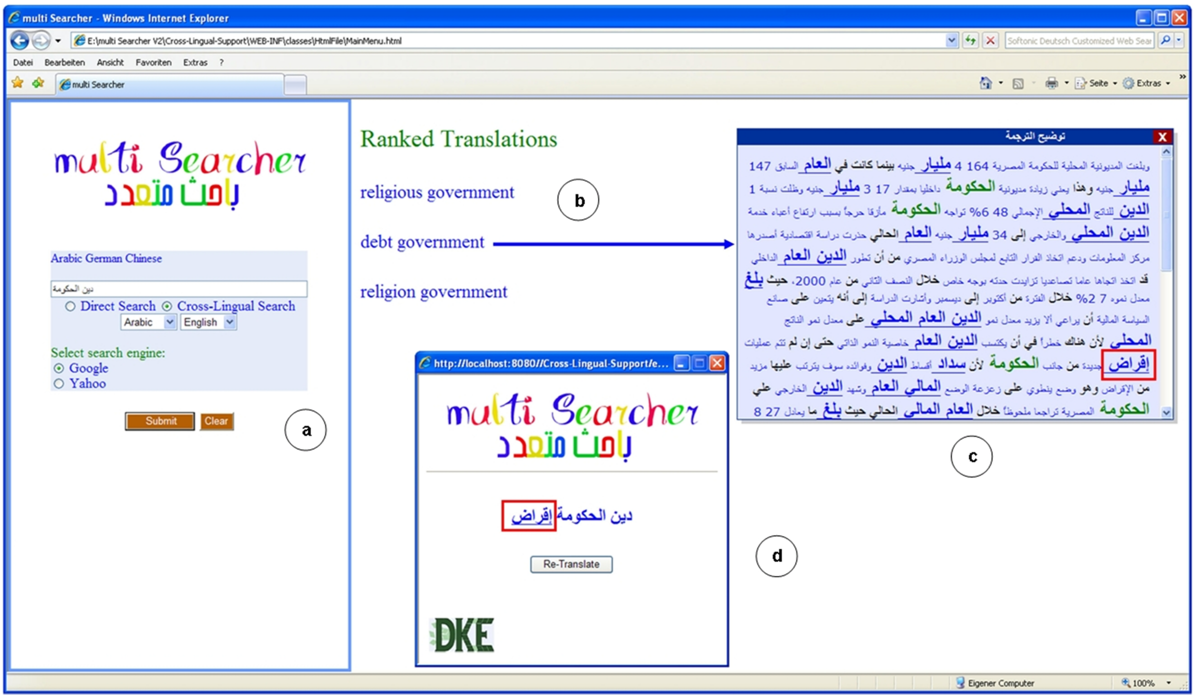
- Can cross-lingual searchers improve the performance of cross-lingual retrieval systems when they have passed control over the query translation? Which type of control should they have and to which extent? Giving users control over a cross-lingual retrieval system means giving them a possibility to review and refine the query translation. This leads to considering these research questions:

- What information from the system do cross-lingual searchers need to refine their queries, how do they obtain this information and how this information presented to them?

In the following, we give a general overview in how the cross-lingual search is performed. Therefore, we start with a short presentation of the first prototype (Ahmed and Nürnberger, 2010; Ahmed, 2010) (see Section 8.3) in order later to identify issues related to it (see Section 8.3.1). Furthermore, Based on the cross-lingual tools literature review (see Chapter 3) and the evaluation of the first prototype, we identified several issues and shortcomings to tackle in the proposed cross-lingual tool in this thesis. In addition, we conducted a broad user study to consider more points of interest and identify more issues in the first prototype which is tackled in the second prototype (see Section 8.4).

8.3 Cross-lingual Interactive Tool: First Prototype

In order to help the user to better understand the meaning of the different query term translations, the tool provides contextual information to clarify the usage - and thus the meaning - of the terms (Ahmed and Nürnberger, 2010). Figure 8.1 (a) shows an example, where the user submits the Arabic query "دين الحكومة dyen alḥkwmh ". The query is automatically translated and the best three translations will be displayed to the user in ranked order (See Figure 8.1 (b)). Each translation is looked up in the target language documents index (one translation after the other) in order to obtain the relevant documents (contextual information), for the translation. In order to get the equivalent documents in the source language, the parallel corpus [4] is queried. Since it is possible that some retrieved documents will be very similar – which would result in duplicate contextual information – the documents retrieved from the source language are automatically grouped and contextual information is selected only once from each cluster. The final selected contextual information is not provided to the user as raw text as it is the case in the state-of-the art cross-lingual tools, but instead, it will be presented as a classified representation of each contextual information term: each term of the contextual information is color-coded according to its related type and can be



وبلغت المديونية المحلية للحكومة المصرية 164 4 مليار جنيه بينما كانت في العام السابق 147
مليار جنيه وهذا يعني زيادة مديونية الحكومة داخليا بمقدار 17 3 مليار جنيه وظلت نسبة 1
الدين للناتج المحلي الإجمالي 48 6% تواجه الحكومة مأزقا حرجاً بسبب ارتفاع أعباء خدمة
الدين المحلي والخارجي إلى 34 مليار جنيه العام الحالي حذرت دراسة اقتصادية أصدرها
مركز المعلومات ودعم اتخاذ القرار التابع لمجلس الوزراء المصري من أن تطور الدين العام الداخلي
قد اتخذ اتجاها عاما تصاعديا تزايدت حدته بوجه خاص خلال النصف الثاني من عام 2000، حيث بلغ
معدل نموه 7 2% خلال الفترة من أكتوبر إلى ديسمبر وأشارت الدراسة إلى أنه يتعين على صانع
السياسة المالية أن يراعي ألا يزيد معدل نمو الدين العام المحلي على معدل نمو الناتج
المحلي لأن هناك خطرا في أن يكتسب الدين العام خاصية النمو الذاتي حتى إن لم تتم عمليات
إقراض جديدة من جانب الحكومة لأن سداد أقساط الدين وفوائده سوف يترتب عليها مزيد
من الإقراض وهو وضع ينطوي على زعزعة الوضع وشهد الدين الخارجي على
الحكومة المصرية تراجعا ملحوظاً خلال العام المالي الحالي حيث بلغ ما يعادل 27 8

Figure 8.1: The translation alternatives with their contextual information (Ahmed and Nürnberger, 2010).

selected as a disambiguating term (the user's query terms are in green, suggested terms, by the tool based on highly frequent co-occurrences in the context of the query are in bold blue and underlined, all remaining terms are blue except stop words that are black and not selectable) (See Figure 8.1 (c)). In order to clarify the interaction scenario, we consider the submitted user query "دين الحكومة dyen alḥkwmh ". The query term " الحكومةalḥkwmh " has two translations ("the government" or "the administration"), while the other term "دينdyen " has several possible translations e.g. ("Religion" or "Debt"). Based on the *MI* score, translation alternatives are displayed in ranked order together with their contextual information (See Figure 8.1 (b) and (c)). Thus the user has the possibility to select the suitable translation. Here, the translations provided by the system ("the government religion") and ("the government debt") are correct even though they are used in a different context. This is due to the fact that ("government") appears frequently in the context of ("religion" or "debt"). As shown in Figure 8.1 (b) and (c), the user is interested in the second ranked translation ("debt government"). Using the contextual information, the user can select one or more terms to improve the translation. To simplify the user's task, the tool automatically proposed relevant terms (highlighted in bold blue and underlined), e.g. ("payment", "financial", "lending", "loan"). Once

[4] www.ldc.upenn.edu/

the user selects, for example, the interactive term "اقراضāqrāḍ " ("lending") (See Figure 8.1 (d)), the tool re-translates the modified query and displays the new translations ("debt government loan", "debt government lending" and "debt administration loan"), to the user. The user can, with a simple mouse click, confirm the translation which will then be sent to his favorite search engine using integrated web services, e.g. Yahoo or Google, retrieving the results and displaying them in the tool interface.

8.3.1 Evaluation

The goal of the tool interface evaluation was to observe current practice on how real cross-lingual information retrieval tasks are accomplished through the proposed tool and to imagine a CLIR system that would fully support cross-lingual information retrieval tasks.

8.3.1.1 Pilot User Study

In the performed pilot user study (Ahmed and Nürnberger, 2010), 5 users were involved. The type of users are students and researchers who have no or little knowledge in the target language. Three of the users were male and two were female. Age ranged from 22 to 31. The differences found between users are more likely to account for the provision of different options to meet more diverse needs. The strength of this study lies in the fusion of different interests and point of view of the test users, whereby even a single user counts in building a broad picture of using the proposed tool. Furthermore, according to the research done by Nielsen and Landauer (1993), this small number of test persons is appropriate to find at least 85% of all usability issues. Most of the remaining 15% usability problems is identified by conducting a second user study with a second group of 15 users (see Section 8.4.2). We have chosen this evaluation layout to identify 98% of the possible usability issues in order to ensure the that the tool targets the user task as good as possible (Nielsen, 1994).

All user sessions were analysed to test a number of points of interest regarding the evaluation of the tool e.g., contextual information usefulness, interactive terms usefulness etc.

- Translation confidence: addressed how useful and accurate was the contextual information that describes the translation in the source language. The translation confidence gained full rate by the users with simple request of improvement. All users found the contextual information which is displayed a long with the translation very helpful in giving them a confidence in the translation. For the user who has no or little knowledge about the target language, the contextual information was very helpful in term of giving them full confidence about the translation that they see but they can not understand. For the improvement request, one user complained about the size of the contextual information. The user suggested decreasing the size of the contextual information (currently, the tool displays 5 documents (sentences) as contextual information). He mentioned one or two short sentences would be enough and will simplify the task of having a confidence in the translation. However, decreasing the contextual information size will lead to insufficiency in the interactive terms that can be used to improve the translation.

- Interactive terms usefulness: addressed the usefulness of the interactive terms in the contextual information that can be used to improve the translation. The suggested interactive terms by the tool, based on highly frequent co-occurrence data, in the context of the query, are in bold blue and underlined. In many cases these were helpful as the user mentioned, however, in some cases the users needed more terms than the ones suggested by the tool. These terms are color-coded blue and are found in the contextual information, which is displayed along with each proposed translation. Although these terms are found in the context of the user query, the users mentioned that these terms in many cases they don't lead to an improvement in the translation.

- Interaction time: addressed how much time needed to interact with the tool in order to improve the translation. The needed time between submitting the query and receiving the ranked translations along with their contextual information is between 2-5 seconds (for query with average length 4 words). However, two users wish to see the translation along with the contextual information in one second if possible. The needed time can be improved in future work. Main part of the delay on performing the task is related to the use of araMorph package, that we use to analyze and translate the Arabic query. We plan in future work to obtain a full dictionary which we can use to speed up the process of finding possible translations for each query term. This will lead to efficiency and accuracy improvement of the tool.

- Tool design: addressed possible improvement of the existing design of the tool e.g., which part of the tool needs redesign or enhancement. Most users were satisfied with the current design of the tool. However, two users suggested some redesign of the tool. These users would like to see the entered query, at any interactive level. For example, in the current design, when the translation is performed and the user would like to input a new query, she/he can only use the back button to enter new query. Furthermore, these users would like to have all past events, with the tool, displayed along with the current event e.g., the translation before the improvement and after the improvement. In future work, we will redesign the tool to take this point into account. This can be useful in case the interaction with the tool does not lead to improve the translation. The user still can interact with the original translation selecting new interactive terms to improve the translation.

8.4 Cross-lingual Interactive Tool: Second Prototype

Based on the literature review (see Chapter 3), we designed the required interface components to tackle each identified research problems (see Section 8.2). These interface components are integrated together in order to perform the cross-lingual task, from submitting the query till getting the relevant documents. In the following, we describe in detail each interface component, how it works and how it tackles each research problem.

8.4.1 Main Components of The Interface

As Figure 8.2 shows, the interface flow starts when the user submits his/her query. The entered user query will pass through several interface components before the cross-lingual search results can be displayed to the user. These interface components are: query pre-processing, automatic translation, contextual information and gloss, query post-processing and Error notification. Figure 8.2, shows how these components are related and how the information flows between the different interface components. In the following, we describe these components in detail.

- Query pre-processing component: Before the query can be translated, it will be pre-processed. The first important pre-processing step is to check whether the query is misspelled or not. If the query is misspelled, the misspelling query term/terms, using the MultiSpell approach (see Chapter 4), will be identified and corrected. MultiSpell is a language-independent spell-checker that is based on an enhancement of a pure n-gram based model. In addition to the correction of the query misspelled terms, we provide a possibility to deal with some special properties for some languages. Currently, we deal with Arabic word form variation problems or German compound word problems. For the Arabic language, it is possible that an Arabic word can be represented in different forms. Therefore, before translating, there is a need to transform the Arabic word to its basic form. The stemming of the user query terms is very important because the dictionary does not include all word forms, instead just the root form. For stemming, we used the araMorph package based on the Buckwalter Arabic morphological analyzer (Buckwalter, 2002). For the German language, compound words can result in having out-of-vocabulary (OOV) problems in cross-lingual information retrieval. Dictionaries usually do not include all compounds words. Therefore, In order to improve cross-lingual information retrieval effectiveness, these compound words need decompounding before translation (see Chapter 2.2.1.3). For decompounding, we use a dictionary-based decompounding approach (Chen and Gey, 2004). Once the user query is pre-processed it will be the input to the automatic translation components.

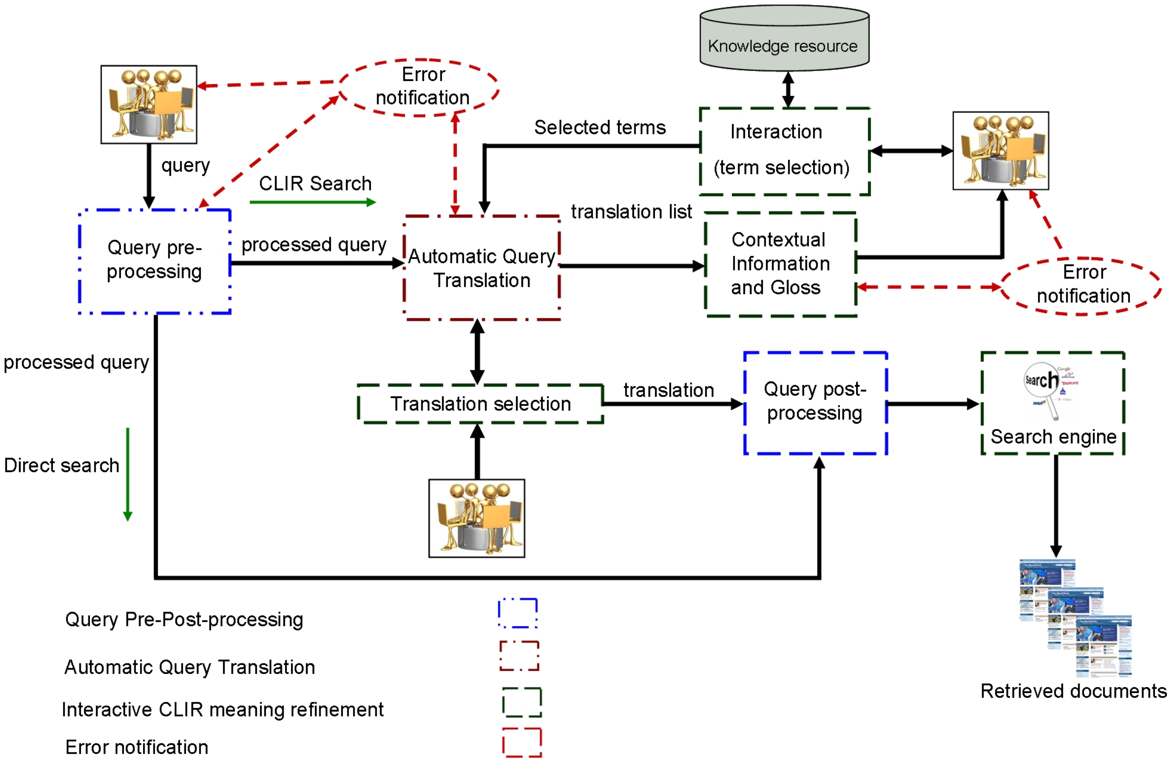


Figure 8.2: The main interface components.

- Automatic translation component: After finding the morphological root of each term in the query (for Arabic) or decompounding the compound words (for German), using the bilingual dictionary, each possible translations for each query term is obtained (translation set). Having a translation set for each query term, the translation combinations between terms in the translation set are generated. The result of this step is having all possible translations for the submitted user query (translation combinations). In order to select and rank the proper translations, statistical methods, based on target (Mutual Information approach) or parallel corpora (Naïve Bayesian approach), are used. The translations that maximize the statistical score measure are selected and ranked (five translations are selected and displayed to the user to interact with) (see Chapter 6).

- Contextual information and gloss component: Once the automatic translation is performed and displayed to the user, new issues as to confidence in the translation can arise. This issue especially affects users who have low or no knowledge in the target language. It is very difficult for those users to deal with the translation confidence without the cross-lingual tool support. In order to give the user a confidence in the translation, that he/she can not understand and in some cases can not even read, a contextual information provider is integrated in the cross-lingual tool. The contextual information is information displayed to the user along with each proposed translation, in a language the user is familiar with. In order to provide this contextual information, parallel corpora can be used. The input for this contextual information provider is the translated user query (five ranked translations). The translated user query is then looked up in the target language documents index (one translation after the other), in order to obtain the relevant documents (contextual information), for the translation. In order to make it easier for the user to respond and understand terms in the contextual information, the contextual information is not delivered to the user as raw text; instead a classified representation for each term, in the contextual information, is generated. For example, an interesting point for the user, to rely on the translation, is to see the query terms in the contextual information. These terms are highlighted in bold black and are displayed with their context in different sentences. In order to improve this feature, synonyms for the given query terms, in the contextual information, are highlighted in light grey (selecting these highlighting mechanisms can avoid issues for people who are color blind). Based on this contextual information, the user then has two possibilities. First, to interact with the interface by confirming the translation with a simple mouse click, which will then be sent to his/her favorite search engine, retrieving the results and displaying them back to the user. Second, if the user is not sure about the translation, he/she can interact with the interface by selecting relevant term/terms proposed by the cross-lingual tool. These terms will be used for re-ranking purposes, which also might result in new translations appearing, different from the initial five displayed translations. Some users, who have a good knowledge in the target language, would like to see information for the translation in the target language "gloss". We made this request available by giving the user a possibility of seeing the information for the translation in the interface. In order to give the user more confidence, the translation terms are highlighted in the gloss, in the same way as in the contextual information.

- Query post-processing component: Once the translation is refined and acknowledged by the user, new issues may arise. The characteristics of highly inflectional languages very often result in poor information retrieval performance. As a result, current search engines suffer from serious performance with the direct query term-to-text-word matching for these languages. Thus, search engines need to be able to distinguish different variants of the same word. In order to tackle this issue, a language-independent conflation approach, based on enhancing the n-gram approach is integrated in the cross-lingual tool (see Chapter 5). The cross-lingual tool suggests possible additional terms to the user's translated query. With a simple mouse click, users have the ability to deselect any one of the additional terms that don't satisfy their need.

- Error notification component: In order to support the user, in using the cross-lingual tool, an error notification component is integrated. The error notification component is responsible for watching all cross-lingual tool components and alerting the user to any failure and its causation. For example, when there is no translation available from the dictionary, for some term/terms, the error notification component will notify the user why his/her query is not translated. Another example, when the cross-lingual tool displays no automatic translation to the user, the error notification will alert the user, for example, that there was not enough statistical data obtained from the corpora to perform the automatic translation. Based on this notification, the user could then reformulate his/her query and perform the translation again. Another example, when the user poorly formulates his/her query i.e., one term is not related to the rest of the terms and thus this term can affect the cohesion scores for the remaining terms. The error notification component, will notify the user about the term that has no significant score with other terms, so the user has the possibility of replacing this term with a suitable one. Another example, when there is a misspelling in the user query, before this misspelling is corrected by the misspelled algorithm, the error notification component will notify the user that there is a misspelled term/terms. Furthermore, the error notification, will notify the user with the suggested correction. Another example, when no contextual information is available, the error notification will notify the user about this tool failure. The user can understand then, the problem is about the availability of the contextual information and not about the translation itself.

In the following, in Section 8.4.2, in addition to the discussed problems which have been revealed based on researching the state-of-the art cross-lingual tools, we conducted a broad user study to consider more points of interest and identify more issues in the first prototype which are tackled in the second prototype. In Chapter 9 an evaluation for English-German language pairs is conducted to evaluate whether the support provided by the proposed cross-lingual tool in this thesis is significant enough to guide the user in improving the translation and thus improving the performance of the cross-lingual retrieval. In the end of this chapter , a conclusion of the proposed cross-lingual interactive tool is presented.

8.4.2 User Study

We conducted a user study (with 15 participants), considering points, such as contextual information usefulness, translation confidence, interactive terms usefulness, the way the user has a control in running the system (how to provide the user with useful information at any level of his/her interaction e.g., showing all interactively selected terms during the cross-lingual retrieval process so the user can deselect any and move back to the initial state), error notification by the system (e.g., when no translation for a query term is found in the dictionary), the delivered information flow (the contextual information for the five ranked translations shouldn't be displayed all at once to the user, instead just the current focus), the design of the system (could the user have the possibility of having a broad overview of all useful information at once e.g., seeing the translation along with the relevant documents obtained by a search engine), highlight more related words in the contextual information or in the gloss (is it possible to identify words in the contextual information with related synonyms) and user support and new language adaptations etc.

In the performed user study, 15 users were involved. The type of users are students and researchers who have no, little or good knowledge in the target language. Ten of the users were male and five were female. Age ranged from 22 to 43.

In the following, we outline each identified problem and the proposed solution to tackle it.

- Translation confidence: addressed how useful and accurate was the contextual information that describes the translation in the source language. The translation confidence gained full rate by the users with simple request of improvement. All users found the contextual information which is displayed a long with the translation very helpful in giving them a confidence in the translation. For the user who has no or little knowledge about the target language, the contextual information was very helpful in term of giving them full confidence about the translation that they see but they can not understand. The improvement request, was about decreasing the size of the contextual information. Currently, the tool displays 5 documents (sentences) as contextual information. Users mentioned one or two short sentences would be enough and will simplify the task of having a confidence in the translation. However, decreasing the contextual information size will lead to insufficiency in the interactive terms that can be used to improve the translation.
  - We tackled this issue in the new design and compensated for this insufficiency. In the new design, the tool provides the user with a list of interactive terms, regardless of the contextual information. The user can select any term/terms to improve the translation, if needed, saving the user time. There will be two benefits from this step: the contextual information will be used only for translation confidence and the list of suggested interactive terms will be used to improve the translation, if needed (see Figure 8.3).
- Lack of information flow control: users complained that intensive-information is displayed at the same time e.g., the user is disturbed by seeing all contextual information for the five ranked translations displayed at once. The user mentioned that it would be helpful to control which information could be seen and when.
  - In the new design (see Figure 8.3), we tackled this issue in that we gave the user a possibility to hide information, which is not of current interest. This is done by displaying only the contextual information for the top ranked translation (first translation) and displays only a few words in the contextual information for other translations. If the user is interested in checking other translations with its contextual information, the user only needs to click on "mehr anzeigen - show more". The tool then displays the full contextual information for the selected translation. At the same time, the tool automatically hides the contextual information for the previously selected translation. The user is then able to see and focus only on the selected translation and its contextual information.

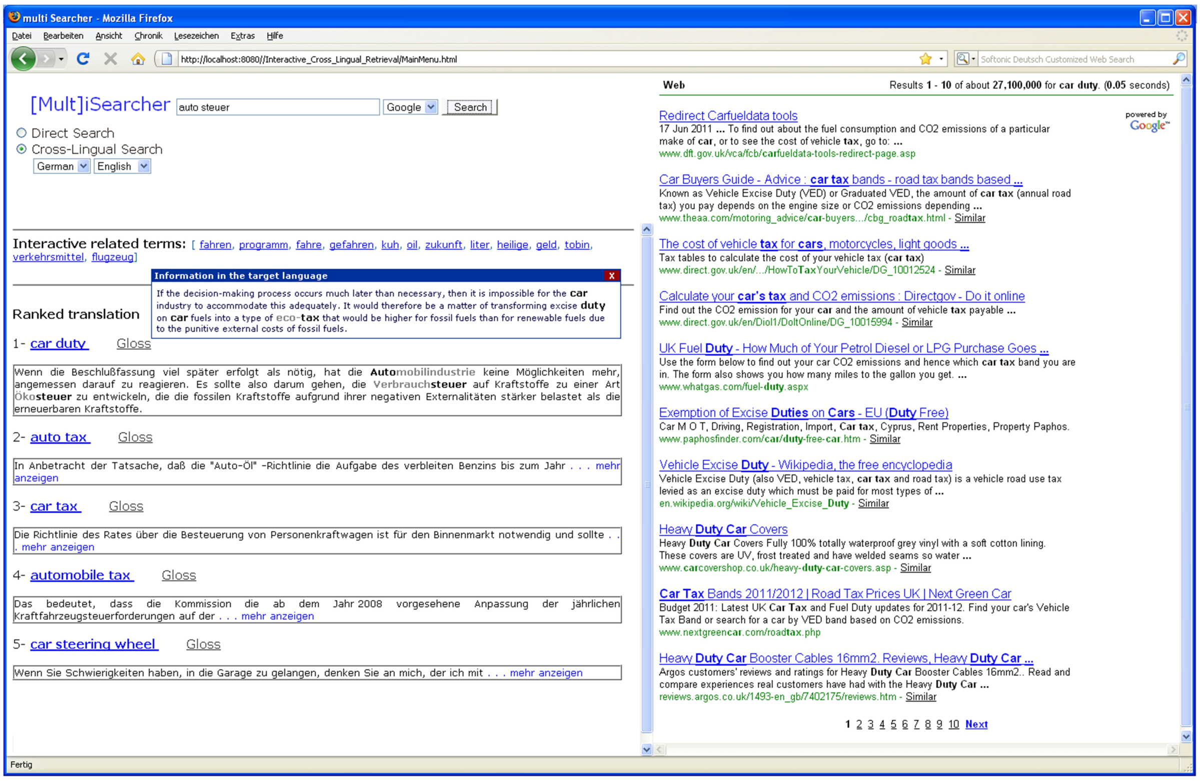
- Lack of control in the interface: The relevant retrieved documents to the translation are not displayed in the same interface along with the translation. In the previous design, the user has to click in the translation so the relevant documents to the translation, which were obtained by a favorite search engine, will be displayed on a new page.

  – As Figure 8.3 shows, in the new design, the user has a greater view of the translation along with the retrieved documents which are displayed in the same interface. Moreover, the tool automatically displays the relevant document to the first translation even before the user performs any action. If the user is interested in seeing the relevant documents to other translations, a simple click on the desired translation and the tool will show the relevant documents to the selected translation on the same page.

- Lack of information in the target language: Some users, who have a good knowledge in the target language, would like to see information for the translation in the target language "gloss".

  – We made this request available in the new design so the user has a possibility of seeing the information for the translation by clicking on the "Gloss" button. In order to give the user more confidence, the translation terms are highlighted in the gloss (see Figure 8.3).

- Interactive terms usefulness / Identifying related terms in the contextual information: Addressed the usefulness of the interactive terms in the contextual information that can be used to improve the translation. The suggested interactive terms by the tool, based on highly frequent co-occurrence data, in the context of the query, are in bold blue and underlined. In many cases these were helpful as the users mentioned, however, in some cases the users needed more terms than the ones suggested by the tool. These terms are color-coded blue and are found in the contextual information, which is displayed along with each proposed translation. Although these terms are found in the context of the user query, the users mentioned that these terms in many cases they don't lead to an improvement in the translation. This issue is interpretable, because currently, we display the contextual information by selecting the most relevant documents to the translation. These documents might have terms which have a very low co-occurrence score with the query and in the corpora as a whole. If the user selects one of these terms,


multi Searcher - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://localhost:8080//Interactive_Cross_Lingual_Retrieval/MainMenu.html
Softonic Deutsch Customized Web Search
[Mult]iSearcher
Google
Search
Fertig
powered by Google

auto steuer

○ Direct Search
● Cross-Lingual Search
German English

Interactive related terms: [fahren, programm, fahre, gefahren, kuh, oil, zukunft, liter, heilige, geld, tobin, verkehrsmittel, flugzeug]

Information in the target language

If the decision-making process occurs much later than necessary, then it is impossible for the **car** industry to accommodate this adequately. It would therefore be a matter of transforming excise **duty** on **car** fuels into a type of eco-**tax** that would be higher for fossil fuels than for renewable fuels due to the punitive external costs of fossil fuels.

Ranked translation

1- car duty Gloss

Wenn die Beschlußfassung viel später erfolgt als nötig, hat die **Auto**mobilindustrie keine Möglichkeiten mehr, angemessen darauf zu reagieren. Es sollte also darum gehen, die Verbrauch**steuer** auf Kraftstoffe zu einer Art Öko**steuer** zu entwickeln, die die fossilen Kraftstoffe aufgrund ihrer negativen Externalitäten stärker belastet als die erneuerbaren Kraftstoffe.

2- auto tax Gloss

In Anbetracht der Tatsache, daß die "Auto-Öl" -Richtlinie die Aufgabe des verbleiten Benzins bis zum Jahr . . . mehr anzeigen

3- car tax Gloss

Die Richtlinie des Rates über die Besteuerung von Personenkraftwagen ist für den Binnenmarkt notwendig und sollte . . . mehr anzeigen

4- automobile tax Gloss

Das bedeutet, dass die Kommission die ab dem Jahr 2008 vorgesehene Anpassung der jährlichen Kraftfahrzeugsteuerforderungen auf der . . . mehr anzeigen

5- car steering wheel Gloss

Wenn Sie Schwierigkeiten haben, in die Garage zu gelangen, denken Sie an mich, der ich mit . . . mehr anzeigen

Web Results **1** - **10** of about **27,100,000** for **car duty**. (**0.05** seconds)

Redirect Carfueldata tools
17 Jun 2011 **...** To find out about the fuel consumption and CO2 emissions of a particular make of **car**, or to see the cost of vehicle **tax**, go to: **...**
www.dft.gov.uk/vca/fcb/**car**fueldata-tools-redirect-page.asp

Car Buyers Guide - Advice : **car tax** bands - road tax bands based **...**
Known as Vehicle Excise Duty (VED) or Graduated VED, the amount of **car tax** (annual road tax) you pay depends on the engine size or CO2 emissions depending **...**
www.theaa.com/motoring_advice/**car**-buyers.../cbg_road**tax**.html - Similar

The cost of vehicle **tax** for **cars**, motorcycles, light goods **...**
Tax tables to calculate the cost of your vehicle tax (**car tax**)
www.direct.gov.uk/en/.../HowTo**Tax**YourVehicle/DG_10012524 - Similar

Calculate your **car's tax** and CO2 emissions : Directgov - Do it online
Find out the CO2 emission for your **car** and the amount of vehicle **tax** payable **...**
www.direct.gov.uk/en/Diol1/DoItOnline/DG_10015994 - Similar

UK Fuel **Duty** - How Much of Your Petrol Diesel or LPG Purchase Goes **...**
Use the form below to find out your car CO2 emissions and hence which **car tax** band you are in. The form also shows you how many miles to the gallon you get. **...**
www.whatgas.com/fuel-**duty**.aspx

Exemption of Excise **Duties** on **Cars** - EU (**Duty** Free)
Car M O T, Driving, Registration, Import, **Car tax**, Cyprus, Rent Properties, Property Paphos.
www.paphosfinder.com/**car**/**duty**-free-**car**.htm - Similar

Vehicle Excise **Duty** - Wikipedia, the free encyclopedia
Vehicle Excise Duty (also VED, vehicle tax, **car tax** and road tax) is a vehicle road use tax levied as an excise duty which must be paid for most types of **...**
en.wikipedia.org/wiki/Vehicle_Excise_**Duty** - Similar

Heavy **Duty Car** Covers
Heavy **Duty Car** Covers Fully 100% totally waterproof grey vinyl with a soft cotton lining. These covers are UV, frost treated and have welded seams so water **...**
www.**car**covershop.co.uk/heavy-**duty**-**car**-covers.asp - Similar

Car Tax Bands 2011/2012 | Road Tax Prices UK | Next Green Car
Budget 2011: Latest UK **Car Tax** and Fuel Duty updates for 2011-12. Find your car's Vehicle Tax Band or search for a car by VED band based on CO2 emissions.
www.nextgreen**car**.com/road**tax**.php

Heavy **Duty Car** Booster Cables 16mm2. Reviews, Heavy **Duty Car ...**
Argos customers' reviews and ratings for Heavy **Duty Car** Booster Cables 16mm2.. Read and compare experiences real customers have had with the Heavy **Duty Car ...**
reviews.argos.co.uk/1493-en_gb/7402175/reviews.htm - Similar

1 2 3 4 5 6 7 8 9 10 Next

Figure 8.3: The new interface design with the initial suggested translation (the retrieved documents obtained by Google).

which will be added to the user query, this term may not have enough statistical co-occurrences data needed to improve the translation.

Another deficit which has been reported by some users is the lack of identifying terms in the contextual information e.g., currently terms related to the user query terms are binary compared e.g., (universitätsabschluss "university degree") wouldn't be recognized as relevant to (universität "university").

– In the new design the contextual information is only used for translation confidence and a list of suggested terms to improve the translation is provided independently from the contextual information. In the old design, the interactive terms are obtained from the contextual information, which is a few documents in size. In the new design, the corpora is used as a whole, to obtain these suggested terms. Only terms that have a significant co-occurrence score, with the query terms, will be suggested.

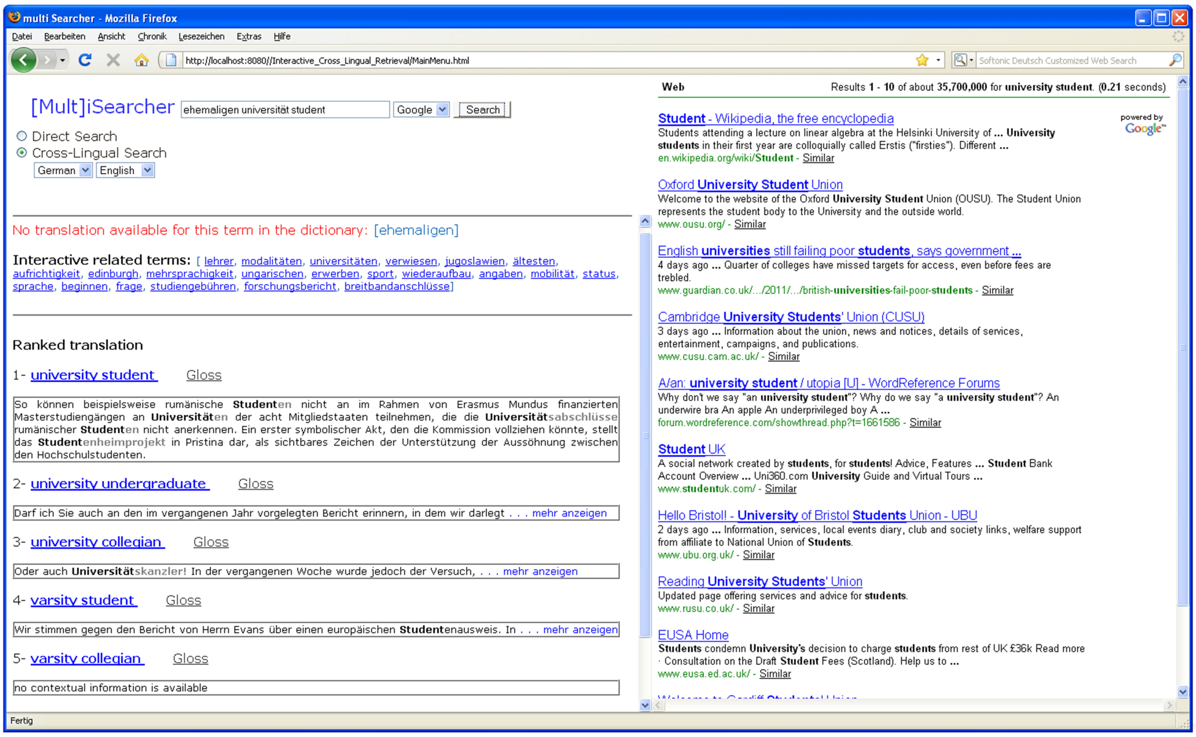
  In order to tackle the second deficit which is the identifying of terms in the contextual information, we highlight the terms which exist in the user query with bold black and the remaining character(s)/word, which forms a synonym for the given query term with light grey (selecting these highlighting mechanism can avoid blind color people issue). As Figure 8.4 shows that (universitätsabschluss "university degree") will be related to (universität "university") and thus "universitätsabschluss" will be highlighted.

- - Lack of detailed error notification: Users mentioned that there is a lack of detailed error notification being displayed when some error occurred e.g., when there is no translation available from the dictionary for some term/terms, the tool wouldn't notify the user. Instead the tool would show a message that there were no translations available. This results in confusion as to whether there is no translation available for a term/terms or whether the translation algorithm couldn't find enough statistical co-occurrence data to perform the translation.

    - In the new design, we tackled this issue by notifying the user that there is no translation available for a term/terms from the dictionary. In order to simplify the user task, the tool will automatically translate the rest of the terms. However, the rest of the terms must be at least two terms so the translation can be performed. As is shown in Figure 8.4, the user submits the German query "ehemaligen Universität Student" , the tool notified the user that there is no translation available for the term "ehemaligen" from the bilingual dictionary and at the same time the tool provides an automatic translation for the rest of the terms which is "university student".

  - Lack of interaction mechanism: Users mentioned that there is a lack of control when they interact with the tool e.g., when the user selects a term/terms to improve the translation, the user has no possibility of removing this term, if he/she discovers that the selected term/terms doesn't improve the translation. In the old design, the user tackled this issue by resubmitting the original query. However, this results in wasted time and effort. Another lack of interaction mechanism is that the user would like to see all terms interacted with, in the past and have the option of removing any term/terms selected, if needed.

    Another lack of interaction mechanism is in order to improve the translation, the user has to select an interactive term/terms from the contextual information. This term is automatically added to the user query. This query will be resubmitted and new translations, based on the selected term/terms, will be provided. This mechanism is not welcomed by the user as it will enlarge the


multi Searcher - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://localhost:8080//Interactive_Cross_Lingual_Retrieval/MainMenu.html
Softonic Deutsch Customized Web Search
Google
Search
Fertig
powered by Google™

[Mult]iSearcher ehemaligen universität student

○ Direct Search
● Cross-Lingual Search
German English

No translation available for this term in the dictionary: [ehemaligen]

Interactive related terms: [lehrer, modalitäten, universitäten, verwiesen, jugoslawien, ältesten, aufrichtigkeit, edinburgh, mehrsprachigkeit, ungarischen, erwerben, sport, wiederaufbau, angaben, mobilität, status, sprache, beginnen, frage, studiengebühren, forschungsbericht, breitbandanschlüsse]

Ranked translation

1- university student Gloss

So können beispielsweise rumänische **Student**en nicht an im Rahmen von Erasmus Mundus finanzierten Masterstudiengängen an **Universität**en der acht Mitgliedstaaten teilnehmen, die die **Universität**sabschlüsse rumänischer **Student**en nicht anerkennen. Ein erster symbolischer Akt, den die Kommission vollziehen könnte, stellt das **Student**enheimprojekt in Pristina dar, als sichtbares Zeichen der Unterstützung der Aussöhnung zwischen den Hochschulstudenten.

2- university undergraduate Gloss

Darf ich Sie auch an den im vergangenen Jahr vorgelegten Bericht erinnern, in dem wir darlegt . . . mehr anzeigen

3- university collegian Gloss

Oder auch **Universität**skanzler! In der vergangenen Woche wurde jedoch der Versuch, . . . mehr anzeigen

4- varsity student Gloss

Wir stimmen gegen den Bericht von Herrn Evans über einen europäischen **Student**enausweis. In . . . mehr anzeigen

5- varsity collegian Gloss

no contextual information is available

Web Results **1** - **10** of about **35,700,000** for **university student**. (**0.21** seconds)

Student - Wikipedia, the free encyclopedia
Students attending a lecture on linear algebra at the Helsinki University of **...** **University students** in their first year are colloquially called Erstis ("firsties"). Different **...**
en.wikipedia.org/wiki/**Student** - Similar

Oxford **University Student** Union
Welcome to the website of the Oxford **University Student** Union (OUSU). The Student Union represents the student body to the University and the outside world.
www.ousu.org/ - Similar

English **universities** still failing poor **students**, says government **...**
4 days ago **...** Quarter of colleges have missed targets for access, even before fees are trebled.
www.guardian.co.uk/.../2011/.../british-**universities**-fail-poor-**students** - Similar

Cambridge **University Students**' Union (CUSU)
3 days ago **...** Information about the union, news and notices, details of services, entertainment, campaigns, and publications.
www.cusu.cam.ac.uk/ - Similar

A/an: **university student** / utopia [U] - WordReference Forums
Why don't we say "an **university student**"? Why do we say "a **university student**"? An underwire bra An apple An underprivileged boy A **...**
forum.wordreference.com/showthread.php?t=1661586 - Similar

Student UK
A social network created by **students**, for **students**! Advice, Features **...** **Student** Bank Account Overview **...** Uni360.com **University** Guide and Virtual Tours **...**
www.**student**uk.com/ - Similar

Hello Bristol! - **University** of Bristol **Students** Union - UBU
2 days ago **...** Information, services, local events diary, club and society links, welfare support from affiliate to National Union of **Students**.
www.ubu.org.uk/ - Similar

Reading **University Students**' Union
Updated page offering services and advice for **students**.
www.rusu.co.uk/ - Similar

EUSA Home
Students condemn **University's** decision to charge **students** from rest of UK £36k Read more · Consultation on the Draft **Student** Fees (Scotland). Help us to **...**
www.eusa.ed.ac.uk/ - Similar

Figure 8.4: The re-ranked translation based on the user interaction (the retrieved documents obtained by Google).

query each time the user interacts with the tool and selects relevant terms to improve the translation. The majority of the users prefer not to revise the original query they submit and only wish to rank the initially obtained translations. In the new design, we tackled this issue by providing the user with more control, in that any term/terms can be selected/deselected by a simple mouse click. The tool will then immediately respond to any action by the user e.g., selecting a term would result in performing the re-ranking process (with the contextual information and the gloss for each translation) and alternatively, deselecting a term would return back to a previous state.

In order to deal with the lack of improving the translation, we offered the user to use the interactive term/terms only for ranking purposes and they will not be added to the original query. This suggestion was welcomed by the user which we took into account in the new design. Figure 8.5 shows that the term "fahren" was used just for the ranking purpose where it gives the translation "car steering wheel" an advantage to move from fifth place (see Figure 8.4) into first place (see Figure 8.5), without adding the term "fahren" to the original query "auto steuer".

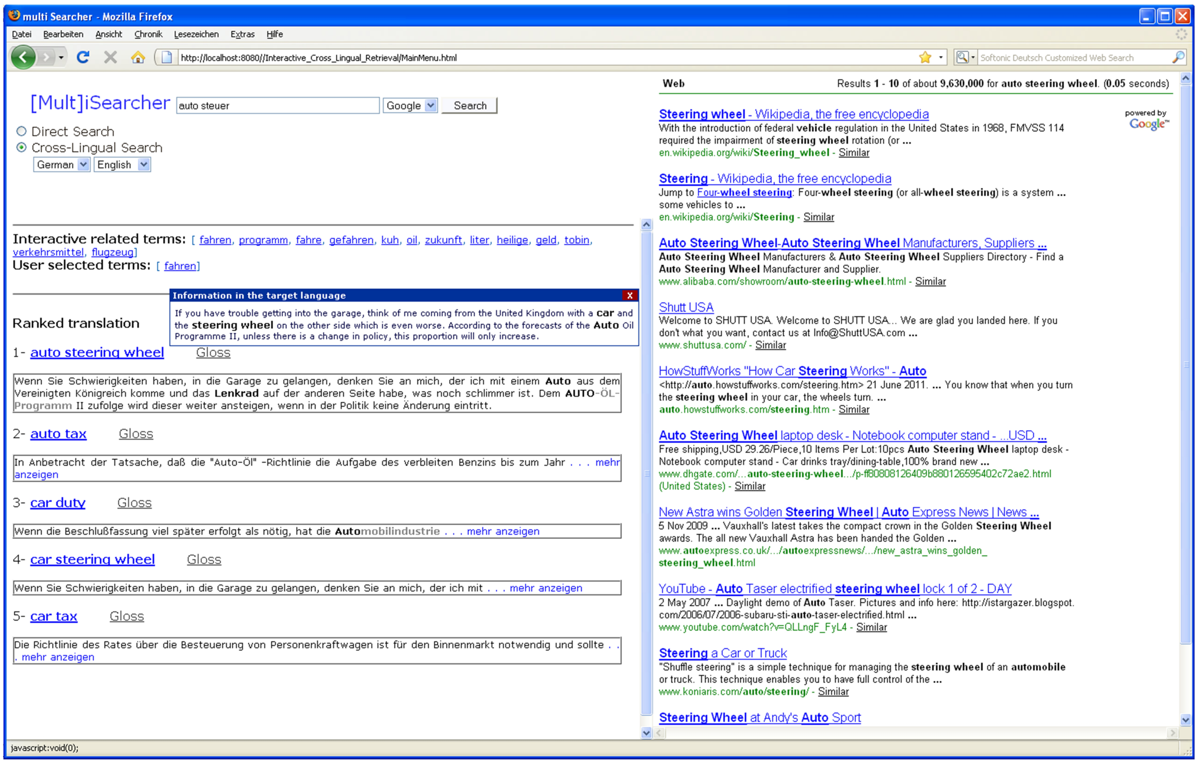
8.5 Conclusion

We designed a cross-lingual interactive tool in order to investigate the feasibility and the validity of utilizing translations for cross-lingual retrieval. To ensure that a user has a certain confidence in selecting a translation, which he/she possibly cannot even read or understand, the designed tool provides sufficient information about translation alternatives and their meaning so that the user has a certain degree of confidence in the translation. Based on the cross-lingual tool literature review, we identified several issues and shortcomings, which we have tackled in the cross-lingual tool proposed in this chapter. We proposed a smooth design that is on the one hand supported by significant back-end components and on the other hand gives the user some control over the query translation. Based on the tackled research issues, we designed the required interface components to tackle each identified research problems. These interface components are integrated together in order to perform the cross-lingual task, from submitting the query till getting the relevant documents. The proposed cross-lingual tool considers the user as an integral part of the cross-lingual process, in that the user can interact with the cross-lingual tool in a way that allows him/her improve the translation and thus improve the cross-lingual retrieval process. We conducted a broad user study to consider more points of interest and identify more issues in the first prototype which are tackled in the second prototype.

Chapter 9

Prototype Evaluation

In addition to the evaluation performed in Chapter 7 for Arabic and English languages, here in this evaluation, we were interested to evaluate the accuracy of the disambiguation algorithm for more languages e.g., English and German. Furthermore, we also were interested in evaluating whether the support provided by our cross-lingual tool is significant enough to guide the user in improving the translation and thus improve the performance of the cross-lingual retrieval.


multi Searcher - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://localhost:8080//Interactive_Cross_Lingual_Retrieval/MainMenu.html
Softonic Deutsch Customized Web Search
Google
Search
powered by Google™
javascript:void(0);

[Mult]iSearcher auto steuer

○ Direct Search
● Cross-Lingual Search
German English

Interactive related terms: [fahren, programm, fahre, gefahren, kuh, oil, zukunft, liter, heilige, geld, tobin, verkehrsmittel, flugzeug]
User selected terms: [fahren]

Information in the target language

If you have trouble getting into the garage, think of me coming from the United Kingdom with a **car** and the **steering wheel** on the other side which is even worse. According to the forecasts of the **Auto** Oil Programme II, unless there is a change in policy, this proportion will only increase.

Ranked translation

1- auto steering wheel Gloss

Wenn Sie Schwierigkeiten haben, in die Garage zu gelangen, denken Sie an mich, der ich mit einem **Auto** aus dem Vereinigten Königreich komme und das **Lenkrad** auf der anderen Seite habe, was noch schlimmer ist. Dem **AUTO**-ÖL-Programm II zufolge wird dieser weiter ansteigen, wenn in der Politik keine Änderung eintritt.

2- auto tax Gloss

In Anbetracht der Tatsache, daß die "Auto-Öl" -Richtlinie die Aufgabe des verbleiten Benzins bis zum Jahr . . . mehr anzeigen

3- car duty Gloss

Wenn die Beschlußfassung viel später erfolgt als nötig, hat die **Auto**mobilindustrie . . . mehr anzeigen

4- car steering wheel Gloss

Wenn Sie Schwierigkeiten haben, in die Garage zu gelangen, denken Sie an mich, der ich mit . . . mehr anzeigen

5- car tax Gloss

Die Richtlinie des Rates über die Besteuerung von Personenkraftwagen ist für den Binnenmarkt notwendig und sollte . . . mehr anzeigen

Web Results **1** - **10** of about **9,630,000** for **auto steering wheel**. (**0.05** seconds)

Steering wheel - Wikipedia, the free encyclopedia
With the introduction of federal **vehicle** regulation in the United States in 1968, FMVSS 114 required the impairment of **steering wheel** rotation (or **...**
en.wikipedia.org/wiki/**Steering_wheel** - Similar

Steering - Wikipedia, the free encyclopedia
Jump to Four-**wheel steering**: Four-**wheel steering** (or all-**wheel steering**) is a system **...** some vehicles to **...**
en.wikipedia.org/wiki/**Steering** - Similar

Auto Steering Wheel-**Auto Steering Wheel** Manufacturers, Suppliers ...
Auto Steering Wheel Manufacturers & **Auto Steering Wheel** Suppliers Directory - Find a **Auto Steering Wheel** Manufacturer and Supplier.
www.alibaba.com/showroom/**auto-steering-wheel**.html - Similar

Shutt USA
Welcome to SHUTT USA. Welcome to SHUTT USA... We are glad you landed here. If you don't what you want, contact us at Info@ShuttUSA.com **...**
www.shuttusa.com/ - Similar

HowStuffWorks "How Car **Steering** Works" - **Auto**
<http://**auto**.howstuffworks.com/steering.htm> 21 June 2011. **...** You know that when you turn the **steering wheel** in your car, the wheels turn. **...**
auto.howstuffworks.com/**steering**.htm - Similar

Auto Steering Wheel laptop desk - Notebook computer stand - ...USD ...
Free shipping,USD 29.26/Piece,10 Items Per Lot:10pcs **Auto Steering Wheel** laptop desk - Notebook computer stand - Car drinks tray/dining-table,100% brand new **...**
www.dhgate.com/...**auto-steering-wheel**.../p-ff80808126409b880126595402c72ae2.html (United States) - Similar

New Astra wins Golden **Steering Wheel** | **Auto** Express News | News ...
5 Nov 2009 **...** Vauxhall's latest takes the compact crown in the Golden **Steering Wheel** awards. The all new Vauxhall Astra has been handed the Golden **...**
www.**auto**express.co.uk/.../**auto**expressnews/.../new_astra_wins_golden_**steering_wheel**.html

YouTube - **Auto** Taser electrified **steering wheel** lock 1 of 2 - DAY
2 May 2007 **...** Daylight demo of **Auto** Taser. Pictures and info here: http://istargazer.blogspot.com/2006/07/2006-subaru-sti-**auto**-taser-electrified.html **...**
www.youtube.com/watch?v=QLLngF_FyL4 - Similar

Steering a Car or Truck
"Shuffle steering" is a simple technique for managing the **steering wheel** of an **automobile** or truck. This technique enables you to have full control of the **...**
www.koniaris.com/**auto**/**steering**/ - Similar

Steering Wheel at Andy's **Auto** Sport

Figure 8.5: The re-ranked translation based on the user interaction (the retrieved documents obtained by Google).

9.1 English-German Evaluation

Different from the evaluation performed in Chapter 7, where the test queries have multiple quite different meanings, here, in order to have a challenged evaluation, we selected 100 test instances of polysemous words from one of the most popular Word Sense Disambiguation evaluation data sets (SemEval 2010)[1] (Lefever and Hoste, 2010). It is very difficult to disambiguate polysemous words as they have separate different meanings that are related to one another. For example, the English polysemous word "plant" can have these related meanings in German "gewächs", "pflanze", "vegetation" etc. Another example is the English polysemous word "passage" that has these separate related meanings in German "Durchgang", "Durchtritt", "Durchfahrt", "Durchlass", "Überfahrt", "Verlauf" , etc. Furthermore, disambiguating polysemous words is a very difficult task with scores being very close to the baseline measure (van Gompel, 2010).

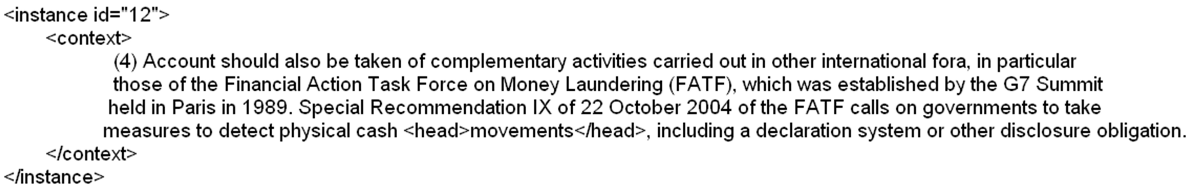
This evaluation is performed in particular, to check whether the proposed cross-lingual tool, in this thesis, is significant to select the correct translation corresponding to the given polysemous word, in the source language.

[1] http://webs.hogent.be/ elef464/lt3_SemEval.html#_subtasks

9.1.1 Experiment Setup

For the cross-lingual word sense disambiguation task in (Lefever and Hoste, 2010), numbers of English nouns were given. For each English noun 20 test instances were provided. For each test instance, possible translations in the target languages were also provided (hand-tagged Gold standards translations). There were two types of scoring the translations, one based on scoring the best translation and the other based on scoring the best 5 translations in the target languages. For our experiment, we used only the first, where we conducted the test only on selecting the best translation of the ambiguous word in the target language. There were two types of tests; one is a bilingual test, where the ambiguous word is translated to one target language or a multilingual test where the ambiguous word is translated into five languages (Dutch, French, German, Spanish and Italian). For our test, we selected the bilingual test where the test instances (queries) are in English and the translations are in German.

The test instances are long sentences where some of them are greater than 63 words in length, (see Figure 9.1) which do not fit into a real life scenario cross-lingual information retrieval, where the search engine queries are usually between 2.4 and 2.7 in length (Gabrilovich et al., 2009). In order to deal with this and have significant evaluation for our disambiguation algorithm, we adapted the test sentences and extracted only important words from each test instance. After removing stop words, for each test instance, important words were selected. This task has been performed by the users, 10 users, each has 10 test instances. They constructed their queries by selecting a few words which describe their needs in the test instances context. For the test instances shown in Figure 9.1, these words are selected: "physical", "cash", "movement". The users were requested to select as few words as possible to express their need. The new test instances (queries) ranged from 2 to 7 words in length with the average being 4.1.



```
<instance id="12">
      <context>
              (4) Account should also be taken of complementary activities carried out in other international fora, in particular
              those of the Financial Action Task Force on Money Laundering (FATF), which was established by the G7 Summit
              held in Paris in 1989. Special Recommendation IX of 22 October 2004 of the FATF calls on governments to take
             measures to detect physical cash <head>movements</head>, including a declaration system or other disclosure obligation.
      </context>
</instance>
```

Figure 9.1: Test instance number "12" for the ambiguous word "movement".

The Gold standards translations (in 5 languages) were extracted from the Europarl parallel corpora[2] (Koehn, 2005). Europarl parallel corpus is a collection of documents for 21 languages. These parallel corpora were extracted from the proceedings of the

[2] http://www.statmt.org/europarl/

European Parliament. To construct the Gold standard translations, human annotators were requested to select one of the automatically provided translations from the corpora. Each number in front of each possible translation reflects the number of times this translation was picked up by the human annotators. Figure 9.2 shows the Gold standard translations for the first 10 test instances for the polysemous word "occupation".

The test sentences (in English) were selected from the JRC-ACQUIS multilingual parallel corpora[3]. The JRC-ACQUIS multilingual parallel corpora is the total body of European law that are applicable in European members states. Currently this corpora is a collection of text written from 1950 up to now, however this text is growing continuously.

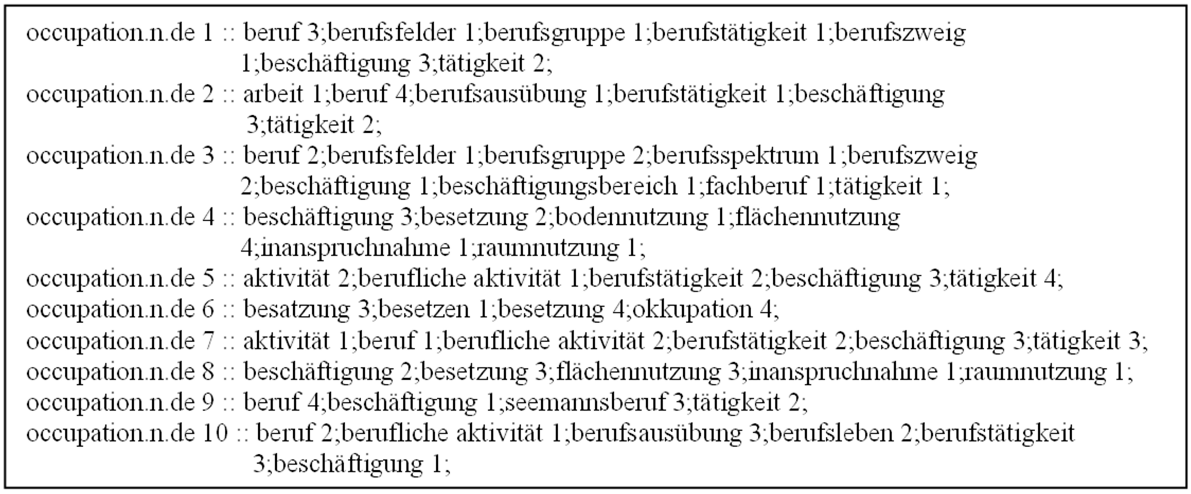
There were 2 test data: first, development test data which contains 5 polysemous nouns (occupation, passage, movement, plant and bank) each were provided with 20 test instances. Second, test data which contain 50 English nouns for each 20 test instances were provided. For our experiment, due to the adaption of this test data to fit for an actual cross-lingual scenario and due to the unavailability of the full test data, we could use only the development test data, so at the end we evaluated our disambiguation algorithm based on 100 test instances (queries).

9.1.2 Disambiguation Algorithm Evaluation

The goal of the evaluation was based on two perspectives, first to evaluate the performance of the disambiguation algorithm in German and English languages. In order to achieve this goal, we smoothly integrated more languages into the proposed tool in this thesis. The integration was a trivial task where only 2 steps were needed. We obtained an English-German dictionary and English-German parallel corpora from Europarl parallel corpora (Koehn, 2005). No modification for the disambiguation algorithm as well as for the contextual information algorithm was required. Second was to evaluate whether user interaction, could improve the performance of the disambiguation algorithm by selecting relevant term/terms proposed by the tool.

In order to evaluate the performance of the disambiguation algorithm, we used the precision measurement which is proposed by many researchers for a word sense disambiguation task e.g., (Dagan and Itai, 1994; Kang, 2003; Fakhrahmad et al., 2011). Precision is the proportion of the correctly disambiguated senses for the ambiguous word. In the gold standard translations, the translation that has a larger number associated with it compared to other possible translations are ranked first (see Figure 9.2).

[3] http://wt.jrc.it/lt/Acquis/

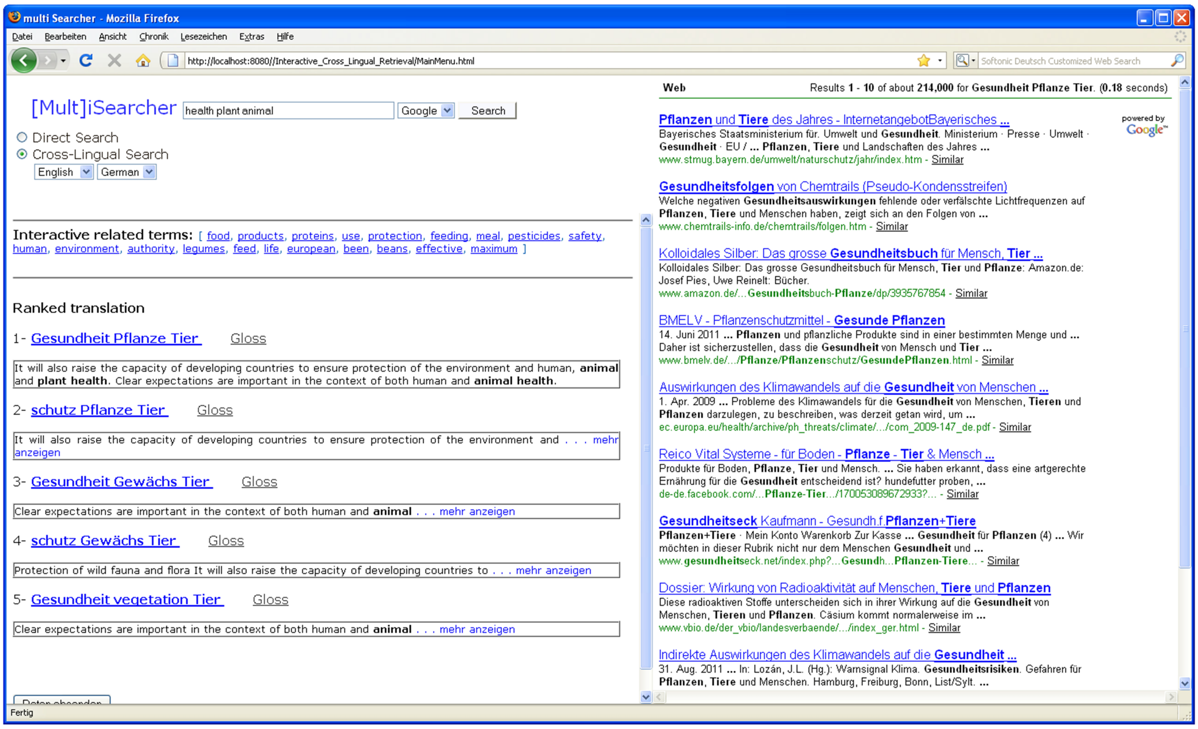


occupation.n.de 1 :: beruf 3;berufsfelder 1;berufsgruppe 1;berufstätigkeit 1;berufszweig 1;beschäftigung 3;tätigkeit 2;
occupation.n.de 2 :: arbeit 1;beruf 4;berufsausübung 1;berufstätigkeit 1;beschäftigung 3;tätigkeit 2;
occupation.n.de 3 :: beruf 2;berufsfelder 1;berufsgruppe 2;berufsspektrum 1;berufszweig 2;beschäftigung 1;beschäftigungsbereich 1;fachberuf 1;tätigkeit 1;
occupation.n.de 4 :: beschäftigung 3;besetzung 2;bodennutzung 1;flächennutzung 4;inanspruchnahme 1;raumnutzung 1;
occupation.n.de 5 :: aktivität 2;berufliche aktivität 1;berufstätigkeit 2;beschäftigung 3;tätigkeit 4;
occupation.n.de 6 :: besatzung 3;besetzen 1;besetzung 4;okkupation 4;
occupation.n.de 7 :: aktivität 1;beruf 1;berufliche aktivität 2;berufstätigkeit 2;beschäftigung 3;tätigkeit 3;
occupation.n.de 8 :: beschäftigung 2;besetzung 3;flächennutzung 3;inanspruchnahme 1;raumnutzung 1;
occupation.n.de 9 :: beruf 4;beschäftigung 1;seemannsberuf 3;tätigkeit 2;
occupation.n.de 10 :: beruf 2;berufliche aktivität 1;berufsausübung 3;berufsleben 2;berufstätigkeit 3;beschäftigung 1;

Figure 9.2: Gold standard translation for the polysemous word "occupation" based on human annotators.

We compared the result of the disambiguation algorithm only with this translation i.e., if the first ranked translation was selected by 4 human annotators and the second was selected by only 3 human annotators, we consider only the first as correct even if for algorithm proposes the second ranked translation as the correct one. In order to give the user wide possibilities to interact with the tool, the tool provides the user with 5 ranked translations along with their contextual information. Furthermore, a list of possible interactive related terms, to the user query, is presented to the user. We assumed the tool translation correct when it is displayed within the 5 ranked translations provided by the tool. Figure 9.3 shows one of the test instances "health plant animal" for the polysemous word "plant". The tool successfully presented to the user for the polysemous word "plant", the tow correct senses (based on human annotators disambiguation) "pflanze" in fist rank, and "Gewächs" in third rank.

Table 9.1 shows the submitted query, number of possible senses, the correct sense (based on the human annotators' selection), the algorithm automatic disambiguation (without user interaction), rank (the disambiguation algorithm provide 5 ranked translations). Table 9.2 shows the list of translations that the tool could successfully provide after the user interaction. It shows, the interactive term (selected by the user to improve the translation), interactive translation (new translations based on the selected user relevant term) and rank (the rank of translation). If we examine the precision shown in Table 9.1, for the ambiguous word (occupation), we find that the disambiguation algorithm gained up to 70% accuracy for the first ambiguous word. With the user interaction, the disambiguation algorithm improved by 5% (see Table 9.2).



http://localhost:8080//Interactive_Cross_Lingual_Retrieval/MainMenu.html

[Mult]iSearcher health plant animal

○ Direct Search
● Cross-Lingual Search
English German

Interactive related terms: [food, products, proteins, use, protection, feeding, meal, pesticides, safety, human, environment, authority, legumes, feed, life, european, been, beans, effective, maximum]

Ranked translation

1- Gesundheit Pflanze Tier Gloss

It will also raise the capacity of developing countries to ensure protection of the environment and human, **animal** and **plant health**. Clear expectations are important in the context of both human and **animal health**.

2- schutz Pflanze Tier Gloss

It will also raise the capacity of developing countries to ensure protection of the environment and . . . mehr anzeigen

3- Gesundheit Gewächs Tier Gloss

Clear expectations are important in the context of both human and **animal** . . . mehr anzeigen

4- schutz Gewächs Tier Gloss

Protection of wild fauna and flora It will also raise the capacity of developing countries to . . . mehr anzeigen

5- Gesundheit vegetation Tier Gloss

Clear expectations are important in the context of both human and **animal** . . . mehr anzeigen

Web Results **1** - **10** of about **214,000** for **Gesundheit Pflanze Tier**. (**0.18** seconds)

Pflanzen und **Tiere** des Jahres - InternetangebotBayerisches ...
Bayerisches Staatsministerium für. Umwelt und **Gesundheit**. Ministerium · Presse · Umwelt · **Gesundheit** · EU / ... **Pflanzen**, **Tiere** und Landschaften des Jahres ...
www.stmug.bayern.de/umwelt/naturschutz/jahr/index.htm - Similar

Gesundheitsfolgen von Chemtrails (Pseudo-Kondensstreifen)
Welche negativen **Gesundheitsauswirkungen** fehlende oder verfälschte Lichtfrequenzen auf **Pflanzen**, **Tiere** und Menschen haben, zeigt sich an den Folgen von ...
www.chemtrails-info.de/chemtrails/folgen.htm - Similar

Kolloidales Silber: Das grosse **Gesundheitsbuch** für Mensch, **Tier** ...
Kolloidales Silber: Das grosse Gesundheitsbuch für Mensch, **Tier** und **Pflanze**: Amazon.de: Josef Pies, Uwe Reinelt: Bücher.
www.amazon.de/...**Gesundheits**buch-**Pflanze**/dp/3935767854 - Similar

BMELV - Pflanzenschutzmittel - **Gesunde Pflanzen**
14. Juni 2011 ... **Pflanzen** und pflanzliche Produkte sind in einer bestimmten Menge und ... Daher ist sicherzustellen, dass die **Gesundheit** von Mensch und **Tier** ...
www.bmelv.de/.../**Pflanze**/**Pflanzen**schutz/**GesundePflanzen**.html - Similar

Auswirkungen des Klimawandels auf die **Gesundheit** von Menschen ...
1. Apr. 2009 ... Probleme des Klimawandels für die **Gesundheit** von Menschen, **Tieren** und **Pflanzen** darzulegen, zu beschreiben, was derzeit getan wird, um ...
ec.europa.eu/health/archive/ph_threats/climate/.../com_2009-147_de.pdf - Similar

Reico Vital Systeme - für Boden - **Pflanze** - **Tier** & Mensch ...
Produkte für Boden, **Pflanze**, **Tier** und Mensch. ... Sie haben erkannt, dass eine artgerechte Ernährung für die **Gesundheit** entscheidend ist? hundefutter proben, ...
de-de.facebook.com/...**Pflanze-Tier**.../170053089672933?... - Similar

Gesundheitseck Kaufmann - Gesundh.f.**Pflanzen**+**Tiere**
Pflanzen+Tiere · Mein Konto Warenkorb Zur Kasse ... **Gesundheit** für **Pflanzen** (4) ... Wir möchten in dieser Rubrik nicht nur dem Menschen **Gesundheit** und ...
www.**gesundheits**eck.net/index.php?...**Gesundh**...**Pflanzen-Tiere**... - Similar

Dossier: Wirkung von Radioaktivität auf Menschen, **Tiere** und **Pflanzen**
Diese radioaktiven Stoffe unterscheiden sich in ihrer Wirkung auf die **Gesundheit** von Menschen, **Tieren** und **Pflanzen**. Cäsium kommt normalerweise im ...
www.vbio.de/der_vbio/landesverbaende/.../index_ger.html - Similar

Indirekte Auswirkungen des Klimawandels auf die **Gesundheit** ...
31. Aug. 2011 ... In: Lozán, J.L. (Hg.): Warnsignal Klima. **Gesundheitsrisiken**. Gefahren für **Pflanzen**, **Tiere** und Menschen. Hamburg, Freiburg, Bonn, List/Sylt. ...

Figure 9.3: The cross-lingual retrieval for one of the test instances "health plant animal" for polysemous word "plant".

For example, the algorithm failed to provide the correct translation for the user query "high occupation rate" but after the user interaction and the selection of the proposed relevant term "unemployment", the disambiguation algorithm could provide the user with the correct translation "besetzung".

S/N	Query	No of senses	Correct sense	Tool automatic disambiguation	Rank
1	workers qualification coal steel occupation	7	(beruf, beschäftigung)	(beruf, beschäftigung)	(2,5)
2	choice occupation training work place	6	beruf	beruf	2
3	employment occupation	9	(beruf, berufsgruppe, berufszweig)	beruf	1

continued on next page

continued from previous page

S/N	Query	No of senses	Correct sense	Tool automatic disambiguation	Rank
4	building land occupation activities	6	(flächennutzung, beschäftigung)	beschäftigung	2
5	occupation repair maintenance	6	tätigkeit	tätigkeit	1
6	military occupation	4	(besetzung, okkupation)	okkupation	1
7	office administrative duties occupation	7	(beschäftigung, tätigkeit)	beschäftigung	1
8	interest premises occupation	5	(besetzung, flächennutzung)	-	-
9	professional education apply occupation	4	beruf	beruf	1
10	rules course occupation	7	berufsausübung	berufsausübung	1
11	name address occupation person	4	(beruf, beschäftigung)	beruf	2
12	farmland occupation	5	bodennutzung	bodennutzung	3
13	flood system occupation area	5	bodennutzung	bodennutzung	5
14	high occupation rate	3	besetzung	-	-
15	occupation territories peace	5	(besetzung, okkupation)	besetzung	1
16	payment provision regular occupation business	8	(beruf, beschäftigung, tätigkeit)	-	-
17	providing occupation cares environment retains population	7	tätigkeit	-	-

continued on next page

continued from previous page

S/N	**Query**	**No of senses**	**Correct sense**	**Tool automatic disambiguation**	**Rank**
18	occupation colonies	6	okkupation	okkupation	2
19	working address occupation exercised	7	(aktivität, beschäftigung)	aktivität	1
20	access protocol occupation rules technical characteristics	3	besetzung	-	-
Overall average Precision: 70%					

Table 9.1: The disambiguation result for the 20 test instances for the ambiguous word "occupation".

S/N	**Query**	**Correct sense**	**Interactive term**	**Tool interactive disambiguation**	**Rank**
14	high occupation rate	besetzung	unemployment	besetzung	3
Precision improved by: 5%					

Table 9.2: Improved translation for the test instance number "14" for the ambiguous word "occupation" after the user interaction.

Table 9.3 shows that for the ambiguous word "plant", the algorithm without the user interaction could provide the correct translation for 11 test instances out of 20 and gained accuracy up to 60%. However, with the user interaction, the disambiguation algorithm is improved by 20 % (see Table 9.4) and gained an overall accuracy average of 80%.

S/N	**Query**	**No of senses**	**Correct sense**	**Tool automatic disambiguation**	**Rank**
1	health plant	8	pflanzenschutz	pflanzenschutz	5
2	health plant animal	4	(pflanze,Gewächs)	(Gewächs, pflanze)	(1,3)
3	products equipment plant technology production	8	anlage	anlage	1

continued on next page

continued from previous page

S/N	**Query**	**No of senses**	**Correct sense**	**Tool automatic disambiguation**	**Rank**
4	water power plant ecosystems	5	(wasserkraft-anlage, wasserkraftwerk)	wasserkraftwerk	2
5	plant pesticide agriculture	10	(pflanzenschutz - mittel, pflanzenschutzprodukt)	pflanzenschutzmittel	1
6	plant installation offshore activities	6	anlage	anlage	1
7	environment water air soil plant animals	4	(gewächs, pflanze)	pflanze	2
8	nuclear plant electricity	5	kernkraftwerk	-	-
9	cutting boning meat plant examination	8	anlage	anlage	1
10	transport carcases processing plant	9	(verbrennungs-anlage, verarbeitungsbetrieb, verarbeitungsanlage)	verarbeitungsanlage	1
11	manufacture plant staff technology	8	anlage	anlage	1
12	seamless tubes power plant	8	kraftwerk	-	-
13	flood system occupation area	5	bodennutzung	bodennutzung	5
14	production plant closures	9	(betriebs-schließung, betriebsstillegung)	betriebsschließung	4

continued on next page

continued from previous page					
S/N	**Query**	**No of senses**	**Correct sense**	**Tool automatic disambiguation**	**Rank**
15	plant firm state	7	fabrik	-	-
16	plant machinery tools	6	fabrik	-	-
17	operator combustion plant	6	großfeuerungs-anlage	-	-
18	pharmacognosy plant animal	4	pflanze	-	-
19	products fish plant human consumption	9	betrieb	-	-
20	wastewater treatment plant	5	abwasserklär- anlage	-	-
Overall average Precision: 60%					

Table 9.3: The disambiguation result for the 20 test instances for the ambiguous word "plant".

S/N	Query	Correct sense	Interactive term	Tool interactive disambiguation	Rank
8	nuclear plant electricity	kernkraftwerk	saftey	kernkraftwerk	2
13	plant machinery production	anlage	imports	anlage	1
16	plant machinery tools	fabrik	directive	fabrik	4
18	pharmacognosy plant animal	pflanze	human	pflanze	5
Precision improved by: 20%					

Table 9.4: Improved translation for the test instance number "8", "13","16" and "18" for the ambiguous word "plant" after the user interaction.

For the ambiguous word "movement" as Table 9.5 shows, the algorithm gained up to 65% accuracy and with the user interaction, the disambiguation algorithm improved by

5% (see Table 9.6) and gained an overall average of 70%.

S/N	Query	No of senses	Correct sense	Tool automatic disambiguation	Rank
1	goods movement frontier traffic	8	(güterverkehr, transport, warenbewegung)	güterverkehr	1
2	student free movement residence	8	freizügigkeit	-	-
3	atmospheric movements effects environment	4	bewegung	bewegung	1
4	items movement transfer trading	9	vekehr	-	-
5	capital movement	5	kapitalbewegung	kapitalbewegung	4
6	entering leaving movements	10	(tiertransport, verbringung)	-	-
7	entering leaving animals movement	10	verkehr	-	-
8	transit movements territory	9	(bewegen, freizügigkeit)	bewegen	1
9	border crossing movement	5	schwankung	schwankung	5
10	variation price movement	5	wechselkursschwankung	-	-
11	rebel movements conflict	5	(rebellenbewegung, rebellenorganisation)	rebellenbewegung	1
12	physical cash movements	4	(geldbewegung, zahlungsverkehr)	geldbewegung	1
13	democratic movement	8	bewegung	bewegung	2
14	harassment political movements	8	bewegung	bewegung	1

continued on next page

continued from previous page					
S/N	**Query**	**No of senses**	**Correct sense**	**Tool automatic disambiguation**	**Rank**
15	future movement	7	bewegung	bewegung	5
16	troops movement	5	truppenbewegung	truppenbewegung	1
17	secondary movements applicants asylum	6	sekundärmigration	-	-
18	monitoring migratory movements	8	(migration, migrationsbewegung, wandering, wanderungsbewegung)	(Wanderungsbewegung, migration)	(1,5)
19	free transport unrestricted movement	10	(bewegen, freizügigkeit)	-	-
20	reinforcement controls movements ovine animals	7	verbringung	verbringung	2
Overall average Precision: 65%					

Table 9.5: The disambiguation result for the 20 test instances for the ambiguous word "movement".

S/N	**Query**	**Correct sense**	**Interactive term**	**Tool interactive disambiguation**	**Rank**
19	free transport unrestricted movement	(bewegen, freizügigkeit)	goods	(bewegen, freizügigkeit)	(2,5)
Precision improved by: 5%					

Table 9.6: Improved translation for the test instance number "19" for the ambiguous word "movement" after the user interaction.

Table 9.7 shows that the disambiguation algorithm for the ambiguous word "passage" gained up to 50% and with the user interaction, the disambiguation algorithm improved by 15% (see Table 9.8)) and gained an overall average of 65%.

For the ambiguous word "bank", the disambiguation algorithm could disambiguate

12 out of 20 test instances with accuracy being 60% (see Table 9.9).

S/N	Query	No of senses	Correct sense	Tool automatic disambiguation	Rank
1	transport document covering passage	5	durchfahrt	durchfahrt	3
2	UNDERGROUND PASSAGE	6	(durchgangsroute, gang)	-	-
3	vessel passage authorities	6	durchreise	durchreise	3
4	frontiers passage fuel	9	(grenzübergang, grenzübertritt, zugang)	grenzübertritt	2
5	veterinarian products passage	6	warenverkehr	-	-
6	fisheries policy passage territorial sea	7	durchfahrt	durchfahrt	5
7	export lading passage transit	5	(durchfahrt, durchreise)	durchfahrt	3
8	products criteria passage metal	2	passieren	passieren	1
9	time events passage provision	2	(laufe,verstrichen)	laufe	1
10	transitional measures passage	3	übergang	-	-
11	certain passages directive	6	passage	-	-
12	rule amendments passage	8	passage	-	-
13	envisaged passage approval process	8	behandlung	-	-
14	references text passage found	8	(passage, textpassage)	-	-
15	situation affected passage directive	9	annahme	-	-

continued on next page

continued from previous page

S/N	Query	No of senses	Correct sense	Tool automatic disambiguation	Rank
16	coasts straits passage	9	(durchfahrt, durchreise)	durchfahrt	4
17	ban torture subsequent passage approved	3	annahme	annahme	1
18	successful passage electronic customs	3	übergang	-	-
19	agricultural sector passage tropical storm	4	(durchfahrt, durchzug, passieren)	durchfahrt	3
20	income passage expert contract staff	3	übergang	-	-
Overall average Precision: 50%					

Table 9.7: The disambiguation result for the 20 test instances for the ambiguous word "passage".

S/N	Query	Correct sense	Interactive term	Tool interactive disambiguation	Rank
10	transitional measures passage	übergang	derogations	übergang	3
12	rule amendments passage	passage	system	passage	1
15	situation affected passage directive	annahme	justifiable	annahme	1
Precision improved by: 15%					

Table 9.8: Improved translation for the test instance number "10", "12" and "15" for the ambiguous word "passage" after the user interaction.

The accuracy could be better because some of the Gold standard translations are not direct translations for the given query terms. For example, we consider the query "palestinian people west bank", the proposed translation in the Gold standard for the "west bank" is "westjordanufer", the word "jordan" does not exist in the query. Therefore, our algorithm could propose only the translation "west ufer" which we consider not correct because it is not proposed in the Gold standard translations. Based on the user interaction, the disambiguation algorithm is improved by 10% (see Table 9.10), so in the end, the overall accuracy for the ambiguous word "bank" against 20 test instances is 70%.

S/N	**Query**	**No of senses**	**Correct sense**	**Tool automatic disambiguation**	**Rank**
1	economic bank international settlements financial	5	bank	bank	1
2	credit agreements creditor countries central bank	6	bank	bank	1
3	palestinian people west bank	6	westjordanufer	-	-
4	economic social development west bank gaza strip	6	westjordanufer	-	-
5	national waters lake bank	2	ufer	ufer	2
6	bank river	2	ufer	ufer	1
7	creditor pay bank credit balance	5	bank	bank	1
8	electronic data bank applications	4	datenbank	-	-
9	regulate bank liquidity	6	bank	bank	5
10	bank river stone wood ponds	2	ufer	-	-
11	west bank gaza strip	6	westjordanufer	-	-
12	budgetary support world bank	5	weltbank	weltbank	3
13	fisheries pay bank account	6	(bankkonto, konto)	-	-

continued on next page

continued from previous page

S/N	Query	No of senses	Correct sense	Tool automatic disambiguation	Rank
14	Hospital blood banks	2	blutbank	blutbank	1
15	private savings bank	4	sparkasse	-	-
16	business local bank holidays	2	(bankfeiertag, feiertag)	bankfeiertag	4
17	electronic money coin bank notes	3	(banknote, geldschein)	banknote	3
18	interest capital bank loan	8	(bankanleihe, bankdarlehen, bankkredit)	bankkredit	1
19	river bank shores lake	2	ufer	-	-
20	commercial investments financial bank	6	bank	bank	1
Overall average Precision: 60%					

Table 9.9: The disambiguation result for the 20 test instances for the ambiguous word "bank".

S/N	Query	Correct sense	Interactive term	Tool interactive disambiguation	Rank
13	fisheries pay bank account	(bankkonto, konto)	fees	bankkonto	4
15	private savings bank	sparkasse	services	sparkasse	1
Precision improved by: 10%					

Table 9.10: Improved translation for the test instance number "13" and "15" after the user interaction.

As Table 9.11 shows, the overall precision average of all test words, without the user interaction, is 62% against 100 test instances. The user interaction could improve the precision of the disambiguation algorithm by 11%.

Ambiguous word	**Precision**	**Improved precision by user interaction**	**Individual overall precision**
occupation	75%	5%	80%
plant	60%	20%	80%
movement	65%	5%	70%
passage	50%	15%	65%
Bank	60%	10%	70%
Overall average Precision	**62 %**	**11 %**	**73 %**

Table 9.11: Overall precision average of the disambiguation algorithm.

This indicates that providing the user with significant information can lead to an improvement in the translation. The disambiguation algorithm gained an overall precision average of 73%, which is a promising result in disambiguating polysemous words.

9.2 Conclusion

In addition to the evaluation performed in Chapter 7 for Arabic and English languages, here in this evaluation, we were interested to evaluate the accuracy of the disambiguation algorithm for more languages e.g., English and German. Furthermore, we also were interested in evaluating whether the support provided by our cross-lingual tool is significant enough to guide the user in improving the translation and thus improve the performance of the cross-lingual retrieval. Therefore, the new prototype has been used to evaluate the performance of the disambiguation algorithm for English/German language pairs. Based on experiments that we performed, our new design achieved significant results and could support the user to improve the performance of the disambiguation algorithm.

Part V

Concluding Remarks and Future Work Perspectives

Chapter 10

Concluding Remarks and Future Work Perspectives

10.1 Summary

The overall theme in this thesis is to advance the state of the art cross-lingual tool, particularly for less-studied languages e.g., Arabic. This has been achieved by researching two main research issues, word sense disambiguation and the user's lack of knowledge in the target languages. Solving or alleviating these two research problems is essential for any cross-lingual retrieval tool. The problem of word sense disambiguation has been considered for two languages pairs Arabic-English and English-German. In order to achieve the research goals, we first identified in-depth, by the literature review, the main research issues related to cross-lingual retrieval, and what has been achieved so far, to tackle these research issues. In order to build better tools to help people understand and use complex cross-lingual retrieval environments, we studied, in detail, the state-of-the art cross-lingual interaction tools that can be used to support the user to perform his/her cross-lingual search. Issues related to each one of the discussed tools were reported and were taken into account when designing the proposed tool in this thesis.

The spotlight of the work presented in this thesis builds on exploiting word correspondence across languages for Word Sense Disambiguation (WSD) and on exploiting parallel linguistic resources to overcome the user's lack of knowledge in the target language. The proposed tool in this thesis provides the user with interactive contextual information in order to involve her/him in the translation process. This contextual information describes the translation in the user's own language so that the user has confidence in the translation.

Experiments dealing with the accuracy of the tool proved that the tool has a certain degree of translation accuracy. Two main evaluations have been conducted: first, pre-post query evaluation and second, cross-lingual evaluation. In the pre-post evaluation, evaluations for different languages in a spelling correction task and evaluation of conflation approaches for Arabic have been performed. The second evaluation has been conducted to test the performance of the cross-lingual tool proposed in this thesis. This was twofold, first evaluation was to evaluate the performance of the disambiguation algorithm for two languages pairs, Arabic/English and English/German. Second, and in order to take the user's point of view, for the proposed tool into account, the tool has been tested in an actual situation in the form of a user study (users who have no knowledge or little knowledge in the target language). The goal of the performed user study was twofold. First, to identify possible weakness in the initial design of the tool in order to tackle them later in Chapter 8 and second, we were interested in evaluating whether the support provided by our cross-lingual tool is significant enough to guide the user in improving the translation and thus improve the performance of the cross-lingual retrieval.

In the following, we describe our future work perspective in regard to the approaches researched and implemented in this thesis.

10.2 Future Work Perspectives

The approaches to tackle the problems of cross-lingual retrieval, which have been proposed in this thesis, are limited to web applications dealing particularly with vagueness in the user query. However, there are some other application domains, where the user query can be very long and thus the approaches proposed in this thesis would not be powerful enough to use in other domains. For example, in the future work perspective, we would like to give some hints in how to adapt the approaches proposed in this thesis to cover other different domains (e.g., cross-lingual prior-art search). Reflecting the efforts of the emerging cross-lingual prior-art research, in the future work perspective, we would like to carefully identify the shortcoming of existing cross-lingual retrieval approaches, in order to propose and implement solutions to tackle these issues in future work. In the following, in Section 10.2.1, we give an overview of the patent information retrieval research. Furthermore, in Section 10.2.2, a discussing about traditional cross-lingual retrieval approach shortcomings is presented i.e., why using only traditional cross-lingual retrieval for cross-lingual prior-art search is not enough. In Section 10.2.3, an overview of which work has been done specifically for cross-lingual prior-art search is presented. In addition, in Section 10.2.4, in the summary, an analysis of the future research directions that need to be researched toward cross-lingual prior-art search.

10.2.1 Prior-Art Search

Patent information retrieval (patent IR), sometimes called patent retrieval or patent search, is a sub branch of information retrieval that aims to support patent experts to retrieve patents that satisfy their information needs and search criteria (Tait, 2008). A common scenario in patent search is prior-art search, which is performed by patent experts to determine whether a new invention can be patentable (Tiwana and Horowitz, 2009). Prior-art search is not a trivial task and is mostly performed by patent experts who need to spend hours and sometimes even days searching potentially relevant patent information. To perform their search tasks, patent experts use information retrieval systems and tools. Prior-art search can be achieved by considering all relevant information found in the patent data that can invalidate the novelty of a patent application claim. Thus, an invention is patentable only when no matched records for this patent claim can be found in the patent data. One missing relevant record in the patent data can cause high material losses due to patent contravention (Bashir and Rauber, 2010). Therefore, patent retrieval is considered as a recall-oriented application domain, where one missing relevant document, can be more important than retrieving a set of top relevant ranking documents.

Monolingual prior-art searching in patent data has specific properties, which set it apart from any traditional information retrieval (IR) system. Patents are generally expressed in grammatically correct language. However, patents are expressed in generic terms and use vague expressions to prevent narrowing down the scope of the inventions which will then lead to the fact that important concepts will be hidden in the patent document. Another reason for this is to extend the coverage of the patents but at same time not allowing people to easily understand the technique behind the invention. Furthermore, different written styles of language can be found in the same patent document which describes an invention i.e., abstract and description fields use a technical terminology while claim fields use legal expressions (Xue and Croft, 2009). The text in these fields is written over different periods of time and may not be in logical order (Atkinson, 2008). In addition to these issues, patent writers tend to use their self-developed terms or intensive use of acronyms to increase their patent acceptance rate during the patent examination procedure. This intensive use of self-developed terminologies and acronyms poses great challenges to any traditional information retrieval system.

In prior-art retrieval systems, a keywords based query is used, where patent users such as patent examiners or law persons construct their query by extracting the query terms from the patent claim field in order to formulate their query and thus get a more relevant search (Konishi et al., 2004). Therefore, the success of the search is highly dependent on the quality of terms which is used to construct the patent query. However, due to the above mentioned patent search issues, constructing a high quality term patent query is not a trivial task. Some documents are easily retrieved by many queries, whereas others may never show up within the top ranked retrieved documents for any reasonable query up to a certain length (Bashir and Rauber, 2010).

Prior-art disclosures are valid regardless of the patent language used. In order for prior-art to be disclosed, it must be cross-lingual. Therefore, cross-lingual retrieval is an indispensable component in prior-art search.

Besides traditional cross-lingual retrieval issues such as translation ambiguity, prior-art cross-lingual search creates even more challenges to retrieve patent documents across languages. For example, how to find the proper translation for vague terms, acronyms and self-developed terms.

10.2.2 Traditional Cross-lingual Retrieval Approach Shortcomings

Traditional cross-lingual retrieval approaches (see Chapter 2) face difficulties when applying them to cross-lingual patent retrieval. For example, using the machine translation approach without any adaption will result in having low performance when applying it to cross-lingual patent retrieval. One issue of using traditional cross-lingual retrieval approaches is the availability of resources. For example, query log data and click through data are available from the web and can be used for traditional cross-lingual retrieval tasks. However, collecting this data for patent retrieval tasks is not a trivial or impractical task. This is because collecting this information from patent searchers is very difficult. Making available what and how patent information is searched may affect the process of finding the patent prior-art search and therefore it should be the responsibility for patent professionals only (Jochim et al., 2010). Another issue, for example, is long sentences which are abundant in patent documents and can prevent the performance of any traditional machine translation system. The problem particularly exists in the claims section where the inventor must write, in a single sentence, a legal monopoly related to the invention. Another issue is that current machine translation systems are not adapted to use the International Patent Classification system (IPC)[1] which is associated with each patent document for training (Ceausu et al., 2011). Integrating the International Patent Classification system in the translation process will improve the machine translation performance for translating each patent domain effectively. Another issue of using cross-lingual retrieval approaches, based on language resources such as WordNet, is the unavailability of such resources for patent domain. Therefore, currently meaning matching in patent retrieval should be performed in a different way (Jochim et al., 2010).

In the following we investigate if some of the existing cross-lingual prior-art search approaches consider some of the traditional cross-lingual retrieval approach shortcomings toward patent retrieval. Furthermore, based on this research, we give a summary of which research directions cross-lingual prior-art search should focus on in the future.

10.2.3 Cross-lingual Prior-Art Search Approaches

Cross-lingual patent retrieval has received more attention in the last few years. Earlier work has been done by Higuchi et al. (2001) who proposed a multi-lingual patent retrieval system called PRIME. PRIME translates a user query into the target language, retrieves patents relevant to the user's information needs, and improves the retrieved patents browsing efficiency by the use of machine translation and clustering techniques. Furthermore, for the out-of-dictionary words, their systems extract new translations from patent families which exist in the patent data collection, to improve the dictionary coverage. Based on the fact that the users are not always sure in which languages the patent they are looking for exists, PRIME system retrieves patents in multiple languages simultaneously and thus PRIME becomes a multi-lingual information retrieval system rather than a cross-lingual retrieval system. PRIME system performs its task as follows:

- First, the entered user query is translated by the query translation module into the foreign language (Japanese or English).
- Second, using the document retrieval model, the user query and its translation is looked up in the patent data collection in order to retrieve the relevant documents.
- Third, among the retrieved documents, only documents which are not in the user query language will be translated using the document translation module.

To improve the browsing efficiency for the retrieved documents, a clustering module is used to divide the retrieved documents into a specific number of groups (clusters)

[1] http://www.wipo.int/classifications/ipc/en/

using Hierarchical Bayesian Clustering (HBC) (Iwayama and Tokunaga, 1995). So far, all explained above is included in any traditional cross-lingual information system. In order to improve traditional cross-lingual retrieval approaches toward cross-lingual patent retrieval, Higuchi et al. (2001) proposed, in their system PRIME, a way of improving the dictionary coverage based on the patent data collection. In an off-line process, the translation extraction module identifies Japanese/English translations in the patent data collection in order to enhance the dictionary coverage and thus enhance the query translation module. Their extraction method works as follows: since patent documents are structured based on a number of fields (e.g., titles, abstracts, and claims), their method first identifies corresponding fragments based on the document's structure in order to improve the extraction accuracy. Since the structure of paired patents is not always the same, only the title and abstract fields, which are usually parallel in the patent data collection, are used. The ChaSen morphological analyzer (Matsumoto et al., 1999) and Brill tagger (Brill and Moore, 2000) are used to extract content words from Japanese and English fragments, respectively. In addition, more than one word into phrases is combined. Finally, the association score is based on the Dice coefficient (Yamamoto and Matsumoto, 2000) for all possible combination phrases, is computed and only those having a higher score will be selected as a final translation. Based on this mechanism a new translation can be produced in order to update the translation dictionary and thus has a possibility of improving the system's performance.

Jochim et al. (2010) studied whether precision and recall of patent retrieval, and more specifically of prior-art retrieval, can be improved by query translation. In particular, they expanded monolingual patent queries with their possible translations. They used patents granted by the EPO (European Patent Office)[2]. After granting a patent, EPO provides manual translations of each patent claim in three languages, English, French and German. These claims parallel translations are used to extract a bilingual dictionary for each language pair. The nature of the EPO data makes it possible to use it as a multilingual corpus. Furthermore, translation can also be found within documents in the corpus where originally patents are written in English and contain sections that are translated to German and French. The availability of this multilingual corpus was the basis to expand the query by using the original translations found in the multilingual corpus. Since the authors have a multilingual corpus in hand, their intuition was, to create queries in the collection language that may be useful for retrieval. Therefore, they have chosen to expand queries with translation terms rather than replacing the original query terms by their translations. This type of query translation is seen as a type of

[2] http://www.epo.org/

query expansion where the original query terms are kept with their possible translations, so in the end, a multilingual query is generated.

To perform the translation, a dictionary is queried in order to obtain possible translations for each query term. In order to improve the retrieval performance, the translation needs to be more accurate. In order to improve the translation, a domain-specific dictionary on patents is proposed. This dictionary should provide more accurate translations than a domain-free dictionary (dict.cc)[3] since it provides better coverage of the patent domain. In their strategy of using a domain-specific dictionary, they faced an obstacle in how to maintain a domain-specific dictionary: the dictionary coverage is affected by the dynamically changing patent sub-domains where duplicating new concepts is common. Furthermore, another weakness for a domain-specific dictionary is the interpretation of the ambiguous language that the patent writers use to deliberately hide details about their patents. In order to tackle such issues Jochim et al. (2010) proposed an approach based on extracting a domain-specific translation dictionary from the patent data collection. Specifically, they use the parallel translations existing between parts of patents in the collection. Firstly, these parallel translations are identified and aligned. However, aligning the parallel translations of the patent claims is not trivial. Patent claims are often composed of a single sentence with 100-200 words and some are up to 600 words in length. These long sentences can cause low performance to the aligning algorithm. In order to deal with this issue, the long sentences are split into small clauses and aligned. For this aligning process, which considers clauses as sentences, they used the freely-available gargantuan [4] sentence aligner. This aligner has a reported F-measure of 98% in sentence aligner task (Braune and Fraser, 2010). Since the aligning process is the core of their approach, they evaluated the proposed aligner gargantuan by conducting manual accuracy evaluation on the patent clauses they built. This evaluation is conducted by two researchers, one of them an expert in sentence alignment. 2898 sentences from randomly chosen patents in the German-English parallel patent claims were taken. In order to create a gold standard for patent clause alignment, the aligned sentences by gargantua were manually edited. In the two conducted evaluations, one by each researcher against the gold standard, gargantua gained F-measure =98% and 99% respectively. Once the aligning process has been performed, translation probabilities between terms in the aligned translations are computed. These translation probabilities between pairs of source and target language terms in the aligned patent claims are computed using the GIZA++ toolkit (Och and Ney, 2003b). They run GIZA++ twice for

[3] http://www.dict.cc/

[4] http://sourceforge.net/projects/gargantua/

each language pair using each of the languages once as the source language. The output of this process is having a table that contains translation candidate terms and their probabilities which are the entries to their patent domain-specific dictionary (PatDict). Not only one single translation is selected, instead a translation probability threshold is defined where only terms that have significant translation probability will be selected. The definition of the translation probability threshold is done based on the translation accuracy or retrieval performance. Selecting the single most probable translation from the dictionary can result in ambiguity where many other possible correct translations will be omitted. Authors pointed out that in their future work they will use other translation methods that allow for contextualisation e.g., returning the top translation or using phrase-based translation which has shown better results compared to word by word translation (Ballesteros and Croft, 1996). Phrase-based translation is expected to improve the translation quality. For example, in German, a compound word such as "Kinderbuch" would be translated as the "children book" instead of just "children".

Leveling et al. (2011) studied the affect of compound words on patent cross-lingual information retrieval. A compound word is a word that is a result of joining two or more words together. Compound words can result in having out-of-vocabulary (OOV) problems in cross-lingual information retrieval. In order to improve cross-lingual information retrieval effectiveness, these compound words need decompounding before translation. Decompounding is the process of splitting compounds into their constituent parts. Decompounding has been found to improve information retrieval effectiveness and multilingual retrieval, because it can tackle the vocabulary mismatch problems (Chen and Gey, 2004). When the search for a patent in compounding languages, such as German or Dutch, cross-lingual retrieval performance is lower than for other language pairs (Piroi, 2010). This issue is due to the presence of compound words in the query or in the patent data collection which will result in a higher rate of OOV compound terms. These OOV, in most cases, can't be translated and will result in poor patent cross-lingual retrieval performance.

Leveling et al. (2011) applied decompounding on German patent topics in the patent cross-lingual search task from the CLEF-IP 2010 track[5] and evaluated machine translation quality by examining the retrieval performance. They used a similar approach for decompounding which has been proposed on (Chen and Gey, 2004) and was applied for domain specific cross-lingual retrieval. The algorithm works as follows:

- A German dictionary which contains non-compound words, in various forms, is

[5] http://www.ir-facility.org/clef-ip-2010-call-for-participation

built.

- A compound German word is decompounded based on the created dictionary in the first step. For example, the German based dictionary contains ball, fuss, fussball, meisterschaft and others, the German compound word "fussballmeisterschaft" (European Football Cup) is decompounded into several compound words based on the German base dictionary. So, based on this step, we have these two compounds (fuss ball europa meisterschaft) and (fussball europa meisterschaft).

- The decomposition with the smallest number of component words is chosen. In the previous example, the decomposition (fussball europa meisterschaft) will be selected as the decompounding for the German compound "fussballmeisterschaft".

- If there is more than one decomposition share with the same number of component words, the one with the highest probability of decomposition will be chosen. The probability is estimated by the product of the relative frequencies of the component words in the training collection.

Leveling et al. (2011) used a training collection that contains English corpora (Leipzig corpora track [6]) with 3 Million sentences and a random sample of 800,000 sentences from German patents in the CLEF-IP collection. The authors evaluated the decompounding based on a gold standard corpus that contained 2000 random sentences extracted from German patents. The Gold standard was manually annotated with the correct decomposition of words and contains 27,932 unique words and 318,000 words in total. The proposed decompounding approach achieved 95.0% accuracy that represented the number of correctly decompounded words by the approach applied over all words in the annotated Gold standard, and achieved 81.4% accuracy for unique words. Decompounding the Gold standard has a clear impact on increasing the number of words by 16.13% while decreasing the number of unique words by 84.8% and thus decomposing is very productive for German. In their experiments over CLEF-IP 2010 patent data, they discovered that often decompounding (decreasing the OOV words) has a positive impact on improving the cross-lingual patent retrieval. For example, with a 50,000 word corpora size, the OOV words were decreased from 20.9% to 3.7%, which resulted in improving the precision from 44.4% to 47.9%. Using a corpora size of 5,000 words, the precision improved by 9% from 36% to 45%.

Another approach for cross-lingual patent retrieval that doesn't require query translation was proposed by Li and Shawe-Taylor (2007). Li and Shawe-Taylor (2007) stud-

[6] http://corpora.uni-leipzig.de/

ied several machine learning techniques for cross-lingual patent retrieval and classification. They proposed a learning algorithm that exploits the bilingual training documents and discover a semantic representation from them. The algorithm was a fully automated cross-lingual information retrieval in which no query translation was required. The method was based on the Kernel Canonical Correlation Analysis (KCCA) method, which can be used to find the maximally correlated projections of documents in two languages. The proposed algorithm has been used for Japanese/English cross-lingual patent retrieval. In order to tackle the problem of handling large training data, the partial Gram-Schmidt orthogonalisation algorithm was used (Cristianini et al., 2002). Several methods for cross-lingual document classification have been investigated. The classification methods were based on the Support Vector Machine (SVM) and on the Kernel Canonical Correlation Analysis (KCCA) that may require different types of training resources. Furthermore, Li and Shawe-Tylor studied two ways of combining the KCCA and SVM and found that the combination gained better results than other algorithms for bilingual or monolingual test documents.

PLuTO (Patent Language Translations Online) provides a rapid solution for the online retrieval and translation of patent documents. This is done by integrating a number of existing state-of-the-art approaches (Ceausu et al., 2011). The Machine Translation (MT) module in PLuTO was implemented based on the MaTrEx[7] (Machine Translation Using Examples) system developed at DCU Stroppa and Way (2006) Tinsley et al. (2008) Penkale et al. (2010).

The translation module in PLuTO was designed to handle the possible necessity of interchanging between novel and previously developed translation modules. This feature is very useful to adapt PLuTO to handle new language pairs and exploring new processing techniques whereas language specific components can be used as a plug-in at any stage of the translation. The hybrid architecture of PLuTO allows the combination of statistical phrase-based, example-based, and hierarchical approaches to translation. Furthermore, MaTrEx operates as a wrapper around existing state-of-the-art components such as a statistical machine translation approach (Moses) (Koehn et al., 2007) and the alignment approach (Giza++) (Och and Ney, 2003b).

The various implemented module components in the MaTrEx system include: word alignment through word packing (Ma et al., 2007), marker-based chunking and chunk alignment (Gough and Way, 2004), treebank-based phrase extraction (Tinsley and Way, 2009), super-tagging (Hassan et al., 2007), and decoding. Furthermore, MaTrEx includes language- specific extensions such as taggers, parsers, etc., which are available on demand

[7] http://www.openmatrex.org

for the pre-and-post processing module. MaTrEx has wide flexibility in that all of these modules can be plugged in or out depending on the language pair used. In PLuTO the user can request translations via a number of ways: through a GUI (Graphical User Interface) as text-based translation, requesting a translation after retrieval is performed or through a number of customized tools.

In order to train the machine translation system, for example for an English-French language pair, all relevant documents are extracted from the MAREC[8] patent corpora collection. MAREC is a first standardized patent corpus which was provided by the Information Retrieval Facility (IRF)[9]. MAREC patent documents include title, an abstract, a description, a drawing and one or more claims. In order to start training the system, the data needs to be cleaned first e.g., deleting duplicate data in case of any, and character encoding normalisation etc.,. In order to create the parallel corpora needed for translation, the processing stages of sentence splitting and alignment had to be adapted to the style of patents. In order to perform this process, a number of shared resources such as abbreviations, segmentation rules etc., are needed. For example, adding abbreviations that are frequent in patent documents. At the end of this process six million parallel sentences for training were extracted.

In order to increase the performance of PLuTO in patent document translations, some of the particular characteristics of patent documents, were taken into account, through the design of the system. For example, references to elements in figures, long sentences and adaptation to the IPC System.

In order to clarify the reference to elements in the figures issue, we consider this example: "Preferably, there is more than one leg (16 , 17 , 18) that is attached to the bottom of the base member (12) " here the language model does not account for the trigram "leg (16 " , and the seventh token in the sequence " (16 , 17 , 18) " the closing parenthesis falls outside the default reordering window of six tokens. PLuTO tackles this issue by applying a number of rules as a pre-processing step. First, the figure reference from the source sentence will be extracted. Second, the sentence will be translated without the figure reference. Third, the reference will be inserted into the correct place in the translated sentences. The correct place for the figure in the translation sentences will be found based on alignment information stored during decoding.

Long sentences which are abundant in patent documents prevent the good performance of any traditional machine translation systems. The problem exists particularly in the claims section where the inventor must write a legal monopoly related to the

[8] http://www.ir-facility.org/prototypes/marec

[9] http://www.ir-facility.org/

invention, in a single sentence.

In order to tackle the long sentences issue in document patent translations, PLuTO splits each input sentence into smaller translatable chunks. In order to perform this process, the resource-light marker-based chunker (Gough and Way, 2004) from MaTrEx was integrated. In order to identify the points at which the sentence should be segmented, the chunker defines a set of marker words such as prepositions, conjunctions, pronouns, etc.,. Additional constraints were defined in PLuTO to avoid over-segmentation of the input. This over-segmentation could result in counterproductivity.

In order to include the possibility of training separate machine translation systems for each patent (sub-) domain, the International Patent Classification system (IPC) were included in PLuTO. There are 8 main categories for the IPC. For example, classification "C" represents "Chemistry", classification "G" represents "Physics", "H" represents "Electricity", etc.,. These 8 patent domains, along with the distribution of the MAREC corpus across each one, were represented to improve the machine translation systems.

In order to measure the performance of the proposed patent translation system PLuTO, an automatic comparative evaluation against two well known commercial systems Google translation and Systran was performed. The evaluation was conducted based on 5000 sentence pairs where PLuTO gained higher translation performance compared to Google and Systran.

10.2.4 Conclusion and Future Work Directions

Reflecting the efforts of the emerging cross-lingual prior-art research, in this chapter, we carefully selected and described what has been achieved, and perhaps even more significantly, what remains to be achieved toward cross-lingual prior-art search. We provided valuable information for cross-lingual prior-art search researchers who are looking for a comprehensive overview of state-of-the art approaches of cross-lingual prior-art. In other words, in this chapter we have investigated some of the significant work carried out on cross-lingual prior-art search. Firstly we gave an overview of the limitations of the state-of-the art approaches for cross-lingual information retrieval in handling prior-art cross-lingual search. Based on our investigation, cross-lingual retrieval approaches are inefficient in handling some of the specific properties for prior-art cross-lingual search; for example, due to the special properties for patent data, machine translation's lack of appropriate resources such as patent web logs, WordNet etc.,. Furthermore, long sentences, which are abundant in patent documents, prevent the performance of any traditional machine translation system. Future work can be directed toward creating appropriate resources specified for patent retrieval such as patent multi WordNet, which can be used to handle patent translations between languages. Due to the abundant use of acronyms, there is a need to define these acronyms with their description in order to have the possibility of translating them into other languages. Patent long sentence issues have been tackled by some of the current cross-lingual prior-art approaches. However, there is a need to improve these approaches to effectively tackle these issues e.g., improving chunker algorithms for long sentence segmentations. Since cross-lingual prior-art search is usually performed by patent experts, there is a deficiency in the existing cross-lingual prior-art approaches in giving the patent experts the possibility of being an integral part of the search process. Giving the patent expert the possibility to interact with the system, to omit or add new translation terms based on selected terms extracted from the patent data would be a vast improvement. Since patents are expressed in generic terms and use vague expressions to prevent narrowing down the scope of the inventions, it leads to important concepts being hidden in the patent document. This makes translation a very difficult task and therefore an ongoing challenge for cross-lingual prior-art search exists.

Part VI

Appendix

Appendix A

Spelling Correction Evaluation Tables

A.1 Results of Word Corrections in English

Misspelling	**Correct Spelling**	**Aspell**	**Microsoft word**	**Google**	**MultiSpell**
abberration	aberration	aberration	aberration	aberration	aberration
accomodation	accommodation	accommodation	accommodation	accommodation	accommodation
acheive	achieve	Achieve	achieve	achieve	achieve
abortificant	abortifacient	**Aficionados**	-	abortifacient	abortifacient
absorbsion	absorption	**absorbsi on**	absorps ion	absorption	absorption
ackward	(awkward, backward)	awkward	(awkward, backward)	awkward	(awkward, backward)
additinally	additionally	additionally	additionally	additionally	additionally
adminstration	administration	administration	administration	administration	administration
admissability	admissibility	admissibility	admissibility	admissibility	admissibility
advertisments	advertisements	advertisements	advertisements	advertisements	advertisements
adviced	advised	advised	advised	**advice**	**advice**
afficionados	aficionados	aficionados	aficionados	aficionados	aficionados
affort	(effort ,afford)	effort	afford	afford	afford
agains	against	**agings**	**agings**	against	against
aggreement	agreement	agreement	agreement	agreement	agreement
agressively	aggressively	aggressively	aggressively	aggressively	aggressively
agriculturalist	agriculturist	-	-	-	agriculturist

continued on next page

continued from previous page

Misspelling	**Correct Spelling**	**Aspell**	**Microsoft word**	**Google**	**MultiSpell**
alcoholical	alcoholic	alcoholically	**alcoholically**	alcoholic	alcoholic
algebraical	algebraic	algebraic	**algebraically**	algebraic	**algebraically**
algoritms	algorithms	algorithms	algorithms	algorithms	algorithms
alterior	(ulterior , anterior)	ulterior	(anterior, ulterior)	ulterior	(anterior, ulterior)
anihilation	annihilation	annihilation	annihilation	annihilation	annihilation
anthromor-phization	anthropomor- phiza-tion	**anthropomor-phizing**	-	-	anthropomor-phization
bankrupcy	bankruptcy	bankruptcy	bankruptcy	bankruptcy	bankruptcy
baout	(about,bout)	bout	(about,bout)	about	bout
basicly	basically	basically	basically	basically	basically
breakthough	breakthrough	**break though**	breakthrough	breakthrough	breakthrough
carachter	character	**crocheter**	character	character	character
cannotation	connotation	connotation	(connotation ,an-notation)	connotation	(connotation ,an-notation)
carismatic	charismatic	charismatic	charismatic	charismatic	charismatic
carmel	caramel	**Carmel**	-	-	caramel
cervial	(cervical, servile)	cervical	cervical	cervical	cervical
clasical	classical	classical	classical	classical	classical
cleareance	clearance	clearance	clearance	clearance	clearance
comissioning	commissioning	commissioning	commissioning	commissioning	commissioning

continued on next page

continued from previous page

Misspelling	**Correct Spelling**	**Aspell**	**Microsoft word**	**Google**	**MultiSpell**
commemerative	commemorative	commemorative	commemorative	commemorative	commemorative
compatabilities	compatibilities	compatibilities	compatibilities	compatibilities	compatabilities
committment	commitment	commitment	commitment	commitment	commitment
debateable	debatable	debatable	debatable	debatable	debatable
determinining	determining	determinining	determinining	determinining	determining
childbird	childbirth	**child bird**	**child bird**	childbirth	childbirth
definately	definitely	definitely	definitely	definitely	definitely
decribe	describe	describe	describe	describe	describe
elphant	elephant	elephant	elephant	elephant	elephant
emmediately	immediately	immediately	immediately	immediately	immediately
emphysyma	emphysema	emphysema	emphysema	emphysema	emphysema
erally	(orally, really)	orally	really	really	orally
eyasr	(years, eyas)	**eyesore**	years	years	eyas
facist	fascist	fascist	fascist	fascist	fascist
fluoroscent	fluorescent	fluorescent	fluorescent	fluorescent	fluorescent
geneology	genealogy	genealogy	genealogy	genealogy	genealogy
gernade	grenade	grenade	grenade	grenade	grenade
girates	gyrates	**grates**	gyrates	**pirates**	gyrates
gouvener	governor	governor	**souvenir**	**gouverneur**	**convener**
gurantees	guarantee	guarantee	guarantee	guarantee	guarantee

continued on next page

continued from previous page

Misspelling	Correct Spelling	Aspell	Microsoft word	Google	MultiSpell
guerrila	(guerilla, guerrilla)	guerrilla	guerrilla	guerrilla	(guerilla, guerrilla)
guerrilas	(guerillas, guerrillas)	guerrillas	guerrillas	guerrillas	(guerillas, guerrillas)
Guiseppe	Giuseppe	Giuseppe	Giuseppe	Giuseppe	Giuseppe
habaeus	(habeas, sabaeus)	habeas	**habitu’es**	habeas	sabaeus
hierarcical	hierarchical	hierarchical	hierarchical	hierarchical	hierarchical
heros	heroes	heroes	heroes	heroes	**herbs**
hypocracy	hypocrisy	hypocrisy	hypocrisy	hypocrisy	hypocrisy
independance	Independence	Independence	-	Independence	Independence
intergration	integration	integration	integration	integration	integration
intrest	interest	interest	interest	interest	interest
Johanine	Johannine	Johannes	Johannes	Johannes	Johannine
judisuary	judiciary	judiciary	judiciary	-	judiciary
kindergarden	kindergarten	kindergarten	kindergarten	kindergarten	kindergarten
knowlegeable	knowledgeable	knowledgeable	knowledgeable	knowledgeable	knowledgeable
labatory	(lavatory, laboratory)	(lavatory, laboratory)	(lavatory, laboratory)	laboratory	(lavatory, laboratory)
lonelyness	loneliness	loneliness	loneliness	loneliness	loneliness
59 legitamate	legitimate	legitimate	legitimate	legitimate	legitimate
libguistics	linguistics	linguistics	linguistics	linguistics	linguistics

continued on next page

continued from previous page

Misspelling	**Correct Spelling**	**Aspell**	**Microsoft word**	**Google**	**MultiSpell**
lisence	(license, licence)	licence	**silence**	licence	licence
mathmatician	mathematician	mathematician	mathematician	mathematician	mathematician
ministery	ministry	ministry	ministry	ministry	ministry
mysogynist	misogynist	misogynist	misogynist	misogynist	misogynist
naturaly	naturally	naturally	naturally	naturally	naturally
ocuntries	countries	countries	countries	countries	countries
paraphenalia	paraphernalia	paraphernalia	paraphernalia	paraphernalia	paraphernalia
Palistian	Palestinian	**Alsatain**	**politian**	Palestinian	Palestinian
pamflet	pamphlet	pamphlet	pamphlet	pamphlet	pamphlet
psyhic	psychic	psychic	psychic	psychic	psychic
Peloponnes	Peloponnesus	Peloponnese	Peloponnese	Peloponnese	Peloponnesus
personell	personnel	personnel	personnel	personnel	personnel
posseses	possesses	possesses	possesses	possesses	possess
prairy	prairie	**priory**	prairie	prairie	**airy**
qutie	(quite, quiet)	quite	quite	**cutie**	**queue**
radify	(ratify,ramify)	ratify	ratify	ratify	ramify
reccommended	recommended	recommended	recommended	recommended	recommended
reciever	receiver	receiver	receiver	receiver	**reliever**
reconaissance	reconnaissance	reconnaissance	reconnaissance	reconnaissance	reconnaissance
restauration	restoration	restoration	restoration	restoration	**instauration**

continued on next page

continued from previous page

Misspelling	**Correct Spelling**	**Aspell**	**Microsoft word**	**Google**	**MultiSpell**
rigeur	(rigueur, rigour, rigor)	**rigger**	rigueur	-	(rigueur, rigour)
Saterday	Saturday	Saturday	Saturday	Saturday	Saturday
scandanavia	Scandinavia	Scandinavia	Scandinavia	Scandinavia	Scandinavia
scaleable	scalable	scalable	-	scalable	scalable
secceeded	(seceded, succeeded)	succeeded	succeeded	seceded	succeeded
sepulchure	(sepulchre, sepulcher)	sepulcher	**sepulchered**	sepulcher	sepulchre
themselfs	themselves	themselves	themselves	themselves	themselves
throught	(thought,through)	(thought, through)	(thought,through)	**throat**	(thought, through)
troups	(trupes, troops)	(troupes, troops)	troupes	troops	troops
simultanous	smultaneous	simultaneous	simultaneous	simultaneous	simultaneous
sincerley	sincerely	sincerely	sincerely	sincerely	sincerely
sophicated	sophisticated	**suffocated**	**supplicated**	-	sophisticate
surrended	(surrounded, surrendered)	surrounded	surrender	surrender	surrounded
unforetunately	unfortunately	unfortunately	unfortunately	-	unfortunately
unnecesarily	unnecessarily	unnecessarily	unnecessarily	-	unnecessarily
usally	usually	usually	usually	usually	usually
useing	using	using	using	using	**seeing**

continued on next page

continued from previous page

Misspelling	**Correct Spelling**	**Aspell**	**Microsoft word**	**Google**	**MultiSpell**
vaccum	vacuum	vacuum	vacuum	vacuum	vacuum
vegitables	vegetables	vegetables	vegetables	vegetables	vegetables
vetween	between	between	between	between	between
volcanoe	volcano	volcano	volcano	volcano	volcano
weaponary	weaponry	weaponry	weaponry	weaponry	weaponry
worstened	worsened	worsened	worsened	-	worsened
wupport	support	support	support	support	support
yeasr	years	years	years	years	yeast
Yementite	(Yemenite, Yemeni)	Yemenite	Yemenite	Yemenite	Yemenite
yuonger	younger	younger	younger	younger	**sponger**

Table A.1: Results of word corrections in English (misspelled are in bold and underlined, "-" refer to no suggestion and recognized as misspelled.

A.2 Results of Word Corrections in Portuguese

Correct Form	Spelling Error	TST	Aspell	MultiSpell
acerca	àcerca	acerca	acerca	acerca
açoriano	açoreano	açoriano	**coreano**	açoriano
alcoolémia	alcoolemia	**alcoolÚmia**	-	alcoolémia
ameixial	ameixeal	ameixial	ameixial	ameixial
antárctico	antártico	catártico	antárctico	antárctico
antepor	antepôr	-	antepor	antepor
árctico	artico	**artigo**	**aórtico**	**aórtico**
artífice	artífece	artífice	artífice	artífice
bainha	baínha	bainha	bainha	bainha
bebé	bébé	bebé	**bebe**	bebé
bege	beje	bege	**beije**	**bejense**
bênção	benção	**bençao**	-	bênção
benefcência	benefciência	beneficência	beneficência	beneficência
biopsia	biópsia	**biópsiu**	-	biopsia
burburinho	borborinho	burburinho	burburinho	burburinho
caiem	caem	-	-	**cabem**
calvície	calvíce	calvície	calvície	calvície
camoniano	camoneano	camoniano	camoniano	camoniano
campeão	campião	campeão	campeão	campeão
chiita	xiita	chiita	**xiitas**	xiitas
comboio	combóio	comboio	comboio	comboio
compor	compôr	-	compor	compor
comummente	comumente	**comovente**	comummente	comummente
constituia	constituía	-	-	constituia
constituiu	constituíu	constituiu	constituiu	constituiu
cor	côr	-	cor	cor
crânio	crâneo	crânio	**cárneo**	crânio
definição	defenição	definição	definição	definição
definido	defenido	definido	-	defendido
definir	defenir	definir	definir	definir
desequilíbrio	desequilibrio	desequilíbrio	desequilíbrio	desequilíbrio
despretensioso	despretencioso	despretensioso	despretensioso	despretensioso

continued on next page

continued from previous page

Correct Form	**Spelling Error**	**TST**	**Aspell**	**MultiSpell**
dignatários	dignitários	dignatários	**digitarias**	dignatários
dispender	despender	dispender	-	**despendes**
dispêndio	dispendio	**dispundio**	**dispundio**	dispendioso
ecrã	ecran	-	écran	écran
emirados	emiratos	**estratos**	**méritos**	emirados
esotérico	isotérico	-	-	esotérico
esquisito	esquesito	esquisito	esquisito	esquisito
estratego	estratega	estratego	-	estratego
feminino	femenino	feminino	feminino	feminino
feminismo	femininismo	-	feminismo	feminismo
fôr	for	-	-	**forcar**
gineceu	geneceu	gineceu	gineceu	gineceu
gorjeta	gorgeta	gorjeta	gorjeta	gorjeta
granjear	grangear	granjear	granjear	granjear
guisar	guizar	guisar	**gizar**	**guinar**
halariedade	hilaridade	hilariedade	-	**polaridade**
hectare	hectar	hectare	-	hectare
hiroshima	hiroxima	**aproxima**	**próxima**	hiroshima
ilacção	elação	ilação	ilação	**delação**
indispensável	indespensável	indispensável	indispensável	indispensável
inflacção	inflação	-	-	**inalação**
interveio	interviu	**intervir**	**Inter viu**	**intervim**
intervindo	intervido	intervindo	-	intervindo
invocar	evocar	invocar	-	**evocai**
ípsilon	ipslon	ípsilon	ípsilon	ípsilon
irisar	irizar	irisar	**razar**	irisar
irupção	irrupção	-	-	irupção
jeropiga	geropiga	jeropiga	**Georgia**	jeropiga
juiz	juíz	-	juiz	Juiz
lampião	lampeão	lampião	sarjeta	**campeão**
lêem	lêm	**lês**	**lema**	lêem
linguista	linguísta	-	linguista	linguista
lisonjear	lisongear	lisonjear	lisonjear	lisonjear
logótipo	logotipo	**logo tipo**	**logo tipo**	logótipo

continued on next page

continued from previous page

Correct Form	Spelling Error	TST	Aspell	MultiSpell
maciço	massiço	mássico	mássico	**massudo**
majestade	magestade	majestade	majestade	majestade
manjerico	mangerico	manjerico	manjerico	manjerico
manjerona	mangerona	**tangerina**	**tangerina**	manjerona
meteorologia	metereologia	meteorologia	meteorologia	meteorologia
miscigenação	miscegenação	miscigenação	miscigenação	miscigenação
nonagésimo	nonagessimo	nonagésimo	nonagésimo	nonagésimo
oceânia	oceania	oceânia	**Oceania**	oceânia
oficina	ofecina	oficina	oficina	oficina
opróbrio	opróbio	**aeróbio**	**próbio**	opróbrio
organograma	organigrama	organograma	-	organograma
paralisar	paralizar	paralisar	paralisar	paralisar
perserverança	preseverança	perserverança	perserverança	perseverance
persuasão	persuação	persuasão	persuasão	persuasão
pirinéus	pirenéus	-	pirinéus	pirinéus
pretensioso	pretencioso	pretensioso	pretensioso	pretensioso
privilégio	previlégio	privilégios	privilégios	privilegios
quadricromia	quadricomia	quadricromia	**quadriculai**	quadricromia
quadruplicado	quadriplicado	quadruplicado	quadruplicado	quadruplicado
quasímodo	quasimodo	-	**quisido**	quasímodo
quilo	kilo	quilo	Nilo	**dilo**
quilograma	kilograma	**holograma**	**holograma**	**holograma**
quilómetro	kilómetro	**milímetro**	**milímetro**	quilómetro
quis	quiz	quis	**qui**	**juiz**
rainha	raínha	rainha	rainha	rainha
raiz	raíz	-	raiz	raiz
raul	raúl	raul	Raul	raul
rectaguarda	retaguarda	rectaguarda	-	rectaguarda
rédea	rédia	rédea	**radia**	**radia**
regurgitar	regurjitar	regurgitar	regurgitar	regurgitar
rejeitar	regeitar	rejeitar	**regatar**	**receitar**
requeiro	requero	requere	requeiro	requer
réstia	réstea	réstia	**resta**	réstia
rubrica	rúbrica	**rúbreca**	rubrica	rubrica

continued on next page

continued from previous page

Correct Form	**Spelling Error**	**TST**	**Aspell**	**MultiSpell**
saem	saiem	**saiam**	saem	**caiem**
saloiice	saloice	**baloice**	saloiice	saloiice
sarjeta	sargeta	sarjeta	sarjeta	Sarjeta
semear	semiar	semear	semear	Semear
suíça	suiça	suíça	suíça	Suíça
supor	supôr	-	supor	Supôs
trânsfuga	transfuga	**transfira**	**transfira**	trânsfuga
transpôr	transpor	-	-	transportar
urano	úrano	-	-	**grano**
ventoinha	ventoínha	ventoinha	ventoinha	ventoinha
verosímil	verosímel	-	-	verosímil
vigilante	vegilante	vigilante	vigilante	vigilante
vôo	voo	-	-	**ovo**
vultuoso	vultoso	vultuoso	-	**vultosos**
xadrez	xadrês	xadrez	**ladres**	xadrez
xamã	chamã	chama	chama	chamá
xelindró	xilindró	**cilindro**	**cilindro**	xelindró
zângão	zangão	zangai	-	**mangão**
zepelin	zeppelin	**zepelim**	**zeplim**	zepelin
zoo	zoô	zoo	**coo**	zoo

Table A.2: Results of word corrections in Portuguese (misspelled are in bold and underlined, "-" refer to no suggestion and recognized as misspelled).

Appendix B

Conflation Approaches Evaluation Tables

B.1 Conflation Using the Revised Bigram and Pure Trigram Approach

S/N	Word	Pure trigram	Revised bigram	Translation
1	مساعدmsāʿd	Relevant	Relevant	Helper
2	بمساعدbmsāʿd	Relevant	Relevant	By helper
3	بمساعدةbmsāʿdh	Relevant	Relevant	By help
4	تساعدtsāʿd	Not retrieved	Relevant	She helps
5	ساعدsāʿd	Relevant	Relevant	He helped
6	ساعدهsāʿdh	Not retrieved	Relevant	He helped him
7	ساعدتsāʿdt	Not retrieved	Relevant	She helped
8	يساعدysāʿd	Not retrieved	Relevant	He helps
9	كمساعدةkmsāʿdh	Relevant	Relevant	As a help
10	ومساعدwmsāʿd	Relevant	Relevant	And helper
11	ومساعدهwmsāʿdh	Relevant	Relevant	And his helper
12	ومساعدةwmsāʿdh	Not retrieved	Relevant	And help
13	وساعدwsāʿd	Not retrieved	Relevant	And he helped
14	لمساعدlmsāʿd	Relevant	Relevant	For helper
15	لمساعدةlmsāʿdh	Relevant	Relevant	For help
16	نساعدnsāʿd	Not retrieved	Relevant	We help

continued on next page

continued from previous page				
S/N	**Word**	**Pure trigram**	**Revised bigram**	**Translation**
17	مساعديmsāʿdy	Relevant	Relevant	My helper
18	مساعدينmsāʿdyn	Relevant	Relevant	Helpers
19	مساعديهmsāʿdyh	Relevant	Relevant	His helpers
20	مساعدوmsāʿdw	Relevant	Relevant	Helpers
21	مساعدونmsāʿdwn	Relevant	Relevant	helpers
22	مساعدوهmsāʿdwh	Relevant	Relevant	His helpers
23	مساعدهmsāʿdh	Relevant	Relevant	His helper
24	مساعدهاmsāʿdhā	Relevant	Relevant	Her helper
25	مساعداmsāʿdā	Relevant	Relevant	A helper
26	مساعدآmsāʿdʾā	Relevant	Relevant	A helper
27	مساعداتmsāʿdāt	Relevant	Relevant	Helps
28	مساعدةmsāʿdh	Relevant	Relevant	Help
29	مساعديmsāʿdy	Not retrieved	Relevant	My helper
30	مساعدتهmsāʿdth	Not retrieved	Relevant	His help
31	آساعدʾāsāʿd	Not retrieved	Relevant	I Help
32	لمساعد lālmsāʿd	Not retrieved	Relevant	The helper
33	مساعدونmsāʿdwn	Not retrieved	Relevant	Helpers
34	ومساعيwmsāʿy	Irrelevant	Irrelevant	-
35	بمساعbmsāʿ	Irrelevant	Irrelevant	-
36	المساعlmsāʿ	Irrelevant	Irrelevant	-
37	مساعيmsāʿy	Irrelevant	Irrelevant	-

Table B.1: The result of the query مساعدmsāʿd (helper) using the revised bigram and pure trigram approach.

B.2 Recall, F-measure Evaluation Example

Query	**Retrieved**	**Similarity**	**Relevant/ Irrelevant**
الشركةālšrkh (the Company)	الشرālšr Evil	60%	Irr.
	الشركālšrk (the trap)	80%	Irr.
	الشركةālšrkh (the company)	100%	Rel.
	بالشركةbālšrkh (by the company)	83.3%	Rel.
	شركةšrkh (company)	60%	Rel.
	ولشركةwlšrkh (and for a company)	66.63%	Rel.
	والشركةwālšrkh (and the company)	83.3%	Rel.
	للشّركةllšrkh (for the company)	66.63%	Rel.
	لشركةlšrkh (for a company)	80%	Rel.
	- - -	- - -	- - -

Table B.2: An example query الشركةālšrkh out of the randomly selected 30 queries for recall and F-measure using revised bigram (only the first 9th word form variations are displayed).

Query	**Retrieved**	**Similarity**	**Relevant/ Irrelevant**
تـدريـبtdryb (Training)	بتدريبbtdryb by training	80%	Rel.
	تدريtdry (she knows)	75%	Irr.
	تدريبtdryb (training)	100%	Rel.
	تـدريـبيtdryby (ongoing training)	79.95%	Rel.
	تـدريـبـينtdrybyn (two training sessions)	66.58%	Rel.
	تدريبيئاtdrybyʾā (ongoing training)	66.58%	Rel.
	تدريبيةtdrybyh (ongoing training)	66.58%	Rel.
	تـدريـبهtdrybh (his training)	79.95%	Rel.
	تـدريـبـهـمtdrybhm (thier training)	66.58%	Rel.
	- - -	- - -	- - -

Table B.3: An example query تـدريـبtdryb out of the randomly selected 30 queries for recall and F-measure using revised bigram (only the first 9th word form variations are displayed).

Query	**Retrieved**	**Similarity**	**Relevant/ Irrelevant**
استقلالesteqlāl (Independence)	استقلāstql independent	66.66%	Irr.
	اتصالetṣāl (call)	80%	Rel.
	استقلالāsteqlāl (independence)	100%	Rel.
	استقلاليةāsteqlālyh (independence)	74.95%	Rel.
	استقلاليتهاāsteqlālythā-(her independence)	60%	Rel.
	استقلالناāsteqlālnā (our independence)	74.95%	Rel.
	استقلالهāsteqlālh (his independence)	85.59%	Rel.
	استقلالهماāsteqlālhmā (thier independence (dual)	66.61%	Rel.
	استقلالهاāsteqlālhā (her independence (dual)	74.95%	Rel.
	- - -	- - -	- - -

Table B.4: An example query استقلالesteqlāl out of the randomly selected 30 queries for recall and F-measure using revised bigram (only the first 9th word form variations are displayed).

Query	**Retrieved**	**Similarity**	**Relevant/ Irrelevant**
آلمانيا ʾālmānyā (Germany)	بالمانيا bālmānyā by Germany	85.59%	Rel.
	برلماني brlmāny (parliamentary)	62.45%	Irr.
	علماني ʿlmāny (secularism)	61.92%	Irr.
	فالمانيا fālmānyā (and so Germany)	85.69%	Rel.
	كالمانيا kālmānyā (as Germany)	85.69%	Rel.
	ولالمانيا wlālmānyā (and for Germany)	74.42%	Rel.
	لّالماني llālmāny (for the German)	62.45%	Rel.
	لّالمانيا llālmānyā (for Germany)	85.69%	Rel.
	المان ālmān (Germans)	66.66%	Rel.
	- - -	- - -	- - -

Table B.5: An example query آلمانيا ʾālmānyā out of the randomly selected 30 queries for recall and F-measure using revised bigram (only the first 9th word form variations are displayed).

Appendix C

Cross-lingual Retrieval Tools Evaluations Tables

C.1 Naïve Bayesian Classifier (NB) Results and Illustrative Tables

C.1.1 Inflectional form examples used by Naïve Bayesian Classifiers (NB) Approach

Sense	**Form**	**Arabic sentence**	English translation
Religion	Basic	لآن الآسلام الذي هو **دين** حوار و انفتاح علي الناس lʾān ālʾāslām ālḏy hw dyn ḥwār w ānftāḥ ʿly ālnās	because Islam, which is a **religion** of dialogue and openness to people
Debt	Basic	الآقتراض من البنوك آو تحويل هذا العجز الى **دين** ālʾāqtrāḍ mn ālbnwk ʾāw tḥwyl hḏā ālʿǧz ā-lā dyn	borrowing from banks or through converting such a deficit into a budget **debt**
Religion	Inflection	ينشر تعاليم **الدين** و الثقافة الاسلامية ynšr tʿālym āldyn w ālṯqāfh ālāslāmyh	promoting tenets of the **religion** and Arabic culture

continued on next page

continued from previous page

Sense	**Form**	**Arabic sentence**	English translation
Debt	Inflection	جدولة **الدين** في نادي باريس ğdwlh āldyn fy nādy bārys	arrangements of **debt** scheduling in Paris Club

Table C.1: Part of sentences examples for the ambiguous word دينdyn for both senses in basic and inflectional form appeared in the training data.

C.1.2 Ranked Translations based on Naïve Bayesian Classifiers(NB)

S/N	Translation Combinations	Score
1	tax,customs,commodities	0,05948
2	tax,customs,goods	0,05539
3	tax,customs,commercial	0,05248
4	tax,customs,crack	0,0484
5	tax,customs,rift	0,0484
6	tax,tariff,commodities	0,01399
7	tax,tariff,goods	0,01283
8	tax,control,commodities	0,01224
9	tax,control,goods	0,01108
10	tax,tariff,commercial	0,007
11	tax,control,commercial	0,00525
12	drawing,customs,commercial	0,0035
13	drawing,control,commercial	0,0035
14	drawing,tariff,commercial	0,0035
15	tax,tariff,crack	0,00175
16	tax,tariff,rift	0,00175
17	indicate,control,goods	0,00175
18	indicate,customs,goods	0,00117
19	indicate,tariff,goods	0,00117
20	drawing,customs,commodities	0,00117
21	drawing,control,commodities	0,00117
22	drawing,tariff,commodities	0,00117

continued on next page

continued from previous page

S/N	**Translation Combinations**	**Score**
23	indicate,control,crack	0,00058
24	indicate,control,rift	0,00058
25	indicate,control,commodities	0,00058
26	indicate,control,commercial	0,00058
27	appoint,customs,commercial	0,00058
28	appoint,control,commercial	0,00058
29	appoint,tariff,commercial	0,00058
30	drawing,customs,crack	0,00058
31	drawing,customs,rift	0,00058
32	drawing,customs,goods	0,00058
33	drawing,control,crack	0,00058
34	drawing,control,rift	0,00058
35	drawing,control,goods	0,00058
36	drawing,tariff,crack	0,00058
37	drawing,tariff,rift	0,00058
38	drawing,tariff,goods	0,00058
39	fee,customs,rift	0
40	fee,customs,commodities	0
41	fee,customs,commercial	0
42	fee,customs,goods	0
43	fee,control,crack	0
44	fee,control,rift	0
45	fee,control,commodities	0
46	fee,control,commercial	0
47	fee,control,goods	0
48	fee,tariff,crack	0
49	fee,tariff,rift	0
50	fee,tariff,commodities	0
51	fee,tariff,commercial	0
52	fee,tariff,goods	0
53	fee,customs,crack	0
54	tax,control,crack	0
55	tax,control,rift	0
56	prescribe,customs,crack	0

continued on next page

continued from previous page

S/N	Translation Combinations	Score
57	prescribe,customs,rift	0
58	prescribe,customs,commodities	0
59	prescribe,customs,commercial	0
60	prescribe,customs,goods	0
61	prescribe,control,crack	0
62	prescribe,control,rift	0
63	prescribe,control,commodities	0
64	prescribe,control,commercial	0
65	prescribe,control,goods	0
66	prescribe,tariff,crack	0
67	prescribe,tariff,rift	0
68	prescribe,tariff,commodities	0
69	prescribe,tariff,commercial	0
70	prescribe,tariff,goods	0
71	indicate,customs,crack	0
72	indicate,customs,rift	0
73	indicate,customs,commodities	0
74	indicate,customs,commercial	0
75	indicate,tariff,crack	0
76	indicate,tariff,rift	0
77	indicate,tariff,commodities	0
78	indicate,tariff,commercial	0
79	appoint,customs,crack	0
80	appoint,customs,rift	0
81	appoint,customs,commodities	0
82	appoint,customs,goods	0
83	appoint,control,crack	0
84	appoint,control,rift	0
85	appoint,control,commodities	0
86	appoint,control,goods	0
87	appoint,tariff,crack	0
88	appoint,tariff,rift	0
89	appoint,tariff,commodities	0
90	appoint,tariff,goods	0

continued on next page

continued from previous page

S/N	**Translation Combinations**	**Score**
91	trace,customs,crack	0
92	trace,customs,rift	0
93	trace,customs,commodities	0
94	trace,customs,commercial	0
95	trace,customs,goods	0
96	trace,control,crack	0
97	trace,control,rift	0
98	trace,control,commodities	0
99	trace,control,commercial	0
100	trace,control,goods	0
101	trace,tariff,crack	0
102	trace,tariff,rift	0
103	trace,tariff,commodities	0
104	trace,tariff,commercial	0
105	trace,tariff,goods	0
106	sketch,customs,crack	0
107	sketch,customs,rift	0
108	sketch,customs,commodities	0
109	sketch,customs,commercial	0
110	sketch,customs,goods	0
111	sketch,control,crack	0
112	sketch,control,rift	0
113	sketch,control,commodities	0
114	sketch,control,commercial	0
115	sketch,control,goods	0
116	sketch,tariff,crack	0
117	sketch,tariff,rift	0
118	sketch,tariff,commodities	0
119	sketch,tariff,commercial	0
120	sketch,tariff,goods	0
121	illustration,customs,crack	0
122	illustration,customs,rift	0
123	illustration,customs,commodities	0
124	illustration,customs,commercial	0

continued on next page

continued from previous page

S/N	Translation Combinations	Score
125	illustration,customs,goods	0
126	illustration,control,crack	0
127	illustration,control,rift	0
128	illustration,control,commodities	0
129	illustration,control,commercial	0
130	illustration,control,goods	0
131	illustration,tariff,crack	0
132	illustration,tariff,rift	0
133	illustration,tariff,commodities	0
134	illustration,tariff,commercial	0
135	illustration,tariff,goods	0

Table C.2: Disambiguation scores for each possible translations sets based on naïve bayesian classifier (NB).

C.2 Ranked Translations based on Mutual Information Approach

S/N	Translation Combinations	Occurrence	MI Score
1	organization AND health AND world	5579	8,62651
2	organization AND health AND international	2457	7,80648
3	organized AND health AND world	415	6,0282
4	organized AND health AND international	328	5,79295
5	organization AND truth AND world	229	5,43367
6	organization AND health AND wide	225	5,41608
7	organization AND truth AND international	205	5,32297
8	arranged AND health AND world	137	4,91995
9	sponsor AND health AND world	116	4,75357
10	organized AND truth AND world	99	4,59511
11	arranged AND health AND international	95	4,55385
12	sponsor AND health AND international	84	4,4308
13	organized AND truth AND international	80	4,38201
14	organizer AND health AND world	57	4,04304

continued on next page

continued from previous page

S/N	Translation Combinations	Occurrence	MI Score
15	organizer AND health AND international	50	3,91201
16	orderly AND health AND world	46	3,82863
17	organized AND health AND wide	44	3,78418
18	orderly AND health AND international	42	3,73766
19	arranged AND truth AND world	34	3,52635
20	arranged AND truth AND international	29	3,36729
21	sponsor AND truth AND world	27	3,29583
22	sponsor AND truth AND international	26	3,25809
23	organization AND truth AND wide	18	2,89037
24	arranged AND health AND wide	14	2,63905
25	orderly AND truth AND world	10	2,30258
26	sponsor AND health AND wide	10	2,30258
27	organized AND truth AND wide	9	2,19722
28	organizer AND truth AND international	7	1,94591
29	organizer AND truth AND world	7	1,94591
30	orderly AND truth AND international	5	1,60944
31	organization AND correctness AND international	4	1,38629
32	orderly AND health AND wide	4	1,38629
33	arranged AND truth AND wide	4	1,38629
34	organizer AND health AND wide	4	1,38629
35	organization AND correctness AND world	3	1,09861
36	organized AND correctness AND world	2	0,69315
37	organization AND health AND universality	0	0
38	organization AND health AND internationalism	0	0
39	organization AND truth AND universality	0	0
40	organization AND truth AND internationalism	0	0
41	organization AND correctness AND universality	0	0
42	organization AND correctness AND internationalism	0	0
43	organization AND correctness AND wide	0	0
44	organized AND health AND universality	0	0
45	organized AND health AND internationalism	0	0
46	organized AND truth AND universality	0	0
47	organized AND truth AND internationalism	0	0
48	organized AND correctness AND universality	0	0

continued on next page

continued from previous page

S/N	Translation Combinations	Occurrence	MI Score
49	organized AND correctness AND internationalism	0	0
50	organized AND correctness AND international	0	0
51	organized AND correctness AND wide	0	0
52	orderly AND health AND universality	0	0
53	orderly AND health AND internationalism	0	0
54	orderly AND truth AND universality	0	0
55	orderly AND truth AND internationalism	0	0
56	orderly AND truth AND wide	0	0
57	orderly AND correctness AND universality	0	0
58	orderly AND correctness AND internationalism	0	0
59	orderly AND correctness AND international	0	0
60	orderly AND correctness AND world	0	0
61	orderly AND correctness AND wide	0	0
62	arranged AND health AND universality	0	0
63	arranged AND health AND internationalism	0	0
64	arranged AND truth AND universality	0	0
65	arranged AND truth AND internationalism	0	0
66	arranged AND correctness AND universality	0	0
67	arranged AND correctness AND internationalism	0	0
68	arranged AND correctness AND international	0	0
69	arranged AND correctness AND world	0	0
70	arranged AND correctness AND wide	0	0
71	organizer AND health AND universality	0	0
72	organizer AND health AND internationalism	0	0
73	organizer AND truth AND universality	0	0
74	organizer AND truth AND internationalism	0	0
75	organizer AND truth AND wide	0	0
76	organizer AND correctness AND universality	0	0
77	organizer AND correctness AND internationalism	0	0
78	organizer AND correctness AND international	0	0
79	organizer AND correctness AND world	0	0
80	organizer AND correctness AND wide	0	0
81	sponsor AND health AND universality	0	0
82	sponsor AND health AND internationalism	0	0

continued on next page

continued from previous page			
S/N	**Translation Combinations**	**Occurrence**	**MI Score**
83	sponsor AND truth AND universality	0	0
84	sponsor AND truth AND internationalism	0	0
85	sponsor AND truth AND wide	0	0
86	sponsor AND correctness AND universality	0	0
87	sponsor AND correctness AND internationalism	0	0
88	sponsor AND correctness AND international	0	0
89	sponsor AND correctness AND wide	0	0
90	sponsor AND correctness AND world	0	0

Table C.3: Ranked translations based on mutual information score.

Appendix D

Preliminary Arabic WordNet Construction

D.1 Introduction

As a repository of lexical information, lexical resources are irreplaceable for every natural language processing (NLP) system. For example, in order to improve the performance of word sense disambiguation applications, an adequate lexical resource is necessary. While in English, and some major European languages, the "lexical bottleneck" problem likely softened e.g., for English WordNet Miller (1995) and for (Dutch, Italian, Spanish, German, French, Czech and Estonian) EuroWordNet (Vossen, 1998), there are no available wide-range lexical resources for other languages such as Arabic. Since it is labor intensive and time consuming to start from scratch and include as much information as possible into a lexical database manually, we have taken an alternative way to build an Arabic WordNet by querying an existing lexical resource e.g., English WordNet and "Arabic English Parallel News Part 1"[1] semi-automatically. Despite

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T18

reducing the time and effort required to manually build such resources, the semi-automatic way may not be accurate enough. For example, the extraction and the SynSet mapping between the English WordNet synset and the planned Arabic WordNet, might be ambiguous. Therefore, we developed an interactive approach where the user plays an essential role, assigning each Arabic word to its equivalent English SynSet.

In this chapter, we give a brief overview about the current development of the Arabic lexical resource, e.g., Arabic WordNet, presenting our contribution in alleviating the acute shortage of such lexical resources. We initially give a brief overview about the Arabic morphological analyzers, followed by a brief overview about Word net, including the Arabic effort for the creation of the Arabic WordNet. In conclusion, we describe our approach in supporting lexicographers in creating Arabic WordNet SynSets. This creation is done query-oriented, where an arabic word is searched and secondly annotated with English SynSets. Parallel corpora are then used to create glosses for every new created Arabic SynSets. A user interface including the functionalities described in our approach is presented and discussed.

D.2 Arabic Morphological Analyzers

In the past few years several studies have been done for automatic morphological analysis of Arabic (Abderrahim and Reguig, 2008). In the following, we restrict our discussion to the two most important Arabic Morphological Analyzer: the finite-state arabic morphological Analyzer and the Tim Buckwalter Arabic morphological analyzer (BAMA) (Buckwalter, 2002).

D.2.1 Finite-State Arabic Morphological Analyzer at Xerox.

In 1996, the Xerox Research Centre Europe produced a morphological analyzer for Modern Standard Arabic. In 1998 a finite-state morphological analyzer of written Modern Standard Arabic words that is available for testing on the Internet was implemented. The system receives online orthographical words of Arabic that can be full diacritics, partial diacritics or without diacritics. The system has wide dictionary coverage. After receiving the words the system analyze them in order to identify affixes and roots from patterns. Beesley (Beesley, 2001) reported that Xerox has several lexicons: the root lexicon which contains about 4390 entries. The second one is a dictionary of patterns which contains about 400 entries. Each root entry is hand-encoded and associated with patterns. The average root participates in about 18 morphologically distinct stems, producing 90000 Arabic stems. When these stems combining with possible prefix and/or suffix by composition, generates 72000000 abstract words.

D.2.2 Tim Buckwalter Arabic morphological analyzer (BAMA).

BAMA is the most well known tool of analyzing Arabic texts (Buckwalter, 2002). It is consist of main database of word forms which interact with other concatenation databases. An Arabic word is considered as concatenation of three regions, a prefix region, a stem region and a suffix region. The prefix and suffix regions can be null. Prefix and suffix lexicon entries cover all possible concatenations of Arabic prefixes and suffixes, respectively. Every word form is entered separately. It takes the stem as the base form. Furthermore it also provides information on the root. (BAMA) morphology reconstructs vowel marks and provides English glossary. It returns all possible compositions of stems and affixes for a word. (BAMA) group together stems with similar meaning with associated it with lemmaID. The (BAMA) contains 38,600 lemmas. For more details about the entire constructions of the (BAMA) we refer the reader to (Habash, 2004).

D.3 WordNet

For better understanding how to create an Arabic lexical resource, we first want to present WordNet, and then give a short introduction about the already existing Arabic WordNet. WordNet is one of the most important English lexical resources available to researchers in the field of text analysis and many related areas. Fellbaum (1998) discussed the design of this electronic lexical database WordNet designed based on psycholinguistic and computational theories of the human lexical memory. WordNet can be used for different applications, like word sense identification, information retrieval, and particularly for a variety of content-based tasks, such as semantic query expansion or conceptual indexing in order to improve information retrieval performance (Vintar et al., 2003). It provides a list of word senses for each word, organized into synonym sets (SynSets), each representing one constitutional lexicalized concept. Every element of a SynSet is uniquely identified by its SynSet identifier (SynSetID). It is unambiguous and a carrier of exactly one meaning. Furthermore, different relations link these elements of synonym sets to semantically related terms (e.g., hyperonyms, hyponyms, etc.). All related terms are also represented as SynSet entries. It also contains descriptions of nouns, verbs, adjectives, and adverbs. WordNet distinguishes two types of linguistic relations. The first type is represented by lexical relations (e.g., synonomy, antonomy and polysemy) and the second by semantic relations (e.g., hyponomy and meronomy). Glosses (human descriptions) are often (about 70% of the time) associated with a SynSet (Ciravegna et al., 1994).

WordNet has been upgraded into different versions. In version of WordNet 2.0 nominalizations, which link verbs and nouns pertaining to the same semantic class were introduced, as well as domain links, based on an "ontology" that should help for the disambiguation process. In the newest version of WordNet 3.0 some changes were made to the graphical interface and

WordNet library with regard to adjective and adverb searches adding “Related nouns" and “Stem Adjectives".

D.4 Arabic WordNet

Black et al. (2006) discussed in their paper an approach to develop an Arabic (WordNet) lexical resource for the Standard Arabic language. The Arabic WordNet project (AWN) bases on the design of the Princeton WordNet (PWN) and is mappable with the PWN version 2.0 and EuroWordNet. The Suggested Upper Merged Ontology (SUMO) and the related domain ontologies are used as the basis for its semantics. The authors already described the “manual” extension and translation of the already existing SynSets from one language (e.g., English) to Arabic (Elkateb, 2005). But it is not clear if and how this manual annotation process is supported by an interactive system.

D.5 Our Approach

In the following, we discuss the *Arabic WordNet Interface* that we implemented, in order to support authors in annotating Arabic words with English SynSets (De Luca et al., 2009). The system can be described through the following steps:

- Arabic Synset Creation
  - The user types an Arabic query word
  - A list of translations in English is retrieved
  - The user checks English translations
  - If a translation is not included, the user can add it through the “other translation” check box
  - An English list of WordNet SynSets related to the chosen translation is retrieved
  - The user checks WordNet SynSets and chooses the correct matching SynSets
  - The SynSetIDs of chosen SynSets are retrieved and assigned to the arabic word
- Arabic Synset Gloss Creation
  - Every word contained in the glosses of every English Synset is retrieved individually
  - The best matching sentences are retrieved from parallel corpora using semantic similarity measures (Patwardhan et al., 2003b)
  - An Arabic list of possible glosses related to the chosen translation are retrieved from the parallel corpora

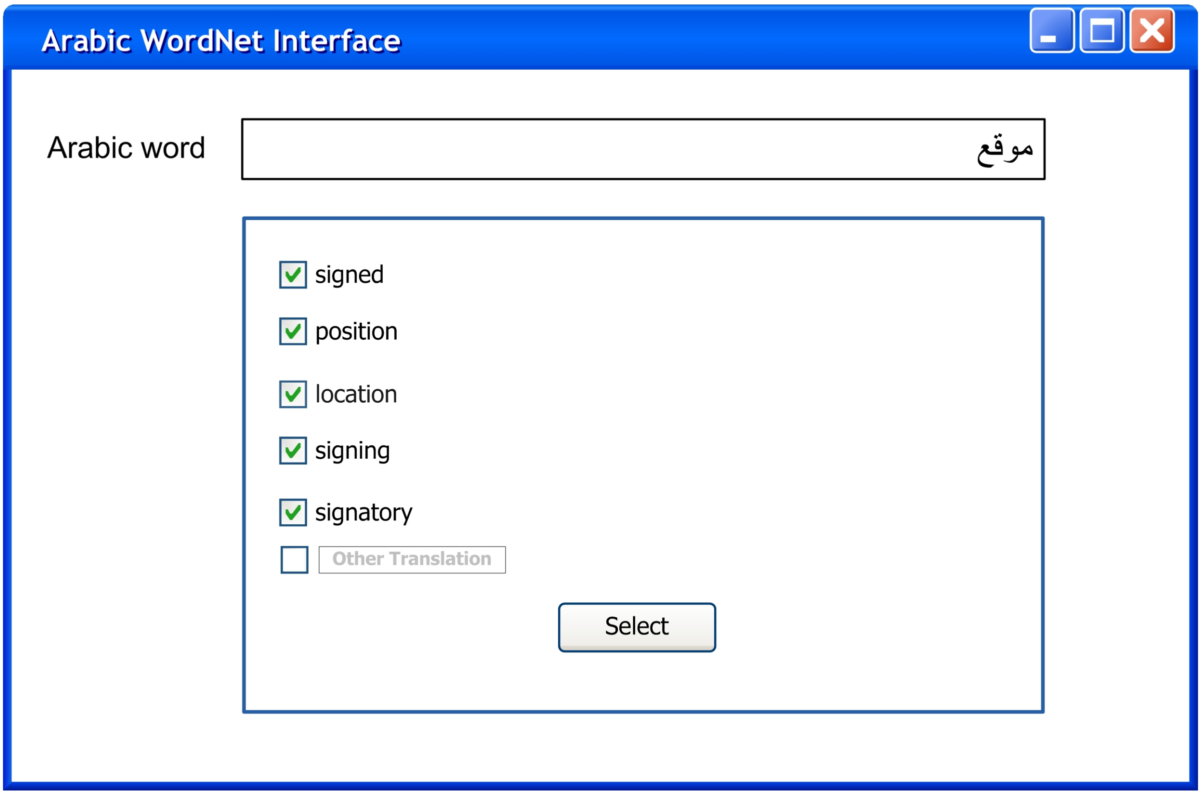
– The user chooses the best matching sentences and thus an Arabic gloss is created

D.5.1 Arabic Synset Creation

TThe process starts after the user submits a query word by means of a client interface (see Figure D.1). In this example, the user is searching for the Arabic word موقعmwqʿ . The system retrieves all matching translations and presents the user with check boxes that the user can activate. The choice of the translations is done by using the araMorph package, that is a sophisticated java-based tool, Buckwalter analyzer (Buckwalter, 2002). This tool includes Java classes for the morphological analysis of Arabic text and the principal Arabic encodings (UTF-8, ISO-8859-6 and CP1256). At this point, the user can decide to maintain all automatically selected translations suggested by the system or choose only the adequate translations from the list, if these conform more to the intended concept described by the query word; the system also gives the possibility of adding a new translation (using the "other" translation check box) that might not be available in the WordNet resource. When these words are selected, the related WordNet SynSets are retrieved and a list of SynSets is presented to the user. Again, in this phase, the user has to choose the best describing SynSet for the searched word (see Figure 6.2). This step is important, in order to retrieve the correct English SynSet that will represent the Arabic word typed at the beginning of the search process. The last step is done when the user has chosen the correct SynSet; the corresponding SynSetID is retrieved and stored together with the Arabic query word. Within this process, we can enrich every Arabic word given as a query, by the user, in a semi-automatic way, creating a new parallel Arabic SynSet with the same SynSetID used in the English WordNet. In this way, we can extend the English WordNet and create an interlingual access through the SynSetID (see Figure D.3).

D.5.2 Arabic Synset Gloss Creation

In order to create the glosses related to the newly created SynSets, different steps have to be considered. The algorithm starts by exploiting the English WordNet glosses and the parallel corpora, in which at least one word contained in the source language (English) matches the translated word (Arabic). Every word contained in the glosses of every English Synset is retrieved and compared with the text included in the relevant English sentences in the "Arabic English Parallel News Part 1" corpora. Semantic similarity measures (Patwardhan et al., 2003b) are applied to compare all words related to the WordNet SynSets with the one contained in the corpora. The best matching sentences, retrieved from the parallel corpora, are presented and an Arabic list of possible glosses related to the chosen translation are presented to the user, who can choose the best matching sentences. These sentences are then added as an Arabic gloss.



Arabic word موقع

☑ signed

☑ position

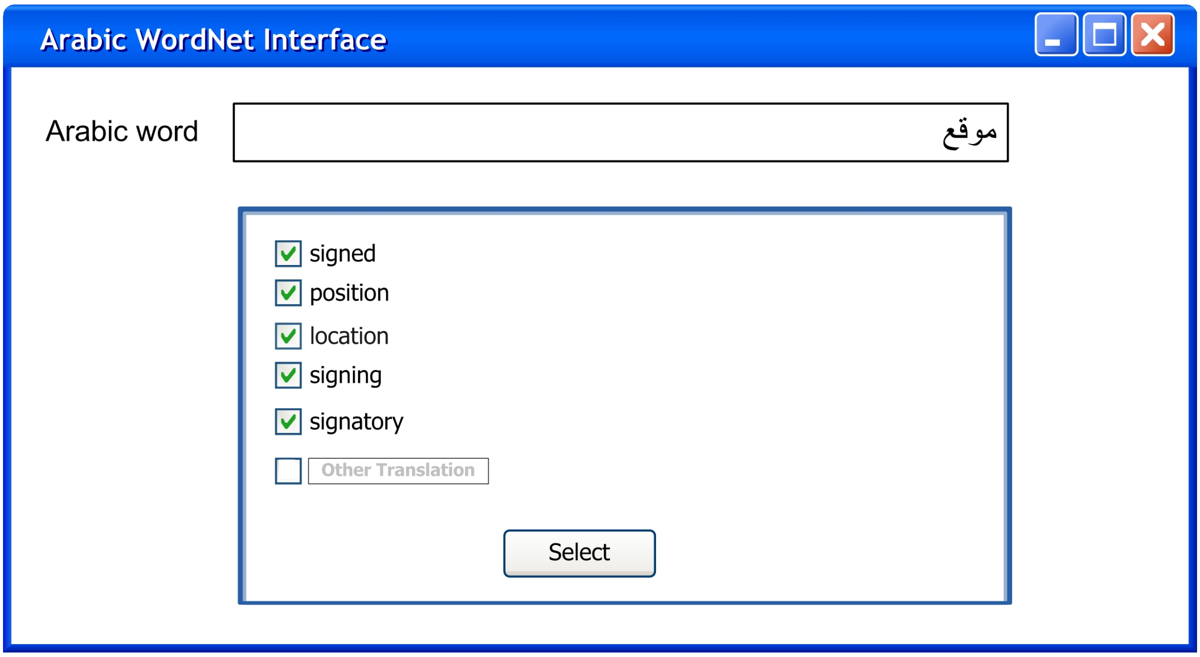
☑ location

☑ signing

☑ signatory

□ Other Translation

Figure D.1: Arabic WordNet Interface - Possible English Translation



Arabic word موقع

☑ signed

☑ position

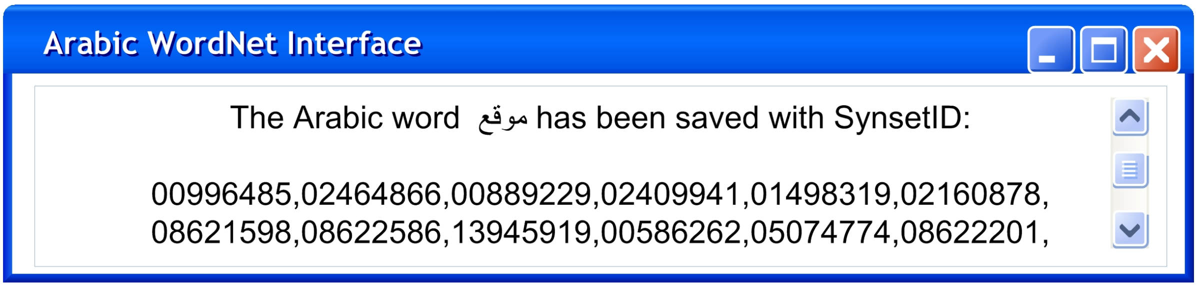
☑ location

☑ signing

☑ signatory

□ Other Translation

Figure D.2: Arabic WordNet Interface - Selecting SynSetIDs for Arabic Word



The Arabic word موقع has been saved with SynsetID:

00996485,02464866,00889229,02409941,01498319,02160878,
08621598,08622586,13945919,00586262,05074774,08622201,

Figure D.3: Arabic WordNet Interface - SynSetID Assignment for Arabic Word

D.6 Conclusion

We presented a tool to support lexicographers in creating Arabic WordNet SynSets. After the discussion of related work, we explained the query-oriented creation of the Arabic SynSets, where an Arabic word is searched and then annotated with English SynSets. Parallel corpora are used to create glosses for every newly created Arabic SynSet. Currently, we are studying how the proposed approach for creating an ArabicWordNet resource can be combined with the approaches presented in (Black et al., 2006). Furthermore, a small user study is planned, in order to evaluate the interface and especially the semi-automatic SynSet and gloss creation process.

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, January 18, 2012

Farag Ahmed

Bibliography

Abdelali, A., Cowie, J. R., Farwell, D., and Ogden, W. C. (2003). Uclir: a multilingual information retrieval tool. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 8(22):103–110.

Abderrahim, M. E. A. and Reguig, F. B. (2008). A morphological analyzer for vocalized or not vocalized arabic language. *Journal of Applied Sciences*, 8(6):984–991.

Abu-Salem, H. (2004). Comparison of stemming and n-gram matching for term-conflation in arabic text. *International Journal of Computer Processing of Oriental languages*, 17(2):61–81.

Abusalah, M., Tait, J., and Oakes, M. P. (2005). Literature review of cross language information retrieval. In *World Academy of Science, Engineering and Technology (2)'05*, pages 175–177.

Adamson, G. W. and Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10(9-10):253–260.

Agirre, E., Atserias, J., Padr, L., and Rigau, G. (2000). Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *In Computers and the Humanities, Special Double Issue on SensEval*, 34(1):103–108.

Ahmed, F. (2010). An interactive system to support cross-lingual retrieval using contextual information. Master's thesis, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany.

Ahmed, F., De Luca, E. W., and Nürnberger, A. (2007). MultiSpell: an n-gram based language-independent spell checker. In *Poster Postproceedings of 8th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007)*, Mexico.

Ahmed, F. and Nürnberger, A. (2007). N-grams conflation approach for arabic text. In *International Workshop on improving Non English Web Searching iNEWS07, In conjunction with The 30th Annual International (ACM SIGIR)*, pages 39–46.

Ahmed, F. and Nürnberger, A. (2008a). Arabic/english word translation disambiguation approach based on naive bayesian classifier. In *Proceedings of the 3th International Multiconfer-*

ence on Computer Science and Information Technology (IMCSIT08), pages 331–338, Wisla, Poland.

Ahmed, F. and Nürnberger, A. (2008b). Arabic/english word translations disambiguation using parallel corpus and matching scheme. In *Proceedings of the 12th European Machine Translation Conference (EAMT08)*, pages 6–11.

Ahmed, F. and Nürnberger, A. (2008c). arasearch: Improving arabic text retrieval via detection of word form variations. In *1st International Conference on Information Systems and Economic Intelligence (SIIE'2008)*, pages 309–323.

Ahmed, F. and Nürnberger, A. (2008d). Corpora based approach for arabic/english word translation disambiguation. *Journal of Speech and Language Technology*, 9:195–213.

Ahmed, F. and Nürnberger, A. (2009). Evaluation of n-gram conflation approaches for arabic text retrieval. *Journal of the American Society for Information Science and Technology (JASIST)*, 60(7):1448–1465.

Ahmed, F. and Nürnberger, A. (2010). multi searcher: can we support people to get information from text they can't read or understand? In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR10)*, pages 837–838, New York, NY, USA. ACM.

Ahmed, F. and Nürnberger, A. (2011). A web statistics based conflation approach to improve arabic text retrieval. In *Proceedings of the Federated Conference Computer Science and Information Systems*, pages 3–9.

Ahmed, F. and Nürnberger, A. (2012). Literature review of interactive cross language information retrieval tools. *The International Arab Journal of Information Technology*, 9(3). (to appear).

Ahmed, F., Nürnberger, A., and De Luca, E. W. (2009a). A corpus-based approach to improve arabic/english cross-language information retrieval. In *Proceedings of the Corpus Linguistics Conference, Liverpool, Uk*.

Ahmed, F., Nürnberger, A., and De Luca, E. W. (2009b). Revised n-gram based automatic spelling correction tool to improve retrieval effectiveness. *Research journal on computer science and computer engineering with applications (Polibits)*, 40:39–48.

Ahmed, F., Nürnberger, A., and Nitsche, M. (2011). Supporting arabic cross-lingual retrieval using contextual information. In Rauber, A. and de Vries (Eds.), A., editors, *Multidisciplinary Information Retrieval*, volume 6653, pages 30–45. Springer-Verlag, Berlin-Heidelberg.

Al-Fedaghi, S. and Al-Anzi, F. (1989). A new algorithm to generate arabic root-pattern forms. In *Proceedings of the 11th National Computer Conference, King Fahd University of Petroleum and Minerals*, pages 04–07, Dhahran, Saudi Arabia.

Al-onaizan, Y. and Knight, K. (2002a). Machine transliteration of names in arabic text. In *ACL Workshop on Comp. Approaches to Semitic Languages*, pages 34–46.

Al-onaizan, Y. and Knight, K. (2002b). Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL02)*, pages 400–408.

Ali, O. M., GadAlla, M., and Abdelwahab, M. S. (2009). Improving machine translation using hybrid dictionary-graph based word sense disambiguation with semantic and statistical methods. *International Journal of Computer and Electrical Engineering (IJCEE)*, 1(5):618–623.

Aljlayl, M. and Frieder, O. (2001). Effective arabic-english cross-language information retrieval via machine readable dictionaries and machine translation. In *ACM Tenth Conference on Information and Knowledge Managemen (CIKM*, pages 295–302. ACM Press.

Aljlayl, M., Frieder, O., and Grossman, D. (2002). On arabic-english cross-language information retrieval: A machine translation approach. *Information Technology: Coding and Computing, International Conference on*, 0:2–7.

Alvarez, C. J., Urrutia, M., DomÃnguez, A., and SÃ¡nchez-Casas, R. (2011). Processing inflectional and derivational morphology: electrophysiological evidence from spanish. *Neuroscience Letters*, 490(1):6–10.

Aqeel, S. U., Beitzel, S. M., Jensen, E. C., Grossman, D., and Frieder, O. (2006). On the development of name search techniques for arabic. *Journal of the American Society of Information Science and Technology (JASIST)*, 57(6):728–739.

Arbabi, M., Fischthal, S. M., Cheng, V. C., and Bart, E. (1994). Algorithms for arabic names transliteration. *IBM Journal of Research and Development*, 38(2).

Atkinson, K. H. (2008). Toward a more rational patent search paradigm. In *Proceedings of the 1st ACM workshop on Patent information retrieval*, PaIR '08, pages 37–40, New York, NY, USA. ACM.

Baldwin, T., Kim, S. N., Bond, F., Fujita, S., Martinez, D., and Tanaka, T. (2008). Mrd-based word sense disambiguation: Further extending lesk. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 775–780.

Ballesteros, L. and Croft, B. (1996). Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pages 791–801.

Ballesteros, L. and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98)*, pages 64–71, New York, NY, USA. ACM.

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02)*, pages 136–145, London, UK. Springer-Verlag.

Bas, D., Broda, B., and Piasecki, M. (2008). Towards word sense disambiguation of polish. In *Proceedings of the 3th International Multiconference on Computer Science and Information Technology (IMCSIT08)*, pages 73–78.

Bashir, S. and Rauber, A. (2010). Improving retrievability of patents in prior-art search. In *ECIR*, pages 457–470.

Beesley, K. (2001). Finite-state morphological analysis and generation of arabic at xerox research: Status and plans in 2001. In *ACL Workshop on Arabic Language Processing*, pages 1–8, Toulouse,France.

Berger, A. L., Brown, P. F., Pietra, S. A. D., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., Printz, H., and Ures, L. (1994). The candide system for machine translation. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 157–162, Plainsboro, New Jersey.

Berlian, V., Vega, S. N., and Bressan, S. (2001). Indexing the indonesian web: Language identification and miscellaneous issues). In *Proceedings of Tenth International World Wide Web Conference*, Hong Kong.

Bian, G.-W. and Teng, S.-Y. (2008). Integrating query translation and text classification in a cross-language patent access system. In *PROCEEDINGS OF NTCIR-7 WORKSHOP MEETING*, pages 16–19.

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C. (2006). Introducing the arabic wordnet project. In *Proceedings of the 3rd International WordNet Conference 2006*.

Braschler, M., Peters, C., and Schäuble, P. (2000). Cross-language information retrieval (clir) track overview. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pages 25–33.

Braune, F. and Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 81–89, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brill, E. and Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL '00)*, pages 286–293, Morristown, NJ, USA. Association for Computational Linguistics.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, CA.

Brown, R. (1998). Automatically-extracted thesauri for cross language ir: When better is worse. In *Proceedings of the 1st Workshop on Computational Terminology (Computerm)*, pages 15–21.

Buckwalter, T. (2002). Arabic morphological analyzer version 1.0. Website. Available online at `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49`; visited on January 8th 2010.

Cabezas, C. and Resnik, P. (2005). Using wsd techniques for lexical selection in statistical. In *Institute for Advanced Computer Studies, University of Maryland.*

Capstick, J., Diagne, A. K., Erbach, G., Uszkoreit, H., Leisenberg, A., and Leisenberg, M. (2000). A system for supporting cross-lingual information retrieval. *Information Processing and Management: an International*, 36(2):275–289.

Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y., and Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In *PROCEEDINGS OF THE FIFTEENTH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 708–714.

Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O. (2001). Improving precision in information retrieval for swedish us-ing stemming. In *Proceedings of NODALIDA '01 - 13th Nordic conference on computational linguistics*, Uppsala, Sweden.

Ceausu, A., Tinsley, J., Zhang, J., and Way, A. (2011). Experiments on domain adaptation for patent machine translation in the pluto project. In *proceedings of the 15th conference of the European Association for Machine Translation (EAMT 2011)*, pages 21–28.

Chan, Y. S. and Ng, H. T. (2007). Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 33–40.

Chang, J. S. and Chert, M. H. C. (1994). Using partial aligned parallel text and part-of-speech information in word alignment. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA94)*, pages 16–23.

Chen, A. and Gey, F. C. (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Inf. Retr.*, 7(1-2):149–182.

Chen, H.-H., Bian, G.-W., and Lin, W.-C. (1999). Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 215–222, Morristown, NJ, USA. Association for Computational Linguistics.

Cheng, P.-C., Chien, B.-C., Ke, H.-R., and Yang, W.-P. (2006). Using ontological chain to resolve the translation ambiguity of cross-language information retrieval. In *Proceedings of the 5th WSEAS international conference on Telecommunications and informatics*, TELE-INFO'06, pages 446–451, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

Ciravegna, F., Magnini, B., Pianta, E., and Strapparava, C. (1994). A project for the construction of an italian lexical knowledge base in the framework of wordnet. Technical Report 9406-15, IRST-ITC.

Cowie, J., Guthrie, J., and Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proceedings of the workshop on Speech and Natural Language*, pages 238–242, Morristown, NJ, USA. Association for Computational Linguistics.

Cristianini, N., Shawe-Taylor, J., and Lodhi, H. (2002). Latent semantic kernels. *J. Intell. Inf. Syst.*, 18(2-3):127–152.

Cucerzan, S. and Brill, E. (2004). Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 293–300.

Dagan, I., Church, K. W., and Gale, W. A. (1993). Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, pages 1–8, Columbus, Ohio.

Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Darwish, K. and Oard, D. W. (2002). Term selection for searching printed arabic. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR–2002)*, pages 261–268, Tampere, Finland.

Davis, M. (1996). New experiments in cross-language text retrieval at nmsus computing research lab. In *Proceedings of the 5th Text Retrieval Conference (TREC-5)*, pages 447–454.

Davis, M. W. and Ogden, W. C. (1997). Quilt: implementing a large-scale cross-language text retrieval system. *SIGIR Forum*, 31(SI):92–98.

Davis, M. W. and Ogden, W. C. (1998). Free resources and advanced alignment for cross-language text retrieval. In Voorhees, E. M. and Harman, D. K., editors, *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, pages 385–394.

de Gispert, A. and Mariño, J. B. (2006). Linguistic tuple segmentation in ngram-based statistical machine translation. In *Proc. of the 9th Int. Conf. on Spoken Language Processing (Interspeech 06)*, pages 1149–1152.

De Luca, E. W., Ahmed, F., and Nürnberger, A. (2009). Annotating arabic words with english wordnet synsets: An arabic wordnet interface. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 61–68.

De Luca, E. W., Hauke, S., Nürnberger, A., and Schlechtweg, S. (2006). MultiLexExplorer - combining multilingual web search with multilingual lexical resources. In *Combined Works. on Language-enhanced Educat. Techn. and Devel. and Eval. of Robust Spoken Dialogue Sys.*, pages 17–21.

Deorowicz, S. and Ciura, M. G. (2005). Correcting spelling errors by modeling their causes. *International Journal of Applied Mathematics and Computer Science*, 15(2):275–285.

Doi, S. and Muraki, K. (1992). Translation ambiguity resolution based on text corpora of source and target language. In *Proceedings of the 15th Conference on Computational Linguistics COLING-92*, pages 525–531.

Ekmekcioglu, F. C., Lynch, M. F., and Willett, P. (1996). Stemming and n-gram matching for term conflation in turkish texts. *Information Research News*, 7(1):2–6.

Elkateb, S. (2005). *Design and implementation of an English Arabic dictionary/editor.* PhD thesis, Manchester University.

Escudero, Gerard, L. M. and Rigau, G. (2000). Boosting applied to word sense disambiguation. In *Proceedings of the 12th European Conference on Machine Learning (ECML)*, pages 129–141, Barcelona, Spain.

Fakhrahmad, S., Rezapour, A., Jahromi, M. Z., and Sadreddini, M. (2011). A new word sense disambiguation system based on deduction. In *Proceedings of the World Congress on Engineering (WCE 2011)*, pages 1271–1281.

Fellbaum, C. (1998). *WordNet, an electronic lexical database.* MIT Press.

Fernandez-Amoros, D., Gil, R. H., Somolinos, J. A. C., and Somolinos, C. C. (2010). Automatic word sense disambiguation using cooccurrence and hierarchical information. In *Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems*, pages 60–67, Berlin, Heidelberg. Springer-Verlag.

Gabrilovich, E., Broder, A., Fontoura, M., Joshi, A., Josifovski, V., Riedel, L., and Zhang, T. (2009). Classifying search queries using the web as a source of knowledge. *ACM Transactions on the Web*, 3(2):1–28.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACLgl)*, pages 177–184.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992a). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992b). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language (HLT '91)*, pages 233–237, Morristown, NJ, USA. Association for Computational Linguistics.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992c). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'92)*, pages 101–112.

Galley, M. and McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. In *Proceedings of the 18th international joint conference on Artificial intelligence(IJCAI'03)*, pages 1486–1488, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gaona, M. A. R., Gelbukh, A., and Bandyopadhyay, S. (2009). Web-based variant of the lesk approach to word sense disambiguation. In *Proceedings of 2009 Eighth Mexican International Conference on Artificial Intelligence (MICAI 2009)*, pages 103–107.

Gearailt, D. N., Gearailt, C. D. N., and College, C. (2005). Dictionary characteristics in cross-language information retrieval. Technical Report - Number 616. Available online at `http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-616.pdf`; visited on October 5th 2010.

Gelbukh, A., Alexandrov, M., and Han, S. (2004). Detecting inflection patterns in NL by minimization of morphological model. In *Progress in Pattern Recognition, Image Analysis and Applications (CIARP 2004)*, LNCS 3287, pages 432–438. Springer.

Ghaoui, A., Yvon, F., Mokbel, C., and Chollet, G. (2005). On the use of morphological constraints in n-gram statistical language model. In *Proceedings of Interspeech-2005*, pages 1281–1284.

Golding, R. and Roth, D. (1999). A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1):107–130.

Gough, N. and Way, A. (2004). Robust large-scale ebmt with marker-based segmentation. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04*, pages 95–104.

Greengrass, M., Robertson, A. M., Robyn, S., and Willett (1996). Processing morphological variants in searches of latin text. *Information research news*, 6(4):2–5.

Habash, N. (2004). Large scale lexeme based arabic morphological generation. In *Proc. of TALN-04*, Fez, Morocco.

Hassan, H., Hearne, M., and Way, A. (2007). Supertagged phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL07)*, pages 288–295.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., and Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from clef 2000/2002. *Information Retrieval*, 7(12):99–119.

Higuchi, S., Fukui, M., Fujii, A., and Ishikawa, T. (2001). Prime: A system for multi-lingual patent retrieval. In *Proceedings of MT Summit VIII*, pages 163–167.

Hodge, V. J. and Austin, J. (2003). A comparison of standard spell checking algorithms and a novel binary neural approach. *IEEE Trans. on Knowl. and Data Eng.*, 15(5):1073–1081.

Hull, D. A. and Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In Frei, H.-P., Harman, D., Schäuble, P., and Wilkinson, R., editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 49–57. ACM.

Ide, N. (1999). Parallel translations as sense discriminators. In *Proceedings of SIGLEX99: Standardizing Lexical Resources, ACL99 Workshop*, pages 52–61, College Park, Maryland.

Indrajit Bhattacharya, Lise Getoor, Y. B. (2004). Unsupervised sense disambiguation using bilingual probabilistic models. In *Proceedings of ACL 2004*, pages 187–194.

Iwayama, M. and Tokunaga, T. (1995). Hierarchical bayesian clustering for automatic text classification. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, pages 1322–1327, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jang, M.-G., Myaeng, S. H., and Park, S. Y. (1999). Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 223–229, Morristown, NJ, USA. Association for Computational Linguistics.

Jochim, C., Lioma, C., Schütze, H., Koch, S., and Ertl, T. (2010). Preliminary study into query translation for patent retrieval. In *Proceedings of the 3rd international workshop on Patent information retrieval*, PaIR '10, pages 57–66, New York, NY, USA. ACM.

Kang, S.-J. (2003). Corpus-based ontology learning for word sense disambiguation. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pages 399–407.

Ker, M. and Chang, J. S. (1997). A class-based approach to word alignment. *Computational Linguistics*, 32(2):313–343.

Khaltar, B.-O., Fujii, A., and Ishikawa, T. (2006). Extracting loanwords from mongolian corpora and producing a japanese-mongolian bilingual dictionary. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44)*, pages 657–664, Morristown, NJ, USA. Association for Computational Linguistics.

Khoja, S. and Garside, R. (1999). Stemming arabic. Website. Available online at `http://zeus.cs.pacificu.edu/shereen/research.htm`; visited on January 15th 2009.

Khreisat, L. (2006). Arabic text classification using n-gram frequency statistics a comparative study. In *Proceedings of 2006 International Conference on Data Mining, Part of the 2006 World Congress in Computer Sciences DMIN 2006*, pages 78–82.

Knight, K. and Graehl, J. (1997). Machine transliteration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (ACL)*, pages 128–135.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Konishi, K., Kitauchi, A., and Takaki, T. (2004). Invalidity patent search system at ntt data. In *Proceedings of NTCIR-4 Workshop Meeting*, Tokyo, Japan.

Kraaij, W. and Pohlmann, R. (1996). Viewing stemming as recall enhancement. In *Proceedings of ACM SIGIR96*, pages 40–48.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439.

K.W., C. and W.A., G. (1991). Probability scoring for spelling correction. *Statistics and Computing*, 1(1):93–103.

Larkey, L., Ballesteros, L., and Connell, M. (2007). Light stemming for arabic information retrieval. In Soudi, A., den Bsch, A. V., and Neumann, G., editors, *Arabic computational morphology*, volume 38, pages 221–243. Springer-Verlag, Netherlands.

Larkey, L. S., AbdulJaleel, N., and Connell, M. (2003). What's in a name?: Proper names in arabic cross language information retrieval. In *ACL Workshop on Comp. Approaches to Semitic Languages*.

Latiri, C., Smaïli, K., Lavecchia, C., Nasri, C., and Langlois, D. (2011). Phrase-based machine translation based on text mining and statistical language modeling techniques. In *Proceedings of the12th International Conference on Intelligent Text Processing and Computational Linguistics - CICLing2011*, Tokyo, Japan.

Lefever, E. and Hoste, V. (2010). Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 15–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation SIGDOC1986*, pages 24–26.

Leveling, J., Magdy, W., and Jones, G. J. (2011). An investigation of decompounding for cross-language patent search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 1169–1170, New York, NY, USA. ACM.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(08):707–710.

Levow, G.-A., Oard, D. W., and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Inf. Process. Manage.*, 41(3):523–547.

Li, Y. and Shawe-Taylor, J. (2007). Advanced learning algorithms for cross-language patent retrieval and classification. *Inf. Process. Manage.*, 43(5):1183–1199.

Lin, D. (2000). Word sense disambiguation with a similarity-smoothed case library. *Computers and the Humanities*, 34(1-2):147–152.

Litkowski, K. C. (2000). Senseval: The cl research experience. *Computers and the Humanities*, 34(1-2):153–158.

Lopez-Ostenero, F., Gonzalo, J., Penas, A., and Verdejo, F. (2002). Interactive cross-language searching: Phrases are better than terms of query formulation and refinement. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *CLEF*, volume 2785 of *Lecture Notes in Computer Science*, pages 416–429. Springer.

Ma, Y., Stroppa, N., and Way, A. (2007). Bootstrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics ACL2007*, pages 304–311.

Mangu, L. and Brill, E. (1997). Automatic rule acquisition for spelling correction. In *Proceedings of the 14th International Conference on Machine Learning*, pages 187–194.

Martins, B. and Silva, M. J. (2004). Spelling correction for search engine queries. In *Proceedings of EsTAL-04, Espana for Natural Language Processing.*

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., and Asahara, M. (1999). Japanese morphological analysis system chasen version 2. Technical report, NaraInstitute of Science and Technology Japan. Technical Report NAIST-IS-TR99009, NAIST.

Mayfield, J., McNamee, P., Costello, C., Piatko, C., and Banerjee, A. (2002). Experiments in filtering and in arabic, video, and web retrieval. In *Proceedings of the Tenth Text Retrieval Conference (TREC 2001), Gaithersburg, Maryland.*

McCarthy, D. and Carroll, J. (2003). Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Comput. Linguist.*, 29(4):639–654.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04)*, page 279, Morristown, NJ, USA. Association for Computational Linguistics.

Meftouh, K., Laskri, T., and SmaÃ¯li, K. (2010). Modeling arabic language using statistical methods. *The Arabian Journal of Science and Engineering*, 35(2C):69–82.

Melamed, I. D. (2000). Models of translational equivalence among words. *Computational Linguistics*, 26(21):221–249.

Mihalcea, R. (2007). Knowledge-based methods for wsd. In Agirre, Eneko; Edmonds, P. E., editor, *Word Sense Disambiguation / Algorithms and Applications*, volume 33, pages 107–131. Springer-Verlag, Netherlands.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

ming Zhan, J., Mou, X., Li, S., and Fang, D. (1998). A language model in a large-vocabulary speech recognition system. In *International Symposium on Chinese Spoken Language Processing (ISCSLP 1998).*

Monz, C. and de Rijke, M. (2002). Shallow morphological analysis in monolingual information retrieval for dutch, german and italian. In *Proc. of Evaluation of Cross-Language Information Retrieval Systems CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 262–277. Springer-Verlag.

Moukdad, H. (2004). Lost in cyberspace: How do search engines handle arabic queries? In *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science, Winnipeg.*

Moukdad, H. and Large, A. (2001). Information retrieval from full-text arabic databases: Can search engines designed for english do the job? *International Journal of Libraries and Information Services*, pages 63–74.

Moulinier, I., McCulloh, A., and Lund, E. (2001). Non-english monolingual retrieval. in cross-language information retrieval and evaluation. In *Proceedings of CLEF 2000 workshop*, pages 176–187. Springer-Verlag.

Mustafa, S. H. (2004). Character contiguity in n-gram-based word matching: the case for arabic text searching. *Processing and Management*, 41(4):819–827.

Nelken, R. and Shieber, S. M. (2007). Lexical chaining and word-sense-disambiguation. Technical Report TR-06-07, School of Engineering and Applied Sciences, Harvard University, Cambridge, MA.

Ng, T. and Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

Nie, J. Y., Simard, M., Isabelle, P., and Durard, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (ACM SIGIR99)*, pages 74–81.

Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann Publishers, San Francisco, California.

Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems (CHI '93)*, pages 206–213, New York, NY, USA. ACM.

Niessen, S., Vogel, S., Ney, H., and Tillmann, C. (1998). A dp-based search algorithm for statistical machine translation. In *Proceedings og COLING-ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 960–967, Montreal, Canada.

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2-3):103–134.

Oard, D. W. (1997a). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on cross-language text and speech retrieval. American Association for Artificial Intelligence*, pages 131–139.

Oard, D. W. (1997b). Cross-language text retrieval research in the usa. In *Proceedings of the Third DELOS Workshop; Cross-Language Information Retrieval, number 97-W003in Ercim Workshop Proceedings, European Research Consortium for Informatics and Mathematics*, pages 7–16.

Oard, D. W. and Diekema, A. R. (1998). Experiments in multi-lingual information retrieval. *Cross-language information retrieval, Annual Review of Information Science and Technology (ARIST)*, 33:223–256.

Oard, D. W., He, D., and Wang, J. (2008). User-assisted query translation for interactive cross-language information retrieval. *Information Processing and Management: an International Journal*, 44(1):181–211.

Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. in coling '00. In *Proceedings of 18th International Conference on Computational Linguistics*, pages 1086–1090.

Och, F. J. and Ney, H. (2003a). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. and Ney, H. (2003b). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Ogden, W. C. and Davis, M. W. (2000). Improving cross-language text retrieval with human interactions. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, page 3044, Washington, DC, USA. IEEE Computer Society.

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003a). Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03)*, pages 241–257.

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003b). Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proc. of the Fourth Int. Conf. on Intell. Text Processing and Computational Linguistics*, pages 241–257, Mexico City, Mexico.

Pedersen, T. (2000). A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. In *Proceedings of 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 63–69.

Penkale, S., Haque, R., Dandapat, S., Banerjee, P., Srivastava, A. K., Du, J., Pecina, P., Naskar, S. K., Forcada, M. L., and Way, A. (2010). Matrex: the dcu mt system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 143–148, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., Herring, P., and Hansen, P. (2004). Observing users, designing clarity: A case study on the user-centered design of a cross-language information retrieval system. *Journal of the American Society for Information Science and Technology (JASIST)*, 55(10):923–934.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63. ACM Press.

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., and Järvelin, K. (2003). Fuzzy translation of cross-lingual spelling variants. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*, pages 345–352.

Piroi, F. (2010). Clef-ip 2010: Retrieval experiments in the intellectual property domain. In Braschler, M., Harman, D., and Pianta, E., editors, *CLEF (Notebook Papers/LABs/Workshops)*.

Pollock, J. J. and Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholary text. *Journal of the American Society for Information Science and Technology (JASIST)*, 34(1):51–58.

Pollock, J. J. and Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. *Commun. ACM*, 27(4):358–368.

Popovic, M. and Willett, P. (1992). The effectiveness of stemming for natural-language access to slovene textual data. *Journal of the American Society for Information Science (JASIS)*, 43(5):384–390.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Preiss, J., Dehdari, J., King, J., and Mehay, D. (2009). Refining the most frequent sense baseline. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions(DEW '09)*, pages 10–18, Morristown, NJ, USA. Association for Computational Linguistics.

Qu, Y., Grefenstette, G., and Evens, D. (2002). Resolving translation ambiguity using monolingual corpora: A report on clairvoyance clef-2002 experiments. In *CLEF-2002 working notes*, pages 115–126.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems Management and Cybernetics*, 19(1):17–30.

Resnik, P. (1995). Disambiguating noun groupings with respect to wordnet senses. In *Proceedings of the third workshop on very large corpora*, pages 54–68.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC.

Roeck, A. N. D. and Al-Fares, W. (2000). A morphologically sensitive clustering algorithm for identifying arabic roots. In *Proceedings of ACL 2000*, pages 199–206, Hong Kong.

Salton, G. (1973). Experiments in multi-lingual information retrieval. *Information Processing Letters*, 2(1):6–11.

Samy, D., Moreno, A., and Guirao, J. M. (2005). A proposal for an arabic named entity tagger leveraging a parallel corpus. In *Proceedings of International Conference on Recent Advances on Natural Language Processing RANLP2005*, pages 459–465.

Sari, S. and Adriani, M. (2008). Using mutual information technique in cross-language information retrieval. In *Proceedings of the 11th International Conference on Asian Digital Libraries (ICADL 08)*, pages 276–284, Berlin, Heidelberg. Springer-Verlag.

Savoy, J. (2002). Report on clef 2002 experiments: Combining multiple sources of evidence. In *CLEF*, pages 66–90.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Semmar, N., Elkateb-Gara, F., Laib, M., and Fluh, C. (2005). A cross-language information retrieval system based on linguistic and statistical approaches. In *Proceedings of 2eme Congres International sur L'Ingenierie de la Langue*, pages 1–12.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell Systems Technical Journal*, 30:50–64.

Shinnou, H. and Sasaki, M. (2003). Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the em algorithm. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 41–48.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Specia, L., Stevenson, M., and das Gracas Volpe Nunes, M. (2007). Learning expressive models for word sense disambiguation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, pages 41–48. The Association for Computer Linguistics.

Stalls, B. G. and Knight, K. (1998). Translating names and technical terms in arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, pages 34–41.

Stroppa, N. and Way, A. (2006). Matrex: Dcu machine translation system for iwslt 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 31–36.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international conference on Information and knowledge management (CIKM '93)*, pages 67–74, New York, NY, USA. ACM.

Tai, S. Y., Ong, C., and Abdullah, N. A. (2000). On designing an automated malaysian stemmer for the malay language. In *Proceedings of the fifth international workshop on information retrieval with Asian languages, Hong Kong*, pages 207–208.

Tait, J., editor (2008). *Proceedings of the 1st ACM workshop on Patent Information Retrieval, PaIR 2008, Napa Valley, California, USA, October 30, 2008*. ACM.

Tinsley, J., Ma, Y., Ozdowska, S., and Way, A. (2008). Matrex: the dcu mt system for wmt 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 171–174, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tinsley, J. and Way, A. (2009). Automatically-generated parallel treebanks and their exploitability in phrase-based statistical machine translation. *Machine Translation*, 34(1):1–22.

Tiwana, S. and Horowitz, E. (2009). Findcite: automatically finding prior art patents. In *Proceedings of the 2nd international workshop on Patent information retrieval*, PaIR '09, pages 37–40, New York, NY, USA. ACM.

Toutanova, K. and Moore, R. C. (2002). Pronunciation modeling for improved spelling correction. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, pages 144–151, Morristown, NJ, USA. Association for Computational Linguistics.

Towell, G. and Voorhees, E. M. (1998). Disambiguating highly ambiguous words. *Computational Linguistics*, 24(1):125–146.

Turba, T. N. (1982). Length-segmented lists. *Communications of the ACM*, 25(8):522–526.

van Gompel, M. (2010). Uvt-wsd1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 238–241, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-sense disambiguation for machine translation. In *Joint Human Language Technology conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 771–778.

Vintar, S., Buitelaar, P., and Volk, M. (2003). Semantic relations in concept-based cross-language medical information retrieval. In *Proceedings of the Workshop on Adapt. Text Extraction and Mining*, Croatia.

Voorhees, E. M. (1993). Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '93)*, pages 171–180, New York, NY, USA. ACM.

Vossen, P., editor (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *J. ACM*, 21(1):168–173.

Wang, Y.-Y. and Waibel, A. (1998). Fast decoding for statistical machine translation. In *Proceedings of the International Conference on Speech and Language Processing*, pages 1357–1363.

Wilks, Y., Fass, D., ming Guo, C., McDonald, J. E., Plate, T., and Slator, B. M. (1993). Providing machine tractable dictionary tools. *Semantics and the Lexicon (Pustejovsky J. ed.)*, pages 341–401.

Wu, D. (1996). A polynomial-time algorithm for statistical machine translation. In *Proceedings of the 34th Annual Conference of the Association for Computational Linguistics (ACL '96)*, pages 152–158.

Xu, J. and Croft, W. B. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, 16(1):61–81.

Xue, X. and Croft, W. B. (2009). Automatic query generation for patent search. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 2037–2040, New York, NY, USA. ACM.

Yamamoto, K. and Matsumoto, Y. (2000). Acquisition of phrase-level bilingual correspondence using dependency structure. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 933–939, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang, Y., Carbonell, J. G., Brown, R. D., and Frederking, R. E. (1998). Translingual information retrieval: learning from bilingual corpora. *Artificial Intelligence*, 103:323–345.

Yang, Y. and Ma, N. (2002). Cmu in cross-language information retrieval at ntcir-3. In *Proceedings of the third NTCIR workshop on research in information retrieval - automatic text summarization and question answering.*

Yannakoudakis, E. J. and Fawthrop, D. (1983). An intelligent spelling error corrector. *Inf. Process. Manage.*, 19(2):101–108.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpus. In *Proceedings of COLING-92*, pages 454–460, Nantes, France.

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the workshop on Human Language Technology(HLT '93)*, pages 266–271, Morristown, NJ, USA. Association for Computational Linguistics.

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Meeting of the Association for Computational Linguistics*, pages 189–196.

Yarowsky, D. and Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 207–216.

Ye, P. (2004). Selectional preference based verb sense disambiguation using wordnet. In *Proceedings of the Australasian Language Technology Workshop*, pages 155–162.