Direct Machine Translation and Formalization Issues of Language Structures and Their Matches by Automated Machine Translation for the Russian-English Language Pair

Anna Novikova
Perm National Research Polytechnic University, Komsomolsky Ave. 29, Perm, Russia
novikova@yandex.ru

Keywords: Machine Translation, Semantic Match, Structural Match, Formalization Of Language Structures, Formalization Language, Distributive Principles, Surface Language Structures, Syntactic Positions, Skeleton Language Structure, In-Depth Model, Pre-Processing Of Text, Referential Text Corpus.

Abstract: The present paper introduces a formalization language for sentences and text corpora that helps tackle the acute problem of formalizing semantic and structural matches of different language systems by direct machine translation. In the paper, a detailed look is taken at the elements of reference and situation-based analysis of the representations of surface and in-depth semantics of semantic contexts in the sphere of business communication relying on the meaning-text theory for automated formalization of language structures and their matches by machine translation. The study is aimed at working out an algorithm that will enhance the quality of machine translation excluding intermediary natural languages and post-editing of translated texts. A distinguishing feature of the suggested approach is structurization and formalization of language structures on the stage of text pre-processing. Such «input filter» of information will enable decoding system create literate messages in compliance with lexical and grammatical distributive principles both in a foreign language and native language, and use them for proofreading of texts in a native language and building structures that can be further translated by existing machine translation systems with high quality.

1 INTRODUCTION

Mankind's history cannot be imagined without intercultural communication that is increasing over decades. Strong evidence of this process is the penetration of a vast number of adopted words in different languages and the appearance of well-known global terms and expressions. Despite that, the issues of automated translation were firstly addressed only in the 20-s of the XX century. The pioneers of this movement are considered to be Estonian (A. Vakher), American (W. Weaver, H.P. Edmundson, P.G. Hays), French (G. Artsrouni), and Russian (P.P. Smirnov-Troyanskii) scientists [1]. The introduction of computing machines contributed to further development of this scientific movement. In the middle of the XX century, the achievements of Warren Weaver of the Rockefeller Foundation and RAND in computerized translation became world-renowned [1]; those forefront achievements laid down a fundamental translation principle when human translation was replaced by computer as a mediator. Since that time, machine translation has distinguished into a separate science intensive direction.

However, despite the rapid development of information technologies, the appearance of artificial intelligence, machine learning, the development of mathematical statistics and the enhancement of computer technics there is still no solution for the problem of machine translation that is able to fully replace a human mediator [2].

Modern implementations in the considered area help solve the problem of translating actual content of simple sentences from 8 up to 30 words, and adjust classifiers for selecting term matches in certain subject areas [3].

Today we have free online solutions from Microsoft, Google, Yandex and Baidu that provide

users with medium-quality translation, that in most cases can be post-edited by an attracted expert. Current situation creates high competitiveness on the market of machine translation, where the main efficiency indicator of machine translation systems adds up to the solution of the problem of semantic and structural matches in different language systems.

2 LANGUAGE STRUCTURES AND APPROACHES FOR THEIR FORMALIZATION

In modern machine translation systems distorted meanings are placed, as a rule, beyond the bounds of the 8th word, when the quality of translation deteriorates due to search errors [3][4][5]. Firstly, this trend can be explained by different syntactic structures of languages – for instance, rigid linear order in English language and frame structure of tenses in German language, - and, secondly, by an increasing word distribution distance. In 2012, the corporation Google made attempts to tackle the problem of language structure matching by integrating a method of cross-lingual word clusters, that performs a direct transfer of kindred language structures; according to the developers, this method helped slightly increase the quality of translation by 26% [6] (see also the Russian system «Crosslator 2.0», that uses a mediator language in order to achieve multilingual translation of terms).

The existing phrase methods of statistical text processing «see» only the local distribution environment and do not consider the lexical-grammatical context of the whole phrase. That is why, scientists continue work on the models and methods that will expand the diapason of hypothetic prediction of machine translation system without the distortion of meanings: for example, an enhanced method for the prediction and estimate of future actions of machine translation system based on algorithm modifications applied in Pharaoh [7], the method of dynamic selection by Arianna Bisazza and Federico Marcello, when the reordering of data is not restricted by decoding by the closest lexical environment (like in the Moses system) but extends to a long distance and, hence, expands the process of hypothetic decision making by translation up to the 15th word [8] (refer also to [9], [10], [11]). One of new approaches for solving the problem of ambiguity by translation is the architecture Transformer, announced by the working group of Google on August 31st, 2017. Underlying this architecture is a mechanism of self-attention, that uses deep neural learning mechanism, lines up and analyzes the interconnection points of all word representations in the context of a sentence. According to the BLEU scores (see [12] and [13]), such decoding algorithm improves the quality of translation by overcoming the difficulty of choosing the single accurate meaning representation of an ambiguous word for a certain semantic context with help of weighted values.

In the paper [14] the problem of machine translation was examined from the standpoint of modelling a speaker's linguistic competence, and suggested a complex functional method for business text translation based on the analysis of semantic features and basic frames. By the approbation of the suggested method for the machine translation from Russian into English the level of semantic match increased in comparison with the translations of the existing machine translation systems. However, this method requires the formalization of language structures (the description of matches among the structures of different languages). The drawback of this methods consists in the necessity to build a separate data base with language structures and their matches for different language pairs. Labor-intensive character of this method can be eliminated by the automation of language structure formalization. To achieve that, we need to classify the types of written forms of meaning representations and formalize a way to present them taking into account the skeleton-semantic and context-dependent components.

Underlying the analysis of language structures is the meaning-text theory [14], that is structurally implemented on the example of the analysis and classification of business communication speech acts. Reference- and situation-based method includes the analysis of surface and in-depth semantic representations (read more about case grammar by Ch. Fillmore in [15]), and namely: the analysis of 1) skeleton-language structures that actualize this or that speech act of business communication 2) valence «core» components, that guide the vector of thought unfolding, and 3) the components of semantic variability to have a possibility to generate *n*-phrases, not depending on the subject content of a phrase. From the standpoint of surface semantics, the analysis of speech acts helps obtain a limited set of surface language structures utilized by speakers in the sphere of business communication in order to describe a certain scenario (for example, invitation, appeal, recommendation, etc.) as well as design the

variability of a semantic context-dependent description of this scenario in form of situations and, finally, describe the possibilities of stylistic tint for the actualized situations.

To test the hypothesis about the possibility to model semantic contexts and formalize language structures we carried out an experiment on the basis of over 500 speech messages in business communication (personal correspondence, analysis of scripts of English-speaking movies and series and their Russian-speaking translated versions, tutorials); the experiment results helped single out 3 big groups of speech acts, i.e. speech acts with pragmatic description, speech acts with emotional (perlocutionary) description, and speech acts with structural-logical description. Each group was further classified from the standpoint of the actualized scenarios. The analysis included the method of structural analysis, skeleton-semantic and context-dependent components of business communication messages.

The structural analysis revealed a group of most frequent scenarios used to implement speech acts with pragmatic description, and namely: appeal, prohibition, request, recommendation, suggestion (offer), invitation, reminding, conviction (persuasion), gratitude, congratulation, hope, wish, apologies).

At the same time, the speech act of a suggestion subdivides into the following types of situations: suggesting an alternative solution, suggesting a way to solve a problem/ problematic situation, suggestion with identifying a desired outcome; and the speech act of congratulations implements the following situations: congratulations with holidays/seasons, congratulations on receiving an award.

The analysis of skeleton-semantic and context-dependent (variable) content of speech acts with pragmatic description helps formalize lexical-grammatical distributive environment of phrases and determine their components.

Let us consider an example of the Russian-English match for the English phrase «**Please go to the following link in order to** print your card» and the Russian phrase «**Для** печати вашей карточки, **пройдите, пожалуйста, по следующей ссылке».** The structural analysis of language structures used to implement the meaning of an appeal and a speech act of an appeal consequently, is presented in the Table 1.

An important observation is the difference in the in-depth syntactic positions in the phrase segmentation in the source and target languages, see Table 2. In this case it is necessary to pay attention to a «free» theme - rheme based character of thought unfolding in certain languages (for example, Russian).

Table 1: Structural analysis of language structures.

Components of structural analysis	**English model**	**Russian model**
Skeleton language structure	Please do something in order to do something.	Для чего-то сделайте, пожалуйста, что-то
Valence core component	go to the following link	пройдите по следующей ссылке
Semantic variability component	print your card	печати вашей карточки

Table 2: The matching analysis of in-depth syntactic positions in a phrase segmentation.

Model of English sentence			
1	2		3
Please	go to the following link		in order to print your card.
Model of Russian sentence			
1=3	2=2	3=1	4=2
Для печати вашей карточки	пройдите	пожалуйста	по следующей ссылке.

In such cases, independent of the order of the train of thoughts in a Russian language the process of sentence decoding should analyze the in-depth semantics of word representations and identify a match for in-depth syntactic positions taking into account the grammar system of the target language (for example, English).

By the consideration of certain stylistic genres of communication, it is possible to narrow a number of linguistic types and structures that actualize

emotional tints. Due to a rigid and laconic character of lexical-grammatical standards of language structures that have historically put together for the stylistic code of business communication, the category of emotionality is poorly represented in this stylistic code and is conventionally represented by 4 basic speech acts, i.e. joy, sadness, bewilderment (feeling confused), respect.

The analysis of speech acts from the standpoint of structural-logical information resulted in 7 basic speech acts: 1) initiating a message, 2) action's plan, 3) attached information, 4) appearance of questions, 5) comments, 6) finalizing, 7) finishing a message. The speech act that describes the beginning of a message characterizes the following sub-situations: writing for the first time, resuming a communication after a break in correspondence, a recurrent reply (ongoing correspondence), a recurrent reply with giving the prehistory of previous replies, a recurrent reply with identifying the goal of an email.

The speech act that realizes action's plan divides into the following sub-situations: 1) a series/ consequence of actions, 2) the current state of business, 3) immediate actions. The speech act that indicates the presence of attached information/materials in an email can actualize 1) a message about the attached information, 2) a message about additional email recipients, 3) a message about forthcoming (consecutive) information, 4) a message that states where additional information can be found. The finalizing speech act can actualize 1) positive attitude to the forthcoming situation, 2) negative attitude to the forthcoming situation, 3) final notes. The speech act that describes the situations with commenting actualizes the following situations: 1) confirmation, 2) clarifying, 3) consent (agreement)/ non-agreement, 4) an intention to understand the reason of a current state, 5) causes and effects, 6) difficulties. In the speech act of questions reasons are described, i.e. if 1) the solution fits, 2) the situation is understood correctly; 3) what is the reason of the current situation; 4) how to solve this or that issue.

Multivariable realizations of speech acts with logical-structural description indicate a relevant character of this subgroup of speech acts for business communication.

Therefore, the presence of a direct match in surface semantics and in-depth syntactic positions in the segmentation of a phrase is evidence of the possibility to formalize the matches of language structure without the introduction of additional conditions.

The determination of skeleton-semantic and context-dependent variable semantic content of business communication messages with stylistic tint description (polite, neutral and informal (rough)) helps classify language expressions and work out a language for describing matches among the structures of the source language and the target language based on the functional method for automated text translation [14].

3 THE FORMALIZATION AND USE OF LANGUAGE STRUCTURES

The following symbols are used for the description of language structures:
() – sentence type definition;
= - clarifying the concept;
{} – merge;
<> - mandatory part of a sentence;
[] – optional part of a sentence;
| - or;
"" – rigidly given context;
\ - clarifying a new variable.

Text description consists then of two stages: sentence type definition and the description of sentence structure.

To describe the types of sentences we use the following classifiers and labels:

1. Classifier of sentence purpose: declarative (1A), interrogatory (1B), imperative (1C).

2. Classifier of mood: indicative (2A), conditional (2B), imperative (2C).

3. Classifier of voice: active (3A), passive (3B).

4 Classifier of tenses: past simple tense (4A), present simple tense (4B), present in process (4C), past in process (4D), present with result (4E), past with result (4F), present with result and process (4G), past with result and process (4I), future simple (4J), future in the past (4H), future in the past with result (4K), near future (4L), near future in the past (4M), immediate future (4N), immediate future in the past (4P).

5. Classifier of the connection between the object of speech and the fact that is reported about the object: assertion (5A), negation (5B).

6. Classifier of predicate type: simple predicate (expressed by notional verb or auxiliary verb) (6A), predicate with modal tint (modal verbs) (6B), predicate with aspect tint (aspect verbs) (6C), predicate with modal and aspect tint (modal verbs + aspect verbs) (6D), predicate with a notional estimate (verbs with meanings of supposition, suggestion, desire, recommendation, advice) (6E),

predicate with modal and notional estimate (modal verbs + a verb with the meaning of supposition, suggestion, desire, recommendation, advice) (6F).

7. Additional types of classifiers can be introduced depending on the language system (for instance, a classifier of words that can change the reordering of a phrase can be set for English language).

The description of sentence structures requires the introduction of such basic variables like: noun phrase (subject) (NP1), object (complement) (NP2), auxiliary verb of a notional verb (always conjugated) (Aux), Gerund (Gerund), infinitive of a notional verb (V), object (complement) (NP3), adjective (Adj), determinative (Det), definite article (Art/def), indefinite article (Art/indef), etc.

These variables can be used to form a new variable or adjust their meanings (see Listing 1), for the specification of sentence types see Listing 2.

Listing 1: An example of describing new variables via basic variables.

```
\AdvP1 = [AdvP1.2][AdvP1.1x]
\NP2v={<Det><N>}|{<Prep><Det><N>}|{<Pro
n/pers>}|{<Adv><Adj>}
\NP2v1g={<Adj>}|{<Det><N>}|{<Prep><Det>
<N>}|{<Adv><Adj>}|{<Pron/pos>}
\AdvP2 =
{"slowly"}|{"quickly"}|{"rapidly"}
\AdvP3 =
{"here"}|{"there"}|{"outside"}|{"inside
"}|{"somewhere"}|{"everywhere"}
```

Listing 2: An example of describing general cross-functional sentence structures.

```
(1A,2A,3A,4B,5A,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1][AdvP2][AdvP3]<".">
(1A,2A,3A,4B,5B,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1][AdvP2][AdvP3]<".">
(1B,2A,3A,4B,5B,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1][AdvP2][AdvP3]<"?">
(1A,2A,3A,4C,5A,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<".">
(1A,2A,3A,4C,5B,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<".">
(1B,2A,3A,4C,5A,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<"?">
(1B,2A,3A,4C,5B,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<"?">
(1A,2A,3A,4E,5A,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<".">
(1A,2A,3A,4E,5B,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<".">
(1B,2A,3A,4E,5A,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<"?">
(1B,2A,3A,4E,5B,6A)=<NP1><VP><NP2>[NP2e
xt][NP3][AdvP1]<"?">
```

The introduction of formal description helps utilize these structures to find matches among languages (use the formalization language as an intermediate language), and develop constructors for building phrases that due to the restrictions, artificially installed by developers into phrase structures, will serve as an input filter and instrument to prepare phrases for translation (see Fig. 1).

The suggested description helps structure any types of sentences. For instance, for English language see the examples in the Listing 3, and the corresponding examples in Russian language are presented in the Listing 4.

Listing 3: An example of structuring sentences for English language.

```
(1A,2A,3A,4B,5A,6A,7A)                  =
{[NP3]<'',''><NP1>[AdvP1.2]<VPvf1><NP2>
[NP2ext][AdvP2] <".">}
(1B,2A,3A,4E,5A,6A,7C)                  =
<VPAuxE><NP1>[AdvP1.3/4]<VP3><NP2>[Np2e
xt][AdvP1.3/2*]<''?''>
```

Listing 4: An example of structuring sentences for Russian language.

```
(1A,2A,3A,4B,5A,6A,7A)                  =
{<NP1>[NP3][AdvP1.2][AdvP2]<VPvf1><NP2>
[NP2ext] <".">}|
{[NP3]<NP1>[AdvP1.2]<VPvf1><NP2>[NP2ext
][AdvP2] <".">}|
{[NP3]<NP1>[AdvP1.2][AdvP2]<VPvf1>|<VPv
f1.2><NP2>[NP2ext] <".">}|
{<NP1>[AdvP1.2][AdvP2]<VPvf1>[NP3]<NP2>
[NP2ext] <".">
(1B,2A,3A,4E,5A,6A,7C)      =      {<NP1>
AdvP1.3/2*][AdvP1.3/4]<VP3><NP2>
[Np2ext]<''?''>}|{[AdvP1.3/2*]<NP1>[Adv
P1.3/4]<VP3><NP2>[Np2ext]<''?''>}
```

The utilization of structures in the source language helps increase the quality of translation by excluding idiomatic expressions, expletive words from the texts, etc. The use of matches to reorder a phrase before using machine translators in many cases helps eliminate errors and distorted meanings (see Table 3) taking advantage of already existing solutions.

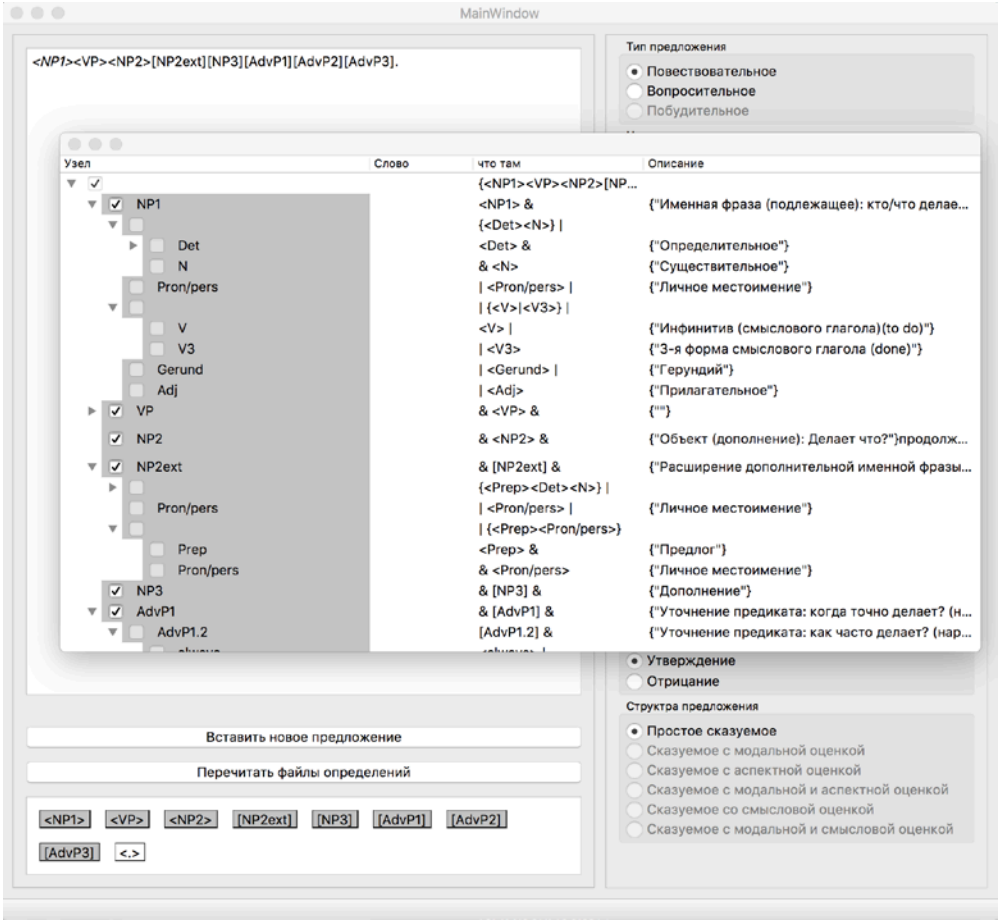

Figure 1: Interface example of phrase construction system.

Table 3: The examples of translation from Russian into English with help of the system Google translate without modifying a sentence structure in an initial sentence and with the modification of sentence structures.

	Sentence structure	**Translation**	**Amount of errors**
Example 1 – «Они в компании всегда быстро проводят обновление программного обеспечения»			
Initial sentence	<NP1>[NP3][AdvP1.2][AdvP2] <VPvf1> <NP2>[NP2ext] <".">	They always update the software in the company.	3
Modified sentence	[NP3]<’’,’’><NP1>[AdvP1.2] <VPvf1><NP2>[NP2ext][AdvP2] <".">	In the company, they always carry out software updates quickly.	0
Example 2 – «Раньше ваша компания когда-либо обновляла программное обеспечение для переводчика?»			
Initial sentence	[AdvP1.3/2*]<NP1>[AdvP1.3/4] <VP3> <NP2>[Np2ext]<’’?’’>	Did your company ever update the software for an interpreter?	3
Modified sentence	<NP1>[AdvP1.3/4]<VP3><NP2> [Np2ext][AdvP1.3/2*]<’’?’’>	Has your company ever updated the software for an interpreter before?	0

Besides, unlike the existing solutions the suggested approach is neither aimed at correcting the grammar which leads to distorted meanings by grammatically correct structures nor requires post-editing of received results.

4 CONCLUSIONS

The use of the described approach helps exclude intermediary natural languages by machine translation (it is known, that by the translation to/from Hebrew German is used as an intermediary language) having replaced them by an artificial language that helps 1. add restrictions into the lexical diversity (if necessary) 2. perform pre-editing of texts for translation having replaced by this operation the attempts to correct errors of automated translation systems (this approach is more attractive as it is easier for the user to make required adjustments in a native source language rather than changing the target language) 3. carry out automated proofreading of texts for their preparation quality. The main method of this approach is based on structural analysis and formalization of texts. At the same time, this approach is not sensitive to the subject area and the prehistory of use.

Further developments can be connected with such directions as language learning, proofreading of the source language, automated translation, text constructor for building phrases in a foreign language, that are implemented in accordance with the scheme presented in the Fig. 2.

Despite all the evident advantages, there are also difficulties connected with the necessity to describe each natural language separately with help of the developed language, which requires high expertise.

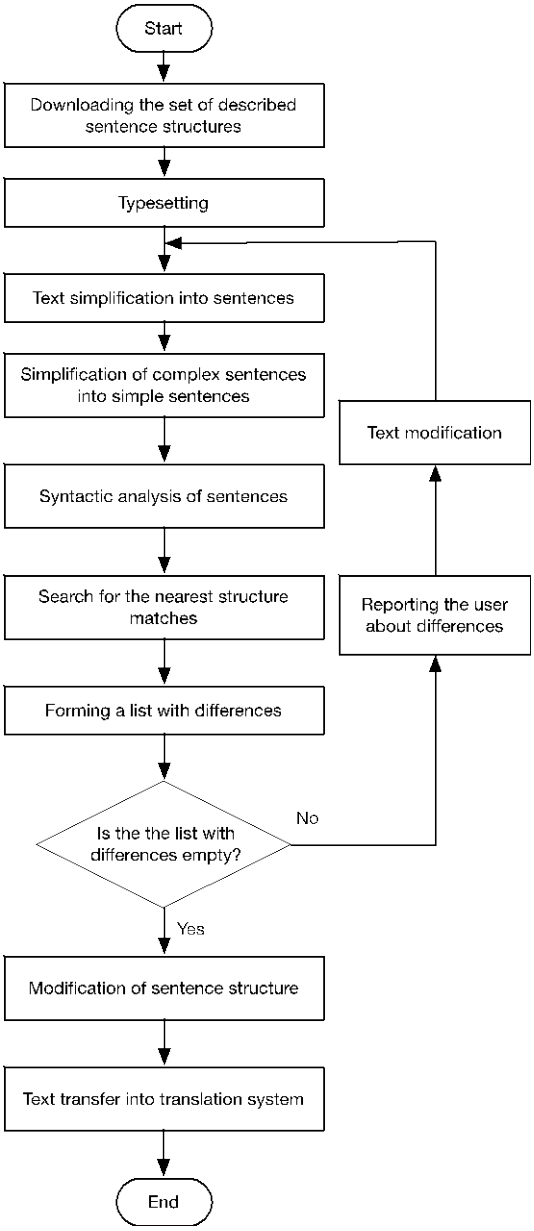

Figure 2: The algorithm of using the description language for language structures.

REFERENCES

[1] B. Henisz-Dostert, R. R. Macdonald, and M. Zarechnak, Machine translation. The Hague ; New York: Mouton, 1979.

[2] P. Sojka, Ed., Text, speech and dialogue: 13th international conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010: proceedings. Berlin; New York: Springer, 2010.

[3] M. Aiken, K. Ghosh, J. Wee, and M. Vanjani, “An Evaluation of the Accuracy of Online Translation Systems,” Commun. IIMA, vol. 09, no. 04, 2009.

[4] A. Kazemi, A. Toral, A. Way, A. Monadjemi, and M. Nematbakhsh, “Syntax- and semantic-based reordering in hierarchical phrase-based statistical machine translation,” Expert Syst. Appl., vol. 84, pp. 186–199, Oct. 2017.

[5] J. B. Mariño et al., “N -gram-based Machine Translation,” Comput. Linguist., vol. 32, no. 4, pp. 527–549, Dec. 2006.

[6] O. Taeckstroem, R. McDonald, and J. Uszkoreit, Proceedings of the 2012 Conference of the North American Chapter Of. Saintoudsburg: Association for Computational Linguistics, 2012.

[7] R. Moore and C. Quirk, “Faster Beam Search Decoding for Phrasal Statistical Machine Translation.,” Proc. 11th Mach. Transl. Summit Cph., pp. 321–327, 2007.

[8] A. Bisazza and F. Marcello, “Efficient Solutions for Word Reordering in German-English Phrase-Based Statistical Machine Translation.,” Assoc. Comput. Linguist., pp. 440–451, 2013.

[9] N. Durrani, H. Schmid, A. Fraser, P. Koehn, and H. Schütze, “The Operation Sequence Model—Combining N-Gram-Based and Phrase-Based Statistical Machine Translation,” Comput. Linguist., vol. 41, no. 2, pp. 185–214, Jun. 2015.

[10] E. Hasler, A. de Gispert, F. Stahlberg, A. Waite, and B. Byrne, “Source sentence simplification for statistical machine translation,” Comput. Speech Lang., vol. 45, pp. 221–235, Sep. 2017.

[11] F. J. Och and H. Ney, “Discriminative training and maximum entropy models for statistical ma- chine translation.,” Proc. 40th Annu. Meet. Assoc. Comput. Lin- Guistics ACL, pp. 295–302, 2002.

[12] Association for Computational Linguistics, P. Isabelle, and Association for Computational Linguistics, Eds., Proceedings of the conference, 40th annual meeting of the Association for Computational Linguistics: Philadelphia, [6 - 13] July 2002, University of Pennsylvania, Philadelphia, Pennsylvania. Hauptbd. ... San Francisco: Morgan Kaufmann, 2002.

[13] P. N. Astya et al., Proceeding, International Conference on Computing, Communication and Automation (ICCCA 2016): 29-30 April, 2016. 2016.

[14] A. V. Novikova and L. A. Mylnikov, “Problems of machine translation of business texts from Russian into English,” Autom. Doc. Math. Linguist., vol. 51, no. 3, pp. 159–169, Jun. 2017.

[15] S. Mazarweh, Fillmore Case Grammar Introduction to the Theory. München: GRIN Verlag GmbH, 2010.