Linking Russian Words to Semantic Frames of FrameNet*

Dana Zlochevskaya
Computational Mathematics Faculty
Lomonosov Moscow State University
Moscow, Russia
dana_zl@mail.ru

Natalia Loukachevitch
Research Computing Center
Lomonosov Moscow State University
Moscow, Russia
louk_nat@mail.ru

Olga Nevzorova
ICMIT Institute
Kazan Federal University
Kazan, Russia
onevzoro@gmail.com

***Abstract*—This work is devoted to the transfer of the FrameNet semantic frames created for English into Russian. The transfer is based on semantic frames identified in English sentences and uses a parallel Russian-English corpus. The resulting set of semantic frames in Russian is utilized to train and evaluate the models for identifying semantic frames in Russian. The results of the work include: 1) a lexicon of Russian words linked to the FrameNet frames, 2) an annotated Russian dataset - sentences with labeled semantic frames, 3) the implementation, testing and analysis of models for identifying semantic frames in Russian.LaTeX.**

***Keywords*—semantic analysis, frames, language transfer, parallel corpus, neural networks**

I. Introduction

Semantic text analysis remains one of the most difficult tasks of natural language processing. Such an analysis requires a theory and a model of semantics representation. One of the most known semantic models is Charles Fillmore's Frame Semantics [1], which was explicated in the FrameNet lexicon of frames [2].

The idea of the Frame Semantics approach is as follows [1]: the senses of words can be represented using situations, their relations and roles. For example, the situation of "cooking" in most cases can be described with the help of the following participants and relations between them as: "the subject who cooks" (the cook), "the object that is cooked" (food), "the source of heat for cooking" (heat source) and "cooking container" (utensils). In this case, it is possible to say that the cooking frame has been introduced, in which "cook", "food", "heat source" and "containers" are elements of the frame. Frame elements, which are "markers" of the frame, that is, they often signal the location of this frame in the text (for example, "fry", "Cook", "stew", etc.) are called lexical units. Frames can be of different complexity, can have different numbers of elements and lexical units. The main task in constructing semantic frames is to show how the elements of the frame are related to each other.

The project is supported by the Russian Science Foundation, grant # 19-71-10056.

FrameNet is a basis of numerous works on the first step of semantic analysis called semantic role labeling [3]–[6]. However, such data are absent for most languages, and development of FrameNet-like resources from scratch is a very laborous process, requiring a lot of resources and time [7], [8]. Therefore various approaches for automatizing of framenet creation for other languages are discussed [9]–[11].

In this paper we consider the transfer of the FrameNet semantic frames into the Russian language. The method is based on the transfer of semantic frames identified in English sentences using a parallel Russian-English corpus. The resulting set of semantic frames in Russian is used to train and evaluate the models for identifying semantic frames in Russian. The results of this work include: 1) a lexicon of Russian words linked to the FrameNet frames, 2) an annotated Russian dataset - sentences with labeled semantic frames, 3) the implementation, testing and analysis of models for the identification of semantic frames in Russian.

II. Related Work

There is an interest of researchers in many countries to have a semantic resource similar to FrameNet for their own languages. However, to create a FrameNet-like resource from scratch is a very difficult, expensive, and time-consuming procedure [7], [8]. Therefore automated methods for generating a framenet for a specific language are used. Such methods can be subdivided into two groups: cross-lingual transfer, and generating frames in unsupervised manner from a large text collection [9], [12], [13]. Cross-lingual transfer can be made via linking FrameNet with WordNet [14], [15] or via parallel corpora [16].

Cross-lingual transfer of frames based on a parallel corpus requires a preliminary extraction of frames in English texts, which usually includes the following main steps [17], [18]:

- recognition of lexical units, which can express frames, in a text. This stage can be reduced to the task of binary classification for each word of the text;

- recognition of semantic frames. For each lexical unit, it is necessary to determine what frames are expressed by this unit. This task can be considered as a problem of multiclass classification (with several labels) for each of the selected lexical units at the previous stage. The classes in this case are the semantic frames themselves, and several frames can correspond to one lexical unit. The main difficulties at this stage are the possible ambiguity of a lexical unit and/or its absence in the training set.
- recognition of the remaining elements of the frame such as roles. Most often, this problem is solved using named entity extraction methods with the condition that the set of roles strongly depends on the frame being processed. In the current work, this stage is not studied.

The solutions for these three tasks can be very different. In the work of the winners of the SemEval 2007 competition [18], to solve the problem of identifying lexical units, morphological and syntactic rules were used. To define semantic frames, classifiers were trained for each frame separately using support vector machine (SVM) method, the features for which were both morphological and syntactic characteristics of a lexical unit, and semantic information about it from WordNet [19].

The widely used SEMAFOR algorithm [20] has become an improvement of this algorithm. To recognize lexical units, an improved version of linguistic rules is used. To identify a frame, a probabilistic model is trained on a similar feature set. In [21], an approach based on recurrent neural networks is proposed for both tasks: recognizing lexical units and recognizing semantic frames. In recent years, works have also appeared that solve this problem using transformers, for example, BERT [22], [23].

III. FrameNet

FrameNet [2] consists of several components:

- The base of semantic frames, containing a description of each frame: its structure, roles and relations between frames;
- Examples of sentences annotated with semantic, morphological and syntactic information. These sentences are examples of the use of semantic frames in natural language;
- Lexical units linked to frames as well as links to examples of annotated sentences.

At the moment, the FrameNet database contains more than 13000 lexical units, of which 7000 are fully annotated with more than 1000 hierarchically related semantic frames. The number of sentences annotated with semantic information is about 200,000. The base exists in several versions and is in the public domain for research purposes. Basic concepts from the terminology of the FrameNet project are as follows:

- Semantic frame is a semantic representation of an event, situation or relationship, consisting of several elements, each of which has its own semantic role in this frame.
- Frame element is a type of participants (roles) in a given frame with certain types of semantic links.
- Lexical unit - a word with a fixed meaning that expressed a given frame or a given frame element.

For example, In the sentence “Hoover Dam played a major role in preventing Las Vegas from drying up." The word "played" is a lexical unit that conveys the presence in the text of the semantic frame "PERFORMERS_AND_ROLES" with the frame elements "PERFORMER", "ROLE" and "PERFORMANCE" (who created, that created, what role the creator assumed in the creation). This example also shows that in one sentence there can be several frames with varying degrees of abstractness.

It is important to note that for the Russian language there is a FrameNet-oriented resource FrameBank [24]. This is a publicly available dataset that combines a lexicon of lexical constructions of the Russian language and a marked-up corpus of their implementations in the texts of the national corpus of the Russian language. The main part of FrameBank consists of 2200 frequent Russian verbs, for which the semantic constructions in which they are used are described, and examples of their implementations in the text are collected. Each construction is presented as a template, in which the morphological characteristics of the participants in their role and semantic restrictions are fixed.

Despite the fact that the semantic constructs of FrameBank are similar to frames from FrameNet, they are methodologically different. FrameNet frames are built around generic events with specific participants and relations between them. FrameBank constructs are built around the senses of specific words. It is supposed that the senses of each lexeme (mainly verbs) in FrameBank form a separate frame. Due to this methodological difference and the orientation of the FrameBank resource towards verbs as a center of constructions, the full use of this data set to solve the problem is impossible.

IV. Methods of Linking Russian Words to FrameNet

The method of linking Russian words to FrameNet is divided into several parts, each of which solves a specific subtask:

1) Training the model of recognizing lexical units and frames in English. At this stage, based on the FrameNet knowledge base, a model is created that can extract lexical units and semantic frames from any sentence in English. The quality of the models is evaluated on the test part of the FrameNet dataset and compared with the existing results.

2) Extracting lexical units and frames from the English part of the parallel corpus using a trained model.
3) Transferring the obtained semantic information into Russian using word matching through a pre-trained embedding model of the Russian language. At this stage, for each lexical unit from an English sentence, its analogue is searched for in a parallel Russian sentence. Transfer quality is evaluated on a test sample, annotated manually.
4) Extracting lexical units for each semantic frame. A set of annotated sentences in Russian is also formed for further recognition of semantic frames in Russian texts.

A. Training models for identification of lexical items and frames in English

As in most studies, two sequential models are trained. The first model predicts potential lexical units that can express a semantic frame, and the second one, based on the predictions of the first model, determines which frames should be generated from the selected lexical units. Both models are trained and tested on the manually annotated FrameNet corpus of sentences.

For training and testing, the annotated corpus FrameNet version 1.5 was used. The entire corpus of annotated sentences contains 158399 example sentences with labeled lexical units and frames. They are presented in 77 documents, 55 texts of which were randomly selected for model training and 23 texts for testing. A modified CONLL09 format is used as a universal format for presenting annotations, in which the following set of tags is attached to each word of the sentence:

- ID - word number in the sentence,
- FORM - the form of a word in a sentence,
- LEMMA - word lemma,
- POS - part of speech for the given word,
- FEAT - list of morphological features,
- SENTID - sentence number,
- LU - lexical unit, if the word is it,
- FRAME - a semantic frame generated by a word if it is a lexical unit.

The task of predicting lexical units in a text is reduced to the task of binary classification for each word in a sentence. In some cases, a lexical unit may not be one word, but a phrase, but in this work, the most common variant with one word as a lexical unit is investigated. To solve this problem, a bidirectional recurrent neural network (BiLSTM) was used. Such models were actively used in previous studies for identification of lexical units and frames [21].

A sentence is sent to the input of the neural network, each word in which is represented by a vector representation of a word from a pretrained embedding model. In addition to this classical representation of words, new features responsible for semantic and morphological information were studied such as: part of speech and the initial form of a word. These features are represented as a one-hot vectors, which is fed to the input of fully connected layers of the neural network. During training, the outputs of these fully connected layers can be considered as vector representations of these features. These features can help the model to memorize morphological and semantic schemes for constructing frames from the FrameNet annotation. All obtained feature vectors for a word are joined by concatenation into the final vector representation, which is fed to the input of the neural network. In the learning process, only vectors of tokens from the pretrained embedding model are fixed, the rest of the vector representations are formed during training.

A sentence is sent to the input of the neural network, each word of which is represented as a vector according to the algorithm described above. The network consists of several BiLSTM layers, to the output of which a fully connected layer with the sigmoid activation function is applied for each word. As a result, at the output of the network, each word of the sentence is matched with the probability of a given word to be a lexical unit in a given sentence.

To optimize the parameters of the neural network, the logistic loss function was used. The adaptive stochastic gradient descent Adam [25] was chosen as an optimizer. To prevent overfitting of the neural network, the Dropout technique [26] was used, based on random switching off of neurons from the layers of the neural network for greater generalization of the trained models.

B. Semantic frame identification

The purpose of the semantic frame identification model is to recognize frames that correspond to lexical items in a sentence. Formally, this is a multi-class classification problem with multiple labels, since the same lexical unit can correspond to several semantic frames in a sentence.

To solve this problem, a model was used that is similar to the model of the selection of lexical units, but with several changes. The main difference is adding information about whether a word is a lexical unit using one-hot coding. If the word is not a lexical unit in a given sentence, the lexical unit representation vector will consist entirely of zeros and will not affect the prediction. It is important to note that the predictions of the model are taken into account only for words that are lexical units in a given sentence. The remaining components of the vector representation of a word are similar to the model for extracting lexical units - a vector of tokens from a pretrained embedding model and one-hot coding that encodes a part of speech.

C. Implementation and results

The following hyperparameters were chosen for training the models:

Model for lexical units	P	R	F1
$SEMAFOR$	74.92	66.79	70.62
$BILSTM$	74.13	66.11	69.89
$BILSTM_{token}$	76.01	67.15	71.3
$BILSTM_{token,lemma}$	76.12	67.98	71.8
$BILSTM_{token,lemma,partofspeech}$	79.47	68.31	73.46

Table I
Results for lexical units recognition in English

Model	P	R	F1
SVM	79.54	73.43	76.36
$SEMAFOR$	86.29	84.67	85.47
$BILSTM$	83.78	79.39	81.52
$BILSTM_{lu}$	88.19	81.54	84.73
$BILSTM_{lu,token}$	88.83	88.12	85.34
$BILSTM_{lu,token,partofspeech}$	89.87	83.91	86.78

Table II
Results of model of semantic frame recognition for English

- The Glove model trained on the English-language Wikipedia[1] was chosen as an embedding model. The vector dimension is 100;
- The size of the vectors encoding the lemma, lexical unit and part of speech is 100, 100, and 20, respectively;
- Number of BiLSTM layers is 3, output dimension is 100;
- The number of training epochs is 40;
- The Dropout coefficient is 0.01.

The results of lexical unit identification obtained on the test sample are presented in Table 1. For comparison, the SEMAFOR algorithm was applied, which determines the lexical units on the basis of linguistic rules. For this, an available author's implementation [2] was used, trained on the same data as the tested models. It can be seen from the results that adding information about the word lemma does not give a significant improvement, however, information about the part of speech allows obtaining quality that is superior to the classical SEMAFOR model.

To evaluate the results of the semantic frame model, the SEMAFOR algorithm was also used, which identifies a frame based on the probabilistic model. In addition, for comparison, the model of the SemEval 2007 [16] was recreated. The obtained results are presented in Table 2. It can be seen that the trained model has a quality comparable to the performance of the SEMAFOR model.

In this way, models for identification of lexical units and semantic frames in English were trained. At this stage, for any sentence in English, a set of frames and corresponding lexical units are identified. To transfer this information to the Russian part of the parallel corpus, it is necessary for each lexical unit from the English sentence to find its analogue in Russian. Since in this study only single-word lexical units are considered, the task is reduced to the comparison of words between sentences in a parallel corpus.

[1]https://nlp.stanford.edu/projects/glove/

[2]https://github.com/Noahs-ARK/semafor

D. Translation of annotations from English into Russian in a parallel corpus

For the study, we used the English-Russian parallel corpus [3], gathered by Yandex. It consists of 1 million pairs of sentences in Russian and English, aligned by lines. The sentences were selected at random from parallel text data collected in 2011-2013.

Like most parallel corpora, this resource is sentence-aligned, not word-aligned, so word-level matching requires further refinement. A simple dictionary translation of an English word into Russian does not provide the desired effect due to the ambiguity of words and differences in structure of languages. Therefore, in addition to direct translation of words, a matching algorithm was implemented based on the similariy of words in an embedding model of the Russian language.

The FastText model in the Skipgram version [27] trained on the Russian National Corpus [4] was chosen for word matching. Thus, the algorithm of matching between languages consists of the following steps:

1) Translation of a lexical unit in English, for which we are looking for an analogue in a parallel sentence in Russian. In this step, all possible translations into Russian are collected using the Google Translate API [5].
2) Further, between each obtained translation and each word of a parallel Russian sentence, the cosine similarity according to the embedding model is calculated.
3) A word in the Russian sentence is considered as an analogue of the original word in English if between the translation of an initial word and the Russian word, the cosine similarity is higher than 0.9.

To evaluate the quality of word matching, the transfer of lexical units in 100 parallel sentences was manually assessed. Both precision (in how many sentences the word was translated correctly) and recall (in how many sentences an analogue of the word in English was found in general) were considered. A simple search for a translation of a word in a parallel sentence was taken as the basic algorithm; the comparison results can be seen in Table 3. It can be seen that the use of the embedding model increases the recall of word matching with a slight decrease in precision.

Word translation search method	P	R
Direct translation matching	91 %	72 %
Distributive word matching	90 %	87 %

Table III
Results of matching words between sentences in a parallel corpus

Thus, applying this algorithm to each pair of sentences in a parallel corpus, it is possible to transfer the selected

[3]https://translate.yandex.ru/corpus

[4]https://rusvectores.org/ru/models/

[5]https://cloud.google.com/translate/docs/

lexical units and the corresponding semantic frames from English into Russian.

E. Characteristics and evaluation of the resulting corpus

The sentences in Russian obtained after the transfer with annotated lexical units and semantic frames form a dataset similar to a of the FrameNet knowledge base.

In it, for each of the 755 frames, lexical units with frequency of use in the context of the frame are presented. This frequency can be interpreted as a certain "reliability" of the lexical unit belonging to the frame. A total of 2.8 million lexical units were analysed out of 1 million sentences. They belong to 755 semantic frames from the FrameNet project. In total, 18150 lexical units have been identified, including 6894 unique words.

Table 5 shows an example of the obtained semantic frames and lexical units assigned to it. Each column refers to a frame, the first line contains its name from the original FrameNet knowledge base, and the second contains lexical units assigned to it in Russian with the frequency of use.

Fear	Labeling	Reason
страх : 1042	термин : 1045	причина : 3287
бояться : 552	понятие : 139	основа : 755
боязнь : 42	этикетка : 119	основание : 533
опасаться : 40	ярлык : 78	мотивация : 247
ужас : 28	терминология : 24	повод : 118
страшиться : 3	марка : 11	мотив : 110
страшно : 3	бренд : 9	поэтому : 2
испуг : 1	клеймо : 2	именно : 1

Table IV
Examples of linking Russian lexical units to the FrameNet frames

The resulting dataset can be used as a separate semantic resource for various natural language processing tasks. In addition, annotated Russian sentences are also valuable. They make it possible to conduct experiments on the selection of lexical units and semantic frames using supervised machine learning methods.

V. Training and testing model for identification of lexical units and frames in Russian

The resulting set of annotated sentences in Russian was used to train models for the selection of lexical units and semantic frames in Russian, similar to the already trained models in English. Out of 970 thousand sentences, 90% were used for training models, the rest were used for testing.

The architecture and method of constructing the vector representation of words are similar to the models for the English language, with the exception of the embedding model - instead of Glove, the FastText model of the Skipgram architecture was used, trained on the National corpus of the Russian language [6]. The vector dimension is 300. The results of the obtained models are presented in Tables 5 and 6.

[6]https://rusvectores.org/ru/models/

Lexical Identification Model	P	R	F1
$BILSTM$	71.67	59.14	64.80
$BILSTM_{token}$	76.09	61.04	67.73
$BILSTM_{token,lemma}$	77.65	61.33	68.53
$BILSTM_{token,lemma,partofspeech}$	78.44	61.72	69.08

Table V
Results of the Russian lexical unit identification model

Semantic Frame Identification Model	P	R	F1
SVM	80.28	72.43	76.15
$BILSTM$	84.10	76.99	80.38
$BILSTM_{lu}$	86.54	78.04	82.07
$BILSTM_{lu,token}$	87.01	78.20	82.40
$BILSTM_{lu,token,partofspeech}$	90.83	83.91	84.66

Table VI
Results of the model for identifying semantic frames for the Russian language

The obtained results for Russian are lower than the results of similar models in English. This can be explained by the fact that during the automatic transfer, there is a loss of data and the introduction of noise at each stage - both when using the models for extracting semantic information in English, and when directly transferring the resulting annotation. In addition, some frames and lexical units can be rarely represented in a parallel corpus, which leads to a low quality of their prediction.

Thus, the results show that the obtained dataset for the Russian language can be used to develop methods for extracting semantic frames and use it for other natural language processing tasks.

VI. Conclusion

In this work, methods of automatic identification of lexical units and semantic frames in Russian have been investigated. The following results were obtained:

- The existing approaches to identification of semantic frames with the use of expert FrameNet annotations have been investigated. Methods for transferring semantic information between languages have been studied, in particular, methods using parallel text corpora.
- Models of lexical units and semantic frames identification have been trained and tested for English. A neural network based on BiLSTM layers, trained on annotated sentences of the FrameNet 1.5 project, was used. Additional morphological information were also used as an input of a neural network. As a result of the experiments, it was possible to achieve the quality F1 73.46 for the selection of lexical units and F1-micro 86.78 for the model identifying semantic frames.
- An algorithm for transferring annotations from English to Russian in a parallel corpus was implemented. The use of the pretrained embedding model allowed increasing the recall of the transfer by 15%, leaving the precision at the same level.

- A set of semantic frames and lexical units linked to them in Russian has been created. A corpus of 970 thousand annotated sentences in Russian with identified lexical units and semantic frames was received. This resource can be used both for further research in the field of automatic extraction of semantic frames, and in other tasks of natural language processing.
- On the obtained set of annotated sentences, we trained models for identification of lexical units and frames in Russian. For the model of identifying lexical units, quality F1 69.08 was obtained, for the model for determining semantic frames F1-micro 84.88.

References

[1] C. J. Fillmore *et al.*, "Frame semantics," *Cognitive linguistics: Basic readings*, vol. 34, pp. 373–400, 2006.

[2] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 1998, pp. 86–90.

[3] M. Palmer, D. Gildea, and N. Xue, "Semantic role labeling," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–103, 2010.

[4] X. Carreras and L. Màrquez, "Introduction to the conll-2005 shared task: Semantic role labeling," in *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, 2005, pp. 152–164.

[5] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, "Deep semantic role labeling: What works and what's next," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 473–483.

[6] J. Cai, S. He, Z. Li, and H. Zhao, "A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware?" in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2753–2765.

[7] J. Park, S. Nam, Y. Kim, Y. Hahm, D. Hwang, and K.-S. Choi, "Frame-semantic web: a case study for korean." in *International Semantic Web Conference (Posters & Demos)*, 2014, pp. 257–260.

[8] K. H. Ohara, S. Fujii, H. Saito, S. Ishizaki, T. Ohori, and R. Suzuki, "The japanese framenet project: A preliminary report," in *Proceedings of pacific association for computational linguistics*. Citeseer, 2003, pp. 249–254.

[9] M. Pennacchiotti, D. De Cao, R. Basili, D. Croce, and M. Roth, "Automatic induction of framenet lexical units," in *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 457–465.

[10] J. C. K. Cheung, H. Poon, and L. Vanderwende, "Probabilistic frame induction," *arXiv preprint arXiv:1302.4813*, 2013.

[11] S. Tonelli, D. Pighin, C. Giuliano, and E. Pianta, "Semi-automatic development of framenet for italian," in *Proceedings of the FrameNet Workshop and Masterclass, Milano, Italy*, 2009.

[12] D. Ustalov, A. Panchenko, A. Kutuzov, C. Biemann, and S. P. Ponzetto, "Unsupervised semantic frame induction using triclustering," *arXiv preprint arXiv:1805.04715*, 2018.

[13] A. Modi, I. Titov, and A. Klementiev, "Unsupervised induction of frame-semantic representations," in *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, 2012, pp. 1–7.

[14] M. L. De Lacalle, E. Laparra, and G. Rigau, "Predicate matrix: extending semlink through wordnet mappings." in *LREC*, 2014, pp. 903–909.

[15] M. L. De Lacalle, E. Laparra, I. Aldabe, and G. Rigau, "Predicate matrix: automatically extending the semantic interoperability between predicate resources," *Language Resources and Evaluation*, vol. 50, no. 2, pp. 263–289, 2016.

[16] T.-H. Yang, H.-H. Huang, A.-Z. Yen, and H.-H. Chen, "Transfer of frames from english framenet to construct chinese framenet: a bilingual corpus-based approach," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[17] C. F. Baker, M. Ellsworth, and K. Erk, "Semeval-2007 task 19: Frame semantic structure extraction," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007, pp. 99–104.

[18] R. Johansson and P. Nugues, "Lth: semantic structure extraction using nonprojective dependency trees," in *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, 2007, pp. 227–230.

[19] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[20] D. Das, D. Chen, A. F. Martins, N. Schneider, and N. A. Smith, "Frame-semantic parsing," *Computational linguistics*, vol. 40, no. 1, pp. 9–56, 2014.

[21] S. Swayamdipta, S. Thomson, C. Dyer, and N. A. Smith, "Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold," *arXiv preprint arXiv:1706.09528*, 2017.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[23] W. Quan, J. Zhang, and X. T. Hu, "End-to-end joint opinion role labeling with bert," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 2438–2446.

[24] O. Lyashevskaya and E. Kashkin, "Framebank: a database of russian lexical constructions," in *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 2015, pp. 350–360.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

Связывание русскоязычной лексики с семантическими фреймами лексикона FrameNet

Злочевская Д.И., Лукашевич Н.В., Невзорова О.А.

Данная статья посвящена исследованию методов выделения семантических фреймов из текстов на русском языке. Рассматривается теория семантических фреймов и ее практическое использования при решении задач обработки естественного языка. Также производится ряд экспериментов по переносу крупнейшего корпуса семантических фреймов проекта FrameNet на русский язык с оценкой качества каждого из подходов. Основой метода является перенос результатов выделения сематических фреймов на английском языке с помощью параллельного русско-английского корпуса. Полученный набор семантических фреймов используется для обучения и оценки моделей выделения семантических фреймов на русском языке.

Результатом данной работы является размеченный набор данных – предложения с выделенными семантическими фреймами, а также реализация, тестирование и анализ характеристик моделей по выделению семантических фреймов на русском языке.

Received 30.05.2021