



Linking FrameNet to Russian Resources

Natalia Loukachevitch¹ , Olga Nevzorova²  , and Dana Zlochevskaya³

¹ Research Computing Center, Lomonosov Moscow State University, Moscow, Russia

² Kazan Federal University, Kremlevskaja str., 18, Kazan 420008, Russia
onevzoro@gmail.com

³ Computational Mathematics Faculty, Lomonosov Moscow State University,
Moscow, Russia

Abstract. In this paper we study three approaches for creating a FrameNet-like resource for Russian, which can be very useful for semantic analysis of Russian texts. These approaches include two methods of linking the FrameNet frames with Russian words: via WordNet and via parallel corpus. Also we study a method for connecting FrameNet and Russian semantic resource FrameBank. We plan to use created bilingual resource for developing a new Russian semantic resource.

Keywords: Semantic analysis · Frames · Language transfer · Parallel corpus · Neural networks

1 Introduction

Semantic analysis aimed at constructing semantic representations of natural sentences is important for developing natural language understanding systems. It remains one of the most difficult tasks in natural language processing. Such an analysis is based on special resources and corpora, such as FrameNet [1], VerbNet [2] or PropBank [3], which are mostly created for English.

One of the most known semantic resources models is the Charles Fillmore's FrameNet lexicon of frames [1] based on his Frame semantics model. The idea of the Frame Semantics approach is as follows [4]: the senses of words can be represented using situations, their relations and roles. For example, the situation of “cooking” (see Fig. 1) 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.

The work was funded by the Russian Science Foundation according to the research project no. 19-71-10056.

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.

Apply_heat

Definition:

A **Cook** applies heat to **Food**, where the **Temperature setting** of the heat and **Duration** of application may be specified. A **Heating instrument** generally indicated to involve the use of a **Medium** (e.g. milk or water) by which heat is transferred to the **Food**. A less semantically prominent **Food** or **Cook** is marked **Co-participant**.

Sally **FRIED** **an egg** **in butter**.

Sally **FRIED** **an egg** **in a teflon pan**.

Ellen **FRIED** **the eggs** **with chopped tomatoes and garlic**.

This frame differs from Cooking_creation in focusing on the process of handling the ingredients, rather than the edible entity that results from the process.

FEs:

Core:

Container [Container]
Semantic Type: Container

The **Container** holds the **Food** to which heat is applied.
BOIL the potatoes **in a medium-sized pan**.
Things that apply the heat directly are Heating_Instruments, e.g. crock-pot, electric skillet.

Cook [Cook]
Semantic Type: Sentient

The **Cook** applies heat to the **Food**.
Drew **SAUTEED** the garlic in butter.

Food [Food]

Food is the entity to which heat is applied by the **Cook**.
Suzy usually **STEAMS** **the broccoli**.
In instructional imperatives, this FE, which would be used for the (missing) object, is tagged CNI:
COOK on low heat for two hours. **CNI**

Heating instrument [Heat_inst]
Semantic Type: Physical_entity

This FE identifies the entity that directly supplies heat to the **Food**.
Jim **BROWNED** the roast **in the oven**.
This FE will take precedence over **Container** when both are expressed in the same constituent. For example:
Kate **COOKED** the rice **in a rice-cooker**.

Fig. 1. Apply_heat frame from FrameNet.

FrameNet is a basis of numerous works on the first step of semantic analysis called semantic role labeling [5–8]. However, such data are absent for most languages, and development of FrameNet-like resources from scratch is a very laborious process, requiring a lot of resources and time [9, 10]. Therefore various approaches for automatizing of framenet creation for other languages are discussed [11–13].

In this paper we describe three approaches for creating a FrameNet-like resource for Russian. These approaches include:

- the method of linking Russian words to FrameNet semantic frames via a chain of semantic resources: FrameNet – English WordNet – Russian WordNet (RuWordNet) – Russian words and expressions;
- the method based on the transfer of semantic frames identified in English sentences using a parallel Russian-English corpus. We train a model on the FrameNet data, then apply it to the English part of a parallel Russian-English corpus, at last we match English frame elements with Russian words in parallel sentences. The resulting set of semantic frames in Russian is used to train and evaluate the models for identifying semantic frames in Russian;

- the method for connecting FrameNet and Russian semantic resource FrameBank.

From the results of this linking, a new Russian semantic resource is planned to be created.

2 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 [9, 10]. 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 [11, 14, 15]. Cross-lingual transfer can be made via linking FrameNet with WordNet [16, 17] or via parallel corpora [18]. 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 [19, 20]:

- 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 [20], 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 [21].

The widely used SEMAFOR algorithm [22] 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 [23], 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 [24,25].

Another approach of creating FrameNet like resources is based on linking between English semantic resources: FrameNet [1], VerbNet [2], PropBank [3] and WordNet [26]. One of the latest integrating resources is the Predicate Matrix project. The developers of this project applied a variety of automatic methods to extend SemLink project coverage [27].

3 FrameNet Structure

FrameNet [1] 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 [28]. 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. Frame-Bank 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.

4 Linking Russian Words to FrameNet via WordNet

The Predicate Matrix Project provides links between several semantic resources including FrameNet frames and WordNet synsets. Russian WordNet-like resource RuWordNet has links from its units (synsets) to the WordNet's synset, therefore it is possible to link Russian words to the FrameNet frames.

4.1 RuWordNet and Its Links to WordNet

Russian wordnet RuWordNet currently contains 135 thousand Russian words and expressions [29]. It is created from the existing Russian re-source for natural language processing RuThes [30] in accordance with principles of wordnet development [20, 26], which includes:

- representation of the lexical system of a language in a form of a semantic net of units and relations between them;
- each unit represents a set of synonyms, so called synset;
- each part of speech has their own semantic net; morphologically-related synsets in semantic nets of different part of speech are connected with special relations;
- main type of relations between synsets for nouns and verbs is hyponym-hypernym relation, connecting more general and more specific synsets.

RuThes is a bilingual resource, its concepts are connected with English and Russian lexical units, these data were used to provide initial linking of the Ru-WordNet synsets, because RuWordNet synsets origin from specific RuThes concepts. The following steps were used to connect RuWordNet and Princeton WordNet synsets [31].

First, automatic linking was implemented, the results of which were checked by linguists. The main problem in automatic linking is ambiguity of English words: for linking with WordNet synsets ambiguity should be resolved. Synsets were connected automatically via:

- non-ambiguous English words and phrases;
- several possibly ambiguous English words from a concept, which have maximum connections with a single WordNet synsets;
- automatic extension of links to other parts of speech: if for example noun synsets are matched, than derivational adjectives and verbs are also matched.

Second, five thousand core WordNet synsets were manually matched.

At the third stage, most frequent Russian words, with absent WordNet links were additionally manually linked with corresponding WordNet synsets.

Currently, about 23.5 thousand links connect RuWordNet and WordNet synsets. Small number of Russian synsets have more then one link to WordNet synsets. WordNet synsets are identified via so-called interlingual index (ili). In the Web-version links go to the rdf version of WordNet (<http://wordnet-rdf.princeton.edu>), where each synset has a separate page.

4.2 Results of Linking Russian Words to FrameNet via PredicateMatrix and Wordnets

The Predicate Matrix is a lexical multilingual resource resulting from the integration of multiple sources of predicate information including FrameNet, VerbNet, PropBank, WordNet and some other resources. The current version of the Predicate Matrix includes verbal predicates for English, Spanish, Basque and Catalan, as well as nominalizations for English and Spanish. Each row of this Predicate Matrix represents the mapping of a role over the different resources and includes all the aligned knowledge about its corresponding verb sense. Fragment of a line from the Predicate Matrix for a English verb abduct looks as follows:

id:eng id:v id:abduct.01 ... wn:abduct%2:35:02 ... fn:Kidnapping fn:abduct....

where wn:abduct%2:35:02 is indication of the WordNet sense, fn: fn:Kidnapping is a link to the FrameNet frame.

This gives possibility to attach Russian words to FrameNet via path RuWordNet synset - WordNet synset - Predicate Matrix - FrameNet frames. Table 1 contains examples of found connections.

One of the problems of this approach is unresolved ambiguity of WordNet-FrameNet links in Predicate Matrix. For example, one synset of the word “boil” (immerse or be immersed in a boiling liquid, often for cooking purposes) is linked to three frames of FrameNet: fn:Absorb_heat, fn:Apply_heat, fn:Cause_harm. Also some links from RuWordNet lead to those RuWordNet synsets, which are not connected to FrameNet.

5 Linking Russian Words to FrameNet via Parallel Corpus

The method of linking Russian words to FrameNet via a parallel corpus 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.

Table 1. Examples of found connections.

Frame identifier	Russian lexical units
fn:Make_agreement_on_action	vrazit' soglasie, soglasht'sya, iz'yavit' soglasie, soglasit'sya
fn:Verdict	vynosit' opravdatel'nyj prigovor, opravdyvat' podsudimyj, opravdat' obvinyaemyj, snyat' obvinenie, opravdyvat' obvinyaemyj, opravdyvat' v sud, opravdyvat', opravdat' podsudimyj, sud opravdat', opravdat' v sud, opravdat', snimat' obvinenie, priznat' nevinovnyj, priznavat' nevinovnyj; priznavat' vinovnyj, prisudit', prisuzhdat', osudit', osudit' na, osudit' prestupnik, osuzhdat', osuzhdat' na, osuzhdat' prestupnik, vynesti obvinenie, prigovarivat', prigovarivat' k, prigovarivat' prestupnik, prigovorit', vynosit' obvinenie, vynosit' obvinitel'nyj prigovor, prigovorit' k, prigovorit' prestupnik, priznat' vinovnyj
fn:Cooking_creation	gotovit', stryapat', izgotovit' eda, gotovit'sya, sostryapat', gotovit' pishcha, gotovit' eda, sgotovit', prigotavlivat', prigotovlyat', nagotovit', prigotovit' pishcha, prigotovit' eda, prigotovit' blyudo, prigotovit', nagotavlivat', izgotovit' pishcha
fn:Apply_heat	svarit', povarit'sya, otvarit', varit', varit'sya, svarit'sya, dovarivat', otvarivat', dovarit', navarivat', navarit', povarit'

5.1 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 [23].

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 [32] was chosen as an optimizer. To prevent overfitting of the neural network, the Dropout technique [33] was used, based on random switching off of neurons from the layers of the neural network for greater generalization of the trained models.

5.2 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.

5.3 Implementation and Results

The following hyperparameters were chosen for training the models:

- the Glove model trained on the English-language Wikipedia¹ 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 2. For comparison, the SEMAFOR algorithm was applied, which determines the lexical units on the basis of linguistic rules. For this, an available author's implementation² 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

¹ <https://nlp.stanford.edu/projects/glove/>.

² <https://github.com/Noahs-ARK/semafor>.

Table 2. Results for lexical units recognition in English.

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

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 [18] was recreated. The obtained results are presented in Table 3.

Table 3. Results of model of semantic frame recognition for English.

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

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.

5.4 Translation of Annotations from English into Russian in a Parallel Corpus

For the study, we used the English-Russian parallel corpus³, 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

³ <https://translate.yandex.ru/corpus>.

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 similarity of words in an embedding model of the Russian language.

The FastText model in the Skipgram version trained on the Russian National Corpus⁴ 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 API⁵.
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 4. It can be seen that the use of the embedding model increases the recall of word matching with a slight decrease in precision.

Table 4. Results of matching words between sentences in a parallel corpus.

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

Thus, applying this algorithm to each pair of sentences in a parallel corpus, it is possible to transfer the selected lexical units and the corresponding semantic frames from English into Russian.

5.5 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.

⁴ <https://rusvectors.org/ru/models/>.

⁵ <https://cloud.google.com/translate/docs/>.

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 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 as-signed 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.

Table 5. Examples of linking Russian lexical units to the FrameNet frames.

Fear	Labeling	Reason
strah : 1042	termin : 1045	prichina : 3287
boyat'sya : 552	ponyatie : 139	osnova : 755
boyazn' : 42	etiketka : 119	osnovanie : 533
opasat'sya : 40	yarlyk : 78	motivaciya : 247
uzhas : 28	terminologiya : 24	povod : 118
strashit'sya : 3	marka : 11	motiv : 110
strashno : 3	brend : 9	poetomu : 2
ispug : 1	klejmo : 2	imenno : 1

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.

5.6 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 language⁶. The vector dimension is 300. The results of the obtained models are presented in Tables 6 and Table 7.

⁶ <https://rusvectors.org/ru/models/>.

Table 6. Results of the Russian lexical unit identification model.

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

Table 7. Results of the model for identifying semantic frames for the Russian language.

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

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.

6 Linking FrameNet and Russian FrameBank

6.1 FrameBank Structure

FrameBank is a bank of annotated samples from the Russian National Corpus. The main task of FrameBank is to identify frame elements in texts, namely to identify participants in situations indicated by predicates (verbs, nouns, adjectives, etc.), and to mark the way they are expressed, regardless of whether the units denoting participants are related syntactically or not.

FrameBank presents information:

- about the lexical constructions and the system of frames in Russian;
- about the semantic-syntactic interface in a more general sense;
- about the ambiguity of predicate tokens and how the system of meanings is related to the constructive potential of tokens.

The bank of lexical structures includes:

- construction patterns of verbs and predicatives;
- morphosyntactic patterns of predicate nouns (include not only control, but also attributive and other syntactic relations);
- morphosyntactic patterns of adjectives, adverbs, introductory phrases, various adjuncts;
- small syntax constructions (phrasemes, idiomatic constructions).

The basic units of FrameBank are individual verb constructions, not generalized frames. The vocabulary of lexical constructions represents each construction as a template, which specifies:

- a) morphosyntactic characteristics of structural elements;
- b) the syntactic rank of the participant;
- c) role of the participant;
- d) lexical and semantic restrictions of structural elements;
- e) participant status (mandatory or optional);
- f) a letter marking the participant in a short pattern, for instance “Y is brought from Z to W”.

For example, Pattern: <Snom V na.PR+ Sloc>, Pjatno vystupilo na platje (ru) ‘A stain appeared on the dress’.

6.2 Methods for Linking FrameNet and FrameBank

The purpose of linking is to establish a semantic correspondence between the predicate of a construction from FB and the semantic frame from FN. The predicate of the FB construction is expressed most often by a verb. Thus, the construction from FB and the frame from FN will be express general semantic meaning.

The data linking of these resources is carried out according to the following algorithm:

1. Corpus of examples of the selected construction from FB is being extracted.
2. The predicate of the construction is translated into English and a set of lexical units (FUs) for FN is formed.
3. Then, for each lexical unit (FU), one need to find a frame from FN that contains this lexical unit. In this way, a set of frame-candidates is formed.
4. For each frame-candidate from FN, a corpus of examples is extracted including the translated lexical unit.
5. For each example from the constructed corpora from FB and FN, one is constructed vector embedding using the word2vec model, and then vectors embeddings of the each corpus.
6. The vectors embeddings of the construction from FB and the frame from FN are compared in pairs using the cosine metric, and the frame with the largest cosine value is selected as a result.

To implement this algorithm, a software module was developed and the experiments on linking the resources were carried out. Fragments of the results obtained are shown in Table 8.

To check the quality of the program, manual testing was carried out, in which the corresponding semantic frames were found on a set of 200 constructions from FB. The accuracy of establishing correct links by this algorithm is 88%, which is a fairly good result.

Table 8. Linking constructions and frames.

Construction from FB	Translated lexical unit	Frame from FN	The result of comparing the vectors embeddings (the frame with the maximum value on the cosine measure is highlighted)
ostat'sya 1.x	stay	Avoiding Temporary_stay Residence State_continue	0.3038 0.6899 0.1057 0.5133
pisat' 1.x	write	Text_creation Contacting Statement Spelling_and_pronouncing People_by_vocation	0.7724 0.5586 0.6551 0.6979 0.5444
vyrazit' 1.x	express	Encoding Expressing_publicly Sending Facial_expression Roadways	0.7236 0.8019 0.5979 0.6978 0.4696
vorovat' 1.x	theft	Self_motion Theft	0.5870 0.7041

7 Conclusion

The FrameNet lexicon created for the English language is one of well-known lexical resources intended for semantic analysis. To create such a resource for other languages is very time-consuming and extensive procedure. In this paper we study three approaches for creating a FrameNet-like resource for Russian. These approaches include:

- the method of linking Russian words to FrameNet semantic frames via a chain of semantic resources: FrameNet – English WordNet – Russian WordNet (RuWordNet) – Russian words and expressions;

- the method based on the transfer of semantic frames identified in English sentences using a parallel Russian-English corpus. We train a model on the FrameNet data, then apply it to the English part of a parallel Russian-English corpus, at last we match English frame elements with Russian words in parallel sentences. The resulting set of semantic frames in Russian is used to train and evaluate the models for identifying semantic frames in Russian;
- the method for connecting FrameNet and Russian semantic resource FrameBank.

From the results of this linking, a new Russian semantic resource is planned to be created.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 86–90 (1998)
2. Kipper, K.: VerbNet: a broad-coverage, comprehensive verb lexicon. Dissertations available from ProQuest (2005). AAI3179808. <https://repository.upenn.edu/dissertations/AAI3179808>. Accessed 14 Nov 2021
3. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. *Comput. Linguist.* **31**(1), 71–106 (2005)
4. Fillmore, C.J., et al.: Frame semantics. In: Geeraerts, D., Dirven, R., Taylor, J. (eds.) *Cognitive Linguistics: Basic Readings*, vol. 34, pp. 373–400 (2006)
5. Palmer, M., Gildea, D., Xue, N.: Semantic role labeling. In: *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–103 (2010)
6. Carreras, X., Marquez, L.: Introduction to the CoNLL-2005 shared task: semantic role labeling. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pp. 152–164 (2005)
7. He, L., Lee, K., Lewis, M., Zettlemoyer, L.: Deep semantic role labeling: what works and what’s next. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1: long papers, pp. 473–483 (2017)
8. Cai, J., He, S., Li, Z., Zhao, H.: A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2753–2765 (2018)
9. Park, J., Nam, S., Kim, Y., Hahm, Y., Hwang, D., Choi, K.-S.: Frame-semantic web: a case study for Korean. In: *ISWC-PD 2014: Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, vol. 1272, pp. 257–260 (2014)
10. Ohara, K.H., Fujii, S., Saito, H., Ishizaki, S., Ohori, T., Suzuki, R.: The Japanese FrameNet project: a preliminary report. In: *Proceedings of the 6th Meeting of the Pacific Association for Computational Linguistics*, Halifax, Nova Scotia, pp. 249–254 (2003)
11. Pennacchiotti, M., De Cao, D., Basili, R., Croce, D., Roth, M.: Automatic induction of FrameNet lexical units. In: *EMNLP 2008: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 457–465 (2008)
12. Cheung, J.C.K., Poon, H., Vanderwende, L.: Probabilistic frame induction. [arXiv:1302.4813](https://arxiv.org/abs/1302.4813) [cs.CL] (2013)

13. Tonelli, S., Pighin, D., Giuliano, C., Pianta, E.: Semi-automatic development of FrameNet for Italian. In: *Proceedings of the FrameNet Workshop and Masterclass*, Milano, Italy (2009)
14. Ustalov, D., Panchenko, A., Kutuzov, A., Biemann, C., Ponzetto, S.P.: Unsupervised semantic frame induction using triclustering. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 55–62 (2018)
15. Modi, A., Titov, I., Klementiev, A.: Unsupervised induction of frame-semantic representations. In: *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pp. 1–7 (2012)
16. De Lacalle, M.L., Laparra, E., Rigau, G.: Predicate matrix: extending SemLink through WordNet mappings. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 903–909 (2014)
17. De Lacalle, M.L., Laparra, E., Aldabe, I., Rigau, G.: Predicate Matrix: automatically extending the semantic interoperability between predicate resources. *Lang. Resour. Eval.* **50**(2), 263–289 (2016). <https://doi.org/10.1007/s10579-016-9348-5>
18. Yang, T.-H., Huang, H.-H., Baker, C.F., Ellsworth, M., Erk, K.: SemEval 2007 Task 19: frame semantic structure extraction. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 99–104 (2007)
19. Johansson, R., Nugues, P.: LTH: semantic structure extraction using nonprojective dependency trees. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 227–230 (2007)
20. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
21. Das, D., Chen, D., Martins, A.F., Schneider, N., Smith, N.A.: Frame-semantic parsing. *Comput. Linguist.* **40**(1), 9–56 (2014)
22. Swayamdipta, S., Thomson, S., Dyer, C., Smith, N.A.: Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold. [arXiv:1706.09528](https://arxiv.org/abs/1706.09528) (2017)
23. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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: Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186 (2019)
24. Quan, W., Zhang, J., Hu, X.T.: End-to-end joint opinion role labeling with BERT. In: *Proceedings of IEEE International Conference on Big Data (Big Data)*, pp. 2438–2446 (2019)
25. Lyashevskaya, O., Kashkin, E.: FrameBank: a database of Russian lexical constructions. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) *AIST 2015. CCIS*, vol. 542, pp. 350–360. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_34
26. Fellbaum, C. (ed.): *WordNet: An Electronic Lexical Database*, p. 423. MIT Press, Cambridge (1998)
27. Palmer, M.: SemLink: linking PropBank, VerbNet and FrameNet. In: *Proceedings of the Generative Lexicon Conference*, pp. 9–15 (2009)
28. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) [cs.LG] (2017)
29. Loukachevitch, N.V., Lashevich, G., Gerasimova, A.A., Ivanov, V.V., Dobrov, B.V.: Creating Russian wordnet by conversion. In: *Computational Linguistics and Intellectual Technologies: Papers From the Annual Conference “Dialogue”*, pp. 405–415 (2016)

30. Loukachevitch, N., Dobrov, B.V.: RuThes linguistic ontology vs. Russian wordnets. In: Proceedings of the Seventh Global Wordnet Conference, pp. 154–162 (2014)
31. Loukachevitch, N., Gerasimova, A.: Linking Russian Wordnet RuWordNet to WordNet. In: Proceedings of the 10th Global Wordnet Conference, pp. 64–71 (2019)
32. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
33. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. [arXiv:1607.04606](https://arxiv.org/abs/1607.04606) [cs.CL] (2016)