Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates

Daniil Larionov
FRC CSC RAS / Moscow, Russia
RUDN University / Moscow, Russia
dslarionov@isa.ru

Artem Shelmanov
Skoltech / Moscow, Russia
FRC CSC RAS / Moscow, Russia
a.shelmanov@skoltech.ru

Elena Chistova
FRC CSC RAS / Moscow, Russia
RUDN University / Moscow, Russia
chistova@isa.ru

Ivan Smirnov
FRC CSC RAS / Moscow, Russia
RUDN University / Moscow, Russia
ivs@isa.ru

Abstract

We build the first full pipeline for semantic role labelling of Russian texts. The pipeline implements predicate identification, argument extraction, argument classification (labeling), and global scoring via integer linear programming. We train supervised neural network models for argument classification using Russian semantically annotated corpus – FrameBank. However, we note that this resource provides annotations only to a very limited set of predicates. We combat the problem of annotation scarcity by introducing two models that rely on different sets of features: one for “known” predicates that are present in the training set and one for “unknown” predicates that are not. We show that the model for “unknown” predicates can alleviate the lack of annotation by using pretrained embeddings. We perform experiments with various types of embeddings including the ones generated by deep pretrained language models: word2vec, FastText, ELMo, BERT, and show that embeddings generated by deep pretrained language models are superior to classical shallow embeddings for argument classification of both “known” and “unknown” predicates.

1 Introduction

Semantic role labeling (SRL) is one of techniques for shallow semantic parsing of natural language texts that produces predicate-argument structures of sentences. Predicates bear the central meaning of a situation expressed by a text. In most semantic theories, predicates are verbs, verbal nouns, and some other verb forms. Arguments are phrases that fill meaning slots of a situation expressed by a predicate and define its essential details. They answer such questions as “who?”, “did what?”, “to whom?”, “with what?”, “where?”, “when?”, etc. It is said that arguments play semantic roles in a situation as roles define meanings of slots. Role meanings and sizes of role inventories vary in different semantic theories and annotated corpora. Converting a text into such shallow semantic structures helps to abstract from syntactic and morphological representations of sentences and is considered to be an important technique for natural language understanding. In (Jurafsky and Martin, 2009), this is demonstrated with the following example for a predicate *break*.

- John [AGENT] broke the window [THEME].
- John [AGENT] broke the window [THEME] with a rock [INSTRUMENT].
- The rock [INSTRUMENT] broke the window [THEME].
- The window [THEME] broke.
- The window [THEME] was broken by John [AGENT].

Note, that despite the surface syntactic representations of these sentences differ, the core predicate-argument structure retains and only adjusts to available situation details.

Semantic role labeling has been shown to be beneficial in a number of tasks, where it is important to compare or query texts by meaning: machine translation (Shi et al., 2016), question answering (Shen and Lapata, 2007), information search (Osipov et al., 2010), information extraction (Bastianelli et al., 2013), sentiment analysis (Marasović and Frank, 2018), and others.

Proceedings of Recent Advances in Natural Language Processing, pages 619–628,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_073

The whole SRL process can be divided in four steps: predicate identification and identification of its frame (disambiguation), argument extraction (for each predicate), argument classification (or labeling of arguments with semantic roles), and global scoring that deals with linguistic constrains. Predicate-argument structures in some notations can be represented as two-level trees, rooted in predicates, with single tokens (nouns, adjectives, pronouns, proper names) as leaves that denote arguments. We adopt this dependency-based notation and treat the problem of semantic role labeling as constructing such trees.

There are two main types of linguistic corpora that are used for training models for SRL: FrameNet-like (Baker et al., 1998) and PropBank-like (Kingsbury and Palmer, 2002). The Russian-language resource that can be used for supervised training is a FrameBank corpus (Lyashevskaya, 2012; Lyashevskaya and Kashkin, 2015). The underlying semantic model of this resource is close to the one FrameNet is based on. The biggest difference from FrameNet besides semantic role inventory lies in the fact that FrameBank does not group several verbs into frames but introduces "frame" structures for each unique verb. The corpus contains partially annotated text samples with predicates, arguments, and their semantic roles.

A notable limitation of this resource is that there are annotations only for a very limited set of predicates. In this work, we combat the problem of annotation scarcity by introducing two classification models that rely on different sets of features: one for "known" predicates that are present in the training set and one for "unknown" predicates that are not seen in the training data. We show that the model for "unknown" predicates can deal with the lack of annotation by using pretrained embeddings. We perform experiments with various types of embeddings including the ones generated by deep pretrained language models: word2vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), and show that embeddings generated by deep pretrained language models are superior to classical shallow embeddings for semantic role labeling in both cases of "known" and "unknown" predicates.

The contribution of this paper is the following:

- We present and evaluate the first full pipeline for semantic role labeling of Russian texts. The developed models and the code are published online[1].
- We show that pretrained embeddings and language models can alleviate the problem of annotation scarcity for predicates.
- We conduct experiments that demonstrate the superiority of using embeddings generated by pretrained language models compared to shallow embeddings like word2vec and FastText.

The rest of the paper is structured as follows. Section 2 discusses the related work on semantic role labeling for Russian and other languages. Section 3 describes the developed pipeline for semantic role labeling of Russian texts. Section 4 presents the results of the experimental evaluation of the developed pipeline. Section 5 concludes and outlines the future work.

2 Related Work

The data-driven methods for semantic role labeling originate from the work (Gildea and Jurafsky, 2002), in which authors propose a statistical model based on various morpho-syntactic features and train it on the FrameNet corpus. The release of the PropBank corpus sparked a notable interest in SRL among researchers. The consecutive works and numerous shared tasks facilitated creation of elaborated machine learning models based on manually engineered lexico-syntactic features (Xue and Palmer, 2004; Punyakanok et al., 2005; Pradhan et al., 2005).

More recent works on semantic role labeling leverage deep neural networks (Collobert et al., 2011) shifting from feature-engineering to architecture-engineering. Several notable approaches suggest doing semantic role labeling in an end-to-end fashion relying only on raw low-level input consisted of characters or tokens and well-known multilayer recurrent networks (He et al., 2017; Sahin and Steedman, 2018; Marcheggiani et al., 2017). State-of-the-art approaches leverage multitask learning (Strubell et al., 2018) and self-attention techniques (Strubell et al., 2018; Tan et al., 2018). Several recent works also report that although the end-to-end approaches have

[1] https://github.com/IINemo/isanlp_srl_framebank/tree/master

managed to show comparable results, syntactic information still significantly helps semantic parsing (He et al., 2018).

It is worth noting a novel approach to creating annotated resources for semantic role labeling. In (He et al., 2015; FitzGerald et al., 2018), instead of annotating a corpora with a scheme grounded in elaborated linguistic theory, which requires highly qualified annotators, researchers suggest question-answer driven approach to construction of annotated resource based on crowd-sourcing. The recently presented QA-SRL Bank 2.0 (FitzGerald et al., 2018) is a large-scale annotated dataset built by non-experts. The construction of such a resource becomes possible due to simplicity of the annotation scheme, which provides an ability to label predicate-argument relationships using question-answer pairs.

There is a number of works devoted to automatic semantic parsing of Russian texts. In (Sokirko, 2001), a rule-based semantic parser is presented that converts a sentence into a semantic graph. The work does not provide numerical evaluation results, and the generated semantic graph is substantially different from predicate-argument structures produced in SRL. In (Shelmanov and Smirnov, 2014), authors present a rule-based semantic parser for Russian that relies on a dictionary of predicates and a set of morpho-syntactic rules created by human experts. They use this parser to automatically annotate representative corpus for supervised training of a transition-based labeler. In (Kuznetsov, 2015; Kuznetsov, 2016), an SVM-based labeling model is trained on FrameBank corpus. Authors rely on feature-engineering approach and suggest to use syntactic features and clusters of lexis. They also implement integer linear programming inference as a post processing step. These works are based on the pre-release version of the FrameBank corpus and do not provide code for data preparation, modeling, and evaluation. They also do not consider argument extraction and the problem with labeling arguments of "unknown" predicates. In (Shelmanov and Devyatkin, 2017), authors experiment with training a neural network models on the FrameBank corpus and suggest using word2vec embeddings for dealing with scarcity of predicate annotations. However, they implement only an argument classification step but not the full SRL pipeline. It is also worth noting that they performed experiments on gold-standard morphological features (POS tags and morphological characteristics), which does not reflect the real-world scenario. In this work, we additionally suggest using embeddings generated by deep pretrained language models, train models on automatically generated linguistic annotations (morphology / syntactic trees), and provide the full pipeline for semantic role labeling including argument extraction. It is also worth noting the Frame-parser project[2], however, it is in an early stage and only implements argument labeling using an SGD classifier.

3 Pipeline for Semantic Role Labeling

The limitations of the FrameBank corpus do not allow to use end-to-end / sequence labeling methods for SRL. Unlike PropBank, its text samples are annotated only partially, so they are not suitable for straightforward training of a supervised argument extractor or a combined pipeline. Therefore, we split our pipeline into multiple stages, some of which leverage rule-based methods.

The pipeline for semantic role labeling assumes that input texts are preprocessed with a tokenizer, a sentence splitter, a POS-tagger, a lemmatizer, and a syntax parser that produces a dependency tree in a Universal Dependencies format (Nivre et al., 2016). The SRL pipeline consists of the following steps: predicate identification, argument extraction, argument classification, and global scoring.

In the predicate identification step, we mark all verbs and some verb forms according to the given POS-tags of sentence tokens. We do not consider verbal nouns as predicates since they are not present in the FrameBank corpus. In the argument extraction step, for each marked predicate, we try to detect its arguments within a sentence by analyzing its syntax dependency tree with a number of manually constructed rules. The arguments in the pipeline are not spans as stated in CoNLL Shared Tasks 2004, 2005, 2012 (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005; Pradhan et al., 2012), but single tokens (nouns, proper names, or pronouns) as stated in CoNLL Shared Tasks 2008, 2009 (Surdeanu et al., 2008; Hajič et al., 2009).

For the argument classification step, we train two neural models that predict roles of arguments

[2] https://github.com/lizaku/frame-parsing

of "known" and "unknown" predicates accordingly. We call a predicate "known" if it appears in a training set within a labeled example, and "unknown" if it does not. During the inference we choose a model by simply checking a presence of a given predicate in a list of predicates appeared in the training corpus. The result of model inference is a set of probabilities for each semantic role in the inventory (above a certain threshold).

In the global scoring step, we enforce the final predicate-argument structure to fulfill the important linguistic constraint: in a single predicate-argument structure, each semantic role can be assigned only once, and each argument can have only one role.

The whole pipeline is schematically depicted in Figure 1.

3.1 Argument Extraction

Base arguments are extracted from a syntax tree of a sentence by rules that take into account POS-tags of tokens and direct syntax dependency links rooted in predicates. Arguments are often also connected to predicates not directly but through simple and complex prepositions that consist of several words. Complex prepositions are detected using the predefined list of triplets <PREP, NOUN, syntactic link>. Name of a syntactic link is used to resolve ambiguity between noun phrases and complex prepositions.

We also take into account tokens that are not related to the predicate directly but are homogeneous to base arguments. The tokens that are linked to the base arguments with a conjunct relation ("conj") are considered as extensions of base arguments and are labeled with the same semantic role as the base argument. The list of predicates is also expanded via adding the syntactic subjects and agents connected to the extracted arguments with a nominal subject ("nsubj") relation, as well as with "name" and "appos" in case it is a person name or a title. The nominal modifier ("nmod") relation is used for nominal dependents and often indicates locations. The adverbial clause modifier ("advcl") helps to find the sequences of participle clauses that have a common subject.

3.2 Argument Classification

For argument classification, we train two feed-forward neural-network models: the model for "known" predicates and the model for "unknown" predicates. The feature set of the model for "known" predicates includes embeddings of argument and predicate lemmas, as well as the following sparse lexical and morpho-syntactic features:

- Various types of morphological characteristics of both an argument and a predicate (case, valency, verb form, etc.)
- Relative position of an argument in a sentence with respect to a predicate.
- Preposition of an argument extracted from a syntax tree (including a complex preposition as a single string).
- Name of a syntactic link that connects an argument token to its parent in the syntax tree.
- Argument and predicate lemmas.

We note that the predicate lemma is one of the essential features for high-quality semantic role labeling, since predicates express a situation in a sentence and determine its meaning slots. Similar morpho-syntactic structures with different predicates can express different meanings. Therefore, the lack of annotation for a predicate in a training set hits hard classifier confidence and overall performance on examples with this "unknown" predicate. We combat this problem by introducing a second model that is trained without predicate lemmas as features, so it should rely on predicate lemma word embeddings and other features instead. This model performs worse than the model for "known" predicates on seen predicates but it is also affected less by an absence of a predicate in a training corpus. This happens because predicate lemma embeddings capture predicate semantics, and similarity between these embeddings can help the model to guess meanings of "unknown" predicates.

In this work, we experiment with various types of word embeddings obtained from shallow models word2vec and FastText, as well as from pretrained language models ELMo and BERT. Recently, it has been shown that pretrained language models provide substantially better generalization to downstream models compared to shallow embeddings built by word2vec or FastText algorithms. This happens because language models can take into account contexts of words, for which they generate an embedding, and capture much longer dependencies in texts. ELMo generates contextual word representations by using a

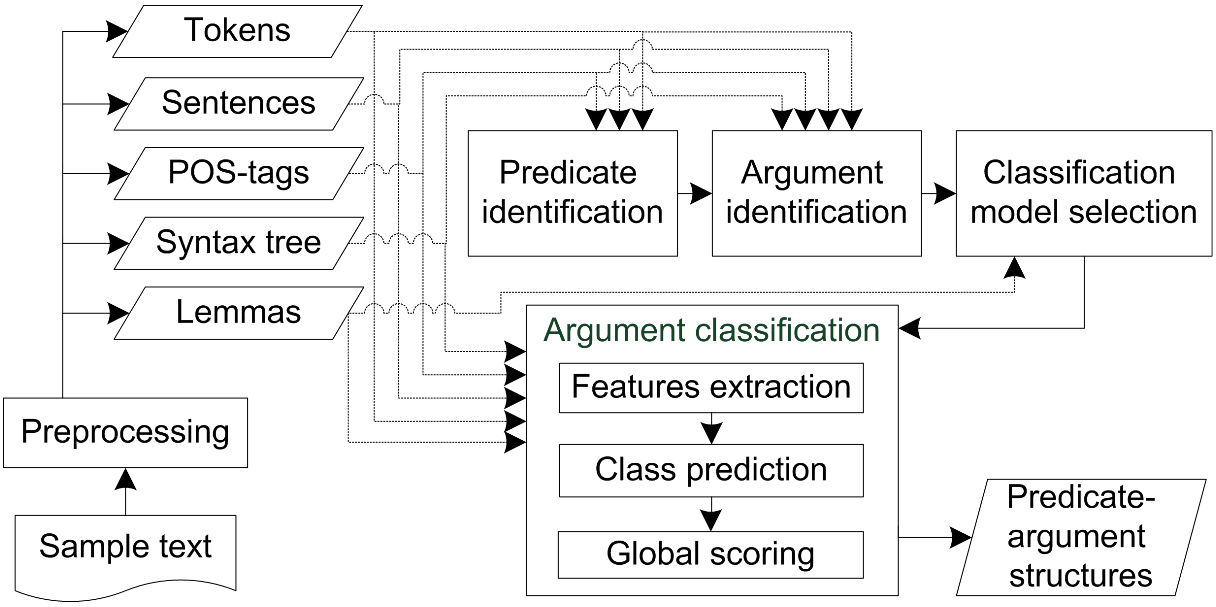

Figure 1: Semantic role labeling pipeline

stack of bidirectional LSTM layers that are trained to predict a following word on the basis of seen context. The output from each layer can be used as a set of features to downstream models. BERT is a masked language model that is trained to predict masked words in a sequence given all other words in the sequence. In addition, it is also trained to simultaneously predict whether two given sentences are consecutive. BERT uses self-attention encoder (Vaswani et al., 2017) that can be much faster than LSTM and can capture longer dependencies across sequences.

The feedforward neural network model for argument classification has three dense layers. Three inputs, namely embedding of a predicate, embedding of an argument, and sparse categorical features are separately passed through the first piecewise layer with ReLU activation. Concatenated outputs of the first layer are then propagated through another ReLU layer and the output layer with softmax activation. Before the activation function, batch normalization is applied on each hidden layer. The network is regularized with dropout. The output of this network is a vector of probabilities for each semantic role in a given inventory.

3.3 Global Scoring

Semantic labels in a single predicate-argument structure are not completely independent from each other. Moreover there is a certain linguistic constraint that requires that there should be no duplicate argument labels, since a meaning slot of a situation can be filled by just a single participant (for the core semantic roles) or a group of homogeneous arguments. In the argument classification step, we use a neural network model to produce a number of probabilities for a list of semantic roles and due to the linguistic constraint, we cannot greedily assign roles with maximum probability. In the global scoring step, we effectively produce the global optimal predicate-argument structure that fulfills the constraint by leveraging an integer linear programming inference (Punyakanok et al., 2004). Formally, it can be described in the following way. Let $x_{ij} \in \{0,1\}$ be a target variable and $x_{ij} = 1$ means an argument j has a semantic role i. Let $p(i,j)$ be a probability of assigning a role i to an argument j estimated by a neural network. Let n and m be the numbers of roles and arguments accordingly. The optimization problem formally:

$$\underset{x_{ij}}{\operatorname{argmax}} \sum_{j}^{m} \sum_{i}^{n} x_{ij} \log p(i,j)$$

$$\sum_{i}^{n} x_{ij} = 1, j = 1..m$$

$$\sum_{j}^{m} x_{ij} = 1, i = 1..n$$

$$x_{ij} \in \{0,1\}, i = 1..n, j = 1..m$$

Table 1: The merging scheme for mixed roles

Original (mixed) role	Destination role
agent – perceiver	perceiver
agent – sbj of mental state	sbj of mental state
result / target	result
location – patient	location
speaker – sbj of psychol. state	sbj of psychol. state

The optimal solution to this problem is used as the final assignment of semantic roles to arguments.

4 Experiments

4.1 Dataset and Preprocessing

FrameBank contains annotated text samples with multiple contextual sentences. Each sentence consists of tokens with their morphological features. We follow preprocessing procedure from (Shelmanov and Devyatkin, 2017) to map annotations to corresponding tokens. We also merged mixed roles annotations from the original dataset into their subsequent roles. See the merging scheme in Table 1.

Unlike (Shelmanov and Devyatkin, 2017), in this work, we do not rely on gold-standard linguistic annotations at all, since the goal of our work is to develop the parser that can process raw texts. To generate linguistic annotations for SRL input, we perform the following linguistic processing steps:

- Tokenization and sentence splitting are performed by NLTK library[3].

- Lemmatization, POS-tagging, and morphological analysis are done by MyStem (Segalovich, 2003).

- Syntax parsing is performed via UDPipe parser (Straka and Straková, 2017) with model trained on SynTagRus (Nivre et al., 2008).

These steps are implemented using the IsaNLP library[4].

The original corpus after preprocessing contains examples for 803 predicates. However, for many predicates there are just few examples, and some semantic roles are also rare. Therefore, we followed (Shelmanov and Devyatkin, 2017) and filtered the dataset keeping only predicates that have at least 10 examples. The filtered dataset retains 643 unique predicates (verbs). We also drop infrequent roles, for which the dataset contains less then 180 examples. The final corpus version contains 52,751 examples for 44 unique semantic roles.

4.2 Embeddings and Pretrained Language Models

In our experiments, we use the following publicly available pretrained word embeddings and language models:

- Word2Vec: RusVectores[5] (Kutuzov and Kuzmenko, 2017). Skip-gram model trained on Russian Wikipedia, dimension: 300.

- FastText: DeepPavlov[6] (Burtsev et al., 2018). Skip-gram model trained on a mixed corpus of Russian Wikipedia and Lenta.ru news texts, dimension: 300.

- ELMo: DeepPavlov. Pretrained on Russian Wikipedia corpus, achieves perplexity of 43.692, 2 layers, dimension: 1024.

- BERT multilingual cased: released by the authors (Devlin et al., 2018). Pretrained on 104 languages, 12 encoder blocks, produces vectors with 768 dimensions.

- RuBERT: DeepPavlov. RuBert is an adaptation of BERT-multilingual with vocabulary enriched with Russian byte-pairs (Arkhipov et al., 2019).

Both ELMo and FastText mitigate out-of-vocabulary problem, so we do not lose any predicates and arguments, while there are some misses for word2vec. BERT models are built upon byte-pair encoding, so we use only the first byte-pair representation for each token as recommended by the authors. It is worth noting that although BERT is a quite large model, it takes only 15 minutes on a single GTX 1080 Ti to encode all examples, compared to 1.5 hours for ELMo.

4.3 Neural Network Hyperparameters

We performed hyperparameter tuning using random search on the task of argument labeling for "known" predicates. The following parameters were selected and used in all further experiments:

[3] https://www.nltk.org/

[4] https://github.com/IINemo/isanlp

[5] http://rusvectores.org/en/

[6] https://deeppavlov.ai/

categorical features layer hidden size – 400, embeddings projection layer hidden size – 100, concatenated vector layer size – 400, dropout – 0.3.

4.4 Experimental Setup

We evaluated the argument extraction step using only the predicate-argument structures labeled in FrameBank and did not take into consideration any other structures in the corpus found by our parser.

For evaluation of argument labeling step, we conducted two experiments using various token representations and dataset splitting schemes.

In the first setup, we evaluate argument classification step on full set of features and test various word representations. Lexical, morphological, and syntax features are encoded in one-hot manner. Macro and micro F1 scores are calculated on a 5-fold cross-validation. Evaluation results are presented in Table 2.

In the second setup, we evaluate the performance of the argument classification models for "unknown" predicates. Thus, we divided the dataset in two parts by leaving 80% of predicates with their examples for training and 20% of predicates for testing. The sets of predicates in training and testing parts do not intersect. The division of predicates was performed at random. For this setup, there was no cross-validation. Instead, we averaged results of 5 models trained with different random seeds. In this experimental setup, we compare models for "unknown" predicates that do not take into account predicate lemma with various types of token embeddings.

To ensure the importance of introducing additional model for "unknown" predicates in our semantic role labeling pipeline and importance of predicate lemma feature, we also evaluate the model for "known" predicates on the test set with "unknown" predicates. The goal of this experiments is to show that the model for "known" predicates overfits on predicate-lemma features and performs worse than models trained specifically for "unknown" predicates, since it is "blind" to its most important feature. We report the results of the model for "known" predicates in this setup only with the embeddings that achieve the best score in the previous experiment. The test results are presented in Table 3.

4.5 Results and Discussion

In the argument extraction step, we achieve 74.48% precision, 85.12% recall, and 79.44% F1-score. We note that many false positives are due to the absence of non-core arguments in our evaluation set. These phrases that bear temporal, locative, and some other types of information are correctly identified by our parser but are considered as mistakes in the evaluation setup resulting in lower precision than it actually is. However, we see that it is the only adequate way to assess extraction quality with partially labeled data.

The results for the argument labeling step presented in Table 2 show that ELMo and BERT outperform all other approaches, including results in (Shelmanov and Devyatkin, 2017), although, unlike them, we do not rely on gold-standard morphological features. In Table 4, we also report the performance per semantic role of the model that uses ELMo word representations.

In many works, BERT outperforms ELMo by a significant margin. However, in our work, there is just an insignificant gap between ELMo and RuBERT. This is probably due to BPE tokenization scheme of BERT, since we take encoded representation only for the first subword unit of each token, with no fine-tuning, leaving a lot of information about words unused.

In the second experimental setup, the gap between RuBERT and ELMo is increased. In this case, the model based on RuBERT shows worse performance than all other approaches. However, there is a certain improvement $\Delta 1.5\%$ of micro F1 score between ELMo and word2Vec-based models. It shows that representations generated by deep pretrained language models can restore semantics of unseen predicates better than shallow models by additionally leveraging the context.

The results of the model for "known" predicates even with the best word representations ELMo are expectedly low. The performance drop compared to the model for "unknown" predicates with the same embeddings is substantial: 10% micro-F1 and more than 8% macro-F1. It is worse than any of other models except RuBERT. This proves that predicate lemma is very important as a feature and SRL pipeline should have two models to process "known" and "unknown" predicates to alleviate the domain shift.

Model	Micro F1	Macro F1
Plain Features Only	76.96 ± 0.67	73.63 ± 0.61
Word2Vec UPOS	79.87 ± 0.34	76.70 ± 0.77
FastText	80.60 ± 0.51	77.39 ± 0.36
ELMo	**83.42** ± 0.60	79.91 ± 0.40
BERT-Multiling	79.04 ± 0.63	75.68 ± 0.72
RuBERT	83.12 ± 0.60	**80.12** ± 0.62

Table 2: Performance of models in the experimental setup with “known” predicates

Model	Micro F1	Macro F1
ELMo (for known pred.)	45.51 ± 0.50	29.31 ± 0.82
Word2Vec UPOS	53.97 ± 0.21	37.29 ± 0.74
ELMo	**55.50** ± 0.51	**37.64** ± 0.41
FastText	49.37 ± 0.43	37.26 ± 0.29
BERT-Multiling	31.81 ± 0.51	21.04 ± 0.13
RuBERT	43.68 ± 0.50	30.84 ± 0.55

Table 3: Performance of models in the experimental setup with “unknown” predicates

5 Conclusion and Future Work

We presented and evaluated the first full pipeline for semantic role labeling of Russian texts. The experiments with various types of embeddings showed that the pretrained language models ELMo and BERT substantially outperform the embeddings obtained with shallow algorithms like word2vec and FastText. We also showed that providing supplementary SRL model for “unknown” predicates can alleviate the problem with annotation scarcity. We note, that in the case of predicting arguments for “uknown” predicates, the deep pretrained language model ELMo also outperformed other types of embeddings. We publish the code of the pipeline that can be used to parse raw Russian texts[7], we also publish the code for model training and experimental evaluation.

In the future work, we are going to apply semi-supervised and unsupervised algorithms to expand the training data and improve the model performance on out-of-domain data.

Acknowledgments

This paper is partially supported by Russian Foundation for Basic Research (project No. 16-29-12937, 17-07-01477).

[7] https://github.com/IINemo/isanlp_srl_framebank/tree/master

Class	Precision	Recall	F-score
agent (11.7%)	76.1	83.3	79.5
patient (10.2%)	85.1	88.7	86.9
theme (6.9%)	84.6	71.6	77.6
sbj of psychol. state (6.2%)	86.7	83.9	85.2
goer (5.7%)	82.9	89.2	85.9
cause (4.7%)	86.2	88.6	87.4
speaker (4.5%)	73.5	78.3	75.8
location (4.1%)	87.4	82.5	84.9
content of action (3.6%)	89.1	83.8	86.3
content of thought (3.4%)	74.6	79.7	77.0
content of speech (3.4%)	75.9	69.5	72.6
final destination (3.4%)	70.3	52.0	59.8
result (2.8%)	63.5	54.0	58.4
patient of motion (2.6%)	88.8	80.4	84.4
stimulus (2.4%)	85.1	72.2	78.1
cognizer (2.3%)	85.1	76.9	80.8
addressee (1.8%)	75.7	79.1	77.4
perceiver (1.7%)	90.5	79.0	84.3
counteragent (1.6%)	56.8	65.6	60.9
effector (1.4%)	77.0	81.0	78.9
subject of social attitude (1.1%)	82.2	79.5	80.8
initial point (1.1%)	76.0	80.4	78.1
topic of speech (1.0%)	58.3	81.5	68.0
manner (1.0%)	84.0	69.3	76.0
recipient (1.0%)	82.3	68.0	74.5
goal (0.9%)	80.0	67.7	73.3
field (0.7%)	90.7	91.8	91.3
attribute (0.7%)	83.5	81.5	82.5
source of sound (0.7%)	73.7	69.5	71.6
behaver (0.6%)	84.8	84.4	84.6
situation in focus (0.6%)	88.2	88.3	88.2
counteragent of social attitude (0.6%)	75.0	58.2	65.5
sbj of physiol. reaction (0.6%)	76.0	85.4	80.4
topic of thought (0.6%)	95.9	88.7	92.2
potential patient (0.5%)	89.3	90.9	90.1
status (0.5%)	89.0	78.4	83.3
patient of social attitude (0.5%)	86.1	76.2	80.8
standard (0.5%)	80.2	85.3	82.7
term (0.5%)	87.5	85.7	86.6
attribute of action (0.5%)	92.5	71.2	80.4
causer (0.4%)	72.6	65.2	68.7
initial possessor (0.4%)	83.7	73.5	78.3
potential threat (0.4%)	73.6	82.7	77.9
path (0.3%)	90.3	80.0	84.9

Table 4: The performance of the model based on ELMo embeddings in the experimental setup with “known” predicates by semantic roles. The frequencies of roles in the training corpus are presented in parentheses

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2013. Textual inference and meaning representation in human robot interaction. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 65–69.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras

Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2051–2060.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 473–483.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2061–2071.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993.

Andrey Kutuzov and Elizaveta Kuzmenko, 2017. *WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models*, pages 155–161. Springer.

Ilya Kuznetsov. 2015. Semantic role labeling for russian language based on russian framebank. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 333–338. Springer.

Ilya Kuznetsov. 2016. *Automatic semantic role labelling in Russian language, PhD thesis (in Russian)*. Ph.D. thesis, Higher School of Economics.

Olga Lyashevskaya and Egor Kashkin. 2015. Framebank: a database of russian lexical constructions. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 350–360.

Olga Lyashevskaya. 2012. Dictionary of valencies meets corpus annotation: a case of russian framebank. In *Proceedings of the 15th EURALEX International Congress*, volume 15.

Ana Marasović and Anette Frank. 2018. SRL4ORL: Improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 583–594.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 411–420.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Joakim Nivre, Igor M Boguslavsky, and Leonid L Iomdin. 2008. Parsing the syntagrus treebank of russian. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 641–648.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

GS Osipov, IV Smirnov, and IA Tikhomirov. 2010. Relational-situational method for text search and analysis and its applications. *Scientific and Technical Information Processing*, 37(6):432–437.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H Martin, and Daniel Jurafsky. 2005. Semantic role chunking combining complementary syntactic views. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 217–220.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2004. Semantic role labeling via integer linear programming inference. In *Proceedings of Coling 2004*, pages 1346–1352.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2005. The necessity of syntactic parsing for semantic role labeling. In *IJCAI*, volume 5, pages 1117–1123.

Gozde Gul Sahin and Mark Steedman. 2018. Character-level models versus morphology in semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 386–396.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280. Citeseer.

A.O. Shelmanov and D.A. Devyatkin. 2017. Semantic role labeling with neural networks for texts in Russian. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017)*, 16, pages 245–256.

A. O. Shelmanov and I. V. Smirnov. 2014. Methods for semantic role labeling of Russian texts. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2014)*, 13, pages 607–620.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21.

Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang. 2016. Knowledge-based semantic embedding for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2245–2254.

Alexey Sokirko. 2001. A short description of dialing project.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.