Classification Models for RST Discourse Parsing of Texts in Russian

Chistova E.V. (chistova@isa.ru)

FRC CSC RAS, Moscow, Russia,
RUDN University, Moscow, Russia

Shelmanov A.O. (shelmanov@isa.ru)

Skoltech, Moscow, Russia,
FRC CSC RAS, Moscow, Russia

Kobozeva M.V. (kobozeva@isa.ru), Pisarevskaya D.B. (dinabpr@gmail.com),
Smirnov I.V. (ivs@isa.ru)

FRC CSC RAS, Moscow, Russia

Toldova S.Yu. (toldova@yandex.ru)

NRU Higher School of Economics, Moscow, Russia

The paper considers the task of automatic discourse parsing of texts in Russian. Discourse parsing is a well-known approach to capturing text semantics across boundaries of single sentences. Discourse annotation was found to be useful for various tasks including summarization, sentiment analysis, question-answering. Recently, the release of manually annotated Ru-RSTreebank corpus unlocked the possibility of leveraging supervised machine learning techniques for creating such parsers for Russian language. The corpus provides the discourse annotation in a widely adopted formalisation – Rhetorical Structure Theory. In this work, we develop feature sets for rhetorical relation classification in Russian-language texts, investigate importance of various types of features, and report results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank corpus. We consider various machine learning methods including gradient boosting, neural network, and ensembling of several models by soft voting.

Key words: RST, word embedding, discourse parsing, machine learning on annotated corpus, feature selection

Классификация риторических отношений для дискурсивного анализа текстов на русском языке

Чистова Е.В. (chistova@isa.ru)

ФИЦ ИУ РАН, Москва, Россия,
Российский университет дружбы народов, Москва, Россия

Шелманов А. О. (shelmanov@isa.ru)

Сколтех, Москва, Россия,
ФИЦ ИУ РАН, Москва, Россия

Кобозева М. В. (kobozeva@isa.ru), Писаревская Д.Б. (dinabpr@gmail.com), Смирнов И.В. (ivs@isa.ru)

ФИЦ ИУ РАН, Москва, Россия

Толдова С.Ю. (toldova@yandex.ru)

НИУ ВШЭ, Москва, Россия

Ключевые слова: дискурсивный анализ, теория риторических структуру, векторные представления слов, отбор признаков, обучение на размеченном корпусе

1 Introduction

There are many natural language processing tasks that require the analysis of text beyond the boundaries of single sentences. Recently, researches have started to approach this problem by leveraging discourse parsing, which made it a very prominent research topic. One of the most widely adopted discourse models of text is Rhetorical Structure Theory (RST), developed by W. Mann and S. Thompson [1]. RST represents a text as a tree of discourse (rhetorical) relations ("Cause", "Condition", "Elaboration", "Concession", "Sequence", "Contrast", etc.) between text segments – discourse units (DUs). These units can play various roles inside a relation: nuclei contain more important information, while satellites give supplementary information. The leaves of the tree are so called elementary discourse units (EDUs), usually clauses. Discourse trees in RST integrate both shallow and deep discourse structure. Discourse units on different levels are combined by the same set of relations. The well-known applications of automatic discourse parsing include the systems for summarization [2], sentiment analysis [3], question-answering [4], natural language generation [5], and dialog parsing [6].

This work is devoted to the problem of developing a system for rhetorical parsing of Russian texts. Recently, the release of manually annotated Ru-RSTreebank corpus [7] unlocked the possibility to use machine learning techniques for this task. In particular, we consider the tasks of classification of discourse relations between DUs into rhetorical types, as well as determining the nuclearity of DUs in a relation.

The contributions of this paper are the following:

- We investigate importance of various types of features for discourse relation classification in Russian-language texts and develop a feature set for this task.
- We report the results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank corpus.
- We publish the models and the code for evaluation.

The rest of the paper is structured as follows: Section 2 presents the background and related work on discourse parsing. Section 3 briefly describes the manually annotated corpus of rhetorical structures Ru-RSTreebank. Section 4 examines features, classification models, and feature selection procedure. Section 5 describes the experimental evaluation of the developed methods, results of feature importance investigation, and results of error analysis. Section 6 concludes the paper and outlines the future work.

2 Background and Related Work

One of the early attempts at data-driven discourse parsing [8] rely to a large extent on syntactic features. The authors leverage lexicalized syntactic trees, probabilistic models, and a bottom-up parser for segmenting and building sentence-level discourse trees. In [9], syntactic features and POS tags are used as features in a shift-reduce discourse parser driven by an averaged perceptron. In HILDA parser [10] the feature set is extended with information about discourse markers, punctuation, and word-level n-grams. In some other works, it is suggested using also syntax and discourse production rules [11, 12], POS tags of the head node and the attachment node, as well as the dominance relationship between DUs, and the distance of each unit to their nearest common ancestor [13]. Some recent studies propose to abandon using any form of syntactic subtrees as features and leverage hidden outputs of a neural syntax parser as implicit features instead [14, 15].

Besides various syntactic features, one can use lexical features, semantic similarities of verbs and nouns [12] in different DUs, tokens and POS tags at the beginning and end of each DU and whether the both of them are in the same sentence [16], bag of words along with the appearing of any possible word pair from both DUs [17]. In [18], neural tensor network with interactive attention was applied to capture the most important word pairs. Authors use them as additional features to word embeddings. In [19], researchers suggest to use some entity-related features to extract implicit discourse relations between sentences of one paragraph, such as whether entities in the current DU were used in previous sentences or not. Authors claim it could be useful for detection of "Expansion"-type relations (e.g., "Restatement"), or occurrence of a topic indication, which is frequent for "Comparison" (e.g., "Contrast", "Concession") and "Temporal" relations. Other representative semantic properties were discovered in [20] for three relation types from Penn Discourse Treebank: "Comparison", "Contingency" (e.g., "Cause", "Condition"), "Expansion". Authors find that "Comparison" relations are usually expressed by negation in one of the two arguments; "Contingency" relation can be discovered if one of the DUs is a subjective judgement, e.g., it can be manifested in the lexical choice of the main verb. "Expansion" relations, being general-specific, can be encoded with pronouns tagging and named entity recognition in a Narrowing Entity Continuity feature by indefinite pronouns detection in DU1 and named entities extraction in DU2 and in a Parallel Entity Continuity feature by comparison of type of named entities in both DUs and detecting any continuity form in the predicate.

Recently, deep learning models that use low-level features were adopted for discourse parsing. In [21], authors propose a transition-based discourse parser that makes use of memory networks to take discourse cohesion into account and benefit discourse parsing, including cases of long span scenarios. Experiments were based on RST Discourse Treebank for English[1]. Several discourse parsing models were created for Chinese. In [22], a framework based on recursive neural network is proposed, it jointly models the subtasks of EDU segmentation, tree structure construction, center labeling, and sense labeling. In [18], researchers use word pairs from two discourse arguments to model pair specific clues, and integrate them as interactive attention into argument representations produced by the bidirectional long short-term memory network (Bi-LSTM). Pair patterns improve recognition of discourse relations. In [23], a text matching network is presented. It encodes the discourse units and the paragraphs by combining Bi-LSTM and CNN to capture both global dependency information and local n-gram information.

In this paper, we primarily rely on feature-engineering approach rather than on deep models for several reasons. The purpose of this work is to set a baseline for the discourse parsing of texts in Russian and investigate importance of various language factors rather to push the performance of the parser to the limit. Although deep models can perform better, they are not transparent enough for feature investigation. We also note that we are still lacking of training data for leveraging deep models. Commonly, these models have a lot of parameters (starting from hundreds of thousands) and tend to overfit on small datasets.

3 Annotated corpus

This study is based on Ru-RSTreebank[2] – first open discourse corpus for Russian [7, 24]. We use an updated version of Ru-RSTreebank that is currently freely available on demand. Currently, it consists of 179 texts, including news, news analytics, popular science, and research articles about linguistics and computer science (203,287 tokens in total). The set

[1] https://catalog.ldc.upenn.edu/LDC2002T07

[2] http://rstreebank.ru/

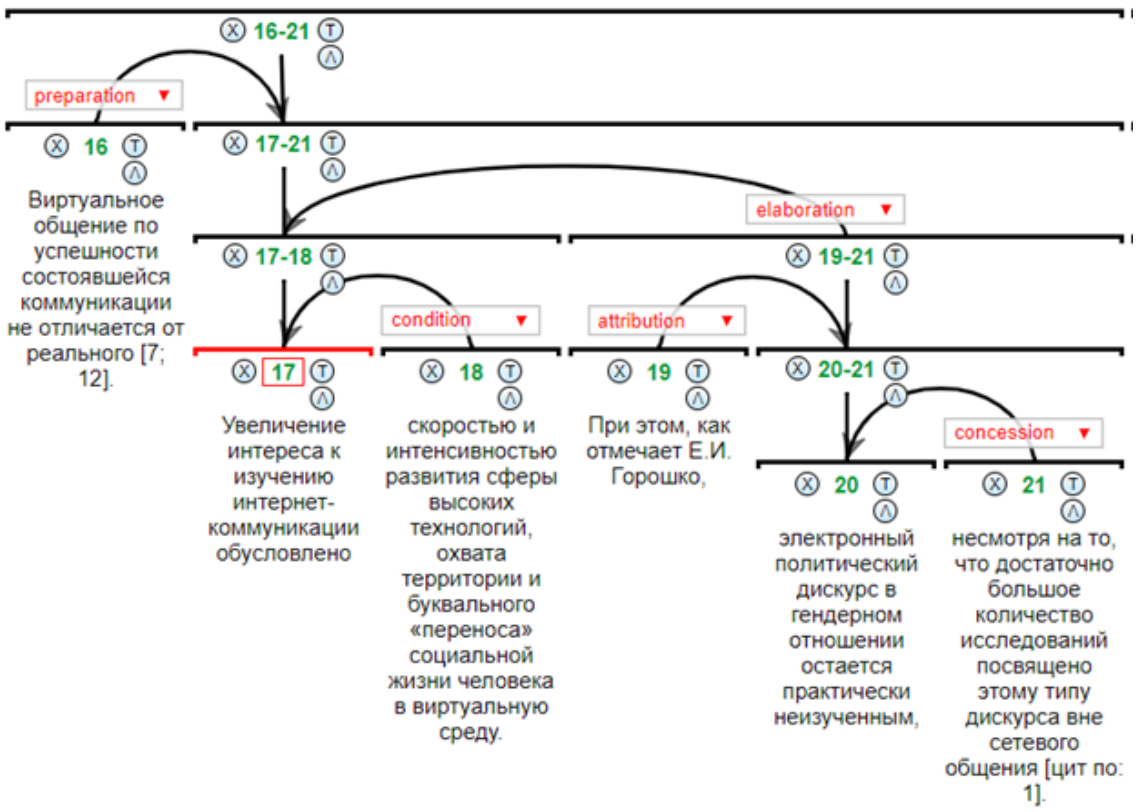

Figure 1: Visualization in rstWeb of an annotated fragment of Ru-RSTreebank

of rhetorical relations was customized to make it more suitable for Russian. The corpus was annotated with an open-source tool called rstWeb[3]. As to inter-annotator agreement, Krippendorff's unitized alpha is 81%.

The corpus contains the following types of annotations: segmentation of EDUs (mostly clauses), nuclearity of discourse units, types of discourse relations, rhetorical tree structures. In addition to ordinary multi-nuclear relation types, there is a relation type "Same-unit", which is used for annotations of cases when one discourse unit is interrupted by another one. A rhetorical tree fragment example is presented in Figure 1.

4 Features and Models for Discourse Parsing

In this work, we focus on two multiclass classification tasks. The objects for classification are pairs of DUs, that are given in the corpus. The first task is classification of DU pairs into 11 rhetorical labels. The second task is nuclearity relationship classification between DUs; there are three types of nuclearity in RST: "Satellite-Nucleus" (SN), "Nucleus-Satellite" (NS), "Nucleus-Nucleus" (NN).

4.1 Features

For both tasks, we consider combinations of various lexical, morphological, and semantic features. As lexical features, we use the list of marker phrases (or discourse connectives), nearly 450 items. It was manually composed on the basis of three sources: expressions

[3] https://corpling.uis.georgetown.edu/rstweb/info/

extracted by experts from the annotated texts, the conjunctions used in complex sentences in Russian described in RusGram[4] and the list of functional MWUs suggested in the Russian National Corpus[5].

The set of features contains various numerical features:

- Number of words.
- Average word length.
- Number of completely uppercase words.
- Number of words starting with an uppercase letter.
- Number of various morphological features. For instance, verbs have person and number.
- Part of speech tags for the first and the last word pairs of each DU.
- Features indicating the similarity between morphological features vectors of both DUs using various similarity measures namely Cosine, Hamming, Canberra, similarity measure for binarized vectors.
- Number of occurrences of stop words.
- Number of occurrences of each marker phrase.
- Occurence of each cue phrase at the beginning and the end of each DU.
- TF-IDF [25] of each DU.
- Cosine similarity between TF-IDFs.
- Jaccard index between lemmatized DUs.
- BLEU similarity measure.
- Averaged word embeddings of each DU. Embedding models were trained using word2vec [26].
- Sample of non-top11 classes examples along with the features described above were supplied to train a regressor, which predicts the probability of appearance of a mononuclear relation between DUs. This prediction is also used as a feature in the relation labeling.

4.2 Classification and Feature Selection Methods

We compared the effectiveness of various widely used supervised learning algorithms, namely, logistic regression, feedforward neural network (NN), support vector machine (SVM) with various kernels [27], and gradient boosting on decision trees (GBT) implemented in LightGBM [28] and CatBoost [29] packages. Feedforward neural network is a 2-layer perceptron regularized with dropout. The first layer activation function is ReLU. The outputs of the first layer are passed through the batch normalization. The activation on the output layer is softmax. As data imbalance highly affect the performance of neural network model, the

[4] http://rusgram.ru

[5] http://ruscorpora.ru/obgrams.html

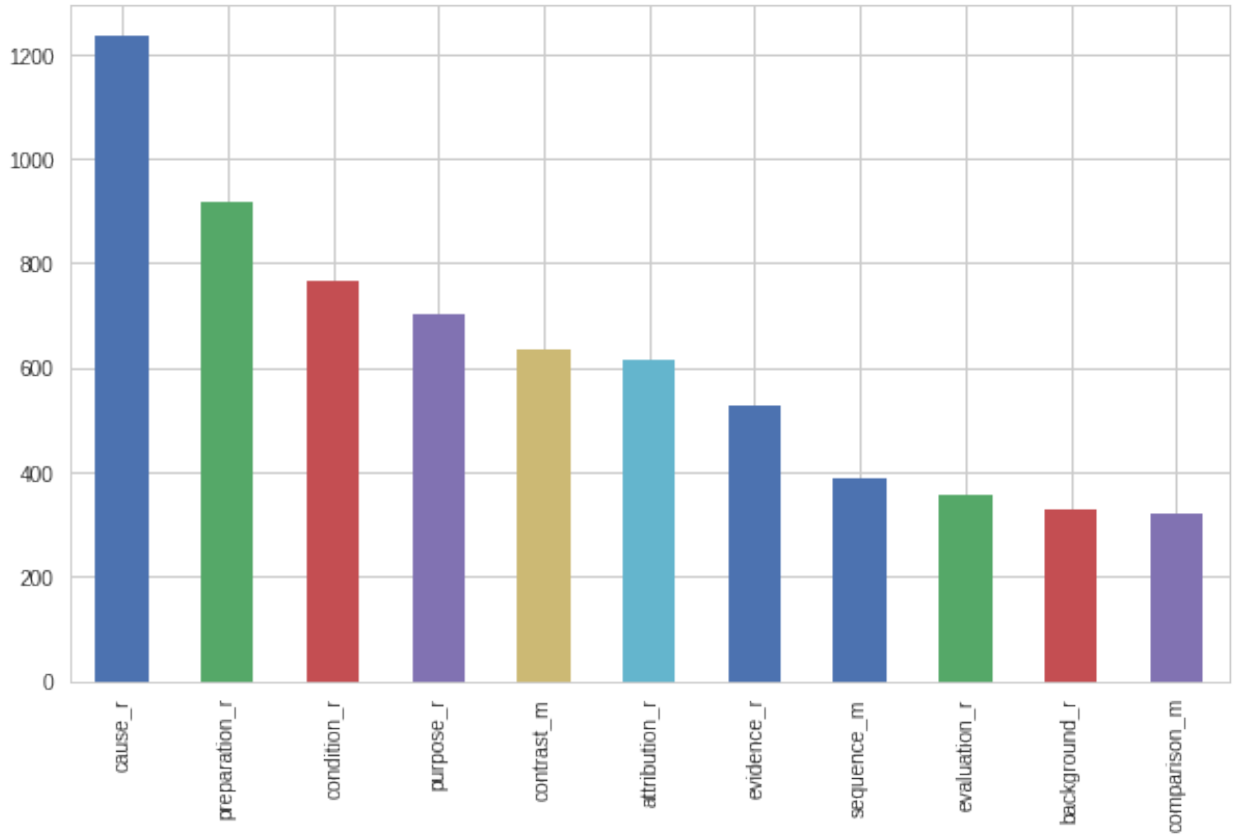

Figure 2: Distribution of rhetorical relation classes in the result dataset

SMOTE technique [30] is incorporated to oversample all classes but the majority class. We also experimented with ensembles by combining several models with soft voting.

The number of features in the original feature space is 3,273. Since only some features are informative, we perform feature selection: in some experiments we pick a strong subset of features selected by L1-regularized logistic regression model. A parameter of regularization is C. The higher C means the lower regularization strength. The best C for feature selector was found using grid search on 5-fold cross validation.

We also build ensembles of classifiers using soft voting. During the preliminary experiments, we found that ensembles of gradient boosting models with feature selection and linear SVM classifiers achieve the best performance.

5 Experiments

5.1 Dataset and Evaluation Procedure

The distribution of the classes in the original Ru-RSTreebank corpus is skewed. For experiments, we excluded "Elaboration" and "Joint" relations, since they are not very informative, although they are the most common. We also excluded "Same-unit" since it is has an utility function. Finally, we took the first 11 most representative classes, for which the dataset contains at least 320 examples. Therefore, we selected 8 mononuclear relations (these relations are marked with postfix "_r") and 3 multinuclear relations (they are marked with postfix "_m"). The result dataset for experimental evaluation contains 6,790 examples, the distribution of the classes is depicted in Figure 2.

Prior to feature extraction, the following text preprocessing steps were taken: tokeniza-

Table 1: Results of rhetorical relation classification models, %

Classifier	Macro F_1		Micro F_1	
	mean	std	mean	std
NN	49.43	1.52	55.78	1.16
Logistic Regression	50.81	1.06	53.81	1.84
LGBM	51.39	2.18	59.91	1.32
Linear SVM	51.63	1.95	56.61	1.54
L_1 Feature selection + LGBM	51.64	2.22	60.29	1.74
CatBoost	53.32	0.96	60.71	0.81
L_1 Feature selection + CatBoost	53.45	2.19	61.09	1.96
voting((L_1 Feature selection + LGBM), Linear SVM)	54.67	1.80	62.39	1.51
voting((L_1 Feature selection + CatBoost), Linear SVM)	54.67	0.38	62.32	0.41

tion, lemmatization, part-of-speech tagging, and morphological analysis using MyStem [31]. The pipeline was implemented via IsaNLP[6] Python library.

For evaluation, we used the standard metrics: precision, recall, and F_1. Macro-averages were employed as our main measurements, and accuracy was omitted, since the distributions of classes are unbalanced. We perform all our experiments using 5-fold cross validation with stratified randomized split of the dataset into 90% for training and 10% for testing.

A randomized grid search algorithm was used to find the optimum logistic regression and SVM parameters: C and type of penalty (L1, L2), and neural network parameters: number of units for each layer, activation function for each layer, dropout rate. Randomized grid search was used for selecting the best hyperparameters for gradient boosting models: number of trees, number of leaves, learning rate, feature sampling ratio, and regularization coefficients. For selection of optimal number of iterations in a CatBoost model, we used its built-in overfitting detector.

After hyperparameter tuning, we get the following best parameters. Logistic regression: inverse regularization strength: 0.001 and L2 penalty. SVM: inverse regularization strength: 0.0001 and L2 penalty, kernel: linear. LightGBM: number of leaves: 36, number of iterations: 1,000, bagging fraction: 0.9, learning rate: 0.1. CatBoost: number of iterations: 2,000, learning rate: 0.1. NN: size of hidden layer: 100, dropout: 0.5, optimization algorithm: Adam, learning rate: 0.01, batch size: 128, number of epochs: 7.

5.2 Main Results

Table 1 summarizes the results of experiments with models for rhetorical relation classification. The results show that gradient boosting models outperform other models. Ensemble of CatBoost model with selected features and a linear SVM model owns the best score.

We evaluated the importance of features related to the word order in the document. There are two types of discourse markers in the feature set: positional, i.e. whether a cue is found at the beginning or at the end of DUs and quantitative, i.e. a number of a cue in each DU. In Table 2, we see a performance drop when removing positional features. At the same time, we can observe that quantitative features do not significantly affect the F_1 score.

The results for distinguishing "Satellite-Nucleus", "Nucleus-Satellite", and "Nucleus-Nucleus" types of relations are presented in Table 3. We used the full set of features described in subsection 4.1. The experiment shows that the gradient boosting models strongly outperform feedforward neural network, SVM and logistic regression classifiers.

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

Table 2: F_1, % for rhetorical relation classification task with different feature sets

Feature set	Macro F_1		
	Logistic Regression	Linear SVM	CatBoost
All features	51.5	50.6	52.4
w/o quantitative features	-0.3	+0.1	-0.1
w/o positional features	-4.0	-4.0	-2.8

Table 3: F_1 for the nuclearity recognition models, %

Classifier	Macro F_1		Micro F_1	
	mean	std	mean	std
Linear SVM	63.01	0.58	64.20	0.52
NN	63.32	0.88	64.59	0.75
Logistic Regression	63.66	0.37	65.02	0.26
L1 Feature selection + LGBM	67.82	0.86	69.17	0.73
CatBoost	68.03	0.45	69.37	0.36
LGBM	68.81	0.77	70.17	0.67
L1 Feature selection + CatBoost	68.82	0.84	70.31	0.76

From the whole set of features (3,624 features), CatBoost model for rhetorical type relation classification selected 2,014 as important features. Analysis of this features is presented in Table 4. We can see that the most important features for this model are related to discourse markers. Table 4 also shows the performance drop when removing features from this model. As we can see, after removing the information about 1,887 features related to discourse markers, this model loses 2.49% of macro F_1.

5.3 Error Analysis

The classification report of the best performed model using a variety of measures is presented in Table 5. In Figure 3, we also provide the confusion matrix generated by this model. Asymmetric relations labeling has relatively better performance, we achieved 74.36% F_1 score for "Attribution" relation.

The worst performance, under 50% F_1 score, was obtained with 4 classes that have least number of training instances: "Comparison" (320 samples), "Evidence" (529 samples), "Evaluation" (356 samples), and "Background" (328 samples). For example, "Evidence", "Evaluation", and "Background" are often recognized as "Cause", the most represented class (1235 samples). The model has a very low recall score on "Background" relation, often labeling it as "Preparation". Macro averaged F_1 score for the classification on the top 7 relations is 72.34±1.37%.

Errors with relation labeling partly occur when there is semantic similarity between true type and predicted type, such as in pairs "Preparation"-"Background", "Comparison"-"Contrast", "Cause"-"Evidence", "Purpose"-"Cause", "Preparation"-"Attribution", "Preparation"-"Sequence". In other cases, such as "Cause"-"Preparation" or "Preparation"-"Attribution", errors can be caused by stylistic difference in news texts/scientific texts that are included in corpus. There are also cases when relation types are not semantically close to each other, these ones need more thorough investigation. For example, if "Cause" is predicted instead of "Contrast", the error can be explained by occurrences of possible cause markers in nucleus or satellite, and corresponding punctuation marks: '[В _основе_ фразеологического сочетания лежат две заимствованные из турецкого языка лексемы_:_] [а сама идиома является

Table 4: Important features selected by CatBoost model per feature type

Type	Features	Number	% in selected	Performance drop, %
Lexical	4 elements of TF-IDF vectors for the first DU; 4 elements of TF-IDF vectors for the second DU;	8	0.4	0.11
Morpho-syntactic	Combinations of punctuation, nouns, verbs, adverbs, conjunctions, adjectives, prepositions, pronouns, numerals, particles at the beginning of a first DU; Combinations of punctuation, verbs, adverbs, nouns, pronouns, adjectives, conjunctions, prepositions, particles, numerals at the end of a first DU; Number of nouns in instrument case, pronouns, adverbs in a first DU; Various combinations of verbs, pronouns, nouns, adverbs, conjunctions, punctuation, particles at the beginning of a second DU; Various combinations of punctuation, nouns, verbs, pronouns, adverbs, adjectives, prepositions, conjunctions, particles at the end of a second DU; Number of occurrences of conjunctions, adverbs, adjectives, pronouns, adpositions in a second DU; Number of passive verbs, gerunds and infinitives in a second DU; Correlation between morphological features vectors of DUs.	119	5.9	0.45
Textual	Number of occurrences of 355 markers in a first DU (18%) Number of occurrences of 331 markers in a second DU (17%) Occurrences of 298 markers at the beginning of X (16%) Occurrences of 326 markers at the end of X (17%) Occurrences of 335 markers at the beginning of Y (19%) Occurrences of 242 markers at the end of Y (13%)	1887	93.69	2.49

Table 5: Relation labeling performance for each class, %

Class	Precision	Recall	F_1-score
attribution	73.11	75.77	74.36
purpose	71.87	73.71	72.70
condition	73.60	65.75	69.36
preparation	57.82	81.09	67.49
cause	51.73	69.96	59.46
contrast	68.43	56.69	56.69
sequence	54.46	54.55	54.22
evidence	44.75	34.53	38.95
comparison	50.43	31.25	38.49
evaluation	31.89	17.46	22.56
background	24.09	5.15	8.41

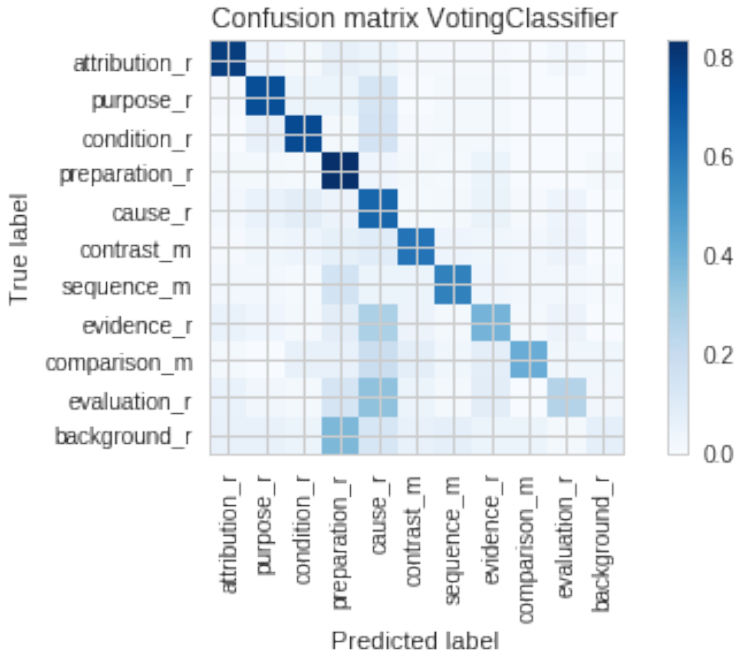

Figure 3: Confusion matrix for the best model

точной калькой турецкого выражения.]' ('[Two lexemes borrowed from Turkish are at the heart of the phraseological unit _:_] [and the idiom itself is a calque of the Turkish expression.]'), '[Текст перевода при этом не является копией или подобием исходного текста.] [Он _порождается_ путем воплощения на языке перевода указанной концептуальной структуры.]' ('[The translated text is not a full copy or a semblance of the original text.] [It is created by emphasizing a specified conceptual structure in the language of translation.]').

6 Conclusion and Future Work

We investigated the performance of different algorithms and features for discourse relations labeling and nuclearity type classification. We found that textual, morpho-syntactic, and lexical features are equally important in the relation labeling; both positional and quantitative textual features improve the quality of classification. Source code of the experiments is

available online[7].

In our future work we are going to implement the complete pipeline for discourse parsing of Russian texts including segmentation and discourse tree construction. We also looking forward to employ state-of-the art deep learning techniques and pretrained language models for relation classification.

Acknowledgments

This paper is partially supported by Russian Foundation for Basic Research (project No. 17-29-07033, 17-07-01477).

We would like to express our gratitude to the corpus annotators T. Davydova, A. Tugutova, M. Vasilyeva and Y. Petukhova.

References

[1] W.C. Mann and S.A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[2] T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, 2013.

[3] S. Somasundaran. *Discourse-level relations for Opinion Analysis*. PhD thesis, University of Pittsburgh, 2010.

[4] Chai J.Y. and Jin R. Discourse structure for context question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL*, 2004.

[5] B. Qin, D. Tang, X. Geng, D. Ning, J. Liu, and T. Liu. A planning based framework for essay generation. *arXiv preprint arXiv:1512.05919*, 2015.

[6] S. Afantenos, E. Kow, N. Asher, and J. Perret. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[7] D. Pisarevskaya, M. Ananyeva, M. Kobozeva, A. Nasedkin, S. Nikiforova, I. Pavlova, and A. Shelepov. Towards building a discourse-annotated corpus of Russian. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue*, number 16, page 23, 2017.

[8] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 149–156, 2003.

[9] Kenji Sagae. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 81–84, 2009.

[7] http://nlp.isa.ru/paper_dialog2019/

[10] H. Hernault, H. Prendinger, and M. Ishizuka. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3), 2010.

[11] Z. Lin, M. Y. Kan, and H. T. Ng. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, 2009.

[12] V. W. Feng and G. Hirst. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 60–68, 2012.

[13] V. W. Feng and G. Hirst. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 511–521, 2014.

[14] M. Zhang, Y. Zhang, and G. Fu. End-to-end neural relation extraction with global optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1730–1740, 2017.

[15] N. Yu, M. Zhang, and G. Fu. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, 2018.

[16] J. Li, R. Li, and E. Hovy. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2061–2069, 2014.

[17] B. Zhang, J. Su, D. Xiong, Y. Lu, H. Duan, and J. Yao. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, 2015.

[18] F. Guo, R. He, D. Jin, J. Dang, L. Wang, and X. Li. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 547–558, 2018.

[19] A. Louis, A. Joshi, R. Prasad, and A. Nenkova. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 59–62, 2010.

[20] W. Lei, Y. Xiang, Y. Wang, Q. Zhong, M. Liu, and M. Y. Kan. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] Y. Jia, Y. Ye, Y. Feng, Y. Lai, R. Yan, and D. Zhao. Modeling discourse cohesion for discourse parsing via memory network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 438–443, 2018.

[22] C. A. Lin, H. H. Huang, Z. Y. Chen, and H. H. Chen. A unified RvNN framework for end-to-end Chinese discourse parsing. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 73–77, 2018.

[23] S. Xu, G. Li, P.and Zhou, and Q. Zhu. Employing text matching network to recognise nuclearity in chinese discourse. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 525–535, 2018.

[24] S. Toldova, M. Kobozeva, and D. Pisarevskaya. Automatic mining of discourse connectives for russian. In *Conference on Artificial Intelligence and Natural Language*, pages 79–87, 2018.

[25] G. Salton and Ch. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR Workshop*, 2013.

[27] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[28] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[29] A. V. Dorogush, V. Ershov, and A. Gulin. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[30] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

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