Towards the Data-driven System for Rhetorical Parsing of Russian Texts

Elena Chistova
FRC CSC RAS / Russia, 119333
RUDN University / Russia, 117198
chistova@isa.ru

Maria Kobozeva, Dina Pisarevskaya
FRC CSC RAS / Russia, 119333
kobozeva@isa.ru,
dinabpr@gmail.com

Artem Shelmanov
Skoltech / Russia, 121205
FRC CSC RAS / Russia, 119333
a.shelmanov@skoltech.ru

Ivan Smirnov
FRC CSC RAS / Russia, 119333
ivs@isa.ru

Svetlana Toldova
NRU Higher School of
Economics / Russia, 101000
toldova@yandex.ru

Abstract

Results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank – first Russian corpus annotated within RST framework – are presented. Various lexical, quantitative, morphological, and semantic features were used. In rhetorical relation classification, ensemble of CatBoost model with selected features and a linear SVM model provides the best score (macro $F_1 = 54.67 \pm 0.38$). We discover that most of the important features for rhetorical relation classification are related to discourse connectives derived from the connectives lexicon for Russian and from other sources.

1 Introduction

One of the widely used discourse models of text is the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). It represents a text as a constituency tree containing discourse (rhetorical) relations between text segments – discourse units (DUs). These units can play different 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), they usually are represented as clauses. Discourse units of different levels are combined by the same set of relations.

The goal of our work is the development of a data-driven system for rhetorical parsing of Russian texts. For training, we use recently released Ru-RSTreebank corpus (Pisarevskaya et al., 2017). In this paper, we describe the pipeline of the parser, present the developed featureset for relation classification task, and present the results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank. Special attention is paid to the importance of discourse connectives.

Discourse connectives are clues signalling that there is a definite relation between two DUs, such as “in consequence of” for “Effect” or “because of” for “Cause”. Some of them are functional words (primary connectives), the rest of them, secondary connectives, are less grammaticalized (Rysova and Rysova, 2014; Danlos et al., 2018), but also should be presented in exhaustive lexicons of connectives. We find that these cue phrases are informative features for rhetorical relation classification.

2 Related Work

First discourse parsers were trained mostly on syntactic features. The authors of (Soricut and Marcu, 2003) experiment with lexicalized syntactic trees for sentence segmentation. In (Subba and Di Eugenio, 2007), authors leverage discourse cue phrases and punctuation in addition to syntactic structure of sentences and POS tags. The same features along with information about n-grams are used to define rhetorical relations in the HILDA parser (Hernault et al., 2010). It is also suggested to use syntax and discourse production rules (Lin et al., 2009; Feng and Hirst, 2012), POS tags of the head node and the attachment node, as well as the dominance relationship between EDUs, and the distance of each unit to their nearest common ancestor (Feng and Hirst, 2014).

In addition to syntactic features, one can use lexical features, semantic similarities of verbs and nouns (Feng and Hirst, 2012) in different EDUs, tokens and POS tags at the beginning and the end of each EDU and whether both of them are in the same sentence (Li et al., 2014), bag of words along with the appearing of any possible word pair from both EDUs (Zhang et al., 2015). In (Joty and Ng., 2015), among other features, authors use discourse cues, lexical chains, and syntactic fea-

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 82–87
Minneapolis, MN, June 6, 2019. ©2019 Association for Computational Linguistics

tures. In (Guo et al., 2018), neural tensor network with interactive attention was applied to capture the most important word pairs from two discourse arguments. These pairs were used as features in addition to word embeddings.

As discourse connectives are important for discourse parsing, recently, lexicons of connectives have been created for several languages. There are lexicons for French (Roze et al., 2012), Czech (Synková et al., 2017), German (Scheffler and Stede, 2016), English (Das and Stede, 2018). For example, DiMLex, a lexicon for German, consists of 275 connectives (Scheffler and Stede, 2016), DiMLex-Eng, the lexicon for English, contains 149 connectives (Das and Stede, 2018). There are also PDTB-based lexicons for French (Laali and Kosseim, 2017) and Portuguese (Mendes and Dombek, 2018).

Recently, deep learning models that use low-level features were adopted for discourse parsing. (Jia et al., 2018) propose a transition-based discourse parser for English that uses memory networks to take discourse cohesion into account. (Chuan-An et al., 2018) propose a framework based on recursive neural network that jointly models several subtasks including EDU segmentation, tree structure construction, as well as center and sense labeling. (Xu et al., 2018) present a text matching network that 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 work, we run several experiments that let investigate the importance of various features for the first data-driven discourse parser for Russian.

3 Corpus Details

Ru-RSTreebank[1] is the first discourse corpus for Russian (Pisarevskaya et al., 2017) annotated within the RST framework. The updated version, used in this research, as well as the guidelines for annotators, are currently freely available on demand. The corpus consists of 179 texts: 79 texts of such genres as news, news analytics, popular science, and 100 research articles about linguistics and computer science (203,287 tokens in total). The corpus was manually annotated with an open-source tool called rstWeb[2]. The customized set of rhetorical relations was adapted for the Russian language. Last value for Krippendorff's unitized alpha, that is used to measuring inter-annotator agreement, is 81%.

Following types of annotations are provided in the corpus: segmentation of EDUs, discourse units nuclearity, types of discourse relations, rhetorical tree construction. Clauses were mostly used as EDUs, with some adaptations for Russian. Verbal adverb phrases are emphasized as EDUs only if they have causal or clarifying meaning. Separate EDUs can occur without verb if they contain prepositional phrases that have cause, effect, contrast, or concession meaning. The release of this corpus unlocked the possibility to use machine learning techniques for discourse parsing.

We created a lexicon of discourse connectives, based on this corpus. The procedure is similar to that described in (Toldova et al., 2017). The connectives from the lexicon were further used as features for discourse parsing of Russian texts.

4 Parsing Approach

4.1 Parsing Pipeline

We divide the task of automated discourse parsing into five subtasks: sentence segmentation, relation prediction, discourse tree construction, classification of connected DU pairs into nuclear-satellite, and labeling relations between DUs.

Sentence segmentation task can be performed with external rule-based tools such as AOT.ru[3] and lies outside the scope of this work. Relation prediction is a simple binary classification task. Positive objects for this task are provided by gold parses of the corpus. Negative objects are generated by considering adjunct unconnected DUs in the gold parses. For construction of the connected discourse tree, we adopt an algorithm presented in (Hernault et al., 2010). The algorithm greedily merges DUs according to probabilities obtained from binary classification on the previous step.

Determining nuclear-satellite relations between DUs according to RST is a three-label classification task: "Satellite-Nucleus" (SN), "Nucleus-Satellite" (NS), "Nucleus-Nucleus" (NN). The final step, in which we predict a label of DU relations, is a multi-label classification task (we select 11 most important relations) that uses results of nuclear-satellite classification.

[1] http://rstreebank.ru/

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

[3] http://aot.ru/

4.2 Classification and Feature Selection Methods

We compare the effectiveness of various widely used supervised learning algorithms: logistic regression, support vector machine with linear kernel, and gradient boosting on decision trees (GBT) implemented in LightGBM[4] and CatBoost[5] packages. Since the feature space is too large and sparse for GBT methods, we perform feature selection in order to keep only the most informative features. For this purpose, we use a wrapper method implemented via logistic regression with L1 regularizer. The regularizer makes the model to aggressively zero feature coefficients during training, which leads to a smaller effective feature space. We also experiment with soft-voting ensembles that combine linear classifiers with GBT models.

4.3 Features

We use combinations of various lexical, quantitative, morphological, and semantic features. Lexical features contain a number of occurrences of cue phrases from a manually composed list of discourse connectives. The list contains nearly 450 items collected from three sources: expressions derived from the connectives lexicon for Russian mentioned above, conjunctions used in complex sentences in Russian described in RusGram[6], and the list of functional multi-word expressions suggested in the Russian National Corpus[7]. Each connective yields a feature according to one-hot encoding. Lexical features also include TF-IDF vectors of bags of words, cosine similarity between these vectors, BLEU, and Jaccard similarity metrics. Quantitative features include number of words, average word length, number of uppercased letters, as well as a number of words that start with uppercase. Morphological features encompass vector of counts of morphological characteristics in each DU, several similarity measures between these vectors and part of speech tags for the first and the last word pairs of each DU. Semantic features include averaged word embeddings of each DU. The word embedding model used in this work is described in (Toldova et al., 2018). The peculiarity of this model is that stop words and punctuation marks were not removed during pretraining, whereby discourse connectives were not lost. For rhetorical relation classification, in addition, we use probabilities obtained in the nuclear-satellite classification step according to a stacking technique.

Classifier	Macro F_1, %	
	mean	std
Linear SVM	63.13	0.39
Logistic Regression	63.65	1.08
CatBoost	67.79	0.57

Table 1: Performance of nuclear-satellite classification models.

5 Experiments

5.1 Evaluation Procedure and Results

For experiments, we excluded “Elaboration” and “Joint” relations, since although they are the most common relations, they are also not very informative. We decided to focus on more specialized relation types. We also excluded “Same-unit”, since it was used in the annotation only for utility purposes to mark discontinuous EDUs. Except aforementioned ones, we took the first 11 most representative classes, for which the dataset contains at least 320 examples. We selected 8 mono-nuclear relations (“Cause”, “Preparation”, “Condition”, “Purpose”, “Attribution”, “Evidence”, “Evaluation”, “Background”) and 3 multi-nuclear relations (“Contrast”, “Sequence”, “Comparison”). The dataset for experimental evaluation contains 6,790 examples. We note that the distribution of the classes is skewed. Before feature extraction, we performed the following preprocessing: tokenization, lemmatization, part-of-speech tagging, and morphological analysis using MyStem tool (Segalovich, 2003). The hyperparameters of our models are tuned using randomized search and overfitting detection tools built in gradient boosting packages. The evaluation scores are obtained using 5-fold cross-validation procedure with macro-averaging.

The results for distinguishing “Satellite-Nucleus”, “Nucleus-Satellite”, and “Nucleus-Nucleus” types of relations are presented in Table 1. The experiment shows that the CatBoost model outperforms linear SVM and logistic regression classifiers.

Table 2 summarizes the results of the exper-

[4] https://lightgbm.readthedocs.io/en/latest/

[5] https://tech.yandex.ru/catboost/

[6] http://rusgram.ru

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

Classifier	Macro F_1, %	
	mean	std
Logistic Regression	50.81	1.06
LGBM	51.39	2.18
Linear SVM	51.63	1.95
L_1 Feature selection + LGBM	51.64	2.22
CatBoost	53.32	0.96
L_1 Feature selection + CatBoost	53.45	2.19
voting((L_1 Feature selection + LGBM), Linear SVM)	54.67	1.80
voting((L_1 Feature selection + CatBoost), Linear SVM)	54.67	0.38

Table 2: Performance of rhetorical relation classification models.

iments with models for rhetorical relation classification. The results show that GBT models strongly outperform other methods. Also, we observe that training on the features selected by L1-regularized logistic regression reduces the variance of GBT models. Ensembles of GBT models with selected features and a linear SVM model own the best score. We should note that the qualitative performances of ensembles with LightGBM and CatBoost are almost the same, however, the computational performance of the latter is significantly better. Therefore, we used CatBoost model for the assessment of the feature importance.

5.2 Feature Importance and Error Analysis

From the whole set of features (3,624 features), CatBoost model for rhetorical type relation classification selected 2,054 informative lexical, morphological, and semantic features (word embeddings).

Important lexical features (1,941) are: occurrences of 318 cue phrases at the beginning and of 326 cue phrases at the end of the first DU; occurrences of 243 cue phrases at the beginning and of 353 cue phrases at the end of the second DU; number of occurrences of 345 cue phrases in the first DU; number of occurrences of 356 cue phrases in the second DU; 5 elements of TF-IDF vectors and 2 elements of averaged word embeddings for the first DU and 9 elements of TF-IDF vectors for the second DU. Important morphological features (97) are: combinations of punctuation, nouns, verbs, adverbs, conjunctions, adjectives, prepositions, pronouns, numerals, particles as the first word pairs of discourse units; combinations of punctuation, verbs, adverbs, nouns, pronouns, adjectives, conjunctions, prepositions, particles, numerals as the last word pairs of discourse units. Therefore, most of the important features are related to discourse connectives.

The 20 least important features include 5 elements of word embeddings of the first DU, 3 elements of TF-IDF vectors and 2 elements of word embeddings of the second DU; average length of the first DU, number of finite verbs in both DUs, one occurrence of a keyword in the second DU; number of nouns in the second DU; Jaccard index between DUs; number of words that start with capital letter in both DUs; number of words in the first DU; occurrence of a period mark at the end of the first DU.

Error analysis of the models for rhetorical relation classification shows that mistakes often occur when there is semantic similarity between true and predicted class for such pairs as: “Comparison”-“Contrast”, “Cause”-“Evidence”. Another reason behind mistakes is the usage of connectives: for instance, if “Cause” is predicted instead of “Contrast”, the error can be explained by occurrences of possible cause cue phrases in a nucleus or a satellite. Relations between long DUs that consist of several EDUs are influenced by the cue phrases inside EDUs, which sometimes results in errors. Especially it concerns the cases of “Evidence” (instead of “Contrast”), “Sequence” (instead of “Comparison”) and “Cause” (instead of “Evidence”).

6 Conclusion

We presented the first RST-based discourse parser for Russian. Rhetorical relation classifier and algorithm for building the RST-tree were implemented for discourse analysis of texts in Russian. Our experiments showed that the ensemble of CatBoost model with selected features and a linear SVM model provides the best results for relation classification. Feature selection procedure showed high importance of discourse connectives. In the future work, we are going to apply an extended version of discourse connectives lexicon for relation classification task, as well as implement more complex deep learning methods.

Acknowledgements

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

Lin Chuan-An, Hen-Hsen Huang, Zi-Yuan Chen, and Hsin-Hsi Chen. 2018. 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.

Laurence Danlos, Katerina Rysova, Magdalena Rysova, and Manfred Stede. 2018. Primary and secondary discourse connectives: definitions and lexicons. *Dialogue & Discourse*, 9(1):50–78.

Tatjana Scheffler Peter Bourgonje Das, Debopam and Manfred Stede. 2018. Constructing a lexicon of english discourse connectives. In *Proceedings of the SIGDIAL 2018 Conference*, pages 360–365.

Vanessa Wei Feng and Graeme Hirst. 2012. 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.

Vanessa Wei Feng and Graeme Hirst. 2014. 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.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. 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.

Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, et al. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3).

Yanyan Jia, Yuan Ye, Yansong Feng, Yuxuan Lai, Rui Yan, and Dongyan Zhao. 2018. 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.

Giuseppe Carenini Joty, Shafiq and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Majid Laali and Leila Kosseim. 2017. Automatic mapping of french discourse connectives to pdtb discourse relations. In *Proceedings of the SIGDIAL 2017 Conference*, pages 1–6.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2061–2069.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. 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.

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

Iria Del Ro Gayo Manfred Stede Mendes, Amlia and Felix Dombek. 2018. A lexicon of discourse markers for portuguese ldm-pt. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 4379–4384.

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Towards building a discourse-annotated corpus of Russian. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2017*, 16, pages 194–204.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LexConn: a French lexicon of discourse connectives. *Revue Discours*, 10.

Magdalena Rysova and Katerina Rysova. 2014. The centre and periphery of discourse connectives. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 452–459.

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

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.

Radu Soricut and Daniel Marcu. 2003. 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.

Rajen Subba and Barbara Di Eugenio. 2007. Automatic discourse segmentation using neural networks. In *Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, pages 189–190.

Pavlína Synková, Magdaléna Rysová, Lucie Poláková, and Jiří Mírovskỳ. 2017. Extracting a lexicon of discourse connectives in czech from an annotated corpus. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 232–240.

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relation markers in Russian RST treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33.

Svetlana Toldova, Dina Pisarevskaya, and Maria Kobozeva. 2018. Automatic mining of discourse connectives for Russian. volume 930, pages 79–87.

Sheng Xu, Peifeng Li, Guodong Zhou, and Qiaoming Zhu. 2018. Employing text matching network to recognise nuclearity in chinese discourse. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 525–535.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. 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.