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

Federal Research Centre Computer Science and Control Russian Academy of Sciences
Skoltech Skolkovo Institute of Science and Technology
National Research University Higher School of Economics

Corpus

The goal is the development of a data-driven system for automated rhetorical parsing of Russian texts. For training, we use the recently published **Ru-RSTreebank** annotated corpus https://rstreebank.ru/eng.

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).

Framework: the Rhetorical Structure Theory. **EDUs**: finite clauses, prepositional phrases, adverbial phrases headed by corresponding connectives (cf. because of, in spite of). 17 rhetorical **relation types**.

Annotation tool: rstWeb (https://corpling.uis.georgetown.edu/rstweb/info/).

Inter-annotator agreement:
Krippendorff's unitized Alpha is 81%.

Types of annotations in the corpus: segmentation of EDUs, discourse units nuclearity, types of discourse relations, rhetorical tree construction.

Lexicon of primary and secondary discourse connectives, based on this corpus and other lexicons – nearly 450 items (cue phrases are informative features for rhetorical relation classification).

Parsing Pipeline

5 subtasks: sentence segmentation, relation prediction, discourse tree construction, classification of connected DU pairs into nuclear-satellite, labeling relations between DUs.

Sentence segmentation: with external rule-based tools such as AOT.ru.

Relation prediction: 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 c**onstruction of the connected discourse tree**, we adopt an algorithm from (Hernault et al., 2010) that greedily merges DUs according to probabilities obtained from binary classification on the previous step.

Classification of connected DU pairs into nuclear-satellite: three-label classification task: "Satellite-Nucleus" (SN), "Nucleus-Satellite" (NS), "Nucleus-Nucleus" (NN).

Labeling relations between DUs: using the results from the 4th step, we predict a label of DU relations - as a multi-label classification task. We select 11 most important relations (for which the dataset contains at least 320 examples).
Feature selection: gradient boosting on decision trees (GBT) + logistic regression with L1 regularizer.

Feature Importance

Features: combinations of various features. 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): occurrences of cue phrases at the beginning and at the end of first and second DUs, 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): 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 DUs.
=> Most of the important features are **related to discourse connectives**. The common reason behind relation classification mistakes is the usage of connectives:

Next steps. The ensemble of CatBoost model with selected features and a linear SVM model provides the best results for relation classification. In the future work, we will develop the pipeline further, apply an extended version of discourse connectives lexicon, as well as implement more complex deep learning methods.

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.

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.

References

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

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), pages 243–281.

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.

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.

The experiments code is available at: http://nlp.isa.ru/paper_dialog2019/

Contacts

Elena Chistova (Russia, FRC CSC RAS; RUDN University chistova@isa.ru)

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

Artem Shelmanov (Russia, Skoltech; FRC CSC RAS a.shelmanov@skoltech.ru)

Ivan Smirnov (Russia, FRC CSC RAS ivs@isa.ru)

Svetlana Toldova (Russia, NRU Higher School of Economics toldova@yandex.ru)

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