Дискурсивный анализ текстов на русском языке

Елена Чистова
ФИЦ ИУ РАН

14 ноября 2019

Теоретические основы риторического анализа текстов на естественном языке

Понятие дискурсивного анализа

Дискурсивная структура текста отражает его связность на уровне выше уровня предложения

Анализ дискурса используется при решении многих задач NLP:

- генерация текстов [Qin et al., 2015]
- анализ тональности [Somasundaran, 2010]
- резюмирование текста [Hirao et al., 2013]
- вопросно-ответный поиск [Galitsky and Ilvovsky, 2017]

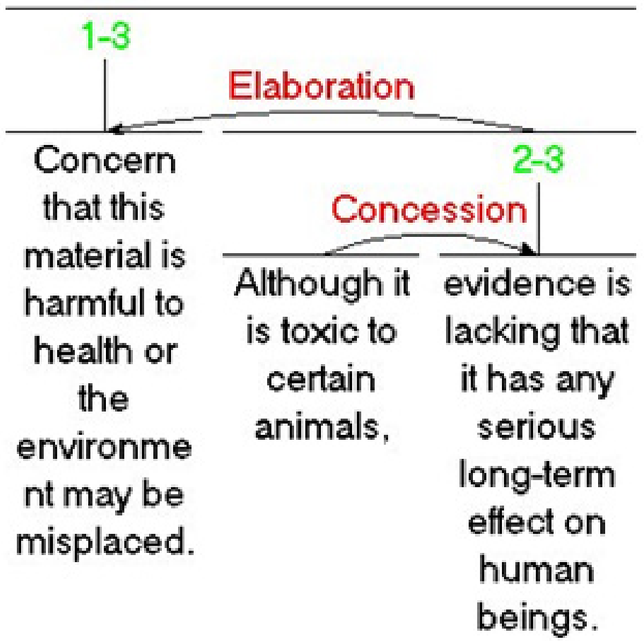
Формальные теории описания дискурсивных структур

- Разметка локальных дискурсивных структур
  - Connective-led annotation (Penn Discourse Treebank (PDTB), TED-Multilingual Discourse Bank (TED-MDB)),
  - Punctuation-led annotation (Chinese Discourse Treebank)
- Недревесные структуры
  - Cohesive relations (Discourse Graphbank)
- Древесные структуры
  - Segment-led annotation (Rhetorical Structure Theory) RST-DT (2003): 78 relations

Теория риторических структур (RST)

William C. Mann & Sandra A. Thompson http://www.sfu.ca/rst/

- Единицы дискурса
  элементарные дискурсивные единицы (EDU),
  дискурсивные единицы (DU)
- Нуклеарность
  ядро - центральная часть риторического отношения
  сателлит - второстепенная часть
- Отношения между дискурсивными единицами
- Иерархия
  дискурсивные единицы входят в отношения более высоких уровней до образования единого дерева



Пример структуры RST
[Mann and Thompson, 1987]

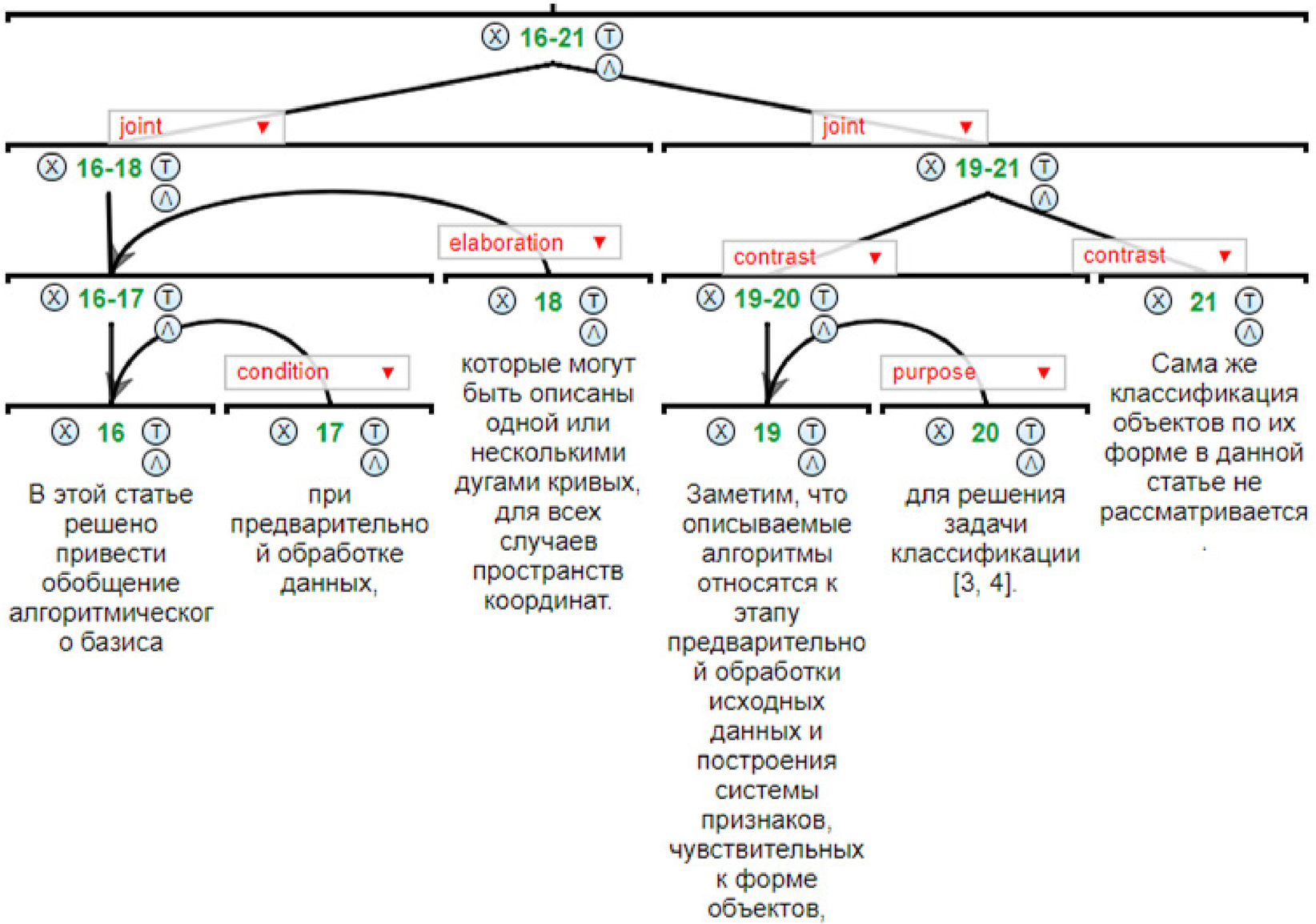
RST-корпусы для разных языков

- English: RST Discourse Treebank [Carlson et al., 2003], 385 новостных статей, 176 383 токенов
- German: Potsdam Commentary Corpus [Stede, Neumann, 2014], 2 900 предложений 175 новостных статей, 32 000 токенов
- Portuguese: CorpusTCC [Pardo et al., 2004], 1 350 предложений из 100 научных текстов, 53 000 tokens
- Spanish: RST Spanish Treebank [da Cunha et al., 2011], 2 256 предложений из 267 документов нескольких жанров
- Другие языки: Dutch: Dutch RUG Corpus [Redeker et al., 2012], Basque: RST Basque Treebank [Iruskieta et al., 2013], Chinese and Spanish-Chinese/Spanish Treebank as a parallel corpus [Cao et al., 2016], etc.

Russian RST-Treebank

- Ru-RSTreebank [Pisarevskaya et al., 2017] - в свободном доступе https://rstreebank.ru/
- 179 текстов; 203287 токенов
- новости, новостная аналитика и научные статьи (лингвистика, компьютерные науки)
- EDU: финитные клаузы (всегда), придаточные предложения, причастные и деепричастные обороты, предложные группы (в зависимости от семантики)
- инструмент разметки: rstWeb (https://corpling.uis.georgetown.edu/rstweb/info/)

Russian RST-Treebank - пример разметки



Методы риторического анализа текстов на русском языке

Пример анализируемого текста

Первый из них - метод статистической оценки, который использует заранее прописанные оценки для каждой из карт. Он прост в реализации, хотя и не быстр - потребуется как минимум поверхностный анализ удельной стоимости карт друг относительно друга.

Этапы дискурсивного анализа текста (RST)

Первый из них - метод статистической оценки,

который использует заранее прописанные оценки для каждой из карт.

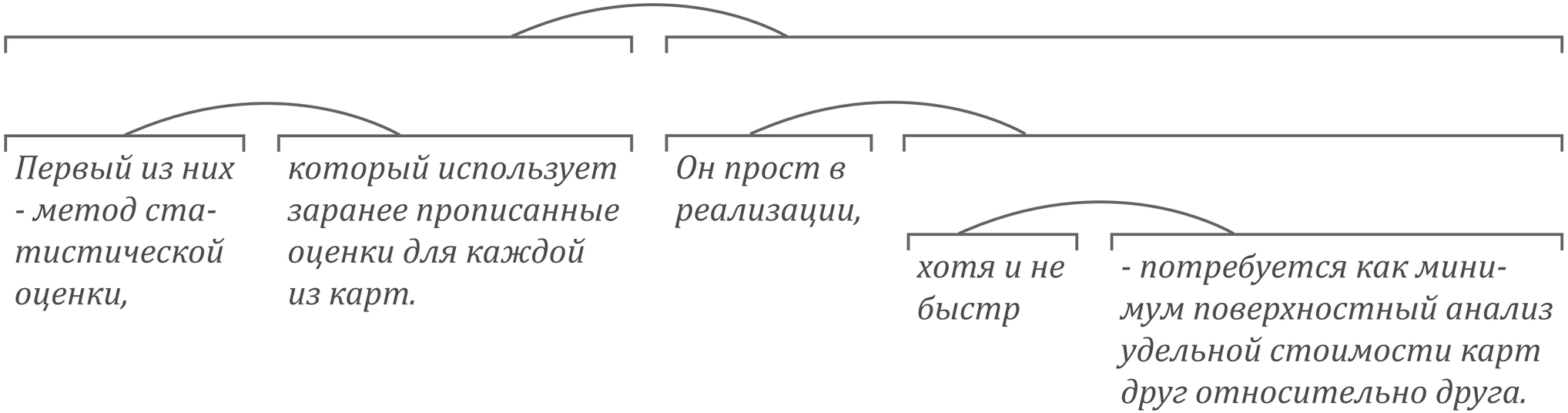
Он прост в реализации,

хотя и не быстр

- потребуется как минимум поверхностный анализ удельной стоимости карт друг относительно друга.

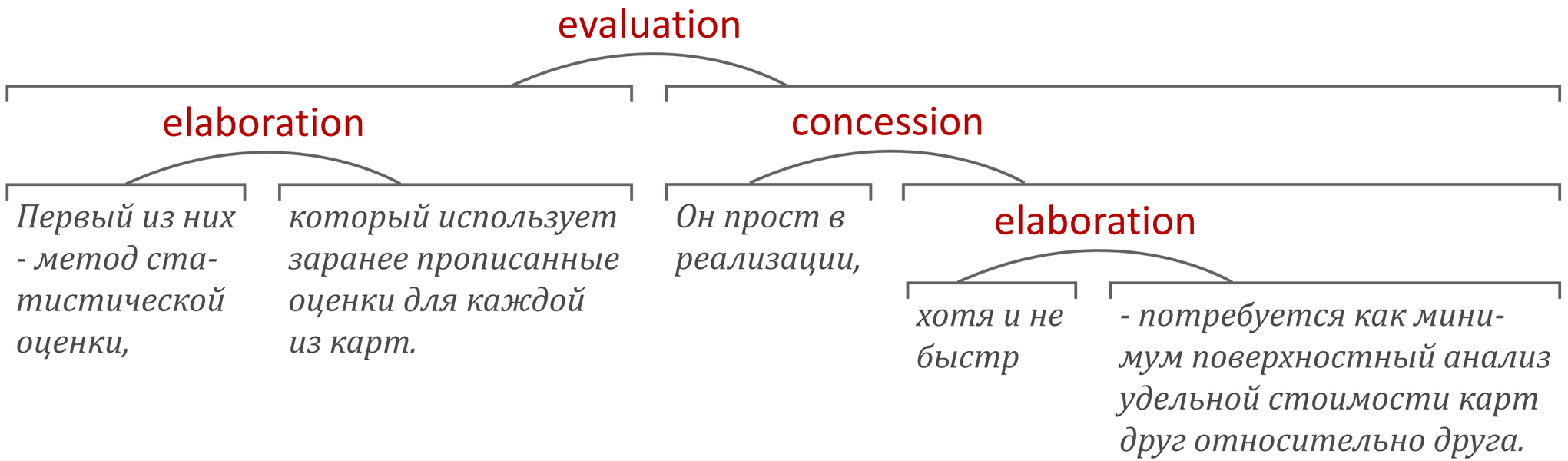
1. Сегментация текста

Этапы дискурсивного анализа текста (RST)



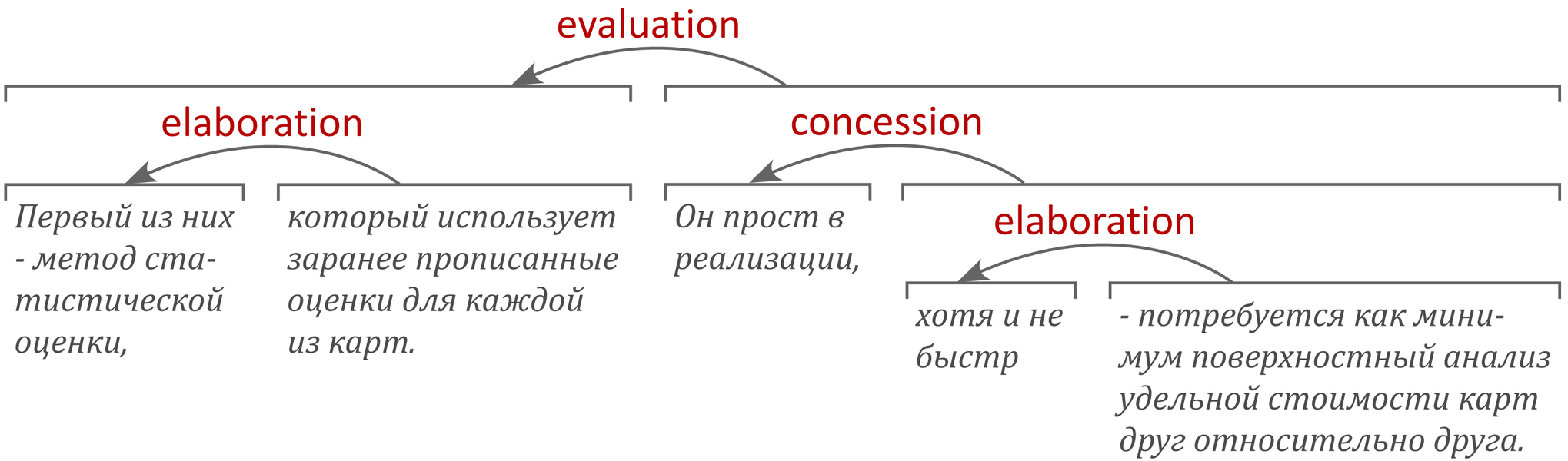
1. Сегментация текста
2. Построение дискурсивного дерева
    a. Обнаружение наличия отношения между двумя ДЕ

Этапы дискурсивного анализа текста (RST)



1. Сегментация текста
2. Построение дискурсивного дерева
   a. Обнаружение наличия отношения между двумя ДЕ
   b. Определение риторического типа отношения

Этапы дискурсивного анализа текста (RST)



1. Сегментация текста
2. Построение дискурсивного дерева
   a. Обнаружение наличия отношения между двумя ДЕ
   b. Определение риторического типа отношения
   c. Определение ядра

Дискурсивная сегментация

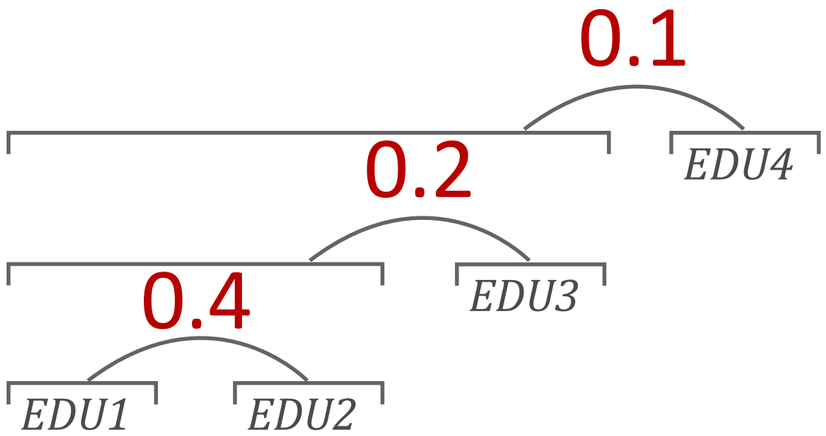
- Бинарная классификация [Soricut and Marcu, 2003]
- Маркировка последовательностей [Hernault et al., 2010]

input	corpus	P	R	F1
tok	deu.rst.pcc	94.88	94.49	94.68
	eng.rst.gum	92.28	82.89	87.33
	eng.rst.rstdt	93.6	93.27	93.43
	eng.sdrt.stac	87.56	80.78	83.99
	eus.rst.ert	87.43	80.94	84.06
	fra.sdrt.annodis	94.31	89.15	91.65
	nld.rst.nldt	94.81	89.97	92.32
	por.rst.cstn	93.04	90.72	91.86
	rus.rst.rrt	83.37	78.44	80.83
	spa.rst.rststb	89.11	90.09	89.6
	spa.rst.sctb	87.16	76.79	81.65
	zho.rst.sctb	66.26	64.29	65.26
	mean	88.65	84.32	86.39

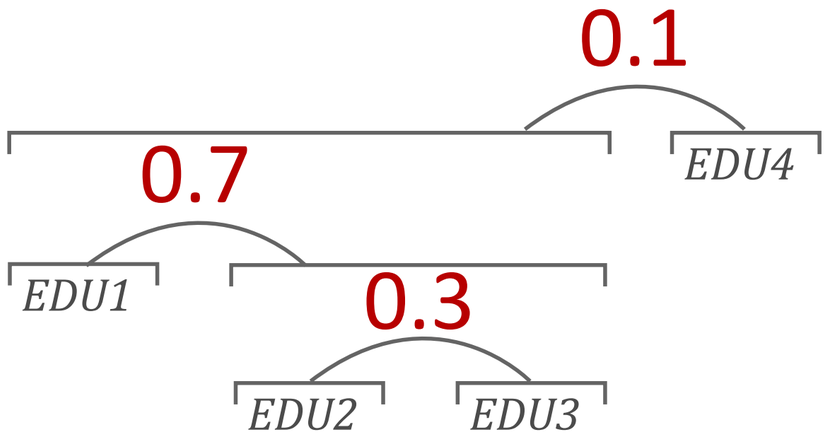
Final detailed scores on segmentation with multilingual BERT [Muller et al., 2019]

Построение дискурсивных деревьев

- Линейный bottom-up алгоритм [Duverle et al., 2009; Feng et al., 2014]
- CYK-подобные стратегии [Joty et al., 2013]



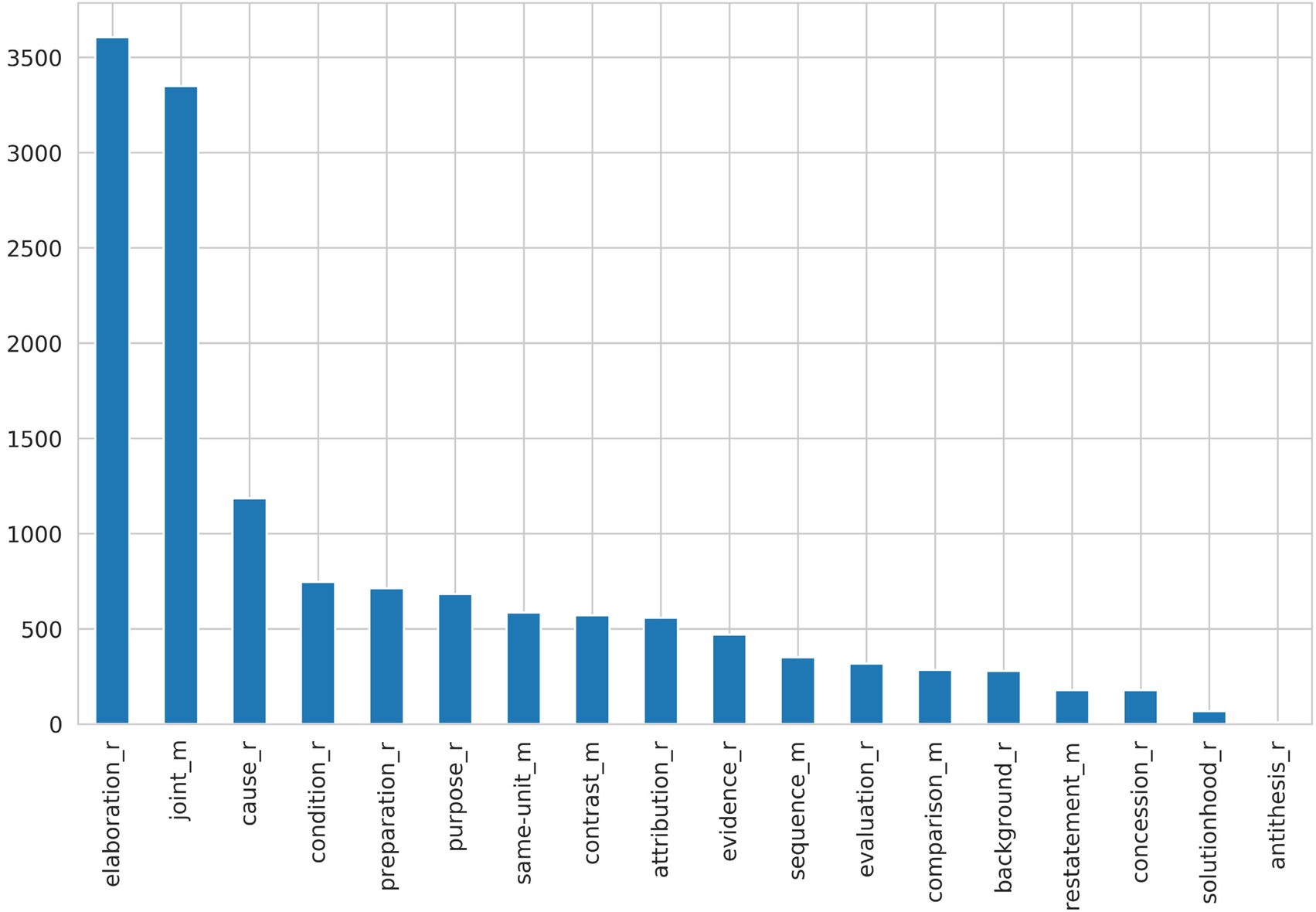
$$\sum_p = 0.7$$



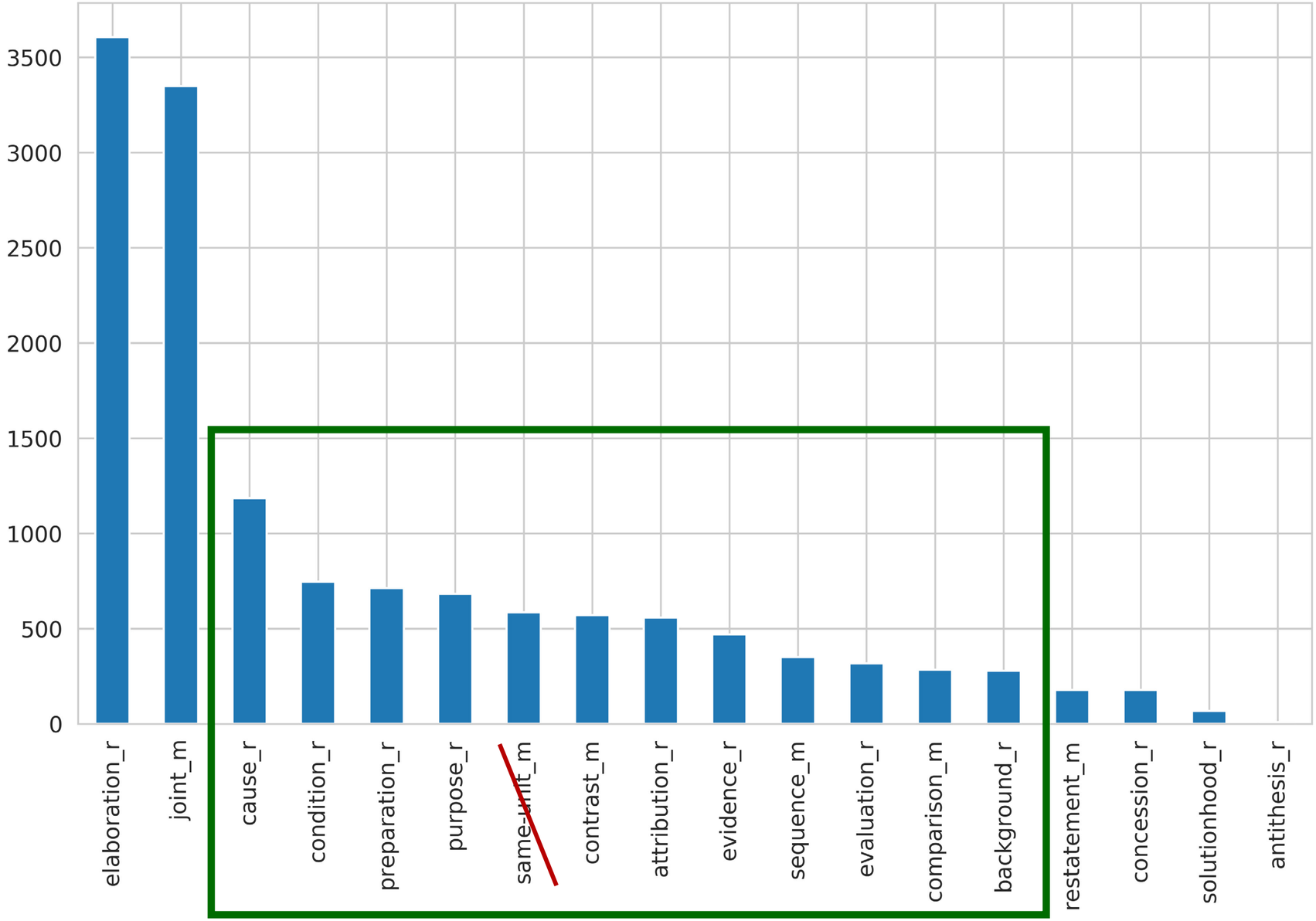
$$\sum_p = 1.1$$

Экспериментальные исследования моделей классификации риторических отношений

Распределение отношений в корпусе Ru-RSTreebank



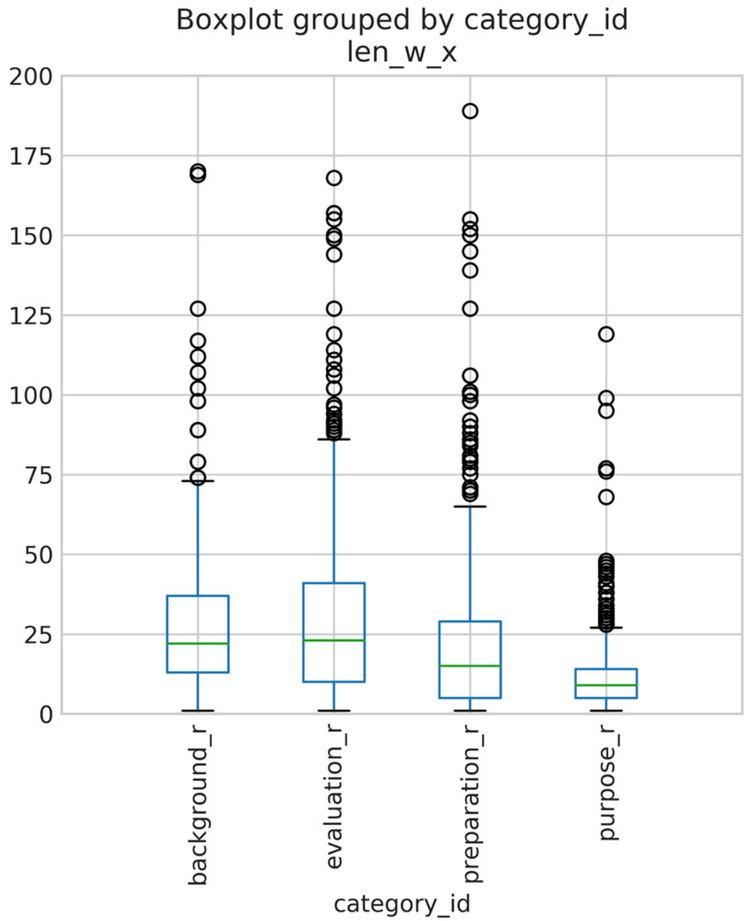
Распределение отношений в корпусе Ru-RSTreebank

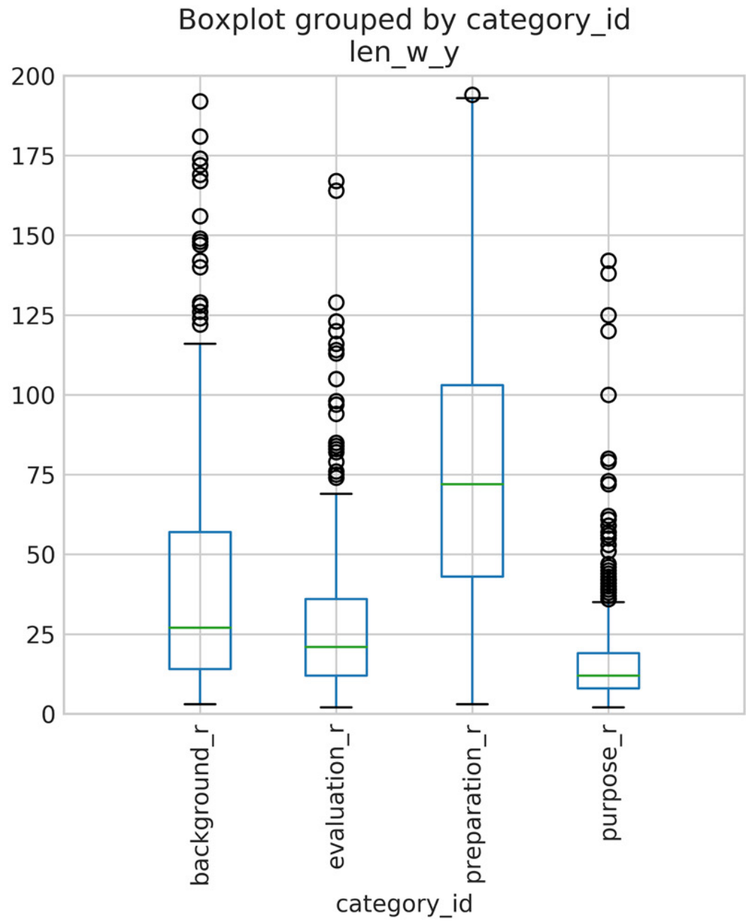


Формальные признаки

- Число слов
- Средняя длина слова
- Доля слов в верхнем регистре
- Доля слов, начинающихся с заглавной буквы
- Количество вхождений различных морфологических признаков

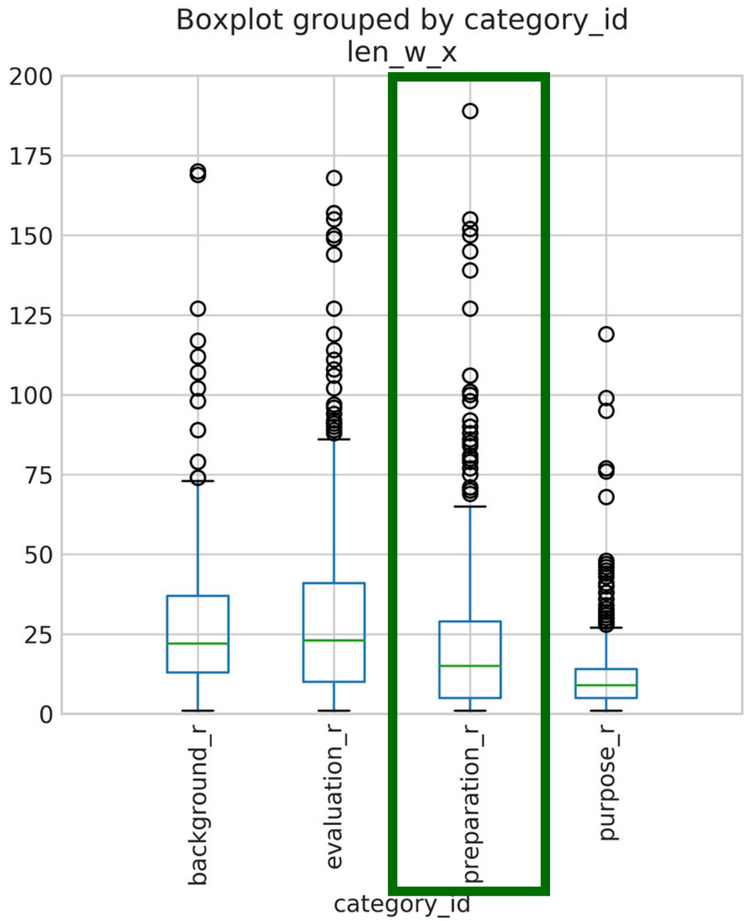
Формальные признаки

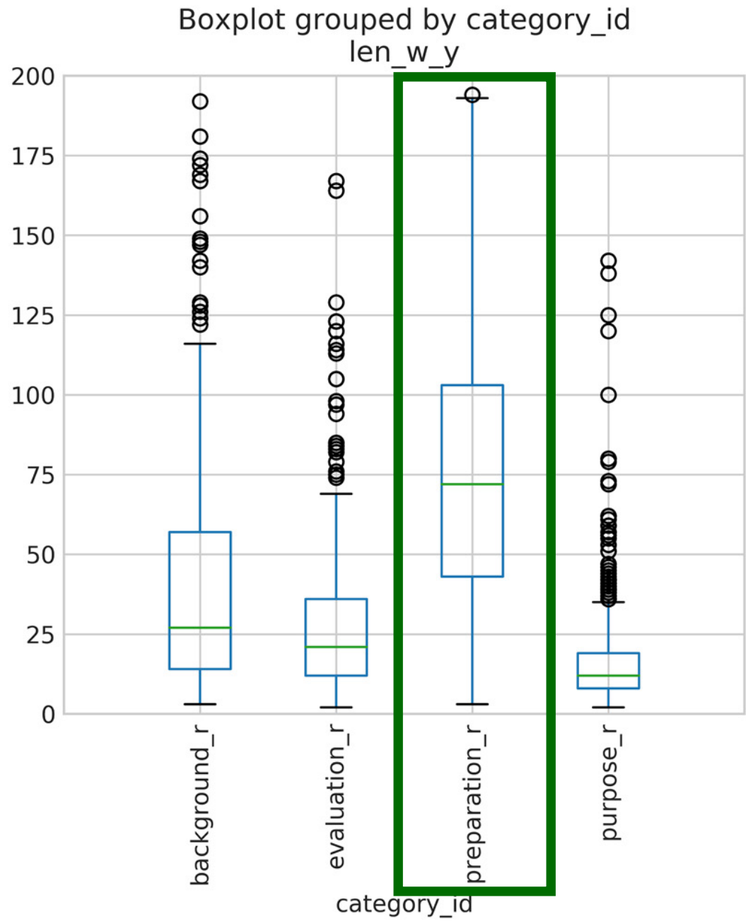




Длина первого (слева) и второго (справа) сегмента в словах в отношениях background, evaluation, preparation, purpose.

Формальные признаки





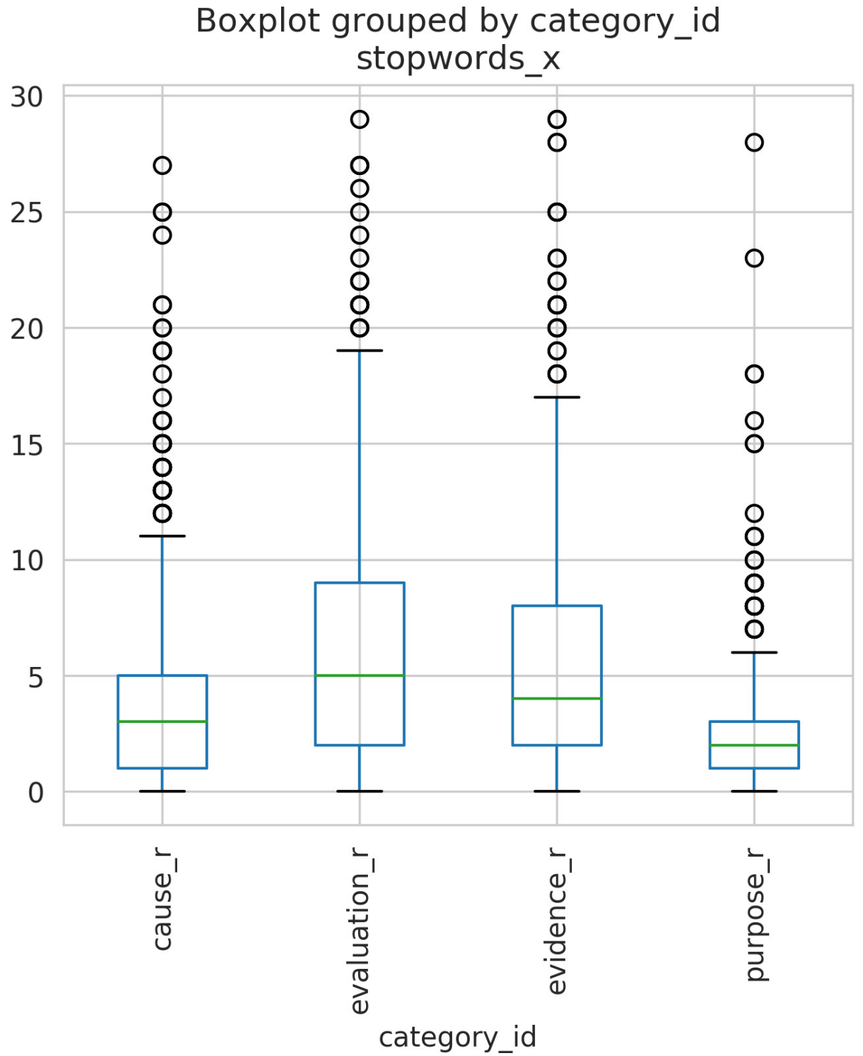
Пример:

[Активные в речи ребёнка ЛСП выявляются:]$_1$

[1) при непрерывном мониторинге (систематически регистрируются регулярно повторяющиеся в устной и письменной речи учащегося лексические единицы); 2) в ходе эксперимента (учитель …]$_2$

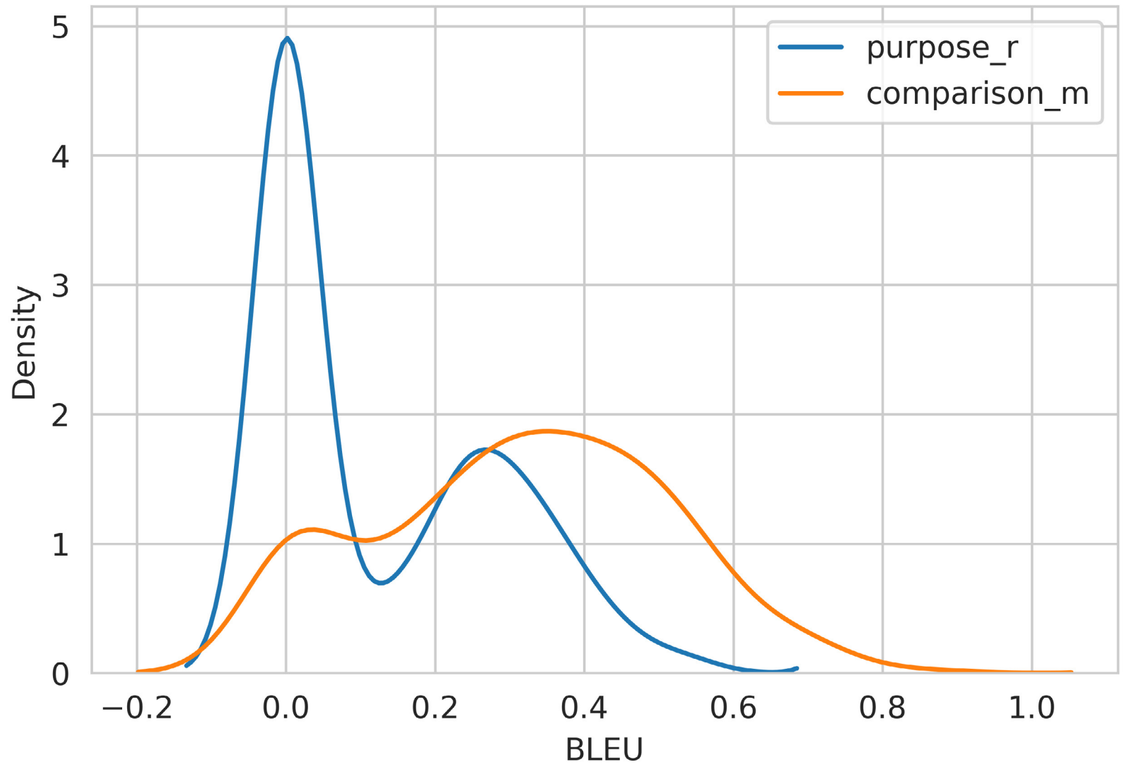
Признаки маркеров и стоп-слов

- Число вхождений стоп-слов
- Число вхождений дискурсивных маркеров
- POS-tag первой и последней пары каждой ДЕ
- Вхождение каждого маркера в начало и в конец каждой ДЕ (*итак*, *то есть*, *очевидно*, *отсюда*, *следовательно*, *в том*, *что* и т. д.)



Семантические признаки

- TF-IDF векторы каждой ДЕ
- усредненные эмбеддинги каждой ДЕ (word2vec)
- признаки близости векторов, образованных морфологическими признаками двух ДЕ: косинусная, Хемминга, Канберра, близость бинаризованных векторов
- косинусная близость векторов TF-IDF
- мера Жаккара между леммами ДЕ
- метрика BLEU



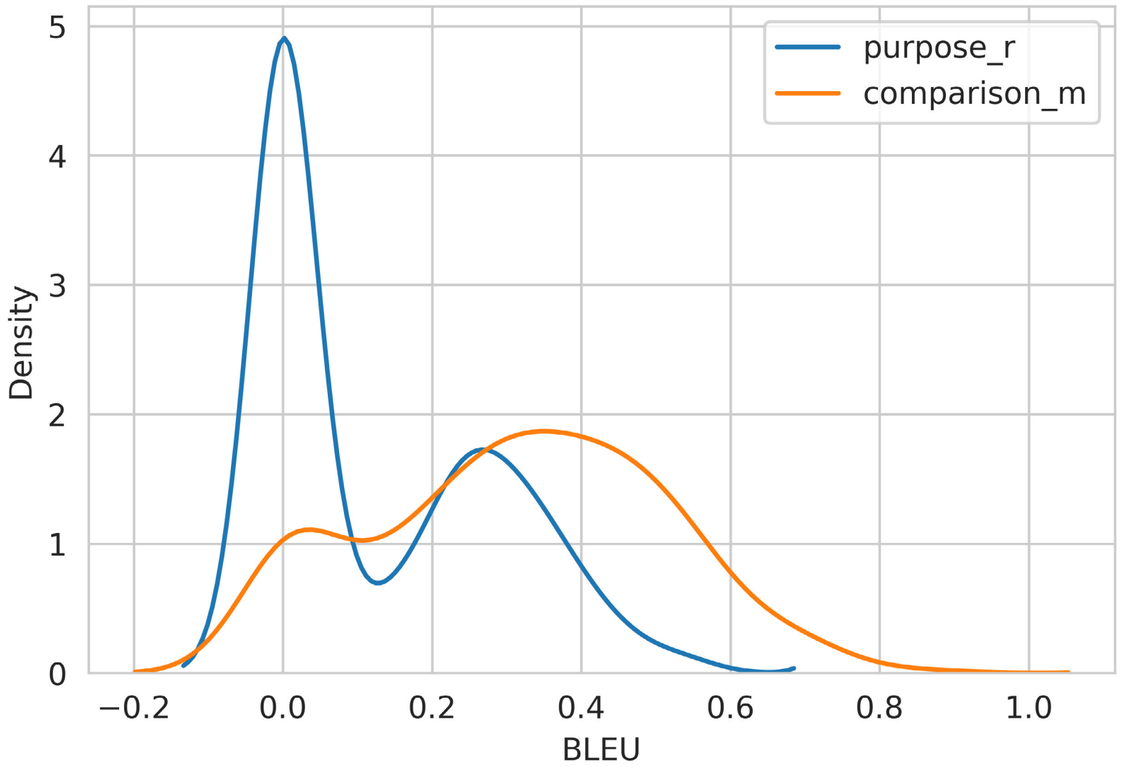
Семантические признаки

Пример:

[Следует отметить, что если известны матрицы А, В, Л, Н , то уравнение (3) решается относительно матрицы М ,]$_1$

[а если известны А,В, Л, М , то уравнение (3) решается относительно матрицы Н в форме [формула].]$_2$

$\mathrm{BLEU}(\mathrm{DU}_1, \mathrm{DU}_2) = 85.49\%$



Методы классификации

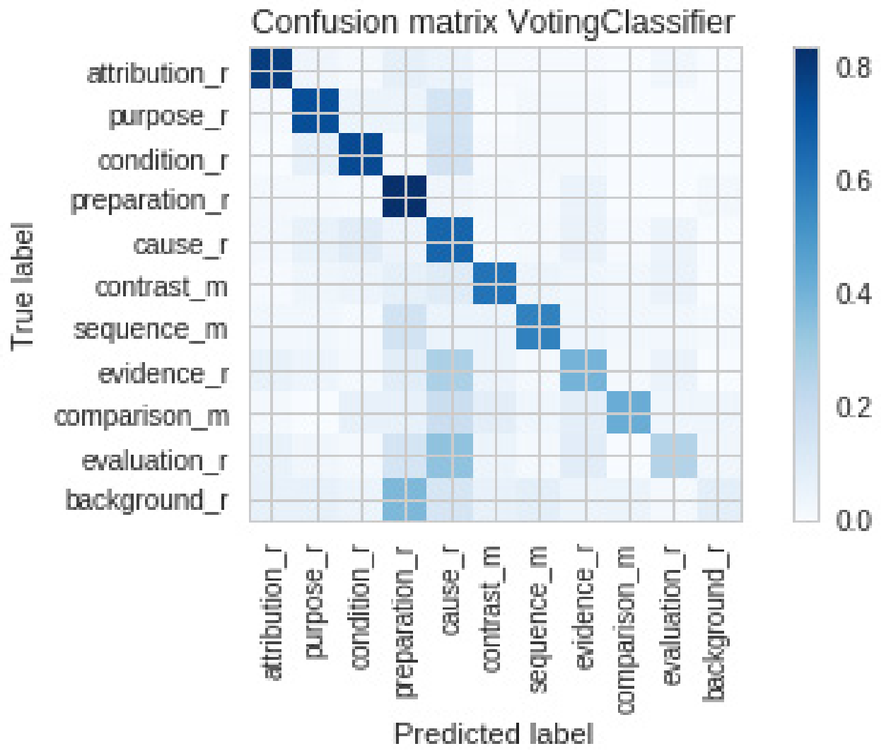
- Логистическая регрессия
- Двухслойная полносвязная нейронная сеть
- Метод опорных векторов
- Градиентный бустинг над решающими деревьями (GBT) (XGBoost, CatBoost)
- GBT на признаках, выделенных при помощи логистической регрессии с L1-регуляризацией
- Ансамбли линейных и GBT-моделей

Результаты: задача маркировки отношений

Классификатор	Macro F1		Micro F1	
	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
L1 Feature selection+LGBM	51.64	2.22	60.29	1.74
CatBoost	53.32	0.96	60.71	0.81
L1 Feature selection+CatBoost	53.45	2.19	61.09	1.96
voting((L1 Feature selection+LGBM), Linear SVM)	**54.67**	1.8	**62.39**	1.51
voting((L1 FS+CatBoost), Linear SVM)	**54.67**	0.38	62.32	0.41

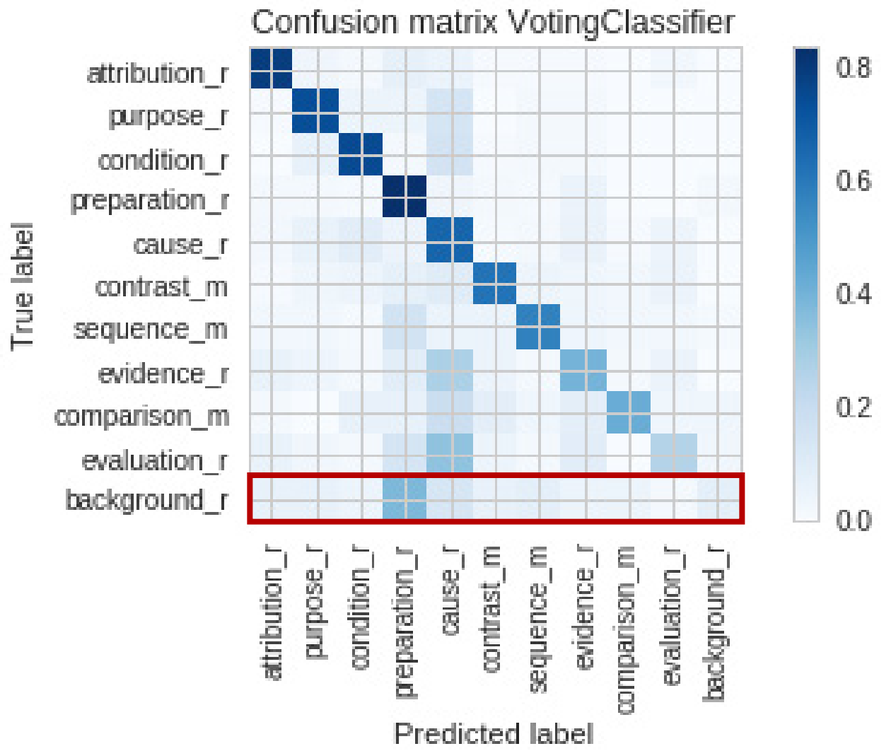
Результаты: задача маркировки отношений

Класс	Precision	Recall	F1
attribution	73.11	75.77	74.41
purpose	71.87	73.71	72.70
condition	73.60	65.75	69.45
preparation	57.82	81.09	67.49
contrast	68.43	56.69	62.01
cause	51.73	69.96	59.46
sequence	54.46	54.55	54.50
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



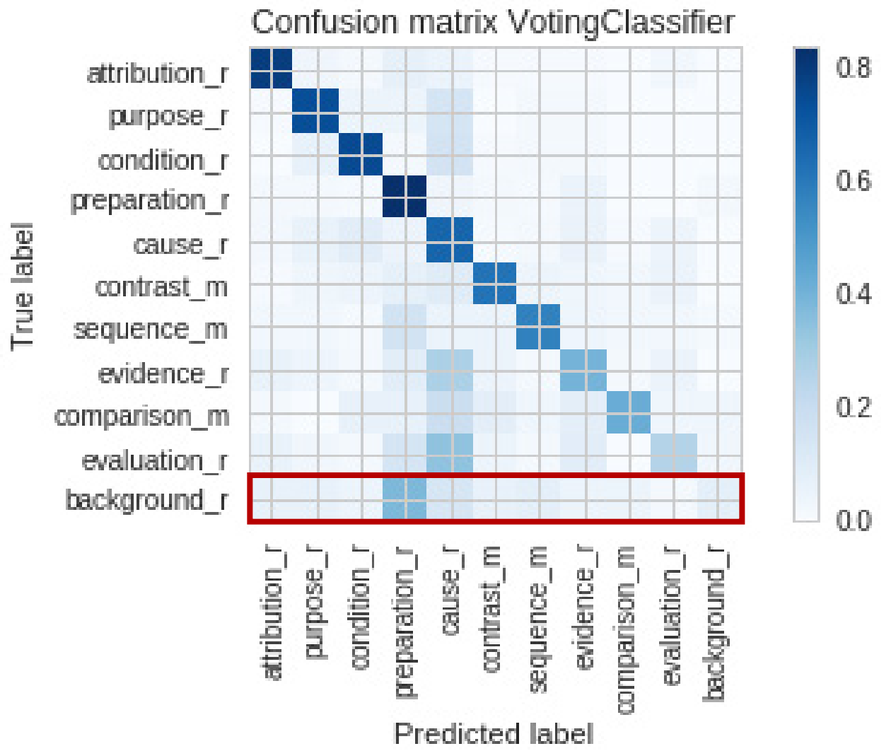
Результаты: задача маркировки отношений

Класс	Precision	Recall	F1
attribution	73.11	75.77	74.41
purpose	71.87	73.71	72.70
condition	73.60	65.75	69.45
preparation	57.82	81.09	67.49
contrast	68.43	56.69	62.01
cause	51.73	69.96	59.46
sequence	54.46	54.55	54.50
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



Результаты: задача маркировки отношений

Класс	Precision	Recall	F1
attribution	73.11	75.77	74.41
purpose	71.87	73.71	72.70
condition	73.60	65.75	69.45
preparation	57.82	81.09	67.49
contrast	68.43	56.69	62.01
cause	51.73	69.96	59.46
sequence	54.46	54.55	54.50
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



Результаты: задача маркировки отношений

Пример:

- *[Мир подводит итоги стодневного пребывания Барака Обамы в кресле президента США.]$_1$ [Американские эксперты и IT-журналисты приходят к выводу, что нынешний президент их страны — самый технологически продвинутый.]$_2$* - preparation
- *[Методов конкурентной разведки очень много.]$_1$ [На такие исследования иногда уходит месяц, иногда год и более.]$_2$* - background

Анализ значимости признаков (маркировка отношений)

Тип	Признаки	Кол-во признаков	Влияние на macro F1, %
Лексические	Элементы TF-IDF векторов	8	0.11
Морфологи-ческие	Парные комбинации различных частей речи в начале и конце ДЕ; Число вхождений существительных и глаголов различной морфологии в ДЕ	119	0.45
Текстовые	Число вхождений различных маркеров в каждую ДЕ; Вхождение маркеров в начало и конец каждой ДЕ	1887	**2.49**

Результаты: задача определения нуклеарности

Классификатор	Macro F1		Micro F1	
	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
CatBoost	67.82	0.45	69.17	0.73
L1 Feature selection+CatBoost	**68.82**	0.84	**70.31**	0.76

Результаты

- Разработаны методы классификации для RST-анализа текстов на русском языке:
  - Для определения риторического типа отношения
  - Для определения нуклеарности отношения
- Произведен анализ признаков
- Опубликован код экспериментов: http://nlp.isa.ru/paper_dialog2019/
- E. Chistova, A. Shelmanov, M. Kobozeva, D. Pisarevskaya, I. Smirnov, S. Toldova. Classification Models for RST Discourse Parsing of Texts In Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”, May 29-June 1, Moscow, Russia. - 2019. - Pages 163-176.
- Elena Chistova, Maria Kobozeva, Dina Pisarevskaya, Artem Shelmanov, Ivan Smirnov, Svetlana Toldova. Towards the Data-driven System for Rhetorical Parsing of Russian Texts // Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, June 6, Minneapolis, MN, USA. - 2019. - Pages 82-87.

Планируемые исследования

- Применение более продвинутых нейронных архитектур для дискурсивной сегментации
- Разработка стратегии для построения деревьев
- Оценка качества всего пайплайна
- Эксперименты с совмещением нескольких подзадач в одной архитектуре

Планируемые исследования

- Применение более продвинутых нейронных архитектур для дискурсивной сегментации
- Разработка стратегии для построения деревьев
- Оценка качества всего пайплайна
- Эксперименты с совмещением нескольких подзадач в одной архитектуре

Источники

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

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.

Galitsky, Boris A., and Dmitry I. Ilvovsky. «On a Chat Bot Finding Answers with Optimal Rhetoric Representation.» In RANLP, pp. 253-259. 2017.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue(pp. 85-112). Springer, Dordrecht.

Pardo T. A. S., Nunes M. G. V., Rino L. H. M. (2004), Dizer: An automatic discourse analyzer for brazilian portuguese, Brazilian Symposium on Artificial Intelligence, Springer Berlin Heidelberg, pp. 224–234.

Da Cunha, I., Torres-Moreno, J. M., & Sierra, G. (2011, June). On the development of the RST Spanish Treebank. In Proceedings of the 5th Linguistic Annotation Workshop (pp. 1-10).

Redeker, G., Berzlánovich, I., Van Der Vliet, N., Bouma, G., & Egg, M. (2012). Multi-layer discourse annotation of a Dutch text corpus. age, 1, 2

Источники

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., & de Lacalle, O. L. (2013). The RST Basque TreeBank: an online search interface to check rhetorical relations. In 4th workshop RST and discourse studies (pp. 40-49).

Cao, S., da Cunha, I., & Bel, N. (2016). An analysis of the Concession relation based on the discourse marker aunque in a Spanish-Chinese parallel corpus. Procesamiento del Lenguaje Natural, (56), 81-88.

Pisarevskaya, Dina, et al. (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.

Soricut R.,Marcu D.Sentence level discourse parsing using syntactic andlexical information // Proceedings of the 2003 Conference of the North Amer-ican Chapter of the Association for Computational Linguistics on HumanLanguage Technology-Volume 1. — Association for Computational Linguis-tics. 2003. — Pp. 149–156.

Hernault H.,Bollegala D.,Ishizuka M.A sequential model for discourse segmentation // International Conference on Intelligent Text Processing andComputational Linguistics. — Springer. 2010. — Pp. 315–326.

Muller, Philippe, Chloé Braud, and Mathieu Morey. «ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents.» Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019. 2019.

Duverle D. A.,Prendinger H.A novel discourse parser based on supportvector machine classification // Proceedings of the Joint Conference of the47th Annual Meeting of the ACL and the 4th International Joint Conferenceon Natural Language Processing of the AFNLP: Volume 2-Volume 2. —Association for Computational Linguistics. 2009. — Pp. 665–673.

Источники

Combining intra-and multi-sentential rhetorical parsing for document-leveldiscourse analysis / S. Joty [et al.] // Proceedings of the 51st Annual Meet-ing of the Association for Computational Linguistics (Volume 1: Long Pa-pers). — 2013. — Pp. 486–496.

Feng V. W.,Hirst G.A linear-time bottom-up discourse parser with con-straints and post-editing // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). —2014. — Pp. 511–521.