Методы извлечения имплицитных дискурсивных отношений

Анна Полянская, НИУ ВШЭ, 2020

Научный руководитель: Светлана Юрьевна Толдова, к.ф.н., доцент ШЛ НИУ ВШЭ

Объект: пары текстовых представлений (элементарных) дискурсивных единиц.

Предмет: способы определения вида дискурсивного отношения.

Цель: разработка автоматической системы с максимальным качеством предсказаний вида дискурсивного отношения.

Инструменты: NLP и ML библиотеки на Python.

Гипотеза: важность признаков, извлеченных из текста, при автоматической классификации будет соответствовать лингвистическому ожиданию об особенностях отдельных дискурсивных отношений и способов их выражения.

Код и данные лежат на https://github.com/polyankaglade/ImpDiscRel

ЭДЕ – минимальный квант дискурса: совокупность информации, которую селективное человеческое сознание может одновременно удерживать в активном состоянии, одна предикация/клауза, описание одного события/состояния, одна произносительная конфигурация [Кибрик & Подлесская 2009: 57].

ДО – отношения между ЭДЕ, связывающие их в единый текст [Mann & Thompson 1988].

Маркер – лексические средства выражения ДО: союзы, союзные слова [https://rstreebank.ru/markersFull]

Имплицитное ДО – между ДЕ есть отношение, но нет специального маркера, его выражающего.

Таргетные отношения:

Contrast – ситуации, содержащиеся в дискурсивных единицах, противопоставлены друг другу, контрастируют относительно некоторой заданной темы (маркеры: *но, а, хотя, несмотря на, однако.* [Соколова & Толдова 2019]).

Cause-effect – ситуация сателлита является причиной ситуации ядра, ядро представляет собой результат действия (маркеры: *потому что, поэтому, из-за чего, по причине, вследствие*).

Данные

Корпус RSTreebank: 333 текста (около 328 000 словоупотреблений) с XML разметкой (границы ЭДЕ и их групп, тип отношения, ядро, сателлит). Жанры: новостные тексты, научно-популярные тексты, научные статьи и тексты блогов [Pisarevskaya et al. 2017].

Итоговая выборка: по 100 пар ЭДЕ на каждое отношение из текстов блогов.

Обработка текстовых данных:

- Исправление ошибок и опечаток
- Нормализация пунктуации (знак между ЭДЕ → к I из них)
- Токенизация с помощью библиотеки Razdel
- POS разметка и лемматизация с помощью библиотеки pyMorphy2
- Векторизация с помощью алгоритма doc2vec из библиотеки gensim

```
А я, между прочим, совершенно спокойна_у меня ж все нуж
ные телефоны записаны, я знаю куда бежать и что делать
                                ↓
а я , между прочим , совершенно спокойна
у меня ж все нужные телефоны записаны , я знаю куда беж
ать и что делать
                                ↓
а я между прочить совершенно спокойный
у я ж весь нужный телефон записать я знать куда бежать
и что делать
```

Признаки

В каждой изученной работе их набор отличается, иногда одинаковые по смыслу, но разные по реализации. Можно сделать следующую классификацию, подчеркнутые = использованные в данной работе:

1) «Технические» признаки – формальные параметры текста
    a. Длина ЭДЕ
    b. Разница в длине первой и второй ЭДЕ
    c. Средняя длина слова
    d. Стоит ли ЭДЕ в начале абзаца
    e. Есть ли между ЭДЕ знак препинания
2) Морфологические признаки
    a. Частеречное представление ЭДЕ
    b. Часть речи вершины ЭДЕ
    c. Время клауз
3) Синтаксические признаки
    a. Кол-во аргументов предиката-вершины
    b. Повторяющиеся синтаксические структуры
4) Лексические признаки
    a. Отдельные слова (обычно как раз маркеры)
    b. Повторяющиеся в обеих ЭДЕ слова
    c. Лексический класс предиката-вершины
    d. Наличие модальных глаголов
    e. Entities Types ядра
    f. Полярность
    g. First-Last-First3
5) Семантические признаки
    a. Эмбединги
    b. Семантическая близость ЭДЕ
    c. Топики

train – 160x44, test – 40x44, с сохранением баланса классов 1:1

Модели

1) Логистическая регрессия с L2 регуляризацией
2) Решающее дерево
3) Случайный лес с 10 деревьями
4) Метод опорных векторов

Метрики качества

В качестве baseline используются результаты лучшей модели из [Chistova et al. 2019], где производилась 11-way классификация по 320+ вхождений на каждый, и среди 3,273 признаков было 350 дискурсивных маркеров.

F1-score	**LReg**	**DTree**	**RForest**	**SVM**	**baseline**
cause	0.723404	0.6250	0.652174	**0.744186**	0.5946
contrast	0.606061	0.4375	0.529412	**0.702703**	0.5669

Precision	**LReg**	**DTree**	**RForest**	**SVM**	**baseline**
cause	0.629630	0.535714	0.576923	**0.695652**	0.5173
contrast	**0.769231**	0.583333	0.642857	0.764706	0.6843

Recall	**LReg**	**DTree**	**RForest**	**SVM**	**baseline**
cause	**0.85**	0.75	0.75	0.80	0.7
contrast	0.50	0.35	0.45	**0.65**	0.57

Тенденции в полученных результатах сопоставимы с бейзлайном, однако абсолютные значения метрик стоит сравнивать с осторожностью, т. к. объем выборки и количество признаков существенно различаются.

Интерпретация признаков

Наиболее важными признаками оказываются:

Для контраста:

1) Количество имен прилагательных
2) Количество междометий
3) Количество одинаковых слов в двух ЭДЕ
4) Количество частиц

Для причины-следствия:

1) Кол-во причастий
2) Кол-во деепричастий
3) Кол-во компаративов
4) Кол-во слов во второй ЭДЕ

Для моделей на основе решающих деревьев:

1) Косинусная близость
2) Длины ЭДЕ
3) Средние длины слов
4) Количество одинаковых слов в двух ЭДЕ

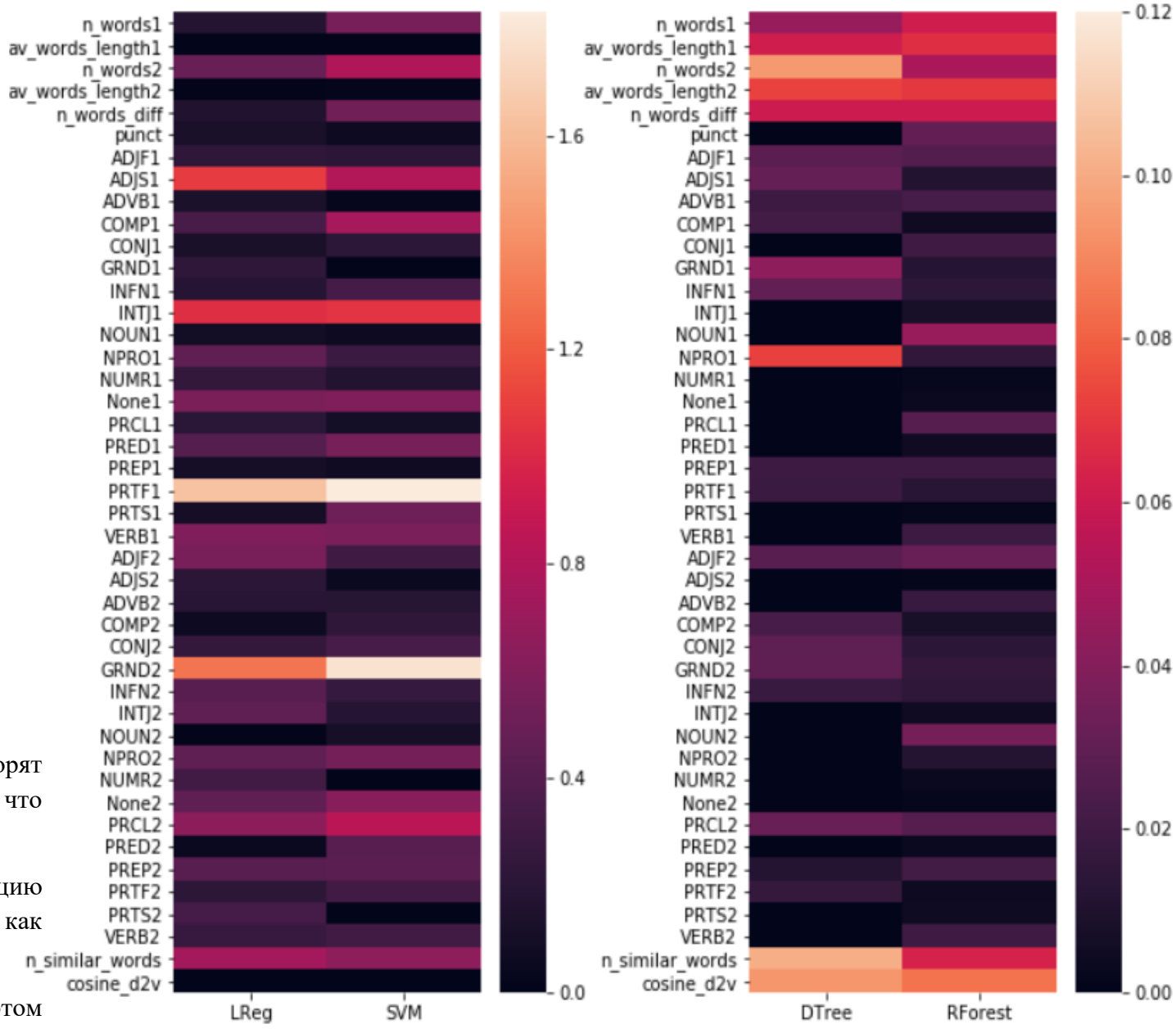
Подтверждение лингвистического ожидания

Повторяющиеся слова и косинусная близость говорят о семантическом параллелизме между двумя ЭДЕ, что характерно для отношений контраста.

Противопоставительные союзы имеет тенденцию семантически сближаться с междометиями, как отмечается в [Соколова & Толдова 2019: 131].

Конструкциям с деепричастным оборотом действительно свойственно выражать причинно-следственные отношения [Евтюхин 1996: 159].

Heatmap весов признаков в моделях



Примечания

Размерность векторов была выбрана равной 10, исходя из распределения длин ЭДЕ, и каждое измерение стало отдельным признаком, однако в дальнейшем эти десять признаков не использовались (без особого изменения в качестве), потому что почти все модели давали части из них большие веса, а интерпретировать такие результаты невозможно.

Особой корреляции между признаками не наблюдалось, хотя и были объяснимые зависимости, например связь между количеством существительным и предлогов и между количеством существительных и общим количеством слов.

Отношения contrast и cause-effect были выбраны по следующим соображениям:

1) имплицитных случаев гораздо меньше, чем эксплицитных, и именно этих двух этих отношений практически больше всего в доступных данных, что позволяет получить достаточный объем итоговой выборки.

span	11570	preparation	779
joint	8086	evaluation	734
elaboration	2873	comparison	591
contrast	2171	background	385
antithesis	1675	restatement	352
attribution	1611	concession	258
condition	1359	evidence	244
cause	1293	cause-effect	201
sequence	1019	solutionhood	126
same-unit	1002	interpretation-evaluation	87
purpose	955	effect	29

2) эти отношения достаточно сильно семантически нагружены – даже когда нет маркера, человек может их определить относительно легко
3) природа этих отношений такова, что ЭДЕ будут связаны между собой по смыслу чуть больше, чем, например, последовательность, конъюнкция или детализация.

Использованная литература

Евтюхин 1996 – В. Б. Евтюхин. Группировка полей обусловленность: причина, условие, цель, следствие, уступка // А.В. Бондарко (ред.). *Теория функциональной грамматики. Локативность. Бытийность. Посессивность. Обусловленности*. СПб.: Наука, 1996. С. 138-160.

Кибрик, Подлесская 2009 – А. А. Кибрик, В. И. Подлесская. *Рассказы о сновидениях: корпусное исследование устного русского дискурса.* М.: Языки славянских культур, 2009.

Соколова, Толдова 2019 – Е. Г. Соколова, С. Ю. Толдова. Особые свойства риторических отношений “Контраст” и “Сравнение” на материале разметки в корпусе Ru-RSTreebank // *Труды международной научной конференции «Корпусная лингвистика-2019».* Санкт-Петербург, Россия, 24-28 июня 2019 г. СПб.: Издательство Санкт-Петербургского государственного университета, 2019.

Chistova et al. 2019 – E. V. Chistova, A. O. Shelmanov, M. V. Kobozeva, D. B. Pisarevskaya, I. V. Smirnov, S. Yu. Toldova. Classification models for RST discourse parsing of texts in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”, Moscow, May 29—June 1, 2019.

Mann & Thompson 1988 – W. Mann, S. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization // *Text-interdisciplinary Journal for the Study of Discourse* 8 (3), 1988. P. 243–281.

Pisarevskaya D. et al. 2017 – D. Pisarevskaya et al. Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017”, Moscow, Russia, May 31— June 3, 2017. P. 194-204