ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Дахина Александра Сергеевна

РАСПОЗНАВАНИЕ РИТОРИЧЕСКИХ ОТНОШЕНИЙ КОНТРАСТА И СРАВНЕНИЯ В РУССКОМ ЯЗЫКЕ

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ152

Академический руководитель
образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

_____

« 04 » ___июня___2019г.

Научный руководитель

К.Ф.Н., Доцент
С.Ю. Толдова

_____

Научный консультант

_____

Москва 2019

Оглавление

Аннотация

Данная работа посвящена исследованию риторических отношений контраста и сравнения в русском языке. Исследование основано на данных Ru-RSTreebank, размеченного в соответствии с теорией Rhetorical Structure Theory (RST) (Mann, Thompson, 1988). Основная задача этой работы выявить значимые признаки, отличающие риторические отношения контраста от отношения сравнения. В данном исследовании рассмотрено влияние различных аспектов на распознавание отношений Contrast и Comparison. Особое внимание в этой работе уделено лексическим и синтаксическим свойствам этих двух отношений. Также, исследуется влияние векторной близости различных частей элементарных дискурсивных единиц (ЭДЕ) с аналогичными частями ЭДЕ, связанными одним риторическим отношением, на распознавание отношений контраста и сравнения. В результате исследования было рассмотрено 13 признаков, из которых только 11 оказались статистически значимыми. Именно из этих 11 признаков и выбирались те, что наиболее характеризуют контраст или сравнение, чтобы можно было автоматически классифицировать эти отношения. Мы провели несколько экспериментов, перебрав все возможные комбинации признаков и протестировав их влияние на классификацию с помощью алгоритма Random Forest. В результате исследования мы выделили три признака, вносящие наибольший вклад в распознавание отношений контраста и сравнения.

Annotation

This work is devoted to the study of rhetorical relations of contrast and comparison in the Russian language. The study is based on data from Ru-RSTreebank, labeled according to the theory of Rhetorical Structure Theory (RST) (Mann, Thompson, 1988). The main objective of this work is to identify significant features that distinguish rhetorical relations of contrast from the relation of comparison. In this paper, the influence of various aspects on the recognition of the relations between Contrast and Comparison is considered. Particular attention is paid to the lexical and syntactic properties of these two relations. Also, examines the impact of spatial proximity of different parts of elementary discourse units (EDU) with the same parts of EDU, related to a rhetorical attitude, to recognize relations of contrast and comparison. As a result of the study, 13 features were considered, of which only 11 were statistically significant. It is from these 11 features that those that most characterize contrast or comparison were chosen so that these relations can be automatically classified. We conducted several experiments, going through all possible combinations of features and tested their impact on the classification using the Random Forest algorithm. As a result of the study, we have identified three features that make the greatest contribution to the recognition of the relations of contrast and comparison.

1. Введение

Автоматический анализ дискурса востребован во многих отраслях обработки текста. Например, для выделения событий. Для представления дискурса существует несколько различных теорий. Их объединяет то, что все теории направлены на установление отношений между дискурсивными единицами, не зависимо от того, как они выражаются. Для некоторых отношений в качестве маркера связи используются распространённые союзы:

Contrast:
*(1) [Те существа, которые лучше приспособились, выживают,][**а** те, кто не приспособился, вымирают,[…]]*

Comparison:
*(2) [К примеру, BMW застраховал риски на североамериканском, то есть долларовом рынке на год вперед,][**а** другая германская компания, Porsche, - аж до 2007 года.]*

В некоторых случаях маркеры могут отсутствовать вовсе:

Comparison:
(3) [У жителей Финляндии средняя масса 69,3 кг], [у монголов и жителей Северного Китая - 55,8,]

В результате, анализ дискурса вызывает трудности не только при автоматической обработке текстов, но и при разборе отношений людьми. Так как далеко не у всех риторических отношений есть уникальные маркеры.

При анализе дискурсивных отношений "контраст" и "сравнение" мы исходим из представления об организации дискурса, основанном на Теории риторических структур (Mann, Thompson, 1988). В ее основе лежит организация текста в формате древовидной структуры, где каждому узлу приписано одно из риторических отношений. Листьями этого дерева будут элементарные дискурсивные единицы (ЭДЕ), чаще всего это предложения. Риторические

отношения могут быть одноядерными и многоядерными. Первый тип связывает ядро и сателлит, то есть главную ЭДЕ и зависимую соответственно. Последний соединяет элементы, которые одинаково важны в анализируемом дискурсе. Типы риторических отношений аналогичны тем, что обычно используются для описания типов сложных предложений.

Одна из основных задач автоматического анализа дискурса – распознавание отношений. Как отмечается в литературе, некоторые языковые средства сигнализируют о каком-то определенном типе отношений, например, союз «но» о контрасте:

(4) В этой ситуации человек свободен выбирать, какие тексты читать, ***но*** *он несвободен осуществлять свой выбор в максимальной степени, поскольку не в состоянии ориентироваться в количестве прочитанного, его оценке и тем более сопоставить с тем, что осталось им не прочитано.*

В настоящей работе речь пойдет об отношениях Contrast и Comparison. Наша задача выявить признаки, важные для указанных отношений и отличающие их друг от друга, а затем проверить, могут ли эти признаки помочь в автоматической классификации отношений контраста и сравнения.

Цель исследования заключается в том, чтобы проверить различные гипотезы о том, какие лингвистически мотивированные признаки позволяют отличать одно отношение от другого (Constrast vs. Comparison). На основе анализа признаков, упомянутых в литературе, выдвинуть гипотезы о том, какие ориентированные могут оказать влияние на классификацию отношений контраста и сравнения.

Также, мы хотим подобрать оптимальный набор выявленных признаков, которые в наибольшей степени влияют на качество распознавания Contrast и Comparison. Основная задача состоит в том, чтобы проверить, действительно ли статистически значимые для отношения признаки помогут в его распознавании.

Для достижения этих целей необходимо решить следующие задачи:

- рассмотреть признаки, которые используются для распознавания отношений контраста и сравнения в теоретических и прикладных работах
- провести корпусной анализ признаков: рассмотреть распределение разных значений признаков в подкорпусе примеров на отношения контраста и сравнения в корпусе, размеченном в терминах теории риторических структур
- построить классификатор
- определить влияние выделенных на предыдущих этапах признаков на качество классификатора

Contrast и Comparison выбраны нами для исследования, потому что при распознавании этих отношений возникает много трудностей. Во-первых, для них сложно выделить специфические маркеры отношений. Еще больше осложняет распознавание то, что некоторые маркеры для этих отношений общие. Во-вторых, контраст и сравнение часто смешиваются при попытке автоматического определения дискурсивных отношений (Толдова и др, 2019).

Мы проводим исследование на материале Ru-RSTreebank, размеченного в соответствии с теорией Rhetorical Structure Theory (RST) (Mann, Thompson, 1988). На данный момент в корпус входит 178 размеченных текстов, из которых нам удалось выделить 832 примера, связанных отношением Contrast и Comparison, из которых 569 относятся к первому, а 263 – ко второму отношению. На основании этих данных мы провели корпусное исследование, а затем использовали алгоритм Random Forest для определения оптимального набора параметров для классификации выбранных отношений.

Подробное описание того, какие признаки мы проверяли и какие из них оказались статистически значимыми, вы можете найти в разделе 4. В разделе 5 описан процесс подбора параметров для классификатора отношений контраста и сравнения и результаты, которые получились.

Текущее исследование не только дополнит знания об отношениях контраста и сравнения в русском языке, но и пригодится для разработки классификаторов других риторических отношений.[1]

2. Обзор существующих подходов

Существует множество задач обработки естественного языка, требующих анализа текста за пределами отдельных предложений. В последнее время исследователи начали подходить к этой проблеме, используя анализ дискурса, что повысило интерес к этой теме среди исследователей. Более того, в последнее время дискурсивный анализ текстов широко применяется при решении задач связанных с компьютерной лингвистикой, так как этот подход позволяет учитывать связи не только внутри предложений, словосочетаний, но и между фразами в тексте целиком.

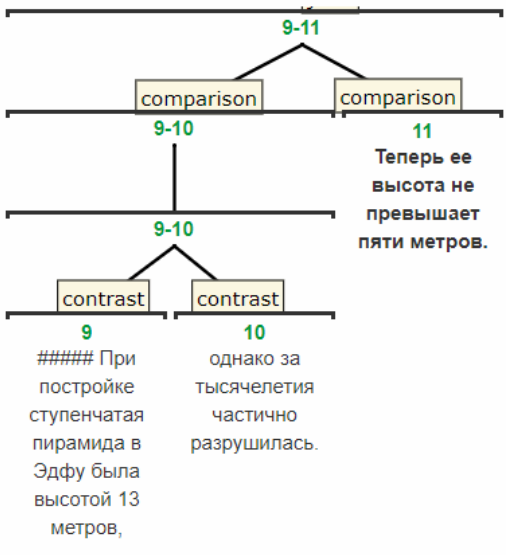
Для решения этих задач используются корпуса с текстами, в которых размечены именно дискурсивные отношения, иначе называющимися - treebanks. Существует несколько теорий разметки treebanks. Одна из них Penn Discourse Treebank (PDTB) (Webber и др., 2016), предполагает разметку не в формате дерева. Согласно этому подходу дискурсивные отношения лексически показываются маркерами, как в корпусе Турецкого языка. Более того Penn Discourse Treebank (PDTB) допускает возможность разметки с учетом важность пунктуации, как это требуется для китайского языка. Также, существуют другие модели не в формате дерева основанные на «cohesive relation», например, Discourse Graphbank (Wolf и др., 2005).

Наиболее подходящая для русского языка модель носит название Rhetorical Structure Theory (RST) (Mann, Thompson, 1988). Этот способ разметки предполагает иерархическую структуру в форме дерева с описанием отношений между частями текста. Листьями дерева являются так называемые элементарные дискурсивные единицы (ЭДЕ), которые соответствуют предикациям. В результате размеченный текст выглядит как единое дерево, где каждому узлу присвоен

[1] Автор работы выражает особенную благодарность научному руководителю, С.Ю. Толдовой, за неоценимую помощь в написании ВКР.

определенный тип риторических отношений. Такой формат позволяет проследить иерархию отношений и взаимосвязь элементов текста. Это позволяет решать задачи, для которых требуется знание того, на каком расстоянии расположены друг от друга некоторые единицы текста и как связаны между собой элементы текста крупнее, чем элементарные дискурсивные единицы. Некоторые элементарные дискурсивные единицы являются более важными и несут более важную информацию (ядро), чем другие (сателлит). Именно модель Rhetorical Structure Theory (RST) была выбрана для создания treebank для русского языка (Писаревская и др., 2017), потому что она хорошо демонстрирует не только отношения между элементами, но и их иерархию в тексте. На рисунке 1 приведен пример разметки текста в RU-Rst treebank:

Рисунок 1. Пример разметки в Ru-RSTreebank



В теории Mann и Thompson выделяются два основных типа риторических отношений: ядро-сателлит (одноядерные) и многоядерные. В то время как первый тип связывает главное предложение и зависимое. Последний включает элементарные дискурсивные единицы, которые одинаково важны в анализируемом дискурсе. Одноядерные отношения включают в себя 16 лейблов: Background (Фон), Cause (Причина), Evidence (Обоснование), Effect (Следствие), Condition (Условие), Purpose (Цель), Concession (Уступка), Preparation

(Подготовка), Conclusion (Вывод), Elaboration (Детализация), Antithesis (Антитезис), Solutionhood (Решение), Motivation (Мотивация), Evaluation (Оценка), Attribution (Источник), Interpretation (Толкование).

В свою очередь многоядерные отношения делятся на 6 типов: Contrast (Контраст), Restatement (Переформулировка), Sequence (Последовательность), Joint (Объединение), Comparison (Сравнение), Same-unit (Тег для прерывающейся единицы).

Для классификации одних ключевую роль играют маркеры риторических отношений, например, Cause - Effect (Толдова и др. 2018), для других необходимо обратить внимание на иные характеристики, как в случае с Contrast и Comparison (Толдова и др. 2018).

В изучении автоматического извлечения риторических отношений на основе маркеров отношений у исследователей есть два подхода:

1. Построение классификатора на основе самостоятельно собранных маркеров отношений (Толдова и др., 2018)

2. На основе имеющегося короткого списка маркеров создать модель, которая могла бы находить новые маркеры для данных риторических отношений (Толдова и др., 2018).

Первый подход позволяет учитывать слабо грамматикализованные маркеры отношений, которые в русском языке в некоторых типах отношений встречаются часто, например, в отношениях типа Cause и Effect. Недостаток этого подхода в том, что количество маркеров, которые учитываются в этой модели ограничено объемом корпуса, а также и его возможностями исследователя вручную извлечь из него маркеры.

В то же время, второй подход решает эту проблему и помогает находить подходящие по смыслу маркеры. Этот способ основан на w2v. Обучается специальная модель w2v, которая в дальнейшем находит подходящие варианты маркеров на основе экспериментов с отношениями типа Причина-следствие. Более того, этот подход более универсален, так как предполагает возможность

использования уже готового extractor'а для обучения на данных других риторических отношений.

В результате стало понятно, что сочетание двух этих подходов может дать хороший результат для отношений Cause и Effect.

Одна из актуальных проблем распознавания риторических отношений заключается в том, что при попытке создать универсальный классификатор исследователи сталкиваются с тем, что некоторые отношения образуют своеобразные пары, внутри которых их может быть трудно различить. Ошибки с маркировкой отношений частично возникают при наличии семантического сходства между реальным типом отношений и прогнозируемым типом, например в паре Contrast – Comparison. В таких случаях требуется более подробное изучение отношений, чтобы понять, по каким признакам их можно отличить друг от друга.

В данной работе мы будем рассматривать и эти критерии в том числе. Также, мы обратим особенное внимание, на значимость для классификации Contrast и Comparison тех черт этих риторических отношений, которые были обнаружены по итогам исследования (Толдова и др., 2019).

Оба отношения, которые мы исследуем – многоядерные, то есть связывают равноправные элементарные дискурсивные единицы. Семантические различия между примерами, относящимися к контрасту или сравнению, на сайте Ru-RSTreebank описываются так:

- «Contrast (Контраст)
  Ситуации, содержащиеся в дискурсивных единицах, противопоставлены друг другу, контрастируют относительно некоторой заданной темы.»

(3)[Возможно, они и знают, что им причитается,][но нет ни сил, ни средств на разбирательства.]

- «Comparison (Сравнение)

Две дискурсивных единицы сравниваются по какому-либо критерию. Отношение может показывать, что некоторые сущности сходны / различны / больше чем / меньше чем и т.д.»

(4) [В одних случаях функции такого прозвища может выполнить нарицательное наименование.][В других случаях прозвища основаны на случайных признаках человека или разнообразных ассоциациях.]

Учитывая то, что различие между этими отношениями и так не всегда прозрачно, осложняет задачу их классификации тот факт, что для этих риторических отношений могут использоваться одни и те же дискурсивные маркеры.

Contrast:
*(1) [Те существа, которые лучше приспособились, выживают,][**а** те, кто не приспособился, вымирают,[…]]*

Comparison:
*(2) [К примеру, BMW застраховал риски на североамериканском, то есть долларовом рынке на год вперед,][**а** другая германская компания, Porsche, - аж до 2007 года.]*

Исследование Contrast и Comparison направлено на выявление лингвистических особенностей, которые могут помочь дифференцировать эти отношения. Как показали данные, значительное количество дискурсивных маркеров, используемых для Contrast, неоднозначны, поскольку они маркируют и другие отношения. Согласно многочисленным исследованиям противительных союзов и конструкций контраста в русском языке (Шведова 1980; Урысон 2004), имеется дополнительный набор признаков, который может помочь отличить контраст и сравнение. К ним относятся, среди прочего:

1. Сходство построения элементарных дискурсивных единиц (ЭДЕ), связанных риторическим отношением – синтаксический параллелизм

Comparison:
(5) [Таким образом, во французском языке превалируют звуки, образуемые в

передней части голосового аппарата.][В английском языке, напротив, преобладают гласные звуки заднего ряда].

2. Выражение отрицания в ЭДЕ.

Contrast:
(3) [*Те существа, которые лучше приспособились, выживают,][а те, кто* ***не*** *приспособился, вымирают,[…]]*

3. Совстречаемость слов в ЭДЕ, связанных риторическим отношением – лексический параллелизм.

(6) [Получаем, что моделирующая программа, выполненная на основе ***алгоритма*** *№2 и функцией розыгрыша судьбы в варианте (б), производит расчеты с первым* ***набором*** *параметров примерно на 70%* ***быстрее****,][а со вторым* ***набором*** *- на 55%* ***быстрее*** *по сравнению с исходным вариантом* ***алгоритма****.]*

Также, исследователи (Толдова и др., 2019) в своей работе применяют несколько других критериев распознавания риторических отношений. Например, длина ЭДЕ в словах, часть речи первой и последней словоформы в элементарных дискурсивных единицах, косинусная близость между взвешенными по TF-IDF ЭДЕ, связанными одним риторическим отношением, количество стоп-слов в дискурсивных единицах и т.д. Классификация риторических отношений по этим критериям дала хорошие результаты для некоторых типов, например для Attribution (F_1-score = 74.36) и Purpose (F_1-score = 72.70). Для Contrast и Comparison значения были ниже: F_1-score = 56.69 и F_1-score = 38.49 соответственно. Более того, как уже было упомянуто, эти отношения показали тенденцию к тому, чтобы смешиваться.

В нашей работе мы будем проверять некоторые из методов, использованных предыдущими исследователями для создания классификатора для всех типов риторических отношений. Мы хотим проверить, имеют ли значимость эти критерии для отношений контраста и сравнения и помогут ли они в том, чтобы предотвратить смешение отношений.

В последнее время, для анализа дискурсивных отношений начали использовать модели глубокого обучения. Была предложена структура, основанная на рекурсивной нейронной сети, которая моделирует подзадачи сегментации EDU, построения древовидной структуру, маркировки центра и маркировки смысла (Lin и др., 2018). Другой подход заключается в создании сети сопоставления текста, то есть сеть кодирует единицы дискурса и абзацы, объединяя Bi-LSTM и CNN для захвата как глобальной информации о зависимостях, так и локальной информации n-граммы.

В нашей работе мы не будем использовать глубокое обучение, потому что нам интересно изучить лингвистические особенности отношений контраста и сравнения. Поскольку различия между Contrast и Comparison размыты, наша задача выявить признаки, значимые для каждого из отношений, чтобы в дальнейших работах исследователям было легче решить проблему смешения этих типов риторических отношений.

3. **Данные**

3.1 Корпус

В Ru-RSTreebank (https://linghub.ru/ru-rstreebank/) сейчас размечено 178 текстов, каждый из которых длинной около 30 предложений. В корпусе представлены следующие жанры текстов:

- Science
- popular science
- news stories
- analytic journalism

Каждый из текстов размечен как единое дерево, где большие элементы текста становятся все меньше и меньше и превращаются в элементарные дискурсивные единицы. Каждому узлу присвоен какой-то тип отношений.

3.2 Данные

В нашем исследовании мы будем использовать 832 примера, из которых 569 связаны отношением Contrast, а 263 – Comparison. Все примеры были размечены вручную. В данной работе используются те же предложения, которые были проанализированы в работе С.Ю. Толдовой и др., 2019.

Основная работа проводилась с дата фреймом, в котором для каждого примера были исходные столбцы, содержащие:

1. текст примера
2. первая ЭДЕ
3. вторая ЭДЕ
4. название риторического отношения (Contrast или Comparison)
5. маркер риторического отношения или “NO”, если явно выраженного маркера в данном примере нет

В результате исследования к исходному фрейму добавились столбцы с нормализованными ЭДЕ и полным текстом. А также, столбцы с данными, полученными в ходе исследования.

4. Нахождение признаков для классификации Contrast и Comparison

Так как, одним из целей исследования является нахождение признаков риторических отношений, которые позволят их явно находить в тексте, мы решили сначала найти несколько особенностей, а затем выбрать из них те, что дают лучшие показатели при классификации.

4.1 Векторная близость

Одно из первых предположений, которое было высказано в начале исследования, что на классификацию отношений может влиять близость двух элементарных дискурсивных единиц, связанных риторическим отношением. Для того, чтобы это проверить, мы использовали нормализованные ЭДЕ, то есть в текстах, которые мы обрабатывали, были удалены стоп-слова русского языка и все словоформы были приведены к начальной форме.

Также, было очевидно, что значимость слов в общей коллекции примеров может повлиять на классификацию, поэтому мы провели два эксперимента:

1. со взвешиванием по TF-IDF
2. без взвешивания

Для векторизации нормализованных ЭДЕ была выбрана модель на основе корпуса Araneum с ресурса RusVectōrēs (Кутузов, Кузьменко, 2017, https://rusvectores.org/ru/). Основной причиной выбора этой модели послужило то, что в ее корпус входит около 10 миллиардов слов. Это самый большой размер корпуса из всех моделей, представленных на сайте.

Затем необходимо было представить слова в векторной форме, для этого была использована технология Word2Vec. Технология основана на том, что вычисляется векторное представление слов, основываясь на контекстной близости, то есть, если какие-то слова встречается в тексте рядом с одинаковыми словами, в векторном представлении их координаты будут близки. В результате работы этого алгоритма мы получили ЭДЕ, представленные в форме чисел.

Затем необходимо выяснить, можно ли утверждать, что какому-то риторическому отношению свойственна близость ЭДЕ, им связанных. То есть, это бы значило, что отличительная черта этого риторического отношения – использование слов, часто употребляющихся в схожих контекстах. Чтобы это выяснить, нам необходимо вычислить близость полученных векторов. Для этого мы использовали формулу косинусной близости, которая используется для измерения угла между векторами. В результате мы получили два массива чисел (со взвешиванием по TF-IDF и без), содержащих близость двух ЭДЕ. Среднее значение для каждого отношения указано в Таблице 1.

Таблица 1. Среднее значение косинусной близость двух ЭДЕ.

	cTF-IDF	Без взвешивания
Contrast	~0.462	~0.445
Comparison	~ 0.404	~0.402

Как видно из приведенных в таблице результатов, взвешивание векторов по TF-IDF увеличивает близость дискурсивных единиц, но при этом результаты для Contrast и Comparison получились практически равны. Основная проблема в классификации этих отношений с помощью метода векторной близости заключается в том, что среднее значение на уровне примерно 0.45 достигнуто из-за того, что близость некоторых частей очень мала, а некоторых равна 0.9. Как распределяется близость двух частей показано на рисунках 2 и 3.

Рисунок 2. Близость частей Contrast и Comparison

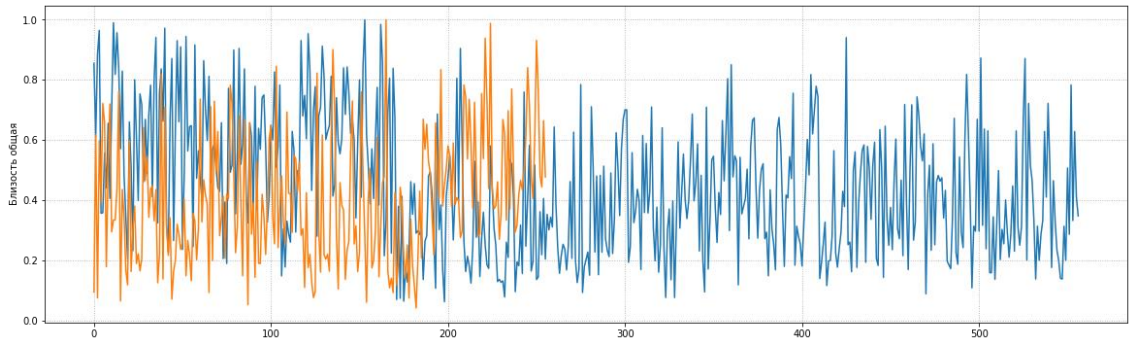
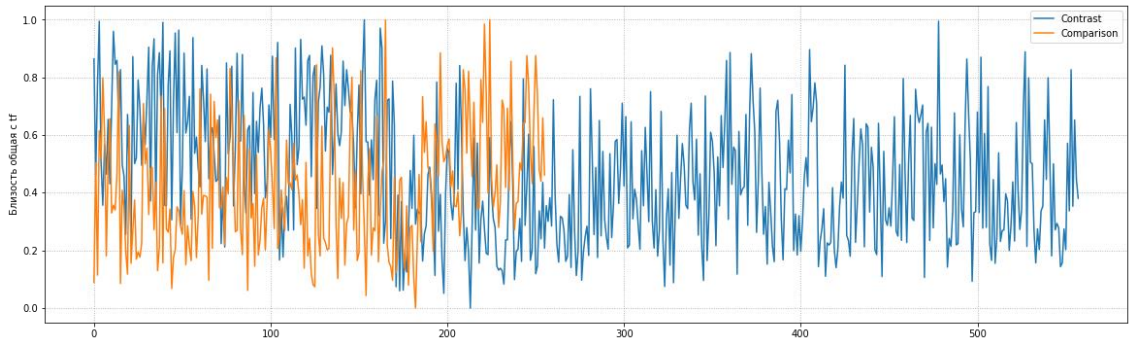


Рисунок 3. Близость частей Contrast и Comparison с TF-IDF



Получается, что ни для одного из отношений не наблюдается общей тенденции, поэтому этот критерий нельзя назвать свойственным отношению Contrast или Comparison.

К вычислению векторной близости мы еще вернемся в нашем исследовании, но позже, потому что нам понадобится еще преобразовать данные.

4.2 Одинаковые субъекты в элементарных дискурсивных единицах.

Мы уже упоминали, что исследуемым дискурсивным отношениям свойственен синтаксический параллелизм. Этот критерий мы исследовали двумя способами.

Первый направлен на то, чтобы найти в двух связанных дискурсивных единицах одинаковые существительные или местоимения в именительном падеже. Для этого мы использовали морфологический анализатор для русского языка MyStem (https://tech.yandex.ru/mystem/). Эта программа анализирует русскоязычный текст и выдает список, каждым элементом которого является морфологический разбор слова. В разборе указаны:

- грамматические характеристики словоформы
- лексема
- сама словоформа

Написанный нами алгоритм получает разбор ЭДЕ от MyStem, а затем, сравнивает найденные в первой и второй части существительные или местоимения в именительном падеже. В результате получилось, что существительные совпадают в 44% предложений Comparison, но только в 29% примерах на Contrast. Разница достаточная для того, чтобы расматривать этот способ как один из маркеров классификации Contrast и Comparison. Для того, чтобы убедиться, что мы не получили такие результаты случайно, необходимо посчитать статистическую значимость результатов. Для этого мы использовали критерий Хи-квадрат. Полученный в результате p-value = 2.5710708595715066e-05 подтверждает значимость исследуемого критерия.

Из этого мы можем сделать вывод, что отношение типа Comparison отличает то, что в двух связанных его частях вероятнее встретить одинаковые существительные в именительном падеже, чем в Contrast. Недостаток этого критерия в том, что изначально нам хотелось проверить сходство субъектов обеих частей, а в данном случае мы захватим лишние примеры:

> *(7)Так, если корневая морфема кард- или пневм- называют анатомическое понятие, обозначающее* ***орган****, <…> указывают не на* ***орган****, а на ткань, состояние, качество, свойство, связанное с отклонением от нормы в функционировании или структуре органа.*

Для того, чтобы проанализировать субъекты предложения, необходим синтаксический анализ предложения, потому что, как мы уже выяснили, морфологический подход не дает точных результатов.

В этом и заключается второй способ анализа параллелизма субъектов. Для того, чтобы получить синтаксический анализ предложения, мы использовали Turku-neural-parser-pipeline (Kanerva и др., 2018, https://turkunlp.org/Turku-neural-parser-pipeline/). Это алгоритм, для синтаксического и морфологического анализа и построения зависимостей, который работает более чем для 50 языков, в том числе и для русского. В результате работы этой программы мы получаем разобранный текст, где для каждого слова есть не только морфологический разбор, но и его синтаксическая роль и указание на связь с каким-то другим членом предложения.

Субъект в предложении обозначается тегом “nsubj”. Слова именно с таким тегом мы и сравнивали. В результате выяснилось, что субъект совпал в 17% примерах Contrast и в 24% - Comparison. Процент уменьшился практичсеки в два раза, но для Comparison, по-прежнему, больше. Хи-квадрат для этих результатов дал p-value = 0.02846884094632826, что позволяет нам считать совпадение субъектов значимым для классификации отношений критерием.

4.2 Одинаковые глаголы в элементарных дискурсивных единицах.

Так как мы проверяем значимость синтаксического параллелизма в классификации Contrast и Comparison, мы должны проверить и совпадение глаголов. Для этого мы использовали те же два способа, что и для субъектов.

Первый способ – MyStem. Но, в данном случае, мы не просто искали совпадающие по инфинитиву глаголы. Важно было проверить наличие параллелизма в именных группах. Поэтому мы сравнивали глаголы, согласовывающиеся с существительным в именительном падеже, найденном в этой же ЭДЕ.

В результате мы 19% совпадений для Contrast и 20% для Comparison. Как и в предыдущих случаях, мы хотим отсеять те критерии, которые на самом деле не влияют на классификацию, поэтому мы применили к данным формулу Хи-квадрат, которая дала нам p-value = 0.8079487136445214. Это значит, что такое распределение мы могли получить случайно. В дальнейшем мы не будем использовать результаты этого эксперимента.

Далее мы проверили совпадение глаголов в текстах, размеченных Turku. Особенность разметки Turku в том, что вершину предложения он помечает тегом «root», что облегчает нам нахождение необходимых глаголов. Также, анализатор предоставляет данные не только синтаксической разметки, но и морфологической что позволит нам обойти предложения с эллипсисом.

Как и в эксперименте с субъектами, процент с использованием Turku уменьшился, так как мы, аналогичным образом, избавились от лишних примеров. В результате процент предложений с совпадающей вершиной-глаголом в Contrast – 4%, а для Comparison это значение упало до 7%. Несмотря на то, что значения очень маленькие, мы проверим статистическую значимость этого критерия и, возможно, будем использовать этот параметр. На данных значениях p-value = 0.0385786255052011. Это значит, что данные, полученные в ходе этого эксперимента, могут пригодиться при подборе параметров для классификации Contrast и Comparison.

4.3 Общая лексика

Одним из критериев для различия Contrast и Comparison может быть то, что одному из этих отношений свойственны повторения лексики в обеих элементарных дискурсивных единицах.

Для того чтобы это проверить, мы опять использовали MyStem, но для этого эксперимента использовался не исходный текст ЭДЕ. Поскольку нам нужно было найти все повторяющиеся лексические единицы, мы использовали примеры с удаленными стоп-словами. Затем алгоритм сравнивал лексемы, входящие в состав

обеих частей, и считал, в скольких предложениях встретились совпадения. Стоит отметить, что мы не учитываем, сколько всего повторяющихся лексем мы нашли, потому что тогда для Comparison у нас бы получилось, что повторяющиеся слова есть в 400% предложений. В нашем эксперименте подсчитывалось количество примеров, в которых нашлась хотя бы одна пара лексем. Несмотря на то, что общее количество совпадений для Contrast так велико, что на одно предложение приходится примерно две пары лексем, процент предложений, в которых повторения встречаются - 58%. Для Comparison это значение равно 70%.

Из этих результатов мы можем сделать вывод, что в наших примерах почти не бывает одиночных повторений лексем. Если слова повторяются, то в словосочетаниях, потому что, если разделить количество найденных повторений на количество предложений, в которых повторения действительно есть, получится, что на каждый такой пример Contrast приходится ~4 слова, а на пример Comparison – почти 6.

Таблица 2. Количество предложений с повторяющейся лексикой и общее количество найденных повторов.

	Количество предложений	Количество повторений
Contrast	331	1418
Comparison	185	1060

Для получившихся значений предложений с повторяющейся лексикой p-value = 0.001015825233428605.

4.4 Величина ЭДЕ

В самом начале исследования у нас появилась теория, что одному из отношений могут быть свойственны более распространенные ЭДЕ, чем другому. Для того, чтобы это проверить, мы использовали исходные тексты примеров и посчитали количество словоформ в каждой части. Результаты отражены в таблице 3.

Таблица 3. Средняя длина ЭДЕ

	Ядро-1	Ядро-2
Contrast	17.85	22.06
Comparison	20.44	20.16

Согласно таблице, первая часть Contrast немного короче, чем вторая, тогда как Comparison обе части равны. Но получившиеся в этом эксперименте

результаты не дают возможности классифицировать Contrast и Comparison, потому что разница между ними слишком мала.

4.5 Эллипсис

При первом взгляде на примеры, с которыми мы работаем, создается впечатление, что одному из отношений более свойственен эллипис, чем другому. Для того чтобы это выяснить, мы использовали уже размеченные с помощью Turku примеры. Мы не могли просто проверить наличие глагола в ЭДЕ, потому что тогда в результат попали бы лишние значения.

В разметке Turku вершина обозначается тегом «root». При анализе примеров мы смотрели словоформы с этим тегом и проверяли: если, согласно морфологическому разбору Turku, это не глагол, значит это предложение засчитывалось как пример с эллипсисом.

В результате у нас получилось, что в примерах Contrast эллипсис встречается в 54%, а в Comparison – 55%. Проверку на статистическую значимость эти данные не прошли, поэтому на этапе подбора оптимального набора значений мы их использовать не будем.

4.6 Отрицание

Предыдущие исследования показали, что отрицание действительно может помочь распознать риторические отношение. Наша задача заключалась в том, чтобы подобрать оптимальный способ их нахождения и проверить, какой из способов лучше.

Первый способ заключался в нахождении определённых слов, маркирующих отрицание: *ни, не, никто, ничто, ничей, никакой, нечего, некого, никогда, нет, нельзя* (Падучева, 2011). Алгоритм сравнивал каждую словоформу со списком ключевых слов и считал, в скольких предложениях встретилось отрицание. Подход в данном случае был такой же, как с повторяющимися лексемами. Нам необходимо было получить количество предложений с отрицанием, а не количество ключевых слов в предложениях.

Второй подход подразумевал работу с регулярными выражениями. Алгоритм искал словоформы, начинающиеся на *не-* или *ни-*. Таким образом мы хотели, чтобы при подсчете учитывалось и отрицание, выраженное префиксами. Таблица показывает, какие результаты получились при каждом из этих подходов.

Таблица 4. Процент предложений с отрицанием в зависимости от способа распознавания.

	список	*не-* или *ни-*
Contrast	50%	73%
Comparison	25%	48%

Мы видим, что второй способ дает больший результат для обоих отношений, потому что он захватывает и слова из списка, использованного в первом случае, и другие слава выражающие отрицание.

Несмотря на то, что разница между первым и вторым подходом достаточно велика, было решено оставить результаты обоих экспериментов для дальнейших исследований, потому что извлечение отрицание с помощью списка ключевых слов дает p-value = 1.1471450826396528e-11, а регулярные выражения – p-value = 9.68908643880129e-13.

4.7 Частицы

Еще один критерий, проверить который мы решили после того, как сами попытались вывести какие-то закономерности для наших данных – количество частиц. Мы решили проверить теорию о том, что в примерах отношения Contrast частиц больше, чем в Comparison.

Для проверки этой гипотезы мы обработали данные с помощью уже упомянутого анализатора MyStem. В системе тегов этого алгоритма есть специальный тег «PART», обозначающий частицы. Так же, как и в предыдущих экспериментах, нам было нужно посчитать не количество частиц, а количество примеров, в которых частицы встречаются.

И, согласно результатам подсчета алгоритма, частицы встречаются в 68% предложений Contrast и в 53% примеров Comparison. P-value для этого эксперимента составило 2.9284817383205424e-49.

4.8 Именные группы

Поскольку вычисление векторной близости двух элементарных дискурсивных единиц не дало какого-то очевидного результата, было принято решение вычислить семантическую близость между именными группами (ИГ) двух дискурсивные единицы, которые входят в отношения.

На первом этапе необходимо было распознать в предложениях ИГ. Для этого использовались уже размеченные с помощью Turku данные. Из них для исследования были выбраны словоформы с тегами «root» и «nsubj» так, чтобы между ними была указана связь.

Затем мы представили ИГ двух частей связанных риторическим отношением в виде вектора с помощью той же модели, которой мы преобразовывали части целиком. Также, для чистоты эксперимента, мы сделали два векторных представления со взвешиванием по TF-IDF и без. Косинусная близость двух ИГ считалась по той же формуле. Результаты получились неожиданными для нас:

Таблица 5. Среднее значение косинусной близость двух ИГ.

	cTF-IDF	Без взвешивания
Contrast	~0.91	~0.796
Comparison	~ 0.51	~0.767

Без взвешивания результаты достаточно высокие и отличаются не сильно. Но чего нельзя сказать о результатах со взвешиванием.

Рисунок 4. Близость ИГ

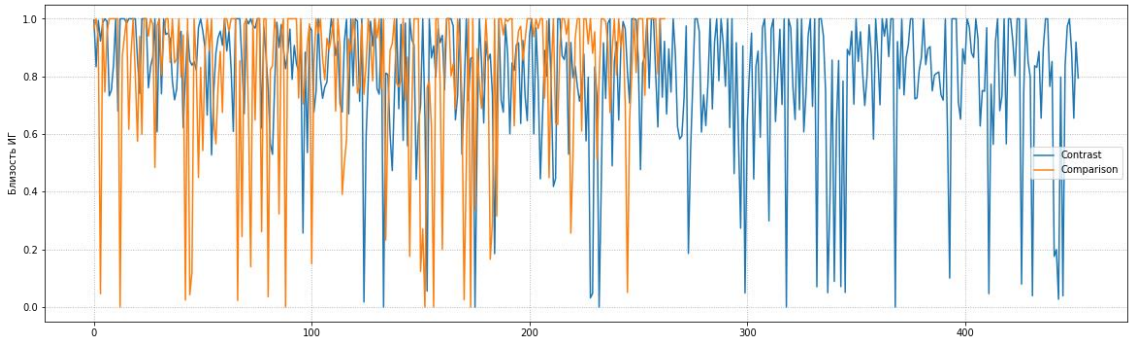
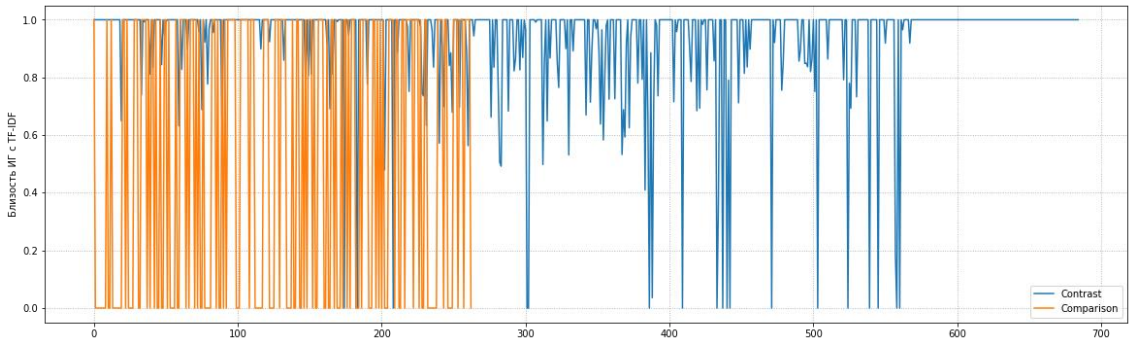


Рисунок 5. Близость ИГ с TF-IDF



Оказалось, что после взвешивания по TF-IDF семантическая близость именных групп в Comparison дает в половине случаев число близкое к 1, а в половине – 0. Поэтому среднее значение равно 0.51. Чего не скажешь о Contrast, где после взвешивания значения только поднялись.

4.9 Промежуточные результаты

После проведенных экспериментов у нас образовалось 11 признаков, которые могут помочь распознаванию Contrast и Comparison. Следующим этапом нашего исследования будет отбор из этих 11 признаков тех, что действительно могут помочь в распознавании риторических отношений. Прежде чем перейти к этому этапу, мы приведем общие таблицы:

Таблица 6. Признаки, значимые для распознавания Contrast и Comparison

	Contrast	Comparison
PART	68%	53%
Отрицание_список	50%	25%
не- или *ни-*	73%	48%
Совпадение лексики	58%	70%
Verb_Turku	4%	7%
Subj_Turku	17%	24%
Subj_mystem	29%	44%
Близость частей cTF-IDF	~0.462	~ 0.404
Близость частей	~0.445	~0.402
Близость ИГ cTF-IDF	~0.91	~ 0.51
Близость ИГ	~0.796	~0.767

5. Классификация

Для того чтобы можно было проверить, как влияет на распознавание риторических отношений тот или иной признак был создан датафрейм, в котором каждому примеру было присвоено какое-то значение по одному из критериев:

- Наличие в примере частиц – 1, отсутвие – 0
- Есть отрицание из списка – 1, нет – 0
- Есть отрицание не- или ни- – 1, нет – 0
- Есть совпадающая лексика – 1, нет – 0
- Есть совпадающие глаголы с тегом «root» – 1, нет – 0

- Есть совпадающие слова с тегом «nsubj» - 1, нет – 0
- Есть совпадающие существительные в им.п. – 1, нет – 0

В случае с векторной близостью значения, присвоенные примерам, соответствовали векторной близости частей или ИГ со взвешиванием или без.

Для того, чтобы подобрать наиболее значимые параметры были выбраны алгоритмы Random Forest. Они подходили для нашей задачи, потому что принимают несколько характеристик объекта, тестирует их, пока не доходит до целевой переменной. Данные алгоритмы часто применяются в задачах с бинарной классификацией. Наша цель была подобрать набор характеристик отношений, который поможет лучше всего распознать Contrast и Comparison и подобрать параметры модели. Для оценки точности модели была выбрана F-мера, так как классы в наших данных не сбалансированы.

На первом этапе мы перебирали все возможные комбинации критериев, чтобы попробовать найти ту, которая даст лучший результат. СамаяВ высокая точность была у комбинаций из 5 или 7 признаков. Промежуточные результаты отражены в таблице:

Таблица 7. Комбинации параметров с высоким результатом

	Macro F1	Micro F1
Близость ИГ, Близость ИГ cTF-IDF, *не-* или *ни-*, Близость частей , Verb_Turku	0.80	0.85
Близость ИГ, Близость ИГ cTF-IDF, Совпадение лексики, Отрицание_список, PART	0.77	0.83
Близость ИГ, Отрицание_список, *не-* или *ни-*, Subj_mystem, Subj_Turku	0.77	0.82
Близость ИГ, Совпадение лексики, *не-* или *ни-*, PART, Близость частей cTF-IDF	0.77	0.81
Близость ИГ, Отрицание_список, Близость частей, Subj_mystem, Verb_Turku	0.77	0.82
Близость ИГ, Отрицание_список, Близость частей, Subj_Turku, Verb_Turku	0.77	0.80
Близость ИГ, Близость ИГ cTF-IDF, *не-* или *ни-*, Близость частей, PART, Subj_Turku, Близость частей cTF-IDF	0.77	0.81

Близость ИГ, Близость ИГ cTF-IDF, Совпадение лексики, Отрицание_список, *не-* или *ни-*, PART, Близость частей cTF-IDF	0.77	0.80

В таблице все значения округлены до двух знаков после запятой, на самом деле значения не идентичны. После анализа полученных данных мы решили, что нужно подробнее изучить результаты, которые дает комбинация с самым высоким качеством и комбинация, которая включает в себя те признаки, которые встречаются минимум в половине комбинаций в таблице. В результате у нас получились два списка с параметрами:

Таблица 7. Списки с признаками для дальнейшей работы

Самый высокий результат	Почти всегда встречаются
Близость ИГ	Близость ИГ
Близость ИГ cTF-IDF	Близость ИГ cTF-IDF
не- или *ни-*	Отрицание_список
Близость частей	Близость частей
Verb_Turku	PART

Мы хотели посмотреть вклад каждого из признаков в результат, поэтому построили деревья для каждого из этих наборов еще раз. Качество комбинации с самым высоким результатом мы уже получили, а для набора из самых частых признаков Macro-F1 и Micro-F1 равны 0.77 и 0.80 соответственно. Алгоритм Random Forest позволяет получить важность каждого из признаков в классификации. Результаты, получившиеся для двух наших наборов, отражены на рисунках 6 и 7.

Рис.6. Вклад в классификацию признаков из комбинации с самым высоким результатом

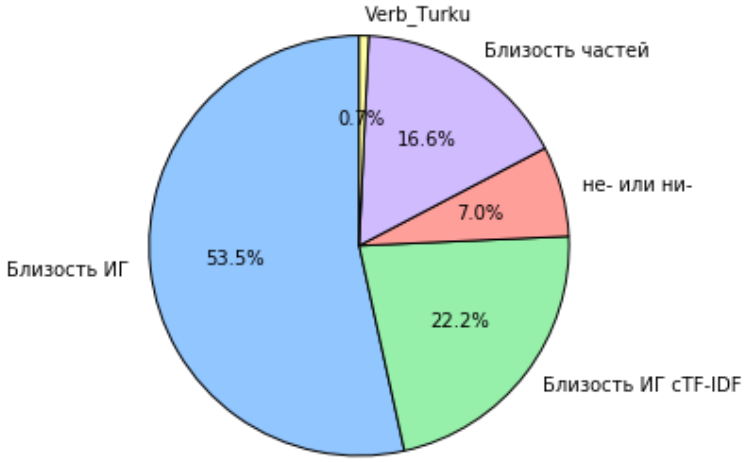
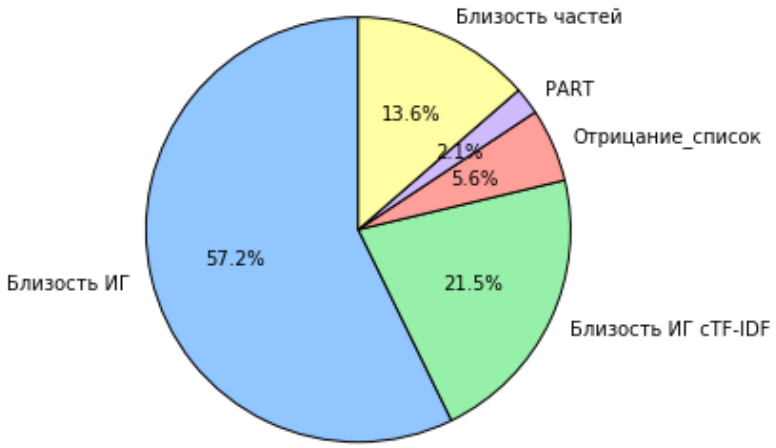


Рис.7. Вклад в классификацию признаков из комбинации с самыми частыми признаками



Из приведенных графиков мы видим, что наибольший вклад в классификацию вносят 3 параметра, которые встречаются и в одном, и в другом наборе: близость ИГ, близость ИГ с TF-IDF, близость частей.

Исходя из этих результатов, мы решили проверить, какое качество получится при классификации Contrast и Comparison с использованием только этих трех характеристик. В результате этого эксперимента у нас получились Macro-F1 и Micro-F1 равные 0.79 и 0.81 соответственно. При этом вклад в классификацию у параметра «близость ИГ» оказался больше 60%.

Если классифицировать отношения контраста и сравнения только по этому параметру Macro-F1 и Micro-F1 равны 0.75 и 0.79 соответственно. Мы видим, что результаты ухудшились, но не очень сильно. Обобщенные результаты представлены в таблице.

Таблица 8. Результаты подбора параметров

Параметры	Macro F1	Micro F1
Близость ИГ, Близость ИГ cTF-IDF, не- или ни, Близость частей,Verb_Turku	0.80	0.85
Близость ИГ, Близость ИГ cTF-IDF, Отрицание_список, Близость частей, PART	0.77	0.80
Близость ИГ, Близость ИГ cTF-IDF, Близость частей	0.79	0.81
Близость ИГ	0.75	0.79

Из проведенных экспериментов мы можем сделать вывод, что основной вклад в классификацию отношений контраста и сравнения вносят только три параметра. Мы видим, что, если к 3 основным параметрам добавить еще несколько, мы можем улучшить результат, но, как показал наш эксперимент, результаты на всех 11 признаках дают качество ниже 0.65.

6. Вывод

В данной работе представлен анализ различных признаков, которые могут сигнализировать о риторических отношениях контраста и сравнения в русском языке. В работе мы исследовали то, как ведут себя различные средства маркирования этих отношений, упомянутые в литературе, на материале корпуса Ru-RSTreebank. Для этой цели мы рассмотрели различные типы сигналов на наших примерах. Наше исследование выявило значимые для отношений контраста и сравнения признаки:

- Количество частиц в примерах
- Наличие выраженного отрицания в дискурсивных единицах
- Лексические повторения в двух ЭДЕ, связанных одним отношением
- Повторение субъектов в ЭДЕ
- Повторение глаголов в частях одного предложения
- Косинусная близость двух дискурсивных единиц, связанных одним отношением, со взвешиванием по TF-IDF и без
- Косинусная близость именных групп двух дискурсивных единиц, связанных одним отношением, со взвешиванием по TF-IDF и без

В результате мы выяснили, что только 3 из них действительно влияют на классификацию отношений контраста и сравнения, а именно: близость ИГ, близость ИГ с TF-IDF, близость частей. Нам не удалось добиться высокой

точности распознавания, но, на наш взгляд, это связано с небольшим набором данных.

В следующих исследованиях мы попробуем улучшить текущий результат путем использования набора признаков с не бинарными значениями. Также попробуем оценить результаты других алгоритмов машинного обучения.

Исходный код экспериментов: https://github.com/tududundra/Dahina_FQW19/

Список литературы

1. Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham
2. Wolf, Florian, et al. Discourse Graphbank LDC2005T08. Web Download. Philadelphia: Linguistic Data Consortium, 2005.
3. Jenna Kanerva and Filip Ginter and Niko Miekka and Akseli Leino and Tapio Salakoski, Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task, Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, 2018
4. 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.
5. Kenji Sagae. Analysis of discourse structure with syntactic dependencies and datadriven shift-reduce parsing. In Proceedings of the 11th International Conference on Parsing Technologies, pages 81–84, 2009.
6. C.A.Lin,H.H.Huang,Z.Y.Chen,andH.H.Chen. AunifiedRvNNframeworkforendto-end Chinese discourse parsing. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 73–77, 2018.
7. 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.

8. Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A Discourse-Annotated Corpus of Conjoined VPs. Proc. 10th Linguistics Annotation Workshop, Berlin: 22–31.
9. D. Zeyrek, I. Demirşahin, A.B. Sevdik Çallı, and R. Çakıcı. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. Dialogue and Discourse, 4(2): 174–184.
10. Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. In Computational Linguistics, 31(2): 249–287.
11. W.C. Mann and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8, 3, 1988: 243–281.
12. Pisarevskaya D. et al. (2017), Towards building a discourse-annotated corpus of Russian, Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf.» Dialogue, Vol. 1, pp. 194204.
13. Prasad R., Miltsakaki E., Dinesh N., Lee A., Joshi A., Robaldo L., Webber B. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report 203, In-stitute for Research in Cognitive Science, University of Pennsylvania.
14. Toldova S., Pisarevskaya D., Kobozeva M. Automatic Mining of Discourse Connectives for Russian, in: Artificial Intelligence and Natural Language, 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings Issue 930. Switzerland : Springer, 2018. P. 79-87.
15. Toldova S., Pisarevskaya D., Vasilyeva M., Kobozeva M. The cues for rhetorical relations in Russian: "Cause-Effect" relation in Russian Rhetorical Structure Treebank, in: Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 30 мая — 2 июня 2018 г.) / Под общ. ред.: В. Селегей, И. М. Кобозева, Т. Е. Янко, И. Богуславский, Л. Л. Иомдин, М. А. Кронгауз, А. Ч. Пиперски. Вып. 17(24). М. : Издательский центр «Российский государственный гуманитарный университет», 2018. P. 747-761.
16. Y. Zhou and N. Xue. 2015. The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. Language Resources and Evaluation: 397–431.
17. Feltracco A., Magnini B., Jezek E. (2018), Lexical Opposition in Discourse Contrast, Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-11.
18. Kobozeva I. M. (2011), Conjunctions as markers of rhetorical relations in discourse: Russian conjunction "i" [Soyuzy kak markery ritoricheskikh otnosheniy v diskurse: russkiy soyuz "i"], L'analisi linguistica e letteraria, 19, № 2, pp. 365-387.
19. Mann W. C., Thompson S. A. (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8, 3, pp. 243–281.

20. Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. (2017), Towards building a discourse-annotated corpus of Russian, Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference “Dialogue 2017”, pp. 194-204.
21. Sannikov V. Z. (1989), Russian coordinate construciotns: semantics, pragmatics, sintaksis [Russkiye sochinitel'nyye konstruktsii: semantika, pragmatika, sintaksis], M, "Nauka".
22. Shvedova, N. YU. ed. (1980), Russian grammar [Russkaya grammatika]. V dvukh tomakh. AN SSSR Institut russkogo jazyka, M.: Nauka, 1980.
23. Taboada M., Das D. (2013), Annotation upon annotation: Adding signalling information to a corpus of discourse relations, Dialogue and Discourse 4(2), pp. 249-281
24. Taboada M., Mann W. C. (2006), Rhetorical structure theory: Looking back and moving ahead, Discourse studies, 8(3), pp. 423-459.
25. Uryson E. V. (2004), Some meanings of conjunction A in the light of modern semantic theory [Nekotoryye znacheniya soyuza A v svete sovremennoy semanticheskoy teorii]. Russian language in scientific coverage [Russkiy yazyk v nauchnom osveshchenii], (2), 17.
26. Uryson E. V. (2012), Conjunctions, connectives, and the valence theory [Soyuzy, konnektory i teoriya valentnostey]. In Computational linguistics and intellectual technologies [Komp'yuternaya lingvistika i intellektual'nyye tekhnologii], pp. 627-638.
27. Zaliznyak, A. A., Mikaelyan, I.L. (2005) Russian Conjunction “a” as a language-specific word [Russkiy soyuz a kak lingvospetsifichnoye slovo], Computational Linguistics and Intellectual Technologies, Proceedings of the International Conference “Dialogue 2005”, pp. 153–159.
28. Toldova S., Pisarevskaya D., Davydova T., Kobozeva M. Contrast and Comparison Relations in RST framework: the case of Russian, in: Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог 2019»
29. Chistova E., Shelmanov A., Kobozeva M., Toldova S., Pisarevskaya D., Smirnov I., Toldova S.Classification models for RST discourse parsing of texts in russian, in: Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог 2019»
30.