Классификация существительных из текстов методами машинного обучения на основе признаков контекстуальной синонимии

Е. Е. Милованова, milovanova.e1997@gmail.com

Московский авиационный институт (национальный исследовательский университет)

Аннотация. *Рассмотрено явление полисемии и ее влияние на автоматический анализ текста. Определено понятие контекстных синонимов и выявлена проблема их автоматического выделения текстов на русском языке. Для решения этой проблемы предложен алгоритм выделения контекстных синонимов на основе вектора признаков, базирующихся на морфологических и семантико-синтаксических данных, полученных в результате предварительного анализа исходного текста.*

Ключевые слова: *Контекстные синонимы, автоматическое выделение синонимов, синонимия, машинное обучение.*

Введение

На сегодняшний день одной из самых сложных проблем в области автоматизированного анализа текста является обработка многозначных слов. Полисемия – это часто встречающееся в естественном языке явление, при котором языковая единица может иметь более одного значения, вследствие чего невозможно полностью автоматизировать процесс обработки текстов на естественном языке. Корректное выделение контекстуальных синонимов – слов, которые схожи по смыслу с другими словами лишь в определенном контексте [1], – является частной задачей данного направления исследований компьютерной лингвистики.

Изучение различных видов семантических отношений языковых единиц и контекстуальной синонимии в частности, примеров их использования в речи важно как лингвистики в целом, поскольку составляют составляет богатство языка и обеспечивают возможность выразить одну и ту же мысль разными способами, указать уникальные смысловые акценты или перефразировать мысли [2], так и для улучшения существующих алгоритмов автоматического семантического анализа текстов, их качественного улучшения или создания новых методов для решения частных задач компьютерной лингвистики.

1. Проблема многозначности слов

Задача анализа текстов на естественном языке основываются на декомпозиции текста с дальнейшим анализом и обработкой полученных данных [3]. Данный процесс осложняется многообразием не только словосочетаний, но и многозначности слов, которая, в свою очередь, в разы увеличивает итоговую вариативность результатов. Количество полученных в итоге данных настолько велико, что существующие информационные системы не способны корректно обработать их и сформулировать соответствующий ответ.

На сегодняшний день явление многозначности, полисемию, разделяют на лексическую и грамматическую [4]:

– Лексическая полисемия – это случай, в котором одно слово служит для обозначения нескольких предметов или явлений. Пример: магнитное поле, маковое поле.

– Грамматическая полисемия – это случай, когда слово может быть употреблено в нескольких значениях. Например, в предложении «Постучали в дверь.» глагол употреблен в неопределенно-личном значении, т. к. в нем не указано, кто осуществил действие. В предложении «Мы постучали в окно.» глагол употребляется в собственно-личном значении, т. к. известно, кто позвонил.

Так же выделяют три разновидности полисемии:

– Радиальная связь – это связь, при которой все значения слова непосредственно связаны с главным.

Например, «разгрузка»:

1. действие по значению глагола разгружать, разгрузить; снятие груза откуда-либо;
2. освобождение от части обязанностей;
3. уменьшение физической или психологической нагрузки; отдых.

– Цепочная связь – это связь, при которой каждое новое значение слова связано с предшествующим, но «крайние» значения могут быть не связаны между собой.

Например, «картина»:

1. произведение живописи;
2. то, что можно видеть, обозревать или представлять себе в конкретных образах;
3. изображение чего-либо в художественном произведении;
4. подразделение акта в драме;
5. то же, что и фильм.

– Смешанная, или радиально-цепочная связь – тип связи, при которой в слове совмещены оба типа связи.

Проектирование интеллектуальной системы, включающей в себя словарь со всеми возможными вариантами связей слов и их значений, на данный момент невозможно. Это связанно не только с несовершенством существующих баз данных и методов их наполнения, но и со свойством человека в случае отсутствия необходимых сведений самостоятельно восполнить их. Для этого выполняется обращение к смежным или ассоциативным связям, основанным на более широких возможностях многозначности [5].

2. Применение методов машинного обучения для решения задач компьютерной лингвистики

На начальном этапе становления компьютерной лингвистики как самостоятельной научной дисциплины она характеризовалась использованием фиксированных лингвистических моделей для решения прикладных задач, однако в дальнейшем фокус сместился на применение математических статистических методов. Это позволило быстро получить достаточно высокие результаты посредством использования относительно простых математических методов, однако на данный момент их использование для задач автоматической обработки текстов уже достигло некоторого своего порога [8].

Направление машинного обучения в компьютерной лингвистике ориентировано на получение новых знаний в результате автоматизации процессов обучения. Однако эффективность данного подхода в рамках компьютерной лингвистики осложняется такими проблемами, как:

– зависимость от объёмов исходных данных и, как следствие, необходимость наличия и обработки больших и сверхбольших размеченных корпусов текстов, онтологий и тезариусов;

– подбор метрики оценки эффективности используемых эмпирических критериев, так как она зависит от сравнения результата человека-эксперта и программного средства [9].

В последние годы сформировалось новое направление извлечения знаний из текстов, связанное с построением онтологий. В его рамках исследователи основываются на дистрибутивной гипотезе, по которой сходство существительных устанавливается на основе сходства их синтаксических контекстов употребления, современные исследователи решают некоторые задачи классификации сущностей из текстов, например, построение классификации существительных на основе предикатно-аргументных структур [10].

3. Синонимы и контекстные синонимы

Синонимы – это слова, равнозначные или похожие по значению. Они могут быть представлены как одной частью речи, так и разными, но

близкими семантически [6]. Также могут включать одно или несколько слов: идти – гулять; большой – огромный; противогаз – резинотехническое изделие номер один. Контекстуальная синонимия – это частное явление синонимии, в котором слова или словосочетания равнозначны другим словам лишь в определенной ситуации. Например: “напряжённая, гнетущая атмосфера”; “пустынный, неприветливый особняк”. В русском языке напряжённый и гнетущий, пустынный и неприветливый – это разные понятия, но в определенном контексте близки по значению.

Контекст – это относительно законченная по смыслу часть текста или высказывания. Общий смысл контекста складывается из значений отдельных слов, в тоже время как сам контекст дополняет и помогает прояснить значение каждого слова. По этой причине часты случаи, когда слова, выступающие в роли контекстных синонимов, таковыми не являются без учета контекста. Например: “Все у них было как-то черство, неотесанно, неладно, негоже, нестройно, нехорошо…” (Н.В. Гоголь “Мёртвые души”)

Синонимическое сближение может возникать между родовыми и видовыми понятиями – гипонимами и гиперонимами, между словами одной тематической группы [7]. Например: “Я позвал собаку. Барбос подошел, он был мне рад. Овчарка села рядом." В данном случае речь идет об одном и том же животном, поэтому слова Барбос, собака, овчарка синонимичны [6]. Кроме того, в данном случае в качестве контекстных синонимов использованы местоимения и имена собственные.

Перифраз, метафора и другие литературные тропы могут порождать контекстные синонимы. Например, у М.Ю. Лермонтова в перечислении “звучал булат, картечь визжала” булат – синоним холодного оружия, сабель и штыков.

Контекстуальные синонимы определяются человеком как продукт индивидуального творческого акта, создающего близость значения слов в рамках определённого контекста. В связи с этим анализ данного языкового явления имеет особое значение при рассмотрении художественного текста: контекстные синонимы возникают из необходимости фиксировать в слове новые оттенки явления, представления или понятия, что определённым образом характеризуют его оценку, отношение к нему [7].

По мнению Т.Б. Радбиля, назначение синонимов заключено в разном представлении одного и того же действия путем выделения и акцентирования разных аспектов и элементов содержания ситуации. Человек в определённом значении “выбирает” из ситуации только часть

информации, создавая собственный способ концептуализации, осмысления этой ситуации. Иными словами, слово представляет собой семантическую модель ситуации. В своей семантической модели ситуации говорящий может что-то выделить, сделать акцент, подчеркивая в слове какой-либо особый признак описываемого объекта, или что-то “затемнять” в слове, отодвигая на задний план или даже искажая семантику ситуации. Поэтому для описания ситуации важна также оставшаяся информация, не вошедшая в часть исходного, основного значения слова [2].

4. Классификация сущностей из текста для выявления контекстных синонимов

В рамках исследования возможностей машинного обучения в области выделения контекстуальных синонимов выдвигается предположение о том, что у каждого осмысленного текста на естественном языке существует уникальный для конкретного контекста набор ключевых сущностей, на которых автор акцентирует внимание [7]. Тогда, основываясь на характерных особенностях явления синонимии, можно предположить, что семантически связанные между собой сущности будут иметь схожие морфологические и семантико-синтаксические характеристики. Следовательно, обученная классифицировать сущности на данных условиях модель может быть использована в дальнейшем как для изучения поведения сущностей в тексте и выделения предполагаемых контекстных синонимов, так и для аналогичного анализа схожих текстов (например, текстов одного автора или на схожую тему).

Для обучения таких моделей в рамках доклада предлагается следующий алгоритм векторизации текста:

1. Получение всех существительных из исходного текста.
2. Приведение существительных в начальную форму (именительный падеж, единственное число).
3. Морфологический анализ каждого существительного с целью определения таких характеристик, как именованность (собственное или нарицательное), род и одушевлённость.
4. Семантико-синтаксический анализ, включающий в себя определение частоты употребления каждого существительного в тексте и количество зависимостей между данным словом и другими частями речи.

Но основе анализа полученных векторов возможно выделение критериев, специфичных для сущностей в контексте исходных данных. После этого выполняется подготовка тренировочной выборки по распределению сущностей на три категории:

“person”: активный персонаж событий, описанных в исходном тексте, обычно человек.

“object”: пассивный предмет, играющий важную роль в исходном тексте.

“something”: объекты, не имеющие высокой информационной ценности ввиду того, что не являются предметами, на которых акцентирует внимание автор.

Преимуществом таким образом обученной модели является установление зависимостей от поведения сущностей в тексте, благодаря чему возможно “считывание” специфичных акцентов автора, например, прозвищ, вне контекста не являющихся именем собственным, или “отсеивание” большого количества сущностей, сопровождающих повествованием на естественном языке. Полученные в рамках каждого класса сущности можно назвать предполагаемыми контекстными синонимами, так как они имеют схожие морфологические и поведенческие характеристики слов-синонимов в исследуемом тексте.

Заключение

Предлагаемый в докладе метод на основе использования машинного обучения для решения задачи выделения контекстных синонимов позволит получить обширный материал для дальнейших исследований в области выделения и изучения многозначности слов в русском языке и, в частности, контекстных синонимов. Также он позволит проводить более глубокий сравнительный анализ использования языковых конструкций разными авторами с точи зрения лингвистики.

Список литературы

1. Реформатский, А.А. Введение в языковедение: учебник для пед. вузов / В.А. Виноградова; 5-е изд., уточн. – М: Аспект-Пресс, 1999. – 536 с.

2. Путятина, Е.И. Контекстуальная синонимия в тексте и его дискурсе / Е.И. Путятина // Вестник КГУ им. Н.А. Некрасова. – 2016. – №4. – С. 148-151.

3. Яцко, В.А. Предметная область компьютерной лингвистики / В.А. Яцко // Вестник Иркутского государственного лингвистического университета. – 2014. – №2. – С. 24-35.

4. Большая российская энциклопедия [Электронный ресурс]: энциклопедия. – Режим доступа : https://bigenc.ru/linguistics/text/3154032

5. Зализняк, А.А. Феномен многозначности и способы его описания / А.А. Зализняк // Вопросы языкознания. – Москва, 2004. – № 2. – С. 20-45.

6. Контекстные синонимы [электронный ресурс] : новостной ресурс. – Режим доступа: https://www.anews.com/p/112585800-kontekstnye-sinonimy-ehto-primery-predlozhenij-s-kontekstnymi-sinonimami

7. Белькова, А.Е. Контекстуальные синонимы как стилистическое средство выразительности в языке поэзии В. А. Мазина / А.Е. Белькова // Вестник НВГУ. – 2014. – №4. – С. 1-7.

8. Толдова, С.Ю. Современные проблемы и тенденции компьютерной лингвистики / С.Ю. Толдова, О.Н. Ляшевская // Вопросы языкознания. – М.: Российская академия наук. – 2014. – №1. С. 120-145.

9. Найдёнова, К.А. Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований / К.А. Найдёнова, О.А. Невзорова // Ученые записки Казанского университета. Серия Физико-математические науки. – 2008. – №4. – С. 5-24.

10. Hindle D. Noun classification from predicate-argument structures / D. Hindle // Proc. Of the Annual Meeting of the Association for Computational Linguistics. – 1990. – P. 268-275.