Построение семантико-синтаксической модели текстов для определения их смысловой близости

А.С. Поречный, email: alex.porechny@mail.ru

НИУ «Московский авиационный институт», кафедра 319

Аннотация. *На сегодня автоматическая обработка текстов на естественном языке является востребованной и не в полной мере решенной задачей. Это подтверждается появлением в последнее десятилетие различных систем по автоматизации обработки естественного языка. Одной из задач обработки естественного языка является определение смысловой близости текстов. Для выделения смысла или семантики текста используется семантический или семантико-синтаксический анализ, причем такой анализ может проводится на разных единицах языка (словах, грамматических формах слова, словосочетаниях, понятиях, предложениях или наборах предложений). Задача определения близости текстов схожа с задачами, на решение которых направлены поисковые системы, системы вопрос-ответ, электронные помощники, системы перевода и т.д. Результаты исследований в области машинного перевода из-за своей специфики являются перспективными в использовании для решения задачи определения близости текстов. В статье предлагается семантико-синтаксическая модель текста, основывающаяся на совокупности метрик, которые рассчитываются по полученному семантико-синтаксическому представлению текста, и представляющая собой вектор признаков, характеризующих различные смысловые аспекты текста. Набор метрик может настраиваться в зависимости от решаемой задачи, что позволит применять предлагаемую модель для определения смысловой близости текстов с разной целью.*

Ключевые слова: *определение семантической близости текстов, естественный язык, семантико-синтаксическая модель, обработка текста, семантико-синтаксический анализ, семантический анализ.*

Введение

С появления первых ЭВМ человечество пытается описать естественных язык, «научить» ЭВМ обрабатывать тексты и речь на естественном языке и автоматизировать различные задачи: распознавание и анализ текста и речи, перевод, извлечение информации и т.д.

Первые попытки, проведённые в 70-х годах прошлого века, показали, что на ограниченной выборке словоформ и других ограничениях, обработка текста возможна. На сегодня же, обработка естественного языка встречается повсеместно в поисковых системах, вопросно-ответных электронных помощниках, автоматических переводчиках, системах определения заимствования, генерации текстов и т.д. Обработка естественного языка (NLP) ЭВМ практически общепринятым требованием к информационным системам.

Однако, уровень семантической обработки естественного языка все еще остается недостаточным, это обуславливается несколькими причинами, в т.ч. неоднозначностью различных единиц языка, начиная со слова, заканчивая набором предложений.

Появление в последнем десятилетии разнообразных инструментов семантической обработки языка, подтверждает отсутствие и востребованность алгоритмов такого рода обработки.

Одной из задач, где семантическая обработка является важнейшим элементом, является оценка смысловой близости текстов. Данная оценка востребована для структуризации неструктурированных данных (текстов, документов и т.д.), для повышения качества решения задач типа «вопрос-ответ», поиска плагиата, поиска текстов по запросу и т.д.

Семантико-синтаксической анализ

Выделение семантики текста – это процесс определения семантических отношений, на основе которых формируется семантическое представление текста. На основе полученного семантического представления возможно определить смысловую нагрузку текста, выделить затронутые темы, определить контекст использования отдельных смысловых единиц языка, что позволяет устранить их неоднозначность. Также семантическое представление текста позволяет определять смысловую близость различных текстов и их отдельных предложений.

В наиболее общем смысле семантический анализ текста представляет собою процесс выделения семантики из текста. Под семантикой обычно подразумевают содержание, информацию, передаваемую текстом или какой-либо единицей языка (словом, грамматической формой слова, словосочетанием, предложением или набором предложений) [1]. Ожидаемым результатом верного семантического анализа является получение смысла текста, который определяется не только текстом, но и контекстом, а также личным опытом, мировоззрением, привычками, интуицией и мироощущением автора текста и т.д., наиболее наглядным примером такой зависимости является аллегория.

Семантически анализ может проводится на разных единицах языка (словах, грамматических формах слова, словосочетаниях, понятиях, предложениях или наборах предложений), при этом в зависимости от единиц языка могут решаться различные задачи.

Семантический анализ на уровне слова может включать устранение двусмысленности или неоднозначности слова, т.е. в зависимости от контекста определяется форма из набора омоформ слова, устранение лексической неоднозначности и т.д.

Семантический анализ на уровне слов, словосочетаний и понятий может включать определение контекстных синонимов, установление связей между словами и местоимениями, которые замещают их, определение идиоматических выражений и т.д.

Семантический анализ на уровне набора предложений может включать в себя определение отношений между предложениями, определение тем или предметных областей, а также установление ссылок между предложениями, отсылок в контекст, в котором используется анализируемый текст или на опыт читателя.

В реальных, а не теоретических системах, не всегда используется полное построение семантического представление, в основном строится синтаксическое представление, реже семантико-синтактические представления текста или его отдельных предложений.

Так, в системе АОТ производится поверхностный семантический анализ на основе синтаксического анализа, который строит синтаксические узлы и отношения между ними [2]. В проекте ЭТАП-3 используется теория «Смысл ↔ Текст», где семантический анализ основан на толково-комбинаторном словаре [3]. В системе Nalaps семантический анализ происходит одновременно с синтаксическим с помощью механизма расширенных сетей переходов [4].

Таким образом, зачастую, нет необходимости полностью строить семантическое представление, достаточно использовать смешанные варианты. Так возможно определение схожести текста на основе слов, линейном поиске на основе пунктов, которые используют поисковые системы, стилистический анализ и т.д. [5].

Использование семантической обработки в схожих задачах с определением близости текстов

В последние годы активно разрабатываются и внедряются различное программное обеспечения, которое направлено на анализ и обработку информации на естественном языке. Так в 2013 году была представлена Word2vec от Google – совокупность моделей нейронной сети, направленные на векторизацию слова на естественном языке, параллельно с этим появляются системы преобразования речи в текст с

последующим анализом, такие как: DeepSpeech, MycroftAI. SemanticVectors, GloVe.

С 2007 года разрабатываются и обучаются модели SemanticVectors, поддерживающиеся участниками из Техасского университета, Технологического университета Квинсленда, Австрийского научно-исследовательского института искусственного интеллекта, Google Inc. и ряда других организаций и частных лиц.

В 2016 году Google презентовал голосового помощника, в основе которого лежит обработка речи, выделение семантической информации для последующей обработки. В 2017 году Яндекс представил голосового помощника для русского языка.

Задача определения близости текстов схожа с задачами, на решение которых направлены поисковые системы, системы вопрос-ответ, электронные помощники, системы перевода и т.д.

Необходимость в оценке близости текстов именно на семантическом уровне, а не на уровне графематики (например, алгоритм Шинглов) или морфологии, можно проследить на развитии поисковых систем. Первые поисковые системы использовали для индексации страниц ключевые слова, однако, быстро было установлено, что ключевые слова не всегда соответствуют содержимому страницы. Самостоятельное автоматическое выявление ключевых слов и словосочетаний повышало качество индексации. Внедрение алгоритма, который основывался на использовании статистических данных (поведение пользователя при выдаче различных вариантов поиска) повысило качество поиска, однако, требуется время для набора статистики для новых запросов, а также возникает задача по сопоставлению различных запросов, которые имеют одинаковую или схожую семантику. Для решения таких задач разрабатываются различные алгоритмы, например, алгоритм Палех от Яндекса, однако, такие алгоритмы все еще не совершенны. К тому же такие алгоритмы направлены на сопоставление большого текста на странице и небольшого запроса (запросы редки бывают более одного предложения).

Анализ систем, которые решают задачу вопрос-ответ в электронных помощниках, показал, что такие системы используют статистику или машинное обучение, например, модели SemanticVectors, которое также основываются на выделение закономерностей.

Популярные системы антиплагиата зачастую используют алгоритм Шинглов и редко применяют семантико-синтаксический или семантический анализ [6, 7, 8].

Семантический перевод текста на разные языки – это обратная задача определения близости текста. По сути семантический перевод

означает перевод семантики текста с одного языка на другой с помощью схожих смыслов слов, словосочетаний, идиом, конструкций предложений и т.д. Необходимо отметить, что слова из разных языков редко имеют тождественную семантику, обычно они имеют некое пересечение смысла в определенном контексте или в сочетании с определенными словами.

Существует подход при переводе текста, который базируется на идее определения текстуальных понятий и их концентрации в тексте. Текстуальные понятия выделяются на основе критериев «единства концепта» (концептуальной связности) и «единства ситуации» (ситуативная связанность), далее в ходе анализа текста выделяются из текста семантически связанные комплексы, а затем определяются, как они связаны друг с другом [9].

Такой подход позволяет при переводе учесть предметные отношения, которые определяются спецификой умственной ситуацией, связанной с культурой языка оригинального текста. Под предметными отношениями имеется ввиду отношения между концептами, которые устанавливаются не внутри семантически связанных комплексов (пример семантического комплекса «молчание/немота/тишина»), а между такими комплексами.

В работе [10] вводится математическая мера близости текстов, которая использует описанные выше текстуальные понятия и их концентрацию в тексте.

Таким образом, наработки для решения задачи машинного перевода из-за своей специфики являются перспективными в использовании для решения задачи определения близости текстов.

Семантико-синтаксическая модель текста для определения смысловой близости

Одной из проблем построения и использования семантических моделей текста является то, что в целом семантический анализ и его результаты сильно отличаются в зависимости от решаемой задачи: поиск, сокращение объема текста (реферирование и аннотирование), сравнение текстов и т.д. Этим обуславливается и наличие множества различных семантических моделей, позволяющих формализовано описать различные аспекты текста.

Но одновременное построение и совместное использование сложных семантических моделей ресурсозатратно и проблематично, особенно для практического использования на произвольных текстах, а не только на ограниченном материале в исследовательских целях.

Для построения семантико-синтаксической модели предлагается использовать результат работы модуля семантико-синтаксического

анализа фреймворка TAWT, который на вход принимает текст в символьном виде, а на выходе возвращает семантико-синтаксическую сеть (сети) текста, в вершинах, которого находятся формы слова с морфологическими характеристиками, а ребрами является семантико-синтаксическая связь.

При этом в модели также включает графы, которые описывают связи между предложениями и абзацами. Такие связи возникают на основе использования тех или иных понятий (понятие, определенное Г.Г. Белоноговым), таким образом возникает контекст связности (когезии) между предложениями в рамках текста. А Использование двухуровневого представления, описанного в работе [11], позволит значительно увеличить коэффициент когезии, за счет абстракций предоставляемым верхнем уровнем данного представления.

Такая модель дает возможность использовать различные метрики и мер, которые ранее совместно не рассчитывались. Для решения отдельных задач анализа текста, в частности сравнения текстов используются различные метрики, например:

1. Сравнение частоты встречаемости слов в тексте.
2. Сравнение число совпадающих по смыслу элементов (от слов до фраз) к общему количеству таких элементов.
3. Выделение n-грамм лексических отношений и их сравнение.
4. Уникальность использования различных терминов друг с другом, учет расстояния между ними.
5. Использование взаимной информации (PMI) для вычисления подобия между парами слов, и латентно-семантического анализа.
6. Сравнение между семантическими схемами текстовых пассажей и семантических классов.
7. Инвертированная длина пути.
8. Метод контекстного окна на данных Google N-Grams.
9. Расстояния между ядрами (темами) текста.
10. Использование лексико-синтаксические шаблона для извлечения и разметки предложений, содержащих слова, находящихся в гипо-гиперонимеческих отношениях.
11. и т.д.

Полученные метрики позволят составить вектор признаков для текста, который возможно применять для определения близости текстов. Метрики отличаются не только по характеристикам текстов, которые позволяют учитывать, но и по сложности и скорости их вычисления, поэтому в зависимости от решаемой задачи используемый набор метрик может быть настраиваемым.

Использование совместно различных метрик должно повысить качество определения близости текстов, однако, наличие нескольких метрик и мер ставит задачу по установлению веса влияния той или иной меры. Для решения этой задачи может быть использована нейросеть для расстановки корректных коэффициентов (весов), на основе которой будет формироваться ответ о смысловой близости текстов.

Заключение

Обработка текста на естественном языке зачастую не использует семантико-синтаксический или семантический анализ. Востребованность новых алгоритмов и подходов в этой области подтверждается появлением большого количество моделей, алгоритмов обработки текста и речи, которые направлены на определение смысловой нагрузки текста.

При анализе текста зачастую необходимо знать контекст для разрешения неоднозначности, причем частично неоднозначность можно снять для одной смысловой единицы языка, если возможно проведение следующих этапов анализа с учетом нескольких вариантов данной единицы языка. Часто возможно делать некоторые допущения при устранении неоднозначностей, используя статистику употребления и сочетаемости смысловых единиц языка между собой. Практическая же реализация различных алгоритмов анализа текста показала, что не всегда необходимо полное семантическое представление текста, а возможно поверхностное построение семантической сети.

Исследование определения близости текстов показало, что данная задача схожа с задачами, решаемыми в поисковых системах, системах типа «вопрос-ответ», электронных помощниках, системах перевода и др. История развития поисковых систем показывает возрастающую необходимость определения семантики. При этом, наработки таких систем всегда могут быть применимы для задачи определения семантический близости текстов, т.к. поисковый запрос как правило не превышает одного предложения. Наработки систем типа «вопрос-ответ» в большей степени используют статистические методы, а не семантический анализ текста.

Наработки по решению задачи машинного перевода можно считать перспективными для использования при разработке алгоритмов или моделей определения семантической близости текстов, т.к. в обеих задачах требуется выделение семантики текста.

В статье предложена семантико-синтаксическая модель текста, основанная на совокупности метрик, которые составляют вектор признаков, характеризующих различные аспекты смысловой нагрузки

текста, этот набор метрик может настраиваться в зависимости от решаемой задачи.

Список литературы

1. Большая российская энциклопедия [Электронный ресурс]: энциклопедия. – Режим доступа: https://bigenc.ru/linguistics/text/3546954
2. Официальный сайт системы «АОТ» [Электронный ресурс]: описание первичного семантического анализа. – Режим доступа: http://aot.ru/docs/seman.html
3. Батура, Т. В. Семантический анализ и способы представления смысла текста в компьютерной лингвистике / Т. Батура // Программные продукты и системы. – 2016. – №4 (116). – С. 45-57.
4. Путилов, Г. П. Создание автоматических тестов по текстовым пользовательским сообщениям средствами системы Nalaps / Г. П Путилов, А. С. Лебедев // Материалы ежегодной Международной конференции «Диалог 2010».
5. Бермудес, С.Х.Г.,. О методе определения текстовой близости, основанном на семантических классах / С.Х.Г. Бермудес, С.У. Керимова // Инженерный вестник Дона. –2016. –№4 (43).
6. Официальный сайт «eTXT» [Электронный ресурс]: eTXT проверка уникальности. – Режим доступа: https://www.etxt.ru/subscribes/etxt-antiplagiat/
7. Официальный сайт «Advego Plagiatus» [Электронный ресурс]: описание Advego Plagiatus. – Режим доступа: https://advego.com/plagiatus/
8. Чехович, Ю. Плагиат в научных статьях: трудности обнаружения перевода. / Ю. Чехович, Р. Кузнецова, О. Бахтеев // Университетская книга. –2017. –№2. –С. 66-67.
9. Марьин, Д. В. Филологический метод представления перевода секвенции текстов (На материале сборника переводов рассказов В. М. Шукшина "V. Shukshin. Stories from a siberian vill // Дис. канд. филол. наук: – Кемерово, 2004. –255с.
10. Сулейманов, А. Ш. Семантическая близость и семантические расстояния между текстами / А. Ш. Сулейманов // Системні технології. – 2007. – №10 (30). –С. 132-136.
11. Балакирев, Н.Е. Использование двухуровневого семантического представления в открытой системе автоматизированного анализа текстов / Н.Е. Балакирев, Е.В. Добрышина // И74 сборник материалов XVI Международной научно-методической конференции «Информатика: проблемы, методология, технологии». Т. 4. Воронеж: Издательство «Научно-исследовательские публикации», – 2016. –С. 184-189.