Применение алгоритмических методов и машинного обучения для решения задач компьютерной лингвистики

Е. В. Полицына, email: kathrin.beaver@mail.ru
С. А. Полицын
М. В. Зеленова

Московский авиационный институт (национальный исследовательский университет)

Аннотация. *Рассмотрено применение алгоритмических методов и методов машинного обучения в компьютерной лингвистике, проанализированы основные проблемы, связанные с компьютерным анализом естественного языка. Предложены методы совместного применения классических алгоритмических методов и методов машинного обучения для решения некоторых практических задач компьютерной лингвистики, проанализированы полученные результаты.*

Ключевые слова: *компьютерная лингвистика, машинное обучение, автоматизированный анализ текста, векторизация текста.*

Введение

Компьютерная лингвистика в настоящее время является одной из быстроразвивающихся областей науки, что обусловлено огромной потребностью в средствах автоматического анализа текстовых данных. Это доказывает как большое количество новых предлагаемых методов, количество публикаций в области автоматической обработки естественного языка (Natural Language Processing, NLP), так и большая востребованность средств компьютерной лингвистики в смежных областях, в первую очередь при разработке различных приложений даже в тех сферах, где на первый взгляд работа с текстом не требуется. Однако практически в любом современной информационной системе присутствуют, например, возможности поиска текстовых данных (имен, фамилий, товаров и др.), средства фильтрации данных и т. д. - всё это требует применения средств NLP, причем, зачастую, с высокими требованиями по скорости работы и точности результата. При этом пользователи, исходя из своего понимания результата анализа текста так, как это делает человек, по умолчанию ожидают достаточно “интеллектуальной” работы информационных систем, например, что при поиске слов будут найдены все их формы, а не будет осуществлен поиск подстроки по точному совпадению. Кроме того, крупные поисковые системы дают пользователям возможность поиска по

источникам на другом языке, выводят результаты по нечеткому совпадению и т.д., т.е. активно используют возможности NLP. Все это делает использование алгоритмов и средств автоматической обработки текстов неотъемлемой частью любой информационной системы, работающей с текстовой информацией.

1. Анализ проблем NLP

При первом взгляде на задачи автоматической обработки естественно-языкового (ЕЯ) текста у разработчиков информационных систем часто складывается мнимое ощущение их простоты, т.к. их простейшие решения для частных случаев конкретной предметной области, конкретного примера на конкретном языке действительно могут сводиться к использованию стандартных средств обработки строк в любом языке программирования. Однако, использование подобных решений даже на чуть более широком наборе данных, делает очевидным их неработоспособность для огромного количества нюансов ЕЯ.

ЕЯ текст является сложным объектом анализа, т.к. содержит множество фраз, допускающих двоякое толкование, имеет слова и выражения в переносном значении или даже эвфемизмы, различные выразительные средства, допускает пропуски частей слов и предложений, которые по мнению автора текста являются “очевидными” или общепринятыми и т.д. Это сильно затрудняет автоматизацию его анализа в целом и снижает качество результатов на каждом из этапов анализа (графематический, морфологический, семантико-синтаксический).

Главной же проблемой, с которой сталкивается любой исследователь ЕЯ текста является богатство средств выражения, которое порождает неоднозначность создаваемого текста, правила составления отдельных слов, предложений и текста целиком сложны, имеют множество нюансов, зависят от культурного контекста и в целом допускают различные способы выражения одной и той же мысли. Эта многозначность также проявляется на всех уровнях анализа, и зачастую, чтобы машине “понять” единицу языка, т.е. разрешить многозначность надо либо обладать какими-то дополнительными знаниями об авторе, предметной области, контексте описываемого события и др. либо, в более простом случае, попытаться разрешить эту неоднозначность на следующих этапах анализа [1], т.к. неоднозначности в ЕЯ могут быть разных видов, например:

- Фонетическая (“скрип колеса” или “скрипка-лиса”).
- Морфологическая (“село”, “мыло”, “мой”, “три”).
- Лексическая (“рожа”).

Синтаксическая (“мужу изменять нельзя”, “критика ученого”). Учет всех особенностей ЕЯ текстов при его автоматическом анализе делает крайне сложной решение глобальной задачи “научить машину понимать текст”, однако без ее решения невозможно в полной мере реализовать в информационных системах действительно “интеллектуальную” обработку текста. Для решения и практических, и исследовательских задач NLP с учетом существующих проблем NLP в настоящее время сложилось два основных подхода: подход, основанный на использовании алгоритмических методов, наборов правил для разрешения многозначностей различных видов и словарей и подход, основанный на применении статистических методов, в частности методов машинного обучения (Machine Learning, ML).

2. Формы представления текста для его автоматического анализа

Однако, для применения любого из подходов необходимо сначала привести ЕЯ текст к “компьютерному” виду, т.е. преобразовать исходный текст в “понятные” компьютеру структуры данных в терминах того или иного языка программирования.

Стандартное для любого языка представление текста в виде строки, состоящей из символов, для большинства задач автоматического анализа текста является бессмысленным. Поэтому первым шагом при решении задач NLP является разделение текста на его отдельные элементы - токенизация. В зависимости от решаемой задачи текст разбивается на какой-либо набор элементов (символы, слова, н-граммы(группы слов), словосочетания, предложения, абзацы). Обычно для решения практических задач представляют интерес слова и предложения.

Однако, даже с этим, на первый взгляд простым этапом, могут возникнуть трудности ввиду неоднозначного написания слов, символов отделения слов друг от друга, определения границ слов и предложений и др. Например, в германских языках принято объединять группы существительных в одну единицу, а в китайском - вообще не выделяются отдельные слова при письме.

При выделении границ предложений и простых предложений внутри сложных отдельные предложения бывает затруднительно отделить от однородных членов, а наличие сокращений в текстах часто способствует неверному определению границ отдельных предложений.

При анализе современных письменных Интернет-источников пунктуация часто используется не по прямому назначению, а для придания эмоциональной окраски тексту (смайлики, эмодзи и т. д.) [2], создавая новые элементы текста, которые часто не менее важны для анализа. Корректное выделение границ единиц текста сильно влияет на

качество результатов работы как методов алгоритмического подхода, так и подхода на основе методов машинного обучения.

3. Алгоритмические методы решения задач КЛ

Алгоритмические методы обработки текста исторически возникли первыми, за десятилетия исследований созданы модели, выдвинуто множество гипотез и построенных на них алгоритмов решения задач КЛ. Например, лексическая и морфологическая неоднозначность в программах Synan, Trigra и Accopost снимается с точностью около 90% [3]. Однако, часто качество работы алгоритмических методов чем-либо ограничено, например, метод, реализованный в системе русско-английского и англо-русского фразеологического машинного перевода RETRANS, позволяет снять омонимию слов с 99% точностью, но ограничением является базовый набор структур, расширение которого требует большого количества ручной работы [4].

Наличие проблемы однозначной обработки ЕЯ влечет за собой создание разнообразных инструментов для анализа текста, как небольших, решающих отдельные задачи (Lemmatizer, Greeb, Stemka, pyMyStem3, TreeTagger, Text Summarization, Tools4noobs и многих других), так и комплексов инструментов: AOT, GATE, LingPipe, UIMA, Texterra, среди коммерчески успешных систем стоит выделить ядро поисковых систем Google и Яндекс, инструменты компаний IBM, Яндекс и ABBYY.

Алгоритмические методы нашли применение в информационных системах для автоматизации работы с большими объемами текстовых данных, в поисковых, новостных, и рекомендательные системах, а также в системах, построенных на применении автоматического анализа текста: системы антиплагиата, спам-фильтры, системы перевода и др.

Разработка новых алгоритмических методов сопряжена с высокими требованиями к знанию ЕЯ, на котором идет обработка и лингвистики в целом. Нормы и правила ЕЯ постоянно меняются и развиваются, для многих правил списки исключений не меньше самого правила и т. д. Кроме того, требуются знания в технических областях: теория алгоритмов, теория графов, математический анализ, математическая статистика, программирование и т. Д. Все это делает проблематичным реализацию не только новых, но и существующих алгоритмов NLP при разработке прикладных информационных систем ввиду высоких требований к специалистам и большой трудоемкости.

4. Применение методов машинного обучения для решения задач КЛ

Акцент на использование статистических методов для автоматического анализа текста делался еще на заре развития NLP, но невозможность количественной оценки большинства особенностей ЕЯ сместила его в сторону алгоритмических методов с использованием некоторых статистических характеристик текстов.

Новой волной мощного применения статистических методов стало машинное обучение, которого представляет собой обширный подраздел области искусственного интеллекта, изучающий методы построения алгоритмов, способных “обучаться”, т.е. в результате работы метода со временем возрастает точность выдаваемого результата. Характерной чертой ML является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Основу методов ML составляет математическая статистика, численные методы, методы оптимизации, методы теории вероятностей, теории графов [5].

Поскольку в основе методов ML лежат методы из математической статистики, они лучше всего подходят для решения задач классификации, кластеризации, прогнозирования, поиска ассоциативных правил и др., в которых допустим ответ с определенным порогом вероятности. Причем, применительно к области КЛ к группе задач классификации относятся определение тональности текста, определение «замусоренности» текста и спама, непосредственно классификация по предметным областям или другим признакам. Примером задачи кластеризации может быть группировка новостей, поиск похожих синтаксических деревьев и др.

5. Векторизация текста

Компьютерное представление текста при использовании методов ML сводится к выделению некоторого “вектора признаков”, т.е. для исходного токена текста или для текста целиком выделяется некоторый набор признаков, с использованием которого происходит дальнейшая обработка. Самый простой способ - «сумка слов» (от англ. bag of words). В этом случае вектор признаков соответствует полному словарю текстовой выборки, для каждого слова считается количество вхождений в текст и это число устанавливается на соответствующую позицию в векторе. При таком подходе даже с применением оптимизации в виде предварительной нормализации слов длина вектора признаков соответствует размеру словника текста, при это для конкретного текста, а тем более для коротких текстов этот вектор сильно разрежен (состоит практически из одних нулей), обработка таких больших векторов является очень ресурсоемкой задачей, поэтому одним из направлений

исследований в ML в КЛ является создание методов, позволяющих уменьшить размер признаков. Среди таких методы выделяются 2 группы: общие, не требующие знаний о предметной области и специальные, которые требуют дополнительной информации о тексте или области и применимы только к конкретной задаче.

Среди общих методов применяются:

- n-граммы — комбинации из n последовательных терминов для упрощения распознавания текстового содержание. Этот метод определяет и сохраняет информацию о смежных последовательности слов в тексте.
- TF-IDF – учитывает соотношение частоты встречаемости конкретного термина и частоты документа, в котором он встречается.
- Word2Vec — набор методов для анализа естественных языков на основе дистрибутивной семантики и векторного представления слов, разработан группой исследователей Google в 2013 году [6]. Сначала создается словарь, «обучаясь» на входных текстовых данных, а затем вычисляется векторное представление слов, основанное на контекстной близости. При этом слова, встречающиеся в тексте рядом, в векторном представлении будут иметь близкие числовые координаты.

Примером использования специальных методов векторизации может являться разработанное приложение для определения мошеннических сообщений социальной сети “ВКонтакте” [7]. Сообщения в социальных сетях обычно короткие, часто отклоняются от норм языка и т.д., применение обычных методов векторизации показало их невысокую эффективность для этой задачи. Однако, если включить в вектор признаков специальную информацию о сообщении (например, содержание картинок, числительных, отдельных слов из словаря), то процент верно определенных мошеннических сообщений возрос до 90%.

6. Результаты применения алгоритмических методов и методов машинного обучения при решении задач КЛ

Результаты работы методов ML сильно зависят как от выбора способа векторизации, так и от качества подготовки обучающей выборки и, особенно, от “замусоренности” исходного текста. Решить эти проблемы позволяют алгоритмические методы КЛ, которые позволяют подготовить текст для последующего применения методов ML:

- провести разбиение длинных участков текста на более мелкие (абзацы, предложения, слова);

- привести текст к единому виду: единый регистр слов, отсутствие знаков пунктуации; произвести расшифровку сокращений и их подстановку в нужной форме слов; заменить численное написание числительных на словесное и т.д.;
- привести слова к его корню (стемматизация) или к начальной форме слова: инфинитив для глагола, именительный падеж единственного числа — для существительных и прилагательных (лемматизация);
- провести “очистку” текста – удалить стоп-слова, не несущие смысловой нагрузки (артикли, междометия, союзы, предлоги и т.д.).

Предварительная подготовка текста значительно уменьшает размер вектора признаков для работы методов ML, при этом повышается их точность работы.

Методы ML могут применяться для проведения исследований в области КЛ, они показывают наилучшие результаты именно на больших массивах данных, обработка которых, зачастую, и необходима для подтверждения выдвинутых гипотез. Такой подход был использован для поиска слов-кандидатов в контекстные синонимы с одинаковыми или похожими морфологическими характеристиками.

Изначально, для поиска таких слов был применен алгоритмический подход: определялись морфологические характеристики слов в тексте в виде битовой шкалы размера 72 бита, где каждой характеристике соответствует от 1 до 4 бит в зависимости от количества вариаций ее значений. Реализация отдельных правил средствами инструментов морфологического анализа не является сложной задачей и позволяет получить точные результаты по написанному набору правил, однако создание набора сложных правил является сложной итерационной исследовательской задачей.

При векторизации вместо исходного слова использовался вектор его морфологических характеристик в численной форме, определенных с помощью алгоритмических методов (часть речи, род, число и т д.). Затем проведена их кластеризация средствами Weka 3.9.4, наилучшие результаты были получены методом HierarchicalClusterer. В результате определено, что при количестве кластеров на 1-2 меньшем, чем количество частей речи в тексте, в один класс попадают слова с похожими наборами ненулевых характеристик, а при количестве классов большем, чем количество различных частей речи в тексте, слова с разными значениями морфологических характеристик распределяются в разные классы, а в одни – с близкими их значениями.

Схожим образом была решена задача поиска одинаковых синтаксических структур. Для ее решения могут применяться чисто

алгоритмические методы, но тогда количество правил, которые надо запрограммировать будет велико, а их расширение потребует привлечения разработчика. Кроме того, рекурсивный обход синтаксических деревьев на больших объемах данных, особенно для текстов с длинными и сложными предложениями, будет занимать длительное время. С использованием же методов машинного обучения и предложенного способа векторизации текста на основе морфологических характеристик слов решение задачи занимает много меньшее время как реализации, так и обработки. Это позволяет получить большой объем данных для дальнейшего более глубокого анализа предложений со сходной синтаксической структурой.

В рамках исследования проанализированы проблемы использования методов ML в NLP, которые могут быть решены за счет совместного применения классических алгоритмических методов и методов машинного обучения. В первую очередь, для всех задач NLP нужны алгоритмические инструменты графематического и морфологического анализа, которые позволяют подготовить исходные данные, а также уменьшить размер вектора признаков. Сам вектор признаков, ориентированный на решение конкретной задачи, как и подготовка обучающей выборки в целом, требуют знаний в области алгоритмов обработки ЕЯ.

Заключение

Проведенный анализ методов решения задач компьютерной лингвистики показал, что богатство и сложность естественных языков порождают огромное количество разноплановых задач. Для их решения применяются как алгоритмические методы, так и методы машинного обучения.

Проведенное исследование позволило выделить ряд критериев для определения возможности и оптимальности применения тех или иных методов. Таким образом, хорошо по качеству результату и оптимально по времени реализации показывает себя применение методов ML при наличии большой выборки текстов с ярко проявленной закономерностью в структуре текста, его форме, лексике и т.д. Это позволяет разработчикам не углубляться в тонкости предметной области и лингвистики, что существенно сокращает время разработки и требования к специалистам ее осуществляющим.

В случае же наличия текстовых данных, отличающихся своим разнообразием или же невозможности подготовки корпуса текстов для применения методов ML, а также при исследовании различных языковых явлений в текстах, безусловно требуется создание новых алгоритмических методов.

Применение существующих алгоритмов NLP, реализованных в программных инструментах, позволяет разрабатывать новые способы векторизации текстов, исходя из знания предметной области и особенностей решаемой задачи. Это практически не способствует развитию новых методов автоматического анализа текста, однако позволяет получить хорошие результаты решения конкретной задачи за приемлемое время, что крайне важно при разработке прикладных систем.

Таким образом, применение и развитие как алгоритмических методов, так и машинного обучения, являются крайне важными и актуальными направлениями развития в области NLP, а также позволяет успешно применять достижения в области автоматического анализа текста для реализации интересных и полезных функций в самых разных информационных системах.

Список литературы

1. Politsyna E. V. The Framework for Hypothesis Verification and Analysis of Natural Language Processing for the Russian Language / E. V. Politsyna, S. A. Politsyn, A. S. Porechny // Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2018), Moscow, Russia, July 5–7, 2018. – CEUR WorkShop Proceedings, ISSN 1613-0073 Aachen, Germany, 2018. – vol. 2268. – pp. 25-33.

2. Кубратова, Ю.А. Использование смайлов в современной коммуникации / Ю.А. Курбатова // Устойчивое развитие науки и образования, №9 2018, сс 202-206

3. Сокирко, А. В. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка / А. В. Сокирко, С.Ю. Толдова // Интернет-математика 2005. - М., 2005. - С. 80–94.

4. Белоногов, Г.Г. Интерактивная система русско-английского и англо-русского машинного перевода политематических научно-технических текстов / Г.Г. Белоногов, Ю.Г. Зеленков, Б.А. Кузнецов, А.П. Новоселов и др. // Научно-техническая информация, сер. 2. - № 3, 1993. - С. 20-27.

5. 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.

6. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR. — 2013

7. Петраш, Я.П. Разработка инструмента для выявления мошеннических сообщений социальной сети «ВКонтакте» / Я.П. Петраш, Д.А. Тихонова // Материалы XIX Международной научно-методической конференции «Информатика: проблемы, методология, технологии». Россия, Воронеж, 2019, С. 1513-1518.