Семантическая декомпозиция текста для решения задачи категоризации обращений

Д. М. Лосева
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
d.m.loseva@gmail.com

***Аннотация*. В докладе приведено исследование вариантов реализации семантической декомпозиции текста и разработка алгоритма семантической декомпозиции для системы поддержки принятия решений. В докладе описано исследование существующих решений задачи семантической декомпозиции текста, описаны теоретические и технические основы проектирования алгоритма, сценарий и примеры использования. Исследованы качественные характеристики разработанного алгоритма: определены и проанализированы время работы и качество результатов.**

Ключевые слова: семантическая декомпозиция; система поддержки принятия решений; ключевые слова; категоризация обращений; лингвистические основы алгоритма; служба технической поддержки

I. Работа технической поддержки

В данной работе рассматриваются процессы, связанные с работой технической поддержки ИТ-службы.

При рассмотрении процесса управления обращениями пользователей ИТ-услуг необходимо ввести следующие понятия:

- Специалист 1-ой линии поддержки – это работник ИТ-подразделения, который принимает, регистрирует, классифицирует, закрывает обращение пользователя исполнителю, согласно зоне ответственности.
- Специалист 2-ой линии поддержки – работник ИТ-подразделения, взаимодействующий с пользователем при выполнении обращения, проверяющий корректность заполнения полей зарегистрированного обращения, при необходимости выполняет внесение корректировок, запрашивает дополнительную информацию, исполняет и закрывает обращение.

Обращение пользователя, после регистрации, категорируется специалистом 1-ой линии поддержки. Категоризация включает в себя определение типа запроса, услуги, затронутого элемента конфигурации, срочности и приоритета, а также исполнителя (это может быть как отдел/группа, так и отдельный сотрудник) обращения в системе управления обращениями.

Так как пользователь формирует обращение в произвольной форме, без выбора категорий, то запросу автоматически присваивается категорий «запрос на обслуживание». Позже специалист 1-ой линии меняет категорию, если есть такая необходимость.

В связи с тем, что обращений с каждым годом становится все больше, а специалистов 1-ой линии технической поддержки обычно не больше 4-х человек на филиал, целесообразно рассмотреть возможные способы «облегчить жизнь» работникам и уменьшить количество ошибок, совершаемых из-за человеческого фактора. В данном случае эта роль выполняется системой поддержки принятия решений, которая решает задачу категоризации обращений пользователей и принятие управленческих решений в зависимости от категории обращения.

II. Алгоритм семантической декомпозиции

Классификаций или категоризацией называют процесс отнесения входных текстовых документов к одной из групп данных (классу или категории).

Классификацию и категоризацию нужно отличать друг от друга. Классификация заключается в определении принадлежности документа некоторому классу. Под классом в данном случае понимается множество документов, обладающими определенными свойствами, признаками, характеристиками. Границы класса определены достаточно точно: документ принадлежит классу, если он обладает достаточным числом признаков, характерных для этого класса.

Задача категоризации является менее определенной, так как категория ограничивает количество признаков некоторыми общими свойствами документов и связями между ними. Границы классов здесь являются нечеткими. Категории обычно формулируются не формально, а только в сравнении с другими категориями

В общем случае категоризация является более сложной процедурой по сравнению с классификацией. В категоризации помимо отнесения документа к какой-либо группе документов, требуется определить сами эти группы, т. е. сформировать категории.

Если рассмотреть общую структуру алгоритма категоризации (рис. 1), можно понять, что алгоритм семантической декомпозиции будет находиться в самом его начале, среди процедур предобработки текста.

Рис. 1. Общая структура алгоритма категоризации

Рассмотрим этапы, предшествующие работе категоризатора, так как алгоритм семантической декомпозиции будет встроен именно в эту часть системы.

Целью предварительной обработки текста является выделение в качестве признаков документа всех значимых слов. В ходе предварительной обработки переводят все буквы к одному регистру (чаще к нижнему) – нормализация, удаление пунктуации, цифр, пробелов, разбивают текст на более мелкие части – токенизация, удаляют так называемые стоп-слова (союзы, предлоги, артикли и пр.), размечают текст по частям речи – морфологический анализ.

К числу классических процедур предварительной обработки относятся:

- Нормализация и токенизация. В самом начале обработки текст приводится к единообразному виду (единый регистр слов, отсутствие знаков пунктуации, расшифрованные сокращения, словесное написание чисел и т.д.).
- Лемматизация и стемминг. Эти процедуры позволяют не различать формы одного и того же слова.
  - лемматизация – приведение каждого слова в документе к его нормальной форме, под которой понимается использование именительного падежа, единственного числа для существительных; именительный падеж, единственное число, мужской род для прилагательных; использование инфинитива для глаголов, причастий, деепричастий;
  - стемминг – отбрасывание изменяемых частей слов (главным образом, окончаний).
- Отбрасывание стоп-слов. Речь идет о предлогах, союзах, числительных, местоимениях, вводных словах. Число таких слов обычно варьируется в пределах нескольких сотен.
- Отбрасывание редких слов, так как они не имеют принципиального значения в тексте.
- Выделение ключевых фраз, т. е. словосочетаний, являющихся устойчивыми оборотами или терминами в данной предметной области [1].

Для построения числовой модели текста используется индексация документов. Самые распространенные модели индексации:

- модель bag-of-words. Здесь документ – многомерный вектор слов и их весов в документе;
- модель индексации Word2vec. Здесь как вектора описываются все слова;
- модель индексации на n-граммах, под которыми понимается последовательность соседних токенов [2].

Этап извлечения признаков из текстовых документов необходим для сопоставления каждому документу набора характеристик, описывающего документ. Основная идея данного этапа состоит в том, чтобы независимо ранжировать токены в соответствии с определенным индексом оценки (обычно это частота встречаемости токенов, зависимости токенов) и выбирать токены с наивысшими баллами [3]. Часто для этого применяются алгоритмы выделения ключевых слов, выделяющие наиболее значимые слова в тексте.

Исходя из вышеизложенной информации, становится ясно, что разрабатываемый алгоритм семантической декомпозиции на ключевые элементы является элементом этапа предобработки.

На вход алгоритму подается «сырой» текст, а на выходе получается список слов или ключевых элементов исходного текста. Этот список – основа для алгоритма выделения ключевых слов. Под ключевыми словами понимаются важные слова или фразы, позволяющие выявить тематику текста.

Не все входящие в состав ключевых фраз слова являются ключевыми. Проблемой здесь является то, что выделение отдельных ключевых слов не всегда позволяет выразить основной смысл содержимого. В этом смысле лучше выделять ключевые фразы [4].

Таким образом, из данного подраздела следует, что из списка ключевых элементов во время работы алгоритма выделения ключевых слов, следующего после алгоритма декомпозиции, будут сформированы ключевые слова (термины), словосочетания или фразы.

Далее подробнее о ключевых элементах.

Как было сказано выше, из-за трудности выражения основного смысла текста отдельными ключевыми словами, стала необходимым возможность выделения ключевых словосочетаний или фраз. Для этого в данной работе предусмотрено расширение понятие ключевого элемента от слова (униграммы) до словосочетания (фразы, n-грамма).

Таким образом, под ключевыми элементами в работе понимаются как отдельные слова, так и словосочетания или n-граммы. Также в данной работе будет употребляться понятие токена в качестве синонима понятию ключевого элемента.

Итак, понятие ключевого элемента обозначено. Следующим шагом работы будет определение понятия семантической декомпозиции, начиная с более общего понятия – семантического анализа.

Целью семантического анализа текста является оценка слов или фраз, которые определяют смысл текста.

Методы анализа текста можно разделить на две группы:

- лингвистический, т.е. определение смысла текста по его семантической структуре;
- статистический, т.е. определение смысла текста по частотному распределению слов в тексте.

Типовые задачи здесь:

а) поиск ключевых слов;

б) поиск цепочек ключевых слов;

в) выявление слов, определяющих контекст текста;

г) подготовка аннотаций к тексту.

Под понятием семантической декомпозиции, в свою очередь, понимают алгоритм, разбивающий значения предложений, фраз или понятий на менее сложные понятия.

В лингвистических экспертизах присутствует огромное множество вариантов компонент, на которые может разбиваться текст. Однако в рамках данной

работы будет вестись исследование семантической декомпозиции на слова и фразы (словосочетания). Поэтому здесь будут рассмотрены три направления работы алгоритма, исходя из типа единиц декомпозиции текста:

- семантическая декомпозиции текста на слова;
- семантическая декомпозиция текста на словосочетания;
- семантическая декомпозиции текста на слова и словосочетания (смешанная декомпозиция).

Схема построения алгоритма декомпозиции представлена на рис. 2.

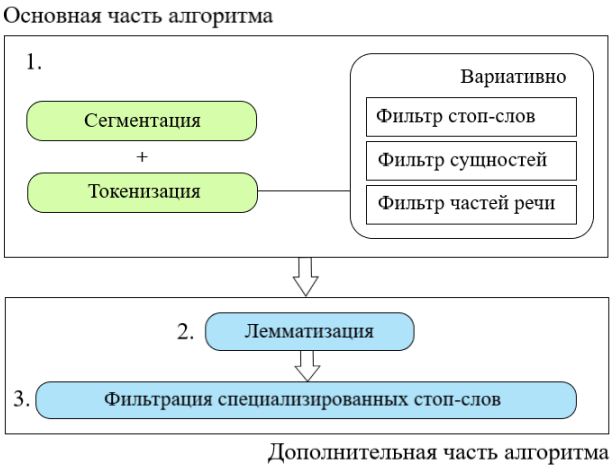

Рис. 2. Схема построения алгоритма декомпозиции текста на ключевые элементы

Процедуры основной части алгоритма декомпозиции (рис. 2):

1. Сегментация предложений: текст делится на предложения. Это вспомогательная процедура для дальнейшего шага. Она нужна для более сложной семантической обработки (учет границ предложений).

2. Токенизация: текст делится на токены. Это основная процедура. Здесь происходит декомпозиция текста на ключевые элементы (токены).

Перед началом работы алгоритма пользователю будет предложено выбрать вариант нужной ему декомпозиции:

1) Чему равно значение n?

2) Быстрая декомпозиция (по умолчанию – с фильтрацией стоп-слов и цифр) или с выбором критериев:

- декомпозиция без фильтрации;
- декомпозиция с фильтрацией стоп-слов и цифр;
- декомпозиция с фильтрацией именованных сущностей;
- декомпозиция с отбором токенов, содержащих определенные части речи.

Помимо основной части алгоритм имеет реализованные дополнительные процедуры обработки токенов. Они могут быть полезны в будущем для выполнения дальнейших шагов задачи категоризации:

- Поиск именованных сущностей поможет фильтровать имена и фамилии клиентов, которые могут идентифицироваться важными словами в тексте заявки, а по своей сути не иметь смысла.
- Лемматизация приводит токены в начальную форму. Это означает, что при необходимости подсчета частоты встречаемости токенов в тексте различные словоформы будут посчитаны вместе, как разные формы одного слова, а не как разные слова. Соответственно, вес данного слова возрастет.
- Определение роли токена и зависимостей в предложении может быть также применимо для определения веса токена при необходимости использования семантических данных.
- Поиск синонимов к заданным словам может быть применим при подготовке размеченных данных в случае, если данных будет мало.
- Фильтрация (удаление стоп-слов и цифр), определение частей речи и поиск n-грамм, состоящих из определенных частей речи, подключены к основной части алгоритма и помогают получать более чистый результат, что сокращает время работы и увеличивает точность следующих шагов алгоритма категоризации.

III. Математические основы алгоритма

Рассмотрим процесс фильтрации подробнее.

Декомпозиция текста на n-граммы основывается на декомпозиции на униграммы. Данный процесс изображен на рисунке 3 ниже:

Рис. 3. Схематическое описание логики поиска униграмм, биграмм, триграмм

На рис. 3 изображена логическая схема формирования n-грамм. Глядя на него, становится понятна важность фильтрации стоп-слов (и других ненужных символов, которые могут быть распознаны как токены) для процесса декомпозиции на n-граммы, так как любое стоп-слово, стоящее в середине предложения, при выделении биграмм появится в двух токенах, при выделении триграмм – в трех и так далее. Так, фильтрация контролирует и размерность выходных данных, и сохраняет пользу процесса предобработки.

Таким образом, выше подробно описаны основные моменты касаемо теоретической и практической части построения алгоритма семантической декомпозиции текста: определены основные и вспомогательные процедуры алгоритма, их назначение, необходимость. Основными процедурами алгоритма являются сегментация (вспомогательная) и токенизация (основная). Алгоритм также содержит ряд дополнительных процедур обработки токенов.

Для реализации алгоритма был выбран язык Python версии 3.10 с использованием инструмента SpaCy для для

разработки алгоритма декомпозиции текста на ключевые элементы. Было проведено исследование свойств программного продукта, таких как время работы алгоритма и качество выходных данных. В ходе исследования было выявлено, что время работы алгоритма зависит от размера входных данных. Качество результатов алгоритма соответствует ожидаемым. Также была проведена успешная проверка соответствия программного продукта выставленным требованиям.

IV. Заключение

По результатам работы основным направлением для дальнейшей научной деятельности в этой прикладной области выбрано исследование влияния типа ключевого элемента при синтаксической декомпозиции текстов на качество результатов системы категоризации. Базой для этого исследования послужит разработанный в данной работе алгоритм.

Список литературы

[1] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: Издательство МГУ имени М. В. Ломоносова, 2011.

[2] Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. 2017. №1 [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/metody-avtomaticheskoy-klassifikatsii-tekstov.

[3] Шагиахмитов Д.Р. Методы извлечения признаков из текстовых документов // E-Scio. 2020. №4 (43) [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/metody-izvlecheniya-priznakov-iz-tekstovyh-dokumentov

[4] Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. (№) 19. С. 85-93. [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov/viewer