МОДИФИКАЦИЯ МЕТОДА РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ В ОБЛАСТИ БИОМЕДИЦИНЫ

Пашук А. В., Гуринович А. Б., Кузнецов А. П., Смирнов В. Л.
Кафедра информатики, кафедра вычислительных методов и программирования, кафедра систем управления, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: pashuk@bsuir.by, gurinovich@bsuir.by, kuznap@bsuir.by

Разрешение лексической многозначности (Word Sense Disambiguation) является промежуточной задачей в процессе поиска и извлечения информации, представляющей собой проблему выбора правильного смысла неоднозначных слов. В данном исследовании предлагается модификация предложенного ранее алгоритма [3] с целью увеличения точности определения верного значения целевого термина.

Введение

Количество научных статей в области биомедицины стремительно увеличивается (рисунок 1), а значит, что все большую актуальность приобретают вопросы качества поиска и структурирования информации. Увеличение объемов информации также замедляет получение новых знаний.

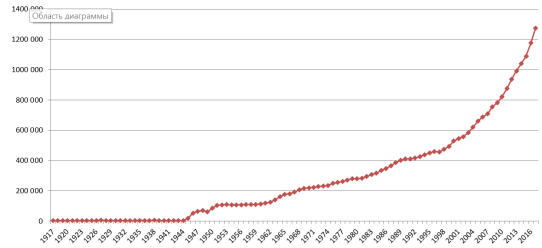

Рис. 1 – Количество публикаций, опубликованных на MEDLINE/PubMed по состоянию на 2018 год [1]

Качество поиска напрямую зависит от степени понимания машиной обрабатываемых научных текстов. Одной из проблем понимания неструктурированной текстовой информации, является проблема разрешения лексической многозначности (Word Sense Disambiguation, WSD). В общем смысле WSD является проблемой классификации – это процесс отнесения слова или фразы к определенному значению, которое отличается от других значений, которые потенциально может принимать это же слово (фраза). Обычно используется одно из двух предположений: слово может принимать только одно значение в рамках рассматриваемого документа, либо слово принимает определенное значение в рамках контекста – соседних слов, предложений.

Проблему разрешение лексической многозначности можно сформулировать следующим образом: дан документ A, содержащий термин t_k из словаря T. Любой термин t_k может быть отнесен с минимум одним из значений S_{ki} из словаря S_k. Задача состоит в том, чтобы определить наиболее вероятное значение s_{k*} для термина t_k, используемого в документе A.

Подробный обзор подходов к разрешению лексической многозначности приведен в обзорной статье [2], в которой автор предлагает использовать три основные категории алгоритмов: основанные на обучении с учителем, основанные на знаниях и основанные на обучении без учителя. Первая категория подходов использует размеченный набор данных с дополнительной информацией, полученной из контекста вокруг термина, для построения моделей машинного обучения, которые предсказывают правильный смысл для заданного контекста.

Подходы, основанные на знаниях [5], [6], не используют какой-либо размеченный корпус данных, а опираются исключительно на словари, такие как Unified Medical Language System (UMLS), которые содержат краткие определения различных смыслов термина и соответствующих синонимов. В подходах без учителя используются методы, основанные на тематическом моделировании [7], для устранения неоднозначности. Системы, использующие такие методы иногда называют Word Sense Discrimination, т.к. происходит группировка нескольких употреблений заданного термина в кластеры, где каждому кластеру соответствует определенное значение целевого слова. Т.е. в этом случае не происходит определение значения и списка предопределенных значений, а только группировка в кластеры.

На рисунке 2 приведена схема модуля, реализующего разработанный алгоритм.

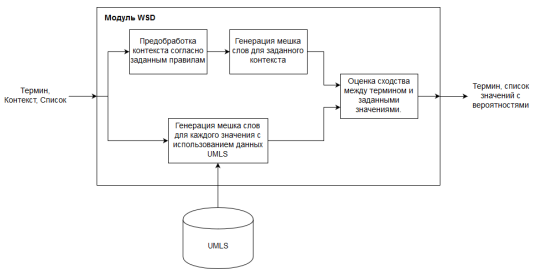

Рис. 2 – Архитектура модуля разрешения лексической многозначности [1]

В [3] были приведены основные сведения о используемом алгоритме разрешения лексиче-

ской многозначности. В данном исследовании будут предложены некоторые улучшения данного алгоритма, затрагивающие блок "Предобработка контекста согласно заданным правилам".

I. Дополнительная обработка контекста

Под дополнительной обработкой текста понимается лемматизация и стемминг слов, входящих в контекст термина, для которого необходимо определить смысл. В текстах используются разные грамматические формы одного и того же слова, а также могут встречаться однокоренные слова. Процедуры лемматизации и стемминга преследуют цель привести все встречающиеся словоформы к одной словарной форме (например, lungs к lung - легкие к легкое), что позволяет увеличить точность сравнения bags-of-words для определения и заданного контекста.

Дополнительной операцией, позволяющей увеличить качество работы алгоритмов сравнения текстов, является удаление стоп-слов (например, предлоги и местоимения).

Стоит отметить, что применительно к обработка биомедицинских текстов следует соблюдать осторожность при применении описанных методик, т.к. существует вероятность потери полезной информации. Так, существующие лемматизаторы могут некорректно преобразовывать специфичные биомедицинские термины, что может привести к ошибкам в работе основного алгоритма. Частично эта проблема решается дополнительными фильтрами, учитывающими регистр написания слов и другие параметры.

В рамках исследования была использована библиотека Natural Language Toolkit (NLTK) [9].

II. Использование графа знаний

Основным источником данных о биомедицинских терминах с используемом алгоритме является словарь UMLS, имеющий графовую структуру. Для улучшения качества работы разрабатываемого алгоритма можно использовать не только информацию о терминах из контекста (например, их определения для построения bags-of-words), а также информацию о взаимосвязях между ними. Таким образом, есть возможность построить граф для заданного контекста и всех входящих в него терминов и оценить вероятность правильности каждого смысла (значения) для разрешения лексической многозначности.

На рисунке 3 приведен пример связей из UMLS для двух значений термина Cold.

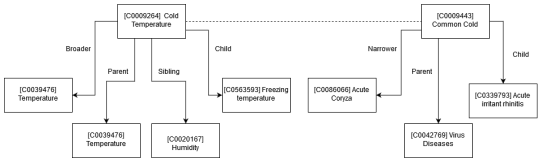

Рис. 3 – Пример графа, сформированного для двух значений термина Cold

В рамках исследования был использован алгоритм PageRank, позволяющий ранжировать все возможные значения многозначных слов на основе их структурной значимости в графе и их связи со словами в заданном контексте.

Заключение

Экспериментальная проверка разработанного алгоритма разрешение лексической многозначности с использованием описанных модификаций показала, что метод дает точность до 87.33

Из возможных улучшений алгоритма можно выделить возможность использования алгоритмов Word Embeddings перед составлением словаря bag-of-words. Также дополнительная информация о возможных улучшениях алгоритма может быть получена после анализа ошибок первого и второго рода (false positives и false negatives соответственно).

Список литературы

1. Statistical Reports on MEDLINE®/PubMed® Baseline Data / Mode of access: https://www.nlm.nih.gov/bsd/licensee/baselinestats.html. – Date of access: 11.09.2019.
2. Navigli, R. Word sense disambiguation: a survey / R. Navigli // ACM Comput. Surv. (CSUR). – 2009. – Vol. 41 (2). – P. 10.
3. Пашук А. В. Анализ методов разрешения лексической многозначности в области биомедицины / А. В. Пашук, А. Б. Гуринович, Н. А. Волорова, А. П. Кузнецов // Доклады БГУИР. – 2019. – №5. – С. 60-65.
4. Disambiguation of Biomedical Text / Mode of access: https://www.slideserve.com/ownah/disambiguation-of-biomedical-text. – Date of access: 10.09.2019.
5. Knowledge-based biomedical word sense disambiguation: comparison of approaches / Mode of access: http://paperity.org/p/56856103/knowledge-based-biomedical-word-sense-disambiguation-comparison-of-approaches. – Date of access: 10.09.2019.
6. DALE: A Word Sense Disambiguation System for Biomedical Documents Trained using Automatically Labeled Examples / Mode of access: https://www.aclweb.org/anthology/N13-3001. – Date of access: 08.09.2019.
7. The effect of word sense disambiguation accuracy on literature based discovery. - BMC Medical Informatics and Decision Making. Mode of access: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-016-0296-1. – Date of access: 08.09.2019.
8. Word Sense Disambiguation (WSD). Test Collection. Collaborations & Outside Resources / Mode of access: https://wsd.nlm.nih.gov/collaboration.shtml. – Date of access: 12.09.2019.
9. Natural Language Toolkit / Mode of access: https://www.nltk.org/. – Date of access: 10.09.2019.