УДК 004.5

АНАЛИЗ МЕТОДОВ РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ В ОБЛАСТИ БИОМЕДИЦИНЫ

А.В. ПАШУК, А.Б. ГУРИНОВИЧ, Н.А. ВОЛОРОВА, А.П. КУЗНЕЦОВ

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь

Поступила в редакцию 4 февраля 2019

Аннотация. Предложен метод разрешения лексической многозначности биомедицинских терминов на основе сравнения «мешков слов», полученных из контекста, определений и информации о связанных терминах из метатезауруса UMLS [1], а также модификация метода с использованием оценки важности слов с помощью статистической меры TF-IDF. Проведена экспериментальная проверка метода на открытом тестовом наборе данных MSH WSD [2], разработанном с целью поддержки исследований в области разрешения лексической многозначности.

Ключевые слова: машинное обучение, обработка текста естественного языка, разрешение лексической многозначности, извлечение информации.

Abstract. A method for resolving the lexical ambiguity of biomedical terms has been proposed. The method is based on a comparison of «word bags» obtained from the context, definitions and information on related terms from the UMLS metathesaurus [1]. Modification of the method using the analysis of word importance using the statistical measure TF-IDF has been proposed. Experimental verification of the method has been performed on the open test MSH WSD data set [2], developed to support research in the field of lexical resolution.

Keywords: machine learning, natural language processing, word sense disambiguation, information retrieval.

Doklady BGUIR. 2019, Vol. 123, No. 5, pp. 60-65
Analysis of the methods of word sense disambiguation in the biomedical domain
A.V. Pashuk, A.B. Gurinovich, N.A. Volorova, A.P. Kuznetsov

DOI: http://dx.doi.org/10.35596/1729-7648-2019-123-5-60-65

Введение

Биомедицинские тексты содержат множество специфичных терминов, используемых только в биомедицинском домене. При этом эти термины, используемые для ссылок на понятия из онтологий или словарей, часто могут иметь несколько значений (биомедицинские термины неоднозначны в 11,7 % случаев [3]). Таким образом, недостаточно извлечь термины из научного текста, возникает проблема разрешения лексической многозначности слов во время текстового анализа.

Качество поиска напрямую зависит от степени понимания машиной обрабатываемых научных текстов. Одной из проблем понимания неструктурированной текстовой информации является проблема разрешения лексической многозначности (WSD – Word Sense Disambiguation). В общем смысле WSD является проблемой классификации – это процесс отнесения слова или фразы к определенному значению, отличающемуся 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.

Модули WSD редко используются в качестве самостоятельной системы и чаще всего используются для решения промежуточной задачи, например, в решении проблемы извлечения информации, поиска информации или машинного перевода.

Теоретический анализ

Задача WSD исторически считается частью NLP и была сформулирована в конце 1940-х годов как одна из задач, необходимых для решения проблемы машинного перевода. В рамках этого исследования были сформированы основные составляющие WSD: использование контекста, статистической информации о словах, внешних источников информации.

До 90-х годов прошлого столетия практически нет статей с исследованиями, посвященными проблеме WSD, так как существовала проблема с недостатком машинно-читаемых источников информации. Развитие информационных технологий привело к появлению множества различных подходов к решению проблемы разрешения лексической многозначности.

Более подробную историю развития систем WSD можно найти в [4, 5].

Все существующие методы разрешения лексической многозначности используют информацию о контексте, в котором используется целевое слово.

Классификация методов следующая.

1. Методы, использующие существующие внешние источники информации [5]:

а) структурированные источники информации (тезаурусы, содержащие информацию об отношениях между словами, например, синонимия, антонимия и др.; машинно-читаемые словари (MRDs); онтологии, содержащие концептуализацию определенной области знаний);

б) неструктурированные источники информации (неразмеченные корпуса текстов (raw corpora); результат синтаксической и/или семантической разметки неразмеченных корпусов (sense-annotated corpora); другие источники, например, словари коллокации слов, словари стоп-слов и др.).

2. Методы, использующие машинное обучение (с учителем или без учителя) на размеченных корпусах текстов.

Проблему разрешения лексической многозначности можно сформулировать следующим образом: дан документ *D*, содержащий термин t_k из словаря *T*. Любой термин t_k может быть соотнесен 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 *D*.

Значение слова – это общепринятый смысл, содержание, соотношение слова с предметами или явлениями объективной действительности. Слова бывают однозначные ($|S_k|=1$) и многозначные ($|S_k|>1$).

Стоит отметить, что задача WSD чувствительна к области применения, например, система, успешно работающая с определением значений слов в художественной литературе, будет работать плохо с технической литературой.

В задаче WSD на нейронной сети на вход обычно подается слово, для которого необходимо определить значение и контекст, в котором слово используется (предложение, абзац, статья). Выходы нейронной сети соответствуют всем возможным значениям слова.

При использовании нейронных сетей для задачи WSD обычно используются очень большие нейронные сети (VLNN), обучаемые на текстах, извлеченных из машинно-читаемых словарей с использованием метода Леска. Суть метода в том, чтобы получить степень пересечения между словарными определениями слов из заданного контекста, содержащего слово, для которого должно быть определено значение [6]. Леск использовал Oxford Advanced Learner's dictionary для получения информации о частоте использования слов в одном контексте с целевым словом.

Основной недостаток нейронных сетей с использованием метод Леска заключается в том, что точность определения значения сильно зависит от качества описаний слов, приведенных в словаре. Преодоление этого недостатка обычно заключается в максимальном расширении словарных определений для нахождения степени пересечения.

В [7, 8] были предприняты попытки генерировать данные для обучения модели с помощью данных, извлекаемых из базы данных WordNet, однако полученные данные уступают по качеству данным, созданным вручную. Авторы рассмотрели только несколько примеров (слова bank, company, word и др.), поэтому результаты данного исследования с большой долей вероятности не могут быть распространены на все множество многозначных слов. Данный подход с использованием ресурсов WordNet не является оправданным для решения проблемы многозначности специализированных терминов ввиду отсутствия большинства аббревиатур или терминов в базе данных.

Основанные на знаниях методы имеют ряд недостатков. Недостаточность структурированных внешних источников информации отрицательно сказывается на эффективности данных методов. Ручное создание источников информации для использования в методах является трудоемким, дорогостоящим и требующим постоянного обновления. В то же время качество автоматически размеченных корпусов значительно ниже качества корпусов, созданных людьми. Это проблема в области WSD получила название «knowledge acquisition bottleneck» (дословно – узкое место приобретения знаний) [5].

Домен биомедицины содержит большое количество многозначных терминов (например, множество аббревиатур, имеющих одинаковое написание, но различное значение в зависимости от контекста (нужен пример)). Поэтому создание тренировочного набора данных для обучения качественной модели, покрывающей большую часть терминов, не представляется возможным и требует длительной подготовки с привлечением специалистов в исследуемой области. Стоит отметить, что существует несколько готовых словарей, которые могут быть использованы для разрешения многозначности и/или обучения модели (MSH-WSD, NLM-WSD [2]), однако они содержат малую часть всего массива биомедицинских терминов и недоступны для бесплатного использования. Эти словари проверяются человеком и содержат фрагменты текстов для каждого смысла многозначного термина. Например, MSH-WSD состоит из описания значений всего 203 многозначных терминов, что не покрывает и 1 % всех многозначных терминов. Обучение с учителем выполняется по одному классификатору для каждого термина для определения правильного значения в конкретном контексте использования термина. Однако при таком подходе остается открытым вопрос возможности использования полученной модели для определения смысла других многозначных терминов, не представленных при обучении.

Основной проблемой извлечения информации из биомедицинской литературы является тот факт, что один термин часто может ссылаться на ген, белок и другие вещества. Также широко используются аббревиатуры, например, MG может ссылаться на миллиграммы (milligram) или магний (magnesium), и акронимы, например, COLD может принимать значение chronic obstructive pulmonary disease (хроническое обструктивное заболевание легких).

Большинство методов WSD в области биомедицины используют базу данных UMLS [1], которая содержит медицинские термины из более чем 100 источников. Последняя версия UMLS (2018AB) включает не менее 100 тысяч многозначных терминов, которые являются частью более чем одного концепта (имеют несколько привязанных CUI, Concept Unique Identifier). Например, аббревиатура ICP относится к 12 различным концептам в базе UMLS. База обновляется дважды в год и отражает последние изменения в терминологии.

Методы разрешения лексической многозначности с помощью машинного обучения делятся на контролируемые и неконтролируемые. В первом случае для работы метода необходим размеченный корпус для обучения модели. Такие методы дают большую точность определения, но имеют очевидный недостаток – обучающее множество должно содержать размеченные слова, для которых в дальнейшем будет определяться значение. На вход таких алгоритмов подается вектор некоторых характеристик, обычно извлекаемых из контекста. На выход подается значение из заданного заранее списка значений. Во втором случае используются неразмеченные данные. Суть таких методов состоит в кластеризации с целью формирования групп связанных терминов. В некоторых источниках задача, решаемая с помощью неконтролируемых методов обучения, называется Word Sense Discrimination, так как на выходе подается список кластеров, не привязанный к значениям из заданного списка значений.

Стоит отметить, что точечные решения проблемы WSD в биомедицинской литературе успешно используются для устранения неоднозначности в аббревиатурах. Для этого автоматически создается корпус, который включает в себя аббревиатуры и полные формы терминов. Затем для решения задачи классификации используется метод опорных векторов (SVM) [9, 10].

Методика

После рассмотрения существующих методов разрешения многозначности было принято решение использовать модификацию подхода, основанного на доступных знаниях (knowledge-based approach [11]) с использованием некоторой метрики сходства между использованием терминов и связанными значениями.

Для каждого контекста составляется мешок слов (Bag of Words). Мешок слов – это модель текстов на натуральном языке, в которой каждый документ или текст выглядит как неупорядоченный набор слов без сведений о связях между ними. Его можно представить в виде матрицы, каждая строка которой соответствует отдельному документу или тексту, а каждый столбец – определенному слову. Ячейка на пересечении строки и столбца содержит количество вхождений слова в соответствующий документ.

Предположим, что имеется словарь терминов $T = \left\{t_1, \ldots, t_{|T|}\right\}$. Для каждого термина t_k из T имеется словарь значений $S_k = \left\{s_{(k,1)}, \ldots, s_{(k,|S_k|)}\right\}$. Каждое значение $s_{k,m}$ связано с мешком слов $B_{(k,m)} = \left\{t_{(k,m,1)}, \ldots, t_{\left(k,m,\left|B_{(k,m)}\right|\right)}\right\}$, содержащим связанные термины. Будем считать, что любой элемент из множества $B_{k,m}$ входит в множество T.

В процессе определения значения термина имеется термин t_i с некоторым контекстом $C = \left\{t_{c1}, t_{c2}, \ldots\right\}$ (представленным в виде множества слов/терминов, окружающих t_i в рассматриваемом тексте) и термин t_i имеет словарь множества возможных значений S_i. Задача заключается в том, чтобы выбрать s^* из S_i, которое, как предполагается, является основным значением термина t_i в рамках контекста C: $s^*\left(t_i, C\right) = \max\limits_{m \in \left[1, |S_i|\right]} \left[\dfrac{C \cdot B_{i,m}}{|C| \cdot |B_{i,m}|}\right]$.

Для составления мешка слов используются собранные на предыдущем этапе исследования данные из Unified Medical Language System – одной из наиболее полной базы данных, содержащей термины, варианты написания (синонимы, антонимы и т. д.), связи между ними и подробные словарные описания.

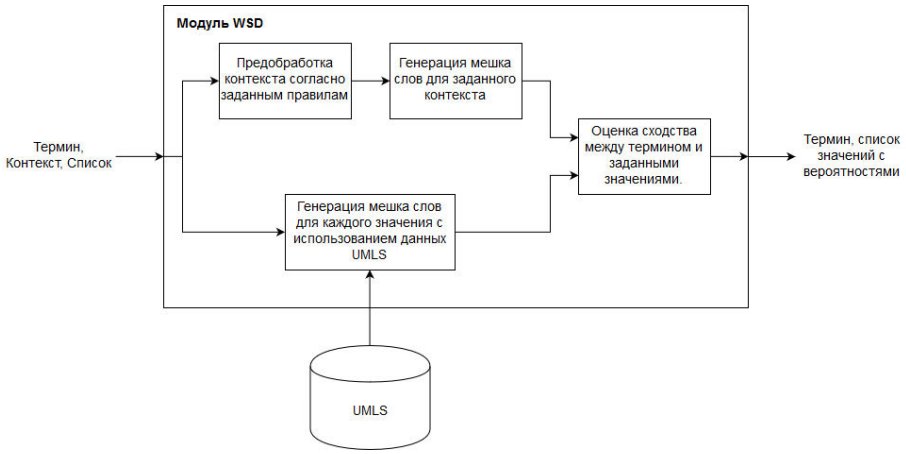
Модификация метода для вычисления оценки подобия возможных значений использует статистическую меру TF-IDF.

TF (Term Frequency) – частота слова, в простейшем случае определяется как число вхождений слова t в документе d – $f_{t,d}$. Однако более распространено вычисление частоты слова как отношения числа вхождений слова t в документе $f_{t,d}$ к общему числу слов в документе $d \in D$: $tf_{t,d} = \dfrac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$.

IDF (Inverse Document Frequency) – мера, позволяющая оценить, какую ценность имеет слово в рамках всего массива документов. Слова, которые встречаются в документах реже, имеют большую ценность и, соответственно, большее значение меры: $idf_t = \ln\left(1 + \dfrac{N}{N_t}\right)$, где N – общее количество документов, $N_t = \left\{d \in D : t \in d\right\}$ – количество документов, содержащих слово t. Итоговая оценка вычисляется путем перемножения двух составляющих: $tf - idf_{t,d,D} = tf_{t,d} \times idf_{t,D}$. Полученные значения используются при вычислении оценки подобия для каждого возможного значения S_i.

Экспериментальная часть

Программная реализация модуля разрешение лексической многозначности была выполнена на языке программирования Python (в частности, модуль nltk). В качестве базы данных для данных, необходимых для работы алгоритма, была выбрана СУБД Mongo DB. Архитектура модуля представлена на рисунке.



Архитектура модуля разрешения лексической многозначности

На вход модуля подается JSON, содержащий текст, размеченный метаданными о биомедицинских концептах, которые были найдены NER-процессором. На выходе модуля получаем модифицированный JSON, в котором каждому концепту добавлен дополнительный атрибут «probability», позволяющий определить наиболее вероятное значение термина.

Экспериментальная проверка разработанного алгоритма проводилась с использованием тестового набора данных [2]. Данный набор разработан для того, чтобы обеспечить объективное сравнение эффективности алгоритмов разрешения лексической неоднозначности в биомедицинской научной литературе. Набор данных содержит от 100 до 200 размеченных контекстов для 203 многозначных биомедицинских терминов (включая аббревиатуры и полные названия), всего около 39 тысяч тестовых примеров контекстов.

Результаты и их обсуждение

В таблице представлены результаты эксперимента.

Результаты эксперимента

Метод разрешения лексической неоднозначности	Accuracy, %
Cosine similarity between bags-of-words with words frequency (context is equal to sentence)	71,99
Cosine similarity between bags-of-words with words frequency ($C = \infty$)	76,38
Cosine similarity between bags-of-words without words frequency (context is equal to sentence)	73,29
Cosine similarity between bags-of-words without words frequency ($C = \infty$)	77,39
TF-IDF ($C = \infty$)	81,30

В системе, для которой разрабатывается модуль WSD, результаты его работы проходят проверку человеком (например автором статьи или исследователем). Соответственно, множество слов, при которых используется то или иное значение целевого слова, может быть расширено с использованием контекста, в котором было использовано конкретное значение слова (фрагмент предложения, предложение, абзац или вся статья) из проверенной статьи.

Заключение

Экспериментальная проверка разработанного алгоритма разрешения лексической многозначности на тестовом наборе данных [2] показала, что метод дает точность до 76,38 %, улучшенная модификация метода с использованием оценки важности слов с помощью TF-IDF – до 81,30 %. Достоинством предлагаемого метода является высокая скорость обработки (в среднем 0,1 с на один термин), что позволяет использовать разработанный модуль в системах реального времени.

Из возможных путей улучшения алгоритма можно выделить:

– использование различных комбинаций для вычисления весов при оценке важности слов;

– внедрение алгоритмов Word Embeddings (word2vec, doc2vec) перед составлением мешка слов;

– расширение мешка слов с использованием информации о контексте, извлеченных из размеченных (проверенных) человеком статей.

Список литературы / References

1. Unified Medical Language System (UMLS) // U.S. National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/ (date of access: 20.11.2018).
2. Word Sense Disambiguation (WSD) Test Collections // U.S. National Library of Medicine. URL: https://wsd.nlm.nih.gov/ (date of access: 30.11.2018).
3. Statistical Reports on MEDLINE/PubMed Baseline Data // U.S. National Library of Medicine. URL: https://www.nlm.nih.gov/bsd/licensee/baselinestats.html (date of access: 16.11.2018).
4. Ide N., Véronis J. Introduction to the special issue on word sense disambiguation: the state of the art // Computational Linguistics – Special issue on word sense disambiguation. 1998. № 24. P. 2–40.
5. Navigli R. Word sense disambiguation: a survey // ACM Computing Surveys. 2009. № 41. P. 1–69.
6. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone // Proceeding SIGDOC '86 Proceedings of the 5th annual international conference on Systems documentation. Toronto, Ontario, Canada: ACM, 1986. P. 24–26.
7. Leacock C., Miller G.A. Using corpus statistics and WordNet relations for sense identification // Computational Linguistics – Special issue on word sense disambiguation. 1998. № 24. P. 147–165.
8. Preiss J., Stevenson M. DALE: A Word Sense Disambiguation System for Biomedical Documents Trained using Automatically Labeled Examples // Proceedings of the 2013 NAACL HLT Demonstration Session. Atlanta, Georgia: Association for Computational Linguistics, 2013. P. 1–4.
9. Liu H., Teller V., Friedman C.A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation // Journal of the American Medical Informatics Association. 2004. № 11. P. 320–331.
10 Word sense disambiguation across two domains: Biomedical literature and clinical notes / G.K. Savova [et al.] // Journal of Biomedical Informatics. 2008. № 41. P. 1088–1100.
11 Jimeno-Yepes A. J., Aronson A. R. Knowledge-based biomedical word sense disambiguation: comparison of approaches // BMC Bioinformatics. 2010. № 11. P. 569–581.

Сведения об авторах

Пашук А.В., аспирант кафедры систем управления Белорусского государственного университета информатики и радиоэлектроники.

Гуринович А.Б., к.ф.-м.н., доцент кафедры вычислительных методов и программирования Белорусского государственного университета информатики и радиоэлектроники.

Волорова Н.А., к.т.н., доцент кафедры информатики Белорусского государственного университета информатики и радиоэлектроники.

Кузнецов А.П., д.т.н., профессор кафедры систем управления Белорусского государственного университета информатики и радиоэлектроники.

Information about the authors

Pashuk A.V., PG student of the control systems department of Belarusian state university of informatics and radioelectronics.

Gurinovich A.B., PhD, associate professor of computational methods and programming department of Belarusian state university of informatics and radioelectronics.

Volorova N.A., PhD, associate professor of the informatics department of Belarusian state university of informatics and radioelectronics.

Kuznetsov A.P., D.Sci, professor of control systems department of Belarusian state university of informatics and radioelectronics.

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул. П. Бровки, 6,
Белорусский государственный университет
информатики и радиоэлектроники
тел. +375-29-875-23-34;
e-mail: pashuk@bsuir.by
Пашук Александр Владимирович

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovka str., 6,
Belarusian state university
of informatics and radioelectronics
tel. +375-29-875-23-34;
e-mail: pashuk@bsuir.by
Pashuk Aleksandr Vladimirovich