ВЫДЕЛЕНИЕ СМЫСЛОВЫХ ПОНЯТИЙ В МЕДИЦИНСКИХ ДИАГНОЗАХ ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ

Д.И. Коваль, И.В. Сушков, А.Б. Тепляков
Томский Политехнический Университет
E-mail: deniskoval12@gmail.com

“Изучив множество публикаций и исследований в области применения методов машинного обучения на основе нейронных сетей, мы выделили несколько наиболее перспективных, на наш взгляд, направлений в создании и развитии систем искусственного интеллекта для здравоохранения”:

1. **“Автоматизированные системы диагностики**
2. **Системы распознавания неструктурированных медицинских записей и понимания естественного языка**
3. **Системы анализа и предсказания событий**
4. **Системы автоматической классификации и сверки информации**

“Исследования в области разработки программного обеспечения для задач обработки естественного языка (Natural Language Processing – NLP, Language Engineering – LE) активно развиваются в различных исследовательских парадигмах. Устойчивые тенденции последнего десятилетия в области LE связаны с широкомасштабными исследованиями в области разработки и применения статистических методов и методов машинного обучения (Machine Learning – ML).

Возрастающее использование статистических методов в задачах LE порождает некоторый отход от методов исследования и моделирования глубинных механизмов, лежащих в основе мышления и языка человека. Статистические методы в NLP позволяют достигнуть определённых результатов в решение ряда задач (распознавание речи, разрешение многозначности, аннотирование текстов и др.), однако представляются перспективным использование гибридных моделей, в которых используется различная техника, в том числе интроспективные методы.

Одним из перспективных направлений исследований в области извлечения информации (Information Extraction – IE) является направление «машинного обучения». Компьютерные системы, реализующие методы ML, ориентированы на получение новых знаний в результате автоматизации процесса обучения. Методы автоматического получения новых знаний на основе эмпирических данных можно успешно применять для формирования баз знаний. Это обстоятельство делается актуальными исследования в области обучения языку (Language Learning), результаты которых применимы в практических приложениях NLP-систем.

Распознавание именованных объектов (NER) назначает тег именованной сущности указанному слову, используя правила и эвристику. Именованный объект, представляющий человека, местоположение и организацию, должен быть распознан. Распознавание именованных объектов - это задача, которая извлекает номинальную и числовую информацию из документа и классифицирует слово на человека, организацию или категорию даты. NER классифицирует все слова в документе на существующие категории и «ни один из вышеперечисленных» [1].

Распознавание биомедицинских названных сущностей очень важно при языковой обработке биомедицинских текстов, особенно при извлечении из документов информации о белках и генах, таких как РНК или ДНК. Поиск названных объектов генов из текстов является очень важной и сложной задачей. Поиск имени гена в текстах соответствует поиску названия компании или имени человека в газетах. Распознавание биомедицинских именованных сущностей представляется более сложным, чем распознавание нормальных именованных сущностей. Многочисленные исследования позволили выявить названные объекты с помощью алгоритмов обучения под наблюдением, основанных на многих правилах [2].

Подходы к обучению с использованием контролируемых методов используют модели Маркова, деревья решений, метод опорных векторов (SVM) и условные случайный поля (CRF). Методы обучения с учителем обычно обучаются с использованием многих функций, основанных на различных лингвистических правилах, и оценивают эффективность с помощью тестовых данных [3].

Распознавание именованных объектов (NER) классифицирует все незарегистрированные слова, встречающиеся в текстах, и является подзадачей для извлечения информации. Обычно NER использует восемь категорий: местоположение, человек, организация, дата, время, процент, денежная стоимость и «ничего из вышеперечисленного». NER сначала находит именованные сущности в предложениях и объявляет категорию сущностей [4].

Распознавание именованных объектов имеет три подхода - на основе словаря, на основе правил и на основе машинного обучения. Подход на основе словаря хранит как можно больше именованных сущностей в списке, называемом справочником. Этот подход кажется очень простым, но в то же время имеет ограничения. NER сложен, потому что целевые слова в основном являются собственными существительными или незарегистрированными словами. Кроме того, новые слова могут генерироваться часто, и даже один и тот же поток слов может распознаваться как разнообразные именованные объекты с точки зрения их текущего контекста. Второй подход NER - подход, основанный

на правилах [5]. Этот подход обычно зависит от правил и шаблонов именованных объектов, появляющихся в реальных предложениях. Хотя подходы, основанные на правилах, могут использовать контекст для решения проблемы нескольких именованных объектов, каждое правило должно быть написано до его фактического использования. Третий подход, основанный на машинном обучении, присваивает именованные объекты словам, даже если слова не перечислены в словаре, а контекст не описан в наборе правил. Для этих подходов в основном используются метод опорных векторов (SVM), скрытые Марковские модели, максимальные энтропийные Марковские модели и условные случайные поля (CRF) [6].

Исследователи по обработке естественного языка были заинтересованы в извлечении информации из генов, рака и белка из биомедицинской литературы [7]. Распознавание биомедицинских названных объектов, которое необходимо для извлечения биомедицинской информации, рассматривается как первый этап интеллектуального анализа текста в биомедицинских текстах. В течение многих лет признание технических терминов в области биомедицины было одной из самых сложных задач в обработке естественного языка, связанной с биомедицинскими исследованиями [8].

Биомедицинская NER сталкивается с трудностями по пяти причинам. Во-первых, из-за текущих исследований количество новых технических терминов быстро увеличивается. Очень сложно создать справочник, который включает все новые термины. Во-вторых, одни и те же слова или выражения могут быть классифицированы как объекты с разными именами с точки зрения их контекста. В-третьих, длина объекта довольно велика, и объект может включать контрольные символы, такие как дефисы (например, «12-о-тетрадеканоилфорбол 13-ацетат»). В-четвертых, выражения аббревиатуры часто используются в биомедицинской области, и они испытывают двусмысленность смысла. Например, «TCF» может относиться к «Т-клеточному фактору» или «Тканевая культуральная жидкость». Наконец, в биомедицинских терминах нормальные термины или функциональные термины объединяются, поэтому биомедицинский термин может стать слишком длинным. Например, «HTLV-I-инфицированный» и «HTLV-I-трансформированный» включают нормальные термины «I», «инфицированный» и «трансформированный». Биомедицинскому NER трудно сегментировать предложение с именованными объектами. Изменения правописания также создают проблему. Кроме того, именованный объект одной категории может включать в себя другой именованный объект другой категории.

Машинные методы обучения концептуальным знаниям представляют собой модель правдоподобных индуктивных и дедуктивных рассуждений, в которых вывод знаний и их использование не отделимы друг от друга. Реализация обучения в режиме правдоподобных рассуждений позволит организовать взаимодействие не только данных и знаний в процессах обработки текстов, но и моделировать процесс взаимодействий учителя и ученика в процессе приобретения знаний в схемах многоагентных взаимодействий.

Список использованных источников

1. Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th international conference on computational linguistics. Association for Computational Linguistics, vol. 1. 2002. p. 1–7.
2. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008;13:652–63.
3. Wilbur J, Smith L, Tanaben L. Biocreative 2 gene mention task. In: Proceedings of second BioCreative challenge evaluation workshop. 2007.
4. Rau LF. Extracting company names from text. In: Proceedings of the conference on artificial intelligence applications of IEEE, vol. 1. 1991. p. 29–32.
5. Zhao S. Named entity recognition in biomedical texts using an HMM model. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics. 2004. p. 84–7.
6. Sekine SN. Description of the Japanese NE system used for Met-2. In: Proceedings of the message understanding conference. 1998. p. 1314–9.
7. Lee KJ, Hwang YS, Rim HC. Two phase biomedical NE recognition based on SVMs. In: Proceedings of the ACL 2003 workshop on natural language processing in biomedicine. Association for Computational Linguistics, vol. 13. 2003. p. 33–40.
8. Song Y, Kim E, Lee GG, Yi B. POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics. 2004. p. 100–3