А. В. Пашук

РАЗРАБОТКА ЭФФЕКТИВНОГО NER-ПРОЦЕССОРА ДЛЯ АНАЛИЗА НАУЧНЫХ ТЕКСТОВ В БИОМЕДИЦИНЕ

Рассматривается проектирование и разработка эффективного NER алгоритма для извлечения научных терминов из текстов научных статей в области биомедицины.

Введение

Name Entity Recognition - это процесс извлечения полезных данных из неструктурированной текстовой информации. Данный процесс включает в себя несколько основных этапов: очистка исходных данных, преобразование "чистых"данных, извлечение полезной информации (терминов).

Прежде всего, текст статьи проходит этап очистки, включающий в себя удаление HTML тегов, другой разметки текста, а также приведение текста к нижнему регистру (для алгоритмов, не учитывающих регистр). Затем текст разбивается на предложения, которые будут подаваться на вход NER-процессоров.

I. NER-процессор с использованием синтаксического разбора

Данный процессор работает с обычным текстом, не содержащим никакой дополнительной разметки или метаинформации. Метод учитывает только порядок слов в предложении и знаки препинания. Название процессора происходит от названия структуры данных, которую он использует (trie – это одно из названий префиксных деревьев). Данная структура позволяет создать индекс, по которому всегда можно извлечь и посчитать количество тестовых фрагментов, начинающихся с заданного текста.

Преимуществом данного метода является его простота и, теоретически, наибольшая точность, если рассматривать нахождение точных совпадений. Однако данный метод не способен обнаружить термины в измененной форме или составные термины, разбитые дополнительными словами или частицами. Другими словами, данный метод не способен извлечь термины из текста, с использованием анализа синтаксической структуры предложений.

II. NER-процессор с использованием POS-тегирования

Название процессора происходит от библиотеки, которая используется для поиска терминов - Natural Language ToolKit (NLTK).

Преимуществом данного метода является использование частей речи, что позволяет значительно быстрее обрабатывать большие фрагменты текста (по сравнению с trieNER). Однако метод имеет существенный недостаток – возможность извлекать термины, состоящие из одного слова, термины, представляющие собой словосочетания либо пропускаются, либо извлекаются частично[2].

III. Сравнение реализованных алгоритмов

Для сравнение эффективности разметки различными процессорами использовались следующие метрики: точность, полнота и F-мера [1].

Таблица 1 – Результаты эксперимента

Алгоритм	Точность	Полнота	F-мера
trieNER	0.873	0.885	0.879
nltkNER	0.955	0.488	0.617

Полученные результаты указывают на то, что алгоритм, использующий синтаксический разбор имеет лучшие характеристики, чем алгоритм, использующий библиотеку NTLK.

IV. Выводы

В ходе исследования были реализованы две модели NER-процессоров, основанные на синтаксическом разборе и POS-тегировании. trieNER алгоритм показал лучшие результаты, чем ntlkNER, что может быть связано с недостаточным объемом словаря, необходимого для точной Part-of-Speech разметки биомедицинских статей.

Список литературы

1. Strzalkowski T. Natural Language Information Retrieval / T. Strzalkowski. - Springer Science Business Media. - 1999. - P.384.
2. Jacquemin C. Spotting and Discovering Terms Through Natural Language Processing / C. Jacquemin. - MIT Press. - 2001. -P. 378.

Пашук Александр Владимирович, аспирант кафедры систем управления БГУИР, pashuk@bsuir.by.

Научный руководитель: Гуринович Алевтина Борисовна, доцент кафедры ВМиП БГУИР, кандидат физико-математических наук, gurinovich@bsuir.by