А. В. Пашук

СИСТЕМА ГЕНЕРАЦИИ ВАРИАНТОВ ДЛЯ БИОМЕДИЦИНСКИХ ТЕРМИНОВ

В работе приводится описание системы генерации вариантов для биомедицинских терминов. Разработанный генератор может быть использован в широком спектре приложений, использующих различные методы извлечения информации из неструктурированного текста.

Введение

Автоматизация извлечения терминов из текста статей сегодня является одной из важнейших задач в области Open Science. Биомедицинские научные статьи включают большое количество названий протеинов, генов, болезней, различные аббревиатуры и другие специфические термины. Сложность извлечения таких терминов заключается в разнообразии написания одних и тех же терминов и большом количестве аббревиатур, имеющих различное значение в зависимости от контекста.

I. Разработка системы правил для генерации вариантов

Стандартный набор правил генератора вариантов включает создание вариантов по следующим правилам: преобразование термина во множественный/единичный вид, удаление/добавление знаков пунктуации (таких как дефисы или апострофы). В рамках исследования был разработан собственный алгоритм, расширяющий список правил стандартного генератора для лучшего определения биомедицинских терминов:

- преобразование чисел в начале термина (5-iodotubercidin, 5iodotubercidin, 5 iodotubercidin и т.д.);
- буквенные индексы в конце термина (penicillin G, penicillin-G и т.д.);
- преобразование числе в конце термина (IL-1, IL 1, IL1 и т.д.);
- преобразование греческих символов в их текстовой выражение ($TNF\ \alpha$, $TNF\ alpha$) и др.

Заключение

Разработан алгоритм, позволяющий генерировать возможные варианты написания биомедицинских терминов. Данный алгоритм может быть использован в различных приложениях, требующих извлечения информации из неструктурированного текста. Экспериментальная проверка результатов работы полученной системы на массиве статей (20 млн. биомедицинских статей) показала, что использование генератора позволяет распознать значительно больше терминов.

Стоит отметить, что необходимо использовать дополнительную проверку сгенерированных вариантов, чтобы отфильтровать только те варианты, которые встречаются в реальных научных статьях.

Список литературы

1. Tsuruoka, Y. Probabilistic term variant generator for biomedical terms / Y. Tsuruoka, J. Tsujii: SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval [Electronic resource]. - Mode of access: http://www.nactem.ac.uk/tsuruoka/papers/sigir03.pdf. - Date of access: 15.03.2017.
2. Strzalkowski T. Natural Language Information Retrieval / T. Strzalkowski. - Springer Science & Business Media. - 1999. - P.384.
3. Jacquemin C. Spotting and Discovering Terms Through Natural Language Processing / C. Jacquemin. - MIT Press. - 2001. -P. 378.

Пашук Александр Владимирович, аспирант кафедры систем управления факультета информационных технологий и управления Белорусского государственного университета информатики и радиоэлекстроники, pashuk@bsuir.by.

Научный руководитель: Гуринович Алевтина Борисовна, доцент кафедры вычислительных методов и программирования Белорусского государственного университета информатики и радиоэлектроники, кандидат физико-математических наук, gurinovich@bsuir.by