0.1. *Мезенцева А.А., Бручес Е.П.* Исследование автоматического связывания сущностей в научных текстах на русском языке

Автоматическое связывание сущностей (англ. entity linking) - задача нахождения соотношения между упоминанием в тексте и уникальной сущностью в структурированной базе знаний (в данной работе используется Wikidata). Актуальность исследования заключается в том, что рассмотренные нами методы не требуют большого количества данных, которого для русского языка, насколько нам известно, нет в открытом доступе. Новизна работы состоит в сравнении различных подходов к решению поставленной задачи и анализе результатов, полученных при тестировании на русскоязычном наборе данных.

Нами были проведены эксперименты, в качестве тестового набора данных использовался корпус научных статей RuSERRC [1]. На вход алгоритму подается единичный токен или последовательность токенов, соответствующих термину. Затем входная последовательность предобрабатывалась - приводилась к начальной форме. Для этого были протестированы две библиотеки - Natasha и MyStem, наилучшие результаты показала вторая. Далее выполнялись два основных шага: создание массива кандидатов для связывания и нахождение наиболее подходящей сущности в полученном множестве кандидатов. Для генерации кандидатов использовалось построковое сравнение и расширение за счёт униграмм, биграмм и триграмм. Последний принёс значительный прирост (с 1.95 до 11.73) среднего количества кандидатов для сущностей и увеличение количества множеств кандидатов, которые содержат нужную сущность, но это привело к значительному снижению точности всей системы с 71% до 19%, так как среди большего количества кандидатов сложнее выбрать подходящий. Для второго шага, ранжирования, было протестировано три метода. Первый, выбор сущности, информация о которой наиболее полно представлена в базе знаний, справлялся с задачей неплохо, но не позволял учитывать контекст. Использование второго подхода, рассчёт векторного расстояния между контекстом упоминания и описанием сущности (как например, в статье [2]), привело к повышению точности (с 19% до 38%). Третий подход, взвешенные векторные расстояния, позволил добиться самого высокого значения точности (54%) для тех терминов, у которых есть связь с сущностями в графе знаний в размеченном корпусе среди всех экспериментов.

Таким образом, финальный набор шагов алгоритма: MyStem для предобработки входной последовательности, n-граммы для генерации кандидатов, взвешенные векторные расстояния для ранжирования. Проведенные эксперименты позволили добиться увеличения значений всех метрик, кроме точности, для всех упоминаний. Улучшить это планируется за счёт использования классификатора на основе сиамской сети для ранжирования и расширения списка кандидатов с помощью синонимов и аббревиатур.

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 19-07-01134.

Научный руководитель — к.ф.-м.н. Батура Т. В.

Список литературы

[1] Мезенцева А. А., Бручес Е. П, Батура Т. В. Автоматическое связывание терминов из научных текстов с сущностями базы знаний. // Вестник НГУ. Серия: Информационные технологии. 2021 Т. 19, № 2. С. 65–75.

[2] Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage // Proceedings of the Section on Survey Research Methods. American Statistical Association. 2020. P. 354–359.