0.1. *Бручес Е.П., Мезенцева А.А.* Извлечение отношений из научных текстов на русском языке

В данной работе рассматривается задача извлечения отношений, которая состоит в нахождении и классификации семантических отношений между научными терминами. Насколько нам известно, для русского языка нет большого количества размеченных данных для этой задачи, поэтому применение стандартного цикла обучения моделей машинного обучения затруднено. Для решения этой задачи мы реализовали и сравнили два подхода: подход, основанный на использовании лексико-синтаксических шаблонов, и подход, основанный на идее обучения без примеров (англ. zero-shot learning). Предложенные методы способны работать в условиях малого количества размеченных данных, что обуславливает актуальность данной работы.

Для реализации подхода, основанного на использовании лексико-синтаксических щаблонов были вручную собраны лексические маркеры, которые однозначно указывают на то или иное семантическое отношение. Сложность этого подхода состоит в том, что очень часто семантические связи выражены имплицитно — это означает, что они могут быть распознаны только при анализе контекста, без опоры на конкретные лексические единицы.

Идея подхода с применением zero-shot learning состоит в том, чтобы взять преобученную модель и дообучить её на данных на том языке, в котором они хорошо представлены, а затем оценить качество модели на русскоязычном корпусе. Гипотеза состоит в том, что информация из другого языка поможет модели делать предсказания в том числе и на данных на целевом языке. В качестве предобученной языковой модели для получения векторных представлений мы взяли BERT bert-base-multilingual-cased. Мы использовали архитектуру модели для классификации отношений R-BERT, которая была предложена в статье [1]. Дообучение моделей было выполнено на англоязычном корпусе SciERC [2], который, в том числе, содержит информацию об отношениях между научными терминами.

Оба подхода были протестированны на корпусе научных текстов на русском языке, который, в том числе, содержит разметку отношений между терминами [3]. Сравнение алгоритмов выполнялось на основе метрики F1-macro. Так, значение данной метрики для каждого из этих подходов составило 0.23. Объединение указанных методов позволило достичь значения F1-macro 0.27.

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] Shanchan W., Yifan H. Enriching pretrained language model with entity information for relation classification // Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019. P. 2361–2364.

[2] Luan Y., He L., Ostendorf M., Hajishirzi H. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 3219–3232.

[3] Батура Т. В., Бручес Е. П., Паульс А. Е., Исаченко В. В., Щербатов Д. Р. Семантический анализ научных текстов: опыт создания корпуса и построения языковых моделей // Программные продукты и системы. 2021. Т. 34. № 1. С. 132–144.