

ПРЕДОБУЧЕННЫЕ ЯЗЫКОВЫЕ МОДЕЛИ В ЗАДАЧЕ ИЗВЛЕЧЕНИЯ ВЛОЖЕННЫХ ИМЕНОВАННЫХ СУЩНОСТЕЙ

Рожков Игорь Сергеевич

*Московский Государственный Университет имени М.В. Ломоносова,
Научно-Исследовательский Вычислительный Центр, лаборатория Анализа
Информационных Ресурсов*

fulstocky@gmail.com

Индекс УДК: 004.89

Аннотация

В данной работе изучаются возможности модели понимания машинного чтения (*MRC, Machine Reading Comprehension*) по передаче знаний в задаче извлечения вложенных сущностей. В такой задаче от модели требуется получить некоторые знания из текстов при помощи подсказок, чтобы извлечь искомые сущности. Обычно для этого требуется обучение на большом наборе данных. Однако в сценариях с ограниченными ресурсами или в сценариях, специфичных для предметной области, создание набора обучающих данных может быть трудным и неэффективным. В таких случаях может быть полезен перенос знаний из более доступных наборов данных, например, из других областей. Для достижения более эффективного извлечения были проведены эксперименты на предобученных моделях BERT различных версий с разными подсказками, которые показали, что тщательная проработка и выбор действительно способствуют улучшению результатов. Были изучены различные подсказки на примере переноса знаний от общего набора данных NEREL к биомедицинскому набору NEREL-BIO, а также малоресурсный набор данных RuNNE.

Ключевые слова: извлечение именованных сущностей, обучение с переносом, малоресурсное обучение, понимание машинного чтения, обучение на подсказках

Введение

Задача распознавания именованных сущностей (*Named Entity Recognition, NER*) — это задача поиска фрагментов текста, связанных с некоторыми физическими или абстрактными объектами именами, и их классификации в соответствии с их смысловым классом, таким как, например, человек или организация. На практике сущности часто вложены друг в друга (т.е. одна последовательность слов является непрерывной подпоследовательностью другой), и это требуется учитывать при создании автоматической системы извлечения сущностей. Примером такого феномена является последовательность *МГУ им. М.В. Ломоносова* со смысловым классом **ORGANIZATION**, внутри которой есть подпоследовательность слов *М.В. Ломоносова*, которые отображают человека, т.е. смысловой класс **PERSON**.

Однако извлечение вложенных именованных сущностей связано с немалым количеством трудностей и ограничений. Среди них:

- Применяемость предобученных моделей в разных областях на практике. Нередко в специальных областях извлечение может иметь иной характер разметки, чем популярные общие обученные модели, из-за чего качество таких моделей может сильно ухудшиться.

- Малоресурсность наборов данных. Большинство нейросетевых моделей требует достаточно большое количество размеченных данных в целях получения требуемого качества. Однако нередко на практике не удаётся собрать достаточное количество таких данных.

Среди существующих наборов данных для русского языка большинство не рассматривало ни одну из вышеуказанных проблем. Совсем недавно возникли наборы данных NEREL [1], NEREL-BIO [2] и RuNNE [3]. Первый набор данных содержит множество вложенных сущностей 29 разных классов. Однако, он не позволял исследовать первое ограничение, в связи с чем на его основе был создан набор данных RuNNE, содержащий несколько специально ограниченных классов сущностей. NEREL-BIO — набор данных биомедицинской области, содержащий 37 разных классов, причём некоторые из них совпадают с классами набора NEREL. В паре эти наборы данных позволяют тщательно исследовать второе ограничение — по переносу знаний из одной, общей области, в другую, более специфичную.

Модель машинного чтения

Среди множества подходов к решению задачи была выбрана т. н. *модель машинного чтения (Machine Reading Comprehension, MRC)* [4]. Данная модель позволяет переформулировать задачу извлечения вложенных именованных сущностей как задачу ответа на вопросы некой предобученной модели, например, BERT [5] по заданной последовательности слов. Выбирая «хорошие» вопросы, можно достичь высокого качества предсказания именованных сущностей. Лучшим типом вопросов в оригинальной работе стали т. н. *определения* классов сущностей, ставшие отправной точкой в исследовании. В работе понятия «вопросы» и «подсказки» взаимозаменяемы — последние отражают суть применения этих вопросов к модели: мы как бы стараемся подсказать модели, на что стоит обращать внимание, и какие знания в ходе машинного чтения полезнее других, что и позволяет улучшить результаты.

Подсказки

В исследовании рассматривались способы улучшения качества предсказаний сущностей модели MRC в обоих ограничениях. В том числе были предложены новые подсказки:

- *N самых частотных сущностей или их компонентов*
- *«Однословные» контекстная, лексическая и структурная*
- *«Полные» подсказки для всех внешних (внутренних, вложенных) и мульти-подходы*

В частности, эксперименты показали, что т. н. *полная лексическая подсказка для всех внешних сущностей* оказалась наилучшей из них. Данная подсказка формируется следующим образом: берётся некоторое предложение из обучающего набора данных, в нём находятся все самые внешние сущности заданного класса, и каждая такая сущность заменяется меткой этого самого класса.

Все представленные подходы просты в создании, т. к. могут быть сгенерированы автоматически, в отличие от использования определений, поскольку требуется их тщательный отбор экспертами из разных словарей.

Результаты

Эксперименты на ограниченных классах набора данных RuNNE показали, что использование определений достигает качества в 61.21% по макро F1-мере, в то время как полная лексическая по всем внешним сущностям позволяет улучшить его до 63.88%. Такой результат позволяет сделать вывод о том, что модель машинного чтения получает больше семантической информации из лексических подсказок. Более того, по сравнению с оригинальным подходом MRC [4] — выбором определений для каждого класса — лексические подсказки создаются полностью автоматически, что значительно проще ручного и тщательного подбора определений.

В задаче переноса знаний большинство подсказок позволили получить результаты лучше, чем при обычном обучении, что говорит о пользе переноса знаний из другой области.

Представленные результаты исследования были опубликованы в [3, 1, 2, 6, 7].

Список литературы

- [1] NEREL: a Russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links / Loukachevitch Natalia, Artemova Ekaterina, Batura Tatiana, Braslavski Pavel, Ivanov Vladimir, Manandhar Suresh, Pugachev Alexander, Rozhkov Igor, Shelmanov Artem, Tutubalina Elena, et al. // Language Resources and Evaluation. — 2023. — P. 1–37.
- [2] NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities / Loukachevitch Natalia, Manandhar Suresh, Baral Elina, Rozhkov Igor, Braslavski Pavel, Ivanov Vladimir, Batura Tatiana, and Tutubalina Elena // Bioinformatics. — 2023. — Vol. 39, no. 4. — P. btad161.
- [3] Соревнование RuNNE-2022: извлечение вложенных именованных сущностей / Артемова Е. Л., Змеев М. В., Лукашевич Н. В., Рожков И.С., Батура Т.В., Иванов В.В. и Тутубалина Е.В. // COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES. — RSUH. — 2022. — Июнь. — Режим доступа: <http://dx.doi.org/10.28995/2075-7182-2022-21-33-41>.
- [4] A Unified MRC Framework for Named Entity Recognition / Li Xiaoya, Feng Jingrong, Meng Yuxian, Han Qinghong, Wu Fei, and Li Jiwei // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 5849–5859.
- [5] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) / ed. by Burstein Jill, Doran Christy, Solorio Thamar. — Minneapolis, Minnesota : Association for Computational Linguistics. — 2019. — June. — P. 4171–4186. — Access mode: <https://aclanthology.org/N19-1423>.
- [6] Rozhkov IS, Loukachevitch NV. Machine Reading Comprehension Model in Domain-Transfer Task // Lobachevskii Journal of Mathematics. — 2023. — Vol. 44, no. 8. — P. 3160–3168.
- [7] Rozhkov IS, Loukachevitch NV. Prompts in Few-Shot Named Entity Recognition // Pattern Recognition and Image Analysis. — 2023. — Vol. 33, no. 2. — P. 122–131.