УДК 004.8

Распознавание именованных сущностей на основе дообученной модели BERT

Плюхин Д.А., Шилин И.А.
(Национальный исследовательский университет ИТМО, г. Санкт-Петербург)
Научный руководитель - к.т.н. Шилин И.А.
(Национальный исследовательский университет ИТМО, г. Санкт-Петербург)

В работе рассматривается задача извлечения именованных сущностей, описывается подготовка датасета для тестирования моделей, описываются архитектуры протестированных моделей, результаты подсчета метрик по полученным данным.

Извлечение именованных сущностей является одной из основных задач обработки естественного языка и тесно связана с автоматическим извлечением фактов из неструктурированных данных. Современные подходы к решению данной проблемы предполагают обучение и использование моделей глубоких нейронных сетей преимущественно архитектуры “Трансформер” с механизмом внимания. Одной из наиболее актуальных моделей для решения задачи извлечения именованных сущностей является BERT. На данный момент существует множество результатов работ по развитию модели BERT в частности и архитектуры “Трансформер” в целом как с акцентом на решение широкого спектра задач, так и с фокусом на решение только задачи извлечения именованных сущностей, при этом для русского языка остается актуальной проблема формирования реестра предобученных моделей с сопоставлением их эффективности, для чего также актуальна и проблема формирования корпусов, имеющих общепринятый формат (например, формат CoNLL-2003). Поэтому актуальной является задача применения модели BERT для извлечения именованных сущностей из текстов на русском языке.

Предлагаемое решение состоит из трех основных частей:

1. Слияние двух существующих корпусов с текстами на русском языке (в частности, factRuEval-2016 и persons-1000) и данными об именованных сущностях для формирования общего датасета, на котором можно в дальнейшем производить кроссвалидацию и сравнивать величины метрик качества работы оцениваемых моделей. Для формирования этой части решения были написаны скрипты для переформатирования входных датасетов, а также автоматизации частеречной и синтаксической разметки при помощи Stanford CoreNLP.
2. Адаптация модели под работу с новыми данными на русском языке - для этого были изучены файлы с параметрами конфигурации и модули стандартных (предобученных) BERT-моделей из пакета deeppavlov - они были соответственно модифицированы и дополнены так, чтобы была возможность выполнения дообучения моделей на новых данных и получения предсказаний в формате, удобном для дальнейшего подсчета метрик.
3. Скрипты для подсчета метрик используют стандартные функции для получения precision, recall и f1-score из пакета scikit-learn, однако расширяют функциональность этого модуля автоматическим формированием списка тегов из предоставляемых файлов с референсной и сгенерированной разметкой, а также производством нескольких типов усреднения как непосредственно самих значений метрик для одной пары исходных файлов, так и для нескольких пар для обеспечения кроссвалидации.

В результате были получены корпус в формате CoNLL-2003, модель BERT, дообученная на собранном корпусе, позволяющая сформировать для каждого словоупотребления один из 7 тегов с учетом BIO-разметки (B-PER, I-PER, B-LOC, I-LOC,

B-ORG, I-ORG, O), и значения метрик оценки качества системы распознавания именованных сущностей с помощью полученной модели. Для сравнения был использован подход, основанный на векторном представлении слов и символов (была использована модель fastText, обученная на корпусе новостей lenta.ru), рекуррентной нейронной сети (lstm) и последующим crf с использованием dropout и batch normalization.

По результатам сравнения были сделаны выводы:

1. Модель BERT позволяет получить большую точность на сформированном корпусе по сравнению с аналогом, поскольку на 49 метриках из 60 (82%) имеет более высокие значения показателей.
2. Дообучение дает положительный эффект - для базовой модели наблюдается улучшение значений 40 из 60 метрик (67%), а для модели BERT по 30 из 60 метрик (50%).

Авторы:
Плюхин Д.А. _____
Шилин И.А. _____

Научный руководитель:
Шилин И.А. _____