

МЕХАНИЗМЫ УЧЕТА СЕМАНТИКИ В ЗАДАЧАХ ПОНИМАНИЯ ЕСТЕСТВЕННОГО ЯЗЫКА

Письменный Алексей Алексеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: pismennyu.aleks@gmail.com

Научный руководитель — Соколов Евгений Андреевич

Использование предобученных языковых моделей, основанных на модели Transformer [3], привело к значительному успеху в решении широкого круга задач обработки естественного языка [1, 2]. При этом современные модели машинного обучения, применяемые для решения данных задач, зачастую обучаются без доступа к информации о структуре текста, связях между словами или их смысловой роли в данном тексте. Модели для решения относительно простых задач (к примеру, фильтрации спама) могут рассматривать текст как набор слов без информации об их порядке, в случае же более сложных задач, таких как построение вопросно-ответных систем или машинный перевод, модели в неявном виде учитывают информацию о структуре текста. При этом решения могут быть дополнительно улучшены путем использования внешних знаний (семантических ролей для слов [5] или бинарных отношений между словами [4]), возможно, также полученных с помощью моделей машинного обучения.

В данной работе предложен подход к учету информации, полученной с помощью анализа зависимостей (dependency parsing). Результатом анализа для конкретного предложения является ориентированный помеченный граф - дерево зависимостей, вершинами которого являются токены (слова, знаки препинания, иногда части слов), а каждое ребро обозначает определенное смысловое отношение между токенами. Был реализован частично основанный на архитектуре Transformer [3] модуль, принимающий на вход и обрабатывающий информацию о дереве зависимостей, который может быть интегрирован с любой моделью, выходом которой являются векторные представления всех токенов текста.

Исследовалась задача поиска диапазона правильного ответа (span-based question answering) на наборе данных SQuAD 2.0 (Stanford Question Answering Dataset). Каждый пример в данном наборе представляет собой поставленный вопрос и фрагмент текста, в котором требуется найти промежуток с ответом на данный вопрос. Дополнительная задача состоит в классификации примеров, в кото-

рых текст не содержит правильного ответа на вопрос. Чаще всего эти вопросы были сформулированы таким образом, чтобы они, во-первых, имели отношение к тексту (по смыслу), и, во-вторых, фрагмент содержал правдоподобный “ответ”. Для данной задачи было предложено решение на основе модели BERT-Base [1], с измененной относительно стандартного подхода [1, 2, 4, 5] функцией ошибки для обучения модели и соответствующей схемой определения диапазона ответа и примеров, в которых нет правильного ответа. Данное решение показало лучшее качество в сравнении с базовым [1] по метрикам EM (Exact Match) и F1.

Добавление к данной модели описанного выше модуля для учета информации из дерева зависимостей позволило также улучшить результат для обоих используемых метрик. Финальная модель включает в себя относительно небольшое число дополнительных параметров. Предложенный модуль может быть также использован для других задач обработки текстов.

Литература

1. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // In 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), Minneapolis, USA, 2019, P. 4171–4186.
2. Lan Z. et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations // International Conference on Learning Representations, Vienna, Austria, 2020
3. Vaswani A. et al. Attention is all you need // In Proceedings of the 31st International Conference on Neural Information Processing Systems, New York, USA, 2017, Long Beach, USA, P. 6000–6010.
4. Zhang Z. et al. Semantics-aware BERT for language understanding // In Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020, P. 9628-9635.
5. Zhang Z. et al. Sg-net: Syntax-guided machine reading comprehension // In Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020, P. 9636–9643.