УДК 004.896

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ, ОТНОСЯЩИХСЯ К НЕПОСРЕДСТВЕННОМУ ОБЪЕКТУ/ЯВЛЕНИЮ, УПОМИНАЮЩЕМУСЯ В ТЕКСТЕ

Сохин Т.Р. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – к.т.н., доцент ИДУ, старший научный сотрудник НЦКР Бутаков Николай Алексеевич (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В данной работе описывается механизм выделения из текста основных ключевых слов, которые являются основным смысловым элементом текста. Особенно нового подхода заключается в возможности работы с короткими текстами, что обеспечивается нейросетевой моделью.

Введение. Задачи идентификации аспектов и извлечения терминов остаются сложными в обработке естественного языка. Хотя методы с учителем достигают высоких результатов, их трудно использовать в реальных приложениях из-за отсутствия размеченных наборов данных. Подходы без обучения превосходят эти методы в ряде задач, но все еще остается проблема извлечения как аспекта, так и соответствующего термина, особенно в постановке с множественным тематикам. В этой работе мы представляем новую нейронную сеть с сверточным механизмом множественного внимания, которая позволяет извлекать пары (аспект, термин) одновременно и демонстрирует эффективность на реальном наборе данных. Мы применяем специальные функции , направленные на улучшение качества многоплановой экстракции. Экспериментальное исследование демонстрирует, что с этой потерей мы увеличиваем точность не только в этом совместном урегулировании, но также и на прогнозировании аспекта.

Извлечение аспектов применяется в различных задачах: анализ настроений, категоризация документов. В 2014 году была начата задача аспектного анализа настроений, и были достигнуты значительные результаты как в извлечении аспектов, так и в извлечении терминов аспектов на разных языках. Однако обычно тема - это общая категория, которая описывает весь документ или предложение. Это серьезная проблема в случае, если нам нужно знать больше деталей. Кроме того, некоторые тексты содержат несколько тем, которые обнаруживаются при моделировании тем, но они не позволяют определить, какая часть документа отвечает за конкретную тему.

Самый простой способ решить эту проблему - двухэтапное извлечение. Сначала мы обнаруживаем аспект, затем извлекаем термин, используя эти знания. Мы заинтересованы в системе, которая включает в себя обе возможности в своем ядре. Кроме того, эти подходы контролируются; это ограничивает их применимость в реальном мире.

Основная часть. Общая модель нейронного внимания для извлечения аспектов основана на идее ABAE, которая значительно превзошла предыдущие методы неконтролируемых предсказаний аспектов. В этой модели набор аспектов изучается в одном и том же пространстве вложения со словами и может быть легко интерпретирован человеком. Мы генерируем закодированное представление предложения, используя механизм внимания, и прогнозируем аспект, используя его. Поскольку мы работаем с неконтролируемой настройкой, в общем, это авто-кодировщик, и для обучения требуется реконструированное предложение. Этот подход предполагает реконструкцию в виде линейной комбинации представления закодированного предложения и матрицы AEM внедрения аспекта, где каждая строка представляет один аспект. Эта матрица дополнительно ограничивается в процессе обучения за счет использования ортогональных потерь. Несмотря на сходство на

уровне генерации матрицы аспекта, наш подход отличается в основе механизма внимания: мы стремимся собрать наиболее релевантную информацию из предложения для каждого аспекта, встречающегося независимо:

1. Предсказать внимание к каждому аспекту, используя CMAM.
2. Построить представления предложений для каждого из этих внимания - представлений предложений с вниманием.
3. Построить представление усредненного предложения с усреднением.
4. После того, как у нас есть вес аспектов, мы можем выбрать аспект аспекта (-ов) из соответствующих вниманий.

Полученная модель позволяет достичь хорошие результаты совместного извлечения аспекта и терма.

Выводы. В этой статье мы представили новую нейронную модель без обучения с учителем с механизмом сверточного множественного внимания для поиска аспектов с выделением связанных термов. В экспериментальном исследовании мы показываем эффективность нашей модели в задачах извлечения аспектов по сравнению с другими современными подходами и возможность правильной идентификации аспекта и аспектного терма.

Дальнейшая работа может быть направлена на усовершенствование процедуры вывода результата для включения множества аспектов с высокой прогнозируемой вероятностью в пределах одной и той же категории и их терминов, а также на модификацию механизма внимания для достижения более плавного распределения весов между словами предложения.

Сохин Т. Р. (автор)

Бутаков Н. А. (научный руководитель)