УДК 004.912

ВЫДЕЛЕНИЕ ОТРИЦАНИЯ И ЕГО КОНТЕСТНЫХ ЗАВИСИМОСТЕЙ С УЧЕТОМ ЛИНГВИСТИЧЕСКИХ ОСОБЕННОСТЕЙ РУССКОГО ЯЗЫКА

Марченко А.М. (Университет ИТМО), **Овчинников И.Д.** (Университет ИТМО; Dasha.AI)

Научный руководитель — профессор Шалыто А.А.

(Университет ИТМО)

Аннотация. В данной работе рассматривается проблема построения алгоритма выделения отрицания в текстах на русском языке. Для решения этой проблемы исследуются основные подходы для поиска отрицания в предложениях на английском языке и лингвистические особенности использования отрицания в русскоязычных предложениях. На основе рассмотренной информации описывается один из возможных подходов к реализации алгоритма выделения отрицания для текстов на русском языке.

Введение. Задача выделения отрицания очень важна в контексте анализа тональности, так как использование отрицания может полностью изменить эмоциональную окраску рассматриваемого текста. Решения данной задачи существуют в основном для текстов на английском языке, для русскоязычных текстов выделение отрицания остается слабо изученным.

В данной работе будут рассмотрены основные методы выделения отрицания для текстов на английском языке и изучены лингвистические особенности построения предложений с отрицанием в русскоязычных текстах. Также будет адаптирован один из рассмотренных методов выделения отрицания, используемый для текстов на английском языке, для использования в текстах на русском языке.

Основная часть. Все рассматриваемые в данной работе методы выделения отрицания для текстов на английском языке можно условно разделить на две категории:

- Методы, основанные на использовании дерева зависимостей
- Методы, основанные на использовании машинного обучения с учителем

Две данные категории методов являются одними из наиболее часто используемых в задачах поиска отрицания.

В методах, основанных на построении дерева зависимостей, применяется поиск ключа отрицания из специального словаря ключей в предварительно построенном с помощью некоторого инструмента (например, SyntaxNet или Stanford Parser) дереве зависимостей. Затем производится поиск родительского узла для найденного на предыдущем шаге ключа. После нахождения данного узла все его дочерние узлы (кроме найденного на первом этапе ключа) помечаются как область отрицания.

В методах, основанных на применении алгоритмов машинного обучения с учителем, используется предварительно размеченный набор данных для обучения некоторой модели (чаще всего LSTM или biLSTM), которая впоследствии будет использоваться для предсказания областей отрицания в неизвестных ей ранее предложениях.

Для адаптации данных методов к предложениям на русском языке следует сначала изучить лингвистические особенности применения отрицания в русскоязычных текстах. Все типы отрицаний в русском языке можно разделить на следующие категории:

- Семантически:
    - Общее
    - Частное
- Синтаксически:
    - Предикативное
    - Присловное

В зависимости от семантики рассматриваемого типа отрицания область его действия будет меняться. Так, например, для предложений с общим отрицанием областью действия будет являться все предложение целиком, а для предложений с частным отрицанием — только отдельно взятая его часть (например, причастный или деепричастный оборот). Разница же между предикативным и присловным отрицанием будет заключаться в типе родительского узла отрицания (в первом случае это, скорее всего, будет глагол, во втором — существительное или наречие). Несмотря на различия между отрицаниями разных типов, можно выделить основное правило определения области отрицания в русскоязычных текстах — отрицание всегда относится к вышестоящему узлу в дереве зависимостей рассматриваемого предложения.

Исходя из этой информации, можно предложить следующий алгоритм адаптации метода поиска ключа и области отрицания с помощью дерева зависимостей для использования с текстами на русском языке:

- Найти в предложении очередной ключ отрицания, принадлежащий словарю возможных ключей
- Для найденного ключа определить родительский узел в дереве зависимостей
- Пометить как область отрицания все дочерние узлы найденного родительского узла (кроме определенного ранее ключа отрицания)
- Повторять предыдущие пункты до тех пор, пока не будут рассмотрены все ключи отрицания, встречающиеся в тексте.

Выводы. Результатом данной работы стала одна из возможных адаптаций метода поиска ключа и области отрицания, основанного на использовании дерева зависимостей, с английского языка на русский. Предложенный адаптированный алгоритм может использоваться для улучшения качества работы различных методов анализа тональности и получения более глубокой информации о структуре предложений на русском языке в целом.

Марченко А.М. (автор)

Шалыто А.А. (научный руководитель)