Министерство образования и науки РФ
Национальный исследовательский
Томский государственный университет
Филологический факультет ТГУ

АКТУАЛЬНЫЕ ПРОБЛЕМЫ ЛИНГВИСТИКИ И ЛИТЕРАТУРОВЕДЕНИЯ

Сборник материалов I (XVI)
Международной конференции молодых ученых
(9—11 апреля 2015 г.)

Выпуск 16

Издательство Томского университета 2015

КОЧЕРГИНА К. С.

Алгоритм поиска оценочной лексики в конфликтных текстах

В данной статье приводится описание алгоритма поиска оценочной лексики. Тип искомой оценочной лексики конкретизируется исходя из заданной цели и специфики объекта исследования. Разрабатываемый алгоритм сопоставляется с методикой анализа тональности текста. Предлагается реализация данного алгоритма при создании экспертной системы для лингвистической экспертизы конфликтных текстов. Намечаются перспективы модификации предложенного алгоритма и экспертной системы в целом.

Ключевые слова: алгоритм поиска, оценочная лексика, конфликтный текст, лингвистическая экспертиза, экспертная система, анализ тональности текста.

Лингвистическая экспертиза относится к группе методов экспертной оценки текстов, часть из которых реализована в формате компьютерных программ[1] (например, автороведческая, почерковедческая). Экспертиза конфликтных текстов проводится преимущественно вручную с обращением к традиционным словарям русского языка. В связи с этим актуально создание *экспертной системы*, цель которой заключается в поддержке проведения лингвистической экспертизы за счёт интеллектуальной обработки естественноязыковой текстовой информации, то есть поиска и маркирования оценочной лексики в конфликтных текстах.

В процессе разработки такой экспертной системы выделяется четыре основных этапа: 1) составление «юрислингвистического» словаря оценочной лексики; 2) создание базы данных на основе этого словаря; 3) алгоритмизация поиска оценочной лексики в конфликтных текстах; 4) реализация алгоритма на языке программирования C++. Данная статья посвящена описанию третьего этапа, в рамках которого разрабатывается алгоритм поиска лексики, представленной в виде таблиц для базы данных, составленных на основе «юрислингвистического» словаря, который представляет собой перечень оценочных слов.

На данный момент в состав этого словаря включён ограниченный круг лексики — только имена существительные, обозначающие номинации лиц и/или совокупности лиц, имеющие в своём значении отрицательную, или потенциально-отрицательную, или положительную/нейтральную семантику (в случае их сочетания со словами-отрицаниями). Такое сужение материала связано с двумя причинами: во-первых, исходя из понятия конфликтного текста актуален, поиск в них *инвективной лексики* (употребляемой субъектом речи при намерении оскорбить адресата), а обозначенная выше группа слов может реализовать задачу оскорбления лица и/или совокупности лиц. Вторая причина заключается в простоте разработки и отладки алгоритма при работе с ограниченной группой слов на начальном этапе.

Суть экспертирования текстов при помощи разрабатываемой системы заключается в поиске и маркировании оценочной лексики, представленной в специальном словаре, что сопоставимо с методикой *анализа тональности текста* (англоязычный эквивалентный термин — sentiment analysis and opinion mining[2,3]), применяемой чаще всего к текстам, представляющим собой размещённые на сайтах отзывы пользователей о тех или иных товарах, услугах. Тональность — это отношение автора текста к объекту описания, выраженное через употребление оценочных слов; можно провести аналогию с актом оскорбления, где инвективное слово выражает отношение автора текста к объекту описания (объекту оценки). Метод анализа тональности текста основан на поиске оценочной лексики в тексте по заранее составленным словарям и правилам, специфичным для различных предметных областей. Каждому слову присваивается соответствующее значение тональности, общая тональность текста определяется как сумма оценочных весов (значений) каждого слова, и текст оценивается как позитивный/негативный/нейтральный.

В предлагаемом в данной статье алгоритме отсутствует последний шаг — вычисление общей тональности текста, что объясняется отличием целей этого метода (и объекта его исследования) от целей лингвистической экспертизы конфликтных текстов. При её проведении главным является не оценивание тональности всего экспертируемого текста, а установление факта наличия в тексте оценочных слов. Употреблённая в тексте инвективная лексика выводится в контекстуальном окружении и маркируется: каждому найденному оценочному слову присваивается вес, который означает принадлежность слова к тому или иному классу инвективной лексики.

Для работы алгоритма поиска лексики требуется *предварительная подготовка* естественноязыкового представления текста к машинной обработке. Процедура поиска оценочной лексики в тексте начинается после работы модуля лингвистического анализа. Модуль лингвистического анализа производит формализацию исходного текста, то есть перевод текста из естественноязыкового представления в формальное, пригодное для машинной обработки, при котором текст хранится в виде совокупности синтаксических графов (как правило, один граф соответствует одному предложению).

Этап собственно поиска. Поиск осуществляется среди вершин графа, которые соответствуют словам из текста, приведённым к начальной форме. Производится последовательный обход всех графов и сравнение каждого слова-вершины со словами из базы данных, причём для сравнения берутся не все слова, а только определённые как имена существительные (в случае отсутствия таковых происходит возврат на предыдущий шаг). Если рассматриваемое слово совпадает с одним

из слов в базе данных, то ему присваивается числовое значение — вес, адекватный предметной области документа (в данном случае — класс инвективности). Граф, содержащий вершину с таким словом, сохраняется отдельно для дальнейшего вывода контекста употребления. Таким образом, формируется новая совокупность графов с маркированными вершинами, которые соответствуют контекстам употребления оценочной лексики. Когда достигнут конец текста (пройден последний граф), пользователю выводится список найденных лексем и контексты их употребления.

Полученные результаты являются конечным продуктом работы программы, но не итогом лингвистической экспертизы, поскольку требуют *последующей интерпретации* специалистом на основании его профессиональных знаний. Такое ограничение работы экспертной системы связано с недостаточным уровнем разработанности исследований в области анализа семантики текста в современной компьютерной лингвистике.

В перспективе планируется разработка и интеграция автоматического семантического анализатора для реализации функции определения адресного/безадресного характера употребления инвективной лексики с дальнейшим установлением объекта оценки. Также доработка интерфейса программы может быть направлена на приведение толкования маркированного слова и/или вывод ссылки на толковый онлайн-словарь, что актуально для оптимизации процесса работы лингвиста-эксперта. Возможно усовершенствование и самого алгоритма поиска оценочной лексики, в том числе расширение спектра искомой лексики и добавление поисковых задач, например, определение характера оскорбления как «юридического» или «бытового».

Примечания

1. *Митина О. В., Евдокименко А. С.* Методы анализа текста: методологические основания и программная реализация // Вестник ЮУрГУ. Серия: Психология.— 2010.— № 40 (216).— С. 29–38.
2. *Pang B., Lee L.* Opinion mining and sentiment analysis.— Foundations and Trends, 2008.— 135 p.
3. *Liu B.* Sentiment analysis and opinion mining.— Morgan & Claypool Publishers, 2012.— 168 p.