Автоматическое построение словаря тональных модификаторов

Нгуен Кхань Кхуен

Студент

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

E-mail: nguenkhan@ya.ru

Научный руководитель *— Ефремова Наталья Эрнестовна*

Тональные модификаторы играют важную роль при определении тональности текстов: с их помощью составляются правила композиции тональности и выделяются признаки для применения методов машинного обучения.

Выделяют следующие виды модификаторов:

- **модификаторы конверсии** преобразуют исходную тональность выражения в обратную, например, не, отсутствие: новость меня обрадовала (+) – новость меня не обрадовала (–);

- **модификаторы-доминаторы** определяют тональность выражения независимо от тональности соседнего слова: красивая/глупая смерть (–);

- **модификаторы-распространители** «распространяют», т.е. усиливают или снижают тональность соседнего слова: увеличение безработицы (–) – уменьшение безработицы (+).

Данная работа посвящена автоматическому составлению словаря модификаторов для русского языка. Для этого была проведена адаптация и доработка метода, описанного в [2]. Метод позволяет автоматически генерировать словарь модификаторов для различных предметных областей и не требует размеченных данных для обучения, кроме тонального словаря, который содержит позитивные и негативные слова.

Генерация словаря включает несколько этапов:

- предсказание тональности слов и выражений с помощью метода опорных векторов (SVM);

- использование правил для определения принадлежности слова к одному из классов модификаторов на основе статистических признаков.

Сначала для всех слов и выражений вычисляется вектор признаков совместной встречаемости с заведомо положительными и отрицательными словами из тонального словаря с использованием меры Positive Pointwise Mutual Information. На основе полученных векторов признаков для слов из тонального словаря обучается SVM, который предсказывает тональности остальных слов и выражений.

Далее, на основе полученных предсказаний, с помощью правил формируются слова-кандидаты в классы и для каждого из них вычисляется вероятность отнесения к выбранному классу. Для каждого класса выбираются пороги значений вероятности для окончательного определения принадлежности слов к классам.

В данной работе предложенный метод был реализован для новостного корпуса, а в качестве тонального словаря использовался словарь РуСентиЛекс [1]. В дальнейшей работе планируется оценить полученный словарь в задаче анализа тональности выражений и рассмотреть вопрос его адаптации к различным предметным областям.

Литература

1. Лукашевич Н. В., Левчик А. В. Создание лексикона оценочных слов русского языка РуСентилекс // Труды конференции OSTIS-2016. 2016. С.377–382.
2. Orith T., Roy B. Learning Sentiment Composition from Sentiment Lexicons // Association for Computational Linguistics, 2018. P. 2230–2241.