Секция «Вычислительная математика и кибернетика»

Кластеризация оценочных слов по тональности на основе Марковских случайных полей

Четвёркин Илья Игоревич
Аспирант
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: ilia2010@yandex.ru

Задача анализа отзывов пользователей (по отношению автора к объекту, тональности, эмоциям) является очень актуальной и имеет ряд возможных приложений, например, сбор и обработка мнений пользователей о новых продуктах по блогам.

Отправной точкой для различных задач анализа мнений является создание словаря оценочной лексики. Каждое слово (или выражение) в таком ресурсе должно обладать некоторой тональностью, например положительной или отрицательной. С помощью такого рода словарей решаются различные задачи, такие как классификация или извлечение мнений [1]. В данной работе описывается двухэтапная модель для извлечения и определения тональности набора оценочных слов. На первом этапе из коллекции отзывов извлекается список оценочных слов, для чего используется совокупность разнообразных признаков слов и методы машинного обучения [1]. Для дальнейшей кластеризации по тональности отбираются наиболее вероятные оценочные слова.

На втором этапе выполняется кластеризация оценочных слов по тональности, основанная на предположении, что оценочные слова, которые идут в тексте друг за другом на небольшом расстоянии, более вероятно имеют одинаковую тональность.

Для формализации данной гипотезы используется модель Изинга [2], в которой вместо электронов представлены оценочные слова, каждое из которых может иметь тональность -1 или +1. Общая энергия системы задается функцией от парных взаимодействий и начальных значений для каждого оценочного слова:

$$E(x, W) = -\sum_{ij} w_{ij} x_j x_j - \sum_i h_i x_i$$

где x_iзадает тональность оценочного слова, w_{ij}вес связи между словами, пропорционален частоте совместной встречаемости и обратно пропорционален среднему расстоянию между словами, h_iнормированное отклонение от средней оценки слова [1], вычисленное по всему корпусу отзывов. Связь между двумя оценочными словами в корпусе добавлялась только в случае если частота совместной встречаемости и среднее расстояние были выше заранее заданного порога. Для поиска согласованного состояния Марковской сети использовался алгоритм распространения доверия.

Описанная двухэтапная модель была применена к коллекции из 28773 отзывов о фильмах, из которой на первом этапе было извлечено 3000 наиболее вероятных слов. На втором этапе из данных слов был построен граф связанности и найдено согласованное состояние сети. Метрикой качества кластеризации являлась правильность классификации (отношение правильных решений к общему числу решений), которая составила 82.7%.

Таким образом, в данной работе предложен алгоритм, который позволяет извлекать оценочные слова с тональностями для заданной предметной области, на основе текстовой коллекции.

Литература

1. Chetviorkin I. I. , Loukachevitch N. V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // In Proceedings of COLING 2012: Technical Papers , pages 593–610

2. 2. Takamura H., Inui T., and Okumura M. Extracting Semantic Orientations of Words using Spin Model. // In ACL, 2005. pp 133–141