Реферат Щенникова Никиты

по статье Дубатовки, Курочкина и Михайловой по теме

“Автоматическая генерация тематических тональных словарей”

В данной статье рассматривается один из способов создания тематического тонального словаря. Одним из больших плюсов данного подхода является то, что для этого подхода не нужен предварительный лейбелинг людьми. Для реализации они используют модель графов и большой корпус данных относящихся к одной теме. Еще одной отличительной особенностью является то, что этот алгоритм можно использовать не только для русского языка.

В качестве сравнения они приводят различные подходы основанные на модели графов, но для них используются предразмеченные данные, что ставит авторов данной статьи в более выигрышное положение в начале. Они приводят в пример метод основанный на переносе тонального словаря из одной темы на другую с помощью графов, метод основанный на словарях других языков, чтобы создать словарь для языка являющегося конечной целью. Таким образом авторы статьи объединяют задачи и пытаются сделать алгоритм который будет максимально универсальным, подходящим для любой темы или языка.

Для своего эксперимента они взяли датасет из 259 023 немаркированных отзыва к отелям. Впоследствие, для проверки точности своих данных они размаркировали 500 отзывов и сделали три тональных словаря: положительный, отрицательный и нейтральный, объёмами 970, 1000, 2591 слов соответственно.

Алгоритм построен на принципе графов, где каждому слову находилось противоположное или ближайшее. Вершинами графа являлись прилагательные, гранями - связи между ними. Связи между словами также зависели от союзов “и” и “но”. За счёт таких связок в том числе строились связи между словами, “и” создавала позитивную, “но” негативную. Связи строились по такому принципу: (ADV | NEG)∗ADJ(,?(AND | BUT)?(ADV | NEG)∗ADJ)+. В Этом процессе проверяется наличие наречий меры или степени, союзов в связке с самими прилагательными. Получившийся граф было необходимо разделить на два кластера, негативный и позитивный. Для этого они инициализировали оба кластера маленьким набором слов: хороший, отличный, замечательный, прекрасный, лучший для позитивного, и для отрицательного: плохой, ужасный, отвратительный, худший, бедный. Потом иттеративно добавлялись слова по одному в один из кластеров для создания словаря. Ближайшие соседи добавленного слова также добавлялись в тот же кластер, после чего пересчитывались связи между словами. Веса для граней считались по этой формуле: 𝑤eight(wor𝑑1,𝑤or𝑑2)=#(word𝑑1𝐴ND 𝑤or𝑑2)−𝐾∗#(𝑤or𝑑1𝐵UT 𝑤or𝑑2).

Для определения в какой из двух кластеров выделять слово они решили сравнить два метода, для одного они добавляли слово в тот кластер с которым у него самая “тяжёлая” грань, для другого они выбирали по сумме связей с вершинами графа в каждом кластере.

Для расчёта точности они решили рассмотреть два варианта программы, с префиксами “не”, и без них. В итоге различия в словаре получились очень незначительными из-за метода графов, который выбрали авторы. В итоге у них получилось, что Precision больше, когда префиксы “не” исключены (79,9%), а Recall больше, когда включены (75,4%). Можно сказать, что для метода без учителя такая точность довольно неплохая, однако для использования в реальных задачах нужна большая точность. С такой точностью, конечно, можно привлечь стороннюю помощь и подкорректировать вручную, однако, это идёт вразрез с идеей работой, созданием тонального словаря автоматически, без помощи человека.

Мне кажется этот подход довольно интересным за счёт того, что не используются предварительно маркированные данные, соответственно упрощается работа аналитика и процесс ускоряется. Однако, точность пока нельзя назвать удовлетворительной для работы при помощи данного алгоритма. Необходимо подумать, как можно улучшить точность весов между вершинами для более точной оценки тональности. Мне показалось, что данная формула имеет интересные идеи, например зависимость веса между прилагательными от союзов, но делать этот фактор основным может быть не очень рационально. Не совсем понятно из статьи, как берётся вес одного слова, по которому считается вес грани, поэтому точно сказать, как можно улучшить формулу, затруднительно.

В заключение, я хотел бы повторить, что идея обучения без учителя для создания тонального словаря, мне кажется очень перспективной и было бы интересно подумать, как можно было бы улучшить данный метод для получения большей точности.