**РЕФЕРАТ НА СТАТЬЮ**

**Automatic Generation of the Domain-Specific Sentiment Russian Dictionaries,**

**Dubatovka A., Kurochkin Yu., Mikhailova E.**

Авторы: Быстрова Ольга, Орлов Александр, Григорьев Эдуард

В данной статье описывается алгоритм создания тонального словаря (sentiment dictionary) для определённых доменов/аспектов. Основным достоинством алгоритма, предлагаемого Дубатковой и коллегами является то, что он не требует предварительной ручной разметки. Для алгоритма нужен только «достаточно большой» (никаких более конкретных значений того, насколько большим должен быть корпус авторами не приводится) корпус текстов и набор базовых оценочных прилагательных, полярность которых одинакова в любых аспектах языка (*хороший/плохой* и т п). Несмотря на то, что алгоритм в статье применяется к русскому языку, при желании его можно использовать и на материалах других языков.

Авторы сравнивают свой алгоритм с как минимум 6 другими алгоритмами для разных языков. Ключевым отличием цитируемых работы от алгоритма Дубатковой и коллег является то, что они все так или иначе используют какие-либо материалы, которые были предварительно (частично )собраны в ручную (собранный в ручную словарь тональной лексики для английского и испанского языка в (Steinberger et al., 2012), собранный в ручную словарь для немецкого в (Clematide, 2010), предобученная модель в (Chetviorkin I. I., Loukachevitch N. V., 2012) и т п )

Работа алгоритма основывается на наблюдении, что сочинительные союзы, как правило, связывают прилагательные и наречия, которые относятся к одному и тому же домену, и либо совпадают по полярности (в случае связи посредством соединительных союза типа *и*), либо противопоставлены по ней (в случае противительных союзов типа *но*).

Для решения поставленной задачи авторы предлагают создать взвешенный граф, где узлы это прилагательные, рёбра между которыми образуются в случае наличия противительных или соединительных связей между ними. Вес рёбер авторы высчитывали по формуле:



, где К – специальная константа (coefficient of the adversative conjunctions relevance), которая добавляется потому, что противительные связи встречаются в текстах значительно реже сочинительных. Чему был равен K в статье авторов не известно.

При этом при вычислении рода связи, следует смотреть не только на союзы. Так, во фразе *вкусная но не дорогая еда,* между прилагательными вкусный и дорогой будет установлена противительная связь, потому что *и не* указывает на то, что понятия относятся к разным тональностям, а не к одной (*но не*, же наоборот, указывает на общую тональность, а не на противоположные). Набор таких “переформулировок” может зависеть от языка (так, к примеру, слова с отрицательной приставкой *un* в английском могут анализироваться не как отдельные слова, а как not + прилагательное без приставки).

Первоначальные позитивный и негативный кластеры графа составляются из базовых универсальных слов, полный список которых для русского языка авторами не приводится. Именно к ним после обработки материала присоединяются последующие узлы.

Для демонстрации работы алгоритма авторы взяли корпус из 259 023 неразмеченных отзывов на отели по всему миру. Тексты содержали опечатки, грамматические ошибки и неформальную лексику. Также авторы упоминают, что часть работ часть текстов содержала информацию о несвязанных вещах (качество перелёта, экскурсий). То, как авторы боролись с этой проблемой, в статье не описывается.

В рамках своего исследования на этом датасете авторы хотели проверить будет ли алгоритм работать лучше, если рассматривать un как часть слова, либо же как отдельный элемент логической структуры[[1]](#footnote-1) .

В работе авторы оценивали точность разделения слов на три класса: позитивные, негативные и нейтральные, и на два: позитивные и негативные. Для измерения точности использовалось соответствие полученных в результате алгоритма данных словарям позитивных, негативных и нейтральных слов, полученным в результате ручной разметки отдельных частей корпуса (500 рандомных отзывов). Такой способ вычисления точности не позволяет оценить точность по каждому слову, выделенному алгоритмом, но при подобном объёме данных иного варианта нет. Ошибки в статье не анализируются.

Авторами было применено несколько алгоритмов вычисления точности, отличающихся в деталях, но в целом все они дают одинаковый вывод: при разделении на три класса, не имеет значения, воспринимаем ли мы un как часть слова или как отдельный элемент структуры, однако при разделении слов на два класса (что во многом, действительно, является иной задачей), отделение приставки un от слова даёт существенный рост качества, особенно при вычислении слов с положительной тональностью. Анализ таблиц результатов из статьи также показывает, что, если мы не отрываем *un,* то слов находится больше, что логично.

Вторая вещь, которая осталось непонятной нам по прочтении статьи – это то, что на стр. 7 авторы упоминают, что существует несколько способов подсчёта дистанции от слова до конечного множества. Так, очевидно, что ряд слов в графе будут связаны более, чем одним ребром. В таком случае расстояние можно считать либо по «самому тяжёлому» ребру, либо по приведённой выше формуле (или же аналогичной ей). На стр. 7 авторы пишут, что они сравнили 2 варианта. В итоге, использовали второй вариант, но почему они считают его лучше не указано.

В целом, работа является инновационной потому, что она предлагает универсальный способ создания тональных словарей для конкретного домена без предварительных затрат на ручную разметку. Алгоритмы обеспечивают построение словарей с точностью (precision) 79,9% и полнотой (recall) 75,4%. Однако оставляет вопросы применимость данного метода к другим частям речи, помимо прилагательных и наречий: к примеру, к оценочным существительным типа *оболтус*, *дурак*, которые, выступая в предикативной позиции, редко связываются сочинительными союзами, а также к длинным словосочетаниям и глаголам. Данный алгоритм также не применим для выявления оценочной лексики в отдельных аспектах домена: авторы выделили положительные и отрицательные прилагательные для отзывов на отели (что является достаточным для большинства коммерческих запросов), но алгоритм в данном виде не применим для выделения прилагательных относящихся к отдельным подаспектам сферы: сервис/еда/интерьер.

Steinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M. A., Lenkova P., Steinberger R., Tanev H., Vazquez S., Zavarella V. (2012), Creating sentiment dictionaries via triangulation, Decision Support Systems, Vol. 53(4), pp. 689–694.

Clematide S., Klenner M. (2010), Evaluation and extension of a polarity lexicon for German, Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), Lisbon, pp. 7–13.

Chetviorkin I. I. and Loukachevitch N. V. (2012), Extraction of Russian sentiment lexicon for product meta-domain, Proceedings of COLING 2012: Technical Papers, Mumbai, pp. 593–610.

1. Этот момент остался не очень нам понятен, так как название статьи включат в себя слово Russian и авторы говорят о том, что будут применять алгоритм к русскому языку. Однако в статье нет ни одного примера на русском, и не очень понятно, что имеется в виду под приставкой un: какая-то русская приставка (типа *без,* *вне (*есть ощущение что такие приставок в русском можно выделить несколько и все они имеют немного иную семантику, чем просто *un)*), или же всё же английская приставка un и результаты в эксперименте были получены на базе английского языка [↑](#footnote-ref-1)