**Заголовок:** Automatic generation of the domain-specific sentiment Russian dictionaries (2016)

**Ссылка:** <http://www.dialog-21.ru/media/3388/dubatovkaaetal.pdf>

**Выполнили:**

* Даша Петрова
* Настя Каприелова
* Настя Кромина
* Андрей Зырянов

В статье авторы представляют алгоритм генерации предметно-ориентированного тонального словаря для русского языка с использованием графовой модели. Из преимуществ можно назвать то, что данный алгоритм не требует ручной человеческой разметки, а также может применяться для разработки словарей на других языках, не только на русском.

Авторы статьи упоминают ряд работ, которые являются основополагающими в написании данной. Среди них расширение существующего словаря с помощью построения графа на корпусе без разметки; улучшение графа с помощью анализа его подграфов с использованием тезауруса и другие. Однако свою работу авторы ни с какой другой не сравнивают. Она принципиально отличается от предыдущих тем, что описываемый в ней алгоритм автоматический и может работать для многих языков.

Алгоритму, о котором идёт речь в статье, на вход подавались деперсонализированные немаркированные отзывы о различных отелях по всему миру. Количество рецензий составило 259 023 оригинальных текстов. Размер датасета получился равен 660 Мб. Для предобработки данных использовался Mystem, позволяющий нормализовать слова и получить необходимую грамматическую информацию.

Помимо этого для того, чтобы оценить точность полученных графовым методом словарей, авторы разметили вручную все прилагательные, которые были выделены рассматриваемыми алгоритмами, на три тональные группы. Тем самым они собрали дополнительные большой словарь с положительной (970), отрицательной (1000) и нейтральной (2591) лексикой.

Для того, чтобы оценить полноту, авторы взяли из исходного датасета 500 случайно выбранных отзывов и разметили их вручную, приписав каждому прилагательному положительную или отрицательную оценку. В итоге “ручной” словарь с позитивной лексикой содержал 173 слова, словарь с негативной лексикой - 127 слов.

Для каждой пары прилагательных при помощи специального шаблона устанавливались тональные связи, и в графе вес ребра между словами в паре определялся количеством контекстов, в которых они встречались с позитивной или негативной связью. Далее итоговый граф делился на два кластера в зависимости от тональности при помощи начальных сетов универсально позитивных и негативных слов. Каждая пара итеративно была отнесена к какой-то группе в зависимости от близости к каждому сету. При этом кандидатом считался тот узел, который имел хотя бы одно общее с группой ребро.

Данный графовый алгоритм уникален тем, что граф строится на неразмеченном корпусе и в дальнейшем может быть построен на любых других корпусах автоматически без субъективной разметки человека. Более того, алгоритм может быть использован на других языках кроме русского, достаточно лишь иметь на этом языке морфологический анализатор, список соединительных и противительных союзов и набор универсальных оценочных слов (слов, которые в не зависимости от домена имеют всегда положительную или всегда отрицательную коннотацию).

Алгоритмы демонстрировали достаточно высокий параметр recall (порядка 0,8), т.е. среди всего массива истинно позитивных результатов (как с отрицательной так и с положительной тональностью) значительная часть была определена верно. Параметр precision был низким (порядка 0,3 - 0,5) при анализе корпуса, включавшего нейтральные слова, и повышался когда нейтральные слова исключались из анализа (достигал уровня 0,7 - 0,8). Таким образом, при наличии нейтральных слов в анализируемом корпусе, среди предсказанных значений большинство являются неверными предсказаниями.

Авторы не проводят непосредственный анализ ошибок алгоритма. Однако в своей работе упоминают трудности, которые могут их повлечь. Например, во фразе “the pool was large but not deep” алгоритм выделит тональные слова “large” и “deep”, каждое из которых будет иметь положительную коннотацию, что, на самом деле, не так. Аналогично алгоритм ошибается во фразах типа “food was delicious and not expensive”, выделяя “expensive” негативным маркированием.

Ключевым преимуществом этой технологии является ее независимость от данных ручной разметки. Всю предобработку данных для анализа с помощью этого алгоритма можно также провести автоматическими методами, что облегчает и ускоряет решение задачи. Но нам кажется неясным, каким образом предполагается решить проблему, вызванную зависимостью между точностью предсказания и наличием в анализируемом словаре нейтральных слов. Точность алгоритма демонстрирует высокие значения только в случае, когда нейтральные слова исключены из анализируемого словаря. Следовательно, для достижения максимальных результатов необходимо решить дополнительную задачу исключения нейтральных слов, для решения которой, возможно, не удастся избежать ручной разметки.

При подсчитывании близости у авторов возникла проблема с тем, что стоит выбирать: “тяжелый вес” одного ребра или сумму множества “легких” ребер, и они решили сравнить эти подходы. Далее они тут же констатируют два вывода, никак их не аргументируя:

1. расстояние между узлом и сетом - вес наиболее тяжелого ребра
2. вес ребра - сумма весов ребер, связанных с рассматриваемым множеством, вычитаемая из суммы весов ребер, связанных с противоположным множеством

В связи с этим осталось непонятным то, почему они выбрали именно тот подход, о котором написали.