DOI: 10.30888/2415-7538.2018-09-1-005

УДК 004.912

SENTIMENT ANALYSIS OF NEWS REPORTS
ОЦЕНКА ТОНАЛЬНОСТИ НОВОСТНЫХ СООБЩЕНИЙ

Asiryan A.K. / Асирян А.К.

Московский государственный университет им. М.В.Ломоносова, факультет вычислительной математики и кибернетики, Москва, Ленинские Горы 1, стр. 52, 119234
Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Moscow, Leninskie Gory 1, str. 52, 119234

Аннотация. В работе представлен алгоритм оценки тональности информационных сообщений по отношению к объектам, адаптирующийся под предметную область. Рассматриваются существующие методы определения эмоциональной оценки текстов. Разработанный метод использует словарь тональных слов, составляемый при анализе конкретной предметной области, на основе множества слов, эмоциональная окраска которых не зависит от контекста.

Ключевые слова: обработка естественного языка, автоматический анализ тональности, лексикон оценочных слов.

Вступление.

В настоящее время задачи определения эмоциональной оценки текста и извлечения мнений пользователей являются одними из самых популярных в области обработки текста. Они нашли применение во многих сферах, например: анализ восприятия товаров и услуг, мониторинг блогов, политические исследования, киноиндустрия, бизнес, сбор мнений о компаниях и т.д. При этом некоторые слова могут принимать разную окраску в зависимости от предметной области. Так, в словосочетании «разбитая на модули программа» слово «разбитая» имеет положительную оценку, а в «разбитый асфальт» – негативную. В связи с этим интерес представляет тональный анализ, учитывающий предметную область. В исследовании не рассматриваются сообщения из блогов и форумов, так как из-за частых ошибок в тексте, употребления жаргонов и отличиях в синтаксических структурах предложений для их обработки необходимы другие подходы. Целью работы является определение эмоциональной окраски новостных сообщений с установлением объектов, подвергшихся оценке.

Существует множество различных работ, посвященных тональному анализу. В статье [1] авторы предлагают использовать набор вручную написанных синтаксических правил, сверточные нейронные сети и их комбинацию. Происходит выделение фактов-мнений с подсчетом их оценки. Далее, каждая сущность связывается с фактом-мнением, и для нее считается оценка настроения. Составлялись дополнительные правила в зависимости от предметной области. В [2] применяется комбинация метода опорных векторов с адаптацией под тематику и двух нейронных сетей: сверточной и рекуррентной. Эти две работы показали наилучшие результаты на соревновании SentiRuEval-2016 [3], проведенном на конференции «Диалог». Также есть подходы, основанные на словарях оценочной лексики. Например, в статье [4] приводится способ построения такого словаря с последующим применением для обучения в методе опорных векторов. Подход, описанный в [5], сопоставляет вручную

написанные шаблоны с деревом зависимостей предложения. Найденные поддеревья с помощью различных эвристик соединяются с объектом, который они описывают. Все методы показали достаточно хорошие результаты относительно других подходов, но еще не сравнимые с разметкой экспертов, что оставляет задачу открытой для исследования.

Основной текст.

Подход, предложенный в данной курсовой работе, основывается на выявлении синтаксических связей между словами, несущими оценочный характер, с как объектами, к которым относится эта оценка, так и другими тональными словами. Для этого необходимо уметь строить дерево зависимостей предложения, отражающее его синтаксическую структуру. Большинство инструментов для русского языка являются закрытыми коммерческими проектами, например, ЭТАП-3 и ABBYY Compreno, показавшие на конференции «Диалог» лучшие результаты [6]. Существует небольшое количество инструментов с открытым исходным кодом, позволяющих решить данную задачу, из них выделяются MaltParser [7], использующий обучение с учителем вместе с грамматикой зависимостей. Он выявляет связи непосредственно между словами: каждой вершине в полученном дереве отвечает ровно одно слово. Недостатком подхода, использующего машинное обучение, является необходимость обучающей выборки, но благодаря корпусу СинТагРус [8], входящему в национальный корпус русского языка (НКРЯ), такую выборку можно получить для исследовательских целей. MaltParser позволяет строить дерево зависимостей, не ограничиваясь одним набором частей речи и соответствующими граммемами, давая право выбрать любой инструмент морфологической разметки. В итоге в качестве инструмента синтаксического анализа текста был выбран MaltParser. Для его обучения было необходимо провести морфологический анализ и выделить нормальные формы слов. Эта информация представлена в корпусе, но так как реальный текст размечается инструментами, то и обучиться необходимо на тексте, размеченном ими. После обучения можно приступать к анализу тональности.

Сначала текст разбивается на предложения с помощью инструмента NLTK [9]. Далее, предложения разбиваются на токены посредством Greeb [10]. Происходит обработка полученных токенов: если три токена стоят друг за другом в тексте, и вторым токеном является дефис, то они конкатенируются в один. После этого путем обработки MyStem [11] и pymorphy2 [12] все токены получают информацию о лемме и граммемах. Если в процессе работы MyStem токен был разбит на еще несколько токенов, то лемма конкатенируется, а часть речи берется из последнего токена, так как почти всегда изначальный токен является сложным словом. Далее, информация о каждом предложении преобразуется в CoNLL-X формат и подается на вход MaltParser для построения деревьев зависимостей.

Основой разработанного алгоритма стала идея, предложенная в статье [13]. В полученных деревьях зависимостей ищутся связи между оценочными словами. В качестве таких слов рассматриваются только прилагательные и

причастия. В рамках одного предложения связь между ними можно описать следующим регулярным выражением:

(PRE|NEG)*W((AND|BUT)?(PRE|NEG)*W)+, где

PRE=ADV|APRO – слово, усиливающее оценку (очень, самый и т. д.),

NEG=PART – отрицание (частица не),

W=A|V,partcp – прилагательное или причастие,

AND=CONJ – соединительный союз (и, также, тоже, ни),

BUT=CONJ – противительный союз (а, но, зато, однако).

По нему можно построить детерминированный конечный автомат (ДКА) для обхода дерева зависимостей. Так как для вершины отвечающей W выражения (PRE|NEG)* и (AND|BUT)? соответствуют двум путям с ней в качестве начальной вершины, то первую ветку можно опустить, исследовав ее только при определении вида связей. Это упрощение превращает отображение регулярного выражения на дерево из поддерева в простой путь. Итоговый автомат приведен на рисунке 1.

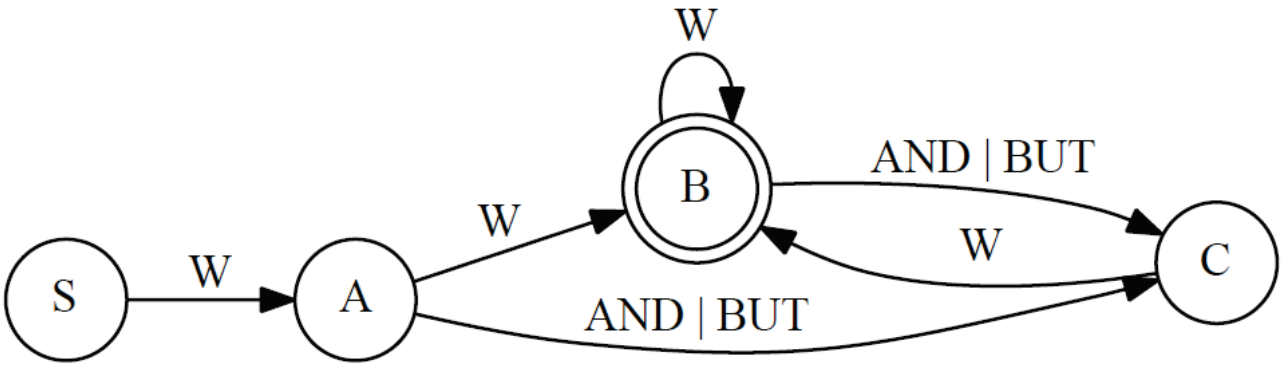

Рис. 1. ДКА для выделения связей между словами с оценкой

Для простоты анализа слова с приставкой «не» и без считаются одинаковыми (например, «хороший» и «нехороший»). Связи делятся на положительные и отрицательные. Первые ставятся между словами, связанными соединительными союзами, вторые – противительными. Также учитывается отрицание самого слова с помощью либо частицы, либо приставки «не». Для ясности следует рассмотреть пример построения связей на предложении «Фильм был хорошим, крайне интересным, но не очень понятным и затянутым.». Между словами «хорошим» и «интересным» связь положительная, но между «понятным» и «затянутым» – отрицательная, так как перед первым из них стоит отрицание. Союз «но» разделяет слова «хорошим» и «затянутым» отрицательной связью, но «хорошим» и «понятным» связаны положительно. Аналогично разбирается связь слова «интересным». Стоит отметить, что, хотя союз «а» и является противительным, его обработка идентична соединительным союзам, так как он имеет смысл только вместе с частицей «не». Каждому слову в соответствие ставится вершина, а ребро отвечает связи (как положительной, так и отрицательной). Таким образом, после обработки текста получается неориентированный граф, отражающий связи оценочных слов.

Инициализация графа происходит с использованием словаря тональной

лексики. Среди всех имеющихся разработок выделяется проект LINIS Crowd SENT [14] – тональный словарь и коллекция текстов с тональной разметкой, в которой может принять участие любой человек через официальный сайт. Главным отличием проекта является не просто набор слов, а информация о множестве оценок, поставленных пользователями. Таким образом, можно выделить слова, тональность которых не меняется при смене контекста. Количество оценок одного типа для слова должно быть больше или равно некоторому порогу.

Вес ребра рассчитывается как разность положительных и отрицательных связей. Расстояние до множества считается как сумма всех ребер, связывающих вершину и вершины из множества. При первом обходе графа строится множество вершин, соединенных с уже оцененными на этапе инициализации. Выбор очередной вершины для оценки выглядит следующим образом:

1. Сначала из построенного множества рассматриваются вершины, соединенные и с набором положительных слов, и с набором отрицательных. Это сделано для того, чтобы один набор не мог расширяться вплоть до границы другого. Если таких вершин нет, то выбор осуществляется среди оставшихся.

2. Среди множества выбранных на предыдущем шаге вершин, берется имеющая большую разность расстояний до положительного и отрицательного множеств.

3. Если на предыдущем этапе под критерий попадает несколько вершин, то из них выбирается та, чье добавление принесет максимальное изменение расстояний для соединенных с ней вершин.

4. Если снова сразу несколько вершин удовлетворяет условию, то выбирается первая вершина.

Выбранной вершине присваивается позитивная оценка, если ее вес неотрицательный, иначе – негативная. Все смежные неоцененные вершины добавляются в обрабатываемое множество. Процесс продолжается до тех пор, пока есть необработанные вершины, соединенные хотя бы с одним из множеств.

Далее, вершины графа отображаются на исходный текст и связываются с оцениваемыми объектами. Под объектом в данной работе рассматривается именованная сущность. Для извлечения сущностей был использован инструмент FreeLing [15]. Затем применяется алгоритм, схожий с подходами, описанными в [1, 5]. Сначала по каждому найденному *тональному слову* (ТС) определяется *тональный участок* (ТУ) – поддерево в дереве зависимостей предложения, корнем которого является тональное слово, а вершины не принадлежат другим тональным участкам. Именованная сущность связывается с участком, если длина пути до него в дереве меньше некоторой константы. Полученный путь вместе с тональным участком определяется как *тональный факт* (ТФ), который переносится на исходный текст и является конечным результатом.

Для проверки работы алгоритма были взяты следующие предметные области (ПО):

1. Новости спорта (http://www.championat.com).

2. Рецензии на фильмы (http://cinemaholics.ru).
3. Новости России (http://lenta.ru).

Информация о собранных текстах представлена в таблицах 1 и 2. Из частей речи были выбраны только важные для анализа. Видно, что имена существительные и глаголы являются основой новостных сообщений. Это объясняется тем, что последние являются повествовательными текстами. Рецензии имеют более выраженную эмоциональную составляющую, что подтверждает статистика для прилагательных и наречий.

Таблица 1. Статистика по текстам

ПО	Кол-во новостей	Кол-во предложений, всего / в среднем	Кол-во слов, всего / в среднем
Спорт	150	1337 / 9	17441 / 116
Кино	49	1006 / 21	23626 / 482
Новости	91	1080 / 12	17945 / 197

Таблица 2. Статистика по частям речи (процент от всех слов)

ПО	Гл.	Сущ.	Прил.	Наречия	Мест.-прил.	Союз
Спорт	14	39	6	4	3	5
Кино	13	37	10	6	5	8
Новости	13	44	9	2	3	4

Порог для словаря тональных слов был взят равным пяти, для связи ТФ и сущностей – трем. Результаты анализа приведены в таблице 3. Так как для подсчета полноты необходим полный анализ текста, что является очень трудоемкой задачей, то вычислялась только точность. К тому же предложенный метод изначально ограничивался тем, что тональными словами являются только прилагательные и причастия. Основными причинами не очень высокой точности являются:

1. Выделение прилагательных в именах собственных, например, название фильма «Живое».
2. Неверная оценка слова в словаре: словосочетания «Исламское государство», «домашний матч».
3. Отсутствие анализа существительных, связанных синтаксически. В предложении «1 апреля “Новая газета” опубликовала статью о массовых задержаниях в Чечне мужчин-гомосексуалов.» часть «массовых задержаниях в Чечне» выделяется как положительно оцененный ТФ, так как слово массовый является позитивным ТС.

Таблица 3. Точность анализа новостей, проценты

ПО	Тональность ТФ	ТФ (связь ТУ с сущностью)
Спорт	44	89
Кино	38	84
Новости	27	89

Учет всех этих замечаний должен значительно повысить точность. Улучшение

алгоритма выделения ТФ позволит получить адекватные значения для полноты. В целом, результаты для спорта и фильмов получились относительно неплохими.

Заключение и выводы.

В рамках данной работы были рассмотрены существующие подходы решения задачи определения тональности текста. Было исследовано представление предложения в виде дерева зависимостей, которое отражает синтаксические связи между словами. Инструмент MaltParser для автоматического построения таких деревьев был обучен на корпусе со снятой омонимией СинТагРус. Полученная модель была использована в дальнейших этапах работы. Был разработан метод построения графа связей слов с эмоциональной оценкой, учитывающий синтаксис. Обработка построенного графа позволила выделить слова, тональность которых зависит от контекста. Они соединяются с извлеченными именованными сущностями для связи с описываемыми объектами. Был проведен анализ трех предметных областей: спорт, кино и новости. Результаты проанализированы, выделены недостатки, которые будут устранены в дальнейшей работе.

Литература:

1. Karpov I. A., Kozhevnikov M. V., Kazorin V. I., Nemov N. R. Entity Based Sentiment Analysis Using Syntax Patterns and Convolutional Neural Network // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – Vol. 15. – M.: RSUH, 2016. – P. 225-236.
2. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K. et al. Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – Vol. 15. – M.: RSUH, 2016. – P. 50-58.
3. Loukachevitch N. V., Rubtsova Y. V. Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – Vol. 15. – M.: RSUH, 2016. – P. 416-426.
4. Kotelnikov E. V., Bushmeleva N. A., Razova E. V., Peskisheva T. A., Pletneva M. V. Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – Vol. 15. – M.: RSUH, 2016. – P. 300-314.
5. Polyakov P. Yu., Kalinina M. V., Pleshko V. V. Automatic Object-oriented Sentiment Analysis by Means of Semantic Templates and Sentiment Lexicon Dictionaries // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – M.: RSUH, 2015. – Vol. 14(2). – P. 44-52.
6. Толдова С. Ю., Соколова Е. Г., Астафьева И., Гарейшина А. и др. Оценка методов автоматического анализа текста 2011-2012: синтаксические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции

«Диалог». – М.: РГГУ, 2012. – Т. 11(2). – С. 77-90.

7. Nivre J., Hall J., Nilsson J., Chanev A. et al. MaltParser: A language-independent system for data-driven dependency parsing // Natural Language Engineering. – Cambridge University Press, 2007. – Vol. 13(2) – P. 95-135.

8. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л., Иомдин Л. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–-2005 гг. (результаты и перспективы). – М.: Индрик, 2005. – С. 193-214.

9. Natural Language Toolkit – NLTK 3.2.5 documentation [Электронный ресурс]. URL: http://www.nltk.org (дата обращения 22.03.2018).

10. dustalov/greeb: Greeb is a simple Unicode-aware regexp-based tokenizer [Электронный ресурс]. URL: https://github.com/dustalov/greeb (дата обращения 22.03.2018).

11. Ilya Segalovich. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. – Las Vegas, 2003. – P. 273-280.

12. Mikhail Korobov. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. – Vol. 542. – Cham: Springer, 2015. – P. 320-332.

13. Dubatovka A., Kurochkin Yu., Mikhailova E. Automatic Generation of the Domain-Specific Sentiment Russian Dictionaries // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – Vol. 15. – M.: RSUH, 2016. – P. 146-158.

14. Koltsova O. Yu., Alexeeva S. V., Kolcov S. N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – Vol. 15. – M.: RSUH, 2016. – P. 277-287.

15. Lluís Padró, Evgeny Stanilovsky. FreeLing 3.0: Towards Wider Multilinguality // Proceedings of the Eight International Conference on Language Resources and Evaluation. – Istanbul, Turkey: European Language Resources Association, 2012. – P. 2473-2479.

Abstract

The paper presents an algorithm for domain-specific sentiment analysis of information messages in relation to objects. Existing methods of text determining emotional evaluation are considered. The developed method uses a vocabulary of tone words, compiled during the analysis of a specific subject, based on a set of words with context-independent emotional coloring.

Key words: natural language processing, automatic sentiment analysis, sentiment lexicon.

Статья отправлена: 25.03.2018 г.