УДК 004.89

А.А. Прокапович, А.А. Егошина
Донецкий национальный технический университет, г. Донецк
кафедра систем искусственного интеллекта

ОБЗОР СОВРЕМЕННЫХ ПОДХОДОВ И СИСТЕМ АНАЛИЗА ТОНАЛЬНОСТИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Аннотация

Прокапович А.А., Егошина А.А. Обзор современных подходов и систем анализа тональности естественно – языковых текстов*. Проведен анализ современных подходов к определению эмоциональной окраски естественно-языковых текстов, рассмотрены особенности существующих систем анализа тональности лексики русского языка. Показана целесообразность использования компонентов анализа тоносльности в веб-приложениях*

Ключевые слова: *естественно-языковые тексты, анализ тональности, эмоциональная окраска лексики, извлечение отношений.*

Постановка проблемы.

В последние годы происходит бурный рост размеров Интернета, в том числе русскоязычного сегмента. Вместе с увеличением числа пользователей сети Интернет, возрастает и количество генерируемого ими контента. Люди оставляют сообщения на форумах, пишут посты в блогах, комментируют товары на страницах интернет-магазинов и пишут в социальных сетях. Согласно исследованиям Всероссийского центра изучения общественного мнения, количество рускоязыного населения, регулярно (не реже раза в месяц) пользующихся интернетом выросло с 38% в 2010 г. до 55% в 2012 г. Число зарегистрированных в социальных сетях за эти 2 года (с 2010 по 2012 гг.) также значительно возросло – с 53% до 82%.[1]

Весь этот контент несет в себе огромное количество информации, которую можно и даже нужно использовать. Существует отдельное направление искусственного интеллекта и математической лингвистики – обработка естественного языка, или компьютерная лингвистика. Оно позволяет извлекать разнообразную информацию, находящуюся в форме текста на естественном языке. Одно из перспективных направлений компьютерной лингвистики – анализ тональности текста.

Анализ тональности текста позволяет извлекать из текста эмоционально окрашенную лексику и эмоциональное отношение авторов по отношению к объектам, о которых идет речь в тексте. Большинство современных систем используют бинарную оценку – «положительный сентимент» или «отрицательный сентимент», однако некоторые системы позволяют выделять силу тональности.

В современном мире на наш выбор в каких-либо ситуациях зачастую влияет мнение других людей – мы читаем отзывы о товаре, прежде чем заказать его в интернет-магазине, узнаем мнение других людей, прежде чем проголосовать на выборах за того или иного кандидата, долго и тщательно выбираем себе ВУЗ, место работы и ресторан, который мы собираемся посетить. Эта информация представляет значительный интерес для маркетологов, социологов и многих других специалистов.

Кроме того, для владельцев интернет-ресурсов жизненно важно знать мнение пользователей – будь это мнение относительно сделанного на вашем портале нововведения, свежей новости на вашем сайте или оценка пользователями товара в вашем интернет-магазине.[2]

Однако, несмотря на перспективность и актуальность этой задачи, существует сравнительно малое число систем, способных анализировать тональность текста на русском языке. Ниже предложен список и описание самых известных на сегодняшний день систем и компонентов анализа тональности текста.

Цель статьи – провести анализ и классификацию методов определения тональности текста и обозначить перспективы использования.

Методы автоматического извлечения отношений.

При решении задачи извлечения отношений наиболее эффективными методами являются: обучение без учителя и статистические методы. Этим методам не нужны размеченные тренировочные данные, отсутствующие в свободном доступе, по сравнению к примеру, с лингвистическими корпусами, созданными для решения классических задач компьютерной лингвистики: определения частей речи, лемматизации и т.д.

Обучение без учителя – один из способов машинного обучения, при решении которых испытуемая система спонтанно обучается выполнять поставленную задачу, без вмешательства со стороны экспериментатора. Как правило, это пригодно только для задач, в которых известны описания

множества объектов (обучающей выборки), и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами. Обучению без учителя можно сопоставить с методом обучения с учителем, для каждого объекта из выборки уже задан правильный ответ, необходимо найти зависимость между ответами и объектами.

Проанализируем один алгоритм методов обучения без учителя и применим к задаче извлечения аспектов. В основе методов распространения лежит следующая идея: с помощью небольшого множества вручную определенных примеров определенного класса итеративно извлекать подобные единицы текста, постепенно накапливая множество.

В применении к задаче извлечения отношений, алгоритм метода распространения может выглядеть так:

Входные данные : k терминов, имеющих оношение
к заданному аспекту: s1,…sk, число итераций n,текстовый документ
0: задать множество U0 = {s1,…,sk}, i=0
1: найти в документе n-граммы n1,…,nm, близкие терминам из U
2: Ui+1=Ui + {n1,…,nm}, i=i+1
3: перейти к 1, если i<n

В исследовании [5] кандидатами в термины могут буть только n – граммы размером от 1 до 3 слов, содержащие только существительные, прилагательные, глаголы и наречия. Близость n – граммы и терминов из множества U_i определена с помощью RlogF метрики [6]:

$$RlogF(ng) = \log freq(ng\ U_i) \times \frac{freg\,(ng\ U_i)}{freg\,(ng)} \quad (1),$$

где Freg(ng,U) – частота совместной встречаемости n – граммы ng, и терминов из U в рамках фрагментов текста, состоящего из фиксированного числа слов.

Задачу извлечения отношений можно рассматривать, как задачу извлечения терминов, часто употребляемых авторами мнений.

В исследователи предполагают, что терминами, описывающими отношения, могут быть одиночные существительные и словосочетания содержащие существительное, часто встречающиеся во мнениях об объектах одного и того же типа. Из всех n – грамм слов, удовлетворяющих этому требованию, выделяются те, с частотой в корпусе более одного процента.

Выделенные n – граммы, состоящие из двух и более слов, проходят проверку на компактность. Если n – грамма компактна как минимум в двух предложениях, то она попадает в список аспектов.

Компактность определяется следующим образом :

- Пусть f – n- грамма из n слов, s – предложение, содержащее все слова из f (возможно расположенные не подряд)
- Если расстояние между любыми двумя словами, смежными в f, в предложении s составляет не более чем три слова, то f компактна в данном конкретном предложении.

Термины, состоящие из одного слова, также проходят статистический тест на чистоту. Отыскиваются все предложения, содержащие термин. Среди найденных предложений подсчитываются предложения, не содержащие прошедший тест на компактность n – граммы, в которые входит этот термин. Если число таких предложений выше некоторого экспериментально определенного порога, то термин попадает в список отношений.

Похожий статистический метод выявления терминов – аспектов, состоящих из двух и более слов, используется в работе и имеет название C-value [5]. Для всех n – грамм, содержащих в себе только определенные части речи, входящие в некоторые множества документов, вычисляется их зависимость, определенная формулой(2).

$$C - value(t) = \begin{cases} log_2\big(len\,(term)\big) \times freg\,(term), если\ |e\_terms| = 0 \\ log_2\big(len\,(term)\big) \times \Big(freg\,(term) - \frac{1}{|e_{terms}|} \sum_{elder\ = e_{terms}} freg\,(elder)\Big), иначе \end{cases} \quad (2)$$

Где term – n – грамма, e-terms – множество, состоящее из всех n – грамм старшего порядка, содержащий term, |e-terms| мощность множества elder – элемент этого множества. Длина термина в символах – len(term).

Рассмотрим пример, иллюстрирующий работу C-value метода. Пусть в корпусе мнений о сотовых телефонах биграмма “retina display” встречается 8 раз, содержащие ее триграммы “great retina display” и “retina display worse” встречаются 3 и 2 раза соответственно. Тогда согласно формуле (2):

$$C - value(retina\ display) = \log(13) \times (8 - \frac{1}{2}(2+3)) \approx 20$$
$$C - value(great\ retina\ display) = \log(18) \times 3 \approx 13 \quad (3)$$
$$C - value(retina\ display\ worse) = \log(18) \times 2 \approx 8$$

Если экспериментально установленный порог C-value для данного корпуса равен 15, то только n – грамма “retina display” попадет во множество терминов – аспектов.

Для решения задачи определения полярности предложений и коротких сообщений эффективны как алгоритмы обучения с учителем, так и методы, основанные на словарях.

Недостатком метода обучения с учителем является составление тренировочного корпуса с примерами из предметной области, в которой будет использоваться классификатор. Однако схожей проблемой обладают и словарные методы: веса терминов словаря, составленного для одной предметной области, могут оказаться малоэффективными для другой.

Задача извлечения аспектов часто решается с помощью методов обучения без учителя и статистическими методами. Для увеличения эффективности этих методов используются лингвистические и частотные фильтры, позволяющие отсеивать слова, не имеющие отношения к аспектам.

Список и описание популярных систем анализа тональности текста на русском языке.

«SentiStrength» [3] — система, разработанная M. Thelwall, K. Buckley, G. Paltoglou и D. Cai. Начальное назначение было, для анализа коротких неструктурированных неформальных текстов на английском языке. Система может быть сконфигурирована для работы с текстом, также и для других языков, в том числе и для текста на русском языке.

Результат выдается в виде двух оценок – оценка позитивной составляющей текста (по шкале от +1 до +5) и оценка негативной составляющей (по шкале от -1 до -5). Также, возможно предоставления оценок в другом виде:

- Бинарная оценка (позитивный/негативный текст)
- Тернарная оценка (позитивный/негативный/нейтральный)
- Оценка по единой шкале от -4 до +4

Алгоритм основан на поиске максимального значения тональности в тексте для каждой шкалы (т.е. поиск слова с максимальной негативной оценкой и слова с максимальной позитивной оценкой). При работе алгоритма учитывается простейшее взаимодействие слов (например, слова-усилители усиливают значение тональности для слова, на которое они действуют – «очень злой» будет иметь более негативную оценку, нежели просто «злой») и идиоматические выражения.[4]

Недостатки системы: система может быть сконфигурирована для русского языка, реализованные в ней алгоритм не учитывают его специфику, в том числе русскую морфологию, что приводит к ряду проблем. Например, для полноценной работы системы с русским языком необходимо в банке данных иметь все словоформы для каждого слова. Кроме того, система считает лишь общую тональность текста, не выделяя субъекты и объекты тональности.

Компонент анализа тональности текста в составе систем «Аналитический курьер» и «X-files» — разработан компанией «Ай-Теко». Компонент определения тональности текста реализует метод, основанный на словарях и правилах.

Данная система выдает пользователю массив размеченных предложений. В предложениях размечаются объекты тональности (при наличии таковых) и цепочка слов, несущая в себе тональность по отношению к ним. Кроме того, на основании найденных цепочек слов подсчитывается общая тональность для каждого предложения. Для подсчета общей тональности используется ряд специальных правил. Например (для предложения «Доктор Смит вылечил больного гриппом»), есть правило, которое говорит, что сочетание позитивного глагола «вылечить» с негативной цепочкой (в данном случай «больной гриппом») приписывает позитив подлежащему глагола (в нашем примере — «доктору Смиту»). Тональность оценивается по тернарной шкале (позитивный/негативный/нейтральный).

Система работает в несколько этапов:

1. Предварительная обработка текста, выделение и классификация найденных слов
2. Объединение найденных слов в связанные друг с другом цепочки
3. Выделение объектов тональности

Недостатки системы: отсутствие количественной оценки текста.

«Ваал» – система, разработанная Шалак Владимиром. Данная система предназначена для оценки «неосознаваемого эмоционального воздействия фонетической структуры текста и отдельных слов на подсознание человека». Работа системы основана на превращении текста в частотный словарь и отнесении некоторых слов к определенным психолингвистическим категориям.

Результат анализа выдается пользователю в виде набора оценок по ряду критериев, относящихся к данному тексту/слову («гладкий – шероховатый», «могучий – хилый») и т.д.

Недостатки системы: система не производит анализ семантики текста, что ведет к сильной ограниченности применимости продукта. Кроме того, использование данного продукта людьми, не являющимися специалистами в области психолингвистики, не представляется возможным.

Компонент анализа тональности в составе системы RCO Fact Extractor – система, разработанная компанией RCO. Для анализа тональности текста система использует подход, основанный на правилах. Данная система учитывает синтаксическую структуру текста и взаимодействие различных типов слов.

Работа компонента происходит в пять этапов:

1. Распознавание всех упоминаний об объекте во всех формах, включая полные, краткие и другие формы упоминаний.
2. Отсев и полный синтаксический разбор конструкций, в которых отражаются все события и признаки, связанные с целевым объектом.
3. Выделение и классификация тех позиций, в которых явно выражается тональность, и тех пропозиций, которые описывают эмоционально-коннотативные ситуации.
4. Для каждой пропозиции принятие решения о тональности «позитив-негатив» с учетом тех мест, которые занимают в её составе эмоционально-коннотативные, тональные и нейтральные слова, средства выражения отрицания.
5. Оценка общей тональности текста на основе тональностей всех входящих в него пропозиций

Для своей работы компонент использует модули синтаксического анализа текста и отождествления наименований, разработанные также в компании RCO.

Недостатки системы: отсутствие количественной оценки текста.

Заключение

В данной работе были рассмотрены и проанализированы самые популярные системы анализа тональности текста для русского языка. Приведенные системы основаны на различных подходах к решению задачи и предназначены для использования в различных условиях. Стоит отметить, что все приведенные системы являются закрытыми и платными. Каждая система имеет ряд преимуществ и недостатков. Выбирать систему для использования нужно, исходя из задачи. Например, при необходимости анализа общей тональности коротких неструктурированных текстов (сообщений в социальной сети) разумнее всего будет использование системы SentiStrength, а при социологических анализах записей блогов – систему «Аналитический Курьер».

При наличие ряда готовых систем, существует острая нехватка решений для анализа тональности текста на русском языке; данная задача к настоящему моменту полностью не решена. Для того, чтобы получить наиболее универсальный и качественный инструмент, необходимо создать систему, удовлетворяющую следующим условиям:

1. Система должна учитывать специфику русского языка – его морфологию, свободный порядок слов и т.д. – в противном случае, эффективность анализа будет снижаться
2. Система должна учитывать семантику текста
3. Оценка должна производиться по более широкой шкале, чем бинарная – зачастую, в тексте интересен не только сам факт наличия эмоциональной окраски, но и его сила
4. Результат пользователю должен выдаваться в простой и понятной форме, доступной к использованию не специалистами

Система, созданная с учетом этих условий, будет сочетать в себе достоинства приведенных в данной статье систем, при этом устраняя их недостатки. Такая система будет иметь высокую эффективность анализа и широкие области применения.

СПИСОК ЛИТЕРАТУРЫ:

[1] РИФ+КИБ: Тренды Рунета-2012: всегда и везде быть в сети [Электронный ресурс]: Всероссийский центр изучения общественного мнения. – Режим доступа: http://wciom.ru/index.php?id=270&uid=112746 28.11.2012

[2] Bo Pang, Lillian Lee Opinion Mining and Sentiment Analysis // Journal Foundations and Trends in Information Retrieval. 2008. С. 1–135

[3] SentiStrength [Электронный ресурс]: SentiStrength – sentiment strength detection in short texts. – Режим доступа: http://sentistrength.wlv.ac.uk/#About 28.11.2012

[4] Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A. Sentiment strength detection in short informal text // Journal of the American Society for Information Science and Technology. 2010.

[5] Frantzi, K., Ananiadou, S. and Mima, H. Automatic recognition of multi-word terms.// International Journal of Digital Libraries 3(2), pp.117-132.,2000.

[6] Thelen, M., & Riloff, E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts.// Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP), 214-221. Morristown, NJ, USA,2002