УДК 004.912

АВТОМАТИЧЕСКИЙ АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ ПО ОТНОШЕНИЮ К ЗАДАННОМУ ОБЪЕКТУ И ЕГО ХАРАКТЕРИСТИКАМ

Н.В. Лукашевич

Московский государственный университет им. М.В. Ломоносова
louk_nat@mail.ru

Аннотация

Статья посвящена рассмотрению подходов к анализу тональности текстов по отношению к заданному объекту, а также его характеристикам (аспектам). Для решения задачи анализа тональности по отношению к характеристикам сущности необходимо решать также задачи извлечения аспектов для сущности, категоризацию или кластеризацию аспектов по аспектным категориям, определение тональности текста по отношению к заданному аспекту или аспектной категории. Также в статье описывается задание по анализу тональности отзывов пользователей в рамках открытого тестирования систем анализа тональности SentiRuEval.

***Ключевые слова*:** анализ тональности, машинное обучение, тематическое моделирование, оценочная лексика, SentiRuEval

1. ВВЕДЕНИЕ

Задача анализа тональности, т. е. выявление мнения автора текста по поводу предмета, обсуждаемого в тексте, является одной из активно развиваемых технологий в сфере автоматической обработки текстов в последнее десятилетие. Актуальность этого приложения во многом связана с развитием социальных сетей, онлайновых рекомендательных сервисов, содержащих большое количество мнений людей по разным вопросам, в частности, о разных товарах, услугах.

Задачей первых подходов к анализу тональности текстов было определить общую тональность документа или его фрагмента [1]. Такой уровень анализа предполагает, что каждый документ выражает единое мнение по поводу некоторой единичной сущности, как, например, в отзыве о некотором товаре.

Поскольку в документе может быть выражена разная тональность по отно-

шению к разным упомянутым в нем сущностям, то на следующем этапе стали решаться задачи анализа тональности по отношению к заданным сущностям, упомянутым в тексте [2, 3].

Наконец, еще более детальным уровнем анализа тональности текстов является анализ мнения по конкретным свойствам или частям (так называемым аспектам) сущности, по которым автор текста может высказывать разную тональность мнения [4–8].

В [5, 9] *мнение* определяется как пятерка (e_i, a_{ij}, s_{ijkl}, h_k, t_l), где e_i – это сущность, к которой относится мнение, a_{ij} – это *аспект* (часть или характеристика) сущности, s_{ijkl}, – это *тональность* мнения относительно этой сущности и данного аспекта, h_k – это автор мнения, t_l – это время, в которое мнение высказано. При этом мнение s_{ijkl} может быть *положительным*, *отрицательным* или *нейтральным* и может выражаться с разной степенью интенсивности, измеряемой, например, по шкале 1–5.

Аспекты могут быть сгруппированы в категории (далее ***аспектные категории***). Для ресторанов – это обычно кухня, обслуживание, интерьер (обстановка). Также в текстах отзывов можно встретить оценку объекта в целом – *прекрасный ресторан.* Эту категорию также можно рассматривать как аспектную (***аспект Объект_в_целом***). Слова и выражения, посредством которых можно сослаться в тексте на аспект сущности, называются ***аспектными терминами***.

В данной статье будут рассмотрены походы к анализу тональности текстов по аспектам. Во втором разделе описаны подходы к классификации аспектных терминов. В третьем разделе представлены походы к автоматическому извлечению аспектных терминов из текстов. В четвертом разделе обсуждаются подходы к автоматическому определению тональности по отношению к заданным аспектам (аспектным категориям, аспектным терминам). В пятом разделе рассматриваются открытое тестирование систем анализа тональности на русском языке SentiRuEval.

2. КЛАССИФИКАЦИЯ АСПЕКТНЫХ ТЕРМИНОВ

Аспектные термины в предметной области могут быть классифицированы по нескольким основаниям.

Наиболее частым видом аспектных терминов являются ***явные аспектные***

термины, которые явно называют объект, его части или характеристики, которые оцениваются автором текста, например, *суп*, *обслуживание*, *зал* – в отзывах о ресторанах.

Явные аспектные термины чаще всего выражаются существительными или группами существительного, но некоторые аспекты могут выражаться и глаголами, например, *встретить* (*хорошо*, *неприветливо*), *ждать* (*слишком долго*, *не пришлось*) при оценке качества сервиса в ресторанах.

Вторым видом аспектных терминов являются так называемые ***неявные аспектные термины*,** которые представляют собой слова с явно выраженным оценочным компонентом значения, которые одновременно указывают и на обсуждаемый аспект (обычно достаточно обобщенную аспектную категорию), например, *вкусный* (*положительный+еда* в отзывах о ресторанах), *комфортный* (*положительный+комфорт* в отзывах об автомобилях). Как и другие оценочные слова, неявные аспектные термины могут сочетаться с т. н. оценочными операторами, которые меняют или усиливают их оценку: *не очень вкусный*, *не слишком комфортный*. Важность таких аспектных терминов для словарей автоматических систем анализа тональности заключается в том, что в ситуациях нераспознавания упомянутых автором эксплицитных терминов (из-за опечаток, новой лексики, сложной референции) неявные аспектные термины дают возможность извлечь позицию пользователя по отношению к некоторой аспектной категории.

Третьим видом выражения своего мнения по поводу некоторой характеристики заданной сущности является сообщение некоторого произошедшего негативного или позитивного факта, который одновременно указывает как на аспектную категорию, так и на его оценку пользователем (далее – *тональные факты*).

Одним из видов тональных фактов являются технические проблемы, упоминаемые в отзывах [10–12]. В [10] указано, что упоминание технических проблем часто включает в себя:

- набор специального вида глаголов, обозначающих, что что-то случилось (*fail, crash, overload, trip, fix, mess, break, overcharge, disrupt*);
- набор глаголов, обозначающих, что что-то не случилось, и часто эти глаголы упоминаются с отрицаниями, а также с глаголами операторами вида (*stop*, *refuse, cease* – прекратить, прекратиться, остановиться и др.),
- некоторыми глаголами с частицами (*knock off, knock out, hang up*),

- а также существительными и словосочетаниями.

Вместе с тем, тональные факты могут включать и значительно более широкий спектр ситуаций, чем технические проблемы, как, например, обнаружение чего-то нежелательного: «Два раза был в этом ресторане, и оба раза **нашел в своей тарелке волос»**. Liu [5] приводит следующий пример тонального факта: «*I bought the mattress a week ago, and a **valley has formed**»* (*«Я купил матрас неделю назад, и **уже образовалась впадина**»*).

Близкие по смыслу тональные факты могут выражаться в тексте разнообразными способами, что затрудняет их обнаружение. Однако частым признаком такого факта является появление в тексте неоценочных слов, имеющих отрицательные или положительные коннотации. Согласно энциклопедии *Кругосвет, «*Коннотации являются разновидностью так называемой прагматической информации, связанной со словом, поскольку отражают не сами предметы и явления действительного мира, а отношение к ним, определенный взгляд на них» (http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/KONNOTATSIYA.html) . Примерами таких слов с отрицательными коннотациями в общественно-политических текстах являются слова *безработица*, *инфляция, стагнация.* В области отзывов о ресторанах слова *волос*, *майонез* несут в себе отрицательные коннотации, т. е. уже появление таких слов в текстах является признаком того, что тональность текста будет скорее отрицательной. В технической области такими словами являются слова, обозначающие поломки (*fail, crash, overload, trip, fix, mess, break*), как это указывалось в работах [10, 11].

В работах [13, 14] для автоматического выявления слов, имеющих отрицательные или положительные коннотации в общественно-политической области, используется специальный набор контекстов вида «бороться с», «предотвратить», «бороться за» и др.

Другой способ выявления слов, имеющих отрицательные или положительные коннотации, обсуждается в работе [15]. Авторы заметили, что слова, имеющие коннотации, практически не могут употребляться с оценочными словами противоположной направленности. Так, практически невозможно сказать: *хорошая безработица, прекрасная преступность* и т. п. Поэтому предлагается для выявления таких аспектных терминов вычислять разность частот встречаемости

слов с положительными или отрицательными словами. Для улучшения качества извлечения таких аспектных терминов учитывались также отрицания, союзы, расстояние от оценочного слова до слова-аспекта.

Также в [16] указано, что есть еще одна категория неявных оценок и аспектов, которые называются авторами «ресурсная проблема». Приводится пример: *This washer uses a lot of water (Эта посудомоечная машина расходует много воды)*. Таким образом, расходование воды является здесь аспектом, а вода – ресурсным термином, чрезмерное расходование которого является отрицательным фактом.

В [17] отмечено, что ресурсные термины должны извлекаться на основе употребления с квантификаторами *много-мало*, а также рядом с глаголами потребления. В этой работе рассматривается итеративный алгоритм, в котором в начале задаются некоторое количество известных глаголов потребления, а также несколько известных ресурсов: газ, вода, электричество, деньги, чернила, моющее средство (detergent), мыло, шампунь.

3. АВТОМАТИЗАЦИЯ ВЫЯВЛЕНИЯ ПРИЗНАКОВ/СВОЙСТВ ТОВАРОВ/УСЛУГ

В качестве аспектных терминов чаще всего рассматриваются существительные и группы существительного [6, 18, 19]. Длина группы существительного предполагается не больше, чем 3–4 слова. При этом указывается, что если извлекать только отдельные существительные как аспектные термины, то они часто могут быть неоднозначными, что, например, приводит к низкому согласию между экспертами [20].

Согласно [5], существует четыре основных подхода к автоматизации извлечения аспектных терминов из текстов:

- подход, основанный на частотных существительных и группах существительного;
- подход, использующий отношения между оценочными выражениями и аспектными терминами;
- подход, основанный на машинном обучении с учителем;
- подход, основанный на статистических тематических моделях.

3.1. Извлечение аспектных терминов на основе частотных

характеристик

Для извлечения кандидатов в аспекты большое значение имеет ***частотность*** их упоминания в анализируемой текстовой коллекции [4, 19]. В [21] подчеркивается, что частотные признаки работают удивительно хорошо для таких простых признаков. Вместе с тем, все-таки среди частотных существительных встречается достаточно много не-аспектов, например, общелитературной лексики, кроме того, плохо улавливаются малочастотные аспектные термины.

В работе [22] для извлечения аспектных терминов используется известный в информационном поиске признак ***tfidf*** [23], который вычисляется как на уровне документов, так и на уровне абзацев. Scaffidi et al. [24] используют для извлечения аспектных терминов ***сравнение частот*** именных групп в коллекции отзывов с частотами этих групп в контрастной коллекции – Национальном британском корпусе.

Если в качестве аспектных терминов извлекаются не только отдельные существительные, но и группы существительного, то необходимо использовать дополнительные признаки для более точного определения длины именной группы. Чаще всего используются так называемые ***контекстные признаки***, которые оценивают частоту встречаемости словосочетания с частотой контекста. Такие признаки позволяют определить границы именной группы.

Например, в [6] используется так называемая мера FLR:

$$FLR(a) = f(a) \cdot LR(a), \qquad LR(a) = \sqrt{l(a) \cdot r(a)},$$

где $f(a)$ – частота аспектного термина, $l(a)$ – количество разных слов, находящихся слева от a, $r(a)$ – количество разных слов, находящихся слева от a. Далее отбираются группы существительного с данной мерой, большей, чем в среднем для словосочетаний. Таким образом, данная мера в первую очередь отбирает группы существительного, которые имеют большое разнообразие слов на своих границах, что показывает, что анализируемый термин a не является фрагментом более длинного словосочетания.

Другим критерием, направленным к этой же цели, является известный признак C-value [25], который снижает вес данного слова или словосочетания, если оно входит в частотное словосочетание большей длины. Тем самым предполагается, что это более длинное словосочетание может рассматриваться как кандидат

на аспект, а текущее представляет его фрагмент. Такой признак для отбора аспектов используется в работе [26].

В работе [27] предлагается считать аспектными терминами только те именные группы, которые появляются в виде подлежащих или объектов глаголов, или в составе предложных групп.

В работе [4] алгоритм исключает из списка потенциальных аспектных терминов те из них, которые не встречаются достаточно часто в заданных шаблонах, обозначающих *часть–целое* (меронимию) с целевым объектом. Для этого на основе поиска в интернете считается показатель PMI (pointwise mutual information) встречаемости предполагаемого аспектного термина с целевым объектом. Например, для цифровых камер проверяется встречаемость кандидатов в термины в образцах вида «*of camera»*, «*camera has»*. Кроме того, в этой работе используется иерархия WordNet для выявления названий компонентов/частей, а также словообразовательные суффиксы типа (*-iness, -ity*). Отметим, что в какой-то мере использование WordNet, фиксированных суффиксов предполагает применение алгоритма именно к техническим областям. Подобный подход (WordNet, суффиксы) представляется неприменимым к фильмам, программному обеспечению, ресторанам. В работе [28] при обзоре работ указывается, что подход [4] является затратным по времени, поскольку идет интенсивное обращение к интернет-поиску.

Отметим, что этот набор характеристик для извлечения аспектов (за исключением проверки на отношение меронимии в работе [4]) очень похож на характеристики, используемые для извлечения терминов в заданной предметной области [29].

3.2. Отношения аспектов с оценочными словами. Итеративные методы *для извлечения аспектных терминов*

Во многих работах указывается, что аспектный термин должен входить в шаблоны с оценочными словами [18] или хотя бы употребляться в одном и том же предложении с оценочными словами [6, 18]; также могут использоваться меры, учитывающие оба эти фактора [18].

В работе [30] для извлечения отношений между аспектными терминами и оценочными словами используется синтаксический анализатор. Отношения

между аспектом и оценочным словом извлекаются на основе заданных путей синтаксической зависимости. Так, например, в предложении «*This movie is not a masterpiece»* слова *movie* и *masterpiece* будут размечены соответственно аспектом и оценочным словом, поскольку между ними существует путь в синтаксическом дереве «*NN – nsubj – VB – dobj – NN».*

Для извлечения аспектных терминов с учетом их отношений с оценочными словами часто используются итеративные методы (bootstrapping). В качестве начального множества могут использоваться частотные именные группы, которые предполагаются аспектами либо задаются вручную.

В известной работе [19] начальное множество аспектных терминов (частотные слова и именные группы) используется для выявления ассоциативных правил, т. е. шаблонов, посредством которых аспекты обычно связаны с оценочными словами. После получения таких правил извлекаются менее частотные аспектные термины, т. е. те именные группы, которые появлялись именно в таких шаблонах с оценочными словами.

В работе [28] рассматривается подход двойного распространения (double propagation) к извлечению аспектных терминов и расширению словаря оценочных слов. В качестве исходного множества задается небольшой словарь оценочных слов, также задаются синтаксические шаблоны, в которые обычно входят оценочные слова и аспектные термины. В итоге вхождение известного оценочного слова в такой шаблон помогает извлекать аспект, а известный аспект, входящий в такой шаблон, помогает извлекать оценочное слово.

Для очистки полученного множества аспектов применяется ряд правил. Например, предполагается, что в одном фрагменте предложения без запятых содержится только один аспектный термин, а другой кандидат должен быть удален, удаляется менее частотный в коллекции.

Оценка этого метода проводилась на пяти областях; была получена средняя F-мера – 0.85. Отметим, что эксперименты проводились на небольшом числе отзывов – в среднем 62.8 отзыва из каждой области [29].

В [21] указано, что итеративные методы, основанные на отношениях с оценочными словами, могут находить низкочастотные аспекты. Вместе с тем, извлекается достаточно много не-аспектов, которые подошли под заданные шаблоны.

При создании комбинированных методов, сочетающих шаблоны и частотность, начинают теряться низкочастотные аспекты и возрастает число параметров для настройки.

В [8] указано, что метод «double propagation» для одновременного извлечения аспектов и оценочных слов, основанный на синтаксическом пути между ними, хорош для коллекции среднего размера: для маленьких коллекций метод дает пониженную полноту, в то время как для больших коллекций – в заданные синтаксические шаблоны проникает много шума.

В работе [31] для оценки значимости аспектных терминов вводятся еще два фактора. Первый фактор рассматривает, насколько разнообразны оценочные слова, применяемые к аспекту-кандидату, – разнообразие обычно свидетельствует о значимости аспектного термина. Во-вторых, в коллекции ищется подтверждение связи аспектного термина с сущностью посредством заданных шаблонов. Например, в области автомобилей можно найти такие фразы, как *the engine of the car (двигатель автомобиля), the car has a big engine (автомобиль имеет большой двигатель),* которые свидетельствуют об отношении *часть–цел*ое между *engine* и *car.* Если слово одновременно встречается и с оценочным словом, и в отношениях с заданной сущностью, то это дает этому аспекту-кандидату сразу высокий вес: например*, there is a bad hole in the mattress (в матрасе имелась большая дыра)*.

В работе [6] для итеративного поиска аспектных терминов используется некоторое начальное множество аспектов, которое пополняется на основе:

- учета меры взаимной информации нахождения аспекта кандидата в одних и тех же предложениях, что и аспекты из начального множества аспектов и частотности аспекта-кандидата,
- при пополнении аспектов полезна очистка избыточных аспектов – например, если в множество аспектов входит и более короткий аспект.

Число вручную выделяемых аспектных терминов товара в данной работе может достигать до 200 аспектов в технических областях. F-мера выделяемых аспектов в данной работе – порядка 72.9%. Обучение проводилось на 45–100 текстов для отдельного объекта [6].

3.3. Использование методов машинного обучения для выявления аспектных терминов

Имеется два направления использования методов машинного обучения с учителем для выявления аспектных терминов:

- методы, основанные на предварительном составлении списка аспектных терминов в некоторой предметной области, и обучение модели, использующей перечисленные в предыдущих разделах признаки, присущие аспектам;
- методы, основанные на разметке последовательности слов в отзывах (разметка аспектных терминов, оценочных слов)

В работе [32] для извлечения аспектных терминов помимо частотности аспектов-кандидатов в отзывах используется сопоставление кандидатов с заголовками словарных статей в Википедии, семантическая близость кандидатов, рассчитанная на основе совокупностей ссылок соответствующих статей Википедии (в итоге 2 признака), а также ассоциирование кандидата в аспекты с именем сущности при поиске в интернете. Результат извлечения аспектов для нескольких объектов оценивается как 72.7% F-меры.

В работе [20] в качестве набора признаков для извлечения аспектных терминов в виде отдельных существительных из отзывов о лэптопах на русском языке рассматривается следующий набор признаков:

- частотность в коллекции отзывов,
- близость к оценочным словам (окно величиной p), в данном случае рассматривалась не близость к оценочным словам в коллекции, а близость на расстоянии 3 к словам *хороший/плохой* в выдаче результатов поиска Яндекса,
- признак странности, вычисляющий относительную частоту слова по сравнению с контрастной коллекцией,
- признак tfidf,
- мера взаимной информации pmi, которая учитывается совместную встречаемость между существительными кандидатами и заявленным типом товара (лэптоп).

На основе различных вариантов каждой из мер авторами работы было получено 23 признака. Указывается, что результат извлечения близок к результатам

англоязычных работ, которые заявляют о F1-мере в интервале 0.76 – 0.86 для разных областей.

Однако наиболее популярными в области извлечения аспектных терминов на основе методов машинного обучения являются подходы, основанные на последовательной разметке, при которой аспекты и не-аспекты аннотируются в корпусе. К размеченным данным применяются методы вида HMM (Hidden Markov models) и CRF (Conditional Random Fields) [33, 34]. В качестве признаков используются такие характеристики, как собственно слова, части речи, синтаксические зависимости, расстояния, предложения с оценочными словами и др. Эти же модели могут применяться и для совместного извлечения аспектов и оценочной лексики.

В [21] указано, что методы, основанные на машинном обучении, могут выявлять и низкочастотные аспекты, но требуют разметки данных. Особенно большие трудозатраты требуются для разметки данных для последовательных методов машинного обучения.

3.4. Использование тематических моделей для извлечения аспектных терминов

Извлечение аспектов может выполняться на основе применения так называемых статистических тематических моделей, т. е. методами, которые предполагают, что каждый текст состоит из набора скрытых тем, а каждая скрытая тема представляет собой вероятностное распределение слов. Обычно рассматриваются два типа тематических моделей: pLSA (probalistic Latent Semantic Analysis) и LDA (Latent Dirichlet Allocation) [35, 36]. В результате применения тематических моделей к коллекции текстов порождается совокупность тем, каждая из которых представляет собой список слов с вероятностями их отнесения к этой теме.

Для извлечения аспектов необходима модификация базовых тематических моделей, направленная на то, чтобы отделить оценочные слова и топики в отдельные темы. При успешном применении таких моделей происходит два одновременных действия: извлечение аспектов и их группирование в обобщенные категории аспектов.

Одна из известных модификаций базовой модели LDA для извлечения аспектных терминов описана в работе [37], в которой показано, что применение базовой модели LDA, которая строится на информации о взаимной встречаемости

слов в одних и тех же текстах, не является эффективной для извлечения аспектов, поскольку во множестве разных отзывов может содержаться один и тот же набор аспектов. Авторы работы применяют глобальную модель для извлечения именований сущностей, а для извлечения аспектных терминов используют скользящее окно из слов или предложений (например, 3 предложения). Собственно, встречаемость слов в таких фрагментах используется для выявления аспектов, при этом они не различают аспектные термины и оценочные слова. В статье приводится следующий пример темы «Обслуживание»: *staff, friendly, helpful, service, desk, concierge, excellent, extremely, hotel, great, reception, English, pleasant, help.*

В работе [38] предложена гибридная модель MaxEnt-LDA (комбинация моделей Maximum Entropy и LDA), в которой производится совместное извлечение аспектных и оценочных слов на основе синтаксических признаков, помогающих разделить аспектные и оценочные слова. Метод Maximum Entropy используется для подбора параметров на размеченных данных.

В [16] указываются следующие проблемы применения тематических моделей для извлечения и группирования аспектных терминов:

- требуются большие объемы данных и тщательная настройка параметров моделей для получения достаточно качественных результатов,
- методы основаны на семплировании Гиббса и поэтому каждый раз дают несколько иной результат,
- тематические модели легко выявляют частотные аспекты, которые выявляются и многими другими методами.

3.5. Группирование аспектов

Выделенные аспектные термины могут быть достаточно разнообразными, и для удобства пользователя они обычно группируются в обобщенные категории. Такими категориями для ресторана могут быть: Кухня, Интерьер, Обслуживание, Местоположение. При этом аспектная категория «Кухня» объединяет множество блюд и продуктов питания, которые могут предлагаться в том или ином ресторане.

В [16] указано, что автоматизация группировки аспектов является критической для многих приложений анализа тональности отзывов.

Использование общезначимых словарей синонимов и тезаурусов имеет в

этой задаче ограниченное применение, поскольку такие группировки аспектных терминов существенно зависят от предметной области. Кроме того, часто аспектные термины выражаются словосочетаниями, которые обычно не описываются в словарях.

В работах [39, 40] предложен алгоритм частичного обучения, который разбивает аспектные термины на предопределенные категории аспектов. При этом предполагается, что сами по себе аспектные термины уже выделены каким-то методом. Сначала авторы вручную относят небольшое количество аспектных терминов к категориям. Затем применяют Expectation Maximization (EM) алгоритм для работы с размеченными и неразмеченными примерами. Кластеризация проводится на базе сходства контекстов упоминания аспектных в окне 15 слов налево и направо. Если в окне встречается другой аспектный термин, то он не включается в окно. Также исключаются стоп-слова.

В методе также применяются два вида дополнительной информации для лучшей инициализации EM-алгоритма: аспектные термины в виде именных групп, имеющие общие слова, обычно относятся к одной категории аспектов (*battery life* и *battery power*), и аспектные термины, являющиеся синонимами в словаре, также чаще всего будут принадлежать одной группе. Эти две эвристики позволяют EM-алгоритму достигать лучших результатов.

Данный алгоритм и различные другие варианты кластеризации аспектных терминов тестируются на нескольких предметных областях. Лучший результат, полученный на основе EM алгоритма в этой работе, достигает качества кластеризации, измеряемого мерой Purity, – 0.55. Purity – мера в кластеризации, замеряющая долю максимального эталонного кластера в автоматических кластере, которая затем усредняется по всем автоматическим кластерам. Таким образом, на текущий момент лучший метод кластеризации в состоянии лишь приблизительно наполовину повторить эталонную кластеризацию.

В работе [41] ставится задача выстроить иерархическую классификацию аспектных терминов, подобно экспертной классификации. Иерархия аспектов строится на основе нескольких признаков сходства:

- контекстный признак: два слова влево и вправо,
- признак совместной встречаемости аспектных терминов, вычисляемый на основе меры взаимной информации PMI,

- длина синтаксического пути между аспектными терминами в предложении, а также синтаксические роли в предложениях (подлежащее, объект, модификатор и т. п.),
- лексические признаки, включая извлеченное из интернета определение аспектного термина.

Иерархия строится итеративно, на основе минимизации нескольких критериев (minimum Hierarchy Evolution, minimum Hierarchy Discrepancy, minimum Semantic Inconsistency), веса признаков подбираются на основе 50 иерархий WordNet и ODP (Open Directory Project).

Результаты показывают, что если начальная иерархия совсем не задана, то качество получаемой иерархии в среднем 0.3–0.4 F-меры. Если задано 20% иерархии, то качество составляет 0.4–0.5 F-меры. Среди признаков максимальный вклад у меры совместной встречаемости.

Ранее обсуждалось, что статистические тематические модели могут одновременно извлекать и группировать аспекты. Для учета в этих моделях знаний о предметной области в работе [42] предложено использовать дополнительные ограничения, извлекаемые из онтологии предметной области, которые могут улучшить качество создаваемых кластеров. Ограничения носят форму must-links и cannot-links. Must-links определяют, что два слова должны быть в одном кластере, cannot-links задают, что два слова не могут быть в одном кластере. Однако предложенный метод приводит к экспоненциальному росту в кодировании cannot-links и имеет сложности в обработке большого количества ограничений.

В работе [43] знание о предметной области сообщается в виде тематической модели в виде исходных (seed) слов для каждой категории аспектов. Кроме того, модель разделяет аспекты и оценочные слова. Приводятся следующие примеры исходных слов:

- *Staff (staff, service, waiter, hospitality, upkeep);*
- *Cleanliness (curtains, restroom, floor, beds, cleanliness);*
- *Comfort (comfort, mattress, furniture, couch pillows)*.

Оценка подхода показывает, что 2 заданных слов в аспекте приводит в среднем к качеству извлечения аспектных слов, измеряемых мерой точности на заданном уровне 30 слов: P@30=70%, 5 заданных слов – P@30=77%.

4. ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ПО ОТНОШЕНИЮ К АСПЕКТНЫМ ТЕРМИНАМ

Как и в общей задаче анализа тональности по документам и предложениям, в задаче определения тональности по отношению к аспектам возможно использование двух основных методов: методов машинного обучения и инженерно-лингвистических методов.

Ключевой вопрос при проставлении оценок тональности аспектов заключается в том, как определить диапазон действия каждого оценочного выражения, относится ли оценочное выражение к аспекту, упомянутому в этом предложении [5]. Одно из основных направлений решения этой проблемы базируется на использовании синтаксической структуры предложений в форме деревьев зависимости [3, 5, 7].

4.1. Методы машинного обучения для определения тональности по отношению к аспектам

В работе [7] на основе заранее собранных и вычитанных оценочных слов и аспектов задача проставления оценок аспектам рассматривается как задача классификации, т. е. для заданного предложения классификатор должен проставить, к какому именно аспектному термину относится данное оценочное слово, что может быть существенным для длинного предложения, в котором упомянуто несколько оценок и несколько аспектов (*хорошая пицца, но лазанья была ужасная*).

В качестве признаков рассматриваются следующие:

- признаки расположения: расстояние между аспектным термином и оценочным словом, число аспектов и оценочных слов в предложении, длина предложения, пунктуация, наличие одних аспектов между другими аспектами и оценочными словами, порядок расположения аспекта и оценочного слова,
- лексические признаки: набор слов между аспектным термином и оценочным словом, наличие союзов и др.,
- части речи оценочного слова и аспектного термина, набор тегов частей речи между аспектом и оценочным словом, части речи соседних слов,
- признаки, основанные на синтаксической структуре: набор тегов по пути между аспектом и оценочным словом, близость по синтаксическому дереву.

В экспериментах было показано, что все четыре типа признаков существенны для выделения пары аспектный термин – оценочное слово, достигнутая F-мера составила 82.2%. Базовый уровень для сравнения, состоявший в том, что оценочное слово приписывается к ближайшему аспекту, составил 76.6% F-меры. Авторы подчеркивают, что они ожидали, что прирост будет больше.

4.2. Лингвистико-инженерные методы проставления оценок аспектов

В лингвистико-инженерных методах предполагается, что на момент классификации известны:

- названия сущностей, их аспектов;
- имеется словарь оценочных слов и выражений, а также правила их преобразования в зависимости от контекста и правила суммирования. Обработка идет обычно по предложениям и включает в себя несколько этапов [16].

Сначала производится проставление в предложении известных аспектных терминов и оценочных слов; оценочные слова имеют проставленную в словаре оценку тональности – в простейшем случае {1, –1}. К оценочным словам применяются операторы, которые могут менять тональность оценочного слова на противоположную.

Далее необходимо учесть структуру предложения для возможной модификации базовых оценок. В частности, в работе [45] указывается на важность обработки союзов типа *но*, *однако*. Если во второй части предложения не обнаружено оценочных слов, но присутствуют союзы *но* или *однако*, то второй части предложения должна быть приписана оценка, противоположная оценке первой части предложения.

В результате должно быть проведено агрегирование оценок по каждой аспектной категории. В работе [45] предложена следующая процедура проставления оценок аспектов в отдельном предложении. Пусть в предложении s содержится набор аспектных терминов $\{a_1, \ldots, a_n\}$ и оценочных выражений $\{sw_1, \ldots, sw_n\}$, для которых оценки из словаря уже модифицированы с учетом операторов и контекста. Тогда оценки тональности каждого аспектного термина вычисляются по следующей формуле:

$$score\,(a_i, s) = \sum_{swj \in s} \frac{sw_j so}{\mathrm{dist}(sw_j, a_i)},$$

где sw_j – оценочное слово или выражение, $sw_j so$ –числовая оценка тональности sw_j, $\mathrm{dist}(sw_j, a_i)$ – расстояние между оценочным словом и аспектом. Таким образом, к каждому аспектному термину в предложении приписываются все оценки, упомянутые в этом предложении, однако их вес падает в зависимости от расстояния между аспектом и оценкой. Если окончательный вес – положительный, то и оценка аспекта положительная, отрицательный вес означает отрицательную оценку, вес 0 – нейтральную оценку.

Результаты, представленные в [45], использующие вышеуказанную формулу, учет операторов, обработку союза *но* и учет контекстно-зависимых оценочных слов достигает F-меры 91% на 5 предметных областях. Система Opine на этих же данных получает 87% [4], алгоритм [20] – 83%.

В работе [45] используется шесть правил композиции оценок для определения тональности по отношению к объекту: *конверсия тональности, агрегация, распространение, доминирование, нейтрализация* и *интенсификация*.

Конверсия – это применение отрицаний и перевод в противоположную тональность. *Агрегация* применяется для синтаксических групп вида *прилагательное-существительное, существительное-существительное, наречие-прилагательное, наречие-глагол,* имеющих противоположную тональность, например, *beautiful fight (прекрасная битва)*. В таком случае этой фразе приписывается доминирующая тональность модификатора: POS(‘*beautiful*’) & NEG(‘*fight*’) =>POSneg(‘*beautiful fight*’).

Правило распространения применяется, когда в предложении употребляется глагол распространения или передачи: PROP-POS(«*to admire»*) & «*his behavior»* => POS(«*his behavior»*); «*Mr. X»* & TRANS(«*supports*») & NEG(«*crime business»*) => NEG(«*Mr. X»*).

Правило доминирования заключается в том, что если полярности глагола и его объекта различны, то полярность глагола преобладает (e.g., NEG(«*to deceive»*) & POS(«*hopes»*) => NEG(«*to deceive hopes»*)); если в сложном предложении фразы соединены союзом «но», то тональность второй части предложения доминирует: например: ‘NEG(«*It was hard to climb a mountain all night long»*), *but* POS(«*a*

magnificent view rewarded the traveler at the morning»).' => POS(предложение))

Правило нейтрализации применяется, когда предлог-модификатор или оператор условия относится к тональному выражению, например, «*despite»* & NEG('*worries*') => NEUT(«*despite worries»*). *Правило интенсификации* усиливает или ослабляет вес тональности, например, Pos_score(«*happy»*) < Pos_score(«*extremely happy»*)).

5. ТЕСТИРОВАНИЕ ОБЪЕКТНО-ОРИЕНТИРОВАННЫХ СИСТЕМ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ

Задача автоматического анализа тональности текстов является сложной комплексной проблемой. Поэтому организуются различные открытые тестирования подходов к анализу тональности текстов. В состав таких тестирований входят такие, как Blog Track, проводимый в рамках конференции TREC, в котором нужно по запросу найти мнение пользователя о сущности, упомянутой в запросе [46]; задания конференции TAC под названием Opinion QA Tasks [47], включающие нахождение ответов на вопросы, содержащие мнения; задания анализа мнений на конференции NTCIR, посвященной обработке текстов на восточных языках [48], анализ сообщений из Твиттера с целью мониторинга репутации заданного объекта [49] и др.

С 2014 года в рамках конференции SemEval организуется тестирование систем анализа тональности по отношению к аспектам сущности [49]. Данные для обучения и тестирования включали изолированные предложения, извлеченные из отзывов в двух предметных областях: ресторанах и лаптопах. Для обучения в каждой из областей было подготовлено около 3 тысяч предложений. Множество аспектных категорий по ресторанам включало: *food (еда), service (обслуживание), price (цена), ambience (обстановка, атмосфера), anecdotes/miscellaneous (другое)***.**

В 2015 году тестирование обработки отзывов в рамках SemEval (http://alt.qcri.org/semeval2015/task12/) включает уже полные отзывы. Аспектные категории усложняются и теперь уже состоят из пар сущность-характеристика (Entity-Atribute pairs – E#A). Набор пар E#A включает в области ресторанов шесть типов сущностей (RESTAURANT, FOOD, DRINKS, SERVICE, AMBIENCE, LOCATION) и 5

типов атрибутов (GENERAL, PRICES, QUALITY, STYLE_OPTIONS, MISCELLANEOUS). Область лаптопов содержит 22 типа сущностей and 9 типов атрибутов (GENERAL, PRICE, QUALITY, OPERATION_PERFORMANCE и др.). Примеры аннотирования предложений в области отзывов о ресторанах выглядят следующим образом:

1) Great for a romantic evening, but over-priced. → {AMBIENCE#GENERAL}, {RESTAURANT#PRICES};

2) The fajitas were delicious, but expensive. → {FOOD#QUALITY}, {FOOD# PRICES}.

Тестирование анализа тональности по аспектам в рамках SentiRuEval

Мероприятие по оценке систем анализа тональности для текстов на русском языке SentiRuEval, которое было организовано в 2014–2015 гг., является вторым после сравнительных исследований систем анализа тональности в рамках сеинара по информационному поиску РОМИП, организованного в 2011–2013 годах. Тестирование в рамках РОМИП было направлено на выявление общей тональности текста (отзыва, поста в блоге, новостной цитаты) [50]. Новое тестирование SentiRuEval направлено на исследование методов анализа текстов по отношению к некоторому заданному объекту или его характеристикам [51].

В 2014–2015 годах в рамках SentiRuEval имеется два типа задания: объектно-ориентированный анализ твитов для двух типов организаций (банки и телекоммуникационные компании) и аспектно-ориентированный анализ отзывов пользователей в двух предметных областях (рестораны и автомобили). Далее будет рассмотрена задача аспектно-ориентированного анализа отзывов в рамках SentiRuEval.

Каждый отзыв содержит мнения пользователя о конкретном объекте. Такие мнения структурируются по заранее заданному набору *целевых аспектов*, т. е. составных частей, либо характеристик оцениваемого объекта. Для ресторанной тематики такими аспектами являются: *кухня*, *интерьер, сервис, цена*. Для автомобилей список аспектов включает в себя: *безопасность*, *комфорт*, *надежность*, *внешний вид*, *цены, ходовые качества*. Набор целевых аспектов дополнен аспектом «*объект в целом*», представляющим общее мнение об объекте.

Для создания обучающей коллекции была осуществлена разметка отзывов, при которой в тексты вносилась следующая информация:

- выделяются аспектные термины, включая эксплицитные, имплицитные и тональные факты;

- выделенным аспектным терминам приписывается их тональность: позитивный, негативный, противоречивый (both) и нейтральный;

- выделенные аспектные термины относятся к аспектной категории;

- отмечается статус выделенного аспектного термина относительно текущего мнения: релевантный (REL), относится к прошлому мнению автора или других людей (PREV), относится к другому объекту (CMPR), относится к гипотетической ситуации (IRR), ирония (IRN); такая разметка помогает выявить аспектные термины, учет которых может ухудшить качество анализа, поскольку они не относятся к текущему мнению автора;

- приписывается оценка аспектной категории в целом по отзыву: нейтральный, положительный, отрицательный, противоречивый, оценка отсутствует.

Участники могли решать следующие задачи на выбор:

Задача А. Выделение ***релевантных отзыву*** эксплицитных аспектных терминов. При этом не должны размечаться как эксплицитные аспектные термины упоминания, относящихся к другим объектам или ситуациям, упоминаемым в отзывах;

Задача Б. Выделение ***релевантных отзыву*** всех аспектных терминов, включая неявные аспектные термины и тональные факты;

Задача В. Присваивание оценки тональности ***эксплицитным*** аспектным терминам;

Задача Г. Присвоение аспектной категории ***эксплицитным*** аспектным терминам;

Задача Д. Заполнение оценок аспектных категорий по отзывам в целом.

Для каждой задачи организаторами были подготовлены прогоны, представляющие базовые уровни (baseline) для сравнения, т. е. представляющие собой очень простые решения поставленных задач.

Базовая система для задач А и Б извлекает список размеченных терминов из обучающей коллекции, лемматизирует их и размечает их в тестовой коллекции на основе ее лемматизированного представления. Если к некоторой последовательности слов применимо более одного термина, то предпочитается более

длинный термин.

Базовая система задачи В приписывает аспектному термину его наиболее частотную аспектную категорию, на основе информации из обучающей коллекции. Если термин отсутствует в обучающей коллекции, то приписывается наиболее частотный аспектная категория. Базовая система задачи Г приписывает аспектным терминам тональности на основе таких же принципов. Базовый уровень для задачи Е представляет собой наиболее частую категорию тональности для каждой аспектной категории (во всех случаях это была положительная тональность).

В тестировании приняли участие 11 участников, причем задача анализа отзывов о ресторанах привлекла значительно больше внимания, чем отзывы об автомобилях. Как указано в [51], лучшие результаты, полученные участниками для задач А и Б по извлечению аспектных терминов, пока ненамного превзошли базовый метод извлечения аспектных терминов, переносящий разметку из обучающего множества в тестовое. Например, при точном сопоставлении эксплицитных аспектов по ресторанам лучший результат составил 0.632 F-меры, а baseline результат – 0.608. Многие участники не смогли превзойти результат baseline системы.

Задачи В и Г являются задачами классификации аспектных терминов, и лучшие результаты были получены на основе методов машинного обучения SVM и Gradient Boosting.

Среди особенностей применяемых подходов для решения разных типов задач можно назвать использование новых, недавно появившихся типов учитываемых факторов, заключающихся в использовании нейронных сетей для представления контекстов слов коллекции в виде более плотных векторов, т. н. word embedding [52], такие факторы использовались в работах [53–55].

Обучающие и тестовые данные, результаты участников, а также скрипты для подсчета результатов доступны по адресу: http://goo.gl/Wqsqit.

ЗАКЛЮЧЕНИЕ

В течение последних 10–15 лет задача автоматического анализа тональности текстов вызывает неизменно высокий интерес у исследователей и имеет разнообразные сферы применения на практике. В данной статье рассмотрены подходы к задачам, связанным с анализом тональности по отношению к заданному объекту, а также к его характеристикам. Также мы описали открытые тестирования, проводившиеся в этой сфере для систем анализ тональности текстов на английском и русском языках. Обучающая и тестовая коллекции, результаты участников и скрипты для подсчета метрик опубликованы для некоммерческого использования

Благодарности

Работа частично поддержана грантом РФФИ, проект № 14-07-00682.

СПИСОК ЛИТЕРАТУРЫ

1. *Pang B., Lee L., Vaithyanathan S.* Thumbs up?: sentiment classification using machine learning techniques // Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002. V. 10. P. 79-86.

2. *Amigo E., Corujo A., Gonzalo J., Meij E., Rijke M.* Overview of RepLab 2012: Evaluating Online Reputation Management Systems // CLEF-2012. 2012.

3. *Jiang Long, Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao*. Target dependent twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011). 2011. P. 151-160.

4. *Popescu A., Etzioni O.* Extracting product features and opinions from reviews // Natural language processing and text mining. Springer: London. 2007. P. 9-28.

5. *Liu B., Zhang L.* A survey of opinion mining and sentiment analysis // Mining Text Data. Springer: US, 2012. P. 415-463.

6. *Bagheri A., Saraee M., de Jong F.* An unsupervised aspect detection model for sentiment analysis of reviews // Natural Language Processing and Information Systems. Springer: Berlin Heidelberg, 2013. P. 140-151.

7. *Glavaš G., Korencic D., Šnajder J.* Aspect-oriented opinion mining from user reviews in Croatian // Proceedings of BSNLP workshop, ACL-2013. 2013. P. 18-23.

8. *Zhang L., Liu B.* Aspect and entity extraction for opinion mining // Data Mining

and Knowledge Discovery for Big Data. Springer: Berlin Heidelberg, 2014. P. 1-40.

9. *Liu B.* Sentiment analysis and Subjectivity // Handbook of Natural Language Processing. CRC Press, Taylor and Francis Group, Boca Raton, 2010. P. 1-38.

10. *Gupta N.K.* Extracting phrases describing problems with products and services from twitter messages // Computación y Sistemas. 2013. V. 17, No 2. P. 197-206.

11. *Ivanov V., Tutubalina E.* Clause-based approach to extracting problem phrases from user reviews of products // Analysis of Images, Social Networks and Texts. Springer International Publishing, 2014. P. 229-236.

12. *Tutubalina E., Ivanov V.* Unsupervised approach to extracting problem phrases from user reviews of products // Proceedings of the Aha! Workshop on Information Discovery in Texts, Coling-2014. 2014. P. 48-53.

13. *Feng S., Bose R., Choi Y.* Learning general connotation of words using graph-based algorithms // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics. 2011. P. 1092-1103.

14. *Feng S., Kang J.S., Kuznetsova P., Choi Y.* Connotation Lexicon: a dash of sentiment beneath the surface meaning // Proceedings of ACL. 2013. P. 1774-1784.

15. *Zhang Lei, Bing Liu.* Identifying noun product features that imply opinions // Proceedings of the Annual Meeting of the Association for Computational Linguistics (short paper) (ACL-2011). 2011. P. 575-580.

16. *Liu B., Zhang L.* A survey of opinion mining and sentiment analysis // Mining Text Data. 2012: Springer US. P. 415-463.

17. *Zhang Lei, Liu B.* Extracting resource terms for sentiment analysis // Proceedings of IJCNLP-2011. 2011. P.1171-1179.

18. *Blair-Goldensohn S., Hannan K., McDonald R., Neylon T., Reis G. A., Reynar J.* Building a sentiment summarizer for local service reviews // Proceedings of WWW Workshop on NLP in the Information Explosion Era. 2008.

19. *Hu M., Liu B.* Mining and summarizing customer reviews // Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004. P. 168-177.

20. *Марчук А.А., Уланов А.В., Макеев И.В., Чугреев А.А.* Автоматическое из-

влечение параметров продуктов из текстов отзывов при помощи интернет-статистик // Труды Международной конференции «Компьютерная лингвистика и информационные технологии, Диалог-2013». 2013. Т. 2. С. 81-91.

21. *Moghaddam S., Ester M.* Aspect-based opinion mining from online reviews. Tutorial at SIGIR-2012. 2012.

22. *Ku Lun-Wei, Yu-Ting Liang, Hsin-Hsi Chen.* Opinion extraction, summarization and tracking in news and blog corpora // Proceedings of AAAI-CAAW'06. 2006.

23. *Manning C.D., Raghavan P., Schütze H.* Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.

24. *Scaffidi Ch., Bierhoff K., Chang E., Felker M., Ng H., Jin Ch.* Red Opal: product-feature scoring from reviews // Proceedings of Twelfth ACM Conference on Electronic Commerce (EC-2007). 2007. P. 182-191.

25. *Frantzi K., Ananiadou S., Mima H.* Automatic recognition of multi-word terms: the C-value/NC-value method // International Journal on Digital Libraries, 2000. V. 3, No 2. P. 115-130.

26. *Zhu J., Wang H., Tsou B., Zhu M.* Multiaspect opinion polling from textual reviews // Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2009). 2009. P. 1799-1802.

27. *Hai Z., Chang K., Cong G.* One seed to find them all: mining opinion features via association // Proceedings of the 21st ACM international conference on Information and knowledge management. 2012. ACM. P. 255-264.

28. *Qiu G., Liu B., Bu J., Chen C.* Opinion word expansion and target extraction through double propagation // Computational Linguistics. 2011. V. 1, No 1. P. 1-18.

29. *Loukachevitch N., Nokel M.* An experimental study of term extraction for real information-retrieval thesauri // Proceedings of Terminology and Artificial Intelligence Conference TIA-2013. 2013. P. 69-78.

30. *Zhuang L., Jing F., Zhu X.* Movie review mining and summarization // Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006), 2006. P. 43-50.

31. *Zhang L., Liu B., Lim S., O'Brien-Strain E.* Extracting and ranking product features in opinion documents // Proceedings of International Conference on Computational Linguistics (COLING-2010). 2010. P. 1462-1470.

32. *Kovelamudi S., Ramalingam S., Sood A., Varma V*. Domain independent model for product attribute extraction from user reviews using Wikipedia // Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2010). 2011. P. 1408-1412.

33. *Niklas J., Gurevych I.* Extracting opinion targets in a single and cross-domain setting with conditional random fields // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010). 2010. P. 1035-1045.

34. *Choi Y., Cardie C.* Hierarchical sequential learning for extracting opinions and their attributes // Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010). 2010. P. 269-274.

35. *Blei D., Ng A., Jordan M.* Latent Dirichlet allocation // The Journal of Machine Learning Research, 2003. No 3. P. 993-1022.

36. *Воронцов К.В., Потапенко А.А.* Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 657-686.

37. *Titov I., McDonald R*. A joint model of text and aspect ratings for sentiment summarization // Urbana, 51, 61801. 2008.

38. *Zhao Wayne Xin, Jing Jiang, Hongfei Yan, Xiaoming Li.* Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010). 2010. P. 56-65.

39. *Zhai Z., Liu B., Xu H., Jia P.* Grouping product features using semi-supervised learning with soft-constraints // Proceedings of Coling-2010. 2010. P. 1272-1280.

40*. Zhai Z., Liu B., Xu H., Jia P.* Clustering product features for opinion mining // Proceedings of the fourth ACM international conference on Web search and data mining. ACM. 2011. P. 347-354.

41. *Yu J., Zha Z.J., Wang M., Wang K., Chua T.S.* Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011. P. 140-150.

42. *Andrzejewski D., Zhu X., Craven M.* Incorporating domain knowledge into topic modeling via Dirichlet forest priors // Proceedings of ICML. 2009. P. 25-32.

43. *Mukherjee A., Liu B.* Aspect extraction through semi-supervised modeling // Proceedings of 50th Annual Meeting of Association for Computational Linguistics

(ACL-2012). 2012. P. 339-348.

44. *Ding X., Liu B., Yu Ph.* A holistic lexicon-based approach to opinion mining // Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008). 2008. P. 231-240.

45. *Neviarouskaya A., Prendinger H., Ishizuka M.* Recognition of affect, judgment, and appreciation in text // Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010). 2010. P. 806-814.

46. *Macdonald C., Santos R. L., Ounis I., Soboroff I.* Blog track research at TREC // SIGIR Forum. 2010. V. 44, No 1. P. 58-75.

47. *Dang H.T., Owczarzak K.* Overview of the tac 2008 opinion question answering and summarization tasks // Proceedings of the First Text Analysis Conference. 2008.

48. *Seki Y. et al.* Overview of multilingual opinion analysis task at NTCIR-7 // Proceedings of the Seventh NTCIR Workshop. 2008. P. 185-203.

49. *Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.* SemEval-2014 Task 4: aspect based sentiment analysis // Proceedings of International Workshop on Semantic Evaluations SemEval-2014. 2014. P. 27-35.

50. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in russian // Proceedings of BSNLP Workshop, ACL 2013. 2013. P. 12-16.

51. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. 2015. V. 2. P. 2-13.

52. *Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J.* Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. 2013. P. 3111-3119.

53. *Blinov P.D., Kotelnikov E.V.* Semantic similarity for aspect-based sentiment analysis // Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. 2015. V. 2. P. 23-33.

54. *Mayorov V., Andrianov I., Astrakhantsev N., Avanesov V., Kozlov I., Turdakov D.* A high precision method for aspect extraction in Russian // Proceedings of In-

ternational Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. V. 2. 2015. P. 34-43.

55. *Tarasov D.S.* Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews // Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. 2015. V. 2. P. 53-64.

AUTOMATIC SENTIMENT ANALYSIS TOWARDS THE ENTITY AND ITS CHARACTERISTICS

N.V. Loukachevitch

Lomonosov Moscow State University
louk_nat@mail.ru

Abstract

The paper considers approaches to sentiment analysis towards a specific entity and its characteristics (aspects). To solve the aspect-oriented sentiment analysis task, it is necessary to extract aspect terms from texts, to classify or cluster aspect terms into aspect categories, to determine the sentiment expressed towards the specfic aspect. The paper also briefly presents SentiRuEval-2015 evaluation of aspect-oriented sentiment analysis systems in Russian.

Keywords: sentiment analysis, machine learning, topic modeling, sentiment lexicon, SentiRuEval

REFERENCES

1. *Pang B., Lee L., Vaithyanathan S.* Thumbs up?: sentiment classification using machine learning techniques // Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002. V. 10. P. 79-86.

2. *Amigo E., Corujo A., Gonzalo J., Meij E., Rijke M.* Overview of RepLab 2012: Evaluating Online Reputation Management Systems // CLEF-2012. 2012.

3. *Jiang Long, Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao*. Target dependent twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011). 2011. P. 151-160.

4*. Popescu A., Etzioni O.* Extracting product features and opinions from reviews

// Natural language processing and text mining. Springer: London. 2007. P. 9-28.

5. *Liu B., Zhang L.* A survey of opinion mining and sentiment analysis // Mining Text Data. Springer: US, 2012. P. 415-463.

6. *Bagheri A., Saraee M., de Jong F.* An unsupervised aspect detection model for sentiment analysis of reviews // Natural Language Processing and Information Systems. Springer: Berlin Heidelberg, 2013. P. 140-151.

7. *Glavaš G., Korencic D., Šnajder J.* Aspect-oriented opinion mining from user reviews in Croatian // Proceedings of BSNLP workshop, ACL-2013. 2013. P. 18-23.

8. *Zhang L., Liu B.* Aspect and entity extraction for opinion mining // Data Mining and Knowledge Discovery for Big Data. Springer: Berlin Heidelberg, 2014. P. 1-40.

9. *Liu B.* Sentiment analysis and Subjectivity // Handbook of Natural Language Processing. CRC Press, Taylor and Francis Group, Boca Raton, 2010. P. 1-38.

10. *Gupta N.K.* Extracting phrases describing problems with products and services from twitter messages // Computación y Sistemas. 2013. V. 17, No 2. P. 197-206.

11. *Ivanov V., Tutubalina E.* Clause-based approach to extracting problem phrases from user reviews of products // Analysis of Images, Social Networks and Texts. Springer International Publishing, 2014. P. 229-236.

12. *Tutubalina E., Ivanov V.* Unsupervised approach to extracting problem phrases from user reviews of products // Proceedings of the Aha! Workshop on Information Discovery in Texts, Coling-2014. 2014. P. 48-53.

13. *Feng S., Bose R., Choi Y.* Learning general connotation of words using graph-based algorithms // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics. 2011. P. 1092-1103.

14. *Feng S., Kang J.S., Kuznetsova P., Choi Y.* Connotation Lexicon: a dash of sentiment beneath the surface meaning // Proceedings of ACL. 2013. P. 1774-1784.

15. *Zhang Lei, Bing Liu.* Identifying noun product features that imply opinions // Proceedings of the Annual Meeting of the Association for Computational Linguistics (short paper) (ACL-2011). 2011. P. 575-580.

16. *Liu B., Zhang L.* A survey of opinion mining and sentiment analysis // Mining Text Data. Springer US, 2012. P. 415-463.

17. *Zhang Lei, Liu B.* Extracting resource terms for sentiment analysis // Proceedings of IJCNLP-2011. 2011. P. 1171-1179.

18. *Blair-Goldensohn S., Hannan K., McDonald R., Neylon T., Reis G. A., Reynar J.* Building a sentiment summarizer for local service reviews // Proceedings of WWW Workshop on NLP in the Information Explosion Era. 2008.

19. *Hu M., Liu B.* Mining and summarizing customer reviews // Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004. P. 168-177.

20. *Marchuk A.A., Ulanov A.V., Makeev I.V., Chugreev A.A.* Extracting product features from reviews with the use of Internet statistics // Proceedings of International Conference on Computational Linguistics and Information Technologies Dialog-2013. 2013. V. 2. P. 81-91.

21. *Moghaddam S., Ester M.* Aspect-based opinion mining from online reviews. Tutorial at SIGIR-2012. 2012.

22. *Ku Lun-Wei, Yu-Ting Liang, Hsin-Hsi Chen.* Opinion extraction, summarization and tracking in news and blog corpora // Proceedings of AAAI-CAAW'06. 2006.

23. *Manning C.D., Raghavan P., Schütze H.* Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.

24. *Scaffidi Ch., Bierhoff K., Chang E., Felker M., Ng H., Jin Ch.* Red Opal: product-feature scoring from reviews // Proceedings of Twelfth ACM Conference on Electronic Commerce (EC-2007). 2007. P. 182-191.

25. *Frantzi K., Ananiadou S., Mima H.* Automatic recognition of multi-word terms: the C-value/NC-value method // International Journal on Digital Libraries, 2000. V. 3, No 2. P. 115-130.

26. *Zhu J., Wang H., Tsou B., Zhu M.* Multiaspect opinion polling from textual reviews // Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2009). 2009. P. 1799-1802.

27. *Hai Z., Chang K., Cong G.* One seed to find them all: mining opinion features via association // Proceedings of the 21st ACM international conference on Information and knowledge management. 2012. ACM. P. 255-264.

28. *Qiu G., Liu B., Bu J., Chen C.* Opinion word expansion and target extraction through double propagation // Computational Linguistics. 2011. V. 1, No 1. P. 1-18.

29. *Loukachevitch N., Nokel M.* An experimental study of term extraction for real

information-retrieval thesauri // Proceedings of Terminology and Artificial Intelligence Conference TIA-2013. 2013. P. 69-78.

30. *Zhuang L., Jing F., Zhu X.* Movie review mining and summarization // Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006), 2006. P. 43-50.

31. *Zhang L., Liu B., Lim S., O'Brien-Strain E.* Extracting and ranking product features in opinion documents // Proceedings of International Conference on Computational Linguistics (COLING-2010). 2010. P. 1462-1470.

32. *Kovelamudi S., Ramalingam S., Sood A., Varma V*. Domain independent model for product attribute extraction from user reviews using Wikipedia // Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2010). 2011. P. 1408-1412.

33. *Niklas J., Gurevych I.* Extracting opinion targets in a single and cross-domain setting with conditional random fields // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010). 2010. P. 1035-1045.

34. *Choi Y., Cardie C.* Hierarchical sequential learning for extracting opinions and their attributes // Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010). 2010. P. 269-274.

35. *Blei D., Ng A., Jordan M.* Latent Dirichlet allocation // The Journal of Machine Learning Research, 2003. No 3. P. 993-1022.

36. *Vorontsov K.V., Potapenko A.A.* Modifikacii EM-algorithma dlya veroyantnostnogo tematicheskogi modelirovaniya // Mashinnoye obuchenie I analys dannych, 2013. V. 1, № 6. P. 657-686.

37. *Titov I., McDonald R*. A joint model of text and aspect ratings for sentiment summarization // Urbana, 51, 61801. 2008.

38. *Zhao Wayne Xin, Jing Jiang, Hongfei Yan, Xiaoming Li.* Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010). 2010. P. 56-65.

39. *Zhai Z., Liu B., Xu H., Jia P.* Grouping product features using semi-supervised learning with soft-constraints // Proceedings of Coling-2010. 2010. P. 1272-1280.

40. *Zhai Z., Liu B., Xu H., Jia P.* Clustering product features for opinion mining // Proceedings of the fourth ACM International Conference on Web search and data

mining. ACM. 2011. P. 347-354.

41. *Yu J., Zha Z.J., Wang M., Wang K., Chua T.S.* Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011. P. 140-150.

42. *Andrzejewski D., Zhu X., Craven M.* Incorporating domain knowledge into topic modeling via Dirichlet forest priors // Proceedings of ICML. 2009. P. 25-32.

43. *Mukherjee A., Liu B.* Aspect extraction through semi-supervised modeling // Proceedings of 50th Annual Meeting of Association for Computational Linguistics (ACL-2012). 2012. P. 339-348.

44. *Ding X., Liu B., Yu Ph.* A holistic lexicon-based approach to opinion mining // Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008). 2008. P. 231-240.

45. *Neviarouskaya A., Prendinger H., Ishizuka M.* Recognition of affect, judgment, and appreciation in text // Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010). 2010. P. 806-814.

46. *Macdonald C., Santos R. L., Ounis I., Soboroff I.* Blog track research at TREC // SIGIR Forum. 2010. V. 44, No 1. P. 58-75.

47. *Dang H.T., Owczarzak K.* Overview of the tac 2008 opinion question answering and summarization tasks // Proceedings of the First Text Analysis Conference. 2008.

48. *Seki Y. et al.* Overview of multilingual opinion analysis task at NTCIR-7 // Proceedings of the Seventh NTCIR Workshop. 2008. P. 185-203.

49. *Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.* SemEval-2014 Task 4: aspect based sentiment analysis // Proceedings of International Workshop on Semantic Evaluations SemEval-2014. 2014. P. 27-35.

50. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in russian // Proceedings of BSNLP Workshop, ACL 2013. 2013. P. 12-16.

51. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. 2015. V. 2. P. 2-13.

52. *Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J.* Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. 2013. P. 3111-3119.

53. *Blinov P.D., Kotelnikov E.V.* Semantic similarity for aspect-based sentiment analysis // Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. 2015. V. 2. P. 23-33.

54. *Mayorov V., Andrianov I., Astrakhantsev N., Avanesov V., Kozlov I., Turdakov D.* A high precision method for aspect extraction in Russian // Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. V. 2. 2015. P. 34-43.

55. *Tarasov D.S.* Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews // Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015. 2015. V. 2. P. 53-64.

СВЕДЕНИЯ ОБ АВТОРЕ

ЛУКАШЕВИЧ Наталья Валентиновна – ведущий научный сотрудник НИВЦ МГУ им. М.В. Ломоносова, кандидат физико-математических наук, louk_nat@mail.ru. В списке трудов – более 150 работ в области автоматической обработки текстов и представления знаний.

Natalia Valentinovna LOUKACHEVITCH *is a leading researcher at Research Computer Center of Lomonosov Moscow State University. She is an author of more than 150 papers in natural language processing and knowledge representation*.

email: louk_nat@mail.ru

Материал поступил в редакцию 15 июля 2015 года