ОТ СОСТАВИТЕЛЯ

Настоящий выпуск журнала «Электронные библиотеки» представляет собой тематический сборник статей, посвященный проблеме автоматического анализа тональности текстов на русском языке.

Задача анализ тональности состоит в автоматическом определении отношения автора текста (позитивном, негативном или нейтральном) к объектам и ситуациям, о которых говорится в анализируемом тексте. В настоящее время автоматический анализ тональности используется в самых различных приложениях, включая мониторинг репутации компаний и публичных персон, анализ общественных настроений в том или ином регионе, анализ сообществ в социальных сетях и многое другое.

В данном тематическом выпуске представлены статьи участников открытого тестирования систем анализа тональности на русском языке SentiRuEval, проведенном в 2014–2015 годах. В данном тестировании участникам были предложены для решения две основные задачи.

Первая задача состояла в автоматическом анализе отзывов пользователей в двух предметных областях (рестораны и автомобили) с целью определить основные характеристики обсуждаемых объектов (так называемые аспекты, например, салат, интерьер для ресторанов) и их оценку пользователем – автором отзыва.

Вторая задача заключалась в анализе постов Твиттера (твитов) для мониторинга репутации организаций в заданной сфере деятельности (банки и телекоммуникационные компании). Данная задача включает как выявление положительного или отрицательного отношения авторов твитов к заданным организациям, так и оценку распространяемых в Твиттере позитивных или негативных новостей об этих организациях.

В статье Н.В. Лукашевич (НИВЦ МГУ им. М.В. Ломоносова) «Автоматический анализ текстов по отношению к заданному объекту и его характеристикам» представлен обзор задач, возникающих в рамках анализа тональности текстов по аспектам. Представлены особенности предлагаемых подходов и достигаемые ими характеристики качества.

Статья П.Д. Блинова и Е.В. Котельникова (Вятский государственный гуманитарный университет) «Семантическое сходство в задаче аспектно-эмоционального анализа» описывает совокупность подходов к анализу тональности текстов по аспектам, начиная с извлечения аспектов, их дальнейшей классификации и

определению тональности. Подход к извлечению аспектов сущности основан на выявлении контекстов употребления слов, представления их в виде векторов и дальнейшем группировании этих слов в аспектные категории.

В статье группы авторов из Казанского федерального университета (Е.В. Тутубалина, В.В. Иванов, М.А. Загулова, Н.Р. Мингазов, И.С. Алимова, В.А. Малых) представлены подходы на основе методов машинного обучения к обеим задачам SentiRuEval: анализ отзывов и анализ твитов. Подробно описаны признаки, используемые в применяемых методах машинного обучения, их модификации в конкретных задачах, а также проведен анализ ошибок.

В статье Ю.В. Адаскиной, П.В. Паничевой и А.М. Попова (ООО «InfoQubes», Санкт-Петербургский государственный университет) исследуется вклад синтаксического анализа в задаче анализа тональности твитов. Для этого проводится синтаксический анализ твитов, получившаяся синтаксическая структура преобразуется в тройки вида (отношение, слово1, слово2), и затем эти тройки используются как дополнительные признаки для системы классификации.

Статья П.Ю. Полякова, М.В. Калининой, В.В. Плешко (ООО «ЭР СИ О») посвящена рассмотрению лингвистико-инженерного подхода к анализу тональности твитов, включающего использование словаря оценочных слов, синтаксического анализатора, а также правил вычисления тональности на основе проведенного анализа.

В статье Ю.В. Рубцовой и С.А. Кошельникова (Институт систем информатики им. А.П. Ершова СО РАН) рассматриваются особенности применения известного метода машинного обучения CRF для анализа тональности твитов, анализируются ошибки полученного классификатора.

Нужно отметить, что мировая практика научных исследований в области автоматической обработки текстов свидетельствует о важности открытых тестирований типа SentiRuEval, в результате которых выявляются и получают большее распространение лучшие подходы, в целом ускоряется развитие автоматических систем. Поэтому практика проведения открытых тестирований становится все более распространенной в мире, в России также будут продолжаться такого рода тестирования автоматического анализа текстов на русском языке.

Н.В. Лукашевич

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

УДК 004.912

СЕМАНТИЧЕСКОЕ СХОДСТВО В ЗАДАЧЕ АСПЕКТНО-ЭМОЦИОНАЛЬНОГО АНАЛИЗА

П.Д. Блинов [1], Е.В. Котельников [2]

Вятский государственный гуманитарный университет
[1] blinoff.pavel@gmail.com, [2]kotelnikov.ev@gmail.com

Аннотация

Исследуется проблема аспектно-эмоционального анализа текста. По сравнению с общим анализом тональности такой вариант является более сложным по причине наличия ряда сопутствующих подзадач, таких, как выделение аспектных терминов, определение тональности по отношению к этим терминам и аспектным категориям. Однако решение данной проблемы значительно расширяет возможности систем автоматического анализа неструктурированного текста.

Приведен обзор предыдущих работ в области аспектно-эмоционального анализа, описаны обучающие и тестовые данные семинара SentiRuEval. Для задачи извлечения аспектных терминов использовано векторное пространство распределенных представлений слов. Тональность аспектных терминов определяется на основе функций совместной информации и семантического сходства. Приведены сравнительные результаты на тестовых данных и заключительные выводы.

Ключевые слова: *аспектно-эмоциональный анализ текста; взаимная информация; распределённые представления слов; машинное обучение; SentiRuEval.*

1. ВВЕДЕНИЕ

Важной задачей в области автоматической обработки текста стала задача анализа тональности. С исследовательской точки зрения она представляет большой интерес, потому что предполагает решение множества нетривиальных задач из области компьютерной лингвистики и машинного обучения. Практическая значимость заключается в том, что автоматический анализ мнений позволит эффективно отслеживать отношение целевой аудитории к продуктам и брендам, своевременно устранять выявленные недостатки и тем самым получать большую прибыль.

За непродолжительный период начальная постановка задачи анализа тональности претерпела значительные изменения. Общая тенденция – более детальный анализ: от определения тональности всего текста и отдельных предложений до конкретных фраз и терминов [1]. В наиболее подробной постановке проблема анализа тональности называется аспектно-эмоциональным анализом [2]. В этом случае мнения исследуются на уровне отдельных аспектов (признаков, характеристик) интересующей сущности. Например, для ресторана такими аспектами или, по-другому, аспектными категориями являются *кухня*, *сервис* и *цена*. В тексте аспекты выражаются своими аспектными терминами, например, для аспекта *кухня* терминами будут названия блюд и продуктов: *хлеб*, *салат Цезарь*, *ролы*, *паста с лососем*, *десерт* и т. д. Такие аспектные термины являются носителями тональности, которую необходимо определить.

Аспектно-эмоциональный анализ часто разбивается на три основные подзадачи: извлечение аспектных терминов; определение тональности аспектных терминов; определение тональности аспектных категорий. Ниже предложены методы решения обозначенных подзадач на основе распределённых представлений слов и семантического сходства между словами.

Статья построена следующим образом: во втором разделе представлен обзор предшествующих работ; описание обучающих и тестовых корпусов приведено в третьем разделе; предлагаемые методы и результаты на тестовых данных содержатся в четвертом разделе; заключительные выводы сделаны в пятом разделе.

2. ПРЕДЫДУЩИЕ РАБОТЫ

Большинство работ по анализу тональности посвящено определению общей тональности текста и гораздо меньше – аспектному варианту такого анализа. Относительно языков исследований, большинство подобных работ выполнено для английского [2] и меньшее количество для русского [3]. Зарубежные исследования в этой области стимулируются проведением специальных мероприятий по оценке качества решения задач анализа тональности, например, соревнования SemEval-2014 [4].

Для извлечения аспектных терминов, как правило, используются два основных подхода [2]: частотный подход; подход на основе машинного обучения.

Работа [5], вероятно, является первой и наиболее известной работой в рамках первого подхода. Общая идея сводится к поиску существительных либо словосочетаний с существительными и применению к найденным лексическим единицам некоторого метода фильтрации для выявления только терминов, релевантных аспекту. Отсев получаемых кандидатов часто выполняется с помощью статистических критериев [6] либо методов, основанных на правилах [7, 8].

Проблема извлечения аспектных терминов представляет собой более конкретную постановку общей задачи извлечения информации, одним из популярных и мощных подходов для решения которой является применение методов разметки последовательностей. Наиболее известный представитель такого подхода – метод условных случайных полей (Conditional Random Fields, CRF) [9–11]. Также для извлечения аспектных терминов применяются другие методы машинного обучения [12, 13].

С целью определения тональности аспектных терминов в подавляющем большинстве случаев используются словари оценочной лексики [14] и методы машинного обучения. Наилучшие результаты в соревновании SemEval-2014 были получены с помощью метода опорных векторов, использующего признаки на основе комбинации четырёх словарей тональности [15].

3. ТЕКСТОВЫЕ ДАННЫЕ

В качестве обучающих и тестовых корпусов использовались материалы российского семинара по тестированию систем анализа тональности SentiRuEval [16].

Корпуса были представлены отзывами пользователей о ресторанах и автомобилях. Каждый из объектов анализировался по некоторому набору аспектных категорий. Рестораны оценивались по четырём категориям: *кухня*, *интерьер*, *сервис* и *цена*. Для автомобилей такое множество состояло из шести аспектных категорий: *комфорт*, *внешний вид*, *надёжность*, *безопасность*, *управляемость* и *цена*. Для учёта мнений относительно всего объекта использовалась категория «*в целом*». В обучающих коллекциях термины указанных категорий были выделены в тексте с указанием их тональности по четырёхбалльной шкале: *позитивный*, *негативный*, *нейтральный* и *конфликтный*. Распределения терминов по шкале тональности представлены в таблице 1. Для каждого отзыва значения тональности в аналогичной шкале были проставлены в целом по аспектным категориям.

Таблица 1. Распределения терминов по шкале тональности

		Рестораны		**Автомобили**	
		Количество терминов	**%**	**Количество терминов**	**%**
Обучающие	Позитивные	1 679	69.5	1 513	48.0
	Негативные	380	13.5	858	27.2
	Нейтральные	714	25.3	690	21.9
	Конфликтные	49	1.7	91	2.9
	Всего	**2 822**	**100**	**3 152**	**100**
Тестовые	Позитивные	2 478	70.7	1 706	54.9
	Негативные	509	14.5	844	27.1
	Нейтральные	440	12.5	454	14.6
	Конфликтные	79	2.3	105	3.4
	Всего	**3 506**	**100**	**3 109**	**100**

Кроме размеченных отзывов, коллекция содержала 19 034 дополнительных отзыва для ресторанов и 8 271 отзыв для автомобилей. Такие отзывы предоставлялись без всякой разметки, но содержали общие оценки тональности, проставленные написавшими их пользователями.

4. АСПЕКТНО-ЭМОЦИОНАЛЬНЫЙ АНАЛИЗ

Существующие векторные модели представления текста обладают существенным недостатком: в них отсутствуют ассоциативные и семантические связи между терминами. Модель представления терминов на основе распределённых представлений слов устраняет такой недостаток. Как показывают эксперименты, такая модель демонстрируют способность к кластеризации семантических схожих слов [17]. Такое свойство оказывается полезным при решении подзадач аспектно-эмоционального анализа.

В предлагаемых методах для построения распределённых представлений использовалась модель с пропусками слов (skip-gram model) [17], реализованная в библиотеке Gensim [18]. Все данные, описанные в разделе 3, использовались для построения пространства распределённых представлений слов размерности 300.

4.1. Метод извлечения аспектных терминов

Из размеченной обучающей коллекции для каждого аспекта может быть получено начальное множество эталонных терминов. Отбираются только однословные существительные и глаголы. Например, в экспериментах с ресторанными отзывами для аспекта *кухня* такое множество состояло из 136 терминов: *меню*, *кухня*, *блюдо*, *еда*, *закуска*, *сок* и др.

После этого для нового проверяемого термина, представленного своим распределённым представлением $\vec{a} = (a_1,...,a_n)$, может быть вычислено его суммарное сходство с конкретным аспектом *asp*, представленным векторами своих начальных терминов $\vec{b_i} = (b_1,...,b_n)$. В качестве меры сходства между векторами использовалось косинусное сходство [19]:

$$sim(\vec{a}, asp) = \sum_{i=1}^{k} \frac{\vec{a} \cdot \vec{b_i}}{\left\|\vec{a}\right\| \cdot \left\|\vec{b_i}\right\|}, \vec{b_i} \in B_{asp}, \quad (1)$$

где B_{asp} – множество начальных терминов аспекта *asp*, $\left|B_{asp}\right| = k$ – количество начальных терминов.

Если значение *sim*, полученное в (1), превосходило заданный порог, проверяемый термин считался аспектным. Пороговые значения для каждой аспектной

категории определялись методом десятикратной перекрёстной проверки на обучающей коллекции.

Однако такой способ выявляет только однословные аспектные термины. Основываясь на данных обучающей коллекции, многословные термины составляют существенную часть всех терминов (около 1/5). Для извлечения таких многословных терминов использовался набор правил:

- объединение последовательно идущих терминов;
- объединение терминов, написанных через предлоги (*котлетки из лосося*, *роллы на гриле*);
- включение в состав термина кавычек или круглых скобок (*салат «Цезарь»*)
- проверка вхождения названия объекта и извлечение его как термина аспекта *в целом* (*кафе «Евразия»*, *ресторан «Моя Италия»*);
- и др.

Базовый алгоритм (baseline) извлечения аспектных терминов, предоставленный организаторами SentiRuEval, выполнял поиск лемматизированных терминов обучающей коллекции в тестовых отзывах [16]. В таблице 2 показаны результаты (точность – P, полнота – R, сбалансированная F_1-мера) базового алгоритма, нашего метода и методов лучших участников. Здесь и далее **полужирным** обозначены лучшие результаты, *курсивом* – результаты предлагаемых методов.

Таблица 2. Результаты извлечения аспектных терминов

		Точное соответствие			**Частичное соответствие**		
	run_id	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Рестораны	baseline	55.70	69.03	60.84	65.80	69.60	66.51
	2_1	72.37	57.38	**63.19**	80.78	61.65	68.91
	4_1	***55.06***	***69.01***	***60.70***	***68.86***	***79.16***	***72.84***
Автомобили	baseline	57.47	62.87	59.41	74.49	67.24	69.66
	2_1	76.00	62.18	**67.61**	85.61	65.51	73.04
	3_1	66.19	65.60	65.13	79.17	72.72	**74.82**
	4_1	*55.77*	*63.55*	*58.63*	*74.17*	*68.87*	*70.16*

Оценки вычислялись по двум критериям: точное и частичное соответствие. При точном соответствии аспектный термин считался выделенным верным, если его границы совпадали с границами термина, указанными асессором. При частичном соответствии верным считалось совпадение на уровне отдельных слов термина.

Согласно критерию частичного соответствия, предложенный метод показал лучший результат для предметной области ресторанов по F_1-метрике. По обоим критериям полнота значительно выше, чем точность, т. е. метод склонен выявлять много аспектных терминов, которые на самом деле не являются таковыми.

Для предметной области автомобилей получившиеся результаты находятся около базового уровня. Вероятно, это связано с недостаточным количеством данных для построения качественного пространства распределённых представлений слов. Неразмеченных отзывов об автомобилях было более, чем в два раза меньше аналогичного количества отзывов о ресторанах. Кроме этого, в автомобильных отзывах присутствуют специфичные термины, учёт которых нашим методом не производился. Например, термины с цифровыми обозначениями: *Двигатель 2.5 литра*, *ваз 2114*, *Мотор 1700 DTI*, *m30b30 двигатель*, *bmw 528i* и т. д.

В общем стоит отметить, что даже результаты относительно простого базового алгоритма оказались недостижимы для многих участников SentiRuEval. Лучшие участники лишь незначительно превзошли установленный базовый уровень (все улучшения по F_1-мере не превосходят 10%). По-видимому, сочетание ограниченного лексикона аспектов и качественной подготовки коллекций стало причиной таких результатов, т. е. обучающие коллекции содержали существенную часть всех терминов, которыми выражались конкретные аспекты. Поэтому простой поиск лемматизированных терминов, выполняемый базовым алгоритмом, позволил обнаружить существенную часть аспектных терминов, что и отражено в его результатах.

4.2. Метод определения тональности аспектных терминов

Очевидно, тональность аспектного термина определяется словами из его контекста. Для того чтобы выразить контекст числовой оценкой, использовались словари эмоциональной лексики для каждой предметной области. Построение

таких словарей выполнялось в два этапа: сначала отбор кандидатов в эмоциональные выражения, затем полученные кандидаты взвешивались для определения тональности.

На роль кандидатов в эмоциональные выражения отбирались все прилагательные и глаголы, а также фрагменты текста, соответствующие шаблону – *<не> + <прилагательное или глагол>*. Для предметной области ресторанов список кандидатов состоял из 34 822 элементов, для автомобилей – 16 416.

Взвешивание полученных кандидатов выполнялось с помощью двух оценок: семантического сходства; взаимной информации (Pointwise Mutual Information, PMI).

Для взвешивания на основе семантического сходства применялась формула (1) с единственным отличием во множестве начальных терминов *B*, которое представлялось эталонными терминами тональности (*позитивной* и *негативной*) вместо начальных терминов аспекта. Такие эталонные термины определялись экспертом и содержали 20 выражений для позитивной и негативной тональностей. Например, негативная тональность задавалась множеством выражений $B_{нег.}$={*уродливый*, *бедный*, *противный*, *ужасный*, *громкий*, *дорогой*, *грубый*, …}. Таким образом, для каждого кандидата получалось два значения суммарных сходств: сходство с позитивной тональностью sim^+ и сходство с негативной тональностью sim^-. Наибольшее значение по модулю с соответствующим знаком становилось итоговой оценкой кандидата. Например, для кандидата *потрясный* значение $sim^+ = 5.7$, а значение $sim^- = 1.6$, следовательно, кандидату приписывается оценка +5.7. В качестве других примеров можно привести: *приятный (+7.1)*, *прекрасный (+6.5)*, *стильный (+5.9)*, *неуместный (–4.8)*, *пошлый (–4.4)*, *жуткий (–4.2)*, *не резаться (–3.69)* и т. д.

Взаимная информация для тех же кандидатов вычислялась на основе дополнительных отзывов с общими оценками тональности. Для более устойчивых результатов такие отзывы были отфильтрованы, чтобы сохранить наиболее позитивные (рестораны: $score \geq 7 \rightarrow +1$ и автомобили: $score \geq 4 \rightarrow +1$) и негативные (рестораны и автомобили: $score \leq 3 \rightarrow -1$) образцы. Итоговая оценка тональности кандидата *w* определялась по следующей формуле [20]:

$$score(w) = PMI(w, pos) - PMI(w, neg). \quad (2)$$

Взаимная информация между кандидатом *w* и, например, *позитивным* классом тональности (для *негативного* класса PMI вычисляется аналогично) определяется формулой [20]:

$$PMI(w, pos) = \log_2 \frac{count(w, pos) \cdot N}{count(w) \cdot count(pos)}, \quad (3)$$

где $count(w, pos)$ – количество раз, которое кандидат *w* встретился в позитивных отзывах, *N* – общее количество терминов в корпусе, $count(w)$ – количество раз, которое кандидат *w* встретился во всех отзывах, $count(pos)$ – количество терминов в позитивных отзывах.

Примеры определённых таким образом тональностей: *классный (+3.1)*, *добротный (+2.6)*, *выдающийся (+1.6)*, *тошнить (–2.7)*, *не дружелюбный (–3.8)*, *хамский (–4.5)* и т. д.

После завершения этапа взвешивания кандидатов получался законченный словарь эмоциональной лексики, сопоставляющий каждой лексической единице (кандидату в эмоциональные выражения) две оценки тональности: на основе семантического сходства и на основе PMI. Фрагмент этого словаря представлен на рисунке 1.

Выражение	*Семантическая оценка*	*PMI оценка*
…	…	…
выгодный	–0.7	+2.5
суховатый	–1.9	+1.4
зажимать	–2.4	+0.2
горчить	–2.7	+0.6
адекватный	+3.8	–0.05
не цеплять	–2.6	+0.4
не сладкий	–1.9	+0.03
добротный	+4.3	+2.6
понятливый	+4.1	–0.4
улыбчивый	+4.5	+1.3
убогий	–4.2	–3.08
…	…	…

Рис. 1. Фрагмент словаря эмоциональной лексики

Диверсификация оценок выражений позволяет более точно оценить истинные значения их тональности. Для некоторых выражений можно проследить взаимодополнение и корректировку полученных оценок. Например, прилагательное

выгодный имеет скорее неверную оценку на основе PMI –0.7, тогда как семантическая оценка +2.5 является более правильной.

С помощью полученных словарей каждый аспектный термин представлялся в ближайшем (три термина слева и справа) и дальнем (шесть терминов слева и справа) контексте, образуя вектор признаков. Далее такие вектора использовались как входные данные для классификатора на основе решающих деревьев (Gradient Boosting Classifier) [21].

Аспектные термины *конфликтной* тональности очень малочисленны (см. табл. 1). Для методов машинного обучения определение таких непредставительных классов довольно проблематично. Путём просмотра обучающей коллекции была выявлена простая закономерность, сохраняющаяся для большинства терминов этой тональности: наличие союза «но» после термина. Поэтому для выявления конфликтной тональности применялось следующее правило: приписывать термину конфликтную тональность, если после него в предложении встречается союз «но».

Базовый алгоритм для этой задачи назначал наиболее часто встречающуюся тональность (*позитивную*) обучающей коллекции всем терминам тестовой коллекции. Результаты базового алгоритма, предлагаемого метода и участников, занявших вторые места, приведены в таблице 3.

Таблица 3. Результаты определения тональности аспектных терминов

		Micro-averaging			**Macro-averaging**		
	run_id	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Рестораны	baseline	71.04	71.04	71.04	32.09	25.06	26.71
	4_1	***82.49***	***82.49***	***82.49***	***58.72***	***55.69***	***55.45***
	3_1	66.96	66.96	66.96	32.23	24.30	26.96
Автомобили	baseline	61.92	61.92	61.92	29.49	26.85	26.48
	4_1	***74.28***	***74.28***	***74.28***	***57.25***	***56.67***	***56.84***
	1_2	65.31	65.31	65.31	35.63	32.97	34.22

Предлагаемый метод показал стабильно высокие результаты для обеих предметных областей.

4.3. Метод определения тональности аспектных категорий

Завершающей задачей аспектно-эмоционального анализа является определение тональности в целом для аспектных категорий. Поскольку в ходе выполнения предыдущих методов извлечены аспектные термины и определены их тональности, остаётся просуммировать полученные значения по каждому из аспектов. Тональности терминов приводились к оценкам путём следующего преобразования: *позитивная: +1*, *негативная: –1*, *конфликтная: 0*. Суммирование по всем терминам аспектной категории определяет тональность всей категории. При положительных значениях итоговой оценки аспектной категории приписывается *позитивная* тональность, при отрицательных значениях – *негативная*. Если хотя бы один термин упомянут с конфликтной тональностью, то вся категория помечалась как *конфликтная*. Отсутствие терминов аспекта указывало на отсутствие мнения по этому аспекту.

Таблица 4. Результаты определения тональности аспектных категорий (F_1-мера)

		run_id		
	Аспект	**baseline**	***4_1***	**10_1**
Рестораны	Кухня	27.89	***45.27***	41.88
	Интерьер	28.45	***48.62***	36.57
	Цена	24.39	***45.40***	34.01
	В целом	27.89	***38.67***	27.98
	Сервис	27.36	***51.09***	45.98
	Среднее	27.20	***45.81***	37.28
Автомобили	Комфорт	22.64	***51.09***	
	Внешний вид	28.37	***44.86***	
	Надёжность	20.93	***42.51***	
	Безопасность	21.79	***43.05***	
	Управляемость	24.38	***44.74***	
	В целом	21.92	***49.61***	
	Цена	25.72	***31.45***	
	Среднее	23.68	***43.90***	

Базовый алгоритм приписывал наиболее распространённую тональность аспектной категории (согласно обучающей коллекции) соответствующим аспектным категориям тестовой коллекции. Результаты метода показаны в таблице 4.

Полученные результаты являются самыми низкими по сравнению с аналогичными значениями для задач извлечения терминов и определения их тональности. Это объясняется высокой сложностью задачи определения тональности аспектных категорий. При вычислении таких интегральных оценок метод оперирует извлечёнными аспектными терминами и их тональностями. При этом появляются два рода ошибок: ошибки, связанные с не извлечёнными или ложно извлечёнными терминами; ошибки определения тональности терминов.

Для предметной области отзывов о ресторанах наиболее сложной категорией являлась категория *в целом*. Аспектная категория *сервис*, напротив, была самой лёгкой. Это связанно с тем, что набор терминов для этой категории довольно ограничен, а значит и вероятность ошибок первого рода меньше.

Для предметной области автомобилей самой простой в определении оказалась категория *комфорт*, а самой сложной – *цена*. Для категории *цена* сложность, вероятно, является следствием того, что во многих случаях выражение мнения по этой категории связано с озвучиванием конкретных цифр, например, «*За такую цену 450000 рублей стоит купить*» или «*Покупал ВА32110 за 86000, а вложил 18000 – не очень получилось заработать*». Относительно таких примеров нужно знать, много это или мало, т. е. явно требуется больше экспертных знаний, помимо тех, которыми располагает система на текущий момент.

ЗАКЛЮЧЕНИЕ

В статье предложен полный набор методов для решения задачи аспектно-эмоционального анализа. Приведены экспериментальные результаты на корпусе отзывов двух предметных областей российского семинара по тестированию систем анализа тональности SentiRuEval.

По критерию частичного соответствия для предметной области ресторанов метод извлечения аспектных терминов показал лучший результат среди 14 методов. По критерию точного соответствия результаты несколько хуже, но по-прежнему среди лучших. Методы определения тональности терминов и аспектных ка-

тегорий показали стабильно высокие результаты для обеих предметных областей. Полученные результаты позволяют заключить, что предлагаемые методы могут быть использованы в практических задачах для выявления мнений пользователей по конкретным аспектам.

Благодарности

Работа выполнена при финансовой поддержке Министерства образования и науки РФ, государственное задание ВятГГУ (код проекта 586).

СПИСОК ЛИТЕРАТУРЫ

1. *Feldman R.* Techniques and applications for sentiment analysis // Communications of the ACM. 2013. V. 56. P. 82- 89.

2. *Liu B.* Sentiment analysis and opinion mining // Synthesis Lectures on Human Language Technologies. 2012. V. 5.

3. *Blinov P.D., Kotelnikov E.V.* Using distributed representations for aspect-based sentiment analysis // Proceedings of International Conference Dialog. 2014. Issue 13 (20). P. 64-75.

4. *Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.* SemEval-2014 Task 4: Aspect Based Sentiment Analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 27-35.

5. *Hu M., Liu B.* Mining and summarizing customer reviews // Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. P. 168-177.

6. *Schouten K., Frasincar F., Jong F.* COMMIT-P1WP3: A Co-occurrence based approach to aspect-level sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 203-207.

7. *Pekar V., Afzal N., Bohnet B.* UBham: lexical resources and dependency parsing for aspect-based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 683-687.

8. *Zhang F., Zhang Z., Lan M.* ECNU: A combination method and multiple features for aspect extraction and sentiment polarity classification // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 252-258.

9. *Kiritchenko S., Zhu X., Cherry C., Mohammad S.* NRC-Canada-2014: Detecting aspects and sentiment in customer reviews // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 437-442.

10. *Chernyshevich M.* IHS R&D Belarus: cross-domain extraction of product features using conditional random fields // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 309-313.

11. *Toh Z., Wang W.* DLIREC: aspect term extraction and term polarity classification system // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 235-240.

12. *Brun C., Popa D., Roux C.* XRCE: hybrid classification for aspect-based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 838-842.

13. *Gupta D., Ekbal A.* IITP: supervised machine learning for aspect based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 319-323.

14. *Bornebusch F., Cancino G., Diepenbeck M., Drechsler R., Djomkam S., Fanseu A., Jalali M., Michael M., Mohsen J., Nitze M., Plump C., Soeken M., Tchambo F., Toni, Ziegler H.* iTac: aspect based sentiment analysis using sentiment trees and dictionaries // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 351-355.

15. *Wagner J., Arora P., Cortes S., Barman U., Bogdanova D., Foster J., Tounsi L.* DCU: aspect-based polarity classification for SemEval Task 4 // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 223-229.

16. *Loukachevitch N.V., Blinov P.D., Kotelnikov E.V., Rubtsova Yu.V., Ivanov V.V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in russian // Proceedings of International Conference Dialog. 2015. P. 2-13.

17. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed representations of words and phrases and their compositionality // Proceedings of NIPS. 2013. P. 3111-3119.

18. Gensim – topic modeling library. URL: http://radimrehurek.com/gensim (дата обращения: 10.04.2015).

19. *Manning C., Raghavan P., Schütze H.* Introduction to information retrieval. Cambridge University Press. New York. 2008.

20. *Islam A., Inkpen D.* Second order co-occurrence PMI for determining the semantic similarity of words // Proceedings of the International Conference on Language Resources and Evaluation. 2006. P. 1033-1038.

21. *Friedman J.* Greedy function approximation: a gradient boosting machine // The Annals of Statistics. 2001. V. 29. P. 1189-1232.

SEMANTIC SIMILARITY FOR ASPECT-BASED SENTIMENT ANALYSIS

P.D. Blinov, E.V. Kotelnikov

Vyatka State Humanities University

[1]blinoff.pavel@gmail.com, [2]kotelnikov.ev@gmail.com

Abstract

The article investigates the problem of aspect-based sentiment analysis. Such version of analysis is more challenging compared to general task of sentiment detection problem. It implies the solutions to the number of related subtasks such as aspect term extraction, aspect term polarity detection and aspect category polarity detection. The solution of aspect-based sentiment analysis problem significantly extends the capabilities of natural language processing systems.

The article gives the overview of previous works in the field and describes the train and test data from the Russian evaluation workshop SentiRuEval. For the task of aspect term extraction the vector space of distributed representations of words was used. Aspect term detection is based on mutual information method and semantic similarity. The paper contains the number of experimental results. At the end the final conclusions are drawn.

Keywords: *aspect-based sentiment analysis; mutual information; distributed representations of words; machine learning; SentiRuEval*.

REFERENCES

1. *Feldman R.* Techniques and applications for sentiment analysis // Communications of the ACM. 2013. V. 56. P. 82-89.

2. *Liu B.* Sentiment analysis and opinion mining // Synthesis Lectures on Human Language Technologies. 2012. V. 5.

3. *Blinov P.D., Kotelnikov E.V.* Using distributed representations for aspect-based sentiment analysis // Proceedings of International Conference Dialog. 2014. Issue 13(20). P. 64-75.

4. *Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.* SemEval-2014 Task 4: Aspect Based Sentiment Analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 27-35.

5. *Hu M., Liu B.* Mining and summarizing customer reviews // Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. P. 168-177.

6. *Schouten K., Frasincar F., Jong F.* COMMIT-P1WP3: A Co-occurrence based approach to aspect-level sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 203-207.

7. *Pekar V., Afzal N., Bohnet B.* UBham: lexical resources and dependency parsing for aspect-based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 683-687.

8. *Zhang F., Zhang Z., Lan M.* ECNU: A combination method and multiple features for aspect extraction and sentiment polarity classification // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 252-258.

9. *Kiritchenko S., Zhu X., Cherry C., Mohammad S.* NRC-Canada-2014: Detecting aspects and sentiment in customer reviews // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 437-442.

10. *Chernyshevich M.* IHS R&D Belarus: cross-domain extraction of product features using conditional random fields // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 309-313.

11. *Toh Z., Wang W.* DLIREC: aspect term extraction and term polarity classification system // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 235-240.

12. *Brun C., Popa D., Roux C.* XRCE: hybrid classification for aspect-based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 838-842.

13. *Gupta D., Ekbal A.* IITP: supervised machine learning for aspect based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 319-323.

14. *Bornebusch F., Cancino G., Diepenbeck M., Drechsler R., Djomkam S., Fanseu A., Jalali M., Michael M., Mohsen J., Nitze M., Plump C., Soeken M., Tchambo F., Toni, Ziegler H.* iTac: aspect based sentiment analysis using sentiment trees and dictionaries // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 351-355.

15. *Wagner J., Arora P., Cortes S., Barman U., Bogdanova D., Foster J., Tounsi L.* DCU: aspect-based polarity classification for SemEval Task 4 // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 223-229.

16. *Loukachevitch N.V., Blinov P.D., Kotelnikov E.V., Rubtsova Yu.V., Ivanov V.V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog. 2015. P. 2-13.

17. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed representations of words and phrases and their compositionality // Proceedings of NIPS. 2013. P. 3111-3119.

18. Gensim – topic modeling library. URL: http://radimrehurek.com/gensim (дата обращения: 10.04.2015).

19. *Manning C., Raghavan P., Schütze H.* Introduction to information retrieval. Cambridge University Press. New York. 2008.

20. *Islam A., Inkpen D.* Second order co-occurrence PMI for determining the semantic similarity of words // Proceedings of the International Conference on Language Resources and Evaluation. 2006. P. 1033-1038.

21. *Friedman J.* Greedy function approximation: a gradient boosting machine // The Annals of Statistics. 2001. V. 29. P. 1189-1232.

СВЕДЕНИЯ ОБ АВТОРАХ

КОТЕЛЬНИКОВ Евгений Вячеславович – кандидат технических наук, доцент Вятского государственного гуманитарного университета.

Evgeny Vyacheslavovich KOTELNIKOV, Candidate of Engineering Sciences (2006). Currently is an Associate Professor at the Department of Applied Mathematics and Computer Science at the Vyatka State Humanities University. Current scientific interests: natural language processing, machine learning.

email: kotelnikov.ev@gmail.com

БЛИНОВ Павел Дмитриевич – инженер-программист факультета информатики, математики и физики Вятского государственного гуманитарного университета.

Pavel Dmitrievich Blinov, software engineer of faculty of computer science, mathematics and physics, Vyatka State Humanities University.

Current scientific interests: data mining, natural language processing, sentiment analysis, machine learning.

email: blinoff.pavel@gmail.com

Материал поступил в редакцию 15 июля 2015 года

УДК 004.912

ТЕСТИРОВАНИЕ МЕТОДОВ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА, ОСНОВАННЫХ НА СЛОВАРЯХ

Е.В. Тутубалина[1], В.В. Иванов[1], М.А. Загулова[1], Н.Р. Мингазов[1], И.С. Алимова[1], В.А. Малых[2]

[1]Высшая школа Информационных технологий и информационных систем Казанского федерального университета

[2]Институт системного анализа РАН, г. Москва

elvtutubalina@kpfu.ru, vivanov@kpfu.ru, lolmariya@gmail.com, icrotek547@gmail.com, alimovaIlseyar@gmail.com, alem.mipt@gmail.com

Аннотация

Технологии анализа тональности текста развиваются интенсивно, что обусловлено ростом объемов открытых источников, представляющих мнения пользователей интернета по различным вопросам. В статье описаны методы для анализа тональности текстов отзывов и коротких сообщений (твитов), приводятся результаты оценки их качества, которая производилась в рамках российского семинара SentiRuEval-2015.

Ключевые слова: *извлечение информации, анализ тональности, классификация текстов, машинное обучение с учителем*

1. ВВЕДЕНИЕ

Задача анализа мнений в документах или отзывах пользователей (opinion mining) является актуальным направлением исследований в области обработки естественного языка, активно развивающимся в настоящее время. Это связано с ростом объема открытых текстовых источников, представляющих данные о мнениях пользователей интернета по различным вопросам. Поскольку учет информации о мнениях потребителей продукта или услуги имеет ценность, как для других пользователей, так и для поставщиков и производителей данного продукта или услуги, задача автоматического извлечения такой информации (задача анализа тональности текста) является актуальной.

Различные постановки задачи анализа тональности (sentiment analysis) можно классифицировать по конечной цели, которая стоит перед системой обработки текста. Обычно задачу классификации текстов в целом по тональности отделяют от задачи определения тональных высказываний относительно аспектных терминов (далее аспектов), о которых высказывание было сделано (aspect-based sentiment analysis). Современные методы решения задач анализа тональности текста принято сравнивать друг с другом в рамках научных семинаров, таких, как SEMEVAL [1, 2], где любому участнику доступны одни и те же наборы данных для обучения методов и постановки задач. Для систем обработки русскоязычных текстов такие «соревнования» проводятся, начиная с 2011 года. Исторически методы решения задачи классификации текстов по их тональности рассматривались ранее в серии семинаров SentiRuEval (в рамках ROMIP [3]). В 2014–2015 годах перед участниками SentiRuEval ставились задачи анализа тональности отзывов относительно аспектов о ресторанах и автомобилях, а также классификации коротких сообщений (твитов) из социальной сети Twitter. Ниже приведен список задач анализа тональности и «звездочкой» отмечены те методы, решения которых представлены в данной статье. Подробное описание исходных данных и задач можно найти в работе [4]. Список задач анализа тональности отзывов, поставленных в рамках SentiRuEval-2015:

A) Задача извлечения явных аспектов;

B) Задача извлечения всех аспектов и тональных фактов;

C) Задача определения тональности относительно явных аспектов (*);

D) Задача категоризации явных аспектов по аспектным категориям (*);

E) Задача определения тональности по каждой из аспектных категорий.

Кроме того, в рамках SentiRuEval решалась задача анализа тональности твитов о банках и телекоммуникационных компаниях. В следующем разделе дан краткий обзор состояния исследований в области анализа тональности. В разделах 3 и 4 приводятся описания методов и полученные результаты.

2. СОВРЕМЕННОЕ СОСТОЯНИЕ ИССЛЕДОВАНИЙ

Извлечение аспектных терминов (аспектов). Это широко известная задача в области анализа тональности, которая требует глубокого понимания каждого

аспекта продукта. Существует несколько наиболее популярных методов, решающих задачу извлечения аспектов как бинарную задачу классификации [5], как задачу классификации последовательностей (sequential classification) [6–8], как задачу тематического моделирования или традиционную задачу кластеризации [9, 10]. Цель классификации – определить, являются ли термины, существительные и словосочетания искомым объектом (аспектом, относительно которого высказывается некоторое мнение). В [5] используются синтаксические шаблоны, соотнесённые с тональностью из словаря общего назначения, для определения наиболее частых словосочетаний с существительными. В [11] предлагается подход, основанный на правилах, используются деревья зависимостей для предложений. Эти подходы не совершенны и дают низкие результаты для аспектов, встречающихся в тексте редко, и для более сложных случаев извлечения аспектов. В работах 2014 года [7, 8] были предложены две модификации метода условных случайных полей (Conditional Random Fields, CRF) для классификации последовательностей в задаче извлечения аспектных терминов.

Анализ тональности аспектов. Большинство ранних подходов к классификации аспектов полагались на вручную созданные словари, содержащие списки позитивных и негативных слов. В работе [12] был предложен метод обучения без учителя, основанный на подсчёте поточечной взаимной информации между фразой и двумя словами-индикаторами для каждой фразы. Методы машинного обучения широко применялись для аспектно-ориентированной классификации [7, 13–15]. Могаддам и Эстер в [9] для извлечения аспектов и их тональности предложили модификацию вероятностных тематических методов, основанных на скрытом распределении Дирихле (Latent Dirichlet Allocation, LDA), используя синтаксические связи зависимостей между аспектами, позитивными и негативными словами в предложении. Однако тематические модели показывают худшие результаты классификации тональных высказываний относительно аспектных терминов по сравнению с классификатором на основе метода опорных векторов (Support Vector Machine, SVM) [16].

Определение категории аспекта. Автоматическая категоризация явных аспектов изучается в рамках задачи резюмирования аспектных терминов по тематическим категориям. Данная задача была исследована в работе [9] как часть за-

дачи аспектного латентного анализа: выделением групп заранее известных ключевых терминов и предсказанием рейтинга каждой из групп. На SemEval-2014 также была проделана работа по разбиению аспектных терминов из отзывов на группы для анализа тональности. Лучшие результаты по F-мере были достигнуты подходами, в которых использовались классификаторы SVM с моделью «мешок слов» и информации из коллекции неразмеченных отзывов [7, 17].

Анализ тональности коротких сообщений. Получение информации из коротких неформальных сообщений, таких, как твиты и смс, вызывает исследовательский интерес в области анализа тональности [7, 18, 19], отслеживания событий [20], извлечения проблем [21], обнаружения сарказма [22] и пр. Традиционный подход к классификации тональности основывался на присутствии негативных и позитивных слов или пиктограмм, изображающих эмоцию (далее – эмотиконов), служащих индикаторами позитивной либо негативной окраски [12, 23, 24]. В современных исследованиях применялись также смешанные подходы, где лексические ресурсы (словари тональности) комбинировали с техниками машинного обучения [7, 25, 26]. Было показано, что биграммы и униграммы модели «мешок слов» важны для машинного обучения и что признаки, характерные для твиттера (хэштеги, ретвиты, ссылки), способствуют улучшению результатов классификации [27]. В [7] описаны эксперименты, показывающие важность определения тональных слов относительно негативного контекста в предложении и необходимость использования нескольких словарей для более точной классификации.

Многие работы по анализу тональности включают в себя формирование лексических ресурсов с указанием тональности слов [28]. Словари создаются различными способами, от ручной разметки до полностью автоматизированных подходов. В [26] был вручную расширен существующий словарь тональных слов, эмотиконов и интернет-сленга. В [25] словарь сформировали автоматически из слов, встреченных по эмоциональным хэштегам, высчитав для каждого слова поточечную взаимную информацию со словами из обучающего множества. Учитывая тот факт, что данная система победила в международном семинаре SemEval, был использован аналогичный подход к генерации словарей.

Анализ тональности сообщений на русском языке изучен на данный момент слабее. В [29] описано первое открытое тестирование анализа тональности отзывов пользователей на русском языке. В нём заняли первые места методы машинного обучения с учителем, использующие метод опорных векторов и словари, и системы, основанные на правилах, классифицирующие отзывы о фильмах, книгах, цифровых камерах. В работе [30] предлагается подход, основанный на специальных словарях и семантических фильтрах. Другие участники использовали созданные вручную словари эмотиконов для каждой из трёх предметных областей [15]. Эксперименты подтвердили, что методы машинного обучения показывают наилучшие результаты относительно традиционных методов, основанных на знаниях, для отзывов пользователей на русском языке. Было показано, что популярные методы машинного обучения не являются универсальными, поскольку каждый классификатор показал наилучшие результаты лишь в одной из предметных областей.

3. ОПИСАНИЕ МЕТОДОВ АНАЛИЗА ТОНАЛЬНОСТИ НА ОСНОВЕ СЛОВАРЕЙ

В этом разделе описан подход к двум заданиям аспектно-ориентированного анализа тональности пользовательских отзывов о ресторанах и автомобилях и к заданию классификации твитов пользователей о банках и телекоммуникационных компаниях. При помощи свободного программного обеспечения для анализа данных были применены техники машинного обучения (см. раздел 3.2), основанные на модели «мешка слов» и признаках, которые подробно описаны в разделе 3.4. Значительная часть признаков строится при помощи тональных словарей двух видов – созданных вручную и автоматически.

3.1. Предобработка текстов

Твиты пишутся неформальным языком и содержат много неформальной лексики, ошибок и специальных символов, поэтому они обрабатывались следующим образом. Все упоминания пользователей нормализуются до @username. Ссылки заменяются @link. Слова, которые используются в отрицательном контексте (например, *не понравилось, не круто, нет прекеканий*), помечаются приписываемым к ним тегом neg_. Текстовые эмотиконы заменяются индикатором той эмоции, которую выражают (например, '*:-)*' заменится на happyEmoticon, '*o_O*' –

на surpriseEmoticon, *';-]*' – на winkEmoticon). Морфологический анализатор MyStem приводит все слова в их начальную форму. Слова в текстах отзывов были также нормализованы при помощи MyStem.

3.2. Методы машинного обучения для анализа тональности

Задача (C) классификации заключалась в предсказании тональной полярности каждого аспекта из отзывов о продуктах (негативный, позитивный, две тональности, нейтральный). Был применён классификатор MaxEnt с параметрами по умолчанию, основанный на модели «мешка слов» и признаках, которые описаны в разделе 3.4. При заданном контексте аспектного термина для извлечения признаков создаются два типа биграмм слов: (1) контекстные биграммы из текста внутри контекстного интервала аспектного термина; (2) аспектные биграммы как комбинация аспектного термина и контекстного слова к нему из контекстного интервала. Контекстный интервал для аспектного термина w_i – последовательность ($w_{i-4}, \cdots, w_{i+4}$).

Задача D заключалась в классифицировании аспекта в одну из установленных заранее категорий. Для ресторанов этими категориями были: еда, сервис, интерьер, цены, общее впечатление. Для автомобилей: маневренность, надёжность, безопасность, внешний вид, комфорт, цена, общее впечатление. Было выполнено обучение классификатора на основе машин опорных векторов (SVM) с последовательной минимальной оптимизацией (SMO). Для каждого аспектного термина был извлечен набор признаков из его контекстного интервала (2 слова до аспектного термина, 2 слова после). Были созданы словари категорий, основанные на оценке каждого слова w в обучающем множестве.

Задача анализа твитов состоит в определении, содержится ли в твите о телекоммуникационных/финансовых компаниях позитивная, негативная либо нейтральная тональность. Применен подход на основе машинного обучения, использующий модель «мешка слов» и вектор признаков, каждый из которых описан ниже. Были протестированы три различных алгоритма обучения (Naive Bayes, MaxEnt, SVM). Лучшие результаты показал SVM, с настройками по умолчанию библиотеки scikit-library на языке Python.

3.3. Два вида словарей тональности

Словари позитивной и негативной лексики строились двумя методами: автоматически и вручную. При создании словаря вручную были собраны отзывы пользователей с сайта otzovik.com. Для точности в позитивный корпус вошли только тексты *Преимуществ*, с оценкой 5, а в негативный корпус – только *Недостатки*, с оценкой 1 или 2. *Преимущества* и *Недостатки* – это те части отзывов, в которых пользователь пишет только то, что он оценил позитивно в продукте, либо, соответственно, только то, что не понравилось. Для каждой предметной области в каждом корпусе было выбрано определённое число глаголов, существительных, прилагательных и наречий, встречающихся максимально часто. Для предметной области телекоммуникаций и банков удалялись те слова, которые являются явно нейтральными для данной предметной области (например, *связь, услуга, платёж, скорость, сотрудник*). Корпусы были расширены формами присутствующих в них слов. В Таблице 1 собрана информация об объёме полученных словарей в сравнении с числом отзывов, для каждой предметной области.

Таблица 1. Информация о словарях, сгенерированных вручную

	Число собранных отзывов	Число позитивных слов, вошедших в словарь	Число негативных слов, вошедших в словарь
Рестораны	7526	741	362
Автомобили	4951	1576	741
Банки	3357	139	131
Телекоммуникации	1928	68	168

Для подсчёта значений словарных признаков для задач C и D использовались взвешенные оценки: каждое слово предложения взвешивается степенью своей удалённости от конкретного аспекта:

$$score(w) = \frac{sc(w)}{e^{|i-j|}}, \quad (1)$$

где i, j – позиции аспектного термина и слова w, $sc(w)$ – тональность слова w, равная 1 для позитивных слов и -1 для негативных, определяется по словарю. Словарь, генерируемый автоматически для твитов, использует обучающее множество размеченных твитов (предоставленное SentiRuEval-2015). Как и в работе [25], подсчитывается тональный балл (score) для каждого слова (w) в этом обучающем множестве:

$$score(w) = PMI(w, pt) - PMI(w, nt), \quad (2)$$

$$PMI(w, pt) = \log_2 \frac{p(w,pt)}{p(w)*p(pt)}, \quad (3)$$

где PMI – поточечная взаимная информация, pt – число позитивных твитов, nt – число негативных твитов, $p(w)$, $p(pt)$ и $p(w, pt)$ – вероятности появления *w* в позитивном корпусе. Тональность слова максимально наглядна: слово «сомнительный» имеет значения -19.68 и 0.28, тогда как «спустя» – только -0.80 и 0.14. Поскольку в твитах встречаются слова, бесполезные с точки зрения задачи, редко встречающиеся слова игнорировались, если они встречались в обучающем множестве менее трёх раз. Автоматический словарь категорий аспектов работает по следующей формуле:

$$score(w) = PMI(w, cat) - PMI(w, oth), \quad (4)$$

$$PMI(w, cat) = \log_2 \frac{p(w,cat)}{p(w)*p(cat)}, \quad (5)$$

где PMI – поточечная взаимная информация, cat – все контексты аспекта в конкретной категории, oth – контексты аспекта во всех прочих категориях, $p(w)$, $p(cat)$ и $p(w, cat)$ – вероятности появления слова w в контексте аспекта конкретной категории.

3.4. Признаковое пространство

Каждый отзыв преобразуется в вектор признаков, которые зависят от аспекта и его контекста в предложении. Каждый твит – в вектор лексических, словарных и характерных для твиттера признаков. Кратко опишем признаки, использованные во всех задачах:

- **n-граммы слов**: униграммы (отдельные слова) и биграммы (словосочетания) из текста твита, если встречаются в множестве твитов/отзывов более двух раз;
- **n-граммы символов**: n=2,..,4 строчных букв, встречающихся более двух раз;

– **признаки словаря, построенного вручную**: количество совпавших позитивных слов; негативных слов;

– **признаки автоматического словаря:** количество слов с со значением score, большим нуля; количество слов с score, меньшим нуля; максимальный тональный score; минимальный тональный score; сумма тональных score; сумма позитивных score; сумма негативных; если с тональным словом в тексте употреблено отрицание, то его тональность меняется на противоположную.

При анализе отзывов дополнительно использовались следующие признаки:

– **контекстные n-граммы:** униграммы и биграммы из контекстного интервала; извлекаются следующими несколькими комбинациями: замена аспектного термина словом *aspect*; замена тональных слов на их тэг тональности (*pos* или *neg)*, на часть речи слов.

– **аспектные биграммы:** биграммы как комбинации самого аспектного термина и слова из контекстного интервала; извлекаются теми же комбинациями, что описаны выше.

Поскольку размеры контекстного интервала ограничены и сложно классифицировать аспект одновременно как позитивный и негативный, для этих случаев было сформулировано своё правило: для предложений, содержащих аспектный термин и союзы *а* или *но,* если классификатор размечает аспект как нейтральный, его оценка меняется на both (двойственная тональность). При анализе твитов дополнительно использовались следующие признаки:

– **слова прописными буквами**: число в твите слов, набранных полностью прописными буквами;

– **пунктуация**: число подряд идущих вопросительных знаков, восклицательных знаков либо их комбинаций. Учитываются последовательности из более, чем одного знака. «*Почему у дебетовой карты списали деньги просто так?!?»* Для подобного твита, с нейтральным выбором слов, признак очень полезен, так как указывает на негатив;

– **последний символ**: является ли последний символ в твите скобкой или восклицательным знаком. №*Сеть прыгает из E в 3G и обратно каждые 5 минут ((*" Встречая такой твит, классификатор не знает, хорошо ли это – прыгать в 3G, но скобки помогают определить, что твит негативный;

– **эмотиконы**: выделяются четыре признака – число позитивных эмотиконов, число негативных эмотиконов, является ли последний униграм твита позитивным эмотиконом, является ли негативным;

– **признаки, специфичные для твиттера**: три бинарных значения – содержит ли твит упоминания юзера, является ли ретвитом, содержит ли твит ссылку.

4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Данный раздел описывает результаты, достигнутые классификаторами твитов, аспектов и категорий аспектов. Каждая из таблиц содержит также *базовые* результаты (baseline) и результаты победителей дорожки SentiRuEval-2015, определённых по макро F-мере, которая выбрана основным критерием соревнования [4]. Раздел содержит эксперименты с удалением различных признаков классификаторов с целью определения значимости каждой группы в рамках задач анализа мнений.

4.1. Результаты и эксперименты в рамках задачи анализа коротких сообщений пользователей

В Таблице 2 приведены результаты экспериментов, проведенных в рамках SentiRuEval-2015, по задаче классификации определения тональности относительно компаний в твитах о банках и телекоммуникациях. Макро F-мера считалась как среднее между F-мерой для позитивного класса и F-мерой негативного класса, нейтральный класс исключается. Предложенные методы показали 4 и 2 результаты среди 10 и 7 команд в «дорожке» по банкам и телекоммуникациям соответственно.

Таблица 2. Метрики качества в задаче классификации твитов в банковской сфере и твитов о телекоммуникационных компаниях

Участник	Телекоммуникации		Банки	
	Micro F,%	Macro F,%	Micro F,%	Macro F,%
Лучший	**53.6**	**48.8**	**34.3**	**35.9**
Предложенный подход	52.8	44.8	33.7	35.2
Офиц. baseline	33.7	18.2	23.8	12.7

Результаты экспериментов по классификации твитов о компаниях описаны в Таблице 3.

Таблица 3. Результаты экспериментов при удалении различных признаков

Набор признаков	Телекоммуникации			Банки		
	macro P, %	macro R, %	macro F, %	macro P, %	macro R, %	macro F, %
Полный вектор	44.3	47.1	44.7	**53.8**	27.9	35.2
Без n-грамм слов	39.0	41.2	**37.3**	50.7	**31.6**	**37.3**
Без n-грамм символов	**44.7**	41.3	40.5	44.4	**23.3**	**30.1**
Без пунктуации	42.9	42.9	41.2	52.2	28.6	35.0
Без прописных	44.6	44.7	43.6	49.8	29.3	34.9
Без эмотиконов	41.3	45.0	40.6	48.9	27.4	33.5
Без последнего символа	45.8	**37.9**	39.0	50.9	27.4	34.0
Без твиттер-признаков	44.7	44.1	44.3	49.1	28.9	35.1
Без словаря, составленного вручную	**37.9**	50.5	43.2	51.6	27.0	34.0
Без словаря, составленного автоматически	42.7	**56.9**	**48.8**	**42.6**	29.2	34.3
Без обоих словарей	41.9	55.3	47.5	49.6	27.6	33.7

Самыми полезными признаками с точки зрения решения задачи оказались n-граммы символов, словари и эмотиконы. Метод показывает улучшение в 0.021% F-меры после удаления n-грамм слов в твитах о банках, и улучшение в 0.041% F-меры после удаления слов автоматического лексикона в области телекоммуникаций. Объяснением этому может служить динамичность контекста сообщений о банках, поскольку твиты обучающего множества датированы 2014 годом, а твиты для тестового множества собирались в 2013 году. Вследствие этого

многие положительные и негативные слова из обучающей выборки либо получили противоположную тональность, либо не были найдены в тестовом корпусе.

После проведения подробного анализа текстов ошибочно классифицированных твитов и того, в какие классы они были ошибочно занесены, были определены следующие типы наиболее частых ошибок классификации:

- орфографические ошибки и транслитерация;
- хэштеги, склеивающие несколько слов;
- эмоциональное обсуждение нейтральных тем;
- недостаточный размер тональных словарей (не замечены ярко-окрашенные тональные слова, незнакомые классификатору).

Таблица 4. Распределение типов ошибок

Предметная область	Орфография и транслитерация	Хэштеги из нескольких слов	Эмоции на нейтральные темы	Слова не из тонального словаря
Телекоммуникации	20.4%	8%	14.9%	43%
Банки	9%	1%	11%	64%

Таблица 4 демонстрирует, что большая часть ошибок возникла по причине недостаточного покрытия эмоционально-окрашенных слов соответствующим словарём тональности. В твите «*Билайну труба короче*» слово «труба» содержит негативное значение, неизвестное классификатору. Негативные твиты, такие, как «*Самый безалаберный банк!»*, классифицированы неверно, поскольку такие негативные слова, как «безалаберный», редко употребляются в речи и ни разу не появились в обучающем множестве.

Однако меры для обработки орфографических ошибок не были приняты, и поэтому слова *ацтой* (отстой) и *чорд* (чёрт) не были узнаны классификатором, хотя их грамотные версии содержатся в словаре. Транслитерированный твит «*Билайн. Дисконнектинг пипл.»* содержит слова, имеющие яркую негативную окраску на английском языке, но для русскоязычного словаря сбор таких лексических единиц представляет сложную задачу. Заметим, что орфографические ошибки способствовали меньшему числу неверных классификаций, чем слова

удлинённые («*ненавижуууу*», «*ураа*»), слова транслитерированные и те ненормативные, в которых часть букв заменялась автором твита на символ «***» или другие специальные символы.

Хэштеги *#отстойсвязь*, *#мтсумри*, *#люблюего* содержат тональную окраску, которая остаётся нераспознанной. Около 8% ошибочных классификаций твитов о телекоммуникациях могло быть исключено путём разделения хэштегов на отдельные слова.

Четвёртый тип ошибок – твиты о событиях, нейтральных для репутации компании, которые, однако, написаны крайне эмоционально (это могут быть твиты о дресс-коде компании, флирте с работником компании, исповедь о потерянной карточке, дружеская беседа). *Похожая ситуация с твитами о празднике или мероприятии, проводимом в офисе компании. «Матч штаб-квартиры Вымпелком – Сибирь. Пока ведем!!! :)»* Во всех приведённых случаях инструкция предписывает размечать твиты как нейтральные. Это стоило классификатору 14.9% ошибок в области телекоммуникаций, 11% – в предметной области банков.

4.2. Результаты и эксперименты в рамках задачи мнений пользователей

В таблицах 5а и 5б приведены результаты экспериментов по задаче определения тональности относительно аспектов. В рамках данной задачи макро F-мера высчитывается как среднее между F-мерами позитивного класса, негативного класса и класса двойственной тональности (без нейтрального класса). В таблицах 5а и 5б классификаторы на основе обучения с учителем показывают вторые результаты по значению макро F-меры со значительным улучшениями классификации относительно baseline методов. Классификатор на основе метода максимальной энтропии показал улучшение над базовыми результатами 14.1% и 13.5% по макро F-мере в рамках задачи для ресторанов и автомобилей соответственно.

Таблица 5а. Метрики качества в задаче классификации аспектов (C), отзывы о ресторанах

Участник	Micro P,%	Micro R,%	Micro F,%	Macro P,%	Macro R,%	Macro F,%
Официальный baseline	71.04	71.04	71.04	32.09	25.06	26.71
1_1	61.94	61.94	61.94	25.17	24.54	23.79
1_2	61.94	61.94	61.94	25.17	24.54	23.79
3_1	66.96	66.96	66.96	32.23	24.30	26.96
4_1	**82.49**	**82.49**	**82.49**	58.72	**55.69**	**55.45**
Предложенный подход	76.71	76.71	76.71	**45.82**	37.29	40.81

Таблица 5б. Метрики качества в задаче классификации аспектов (C), отзывы об автомобилях

Участник	Micro P,%	Micro R,%	Micro F,%	Macro P,%	Macro R,%	Macro F,%
Офиц. baseline	61.92	61.92	61.92	29.49	26.85	26.48
1_1	64.71	64.71	64.71	33.99	31.94	32.93
1_2	65.31	65.31	65.31	35.63	32.97	34.22
3_1	55.89	55.89	55.89	30.16	26.21	27.94
4_1	**74.28**	**74.28**	**74.28**	**57.25**	**56.67**	**56.84**
1_3	62.52	62.52	62.52	35.07	32.62	33.45
Предложенный подход	71.11	71.11	71.11	44.81	37.61	40.01

В Таблице 6 приведены результаты предложенного в статье метода, результаты метода, занявшего второе место по макро F-мере, а также официальные базовые результаты в рамках задачи категоризации явных аспектов по аспектным категориям. Предложенный метод показал наилучшие результаты среди 4 систем в обеих предметных областях. Лучший подход даёт улучшения над базовыми результатами 6,57% и 8,85% по макро F-мере для ресторанов и автомобилей, соответственно.

Таблица 6. Метрики качества в задаче категоризации аспектов (D)

Участник	Рестораны			Автомобили		
	Macro P,%	Macro R,%	Macro F,%	Macro P,%	Macro R,%	Macro F,%
Предложенный подход	**89.60**	**84.14**	**86.53**	**68.54**	**63.55**	**65.21**
Второе место	86.27	79.63	81.10	71.46	57.50	60.77
Офиц. baseline	87.42	77.37	79.96	66.72	51.90	56.36

Результаты классификации тональных высказываний относительно аспектов на основе различных наборов признаков представлены в Таблице 7 по следующей схеме: каждый эксперимент по классификации проведен на всем наборе признаков, за исключением признака, указанного в строке таблицы. Самым эффективным признаком показали себя аспектные биграммы, которые являются комбинацией аспектного термина и слова из контекстного интервала.

Таблица 7. Результаты экспериментов с признаками определения тональности относительно явных аспектов (C)

Набор признаков	Рестораны			Автомобили		
	Macro P,%	Macro R,%	Macro F,%	Macro P,%	Macro R,%	Macro F,%
Все признаки	45.82	**37.29**	**40.81**	**44.81**	**37.61**	**40.01**
Без n-грамм символов	44.79	36.59	40.00	44.80	37.50	39.94
Без словарных униграмм	42.59	36.51	39.21	**42.13**	**36.69**	**38.69**
Без аспектных биграмм	**42.61**	**33.96**	**37.28**	43.80	37.46	39.51
Без контекстных n-грамм	43.55	35.86	39.06	43.70	37.17	39.41
Без словарных score	**46.29**	36.81	40.50	43.74	37.47	39.59

Эксперименты в рамках задачи категоризации явных аспектов по аспектным категориям описаны в Таблице 8. Как показывают результаты, самыми важными признаками являются признаки, основанные на поточечной взаимной информации для категорий и включающие в себя максимальную и минимальную оценки, среднее значение оценок и сумму оценок контекста для аспектного термина.

Таблица 8. Результаты экспериментов с признаками для категоризации аспектов (D)

Набор признаков	Рестораны			Автомобили		
	P,%	R,%	F,%	P,%	R,%	F,%
n-граммы слов	76.50	71.93	73.88	65.54	60.60	62.19
n-граммы слов + единая кумулятивная оценка	81.85	77.05	79.14	68.00	62.96	64.61
n-граммы слов + все оценки	**89.60**	**84.14**	**86.53**	**68.54**	**63.55**	**65.21**

ЗАКЛЮЧЕНИЕ

В статье описаны методы анализа тональности текстов отзывов и коротких сообщений (твитов), приведены результаты оценки их качества, которая производилась в рамках российского семинара SentiRuEval-2015. Предложенные классификаторы на основе метода опорных векторов показали четвертый и второй результаты среди 10 и 7 систем в «дорожке» по задаче классификации твитов о банках и телекоммуникациях соответственно. В рамках задачи анализа мнений пользователей о ресторанах и машинах в статье описаны два метода машинного обучения: (i) метод определения тональности о конкретном объекте (аспекте), упоминание которого содержится в отзыве пользователя; (ii) метод категоризации аспектов по категориям. Классификатор на основе метода максимальной энтропии показал улучшение над базовыми результатами 14.1% и 13.5% по макро F-мере в рамках задачи определения тональности для ресторанов и автомобилей соответственно. Классификатор на основе метода опорных векторов в рамках задачи категоризации показал наилучшие результаты среди других систем, участвующих в «дорожке».

Результаты экспериментов по анализу мнений относительного значимости признаков в задаче классификации аспектов показали, что наиболее эффективными признаками являются аспектные биграммы, которые являются комбинацией аспектного термина и слова из контекстного интервала. В классификации твитов лучшие результаты показывают классификаторы, использующие следующие признаки: n-граммы символов, словари и набор символов, изображающие эмоцию.

В статье приведены анализ и классификация ошибок классификации твитов, среди которых наиболее частотными являются орфографические ошибки, транслитерация, наличие хэштегов, а также эмоциональное обсуждение нейтральных тем в твиттере.

СПИСОК ЛИТЕРАТУРЫ

1. *Nakov P., Kozareva Z., Ritter A., Rosenthal S., Stoyanov V., Wilson T.* Semeval-2013 Task 2: sentiment analysis in Twitter // Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013). 2013. P. 312-320.

2. *Rosenthal S., Ritter A., Nakov P., Stoyanov V.* SemEval-2014 Task 9: sentiment analysis in Twitter // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 73-80.

3. *Chetviorkin I.I., Braslavski P.I., Loukachevitch N.V*. Sentiment analysis track at ROMIP 2011 //Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». 2011. С. 739-746.

4. *Loukachevitch N.V., Blinov P.D., Kotelnikov E.V., Rubtsova Yu.V., Ivanov V.V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». 2015. Вып. 14 (21). С. 2-13.

5. *Popescu A.M., Etzioni O.* Extracting product features and opinions from reviews // Natural language processing and text mining. ACL, 2007. P. 9-28.

6. *Jakob N., Gurevych I.* Extracting opinion targets in a single-and crossdomain setting with conditional random fields // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. ACL, 2010. P. 1035-1045.

7. *Kiritchenko S., Zhu X., Cherry C., Mohammad S.M.* NRC-Canada-2014: Detecting aspects and sentiment in customer reviews // SemEval 2014. NAACL, 2014. P. 437-442.

8. *Chernyshevich M.* IHS R&D Belarus: cross-domain Extraction of product features using conditional random fields // SemEval 2014. Dublin, 2014. P. 309-313.

9. *Moghaddam S., Ester M.* On the design of LDA models for aspect-based opinion mining // Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, 2012. P. 803-812.

10. *Zhao Y., Qin B., Liu T.* Clustering product aspects using two effective aspect relations for opinion mining // Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer International Publishing, 2014. P. 120-130.

11. *Poria S., Cambria E., Ku L. W., Gui C., Gelbukh A.* A rule-based approach to aspect extraction from product reviews // Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP). 2014. P. 28-37.

12. *Turney P.D.* Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th Annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002. P. 417-424.

13. *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-Volume 10. – Association for Computational Linguistics, 2002. P. 79-86.

14. *Pang B., Lee L.* Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008. V. 2, No 1-2. P. 1-135.

15. *Blinov P., Klekovkina M., Kotelnikov E., Pestov O.* Research of lexical approach and learning methods for sentiment analysis // Computational Linguistics and Intellectual Technologies. 2013. No 2 (12). P. 48-58.

16. *Lu B., Ott M., Cardie C., Tsou B.K.* Multi-aspect sentiment analysis with topic models // Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference. IEEE, 2011. P. 81-88.

17. *Pontiki M., Papageorgiou H., Galanis D., Androutsopoulos I., Pavlopoulos J., Manandhar S.* Semeval-2014 task 4: aspect based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 27-35.

18. *Go A., Bhayani R., Huang L.* Twitter sentiment classification using distant supervision // CS224N Project Report, Stanford. 2009. V. 1. P. 12.

19. *Sidorov G., Miranda-Jimenez S., Viveros-Jimenez F., Gelbukh A., Castro-Sanchez N., Velasquez F., Gordon J.* Empirical study of machine learning based approach for opinion mining in tweets // Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2013. P. 1-14.

20. *Sakaki T., Okazaki M., Matsuo Y.* Earthquake shakes Twitter users: real-time event detection by social sensors // Proceedings of the 19th International Conference on World Wide Web. ACM, 2010. P. 851-860.

21. *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.

22. *Davidov D., Tsur O., Rappoport A.* Semi-supervised recognition of sarcastic sentences in twitter and amazon // Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2010. P. 107-116.

23. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* Lexicon-based methods for sentiment analysis // Computational Linguistics. 2011. V. 37, No 2. P. 267-307.

24. *O'Connor B., Balasubramanyan R., Routledge B.R., Smith N.A.* From tweets to polls: linking text sentiment to public opinion time series // ICWSM. 2010. V. 11. P. 122-129.

25. *Mohammad S.M., Kiritchenko S., Zhu X.* NRC-Canada: building the state-of-the-art in sentiment analysis of tweets // Second Joint Conference on Lexical and Computational Semantics (* SEM). 2013. V. 2. P. 321-327.

26. *Evert S., Proisl T., Greiner P., Kabashi B.* SentiKLUE: updating a polarity classifier in 48 hours // SemEval 2014. 2014. P. 551.

27. *Barbosa L., Feng J.* Robust sentiment detection on twitter from biased and noisy data // Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010. P. 36-44.

28. *Martinez-Camara E., Martin-Valdivia M.T., Urena-Lopez L.A., Montejo-Raez A.R.* Sentiment analysis in twitter // Natural Language Engineering. 2014. V. 20, No 01. P. 1-28.

29. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in Russian // ACL 2013. 2013. P. 12-17.

30. *Frolov A.V., Polyakov P.Yu., Pleshko V.V.* Using semantic filters in application to book reviews sentiment analysis // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». 2013. No 12 (19).

SENTIMENT CLASSIFICATION OF REVIEWS AND TWITTER POSTS IN RUSSIAN BASED ON DICTIONARIES

E.V. Tutubalina[1], V.V. Ivanov[1], M.A. Zagulova[1], N.R. Mingazov[1], I.S. Alimova[1], V.A. Malykh[2]

elvtutubalina@kpfu.ru, vivanov@kpfu.ru, lolmariya@gmail.com, nicrotek547@gmail.com, alimovaIlseyar@gmail.com, alem.mipt@gmail.com

[1] High School of Information Technology and Information Systems of Kazan Federal University

[2]Institute for Systems Analysis of Russian Academy of Sciences

Abstract

Sentiment analysis and opinion mining technologies are growing fast. This is mostly due to a rapid grow of the data sources consisting a vast amount of user opinions and reviews on a wide set of topics. In this paper we describe methods for sentiment analysis of reviews and short messages (tweets), as well as evaluation of results obtained during SentiRuEval-2015.

Keywords: *information extraction, sentiment analysis, text classification, supervised learning*

REFERENCES

1. *Nakov P., Kozareva Z., Ritter A., Rosenthal S., Stoyanov V., Wilson T.* Semeval-2013 Task 2: sentiment analysis in Twitter // Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013). 2013. P. 312-320.

2. *Rosenthal S., Ritter A., Nakov P., Stoyanov V.* SemEval-2014 Task 9: sentiment analysis in Twitter // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 73-80.

3. *Chetviorkin I.I., Braslavski P.I., Loukachevitch N.V*. Sentiment analysis track at ROMIP 2011 // Kompyuternaya lingvistika i intellektualnyie tehnologii: po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog». 2011. S. 739-746.

4. *Loukachevitch N.V., Blinov P.D., Kotelnikov E.V., Rubtsova Yu.V., Ivanov V.V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in russian // Kompyuternaya lingvistika i intellektualnyie tehnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog». 2015. Vyip. 14 (21). S. 2-13.

5. *Popescu A.M., Etzioni O.* Extracting product features and opinions from reviews // Natural language processing and text mining. ACL, 2007. P. 9-28.

6. *Jakob N., Gurevych I.* Extracting opinion targets in a single-and crossdomain setting with conditional random fields // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. ACL, 2010. P. 1035-1045.

7. *Kiritchenko S., Zhu X., Cherry C., Mohammad S.M.* NRC-Canada-2014: Detecting aspects and sentiment in customer reviews // SemEval 2014. NAACL, 2014. P. 437-442.

8. *Chernyshevich M.* IHS R&D Belarus: cross-domain Extraction of product features using conditional random fields // SemEval 2014. Dublin, 2014. P. 309-313.

9. *Moghaddam S., Ester M.* On the design of LDA models for aspect-based opinion mining //Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, 2012. P. 803-812.

10. *Zhao Y., Qin B., Liu T.* Clustering product aspects using two effective aspect relations for opinion mining // Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer International Publishing, 2014. P. 120-130.

11. *Poria S., Cambria E., Ku L. W., Gui C., Gelbukh A.* A rule-based approach to aspect extraction from product reviews // Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP). 2014. P. 28-37.

12. *Turney P.D.* Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th Annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002. P. 417-424.

13. *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-Volume 10. – Association for Computational Linguistics, 2002. P. 79-86.

14. *Pang B., Lee L.* Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008. V. 2, No 1-2. P. 1-135.

15. *Blinov P., Klekovkina M., Kotelnikov E., Pestov O.* Research of lexical approach and learning methods for sentiment analysis // Computational Linguistics and Intellectual Technologies. 2013. No 2 (12). P. 48-58.

16. *Lu B., Ott M., Cardie C., Tsou B.K.* Multi-aspect sentiment analysis with topic models // Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference. IEEE, 2011. P. 81-88.

17. *Pontiki M., Papageorgiou H., Galanis D., Androutsopoulos I., Pavlopoulos J., Manandhar S.* Semeval-2014 task 4: aspect based sentiment analysis // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014. P. 27-35.

18. *Go A., Bhayani R., Huang L.* Twitter sentiment classification using distant supervision // CS224N Project Report, Stanford. 2009. V. 1. P. 12.

19. *Sidorov G., Miranda-Jimenez S., Viveros-Jimenez F., Gelbukh A., Castro-Sanchez N., Velasquez F., Gordon J.* Empirical study of machine learning based approach for opinion mining in tweets // Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2013. P. 1-14.

20. Sakaki T., Okazaki M., Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors // Proceedings of the 19th International Conference on World Wide Web. ACM, 2010. P. 851-860.

21. *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.

22. *Davidov D., Tsur O., Rappoport A.* Semi-supervised recognition of sarcastic sentences in twitter and amazon // Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2010. P. 107-116.

23. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* Lexicon-based methods for sentiment analysis // Computational Linguistics. 2011. V. 37, No 2. P. 267-307.

24. *O'Connor B., Balasubramanyan R., Routledge B.R., Smith N.A.* From tweets to polls: linking text sentiment to public opinion time series // ICWSM. 2010. V. 11. P. 122-129.

25. *Mohammad S.M., Kiritchenko S., Zhu X.* NRC-Canada: building the state-of-the-art in sentiment analysis of tweets // Second Joint Conference on Lexical and Computational Semantics (* SEM). 2013. V. 2. P. 321-327.

26. *Evert S., Proisl T., Greiner P., Kabashi B.* SentiKLUE: updating a polarity classifier in 48 hours // SemEval 2014. 2014. P. 551.

27. *Barbosa L., Feng J.* Robust sentiment detection on twitter from biased and noisy data // Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010. P. 36-44.

28. *Martinez-Camara E., Martin-Valdivia M.T., Urena-Lopez L.A., Montejo-Raez A.R.* Sentiment analysis in twitter // Natural Language Engineering. 2014. V. 20, No 01. P. 1-28.

29. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in Russian // ACL 2013. 2013. P. 12-17.

30. *Frolov A.V., Polyakov P.Yu., Pleshko V.V.* Using semantic filters in application to book reviews sentiment analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog". 2013. No 12 (19).

СВЕДЕНИЯ ОБ АВТОРАХ

ТУТУБАЛИНА Елена Викторовна – аспирант Высшей школы информационных технологий и информационных систем Казанского федерального университета.

Elena Viktorovna TUTUBALINA received Diploma in applied mathematics and informatics from Kazan Federal University (2012). She is a graduate student at the High School of Information Technology and Information Systems of Kazan Federal University. Current scientific interests: natural language processing, opinion mining, topic modeling.

email: ElVTutubalina@kpfu.ru

ИВАНОВ Владимир Владимирович – старший преподаватель кафедры интеллектуальных технологий поиска Высшей школы информационных технологий и информационных систем Казанского федерального университета.

Vladimir Vladimirovich IVANOV is a head of Big Data and Textual Analysis Lab at the High School of Information Technology and Information Systems of Kazan Federal University. Current scientific interests: natural language processing, information extraction.

email: vivanov@kpfu.ru

ЗАГУЛОВА Мария – студент Высшей школы информационных технологий и информационных систем Казанского федерального университета.

Maria ZAGULOVA is a student at the High School of Information Technology and Information Systems of Kazan Federal University. Current scientific interests: sentiment analysis.

email: lolmariya@gmail.com

МИНГАЗОВ Никита – лаборант-исследователь НИЛ «Большие данные и анализ текста» Высшей школы информационных технологий и информационных систем Казанского федерального университета.

Nikita MINGAZOV is an assistant engineer at the Big Data and Textual Analysis Lab of High School of Information Technology and Information Systems of Kazan Federal University. Current scientific interests: information retrieval.

email: nicrotek547@gmail.com

АЛИМОВА Ильсеяр – аспирант Высшей школы информационных технологий и информационных систем Казанского федерального университета.

Ilseyar Alimova received Diploma in applied mathematics and informatics from Kazan Federal University (2014). She is a graduate student at the High School of Information Technology and Information Systems of Kazan Federal University. Current scientific interests: time series.

email: alimovaIlseyar@gmail.com

МАЛЫХ Валентин – аспирант Института системного анализа РАН.

Valentin MALYKH is a student at the Institute for Systems Analysis of Russian Academy of Sciences. Current scientific interests: natural language processing.

email: alem.mipt@gmail.com

Материал поступил в редакцию 15 июля 2015 года

УДК 004.912

ИСПОЛЬЗОВАНИЕ СИНТАКСИСА ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТВИТОВ НА РУССКОМ ЯЗЫКЕ

Ю.В. Адаскина[1], П.В. Паничева[2], А.М. Попов[3]

ООО «InfoQubes», Санкт-Петербургский государственный университет
[1]adaskina@gmail.com, [2]p.panicheva@spbu.ru, [3]hedgeonline@gmail.com

Аннотация

Представлен подход к решению задачи анализа тональности в рамках тестирования SentiRuEval – открытого соревнования систем анализа тональности на русском языке. Описанный алгоритм был применен в дорожке по анализу тональности твитов о банках и телекоммуникационных компаниях. Для этих данных была разработана и оценена классификация на три класса: положительный, отрицательный и нейтральный.

Для решения поставленной задачи использовались различные алгоритмы машинного обучения. Признаками для классификатора являлись лингвистические данные, полученные из текста с помощью разработанного нами морфо-синтаксического анализатора. Нормализованные слова, а также синтаксические связи, оказались решающими признаками для достижения наилучшего результата, который был получен с помощью статистического алгоритма опорных векторов.

Оценка, проведенная организаторами конкурса, выявила высокое качество предложенного подхода, который занял первую строчку по трем из четырех мерам качества.

Ключевые слова*: анализ тональности, синтаксические связи, русский язык, статистические методы, классификация текстов.*

ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

Будучи одним из наиболее изученных направлений прикладной лингвистики, анализ тональности остается одной из самых востребованных задач как для теоретических исследований, так и для бизнес-приложений. Анализ тональности применялся на разных уровнях, начиная от документа целиком и постепенно сужаясь к отдельному предложению. Тональность на уровне предложений распознается, исходя из предположения, что одним предложением в языке обычно выражается одно мнение. В последнее время основной фокус сдвинулся на более мелкие единицы внутри предложения, в сферу анализа попадают случаи, когда в предложении есть оценка нескольких сходных объектов (например, нескольких брендов), а также случаи оценки разных аспектов одного и того же объекта (например, таких параметров товара, как прочность, цена, дизайн и т. п.). Основные усилия лингвистов сегодня направлены на создание и развитие высокоточных автоматических методов анализа тональности, что в свою очередь поднимает вопрос о методах оценки качества таких систем. Многие независимые организации проводят тестирования различных методов автоматического анализа естественного языка, самым влиятельным среди российских можно считать соревнование Dialogue Evaluation, проводимое в рамках международной конференции по компьютерной лингвистике «Диалог». В 2015 году состоялось третье тестирование систем анализа тональности SentiRuEval; первые два обсуждаются в [1, 2]. В этом году оценивалось в том числе предметно-ориентированное распознавание тональности на различных типах данных (см. [3]).

В данной статье описан подход к заданиям SentiRuEval, а именно, в двух дорожках, посвященных объектно-ориентированному анализу мнений в твитах о банках и телекоммуникационных компаниях. От участников этих дорожек требовалось произвести трехклассовую классификацию тестовых данных, разделив их на негативный, позитивный и нейтральный классы.

Полученные результаты основаны на классификаторе SVM, хотя предварительные эксперименты показали незначительные различия между ним и классификатором Naïve Bayes. В качестве признаков для обучения использовались нормализованные формы слов в комбинации с синтаксическими связями, где под последними понимаются тройки из нормальных форм двух связанных слов и типа

синтаксического отношения между ними. Для всех предварительных экспериментов добавление синтаксических связей существенно улучшало результаты классификации; в оценке, использованной организаторами соревнования, влияние синтаксиса было чуть менее значительно. Тем не менее, результаты соревнования подтвердили эффективность разработанного метода: по трем из четырех метрик качества созданной системе удалось занять первое место среди участников.

ПРЕДПОСЫЛКИ ИССЛЕДОВАНИЯ

Одним из наиболее распространенных подходов к решению задач анализа тональности является применение машинного обучения. Надо отметить, что анализ тональности хорошо ложится на такую стандартную задачу, часто решаемую с привлечением машинного обучения, как классификация, где документ классифицируется по трем классам тональности: положительному, отрицательному и нейтральному. С одной стороны, машинное обучение, как вероятностный метод, позволяет свести к минимуму лингвистическую составляющую, сохраняя при этом относительно высокие показатели качества [4]. С другой стороны, необходимым условием применения любых алгоритмов машинного обучения является потребность в параметризации обучающих и анализируемых данных. Обычно текст параметризуется как «мешок слов» [4], реже – как n-грамм. В некоторых случаях такой подход оправдывает себя, однако для языков с развитой морфологией число таких признаков может быть очень велико, а их абсолютная встречаемость, наоборот, очень низкой. Это будет препятствовать любым попыткам обобщения, которые проводят алгоритмы машинного обучения. Зачастую, чтобы, с одной стороны, уменьшить число признаков, а с другой, — повысить их встречаемость, применяются различные лингвистические приемы, например, приведение слов к нормальной форме [5], добавление в признаки семантической [6] или синтаксической [7] информации и т. д.

Одним из наиболее известных исследований по анализу тональности с использованием синтаксической информации является работа [8]. В ней описано применение SVM-классификатора для анализа тональности с использованием различных признаков, в том числе, лемматизации и синтаксических поддеревьев. В работе [9] описано использование синтаксической информации в системе ана-

лиза тональности текстов на русском языке. Авторы применяют подход, основанный исключительно на правилах, при котором текст рассматривается не как «мешок слов», а как набор синтаксических деревьев. Данный метод позволяет проводить так называемый «объектно-ориентированный» анализ тональности, когда мнение высказывается относительно какого-то объекта в тексте. В работе [10] используются различные алгоритмы машинного обучения (SVM, Naïve Bayes) для анализа тональности; исследуется влияние лемматизации и применения другого вида лингвистических ресурсов, словарей синонимов, на качество анализа тональности. В частности, сделан вывод о том, что для русского языка лемматизация и словари синонимов оказывают положительное влияние на качество.

Таким образом, список наиболее часто используемых признаков для машинного обучения выглядит следующим образом:

- словоформы (униграммы);
- леммы (нормализованные униграммы);
- n-граммы;
- нормализованные n-граммы;
- бинарная встречаемость слов;
- синтаксические связи/поддеревья.

Следует отметить, что использование синтаксических признаков само по себе подразумевает сложную и длительную процедуру синтаксического анализа. Однако исследования, в которых применяются синтаксические признаки, показывают, что синтаксическая информация позволяет существенно повысить как полноту, так и точность (см., например, [11–13]) алгоритмов классификации текстов. Так, в работе [14], посвященной задаче автоматического извлечения контекста, синтаксические признаки оказывают решающий вклад в достижение F-меры в 70%.

В заключение обзора предпосылок перечислим несколько работ, посвященных анализу тональности на материале твитов — сообщений, представляющих собой отдельный подтип данных [15–19]. Особенности Твиттера — ограничение на длину сообщения и ориентация на жанр мгновенных реакций на происходящее – сказываются на особенностях методов их анализа.

В задачи данного исследования входил поиск ответа на вопрос, как применение синтаксической информации в качестве признаков для машинного обучения повлияет на качество анализа тональности текстов на русском языке.

ДАННЫЕ И ПОСТАНОВКА ЗАДАЧИ

Наша компания принимала участие в дорожках по анализу тональности в твитах, посвященных банкам и телекоммуникационным компаниям. Детальное описание заданий представлено в [3]. Организаторы предоставили участникам обучающие и тестовые выборки размером около 10 тысяч текстов каждая; обе текстовые коллекции, в свою очередь, делились примерно пополам на твиты о двух типах брендов (банки и телекоммуникационные компании). Обучающие данные были вручную размечены экспертами SentiRuEval, каждому тексту было проставлено значение тональности или помечено его отсутствие. Твиты, для которых не было согласия в оценках хотя бы у двух из трех экспертов, исключались из корпуса, в результате чего размер обучающей коллекции для банков составил 4549 документов, для телекоммуникационных компаний – 3845 документов. Тестовый корпус был размечен нейтральными значениями для каждой из компаний, упомянутых в твите, от участников требовалось заменить эти значения на положительные или отрицательные, или же сохранить нейтральное.

АЛГОРИТМ

В основе метода лежит машинное обучение с использованием различных признаков, полученных нашим лингвистическим модулем. Остановимся на этих аспектах подробнее.

Модуль лингвистического анализа

Для анализа текстов использовался морфосинтаксический парсер InfoQubes, который ранее показал свою эффективность для решения задачи полуавтоматического пополнения лексических классов (см. [20]). Эта платформа является коммерческой разработкой нашей компании. Анализатор состоит из нескольких модулей, среди них важно отметить модули: морфологического анализа; распознавания неизвестных слов и слов с опечатками; поверхностного синтаксиса; основного синтаксиса; пост-синтаксической обработки.

Морфологический модуль (модуль приведения слов к нормальным формам) основан на словоизменительном словаре А.А. Зализняка [21], этот модуль осуществляет лемматизацию и приписывает словоформам наборы значений грамматических категорий. Модуль распознавания неизвестных слов и слов с опечатками анализирует фрагменты текста, которые отсутствуют в морфологических словарях. На основе выделения суффиксов и приставок, а также степени схожести неизвестных слов со словами, имеющимися в словарях, модуль может приписать неизвестному слову грамматические значения. Такая возможность играет особую роль при работе с данными из социальных сетей, особенно короткими текстами Твиттера, которые зачастую пишутся в спешке, что увеличивает вероятность появления опечаток.

Модуль поверхностного синтаксиса собирает основные фразовые категории: имена существительные, имена прилагательные, глаголы и их зависимые. Кроме того, здесь реализованы некоторые вспомогательные функции, например, распознавание именованных сущностей; частично именно этот модуль проставляет маркер отрицания.

Синтаксический модуль представляет собой конечный автомат, который на вход получает текст, обработанный морфологически и поверхностно-синтаксически, а на выходе возвращает синтаксическое дерево. В качестве входной контекстно-свободной грамматики для парсера используется сложная система из 515 синтаксических правил. Обычно синтаксическое правило соединяет два слова или фразовые категории в категорию более высокого уровня и проставляет синтаксическое отношение. Таким образом, из грамматики непосредственных составляющих выводится структура зависимостей. В грамматике разрешены только бинарные связи, каждое синтаксическое отношение характеризуется исходным словом, целевым словом и типом связи между ними. В системе используется 16 синтаксических связей, одна из которых имеет 4 разновидности, поэтому в нашем признаковом пространстве рассматривается как 4 различных типа связи. В Таблице 1 представлены частоты встречаемости 19 типов синтаксических связей в обучающих корпусах дорожек по оценке тональности:

Таблица 1. Синтаксические связи, полученные системой на обучающем корпусе

Название	**Встречаемость в корпусе о телекоме**	**Встречаемость в корпусе о банках**
Argument:DirectObject	2778	2372
Argument:IndirectObject	5748	3585
Argument:PassiveSubject	291	232
Argument:Subject	3148	1805
Attribute	6814	6682
Auxiliary	578	208
Circumstance	3033	1211
Coordinate	1008	1698
Determiner	687	239
Genitive	3963	3355
Identity	2200	4937
Infinitive	772	465
Modifier	707	294
Phrasal	1519	959
Possessive	368	126
Preposition	6582	4554
Quantifier	501	605
Subordinate	226	77
Undefined	1050	1159

Модуль пост-синтаксической обработки анализирует собранные поддеревья и может по необходимости редактировать узлы или связи между ними. На этом этапе устанавливаются недостающие синтаксические связи, например, во фразовых категориях, которые были собраны модулем поверхностного синтаксиса. Кроме того, этот модуль проставляет маркер отрицания и некоторые семантические теги.

Настройка признаков машинного обучения

Эксперименты проводились с единичными леммами (униграммами), сочетаниями лемм (биграммами) и синтаксическими связями в качестве признаков для классификаторов на основе опорных векторов и наивного байесовского (см.

[22]), с использованием трехклассовой классификации (нейтральный, позитивный и негативный классы). В каждом из экспериментов были использованы правила нормализации морфологического модуля. Так как сообщения в Твиттере характеризуются ограниченной длиной, ожидалось, что построение полных деревьев синтаксического разбора будет затруднено. Поэтому в признаковое пространство были включены синтаксические связи как пары связанных слов и тип отношения между ними, иными словами, синтаксические связи представляются как тройки «главное слово – тип связи – зависимое слово». В качестве опциональных параметров также использовалось отрицание, проставляемое нашим морфологическим анализатором: маркер, который получает слово, связанное с одной из отрицательных единиц (в первую очередь, частица «не», кроме того, предлог «без», существительное «отсутствие» и др.). Кроме того, в качестве параметра опционально исключались слова, обозначающие один из исследуемых брендов, так как подразумевалось, что общая направленность на бренды может отрицательно повлиять на результаты. Все использованные признаков и опциональные параметры представлены в Таблице 2.

Таблица 2. Характеристики признаков

№	Пример признака	Тип признака	Опциональные параметры	Расшифровка	Комментарий
1	ВАРИАНТ	Лемма	Маркер отрицания не учитывается	Лемма *ВАРИАНТ*	Нормализованное слово
2	ВАРИАНТ|Argument|НЕТ|PassiveSubject	Синтаксическая связь	Маркер отрицания не учитывается	Связь «субъект в пассивной конструкции» *ВАРИАНТА НЕТ*	Определенный тип синтаксической связи между двумя словами (в данном случае с подтипом, так как связь «аргумент» имеет 4 разновидности)
3	ВАРИАНТ|Attribute|ЭТОТ|	Синтаксическая связь	Маркер отрицания не учитывается	Связь «атрибут» *ЭТОТ*	Определенный тип синтаксической связи между двумя словами

				ВАРИАНТ, отрицание отсутствует	
4	КРУТОЙ|ВАРИАНТ	Биграмма	Маркер отрицания не учитывается	Биграмма *КРУТОЙ ВАРИАНТ*	Два смежных слова
5	ДРУГОЙ|ВАРИАНТ	Биграмма	Маркер отрицания не учитывается	Биграмма *ДРУГОЙ ВАРИАНТ*	Два смежных слова
6	ВАРИАНТ|0	Лемма	Маркер отрицания учитывается	Лемма *ВАРИАНТ*, на обоих словах нет отрицания	Сочетание нормализованных слов с информацией об отрицании, в данном случае отрицание отсутствует
7	ВАРИАНТ|1	Лемма	Маркер отрицания учитывается	Лемма *ВАРИАНТ*, слово с отрицанием	Сочетание нормализованных слов с информацией об отрицании, в данном случае отрицание присутствует на одном из слов
8	ВАРИАНТ|1|Argument|НЕТ|0|PassiveSubject	Синтаксическая связь	Маркер отрицания учитывается	Связь «субъект в пассивной конструкции» *ВАРИАНТА НЕТ*, слово *ВАРИАНТ* с отрицанием	Сочетание синтаксической связи с информацией об отрицании, в данном случае отрицание присутствует на одном из слов
9	ВАРИАНТ|0|Attribute|ЭТОТ|0|	Синтаксическая связь	Маркер отрицания учитывается	Связь «атрибут» *ЭТОТ ВАРИАНТ*, на обоих словах нет отрицания	Сочетание синтаксической связи с информацией об отрицании, в данном случае отрицание отсутствует

10	КРУТОЙ|0|ВАРИАНТ|0	Биграмма	Маркер отрицания учитывается	Биграмма *КРУТОЙ ВАРИАНТ*, на обоих словах нет отрицания	Сочетание биграммы с информацией об отрицании, в данном случае отрицание отсутствует
11	ДРУГОЙ|0|ВАРИАНТ|1	Биграмма	Маркер отрицания учитывается	Биграмма *ДРУГОЙ ВАРИАНТ*, слово *ВАРИАНТ* с отрицанием	Сочетание биграммы с информацией об отрицании, в данном случае отрицание присутствует на одном из слов

Отметим также, что организаторы ставили перед участниками задачу связывать оценку, содержащуюся в твите, с брендом, к которому она относится. Были проанализированы документы обучающего корпуса, в которых содержатся несовпадающие значения тональности, и выявлено крайне малое их количество: менее 1% для корпусов обеих тематик. Поэтому было принято решение пренебречь этими документам и упростить модель данных, распространяя найденную в документе тональность на все бренды, содержащиеся в тексте.

РЕЗУЛЬТАТЫ ПРЕДВАРИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Предварительные эксперименты были проведены с применением десятикратной кросс-валидации на обучающей текстовой коллекции. Описанный выше алгоритм анализа тональности был совмещен с алгоритмом извлечения названий брендов, основанном на правилах. Для того чтобы оценить результаты, каждому документу были сопоставлены его идентификационный номер, идентификатор бренда и значение тональности. На основании этой информации и предоставленной разметки была подсчитана общая Полнота, Точность и F1-мера. В расчетах учитывался нейтральный класс, а также проверялось качество определения бренда, что отличает полученные нами метрики от метрик, использованных организаторами. Оценки, полученные на основе предварительного эксперимента, представлены в следующих таблицах, наивысший результат выделен жирным

шрифтом. Таблица 3 относится к корпусу твитов о телекоммуникационных компаниях, Таблица 4 – о банках.

Таблица 3. Предварительные результаты для данных «Телеком», SVM

Признаки	**Опциональные параметры эксперимента**		**Оценки**		
	Маркер отрицания	**Удаление названия бренда**	**Точность**	**Полнота**	**F1-мера**
Леммы	–	–	0,7464	0,7482	0,7473
	+	–	0,7549	0,7567	0,7558
	–	+	0,7554	0,7571	0,7563
	+	+	0,7608	0,7625	0,7616
Синтаксические связи	–	–	0,7275	0,5567	0,6308
	+	–	0,7228	0,5532	0,6267
	–	+	0,7196	0,5470	0,6216
	+	+	0,7215	0,5484	0,6231
Леммы + синтаксические связи	–	–	**0,7715**	**0,7734**	**0,7725**
	+	–	0,7692	0,7710	0,7701
	–	+	0,7675	0,7692	0,7684
	+	+	0,7632	0,7648	0,7640
Леммы + синтаксические связи, χ2 распределение для 5000 лучших признаков	–	–	0,5865	0,5879	0,5872
Биграммы	–	–	0,7242	0,7077	0,7158
Биграммы + связи	–	–	0,7204	0,7220	0,7212
Биграммы + леммы	–	–	0,7650	0,7668	0,7659
Биграммы + леммы + синтаксические связи	–	–	0,7684	0,7702	0,7693

Таблица 4. Предварительные результаты для данных «Банки», SVM

Признаки	**Опциональные параметры эксперимента**		**Оценки**		
	Маркер отрицания	**Удаление названия бренда**	**Точность**	**Полнота**	**F1-мера**
Леммы	–	–	0,9046	0,9061	0,9053
	+	–	0,9021	0,9036	0,9029
	–	+	0,9073	0,9087	0,9080
	+	+	0,9032	0,9046	0,9039
Синтаксические связи	–	–	0,9040	0,8184	0,8591
	+	–	0,9080	0,8220	0,8628
	–	+	0,9040	0,8171	0,8583
	+	+	0,9066	0,8194	0,8608
Леммы + синтаксические связи	–	–	0,9059	0,9074	0,9066
	+	–	0,9047	0,9062	0,9055
	–	+	0,9083	0,9097	0,9090
	+	+	**0,9095**	**0,9108**	**0,9101**
Биграммы	–	–	0,8968	0,8949	0,8959
Биграммы + связи	–	–	0,8957	0,8971	0,8964
Биграммы + леммы	–	–	0,9021	0,9036	0,9029
Биграммы + леммы + синтаксические связи	–	–	0,9026	0,9041	0,9033
Леммы + синтаксические связи, $\chi 2$ распределение для 5000 лучших признаков	–	–	0,8257	0,8269	0,8263

Предварительные результаты показали, что комбинация лемм и синтаксических связей обеспечивает наилучшие результаты для обоих тестовых корпусов, а добавление отрицания и исключение брендов не оказывают существенного влияния на результат. Этот результат подтверждает исходную гипотезу, что синтаксические связи должны улучшить показатели. Биграммы и леммы показывают почти такие же высокие результаты, как леммы и связи. Наивный байесовский классификатор подтвердил эти тенденции с небольшим понижением абсолютных

значений показателей. Также был проведен эксперимент с исключением некоторых низкочастотных признаков, с применением алгоритма отбора признаков (feature selection), но это привело к неудовлетворительным результатам. В таблицы результатов выше были включены значения, полученные при помощи отбора признаков, и можно увидеть значительное падение показателей. Кроме того, применение меры TF-IDF также существенно ухудшило результаты. Представляется, что обучающие данные слишком рассеяны, чтобы на них могли сработать алгоритмы отбора признаков, они, возможно, были бы полезны на большем обучающем корпусе, где была бы выше частотность каждого отдельного признака.

РЕЗУЛЬТАТЫ СОРЕВНОВАНИЯ

Организаторы предоставляли участникам право отправить результаты нескольких прогонов, поэтому нами была выбрана SVM-классификация на основе биграмм и лемм в комбинации с синтаксическими связями. Также из параметрической модели опционально удалялись названия брендов. Для более полного анализа предложенных алгоритмов также был проведен внеконкурсный прогон SVM-классификатора на леммах. Таблица ниже представляет собой несколько модифицированную таблицу лучших результаты участников из обзорной статьи организаторов [3]. В нее добавлены результаты всех наших прогонов, их идентификаторы заменены на названия признаков соответствующих экспериментов в правой колонке. Номера других участников оставлены без изменения. Жирным шрифтом выделен лучший результат, курсивом – наш внеконкурсный прогон. В качестве оценочной метрики организаторы использовали две разновидности F-меры: F-микро и F-макро (подробнее см. [3]). Как видно из таблицы, на основе предложенных алгоритмов были получены лучшие результаты по трем метрикам качества из четырех.

Таблица 5. Результаты соревнования

Область	Мера	Базовый уровень	Результат	Идентификатор участника
Телеком	Macro F	0,182	**0,488**	**леммы+связи**
			0,483	леммы+связи, бренды удалены
			0,480	3
			…	…
			0,469	*леммы*
			0,465	*леммы, бренды удалены*
	Micro F	0,337	**0,536**	**леммы+связи**
			0,536	леммы+связи, бренды удалены
			0,528	10
			…	
			0,512	*леммы*
			0,514	*леммы, бренды удалены*
Банки	Macro F	0,127	**0,360**	**4**
			0,352	10
			0,345	*леммы*
			0,345	леммы, бренды удалены
			0,343	леммы+связи, бренды удалены
	Micro F	0,238	**0,366**	**леммы+связи, бренды удалены**
			0,364	леммы+связи
			0,363	*леммы*
			0,362	*леммы, бренды удалены*
			0,343	8

Результаты организаторов оценки существенно отличаются от наших предварительных результатов, что объясняется различными подходами к оценке: нами использовалась только одна из F-мер (F-микро), а также организаторы исключили из подсчета нейтральный класс документов.

Эти результаты только отчасти соотносятся с нашими предварительными результатами и нашей исходной гипотезой: на данных о телекоммуникационных компаниях леммы в комбинации с синтаксическими связями работают лучше,

чем одни леммы приблизительно на 2% в микро- и макро- F-мерах. На корпусе о банках результаты неубедительны: добавление синтаксических связей улучшает F-микро на 0,25%, но ухудшает F-макро примерно на 0,2%. В наших предварительных экспериментах результаты на корпусе о банках были выше, чем на корпусе о телекоме, а результат соревнования говорит об обратном. Показатели для банков оказались ниже показателей для телекома у всех участников.

Принятое решение не учитывать документы с упоминанием несовпадающих значений тональности оказалось удачным: в тестовом корпусе их количество тоже было минимальным.

Другие высокие результаты были получены участниками, использовавшими алгоритм, основанный на правилах, классификаторы методом максимальной энтропии и SVM на различных наборах признаков, главным образом, на словах и буквенных n-граммах.

ЗАКЛЮЧЕНИЕ

К задаче анализа тональности на двух предметных областях был применен статистический алгоритм с использованием синтаксических связей, что позволило получить высокие результаты, опередившие другие методы. Использовалась трехклассовая классификация, показатели классификации на леммах в качестве признаков были улучшены за счет добавления синтаксической информации. В некоторых случаях улучшения не происходило за счет высокой разреженности данных, этот вопрос требует дополнительного анализа и исследования. Признаки для классификации получены при помощи нашего морфосинтаксического парсера, уже продемонстрировавшего свою эффективность на другой задаче, связанной с семантикой (см. [20]). Так как число документов, включавших несовпадающие значения тональности, было очень маленьким, модель данных была упрощена до представления «один документ — одна оценка тональности». Для разреженного корпуса небольшого размера классификатор SVM представляется оптимальным методом. Добавление отрицания или удаление брендов не оказывает существенного влияния на результат: можно предположить, что вся информация, которую могут привнести показатели отрицания, уже содержится в синтаксических поддеревьях.

В качестве направлений дальнейших исследований можно отметить несколько пунктов. В процессе подготовки к соревнованию возможности нашего лингвистического анализатора были использованы не до конца, в качестве признаков можно было бы использовать другие результаты обработки, например, семантические теги, синонимические ряды, смайлы, а также более развернутые результаты синтаксического модуля. Для коротких сообщений из Твиттера можно ожидать большое количество ошибок и неполных деревьев в синтаксисе, поэтому применение пар синтаксически связанных слов кажется наиболее перспективным. Для других типов данных возможно расширение синтаксических признаков и улучшение результатов классификации за счет этого. Кроме того, для твитов эффективным может быть анализ смайлов, символов эмодзи и хештегов, что было проделано некоторыми участниками соревнования. Данные также не позволили успешно применить те или иные способы фильтрации признаков, что, как представляется, было бы возможно, если бы частотность признаков была выше, например, за счет замены конкретных слов идентификаторами семантических классов.

СПИСОК ЛИТЕРАТУРЫ

1. *Chetviorkin I., Braslavskiy P., Loukachevich N.* Sentiment analysis track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2012». 2012. P. 1-14.

2. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in Russian // Proceedings of BSNLP workshop, ACL, Prague. 2013. P. 12-17.

3. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Ju., Ivanov V., Tutubalina H.* Sentirueval: testing object-oriented sentiment analysis systems in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». 2015. Issue 14. V. 2. P. 13-24.

4. *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.

5. *Mullen T., Collier N.* Sentiment analysis using support vector machines with diverse information sources // Proceedings of 9th EMNLP. 2004. P. 412-418.

6. *Turney P.* Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th ACL. 2002. P. 417-424.

7. *Kudo T., Matsumoto Y.* A boosting algorithm for classification of semi-structured text // Proceedings of 9th EMNLP. 2004. P. 301-308.

8. *Matsumoto S., Takamura H., Okumura M.* Sentiment classification using word sub-sequences and dependency sub-trees // Ho T.-B., Cheung D., Liu H. (eds.) PAKDD 2005. V. 3518. P. 301-311.

9. *Mavljutov R.R., Ostapuk N.A.* Using basic syntactic relations for sentiment analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2013». 2013. P. 91-100.

10. *Yussupova N., Bogdanova D., Boyko M.* Applying of sentiment analysis for texts in russian based on machine learning approach // Proceedings of The Second International Conference on Advances in Information Mining and Management, Italy. 2012. P. 8-14.

11. *Furnkranz J., Mitchell T. M., Rilof E.* A case study in using linguistic phrases for text categorization on the WWW // Proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, US. 2998. P. 5-12.

12. *Caropreso M.F., Matwin S., Sebastiani F.A.* Learner-independent evaluation of the usefulness of statistical phrases for automated text categorization // Amita G. Chin (ed.), Text Databases and Document Management: Theory and Practice. 2006. P. 78-102.

13. *Nastase V., Shirabad J.S., Caropreso M.F.* Using dependency relations for text classification // Proceedings of the 19th Canadian Conference on Artificial Intelligence, Quebec City. 2006. P. 12-25.

14. *Zhao S., Grishman R.* Extracting relations with Integrated Information using kernel methods // Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, US. 2005. P. 419-426.

15. *Jansen B.J., Zhang M., Sobel K., Chowdury A.* Twitter power: tweets as electronic word of mouth // Journal of the American Society for Information Science and Technology. 2009. V. 60, No 11. P. 2169-2188.

16. *Go A., Bhayani R., Huang L.* twitter sentiment classification using distant supervision // Technical report, Stanford. 2009.

17. *Jiang L., Yu M., Zhou M., Liu X., Zhao T.* Target-dependent Twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, US. 2011. P. 151-160.

18. *Kouloumpis E., Wilson, T., Moore J.* Twitter sentiment analysis: the good the bad and the omg! // Artificial Intelligence. 2011. P. 538-541.

19. *Pak A., Paroubek P.* Twitter as a corpus for sentiment analysis and opinion mining // Proceedings of LREC, Valetta. 2010. P. 75-100.

20. *Адаскина Ю.В., Паничева П.В., Попов А.М.* Полуавтоматическое пополнение словарей на основе синтаксических связей // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014), Санкт-Петербург, 19 – 20 ноября 2014 г. 2014. С. 271-276.

21. *Зализняк А.А.* Грамматический словарь русского языка. М.: Русский язык, 1980.

22. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R,, Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* Scikit-learn: machine learning in Python // Journal of Machine Learning Research. 2011. V. 12 (Oct). P. 2825-2830.

USING SYNTAX FOR SENTIMENT ANALYSIS OF RUSSIAN TWEETS

Yu.V. Adaskina[1], P.V. Panicheva[2], A.M. Popov[3]

«InfoQubes», Sanct-Petersburg State University

[1]adaskina@gmail.com, [2]p.panicheva@spbu.ru, [3]hedgeonline@gmail.com

Abstract

The paper describes our approach to the task of sentiment analysis of tweets within SentiRuEval – an open evaluation of sentiment analysis systems for the Russian language. We took part in the task of sentiment analysis of Russian tweets concerning two types of organizations: banks and telecommunications companies. On both datasets, the participants were required to perform a three-way classification of tweets: positive, negative or neutral.

We used various statistical methods as basis for our machine learning algorithms. Linguistic features produced by our morpho-syntactic analyzer are applied to the classification. Syntactic relations proved to be a crucial feature for any statistical method evaluated, and SVM-based classification performed better than the others. Normalized words are another important feature for the algorithm.

The evaluation revealed that our method proved to be rather successful: we scored the first in three out of four evaluation measures.

Keywords: *sentiment analysis, syntactical relations, Russian language, statistical methods, text classification*.

REFERENCES

1. *Chetviorkin I., Braslavskiy P., Loukachevich N.* Sentiment analysis track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2012». 2012. P. 1-14.

2. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in Russian // Proceedings of BSNLP workshop, ACL, Prague. 2013. P. 12-17.

3. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Ju., Ivanov V., Tutubalina H.* Sentirueval: testing object-oriented sentiment analysis systems in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». 2015. Issue 14. V. 2. P. 13-24.

4. *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.

5. *Mullen T., Collier N.* Sentiment analysis using support vector machines with diverse information sources // Proceedings of 9th EMNLP. 2004. P. 412-418.

6. *Turney P.* Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th ACL. 2002. P. 417-424.

7. *Kudo T., Matsumoto Y.* A boosting algorithm for classification of semi-structured text // Proceedings of 9th EMNLP. 2004. P. 301-308.

8. *Matsumoto S., Takamura H., Okumura M.* Sentiment classification using word sub-sequences and dependency sub-trees // Ho T.-B., Cheung D., Liu H. (eds.) PAKDD 2005. V. 3518. P. 301-311.

9. *Mavljutov R.R., Ostapuk N.A.* Using basic syntactic relations for sentiment analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2013». 2013. P. 91-100.

10. *Yussupova N., Bogdanova D., Boyko M.* Applying of sentiment analysis for texts in russian based on machine learning approach // Proceedings of The Second International Conference on Advances in Information Mining and Management, Italy. 2012. P. 8-14.

11. *Furnkranz J., Mitchell T. M., Rilof E.* A case study in using linguistic phrases for text categorization on the WWW // Proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, US. 2998. P. 5-12.

12. *Caropreso M.F., Matwin S., Sebastiani F.A.* Learner-independent evaluation of the usefulness of statistical phrases for automated text categorization // Amita G. Chin (ed.), Text Databases and Document Management: Theory and Practice. 2006. P. 78-102.

13. *Nastase V., Shirabad J.S., Caropreso M.F.* Using dependency relations for text classification // Proceedings of the 19th Canadian Conference on Artificial Intelligence, Quebec City. 2006. P. 12-25.

14. *Zhao S., Grishman R.* Extracting relations with Integrated Information using kernel methods // Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, US. 2005. P. 419-426.

15. *Jansen B.J., Zhang M., Sobel K., Chowdury A.* Twitter power: tweets as electronic word of mouth // Journal of the American Society for Information Science and Technology. 2009. V. 60, No 11. P. 2169-2188.

16. *Go A., Bhayani R., Huang L.* twitter sentiment classification using distant supervision // Technical report, Stanford. 2009.

17. *Jiang L., Yu M., Zhou M., Liu X., Zhao T.* Target-dependent Twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, US. 2011. P. 151-160.

18. *Kouloumpis E., Wilson, T., Moore J.* Twitter sentiment analysis: the good the bad and the omg! // Artificial Intelligence. 2011. P. 538-541.

19. *Pak A., Paroubek P.* Twitter as a corpus for sentiment analysis and opinion mining // Proceedings of LREC, Valetta. 2010. P. 75-100.

20. *Adaskina Yu.V., Panicheva P.V., Popov A.M.* Poluavtomaticheskoie popolnenie slovarei na osnove sintaksicheskih svyazei // Tehnologii informacionnogo obtschestva v nauke, obrazovanii i kul'ture. Trudy XVII Vserossiiskoi ob'edinennoi konferencii «Internet i sovremennoe obschestvo» (IMS-2014), Sankt-Petersburg. 2014. S. 271-276.

21. *Zaliznyak A.A.* Grammaticheskii slovar russkogo yazika. M.: Russkii yazik, 1980.

22. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R,, Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* Scikit-learn: machine learning in Python // Journal of Machine Learning Research. 2011. V. 12 (Oct). P. 2825-2830.

СВЕДЕНИЯ ОБ АВТОРАХ

АДАСКИНА Юлия Владимировна – кандидат филологических наук, лингвист-эксперт компании «Инфо-Кьюбс».

Yulia ADASKINA received her Masters and PhD degrees in Theoretical and Applied Linguistics from Moscow State University. Currently is an expert linguist at InfoQubes, Moscow. Her scientific interests include syntax, data mining and distributional semantics.

email: adaskina@gmail.com.

ПАНИЧЕВА Полина Вадимовна – аспирант кафедры теоретической и прикладной лингвистики Санкт-Петербургского государственного университета.

Polina Vadimovna PANICHEVA, received her MSc degree in Information Technology from ITT Tallaght, Dublin, Ireland (2011). Currently is a PhD student at the Department of Theoretical and Applied Linguistics of St. Petersburg State University, Russia. Current scientific interests: distributional semantics, cognitive semantics, affective language, linguistic psychological profiling.

email: p.panicheva@spbu.ru

ПОПОВ Андрей Михайлович – аспирант кафедры математической лингвистики Филологического факультета Санкт-Петербургского государственного университета.

Andrei Mikhailovich POPOV, received MS degree in linguistics from Saint-Petersburg State University (2014). Currently is a graduate student at the Saint-Petersburg State University. Current scientific interests: machine learning, syntax parsing, fact extraction, sentiment analysis.

email: hedgeonline@gmail.com

Материал поступил в редакцию 15 июля 2015 года

УДК 004.912

ОПЫТ ПОСТРОЕНИЯ СИСТЕМЫ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ОБЪЕКТОВ НА ОСНОВЕ СИНТАКТИКО-СЕМАНТИЧЕСКОГО АНАЛИЗАТОРА

П.Ю. Поляков[1], М.В. Калинина[2], В.В. Плешко[3]

ООО «ЭР СИ О»

[1] pavel@rco.ru, [2] kalinina_m@rco.ru, [3] volodia@rco.ru

Аннотация

Исследуется применение лингвистического подхода для решения задачи автоматического определения тональности объекта. Исследование проводилось в рамках цикла тестирования систем автоматического анализа тональности SentiRuEval. Задание, предложенное организаторами дорожки, заключалось в том, чтобы определить мнение пользователя (положительное, отрицательное или нейтральное) по отношению к операторам сотовой связи на материале сообщений социальной сети Twitter и новостей. Авторы настоящей работы исключили новостные сообщения из тестовой коллекции, так как формальные тексты существенно отличаются от неформальных по своей структуре и лексике и, следовательно, требуют другого подхода. При решении поставленной задачи был использован лингвистический метод, основанный на синтактико-семантическом анализе. Согласно этому подходу тональная лексика привязывается к объекту на одной из двух последовательных стадий. Первая стадия включает в себя использование семантических шаблонов, которые сравниваются с деревом синтаксического разбора предложения; вторая стадия использует эвристики для связывания тональной лексики с объектом оценки в случае, когда синтаксические связи между ними отсутствуют. Машинное обучение не применялось. Метод продемонстрировал очень хорошие результаты, которые примерно совпадают с лучшими результатами методов с использованием машинного обучения и гибридных методов.

***Ключевые слова**: определение тональности, анализ мнений, тональность объектов, тональность атрибутов, синтактико-семантический анализ, семантические шаблоны*

1. ВВЕДЕНИЕ

Задача автоматического определения тональности в текстах на естественном языке в настоящее время является весьма актуальной. Многие производители товаров и услуг заинтересованы в мониторинге социальных сетей и блогов на предмет выявления отзывов потребителей. Тем не менее, до недавнего времени не существовало размеченного корпуса текстов на русском языке, с помощью которого разработчики могли бы тестировать свои решения и оценивать их качество. Данный пробел были призваны восполнить дорожки семинара РОМИП, а позже SentiRuEval, по автоматическому определению тональности. Однако задания дорожек предыдущих семинаров заключались в определении общей тональности текста (см., к примеру, [1]), в то время как на SentiRuEval 2015 постановка задачи была принципиально новой: определение тональности объекта. Данная задача является более сложной и требует более высокоточных алгоритмов, так как в случае определения общей тональности текста важно только соотношение положительно и отрицательно окрашенных терминов в тексте, в то время как при определении тональности объекта большое значение также имеет синтаксическая связанность объекта с тональной лексикой.

Объектно-ориентированный подход не является новым для авторов данной работы: подобный метод уже применялся в предыдущих исследованиях. В частности, была разработана экспериментальная автоматизированная система для анализа положительных и отрицательных отзывов об автомобилях и классификации их: «за что хвалят/ругают?» [2]. Для извлечения знаний было выбрано крупнейшее из нескольких десятков автомобильных сообществ «Живого журнала» – сообщество AUTO_RU «Все об автомобилях». Для оценки высказываний об автомобилях с точки зрения характеристик их потребительских свойств (положительная/отрицательная) была разработана экспериментальная онтология, содержащая:

1. более 700 различных наименований марок автомобилей и фирм-производителей; первоначальный список был расширен синонимами с вариантами

написаний; символьно-цифровые модели, упоминаемые в тексте (*BMW 325i, ВАЗ 21053*), не включались в словарь, а распознавались по формальным правилам;

2. более 1200 терминов в 24 группах, среди которых:

- 211 наименований узлов автомобиля (*движок, коробка передач, ходовая часть*);

- 71 наименование свойств (*ходовые качества, комфорт, безопасность, надежность* и т. д.);

- 882 наименования оценок характеристик узлов и свойств, включающие прилагательные, существительные, глаголы и наречия (*крутой, поломка, глючить, отстойно*);

- 37 эмоциональных характеристик (*любить, жалоба, плеваться*);.

3. около 100 семантических шаблонов, описывающих возможные синтаксические связи в предложении между группами терминов из онтологии.

Автоматизированная разработка онтологии проводилась на базе анализа языкового материала сообщества AUTO_RU «Живого журнала» при помощи средств компьютерного анализа текста. В итоге из 500 000 сообщений (60 Мбайт текста) было извлечено всего более 5000 оценок автомобилей, их узлов и характеристик, из которых более 1000 (795 хороших и 328 плохих) оценок было привязано к маркам автомобилей, а более 4000 оценок узлов и характеристик программе не удалось привязать к конкретным маркам. В результате была достигнута точность 84%, а полнота извлечения около 20%.

В настоящем исследовании были учтены ошибки и проблемы предыдущих работ, и метод семантических шаблонов был дополнен методом, позволяющим учитывать характеристики, не привязанные синтаксически к объекту интереса; что значительно повысило полноту анализа.

Следует также упомянуть, что во всех предыдущих случаях авторы оценивали полученные результаты самостоятельно. Участие в семинаре SentiRuEval дало возможность получить независимую оценку настоящего метода и сравнить результаты с другими участниками.

В данной работе представлены результаты применения лингвистического метода, включающего в себя синтактико-семантический анализ (также в литера-

туре используется термин «семантико-синтаксический анализ»), к задаче автоматического определения тональности объекта. Поставленная задача заключалась в выявлении мнения пользователей (положительного и отрицательного) по отношению к операторам сотовой связи на материале сообщений социальной сети Twitter. При решении данной задачи авторы настоящей работы ограничились лингвистическим методом, исключив машинное обучение, так как было интересно посмотреть, какие результаты даст чисто лингвистический подход.

2. ИСТОРИЯ ВОПРОСА

Как правило, методы выявления тональности по отношению к объекту или его характеристикам при нахождении объекта оценки опираются либо на исключительно статистические алгоритмы, расстояние в словах, машинное обучение и т. п. (начиная с первой работы по определению тональности объектов [3]), либо используют элементы поверхностно-синтаксического анализа для сегментации предложения, нахождения значимых союзов, отрицания и модификаторов (например, [4]). В рамках других подходов ищутся синтаксические связи между тонально окрашенным термином и объектом оценки (например, [5]), но упускается из рассмотрения тональная лексика, синтаксически не связанная с целевым объектом. Отличительной особенностью настоящего подхода является то, что при применении глубокого лингвистического анализа учитываются не только синтаксически связанные с объектом тональные слова (что обеспечивает высокую точность), но и независимая тонально окрашенная лексика и фразы (что дает высокую полноту).

Некоторые исследователи пытаются совмещать статистические и лингвистические методы для достижения лучших результатов, например, в [6] авторы, среди прочего, используют дерево синтаксических зависимостей для связывания лексики, выражающей мнение, с объектами оценки; как показывают эксперименты, учет синтаксических связей значительно повышает показатели их метода. Однако их алгоритм ищет только прямой и кратчайший путь в дереве зависимостей, таким образом, данный метод испытывает затруднения при анализе более длинных и сложных предложений. Кроме того, авторы не делают различий между объектом оценки (например, *фотокамера*), его составляющими частями

(например, *линза, ремешок*) и его характеристиками (например, *удобство применения*); и, следовательно, помечают ближайшую именную группу в качестве объекта оценки. В отличие от данного подхода в настоящей работе используется элементарная онтология для разделения объекта оценки, его составных частей и качеств; и при выявлении в тексте оценки атрибута или качества алгоритм продолжает поиск конечного целевого объекта, идя по дереву зависимостей. Если не удается установить конечный объект оценки синтаксическим путем, его поиск продолжается при помощи эвристик, основанных на учете расстояния в клаузах. При нахождении конечного объекта оценки тональность, приписанная его атрибуту, переносится на данный объект.

3. МЕТОДЫ

При выполнении поставленной задачи был учтен опыт более ранних исследований авторов данной статьи и использованы имеющиеся наработки. Подробное описание методов можно найти в работах [7] и [2]. Новизной по отношению к описанным методам был учет так называемой свободной тональности, подробнее о которой рассказывается в пункте 3.2.

Алгоритм анализа текста применительно к задаче определения тональности имеет следующие этапы:

1) токенизация – разбиение текста на абзацы, предложения, токены; определение типа токенов (русское слово, латинское слово, знак препинания, специальная конструкция);

2) морфологический анализ – определение грамматических характеристик слова (часть речи, падеж, число, род, лицо и т. д.). Основной словарь содержит: 110 тыс. слов (52 тыс. существительных, 24 тыс. глаголов, 33 тыс. прилагательных, остальное – наречия, служебные, наименования, имена, фамилии, география), 743 приставки для правил точного анализа неизвестных слов, 162 окончания для правил точного анализа неизвестных слов. Дополнительный словарь содержит 27 тыс. фамилий и 23 тыс. имен. Неизвестные слова анализируются в приближенной морфологии по правилам на известные приставки/окончания и на основе частоты суффиксов и окончаний известных слов. Подробнее с описанием морфоанализатора можно ознакомиться в [8];

3) извлечение объектов интереса – в тексте по общим правилам, опирающимся на морфологию и ключевые слова, выделяются имена персон, названия организаций и географические наименования; происходит поиск референтных упоминаний объектов, устанавливается кореферентность и анафорические связи, отождествляются упоминания одного и того же объекта в разных местах текста; идентифицируются объекты, описанные в формате XML по специальным правилам. Подробнее см. в [9];

4) синтаксический анализ – синтаксический разбор предложения в терминах дерева зависимостей, установление синтактико-семантических связей между словами и их ролей (субъект, объект, предикат и т. д.);

5) извлечение фактов (применение семантических шаблонов) – поиск в синтактико-семантической сети разбора предложения такой подсети, которая изоморфна шаблону, с заполнением слотов соответствующего фрейма именами участников ситуации из текста в соответствии с ролями, указанными в узлах шаблона [7];

6) поиск свободной тональности, привязка ее к объектам интереса.

Этапы 1, 2 и 4 были реализованы с помощью стандартных инструментов анализа текста, входящих в состав RCO Fact Extractor SDK (подробнее см. [10]). На этапе 3 особое внимание было уделено описанию объектов, представляющих интерес в рамках решаемой задачи (названия мобильных операторов, телекоммуникационная терминология и т. д.). Этапы 5 и 6 являются основными для решения задачи определения тональности и, следовательно, будут описаны подробно далее.

3.1. Семантические шаблоны

Основной метод автоматического определения тональности включал в себя использование семантических шаблонов. Семантический шаблон – это ориентированный граф, представляющий собой фрагмент дерева синтаксической зависимости с ограничениями, наложенными на его вершины. Дерево синтаксического разбора предложения содержит синтактико-семантические связи между словами, которые определяются синтаксическим анализатором. Ограничения в узлах шаблона могут накладываться на часть речи, имя сущности, семантический тип,

синтаксическую связь, морфологическую форму и т. д. Поиск фактов осуществляется путем поиска подграфа в дереве синтаксической зависимости, совпадающего с шаблоном (с учетом всех ограничений).

Для имплементации синтактико-семантического анализа использовался синтаксический парсер RCO, основанный на грамматике зависимостей. Семантическая сеть, построенная анализатором, инвариантна к порядку слов и залогу глагола; например, предложения (1) *Оператор украл деньги со счета* и (2) *Деньги украдены оператором со счета* будут иметь одинаковое представление. Подобная семантическая сеть представляет собой промежуточный уровень представления между собственно семантической схемой ситуации и ее конкретным языковым выражением, т. е. представлением глубинно-синтаксического уровня, абстрагированным от особенностей поверхностного синтаксиса.

Настройки семантического интерпретатора позволяют отфильтровывать отрицание и «ирреальные» предложения (повелительное, сослагательное наклонения и т. д.), которые не соответствуют реальным событиям и фактам, и потому не представляют интереса для фактографического анализа. Как результат, примеры вроде: (3) *если Билайн будет плохо работать; сеть якобы падает; связь бы обрывалась; не Билайн плохо работает* можно исключать из анализа тональности.

Для сокращения количества шаблонов, описывающих семантические фреймы, имеются так называемые служебные шаблоны, которые добавляют новые узлы и связи в синтактико-семантическую сеть. В процессе семантического анализа и извлечения фактов служебные шаблоны срабатывают в первую очередь, и, таким образом, семантические шаблоны опираются на сеть, построенную синтаксическим анализатором и модифицированную служебными шаблонами. Например, если мы интерпретируем высказывания: (4) *X делает Y,* (5) *X начинает делать Y* и (6) *X решил сделать Y* как тождественные в рамках описания определенной ситуации, вместо того, чтобы создавать семантический шаблон для каждого из этих примеров, можно написать один служебный шаблон, который будет маркировать субъект вспомогательного глагола как субъект смыслового глагола, и один простой семантический шаблон вида: (4) *X делает Y.*

Семантические шаблоны могут иметь так называемые запрещающие вершины, которые накладывают ограничения на контекст, определяя, при наличии

какого контекста шаблон не должен срабатывать. Например, высказывание (7) *У Билайна надежная связь* выражает позитивную характеристику объекта, в то время как добавление наречия *наименее* меняет его тональность на противоположную: (8) *У Билайна наименее надежная связь*. С помощью запрещающих вершин можно делать различия между двумя этими высказываниями, добавив условие, что прилагательное, выражающее оценку, не должно быть модифицировано наречием *наименее*. Использование запрещающих вершин помогает значительно повысить точность определения тональности.

На Рис. 1 представлен семантический шаблон, который используется для определения тональности объекта, выраженной глаголом или наречием, в предложениях вида: (9) *Билайн ловит хорошо;* (10) *Интернет летает*.

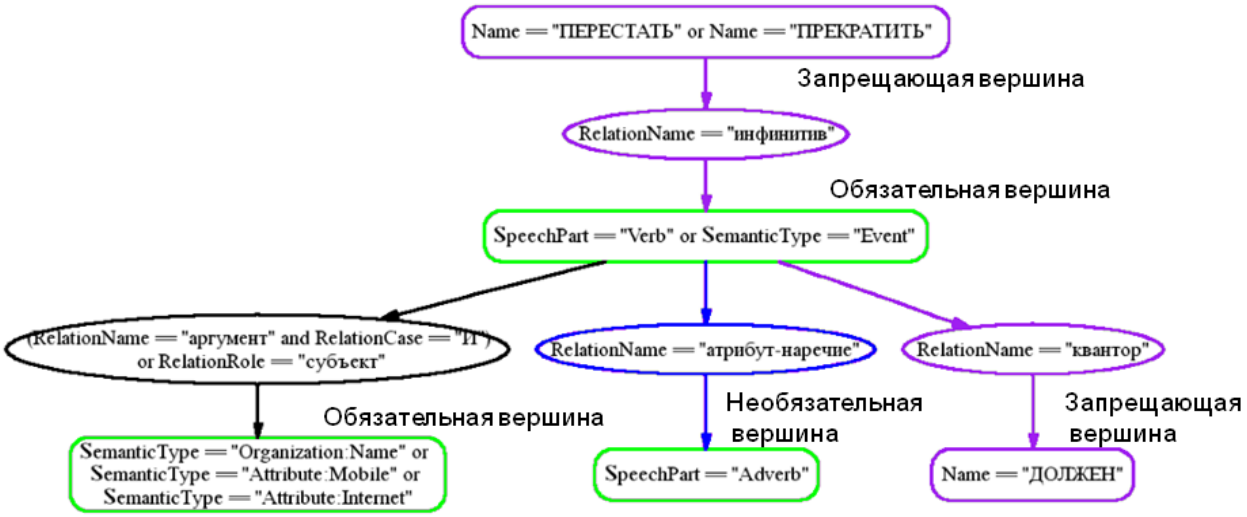


Рис. 1. Пример семантического шаблона

Узлы шаблона содержат ограничения на часть речи (*SpeechPart == "Verb"; SpeechPart == "Adverb"*), конкретные слова (*Name == "ПЕРЕСТАТЬ" or Name == "ПРЕКРАТИТЬ"*), семантические категории (*SemanticType == "Organization:Name" or SemanticType == "Attribute:Mobile"*). Ограничения на синтактико-семантические связи между словами включают в себя: RelationName – тип синтактико-семантической связи между узлами (*RelationName == "аргумент"; RelationName == "квантор"*), RelationRole – семантическую роль (*RelationRole == "субъект*"), RelationCase – падеж (*RelationCase == "И"*). Запрещающие вершины говорят о том, что

глагол, выражающий оценку, не должен быть подчинен фазисным глаголам *перестать* или *прекратить* и не должен быть модифицирован предикативом *должен*. Таким образом, шаблон сработает на предложении (9) *Билайн хорошо ловит* (которое выражает положительную оценку), но не сработает на примерах (11) *Билайн перестал хорошо ловить* (выражает негативную оценку) и (12) *Билайн должен хорошо ловить* (оценка не определена).

Ограничения, накладываемые на узлы семантических шаблонов, были дополнены специальными словарями (фильтрами), содержащими лексику, выражающую позитивную или негативную оценку. Данные словари содержат существительные, прилагательные, глаголы, наречия, а также словосочетания. Термин из фильтра должен быть синтаксически связан с объектом оценки. Отбор лексики для фильтров производился вручную лингвистом-экспертом. Примеры положительных терминов: *супербыстрый, шустро, красота, крутяк, блистать, радовать, обеспечивать уверенный прием*. Примеры отрицательных терминов: *завышенный, препротивнейший, позорище, тормознутость, обдирать, терять соединение, фигово.*

Например, в качестве ограничений семантические фильтры накладываются на следующие узлы шаблона на Рис.1 : глаголы и отглагольные существительные параметризуют вершину с ограничением: *SpeechPart == "Verb" or SemanticType == "Event"*; наречия параметризуют вершину с ограничением: *SpeechPart == "Adverb"*; обе эти вершины имеют семантическую роль «Оценка».

Целевыми объектами оценки являлись крупнейшие российские операторы сотовой связи (Билайн, Мегафон, МТС, Ростелеком и Tele2), но также учитывалась оценка пользователями атрибутов сотовых операторов (качество связи, мобильный интернет, абонентская поддержка и т. п.).

Анализируя отзывы пользователей в социальных сетях и на форумах, эксперты определили набор атрибутов, на которые пользователи наиболее часто обращают внимание. Таким образом был составлен список наиболее важных для пользователей характеристик. Данные атрибуты были поделены на три класса: 1) атрибуты мобильной связи – термины, имеющие отношение исключительно к мобильной телефонии: *SMS, MMS, 3G, LTE, SIM-карта, роуминг* и т. д.; 2) интернет-атрибуты – термины, имеющие отношение исключительно к интернету (любому):

интернет, пинг и т. д.; 3) общие атрибуты – термины, часто используемые в связи с мобильной телефонией, но могущие описывать и другие предметные области: *колл-центр, сигнал, сеть, техподдержка, баланс* и т. д. Каждый из трех классов был пополнен синонимами и вариантами написания (*интернет=инет=и-нет; lte=лте =lтешечка =лте-шечка; баланс счета=состояние счета=средства на счету=деньги на счету* и т. п.). При нахождении в тексте оценки атрибута данная оценка переносилась также на соответствующего мобильного оператора.

На Рис. 1 вершина с ограничением *SemanticType == "Organization:Name" or SemanticType == "Attribute:Mobile" or SemanticType == "Attribute:Internet"* параметризуется названиями сотовых операторов, атрибутов мобильной связи или интернет-атрибутами; данная вершина имеет роль «Объект оценки».

Описанный метод обеспечивает очень высокую точность, однако его полнота оставляет желать лучшего.

3.2. «Свободная» тональность

Хотя использование семантических шаблонов дает очень высокую точность, применение этого метода имеет определенные ограничения: слово, выражающее оценку, должно быть в том же предложении, что и объект оценки, и должно быть синтаксически с ним связано. Так как в реальных текстах дело далеко не всегда обстоит таким образом, некоторые случаи явно выраженной тональности будут упущены данным подходом, и полнота пострадает. Особенно ощутима эта проблема при анализе неформальных текстов – сообщений форумов, социальных сетей, блогов и т. д. При написании неформальных сообщений пользователи часто пренебрегают правилами орфографии и пунктуации, делают опечатки, из-за чего синтаксический анализатор может ошибаться при построении связей в предложении, или синтактико-семантическая сеть может вовсе развалиться. Кроме того, пользователи могут выражать свои эмоции через междометия, которые не являются частью синтаксической структуры дерева зависимостей, и, следовательно, не могут быть выловлены при помощи семантических шаблонов. Термины, которые выражают оценку, но синтаксически не связаны с объектом оценки (или анализатор не смог построит такую связь), в рамках данной работы получили условное название «свободная тональность».

Для решения проблемы свободной тональности был применен подход, основанный на алгоритме, ищущем в тексте тонально окрашенную лексику, опираясь на словари (или профили) положительной и отрицательной лексики, и, если такая лексика найдена, пытающемся привязать ее к объекту оценки.

Оба этих метода (семантических шаблонов и свободной тональности) дополняют друг друга, работая последовательно, при этом метод семантических шаблонов работает первым. Алгоритм поиска свободной тональности «игнорирует» тональную лексику, уже привязанную к объекту шаблонами, так как предполагается, что точность, обеспечиваемая шаблонами, близка к стопроцентной.

В качестве профилей, содержащих положительную и отрицательную тональные лексики, были использованы соответствующие фильтры с небольшой модификацией: были удалены контекстно зависимые термины, и оставлена только явная оценочная или эмоциональная лексика. Например, были убраны глаголы *УМЕРЕТЬ, ПРОИГРЫВАТЬ*, так как, хотя они, бесспорно, выражают негативную оценку в следующих примерах: (13) *интернет умер;* (14) *оператор X проигрывает оператору Y*; в другом контексте, не связанном с мобильной телефонией, они могут не иметь оценочного значения, а просто констатировать факт. Одновременно с этим профили были пополнены междометиями и устойчивыми экспрессивными выражениями, которые нельзя синтаксически привязать к объекту оценки, например: (15) *не надо так! что за нах; ни фига себе; ну как так можно* и т. п.

При обнаружении тонально окрашенных терминов алгоритм ищет в данном тексте объект оценки – название сотового оператора – и приписывает ему тональность. В случае, если в тексте упоминается несколько сотовых операторов, оценка приписывается ближайшему из них. В случае, если в одном тексте обнаружены и положительные, и отрицательные термины, относящиеся к одному оператору, предпочтение отдавалось отрицательной оценке, так как предполагалось, что позитивная лексика в данном контексте выражает сарказм.

В рамках решения данной задачи машинное обучение не применялось. Описанные методы опирались исключительно на лингвистический анализ.

4. ТЕСТОВАЯ ВЫБОРКА

Обучающая и тестовая выборки, предоставленные организаторами, состояли из 5000 размеченных и 5000 неразмеченных сообщений социальной сети Twitter, содержащих оценочные суждения пользователей либо положительные или отрицательные информационные поводы, касающиеся сотовых операторов.

Так как основной задачей определения тональности в социальных сетях является выявление мнения пользователей, были отобраны сообщения, содержащие перепечатки новостей, после чего было дополнительно измерено качество определения тональности на тестовой выборке без новостных сообщений. В результате новостные сообщения были исключены из итоговой тестовой коллекции, так как разница в синтаксической структуре и лексике между формальными (новостными) и неформальными (посты, блоги, твиты) текстами представляется принципиальной. Как правило, авторы новостных сообщений явно не выражают свое отношение: новости содержат простое описание событий и фактов, которые можно трактовать как положительный или отрицательный информационный повод по отношению к объекту интереса, в том время как явное оценочное суждение в них не содержится. Кроме того, лексическое наполнение неформальных сообщений значительно отличается от лексики, встречающейся в формальных текстах. Следовательно, анализ новостных сообщений требует другого подхода.

С учетом вышесказанного было дополнительно оценено качество работы настоящего метода после исключения из тестовой коллекции перепечаток новостей и пресс-релизов компаний. Так как данный метод основан исключительно на лингвистическом подходе, не было необходимости использовать обучающую выборку.

РЕЗУЛЬТАТЫ

Для начала, с целью оценки уровня согласованности между экспертами, наш эксперт разметил тестовую коллекцию вручную и пометил каждое упоминание сотового оператора как позитивное, негативное или нейтральное. Результаты оценки эксперта представлены в Таблице 1. Качество результатов оценивалось с помощью F1-меры с макро- и микроусреднением. Дополнительно, для наглядности, в таблицах приведены точность и полнота. Как видно из Таблицы 1, разметка сообщений нашим экспертом отличалась от разметки организаторов. Цифры, по-

лученные нашим экспертом, представляются максимально возможным результатом для системы автоматического определения тональности на данной коллекции. Согласованность между нашим экспертом и разметкой организаторов стала выше, когда из выборки были исключены новости, что подтверждает предположение о том, что для анализа тональности информационных поводов требуется иной подход.

Таблица 1. Оценка согласованности между нашим экспертом и организаторами

Macro-average			**Micro-average**		
Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
0.785	0.694	0.737	0.831	0.735	0.780

Подробные результаты применения настоящего метода представлены в Таблице 2. Для сравнения в таблицу включен лучший из всех результатов, показанных участниками дорожки. Результат, полученный настоящим методом, оказался одним из лучших.

Таблица 2. Результат оценки настоящего метода и лучшая F1-мера среди методов всех участников

	Macro-average			**Micro-average**		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
RCO	0.465	0.562	0.492	0.475	0.583	0.524
Лучший результат			0.492			0.536

Интересно отметить, что несколько методов, основанных на различных подходах (полностью машинное обучение, гибридный подход – машинное обучение с элементами синтаксиса), показали очень близкую величину F1 – около 0.5 (подробнее о сравнении с результатами других участников см. [11]); и, тем не менее, эти результаты значительно ниже теоретического максимума, который соответствует уровню согласованности между экспертами (см. Таблицу 1). Данный факт

служит лишним подтверждением того, что автоматическое определение тональности до сих пор является трудной и интересной задачей.

ЗАКЛЮЧЕНИЕ

Описанный лингвистический подход продемонстрировал очень высокое качество, которое примерно соответствует лучшим результатам, показанным методами с использованием машинного обучения и гибридными методами (сочетающими в себе машинное обучение с элементами синтаксического анализа).

В будущем планируется дополнить лингвистический подход методами машинного обучения:

- в части генерации словарей тональной лексики;
- в части генерации шаблонов;
- в части генерации правил связывания объектов с атрибутами и отнесения свободной тональности.

Также планируется провести оценку полноты и точности определения тональности по отдельности для каждого способа выражения мнения пользователя об объекте и учитывать вес достоверности определения тональности.

СПИСОК ЛИТЕРАТУРЫ

1. *Четверкин И.И., Браславский П.И., Лукашевич Н.В.* Дорожки по анализу мнений на РОМИП // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’ 2012. Бекасово, 2012.

2. *Ермаков А.Е.* Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии. 2009. № 7. С. 50-55.

3. *Hu M., Liu B.* Mining and summarizing customer reviews // International Conference on Knowledge Discovery and Data Mining (ICDM), 2004.

4. *Kan D.* Rule-based approach to sentiment analysis at ROMIP’11 // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’ 2012. Бекасово, 2012.

5. *Popescu A., Etzioni O.* Extracting product features and opinions from reviews // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2005.

6. *Jakob N., 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.

7. *Ермаков А.Е., Плешко В.В.* Семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии. 2009. № 6. С. 2-7.

8. *Ермаков А.Е., Плешко В.В.* Компьютерная морфология в контексте анализа связного текста // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. М.: Наука, 2004.

9. *Ермаков А.Е., Плешко В.В., Митюнин В.А.* RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов. Москва, 2003. URL: http://www.rco.ru/?p=4599.

10. RCO Fact Extractor SDK (Rus.), URL: http://www.rco.ru/?page_id=3554.

11. *Поляков П.Ю., Калинина М.В., Плешко В.В.* Автоматическое определение тональности объектов с использованием семантических шаблонов и словарей тональной лексики // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2015. Москва, 2015.

EXPERIMENT IN BUILDING AN AUTOMATIC OBJECT-ORIENTED SENTIMENT DETECTION SYSTEM BASED ON THE SYNTACTIC AND SEMANTIC ANALYZER

P.Yu. Polyakov[1], M.V. Kalinina[2], V.V. Pleshko[3]

RCO
[1] pavel@rco.ru, [2] kalinina_m@rco.ru, [3] volodia@rco.ru

Abstract

This paper focuses on the use of a linguistics-based method for automatic object-oriented sentiment analyses. The study was conducted as part of SentiRuEval automatic sentiment analysis system testing cycle. The original task was to extract users'

opinions (positive, negative, neutral) about telecom companies, expressed in tweets and news. In this study news was excluded from the dataset because, being formal texts, news significantly differs from informal ones in its structure and vocabulary and therefore demands a different approach. Only linguistic approach based on syntactic and semantic analysis was used. In this approach, a sentiment-bearing word or expression is linked to its target object at either of two stages, which perform successively. The first stage includes usage of semantic templates matching the dependence tree, and the second stage involves heuristics for linking sentiment expressions and their target objects when syntactic relations between them do not exist. No machine learning was used. The method showed a very high quality, which roughly coincides with the best results of machine learning methods and hybrid approaches.

Keywords*: sentiment analysis, object-oriented sentiment analysis, aspect-based sentiment analysis, opinion mining, syntactic and semantic analysis, semantic templates*

REFERENCES

1. *Chetviorkin I.I., Braslavski P.I., Loukachevitch N.V*. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: International Conference Dialog-2011. 2011. P. 739-746.

2. *Ermakov A.E.* Izvlechnie znanii iz teksta i ih obrabotka: sostoyanie I perspektivy // Informacionnye tehnologii. 2009. № 7. P. 50-55.

3. *Hu M., Liu B.* Mining and summarizing customer reviews // International Conference on Knowledge Discovery and Data Mining (ICDM), 2004.

4. *Kan D.* Rule-based approach to sentiment analysis at ROMIP'11 // Computational linguistics and intellectual technologies: proceedings of International conference Dialog-2012, 2012.

5. *Popescu A., Etzioni O.* Extracting product features and opinions from reviews // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2005.

6. *Jakob N., 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.

7. *Ermakov A.E., Pleshko V.V.* Semanticheskaya interpretaciya v sistemah komp'yuternogo analysa teksta // Informacionnye tehnologii. 2009. № 6. S. 2-7.

8. *Ermakov A.E., Pleshko V.V.* Komp'yuternaya morfologia v kontekste analysa svyaznogo teksta // Computational Linguistics and Intellectual Technologies: International Conference Dialog-2004. 2004.

9. *Ermakov A.E., Pleshko V.V., Mityunin V.A.* RCO Pattern Extractor: komponent vydeleniya osobih ob'ektov v tekste // Informatizaciya i informacionnaya bezopasnost pravoohranitelnih organov: XI Mezdunarodnaia nauchnaya konferenciya. Sbornik trudov. Moskva, 2003. URL: http://www.rco.ru/?p=4599.

10. RCO Fact Extractor SDK (Rus.), URL: http://www.rco.ru/?page_id=3554.

11. *Polyakov P. Yu., Kalinina M.V., Pleshko V.V.* Avtomaticheskoe opredelenie tonalnosti ob'ektov s ispolzovaniem semanticheskih shablonov I slovarei tonalnoi leksiki // Computational Linguistics and Intellectual Technologies: International Conference Dialog-2015. 2015.

СВЕДЕНИЯ ОБ АВТОРАХ

ПОЛЯКОВ Павел Юрьевич – ведущий программист компании ООО «ЭР СИ О» (RCO), аспирант Остравского технического университета.

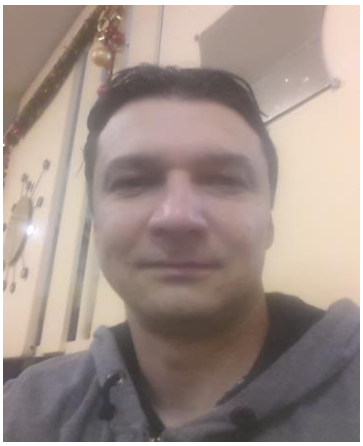
Pavel Yurjevich POLYAKOV, received Master's degree in applied physics and mathematics from Moscow Institute Physics and Technology (2004). Currently is a lead programmer at RCO LLC and PhD student at the Technical University of Ostrava. Current scientific interests: text mining, computational linguistics, knowledge extraction technologies, data mining, artificial intelligence, Boolean factor analysis, recurrent neural networks.

e-mail: pavel@rco.ru

КАЛИНИНА Мария Викторовна – ведущий лингвист компании ООО «ЭР СИ О» (RCO).

Maria Viktorovna KALININA, received Master's degree in theoretical and applied linguistics from Lomonosov Moscow State University (2003). Currently is a lead linguist at RCO LLC, a leading company of the Russian market in the field of computational linguistics and processing of unstructured information. Current scientific interests: data mining, text mining, computational linguistics, knowledge extraction technologies.

e-mail: kalinina_m@rco.ru



ПЛЕШКО Владимир Владимирович – генеральный директор компании ООО «ЭР СИ О» (RCO).

Vladimir Vladimirovich PLESHKO, received Master's degree in applied mathematics from Lomonosov Moscow State University (1996). Currently is CEO at RCO LLC, a leading company of the Russian market in the field of computational linguistics and processing of unstructured information. Current scientific interests: data mining, text mining, computational linguistics, knowledge extraction technologies, artificial intelligence.

e-mail: vp@rco.ru

Материал поступил в редакцию 15 июля 2015 года

УДК 004.912

ИЗВЛЕЧЕНИЕ АСПЕКТОВ ТОВАРОВ ИЛИ УСЛУГ ИЗ ОТЗЫВОВ ПОТРЕБИТЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ

Ю.В. Рубцова[1], С.А. Кошельников[2]

Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск

[1]yu.rubtsova@gmail.com, [2]koshelnikovsa@gmail.com

Аннотация

Описана система, принимавшая участие в соревновании SentiRuEval-2015 по автоматическому извлечению аспектов из отзывов и оценке этих аспектов по тональности. В основе разработанной системы лежит алгоритм условных случайных полей (CRF), она использовалась в решении двух подзадач и тестировалась на двух предметных областях: рестораны и автомобили. Для обеих задач и обеих предметных областей показаны высокие показатели метрики полноты. Это означает, что система может вполне успешно находить аспектные термины. Вместе с тем, полученный низкий показатель точности свидетельствует о том, что система принимает за аспектные достаточно много терминов, которые аспектными не являются. В целом же система показала сравнительно хорошие результаты по сравнению с другими участниками соревнования.

Ключевые слова: *извлечение знаний, извлечение аспектов, CRF.*

1. ВВЕДЕНИЕ

Год от года растет популярность блогов, социальных сетей, сайтов отзывов о товарах и услугах, вместе с этим растет и количество отзывов, которые пишут пользователи интернета. Таким образом, за последнее время в различных предметных областях накоплен огромный объем отзывов, оценок, рекомендаций, привлекающих внимание как исследователей, которые занимаются извлечением

мнений из текстов, анализом тональности отзывов, поиском трендов, так и представителей бизнеса, которые занимаются практическими задачами репутационного маркетинга.

Чаще всего автоматический анализ тональности исследуется на следующих уровнях: всего документа [1–3]; предложения или фразы [4]; аспектов [5–7].

Как правило, человек высказывает мнение не относительно товара или сервиса целиком, а относительно части товара, некоторого свойства или характеристики, это и есть аспект, который требуется извлечь из текста и оценить его тональность. Исследуя анализ тональности на уровне аспектов, мы можем получить гораздо больше полезной информации об отношении автора текста к разным свойствам исследуемого товара или услуги, чем проводя анализ тональности всего текста целиком.

В рамках конференции «Диалог–2015» проводилось тестирование SentiRuEval [8] технологий автоматической тоновой классификации отзывов на уровне аспектов. Участникам тестирования предлагалось решить пять подзадач:

A. выделение явных аспектных терминов в отзыве;

B. выделение всех аспектных терминов в отзыве;

C. нахождение оценок тональности для аспектных терминов (явных и неявных);

D. классификация аспектных терминов по категориям;

E. проставление оценок аспектным категориям по отзыву в целом.

В данной работе описана система, которая принимала участие в соревновании SentiRuEval в подзадачах A и B.

Статья структурирована следующим образом: в следующем разделе приведен обзор работ по извлечению аспектов. В разделе 3 описана разработанная система. В разделе 4 представлены результаты системы в сравнении с результатами других участников SentiRuEval, а в пятом разделе приведен анализ ошибок. В заключении сделаны выводы и описаны возможные перспективы работы.

2. ОБЗОР РАБОТ ПО ИЗВЛЕЧЕНИЮ АСПЕКТОВ

Существует четыре основных подхода к извлечению аспектных терминов из текстов. Первый подход основан на частоте использования существительных и словосочетаний. Как правило, оставляя комментарии, разные люди используют

одни и те же термины для описания свойств объекта и своего отношения к этим объектам и другие термины, отличные от первых, для всего остального текста (описания ситуаций, дополнительная сопроводительная информация). Таким образом, имея достаточно большое количество комментариев и отзывов, можно извлекать эксплицитные аспектные термины с показателями полноты до 72% и точностью до 80% – в основе алгоритма лежит подсчет частоты употребления существительных и словосочетаний, которые часто встречаются в текстах одной предметной области [9]. Позже метрика полноты этого алгоритма была усовершенствована на 22% при снижении метрики точности всего на 3% за счет подключения дополнительных текстовых корпусов и использования другого алгоритма разметки признаков [10]. Так как общеупотребимые слова часто встречаются в текстах, они ошибочно определялись как аспектные термины, поэтому была придумана фильтрация для исключения высокочастотных неаспектных существительных и словосочетаний [11].

Второй подход основан на извлечении одновременно и тонального термина (или выражение отношения пользователя), и аспекта. Так как любое мнение высказывается по отношению к какому-либо объекту, то, находя тональные термины (sentiment word), мы можем находить аспекты, к которым относятся найденные тональные термины. Hu and Liu использовали этот подход для нахождения низкочастотных аспектных терминов [9].

Следующий подход – машинное обучение с учителем (supervised machine learning). Как правило, в задачах извлечения аспектных терминов машинное обучение с учителем сводится к задаче маркировки последовательностей, потому что аспекты продукта и выражения мнения о продукте зачастую являются взаимозависимыми и образуют последовательность слов. Наиболее часто используемые методы машинного обучения с учителем – это скрытые марковские модели (Hidden Markov Modeling – HMM) [12] и метод условных случайных полей (Conditional random fields – CRF) [13].

Четвертый метод – это машинное обучение без учителя или тематическое моделирование (topic modeling). Тематическое моделирование предполагает, что каждый документ состоит из нескольких тем, каждая из тем имеет свою веро-

ятность относительно выбранного документа [14, 15]. Большинство работ по извлечению аспектных терминов с использованием тематического моделирования опирается на следующие методы: вероятностный латентно-семантический анализ (pLSA) [16] и Латентное размещение Дирихле (LDA) [17].

При решении сложных задач, таких, как одновременное извлечение аспектных терминов и их классификация по тональности или извлечение аспектов и автоматическая группировка по аспектным категориям, могут использоваться комбинированные подходы, например, комбинация методов максимальной энтропии и латентного размещения Дирихле [18] или системы с частичным обучением, которые используют тематическое моделирование и предоставленные пользователем термины, размеченные для некоторых категорий [19].

3. ОПИСАНИЕ СИСТЕМЫ

Разработанная система принимала участие в решении двух задач:

- определение эксплицитных (явных) аспектных терминов, т. е. требовалось извлечь часть исследуемого объекта или его характеристику, например, «движок» для предметной области «автомобили» или «обслуживание» для предметной области «рестораны»;
- определение всех аспектных терминов исследуемого объекта; к поиску эксплицитных терминов добавляется поиск имплицитных терминов (термин + однозначное тональное отношение автора к этому термину) и поиск тонально окрашенных фактов (автор не использует тонально-окрашенные слова, но указывает на некоторый факт, однозначно определяющий отношение автора к объекту).

Для извлечения аспектов из предложений, которые содержат мнения автора о товаре или услуге, был использован метод CRF. В качестве входных данных CRF использует последовательности лексем, далее алгоритм вычисляет вероятности различных возможных последовательностей меток и выбирает одну с максимальной вероятностью.

CRF – это графическая модель, предназначенная для оценки условных вероятностей событий, соответствующих вершинам некоторого графа Г, при усло-

вии наблюдаемых данных. Пусть x = {x_1, … , x_N} – последовательность наблюдаемых данных. В нашем случае это токены одного отзыва. Пусть Y = {y_1, … , y_N} – последовательность случайных величин, связанных с вершинами графа Г.

В нашем случае графическая модель выглядит приведенном на рисунке 1 образом, а случайные величины y_N – метки токенов, которые хотим научиться предсказывать.

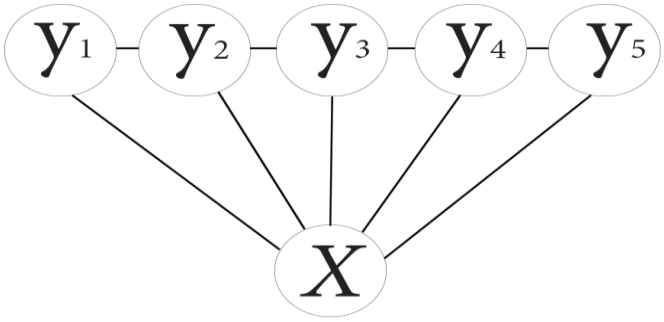


Рис. 1. Пример представления условного случайного поля

Такой набор данных называется CRF, если для каждой вершины υ графа Г выполнено марковское условие

$$P(y_v \mid Y_{V\setminus\{v\}}, X) = P(y_v \mid Y_{o(v)}, X), \tag{1}$$

где *V* – множество вершин графа Г, а *O(υ)* – множество соседних с υ вершин в графе Г. Таким образом, метка зависит только от близкого контекста. Тогда, согласно [20], линейным условным случайным полем называется распределение вероятностей вида

$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left(\sum_{c \in C} \lambda_c f_c(y_c, X)\right), \tag{2}$$

где *Z*(*x*) − нормализующий множитель, *C* – множество всех клик графа Г, f_c – признаки, а λ_i – коэффициенты. Эти коэффициенты подбираются в процессе обучения данной модели так, чтобы максимизировать логарифм функции правдоподобия на обучающем наборе *X, Y*:

$$Z(X) = \sum_{y} \exp\left(\sum_{c \in C} \lambda_c f_c(y_c, X)\right), \tag{3}$$

Среди преимуществ CRF выделяют:

- отсутствие презумпции условной независимости наблюдаемых переменных,
- отсутствие ситуации, когда преимущество получают состояния с меньшим количеством переходов, так как строится единое распределение вероятностей и нормализация (коэффициент $Z(x)$) производится в целом, а не в рамках отдельного состояния.

В качестве программного комплекса был использован Mallet [21].

3.1. Предобработка

Так как мы имеем дело с последовательностями слов, то каждое слово в этой последовательности было помечено меткой. Например, “s-e” обозначает начало эксплицитного аспектного термина, “c-e” – продолжение эксплицитного термина, “s-i” обозначает начало имплицитного аспекта, продолжение имплицитного аспекта обозначается как “c-i”, аналогично s-f и c-f для фактического аспектного термина. Меткой “O” снабжаются все неаспектные термины.

Для выделения частей речи и лемматизации слов в отзыве был использован TreeTagger для русского языка [22].

Марки автомобилей часто пишут латиницей или содержат цифры, например, Nissan Micra или ВАЗ 2109. Поэтому для коллекции автомобилей были добавлены правила, позволяющее выделить полное название (марку) автомобиля в один эксплицитный термин. Как видно из таблицы 3, это дало положительные результаты: система заняла 3-е место согласно F-мере (точное соответствие).

Перед началом работы все заглавные буквы были преобразованы в строчные, так как программные средства могут посчитать, что аспект “Мотор” и “мотор” – это два разных аспекта, что на самом деле таковым не является.

3.2. Признаки для CRF

Слово (Word). Текущее слово в исходной словоформе использовалось в качестве признака. Для того чтобы получить больше информации о контексте употребления слова, были извлечены предыдущее и последующее слова, они были использованы как дополнительные признаки.

Часть речи (POS) текущего токена использовалась в качестве признака. Аспектные термины часто бывают существительными. Применение разметки по частям речи добавляет полезную информацию о принадлежности слова к той или иной части речи. Для определения части речи был использован TreeTagger. TreeTagger проводит полный морфологический анализ слова, который излишен для описанной задачи, поэтому полный морфологический разбор был сокращен до названия части речи. Например, N (noun) для “мотор” или V (verb) для “ехать”.

Лемма (Lemma). Нормальная форма слова текущего токена. Русский язык богат на словоформы, поэтому нормальная форма слова была добавлена в качестве признака.

3.3. Архитектура системы

Были построены две системы:

Система 1. Применение CRF со всеми вышеуказанными метками. Для задачи A – поиска эксплицитных терминов использовались: s-e, c-e, O; для задачи B – поиска всех терминов: s-e, c-e, s-i, c-i, s-f, c-f, O.

Система 2. Это объединение двух результатов, полученных от двух CRF: CRF для поиска только эксплицитных терминов и CRF для поиска имплицитных терминов + фактических терминов, но не эксплицитных.

Для решения задачи A применялась только первая система, для задачи B – обе. В основе обеих систем лежат CRF и одинаковые признаки, различие есть только для задачи B.

4. РЕЗУЛЬТАТЫ

Результаты участников в задачах A и B оценивались с помощью F-меры. Рассчитывались два варианта F-меры: точное соответствие и частичное соответствие. F-мера точного соответствия рассчитывалась для каждого отзыва отдельно, усреднение всех полученных значений является результирующей F-мерой. Для измерения частичного соответствия рассчитывалось пересечение между золотым стандартом и термином, извлеченным системой. В таблицах 1–4 представлен результат работы системы при решении задачи A, в таблицах 5–8 – результат задачи B. Результаты, полученные построенной системой, сравнивались с Baseline и с двумя первыми лучшими результатами участников SentiRuEval.

Из таблиц 1–4 видно, что для обеих предметных областей система показывает хорошие результаты по метрике полнота (2-е место и для ресторанов, и для автомобилей в задаче А), причем для предметной области «автомобили» результат выше у системы без применения признаков “lemma”, вероятно, это связано с использованием правил предобработки для коллекции автомобилей.

Аналогично для задачи В обе построенные системы показали достаточно высокий результат по метрике полнота (таблицы 5–8). В предметной области «рестораны» система 1 с использованием признаков word+pos+lemma заняла третье место среди всех участников согласно F-мере в случае частичного соответствия.

Таблица 1. Результаты для задачи А (предметная область «рестораны»), точное соответствие

	Precision	**Recall**	**F-measure**
baseline	0,557	0,6903	0,6084
№1	0,7237	0,5738	0,6319
№2	0,6358	0,6327	0,6266
Word+POS	0,661	0,515	0,5704
+lemma	0,6674	0,5417	0,5899

Таблица 2. Результаты для задачи А (предметная область «рестораны»), частичное соответствие

	Precision	**Recall**	**F-measure**
baseline	0,658	0,696	0,6651
№1	0,8078	0,6165	0,689
№2	0,7458	0,7114	0,7191
Word+POS	0,738	0,563	0,6277
+lemma	0,7485	0,5937	0,652

Таблица 3. Результаты для задачи А (предметная область «автомобили»), точное соответствие

	Precision	**Recall**	**F-measure**
baseline	0,5747	0,6287	0,5941
№1	0,76	0,6218	0,6761

№2	0,6619	0,656	0,6513
Word+POS	0,7109	0,5454	0,6075
+lemma	0,704	0,5785	0,6256

Таблица 4. Результаты для задачи А (предметная область «автомобили»), частичное соответствие

	Precision	**Recall**	**F-measure**
baseline	0,7449	0,6724	0,6966
№1	0,7917	0,7272	0,7482
№2	0,8561	0,6551	0,7304
Word+POS	0,797	0,6047	0,6747
+lemma	0,7908	0,6485	0,6991

Таблица 5. Результаты для задачи В (предметная область «рестораны»), точное соответствие

	Precision	Recall	F-measure
baseline	0,546577	0,647729	0,587201
№1	0,609432	0,600621	0,600128
№2	0,733599	0,513197	0,596179
Система 1 Word+POS	0,639256	0,456334	0,52577
+lemma	0,639798	0,487202	0,546905
Система 2 Word+POS	0,652145	0,458471	0,531644
+lemma	0,67152	0,491622	0,56153

Таблица 6. Результаты для задачи В (предметная область «рестораны»), частичное соответствие

	Precision	Recall	F-measure
baseline	0,671626	0,593093	0,619285
№1	0,756213	0,610754	0,667928

№2	0,668677	0,637097	0,645234
Система 1 Word+POS	0,710428	0,493393	0,5692
+lemma	0,709915	0,529354	0,595303
Система 2 Word+POS	0,724649	0,457863	0,547813
+lemma	0,752364	0,493553	0,585126

Таблица 7. Результаты для задачи B (предметная область «автомобили»), точное соответствие

	Precision	Recall	F-measure
baseline	0,597886	0,589612	0,588623
№1	0,7701	0,553546	0,636623
№2	0,656321	0,616423	0,630149
Система 1 Word+POS	0,690826	0,476309	0,556107
+lemma	0,670594	0,518742	0,578086
Система 2 Word+POS	0,718995	0,482064	0,568331
+lemma	0,701193	0,520375	0,589311

Таблица 8. Результаты для задачи B (предметная область «автомобили»), частичное соответствие

	Precision	Recall	F-measure
baseline	0,783254	0,605976	0,674288
№1	0,814283	0,650998	0,714762
№2	0,795431	0,646999	0,704189
Система 1 Word+POS	0,793637	0,53216	0,625502
+lemma	0,777257	0,584768	0,656113
Система 2 Word+POS	0,808562	0,509979	0,61308
+lemma	0,782394	0,558153	0,638947

5. АНАЛИЗ ОШИБОК

Выделяют несколько типов ошибок: система не извлекла аспектный термин, система приняла за аспектный термин обычное слово или фразу. Существует еще один тип ошибок: частично извлеченный аспектный термин. С помощью имеющихся средств довольно затруднительно провести подробный анализ третьего типа ошибок. Как видно из таблицы 9, большинство ошибок системы связано с ненайденными аспектными терминами.

Таблица 9. Распределение ошибок системы для задачи А (точное соответствие)

	Рестораны	Автомобили
Word+POS		
Не распознано	65%	68%
Избыточно распознано	35%	32%
Word+POS+Lemma		
Не распознано	63%	65%
Избыточно распознано	37%	35%

Было выделено 4 типа ошибок, которые допустила разработанная система.

1. Технические ошибки

1.1. Специальные символы: не было учтено, что некоторые специальные символы могут оказаться в неправильной кодировке или отображаться в виде кода, а не в виде символа. Например, система не могла корректно обработать и извлечь аспектный термин: «Салат "Цезарь "», система извлекала только «Салат "Цезарь», не учитывая последнюю кавычку.

***1.2. Преобразование всех заглавных букв к строчным*.** В разделе 3.1 описаны этап предобработки и приведение всех заглавных букв к строчным, не было учтено, что некоторые аббревиатуры могут иметь иное значение при написании строчными буквами. Например, в предметной области автомобили «ТО» (техническое обслуживание) является аспектным термином, но после преобразований, получили «то» (частица), которая аспектным термином уже не является. Словарь

аббревиатур и специализированных терминов мог бы помочь избежать этой ошибки.

2. Нераспознанные аспектные термины

***2.1. Сокращения*.** Обе построенные системы не смогли справиться с сокращениями, например, могли извлечь «рубли», но не извлекали «руб.» и «р.».

***2.2. Перечисления*.** Системы плохо справлялись с аспектными терминами, которые перечислялись через запятую. Например, из трех аспектных терминов «Овощи», «Салат Цезарь», «лосось» система могла найти только один («Салат Цезарь») или объединить два или все три аспектных термина в один. Есть предположение, что маркировка знаков препинания могла бы помочь преодолеть эту ошибку.

3. Частично извлеченные аспектные термины

3.1. Не распознает слова до «главного» слова. Сначала было предположение, что система хорошо справляется с существительными и достаточно точно извлекает аспектный термин, если он является существительным, например, «Добавляла **вина**» (система извлекла только «вина», но не «добавляла»). Позже был обнаружен класс частично распознанных аспектных терминов, которые не являются существительными, например, «Официант **хамил**» (система не извлекла «официант»).

3.2. Не распознает слова после «главного» слова. Например, «**местечко** в углу» (извлечено только «**местечко**»). Стоит заметить, что ошибок типа 3.2 было значительно меньше, чем ошибок типа 3.1.

4. Избыточно извлеченные аспектные термины

Системы не очень хорошо отработали на именованных сущностях, например, система могла извлечь «Александр» в качестве аспекта, что таковым не является.

Из таблицы 9 видно, что добавление лемм в качестве признаков ведет к увеличению избыточно извлеченных аспектных терминов. Было проведено сравнение двух систем между собой и обнаружено, что система 2 лучше справляется со словосочетаниями. Например, система 2 извлекала аспект «суп из утки», вместо просто «суп», который извлекла система 1. Однако словосочетание не только

положительно сказывается на результатах системы 2, но также влечет к увеличению избыточно извлеченных аспектов. Например, система 2 извлекла в качестве аспекта: «пасту с морепродуктами, мужу».

ЗАКЛЮЧЕНИЕ

В статье представлены системы по извлечению аспектных терминов из текста, построенные методом условных случайных полей. Показано, что использование даже небольшого количества признаков для CRF дает неплохие результаты. Таким образом, системы показали результаты, сравнимые с лучшими результатами участников SentiRuEval, особенно по метрике полноты для извлеченных аспектов.

В дальнейшем планируется добавить статистические методы в качестве признаков для CRF, а также провести исследование и получить ответ на вопрос, как улучшить результаты точности, не снижая результатов полноты.

СПИСОК ЛИТЕРАТУРЫ

1. *Turney P.D.* Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002. P. 417-424.

2. *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. Association for Computational Linguistics, 2002. V. 10. P. 79-86.

3. *Рубцова Ю.В.* Разработка и исследование предметно независимого классификатора текстов по тональности // Труды СПИИ РАН. 2014. Т. 5, № 36. С. 59-77.

4. *Wilson T., Wiebe J., Hoffmann P.* Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis // Computational linguistics. 2009. V. 35, No 3. P. 399-433.

5. *Liu B.* Sentiment analysis and opinion mining // Synthesis Lectures on Human Language Technologies. 2012. V. 5, No 1. P. 1-167.

6. *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.

7. *Marrese-Taylor E., Velásquez J.D., Bravo-Marquez F.* A novel deterministic approach for aspect-based opinion mining in tourism products reviews // Expert Systems with Applications. 2014. V. 41, No 17. P. 7764-7775.

8. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Yu.V., Ivanov V.V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog–2015. 2015. P. 3-9.

9. *Hu M., Liu B.* Mining and summarizing customer reviews // Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004. P. 168-177.

10. *Popescu A.M., Etzioni O.* Extracting product features and opinions from reviews // Natural Language Processing and Text Mining. Springer London, 2007. P. 9-28.

11. *Moghaddam S., Ester M.* ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews // Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011. P. 665-674.

12. *Jin W., Ho H.H., Srihari R.K.* OpinionMiner: a novel machine learning system for web opinion mining and extraction // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009. P. 1195-1204.

13. *Jakob N., Gurevych I.* Extracting opinion targets in a single-and cross-domain setting with conditional random fields // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. P. 1035-1045.

14. *Titov I., McDonald R.* Modeling online reviews with multi-grain topic models // Proceedings of the 17th International Conference on World Wide Web. ACM, 2008. P. 111-120.

15. *Brody S., Elhadad N.* An unsupervised aspect-sentiment model for online reviews // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. P. 804-812.

16. *Hofmann T.* Unsupervised learning by probabilistic latent semantic analysis // Machine learning. 2001. V. 42, No 1-2. P. 177-196.

17. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. V. 3. P. 993-1022.

18. *Zhao W.X. et al.* Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. P. 56-65.

19. *Mukherjee A., Liu B.* Aspect extraction through semi-supervised modeling // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012. P. 339-348.

20. *Sutton C., McCallum A.* An introduction to conditional random fields for relational learning // Introduction to Statistical Relational Learning. 2006. P. 93-128.

21. *McCallum A.K.* MALLET: A Machine Learning for Language Toolkit. 2002.

22. *Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, Dagmar Divjak*. Designing and evaluating a russian tagset // LREC. 2008.

EXTRACTION OF ASPECTS OF GOODS AND SERVICES FROM CONSUMERS REVIEWS USING CONDITIONAL RANDOM FIELDS MODEL

Yu.V. Rubtsova[1], S. A. Koshelnikov [2]

The A.P. Ershov Institute of Informatics Systems
[1]yu.rubtsova@gmail.com, [2] koshelnikovsa@gmail.com

Abstract

This paper describes the Information extraction system that was presented at SentiRuEval-2015: aspect-based sentiment analysis of users' reviews in Russian. The proposed system uses a conditional random field algorithm to extract aspect terms mentioned in the text. A set of morphological features was used for machine learning. The system intent to perform two subtasks, Task A – automatic extraction of explicit aspects and Task B – automatic extraction of all aspects (explicit, implicit and sentiment facts), and tested on two domains: restaurants and automobiles. Our systems performed competitively and showed the results comparable to those of the other 10 participants.

Keywords: *information retrieval, CRF, aspect extraction, content analysis.*

REFERENCES

1. *Turney P.D.* Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002. P. 417-424.

2. *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. Association for Computational Linguistics, 2002. V. 10. P. 79-86.

3. *Rubtsova Yu.V.* Rasrabotka I issledovanie predmetno nezavisimogo classifikatora tekstov po tonalnosti // Trudi SPII RAN. T. 5, № 36. S. 59-77.

4. *Wilson T., Wiebe J., Hoffmann P.* Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis // Computational linguistics. 2009. V. 35, No 3. P. 399-433.

5. *Liu B.* Sentiment analysis and opinion mining // Synthesis Lectures on Human Language Technologies. 2012. V. 5, No 1. P. 1-167.

6. *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.

7. *Marrese-Taylor E., Velásquez J. D., Bravo-Marquez F.* A novel deterministic approach for aspect-based opinion mining in tourism products reviews // Expert Systems with Applications. 2014. V. 41, No 17. P. 7764-7775.

8. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Yu.V., Ivanov V.V., Tutubalina E.* SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog–2015. 2015. P. 3-9.

9. *Hu M., Liu B.* Mining and summarizing customer reviews // Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004. P. 168-177.

10. *Popescu A.M., Etzioni O.* Extracting product features and opinions from reviews // Natural Language Processing and Text Mining. Springer London, 2007. P. 9-28.

11. *Moghaddam S., Ester M.* ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews // Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011. P. 665-674.

12. *Jin W., Ho H.H., Srihari R.K.* OpinionMiner: a novel machine learning system for web opinion mining and extraction // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009. P. 1195-1204.

13. *Jakob N., Gurevych I.* Extracting opinion targets in a single-and cross-domain setting with conditional random fields // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. P. 1035-1045.

14. *Titov I., McDonald R.* Modeling online reviews with multi-grain topic models // Proceedings of the 17th International Conference on World Wide Web. ACM, 2008. P. 111-120.

15. *Brody S., Elhadad N.* An unsupervised aspect-sentiment model for online reviews // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. P. 804-812.

16. *Hofmann T.* Unsupervised learning by probabilistic latent semantic analysis // Machine learning. 2001. V. 42, No 1-2. P. 177-196.

17. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. V. 3. P. 993-1022.

18. *Zhao W.X. et al.* Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. P. 56-65.

19. *Mukherjee A., Liu B.* Aspect extraction through semi-supervised modeling // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012. P. 339-348.

20. *Sutton C., McCallum A.* An introduction to conditional random fields for relational learning // Introduction to Statistical Relational Learning. 2006. P. 93-128.

21. *McCallum A.K.* MALLET: A Machine Learning for Language Toolkit. 2002.

22. *Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, Dagmar Divjak*. Designing and evaluating a russian tagset // LREC. 2008.

СВЕДЕНИЯ ОБ АВТОРАХ

РУБЦОВА Юлия Владимировна – аспирант Института систем информатики им. А.П. Ершова СО РАН, г. Новосибирск.

Yuliya Vladimirovna RUBTSOVA – Currently is a graduate student at The A.P. Ershov Institute of Informatics Systems (IIS), Siberian Branch of the Russian Academy of Sciences. Current scientific interests: sentiment analysis, automatic text labeling, information extraction, morphological and syntactic analysis, social networks analysis.

email: yu.rubtsova@gmail.com

КОШЕЛЬНИКОВ Сергей Андреевич – разработчик программного обеспечения.

Serge Andreevish KOSHELNIKOV, current Software developer, Project Manager. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: koshelnikovsa@gmail.com

Материал поступил в редакцию 15 июля 2015 года