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