SentiRuEval: тестирование системы анализа тональности текстов на английском языке по запросу к заданному объекту

Лукашевич Н. В. (louk_nat@mail.ru)1,
Блинов П. Д. (blinoff.pavel@gmail.com)2,
Котельников Е. В. (kotelnikov.ev@gmail.com)2,
Рубцова Ю. В. (yu.rubtsova@gmail.com)3, Иванов
В. В. (nomemm@gmail.com)4, Тутубалина Е.
(тленусик@gmail.com)4

1МГУ им. М. В. Ломоносова, Москва, Россия;
2Вятский государственный гуманитарный университет,
г. Киров, Россия; 3Институт систем информатики им. А.
П. Ершова СО РАН, Новосибирск, Россия; 4Казанский
федеральный университет, Казань, Россия

Статья описания данных, правил и результатов SentiRuEval — тестирование системы автоматического анализа тональности русскоязычных текстов по отношению к заданному объекту или его свойствам. Участникам были предложены два задания. Первое задание было аспектно-ориентированным анализом отзывов о ресторанах и автомобилях; Основная цель этого задания заключалась в том, чтобы найти слова и выражения, выявить важные характеристики сущности (аспектные термины), и классифицировать их по тональности и обобщенным категориям. Второе задание заключалось в анализе истории о репутации заданных компаний. Такие твиты стали либо выражением мнения пользователя о компании, ее продукции или предъявленных услуг, либо негативных или позитивных фактов, которые обнаруживаются об этой компании.

Ключевые слова: анализ тональности текстов, оценка качества, раз метка коллекций, оценочные слова.

SentiRuEval: Тестирование Объектно-ориентированное чувство Системы анализа на русском языке

Лукашевич Н.В. (louk_nat@mail.ru)1, Блинов П.Д. (blinoff.pavel@gmail.com)2, Котельников Е.В. (kotelnikov.ev@gmail.com)2, Рубцова Ю.В. (yu.rubtsova@gmail.com)3, Иванов В.В. (nomemm@gmail.com)4, Тутубалина Е. (tlenusik@gmail.com)4

1Московский государственный университет им. М.В. Ломоносова, Москва, Россия; 2Вятский государственный гуманитарный университет, Киров, Россия; 3А. Институт систем информатики им. П. Ершова, Новосибирск, Россия; 4Казанский федеральный университет, Казань, Россия

В статье описаны данные, правила и результаты SentiRuEval, оценки российских систем объектно-ориентированного анализа настроений. Участникам были предложены две задачи. Первой задачей был аспектно-ориентированный анализ отзывов о ресторанах и автомобилях, то есть первоочередной задачей было найти слова и выражения, обозначающие важные характеристики объекта (аспектные термины), а затем классифицировать их по классам полярности и аспектным категориям.
Второй задачей был репутационный анализ твитов, касающихся банков и телекоммуникационных компаний. Целью данного анализа было классифицировать твиты в зависимости от их влияния на репутацию указанной компании. Такие твиты могут выражать мнение пользователя или положительный или отрицательный факт об организации.

Ключевые слова: анализ настроений, отзывы пользователей, маркировка коллекций, аспектные слова, оценка.

1. Введение

В последние годы большое внимание исследователей и практиков привлекает задача автоматического анализа тональности текстов на естественном языке, то есть автоматического извлечения мнений, выраженных в текстах. Это связано с тем, что у этой задачи много полезных приложений. Поэтому анализ и представление мнений пользователей о продуктах и услугах представляет интерес как для их производителей и конкурентов, так и для новых пользователей. Обработка общественного мнения важна для властей для лучшего управления.

Первоначальные подходы к автоматическому анализу тональности пытались определить общую тональность всего текста или предложения (Pang et al., 2002). Этот уровень анализа предполагает, что каждый документ выражает мнение об одном объекте (например, об одном продукте). Позже появилась задача объектно-ориентированного анализа настроений, когда система должна выявлять настроения по отношению к конкретной сущности, упомянутой в тексте (Amigo et al., 2012; Jiang et al., 2011).

Наконец, у автора текста могут быть разные мнения относительно конкретных свойств (или аспектов) объекта. Чтобы выявить эти мнения, необходимо провести так называемый аспектный анализ настроений (Liu, 2012; Bagheri et al., 2013; Glavaš et al., 2013; Popescu, Etzioni, 2005; Zhang, Liu, 2014). Аспекты выражаются в текстах с аспектными терминами и обычно могут быть классифицированы по категориям. Например, категория аспекта «Обслуживание» в отзывах о ресторанах может быть выражена такими терминами, как персонал, официант, официантка, официант.

Автоматический анализ настроений представляет собой сложную проблему обработки естественного языка. Несколько инициатив по оценке были посвящены изучению лучших методов анализа настроений и связанных с ними приложений. Эти инициативы включают отслеживание блога в рамках конференции TREC (Macdonald et al., 2010), задачи TAC Opinion QA Tasks (Dang, Owczarzak, 2008), отслеживание мнений на конференциях NTCIR (Seki et al., 2008), отслеживание управления репутацией на конференции CLEF (Amigo et al., 2012), Twitter и задачи анализа настроений в рамках инициативы SemEval (Nakov et al., 2013; Rosenthal et al., 2014) и др.

В этой статье мы представляем результаты оценки SentiRuEval, ориентированной на сущностно-ориентированный анализ настроений в Твиттере и аспектно-ориентированный анализ отзывов пользователей на русском языке. Эта оценка является вторым российским мероприятием по оценке анализа настроений на русском языке после отслеживаний анализа настроений ROMIP в 2011–2013 гг. В этом году в Sen tiRuEval у нас было два типа заданий. Первая задача — аспектно-ориентированный анализ настроений пользователей. Данные включали отзывы о ресторанах и автомобилях.

Второй задачей был объектно-ориентированный анализ тональности российских твитов, касающихся двух разновидностей организаций: банков и телекоммуникационных компаний.

Структура работы выглядит следующим образом. В разделе 2 мы рассматриваем связанные инициативы оценки в анализе настроений. Раздел 3 описывает задачи, данные и принципы маркировки в обзорном анализе на основе аспектов. Раздел 4 описывает данные и задачу в сущностно-ориентированном анализе настроений в Твиттере. В разделе 5 обсуждаются результаты, полученные участниками.

2. Связанная работа

Несколько инициатив по оценке были посвящены задачам анализа тональности, аналогичным текущей оценке SentiRuEval.

В последние годы в рамках конференции SemEval были организованы два типа оценок сентимент-анализа: сентимент-анализ в Twitter и аспектный сентимент-анализ отзывов. В задаче Twitter одна из подзадач была задачей уровня сообщения, то есть участвующие системы должны классифицировать, имеет ли сообщение положительное, отрицательное или нейтральное настроение (Nakov et al., 2013; Rosenthal et al., 2014). Задача направлена на выявление, а именно, авторского мнения в отличие от нейтральной или объективной информации.

В рамках инициативы CLEF (http://www.clef-initiative.eu/) в 2012–2014 гг. были организованы оценки RebLab, посвященные мониторингу репутационных твитов. В задачи входило определение полярности репутационной классификации.

Цель состояла в том, чтобы решить, имеет ли содержание твита положительное или отрицательное влияние на репутацию компании. Организаторы подчеркивают, что полярность для репутации существенно отличается от стандартного анализа настроений, который должен различать субъективные оценки.

из объективной информации. При анализе полярности для репутации необходимо учитывать как факты, так и мнения, чтобы определить, какое значение может иметь часть информации для репутации данного объекта (Amigo et al., 2012; Amigo et al., 2013).

Оценка аспектно-ориентированного обзорного анализа в SemEval впервые была организована в 2014 г. (Pontiki et al., 2014). Набор данных включал отдельные, вырванные из контекста предложения (не полные обзоры) в двух доменах: рестораны и ноутбуки. Для обучения в каждом домене было подготовлено 3 тыс. предложений. Набор категорий аспектов для ресторанов: еда, обслуживание, цена, атмосфера, анекдоты/разное.

В 2015 году SemEval оценивает анализ отзывов на основе аспектов (http://alt.qcri.org/semeval2015/task12/) и фокусируется на обзорах целиком. Аспектные категории терминов усложнились и теперь состоят из пар Сущность-Атрибут (E#A). Реестр E#A для домена ресторанов содержит 6 типов объектов (RES TAURANT, FOOD, DRINKS, SERVICE, AMBIENCE, LOCATION) и 5 атрибутов (GENERAL, PRICES, QUALITY, STYLE_OPTIONS, MISCELLANEOUS). Домен ноутбуков содержит 22 типа сущностей и 9 меток атрибутов.

В 2011–2013 годах было организовано два мероприятия по оценке российских систем анализа настроений. Первая оценка была посвящена извлечению общего настроения из обзоров пользователей в трех областях: фильмы, книги и цифровые камеры. За обучение участникам были предоставлены отзывы от рекомендательных сервисов. Оценка проводилась на постах в блогах, извлеченных с помощью блог-сервиса Яндекса (Четвиоркин и др., 2012). Вторая оценка предложила участникам две новые задачи, а именно: извлечение общей тональности цитирования (прямой или косвенной речи) из новостных статей и поиск ориентированной на тональность информации в блогах, когда на запрос (из вышеуказанных доменов) мнения пользователей в блоге сообщения должны быть найдены (Четвиоркин, Лукашевич, 2013).

3. Способы выражения мнения об аспектах

Аспектные термины также можно разделить на несколько категорий. Их можно разделить на три подтипа: эксплицитные аспекты, имплицитные аспекты и сентиментальные факты.

Явные аспекты обозначают некоторую часть или характеристики описываемого объекта, например, персонал, пасту, музыку в обзорах ресторанов. Эксплицитные аспекты обычно представляют собой существительные или группы существительных, но в некоторых категориях аспектов мы можем встретить явные аспекты, выраженные в виде глаголов. Например, в ресторанах важной характеристикой качества обслуживания является время ожидания заказа, поэтому эта характеристика может быть указана с глаголом ждать (ждать): ждали больше часа — ждали больше часа.

Неявные аспекты — это отдельные слова или отдельные слова с операторами настроений, которые содержат в себе как специфические настроения, так и четкое указание на категорию аспекта. В отзывах о ресторанах частыми имплицитными аспектами являются такие слова, как вкусно (положительно+еда), комфортно (положительно+интерьер), некомфортно (негативно+интерьер). Важность этих слов для автоматических систем состоит в том, что имплицитные аспекты позволяют системе настроений выявлять мнение пользователя о характеристиках объекта, даже если термин явного аспекта неизвестен, написан с ошибкой или сложно упомянут.

Факты настроения не упоминают настроение пользователя напрямую, формально они формируют нас только о реальном факте, однако этот факт передает нам настроение пользователя, а также категорию аспекта, к которой оно относится. Например, сантимент-факт использовала на все во просы (ответил на все вопросы) означает положительную характеристику ресторанного обслуживания; это выражение достаточно часто встречается в отзывах о ресторанах.

В маркировке SentiRuEval мы аннотировали эти три подтипа аспектных терминов, и наши задачи для участников заключались не только в извлечении явных аспектных терминов, но и в извлечении всех аспектных терминов (см. Раздел 4).

Мнение об аспектах может быть выражено несколькими способами.

Прямой способ передачи мнения заключается в использовании слов-мнений, таких как как хороший, плохой, отличный, ужасный, нравится, ненавижу и т. д.

Мнения могут быть сформулированы как сравнения с другими объектами, предыдущими случаями или мнениями других людей (Liu, 2012; Jindal, Liu, 2006). Проблема автоматического анализа в этих случаях возникает из-за того, что используемые положительные или отрицательные слова могут не иметь отношения к текущему отзыву. Кроме того, сравнение может быть поставлено различными способами, не только с помощью сравнительных конструкций. Например, в следующей выдержке из обзора ресторана сравнение отмечено словом « другой», а положительные слова «насладился » и «замечательный» характеризуют ресторан, отличный от ресторана, о котором идет речь:

> Мы решили не есть там десерт и кофе, а вместо этого пошли в другой ресторан, где прекрасно завершили наш вечер.

Мы можем сформулировать свое мнение как рекомендацию (конструктивное или суггестивное мнение — см. (Arora, Srinivasa, 2014)) или описание желаемой ситуации или характеристик объекта, так называемые ирреальные факторы (Taboada et al., 2011; Kusnetsova et al., 2011; др., 2013). В этих случаях упомянутые положительные слова могут скрывать отрицательное мнение.

Наконец, мнение можно выразить с помощью иронии или сарказма (Barbieri, Saggion, 2014; Riloff et al., 2013). В таких случаях мнение может выглядеть как положительное или хотя бы среднее, а на самом деле оно резко отрицательное, как в следующем примере: «Отличный перевод, ничего не понимаю».

В маркировке SentiRuEval мы отметили эти подтипы мнений для дальнейшего исследования (см. Раздел 4).

4. Маркировка и задачи аспектного анализа отзывов на SentiRuEval

Для оценки систем аспектно-ориентированного анализа тональности мы выбрали два основных направления: обзоры ресторанов и обзоры автомобилей. В обзорах ресторанов к аспектным категориям относятся: ЕДА, СЕРВИС, ИНТЕРЬЕР (включая атмосферу), ЦЕНА, ОБЩИЕ СРЕДСТВА. Для автомобилей категории аспектов: УПРАВЛЯЕМОСТЬ, НАДЕЖНОСТЬ, БЕЗОПАСНОСТЬ, ВНЕШНИЙ ВИД, КОМФОРТ, ЗАТРАТЫ, ОБЩИЕ СВЕДЕНИЯ.

Длина обзоров может сильно варьироваться от одного короткого предложения до длинного повествования. Могут быть и сдвиги в ту или иную сторону. В качестве эксперимента для маркировки в ресторанном домене мы попытались извлечь наиболее типичные отзывы.

из нашей коллекции. Для ее достижения была проведена следующая процедура. Мы представили каждый обзор в виде вектора набора слов и рассчитали вектор глобальной коллекции, усредняя все отдельные векторы. Затем мы наложили ограничения на минимальную и максимальную длину обзора и выбрали наиболее похожие обзоры в соответствии с косинусным сходством между глобальным вектором и векторами одиночного обзора. В результате для маркировки были выбраны наиболее типичные представители обзоров.

Разметка обучающих и тестовых данных проводилась с помощью инструмента аннотирования BRAT (Stenetorp et al., 2012). Аннотаторы имели доступ к обзорам коллекций через веб-интерфейс. Для унификации и согласования процедуры аннотирования было подготовлено руководство для оценщика1 .

Он основан на рекомендациях по аннотации SemEval-2014 (Pontiki et al., 2014).

Задача аннотации состояла в том, чтобы разметить два основных типа токенов: термины аспектов в обзоре и категории аспектов, прикрепленные ко всем обзорам. Категории аспектов были помечены общей оценкой настроений, выраженных в тексте: положительное, отрицательное, и то, и другое или отсутствующее.

В соответствии с описанной выше категоризацией мнений и аспектов, аннотация аспектных терминов в тексте включала несколько измерений:

1. Сначала аннотаторы должны указать явные аспекты, неявные аспекты или факты настроений в обзорных текстах и присвоить им соответствующий тип (явный, неявный или факт).

2. Все термины аспектов должны быть отнесены к категориям аспектов целевого объекта.
3. Аннотаторы отметили полярность термина аспекта: положительный, отрицательный, нейтральный, или оба.
4. Аннотаторы отметили актуальность термина для обзора: a. Rel – релевантный (к текущему обзору), b. Cmpr — сравнение, т. е. термин относится к другому объекту, c. Prev — предыдущий, то есть термин связан с предыдущими мнениями, d. Irr — irrealis, то есть термин является частью рекомендации или описания.

желаемой ситуации, т.е. Ирн – ирония.

Так, например, аннотация слова девушка в контексте милая де вушка (милая девушка) в обзоре ресторана включает направленность настроения — положительную, категорию аспекта — услугу, знак аспекта — релевантный, тип аспекта — явный.

Такой подробный процесс аннотирования очень трудоемкий. Таким образом, каждый обзор был помечен только одним оценщиком. Однако для проверки качества маркировки аспектов после завершения маркировки были выполнены две процедуры. Сначала из разметки были извлечены все размеченные аспектные термины в соответствии с их типами и категориями и просмотрены; поэтому некоторые случайные ошибки были найдены и исправлены.
Во-вторых, мы сравнили тональность аспекта, присвоенную обзору в целом, и тональность конкретных терминов в этом обзоре. В случаях различий между этими двумя видами маркировки дополнительно проверялась разметка рецензии.

Во время процедуры аннотации балансировка по тональности или аспектным терминам не производилась; мы старались, чтобы естественные дистрибутивы были специфичны для обзоров в данном домене. Некоторая статистика по релевантным терминам (Rel) представлена в таблице 1.

---

1 Руководство доступно по адресу http://goo.gl/Wqsqit.

Таблица 1. Статистика корпуса

		Рестораны		Автомобили	
		Тренироваться	Тест	Тренироваться	Тест
Количество отзывов		201	203	217	201
Количество терминов, которые	явный	2822	3506	3152	3109
	неявный	636	657	638	576
	факт	523	656	668	685
Количество терминов, которые	положительный	2 530	3 424	2 330	2 499
	отрицательный	684	865	1 337	1 300
	нейтральный	714	445	691	456
	оба	53	85	100	115

Размеченные данные позволили предложить участникам следующие задачи:

- Задача A: автоматическое извлечение явных аспектов,
- Задача B: автоматическое извлечение всех аспектов, включая факты тональности, • Задача C: извлечение настроений по отношению к явным аспектам, • Задача D: автоматическая категоризация явных аспектов в категории аспектов, • Задача E : анализ тональности всего обзора по категориям аспектов.

Для оценки автоматических систем использовались следующие меры качества.

Для задач A и Б применялся макрос F1-мера в двух вариантах: точное совпадение и частичное совпадение. Макро F1-мера в данном случае означает вычисление F1-меры для каждого обзора и усреднение полученных значений.

Чтобы измерить частичное совпадение для каждого термина аспекта золотого стандарта t , мы вычисляем точности и отзыва следующим образом:

$$\text{Точность} \quad \text{"="} \ | \tau \quad \tau_c \ |$$

$$\text{Точность} \quad \text{"="} \ \frac{| \tau | \tau_\tau \ |}{| \ \tau_c \ c \ c \ |}, \quad | \ \tau_c \ |$$

$$\text{Вспомнить} \quad \text{"="} \ \frac{| \tau\tau \quad c \ |}{\text{"="} \ | \ \tau\tau\tau \ | \ _c \ |}$$

Вспомните , где ts — извлеченный термин аспекта, котöрый пересекается с термином t, t ts — пересечение между терминами t и ts , |t| длина срока в токенах. Таким образом, F1-мера вычисляется для каждого термина, а затем мы усредняем значения для всех терминов золотого стандарта.

Для оценки тональности терминов-аспектов (задача C) использовались оба варианта F1-меры (макро- и микро-). Расчет макро-F1-меры основан на отдельном расчете точности, полноты и F-меры для каждой рассматриваемой категории, затем полученные значения усредняются. Это позволяет нам оценивать качество категоризации одинаково для каждой категории. Микро-мера F1 рассчитывается по глобальной матрице путаницы, эта мера сильно зависит от дисбаланса в распределении классов.

Для аспектной категоризации терминов (задача D) и анализа тональности отзывов в целом (задача E) использовался макрос F1-мера.

Таблица 2. Результаты аспектно-ориентированного анализа отзывов (домен ресторана)

Мера задачи		Базовый уровень	Результаты участников	Идентификатор участника
А	Точное соответствие, Макро F	0,608	0,632	2
			0,627	1
А	Частичное совпадение, Макро F	0,665	0,728	4
			0,719	1
Б	Точное соответствие, Макро F	0,587	0,600	1
			0,596	2
Б	Частичное совпадение, Макро F	0,619	0,668	1
			0,645	1
С	Макро F	0,267	0,554	4
			0,269	3
С	Микро Ф	0,710	0,824	4
			0,670	3
Д	Макро F	0,800	0,865	8
			0,810	4
Е	Макро F	0,272	0,458	4
			0,372	10

Для всех задач мы подготовили базовые прогоны. Базовая система для задач А и В извлекает список помеченных терминов из обучающей коллекции, лемматизирует их и применяет к лемматизированному представлению тестовой коллекции. Если несколько терминов соответствуют одной и той же последовательности слов, то предпочтительным является более длинный термин.

Базовые системы задач С и D относят термин аспекта к его наиболее часто встречающейся категории в обучающей коллекции. Если термин отсутствует в обучающей коллекции, то применяется наиболее часто встречающаяся категория аспекта. Базовый уровень задачи Е является наиболее частой категорией тональности для данной категории аспектов (положительной во всех случаях).

В общей сложности 12 участников с 21 прогоном участвовали в задачах анализа настроений обзора. Из-за ограничений по объему здесь представлены только два лучших результата в каждой задаче и только первичная F-мера, полные результаты доступны на http://goo. гл/Wqsqit. В таблице 2 представлены результаты участников по отзывам о ресторанах, а в таблице 3 — результаты по отзывам об автомобилях. Автомобильные обзоры привлекли гораздо меньше внимания участников.

Из таблиц 2, 3 видно, что базовые уровни для выделения аспектных терминов (задачи А и Б) достаточно высоки, что означает значительное совпадение аннотаций обучающей и тестовой коллекций. Лучшие методы в этих задачах были основаны на распределенных подходах, дополненных набором правил (участник 4) и рекуррентными нейронными сетями (участник 1). Для точного сопоставления аспектов наилучшие результаты были достигнуты при мечении последовательностей с помощью SVM по богатому набору морфологических, синтаксических и семантических признаков (участник 2).

Таблица 3. Результаты аспектно-ориентированного анализа отзывов (автомобильная область)

Мера задачи		Базовый уровень	Результаты участников	Идентификатор участника
A	Точное соответствие, Макро F	0,594	0,676	2
			0,651	1
A	Частичное совпадение, Макро F	0,697	0,748	1
			0,730	2
Б	Точное соответствие, Макро F	0,589	0,636	2
			0,630	1
Б	Частичное совпадение, Макро F	0,674	0,714	1
			0,704	1
С	Макро F	0,264	0,568	4
			0,342	1
С	Микро Ф	0,619	0,742	4
			0,647	1
Д	Макро F	0,564	0,652	8
			0,607	4
Е	Макро F	0,237	0,439	4

Наилучший результат в анализе отношения к терминам аспекта (задача C) был получен с классификатором Gradient Boosting Classifier (участник 4). Функции были основаны на модели скип-грамм, использующей контексты слов для лучшего изучения векторных представлений и точечной взаимной информации. В задаче категоризации явных аспектных терминов (задача D) наилучшие результаты показал метод SVM с признаками, основанными на точечной взаимной информации (участник 8). Второй результат дает метод, основанный на сходстве терминов в пространстве распределенных представлений слов (участник 4). Для задания E наилучшие результаты были достигнуты при объединении результатов, полученных в заданиях A, C и D (участник 4).

5. Объектно-ориентированный анализ тональности твитов

Цель анализа настроений в Твиттере на SentiRuEval состояла в том, чтобы найти ориентированные на настроения мнения или положительные и отрицательные факты о двух типах организаций: банках и телекоммуникационных компаниях. Эта задача очень похожа на задачу полярности репутации при оценке Re pLab (Amigo et al., 2013).

Обучающая и тестовая коллекции твитов были снабжены полями, соответствующими всем возможным организациям, для которых были извлечены твиты. Конкретная организация, упомянутая в данном твите, была отмечена меткой «0», что означает «нейтральность» в качестве значения по умолчанию. Аннотаторы и участвующие системы должны оставить это значение без изменений, если твит считается нейтральным, или заменить значение на «1» (положительное) или «-1» (отрицательное). Аннотаторы также могут помечать твиты знаком «--», что означает =бессмысленность=, или знаком «+-», что означает положительные и отрицательные настроения в одном и том же твите. Оба последних случая были исключены из оценки.

Для обучения и тестирования коллекций асессоры пометили 5000 твитов в каждом домене (всего было помечено 20000 твитов). Важно подчеркнуть, что обучающая и тестовая коллекции были выпущены в разные промежутки времени. Твиты обучающей коллекции были написаны в 2014 г., твиты тестовой коллекции опубликованы в 2013 г.

Таблица 4. Результаты процедуры голосования по маркировке тестовой коллекции твитов

Домен	Количество твитов с одинаковыми метками не менее чем от 2 оценщиков	Полное совпадение маркировки 3 816	Окончательное количество твитов в тестовой коллекции
банки	4 915 (98,30%) 4	(76,36 %) 2 233	4 549
Телекоммуникационные компании	503 (90,06%)	(44,66 %)	3 845

Анализируя разметку обучающей коллекции, мы обнаружили, что оценка некоторых твитов может вызвать серьезное обсуждение их тональности. Для уменьшения субъективности маркировки, а также случайных ошибок, тестируемая коллекция маркировалась тремя асессорами, а для получения результатов маркировки применялась схема голосования. Наконец, из коллекции были удалены неактуальные твиты. Результаты подготовки коллекции представлены в табл. 4.

Участвующие системы должны были выполнять трехстороннюю классификацию твитов: положительные, отрицательные или нейтральные. В качестве основной меры качества использовали макросреднюю F-меру, рассчитываемую как среднее значение между F-мерой положительного класса и F-мерой отрицательного класса. Поэтому мы проигнорировали Fneutral, потому что эта категория обычно никому не интересна. Но это не сводит задачу к двухклассовому прогнозу, так как ошибочная маркировка нейтральных твитов негативно влияет на Fpos и Fneg. Кроме того, для двух классов тональности были рассчитаны микроусредненные F-меры.

Таблица 5. Результаты участников в задачах классификации твитов. Идентификаторы участников обзора и задач Твиттера различаются

Домен Измерение базовых показателей Результаты участников Идентификатор участника				
Телекоммуникационный макрос F		0,182	0,488	2
			0,483	2
			0,480	3
Телеком Микро Ф		0,337	0,536	2
			0,528	10
			0,510	3
банки	Макро F	0,127	0,360	4
			0,352	10
			0,335	2
банки	Микро Ф	0,238	0,366	2
			0,364	2
			0,343	8

В таблице 5 мы представляем лучшие результаты анализа тональности твитов для каждого домена и показателя. В большинстве лучших подходов к этой задаче использовался метод классификации SVM. Признаками участника 2 были синтаксические связи, представленные в виде троек (главное слово, зависимое слово, тип отношения). Участник 3 применил основанный на правилах метод учета синтаксических отношений между эмоциональными словами и целевыми объектами без какого-либо машинного обучения.

Дополнительно один из участников выполнил независимую экспертную маркировку телекоммуникационных твитов и получил Макро-Ф — 0,703, а Микро Ф — 0,749, что можно считать максимально возможной производительностью автоматизированных систем.

Анализ полученных результатов показал, что большинство участников решили общую (не сущностно-ориентированную) задачу классификации твитов; Сущностно-ориентированные подходы не дали лучших результатов по сравнению с общими подходами в твитах, в которых упоминалось несколько сущностей.

6. Заключение

В этой статье мы описали данные, правила и результаты SentiRuEval, оценки российских систем объектно-ориентированного анализа настроений. Мы предложили участникам две задачи. Первой задачей был аспектно-ориентированный анализ отзывов о ресторанах и автомобилях, то есть первоочередной задачей было найти слова и выражения, обозначающие важные характеристики объекта (аспектные термины), а затем классифицировать их по классам полярности и аспектным категориям.

Второй задачей был репутационный анализ твитов о банках и телекоммуникационных компаниях. Целью данного анализа было классифицировать твиты в зависимости от их влияния на репутацию указанной компании.
Такие твиты могут выражать мнение пользователя или положительный или отрицательный факт об организации.

В каждом задании принимало участие около десяти участников из вузов и промышленности. Они применили различные подходы к машинному обучению, включая SVM, повышение градиента, CRF, рекуррентные нейронные сети и другие. Учитывая результаты участников, можно сделать вывод, что объектно-ориентированный анализ настроений плохо решается применяемыми методами. И большинство систем и методов нуждаются в значительном улучшении, чтобы лучше справляться с такими задачами.

В обзорных сборниках также были отмечены интересные языковые явления. В частности, мы назвали сравнения с другими сущностями или с предыдущими мнениями, желательными, но не существующими ситуациями, иронией. Так что изучение разметки может быть полезно и лингвистам. Все подготовленные материалы доступны для исследовательских целей (обзоры: http://goo.gl/Wqsqit и твиты: http://goo.gl/qHeAVo).

Благодарности

Работа частично поддержана грантами РФФИ № 14-07-00682, № 15-07-09306
Минобрнауки России, НИЧ № 586.

Рекомендации

1. Amigo E., Corujo A., Gonzalo J., Meij E., de Rijke M. (2012), Обзор RepLab 2012: Оценка систем управления онлайн-репутацией, CLEF 2012 Evaluation Labs and Workshop Notebook Papers, Рим.
2. Амиго Э., Альборнос Дж. К., Чугур И., Корухо А., Гонсало Дж., Мартин Т., Мейдж Э., де Райке М., Спина Д. (2013), Обзор RepLab 2013: Оценка мониторинга онлайн-репутации Системы, CLEF 2013, Конспект лекций по информатике, том 8138, стр. 333–352.

3. Арора Р., Сриниваса С. (2014), Многогранная характеристика ландшафта добычи мнений, Семинар COMSNETS по науке и разработке социальных сетей, Бангалор, стр. 1–6.

4. Багери А., Сараи М., де Йонг Ф. (2013), Модель обнаружения аспектов без присмотра для анализа настроений обзоров, в информационных системах и системах обработки естественного языка, Springer, Берлин, Гейдельберг, стр. 140–151.
5. Барбьери Ф., Сагион Х. (2014), Моделирование иронии в Твиттере: анализ и оценка характеристик, Труды LREC, стр. 4258–4264.
6. Четверкин И.И., Браславский П.И., Лукачевич Н.В. (2012), Трек анализа настроений на РОМИП 2011, Материалы международной конференции «Диалог», стр. 739–746.
7. Четвирокин И., Лукашевич Н. (2013), Трек анализа настроений на РОМИП-2012, Материалы международной конференции «Диалог», том 2, стр. 40–50.
8. Данг Х.Т., Овчарзак К. (2008), Обзор задач TAC 2008, ответы на вопросы и подведение итогов, Материалы первой конференции по анализу текста.

9. Главаш Г., Коренчич Д., Шнайдер Й. (2013), Аспектно-ориентированный анализ мнений на основе отзывов пользователей на хорватском языке, Материалы 4-го семинара по балто-славянской обработке естественного языка, стр. 18–22.
10. Цзян Л., Ю М., Чжоу М., Лю С., Чжао Т. (2011), Целезависимая классификация настроений в Твиттере, Труды 49-го ежегодного собрания Ассоциации компьютерной лингвистики, стр. 151– 160.
11. Джиндал Н., Лю Б. (2006), Анализ сравнительных предложений и отношений, Материалы 21-й Национальной конференции по искусственному интеллекту, Бостон, стр. 1331–1336.

12. Кузнецова Е., Лукачевич Н., Четверкин И. (2013), Правила тестирования системы анализа настроений, Материалы международной конференции «Диалог», стр. 71–80.
13. Лю Б. (2012), Анализ настроений и изучение мнений, Обобщающие лекции. по технологиям человеческого языка, Vol. 5(1).
14. Макдональд К., Сантос Р., Оунис И., Соборофф И. (2010), Исследование блога в TREC, Форум ACM SIGIR, Vol. 44 (1), стр. 58–75.
15. Мохаммад С.М., Кириченко С., Чжу С. (2013), NRC-Канада: Создание современного состояния анализа тональности твитов, Материалы 7-го Международного семинара по упражнениям семантической оценки (SemEval-2013), Атланта, стр. 321–327.
16. Наков П., Козарева З., Риттер А., Розенталь С., Стоянов В., Уилсон Т. (2013), Se meval-2013 Task 2: Sentiment Analysis in Twitter, Материалы 7-го Международного семинара по семантике. Оценка (SemEval-2013), Атланта, стр. 312–320.

17. Панг Б., Ли Л., Вайтьянатан С. (2002), Недурно? Классификация настроений с использованием методов машинного обучения, Материалы конференции ACL-02 по эмпирическим методам обработки естественного языка, Vol. 10, стр. 79–86.
18. Понтики М., Галанис Д., Павлопулос Дж., Папагеоргиу Х., Андруцопулос И., Манандхар С. (2014), SemEval-2014 Задача 4: Аспектный анализ тональности, Материалы 8-го Международного семинара по семантической оценке (Se mEval 2014), Дублин, стр. 27–35.

19. Попеску А. М., Этциони О. (2005), Извлечение характеристик продукта и мнений из обзоров, Труды конференции по технологии человеческого языка и конференции по эмпирическим методам обработки естественного языка (HLT/EMNLP), Ванкувер, стр. 339–346. .

20. Рилофф Э., Кадир А., Сурв П., Де Сильва Л., Гилберт Н., Хуанг Р. (2013), Сарказм как контраст между позитивным настроением и негативной ситуацией, Материалы конференции по эмпирическим методам в Обработка естественного языка (EMNLP), стр. 704–714.

21. Розенталь С., Риттер А., Наков П., Стоянов В. (2014), SemEval-2014 Task 9: Sentiment Analysis in Twitter, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Дублин, стр. 73–80.
22. Секи Ю., Эванс Д.К., Ку Л.В., Сан Л., Чен Х.Х., Кандо Н. (2008), Обзор многоязычной задачи анализа мнений на NTCIR-7, Материалы рабочей встречи NTCIR-7, Токио, стр. 185–203.

23. Стенеторп П., Пюйсало С., Топич Г., Охта Т., Ананиаду С., Цуджи Дж. (2012), BRAT: веб-инструмент для аннотации текста с помощью НЛП, Материалы демонстраций в 13-я конференция Европейского отделения Ассоциации компьютерной лингвистики, Авиньон, стр. 102–107.

24. Табоада М., Брук Дж., Тофилоски М., Фолл К., Стеде М. (2011), Методы анализа настроений на основе лексикона, Вычислительная лингвистика, Том. 37(2), стр. 267–307.
25. Чжан Л., Лю Б. (2014), Извлечение аспектов и сущностей для интеллектуального анализа мнений // Интеллектуальный анализ данных и обнаружение знаний для больших данных, Springer, Берлин, Гейдельберг, стр. 1–40.