Извлечение аспектных терминов на основе условных случайных полей и векторных представлений слов*

Д.О. Машкин <danil.mashkin@gmail.com>
Е.В. Котельников <kotelnikov.ev@gmail.com>
Вятский государственный университет,
610000, г. Киров, ул. Московская, д.36

Аннотация. В интернете существует множество площадок, которые предоставляют пользователям возможность обмениваться своими мнениями и оставлять отзывы о всевозможных товарах и услугах. Эти мнения могут быть полезны не только для других пользователей, но и для компаний, которые хотят отслеживать собственную репутацию и получать своевременные отзывы о своих продуктах и услугах. Наиболее детальная постановка задачи в данной области ставится при аспектно-ориентированном анализе тональности, в котором определяется отношение пользователя не только к объекту в целом, но и к отдельным его аспектам. В настоящей работе рассмотрено решение подзадачи извлечения аспектных терминов при аспектно-ориентированном анализе тональности. Представлен обзор исследований в данной области. Подзадача извлечения аспектных терминов рассматривается как проблема разметки последовательности; для её решения применяется модель условных случайных полей (CRF). Для составления признакового описания последовательности используются векторные представления слов, полученные на основе нейросетевых моделей для русского языка, а также части речи анализируемых слов. Представлены этапы работы программной системы извлечения аспектных терминов. Эксперименты с разработанной программной системой проводились на размеченном корпусе отзывов о ресторанах, созданном в рамках международного тестирования SemEval-2016 Task 5. Исследованы зависимости качества решения подзадачи извлечения аспектных терминов от различных нейросетевых моделей и вариаций признаковых описаний. Наилучшие результаты (F1-мера = 69%) демонстрирует вариант системы, учитывающий контекст и части речи. Работа содержит подробный анализ ошибок, допущенных системой, а также предложения по возможным вариантам их коррекции. В заключении представлены направления дальнейших исследований.

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

Ключевые слова: аспектно-ориентированный анализ тональности; извлечение аспектных терминов; машинное обучение; разметка последовательностей слов; векторное представление слов; word2vec; SemEval-2016.

DOI: 10.15514/ISPRAS-2016-28(6)-16

Для цитирования: Машкин Д.О., Котельников Е.В. Извлечение аспектных терминов на основе условных случайных полей и векторных представлений слов. Труды ИСП РАН, том 28, вып. 6, 2016 г., стр. 223-240. DOI: 10.15514/ISPRAS-2016-28(6)-16

1. Введение

Интернет-ресурсы являются удобной платформой для коммуникаций, обсуждения и поиска новой информации. Мнения пользователей, содержащие определенную тональность, представляют огромную ценность для компаний, которые хотят отслеживать собственную репутацию и получать своевременные отзывы о своих продуктах и услугах.

Наиболее детальной постановкой задачи в данной области является аспектно-ориентированный анализ тональности (АОАТ) [1], в котором определяется отношение пользователя не только к объекту в целом, но и к отдельным его аспектам. Например, в отзыве о ресторане *«К сервису действительно никаких нареканий и интерьер такого рода мне нравится, но говядина по-французски была ужасна»* можно выделить три аспекта – это *обслуживание*, *интерьер* и *еда*. Тональность аспектов может быть разной. Каждый аспект выражается различными словами или словосочетаниями, которые называются аспектными терминами (АТ). В приведенном примере АТ являются слова *«сервису»*, *«интерьер»* и словосочетание *«говядина по-французски»*. Таким образом, можно выделить четыре основные подзадачи АОАТ [1]:

- извлечение аспектных терминов;
- категоризация аспектных терминов по заданным аспектам;
- определение тональности аспектных терминов;
- определение тональности по отношению к аспектам в целом.

Первое тестирование систем анализа в этой области проходило в рамках семинара SemEval-2014 [2]. Первым российским стало тестирование SentiRuEval в рамках конференции Dialogue-2015 [3]. В 2016 году SemEval стал многоязычным, включив дорожки для английского, русского и других языков [4].

В данной статье описывается программная система для решения первой подзадачи АОАТ – извлечения аспектных терминов на материале русского языка. Подзадача рассматривается как проблема разметки последовательности слов и решается с применением условных случайных полей (Conditional Random Fields, CRF). При этом для обучения модели используются векторные представления слов (ВПС), а также дополнительные морфологические и

словарные признаки. Эксперименты проводятся на корпусе отзывов, предоставленном организаторами международного тестирования SemEval-2016.

Статья организована следующим образом. Во втором разделе приведен краткий обзор исследований в области извлечения аспектных терминов. В третьем разделе описываются текстовые данные, предоставленные организаторами тестирования SemEval-2016. Четвертый раздел посвящен описанию предложенной программной системы. В пятом разделе описываются эксперименты и приводится сравнение метрик качества программной системы с официальными результатами SemEval-2016. Шестой раздел содержит анализ и классификацию ошибок. Седьмой раздел посвящен выводам и описанию перспективных направлений дальнейших исследований.

2. Обзор исследований в области АОАТ

Исследования в области АОАТ получили широкое развитие в большой степени благодаря открытым тестированиям SemEval [2], [4] и SentiRuEval [3]. В работе [5] приведен обзор решений участников тестирований последних лет. В работе [1] представлен подробный обзор подходов и методов решения всех подзадач АОАТ. Для решения подзадачи извлечения АТ, которая является предметом настоящего исследования, используются следующие подходы:

- извлечение аспектов с помощью методов машинного обучения с учителем (supervised learning) [6], [7], [8], [9], [11], [12];
- поиск часто встречающихся имен существительных и именных групп (frequent nouns and noun phrases) [13], [14], [16];
- извлечение аспектов на основе связи между мнением и объектом (relation-based methods) [13];
- извлечение аспектов с помощью тематического моделирования (topic modeling) [19], [20].

Подзадача извлечения АТ подразумевает нахождение слов или словосочетаний в предложении, по отношению к которым выражено мнение пользователя. Определенные слова могут с большей вероятностью сигнализировать о том, что следующие за ними слова входят в состав АТ. Поэтому подход к решению подзадачи извлечения АТ как к проблеме разметки последовательности с помощью модели машинного обучения условные случайные поля является одним из самых распространенных [6], [7], [8], [9], [11], [12].

В работе [6] для извлечения АТ авторы использовали CRF со статистическими и синтаксическими признаками. ВПС [10], ранее использованные для нахождения семантической близости, также широко применяются для задачи извлечения АТ в качестве признаков для CRF [11], [8].

Система, описанная в работе [11], показала лучший результат в тестировании систем извлечения АТ для русского языка SentiRuEval-2015. Авторы

используют семантическую близость слов. Каждому слову из тестовой выборки ставился в соответствие вектор из нейросетевой языковой модели. Слова помечались АТ в зависимости от близости по косинусной мере их ВПС, входящим в состав АТ в обучающей выборке.

В работе [8] описана программная система, занявшая второе место в решении подзадачи извлечения АТ на SemEval-2016 среди неограниченных обучающей выборкой систем. Признаковое описание для слов из последовательности составлялось следующим образом. Использовалось контекстное окно из пяти слов – текущего и по два слова слева и справа от него. Каждому слову из контекстного окна сопоставлялось ВПС из нейросетевой модели. Если ВПС не было найдено в модели, то система назначала слову первое найденное ВПС, сопоставленное соседнему слову слева или справа. Если предыдущим словам также не было назначено ВПС, то текущему слову сопоставлялся один из четырех специально введенных векторов для первых двух слов с начала и двух слов с конца предложения. Также признаковое описание дополнялось морфологическими и лексическими признаками.

Система, показавшая лучшие результаты на SemEval-2016 [12], также основана на CRF и ВПС. Авторы использовали выходные данные рекуррентной нейронной сети в качестве дополнительных признаков.

3. Текстовый корпус

Одна из задач международного тестирования SemEval-2016 была посвящена АОАТ и заключалась в определении мнений, выраженных в отзывах пользователей по отношению к определенным сущностям и их атрибутам. Для русского языка организаторами тестирования был предоставлен корпус отзывов о ресторанах. Отзывы этой предметной области могли содержать мнения о шести сущностях (ресторан, еда, напитки, атмосфера, обслуживание, местонахождение) и пяти их атрибутах (цены, качество, стиль, разное и в целом).

Пятая задача SemEval-2016 делилась на две подзадачи – АОАТ на уровне предложения и АОАТ на уровне текста. Первая подзадача подразумевала поиск мнений пользователей в рамках каждого предложения отзыва отдельно, а вторая подзадача – анализ всего отзыва целиком. Последняя задача – АОАТ на уровне текста – в данной работе не рассматривается. Первая подзадача включала три этапа:

- этап 1: определение категории аспекта (Aspect Category Detection, ACD). Цель этого этапа – найти все пары сущностей и атрибутов, по отношению к которым выражено мнение пользователя. В данной работе пара *сущность#атрибут* называется *аспектом*;
- этап 2: определение объекта, по отношению к которому выражено мнение (Opinion Target Expression, OTE). Данный этап подразумевал нахождение слова или словосочетания, использованного в тексте для

явного указания на сущность, по отношению к которой выражено мнение. В настоящей работе такие объекты называются *аспектными терминами* (АТ);

- этап 3: семантическая полярность (Sentiment Polarity). Каждой паре *сущность#атрибут* необходимо было сопоставить одну из следующих меток полярности: *позитивная*, *негативная* или *нейтральная*.

Предоставленный организаторами размеченный корпус отзывов состоит из обучающей и тестовой выборок. Выборки представляют собой XML-файлы, в которых каждый отзыв разделен на предложения. Для каждого предложения в обучающей выборке выделены АТ (слово или словосочетание с указанием граничных символов), их категории и полярности.

В Табл. 1 приведены характеристики корпуса. Количество отзывов, предложений, слов и АТ в обучающей выборке примерно втрое больше, чем в тестовой.

Табл. 1. Количественные характеристики отзывов SemEval-2016

Table 1. Quantitative characteristics of the reviews SemEval-2016

Характеристика корпуса	Обучающая выборка	Тестовая выборка
Количество отзывов	312	103
Количество предложений	3 548	1 168
Количество слов	40 094	13 181
Количество АТ	3159	972
однословные АТ	2 580	785
многословные АТ	579	187

Следует отметить, что количество АТ, являющихся словосочетаниями в выборках, значительно меньше, чем количество однословных АТ, и составляют соответственно лишь 19.3% и 18.2% от общего числа АТ в выборках. Подобная разница сказывается на качестве обучения модели и ее прогнозам по отношению к АТ, выраженным словосочетаниями (см. раздел 5).

В данной работе описано решение второго этапа (OTE) первой подзадачи (sentence-level ABSA) пятой задачи SemEval-2016, которая подразумевает извлечение АТ, т. е. определение аспектных слов или словосочетаний с указанием их граничных символов.

4. Описание программной системы

В настоящей работе подзадача извлечения АТ рассматривается как проблема разметки последовательностей слов. Идея данного подхода состоит в том, чтобы для каждого предложения построить последовательность признаковых описаний слов. В качестве модели машинного обучения с учителем была

выбрана модель условных случайных полей CRF и ее реализация для языка Python – *pyStruct* [21].

Для представления слова использовались ВПС, часть речи, присутствие слова в словаре терминов, являющихся АТ в обучающей выборке, наличие заглавной буквы в слове и синтаксическая связь с предком. Указанные признаки формируют признаковое описание каждого слова в последовательности.

Для решения подзадачи извлечения АТ программная система должна выполнять следующие функции:

- загрузка обучающей и тестовой выборок отзывов из XML-файлов;
- токенизация – разделение предложений на слова;
- генерация признакового описания для каждого слова обучающей и тестовой выборок;
- присвоение меток признаковым описаниям обучающей выборки;
- обучение модели CRF на последовательности признаковых описаний и меток из обучающей выборки;
- предсказание меток для признаковых описаний из тестовой выборки;
- сопоставление полученным меткам слов из тестовой последовательности для преобразования их в АТ;
- сохранение отзывов тестовой выборки вместе с АТ в XML-файл.

Токенизация каждого предложения проводилась с помощью морфологического анализатора *mystem* [22]. Получая на вход целое предложение, *mystem* возвращает массив слов вместе с леммой и грамматической информацией каждого слова. Лемма используется для поиска ВПС в нейросетевой языковой модели, а часть речи добавляется в признаковое описание.

Основой для создания признакового описания стал программный инструмент *word2vec*, предназначенный для построения ВПС, основанный на нейросетевых языковых моделях. Он способен определять семантическую близость слов путем максимизации косинусной меры сходства между векторами слов, находящихся в похожих контекстах [10].

Для обучения модели на вход ей подается последовательность всех признаковых описаний из обучающей выборки, которым сопоставлены соответствующие метки АТ. В системе использовалась распространенная схема BIO-разметки [23], где B – метка токена, с которого начинается АТ, I – метка токена, входящего в АТ, но не являющегося первым, O – метка токена, не входящего в состав АТ. При обработке тестовой последовательности признаковых описаний модель CRF должна сопоставить ей наиболее вероятную последовательность меток АТ.

В нашей системе использовались общедоступные нейросетевые модели для русского языка [24], разработанные для определения семантической близости

в тестировании Russian Semantic Similarity Evaluation (RUSSE) [25]. Модели создавались с помощью нескольких корпусов:

- *News* – корпус текстов российских новостей, собранных коммерческими новостными агрегаторами. После предобработки объем корпуса составил 1300 миллионов токенов.
- *Web* – корпус текстов, составленный из русскоязычных веб-страниц. Около 9 миллионов страниц были выбраны случайным образом из индекса крупнейших поисковых систем. После предобработки объем корпуса составил примерно 620 миллионов токенов.
- *Ruscorpora* – Национальный корпус русского языка [26]. После предобработки объем корпуса составил 107 миллионов токенов.
- *RuwikiRuscorpora* – корпус, содержащий *Ruscorpora* и русскоязычную Википедию. Размер корпуса – более 280 миллионов токенов.

При создании корпуса *News* была использована архитектура нейронной сети skip-gram, для остальных моделей – CBOW. Лемматизация производилась с помощью *mystem*, после чего из корпусов были убраны стоп-слова.

5. Эксперименты

В данном разделе рассмотрен ряд экспериментов, целью которых было составить признаковое описание, наиболее подходящее для модели CRF в задаче извлечения АТ. Метрикой оценки качества была F1-мера, полученная с помощью предоставленного организаторами SemEval-2016 программного инструмента для тестирования [4]. Проведенные эксперименты были направленны на решение следующих задач:

- определение оптимального размера контекстного окна ВПС и частей речи в признаковом описании;
- определение влияния дополнительных признаков (наличие слова в словаре, заглавная буква, синтаксическая связь) на качество системы;
- определение возможности совместного использования дополнительных признаков;
- поиск нейросетевой языковой модели, обеспечивающей наилучшее качество решения подзадачи извлечения АТ.

Целью первой серии экспериментов стало определение влияния контекста слова в признаковом описании на результаты системы. Контекстом считаются слова, расположенные слева и справа в окне заданного размера от анализируемого слова. В Табл. 2 показано изменение F1-меры после расширения контекстного окна. В признаковом описании использовались ВПС из нейросетевой модели на основе корпуса *News* (этот корпус оказался лучше, см. ниже) и части речи слов. В подзадаче извлечения АТ в рамках SemEval-2016 для русского языка участниками не было представлено методов,

которые превзошли бы базовый метод (baseline) организаторов, поэтому в Табл. 2 показано сравнение с официальным базовым методом [4].

Табл. 2. Расширение контекстного окна в признаковом описании слова

Table 2. Feature set context window expansion

Новый признак в признаковом описании	Точность, %	Полнота, %	F1-мера, %	Изменение F1-меры по отношению к базовому методу, %
Базовый метод (baseline)	57.66	43.06	49.31	–
ВПС	62.23	46.74	53.38	4.07
Часть речи слова	61.77	49.05	54.68	5.37
ВПС контекста [−1; +1]	65.66	52.83	58.55	9.24
Часть речи слов контекста [−1; +1]	69.81	62.18	65.77	16.44
ВПС контекста [−2; +2]	71.68	64.07	67.66	18.35
Часть речи слов контекста [−2; +2]	71.59	64.60	67.91	18.60
ВПС контекста [−3; +3]	71.42	65.65	68.41	19.10
Часть речи слов контекста [−3; +3]	**72.00**	**66.17**	**68.96**	**19.65**
ВПС контекста и часть речи [−4; +4]	71.15	65.54	68.23	18.92

Использование только ВПС в качестве признакового описания уже превосходит базовый метод, предоставленный организаторами SemEval-2016, по F1-мере на 4.07%. Максимальное изменение F1-меры достигается после добавления к признаковому описанию ВПС (+3.87%) и частей речи (+7.22%) соседних слов. По мере расширения контекстного окна F1-мера незначительно увеличивается и достигает максимального значения 68.96% (на 19.65% больше базового метода) при использовании в признаковом описании трех соседних слов слева и справа.

Размер контекстного окна значительно сказывается на времени генерации признакового описания и потребляемой памяти, поэтому для дальнейших экспериментов использовались окна [−1; +1] и [−2; +2].

В **Error! Not a valid bookmark self-reference.** сравниваются результаты использования в признаковом описании дополнительных признаков. В отличие от предыдущей серии экспериментов, в табл. 4 приведено изменение F1-меры при добавлении нового признака к признаковому описанию контекстного окна [−1; +1] и исключению всех остальных признаков. Первый из исследуемых признаков – наличие в слове заглавной буквы. Второй – присутствие слова в словаре аспектной лексики. В словарь аспектной лексики

вошли 552 слова, являющиеся АТ в обучающей выборке. Третьим исследуемым признаком стала синтаксическая связь с предком. Для построения дерева зависимостей использовался программный инструмент MaltParser [27], обученный на корпусе SynTagRus [26].

Табл. 3. Сравнение влияния дополнительных признаков

Table 3. Comparison of the additional features impact

Исследуемый признак	Точность, %	Полнота, %	F1-мера, %	Изменение F1-меры, %
ВПС контекста и части речи [−1; +1]	69.81	62.18	65.77	–
Есть ли заглавная буква?	70.34	62.28	66.07	0.30
Содержится ли слово в словаре лексики АТ?	68.45	**64.28**	66.30	0.53
Синтаксическая связь слова.	**73.07**	60.71	**66.32**	**0.55**

Использование дополнительных признаков в признаковом описании привело к незначительному увеличению F1-меры. Приведенное в Табл. 4 совместное использование этих признаков также не смогло увеличить F1-меру больше, чем на 0.78%.

Табл. 4. Совместное использование дополнительных признаков

Table 4. Using of additional feature set

Используемые признаки	Точность, %	Полнота, %	F1-мера, %	Изменение F1-меры %
ВПС контекста и части речи [−1; +1]	**69.81**	62.18	65.77	–
Первые два признака.	68.52	**64.70**	**66.55**	**0.78**
Все три признака.	69.64	63.13	66.22	0.45

При этом отметим, то добавление первых двух дополнительных признаков к признаковому описанию, содержащему ВПС и частей речи из контекстного окна [−3; +3], привело к ухудшению оценки качества по F1-мере на 0.39%.

В Табл. 5 приведено сравнение производительности системы при использовании разных нейросетевых моделей из работы [24]. Признаковое описание элементов последовательности содержало векторные представления слова и слов из контекстного окна [−2; +2], части речи слова и слов из контекста, а также дополнительные признаки – наличие заглавной буквы и присутствие слова в корпусе слов, являющихся АТ.

Табл. 5. Сравнение результатов программной системы с базовым методом SemEval-2016

Table 5. Comparison between system results and SemEval-2016 baseline

Метод/модель	Точность, %	Полнота, %	F1-мера, %
Базовый метод (baseline)	57.66	43.06	49.31
Модель на основе *Ruscorpora*	72.29	63.86	67.81
На основе *News*	**72.97**	**65.23**	**68.88**
На основе R*uwikiruscorpora*	71.44	63.86	67.44
На основе *Web*	71.01	64.07	67.36

Все варианты CRF-классификатора, обученного на признаковых описаниях, основанных на общедоступных моделях word2vec [24], показали примерно равные результаты и превзошли официальный базовый метод по F1-мере. Несмотря на то, что word2vec-модели готовились для тестирования по определению семантической близости, содержащиеся в них ВПС оказались применимы и для задачи извлечения АТ в АОАТ. Лучшей из представленных word2vec-моделей в этой задаче оказалась модель *News*, обученная на корпусе новостных статей.

Как было отмечено выше, отзывы могут содержать АТ, являющиеся как отдельными словами, так и словосочетаниями. В Табл. 6 приведены сравнительные показатели извлечения АТ для обоих случаев.

Табл. 6. Сравнение производительности системы для АТ, являющихся отдельными словами, и АТ из словосочетаний

Table 6. System performance comparison between one word AT and multiword AT

АТ	Точность, %	Полнота, %	F1-мера, %
Однословные	76,11	73,05	74,55
Многословные	51,81	30,64	38,51

Как и предполагалось, недостаточное количество АТ из словосочетаний в обучающей выборке сказалось на извлечении подобных АТ в тестовой последовательности. Система смогла найти меньше трети АТ из словосочетаний, а точность составила лишь 51,81%, что в итоге дало F1-меру равную 38,51%. F1-мера для АТ, выраженных отдельными словами, оказалась почти в два раза выше и составила 74.55%.

6. Анализ ошибок

Можно выделить несколько типичных ошибок, допускаемых системой. Ниже приведены примеры отзывов из тестовой выборки, распознанные с ошибками, возможные причины ошибок и примеры из обучающей выборки. Все примеры взяты из корпуса отзывов тестирования SemEval-2016.

Главной проблемой системы является пропуск АТ, при этом пропускаются как отдельные слова, являющиеся АТ, так и словосочетания: система позволяет найти лишь 30,64% АТ из словосочетаний с точностью 51,81%. Для АТ из одного слова полнота равна 73,05%, а точность – 76,11%. Вот некоторые примеры предложений из отзывов, АТ в которых не были распознаны:

1) ***«Чурчхела** не понравилась, была замороженная и безвкусная».*
2) *«зато меня порадовал **тирамису** − вкусно, но подан некрасиво в стаканчике».*
3) ***«"Пицца Pro Тесто"** − поскольку первый раз, решили сразу понять "да" или "нет" заказав фирменную пиццу самого большого размера и опять же, не ошиблись».*
4) ***«Меню** продуманное и разнообразное, замечательные десерты, есть отличные предложения торжественных **блюд** для банкетов, на заказ сделают **торт** любой сложности».*
5) *«В такой **ресторан** хочется вернуться вновь».*
6) *«Вкусно! Заказывали и **итальянскую и японскую кухню**, всем понравилось».*

В первых двух примерах АТ являются редкие слова *чурчхела* и *тирамису*, которых нет в нейросетевой языковой модели *News*. Латиница из третьего примера также не содержится в нейросетевой языковой модели. Исправить подобные ошибки можно, использовав другую нейросетевую модель, содержащую в себе необходимые термины предметной области.

В четвертом и пятом примерах пропуск АТ *меню*, *блюд, торт* и *ресторан* может быть следствием специфики предметной области и обучающей выборки, представленной организаторами. В обучающей выборке классификатор получил множество примеров, где эти слова не являются АТ. Некоторые из них приведены далее. *«Милая девушка принесла меню»* – в данном примере АТ является слово *девушка*, а не *меню*. *«И даже когда пытались его подозвать, чтобы он собрал нам оставшиеся блюда с собой, он просто отворачивался и уходил в другой конец зала»* – в этом предложении пользователь не выражает какого-либо мнения к аспекту *еда* и, в частности, к блюдам, поэтому слово *блюда* не является АТ. *«Каравай и торт мы заказали там же, нам все привезли»* – нет выражения тональности по отношению слову *торт*, поэтому оно не является АТ. Слово *ресторан* в четвертом примере является АТ и принадлежит аспекту, характеризующему ресторан в целом, однако часто это слово употребляют для описания какого-то другого аспекта. Так, например, в предложении *«Для меня это просто идеальный интерьер небольшого ресторана»* – слово *ресторан* используется для описания аспекта *интерьер*. Для исправления подобного типа ошибок система должна учитывать не только близкий контекст текущего слова, но и все предложение.

Словосочетание *итальянскую и японскую кухню* из шестого примера должно было быть распознано как АТ, однако последовательности слов в обучающей выборке редко содержали АТ, начинающиеся на слово *итальянский*. Так,

например, в предложениях *«Кухня до итальянской не дотягивает − если только размером порции»* и *«Обычный совковый ресторанчик с итальянским названием»* слово *итальянский* не является АТ. Причиной подобных ошибок программной системы является недостаточное содержание АТ из словосочетаний в обучающей выборке и специфика подхода разметки последовательностей. Часто порядок слов в АТ может быть изменен без потери смысла, это означает, что классификатор должен присваивать одним и тем же словам различные метки, в зависимости от контекста. В отличие от шестого примера, в предложении из обучающей выборки *«Очень радует, что есть японская и итальянская кухни»* другой порядок слов и присутствует АТ *японская и итальянская кухни*, а не *итальянскую и японскую кухню*. Это значит, что обучаясь, классификатор присваивает слову *японский* метку начала АТ – *B*, а слову итальянский – метку слова, входящего в состав АТ, – *I*. При разметке классификатором тестовой последовательности эти метки оказываются неправильными.

Наряду с пропуском слов, являющихся АТ, встречаются случаи неверного распознавания классификатором слов как АТ. В следующих примерах показаны подобные случаи:

7) *«**Ресторан** на Энергетиков».*
8) *«**Ресторан** был пуст!».*
9) *«Оказывается пара за соседним столиком заказала такое же **блюдо** и оно было предназначено для них».*
10) *«Заказала горячую **закуску из овощей**».*

Слова *ресторан* и *блюдо* уже обсуждались в четвертом и пятом примерах. В седьмом, восьмом и девятом примерах продемонстрирована обратная ситуация. При обучении классификатор получал противоречивые примеры для слова *ресторан*. Поэтому при назначении метки слову из тестовой выборки большую роль сыграл контекст слов, и седьмой и восьмой примеры демонстрируют, что контекст слов в признаковом описании стал причиной ошибки классификатора. В этих примерах слово *ресторан* расположено на первом месте в предложении, следовательно, в признаковом описании левая часть контекстного окна пустая (признаковое описание каждого элемента последовательности должно быть одинаковой длины, поэтому левая часть, отведенная для соседних слева слов, заполняется нулями). Обучающая выборка содержит несколько предложений, в которых слово *ресторан*, находящееся на первом месте в предложении, является АТ. Вот некоторые из них: *«Ресторан замечательный!»*, *«Ресторан хорошо подходит для романтического ужина»*. В отличии от примеров 7–10, все перечисленные предложения содержат мнения пользователей. Для того, чтобы учитывать подобные случаи, можно использовать подход извлечения АТ на основе связи между мнением и объектом [1].

В примерах 11–12 приведены примеры АТ из словосочетаний, в которых программная система вместо выделения одного АТ, выраженного словосочетанием, распознала несколько АТ из одиночных слов:

11) *«Интерьер супер, очень удобные **кабинки с мягкими подушками**».*
12) *«Понравились **с креветкой, филадельфией и овощами**».*

В примере 11 системой были выделены АТ *кабинки* и *подушками*. В примере 12 отдельными АТ стали все три слова – *креветкой*, *филадельфией*, *овощами*. Обучающая выборка содержала мало примеров с этими словами. Классификатор мог допустить ошибку из-за недообучения или сходства ВПС этих слов с представлениями слов, часто являющимися АТ.

В примерах 13–14 приведены редкие ошибки, для обработки которых необходимо использование дополнительных подходов к извлечению АТ:

13) *«Цены не кусаются, учитывая что это центр города и сравниваю с другими **суши барами** =)».*
14) *«Все оказалось вкусным, салаты, суп, колбаски из баранины – просто **песня**».*

Пример 13 содержит сравнение цен объекта с другими *суши барами*. Сравнение сложно обработать с помощью машинного обучения, для обработки подобных отзывов может быть применен подход, основанный на правилах.

В примере 14 классификатор ошибочно определил слово *песня* как АТ. В данном контексте словосочетание *просто песня* является оценочной фразой, выраженной идиомой, т. е. словосочетанием, значение которого напрямую не следует из значений составляющих его слов. Идиомы могут выражать как позитивные, так и негативные эмоции, и, как правило, достаточно сильные. Обработка подобных словосочетаний требует предварительно собранного списка всевозможных идиом.

7. Заключение

В рамках данного исследования была разработана программная система для извлечения АТ в АОАТ, основанная на модели CRF. В ходе экспериментов было составлено оптимальное признаковое описание, включающее ВПС из контекстного окна [–3; 3] и их части речи. При тестировании на корпусе отзывов семинара SemEval-2016 программная система показала качество 68.96% по F1-мере. Данный результат превосходит базовый метод, предоставленный организаторами тестирования. Анализ ошибок показал, что подобный подход не всегда обрабатывает ситуации, когда одно и то же слово зависит от контекста и может являться или не являться АТ, поэтому направлениями дальнейших исследований будут обработка подобных исключений и объединение различных подходов к решению подзадачи извлечения АТ.

Список литературы

[1]. Liu B., Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 2012, pp. 1–167.

[2]. 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, pp. 27-35.

[3]. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: Testing Object-oriented Sentiment Analysis Systems in Russian. Proceedings of the 21st International Conference on Computational Linguistics (Dialog-2015), 2015, volume 2, pp. 12–24.

[4]. Pontiki M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S., AL-Smadi M., Al-Ayyoub M., Zhao Y., Qin B., De Clercq O., Hoste V., Apidianaki M., Tannier X., Loukachevitch N., Kotelnikov E., Bel N., Zafra S. M. J., Eryigit G. Semeval-2016 task 5: Aspect based sentiment analysis. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 19–30.

[5]. Андрианов И., Майоров В., Турдаков Д. Современные методы аспектно-ориентированного анализа эмоциональной окраски. Труды ИСП РАН, том 27, вып. 5, 2015 г., стр. 5-22. DOI: 10.15514/ISPRAS-2015-27(5)-1.

[6]. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars. Proceedings of the 21st International Conference on Computational Linguistics (Dialog-2015), 2015, volume 2, pp. 46–57.

[7]. 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, 2010, pp. 1035-1045.

[8]. Xenos D., Theodorakakos P., Pavlopoulos J., Malakasiotis P., Androutsopoulos I. AUEB-ABSA at SemEval-2016 Task 5: Ensembles of Classifiers and Embeddings for Aspect Based Sentiment Analysis. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 312–317.

[9]. Hamdan H. SentiSys at SemEval-2016 Task 5: Opinion Target Extraction and Sentiment Polarity Detection. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 350–355.

[10]. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 2013, pp. 3111–3119.

[11]. Blinov P. D., Kotelnikov E. V. Semantic Similarity for Aspect-Based Sentiment Analysis. Proceedings of the 21st International Conference on Computational Linguistics Dialog-2015, 2015, volume 2, pp. 36–45.

[12]. Toh Z., Su J. NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 282–288.

[13]. Hu M., Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, 2004, pp. 168–177.

[14]. Popescu A. M., Nguyen B., Etzioni O. OPINE: Extracting product features and opinions from reviews. Proceedings of HLT/EMNLP on interactive demonstrations, 2005, pp. 32–33.

[15]. Turney P. D. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning, 2001, pp. 491–502.

[16]. Scaffidi C., Bierhoff K., Chang E., Felker M., Ng H., Jin C. Red Opal: product-feature scoring from reviews. Proceedings of the 8th ACM conference on Electronic commerce, 2007, pp. 182–191.

[17]. Hofmann T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, 1999, pp. 50–57.
[18]. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. Journal of machine Learning research, 2003, pp. 993-1022.
[19]. Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, volume 1, pp. 339-348.
[20]. Titov I., McDonald R., Modeling online reviews with multi-grain topic models. Proceedings of the 17th international conference on World Wide Web. ACM, 2008, pp 111–120.
[21]. Müller A.C., Behnke S. PyStruct: learning structured prediction in python. Journal of Machine Learning Research 15(1), 2014, pp. 2055–2060.
[22]. Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In MLMTA, 2003, pp. 273–280.
[23]. Ramshaw L. A., Marcus M. P. Text chunking using transformation-based learning. Natural language processing using very large corpora, Springer Netherlands, 1999, pp. 157-176.
[24]. Kutuzov A, Andreev I. Texts in, meaning out: neural language models in semantic similarity task for Russian. Proceedings of the 21st International Conference on Computational Linguistics Dialog-2015, 2015, volume 2, pp. 133–144.
[25]. Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N. Russe: The first workshop on russian semantic similarity. Proceedings of the 21st International Conference on Computational Linguistics Dialog-2015, 2015, volume 2, pp. 89-105.
[26]. Plungian V. A. Why we make Russian National Corpus? Otechestvennye Zapiski 2, 2005.
[27]. Sharov S, Nivre J. The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. Proceedings of the 17th International Conference on Computational Linguistics Dialog-2011, 2011, pp. 657–670.

Aspect term extraction based on word embedding and conditional random fields

D.O. Mashkin <danil.mashkin@gmail.com>
E.V. Kotelnikov <kotelnikov.ev@gmail.com>
Vyatka State University,
36, Moskovskaya, Kirov, 610000, Russia.

Abstract. There are many sites in the Internet that allow users to share their opinions and write reviews about all kinds of goods and services. These views may be useful not only for other users, but also for companies which want to track their own reputation and to receive timely feedback on their products and services. The most detailed statement of the problem in this area is an aspect-based sentiment analysis, which determines the user attitude not only to the object as a whole, but also to its individual aspects. In this paper we consider the solution of subtask of aspect terms extraction in aspect-based sentiment analysis. A review of research in this area is given. The subtask of aspect terms extraction is considered as a problem of sequence labeling; to solve it we apply the model of conditional random fields (CRF). To create the sequence feature description, we use distributed representations of words derived from neural network models for the Russian language and parts of speech of the analyzed words. The stages of the aspect terms extraction software system are shown. The experiments with the developed software system were carried out on the corpus of labeled reviews of restaurants, created in the International Workshop on Semantic Evaluation (SemEval-2016). We describe the dependence of the quality of aspect terms extraction subtask on various neural network models and the variations of feature descriptions. The best results (F1-measure = 69%) are shown by a version of the system, which takes into account the context and the parts of speech. This paper contains a detailed analysis of errors made by the system, as well as suggestions on possible options for their correction. Finally, future research directions are presented.

Keywords: aspect-based sentiment analysis; sentiment analysis; aspect term extraction; machine learning; sequential labeling;, word embedding; word2vec; SemEval-2016.

DOI: 10.15514/ISPRAS-2016-28(6)-16

For citation: Mashkin D.O., Kotelnikov E.V. Aspect term extraction based on word embedding and conditional random fields. Trudy ISP RAN/Proc. ISP RAS, vol. 28, issue 6, 2016. pp. 223-240 (in Russian). DOI: 10.15514/ISPRAS-2016-28(6)-16

References

[1]. Liu B., Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 2012, pp. 1–167.
[2]. 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, pp. 27-35.

[3]. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: Testing Object-oriented Sentiment Analysis Systems in Russian. Proceedings of the 21st International Conference on Computational Linguistics (Dialog-2015), 2015, volume 2, pp. 12–24.

[4]. Pontiki M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S., AL-Smadi M., Al-Ayyoub M., Zhao Y., Qin B., De Clercq O., Hoste V., Apidianaki M., Tannier X., Loukachevitch N., Kotelnikov E., Bel N., Zafra S. M. J., Eryigit G. Semeval-2016 task 5: Aspect based sentiment analysis. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 19–30.

[5]. Andrianov I., Mayorov V., Turdakov D. Modern Approaches to Aspect-Based Sentiment Analysis. Trudy ISP RAN/Proc. ISP RAS, vol. 27, issue 5, 2015, pp. 5–22 (in Russian). DOI: 10.15514/ISPRAS-2015-27(5)-1.

[6]. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars. Proceedings of the 21st International Conference on Computational Linguistics (Dialog-2015), 2015, volume 2, pp. 46–57.

[7]. 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, 2010, pp. 1035-1045.

[8]. Xenos D., Theodorakakos P., Pavlopoulos J., Malakasiotis P., Androutsopoulos I. AUEB-ABSA at SemEval-2016 Task 5: Ensembles of Classifiers and Embeddings for Aspect Based Sentiment Analysis. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 312–317.

[9]. Hamdan H. SentiSys at SemEval-2016 Task 5: Opinion Target Extraction and Sentiment Polarity Detection. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 350–355.

[10]. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 2013, pp. 3111–3119.

[11]. Blinov P. D., Kotelnikov E. V. Semantic Similarity for Aspect-Based Sentiment Analysis. Proceedings of the 21st International Conference on Computational Linguistics Dialog-2015, 2015, volume 2, pp. 36–45.

[12]. Toh Z., Su J. NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 282–288.

[13]. Hu M., Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, 2004, pp. 168–177.

[14]. Popescu A. M., Nguyen B., Etzioni O. OPINE: Extracting product features and opinions from reviews. Proceedings of HLT/EMNLP on interactive demonstrations, 2005, pp. 32–33.

[15]. Turney P. D. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning, 2001, pp. 491–502.

[16]. Scaffidi C., Bierhoff K., Chang E., Felker M., Ng H., Jin C. Red Opal: product-feature scoring from reviews. Proceedings of the 8th ACM conference on Electronic commerce, 2007, pp. 182–191.

[17]. Hofmann T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, 1999, pp. 50–57.

[18]. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. Journal of machine Learning research, 2003, pp. 993-1022.

[19]. Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, volume 1, pp. 339-348.

[20]. Titov I., McDonald R., Modeling online reviews with multi-grain topic models. Proceedings of the 17th international conference on World Wide Web. ACM, 2008, pp 111–120.

[21]. Müller A.C., Behnke S. PyStruct: learning structured prediction in python. Journal of Machine Learning Research 15(1), 2014, pp. 2055–2060.

[22]. Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In MLMTA, 2003, pp. 273–280.

[23]. Ramshaw L. A., Marcus M. P. Text chunking using transformation-based learning. Natural language processing using very large corpora, Springer Netherlands, 1999, pp. 157-176.

[24]. Kutuzov A, Andreev I. Texts in, meaning out: neural language models in semantic similarity task for Russian. Proceedings of the 21st International Conference on Computational Linguistics Dialog-2015, 2015, volume 2, pp. 133–144.

[25]. Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N. Russe: The first workshop on russian semantic similarity. Proceedings of the 21st International Conference on Computational Linguistics Dialog-2015, 2015, volume 2, pp. 89-105.

[26]. Plungian V. A. Why we make Russian National Corpus? Otechestvennye Zapiski 2, 2005.

[27]. Sharov S, Nivre J. The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. Proceedings of the 17th International Conference on Computational Linguistics Dialog-2011, 2011, pp. 657–670.