Выделение аспектных терминов в отзывах с использованием моделей условных случайных полей

Рубцова Ю.В. (yu.rubtsova@gmail.com),
Кошельников С.А. (кошельниковса@gmail.com)

Ключевые слова: извлечение аспектов, CRF, извлечение мнений, отзовы пользователей

Извлечение аспекта с помощью Условные случайные поля

Рубцова Ю.В. (yu.rubtsova@gmail.com),
Кошельников С.А. (koshelnikovsa@gmail.com)

В данной статье описывается система извлечения аспектов, которая была представлена на SentiRuEval-2015: анализ тональности отзывов пользователей на русском языке на основе аспектов. Предлагаемая система использует алгоритм условного случайного поля для извлечения аспектов, упомянутых в тексте. Мы использовали набор морфологических и синтаксических признаков для машинного обучения и продемонстрировали, что использование лемм в качестве признака может улучшить результаты извлечения аспектов. Система использовалась для выполнения двух подзадач: Задача А — автоматическое извлечение явных аспектов и Задача Б — автоматическое извлечение всех аспектов (явных, неявных и сентиментальных фактов) и тестировалась на двух доменах — ресторанах и автомоби Обе подзадачи, А и В, в обеих областях были выполнены с достаточно высоким уровнем точности, что означало, что система была способна достаточно точно распознавать аспектные термины. Но более низкие результаты отзыва подразумевают, что система нашла достаточно терминов аспектов, которые нельзя рассматривать как аспекты в соответствии с золотым стандартом. Наши системы выступили конкурентоспособно и показали результаты, сравнимые с результатами других 10 участников.

Ключевые слова: обнаружение аспектов, извлечение аспектов, CRF, анализ мнений, обзоры.

1. Введение

С ростом популярности блогов, социальных сетей и сайтов с отзывами пользователей о продуктах и услугах с каждым годом пользователи сети публикуют все больше отзывов. В результате накопился огромный пул обзоров, оценок и рекомендаций в различных областях, что привлекает внимание как исследователей, занимающих сбором мнений, анализом настроений и выявлением тенденций, так и бизнесменов, более заинтересованных в практическом применении. репутационного маркетинга а

Автоматический анализ тональности в основном используется на следующих уровнях: • Уровень документа (Терни, 2002; Панг и др., 2002; Рубцова, 2014), • Уровень предложения или фразы (Уилсон и др., 2009), • Уровень аспекта (Лю 2012; Чжан, Лю, 2014; Маррез-Тейлор и др., 2014).

Как правило, люди высказывают свое мнение не о товаре или услуге в целом, а о какой-то их части, свойстве или характеристике, и именно этот аспект необходимо извлечь из текста и подвергнуть сентимент-анализу. Анализ тональности на уровне аспектов может дать нам гораздо больше полезной информации об мнении автора о различных характеристиках анализируемого продукта или услуги, чем анализ тональности всего текста.

Конференция-диалог включала в себя раздел «Оценка диалога: оценка систем анализа настроений для российского SentiRuEval» (Лукачевич и др., 2015). Участники оценки должны были выполнить следующие 5 подзадач: A. Извлечь явные аспекты из предложенного обзора, B. Извлечь все аспекты из предложенного обзора, C. Выполнить анализ настроений явных аспектов, D. Классифицировать термины аспектов. по предопределенным категориям, E. Оцените категории аспектов, связанные с предлагаемым обзором в целом.

В этом документе описывается система, которая использовалась для выполнения Заданий A и B во время соревнований SentiRuEval.

Остальная часть статьи структурирована следующим образом. В разделе 2 мы обсуждаем современное состояние дел и различные механизмы извлечения аспектов из обзоров продуктов. В разделе 3 мы описываем нашу систему. Раздел 4 демонстрирует производительность нашей системы по сравнению с результатами систем других участников SentiRuEval. В разделе 5 представлены подробные выводы и перспективы дальнейшего развития.

2. Связанная работа

Существует четыре основных подхода к извлечению аспектов из текстов. Первый основан на частоте употребления существительных и/или именных словосочетаний. Обычно люди используют схожие термины для описания характеристик и своего отношения к продуктам и разные термины для описания других деталей (ситуация, необходимая сопроводительная информация) в своих комментариях. Таким образом, подсчет частоты наиболее часто встречающихся существительных и/или словосочетаний в текстах одной и той же предметной области помогает извлечь термины-аспекты из большого количества обзоров (Hu and Liu, 2004). Позже уровень точности этого алгоритма был улучшен на 22% (Popescu and Etzioni, 2005). Поскольку общеупотребительные слова часто встречаются в текстах и часто определяются как аспекты, был изобретен механизм фильтрации, чтобы исключить из результатов анализа наиболее распространенные существительные и/или фразы без аспектов (Moghaddam and Ester, 201

Второй подход основан на одновременном извлечении как сентиментальных слов (мнений пользователей), так и аспектов. Поскольку любое мнение выражается по отношению к объекту, ища слова настроения, мы можем найти аспекты, к которым они относятся. Ху и Лю использовали этот подход для поиска низкочастотных аспектов (Ху и Лю, 2004). Другой подход — контролируемое машинное обучение. Как правило, для целей извлечения аспектов

машинное обучение с учителем ориентировано на задачи маркировки последовательности, поскольку аспекты и мнения о продуктах часто взаимосвязаны и представляют собой последовательность слов. Наиболее распространенными методами контролируемого машинного обучения являются скрытое марковское моделирование (HMM) (Jin et al., 2009) и условные случайные поля (CRF).
(Лафферти и др., 2001; Саттон и МакКаллум, 2006; Якоб и Гуревич, 2010). Четвертый подход — неконтролируемое машинное обучение или тематическое моделирование. Тематическое моделирование предполагает, что каждый документ состоит из смеси тем, и каждая тема представляет собой распределение вероятностей (Titov and McDonald, 2008; Brody and Elhadad, 2010). Большинство работ по извлечению аспектов с использованием подхода тематического моделирования основаны на методах модели расширенного вероятностного латентного семантического анализа (pLSA) (Hofmann, 2001) и модели латентного распределения Дирихле (LDA) (Blei et al., 2003).

Для выполнения сложных задач, таких как одновременное извлечение аспектов и анализ тональности или одновременное выделение аспектов и категоризация, можно использовать комбинацию различных подходов, таких как максимальная энтропия и латентное распределение Дирихля (Zhao WX et al, 2010) или полуконтролируемая модель с тематическим моделированием. подход, когда пользователь предоставляет исходные слова для нескольких категорий аспектов (Mukherjee and Liu, 2012).

3. Описание системы

Мы участвовали в двух оценках:

- Извлечение явных аспектов, т.е. извлечение части анализируемого объекта или одной из его характеристик, таких как двигатель для домена автомобилей или услуга для домена ресторанов,

- Выделение всех аспектов анализируемого объекта, включающее в себя выделение явных аспектов, имплицитных аспектов (аспект + однозначное мнение автора об аспекте) и фактов настроений (когда автор не использует выражения мнения, а указывает факт, который однозначно раскрывает его отношение к объекту).

Для извлечения целей или аспектов мнений из предложений, содержащих выражения мнений, мы использовали CRF. CRF показывает сравнительно хорошие результаты для задачи извлечения аспектов из отзывов. Например, для общего задания SemEval-2014, связанного с анализом настроений на основе аспектов, два лучших результата были получены с системами, основанными на CRF (Pontiki rt al., 2014).

Условные случайные поля предлагаются в качестве модели ненаправленной последовательности, которая моделирует условную вероятность p(Y|X) над скрытой последовательностью Y при заданной последовательности наблюдений X. То есть условная модель обучается маркировать неизвестную последовательность наблюдений X путем выбора скрытой последовательность Y, которая максимизирует p(Y|X). В качестве программной реализации CRF мы использовали инструмент Mallet (McCallum, 2002).

3.1. Предварительная обработка

Якоб и Гуревич (Якоб и Гуревич, 2010) представили возможные метки по схеме маркировки Inside-Outside-Begin (IOB): B-Target, идентифицирующая

начало мишени мнения; I-Target, определяющий продолжение цели, и O для других (нецелевых) токенов. Поэтому, поскольку мы использовали последовательную маркировку, мы присваивали метку каждому слову в предложении, где se указывало на начало явного термина аспекта, ce указывало на продолжение явного термина аспекта, si указывало начало неявного термина аспекта, ci указывало продолжение термина неявного аспекта (точно так же, как для терминов-фактов: sf для факта начала, cf для факта продолжения), а O указывает термин, не являющийся аспектом.

Чтобы извлечь синтаксические признаки (например, POS и лемму), описанные в следующем разделе, мы использовали TreeTagger для русского языка (Sharoff et al., 2008).

Также мы обратили внимание, что марки автомобилей часто пишутся латиницей и/или содержат такие цифры, как Nissan Micra или ВАЗ 2109. Поэтому для коллекции автомобилей мы добавили правила, позволяющие распознавать полное название автомобиля (или марку) как один явный термин. Как видно из таблицы 3, это дало некоторые положительные результаты — Система заняла 3-е место по варианту точного соответствия F-меры.

Мы также преобразовали все заглавные буквы в строчные, поскольку программные инструменты могут воспринимать Engine и Engine как два разных аспекта, что неверно.

3.2. Функции

B

качестве признаков использовались строки слов текущего токена. Мы извлекли одно предыдущее и одно последующее слово и использовали их в качестве дополнительных признаков слова, чтобы получить больше информации о контексте, в котором это слово используется.

POS

В качестве функции использовался тег части речи (POS) текущего токена. Аспекты терминов часто выражаются существительными. Маркировка POS добавляет полезную информацию о части речи, к которой принадлежит слово. Для определения части речи мы использовали TreeTagger — инструмент, выполняющий полный синтаксический анализ. Мы сводим полный морфологический анализ к таким частям речи, как N для двигателя и V для вождения.

Лемма

В качестве признака использовалась лемма текущего токена. В связи с огромным количеством словоформ в русском языке мы добавили в качестве признака нормальную форму слова. Для извлечения лемм мы также используем TreeTagger.

3.3. Архитектура

Мы построили две системы:

- Система 1: CRF со всеми вышеупомянутыми метками. Мы использовали метки se, ce и O для явного извлечения аспектов для выполнения задачи A и se, ce, si, ci, sf, cf, O для извлечения всех аспектов для задачи B.

- Система 2: Комбинация результатов двух CRF — CRF для извлечения явных терминов аспекта и CRF для извлечения неявных терминов аспекта + терминов фактов тональности (не явных).

Задача А выполнялась с использованием Системы 1, а Задача Б — с использованием обеих систем.

4. Результаты

Результаты задач А и Б оценивались по F-мере. Были рассчитаны два случая меры F: 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, было рассчитано пересечение между золотым стандартом и извлеченным термином. В таблицах 1–4 показано, как эффективность Системы в Задаче А, а в Таблицах 5–8 соотносится с выполнением Задачи В. Результаты Системы сравнивались с исходным уровнем и двумя лучшими результатами участников SentiRuEval.

Как видно из Табл. 1-4, Система продемонстрировала высокий уровень точности в обеих областях (2-е место в Задаче А как для автомобилей, так и для ресторанов по показателям Точности). Следует отметить, что в области автомобилей результаты были лучше, когда свойство леммы не использовалось — это может быть связано с правилами предварительной обработки коллекции автомобилей. В Задаче В обе построенные системы также показали достаточно высокий уровень точности (см. Таблицу 5–8). В домене ресторанов система 1 с признаками слово+пос +лемма поставила всех участников на 3-е место по случаю частичного совпадения F-меры.

Таблица 1. Результаты задачи А, домен ресторана, точное соответствие

Система	Точность	Отзывать	F-мера
исходный уровень	0,557	0,6903	0,6084
№1	0,7237	0,5738	0,6319
№2	0,6358	0,6327	0,6266
Word+POS	0,661	0,515	0,5704
+лемма	0,6674	0,5417	0,5899

Таблица 2. Результаты задачи А, домен ресторана, частичное совпадение

Система	Точность	Отзывать	F-мера
исходный уровень	0,658	0,696	0,6651
№1	0,8078	0,6165	0,728
№2	0,7458	0,7114	0,7191
Word+POS	0,738	0,563	0,6277
+лемма	0,7485	0,5937	0,652

Таблица 3. Результаты задачи A, автомобильный домен, точное соответствие

Система	Точность	Отзывать	F-мера
исходный уровень	0,5747	0,6287	0,5941
№1	0,76	0,6218	0,6761
№2	0,6619	0,656	0,6513
Word+POS	0,7109	0,5454	0,6075
+лемма	0,704	0,5785	0,6256

Таблица 4. Результаты задачи A, автомобильный домен, частичное совпадение

Система	Точность	Отзывать	F-мера
исходный уровень	0,7449	0,6724	0,6966
№1	0,7917	0,7272	0,7482
№2	0,8561	0,6551	0,7304
Word+POS	0,797	0,6047	0,6747
+лемма	0,7908	0,6485	0,6991

Таблица 5. Результаты задачи B, домен ресторана, точное соответствие

Система	Точность	Отзывать	F-мера
исходный уровень	0,546577	0,647729	0,587201
№1	0,609432	0,600621	0,600128
№2	0,733599	0,513197	0,596179
Система 1 Слово+POS	0,639256	0,456334	0,52577
+лемма	0,639798	0,487202	0,546905
Система 2 Word+POS	0,652145	0,458471	0,531644
+лемма	0,67152	0,491622	0,56153

Таблица 6. Результаты задачи B, домен ресторана, частичное совпадение

Система	Точность	Отзывать	F-мера
исходный уровень	0,671626	0,593093	0,619285
№1	0,756213	0,610754	0,667928
№2	0,668677	0,637097	0,645234
Система 1 Word+POS	0,710428	0,493393	0,5692
+лемма	0,709915	0,529354	0,595303
Система 2 Word+POS	0,724649	0,457863	0,547813
+лемма	0,752364	0,493553	0,585126

Таблица 7. Результаты задачи B, Car domain, точное соответствие

Система	Точность	Отзывать	F-мера
исходный уровень	0,597886	0,589612	0,588623
№1	0,7701	0,553546	0,636623
№2	0,656321	0,616423	0,630149
Система 1 Word+POS	0,690826	0,476309	0,556107
+лемма	0,670594	0,518742	0,578086
Система 2 Word+POS	0,718995	0,482064	0,568331
+лемма	0,701193	0,520375	0,589311

Таблица 8. Результаты задачи B, домен автомобиля, частичное совпадение

Система	Точность	Отзывать	F-мера
исходный уровень	0,783254	0,605976	0,674288
№1	0,814283	0,650998	0,714762
№2	0,795431	0,646999	0,704189
Система 1 Word+POS	0,793637	0,53216	0,625502
+лемма	0,777257	0,584768	0,656113
Система 2 Word+POS	0,808562	0,509979	0,61308
+лемма	0,782394	0,558153	0,638947

4.1. Анализ ошибок

Анализ ошибок указал на некоторые распространенные ошибки: не распознанные и чрезмерно распознанные. Вообще есть еще один тип ошибок для задачи извлечения аспекта — частично распознанные термины аспекта. Благодаря предоставленным сценариям оценки мы не сможем наблюдать ошибки третьего типа. Из Таблицы 9 видно, что основная масса ошибок связана с нераспознанными аспектными терминами.

Табл. 9. Распределение типов ошибок для задачи A (точное совпадение)

	Рестораны	Легковые автомобили
Word+POS		
не распознано	65%	68%
чрезмерно распознано	35%	32%
Word+POS+лемма		
не распознано	63%	65%
чрезмерно распознано	37%	35%

Мы также можем заметить, что добавление лемм в качестве функции CRF приводит к увеличению чрезмерно распознаваемых терминов. Мы сравнили две наши системы и выяснили, что вторая лучше справляется с коллокацией. Например, она извлекает «утиный суп» («суп из утки») вместо просто «суп» («суп»), извлекаемый системой 1. Однако извлечение словосочетаний также является недостатком системы 2, поскольку иногда она извлекает много нерелевантных терминов. Например, «паста с морепродуктами мужу» («пасту с морепродуктами, мужу»).

В будущем мы хотели бы поэкспериментировать с дополнительными статистическими и лексическими функциями CRF. Использование дополнительных текстовых коллекций и предварительная обработка тематического моделирования также могут внести дополнительные улучшения.

5. Выводы

Мы представили две системы выделения аспектов, построенные на основе алгоритма условного случайного поля. Реализация этих систем показала, что предобработка и использование лемм для русского языка в качестве признака И РК сравнительно хорошо показывает общую F-меру. Производительность наших систем была сопоставима с лучшими результатами участников SentiRuEval. Впоследствии мы собираемся добавить статистические методы в качестве функции CRF. Мы также планируем провести исследование и найти способ улучшить результаты отзыва без снижения точности.

Рекомендации

1. Блей Д.М., Нг А.И., Джордан М.И. (2003). Скрытое распределение Дирихле. Журнал исследований в области машинного обучения, 3, 993–1022.
2. Броуди С., Эльхадад Н. (2010). Неконтролируемая модель аспекта-настроения для онлайн-обзоров. В технологиях человеческого языка: Ежегодная конференция Североамериканского отделения Ассоциации вычислительной лингвистики 2010 г., стр. 804–812.

3. Хофманн Т. (2001). Обучение без учителя с помощью вероятностного латентного семантического анализа Мис. Машинное обучение, 42 (1–2), стр. 177–196.
4. Ху М., Лю Б. (2004). Сбор и обобщение отзывов клиентов. В материалах десятой международной конференции ACM SIGKDD по открытию знаний и интеллектуальному анализу данных, стр. 168–177.

5. Якоб Н., Гуревич И. (2010). Извлечение целевых мнений в одном и нескольких основных параметрах с условными случайными полями. В материалах конференции 2010 г. по эмпирическим методам обработки естественного языка, ACL, стр. 1035–1045.
6. Джин В., Хо Х.Х., Шрихари Р.К. (2009, июнь). OpinionMiner: новая система машинного обучения для сбора и извлечения мнений из Интернета. В материалах 15-й международной конференции ACM SIGKDD по открытию знаний и интеллектуальному анализу данных, стр. 1195–1204.

7. Лафферти Дж., МакКаллум А., Перейра Ф. (2001). Условные случайные поля: вероятностные модели для сегментации и маркировки данных последовательности. В материалах Международной конференции по машинному обучению (ICML-2001).

8. Лю Б. (2012). Анализ настроений и добыча мнений. Обобщающие лекции о Х у Языковые технологии человека, 5 (1), стр. 1–167.

9. Лукашевич Н.В., Блинов П.Д., Котельников Е.В., Рубцова Ю.В., Иванов В.В., Тутубалина Е. (2015), SentiRuEval: Тестирование объектно-ориентированных систем анализа настроений на русском языке, Материалы международной конференции Dialog.

10. Маррез-Тейлор Э., Веласкес Дж. Д., Браво-Маркес Ф. (2014). Новый детерминированный подход к анализу мнений на основе аспектов в обзорах туристических продуктов. Экспертные системы с приложениями, 41(17), стр. 7764–7775.

11. МакКаллум А.К. (2002). MALLET: машинное обучение для набора языковых инструментов.

12. Могаддам С., Эстер М. (2011). ILDA: взаимозависимая модель LDA для изучения скрытых аспектов и их оценок из онлайн-обзоров продуктов. В материалах 34-й международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, стр. 665–674. 13. Мукерджи А., Лю Б. (2012). Извлечение аспектов посредством

полууправляемого моделирования. В материалах 50-го ежегодного собрания Ассоциации компьютерной лингвистики: длинные статьи, том 1, стр. 339–348.

14. Панг Б., Ли Л., Вайтьянатан С. (2002). Недурно?: классификация настроений с использованием методов машинного обучения. В материалах конференции ACL-02 по эмпирическим методам обработки естественного языка, том 10, стр. 79–86.

15. Понтики М., Папагеоргиу Х., Галанис Д., Андруцопулос И., Павлопулос Дж., Манандхар С. (2014). Семеваль-2014, задание 4: Аспектный анализ тональности. В материалах 8-го Международного семинара по семантической оценке, SemEval 2014, стр. 27–35.

16. Попеску А.М., Эциони О. (2007). Извлечение характеристик продукта и мнений из отзывов. Обработка естественного языка и анализ текста, стр. 9–28.

17. Рубцова Ю.В. (2014). Классификатор настроений, независимый от предметной области разработки и исследования // Труды СПИИРАН. Т. 5(36). С. 59–77.

18. Шарофф С., Копотев М., Ержавец Т., Фельдман А., Дивьяк Д. (2008) Разработка и оценка наборов тегов для русского языка. В ЛРЭК.

19. Саттон К., МакКаллум А. (2006). Введение в условные случайные поля для реляционного обучения. Введение в статистическое реляционное обучение. Массачусетский технологический институт Пресс.

20. Титов И., Макдональд Р. (2008). Моделирование онлайн-обзоров с помощью многогранных тематических моделей. В материалах 17-й международной конференции World Wide Web, ACM, стр. 111–120.

21. Терни П.Д. (2002). Большой палец вверх или большой палец вниз?: семантическая ориентация применяется к неконтролируемой классификации отзывов. В материалах 40-го ежегодного собрания ассоциации компьютерной лингвистики, ACL, стр. 417–424.

22. Уилсон Т., Вибе Дж., Хоффманн П. (2009). Распознавание контекстуальной полярности: исследование особенностей анализа настроений на уровне фраз. Компьютерная лингвистика, 35(3), 399–433.

23. Чжан Л., Лю Б. (2014). Извлечение аспектов и сущностей для сбора мнений. В данных интеллектуальный анализ и обнаружение знаний для больших данных, стр. 1–40.

24. Чжао В.С., Цзян Дж., Ян Х., Ли С. (2010). Совместное моделирование аспектов и мнений с помощью гибрида MaxEnt-LDA. В материалах конференции 2010 г. по эмпирическим методам обработки естественного языка, ACL, стр. 56–65.