Модель извлечения пользовательских мнений о потребительских свойствах товара на основе рекуррентной нейронной сети¹

Ю.П. Ехлаков

доктор технических наук
профессор кафедры автоматизации обработки информации
Томский государственный университет систем управления и радиоэлектроники
Адрес: 634050, г. Томск, пр. Ленина, д. 40
E-mail: upe@tusur.ru

Е.И. Грибков

аспирант кафедры автоматизации обработки информации
Томский государственный университет систем управления и радиоэлектроники
Адрес: 634050, г. Томск, пр. Ленина, д. 40
E-mail: drnemor@gmail.com

Аннотация

В статье на основе анализа существующих подходов к задаче разметки последовательности предложена модель структурного предсказания с использованием рекуррентной нейронной сети долгой краткосрочной памяти (long short-term memory, LSTM), позволяющая извлекать мнения пользователей из текстов отзывов. Представлены конфигурация модели и набор правил изменения ее состояния, позволяющие использовать как признаки из обрабатываемого предложения, так и прошлые результаты собственных предсказаний. Для обучения и оценки качества работы модели создан размеченный набор текстов отзывов о мобильных телефонах из интернет-магазина Amazon. Описана процедура обучения модели извлечения мнений на наборе отзывов, предложены конкретные значения гиперпараметров модели. Проведено экспериментальное сравнение разработанной модели с моделью на основе условного случайного поля (conditional random field, CRF) с использованием LSTM. Для сравнения моделей использовался критерий F1, характеризующий соотношение полноты и точности извлечения мнений. Результаты экспериментального исследования модели показывают, что предложенная модель дает более высокие по сравнению с аналогом результаты: в случае строго совпадения F1 для аспектов выше на 4,51%, для оценочных высказываний – на 5,44%.

В качестве практической апробации предложенной модели было проведено извлечение мнений о потребительских свойствах мобильных телефонов из текстов отзывов, отсутствующих в размеченной выборке. Кроме того, дополнительно были извлечены и представлены мнения из отзывов на другие категории товаров: телевизоры, средства для ухода за кожей, планшеты. Результаты апробации позволяют говорить о том, что модель показала хорошие результаты как для извлечения мнений о мобильных телефонах, так и о других категориях товаров. Полученные результаты могут быть полезны специалистам в области компьютерной лингвистики и машинного обучения, а также руководителям и менеджерам интернет-магазинов для определения потребительских предпочтений потенциальных пользователей, выдачи рекомендаций по приобретению товаров и предоставления развитых средств поиска по каталогу товаров.

Ключевые слова: обратная связь пользователей; глубокое обучение; машинное обучение; обработка естественного языка; обработка мнений.

Цитирование: Ехлаков Ю.П., Грибков Е.И. Модель извлечения пользовательских мнений о потребительских свойствах товара на основе рекуррентной нейронной сети // Бизнес-информатика. 2018. № 4 (46). С. 7–16. DOI: 10.17323/1998-0663.2018.4.7.16

¹ Работа выполнена в рамках государственного задания Министерства образования и науки РФ, проект № 8.8184.2017/8.9

Введение

Покупка любого достаточно сложного технического устройства (например, телефона, холодильника или телевизора) представляет собой нетривиальную проблему выбора для потенциального потребителя. Присутствие на рынке большого количества производителей и модельных рядов, разнообразие технических характеристик товаров объективно приводят к комбинаторному росту числа возможных вариантов потребительских свойств продуктов-аналогов (товаров). Покупатель либо приобретает товар «на удачу», надеясь на бренд производителя, либо принимает решение о покупке на основе рекламной информации. С развитием Интернета и социальных сетей большинство покупателей-прагматиков при принятии решений о покупке руководствуются не только рекламой, но и мнением покупателей о потребительских свойствах товара, которые выявляются при его длительном использовании.

Компании-производители товаров также заинтересованы в лучшем понимании своих клиентов: какие товары они предпочитают, какие положительные и отрицательные характеристики приобретаемых товаров они выделяют. Опираясь на эти данные, руководство может формировать ассортимент, осуществлять индивидуальный подбор товаров и услуг, делать специальные предложения клиентам и предпринимать другие действия, направленные на повышение лояльности покупателей и улучшение конкурентоспособности.

Существует большое количество источников, из которых можно узнать мнения пользователей о конкретных товарах. Это могут быть тематические форумы, статьи и видеоролики с обзорами, а также группы в социальных сетях. Многие сетевые магазины позволяют своим покупателям оставлять мнения о приобретенных товарах на их страничках. Облегчить задачу поиска информации могут агрегаторы типа «Яндекс.Маркет», которые собирают в одном месте комментарии пользователей, а также предоставляют сообществу возможность оценивать полезность содержащейся в них информации.

Однако большинство подобных платформ решают только проблему сбора информации в одном месте, без ее автоматического анализа и обобщения. Посетители вынуждены самостоятельно изучать и анализировать отзывы, что при большом объеме информации весьма проблематично. Качественное решение задачи по извлечению пользовательских мнений о потребительских свойствах товаров и представление

информации в удобном для анализа виде возможны при использовании современных методов машинного обучения в области обработки естественного языка.

В настоящее время большинство работ, посвященных обработке текстов отзывов пользователей, основывается на анализе тональности (эмоционального настроения) текста. При этом тональность рассматривается как атрибут целого текста или его крупных частей (параграфов и предложений) [1, 2], чего явно недостаточно для извлечения пользовательских мнений. Исследования в области аспектно-ориентированного анализа тональности посвящены вопросам поиска отдельных упоминаний аспектов (потребительских свойств и характеристик) продуктов и определению отношения пользователей к ним «в целом», путем отнесения к одной из категорий: «хорошее», «плохое», «нейтральное», «неизвестное» [3, 4]. Наиболее проработанной постановкой задачи определения отношения пользователя к продукту является детальный анализ тональности [5], где предлагается выделять в тексте отзывов аспекты и высказывания, в которых пользователь выражает свое отношение к данному аспекту. В ряде работ для детального анализа тональности применяются модели сегментации последовательности на основе условных случайных полей (conditional random field, CRF) [3, 6, 7]. В качестве входных признаков для сегментирования используются результаты работы синтаксических анализаторов: частеречная разметка, деревья зависимостей, деревья непосредственных составляющих и др. Однако такие анализаторы недоступны для многих естественных языков. Кроме того, точность таких анализаторов сильно зависит от природы обучающих данных. Помимо этого, в данных моделях используются словари специальной лексики: эмоциональной, оценочной и т.д. Последние достижения в области статистического моделирования естественного языка с помощью глубокого обучения позволяют избежать использования дополнительных признаков при обучении этих моделей [8, 9]. Такие модели самостоятельно формируют необходимые для решения задачи признаки, используемые во время обучения.

В данной работе для извлечения пользовательских мнений предлагается модель структурного предсказания на основе рекуррентной нейронной сети долгой краткосрочной памяти (long short-term memory, LSTM), которая не требует наличия таких анализаторов и полагается только на предобученные векторные представления. Для обучения и проверки качества модели используется размеченный авторами корпус (набор данных) отзывов о мобильных телефонах.

1. Формирование обучающей выборки для извлечения пользовательских мнений из текстов

В настоящее время существует небольшое количество наборов данных, предназначенных для обучения моделей детального анализа тональности. Так, в работе [10] предлагается размеченный набор отзывов из интернет-магазина Amazon — USAGE, состоящий из 800 отзывов по 8 категориям относительно простых товаров, вроде гостеров и кофе-машин. В основу предлагаемой схемы разметки текстов отзывов пользователей положены следующие сущности:

♦ **аспект** — важная характеристика или упоминание продукта (дополнительно приводится информация о том, принадлежит ли аспект продукту, о котором написан отзыв);

♦ **оценочное высказывание** — текст, содержащий высказанное пользователем мнение об аспекте (отмечается тональность высказывания);

♦ **разрешение кореференции** — для случаев, когда аспект является ссылкой на сущность из другого предложения в отзыве;

♦ **связь между аспектом и оценочным высказыванием** — если оценочное высказывание было связано с определенным аспектом.

Несмотря на доступность набора данных, авторами было решено сделать дополнительный размеченный корпус. Это обусловлено двумя причинами. Во-первых — желанием проверить качество работы алгоритмов извлечения мнений на потенциально более технически сложных и многофункциональных товарах (поэтому разметка проводилась на категории мобильных телефонов). Во-вторых — в силу ограниченности человеческих ресурсов авторы использовали более простую схему разметки по сравнению со схемой USAGE (в частности, не проводилось разрешение кореференции и в разметку вносились только пары «аспект — оценочное высказывание»).

Для разметки использовался корпус отзывов пользователей интернет-магазина Amazon [11]. Этот корпус содержит порядка 143 миллионов текстов отзывов о 25 категориях товаров, написанных в период с мая 1996 по июль 2014 года, вместе с метаданной, включающей названия, идентификаторы и описания продуктов, категории товаров, информацию о бренде и цене, идентификатор автора и оценку продукта пользователем. Аннотаторам предлагалось выделить в текстах отзывов спаны (непрерывные

последовательности слов), содержащие аспекты и оценочные высказывания. В отличие от набора данных USAGE, аспекты и оценочные высказывания выделялись строго парами «аспект — оценочное высказывание», которые вместе формируют пользовательские мнения. Далее мнением будем называть пару

$$O = \langle A(a_{begin}, a_{end}), D(d_{begin}, d_{end}) \rangle,$$

где $A(a_{begin}, a_{end})$ — аспект мнения, начинающийся на слове с порядковым номером a_{begin} и заканчивающийся на слове с номером a_{end} ;

$D(d_{begin}, d_{end})$ — оценочное высказывание, начинающийся на слове с порядковым номером d_{begin} и заканчивающийся на слове с номером d_{end} .

При этом спаны разных отзывов не должны пересекаться. Таким образом, размеченное предложение из текста отзыва с мнением O_1 , состоящего из аспекта с соответствующим спаном $A(1, 2)$ и оценочным высказыванием $D(4, 5)$ будет выглядеть следующим образом:

$$[Battery_1 life_2]_{A-O_1} is_3 [quite_4 impressive_5]_{D-O_1}!$$

Всего было размечено 3 232 отзыва. Размеченный корпус содержит 9 344 мнения, 1 994 уникальных аспекта, 5 124 уникальные оценочные выражения. Количественные характеристики размеченного корпуса приведены в таблице 1.

Таблица 1.

Количественные характеристики корпуса пользовательских отзывов о мобильных телефонах

Характеристика	Значение
Количество отзывов	3 232
Количество мнений в корпусе	9 344
Количество уникальных аспектов	1 994
Количество уникальных оценочных выражений	5 124

2. Модель извлечения пользовательских мнений

Задача извлечения пользовательских мнений может быть представлена как задача разметки последовательности, где для каждого элемента w_i входной последовательности $\{w_1, \dots, w_n\}$ необходимо определить метку класса (tag) y_i . Для этого необходимо преобразовать спаны мнений в последовательность тегов. Одним из наиболее популярных способов за-

дания спанов в виде последовательности является формат IOB [12]. В нем используется три типа тегов: **O** — отсутствие тега; **B-X** — начало новой сущности **X**; **I-X** — продолжение сущности **X**. В данной работе предложено выделять два типа сущностей — аспекты и оценочные высказывания, обозначенные словами Aspect и Description. Тогда множество доступных тегов **Y** будет состоять из следующих элементов: **O** — отсутствие сущности; **B-Aspect** — начало аспекта; **I-Aspect** — продолжение аспекта; **B-Description** — начало оценочного высказывания; **I-Description** — продолжение оценочного высказывания. Таким образом, предложению с набором мнений мы можем однозначно поставить в соответствие набор тегов, как это показано на рисунке 1.

$[Battery_1]$	$life_2]$	$Aspect-O_1$	is_3	$[good_4]$	$Description-O_1$
B-Aspect	I-Aspect	O		B-Description	

Рис. 1. Пример преобразования спанов мнений в последовательность тегов в формате IOB

В качестве формального аппарата решения задачи предлагается использовать рекуррентные нейронные сети (РНС). Данный тип нейронных сетей широко применяется для решения самых разнообразных задач машинного обучения: построения лингвистических моделей естественного языка [13], частеречной разметки [8], классификации последовательностей [14], распознавания аудиосигналов [15], прогнозирования временных рядов [16] и др.

На вход РНС на каждом шаге t подается произвольная последовательность, которая на основе рекуррентных связей между последовательностью скрытых состояний преобразуется в последовательность выходов:

$$h_t = f(U x_t + W h_{t-1} + b),$$

где x_t — текущий вход;

h_{t-1} — предыдущее скрытое состояние;

U и W — матрицы входного и рекуррентного взаимодействия;

b — вектор смещений;

f — нелинейная функция.

Наличие рекуррентных связей позволят сети передавать информацию об общем «контексте» обрабатываемой последовательности и учитывать ее при формировании выходов h_t . Таким образом, можно рассматривать h_t как промежуточное представле-

ние последовательности, в котором аккумулирована информация о прошлых шагах. В данном случае используем рекуррентную нейронную сеть типа LSTM [17], поскольку она менее подвержена проблеме затухающего градиента (vanishing gradient) [18] по сравнению с обычными РНС:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \odot \tanh(c_t). \end{aligned} \quad (1)$$

Для определения меток слов используется классификатор, работающий следующим образом: к каждому выходному значению применяется линейное преобразование с функцией софтмакс, результатом которого является распределение вероятности по возможным меткам:

$$P(\hat{y}_i | x_i)_i = \frac{\exp(W_{tag} h_i + b_{tag})_i}{\sum_j \exp(W_{tag} h_i + b_{tag})_j}. \quad (2)$$

Из определения общего вида РНС видно, что информация распространяется по сети в направлении слева направо. В некоторых случаях для повышения эффективности классификации конкретного слова необходимо знать контекст следующих за ним слов. Для того чтобы на каждом шаге предсказания была доступна информация из обоих направлений, применяются двунаправленные нейронные сети, одна из которых обрабатывает входную последовательность слева направо, а другая — справа налево, после чего соответствующие одной позиции скрытые состояния объединяются:

$$\begin{aligned} \bar{h}_t &= f(\bar{U} x_t + \bar{W} \bar{h}_{t-1} + \bar{b}), \\ h_t &= [\bar{h}_t; \bar{h}_t]. \end{aligned}$$

Следует отметить, что в задачах структурного предсказания (к которым относится и задача разметки последовательности) существуют зависимости между метками классов в выходной последовательности. Модели, которые не учитывают данные зависимости, могут получать на выходе некорректные последовательности меток. Например, в случае предсказания меток последовательности в формате IOB последовательность предсказаний **O I-Aspect** является ошибочной, т.к. элемент класса **I-Aspect** может следовать только после соответствующего ему **B-Aspect**.

Для обработки зависимостей между отдельными предсказаниями предлагается использовать модель условного случайного поля (conditional random field, CRF) предложенную в [19]:

$$s(W, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i}, \quad (3)$$

где A — матрица вероятностей переходов от метки y_i к метке y_{i+1} ;

P_{i, y_i} — вероятность метки y_i в позиции i ;

$W = \{w_1, w_2, \dots, w_n\}$ — входная последовательность;

$y = \{y_1, \dots, y_n\}$ — последовательность предсказаний.

Тогда вероятность последовательности y рассчитывается следующим образом:

$$p(y | W) = \frac{\exp(s(W, y))}{\sum_y \exp(s(W, y))}, \quad (4)$$

где суммирование в знаменателе происходит по всем возможным последовательностям y .

При обучении модели условного случайного поля максимизируется логарифм вероятности истинной последовательности:

$$\log(p(y | W)) = s(W, y) - \log\left(\sum_y \exp(s(W, y))\right). \quad (5)$$

Для поиска оптимальной последовательности предсказаний предлагается использовать метод динамического программирования. При этом оптимальная последовательность меток должна соответствовать максимуму выражения $p(y | W)$, а именно:

$$\hat{y} = \arg \max_y p(y | W).$$

Использование модели условных случайных полей позволяет получить глобально оптимальную структуру предсказания только для случая, когда результатом предсказания является линейная структура, использующая в каждом узле локальные признаки. Данные ограничения привели к развитию методов структурного предсказания, способных обрабатывать более сложные виды структур (например, деревья) и использовать нелокальные признаки (например, результаты классификации слов на предыдущих шагах) [20, 21]. С учетом этого для определения меток слов в предлагается использовать следующее выражение:

$$P(\hat{y}_i | \phi(c_i))_i = \frac{\exp(W_{tag} \phi(c_i) + b_{tag})_i}{\sum_j \exp(W_{tag} \phi(c_i) + b_{tag})_j}, \quad (6)$$

где c_t — конфигурация модели в момент t ;

$\phi(c_t)$ — функция, отображающая конфигурацию c_t в набор признаков.

Практическое применение выражения (6) для извлечения пользовательских мнений требует задания вида конфигурации c_t и функции получения признаков $\phi(c_t)$. Для этого предлагается использовать идеи из модели для построения деревьев зависимостей, представленные в работе [22]. Конфигурация определяется четверкой вида $c = (S, B, l, Y)$, состоящей из буфера B , содержащего необработанные на текущий момент элементы входной последовательности, стека S , в который складываются слова из найденных на текущий момент сущностей, метки последней найденной сущности l (Aspect или Description) и выходной последовательности меток Y . На каждом шаге t формируется вектор признаков $\phi(c_t)$, который используется для определения метки текущего слова $\hat{y}_t$ и изменения текущей конфигурации в соответствии с этой меткой. Конфигурация изменяется в соответствии с правилами, представленными в таблице 2. Символом «точка с запятой» в таблице обозначена конкатенация последовательностей.

Таблица 2.

Правила изменения элементов конфигурации

$\hat{y}_t$	S^{t+1}	l^{t+1}	Y^{t+1}	Предусловие
$B - y$	$b_t; S^t$	y	$Y^t; \hat{y}_t$	—
$I - y$	$b_t; S^t$	l^t	$Y^t; \hat{y}_t$	$\hat{y}_{t-1} \in \{B - y, I - y\}$
O	S^t	l^t	$Y^t; \hat{y}_t$	—

Исходя из предложенной формы конфигурации c и правил ее изменения, предлагается использовать $\phi(c_t)$ следующего вида:

$$\phi(c_t) = [B_t, \dots, B_{t+M_B}, S_1^{t-1}, \dots, S_{M_S}^{t-1}, Y_{t-1}^{t-1}, Y_{t-2}^{t-1}, E_{l^t}^t],$$

где B_t — i -й элемент буфера B ;

S_j^{t-1} — j -й элемент стека S на шаге t ;

E_k^t — k -я строка матрицы E ;

Y_n^t — n -й элемент последовательности меток на шаге t .

В качестве элементов буфера используются скрытые состояния, получаемые из последнего слоя многослойной двунаправленной LSTM. Так, каждый элемент в буфере будет содержать информацию не только о слове на соответствующей позиции, но и о предшествующем и последующих контекстах. Авто-

ры полагают, что информация о словах из стека найденных сущностей будет способствовать корректному обнаружению начала последующих сущностей. Например, найденный аспект «время жизни от батареи» может подсказать модели, что следующее слово «отличное» является оценочным высказыванием. Кроме того, дополнительной подсказкой является метка предыдущей найденной сущности l^{t+1} . При знаками для l являются строки матрицы $E \in \mathbb{R}^{2 \times d}$. Входными данными для LSTM являются векторные представления слов, полученные с помощью модели FastText [23]. В данной работе использовались предобученные векторные представления².

При вычислении вероятностей классов $P(\hat{y}_t | \phi(c_t))_i$ рассчитываются только вероятности меток класса, которые не противоречат предположениям из *таблицы 2*. Это позволяет избежать получения некорректной последовательности предсказаний.

3. Алгоритм разметки последовательности

Исходя из вышесказанного, алгоритм разметки последовательности можно представить следующим образом.

Пусть исходный текст задан в виде последовательности слов $W\{w_1, w_2, \dots, w_n\}$. Требуется определить выходную последовательность меток слов $Y\{y_1, y_2, \dots, y_n\}$. Для обучения модели используется алгоритм оптимизации Adam [24] с параметрами $lr = 0,001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$ и нормировкой градиента на уровне 3,0.

Шаг 1. Инициализируем начальное состояние $B = \emptyset$, $C = \emptyset$, $l = 0$, $Y = \emptyset$, $t = 0$.

Шаг 2. Наполняем буфер скрытыми представлениями входной последовательности, полученными с помощью сети LSTM: $\forall t \ B_t = LSTM(w_t, B_{t-1})$.

Шаг 3. Если $t < n$, то переход на шаг 4, иначе переход на шаг 6.

Шаг 4. Формируем вектор признаков $\phi(c_t)$ и определяем тег для текущей позиции в буфере: $\hat{y}_t = \arg \max P(\hat{y}_t | \phi(c_t))$.

Шаг 4. Изменяем S , i , Y в соответствии с правилами из *таблицы 2* в зависимости от метки $\hat{y}_t$.

Шаг 5. $t = t + 1$, переход на шаг 3.

Шаг 6. Конец.

Результаты работы предложенных модели и алгоритма на примере обработки отзывов о мобильных телефонах из интернет-магазина Amazon приведены

в *таблице 3*. Например, в предложении из отзыва на мобильный телефон «The screen is fantastically large while the overall dimensions of the phone are manageable for those without giant hands» были извлечены мнения «screen is fantastically large» и «dimensions of the phone are manageable».

Таблица 3.

Мнения, извлеченные из отзывов на мобильные телефоны

Товар	Мнения
Sony Xperia XA	Phone is awesome; phone is easy to use; phone is perfect for those who need extra storage; battery life is mediocre; battery life is absolutely terrible compared; no great sound
Apple iPhone 6S	It didn't work properly from the beginning; it's a decent; bad charger; worry free product
Huawei P20	Phone is a flagship performer; phone stopped receiving phone calls; phone is absolutely amazing for the price; the screen is fantastically large; camera is simply amazing; fantastic camera; camera produces great photos

Кроме того, на базе размеченного набора данных были обработаны мнения пользователей о других категориях товаров (*таблица 4*). Результаты апробации позволяют говорить о том, что модель показала хорошие результаты как для извлечения мнений о мобильных телефонах, так и о товарах других категорий.

Таблица 4.

Мнения, извлеченные из отзывов на товары других категорий

Товар	Мнения
EltaMD PM Therapy Facial Moisturizer	Is a great night cream product; product highly emollient without being greasy; product recommended by my dermatologist; very moisturized skin feel
Samsung UN55MU6500 Curved 55-Inch 4K Ultra HD Smart LED TV	Easy setup TV; TV is not worth the money or aggravation; wonderful picture; picture is so clear; remote is easy to use; remote is ergonomic and a breeze to use; color is unbelievable
Fire 7 Tablet with Alexa	Small tablet; amazing little tablet; screen does not react well to water; screen is freezing; battery doesn't hold a charge; battery dies to quickly; charging port its weakness

4. Экспериментальные исследования модели

Экспериментальные исследования предложенной модели проводились в сравнении с моделью на основе двунаправленной LSTM и CRF из работы [22], без использования знаковых представлений. Модели обучались методом обратного распростра-

² Источник: <https://github.com/plasticityai/magnitude>

Таблица 5.

**Результаты извлечения мнений из набора данных
о мобильных телефонах (строгое соответствие)**

Модель	Аспекты			Оценочные высказывания		
	R	P	F1	R	P	F1
Bi-LSTM + CRF	39,20	50,58	44,17	41,30	54,03	46,82
Предложенная модель	47,87	49,51	48,68	49,70	52,93	52,26

Таблица 6.

**Результаты извлечения мнений из набора данных
о мобильных телефонах (мягкое соответствие)**

Модель	Аспекты			Оценочные высказывания		
	R	P	F1	R	P	F1
Bi-LSTM + CRF	53,93	63,05	57,83	56,09	64,93	60,19
Предложенная модель	61,08	62,74	61,9	62,49	64,98	63,71

нения ошибки. Для оптимизации параметров использовался алгоритм Adam [24] с параметрами l $r = 0,001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$ и нормировкой градиента на уровне 3,0.

Для извлечения пользовательских мнений использовался описанный в разделе 1 набор данных. Из-за небольшого объема данных оценки рассчитывались процедурой скользящего контроля с разбиением на пять выборок. Для исключения влияния лексики из обучающей выборки на проверочную выборку для скользящего контроля разделились на уровне отдельных документов.

В модели CRF и предлагаемой модели использовалась двухслойная двунаправленная LSTM со скрытым состоянием из 100 элементов, на вход которой подавались фиксированные векторы размера 100. Для задания конкретного вида функции $\phi(t)$ в качестве исходных были использованы следующие значения гиперпараметров: $C_{BT} = 2$, $C_{ST} = 4$, $e_{label} \in \mathbb{R}^{20}$.

При проведении эксперимента использовались стандартные для такого типа задач критерии: точность (precision), полнота (recall) и F-мера (F1-measure), определяющая общее качество модели. Численные значения критериев рассчитывались двумя способами: строгим (strict) — когда совпадение засчитывается только в случае полного соответствия найденных спанов с истинными; мягким — когда соответствие засчитывается, если истинный и найденный спан имеют хотя бы одно общее слово.

Анализ таблиц 5 и 6 показывает, что для обоих способов расчета критериев предлагаемая модель демонстрирует лучшие по сравнению с моделью на основе Bi-LSTM и CRF результаты. Для строгого соответствия улучшение обнаружения спанов аспектов составляет 4,51%, а спанов оценочных высказываний — 5,44%. Для мягкого соответствия улучшение обнаружения спанов аспектов составляет 3,77%, а спанов оценочных высказываний — 3,52%.

Заключение

Предложенный подход, как совокупность размеченного набора данных и модели структурного предсказания с использованием рекуррентной сети типа LSTM, позволяет извлекать из текстов отзывы покупателей мнения о потребительских свойствах как товара в целом, так и его отдельных элементов. Разработанная модель структурного предсказания с использованием рекуррентной нейронной сети способна использовать нелокальные признаки для предсказания сущностей и не требует использования дополнительных синтаксических признаков.

Модель, обученная на наборе данных, показала лучшие по сравнению с моделью на основе CRF результаты: F1 для аспектов выше на 4,51%, F1 для оценочных высказываний выше на 5,44%. Экспериментальные исследования показали, что введение дополнительных признаков в базовую модель положительным образом влияет на результаты.

Полученные результаты могут быть полезны специалистам в области обработки естественного языка и компьютерной лингвистики, а также бизнес-сообществу при реализации товаров и услуг и развитии их потребительских свойств.

Вопросы предсказания связей между аспектами

и оценочными высказываниями для улучшения качества извлекаемых мнений, а также внесение в разметку информации о тональности мнений и принадлежности аспекта главному объекту отзыва и кореференции будут рассмотрены в последующих работах. ■

Литература

1. Sadegh M., Ibrahim R., Othman Z.A. Opinion mining and sentiment analysis: A survey // *International Journal of Computers & Technology*. 2012. Vol. 2. No. 3. P. 171–178.
2. Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: A survey // *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2018. Vol. 8. No. 4. P. 1942–4787.
3. SemEval-2016 task 5: Aspect based sentiment analysis / Pontiki M. [et al.] // *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, CA, USA, 16–17 June 2016. P. 19–30.
4. Jo Y., Oh A.H. Aspect and sentiment unification model for online review analysis // *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011)*. Hong Kong, China, 9–12 February 2011. P. 815–824.
5. Zirn C., Niepert M., Stuckenschmidt H., Strube M. Fine-grained sentiment analysis with structural features // *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*. Chiang Mai, Thailand, 8–13 November 2011. P. 336–344.
6. Yang B., Cardie C. Extracting opinion expressions with semi-Markov conditional random fields // *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*. Jeju Island, Korea, 12–14 July 2012. P. 1335–1345.
7. Yang B., Cardie C. Joint inference for fine-grained opinion extraction // *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, 4–9 August 2013. Vol. 1. P. 1640–1649.
8. Natural language processing (almost) from scratch / Collobert R. [et al.] // *Journal of Machine Learning Research*. 2011. No. 12. P. 2493–2537.
9. Zhai F., Potdar S., Xiang B., Zhou B. Neural models for sequence chunking // *Proceedings of the Thirty-First Conference on Artificial Intelligence (AAAI-17)*. San Francisco, CA, USA, 4–9 February 2017. P. 3365–3371.
10. Klinger R., Cimiano P. The USAGE review corpus for fine-grained, multi-lingual opinion analysis // *Proceedings of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland, 26–31 May 2014. P. 2211–2218.
11. He R., McAuley J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering // *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)*. Montreal, Canada, 11–15 April 2016. P. 507–517.
12. Sang E.F., Veenstra J. Representing text chunks // *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics (EACL 1999)*. Bergen, Norway, 8–12 June 1999. P. 173–179.
13. Recurrent neural network based language model / Mikolov T. [et al.] // *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*. Makuhari, Chiba, Japan, 26–30 September 2010. Vol. 2. P. 1045–1048.
14. Ghosh M., Sanyal G. Document modeling with hierarchical deep learning approach for sentiment classification // *Proceedings of the 2nd International Conference on Digital Signal Processing (ICDSP 2018)*. Tokyo, Japan, 25–27 February 2018. P. 181–185.
15. Graves A., Jaitly N., Mohamed A. Hybrid speech recognition with deep bidirectional LSTM // *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013)*. Olomouc, Czech Republic, 8–12 December 2013. P. 273–278.
16. Robust online time series prediction with recurrent neural networks / Guo T. [et al.] // *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA 2016)*. Montreal, Canada, 17–19 October 2016. P. 816–825.
17. Hochreiter S., Schmidhuber J. Long short-term memory // *Neural Computation*. 1997. Vol. 9. No. 8. P. 1735–1780.
18. Bengio Y., Simard P., Frasconi P. Learning long-term dependencies with gradient descent is difficult // *IEEE Transactions on Neural Networks*. 1994. Vol. 5. No. 2. P. 157–166.
19. Lafferty J., McCallum A., Pereira F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*. Williamstown, MA, USA, 28 June – 1 July 2001. P. 282–289.
20. Chen D., Manning C.D. A fast and accurate dependency parser using neural networks // *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Doha, Qatar, 25–29 October 2014. P. 740–750.
21. Transition-based dependency parsing with stack long short-term memory / Dyer C. [et al.] // *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, 27–31 July 2015. Vol. 1. P. 334–343.
22. Neural architectures for named entity recognition / Lample G. [et al.] // *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL 2016)*. San Diego, CA, USA, 12–17 June 2016. P. 260–270.
23. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information // *Transactions of the Association for Computational Linguistics*. 2017. No. 5. P. 135–146.
24. Kingma D.P., Ba J.L. Adam: A method for stochastic optimization // *arXiv:1412.6980v9 [cs. LG]*. 2017. [Электронный ресурс]: <https://arxiv.org/pdf/1412.6980.pdf> (дата обращения 10.10.2018).

User opinion extraction model concerning consumer properties of products based on a recurrent neural network³

Yuri P. Yekhlakov

Professor, Department of Data Processing Automation
Tomsk State University of Control Systems and Radioelectronics
Address: 40, Prospect Lenina, Tomsk, 634050, Russia
E-mail: upe@tusur.ru

Egor I. Gribkov

Doctoral Student, Department of Data Processing Automation
Tomsk State University of Control Systems and Radioelectronics
Address: 40, Prospect Lenina, Tomsk, 634050, Russia
E-mail: drnemor@gmail.com

Abstract

This article offers a long short-term memory (LSTM) based structured prediction model taking into account existing approaches to sequence tagging tasks and allowing for extraction of user opinions from reviews. We propose a model configuration and state transition rules which allow us to use past predictions of the model alongside sentence features. We create a body of annotated user reviews about mobile phones from Amazon for model training and evaluation. The model trained on reviews corpus with recommended hyperparameter values. Experiment shows that the proposed model has a 4.51% increase in the F1 score for aspects detection and a 5.44% increase for aspect descriptions compared to the conditional random field (CRF) model with the use of LSTM when F1 spans are matched strictly.

The extraction of user opinions on mobile phones from reviews outside of the collected corpus was conducted as practical confirmation of the proposed model. In addition, opinions from other product categories like skin care products, TVs and tablets were extracted. The examples show that the model can successfully extract user opinions from different kinds of reviews. The results obtained can be useful for computational linguists and machine learning professionals, heads and managers of online stores for consumer preference determination, product recommendations and for providing rich catalog searching tools.

Key words: user feedback; deep learning; machine learning; natural language processing; opinion processing.

Citation: Yekhlakov Yu.P., Gribkov E.I. (2018) User opinion extraction model concerning consumer properties of products based on a recurrent neural network. *Business Informatics*, no. 4 (46), pp. 7–16. DOI: 10.17323/1998-0663.2018.4.7.16

References

1. Sadegh M., Ibrahim R., Othman Z.A. (2012) Opinion mining and sentiment analysis: A survey. *International Journal of Computers & Technology*, vol. 2, no. 3, pp. 171–178.
2. Zhang L., Wang S., Liu B. (2018) Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. 1942–4787.
3. Pontiki M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S., Mohammad Al-S., Al-Ayyoub M., Zhao Y., Qin B., De Clercq O. (2016) SemEval-2016 task 5: Aspect based sentiment analysis. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, CA, USA, 16–17 June 2016, pp. 19–30.
4. Jo Y., Oh A.H. (2011) Aspect and sentiment unification model for online review analysis. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011). Hong Kong, China, 9–12 February 2011, pp. 815–824.
5. Zirn C., Niepert M., Stuckenschmidt H., Strube M. (2011) Fine-grained sentiment analysis with structural features. Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011). Chiang Mai, Thailand, 8–13 November 2011, pp. 336–344.

³ This study was conducted under government order of the Ministry of Education and Science of Russia, project No. 8.8184.2017/8.9

6. Yang B., Cardie C. (2012) Extracting opinion expressions with semi-Markov conditional random fields. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*. Jeju Island, Korea, 12–14 July 2012, pp. 1335–1345.
7. Yang B., Cardie C. (2013) Joint inference for fine-grained opinion extraction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, 4–9 August 2013. Vol. 1, pp. 1640–1649.
8. Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P. (2011) Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, no. 12, pp. 2493–2537.
9. Zhai F., Potdar S., Xiang B., Zhou B. (2017) Neural models for sequence chunking. *Proceedings of the Thirty-First Conference on Artificial Intelligence (AAAI-17)*. San Francisco, CA, USA, 4–9 February 2017, pp. 3365–3371.
10. Klinger R., Cimiano P. (2014) The USAGE review corpus for fine-grained, multi-lingual opinion analysis. *Proceedings of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland, 26–31 May 2014, pp. 2211–2218.
11. He R., McAuley J. (2016) Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)*. Montreal, Canada, 11–15 April 2016, pp. 507–517.
12. Sang E.F., Veenstra J. (1999) Representing text chunks. *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics (EACL 1999)*. Bergen, Norway, 8–12 June 1999, pp. 173–179.
13. Mikolov T., Karafiat M., Burget L., Cernocky J., Khudanpur S. (2010) Recurrent neural network based language model. *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*. Makuhari, Chiba, Japan, 26–30 September 2010, vol. 2, pp. 1045–1048.
14. Ghosh M., Sanyal G. (2018) Document modeling with hierarchical deep learning approach for sentiment classification. *Proceedings of the 2nd International Conference on Digital Signal Processing (ICDSP 2018)*. Tokyo, Japan, 25–27 February 2018, pp. 181–185.
15. Graves A., Jaitly N., Mohamed A. (2013) Hybrid speech recognition with deep bidirectional LSTM. *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013)*. Olomouc, Czech Republic, 8–12 December 2013, pp. 273–278.
16. Guo T., Xu Z., Yao X., Chen H., Aberer K., Funaya K. (2016) Robust online time series prediction with recurrent neural networks. *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA 2016)*. Montreal, Canada, 17–19 October 2016, pp. 816–825.
17. Hochreiter S., Schmidhuber J. (1997) Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735–1780.
18. Bengio Y., Simard P., Frasconi P. (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166.
19. Lafferty J., McCallum A., Pereira F.C.N. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*. Williamstown, MA, USA, 28 June – 1 July 2001, pp. 282–289.
20. Chen D., Manning C.D. (2014) A fast and accurate dependency parser using neural networks. *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Doha, Qatar, 25–29 October 2014, pp. 740–750.
21. Dyer C., Ballesteros M., Ling W., Matthews A., Smith N.A. (2015) Transition-based dependency parsing with stack long short-term memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, 27–31 July 2015, vol. 1, pp. 334–343.
22. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. (2016) Neural architectures for named entity recognition. *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL 2016)*. San Diego, CA, USA, 12–17 June 2016, pp. 260–270.
23. Bojanowski P., Grave E., Joulin A., Mikolov T. (2017) Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, no. 5, pp. 135–146.
24. Kingma D.P., Ba J.L. (2017) *Adam: A method for stochastic optimization*. arXiv:1412.6980v9 [cs. LG]. Available at: <https://arxiv.org/pdf/1412.6980.pdf> (accessed 10 October 2018).