А.С. Скворцова (Минск, МГЛУ)

ПОДХОДЫ К АВТОМАТИЧЕСКОМУ ВЫЯВЛЕНИЮ ОЦЕНОЧНЫХ СУЖДЕНИЙ

В последнее время широкое распространение получили задачи анализа субъективности текста и извлечения из него не фактов, а мнений (оценок) автора относительно какого-то объекта или его характеристики, например, репутации организации. Это связано, в первую очередь, с возросшей популярностью блогов, форумов, социальных сетей, специализированных сайтов отзывов, а также с увеличением объемов контента пользователей сети Интернет. Необходимо отметить, что поиск и анализ оценочных суждений часто затруднен, поскольку полезной может оказаться даже небольшая часть длинного сообщения или одно сообщение, выделенное из нескольких страниц форума, блога или социальной сети. При этом наибольшую трудность в выявлении субъективных конструкций представляет отделение оценочной информации от фактической.

Мнения пользователей о каком-либо объекте часто выражаются посредством оценочных слов и выражений, которые несут в себе некоторую положительную или отрицательную оценку. Препятствием для решения этой задачи является высокая вариативность оценочной лексики, не позволяющая сформулировать достаточное количество универсальных закономерностей, выполняющихся для каждой предметной области [1, с. 3]. Невозможно заранее сформировать полный список оценочных слов и выражений, которые будут применимы для всех предметных областей, поскольку некоторые оценочные единицы употребляются только в конкретных предметных областях, а другие являются оценочными в одной области и не являются оценочными в другой. Определенную часть оценочных слов нельзя найти в словарях, например, слово *никакущий*. Иногда то или иное слово в определенном контексте может употребляться как оценочное, например, слово *скомканный* (о фильмах или книгах) либо выраженная одной и той же фразой оценка может быть положительной в одном случае (*непредсказуемый сюжет фильма*) и отрицательной в другом (*непредсказуемое поведение программы*), либо вовсе не являться оценкой.

Существует много различных подходов к автоматическому извлечению оценочных суждений из текста. Рассмотрим некоторые из них. Один из основных подходов к автоматическому извлечению оценочных слов из текстов базируется на информации из словарей, чаще всего — тезаурусов. При этом вручную выбирается небольшое начальное множество слов, которое затем дополняется с помощью словарей. Основной принцип заключается в том, что если слово оценочное, то его синонимы и антонимы также будут оценочными (возможна только смена ориентации). Поэтому, имея слова из начального множества, можно с помощью этих связей составить новое множество, которое будет более полным [2]. В работе [3] на основе толкований слов в словаре определяется их ориентация

(положительная или отрицательная). Основная идея заключается в том, что слова с одинаковой ориентацией имеют похожие толкования. Таким образом, создается классификатор, который делит слова на положительно ориентированные и отрицательно ориентированные.

Корпусный подход к автоматическому извлечению оценочных слов из текстов основан на поиске правил и закономерностей в текстах. В работе [2] оценочная характеристика слова вычисляется путем сопоставления совместной встречаемости данного слова со словами «отличный» и «плохой» в конкретной предметной области. Полученная оценочная направленность слов используется для классификации отзывов на положительные и отрицательные.

В работе [4] предлагается следующая модель структуры оценочной конструкции. Объект O — это сущность, которая может быть продуктом, человеком, событием, организацией или темой. Она связана с парой (T, A), где T — иерархия или таксономия компонентов и подкомпонентов, A — список атрибутов O. Каждый компонент имеет свое собственное множество компонентов и атрибутов. Например, конкретная модель фотоаппарата является объектом. У него есть определенное число компонентов — объектив, батарея, видоискатель и т.д., а также множество атрибутов — качество картинки, вес, размер и т.п. У объектива также есть множество атрибутов — светосила, искажения, и т.д. То есть, структуру объекта можно представить в виде дерева, корнем которого является сам объект, а вершинами — его компоненты или подкомпоненты. Каждому ребру дерева можно поставить в соответствие отношение включения, каждой вершине — набор атрибутов. Мнение может быть высказано о любой вершине и любом атрибуте вершины. С целью упрощения дальнейших описаний для обозначения, как компонентов, так и атрибутов используют понятие *свойство*. Отметим, что в этой модели сам объект также рассматривается как свойство. Если свойство встречается в оценочном тексте (*Качество картинки потрясающее*), оно называется явным свойством. Если свойство не встречается в тексте, но подразумевается (*Наушники сломались через два дня после покупки*), оно называется неявным свойством [5].

Методы выделения оценочных суждений из текста можно разделить на группы в зависимости от типа исходной информации и поставленных целей. В работе [5] представлены следующие три основных подхода:

1. Классификация мнений. В данном случае задача извлечения мнений рассматривается как задача классификации текстов. Детали того, что именно понравилось или не понравилось пользователю, не рассматриваются. Главная цель классификации мнений — быстрое определение эмоциональной направленности текста, общего впечатления, оценка преобладающего мнения об объекте. Выделяют два основных подхода к классификации мнений об объекте: лексический и машинный. При машинном подходе компьютерная система обучается на очень большом корпусе размеченных текстов, в основном без использования словарей. Такие системы имеют достаточно высокую точность, однако эта точность может быть достигнута лишь в одной

предметной области. Лексический подход предполагает индивидуальный анализ слов и/или фраз с использованием словарей эмоциональной/тональной лексики. В зависимости от направления тональности слову из текста приписывается определенный эмоциональный вес, веса всех слов суммируются и выводится общий эмоциональный вес текста.

2. Извлечение свойств. При использовании данного подхода из текста выделяются отдельные фрагменты, относящиеся к объекту и его свойствам, а затем вычисляется их семантическая ориентация. Данное понятие, введенное в 1997 г. В. Хатзивассилоглоу и К. МакКеоуном [6] означает процедуру измерения степени «эмоциональности», «эмоциональной направленности» слова. В 1957 г. американский психолог Ч. Осгуд ввел понятие семантического дифференциала для выявления коннотаций слов. Группе испытуемых предлагалось оценить множество понятий по набору биполярных градуированных шкал. Как наиболее значимые выделялись шкалы оценки (хороший/плохой), силы (сильный/слабый) и активности (активный/пассивный). В. Хатзивассилоглоу и К. МакКеоун взяли это понятие за основу, сохранив только шкалу оценки. Расположение слова на этой шкале получило название «семантической ориентации». Положительная семантическая ориентация означает положительную оценку, похвалу, отрицательная — отрицательную оценку, критику.

Разновидностью второго подхода является метод лексических шаблонов. Использовать лексические шаблоны в данном случае позволяет то, что синтаксическая структура оценочных конструкций, в отличие от лексики, практически не зависит от предметной области. Предлагаемый метод состоит из трех шагов.

Шаг 1. В анализируемом тексте производится разметка частей речи и разбивка предложений на сегменты. Сегментом называется предложение или фрагмент предложения, не содержащий противопоставлений. Отсутствие противопоставлений внутри сегмента в большинстве случаев позволяет однозначно оценить его семантическую направленность. В качестве сегментов выделяют простые предложения, сложные предложения, не содержащие противопоставлений, и части сложных предложений, разделенные союзами *но, зато, несмотря, хотя, однако*. Каждый сегмент представляется как последовательность, являющаяся «заготовкой» для шаблона. Например, предложение *Это плохие наушники* превращается в следующую последовательность: {*это*, PRN}{*плохие*, ADJ} {*наушники*, NN}.

Шаг 2. Чтобы разбить сегменты на более короткие, используются *n*-граммы. Обычно оказывается достаточно триграмм. Если не использовать *n*-граммы, длинные сегменты приводят к порождению неподходящих, редко встречающихся или просто некорректных шаблонов.

Шаг 3. Все слова, обозначающие свойства, заменяются меткой свойства *$feature*. Эта замена необходима, так как разные объекты обладают разными свойствами, а замена гарантирует, что можно найти общий для всех шаблон. После замены указанная выше последовательность приобретает следующий

вид: {*это*, PRN}{*плохие*, ADJ}{$feature, NN}. Такая последовательность называется лексическим шаблоном.

Все лексические шаблоны накапливаются в базе данных и используются для анализа текстов новых отзывов. При этом совпадения с лексическими шаблонами базы данных определяются в каждом сегменте текста отзыва. Слово, совпадающее с меткой $feature, считается свойством. При этом возможны три ситуации:

1) если сегмент удовлетворяет нескольким правилам, то устанавливается очередность правил, исходя из того, какие части речи чаще оказываются свойствами;

2) если к сегменту не применимо ни одно из правил, из него извлекаются имена существительные, как будто они были свойствами;

3) если сегмент состоит из одного слова, шаблоны не применяются, а это слово сразу рассматривается как свойство [1].

3. Анализ сравнительных конструкций. Прямое выражение положительного или отрицательного мнения — только одна из форм оценки. Сравнение объекта с другими объектами также является оценкой, иногда более убедительной для пользователя. По семантике и синтаксической структуре сравнительные конструкции отличаются от обычных оценочных высказываний.

Таким образом, с развитием технологий глубинного семантического анализа текста стало возможным автоматически извлекать и анализировать оценочные суждения пользователей Сети, касающиеся различных объектов, в том числе и репутации организации.

ЛИТЕРАТУРА

1. Кравченко, А.Н. Генерация шаблонов оценочных выражений на основе неразмеченного текста // Компьютерная лингвистика и компьютерные технологии [Электронный ресурс]. — Режим доступа: http://www.dialog-21.ru/digests/dialog2012/materials/pdf /129.pdf. — Дата доступа: 10.02.2017.
2. Hu, M., Liu, B. Mining and Summarizing Customer Reviews [Electronic resource].— Mode of acess: http://users.cis.fiu.edu/~lli003/Sum/ KDD/2004/p168-hu.pdf. — Date of access: 15.02.2017.
3. Минина, М.А. Психолингвистический анализ семантики оценки: автореф. дис. …канд. фил. наук: 10.02.04; АН Росии, Институт языкознания. — М., 1995.
4. Liu, B. Web Data Mining. — Berlin, 2007.
5. Liu, B. Sentiment Analysis and Subjectivity in Handbook of Natural Language Processing. — NY, 2010. — P. 257–282.
6. Hatzivassiloglou, V., McKeown, K. Predicting the Semantic Orientation of Adjectives in Proceedings of ACL/EACL. — Madrid, 1997. — P. 174–181.