ГЕНЕРАЦИЯ ШАБЛОНОВ ОЦЕНОЧНЫХ ВЫРАЖЕНИЙ НА ОСНОВЕ НЕРАЗМЕЧЕННОГО ТЕКСТА

Кравченко А. Н. (etercian@gmail.com)

Высшая Школа Экономики (ГУ), Москва, Россия

Одной из ключевых задач при извлечении мнений является отделение оценочных конструкций от фактической информации. Препятствием для решения этой задачи является высокая вариативность оценочной лексики для различных предметных областей, не позволяющая сформулировать достаточное количество универсальных закономерностей, выполняющихся для каждой области.

В данной работе предлагается алгоритм выделения шаблонов оценочных конструкций для заданной предметной области, решающий эту проблему. Алгоритм основан на методе синтаксических шаблонов и включает в себя автоматические разметку обучающего корпуса и фильтрацию сформированных шаблонов, позволяющие свести человеческое участие к минимуму.

Работа метода была проанализирована для трех различных предметных областей.

Данное исследование проводилось при финансовой поддержке Правительства Российской Федерации (Минобрнауки России) в рамках договора № 13.G25.31.0096 о «Создании высокотехнологичного производства кросс-платформенных систем обработки неструктурированной информации на основе свободного программного обеспечения для повышения эффективности управления инновационной деятельностью предприятия в современной России».

Ключевые слова: извлечение мнений, определение тональности, лексические шаблоны, тональность

AUTOMATIC GENERATION OF EXTRACTION PATTERNS FOR SUBJECTIVE EXPRESSIONS FROM UNTAGGED TEXT

Kravchenko A. N. (etercian@gmail.com)

Higher School of Economics, Moscow, Russian Federation

The goal of opinion mining is to extract and summarize opinionated contents from news, blogs, comments and reviews. One of the main tasks in opinion mining is detecting the boundaries of opinionated expressions and distinguishing between subjective expressions and factual information. High lexicon diversity for different domains excludes the possibility of formulating universal extraction rules that would work for any area of knowledge.

In this paper we suggest a solution for this problem, reviewing a classification of subjective expressions in Russian and proposing an algorithm for automatic generation of extraction patterns for subjective expressions from untagged text based on label sequential rules (LSR). The algorithm also includes automatic tagging of the training corpora and result filtering to minimize the need for human participation.

At first the proposed algorithm uses an assortment of domain-independent pivots to distinguish opinionated sentences from the factual ones, which allows to avoid manual tagging. Possible subjective expressions are then extracted from selected sentences using a set of syntactic patterns. The applicability of this method is based on the fact that syntactic structure of subjective expressions is domain-independent as well. The resulting subjective expressions are, on the contrary, domain-specific.

After that, the expressions are filtered with a use of probabilistic algorithm, increasing precision and therefore minimizing the need for human participation

The effectiveness of the proposed approach was evaluated on an collection of approximately 300 000 sentences, gathered from three different domains user reviews on movies, headphones and photo cameras. The best results (80 % precision) were shown on domains with existing objective criteria and low lexical variability, such as reviews on cameras and headphones. For movie reviews precision reached 64,3 % after filtering.

Key words: opinion mining, sentiment analysis, extraction patterns, subjective expressions, text mining

1. Задача отделения эмоциональной информации от фактической.

В последнее время получили широкое распространение задачи оценки тональности текста и извлечения мнений (opinion mining, sentiment analysis), фокусирующихся на выделении в тексте эмоций, а не фактов. Это связано в первую очередь с возросшей популярностью блогов и социальных сетей и увеличением объемов пользовательского контента. Мнения — источник ценной информации как для социологов, маркетологов и журналистов, так и для самих пользователей, при этом поиск подобной информации часто затруднен полезной может оказаться небольшая часть длинного сообщения или одно сообщение из нескольких страниц темы форума.

Наибольшую сложность в решении этой задачи представляет отделение эмоциональной информации от фактической и выделение субъективных конструкций. Эффективным методом являются лексические шаблоны, но на практике возникает препятствие, связанное с высокой вариативностью оценочной лексики для различных предметных областей. Например, выраженная одной и той же фразой оценка может быть положительной в одном случае и отрицательной в другом («непредсказуемый сюжет фильма» и «непредсказуемое поведение программы»), или не являться оценкой вовсе. Таким образом, составленный для одной предметной области набор шаблонов нельзя использовать даже для смежных областей, что делает ручное составление шаблонов неоправданным. Другим серьезным препятствием является недостаток обучающих данных.

В данной работе предлагается решение этой проблемы с помощью алгоритма извлечения шаблонов оценочных выражений для заданной предметной области. Алгоритм основывается на выявленных при изучении тонально окрашенных текстов закономерностях, таких как схожесть синтаксической структуры оценочных текстов и повторяемость типичных конструкций. Благодаря использованию этих закономерностей, работа алгоритма не требует предварительной разметки корпуса, составления словаря предметной области и сводит к минимальной необходимость фильтрации результата.

Полученные шаблоны могут использоваться для получения отзывов в интернет-магазинах для заданной категории продуктов или для выявления тенденций в блогах или новостных обзорах.

2. Основные подходы к извлечению оценки

2.1. Модель оценочной конструкции

Существуют разные подходы к формализации оценки. Например, в своей книге Bing Liu [Liu, 2007] предлагает следующую модель мнения или оценочной конструкции:

Определение: *объект* O — это сущность, которая может быть продуктом, человеком, событией, организацией или темой. Она связана с парой (T, A), где A иерархия или таксономия ***компонентов*** и ***подкомпонентов***, а A список ***атрибутов*** O. Каждый компонент имеет свое собственное множество компонентов и атрибутов.

Конкретная модель фотоаппарата является объектом. У него есть множество компонентов: объектив, батарея, видоискатель, итд, и также множество атрибутов: качество картинки, вес, размер, итд. У объектива также есть множество атрибутов: светосила, искажения, и т. д.

То есть, объект можно представить как дерево, где корнем дерева является сам объект, а все прочие вершины компоненты или подкомпоненты объекта. Каждому ребро дерева можно сопоставить отношение включения, каждой вершине — набор атрибутов. Мнение может быть высказано о любой вершине и любом атрибуте вершины.

В данной работе для того чтобы упростить дальнейшие описания, будем использовать слово «***свойство***» для обозначения как компонентов, так и атрибутов, что позволит нам опустить иерархию. 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***.

2.2. Определение направленности оценки

Для определения направленности оценки используется понятие семантической ориентации (semantic orientation). Это метод измерения степени «эмоциональности», «эмоциональной направленности» слова, введенный в 1997г Hatzivassiloglou и McKeown [Hatzivassiloglou& McKeown, 1997].

В 1957 американский психолог Чарльз Осгуд ввел для выявления коннотаций слов метод семантического дифференциала. Группе испытуемых предлагалось оценить множество понятий по набору биполярных градуированных шкал. В исходной работе как наиболее значимые выделялись шкалы «оценки» (хороший/плохой), «силы» (сильный/слабый) и «активности» (активный/пассивный).

Hatzivassiloglou и McKeown взяли это понятие за основу, оставив только шкалу оценки. Расположение терма на этой шкале получило название «семантической ориентации» (semantic orientation). Положительная семантическая ориентация означает положительную оценку, похвалу, отрицательная отрицательную оценку, критику.

Оценка семантической ориентации вручную слишком трудоемка, поэтому создаются автоматизированные способы, например оценка с помощью WordNet или взаимной поточечной информации. Подробный обзор существующих методов можно найти в работе Turney [Turney, 2003].

2.3. Методы выделения оценочных конструкций.

Методы выделения оценки из текста можно разделить на группы, в зависимости от типа исходной информации и поставленных целей: Liu в своей книге выделяет три основных подхода

1. Классификация мнений (sentiment classification). Задача извлечения мнений рассматривается как задача классификации текстов. Детали того, что именно понравилось или не понравилось пользователю не рассматриваются. Главная цель классификации мнений быстро определить эмоциональную направленность текста, дать общее впечатление, оценить преобладающее мнение об объекте.
2. Извлечение свойств (feature extraction). При данном подходе из текста выделяются отдельные фрагменты, относящиеся к объекту и его свойствам, затем вычисляется их семантическая ориентация.

Подход к задаче извлечения мнений как к задаче классификации текстов полезен во многих случаях, но часто его оказывается недостаточно. Автор отзыва может быть в целом доволен объектом, но критиковать отдельные его свойства, или наоборот. Точно так же, негативный отзыв еще не значит, что автору не нравится абсолютно все. Кроме того, не все тексты являются полностью оценочными или фокусируются на одном объекте. Например, в сообщениях в блогах необходимо сначала выделить оценочные предложения.

3. Анализ сравнительных конструкций. Прямое выражение положительного или отрицательного мнения это только одна из форм оценки. Сравнение объекта с другими тоже является оценкой, зачастую оно даже более убедительно для пользователя. Сравнительные конструкции отличаются по семантике и синтаксической структуре от обычных оценочных и требуют особого подхода.

3. Метод лексических шаблонов

Метод, предлагаемый в данной работе, относится ко второму типу задач извлечения мнений и ориентирован на извлечение свойств из текстов произвольного формата. Использовать лексические шаблоны в данном случает 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: Производится разметка частей речи и разбивка предложений на сегменты.

Определение: Сегментом будем называть предложение или фрагмент предложения, не содержащие противопоставлений.

Отсутствие противопоставлений внутри сегмента в большинстве случаев позволяет однозначно оценить его семантическую направленность.

На первом шаге в качестве сегментов будем выделять 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:

«Это плохие наушники»

Превращается в последовательность:
{это, **PRN}{**плохие, **ADJ}{**наушники, **NN}**

Шаг 2: Используются n-граммы чтобы разбить сегменты на более короткие. Триграмм обычно оказывается достаточно. 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, т. к. свойства разных продуктов разные, и замена гарантирует, что мы сможем найти общий для всех шаблон.

После замены указанная выше последовательность превращается в:
{это, **PRN}{**плохие, **ADJ}{$feature**, **NN}**

Результат сохраняется в базе, к нему также может применяться стемминг. Такая последовательность называется ***лексическим шаблоном***.

После этого можно использовать полученные последовательности для извлечения данных. Совпадения с лексическими шаблонами ищуются в каждом сегменте отзыва. Слово, совпадающее с меткой **$feature** считается свойством.

Возможны три ситуации:

- если сегмент удовлетворяет нескольким правилам, мы устанавливаем очередность правил, исходя из того, какие части речи чаще оказываются свойствами.
- если к сегменту не применимо ни одно из правил, из него извлекаются существительные, так, как если бы они были свойствами
- если сегмент состоит из одного слова, шаблоны не применяются, это слово сразу рассматривается как свойство.

4. Классификация оценочных конструкций

Тонально окрашенными будем называть такие элементы текста (синтагмы, фразы), которые несут в себе оценочную семантику.

С лингвистической точки зрения смысл текста (его субъективное содержание) характеризуется следующими группами факторов:

- лексико-грамматическими средствами, выражающими модальные характеристики ситуации, модусные смыслы и явное отношение автора к описываемой ситуации, в том числе выбор тонально окрашенного слова вместо нейтрального из синонимического ряда. Эти факторы можно назвать лексической компонентой, лексические компоненты разных областей, как правило, сильно отличаются.
- трансформации «нейтральной» структуры предложения, связанные с изменением порядка слов, осложнением, трансформацией залога, введением показателей смысловых отношений и других элементов. Эти факторы будем называть синтактической компонентой. Как правило, они независимы от предметной области.

Важными также являются понятия ***пропозиции,*** в том числе ***логической*** и ***событийной***, *и* ***семантических ролей,*** используемые в формальной семантике. ***Пропозицией*** называется обозначенное в речи действительное или возможное положение дел, семантическое ядро высказывания. Основой пропозиции является предикатное выражение, которое состоит из предиката, выражающего действия, состояния, свойства или отношения, и набора его аргументов, которые представляют собой знаки вещей, то есть имена, чьи свойства обозначаются.

Предикат ***событийной пропозиции*** связан со сферой бытия, движения, деятельности (физической или социальной); **логической пропозиции** — с отражением отношений, устанавливаемых в процессе мыслительной деятельности, логических рассуждений (отношения идентификации, тождества и т. п.).

Семантической ролью имени называется часть семантики предиката, отражающая общие свойства аргумента предиката — участника называемой предиктом ситуации. Описание в терминах семантических ролей отражает сходства моделей управления различных предикатных слов. Это понятие было предложено Чарльзом Филлмором, вначале использовавшим термин «глубинный падеж».

Теория семантических ролей слишком объемна для того, чтобы приводить ее здесь полностью. За подробным изложением этой теории можно обратиться к книге [Гриднева, 2009] или [Сусов, 2006].

В работе Ермакова и Киселева [Ермаков А. Е., Киселев С.Л, 2005] была предложена следующая классификация оценочных конструкций:

Тип 1. Явная тональная характеристика.

Объект или инициированное им событие наделяется признаком, имеющим оценочную семантику. Выделяются следующие роли:

- Объект оценки целевой объект;
- Атрибут существительное или именная группа, прилагательное, наречие, тонально окрашенный предикат-глагол.

Типовые пропозиции, которыми выражаются ситуации этого класса, курсивом выделены тонально окрашенные участники:

1. Логическая пропозиция полная: наушники *дрянь*, Сидоров *плохой руководитель*. Тональность выражается именной группой, образуемой существительным.

2. Логическая пропозиция свернутая с существительным: *гениальный авантюрист* Петров; *кристальная чистота* звука. Тональность выражается именной группой, образуемой существительным.

3. Логическая пропозиция свернутая с прилагательным: *запоминающийся* дизайн, *качественное* изображение, руководитель *нерешителен*. Тональность выражается прилагательным.

4. Свернутая логическая пропозиция в составе событийной, отражающая оценку события, в котором целевой объект выступает в роли протагониста: телефон *быстро сломался*, Иванов *бездумно согласился,* президент *принял авантюрное решение*. Тональность может выражаться наречием при глаголе, прилагательным при событийном существительном, самим глаголом.

Тип 2. Прямая эмоционально-коннотативная характеристика.

Класс эмоционально «заряженных» ситуаций, отражающих отношение целевого объекта к тонально окрашенным сущностям, их оценку целевым объектом, или наоборот отношение этих сущностей к объекту, оценку объекта ими. Выражается событийными пропозициями: президент борется с преступностью, народ выносит осуждение власти. Выделяются участники в следующих ролях:

- Субъект активный участник, в приведенных примерах «президент» и «народ»;
- Объект пассивный участник ситуации, в приведенных примерах «преступность» и «политика власти»;
- Предикат глагол или существительное, выражающее отношение Субъекта к Объекту (в приведенных примерах «бороться с» и «осуждение»).

Ниже перечислены типовые пропозиции, которыми выражаются ситуации этого класса, курсивом выделены тонально окрашенные участники.

1. Событийная пропозиция полная или свернутая, в которой роль субъекта занимает целевой объект: власть *борется с олигархами*; президент *ведет борьбу с коррупцией*, *борьба* президента *за права народа*. Тональность складывается из семантики именной группы в роли объекта и семантики предиката по принципу «положительное отношение к положительному позитив» и наоборот. Если эмоциональный коннотат объекта или предиката не определен, тональность считается нейтральной (президент встретился с олигархами, Иванов борется с сорняками на даче).

2. Событийная пропозиция полная или свернутая, в которой роль объекта занимает целевой объект: олигархи испугались президента, страна выражает

недоверие к власти, ненависть преступников к власти. В случае, если семантика субъекта имеет положительный эмоциональный коннотат, общая тональность складывается по тому же принципу, что и в (1). Если же семантика субъекта имеет отрицательный коннотат, то общая тональность не определена: олигархи полюбили президента.

Тип 3. Ассоциированный эмоциональный коннотат

Класс эмоционально заряженных ситуаций, фигурирующих в одном предложении с целевым объектом, но не связанных с ним напрямую (в ряде случаев эту связь просто не удается идентифицировать средствами автоматического анализа текста). Выражаются событийными пропозициями: обнищание пенсионеров, купил новую пару, пристраститься к пиву. Выделяются участники в следующих ролях:

- Участник участник, на состояние которого влияет событие;
- Предикат событие, которое влияет на Участника.

Тональность складывается из семантики именной группы в роли участник и семантики предиката по принципу «хорошо для хорошего --- позитив» и наоборот. Если эмоциональный коннотат Участника или Предиката не определен, тональность считается нейтральной.

Анализ типов 2 и 3 методом шаблонов сильно затруднен, так как их обработка требует глубокого синтаксического и семантического анализа. В дальнейшем в данной работе будут рассматриваться только оценки первого типа.

5. Определение границ мнения и преобразование полученных оценочных конструкций в шаблоны

На основании приведенной в предыдущем разделе информации были разработаны синтаксические шаблоны, используемые для выделения оценочных конструкций из текста.

Приведем примеры для каждого типа оценки:

Пример:

(1) *Логическая пропозиция полная:*
наушники — дрянь — {Nn-Nom, Nn-Nom}
Иванов — плохой руководитель — {Prop-Nom, Adj-Nom, Nn-Nom}

Пример:

(2) *Логическая пропозиция свернутая с существительным:*
кристальная чистота звука — {Nn-Nom, Adj-Nom, Nn-Gen}

Пример:

(3) *Логическая пропозиция свернутая с прилагательным:*
качественное изображение — {Adj-Nom, Nn-Nom}
руководитель нерешителен {Nn-Nom, Adj-Brf}

Пример:

(4) *Свернутая логическая пропозиция в составе событийной, отражающая оценку события, в котором целевой объект выступает в роли протагониста:*
телефон быстро сломался — {Nn-Nom, Adv, Verb}
Иванов бездумно согласился {Prop-Nom, Adv, Verb}

При обучении алгоритма находятся и учитываются все совпадения синтаксических шаблонов с сегментами, определенными как оценочные, затем полученный сегмент преобразуется в лексический шаблон. Как правило, это происходит с помощью замены существительного или глагола на метку атрибута.

Заметим, что применение этого метода требует предварительного морфологического анализа текста

6. Подготовка обучающего корпуса

Для обучения разработанного алгоритма необходим корпус, к котором каждое предложение размечено как оценочное или безоценочное. Определения тональной направленности документа в целом недостаточно, так как то, что текст в целом является оценочным, не означает, что оценочным является каждое его предложение. Субъективные тексты обычно содержат какое-то количество фактической информации.

Одним из основных препятствий для создания алгоритмов, подобных описанному, является недостаток обучающих данных. Для обучения алгоритма, классифицирующего тексты, легко составить коллекцию, используя, например, отзывы на соответствующих сайтах, многие из них помимо отзыва включают в себя оценочные баллы.

Гораздо тяжелее составить коллекцию отдельных предложений, которые могут быть однозначно определены как объективные или субъективные. Большая часть работ, посвященных алгоритмам, работающим на уровне предложений, использовали вручную размеченный корпус. Это требует больших временных и трудозатрат, поэтому объем подобных корпусов сравнительно мал.

Elen Riloff [Riloff&Wiebe, 2003] был предложен метод автоматизированной разметки, позволяющий создать корпус гораздо большего объема, чем любой из размеченных вручную, так как количество неразмеченных оценочных текстов, например, в Интернете огромно. Он предполагает использование контекстно-независимых оценочных слов для поиска оценочных предложений. Если предложение содержит такое слово, оно считается оценочным, в противном случае предложение считается нейтральным.

Заметим, что в предложении могут одновременно содержаться как контекстно-зависимые, так и контекстно-независимые оценочные слова, что позволяет, распознав оценочное предложение по контекстно-независимому слову, выделять шаблоны, специфичные для заданной предметной области.

7. Фильтрация полученных шаблонов

Отбор работающих шаблонов обычно предполагает участие человека, данный алгоритм позволяет его минимизировать при наличии достаточно большого обучающего корпуса.

Метод основывается на том наблюдении, что оценочные конструкции, как правило, встречаются повторно в других оценочных текстах. Если конструкция была выделена как оценочная ошибочно, скорее всего, она не встретится или будет встречаться реже настоящих оценочных.

Для каждого шаблона по следующей формуле вычисляется условная вероятность того, что предложение, содержащее этот шаблон, будет оценочным:

$$P(subjective/pattern\_i) = subjfrec(pattern\_i)/frec(pattern\_i),$$

где subjfreq(pattern) — количество раз, которые шаблон встретился в оценочных сегментах, freq(pattern) — количество всех использований шаблона.

Если вероятность выше порогового значения, шаблон принимается, если ниже отбрасывается. Пороговое значение выбирается экспериментально.

8. Схема работы алгоритма

Таким образом, работа алгоритма состоит из 6 этапов:

1. разметка частей речи
2. обобщение морфологических категорий
3. классификация сегментов
4. синтаксический анализ и выделение оценочных фрагментов
5. преобразование текстовых фрагментов в шаблоны
6. фильтрация шаблонов

Для разметки частей речи документ пропускается через морфологический парсер с параметрами сохранения пунктуации и выдачи грамматических категорий. После этого каждая лексема представляется в виде структуры <лексема, базовая форма, тип лексемы>, а сам документ разбивается на сегменты. Окончанием сегмента считается точка, многоточие, точка с запятой, восклицательный или вопросительный знак, или запятая за которой следует разделительный союз «но» или «зато».

Тип лексемы является обобщением выданной стеммером грамматической информации, так как в необработанном виде эта информация неоднородна,

избыточна и неудобна для использования. Для его выделения используется язык регулярных выражений, где каждое выражение представляет собой грамматический образ некоторого типа лексем.

Классификация лексем основана на диссертации И. Ножова [Ножов И., 2003].

Каждый сегмент классифицируется в соответствии с описанным в разделе 4 методом. Если сегмент содержит оценочное слово, он считается оценочным, если нет нейтральным. Список оценочных слов составлен вручную с помощью словарей синонимов и содержит около 500 слов, все они являются контекстно-независимыми.

После этого к полученным сегментам применяются синтаксические шаблоны, представляющие из себя последовательности типов лексем, сопоставляемых с типами лексем в тексте. Находятся и учитываются все совпадения. Экспериментальная оценка показала, что не имеет смысла использовать шаблоны, содержащие более четырех элементов.

Синтаксический анализ реализуется в этой части программы за счет шаблонов. Тип лексемы содержит грамматическую информацию, а строение шаблонов учитывает основные типы структур именных и глагольных групп.

Затем из каждого полученного сегмента в соответствии с типом синтаксического шаблона выделяется лексический шаблон оценочного выражения.

Далее по приведенной в предыдущей главе формуле вычисляется вероятность того, что шаблон является оценочным. Каждое слово шаблона сопоставляется с базовой формой встреченного в тексте слова, сохраняется согласование падежей внутри шаблона. Находятся все совпадения.

$$Pr(subjective|patter\_i) = subjfrec(pattern\_i)/freq(pattern\_i),$$

где subjfreq(pattern) количество совпадений шаблона с оценочными сегментами, freq(pattern) количество всех совпадений шаблона с текстом.

Если вероятность не превышает заданное пороговое значение, шаблон отбрасывается. Эксперименты показали, что для более формальных областей, например, бытовой техники оптимальное пороговое значение может превышать 0.9, для областей, предполагающих менее четкую терминологию (таких как, например, отзывы о фильмах) снижаться до 0.6.

Заметим, что для корректной работы фильтра корпус должен быть достаточно большим, не менее 100 000 предложений.

Пример работы алгоритма:

Исходный текст (комментарий в интернет-магазине, пунктуация сохранена):

«Купил вчера эти наушники. Звук — сказка хирургическая точность. Сочные обволакивающие басы, отличные верха и громкость. Честно непередаваемое ощущение полета.

Один недостаток — конструкция немного хрупковатая.»

Этапы 1–3: Морфологическая разметка и сегментирование текста (жирным шрифтом выделены сегменты, отмеченные как оценочные):

[< Купил купить Verb-Fin > < вчера вчера Adv > < эти этот Det-Nom > < наушники наушник Nn-Nom >]

[< Звук звук Nn-Nom > < сказка сказка Nn-Nom > < хирургическая хирургический Adj-Nom > < точность точность Nn-Nom >]

[< Сочные сочный Adj-Nom > < обволакивающие обволакивать Adj-Nom > < басы бас Nn-Nom >

< отличные отличный Adj-Nom > < верха верх Nn-Nom >

< и и Conj > < громкость громкость Nn-Nom >]

[< Честно честно Adj-Brf > < непередаваемое непередаваемый Adj-Nom > < ощущение ощущение Nn-Nom > < полета полет Nn-Gen >]

[< Один один Det-Nom > < недостаток недостаток Nn-Nom >]

[< конструкция конструкция Nn-Nom > < немного немного Adv >

< хрупковатая хрупковатый Adj-Nom >]

Этапы 4–5: Выделение оценочных фрагментов с помощью синтаксических шаблонов, преобразование их в лексические шаблоны:

[Adj-Nom, Nn-Nom, Conj, Nn-Nom] -> отличные верха и громкость -> **отличные $feature**

[Adj-Nom, Adj-Nom, Nn-Nom] -> сочные обволакивающие басы -> **сочные обволакивающие $feature**

[Nn-Nom, Nn-Nom] -> звук сказка -> **$feature сказка**

[Nn-Nom, Adj-Nom] -> сказка хирургическая -> **$feature хирургическая**

[Adj-Nom, Nт-Nom] -> хирургическая точность -> **хирургическая $feature**

[Adj-Nom, Nn-Nom, Nn-Gen] -> честно непередаваемое ощущение -> **честно непередаваемое $feature**

[Adj-Nom, Nn-Nom, Nn-Gen] -> непередаваемое ощущение полета -> **непередаваемое $feature**

[Nn-Nom, Adv, Adj-Nom] -> конструкция немного хрупковатая -> **$feature немного хрупковатая**

Этап 6: Фильтрация шаблонов

оставлены:

отличные $feature

сочные обволакивающие $feature

непередаваемое $feature

$feature немного хрупковатая

отфильтрованы:

честно непередаваемое $feature

$feature хирургическая

хирургическая $feature

9. Оценка результата

В настоящее время методов объективного тестирования систем тональной разметки текстов еще не разработано. Поэтому применяемый нами метод тестирования основывается на субъективных оценках небольших текстовых подборок экспертом. В качестве тестовой коллекции был собран корпус объемом приблизительно 300 000 предложений, состоящий из трёх частей — отзывы о фотокамерах, отзывы о наушниках и рецензии зрителей на фильмы.

Эксперт получает тональную разметку текстов при помощи системы и затем оценивает, насколько он в каждом конкретном случае согласен или не согласен с результатом. В случае несогласия эксперт отмечает, что нетональная конструкция была размечена как тональная. Затем на основании этих оценок вычисляется точность тональной разметки по формуле:

$$P= N\_subj/N\_all,$$

где N_subj количество верных шаблонов N_all количество всех выделенных шаблонов.

При оценке мы не учитываем полноту, т. к. из-за самой конструкции алгоритма она является невысокой (прежде всего из-за использования автоматической фильтрации), что может быть компенсировано размером обучающей коллекции благодаря автоматической разметке.

Примеры извлеченных конструкций :

(5) *отзывы о фотоаппаратах:*
просто отличная $feature
плох стандартный $feature
очень удобное $feature
довольно живучий $feature
$feature очень быстрая
$feature ИМХО дорогие
$feature недостаточно светосильный
хороший резкий $feature
прочный удобный $feature
удобная быстрая $feature
дополнительный монохромный $feature
яркий большой $feature
неравномерное искусственное $feature
прекрасная цветопередача $feature
удачная реализация $feature
избыточное число $feature
хорошая система $feature
низкий уровень $feature

(6) ***рецензии на кинофильмы:***
непередаваемо радостную $feature
слишком поверхностные $feature
очень жесткий $feature
неприятный $feature
очень актуальный $feature
весьма странный $feature
страшный $feature
довольно тяжелый $feature
$feature очень тяжелый
$feature очень личный
$feature немного нудный
больное воображение $feature
$feature полное фуфло
великолепная игра $feature
$feature гениальный
$feature клевый
$feature хороший
$feature отвратительный
смешное такое $feature

(7) ***отзывы о наушниках:***
кристально чистый $feature
более объемный $feature
очень интересная $feature
очень удобный $feature
$feature немного хрупковатая
$feature вполне приличный
$feature весьма ничего
$feature просто супер
$feature правда дороговатые
$feature ничего особенного
профессиональная звуковая $feature
пластиковый прочный $feature

Точность выдачи алгоритма оказалась зависящей от предметной области.

Для областей, подразумевающих большое количество объективных критериев и относительно невысокую вариативность лексики, таких как отзывы о наушниках и фотокамерах, алгоритм показал очень хорошие результаты до 52% до фильтрации и 80% после фильтрации с пороговым значением вероятности сохранения шаблона $\theta=0.9$.

Для рецензий на фильмы области с большей вариативностью лексики и небольшим количеством объективных критериев точность была гораздо ниже 29% до фильтрации и 64,3% после фильтрации с пороговым значением вероятности сохранения шаблона $\theta=0.6$.

Точность можно повысить, увеличивая для алгоритма фильтрации пороговое значение вероятности сохранения шаблона, но это сильно снижает полноту, особенно для плохо формализованных областей.

Также можно выделить два класса ошибок, возникающих при определении субъективности:

1. Ошибки работы модуля морфологической разметки текста.
2. Ошибки, связанные со сходством фактических конструкций с оценочными.

10. Заключение

Как показывают результаты, предложенный метод достигает достаточно высокой (80%) точности на текстах определенной тематики и может быть в дальнейшем использован в качестве компонента системы извлечения мнений. Точность так же может быть повышена с помощью улучшения качества разметки обучающего корпуса и использованиея более глубокого синтаксического анализа, что является направлением дальнейшей работы.

References

1. *Carenini G., Pauls A.* (2006), Multi-Document Summarization of Evaluative Text in Proceedings of EACL 2006, Trento, Italy, 2006, pp. 305–312.
2. *Ermakov A. E., Kiselev S. L.* Linguistic Model For Computational Sentiment Analysis of Media [Lingvisticheskaya model dla komputernogo analiza tonalnosti publikacii v SMI]. Kompjuternaia Lingvistika i Intellektualnye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii Dialog 2005 [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialog 2005, Moskva, 2005.
3. *Hatzivassiloglou V., McKeown K.* (1997), Predicting the semantic orientation of adjectives in Proceedings of ACL/EACL 1997, Madrid, Spain, Complutense University of Madrid, 2007, pp. 174–181.
4. *Hu M. and Liu B.* (2004), Mining and Summarizing Customer Reviews in Proceedings of KDD-2004, Seattle, WA, 2004, pp. 168–177,
5. *Hu M. and Liu B.* (2004), Mining Opinion features in Customer Reviews in Proceedings of AAAI'04, Boston, Massachusetts, USA: AAAI Press, 2004, pp. 755–760.
6. *Kobayashi N., Inui K., Tateishi K., Fukushima T.* (2004), Collecting Evaluative Expressions for Opinion Extraction in Proceedings of IJCNLP-2004, Berlin, Germany: Springer, 2004, pp. 596–605.
7. *Liu, B.* (2007), Web Data Mining, Springer, Berlin.
8. *Liu, B.* (2010), Sentiment Analysis and Subjectivity in Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC 2010, NY, USA, pp. 257–282.

9. *Nozhov I.* (2003) Morphologic and Syntactic Text Processing (Models and Computations), available at http://www.aot.ru/docs/Nozhov/msot.pdf
10. *Popescu A. M., Etzioni O.* (2005), Product features and Opinions from Reviews in Proceedings of HLT-EMNLP 2005. Vancouver, Canada: ACL, 2005, pp. 339–346.
11. *Riloff E, Wiebe J.* (2003) Learning Extraction Patterns for Subjective Expressions, in Proceedings of EMNLP-03, Sapporo, Japan, 2003, pp. 97–104.
12. *Sibiriakov A.* Extracting opinions about products from blogs and forums according to sentiment [Izvlecheniye mneniy o tovarah iz forumov i blogov s uchetom tonalnosti], available at http://elar.usu.ru/bitstream/1234.56789/2064/1/RuSSIR_2008_07.pdf
13. *Turney P.* (2003), Inference of Semantic Orientation from Association, available at http://cogprints.org/3164/01/turney-littman-acm.pdf
14. *Turney P.* (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews in Proceedings of ACL 2002, Philadelphia, PA, U.S.A., 2002, pp. 417–424.

Русскоязычные статьи:

1. *Сибиряков А.*, Извлечение мнений о товарах из форумов и блогов с учетом тональности [PDF] (http://elar.usu.ru/bitstream/1234.56789/2064/1/RuSSIR_2008_07.pdf).
2. *Гриднева Н. Н.*, Основные семантики синтаксиса. Санкт-Петербург, Издательство СпбГУЭФ, 2009. 48 с.
3. *Ермаков А. Е., Киселев С. Л.*, Лингвистическая модель для компьютерного анализа тональности публикаций СМИ Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог 2005. Москва, Наука, 2005
4. *Ножов И.*, Морфологическая и синтаксическая обработка текста (модели и программы), 2003, [PDF] (http://www.aot.ru/docs/Nozhov/msot.pdf).
5. *Сусов И. П.*, Введение в языкознание. Москва, Восток — Запад, 2006. 382 с.