

ФОРМАЛИЗАЦИЯ ПРОЦЕССА АСПЕКТНО-ОРИЕНТИРОВАННОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ НА МАТЕРИАЛЕ ОЦЕНОЧНЫХ ОТЗЫВОВ НА ПРЕДПРИЯТИЯ ГОСТИНИЧНОГО ТИПА

Анализ тональности (или *Sentiment analysis*) является новым направлением компьютерной лингвистики. Данное направление решает целый ряд задач, связанных с определением отношения субъекта высказывания к некоему объекту действительности по шкале *позитивное-негативное*. В зависимости от типа задачи оценочная шкала может быть более детальной. Отношение субъекта носит оценочный характер и выражается в лингвистических и экстралингвистических средствах. Определить тональность высказывания — значит определить данное оценочное отношение.

Процесс определения тональности не имеет единого общепринятого метода решения из-за многозадачности данной области, большого количества объектов действительности и трудности процесса формализации. Задачи определения тональности можно условно разделить на объектно-ориентированные и аспектно-ориентированные. Под первыми понимается определение общего оценочного отношения к одному субъекту, под вторыми — определение оценочного отношения к набору характеристик одного субъекта. Рассмотрим данную разницу на примере следующего фрагмента оценочного отзыва:

*The **hotel** was absolutely horrible. First, it is **located** in a noisy part of town. Second, the **room** is spacious but badly cleaned. Third, coffee is disgusting...*

В данном фрагменте в качестве объекта тональности выступает отель. Слова, выделенные жирным шрифтом, относятся к его характеристикам. Данные характеристики называются аспектами. Подчеркнутые слова выражают тональность. В процессе выполнения объектно-ориентированного анализа аспекты, как правило, игнорируются и происходит подсчет оценочных выражений. В нашем случае это четыре негативных (*horrible*, *badly cleaned*, *disgusting*, *noisy*) и одно позитивное слово (*spacious*). Таким образом, общая тональность будет негативной. В случае аспектно-ориентированного анализа оценочная информация будет представлена пользователю вместе с аспектными словами.

Системы определения тональности можно разделить на однодоменные и мультидоменные. Однодоменные системы определяют тональность объектов одной предметной области, а мультидоменные не зависят от предметной области. Как правило, аспектно-ориентированный анализ проводится в однодоменных системах, а объектно-ориентированный в мультидоменных. Это связано с спецификой некоторых доменов. Например,

слово *terrifying* будет негативным для мобильного телефона, но позитивным в отзыве на фильм ужасов.

Предлагаемая система определения тональности является аспектно-ориентированной и однодоменной. Объектом тональности выступают предприятия гостиничного типа. Субъектом тональности — авторы оценочных отзывов. Оценочный интернет-отзыв представляет собой оформленный тип текста, содержащий авторскую оценку о некоем явлении действительности, размещенный в сети Интернет с целью обмена информацией между пользователями. Материалом для исследования послужили пятьдесят оценочных отзывов с сайтов *Tripadvisor.com* и *Booking.com*. Причиной выбора этих сайтов послужила политика администрации, направленная на предотвращение публикаций рекламного характера. Например, на сайте *Booking.com* можно оставить отзыв об отеле только после оплаты пребывания в нем. Все отзывы подверглись статистической обработке, в результате которой был получен алфавитный словарь. Словарь содержит 3031 словоформу. К каждой словоформе приписан семантический подкласс и частеречная принадлежность. Под семантическим подклассом понимается отнесенность слова к некой области действительности. Например, словарная запись *Dubai_N10* значит, что слово *Dubai* является существительным (N) с семантическим подклассом 10 (географические названия).

В состав алфавитного словаря входят *аспектные слова*. Под аспектными словами понимаются те слова, которые называют определенный аспект объекта тональности. Всего было выделено семь семантических аспектных подклассов существительных: N1 (компонент номера: *bed, room, view...*), N2 (компонент обслуживания: *staff, waiter, service...*), N3 (компонент питания: *breakfast, coffee...*), N4 (компонент месторасположения: *location, beach...*), N5 (компонент цены: *price, expenses...*), N6 (компонент инфраструктуры: *pool, Jacuzzi...*), N7 (архитектурные элементы отеля: *corridor, reception...*). Также был выделен один аспектный подкласс глагола, обозначающий расположение отеля (*issituated, waslocated...*).

Таким образом, в алфавитном словаре содержится информация для последующего синтаксического анализа и идентификации аспектов. Для определения тональности используется тональный словарь. Тональный словарь состоит из 617 лексических единиц. Каждой единице приписана частеречная принадлежность и положительная (P) или отрицательная (N) тональность. Например, *beautiful_A(P)*, *disgusting_A(N)* и т.д. В состав тонального словаря включены устойчивые выражения (*thumbsup_(P)*) и оценочные словосочетания, характерные для данной предметной области (*memorablestay_(P)*). К определенным лексическим единицам приписаны правила, уточняющие тональность и аспект. Например, *если после слова old следует существительное с аспектным семантическим подклассом N3 (компонент питания), то слово обладает негативной тональностью*. Так же

в состав словаря входят слова-инвенторы (т.е. слова, меняющие тональность на противоположную) и слова-интенсификаторы (т.е. слова, которые усиливают тональность). К первым относятся всевозможные отрицания, ко вторым, в основном, наречия (*absolutely, really, generally...*).

Таким образом, два словаря формируют основу базы данных для дальнейшего процесса определения тональности. В. Хатзивассилоглуу называет *субъективность* обязательным критерием наличия оценочной информации. К субъективности он относит «всю информацию, которая относится к мнению и оценке» и противопоставляет ее объективной информации, которая «сообщает факты об объектах действительности» [1, с. 299]. Очевидно, что оценочная информация содержится в семантике слов, поэтому для определения тональности требуется определить в тексте субъективную информацию путем выделения оценочных слов. Проведя анализ отобранных 50 отзывов, можно утверждать, что позиционно тональные и аспектные слова условно распределяются следующим образом:

- 1) в составе конструкции A+N (*bad room, delicious breakfast*);
- 2) аспектное слово находится в группе подлежащего, а тональное — в группе сказуемого. (*The location is very convenient*);
- 3) тональность выражается за счет семантики целой фразы (*We will most likely stay in this hotel*).

Для формализации процесса представления тональности в тексте компьютерная система для начала отбирает только те предложения, в которых есть тональные слова, т.е. там, где присутствует субъективность. Далее каждому слову в предложении приписывается частеречная принадлежность и семантический подкласс, согласно информации в алфавитном словаре. Затем, используя информацию из тонального словаря, словам приписывается тональность, определяются инвенторы и интенсификаторы. Например, первое предложение из приведенного выше примера будет выглядеть так:

The (L2) hotel (N7) was (V1) absolutely (D2_Int) horrible (A3_N).

Далее идет сверка с позиционным шаблоном A+N. Если он отсутствует в предложении, либо в предложении еще есть тональные слова, то происходит синтаксический анализ предложения для выделения групп подлежащего и сказуемого. Синтаксический анализ проводится на основе списка граничных сигналов. В результате синтаксического анализа конструкция *the hotel* будет выделена в группу подлежащего, а *was absolutely horrible* в группу сказуемого. Аспектное слово (*hotel*) присутствует в группе подлежащего, а тональное слово (*horrible*) с интенсификатором (*absolutely*) — в группе сказуемого. Аспектное и тональное слово с интенсификатором будут извлечены для представления пользователю.

Формальное представление оценочной информации, которая выражается за счет семантики всего высказывания, не является окончательно решенной задачей. Трудность формализации такого типа оценочной

информации заключается в отсутствии формальных признаков наличия оценки. Позиционный и статистический анализ таких высказываний в составе оценочных отзывов показал, что их набор ограничен и они фактически всегда находятся в составе последнего предложения или являются им. В составе рассматриваемой системы определения тональности были выделены лингвистические признаки таких высказываний, которые можно использовать для извлечения. Например, это словосочетания и конструкции *would recommend, would stay If +N +was/were...* Если в последнем предложении отзыва встречаются подобные конструкции или словосочетания, то предложение извлекается целиком.

Описанная система позволяет определить тональность оценочных отзывов на предприятия гостиничного типа.

ЛИТЕРАТУРА

1. Hatzivassiloglou, V. Effects of adjective orientation and gradability on sentence subjectivity / V.Hatzivassiloglou, J. Wiebe // Proceedings of the 18th conference on Computational linguistics, 2000. — P. 299–305.