А.Л. Бочкова (Минск, МГЛУ)

ЛИНГВИСТИЧЕСКАЯ БАЗА ДАННЫХ КАК ОСНОВА СИСТЕМЫ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ МНЕНИЙ УЧАСТНИКОВ ИНТЕРНЕТ-КОММУНИКАЦИИ

В настоящее время проблема извлечения и резюмирования мнений становится все более востребованной в связи с быстрым развитием интернет-коммуникации. Глобальная сеть Интернет располагает различными сервисами, позволяющими пользователям обмениваться мнениями об объектах, явлениях, процессах и событиях. К числу наиболее популярных сервисов относятся чаты, форумы, блоги, гостевые книги, социальные сети. С другой стороны, у определенных лиц и организаций все чаще возникает необходимость выявить в огромном потоке мнений ту информацию, которая будет наиболее полезной. Понимание смысла, как отдельного предложения, так и всего текста, подразумевает не только определение его информационного содержания, но и выделение прагматической направленности, выраженной категорией оценочности. Оценочно-окрашенными принято считать такие элементы текста, которые несут в себе оценочную семантику: позитивную, нейтральную или негативную. Во мнениях участников Интернет-коммуникации всегда присутствует оценочный аспект, определяющий выбор конкретных лексических средств.

Методы выделения оценки из текста можно разделить на группы в зависимости от типа исходной информации и поставленных целей. В работе [1] представлены следующие три основных метода:

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 впечат-ления, оценка преобладающего мнение об объекте.

2. Извлечение свойств. При использовании данного подхода 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 во многих случаях полезен, но часто оказывается недостаточным. Автор отзыва может быть в целом доволен объектом, но критикует отдельные его свойства. 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.

3. Анализ сравнительных конструкций. Прямое выражение положительного или отрицательного мнения — это только одна из форм оценки. Сравнение объекта с другими объектами тоже является оценкой, зачастую оно даже более убедительно для пользователя. По семантике и

синтаксической структуре сравнительные конструкции отличаются от обычных оценочных высказываний.

В рамках первого метода выделяют два основных подхода к классификации мнений об объекте: лексический и машинный [2]. 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 [3]. Лексический подход предполагает индивидуальный анализ слов и/или фраз с использованием словарей тональной / эмоциональной лексики. В зависимости от направления тональности слову из текста приписывается определенный вес, веса всех слов суммируются, и выводится общий вес текста [4]. При разработке системы автоматического извлечения мнений участников Интернет-коммуникации автором статьи был выбран лексический подход. Таким образом, лингвистическая база данных системы представляет собой совокупность нескольких словарей.

Материалом для исследования послужили 180 текстов отзывов о фильме, взятых с сайта http://movies.yahoo.com. Такие тексты являются наиболее сложными для анализа с точки зрения эмоциональности, поскольку в одном отзыве его автор может высказываться как положительно, так и отрицательно. Зачастую он дает неоднородную оценку фильму, указывая его сильные и слабые стороны, критикуя, и, при этом, отмечая, что именно ему понравилось в данном фильме.

В ходе анализа текстов отзывов были составлены два словаря тональной лексики: словарь чисто тональной лексики и словарь контекстно-зависимых слов. Словарь контекстно-зависимых слов необходим, поскольку одно и то же слово с разным контекстом (правым или левым) будет иметь разное направление тональности. Например, прилагательное *huge* в сочетании *huge success* будет положительным, в то время как в сочетании *huge failure* — отрицательным. В табл.1 представлен фрагмент этого словаря.

Таблица 1

Фрагмент словаря контекстно-зависимых слов

№ п/п	Лексическая единица	Пример
1.	*little*	*little experience*
2.	*high*	*high performance; high price*
3.	*new*	*new year; new ideas*
4.	*huge*	*huge success; huge man*
5.	*big*	*big fan; big problem*
6.	……..	……..

Словарь тональной лексики является словарем словоформ и состоит из двух частей: положительных тональных слов и отрицательных тональных слов. В качестве шкалы оценки было выбрана шкала от +3 до –3, т.е. [+3, +2,

+1; –1, –2, –3]. Каждое слово было размечено вручную. В табл. 2 и табл. 3 представлены фрагменты двух частей словаря.

Таблица 2

Фрагмент словаря положительных тональных слов

№ п/п	Положительное тональное слово	Вес
1.	*good*	1
2.	*great*	3
3.	*like*	1
4.	*enjoy/enjoyed/enjoying*	1
5.	*love/loved/loving*	2
6.	*perfect*	3
7.	*beautiful*	1
8.	*interesting*	1
9.	*favourite*	1
10.	*wonderful*	2
11.	*stunning*	2
12.	*entertaining*	1
13.	……	

Таблица 3

Фрагмент словаря отрицательных тональных слов

№ п/п	Отрицательное тональное слово	Вес
1.	*bad*	–1
2.	*wrong*	–1
3.	*boring*	–2
4.	*terrible*	–3
5.	*disappointing/disappointment/disappointed*	–1
6.	*ruin/ruined*	–2
7.	*hate/hated*	–3
8.	*flat*	–1
9.	*awfull*	–3
10.	*ridiculous*	–1
11.	……	

Для извлечения из текста оценочных мнений также необходим словарь интенсификаторов, т.е. слов, увеличивающих или уменьша-ющих вес тонального слова. Например, *very, really, rather, absolutely* и т.д. Интенсификаторы ранжируются по следующей шкале: высокий, средний и низкий. Фрагмент словаря интенсификаторов представлен в табл.4.

Таблица 4

Фрагмент словаря интенсификаторов

Высокий интенсификатор	Средний интенсификатор	Низкий интенсификатор
very	*actually*	*low*
amazingly	*frankly*	*fairly*
awfully	*honestly*	*merely*
really	*reasonably*	*quite*
certainly	*pretty*	*relatively*
……	……	……

Помимо словарей тональной лексики и интенсификаторов лингвистическая база данных содержит список слов, инвертирующих вес тонального слова, типа *not, never, no, without.* Например, *good +1 — not good -1; pleasure +1 — without pleasure -1; disappoints -1 — never disappoints +1.* Не лишним будет включить в лингвистическую базу данных слова, являющиеся ключевыми для выбранной предметной области (рецензии или отзывы о фильмах). Это позволит избежать анализа той лексики, которая по своей сути является тональной, но не имеет отношения к фильмам. Например, *acting, scene (scenes), soundtrack, character (characters), movie (movies), cast, plot, actor (actors), actress (actresses), performance, film, interpretation, adaptation.* В случае, если в одном отзыве встретятся словосочетания *good weather* и *good film* (*The weather was good, so I decided to go to the cinema and watch a good film),* оценочное прилагательное *good* в контексте погоды не повлияет на общую оценку фильма.

ЛИТЕРАТУРА

1. Liu, B. Sentiment Analysis and Subjectivity in Handbook of Natural Language Processing. – Second Edition. – NY, Chapman and Hall, 2010. – P. 257–282.

2. Blinov, P.D. Research of Lexical Approach and Machine Learning Methods for Sentiment Analysis / P.D. Blinov, M.V. Klekovkina, E.V. Kotelnikov, O.A. Pestov // Компьютерная лингвистика и компью-терные технологии [Электронный ресурс]. – Режим доступа http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BlinovPD.pdf. – Дата доступа: 28.02.2014.

3. Agarwal, A. Sentiment Analysis of Twitter Data / Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R // Proceedings of the Workshop on Language in Social Media. – LSM 2011. – P. 30–38.

4. Ding, X. A Holistic Lexicon Based Approach to Opinion Mining / Ding X., Liu B., Yu P.S. // Proceedings of the Conference on Web Search and Web Data Mining. – WSDM 2008. – P. 231–240.