Е.Г. Брунова
Тюменский государственный университет
Кафедра иностранных языков
и межкультурной профессиональной
коммуникации ИМЕНИТ
Заведующий кафедрой
Доктор филологических наук
egbrunova@mail.ru

МЕТОДИКА СОСТАВЛЕНИЯ ОЦЕНОЧНОГО ЛЕКСИКОНА ДЛЯ КОНТЕНТ-АНАЛИЗА МНЕНИЙ
THE METHODOLOGY OF BUILDING A LEXICON FOR SENTIMENT ANALYSIS

Контент-анализ субъективности, в том числе – мнений пользователей о качестве товаров и услуг, является одной из наиболее перспективных задач в области автоматической обработки текста (Natural Language Processing). Большинство работ в данной области появились за последние 10-12 лет, что связано с развитием социальных сетей, блог-платформ и других технологий, результатом которого стало появление большого объема пользовательского текста (user-generated text). Величина объема пользовательского текста позволяет применить разнообразные статистические методы с достаточно достоверными результатами.

Контент-анализ субъективности необходим для успешного функционирования вопросно-ответных систем и систем извлечения информации, где требуется отличать факты от мнений. Кроме того, производители товаров и поставщики услуг заинтересованы в получении обработанной информации о настроениях потребителей. Потребителей, в свою очередь, при выборе товара или услуги, интересуют мнения других людей, основанные на их личном опыте.

Целью данного исследования является описание методики построения оценочного лексикона для контент-анализа мнений.

В англоязычных источниках для обозначения подобных видов анализа используются термины *sentiment analysis, opinion mining*. Для русскоязычных источников характерны разнообразные переводы английских терминов –*анализ эмоциональной окраски текста* (Ермаков, Ермакова, 2012: 85), *анализ тональности текста* (Пазельская, Соловьев, 2011), вплоть до полного заимствования из английского языка – *сентимент-анализ* (Ермаков, Ермакова, 2012: 85). По нашему мнению, ни один из предлагаемых переводов устоявшихся английских терминов не представляется удачным. Англ. *sentiment* означает, не только «чувство» (а именно такие ассоциации связаны с рус. *сентимент*), но и «мнение», и в термине *sentiment analysis* реализуется именно это второе значение. Термин *тональность* пришел из музыки и фонетики, и его метафорическое применение по отношению к лексике является малоинформативным. Такая же метафоричность характерна и для *эмоциональной окраски текста.* Термин *классификация отзывов* (Четверкин, Лукашевич, 2011: 73) представляется слишком общим для наших задач. Поэтому мы предлагаем собственный термин *контент-анализ мнений (КАМ).* Под *контент-анализом мнений (КАМ)* подразумевается группа методов изучения субъективности в естественном языке путем извлечения из текста мнений и эмоций, а также их последующей обработки.

КАМ принадлежит к семейству методов классификации корпуса текстов. При этом корпус текстов чаще всего разбивается на два класса: с положительной (хорошо, нравится) и отрицательной (плохо, не нравится) оценкой (Hu, Liu, 2004; Nasukawa, Yi, 2003; Pang et al., 2002: Turney, 2002) Некоторые исследователи добавляют третий класс – с нейтральной оценкой (Четверкин, Лукашевич, 2011).

Основным инструментом для такой классификации является оценочный лексикон, под которым понимается множество слов естественного языка, в семантике которых содержится оценка. Такое множество разбивается, как

правило, на два подмножества (положительный лексикон и отрицательный лексикон), которые впоследствии используются для классификации текстов.

Грамотно составленный оценочный лексикон имеет особую значимость для методов КАМ, рассматривающих текст как набор слов (bag of words), не требующих синтаксического разбора и устраняющих избыточность, характерную для пользовательских текстов (Ермаков, Ермакова, 2012: 87). К таким методам относится, в частности, наивный Байесовский классификатор, который показывает весьма неплохие результаты, несмотря на предельно упрощенное представление о тексте (Webb, Boughton, Wang, 2005).

В качестве примера создания оценочного лексикона можно упомянуть исследование М. Ху и Б.Лью (Hu, Liu, 2004). Лексикон, созданный данными авторами, представляет собой два множества слов английского языка, которые используются для оценки свойств товаров, продаваемых онлайн. При этом такой товар представляется как объект КАМ (opinion target), т.е. сущность, по поводу которой высказывается мнение. Такая сущность, в свою очередь, представляется как набор частей или свойств и набор атрибутов, а каждая часть или свойство также может быть представлена как набор атрибутов. Таким образом, объект КАМ представляется как иерархическая структура. Например, у объекта КАМ *digital camera* (цифровой фотоаппарат) выделяются такие атрибуты как *picture quality* (качество изображения) и *size* (размер), мнение по поводу которых потребители выражают с помощью оценочного лексикона, например, «качество изображения – хорошее». Значительную часть лексикона составляют имена прилагательные, выбранные из пользовательского текста, а также их синонимы и антонимы. Однако при выражении своего мнения пользователи не склонны ограничиваться прилагательными, например, при оценке работоспособности пользователь может использовать глагол – «не работает».

На примере исследования М. Ху и Б.Лью можно выделить основные проблемы при составлении оценочного лексикона:

1) большой объем ручной работы: базовый лексикон (seed lexicon),

который впоследствии может быть расширен автоматизированными средствами, выделяется вручную из пользовательского текста.

2) богатство естественного языка и разнообразие языковых средств оценки: исследователю всегда приходится иметь в виду, что пользователь может использовать редкое, нестандартное средство, не отраженное в оценочном лексиконе.

3) необходимость учета орфографических ошибок пользователей: лексикон М. Ху и Б.Лью включает слова с орфографическими ошибками, извлеченные из пользовательского текста.

По поводу большого объема ручного труда было бы естественным ожидать, что, однажды проделав такую работу, ее можно было бы впоследствии использовать для новых исследований. Более того, лексикон, составленный на одном языке, можно было бы перевести на другие языки. Насколько оправданны такие ожидания? К сожалению, создание универсального оценочного лексикона представляется малорезультативным, поскольку он был бы не только заведомо неполным, но и противоречивым. За исключением весьма ограниченного набора оценочных прилагательных (*хороший – плохой* и т.п.), применение одних и тех же лексем для разных предметных областей может давать противоположные результаты. Например, слово «долго» входит в положительный лексикон для оценки парфюмерии («аромат долго держится»), и в отрицательный лексикон для оценки банковской деятельности («операционистка долго обслуживает»). Иногда такое противоречие может встретиться даже в пределах одной предметной области при оценке разных атрибутов, например, для предметной области «цифровой фотоаппарат»: «батарея долго служит» (положительная оценка) и «фокус фотоаппарата долго устанавливать» (отрицательная оценка).

Единицы лексикона, имеющие противоречивый характер, как правило, относятся к лексическим усилителям, в частности, – к «параметрическим определениям с усилительным значением» [*Стилистический энциклопедический словарь русского языка*, 2003: 550], например, *самый,*

такой, абсолютно, *большой, маленький, долгий, короткий, высокий, низкий* и т.п. Н.В.Лукашевич и И.И.Четверкин предлагают выделять параметрическую лексику в качестве операторов, влияющих на степень оценки (Лукашевич, Четверкин, 2011: 77), однако в их статье под операторами понимаются отрицательные частицы и лексические усилители прилагательных (*не, нет, полный, очень, самый* и т.п.), а не собственно прилагательные или наречия. Мы полагаем, что лексические усилители могут включать также прилагательные и наречия, выражающие параметрическое значение того или иного атрибута предметной области, например, *скорость, цена, устойчивость* и т.п. В зависимости от атрибута большие или малые параметры получают положительную или отрицательную оценку, так, *большой* применительно к скорости, надежности или устойчивости означает положительную оценку, а по отношению к цене или затраченному времени – отрицательную. Именно атрибуты определяют зависимость данных единиц лексикона от предметной области. Решением в данном случае может служить включение в качестве единиц лексикона биграмм или триграмм (двух или трех соседних слов), в которые должно входить наименование самого атрибута (Pang, Lee, Vaithyanathan, 2002). В качестве альтернативы можно предложить выделение отдельных классов для лексических усилителей и для атрибутов, чувствительным к таковым. Для таких классов можно было бы сформулировать специфические правила классификации и продвинуться на пути универсализации лексикона.

Сложность и трудоемкость построения оценочного лексикона не означает, что этот процесс нельзя оптимизировать. Процесс оптимизации может быть организован в 4 этапа:

Первый этап: вручную составляется базовый лексикон (seed lexicon), содержащий относительно небольшое число единиц. Основной единицей лексикона в англоязычных разработках является лемма (слово в своей основной форме). Применительно к русскому языку представляется целесообразным использовать усеченное слово, т.е. основу без окончаний

(стемминг), например, *хорош* вместо *хороший* и *хорошо*.

Второй этап: базовый лексикон расширяется за счет синонимов и антонимов, например, *хорош(ий)→превосходн(ый), отличн(ый), замечательн(ый), плох(ой), нехорош(ий)* и т.д. Для этой цели используются словари синонимов и антонимов.

Третий этап: базовый лексикон расширяется с помощью поисковых запросов с булевыми операторами.

Четвертый этап: лексикон пополняется и уточняется в процессе эксплуатации системы.

Методика расширения лексикона с помощью поисковых запросов была предложена В. Хацивасилоглу и К. МакКьюном (Hatzivassiloglou, McKewn, 1997). В ее основе лежит предположение о том, что прилагательные или наречия, соединенные союзом *И* имеют одинаковую полярность, а прилагательные или наречия, соединенные союзом *НО* имеют противоположную полярность. Рассмотрим реальный пример:

Допустимые высказывания:

(1) Ипотечный кредит в Сбербанке: все быстро **и** грамотно.

(2) Ипотечный кредит в Сбербанке: все грамотно, **но** долго.

Недопустимые высказывания:

(3) * Ипотечный кредит в Сбербанке: все быстро, **но** грамотно.

(4) * Ипотечный кредит в Сбербанке: все долго **и** грамотно.

Наречия *быстро* и *грамотно* имеют одинаковую (положительную) оценку, в то время как у наречий *долго* и *грамотно* оценка противоположная (отрицательная и положительная соответственно). Следовательно, имея сравнительно небольшой объем прилагательных и наречий с известной оценкой и используя поисковую систему (например, запросы *быстро и,* а также *быстро но*), мы можем пополнить оценочный лексикон. При этом необходимо учитывать следующее:

1. Поиск должен осуществляться только в рамках исследуемого корпуса.
2. Результатомпоискаявляетсяграф,гдеузламибудутприлагательные

или наречия, а ребрами – сходство или различие оценки.

3. При кластеризации результатов поиска получаются два подмножества оценочного лексикона, имеющих противоположную оценку. При этом оценка внутри каждого подмножества является одинаковой (положительной или отрицательной).

Для создания базового лексикона нами было взято 20 случайных документов – отзывов о качестве банковского обслуживания с сайта www.banki.ru (10 положительных и 10 отрицательных). Из данных документов вручную было отобрано около 100 лексем, составивших базовый лексикон.

Базовый лексикон был расширен с помощью синонимов, антонимов, а также с помощью поисковых запросов, запрашивались слова из базового лексикона с операторами *И* и *НО* (поисковая система www.google.com с ограничением поиска по сайту www.banki.ru).

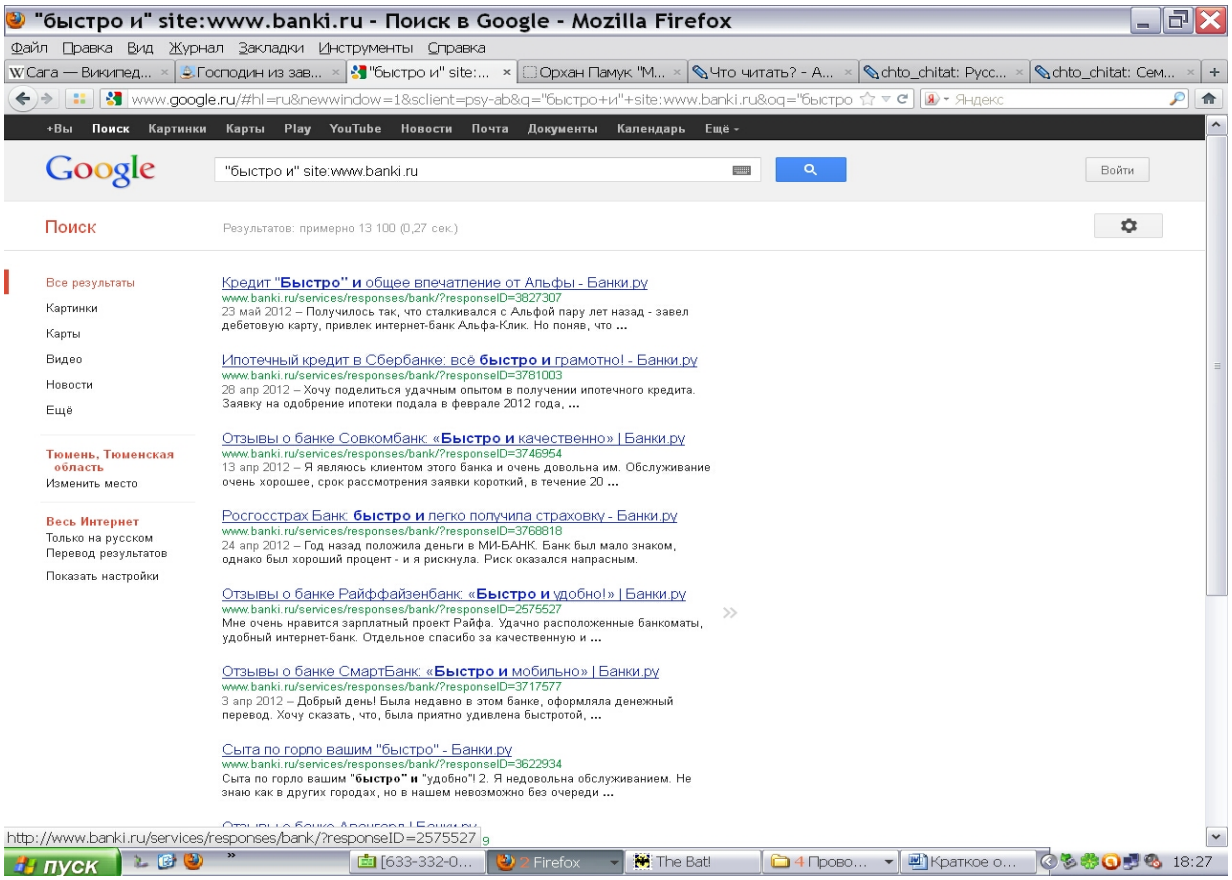

Рис. 1 Пример расширения лексикона с помощью оператора *И*.

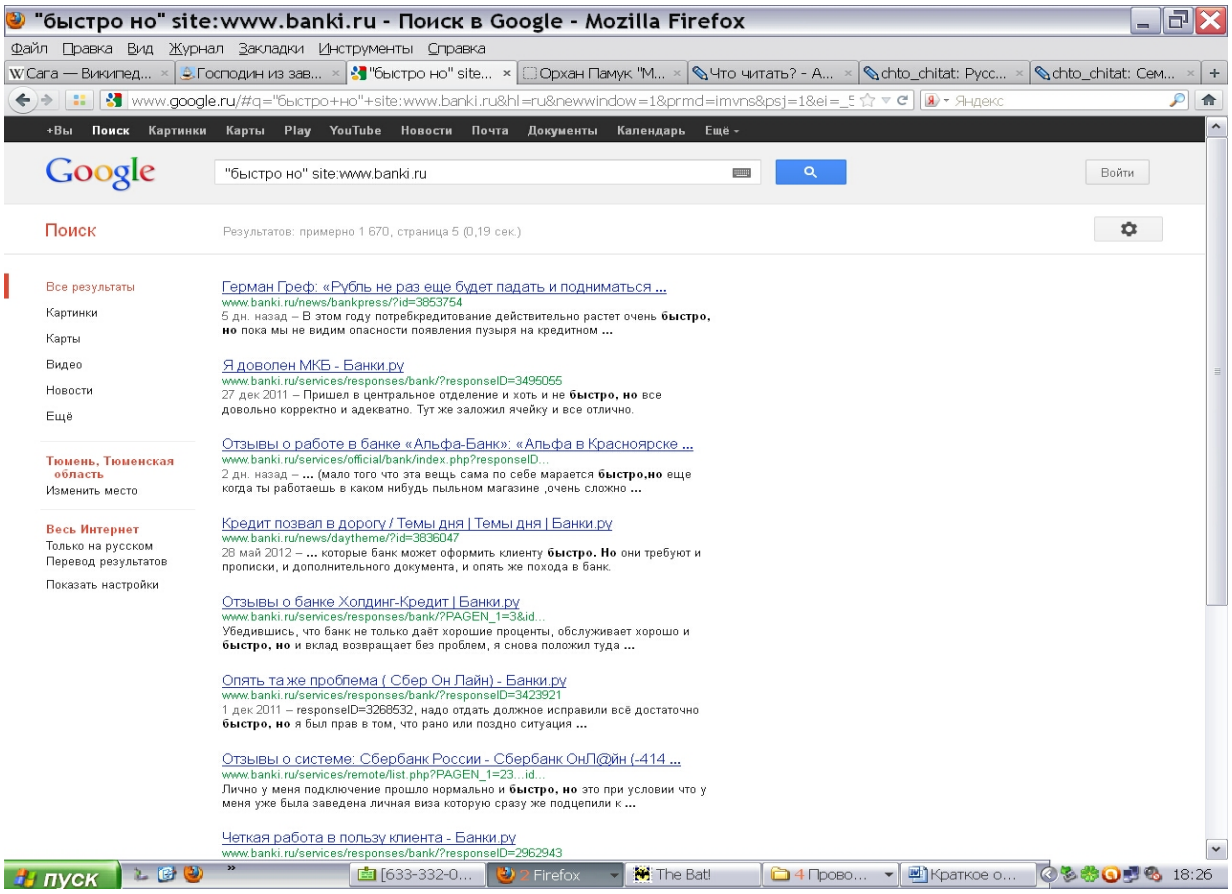

Рис. 2 Пример расширения лексикона с помощью оператора *НО*.

После расширения объем оценочного лексикона составил 500 единиц (207 положительных и 293 отрицательных).

Таблица 1. Состав оценочного лексикона

	+	-
Прилагательные и наречия	118	119
Глаголы	50	115
Существительные	39	59
Всего	207	293

Фрагменты оценочного лексикона

Положительный лексикон:

1. Прилагательные и наречия: Безопасный, бесплатный, быстрый,

вежливый, грамотный, качественный, компетентный, четкий, эффективный…

2. Глаголы: Возместить, впечатлить, выслушать, доверить, обрадоваться, отреагировать, поблагодарить, помогать, стараться, улыбаться…
3. Существительные: Благодарность, внимание, гарантия, доброжелательность, доверие, защита, инициатива, качество, компетентность, молодец, оптимизм, плюс, помощь …

Отрицательный лексикон:

4. Прилагательные и наречия: Агрессивный, бедный, безвыходный, бюрократичный, грубый, досадный, конфликтный, напряженный, небрежный, обидный, тесный, трудный, тяжелый…

5. Глаголы: Воровать, выплюнуть, заблокировать, материться, накричать, отказать, потерять, проглотить, ругать, рыдать, рыться, шипеть

6. Существительные: бред, вина, вор, гадюка, дурак, жалоба, конфликт, косяк, мрак, нервотрепка, обида, очередь, сбой, ступор, убогость, хамство, шок…

Таким образом, построение оценочного лексикона является трудоемким процессом, который приходится производить вручную с учетом предметной области, и использование лексикона одной предметной области для исследования другой представляется на данном этапе проблематичным. Определенным шагом на пути к универсализации оценочного лексикона могло бы стать использование биграмм и триграмм как единиц лексикона, а также выделение отдельных классов для лексических усилителей и для атрибутов, чувствительным к таковым. Составленный вручную базовый лексикон можно существенно расширить за счет синонимов, антонимов и поисковых запросов с булевыми операторами, а также пополнять и уточнять в процессе эксплуатации.

БИБЛИОГРАФИЯ

1. *Ермаков С.А., Ермакова Л.М.* Методы оценки эмоциональной окраски текста // Вестник Пермского университета. Вып. 1(19). 2012. С. 85-89.

2. *Пазельская А.Г., Соловьев А.Н.* Метод определения эмоций в текстах на русском языке: Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». Вып. 10 (17). – М.: Изд-во РГГУ, 2011. С.510-522.
3. *Стилистический энциклопедический словарь русского языка*. – М: Флинта, Наука. Под редакцией М.Н. Кожиной. 2003. 696 с.
4. *Четверкин И.И., Лукашевич Н.В. Извлечение и использование оценочных слов в задаче классификации отзывов на три класса // Вычислительные методы и программирование. 2011. Т.12. С. 73-81.*
5. *Liu, B.* Sentiment Analysis and Subjectivity // Handbook of Natural Language Processing**,** Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.
6. *Hatzivassiloglou V., McKeown K.* Predicting the Semantic Orientation of Adjectives // // Proc. of the 35th Annual Meeting of ACL, Madrid. 1997. P. 174-181.
7. *Hu M., Liu B.* Mining and summarizing customer reviews // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, 2004.
8. *Manning C., Raghavan P., Schütze H.* Introduction to Information Retrieval. Cambridge University Press, 2008.
9. *Nasukawa T., Yi J.,* Sentiment Analysis: Capturing Favorability Using Natural Language Processing // Proceedings of the 2nd International Conference on Knowledge Capture. Florida, 2003. P. 70-77.
10. *Pang B., Lee L., Vaithyanathan S.* Thumbs up? Sentiment Classification using Machine Learning Techniques // Proceedings of EMNLP, 2002. http://www.cs.cornell.edu/people/pabo/papers/sentiment.pdf.
11. *Turney P.* Thumbs Up or Thumbs Down? Semantic Orientation Applied to Supervised Classification of Reviews // Proceedings of the ACL, 2002. pp.417-424.
12. *Webb G., Boughton J., Wang Z.* Not So Naive Bayes: Aggregating One-Dependence Estimators // Machine Learning, 58, 2005. P. 5-24.