ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

М. В. Чернышевич

АКТУАЛЬНЫЕ АСПЕКТЫ РЕШЕНИЯ ЗАДАЧИ АВТОМАТИЧЕСКОГО СЕНТИМЕНТ-АНАЛИЗА ТЕКСТА

В статье рассмотрены некоторые аспекты решения задачи автоматического сентимент-анализа текста, сглажены терминологические разногласия, формально определена структура мнения, включающая наиболее важные с точки зрения конечного пользователя компоненты, определены основные требования к качественному решению целевой задачи.

Мнения, эмоции и сентименты выражают психическое состояние людей в зависимости от влияния разных обстоятельств и предметов окружающего мира. Их можно разделить на позитивные (радость, счастье, привязанность и пр.), негативные (печаль, гнев, страх и пр.) и нейтральные. Мнения и эмоции являются важной частью общения, которые во многом обусловливают действия людей.

Хотя компьютерная лингвистика и обработка естественного языка (ЕЯ) имеют богатую историю, автоматический анализ мнений в тексте стал предметом активных исследований только в последнее десятилетие. Для этого есть несколько причин. Во-первых, данный анализ имеет широкий спектр коммерческих применений и, прежде всего такого, как предоставление обратной связи от клиентов к производителям продуктов и услуг, а бизнес-потребности всегда провоцируют высокую мотивацию для исследований и экспериментов в соответствующей области. Во-вторых, технический прогресс и современные информационные технологии сделали возможным генерацию, сбор и обработку большого количества данных, требуемых для эффективного решения задачи и широко представленных, в первую очередь, в социальных сетях в виде мнений и оценок пользователей о поставляемых продуктах (товарах) и оказываемых услугах.

Так как обозначенная задача является относительно новой в сфере автоматической обработки ЕЯ, то в связанной с ней терминологии в различных источниках все еще наблюдаются разночтения. Так, общепринятыми переводами оригинального термина *sentiment analysis of text* на русский язык являются ‘анализ эмоциональной окраски текста’ или ‘анализ тональности текста’. Достаточно широко используемым стал и вариант ‘сентимент-анализ текста’.

Анализ эмоциональной окраски текста – задача автоматического анализа мнений и эмоциональной лексики, выраженных в тексте [1]. Цель решения данной задачи – это, прежде всего, распознавание эмоционально окрашенных фрагментов текста и далее распознавание объектов (собственно объектов, фактов, процессов, событий и т.п., их атрибутов и свойств), в отношении

которых в анализируемом тексте высказано субъективное мнение, и формирование их по определенным критериям оценки. При этом общее описание задачи сентимент-анализа в существующей литературе неоднозначно и это во многом обусловлено тем, что в качестве центральной единицы сентимент-анализа в разных источниках рассматриваются такие разные категории, как мнение, оценка, эмоции или чувства автора сообщения.

Рассмотрим подробнее понятия мнения и оценки.

Согласно Н. Д. Арутюновой, оценка – это признак значимости отдельного объекта для субъекта. Автор дает следующее определение: оценка – «это умственный акт, в результате которого устанавливается отношение субъекта к оцениваемому объекту с целью определения его значения для жизни и деятельности субъекта» [2]. Таким образом, оценка является абстрактной категорией и реализуется в языке в виде оценочного предложения.

Словарь С. И. Ожегова дает следующее определение понятия *мнение* (о фактах, событиях, лицах) – это суждение, выражающее чью-нибудь субъективную точку зрения, субъективное (личное) отношение к чему-нибудь [3].

Таким образом, с точки зрения задачи АСАТ, которая в первую очередь предполагает распознавание субъективного отношения автора к объектам окружающего мира, понятия оценки (оценочного предложения) и мнения являются в какой-то степени эквивалентными.

Нет однозначного мнения и о структуре оценки. Так, например, Т. В. Маркелова предлагает выделять в структуре оценки четыре основных компонента: субъект, объект, характер оценки и основание [4, с. 10].

Субъект оценки – индивид, приписывающий ценность определенному объекту путем выражения оценки, исходя из эмпирического опыта и общепринятых норм.

Объект оценки – предмет окружающего мира, по отношению к которому дается оценка.

Характер оценки – выражение определенного отношения между субъектом и объектом. Это отношение может быть определено категорией, как положительное или отрицательное, или позицией на шкале оценок, состоящей из зоны положительного и зоны отрицательного. Высказывание посетителя ресторана *“The ice cream in this restaurant was amazing!”* ‘Мороженое в этом ресторане было изумительным’ содержит положительную оценку (положительный характер оценки), близкую к максимальному значению на шкале оценок.

Основание оценки – это компонент высказывания, который выражает ее суть и является реальной основой оценочной конструкции.

Помимо обязательных элементов, структура оценки предполагает и наличие факультативных. К ним относят аксиологические предикаты мнения, ощущения, восприятия, а также мотивировки, различные средства интенсификации и деинтенсификации [5, с. 12].

В области прикладной лингвистики, сложилось немного иное представление оценки (мнения). Так, у Б. Лью [6, с. 19] оно состоит из следующих компонентов.

E n t i t y – объект, по отношению к которому высказано мнение. Он может относиться к одному или к целому классу объектов реального мира, процессов, ситуаций и т.д. В качестве объектов чаще всего выступают продукты, услуги, люди, организации, события или любой объект конкретного обсуждения.

Всякий объект действительности, выступающий в роли предмета оценки, обладает не определенным по числу и составу набором аксиологически релевантных свойств и компонентов, каждый из которых может в свою очередь оцениваться. В то время как Т. В. Маркелова не разделяет объект и его свойства на отдельные категории [4, с. 10], Б. Лью разграничивает свойства и характеристики объекта в особые компоненты оценки – аспекты объекта [6, с. 19]. Так, в предложении *This phone has great camera* ‘У этого телефона отличная камера’ можно выделить объект *phone* и аспект объекта *camera*.

V a l u e – оценка, высказанная в мнении, также – тональность мнения. Оценка может быть категориальной, то есть относиться к одной из определенных заранее категорий, например, положительная, негативная или нейтральная. Кроме того, оценка может ранжироваться по некоторой шкале, например от 1 до 5, с заданным шагом. В данном случае значение оценки будет указывать на ее интенсивность от крайне негативной (1) до крайне позитивной (5).

H o l d e r – автор или источник мнения.

T i m e – время, когда мнение было высказано.

Заметим, что представленная выше структура Б. Лью относится к понятию «мнение», в то время как близкая к ней структура, рассматриваемая Т. В. Маркеловой, – к понятию «оценка» [6, с. 51]. У Б. Лью эта категория также присутствует, но только как один из компонентов мнения (value), указывающий на тональность отношения между субъектом и объектом. Такие и другие различия в терминологии, как показывает анализ, действительно существуют в рассматриваемой задаче, что, как правило, характерно на начальных этапах развития той или иной области научных исследований. Мы в данной работе полагаем эквивалентными, во-первых, понятия мнения, эмоции и сентимента, а во-вторых – понятия оценки и тональности. Такая позиция ближе к позиции Б. Лью, о чем, собственно, подтверждает сформулированная нами цель решения задачи АСАТ, то есть мы исходим из того, что это решение, как эффективное, должно строиться на строгом структурном формальном представлении содержащихся в текстах мнений (эмоций, сентиментов) и распознавании, прежде всего, их тональности (оценки) и объектов (включая отдельные аспекты), в отношении которых высказано мнение. При этом тональность мнения в итоге может быть представлена либо теми же лексическими единицами, которые распознаны в оценочном предложении

в качестве таковых, либо, что характерно именно для современного подхода к решению указанной задачи, формализована на основе предварительной классификации тональности мнений.

На примере небольшого отзыва проиллюстрируем приведенные выше категории: *I bought an iPhone a few days ago. The touch screen is cool. But the camera resolution is really disappointing. 10.02.2017* ‘Я купил iPhone несколько дней назад. Отличный сенсорный экран. Но разрешение камеры разочаровало’.

Текст данного сообщения содержит несколько предложений, описывающих впечатление пользователя о телефоне. Предложение (1) содержит только фактическую информацию. Предложения (2) и (3) содержат мнение автора об отдельных аспектах телефона. Таким образом, это мнение, которое мы условно обозначим C, содержит следующие компоненты: субъект мнения h – сам автор сообщения, объект мнения o – телефон *iPhone*, аспекты объекта a – *touch screen* и *camera resolution*. Тональность v мнения по отношению к аспектам объекта разная: позитивная по отношению к *touch screen* и негативная по отношению к *camera resolution*.

Описанный опыт был получен пользователем в определенную дату t, вероятно, довольно близкую к дате публикации сообщения.

Таким образом, в компьютерной лингвистике принято рассматривать мнение в виде кортежа из пяти компонентов (субъект, объект, аспект, тональность, время):

$$C = (h,o,a,v,t),$$

где h – субъект, o – объект, a – аспект, v – тональность, t – время.

Если проанализировать вышеперечисленные компоненты мнения с точки зрения поставленной задачи АСАТ, то следует отметить, что такие компоненты, как субъект (автор) и время, во-первых, часто не имеют четкого формального выражения в тексте, а, во-вторых, не столь актуальны для массовых приложений, поэтому они остаются за рамками нашего анализа. Таким образом, наиболее важными компонентами мнения в рамках поставленной задачи являются: объект, его аспекты и тональность:

$$C = (o,a,v).$$

Определенная нами структура мнения предопределяет также ряд подзадач, решение которых необходимо в рамках комплексной задачи АСАТ:

1) распознавание эмоционально окрашенных текстовых фрагментов (оценочных предложений);

2) определение и категоризация объектов и их аспектов (характеристик), о которых высказывается мнение;

3) оценка тональности мнений по отношению к объектам и аспектным терминам в соответствии с разработанной шкалой.

За последнее десятилетие было проведено большое количество исследований и предложены различные методы решения данных подзадач,

которые условно можно разделить на две основные группы: построенные на использовании лингвистических правил (шаблонов, паттернов) и вероятностно-статистические методы.

Первые предполагают разработку экспертами специальных паттернов, учитывающих взаимосвязи слов внутри предложения. Во многих работах, например [7; 8; 9], правила применяются к структуре текста, полученной в результате его морфологической и синтаксической предобработки. Кроме этого, обычно требуется разработка большого количества таких лингвистических ресурсов, как словари тонально-окрашенной лексики, словари характеристик и параметров объектов и др.

Вторая группа методов основана на использовании различного рода частотных характеристик текста [10; 11], построении вероятностных моделей 13], машинном обучении с учителем [14; 15; 16; 17]. Эти методы предполагают наличие заранее размеченной коллекции (корпуса) текстов, на которой происходит обучение моделей.

Проведенный анализ показал, что каждая из указанных выше групп методов, наряду с достоинствами, обладает определенными недостатками, которые либо затрудняют, либо делают даже неэффективным их использование в промышленных системах обработки текстовых документов. Так, методы, основанные на лингвистических правилах, наиболее трудоемки по сравнению с методами в рамках других подходов, однако при хорошем наполнении тональных словарных списков и покрытии паттернами различных способов выражения сентимента по отношению к объекту, они позволяют достичь хороших качественных показателей. Недостаток этих методов, помимо их трудоемкости, в том, что довольно трудно «покрыть» правилами все возможные способы выражения целевых отношений в ЕЯ, и таким образом данный подход показывает обычно невысокую полноту, но высокую точность решения задачи.

Методы на основе машинного обучения демонстрируют сегодня наилучшие результаты при решении многих задач автоматической обработки ЕЯ. Кроме того, важным преимуществом является универсальность этих методов по отношению, прежде всего, к ЕЯ и типу решаемой задачи сентимент-анализа текста, обучение непосредственно на текстовых единицах, без необходимости создания широкого пространства признаков и разработки лингвистических ресурсов. К недостаткам статистических методов можно отнести необходимость аннотирования большого корпуса текстов для тренировки и тестирования модели. Кроме того, существующая сложность внутренней структуры классифицирующих моделей делает процесс интерпретации результатов здесь слабо контролируемым. Наконец, классификатор, обученный на текстах одной предметной области, может не «справляться» со своей задачей для текстов из другой предметной области.

В целом проведенный анализ показал, что современное качественное решение комплексной задачи АСАТ, пригодное для использования в промышленных системах обработки текстов, может быть построено на основе

комбинированного метода решения задачи, объединяющего технологии машинного обучения и экспертных лингвистических правил. Кроме того, необходим достаточно глубокий лингвистический анализ текстовых документов, включающий, кроме грамматического и синтаксического, также семантический уровень их анализа. Именно семантический анализ обеспечивает высокую степень унификации фактов, выраженных различными синтаксическими конструкциями, и по этой причине позволяет распознавать самый широкий спектр аспектов тональности и самой тональности в соответствии с гибкой шкалой.

Система ACAT должна выявлять оценочные высказывания в текстах произвольной предметной области и источников с высокими показателями полноты и точности работы. Для этого необходимо разработать модули фильтрации недостоверных мнений, коррекции слов, не соответствующих нормам ЕЯ, разграничения фактической и субъективной информации. Кроме того, промышленная система должна быть универсальной (легко адаптироваться) по отношению к различным языкам и индивидуальным особенностям языка субъекта.

Таким образом, в данной статье сглажены некоторые существенные с точки зрения поставленной задачи терминологические разногласия, все еще существующие в рассматриваемой предметной области, формально определена структура мнения, включающая наиболее важные, с точки зрения конечного пользователя, компоненты, определены основные требования к качественному решению целевой задачи.

ЛИТЕРАТУРА

1. *Pang, B.* Opinion mining and sentiment analysis / B. Pang, L. Lee // Foundations and Trends in Information Retrieval. – 2008. – Vol. 2. – С. 1–135.
2. *Арутюнова, Н. Д.* Типы языковых значений: Оценка. Событие. Факт / Н. Д. Арутюнова. – М. : Наука – 1988. – 337.
3. *Ожегов, С. И.* Толковый словарь русского языка / С. И. Ожегов, Н. Ю. Шведова. – М. : ИТИ «Технологии», 2003.
4. *Маркелова, Т. В.* Семантика оценки и средства ее выражения в русском языке / Т. В. Маркелова. – М. : МПУ, 1993. – 125 с.
5. *Вольф, Е. М.* Функциональная семантика оценки / В. М. Вольф. – М. : Наука, 1985. – 246 с.
6. *Liu, B.* Sentiment Analysis and OpinionMining / B. Liu // Morgan & Claypool Publishers. – 2012. – 180 с.
7. *Khan, A.* Sentiment classification from online customer reviews using lexical contextual sentence structure / A. Khan, B. Baharudin, K. Khan // ICSECS 2011 : Software Engineering and Computer Systems. – 2011. – С. 317–331.
8. *Паничева, П.* Система сентиментного анализа ATEX, основанная на правилах, при обработке текстов различных тематик / П. Паничева // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». – 2013. – С. 101–113.

9. *Bancken, W.* Automatically Detecting and Rating Product Aspects from Textual Customer Reviews / W. Bancken, D. Alfarone, J. Davis // Proceedings of DMNLP workshop at ECML/PKDD. – 2004. – С. 1–16.
10. *Hu, M.* Mining and summarizing customer reviews / M. Hu, B. Liu // Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004). – 2004. – С. 168–177.
11. *Long, C.* A review selection approach for accurate feature rating estimation / C. Long, J. Zhang, X. Zhu // Proceedings of Coling 2010: Poster Volume. – 2010. – С. 766–774.
12. *Ku, L.* Opinion extraction, summarization and tracking in news and blog corpora / L. Ku, Y. Liang, H. Chen // Proceedings of AAAI-CAAW'06. – 2006. – С. 100–107.
13. *Popescu, A.-M.* Extracting product features and opinions from reviews / A.-M. Popescu, O. Etzioni // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2005). – 2005. – С. 339–346.
14. *Chernyshevich, M.* IHS R&D Belarus: Cross-domain Extraction of Product Features using Conditional Random Fields / M. Chernyshevich // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). – Ireland. – 2014. – С. 309–313.
15. *Patra, B.G.* JU_CSE: A Conditional Random Field (CRF) Based Approach to Aspect Based Sentiment Analysis / B. G. Patra, S. Mandal, D. Das, S. Bandyopadhyay // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). – Ireland. – 2014. – С. 370–374
16. *Irsoy, O.* Opinion Mining with Deep Recurrent Neural Networks / O. Irsoy, C. Cardie // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – Doha, Qatar. – 2014. – С. 720–728.
17. *Socher, R.* Recursive deep models for semantic compositionality over a sentiment treebank / R. Socher, A. Perelygin, J. Wu, J. Chuang, D. Manning, A. Y. Ng, C. Potts // Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA. – 2013. – С. 1631–1642.

The article explores some aspects of sentiment analysis, resolves terminological ambiguities, defines the formal structure of an opinion that includes the most important for the end-users components and sets the requirements for a sentiment-analysis system of high quality.

Поступила в редакцию 22.11.17