Брунова Елена Георгиевна

АВТОМАТИЗИРОВАННЫЙ КОНТЕНТ-АНАЛИЗ МНЕНИЙ ТРЕХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

Исследование, выполненное в рамках прикладной лингвистики, посвящено анализу субъективной информации в пользовательском контенте. Проанализированы отзывы на русском языке из трех предметных областей, в качестве критерия эффективности применялась мера Ван Ризбергена. Установлено, что эффективность применяемого алгоритма не снижается при анализе отзывов из других предметных областей. Доказано, что система распознает положительные отзывы лучше, чем отрицательные.

Адрес статьи: www.gramota.net/materials/2/2014/12-2/9.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2014. № 12 (42): в 3-х ч. Ч. II. С. 43-47. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2014/12-2/

Эпизодическим персонажам рассказа А. О. Новодворский дает обычные имена, ничем не выделяющиеся на фоне других. Надя, Катя, Марья – переводные, в рассказе выполняют *информационно-стилистическую функцию*, поскольку именно они позволяют понять мотивы поведения героинь.

Список литературы

1. **Алексеев М. П., Бельчиков Н. Ф.** История русской литературы. М.: Академия наук, 1956. 531 с.
2. **Бондалетов В. Д.** Русская ономастика: учебное пособие для студентов пед. институтов по спец. «Рус. яз. и лит.». М.: Просвещение, 1983. 224 с.
3. **Даль В. И.** Толковый словарь живого великорусского русского языка: в 4-х т. М.: Русский язык, 1998. Т. 2. 450 с.
4. **Медведева М. В., Рычкова Н. П.** Словарь современного русского литературного языка: в 17-ти т. М.: Академия наук, 1961. Т. 11. 1302 с.
5. **Миронов Г. М.** Краткая литературная энциклопедия: в 9-ти т. М.: Сов. энцикл., 1978. Т. 5. 528 с.
6. **Митрофанова О. Д.** Словарь русских личных имен. М.: Наука, 1980. 406 с.
7. **Ожегов С. И., Шведова Н. Ю.** Толковый словарь русского языка: 80 000 слов и фразеологических выражений. Изд-е 4-е, доп. М.: АТЕМП, 2004. 944 с.
8. **Суперанская А. В.** Общая теория имени собственного. М.: Наука, 1973. 290 с.
9. **http://az.lib.ru/o/osipowichnowodworskij_a_o/** (дата обращения: 12.10.2013).
10. **http://feb-web.ru/feb/irl/il0/i92/I92-1682.HTM** (дата обращения: 08.02.2014).

CATEGORY OF A PROPER NAME IN THE STORY BY O. A. NOVODVORSKII “CAREER”

Bozhkova Galina Nikolaevna, Ph. D. in Philology
Kazan (Volga region) Federal University (Branch) in Elabuga
bozhkova.galina@mail.ru

Literary proper names present itself a very interesting object for research both in terms of linguistics and literary criticism. Meaningful aspect of names in the fiction is huge: clearness, vividness of a personage‘s characteristic is sometimes achieved by his name as itself; poetonyms within a literary text help to reveal storyline collisions, emotional experiences. Surnames of literary personages as well as their first names can give expressive and emotional, ironic or satirical characteristic to a person.

Key words and phrases: story; name; pseudonym; nickname; satirical function; characterological function; informational and stylistic function.

УДК 81'322
Филологические науки

Исследование, выполненное в рамках прикладной лингвистики, посвящено анализу субъективной информации в пользовательском контенте. Проанализированы отзывы на русском языке из трех предметных областей, в качестве критерия эффективности применялась мера Ван Ризбергена. Установлено, что эффективность применяемого алгоритма не снижается при анализе отзывов из других предметных областей. Доказано, что система распознает положительные отзывы лучше, чем отрицательные.

Ключевые слова и фразы: прикладная лингвистика; обработка естественного языка; алгоритм; контент-анализ мнений; предметная область; пользовательский контент.

Брунова Елена Георгиевна, д. филол. н., доцент
Тюменский государственный университет
egbrunova@mail.ru

АВТОМАТИЗИРОВАННЫЙ КОНТЕНТ-АНАЛИЗ МНЕНИЙ ТРЕХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ©

1. Введение

Задачу извлечения и обработки субъективной информации можно свести к задаче классификации корпуса текстов на два класса: с положительной (хорошо, нравится) и отрицательной (плохо, не нравится) оценками [14; 16]. Некоторые исследователи добавляют еще третий класс – с нейтральной оценкой [5] и даже четвертый класс – со смешанной оценкой [12], однако, в основе любой гибридной классификации мы обнаруживаем бинарный принцип.

Контент-анализ мнений (англ. *sentiment analysis*) – это группа методов для извлечения и последующей обработки мнений и эмоций из текстов на естественных языках. Первые публикации по контент-анализу мнений появились в конце 1990-х – начале 2000-х гг. [14-16], и с тех пор в этой области сделано достаточно много: составлены оценочные лексиконы, разработаны алгоритмы [9-11; 14]. Все эти успешные исследования посвящены анализу англоязычных текстов, и казалось логичным применить их результаты для других

языков, перевести лексиконы и модифицировать инструменты для синтаксического анализа. Что касается анализа текстов на русском языке, то публикации начали появляться с 2010 г. [3-8].

Практически сразу стало очевидно, что контент-анализ мнений должен специализироваться по конкретному естественному языку ввиду существенных особенностей морфологии и синтаксиса. Таким образом, инструменты для синтаксического анализа, разработанные для английского языка, не могут быть применены для других языков, в частности, – русского. Более того, попытки построить универсальный оценочный лексикон, основной и самый трудоемкий инструмент контент-анализа мнений, также иногда приводят к противоречивым результатам. В частности, одно и то же слово может относиться к разным классам в зависимости от контекста [9, p. 242].

Цель данного исследования заключается в рассмотрении возможностей и ограничений при контент-анализе мнений в нескольких предметных областях.

2. Материал и методика исследования

Для эксперимента случайным образом были отобраны 80 текстов на русском языке по предметным областям: *качество банковского обслуживания* с сайта [17], *качество обслуживания в гостиницах* и *достопримечательности* с сайта [18]. Шесть отзывов были изъяты из корпуса, поскольку они являются результатом автоматического перевода. Таким образом, корпус исследуемых текстов составил 74 отзыва.

Структура оценочного лексикона представлена ниже.

Главные классы:

Положительный лексикон: *Безопасный*, *бесплатный*, *вежливый*, *компетентный*, *четкий*, *эффективный* …

Отрицательный лексикон: *Агрессивный*, *безвыходный*, *грубый*, *досадный*, *обидный*, *трудный* …

Вспомогательные классы:

Инкременты: *Очень*, *совершенно*, *максимум, никогда, никто* …

Декременты: *немного*, *мало, небольшой, минимум* …

Модификаторы полярности: *Не*, *нет*, *без* …

Антимодификаторы полярности: *Так*, *такой* …

Эксперимент проводился с помощью программного комплекса SENTIMENTO, реализованного в виде Интернет-приложения на базе *Apache* [2]. Система предусматривает возможность пополнения лексикона с помощью импортирования файлов *Excel* и добавления отдельных слов.

Пользователь подтверждает заключение системы или опровергает его. Эти данные используются для дальнейшего совершенствования системы и определения ее эффективности. По сути, и система, и пользователь отвечают на один и тот же вопрос: относится ли текст к определенному классу, например, является ли данный текст положительным отзывом. Для сопоставления заключений системы с заключением пользователя применяется таблица сопряжения, см. Табл. 1.

Таблица 1.

Таблица сопряжения заключений системы и пользователя

	Релевантно (пользователь отвечает да)	**Нерелевантно (пользователь отвечает нет)**
Отобрано (система отвечает да)	*tp* (истинные да)	*fp* (ложные да)
Не отобрано (система отвечает нет)	*fn* (ложные нет)	*tn* (истинные нет)

Далее вычисляются точность, полнота и мера Ван Ризбергена. Точность (англ. *Precision*) – это доля отобранных текстов, которые являются релевантными $P = tp/(tp + fp)$. Полнота (англ. *Recall*) – это доля релевантных документов, которые были отобраны $R = tp/(tp + fn)$. Мера Ван Ризбергена вычисляется по формуле: $F1 = (2P \times R) / (P + R)$ [11, p. 155].

3. Эксперимент

На первом этапе эксперимента алгоритм *REGEX* был применен только для одной предметной области. Результаты представлены в Табл. 2.

Таблица 2.

Эффективность контент-анализа мнений на материале одной предметной области

Предметная область	**Качество банковского обслуживания (20 текстов)**
Точность (*POS*)	0,9
Полнота (*POS*)	0,93
***F*1 (*POS*)**	**0,92**
Точность (*NEG*)	1,00
Полнота (*NEG*)	0,77
***F*1 (*NEG*)**	**0,87**
***F*1 (среднее)**	**0,895**

Затем алгоритм был проверен на зависимость от предметной области. Для этого он был протестирован на большем количестве отзывов из предметной области *качество банковского обслуживания*, а также еще на двух предметных областях. Результаты представлены в Табл. 3.

Таблица 3.

Эффективность контент-анализа мнений на материале трех предметных областей

Предметная область	Качество банковского обслуживания (32 текста)	Качество обслуживания в гостиницах (22 текста)	Достоприм. (20 текстов)
Точность (*POS*)	0,82	0,92	0,89
Полнота (*POS*)	0,93	1,00	1,00
***F*1 (*POS*)**	**0,88**	**0,96**	**0,94**
Точность (*NEG*)	1,00	1,00	1,00
Полнота (*NEG*)	0,77	0,90	0,92
***F*1 (*NEG*)**	**0,87**	**0,95**	**0,96**
***F*1 (среднее)**	**0,875**	**0,955**	**0,95**

4. Результаты и их интерпретация

Как видно из Табл. 2 и 3, среднее значение *F*1 для предметной области *качество банковского обслуживания* несколько снизилось после расширения экспериментального корпуса. Причиной является то обстоятельство, что некоторые оценочные слова, определяемые пользователем, не входили в оценочный лексикон и, следовательно, не опознавались системой. Это легко исправить с помощью пополнения оценочного лексикона.

Эксперимент показал, что значения точности, полноты и *F*1 не снизились, после того как алгоритм был применен для других предметных областей. Более того, средние значения *F*1 для предметных областей *качество обслуживания в гостиницах* и *достопримечательности* оказались даже несколько выше, чем *F*1 для предметной области *качество банковского обслуживания* (соответственно 0,955; 0,95 и 0,875).

Было установлено, что алгоритм в целом и оценочный лексикон в частности существенно не зависят от предметной области. Особое внимание было уделено ошибкам системы (ложным да и нет). В этих случаях тексты анализировались вручную, чтобы выявить проблемы алгоритма, приводящие к некорректному анализу. Их можно обобщить следующим образом:

- **Неоднозначность**:

– **Полисемия**: некоторые слова могут иметь положительное или отрицательное значение наряду с нейтральным, ср. *С меня довольно,* но *Комната довольно удобная. Ситуация стала довольно неприятной.* В первом случае *довольно* отрицательное слово, во втором и третьем – нейтральное;

– **Параметрическая лексика**: некоторые слова из оценочного лексикона демонстрируют зависимость от предметной области, например, *долго* относится к положительному лексикону при оценке времени работы батареи (предметная область *смартфон*), и к отрицательному – при оценке затрат времени клиента (предметная область *качество банковского обслуживания*). Это параметрическая лексика, т.е. слова, обозначающие возрастание или убывание какого-либо параметра, специфичного для данной предметной области;

– **Устойчивые выражения**: поведение отдельного слова отличается от его поведения в составе устойчивого выражения, например, *так* – это антимодификатор полярности, *так себе* относится к отрицательному лексикону, а *так держать* – к положительному;

– **Неоднозначное *и***: союз *и*, как правило, соединяет сложносочиненные части предложения, например, *Девушка предложила карту Visa Gold, не объяснив мне условия обслуживания **и** не показав тарифы* [17]. Определение этого союза очень важно для ряда правил формальной грамматики нашего алгоритма, в том числе – для примера, приведенного выше (модификатор *не* изменяет полярность двух положительных слов: *объяснив*, *показав,* соединенных *и*. Однако *и* иногда ведет себя как инкремент, ср. *Могли бы оставить вход **и** бесплатным.)*

- **Ирония и сарказм**: автор отзыва может использовать слова из положительного лексикона, хотя смысл предложения оказывается отрицательным. Если такие слова заключены в кавычки, можно определить их с помощью одного из правил нашего алгоритма, например, *да уж, _привлекательные условия‘*. Если кавычки не используются, алгоритм не способен определить иронию, например, *ВОТ ЭТО СЕРВИС! У нас многотысячный город! И нет возможности внести наличные! Красота! А если людям кредит надо платить* [Ibidem]?!
- **Неоправданные ожидания**: в некоторых отзывах мы наблюдаем положительное начало и отрицательное заключение, поворотной точкой обычно служит союз *но* или *однако,* например: *Когда открываешь номер, заходишь, то хочется сказать «О! Круто!». Но потом понимаешь, что всё очень неудобно и не продумано* [18].
- **Отрицательный лексикон используется для описания самой предметной области, а не мнения**: Эта проблем часто возникает при описании фильмов ужасов и т.п. Мы столкнулись с ней при описании гостиницы: *На коврах в коридоре фразы из _Преступления и наказания‘, мрачные тона, полумрак. Все Вам дает понять, что отель сделан _по Достоевскому‘. Даже то, что в центральном корпусе потолки не отделаны, а просто покрашены в серый цвет позволяет Вам проснуться и увидеть этот мир глазами Раскольникова* [Ibidem].
- **Ошибки стемминга**: В случае некорректного стемминга слово не определяется системой или относится к неверному классу, например, *люб-* может означать не только положительное *любить*, но и нейтральное *любой*.

• **Ошибки орфографии и / или пунктуации**: В случае орфографических ошибок слово не определяется системой или относится к неверному классу. В случае погрешностей в пунктуации некорректно работают правила формальной грамматики.

5. Заключение

Правила формальной грамматики, входящие в алгоритм REGEX, могут считаться универсальными для анализа любого текста на русском языке, поскольку они были сформулированы, исходя из особенностей грамматического строя русского языка. Второстепенные классы лексикона (инкременты, декременты, модификаторы и антимодификаторы полярности) также не зависят от предметной области. Главные классы (положительный и отрицательный лексиконы) подвергаются бóльшему риску зависимости от предметной области. Однако эксперимент на трех предметных областях не показал зависимости от предметной области, а эффективность применяемого алгоритма не снижалась.

Были проанализированы ошибки системы, расширен оценочный лексикон, исправлены ошибки стемминга. В алгоритм была добавлена проверка орфографии и пунктуации. Что касается иронии и сарказма, то в большинстве случаев представляется невозможным их определение с помощью инструментов контент-анализа мнений.

Эксперимент показал, что эффективность анализа положительных отзывов при анализе одной предметной области выше, чем отрицательных, т.е. система лучше распознает положительные отзывы. Автор положительного отзыва склонен использовать стандартные слова и фразы. Автор отрицательного отзыва, напротив, склонен к использованию нестандартных языковых средств. Вариативность при выражении отрицательных эмоций является той особенностью, о которой писал Лев Толстой: «Все счастливые семьи похожи друг на друга; каждая несчастливая семья несчастлива по-своему». Иными словами, норма, вызывающая положительные эмоции, всегда стереотипна, а отклонений от нее бесконечно много. При возрастании вариативности языковых средств риск отсутствия того или иного слова в оценочном лексиконе также возрастает.

С другой стороны, тенденция более высокой эффективности при распознавании положительных отзывов не наблюдается в предметной области *достопримечательности*. По-видимому, при высказывании мнения по поводу достопримечательностей авторы менее эмоциональны, чем при описании взаимоотношений с работниками банка или гостиницы, поскольку они не вовлечены в межличностные конфликты.

Автор выражает свою признательность к. филол. н. Юлии Владимировне Бидуле за помощь в проведении эксперимента.

Список литературы

1. **Брунова Е. Г.** Составление лексикона для контент-анализа мнений // Теоретические и прикладные аспекты изучения речевой деятельности. Н. Новгород: НГЛУ им. Н. А. Добролюбова, 2013. Вып. 1 (8). С. 24-29.
2. **Брунова Е. Г., Бидуля Ю. В.** Алгоритм с элементами формальной грамматики для контент-анализа мнений // Вестник Тюменского государственного университета. Серия «Физико-математические науки. Информатика». 2014. № 7. С. 242-250.
3. **Ермаков С. А., Ермакова Л. М.** Методы оценки эмоциональной окраски текста // Вестник Пермского университета. 2012. Вып. 1 (19). С. 85-89.
4. **Куликов С. Ю.** Автоматизация составления оценочного словаря широкой предметной области (опыт использования неспециализированного корпуса текстов) // Вестник Иркутского государственного технического университета. 2014. № 8 (91). С. 240-243.
5. **Лукашевич Н. В., Четверкин И. И.** Извлечение и использование оценочных слов в задаче классификации отзывов на три класса // Вычислительные методы и программирование. 2011. Т. 12. С. 73-81.
6. **Оробинская Е. А., Кочуева З. А.** Технологии *Text Mining*: Обзор методов и задач обработки смысловой информации // Вестник Херсонского национального технического университета. 2010. № 2 (38). С. 348-353.
7. **Пазельская А. Г., Соловьев А. Н.** Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». М.: Изд-во РГГУ, 2011. Вып. 10 (17). С. 510-522.
8. **Полякова Е. В.** Когнитивные особенности выражения моральных чувств «Любовь» и «Страх» в идиоматике русского и английского языков // Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2013. № 9 (27): в 2-х ч. Ч. 2. С. 157-163.
9. **Ganapathibhotla M., Liu B.** Mining Opinions in Comparative Sentences // Proc. of the 22nd International Conference on Computational Linguistics. Manchester: Brighton, 2008. P. 241-248.
10. **Liu B.** Sentiment Analysis and Subjectivity [Электронный ресурс] // Handbook of Natural Language Processing. 2010. URL: http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf (дата обращения: 27.09.2014).
11. **Manning Ch., Raghavan P., Schütze H.** Introduction to Information Retrieval. Cambridge: Cambridge UP, 2009. 544 p.
12. **Pal J., Saha A.** Identifying Themes in Social Media and Detecting Sentiments // International Journal of Statistics and Applications. 2011. Vol. 1. No. 1. P. 14-19.
13. **Pang B., Lee L.** Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. 2008. Vol. 2. No 1-2. P. 1-135.
14. **Pang B., Lee L., Vaithyanathan S.** Thumbs up? Sentiment Classification Using Machine Learning Techniques [Электронный ресурс] // Proc. of EMNLP. 2002. URL: http://www.cs.cornell.edu/people/pabo/papers/sentiment.pdf (дата обращения: 27.09.2014).
15. **Turney P.** Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: University of Pennsylvania, 2002. P. 417-424.
16. **Wiebe J., Bruce R., O'Hara T.** Development and Use of a Gold-Standard Data Set for Subjectivity Classifications // Proc. of the 37th Annual Meeting of the Association for Computational Linguistics. Maryland: University of Maryland, 1999. P. 246-253.
17. **www.banki.ru** (дата обращения: 26.10.2014).
18. **www.tripadvisor.ru** (дата обращения: 26.10.2014).

AUTOMATIZED CONTENT ANALYSIS OF COMMENTS IN THREE SUBJECT AREAS

Brunova Elena Georgievna, Doctor in Philology, Associate Professor
Tyumen State University
egbrunova@mail.ru

The research executed within the framework of applied linguistics is devoted to the analysis of subjective information in the user's content. The author analyzes the comments in the Russian language from the three subject areas applying Van Rijsbergen's effectiveness measure as a criterion of efficiency. It is shown that the efficiency of the applied algorithm does not decrease while analyzing the fragments from the other subject areas. The researcher argues that the system recognizes positive comments better than negative.

Key words and phrases: applied linguistics; processing natural language; algorithm; content analysis of comments; subject area; user's content.

УДК 81.008(045)
Филологические науки

В статье на основе сопоставительного анализа содержания концепта БЕЗОПАСНОСТЬ выявлена этнокультурная специфика его репрезентации в русской и английской языковых картинах мира. Наряду со сходством установлена асимметричность представленности лексических единиц и их значений в сопоставляемых языках. Обнаружено наличие общего и особенного в выражении представлений о безопасности представителями двух различных этнокультур.

Ключевые слова и фразы: языковая картина мира; концепт; концепт БЕЗОПАСНОСТЬ; вербализация концепта; значение; этнокультурная специфика.

Варданян Людмила Валерьевна, к. филол. н.
Щукина Екатерина Сергеевна
Мордовский государственный педагогический институт имени М. Е. Евсевьева
ljudmila_v@mail.ru; shchukina.ekaterinka@mail.ru

ЭТНОКУЛЬТУРНАЯ СПЕЦИФИКА РЕПРЕЗЕНТАЦИИ КОНЦЕПТА «БЕЗОПАСНОСТЬ» В РУССКОЙ И АНГЛИЙСКОЙ ЯЗЫКОВЫХ КАРТИНАХ МИРА©

Работа проводилась при поддержке Минобрнауки РФ в рамках Программы стратегического развития ФГБОУ ВПО «Мордовский государственный педагогический институт им. М. Е. Евсевьева» на 2012-2016 гг. «Педагогические кадры для инновационной России».

В настоящее время не вызывает сомнения тот факт, что язык и культура взаимосвязаны и оказывают друг на друга взаимовлияние. Анализируя роль и место языка в культуре, Н. Д. Арутюнова высказала мысль о его инструментальных возможностях, которые обеспечивают создание и приумножение культуры, ее применение, сохранение, выражение. Она пишет, что «Язык формирует концепты и суждения, осуществляет коммуникацию …, хранит историческую и культурную память народов, выражает и сохраняет знания о мире и человеке» [3, с. 3].

Язык, который «становится средством отражения … внутреннего мира» [1, с. 96], реализует определенный способ отображения человеком действительности в соответствии с конкретным историческим опытом данного народа, его культурой, условиями жизни. По наблюдению Л. П. Водясовой, «Каждая нация имеет свое мировосприятие и воплощает его в своей особенной, неповторимой проекции, иными словами, имеет свой способ концептуализации» [7, с. 120]. На разнообразие языковых средств, используемых для вербализации концептов, оказывает влияние как историческая эпоха, так и уровень культурного развития социума: оно объясняется «традициями того или иного народа, спецификой его мышления, принципами кодирования информации и т.п.» [6, с. 61].

Исходя из того, что культура находит уникальное воплощение в языке, анализ языкового выражения концептов является одним из наиболее эффективных путей познания системы ценностей, мироощущения и картины мира представителей той или иной культуры. В поле зрения отдельных этносов «попадают лишь отдельные фрагменты цельного образа мира» [14, с. 156]. В результате этого, концепты могут варьироваться в разных культурах и их семантический состав также «меняется от культуры к культуре, от этноса к этносу, от одной социальной группы к другой и от одной личности к другой» [8, с. 7]. Особый интерес в плане исследования представляют концепты, которые «отражают специфику национально-культурной картины мира» [11, с. 130], раскрывают ценностные приоритеты культуры и занимают важное место в жизни этноса.

Н. Ф. Алефиренко отмечает, что «попытки некоторых авторов показать этнокультурное своеобразие языковой картины мира на материале одного языка не имеют доказательной базы» [2, с. 6]. В нашем исследовании национальная специфика различных менталитетов прослеживается на примере репрезентации одного и того же концепта в разных языках и культурах, «что способствует выявлению общего и особенного в восприятии, понимании и концептуализации мира представителями различных этнокультур» [5, с. 122].