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

КОНТЕНТ-АНАЛИЗ МНЕНИЙ КАК ЗАДАЧА КЛАССИФИКАЦИИ

Введение. Информацию, содержащуюся в текстах на естественном языке, можно условно отнести к одному из двух типов: факты и мнения. Факты – это объективная информация, описывающая сущности и события, а также их свойства. Мнения – это субъективная информация, описывающая оценку (одобрение или неодобрение) и эмоции человека по отношению к сущностям и событиям, а также их свойствам. Успешные разработки в области поиска, извлечения и обработки информации из текстов на естественном языке, как правило, сосредоточены на задачах, связанных с фактами, и только в последнее десятилетие стали появляться исследования, посвященные поиску, извлечению и обработке мнений [Carenini et al., 2005], [Gamon et al., 2005], [Hu & Liu, 2004], [Liu, 2010], [Nasukawa & Yi, 2003], [Pang & Lee, 2008], [Turney, 2002], [Ермаков, Ермакова, 2012], [Лукашевич, Четверкин, 2011], [Брунова, 2012].

В настоящее время анализ субъективной информации является одной из наиболее перспективных задач в области обработки текста на естественном языке [Топтыгина, 2011], [Черкасс, 2009]. Задачу контент-анализа мнений (КАМ) можно свести к задаче классификации корпуса текстов на два класса: с положительной (хорошо, нравится) и отрицательной (плохо, не нравится) оценкой [Павлов, Добров, 2011], [Hehery, 1994], [Wiebe

et al., 1999]. Некоторые исследователи добавляют еще третий класс – с нейтральной оценкой [Лукашевич, Четверкин, 2011] и даже четвертый класс – со смешанной оценкой [Pal & Saha, 2011], однако в основе любой гибридной классификации мы обнаруживаем бинарный принцип. Обязательные и необязательные классы при КАМ указаны в Таблице 1.

Таблица 1

Возможные классы при контент-анализе мнений

	Обязательные классы		Необязательные классы	
Класс	Положительная оценка (+)	Отрицательная оценка (-)	Смешанная оценка (+ и -)	Нейтральная оценка (ни +, ни -)
Интерпретация	Нравится	Не нравится	Что-то нравится, что-то не нравится	Отсутствие оценки

Модель объекта контент-анализа мнений. Объект КАМ (opinion target) – это сущность, по поводу которой высказывается мнение. Такая сущность может быть представлена товаром, услугой, личностью, организацией, событием, темой, идеологической платформой и т.д. Объект КАМ может состоять из ряда частей и обладать набором атрибутов. Каждая часть объекта, в свою очередь, может обладать своим набором атрибутов. Таким образом, объект КАМ может быть представлен как дерево, иерархия или таксономия [Liu, 2010].

Рассмотрим объект КАМ на примере отзыва о смартфоне Apple Iphone v. 4:

10 августа 2011 г. 20:05 (1) Айфон мне очень нравится! (2) Огромный сенсорный экран - это просто наслаждение! (3) Сам телефон легкий, тонкий и красивый! (4) Долго не требует подзарядки, что для меня очень удобно. (5) Можно работать в интернете, при таком огромном экране и быстром срабатывании телефона поиск информации в сети – одно удовольствие! (6) Работает быстро как компьютер! (7) Камера делает прекрасные снимки, можно тут же в телефоне редактировать фото! (8)

До этого хотела купить маленький ноутбук, а теперь думаю с таким телефоном и компьютера не надо, он все может! (9) Однако мама сказала, что это очень дорого.

В данном примере объектом КАМ является модель смартфона Apple Iphone v. 4. У данного объекта имеется множество частей (экран, камера, аккумулятор, операционная система) и множество атрибутов (размер, вес, качество голосовой связи) (Рис.1). Каждая часть может иметь собственное множество атрибутов, например, часть *аккумулятор* имеет атрибуты: размер, вес, время работы без подзарядки. Таким образом, данный объект может быть представлен как дерево, корневым узлом которого является сам объект КАМ, а другими узлами – его части или подчасти. Связь между узлами является отношением *целое – часть*. Мнение может быть высказано по поводу любого узла объекта КАМ или любого его атрибута.

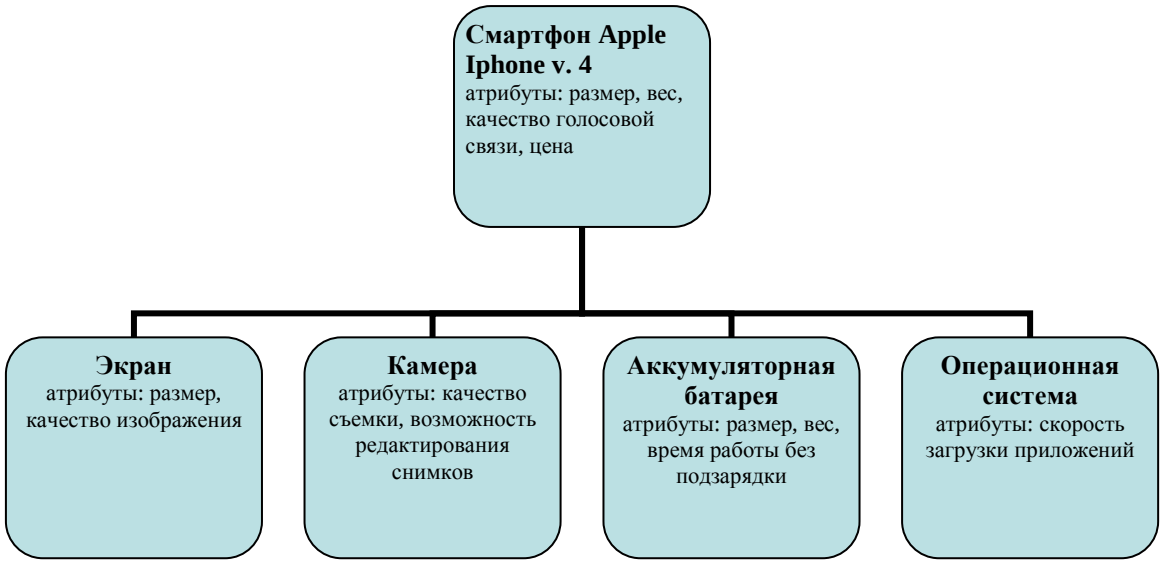

Рис. 1. Структура объекта КАМ (на примере модели смартфона Apple Iphone v. 4).

В нашем примере можно извлечь мнения по поводу объекта КАМ как корневого узла (предложения 1, 3, 8, 9), узла *экран* (предложения 2, 5), узла *батарея* (предложение 4), узла *камера* (предложение 7) и узла *операционная система* (предложения 5, 8).

Модель объекта КАМ можно представить следующим образом:

O: (T, A)

где O – объект КАМ

T – иерархия частей и подчастей объекта

A – множество атрибутов

Однако для практического применения КАМ, особенно для обычных пользователей – читателей отзывов, иерархическая структура представляется излишне сложной. Поэтому Б. Лью предлагает упростить иерархическую структуру до плоской модели, где свойство объекта (object feature) обозначает как собственно атрибут объекта или его части, так и сам объект в целом (специфическое свойство) или какую-либо из его частей [Liu, 2010]. Модель объекта КАМ при таком упрощении будет являться моделью, основанной на свойствах (feature-based model).

Свойство объекта КАМ (object feature) выражается словом или группой слов естественного языка и рассматривается в контексте документа КАМ. Оно может быть эксплицитным или имплицитным. Свойство является эксплицитным, если оно выражается словом, включая синонимы, или группой слов естественного языка в документе КАМ, например, *Камера делает прекрасные снимки* (свойство – *камера*). Свойство является имплицитным, если оно не выражается напрямую в документе КАМ, а подразумевается в его контексте, например, *Сам телефон легкий, тонкий и красивый!* (свойства – *вес*, *размер*, *внешний вид*).

Упрощенная модель объекта КАМ будет иметь следующий вид:

O: (F)

где O – объект КАМ

F – множество свойств, включая сам объект (специфическое свойство), его части и атрибуты

Модель документа контент-анализа мнений. Под документом КАМ (opinionated document) мы понимаем текст на естественном языке, выражающий мнение по поводу объекта КАМ. Документ КАМ может

выражать как общее мнение об объекте (*Айфон мне очень нравится*), так и мнение о любой его части (*Огромный сенсорный экран – это просто наслаждение*). Примерами документов КАМ могут быть отзывы о товарах или услугах, посты или комментарии на форумах или блогах.

Модель документа КАМ можно представить следующим образом:

$d = (s_1, s_2, ..., s_n)$,

где d – документ КАМ,

s – предложение

Необходимо отметить, что не каждое предложение документа КАМ может содержать мнение. Автор документа КАМ может добавить информацию о фактах, что является важным для определения достоверности отзыва, но представляет собой избыточную информацию для самой процедуры КАМ. С другой стороны одно предложение может содержать два мнения, например, *Экран замечательный, но батарея быстро разряжается.* Тем не менее, чаще всего, одно предложение содержит одно мнение или не содержит такового, поэтому именно предложение документа КАМ мы считаем основной единицей анализа.

Информация, которую мы можем извлечь из документа КАМ, не ограничивается объектом КАМ. Для КАМ имеет значение также субъект, предметная область, полярность, дата и время документа КАМ.

Субъект КАМ (opinion holder) – это личность или организация, выражающая свое мнение, эмоции и настроение по поводу объекта КАМ. Субъект КАМ может не совпадать с автором текста, например, в случае цитирования мнения другой личности: в нашем примере субъектами КАМ являются пользователь смартфона: «я» (предложения 1, 2, 3, 4, 5, 6, 7, 8) и «моя мама» (предложение 9).

Предметная область (domain) – предметная область, к которой относится объект КАМ. В нашем примере предметной областью является смартфон.

Полярность или семантическая ориентация (polarity, semantic orientation) – одобрение объекта КАМ (положительная полярность), осуждение (отрицательная полярность) или нейтральное отношение (нулевая полярность). В нашем примере мы наблюдаем положительную полярность (предложения 1, 2, 3, 4, 5, 6, 7, 8) и отрицательную полярность (предложение 9).

Сила полярности – условная шкала полярности. Чаще всего используются две шкалы: трехбалльная (нравится, безразлично, не нравится) и пятибалльная или пятизвездочная (очень хорошо, хорошо, без оценки, плохо, очень плохо). Последняя шкала связана со сложившейся традицией оценки разнообразных продуктов и сообщений на веб-сайтах, а также системы оценки гостиничного сервиса.

Дата и время документа КАМ имеют значение для динамического анализа и достаточно легко извлекаются, поскольку, как правило, они автоматически фиксируются на сайтах отзывов или форумах.

Таким образом, корпус документов КАМ содержит мнения, высказанные множеством субъектов $\{h_1, h_2, \ldots, h_n\}$ по поводу множества объектов $\{o_1, o_2, \ldots, o_n\}$, а каждый объект выражается подмножеством свойств Fj. Мнение об объекте КАМ включает пять элементов $(o_j, f_k, oo_{ijkl}, h_i, t_l)$, где o_j – объект КАМ , f_k – свойство объекта o_j, oo_{ijkl} – полярность мнения о свойстве f_k объекта o_j, h_i – субъект КАМ, а t_l – время и дата мнения, высказанного h_i [23], [26], [27].

Для решения задач КАМ используются различные методы и алгоритмы: наивный Байесовский классификатор (Naïve Bayes) [Ермаков, Ермакова, 2012], [Лукашевич, Четверкин, 2011], [Manning et al., 2008], алгоритм PMI-IR [Manning et al., 2008], [Turney, 2002], [Webb et al., 2005], алгоритм k-ближайших соседей (KNN, k-nearest neighbor algorithm) [Ермаков, Ермакова, 2012], [Лукашевич, Четверкин, 2011], метод опорных векторов (SVM, support vector machine) [Salton et al., 1975], латентный семантический анализ (Latent Semantic Analysis) [Blei et al., 2003], метод нейронных сетей

[Manning et al., 2008], метод логистической регрессии (Logistic Regression) [Manning et al., 2008], метод близости косинусов угла (Cosine Similarity) [Dittenbach], Good Grief Algorithm [Snyder & Barzilay, 2007] и др.

Данные методы и алгоритмы относятся к двум крупным категориям: методы обучения с учителем (supervised learning) и методы обучения без учителя (unsupervised learning)

Методы обучения с учителем сортируют полнотекстовые документы по заранее известным категориям (классам). Обычно множество документов делят на две части: одна часть является обучающим множеством, т.е. данными для обучения алгоритма, вторая – тестовым множеством, т.е. данными для оценки качества полученного классификатора. В роли учителя выступает выборка документов, для которых заранее известна принадлежность той или иной категории (обучающее множество). Множество категорий и обучающее множество документов формируют эксперты. Типичным примером метода обучения с учителем для КАМ является наивный Байесовский классификатор (Naïve Bayes) [Webb et al., 2005].

Методы обучения без учителя предусматривают спонтанное обучение системы выполнению поставленной задачи, т.е. без вмешательства со стороны человека. Такие методы используются, как правило, для задач, в которых известны описания множества объектов (обучающей выборки), и необходимо обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами. Типичным примером метода обучения без учителя является алгоритм PMI-IR (англ. *Pointwise Mutual Information* – межточечная взаимная информация, и *Information Retrieval* – извлечение информации), предложенный П. Терни [Turney, 2002].

Оценка эффективности методов контент-анализа мнений. Для оценки эффективности методов КАМ традиционно используют матрицу неточностей (confusion matrix), применяемую в поисковых системах при оценке степени соответствия поискового запроса найденной информации

(Таблица 2). В основе такой оценки лежит сопоставление результатов оценки поисковой системы с оценкой эксперта (человека) [Manning et al., 2008].

Таблица 2. Матрица неточностей

	Система говорит да	Система говорит нет
Человек говорит да	tp	fn
Человек говорит нет	fp	tn

где – tp - количество найденных релевантных документов, «истинные да» (true positive)

tn – количество найденных нерелевантных документов, «истинные нет» (true negative)

fp – количество ненайденных нерелевантных документов, «ложные да» (false positive)

fn – количество ненайденных релевантных документов, «ложные нет» (false negative)

Левый столбец матрицы (tp + fp) – это общее количество документов, найденных системой. Первая строка (tp + fn) – это общее количество релевантных документов. Вторая строка (fp + tn) – это общее количество нерелевантных документов.

Применительно к оценке классификаций релевантным считается результат правильного отнесения документа к тому или иному классу, нерелевантным – результат ошибочного отнесения документа к тому или иному классу.

Для измерения эффективности используются *точность* P (precision), *полнота* R (recall), *правильность* A (accuracy) и *мера* F_1 (мера Ван Ризбергена).

Точность представляет собой отношение количества найденных релевантных документов к общему количеству документов, найденных системой.

$$P = \left[\frac{tp}{tp + fp} \right] \quad (1)$$

Полнота представляет собой отношение количества найденных релевантных документов к общему количеству релевантных документов.

$$R = \left[\frac{tp}{tp + fn} \right] \quad (2)$$

Правильность представляет собой отношение общего количества релевантных документов к сумме общего количества релевантных документов и общего количества нерелевантных документов.

$$A = \left[\frac{tp + fn}{tp + tn + fp + fn} \right] \quad (3)$$

Мера F1 (мера Ван Ризбергена) вычисляется по формуле:

$$F1 = \left[\frac{2P * R}{P + R} \right] \quad (4)$$

Необходимо отметить, что оценка эффективности метода классификации может быть названа объективной лишь условно, поскольку она основана на оценке человека, и, следовательно, – содержит элемент субъективности.

Заключение. Задачу контент-анализа мнений можно свести к задаче классификации корпуса текстов на два класса: с положительной и отрицательной оценкой. Исходя из особенностей модели объекта КАМ и модели документа КАМ возможно использование различных методов и алгоритмов.

Типичным представителем группы методов обучения с учителем является наивный Байесовский классификатор. Его достоинствами являются быстрота исполнения при относительно высокой точности измерения, небольшие потребности для хранилища данных, малое количество данных, необходимых для обучения и устойчивость к несущественным атрибутам. Недостатками являются проблема потери значимости и искажение результатов из-за пренебрежения синтаксическими отношениями между словами.

Типичным представителем группы методов обучения с учителем является алгоритм PMI-IR. Его достоинствами является обучение системы

без вмешательства человека, а также возможность применения к любой предметной области. Недостатком является невысокая точность измерения.

Для сравнения эффективности работы алгоритмов используется мера F1 (мера Ван Ризбергена), которая вычисляется с учетом точности (отношения количества найденных релевантных документов к общему количеству документов, найденных системой) и полноты (отношение количества найденных релевантных документов к общему количеству релевантных документов). Оценка эффективности метода классификации может быть названа объективной лишь условно, поскольку она основана на оценке человека, и, следовательно, – содержит элемент субъективности.

БИБЛИОГРАФИЯ

1. *Брунова Е.Г.* Методика составления оценочного лексикона для контент-анализа мнений [Электронный ресурс] / Е. Г. Брунова // Language and Science. – 2012. – Вып.1. – Электрон. дан. – [2012]. – Режим доступа: http://www.utmn.ru/docs/9317.pdf– Дата обращения: 03.11.2013.
2. *Ермаков, С.А., Ермакова, Л.М.* Методы оценки эмоциональной окраски текста [Текст] / С.А. Ермаков, Л.М. Ермакова // Вестник Пермского университета. – 2012. – Вып. 1(19). – С. 85-89.
3. *Ермаков, А.Е., Киселев, С.Л.* Лингвистическая модель для компьютерного анализа тональности публикаций СМИ [Текст] / А.Е. Ермаков, С.Л. Киселев // Компьютерная лингвистика и интеллектуальные технологии: Диалог 2005. – М., 2005. – С. 312-313.
4. *Лукашевич, Н.В., Четверкин, И.И.* Извлечение и использование оценочных слов в задаче классификации отзывов на три класса [Текст] / Н.В. Лукашевич, И.И. Четверкин // Вычислительные методы и программирование. – 2011. – Т.12. – С. 73-81.

5. *Павлов, А., Добров, Б.* Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры [Текст] / А. Павлов, Б. Добров // Вычислительные методы и программирование. – 2011 – Т. 12. – С. 58-72.
6. *Топтыгина, Е.Н.* О конструктивно-синтаксическом способе выражения субъективной модальности в политическом дискурсе [Текст] / Е.Н. Топтыгина// Политическая лингвистика. – 2011. – Вып. 2(36). – С. 176-179.
7. *Черкасс, М.И.* Понятие тональности в лингвистике [Текст] / М.И. Черкасс // Идеи. Поиски Решения. Т.1. – Минск: БГУ, 2009. – С. 147-149.
8. *Blei, D., Ng, A., Jordan, M.* Latent Dirichlet Allocation [Текст] / D. Blei, A. Ng, M. Jordan // Journal of Machine Learning Research. – 2003. – No.3. – P. 993-1022.
9. *Carenini, G., Ng, R., Zwart, E.* Extracting Knowledge from Evaluative Text [Текст] / G. Carenini, R. Ng, E. Zwart // Proc. of the 3rd International Conference on Knowledge Capture. – 2005. – P. 11-18.
10. *Dittenbach, M.* Scoring and Ranking Techniques - tf-idf Term Weighting and Cosine Similarity [Электронный ресурс] / M. Dittenbach. Электрон. дан. – Режим доступа: //http://www.ir-facility.org/scoring-and-ranking-techniques-tf-idf-term-weighting-and-cosine-similarity. Дата обращения: 10.06.2013 г.
11. *Gamon, M., et al.* Pulse: Mining Customer Opinions from Free Text [Текст] / M. Gamon et al. // Proc. of the 6th International Symposium on Intelligent Data Analysis (IDA). – 2005. – P. 121-132.
12. *Hu, M., Liu, B.* Mining and summarizing customer reviews [Текст] / M. Hu, B. Liu // Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2004. – P. 168-177.

13. *Liu, B.* Sentiment Analysis and Subjectivity [Текст] / B. Liu // Handbook of Natural Language Processing, ed. by N. Indurkhya and F. J. Damerau. – 2010.
14. *Liu, B.* Web Data Mining [Текст] / B. Liu // Exploring Hyperlinks, Contents, and Usage Data. – Springer, 2006.
15. *Hehery, R.* Classification [Текст] / R. Hehery // Machine Learning, Neural and Statistical Classification, ed. by D.Michie et al. – 1994. – P. 6-17
16. *Manning, Ch., Raghavan, P, Schütze, H.* Introduction to Information Retrieval. [Текст] / Ch. Manning, P. Raghavan, H. Schütze. – Cambridge: Cambridge University Press, 2008. – 504 p.
17. *Nasukawa, T., Yi, J.* Sentiment Analysis: Capturing Favorability Using Natural Language Processing [Текст] / T. Nasukawa, J. Yi // Proc. of the 2nd International Conference on Knowledge Capture. – Florida, 2003. – P. 70-77.
18. *Pal, J., Saha, A.* Identifying Themes in Social Media and Detecting Sentiments [Текст] / J. Pal, A. Saha // International Journal of Statistics and Applications. – 2011. – Vol . 1. – No. 1. – P. 14-19.
19. *Pang, B., Lee, L.* Opinion Mining and Sentiment Analysis [Текст] / B. Pang, L. Lee. – 2008. – 135 p.
20. *Salton, G., Wong, A., Yang, C.* A Vector Space Model for Automatic Indexing [Текст] / G. Salton, A. Wong, C. Yang // Communications of the ACM. – 1975. – Vol. 18. – No. 11. – P. 613–620.
21. *Snyder, B., Barzilay, R.* Multiple Aspect Ranking using the Good Grief Algorithm [Текст] / B. Snyder, R. Barzilay // Proc. of the Joint Human Language Technology, North American Chapter of the ACL Conference HLT- NAACL. – 2007. – P. 300-307.
22. *Turney, P.* Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews [Текст] / P. Turney // Proc. of

the 40th Annual Meeting on Association for Computational Linguistics. – 2002. – P. 417-424.

23. *Webb, G., Boughton, J., Wang, Z.* Not So Naive Bayes: Aggregating One-Dependence Estimators [Текст] / G. Webb, J. Boughton, Z. Wang // Machine Learning. – 2005. – No. 58. – P. 5-24.

24. *Wiebe, J., Bruce, R., O'Hara, T.* Development and Use of a Gold-Standard Data Set for Subjectivity Classifications [Текст] / J. Wiebe, R. Bruce, T. O'Hara // Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. – 1999. – P.246-253.