АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ МНЕНИЙ НА ОСНОВЕ СЛОВАРЯ МАРКЁРОВ (ДЛЯ МЕДИЙНЫХ ТЕКСТОВ НА АНГЛИЙСКОЙ И БЕЛОРУССКОМ ЯЗЫКАХ)

OPINION MINING IN MEDIA TEXTS IN ENGLISH AND BELARUSIAN (MARKERS DICTIONARY APPROACH)

В.В. Козлова

V.V. Kazlova

Минский государственный лингвистический университет
Минск, Беларусь

Minsk State Linguistic University
Minsk, Belarus

e-mail: ruis1@yandex.ru

Рассматриваются принципы автоматического определения и сортировки мнений, функционирующих в информационных и аналитических жанрах медийного дискурса на английском и белорусском языках, основанный на словарях языковых единиц, маркирующих высказывания-мнения. Преимущество предложенной модели заключается в том, что она учитывает семантику выявленных единиц, их частотность, а также типологию высказываний-мнения.

The principles of automatic identification and sorting of opinions that operate in the informational and analytical genres of media discourse in English and Belarusian (based on the dictionary approach that include linguistic markers of opinions). The advantage of the proposed model is that it takes into account the semantics of the identified markers, their frequency, as well as the typology of opinions.

Ключевые слова: поиск мнений; словарь языковых маркёров; медийный дискурс; автоматическая сортировка мнений; медиатекст.

Keywords: opinion search; dictionary of opinion markers; media discourse; automatic sorting of opinions; media text.

На сегодняшний день активное развитие получает такое направление автоматической обработки текста и интеллектуального анализа данных, как sentiment analysis and opinion mining – автоматическая идентификация и анализ эмотивности и мнений (L. Lee, B. Pang, P. Turney, H. Tang, J. Wiebe, E. Riloff; Т.П. Карпилович, Т.Н. Гребень, М.В. Чернышевич и др.). Существует несколько методов, которые наиболее частотны и эффективны при проведении подобного рода исследований: метод векторного анализа, метод, основанный на составлении словарей маркёров (тональной, или эмотивной, лексики; единиц выра-

жения мнения), а также гибридный метод, который является смесью названных выше двух методов.

Сутью метода векторного анализа (B. Pang, M. Gamon) является разработка эталонного корпуса текстов, последующая его разметка и сравнение новых текстов с данных эталонным корпусом. На основании полученного результата сравнения исследуемому тексту приписывается положительное или отрицательное значение (по заранее определённой шкале тональности) [1, с. 5]. Преимущество данного подхода заключается в его относительной быстроте применения и обучении программы с каждым введением каждого нового текста. Недостатки же очевидны: метод трудоёмок, корпус текстов разнородный (что приводит к неполноте лексического покрытия), также он более подходит для идентификации тональности высказывания, чем для поиска и анализа мнений, и при этом не позволяет определить эмотивность на уровне целого предложения [2, с. 577].

Метод, основывающийся на составлении словарей эмотивных лексических единиц и единиц, выражающих мнение, подразумевает анализ текстовых контекстов по совокупности найденных в нём маркёров (L. Lee, T. Nasukawa, M. Tsystarau, Т.Н. Гребень). Контекст оценивается по шкале, которая отражает количество релевантных лексических единиц [3]. Данный подход менее трудоёмок, он применим к отдельным предложениям, в соединении с методами лингвистического анализа текста позволяет проводить не только количественный, но и более глубокий семантический и прагматический анализ высказывания [4]. Среди недостатков можно отметить то, что при использовании данного метода достаточно сложно дать количественную позитивную или негативную оценку тональности всего текста. Однако для исследования высказываний мнения данный метод работает достаточно хорошо, поскольку позволяет идентифицировать контексты мнения из текста и проанализировать их.

Следует оговорить, что в проводимом нами исследовании под «контекстом мнения» или «высказыванием мнения» понимается фрагмент текста, содержащий точку зрения некоторого коммуниканта, т.е. фрагмент текста, достаточный для понимания данной точки зрения, мнения по какому-либо вопросу. Так, контекстом мнения может выступать часть предложения (*Пэўныя праблемы могуць узнікнуць і з «камуналкай»* […]» «Звязда»), целое предложение (*Unless we change the way we fund universities, our system will collapse* 'Если только мы не изменим финансирование университетов, наша система рухнет' «The Guardian»), несколько предложений (*But there's no avoiding the real question. Are adult novels larger than children's novels largely because they seek to con-*

front all these issues? Of course they are. 'Но избежать действительно важного вопроса невозможно. Романы для взрослых весомее романов для детей в большей мере потому, что они касаются всех этих проблем? Конечно же, нет' «The Guardian»), абзац или целый текст. И метод, основанный на составлении словарей маркёров позволяет изучать эксплицитные высказывания мнения (т.е. те, в которых фигурируют маркёры). И если брать в качестве материала исследования тексты СМИ, то необходимо отметить, что они насыщены именно эксплицитно выраженными контекстами мнения, а значит в этом случае, метод, основанный на составлении словарей маркёров, является эффективным.

Тем не менее для более точных результатов необходимо сочетать возможности обоих методов. Например, для анализа мнений имеет смысл не только составлять словари маркёров, но также изучать данные единицы (вычленять их и соотносить) в соответствии с определённой предметной областью, конкретным жанром.

Публицистический текст отличается вторичностью к некоторому прототексту. Последний может представляться в качестве текста, так и ситуации. Современные медиа-тексты насыщены контекстами мнения, даже в сфере информационных жанров («Новостная заметка», «Репортаж», «Информационное интервью» и т.д.). При этом наибольшее количество высказываний мнения обнаружено в текстах аналитических жанров «Мнения» и «Комментарии». В ходе исследования рассмотрено более 6300 высказываний мнения (3205 высказываний мнения на английском языке и 2941 на белорусском). Так, среднее количество высказываний, содержащих мнение, в аналитических статьях обозначенных жанров достигает 74 – 87% (данные для белорусского и английского языка, соответственно; Таблица 1). Это могут быть собственно высказывания мнения автора, так и цитаты. Остальная часть текста является либо описанием проблемной ситуации, либо представляет факты.

Таблица 1

Частотность высказываний мнения в информационных и аналитических жанрах медийного дискурса

Параметр	Информационные жанры		Аналити-ческие жанры		Жанры «Мнения» и «Комментарии»	
	анг. яз.	бел. яз.	анг. яз.	бел. яз.	анг. яз.	бел. яз.
Высказывания фактологического характера	41%	60%	15%	23%	13%	26%
Высказывания мнения	59%	40%	85%	77%	87%	74%

Как отмечалось выше высказыванием мнения может служить часть предложения, если в ней фигурирует хотя бы один маркёр и если из этой части предложения можно извлечь семантически завершённое высказывание мнения. В зависимости от типа маркёра, выделяемых на основе контекстуального и логико-семантического анализа, все высказывания мнения мы делим на 3 больших вида (мнения интерпретационного характера, в которых фигурируют такие маркёры, как to mean ‘обозначать’, explanation ‘объяснение’, result ‘результат’, because ‘потому что’, атрымліваецца, тлумачыць, падстава, выснова, бо, калі… то и т.д.; идееполагающего характера с маркёрами should ‘следует’, need ‘необходимо’, idea ‘идея’, problem ‘проблема’, трэба, неабходна, мэта, пажаданне, рэкамендацыя и т.д.; и и прогностического характера – формы будущего времени, условного наклонения и т.д.). Но даже если принимать за контекст мнения часть предложения, всё равно существует ряд примеров, когда невозможно свести контекст к 1 маркёру (или же хотя бы к одному типу маркёров). Такие высказывания мы относим к комбинированному, или гибридному, виду – это около 10% выборки.

Тем не менее для облегчения задачи автоматического распознания мнений более удобно рассматривать именно предложение как минимальное высказывание мнения. При этом в каждом предложении обычно фигурирует 2-3 маркёра. Реже – 1 или 4 и более. Соответственно, если тексты СМИ насыщены эксплицитными контекстами мнения, в каждом из которых функционируют маркёры мнения, мы при приходим к проблеме сортировки выдачи.

Например, пользователь сформулировал запрос: «Что нужно делать, чтобы сохранить белорусские болота?» Можно сделать вывод, что его интересуют мнения идееполагающего характера (не прогностического или интерпретационного характера – их мы сразу исключаем в выдаче при первичной сортировке маркёров). Но автор мог оформить свою точку зрения по-разному. Например, *Балоты трэба ахоўваць* (категоричное мнение; маркёр – модальный глагол с оттенком обязанности – “*трэба*”), или *Ахоўвайце балоты* (также достаточно категоричное мнение; маркёр – глагол в форме императива), или *Я жадаю, каб балоты ахоўвалі* (менее кагоричное мнение, автор делится своими желаниями, идеями, мечтами; маркёр – “*жадаць (каб)*”), или *Пытанне ў наступным: як балоты ахоўваць?* (некатегоричное мнение, выраженное в форме вопроса, также в структуре отмечаем маркёр “*пытанне*”, который лишь указывает на наличие проблемы) И таким образом, для составления алгоритма мы приходим к необходимости присвоить каждому маркёру количественное значение – удельный вес. Он присваивается, исходя из семантического, прагматического и коли-

чественного анализа маркёров, фигурирующих в текстах СМИ (от 1 до 6). Так, модальным глаголам с оттенком обязанности присваивается значение 6, императивам – 5, глаголу "*жадаць*", выражающим желания, идеи, мечты – 4, существительному "*пытанне*" – 3. Вес каждого маркёра сверяется по заранее составленному словарю маркёров мнения в медийных текстах определённых жанров (фрагмент словаря представлен в Таблице 2).

Таблица 2

Фрагмент словаря маркёров мнения для медийных текстов на английском и белорусском языках

Языковые единицы с модальным значением			
(more) likely	5	Shall	6
Can	6	Should	6
Could	6	To be able	4
Have to	6	To be going	4
It is time	4	To be to	4
May	6	(З) магчы / (Не) можа	6
Maybe	6	(не) магчыма / мо	6
Might	6	(ня) хай	6
Must	6	Варта	4
Necessarily / necessary	4	Можна	6
Need	6	Мусіць /мусяць	4
Ought to	6	Неабходна	6
Perhaps	5	Павінна / павінен	6
Possible / possibly	4	Патрэбна / трэба	6
Probably	4		

Таблица 3

Интерпретация значений веса высказывания-мнения

Значение веса	**Интерпретация**
≤ 3	Низкое значение высказывания-мнения
= от 4 до 6	Среднее значение высказывания-мнения
≥ 7	Высокое значение высказывания-мнения

Далее в выдаче необходимо определить удельный вес каждого предложения. Делается это по сумме значений удельного веса маркёров, функционирующих в определённом высказывании мнения. Соответст-

венно, получаем контексты мнения с низким, средним и высоким значением (Таблица 3). Выдача формируется от предложения с наиболее высоким показателем к наименьшему.

Таким образом, метод, основывающийся на заранее составленных словарях маркёров, эффективен для составления принципиального алгоритма поиска мнений в медийных текстах. При этом применение семантического, прагматического и количественного анализа позволяет значительно улучшить результаты по автоматическому анализу высказываний мнения и, соответственно, предоставить более качественную выдачу на запрос пользователя. Эту задачу решает включение в словарь маркёров мнения такого значения, как «удельный вес» каждого маркёра с последующим подсчётом удельного веса каждого высказывания мнения.

Библиографические ссылки

1. Bing Liu. Sentiment Analysis and Opinion Mining. California: Morgan & Claypool Publishers, 2012.

2. Пазельская А.Г. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной междунар. конф. «Диалог». М.: Изд-во РГГУ, 2011.

3. Nasukawa T. Sentiment analysis: capturing favorability using natural language processing // In Proceedings of the 2nd international conference on Knowledge capture, Florida, USA, October 23-25, 2003.

4. Yi J. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques // In Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003), Florida, USA, November 19-22, 2003.

СТРУКТУРНА-ТЭМАТЫЧНЫЯ АСАБЛІВАСЦІ МЕДЫЯПЛАТФОРМЫ «ЖИВОЙ ЖУРНАЛ»

STRUCTURAL AND THEMATIC FEATURES OF THE LIVEJOURNAL MEDIA PLATFORM

Г.Ч. Рыжковіч

G.Ch. Ryzhkovich

Гродзенскі дзяржаўны ўніверсітэт імя Янкі Купалы
Гродна, Беларусь

Yanka Kupala Grodno State University
Grodno, Belarus

e-mail: annamichalouskaya@mail.ru

У артыкуле раскрываюцца структурныя асаблівасці медыяплатформы Живой журнал», апісваюцца асноўныя мадэлі камунікацыі блог-платформы, вызначаюцца асноўныя тэматычныя кірункі ў блогах медыяплатформы.