Автоматическое определение тональности объектов с использованием семантических шаблонов и словарей тональной лексики

Поляков П. Ю (pavel@rco.ru),
Калинина М. В. (kalinina_m@rco.ru),
Плешко В. В. (vp@rco.ru) ООО
«ЭРСИО», Москва, Россия

Ключевые слова: определение тональности, анализ мнений, тональность объектов, тональность атрибутов, синтактико-семантический анализ, семантические шаблоны.

Автоматический объектно-ориентированный Анализ настроений с помощью семантических шаблонов и Словари лексики чувств

Поляков П.Ю (pavel@rco.ru),
Калинина М.В. (kalinina_m@rco.ru),
Плешко В.В. (vp@rco.ru)
ООО «РКО», Москва, Россия

В этой статье изучается использование основанного на лингвистике подхода к автоматическому объектно-ориентированному анализу настроений. Первоначальной задачей было извлечь мнения пользователей (положительные, отрицательные, нейтральные) о телекоммуникационных компаниях, высказанные в твитах и новостях. Мы исключили новости из набора данных, поскольку считаем, что формальные тексты существенно отличаются от неформальных по структуре и лексике и поэтому требуют другого подхода. Мы ограничились лингвистическим подходом, основанным на синтаксическом и семантическом анализе. В этом подходе слово или выражение, несущие настроение, связываются со своим целевым объектом на любом из двух этапов, которые выполняются последовательно. На первом этапе используются семантические шаблоны, соответствующие дереву зависимостей, а на втором этапе используются эвристики для связывания выражений тональности и их целевых объектов, когда между ними не существует синтаксических отношений. Машинное обучение не использовалось. Метод показал очень высокое качество, примерно совпадающее с лучшими результатами методов машинного обучения и гибридных подходов (сочетающих машинное обучение с элементами синтаксического анали

Ключевые слова: анализ настроений, объектно-ориентированный анализ настроений, анализ настроений на основе аспектов, анализ мнений, синтаксический и семантический анализ, семантические шаблоны.

1. Введение

Задача автоматического анализа тональности текстов на естественном языке стала чрезвычайно востребованной. Многие коммерческие компании, производящие товары и услуги, заинтересованы в мониторинге социальных сетей и блогов на предмет мнения пользователей об их продуктах и услугах. Однако до недавнего времени не существовало размеченных корпусов текстов на русском языке, на которых разработчики могли бы протестировать и сравнить качество своих методов. Этот пробел был заполнен ROMIP и более поздними конференциями по анализу настроений SentiRuEval с их треками анализа настроений. Однако задача предыдущих конференций заключалась в выявлении общего настроения текста (например, см. Четверкин И., Браславский П.И., Лукашевич Н. [2]), а на нынешней конференции задача была совершенно новая — предметно-ориентированная. анализ настроений, который сложнее и требует более сложных алгоритмов; так как в случае обнаружения общего настроения важны выбор положительных и отрицательных терминов и определение их весов, в то время как в случае объектно-ориентированного обнаружения настроения также большое значение имеют синтаксические отношения между целевым объектом и словом, выражающим настроение. .

Такой объектно-ориентированный метод для нас не нов; мы уже использовали подобный подход в наших предыдущих исследованиях. Например, мы оценивали сентиментально-ориентированные мнения о марках автомобилей на материале ЖЖблога AUTO_RU (см. описание метода у Ермакова А.Е. [4]). Однако следует отметить, что во всех предыдущих случаях результаты оценивались только нами. Участие в SentiRuEval дало нам возможность провести независимую оценку нашего метода и сравнить наши результаты с результатами других участников.

В этой статье мы представляем результаты применения лингвистического подхода, включающего синтаксический и семантический анализ, к задаче автоматического объектно-ориентированного анализа настроений. Мы ограничились только лингвистическим методом, исключив машинное обучение, потому что было интересно посмотреть, какие результаты даст чисто лингвистический подход без методов машинного обучения.

Задача состояла в том, чтобы найти сентиментально-ориентированные мнения (положительные и отрицательные) о телекоммуникационные компании в твитах.

2. Сопутствующая работа

Обычно объектно-ориентированные или аспектно-ориентированные подходы либо полагаются только на алгоритмы, основанные на статистике, подсчет расстояния до слов, машинное обучение и т. д. для поиска целей мнений (начиная с первой работы по извлечению целей мнений Ху и Лю [5]). ; или они могут использовать неглубокий синтаксический анализ, чтобы сегментировать предложение, найти значимые союзы, отрицания и модификаторы (например, Кан Д. [7]). Другие подходы ищут синтаксическую зависимость между термином тональности и его целью (например, Попеску А., Этциони О. [9]), игнорируя слова, несущие тональность, которые синтаксически не связаны с каким-либо целевым объектом. Отличительной особенностью нашего подхода является то, что с помощью глубокого лингвистического метода мы учитываем не только синтаксически связанные термины тональности (что обеспечивает высокую точность), но и независимые слова и фразы, несущие тональность (что обеспечивает высокую полноту).

Некоторые исследователи пытаются комбинировать статистические и лингвистические методы для достижения наилучших результатов; например, в Якоб Н., Гуревич И. [6] авторы используют, среди прочего, дерево разбора зависимостей для связи выражений мнений и соответствующих целей; и эксперименты показывают, что добавление функции, основанной на пути зависимости, приводит к значительному улучшению их метода. Однако их алгоритм ищет только короткие и прямые отношения зависимости; поэтому у их подхода есть трудности с более сложными предложениями. Более того, они не делают различия между целевым объектом (например, камерой), его атрибутами или частями (например, крышкой объектива, ремешком) и его качествами (например, удобством использования); и, следовательно, они обозначают ближайшее именное словосочетание как цель мнения. Напротив, мы используем очень простую онтологию, чтобы различать целевой объект, атрибуты и качества; и, обнаружив настроение, связанное с атрибутом или качеством, наш алгоритм спускается вниз по дереву разбора зависимостей в поисках целевого объекта. Если он не найден синтаксически, целевой объект ищется с помощью эвристики, основанной на расстоянии предложения. Когда целевой объект найден, тональность, помеченная для его атрибута, присваивается объекту.

3. Методы

Для выполнения задачи мы опирались на наши предыдущие исследования и решения. Подробное описание этих методов можно найти у Ермакова А.Е., Плешко В.В. [3] и Ермакова А.Е. [4]. Новым в подходах, описанных в [3] и [4], было добавление так называемого «обнаружения свободных настроений», которое будет описано в разделе 3.2.

Алгоритм анализа текста имеет следующие этапы в отношении тональности.
задача обнаружения:

1) Токенизация; 2)
Морфологический анализ; 3)
извлечение объекта; 4)
синтаксический анализ; 5)
извлечение фактов (использование семантических шаблонов); 6)
Бесплатное обнаружение настроений.

Этапы 1, 2 и 4 были реализованы стандартными инструментами RCO для общего анализа текста. На третьем этапе больше внимания уделялось объектам, относящимся к данной тематике (названия мобильных компаний, телекоммуникационная терминология и т.п.). Этапы 5 и 6 были основными в задаче определения тональности и поэтому будут подробно описаны.

3.1. Семантические шаблоны

Основной метод анализа настроений заключался в использовании семантических шаблонов.

Семантический шаблон представляет собой ориентированный граф, представляющий собой фрагмент синтаксического дерева с определенными ограничениями, наложенными на его узлы. Синтаксическое дерево предложения содержит семантические и синтаксические отношения между словами, которые определяются синтаксическим анализатором. Ограничения в шаблонах могут применяться к части речи, имени, семантическому типу, синтаксическим отношениям, морфологическим формам и т. д. Извлечение фактов осуществляется путем нахождения в синтаксическом дереве предложения подграфа, изоморфного шаблону (со всеми ограничениями).

Был ис пользован с интакс ичес кий анализатор RCO, ос нованный на подх оде дерева завис имос тей. С емантичес кая с еть, пос троенная с интакс ичес ким парс ером, инвариантна к поря дку с лов и г олос у; например, предложения (1) Оператор укра ла денег с о с чета и (2) Деньг и укра дены оператором с о с чета будут иметь одинаковую с емантичес кую с еть. Такая с емантичес кая с еть предс тавля ет с обой промежуточный уровень предс тавления между с мыс ловой с х емой с итуац ии и ее с ловес ным выражением, т. е. г лубинно-с интакс ичес кое предс тавление, абс траг ированное от поверх нос тног о с интакс ис а.

Нас тройки с емантичес ког о интерпретатора позволя ют отфильтровывать нег ативные и «ненас тоя щие» (импера тивные, ус ловные и т.п.) выс казывания , которые не с ответс твуют реальным с обытия м и не подлежат анализу. В результате такие примеры, как (3) ес ли Билайн будет плох о работать; с еть признаков падает; с вя зь бы обрывалас ь; не Билайн плох о работает может быть ис ключ ен из определения тональнос ти.

Для уменьшения количес тва шаблонов, опис ываю щих с емантичес кие фрейм ы, с ущес твуют так называемые вс помог ательные шаблоны, которые добавля ют в с емантичес кую с еть новые узлы и отношения . В проц ес с е с емантичес ког о анализа и извлечения фактов вс помог ательные шаблоны работаю т раньше вс ех ос тальных шаблонов, так что с емантичес кие шаблоны мог ут опираться я на с еть, пос троенную как с интакс ичес ким анализатором, так и вс помог ательными шаблонами. Например, ес ли мы интерпретируем такие фразы, как (4) X делает Y, X начинает делать Y и (5) X решает делать Y как равные для определенног о с емантичес ког о фрейм а, вмес то с оздания с емантичес ког о шаблона для каждог о примера мы можем иметь один вс помог ательный шаблон, который будет обознача ть подлежащее г лавног о г лаг ола как подлежащее подчиненног о г лаг ола, и один прос той с емантичес кий шаблон — (4) X делает Y.

С емантичес кие шаблоны мог ут иметь так называемые «запрещаю щие узлы», которые накладываю т ог раничения на контекс т, опреде ля я , в каком контекс те шаблон не должен с овпадать. Например, (6) У Билайна надежная с вя зь я вля ет с я положительным утверждением, а добавление наречия ог раничения меня ет ег о значение на противоположное: (7) У Билайна мало надежная с вя зь. С помощью запрещаю щих узлов мы можем различать эти два предложения , утверждая , что прилаг ательное не должно модифиц ировать с я наречием меньше.

Ис пользование запрещаю щих узлов значительно повыша ет точнос ть анализа нас троений.

На рис . 1 показан с емантичес кий шаблон, ис пользуемый для определения нас троения , выраженног о г лаг олом или наречием в таких предложения х , как: (8) Билайн ловит х орошо; Интернет летает.

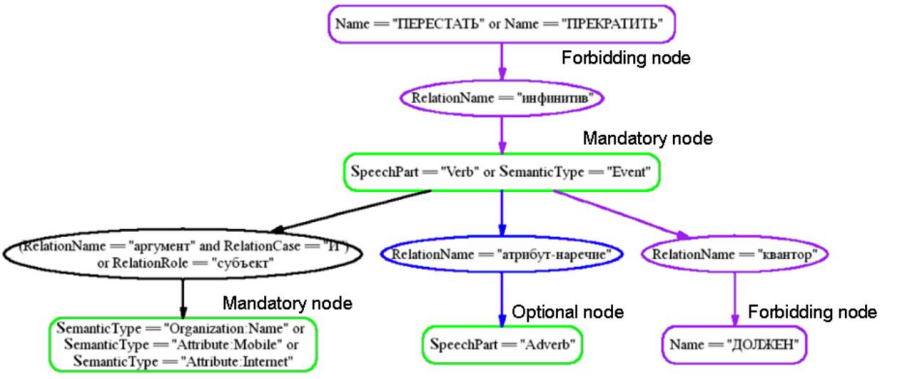

Рис . 1. Пример с емантичес ког о шаблона

Узлы содержат ограничения по части речи (SpeechPart == «Глагол»; SpeechPart == «Наречие»), лексическим единицам (Имя == «ПЕРЕСТАТЬ» или Имя == «ПРЕКРАТИТЬ»), семантическим категориям (SemanticType == «Организация: Имя» или SemanticType == «Атрибут: Мобильный»). Ограничения на семантические и синтаксические отношения между словами включают: имя отношения (RelationName == «аргумент»; RelationName == «квантор»), семантическую роль (RelationRole == «субъект»), падеж (RelationCase == «И»). Для узлов торговых указывается, что глагол должен, выражающий чувство, не должен контролироваться глаголами перестать или прекратить или модифицироваться предикативом. Таким образом, этот шаблон будет соответствовать предложению (8) Билайн хорошо ловит (что является положительным), но не (9) Билайн перестал хорошо ловить (что является отрицательным) или (10) Билайн должен хорошо ловить (что мы считаем нейтральным).

Ограничения семантических шаблонов обогащались за счет использования специальных словарей (так называемых фильтров), содержащих лексику для положительных и отрицательных оценок. Этот словарь включает существительные, прилагательные, глаголы, наречия и словосочетания. Слово из фильтра должно быть синтаксически связано с целью оценки. Подбор терминов для фильтров осуществлялся вручную экспертом-лингвистом. Примеры положительных терминов: супербыстрый, шустро, красота, крутяк, блистать, радовать, обеспечить уверенный прием. Примеры отрицательных терминов: завышенный, препротивнейший, позорище, тормознутость, обдирать, терять соединение, фигово.

Например, набор определенных слов из семантических фильтров применяется к шаблону на рис.1 в качестве ограничений: глаголы или отглагольные существительные параметризуют узел ограничением SpeechPart == «Verb» или SemanticType == «Event»; параметр adverbs задает узел с ограничением SpeechPart == «Наречие», оба эти узла имеют семантическую роль «Оценка».

Конечными объектами оценки были основные российские операторы сотовой связи (Билайн, Мегафон, МТС, Ростелеком, Теле2), но также учитывались оценки пользователями достоинств провайдеров (качество связи, мобильный интернет, сервис и т.д.)..

Анализируя комментарии и мнения пользователей в социальных сетях и на форумах, эксперты определили набор признаков, наиболее часто упоминаемых пользователями мобильных телефонов. Таким образом, был составлен список самого важного для пользователей. Данные атрибуты были разделены на три класса: 1) Мобильные атрибуты — термины, строго связанные с мобильной телефонией: SMS, MMS, 3G, LTE, SIM-карта, роуминг и т. д.; 2) Атрибуты Интернета — термины, строго связанные с Интернетом: Интернет, пинг и т. д.; 3) Общие атрибуты — термины, часто используемые в связи с мобильной телефонией, но которые могут относиться и к другим областям: колл-центр, сигнал, сеть, поддержка клиентов, баланс и т. д. Каждый список дополнен синонимами и вариантами написания (интернет=инет). =инет; lte=лте =lteшечка=лте-шечка; При обнаружении настроения, связанного с определенным атрибутом, данное мнение также приписывалось соответствующему оператору мобильной связи.

На рис.1 узел с ограничением SemanticType == «Organization:Name» или SemanticType == «Attribute:Mobile» или SemanticType == «Attribute:Internet» параметризуется именами мобильных операторов, мобильными атрибутами или интернет-атрибутами; семантическая роль узла - «Цель оценки».

Этот метод обеспечивает очень высокую точность, хотя и не столь высокую полноту.

3.2. «С вободное» нас троение

Х отя ис пользование с емантичес ких шаблонов обес печивает очень х орошую точнос ть, у этог о метода ес ть недос таток — с лово, выражающее нас троение, должно нах одиться в том же предложении, что и объект оц енки, и должно быть с интакс ичес ки с вя зано с ним. Пос кольку это не вс егда так в ес тес твенных текс тах , нек оторые с лучаи я с но выраженног о чув ства будут опущены этим методом, и припоминание пос традает. Эта проблема с тановится ч резвычайно актуальной при анализе неформальных текс тов — форумов, с айтов с оц иальных с етей, блог ов и т. д. При напис ании неформальног о текс тового с ообщения пользователи час то иг норируют пунктуац ионные и орфог рафичес кие правила, опечатываются , из-за чег о с интакс ичес кий анализатор может не правильно проанализировать с труктуру предложения и пос троить с емантичес кую с еть. Пользователи час то выражают с вое мнение ч рез междометия , к оторые не я вля ю тся час тью с интакс ичес ког о дерева; с ледовательно, с емантичес кие шаблоны в этом с лучае бес полезны. Мы называем с лова, к оторые выражают нас троение, но не имеют с интакс ичес ког о отношения к объекту оц енки (или такое отношение не было пос троено с интакс ичес ким анализатором), «с вободным нас троен

Для решения этой проблемы был применен друг ой метод. Мы ис пользовали алг оритм, к оторый ищет с вободную тональнос ть в текс те, ис пользуя с ловари (или профили) позитивной и нег ативной лек с ики, и, ес ли такая тональнос ть най дена, пытается с вя зать ее с ц елевым объектом.

Эти два метода дополня ю т друг друг а, при этом с начала работает метод с емантичес ког о шаблона. В с вя зи с этим клас с ификатор «иг норирует» уже най денные термины, относ я щиеся к ц елевому объекту по шаблонам, пос кольку мы предполаг аем, что точнос ть, обес печиваемая с емантичес кими шаблонами, близка к 100%.

В качес тве профилей для позитивных и нег ативных клас с ов мы ис пользовали с оответс твующие фильтры, у брав контекс тно-завис имые с ентиментальные с лова и ос тавив только экс плиц итно-эмоц иональную или оц еночную лек с ику. Например, мы у далили г лаг олы УМЕРЕТЬ, ПРОИГ РЫВАТЬ, потому что, х отя они я вно отриц ательные в контекс те, например: (11) интернет умер; (12) оператор X проиг рывает оператору Y; но в друг ом контекс те, не с вя занном с мобильной с вя зью , они мог ут быть нейтральными и прос то конс татировать факт. В то же время мы обог атили наши профили междометия ми и друг ими эмоц иональными выражения ми, к оторые не мог ут быть с интакс ичес ки с вя заны с объектом оц енки, например: (13) не надо так! что за нах ; ни фиг а с ебе; ну как так можно и т.д.

Най дя тональнос ть, наш алг оритм ис кал в заданном текс те объект оц енки – название мобильной компании и припис ывал эту тональнос ть мишени. Е сли в текс те у поминалос ь нес колько мобильных операторов, оц енка прис ваивалас ь ближайшему оператору. Е сли были обнаружены как положительные, так и отриц ательные нас троения , с вя занные с одним и тем же у помя нутым оператором мобильной с вя зи, мы отдавали предпочтение отриц ательным нас троения м, с читая положительные выражения с арказмом.

Машинное обучение не ис пользовалос ь. Приме ня емые методы ос новывалис ь только на линг вис тичес ком анализе.

4. Набор данных

У чебно-тес товая коллекц ия , предос тавленная орг анизаторами, с ос тоя ла из 5000 помеченных и 5000 не помеченных твитов, с одержащих с ентиментальные мнения или положительные и отриц ательные факты о телекоммуникац ионных компания х .

Поскольку основной целью анализа тональности в социальных сетях является поиск сентимент-ориентированных мнений, мы помечали тексты, содержащие репринты новостей, и дополнительно измеряли качество определения тональностей для обучающей коллекции, исключая репринты новостей. Мы исключили новостные тексты из окончательного набора данных, потому что считаем, что разница в структуре и лексике между формальными (новости) и неформальными (посты, блоги, твиты) текстами имеет решающее значение. Как правило, в новостных текстах авторы открыто не выражают свое отношение; новости чаще содержат освещение событий и фактов, которые могут быть интерпретированы как позитивные, так и негативные для ньюсмейкера, а не откровенные настроения; поэтому анализ новостей требует другого подхода. Кроме того, лексика неформальных текстов сильно отличается от лексики формальных те

Поэтому мы дополнительно оценили эффективность метода на коллекции с исключенными из набора данных репринтами новостей и пресс-релизами компаний. Поскольку наш метод основан только на лингвистическом анализе, мы не использовали обучающую коллекцию.

5. Результаты

Первоначально для оценки совпадения между ассесорами мы попросили нашего эксперта вручную оценить тестовую коллекцию и отметить каждое упоминание о компании х мобильной связи как положительное, отрицательное или нейтральное. Результаты оценки нашего эксперта представлены в табл. 1. В качестве первичной оценочной метрики использовалась F1-мера, усредненная макро- и микросредой [1]. Кроме того, для удобства в таблицах также представлены полнота и точность. Как видно из таблицы 1, оценка твитов нашим экспертом отличалась от оценки организаторов. Мы считаем оценку, выставленную нашим экспертом, максимально возможной для системы автоматического определения настроений для данной коллекции. Согласие между нашим экспертом и маркировкой организаторов было выше, когда мы исключили новости из набора данных, что подтверждает наше предположение о том, что для анализа настроений новостей следует использовать другой подход.

Таблица 1. Оценка совпадения
между экспертом и оценщиком

	Макро-средний			микросреднее		
	Напомним,	точность F1		Напомним,	точность F1	
С новостями	0,722	0,686	0,703	0,771	0,728	0,749
Без новостей	0,785	0,694	0,737	0,831	0,735	0,780

Результаты всех участников показаны на рис. 1, наши результаты выделены жирными линиями и помечены как «RCO». Интересно, что несколько методов, вероятно, основанных на разных подходах, демонстрируют очень близкие высокие значения F1 (около 0,5), тем не менее, эти значения значительно меньше теоретического максимума, что соответствует совпадению между оценщиками (см. столбцы «Эксперт» на рис. 1).). Это может доказать, что задача автоматического определения тональности все еще остается сложной задачей.

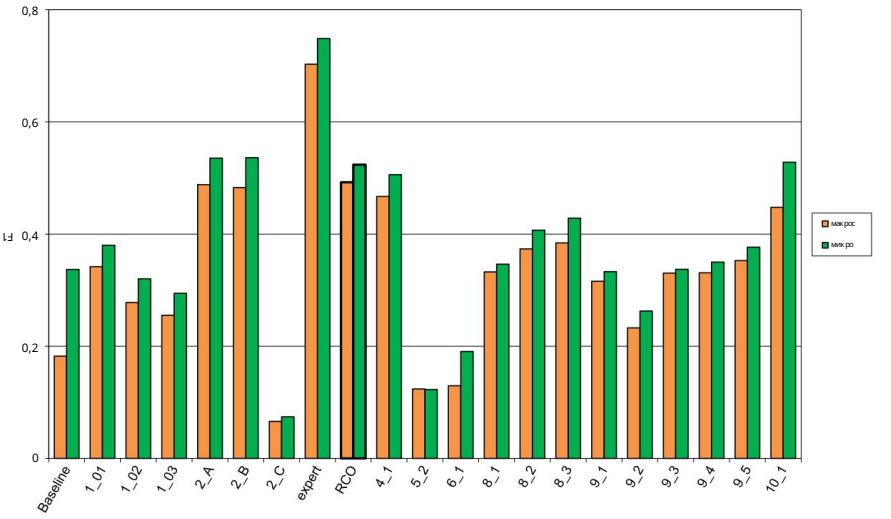

Рис. 2. Макро- и микроусредненные показатели F1, рассчитанные по тестовой коллекции для всех участников. Баллы для нашего метода помечены как «RCO». Баллы экспертной оценки помечены как «эксперт».

Подробные результаты нашего метода представлены в таблице 2. Мы рассчитали полноту, точность и F1 для оригинальной коллекции (помеченной как «С новостями») и для коллекции с исключением сообщений, содержащих новости и пресс-релизы (помеченных как «Без новостей»). . Для сравнения представлены лучшие баллы среди методик всех участников.

Таблица 2. Производительность нашего метода и лучшая мера F1 среди методов всех участников

	Макро-средний			микросреднее		
	Напомним,	точность	F1	Напомним,	точность	F1
С новостями	0,436	0,566	0,480 0,451	0,562	0,585	0,509
Без новостей	0,465	0,492	0,475		0,583	0,524
Лучший результат			0,492			0,536

6. Заключение

Наш комбинированный лингвистический метод показал очень высокое качество, что примерно совпадает с лучшими результатами методов машинного обучения и гибридных подходов (сочетание машинного обучения с элементами синтаксического анализа). В будущем мы планируем добавить машинное обучение к нашему лингвистическому подходу.

Рекомендации

1. Блинов П.Д., Котельников Е.В. (2014), Использование распределенных представлений для основе сентимент-анализа, Диалог '14, Бекасово.
2. Четверкин И., Браславский П.И., Лукашевич Н. (2012), Трек анализа настроений на РОМИП2011, Бекасово.
3. Ермаков А.Е., Плешко В.В. (2009), Абстрактная семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии. 6, стр. 2–7.
4. Ермаков А. Е. Извлечение знаний из текста и их обработка: современное состояние и перспективы // Информационные технологии. 7, стр. 50–55.
5. Ху М., Лю Б. (2004), Анализ и обобщение отзывов клиентов, Международная конференция по обнаружению знаний и анализу данных (ICDM).
6. Якоб Н., Гуревич И. (2010), Извлечение целевых мнений в одно- и междоменной обстановке с условными случайными полями, Материалы конференции по эмпирическим методам обработки естественного языка (EMNLP-2010).
7. Кан Д. (2012), Подход к анализу настроений на основе правил на РОМИП11 8. , Бекасово.
Лукашевич Н., Блинов П., Котельников Е., Рубцова Ю., Иванов В., Тутубалина Е.
(2015), SentiRuEval Тестирование систем объектно-ориентированного анализа настроений на русском языке.
9. Попеску А., Этциони О. (2005), Извлечение характеристик продукта и мнений из обзоров, Материалы конференции по эмпирическим методам обработки естественного языка (EMNLP).
10. Поляков П.Ю, Калинина М.В., Плешко В.В. (2012), Исследование применимости тематической классификации к проблеме классификации рецензий. Диалог '12.
Наро-Фоминск.
11. Поляков П. Ю, Фролов А. В., Плешко В. В. (2013), Использование семантических категорий в приложении к сентимент-анализу рецензий на книги, Диалог '13, Бекасово.