УДК 004.912

ОПЫТ ПОСТРОЕНИЯ СИСТЕМЫ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ОБЪЕКТОВ НА ОСНОВЕ СИНТАКТИКО-СЕМАНТИЧЕСКОГО АНАЛИЗАТОРА

П.Ю. Поляков[1], М.В. Калинина[2], В.В. Плешко[3]

ООО «ЭР СИ О»

[1] pavel@rco.ru, [2] kalinina_m@rco.ru, [3] volodia@rco.ru

Аннотация

Исследуется применение лингвистического подхода для решения задачи автоматического определения тональности объекта. Исследование проводилось в рамках цикла тестирования систем автоматического анализа тональности SentiRuEval. Задание, предложенное организаторами дорожки, заключалось в том, чтобы определить мнение пользователя (положительное, отрицательное или нейтральное) по отношению к операторам сотовой связи на материале сообщений социальной сети Twitter и новостей. Авторы настоящей работы исключили новостные сообщения из тестовой коллекции, так как формальные тексты существенно отличаются от неформальных по своей структуре и лексике и, следовательно, требуют другого подхода. При решении поставленной задачи был использован лингвистический метод, основанный на синтактико-семантическом анализе. Согласно этому подходу тональная лексика привязывается к объекту на одной из двух последовательных стадий. Первая стадия включает в себя использование семантических шаблонов, которые сравниваются с деревом синтаксического разбора предложения; вторая стадия использует эвристики для связывания тональной лексики с объектом оценки в случае, когда синтаксические связи между ними отсутствуют. Машинное обучение не применялось. Метод продемонстрировал очень хорошие результаты, которые примерно совпадают с лучшими результатами методов с использованием машинного обучения и гибридных методов.

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

1. ВВЕДЕНИЕ

Задача автоматического определения тональности в текстах на естественном языке в настоящее время является весьма актуальной. Многие производители товаров и услуг заинтересованы в мониторинге социальных сетей и блогов на предмет выявления отзывов потребителей. Тем не менее, до недавнего времени не существовало размеченного корпуса текстов на русском языке, с помощью которого разработчики могли бы тестировать свои решения и оценивать их качество. Данный пробел были призваны восполнить дорожки семинара РОМИП, а позже SentiRuEval, по автоматическому определению тональности. Однако задания дорожек предыдущих семинаров заключались в определении общей тональности текста (см., к примеру, [1]), в то время как на SentiRuEval 2015 постановка задачи была принципиально новой: определение тональности объекта. Данная задача является более сложной и требует более высокоточных алгоритмов, так как в случае определения общей тональности текста важно только соотношение положительно и отрицательно окрашенных терминов в тексте, в то время как при определении тональности объекта большое значение также имеет синтаксическая связанность объекта с тональной лексикой.

Объектно-ориентированный подход не является новым для авторов данной работы: подобный метод уже применялся в предыдущих исследованиях. В частности, была разработана экспериментальная автоматизированная система для анализа положительных и отрицательных отзывов об автомобилях и классификации их: «за что хвалят/ругают?» [2]. Для извлечения знаний было выбрано крупнейшее из нескольких десятков автомобильных сообществ «Живого журнала» – сообщество AUTO_RU «Все об автомобилях». Для оценки высказываний об автомобилях с точки зрения характеристик их потребительских свойств (положительная/отрицательная) была разработана экспериментальная онтология, содержащая:

1. более 700 различных наименований марок автомобилей и фирм-производителей; первоначальный список был расширен синонимами с вариантами

написаний; символьно-цифровые модели, упоминаемые в тексте (*BMW 325i, ВАЗ 21053*), не включались в словарь, а распознавались по формальным правилам;

2. более 1200 терминов в 24 группах, среди которых:

- 211 наименований узлов автомобиля (*движок, коробка передач, ходовая часть*);

- 71 наименование свойств (*ходовые качества, комфорт, безопасность, надежность* и т. д.);

- 882 наименования оценок характеристик узлов и свойств, включающие прилагательные, существительные, глаголы и наречия (*крутой, поломка, глючить, отстойно*);

- 37 эмоциональных характеристик (*любить, жалоба, плеваться*);.

3. около 100 семантических шаблонов, описывающих возможные синтаксические связи в предложении между группами терминов из онтологии.

Автоматизированная разработка онтологии проводилась на базе анализа языкового материала сообщества AUTO_RU «Живого журнала» при помощи средств компьютерного анализа текста. В итоге из 500 000 сообщений (60 Мбайт текста) было извлечено всего более 5000 оценок автомобилей, их узлов и характеристик, из которых более 1000 (795 хороших и 328 плохих) оценок было привязано к маркам автомобилей, а более 4000 оценок узлов и характеристик программе не удалось привязать к конкретным маркам. В результате была достигнута точность 84%, а полнота извлечения около 20%.

В настоящем исследовании были учтены ошибки и проблемы предыдущих работ, и метод семантических шаблонов был дополнен методом, позволяющим учитывать характеристики, не привязанные синтаксически к объекту интереса; что значительно повысило полноту анализа.

Следует также упомянуть, что во всех предыдущих случаях авторы оценивали полученные результаты самостоятельно. Участие в семинаре SentiRuEval дало возможность получить независимую оценку настоящего метода и сравнить результаты с другими участниками.

В данной работе представлены результаты применения лингвистического метода, включающего в себя синтактико-семантический анализ (также в литера-

туре используется термин «семантико-синтаксический анализ»), к задаче автоматического определения тональности объекта. Поставленная задача заключалась в выявлении мнения пользователей (положительного и отрицательного) по отношению к операторам сотовой связи на материале сообщений социальной сети Twitter. При решении данной задачи авторы настоящей работы ограничились лингвистическим методом, исключив машинное обучение, так как было интересно посмотреть, какие результаты даст чисто лингвистический подход.

2. ИСТОРИЯ ВОПРОСА

Как правило, методы выявления тональности по отношению к объекту или его характеристикам при нахождении объекта оценки опираются либо на исключительно статистические алгоритмы, расстояние в словах, машинное обучение и т. п. (начиная с первой работы по определению тональности объектов [3]), либо используют элементы поверхностно-синтаксического анализа для сегментации предложения, нахождения значимых союзов, отрицания и модификаторов (например, [4]). В рамках других подходов ищутся синтаксические связи между тонально окрашенным термином и объектом оценки (например, [5]), но упускается из рассмотрения тональная лексика, синтаксически не связанная с целевым объектом. Отличительной особенностью настоящего подхода является то, что при применении глубокого лингвистического анализа учитываются не только синтаксически связанные с объектом тональные слова (что обеспечивает высокую точность), но и независимая тонально окрашенная лексика и фразы (что дает высокую полноту).

Некоторые исследователи пытаются совмещать статистические и лингвистические методы для достижения лучших результатов, например, в [6] авторы, среди прочего, используют дерево синтаксических зависимостей для связывания лексики, выражающей мнение, с объектами оценки; как показывают эксперименты, учет синтаксических связей значительно повышает показатели их метода. Однако их алгоритм ищет только прямой и кратчайший путь в дереве зависимостей, таким образом, данный метод испытывает затруднения при анализе более длинных и сложных предложений. Кроме того, авторы не делают различий между объектом оценки (например, *фотокамера*), его составляющими частями

(например, *линза, ремешок*) и его характеристиками (например, *удобство применения*); и, следовательно, помечают ближайшую именную группу в качестве объекта оценки. В отличие от данного подхода в настоящей работе используется элементарная онтология для разделения объекта оценки, его составных частей и качеств; и при выявлении в тексте оценки атрибута или качества алгоритм продолжает поиск конечного целевого объекта, идя по дереву зависимостей. Если не удается установить конечный объект оценки синтаксическим путем, его поиск продолжается при помощи эвристик, основанных на учете расстояния в клаузах. При нахождении конечного объекта оценки тональность, приписанная его атрибуту, переносится на данный объект.

3. МЕТОДЫ

При выполнении поставленной задачи был учтен опыт более ранних исследований авторов данной статьи и использованы имеющиеся наработки. Подробное описание методов можно найти в работах [7] и [2]. Новизной по отношению к описанным методам был учет так называемой свободной тональности, подробнее о которой рассказывается в пункте 3.2.

Алгоритм анализа текста применительно к задаче определения тональности имеет следующие этапы:

1) токенизация – разбиение текста на абзацы, предложения, токены; определение типа токенов (русское слово, латинское слово, знак препинания, специальная конструкция);

2) морфологический анализ – определение грамматических характеристик слова (часть речи, падеж, число, род, лицо и т. д.). Основной словарь содержит: 110 тыс. слов (52 тыс. существительных, 24 тыс. глаголов, 33 тыс. прилагательных, остальное – наречия, служебные, наименования, имена, фамилии, география), 743 приставки для правил точного анализа неизвестных слов, 162 окончания для правил точного анализа неизвестных слов. Дополнительный словарь содержит 27 тыс. фамилий и 23 тыс. имен. Неизвестные слова анализируются в приближенной морфологии по правилам на известные приставки/окончания и на основе частоты суффиксов и окончаний известных слов. Подробнее с описанием морфоанализатора можно ознакомиться в [8];

3) извлечение объектов интереса – в тексте по общим правилам, опирающимся на морфологию и ключевые слова, выделяются имена персон, названия организаций и географические наименования; происходит поиск референтных упоминаний объектов, устанавливается кореферентность и анафорические связи, отождествляются упоминания одного и того же объекта в разных местах текста; идентифицируются объекты, описанные в формате XML по специальным правилам. Подробнее см. в [9];

4) синтаксический анализ – синтаксический разбор предложения в терминах дерева зависимостей, установление синтактико-семантических связей между словами и их ролей (субъект, объект, предикат и т. д.);

5) извлечение фактов (применение семантических шаблонов) – поиск в синтактико-семантической сети разбора предложения такой подсети, которая изоморфна шаблону, с заполнением слотов соответствующего фрейма именами участников ситуации из текста в соответствии с ролями, указанными в узлах шаблона [7];

6) поиск свободной тональности, привязка ее к объектам интереса.

Этапы 1, 2 и 4 были реализованы с помощью стандартных инструментов анализа текста, входящих в состав RCO Fact Extractor SDK (подробнее см. [10]). На этапе 3 особое внимание было уделено описанию объектов, представляющих интерес в рамках решаемой задачи (названия мобильных операторов, телекоммуникационная терминология и т. д.). Этапы 5 и 6 являются основными для решения задачи определения тональности и, следовательно, будут описаны подробно далее.

3.1. Семантические шаблоны

Основной метод автоматического определения тональности включал в себя использование семантических шаблонов. Семантический шаблон – это ориентированный граф, представляющий собой фрагмент дерева синтаксической зависимости с ограничениями, наложенными на его вершины. Дерево синтаксического разбора предложения содержит синтактико-семантические связи между словами, которые определяются синтаксическим анализатором. Ограничения в узлах шаблона могут накладываться на часть речи, имя сущности, семантический тип,

синтаксическую связь, морфологическую форму и т. д. Поиск фактов осуществляется путем поиска подграфа в дереве синтаксической зависимости, совпадающего с шаблоном (с учетом всех ограничений).

Для имплементации синтактико-семантического анализа использовался синтаксический парсер RCO, основанный на грамматике зависимостей. Семантическая сеть, построенная анализатором, инвариантна к порядку слов и залогу глагола; например, предложения (1) *Оператор украл деньги со счета* и (2) *Деньги украдены оператором со счета* будут иметь одинаковое представление. Подобная семантическая сеть представляет собой промежуточный уровень представления между собственно семантической схемой ситуации и ее конкретным языковым выражением, т. е. представлением глубинно-синтаксического уровня, абстрагированным от особенностей поверхностного синтаксиса.

Настройки семантического интерпретатора позволяют отфильтровывать отрицание и «ирреальные» предложения (повелительное, сослагательное наклонения и т. д.), которые не соответствуют реальным событиям и фактам, и потому не представляют интереса для фактографического анализа. Как результат, примеры вроде: (3) *если Билайн будет плохо работать; сеть якобы падает; связь бы обрывалась; не Билайн плохо работает* можно исключать из анализа тональности.

Для сокращения количества шаблонов, описывающих семантические фреймы, имеются так называемые служебные шаблоны, которые добавляют новые узлы и связи в синтактико-семантическую сеть. В процессе семантического анализа и извлечения фактов служебные шаблоны срабатывают в первую очередь, и, таким образом, семантические шаблоны опираются на сеть, построенную синтаксическим анализатором и модифицированную служебными шаблонами. Например, если мы интерпретируем высказывания: (4) *X делает Y,* (5) *X начинает делать Y* и (6) *X решил сделать Y* как тождественные в рамках описания определенной ситуации, вместо того, чтобы создавать семантический шаблон для каждого из этих примеров, можно написать один служебный шаблон, который будет маркировать субъект вспомогательного глагола как субъект смыслового глагола, и один простой семантический шаблон вида: (4) *X делает Y.*

Семантические шаблоны могут иметь так называемые запрещающие вершины, которые накладывают ограничения на контекст, определяя, при наличии

какого контекста шаблон не должен срабатывать. Например, высказывание (7) *У Билайна надежная связь* выражает позитивную характеристику объекта, в то время как добавление наречия *наименее* меняет его тональность на противоположную: (8) *У Билайна наименее надежная связь*. С помощью запрещающих вершин можно делать различия между двумя этими высказываниями, добавив условие, что прилагательное, выражающее оценку, не должно быть модифицировано наречием *наименее*. Использование запрещающих вершин помогает значительно повысить точность определения тональности.

На Рис. 1 представлен семантический шаблон, который используется для определения тональности объекта, выраженной глаголом или наречием, в предложениях вида: (9) *Билайн ловит хорошо;* (10) *Интернет летает*.

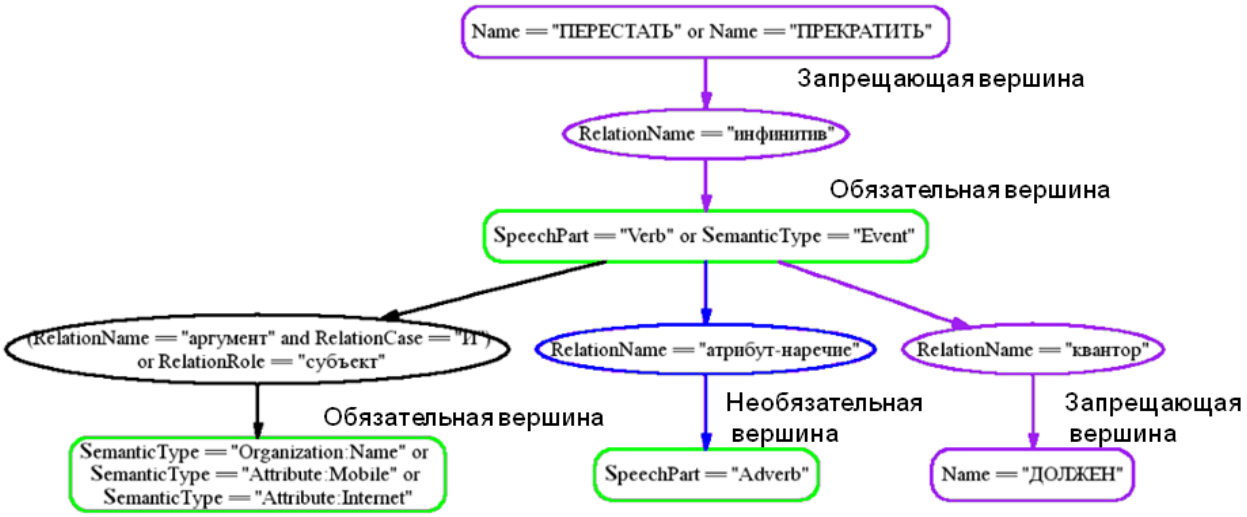


Рис. 1. Пример семантического шаблона

Узлы шаблона содержат ограничения на часть речи (*SpeechPart == "Verb"; SpeechPart == "Adverb"*), конкретные слова (*Name == "ПЕРЕСТАТЬ" or Name == "ПРЕКРАТИТЬ"*), семантические категории (*SemanticType == "Organization:Name" or SemanticType == "Attribute:Mobile"*). Ограничения на синтактико-семантические связи между словами включают в себя: RelationName – тип синтактико-семантической связи между узлами (*RelationName == "аргумент"; RelationName == "квантор"*), RelationRole – семантическую роль (*RelationRole == "субъект*"), RelationCase – падеж (*RelationCase == "И"*). Запрещающие вершины говорят о том, что

глагол, выражающий оценку, не должен быть подчинен фазисным глаголам *перестать* или *прекратить* и не должен быть модифицирован предикативом *должен*. Таким образом, шаблон сработает на предложении (9) *Билайн хорошо ловит* (которое выражает положительную оценку), но не сработает на примерах (11) *Билайн перестал хорошо ловить* (выражает негативную оценку) и (12) *Билайн должен хорошо ловить* (оценка не определена).

Ограничения, накладываемые на узлы семантических шаблонов, были дополнены специальными словарями (фильтрами), содержащими лексику, выражающую позитивную или негативную оценку. Данные словари содержат существительные, прилагательные, глаголы, наречия, а также словосочетания. Термин из фильтра должен быть синтаксически связан с объектом оценки. Отбор лексики для фильтров производился вручную лингвистом-экспертом. Примеры положительных терминов: *супербыстрый, шустро, красота, крутяк, блистать, радовать, обеспечивать уверенный прием*. Примеры отрицательных терминов: *завышенный, препротивнейший, позорище, тормознутость, обдирать, терять соединение, фигово.*

Например, в качестве ограничений семантические фильтры накладываются на следующие узлы шаблона на Рис.1 : глаголы и отглагольные существительные параметризуют вершину с ограничением: *SpeechPart == "Verb" or SemanticType == "Event"*; наречия параметризуют вершину с ограничением: *SpeechPart == "Adverb"*; обе эти вершины имеют семантическую роль «Оценка».

Целевыми объектами оценки являлись крупнейшие российские операторы сотовой связи (Билайн, Мегафон, МТС, Ростелеком и Tele2), но также учитывалась оценка пользователями атрибутов сотовых операторов (качество связи, мобильный интернет, абонентская поддержка и т. п.).

Анализируя отзывы пользователей в социальных сетях и на форумах, эксперты определили набор атрибутов, на которые пользователи наиболее часто обращают внимание. Таким образом был составлен список наиболее важных для пользователей характеристик. Данные атрибуты были поделены на три класса: 1) атрибуты мобильной связи – термины, имеющие отношение исключительно к мобильной телефонии: *SMS, MMS, 3G, LTE, SIM-карта, роуминг* и т. д.; 2) интернет-атрибуты – термины, имеющие отношение исключительно к интернету (любому):

интернет, пинг и т. д.; 3) общие атрибуты – термины, часто используемые в связи с мобильной телефонией, но могущие описывать и другие предметные области: *колл-центр, сигнал, сеть, техподдержка, баланс* и т. д. Каждый из трех классов был пополнен синонимами и вариантами написания (*интернет=инет=и-нет; lte=лте =lтешечка =лте-шечка; баланс счета=состояние счета=средства на счету=деньги на счету* и т. п.). При нахождении в тексте оценки атрибута данная оценка переносилась также на соответствующего мобильного оператора.

На Рис. 1 вершина с ограничением *SemanticType == "Organization:Name" or SemanticType == "Attribute:Mobile" or SemanticType == "Attribute:Internet"* параметризуется названиями сотовых операторов, атрибутов мобильной связи или интернет-атрибутами; данная вершина имеет роль «Объект оценки».

Описанный метод обеспечивает очень высокую точность, однако его полнота оставляет желать лучшего.

3.2. «Свободная» тональность

Хотя использование семантических шаблонов дает очень высокую точность, применение этого метода имеет определенные ограничения: слово, выражающее оценку, должно быть в том же предложении, что и объект оценки, и должно быть синтаксически с ним связано. Так как в реальных текстах дело далеко не всегда обстоит таким образом, некоторые случаи явно выраженной тональности будут упущены данным подходом, и полнота пострадает. Особенно ощутима эта проблема при анализе неформальных текстов – сообщений форумов, социальных сетей, блогов и т. д. При написании неформальных сообщений пользователи часто пренебрегают правилами орфографии и пунктуации, делают опечатки, из-за чего синтаксический анализатор может ошибаться при построении связей в предложении, или синтактико-семантическая сеть может вовсе развалиться. Кроме того, пользователи могут выражать свои эмоции через междометия, которые не являются частью синтаксической структуры дерева зависимостей, и, следовательно, не могут быть выловлены при помощи семантических шаблонов. Термины, которые выражают оценку, но синтаксически не связаны с объектом оценки (или анализатор не смог построит такую связь), в рамках данной работы получили условное название «свободная тональность».

Для решения проблемы свободной тональности был применен подход, основанный на алгоритме, ищущем в тексте тонально окрашенную лексику, опираясь на словари (или профили) положительной и отрицательной лексики, и, если такая лексика найдена, пытающемся привязать ее к объекту оценки.

Оба этих метода (семантических шаблонов и свободной тональности) дополняют друг друга, работая последовательно, при этом метод семантических шаблонов работает первым. Алгоритм поиска свободной тональности «игнорирует» тональную лексику, уже привязанную к объекту шаблонами, так как предполагается, что точность, обеспечиваемая шаблонами, близка к стопроцентной.

В качестве профилей, содержащих положительную и отрицательную тональные лексики, были использованы соответствующие фильтры с небольшой модификацией: были удалены контекстно зависимые термины, и оставлена только явная оценочная или эмоциональная лексика. Например, были убраны глаголы *УМЕРЕТЬ, ПРОИГРЫВАТЬ*, так как, хотя они, бесспорно, выражают негативную оценку в следующих примерах: (13) *интернет умер;* (14) *оператор X проигрывает оператору Y*; в другом контексте, не связанном с мобильной телефонией, они могут не иметь оценочного значения, а просто констатировать факт. Одновременно с этим профили были пополнены междометиями и устойчивыми экспрессивными выражениями, которые нельзя синтаксически привязать к объекту оценки, например: (15) *не надо так! что за нах; ни фига себе; ну как так можно* и т. п.

При обнаружении тонально окрашенных терминов алгоритм ищет в данном тексте объект оценки – название сотового оператора – и приписывает ему тональность. В случае, если в тексте упоминается несколько сотовых операторов, оценка приписывается ближайшему из них. В случае, если в одном тексте обнаружены и положительные, и отрицательные термины, относящиеся к одному оператору, предпочтение отдавалось отрицательной оценке, так как предполагалось, что позитивная лексика в данном контексте выражает сарказм.

В рамках решения данной задачи машинное обучение не применялось. Описанные методы опирались исключительно на лингвистический анализ.

4. ТЕСТОВАЯ ВЫБОРКА

Обучающая и тестовая выборки, предоставленные организаторами, состояли из 5000 размеченных и 5000 неразмеченных сообщений социальной сети Twitter, содержащих оценочные суждения пользователей либо положительные или отрицательные информационные поводы, касающиеся сотовых операторов.

Так как основной задачей определения тональности в социальных сетях является выявление мнения пользователей, были отобраны сообщения, содержащие перепечатки новостей, после чего было дополнительно измерено качество определения тональности на тестовой выборке без новостных сообщений. В результате новостные сообщения были исключены из итоговой тестовой коллекции, так как разница в синтаксической структуре и лексике между формальными (новостными) и неформальными (посты, блоги, твиты) текстами представляется принципиальной. Как правило, авторы новостных сообщений явно не выражают свое отношение: новости содержат простое описание событий и фактов, которые можно трактовать как положительный или отрицательный информационный повод по отношению к объекту интереса, в том время как явное оценочное суждение в них не содержится. Кроме того, лексическое наполнение неформальных сообщений значительно отличается от лексики, встречающейся в формальных текстах. Следовательно, анализ новостных сообщений требует другого подхода.

С учетом вышесказанного было дополнительно оценено качество работы настоящего метода после исключения из тестовой коллекции перепечаток новостей и пресс-релизов компаний. Так как данный метод основан исключительно на лингвистическом подходе, не было необходимости использовать обучающую выборку.

РЕЗУЛЬТАТЫ

Для начала, с целью оценки уровня согласованности между экспертами, наш эксперт разметил тестовую коллекцию вручную и пометил каждое упоминание сотового оператора как позитивное, негативное или нейтральное. Результаты оценки эксперта представлены в Таблице 1. Качество результатов оценивалось с помощью F1-меры с макро- и микроусреднением. Дополнительно, для наглядности, в таблицах приведены точность и полнота. Как видно из Таблицы 1, разметка сообщений нашим экспертом отличалась от разметки организаторов. Цифры, по-

лученные нашим экспертом, представляются максимально возможным результатом для системы автоматического определения тональности на данной коллекции. Согласованность между нашим экспертом и разметкой организаторов стала выше, когда из выборки были исключены новости, что подтверждает предположение о том, что для анализа тональности информационных поводов требуется иной подход.

Таблица 1. Оценка согласованности между нашим экспертом и организаторами

Macro-average			**Micro-average**		
Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
0.785	0.694	0.737	0.831	0.735	0.780

Подробные результаты применения настоящего метода представлены в Таблице 2. Для сравнения в таблицу включен лучший из всех результатов, показанных участниками дорожки. Результат, полученный настоящим методом, оказался одним из лучших.

Таблица 2. Результат оценки настоящего метода и лучшая F1-мера среди методов всех участников

	Macro-average			**Micro-average**		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
RCO	0.465	0.562	0.492	0.475	0.583	0.524
Лучший результат			0.492			0.536

Интересно отметить, что несколько методов, основанных на различных подходах (полностью машинное обучение, гибридный подход – машинное обучение с элементами синтаксиса), показали очень близкую величину F1 – около 0.5 (подробнее о сравнении с результатами других участников см. [11]); и, тем не менее, эти результаты значительно ниже теоретического максимума, который соответствует уровню согласованности между экспертами (см. Таблицу 1). Данный факт

служит лишним подтверждением того, что автоматическое определение тональности до сих пор является трудной и интересной задачей.

ЗАКЛЮЧЕНИЕ

Описанный лингвистический подход продемонстрировал очень высокое качество, которое примерно соответствует лучшим результатам, показанным методами с использованием машинного обучения и гибридными методами (сочетающими в себе машинное обучение с элементами синтаксического анализа).

В будущем планируется дополнить лингвистический подход методами машинного обучения:

- в части генерации словарей тональной лексики;
- в части генерации шаблонов;
- в части генерации правил связывания объектов с атрибутами и отнесения свободной тональности.

Также планируется провести оценку полноты и точности определения тональности по отдельности для каждого способа выражения мнения пользователя об объекте и учитывать вес достоверности определения тональности.

СПИСОК ЛИТЕРАТУРЫ

1. *Четверкин И.И., Браславский П.И., Лукашевич Н.В.* Дорожки по анализу мнений на РОМИП // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’ 2012. Бекасово, 2012.

2. *Ермаков А.Е.* Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии. 2009. № 7. С. 50-55.

3. *Hu M., Liu B.* Mining and summarizing customer reviews // International Conference on Knowledge Discovery and Data Mining (ICDM), 2004.

4. *Kan D.* Rule-based approach to sentiment analysis at ROMIP’11 // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’ 2012. Бекасово, 2012.

5. *Popescu A., Etzioni O.* Extracting product features and opinions from reviews // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2005.

6. *Jakob N., Gurevych I.* Extracting opinion targets in a single-and cross-domain setting with conditional random fields // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010), 2010.

7. *Ермаков А.Е., Плешко В.В.* Семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии. 2009. № 6. С. 2-7.

8. *Ермаков А.Е., Плешко В.В.* Компьютерная морфология в контексте анализа связного текста // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. М.: Наука, 2004.

9. *Ермаков А.Е., Плешко В.В., Митюнин В.А.* RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов. Москва, 2003. URL: http://www.rco.ru/?p=4599.

10. RCO Fact Extractor SDK (Rus.), URL: http://www.rco.ru/?page_id=3554.

11. *Поляков П.Ю., Калинина М.В., Плешко В.В.* Автоматическое определение тональности объектов с использованием семантических шаблонов и словарей тональной лексики // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2015. Москва, 2015.

EXPERIMENT IN BUILDING AN AUTOMATIC OBJECT-ORIENTED SENTIMENT DETECTION SYSTEM BASED ON THE SYNTACTIC AND SEMANTIC ANALYZER

P.Yu. Polyakov[1], M.V. Kalinina[2], V.V. Pleshko[3]

RCO
[1] pavel@rco.ru, [2] kalinina_m@rco.ru, [3] volodia@rco.ru

Abstract

This paper focuses on the use of a linguistics-based method for automatic object-oriented sentiment analyses. The study was conducted as part of SentiRuEval automatic sentiment analysis system testing cycle. The original task was to extract users'

opinions (positive, negative, neutral) about telecom companies, expressed in tweets and news. In this study news was excluded from the dataset because, being formal texts, news significantly differs from informal ones in its structure and vocabulary and therefore demands a different approach. Only linguistic approach based on syntactic and semantic analysis was used. In this approach, a sentiment-bearing word or expression is linked to its target object at either of two stages, which perform successively. The first stage includes usage of semantic templates matching the dependence tree, and the second stage involves heuristics for linking sentiment expressions and their target objects when syntactic relations between them do not exist. No machine learning was used. The method showed a very high quality, which roughly coincides with the best results of machine learning methods and hybrid approaches.

Keywords*: sentiment analysis, object-oriented sentiment analysis, aspect-based sentiment analysis, opinion mining, syntactic and semantic analysis, semantic templates*

REFERENCES

1. *Chetviorkin I.I., Braslavski P.I., Loukachevitch N.V*. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: International Conference Dialog-2011. 2011. P. 739-746.

2. *Ermakov A.E.* Izvlechnie znanii iz teksta i ih obrabotka: sostoyanie I perspektivy // Informacionnye tehnologii. 2009. № 7. P. 50-55.

3. *Hu M., Liu B.* Mining and summarizing customer reviews // International Conference on Knowledge Discovery and Data Mining (ICDM), 2004.

4. *Kan D.* Rule-based approach to sentiment analysis at ROMIP'11 // Computational linguistics and intellectual technologies: proceedings of International conference Dialog-2012, 2012.

5. *Popescu A., Etzioni O.* Extracting product features and opinions from reviews // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2005.

6. *Jakob N., Gurevych I.* Extracting opinion targets in a single-and cross-domain setting with conditional random fields // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010), 2010.

7. *Ermakov A.E., Pleshko V.V.* Semanticheskaya interpretaciya v sistemah komp'yuternogo analysa teksta // Informacionnye tehnologii. 2009. № 6. S. 2-7.

8. *Ermakov A.E., Pleshko V.V.* Komp'yuternaya morfologia v kontekste analysa svyaznogo teksta // Computational Linguistics and Intellectual Technologies: International Conference Dialog-2004. 2004.

9. *Ermakov A.E., Pleshko V.V., Mityunin V.A.* RCO Pattern Extractor: komponent vydeleniya osobih ob'ektov v tekste // Informatizaciya i informacionnaya bezopasnost pravoohranitelnih organov: XI Mezdunarodnaia nauchnaya konferenciya. Sbornik trudov. Moskva, 2003. URL: http://www.rco.ru/?p=4599.

10. RCO Fact Extractor SDK (Rus.), URL: http://www.rco.ru/?page_id=3554.

11. *Polyakov P. Yu., Kalinina M.V., Pleshko V.V.* Avtomaticheskoe opredelenie tonalnosti ob'ektov s ispolzovaniem semanticheskih shablonov I slovarei tonalnoi leksiki // Computational Linguistics and Intellectual Technologies: International Conference Dialog-2015. 2015.

СВЕДЕНИЯ ОБ АВТОРАХ

ПОЛЯКОВ Павел Юрьевич – ведущий программист компании ООО «ЭР СИ О» (RCO), аспирант Остравского технического университета.

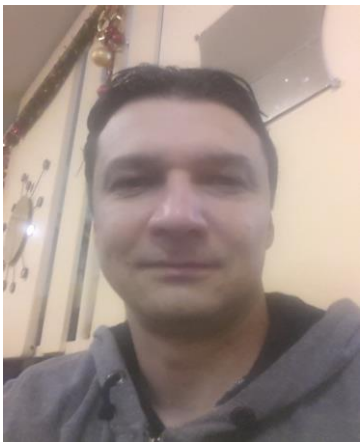
Pavel Yurjevich POLYAKOV, received Master's degree in applied physics and mathematics from Moscow Institute Physics and Technology (2004). Currently is a lead programmer at RCO LLC and PhD student at the Technical University of Ostrava. Current scientific interests: text mining, computational linguistics, knowledge extraction technologies, data mining, artificial intelligence, Boolean factor analysis, recurrent neural networks.

e-mail: pavel@rco.ru

КАЛИНИНА Мария Викторовна – ведущий лингвист компании ООО «ЭР СИ О» (RCO).

Maria Viktorovna KALININA, received Master's degree in theoretical and applied linguistics from Lomonosov Moscow State University (2003). Currently is a lead linguist at RCO LLC, a leading company of the Russian market in the field of computational linguistics and processing of unstructured information. Current scientific interests: data mining, text mining, computational linguistics, knowledge extraction technologies.

e-mail: kalinina_m@rco.ru



ПЛЕШКО Владимир Владимирович – генеральный директор компании ООО «ЭР СИ О» (RCO).

Vladimir Vladimirovich PLESHKO, received Master's degree in applied mathematics from Lomonosov Moscow State University (1996). Currently is CEO at RCO LLC, a leading company of the Russian market in the field of computational linguistics and processing of unstructured information. Current scientific interests: data mining, text mining, computational linguistics, knowledge extraction technologies, artificial intelligence.

e-mail: vp@rco.ru

Материал поступил в редакцию 15 июля 2015 года