OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822

СОЗДАНИЕ ЛЕКСИКОНА ОЦЕНОЧНЫХ СЛОВ РУССКОГО ЯЗЫКА РУСЕНТИЛЕКС

Лукашевич Н.В., Левчик А.В.

Московский государственный университет им. М.В. Ломоносова,
г. Москва, Россия
louk_nat@mail.ru

[*]RJ Games
endight@gmail.com

В данной статье описан новый лексикон оценочных слов и выражений русского языка РуСентиЛекс. Данный лексикон был собран из нескольких источников: оценочные слова из тезауруса русского языка РуТез, сленговые слова из Твиттера и слова с позитивными или негативными ассоциациями (коннотациями) из корпуса новостей. Для многозначных слов, имеющих различную оценочную направленность (тональность) при использовании в разных значениях, установлены связи значений с соответствующими понятиями в тезаурусе русского языка РуТез, что может облегчить выбор соответствующего значения слова в конкретной предметной области или конкретном контексте.
Ключевые слова: анализ тональности; оценочная лексика; лексическое значение; тезаурус; лексикон; машинное обучение

Введение

Автоматический анализ тональности текстов используется во многих практических приложениях, включая анализ пользовательских отзывов, постов в социальных сетях, новостных статей и др. Важным компонентом таких систем являются словари оценочных слов и выражений, которые могут быть использованы как в инженерных подходах к анализу тональности, основанных на словарях и правилах [Taboada et al., 2011], так и быть источниками признаков в подходах, основанных на машинном обучении [Mohammad et al., 2013; Severyn, Moschitti, 2015].

Для английского языка было создано большое количество различных словарей, которые разрабатывались экспертами [Wilson et al., 2005] или в процессе краудсорсинга [Mohammad, Turney, 2013]. Для других языков использовались в основном различные подходы автоматического порождения словарей оценочной лексики [Chetviorkin, Loukachevitch, 2012; Perez-Rosas et al., 2012; San Vicente et al., 2014; Yang et al., 2013].

Известно, что состав тональных словарей в значительной степени зависит от предметной области, поэтому множество работ посвящено извлечению или настройке тональных словарей на конкретную предметную область [Blitzer et al., 2007; Choi, Cardie 2009; Lau et al., 2011].

При этом авторы других работ показывают [Mansour et al., 2013], что комбинация обучающих данных из нескольких областей в подходах, основанных на машинном обучении, повышает качество классификатора текстов по тональности в каждой из областей. Это означает, что существует относительно стабильное множество общих (межпредметных) оценочных слов с относительно стабильной оценочной ориентацией. Кроме того, как было показано в [Mohammad et al., 2013], признаки для тональной классификации текстов, порождаемые на основе существующих словарей, полезны для улучшения качества работы систем анализа тональности, снижают зависимость от обучающих данных. Таким образом, для любого естественного языка полезно наличие словаря оценочной лексики, созданного вручную.

При этом создание такого общего словаря оценочной лексики не так просто осуществить, поскольку, несмотря на то, что есть очевидные оценочные слова (*хороший*, *плохой* и т.п.), для анализа большого объема лексики с этой точки зрения может понадобится просмотреть много слов,

поэтому оправданным является подход на основе частичной автоматизации этого процесса.

Данная статья описывает новый лексикон оценочных слов и выражений русского языка РуСентиЛекс. Лексикон был создан автоматизированно, на основе комбинации автоматических методов извлечения оценочных слов из текстов и последующего их ручного просмотра и описания. Созданный словарь содержит более десяти тысяч русских слов и выражений, выражающих некоторую оценку. Для многозначных слов, которые имеют в разных значениях разную тональность, указаны отсылки на соответствующие понятия тезауруса русского языка РуТез, что может помочь разрешать тональную неоднозначность слова в конкретной предметной области или контексте.

Статья организована следующим образом. В первой секции представлены близкие работы по созданию словарей оценочной лексики. Во второй секции описывается структура созданного словаря оценочной лексики русского языка РуСентиЛекс. Третий раздел представляет методы и источники, на основе которых был собран данный лексикон.

1. Обзор близких работ

Вручную созданные словари оценочной лексики могут быть представлены в виде простых списков слов с некоторыми атрибутами. Также разметка тональности слов может быть выполнена с учетом значений слов, так, что каждое значение слова получает свою отдельную оценку тональности.

Один из известных словарей оценочной лексики английского языка MPQA [Wilson et al., 2005] был составлен из нескольких источников (ручных и автоматически порожденных словарей оценочных слов) и содержит свыше 8000 отдельных слов. Слова в словаре размечены метками полярности (позитивный, негативный или нейтральный) и оценочные слова снабжены пометами силы оценочного содержания (сильный или слабый).

Вручную созданный английский оценочный словарь AFINN [Nielsen, 2009] был специально дополнен ругательными и сленговыми словами для адаптации его к автоматическому анализу сообщений в социальных сетях. Он содержит около 2400 слов, помеченных числовым весом полярности, изменяющегося от -5 (очень негативный) до +5 (очень позитивный).

В работе [Baccianella et al., 2010] описывается словарь SentiWordNet, который основан на тезаурусе английского языка WordNet. Он получен в результате автоматической разметки синсетов (=наборов синонимов) WordNet в результате чего каждому синсету поставлено в соответствие три числа, которые обозначают долю позитивности, негативности и нейтральности слов из данного синсета. Таким образом, разные значения одного и того же слова могут иметь различные оценки тональности.

В словаре SenticNet [Cambria et al., 2010] слова и выражения размечены по четырем измерениям: приятность (pleasantness), внимание (attention), восприимчивость (sensitivity), склонность (aptitude). Для получения числовых оценок авторы использовали оценочные слова и соответствующие веса, определенные в Hourglass of Emotions [Cambria et al., 2012] как начальное множество для получения оценок тональности для остальных понятий. Авторы словаря данного словаря уделяют особое внимание выражениям, в состав которых входят градуальные прилагательные, которые не имеют априорной тональности (*большой* и др.). Последняя версия SenticNet содержит около 30 тысяч слов и выражений.

В работе [Zasko-Zielinska et al., 2015] описывается процесс аннотирования по тональности лексических единиц, описанных в польском ворднете plWordNet. Лексическая единица в данном случае представляет собой пару (лемма, номер_значения). Для примерно 30 тысяч лексических единиц (прилагательных и существительных) были указаны полярность (позитивно, негативно, нейтрально) и оценка интенсивности (сильная и слабая). Кроме того, лексическим единицам были приписаны идентификаторы основных эмоций: радость (joy), доверие (trust), страх (fear), удивление (surprise), грусть (sadness), отвращение (disgust), гнев (anger), предвкушение (anticipation). Из размеченного набора лексических единиц 30% были позитивными или негативными.

Лексикон оценочных ассоциаций (Word-Emotion Association) Исследовательского центра Канады (NRC Canada) был создан с помощью краудсорсинга и содержит слова и выражения, которые имеют ассоциации с тональностью и определенными эмоциями [Mohammad, Turney, 2013]. Эмоциональная разметка осуществлялась по шести категориям.

В работе [Chetviorkin, Loukachevitch, 2012] описывается подход к автоматическому созданию словаря оценочной лексики в области товаров и услуг для русского языка ProductSentiRus[1]. Словарь ProductSentiRus был получен применением обученной модели к наборам отзывов в нескольких предметных областях. Словарь представлен как список 5 тысяч слов, упорядоченных по мере снижения вычисленной вероятности их оценочности без указания позитивной или негативной тональности.

2. Структура лексикона РуСентиЛекс

По своей структуре лексикон РуСентиЛекс представляет собой упорядоченный по алфавиту список слов и выражений. Он содержит следующие

[1] http://www.cir.ru/SentiLexicon/ProductSentiRus.txt

типы русскоязычных слов, значения которых связаны с тональностью:

– слова (выражения) литературного русского языка, для которых хотя бы одно значение имеет оценочный компонент, что означает, что слово в этом значении либо явно выражает отношение к обсуждаемому объекту (*отличный*), либо передается через выражаемую эмоцию (*грустно*);
– слова (выражения), не передающие оценочное отношения автора, но имеющие положительную или отрицательную коннотацию [Feng et al., 2013], например, *безработица, терроризм, болезнь, спам* и др.;
– сленговые и ругательные слова из Твиттера.

Все лексические единицы, описанные в РуСентиЛекс и их значения, рассматриваются с трех точек зрения. Во-первых, указывается полярность слова: позитивная или нейтральная; возможны также приписывание пар полярностей. Во-вторых, проставляется источник тональности: прямо выраженная оценка, эмоция или коннотация.

В-третьих, представлены тональные различия между значениями многозначного слова. Если все значения многозначного слова имеют одну и ту же тональность во всех значениях, то указывается просто тональность слова. Если слово имеет различные характеристики тональности в своих разных значениях, то описываются особенности каждого значения. Для идентификации значений устанавливается ссылка на понятия тезауруса русского языка РуТез.

Тезаурус РуТез[2] представляет собой лингвистическую онтологию для автоматической обработки текстов, т.е. онтологию, в которой большинство понятий введены на основе значений реально существующих языковых выражений. Опубликованная версия тезауруса РуТез содержит около 100 тысяч русскоязычных слов и выражений.

Если сравнивать с лингвистическими ресурсами типа WordNet, то можно отметить, что тезаурус РуТез представлен как единая сеть понятий, не разделяемая по отдельным частям речи. В качестве текстовых входов понятия могут быть слова разных частей речи, а также словосочетания различной синтаксической структуры (именные группы, глагольные группы, группы прилагательного и др.). Каждое понятие имеет уникальное имя. Понятия соединяются между собой несколькими типами концептуальных отношений [Loukachevitch, Dobrov, 2014].

При подготовке словаря РуСентиЛекс было замечено, что в русском языке имеется значимое количество слов, которые во всех своих зафиксированных в тезаурусе значениях имеют одну и ту же тональность (например, *грязный*). Поэтому было принято решение не расписывать подробно тональность таких слов по отдельным значениям, а указывать общую тональность слова. Таким образом, значения таких слов в тезаурусе могут пополняться, но тональность у них уже приписана.

Словарь РуСентиЛекс хранится в простом текстовом формате, подобном формату словаря MPQA [Wilson et al., 2005]. Каждой единице словаря, которая может быть словом, словосочетанием или лексической единицей (т.е. парой слово-понятие тезауруса РуТез) приписываются следующие атрибуты:

- слово или фраза,
- часть речи,
- слово или фраза, в которой каждое слово стоит в лемматизированной форме, что необходимо для сопоставления фраз с текстами, в которых фраза может стоять в разных словоизменительных формах,
- тональность. Тональность может быть позитивная (positive), негативная (negative), нейтральная (neutral) или двойная, например, positive/negative. В последнем случае такая отметка означает, что слово (фраза) обычно употребляется с какой-то оценкой, но эта оценка в значительной степени зависит от контекста употребления слова;
- источник тональности (явно выраженная оценка, эмоция, или факт);
- отсылки к понятиям тезауруса РуТез для значений тех слов, которые имеют различающуюся тональность в разных значениях. Для этого производится указание на имя соответствующего понятия в тезаурусе. Отметим, что в таких случаях описывается тональность для всех значений многозначного слова.

Например, слово *пресный* имеет три различных значения в тезаурусе РуТез. Два из них (значение как безвкусный о еде и значения неинтересный) имеют негативную тональность. Еще одно значение слова, связанное с пресной водой, имеет позитивную коннотацию, поскольку обладание пресной водой – это хорошо, ее истощение – это плохо и т.п.

Таким образом, описание значений слова *пресный* выглядит следующим образом (метки в кавычках соответствуют именам понятий в тезаурусе РуТез):

пресный, Adj, *пресный*, negative, emotion, "НЕВКУСНЫЙ"
пресный, Adj, *пресный*, negative, opinion, "НЕИНТЕРЕСНЫЙ"
пресный, Adj, *пресный*, positive, fact, "ПРЕСНАЯ ВОДА"

Другое оценочное русское слово *грязный* имеет два значения в тезаурусе РуТез, но оно описывается в РуСентиЛекс без ссылок на значения, поскольку оба этих значения являются негативными:

грязный, Adj, *грязный*, negative, opinion.

Слова-кандидаты на включение в лексикон были извлечены автоматически (см. п. 3). Далее эксперты-

[2] http://www.labinform.ru/pub/ruthes/index.htm

лингвисты анализировали употребление каждого слова в современных новостных текстах. Новостные тексты выбраны потому, что они адресованы максимальной аудитории современных русскоязычных людей и поэтому в среднем передают норму современного русского языка.

Например, при анализе слова *аккуратист* (*аккуратный человек*) некоторым экспертам казалось, что употребление этого слова несет негативную оценку. Но проанализированные контексты употребления этого слова показали, что оценка скорее позитивная, например:

*Страховщики заплатят штраф за отсутствие скидок водителям-**аккуратистам*** (Вести-ФМ 29.10.2015);

*Шамардин: Это интеллигентнейший человек, который показал себя с первых дней своего обучения. **Аккуратист** во всем, в одежде, в поведении* (Спорт FM, 13.10.2015).

В Таблице 1 приведены количественные характеристики разных категорий слов и выражений в словаре РуСентиЛекс.

Таблица 1. Количественные характеристики лексикона РуСентиЛекс

Категория словарного входа	**Количество**
Негативный	9744
Positive/negative	241
Позитивный	3585
Нейтральные	1394
Фактические	4607
Слова из Твиттера, отсутствующие в тезаурусе РуТез	798
Словосочетаний	2545
Всего разных текстовых входов	10467
Всего значений	14492

3. Источники для порождения лексикона

Словарные входы лексикона РуСентиЛекс были получены автоматизированно, на основе извлечения оценочной лексики из нескольких источников.

3.1. Использование существующих списков оценочных слов в конкретных предметных областях

Для получения начального списка оценочных слова и соответствующего списка понятий из тезауруса РуТез были использованы списки оценочных слов, которые были составлены в конкретных предметных областях в рамках лингвистико-инженерного подхода к анализу тональности.

Лингвистико-инженерный подход включает в себя создание словарей оценочных слов и выражений с проставленными весами оценочности, которые обычно обозначают положительными числами – положительную тональность, а отрицательными числами – негативную тональность, а также составления списка правил, которые комбинируют эти оценки между собой и со словами-операторами, усиливающими (*очень*) или модифицирующими (*не, нет*) исходную оценку слов [Kuznetsova et al., 2013].

Слова и выражения, упомянутые в этих списках были сопоставлены с тезаурусом РуТез, и все понятия, текстовые входы которых сопоставились, были извлечены для дальнейшего анализа экспертом. Также была проставлена средняя оценка понятия по его текстовым входам.

Задачей эксперта было проверить тональную оценку слова, проставленную в исходном списке, уточнить оценку значений слова, если они отличаются, а также проверить полноту создаваемого словаря за счет синонимов и текстовых входов близких понятий тезауруса.

3.2. Извлечение слов с положительными и отрицательными коннотациями

Неоценочные слова с оценочными коннотациями, упомянутые в материалах прессы, обычно ссылаются на негативные или позитивные явления в общественной или личной жизни человека. Оценочность употребления таких слов сильно зависит от контекста. Например, рост безработицы, упомянутый в статистическом отчете, является просто фактографической информацией. Если же это же выражение упомянуто в аналитическом документе, оценивающем последствия некоторых решений, то выражение будет оценочным, отрицательным.

Для автоматического извлечения слов-фактов было сделано предположение, что эти слова могут упоминаться в характерных контекстах, поскольку с негативными фактами обычно борются, уничтожают, а позитивные явления поддерживаются и защищаются. Поэтому были созданы списки шаблонов для выявления слов с негативными (35 шаблонов) и позитивными (20 шаблонов) коннотациями. Эти шаблоны были применены к новостной коллекции размером два миллиона статей.

Извлеченные слова и выражения были разделены на три класса: слова с негативными коннотациями (большинство вхождений относилось к негативными шаблонам), слова с позитивными коннотациями (большинство найденных шаблонов были позитивными), и нейтральные слова, которые встречались в обоих типах шаблонов. Например, наиболее часто встречались в негативных шаблонах такие слова, как: *коррупция, терроризм, преступление, экстремизм, наркотик, инфляция, барьер, угроза, кризис, безработица* и др. В результате этой процедуры было извлечено 4879 негативных слов и выражений, 3249 позитивных и 596 нейтральных с частотой появления в шаблонах более пяти раз.

Очевидно, что нейтральных слов должно быть значительно больше. Было сделано предположение,

что верхние уровни иерархии тезауруса нейтральны. Поэтому, двигаясь по связям с самого верхнего уровня понятий, соответствующие им текстовые входы, добавлялись в список нейтральных слов, до тех пор, пока не было достигнуто понятия, с негативным или позитивным текстовым входом. В результате список нейтральных слов стал включать 4434 элемента.

После этого все размеченные слова и словосочетания были использованы в качестве исходных слов для так называемой процедуры распространения меток (label propagation) [Zhu, Ghahramani, 2002] по отношениям тезауруса РуТез. Данная процедура использовала предположения о структуре тезауруса. Так, распространение оценки вверх (по отношениям гипероним, целое) имело меньший вес, чем по отношениям вниз.

Полученные слова и фразы были проанализированы экспертом для пополнения лексикона РуСентиЛекс.

3.3. Использование модели машинного обучения с учителем для извлечения оценочных слов из Твиттера.

Для анализа текстов социальных сетей недостаточно использовать оценочные слова, употребляемые в литературном русском языке. Поэтому были предприняты специальные усилия, чтобы извлечь наиболее вероятные оценочные слова из Твиттера. Для этого была применен метод машинного обучения, модель извлечения оценочных слов для которого была построена на размеченных данных отзывов о фильмах [Chetviorkin, Loukachevitch, 2012]. Модель хорошо показала себя при переносе на другие предметные области. Размер коллекции русскоязычных твитов, использовавшейся в этой процедуре, составлял более миллиона неразмеченных твитов.

Применяемая модель основана на использовании статистических данных, полученных из трех коллекций: коллекции с высокой концентрацией оценочных слов (А), контрастной коллекции той же предметной области с низкой концентрацией оценочных слов (В), и с контрастной общей коллекцией (использовалась новостная коллекция). В данном случае в качестве коллекции (А) использовался набор твитов, в котором нашлось хотя бы одно слово из автоматически собранного словаря ProductSentiRus [Chetviorkin, Loukachevitch, 2014]. Остальные твиты использовались в качестве коллекции (В) с низким содержанием оценочных слов.

В результате из твитов были извлечены слова, упорядоченные по мере снижения их вероятности быть оценочными словами [Chetviorkin, Loukachevitch, 2014]. Точность данного списка на уровне первой 1000 слов была оценена как 79.9%. Из первых пяти тысяч слов этого списка были исключены уже ранее известные оценочные слова, а остальные были рассмотрены лингвистом для включения в лексикон РуСентиЛекс.

Заключение

В данной статье был описан новый лексикон оценочных слов и выражений русского языка РуСентиЛекс. Данный лексикон был собран из нескольких источников: оценочные слова из тезауруса русского языка РуТез, сленговые слова из Твиттера и слова с позитивными или негативными ассоциациями (коннотациями) из корпуса новостей.

Для многозначных слов, имеющих различную оценочную направленность при использовании в разных значениях, установлены связи значений с соответствующими понятиями в тезаурусе русского языка РуТез, что может облегчить выбор соответствующего значения слова в конкретной предметной области или конкретном контексте. Все единицы лексикона расклассифицированы по четырем категориям тональности и трем источникам тональности (мнение, эмоция, факт).

Созданный лексикон может служить основой для создания оценочных словарей в конкретной предметной области или использоваться для порождения признаков в подходах к анализу тональности, основанных на машинном обучении.

Работа выполнена при частичной поддержке фонда РФФИ, грант 14-07-00682.

Библиографический список

[Baccianella, 2010] Baccianella, S., Esuli, A., Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of LREC-2010, Vol. 10, P. 2200-2204.

[Blitzer, 2007] Blitzer, J., Dredze, M., Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of ACL-2007, P. 440–447.

[Cambria, 2012] Cambria E., Livingstone A., and A. Hussain. The hourglass of emotions. Cognitive Behavioural Systems, Lecture Notes in Computer Science, vol. 7403, Springer, P. 144–157

[Cambria, 2010] Cambria E., Hussain A., Havasi C., Eckl C. Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems. Springer, LNCS, vol. 5967, P. 148-156.

[Chetviorkin, 2012] Chetviorkin, I., Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain. Proceedings of COLING-2012, P. 593-610.

[Chetviorkin, 2014] Chetviorkin, I, Loukachevitch, N. Two-Step Model for Sentiment Lexicon Extraction from Twitter Streams. Proceedings of WASSA workshop in conjugation with ACL-2014, P. 67-72.

[Choi, 2009] Choi Y., Cardie C. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. Proceedings of EMNLP '09, P. 590–598.

[Feng, 2013] Feng S., Jun S. Kang, Polina Kuznetsova, Yejin Choi. Connotation lexicon: a dash of sentiment beneath the surface meaning. Proceedings of ACL-2013, P. 1774-1784.

[Kuznetsova, 2013] Kuznetsova, E. S., Loukachevitch, N. V., Chetviorkin, I. I. Testing rules for a sentiment analysis system. Proceedings of International Conference Dialog-2013, P. 71-80.

[Lau, 2011] Lau R., Lai C., Bruza P., Wong K. Pseudo Labeling for Scalable Semi-supervised Learning of Domain-specific Sentiment Lexicons. Proceedings of 20th ACM Conference on Information and Knowledge Management..

[Loukachevitch, 2014] Loukachevitch, N., Dobrov, B. RuThes Linguistic Ontology vs. Russian Wordnets. Proceedings of Global WordNet Conference GWC-2014, Tartu.

[Mansour, 2013] Mansour, R., Refaei, N., Gamon, M., Abdul-Hamid, A., Sami, K. Revisiting The Old Kitchen Sink: Do We Need Sentiment Domain Adaptation? Proceedings of RANLP-2013, P. 420-427.

[Mohammad, 2013a] Mohammad, S. M., Turney, P. D. Crowdsourcing a word–emotion association lexicon. Computational Intelligence. – 2013. – *29*(3). – P. 436-465.

[Mohammad, 2013b] Mohammad, S. M., Kiritchenko, S., Zhu, X. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. Proceedings of Second Joint Conference on Lexical and Computational Semantics (SEM),Vol. 2, P. 321-327.

[Nielsen, 2011] Nielsen F. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, P. 93–98.

[Perez-Rosas, 2012] Perez-Rosas, V., Banea, C., Mihalcea, R. Learning Sentiment Lexicons in Spanish. Proceedings LREC-2012, P. 3077-3081.

[San Vicente, 2014] San Vicente, I., Agerri, R., Rigau, G., Sebastián, D. S. Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages. Proceedings of EACL-2014, 88.

[Severyn, 2015] Severyn, A., Moschitti, A. On the automatic learning of sentiment lexicons. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2015).

[Taboada, 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. Lexicon-based methods for sentiment analysis. Computational linguistics. – 2011. – 37(2). – P. 267-307.

[Wilson, 2005] Wilson, T., Wiebe, J., Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the conference on human language technology and empirical methods in natural language processing, P. 347-354.

[Yang, 2013] Yang, A. M., Lin, J. H., Zhou, Y. M., Chen, J. Research on building a Chinese sentiment lexicon based on SO-PMI. Applied Mechanics and Materials. – 2013. – Vol. 263. – P. 1688-1693.

[Zasko-Zielinska, 2015] Zasko-Zielinska M., Piasecki M., Szpakowicz S. A Large Wordnet-based Sentiment Lexicon for Polish. Proceedings Recent Advances in Natural Language Processing Conference (RANLP-2015). P. 721-728.

[Zhu, 2002] Zhu X., Ghahramani, Z. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Ca№ 22(4).

CREATING RUSSIAN SENTIMENT LEXICON

Loukachevitch N.V., Levchik A.V.

Lomonosov Moscow State University, Moscow, Russia

louk_nat@mail.ru

RJ Games, Moscow, Russia

endight@gmail.com

The paper describes the new Russian sentiment lexicon - RuSentiLex. The lexicon was gathered from several sources: opinionated words from domain-oriented Russian sentiment vocabularies, slang and curse words extracted from Twitter, objective words with positive or negative connotations from a news collection. The words in the lexicon having different sentiment orientations in specific senses are linked to appropriate concepts of the thesaurus of Russian language RuThes. All lexicon entries are classified according to four sentiment categories and three sources of sentiment (opinion, emotion, and fact). The lexicon can serve as the first version for the construction of domain-specific sentiment lexicons and be used for feature generation in machine-learning approaches.

Introduction

Automatic sentiment analysis is useful in many practical applications, such as analysis of users' reviews, posts in social networks, newspaper articles, etc. Sentiment lexicons are important components of sentiment analysis systems. They can be applied in lexicon-based approaches or be sources of features in the machine-learning framework

In this paper we present a new manually created general Russian Sentiment Lexicon − RuSentiLex. The lexicon contains about 10 thousand Russian sentiment-related words and expressions. Ambiguous words that have different sentiment polarity in different senses are provided with links to appropriate concepts of the Thesaurus of Russian language, which can help disambiguate sentiment ambiguity in specific domains or contexts.

Main Part

The RuSentiLex lexicon is an alphabet-ordered Russian sentiment vocabulary. It contains the following types of Russian sentiment-related words:

- words from general Russian for that at least one sense has a positive or negative polarity what means that it conveys negative/ positive attitude (*excellent*) or negative/positive emotion, (*sadness*);
- non-opinionated words with negative or positive connotations such as *unemployment*, *terrorism, disease*, *cancer*, *explosion*, etc.;
- slang and curse words from Twitter.

All words and their senses are considered from three points of views: polarity (negative, positive, or neutral); source (attitude, emotion, or non-opinionated fact); sentiment differences between word senses. If a word has different sentiment orientations or sources in its different senses then links between the senses and RuThes concepts are established.

RuSentiLex lexicon was obtained from several sources using semi-automatic techniques.

Conclusion

In the paper we describe the new Russian sentiment lexicon - RuSentiLex. The current size of the lexicon is about ten thousand words and phrases. The lexicon was gathered from several sources: opinionated words from general Russian thesaurus RuThes, slang and curse words extracted from Twitter, objective words with positive or negative connotations from news.