Высокоточный метод извлечения аспектных терминов для русского языка

Майоров В. (vmayorov@ispras.ru),
Аванесов В. (avanesov@ispras.ru),
Андрианов И. (ivan.andrianov@ispras.ru),
Астраханцев Н. (astrakhantsev@ispras.ru), Козлов
И. (kozlov-ilya@ispras.ru), Турдаков Д.
(turdakov@ispras.ru)

Институт Системного Программирования РАН,
Москва, Россия

Ключевые слова: извлечение аспектных терминов, анализ эмоциональной окраски, извлечение именованных сущностей, автоматическое извлечение терминов

Высокоточный метод для Извлечение аспекта на русском языке

Майоров В. (vmayorov@ispras.ru),
Андрианов И. (ivan.andrianov@ispras.ru),
Астраханцев Н. (astrakhantsev@ispras.ru), Аванесов
В. (avanesov@ispras.ru), Козлов И.
(kozlov-ilya@ispras.ru), Турдаков Д.
(turdakov@ispras.ru)

Институт системного программирования РАН, Москва, Россия

В этом документе представлена работа, выполненная ISPRAS над задачей извлечения аспектов на SentiRuEval 2015. Наша команда представила по одному прогону для задачи A и задачи B и получила лучшую точность для обеих задач для всех доменов среди всех участников. Наш метод также показал наилучшую F1-меру для точного сопоставления терминов аспекта для задачи A для автомобильной области и как для задачи A, так и для задачи B для ресторанной области.

Метод основан на последовательной классификации токенов с помощью SVM.
Он использует локальные, глобальные, синтаксические функции, GloVe, тематическое моделирование и функции автоматического распознавания терминов. В этой статье мы также представляем оценку значимости различных групп признаков для задачи.

Ключевые слова: извлечение аспектов, анализ тональности, NERC, синтаксические деревья,
Тематическое моделирование, GloVe, автоматическое распознавание терминов

Введение

В этой статье описывается участие в задачах извлечения аспектов SentiRuEval.
2015, в котором основное внимание уделяется выявлению аспектов в отзывах о ресторанах и автомобилях.

Извлечение аспектов является частью объектно-ориентированного анализа настроений. У автора текста могут быть разные мнения относительно конкретных свойств объекта, называемых аспектами. Термины-аспекты представляют эти аспекты в конкретном тексте.

Организаторы конкурса разделили все термины-аспекты на три типа: Эксплицитные аспекты, Имплицитные аспекты, Сентиментальные факты (Лукашевич Н.В. и др., 2015). Согласно постановке задачи, «Явные аспекты обозначают какую-то часть или характеристики описываемого объекта, например, персонал, макаронные изделия, музыку в обзорах ресторанов. [...] Неявные аспекты - это отдельные слова или отдельные слова с операторами настроений, которые содержат в себе как определенные настроения, как четкое указание на категорию аспекта. В отзывах о ресторанах частыми имплицитными аспектами являются такие слова, как вкусно (положительно+еда)

[...] Факты настроения не упоминают настроение пользователя напрямую, формально они информируют нас только о реальном факте, однако этот факт передает нам настроение пользователя, а также категорию аспекта, к которой он относится. Например, сентиментальный факт «использовала на все вопросы» (отвечал на все вопросы) означает положительную характеристику ресторанного сервиса».

Набор данных SentiRuEval был аннотирован этими тремя подтипами аспектных терминов, и участников попросили выделить отдельно только явные аспектные термины и все аспектные термины. В оставшейся части статьи мы будем называть задачу извлечения явных аспектов «Задачей А», а задачу извлечения всех аспектов — «Задачей В».

Наша система извлечения аспектов использует контролируемое машинное обучение с машинами опорных векторов (SVM), чтобы классифицировать каждый токен обзора по классам, которые обозначают начало или середину аспектов или терминов вне аспекта. Мы обучаем наш классификатор только на явных терминах аспектов, чтобы выполнить задачу А, и используем объединение результатов трех разных классификаторов, обученных для извлечения каждого типа аспектов в отдельности.

Основной задачей был поиск хорошего функционального пространства. Мы определяем три группы признаков: локальные признаки, вычисляемые в пределах одного предложения; глобальные признаки рассчитаны для одного документа; и функции, использующие внешние ресурсы.

Документ организован следующим образом: в разделе 1 дается краткий обзор соответствующей работы; в разделе 2 мы представляем полное описание нашего метода и пространства признаков, которое он использует; Раздел 3 обеспечивает оценку различных комбинаций функций для каждой задачи; в заключительном разделе мы делаем заключение по данной работе.

1. Связанная работа

Задача извлечения аспектов широко изучается в последние годы. Существует четыре основных подхода (Liu, 2012) к этой задаче. Первый подход заключается в извлечении часто встречающихся существительных и именных словосочетаний (Hu & Liu, 2004) (Popescu & Etzioni, 2007) (Scaffidi et al., 2007). Второй использует слово мнения и целевые отношения (Hu & Liu, 2004) (Qiu et al., 2011).
(Пориа и др., 2014). Эти методы основаны на идее, что слова-мнения (т. е. слова или фразы, определяющие настроение) связаны с аспектными выражениями в обзорах. Третий подход использует тематическое моделирование (Mei et al., 2007) (Branavan et al., 2008) (Li, Huang & Zhu,

2010). Последний подход основан на контролируемом машинном обучении. Было показано, что наиболее эффективными методами являются последовательное обучение, а именно скрытые марковские модели (Jin & Ho, 2009) и условные случайные поля (Jakob & Gurevych, 2010) (Choi & Cardie, 2010).

2. Описание метода

2.1. Обзор

Мнение пользователя может быть выражено несколькими способами. Каждый аспект в наборах данных, предоставленных организаторами, был помечен одним из пяти типов выражения: релевантным (упоминание термина аспекта актуально для текущего объекта обзора), сравнением (термин аспекта упоминается в сравнении с другим объектом), предыдущим (термин аспекта упоминается в сравнении с предыдущим опытом), ирреалис (термин аспекта упоминается для описания гипотетического, а не материализованного положения вещей) и ирония (термин аспекта упоминается с иронией). Мы объединили все оценки, кроме относящихся к одному классу «другое» из-за относительно небольшого количества аспектов с сравнением оценок, иронией и т. д.

Сначала мы токенизируем все обзоры и превращаем задачу в задачу маркировки последовательности: заданный список токенов присваивает последовательность тегов каждому элементу последовательности. Наш метод присваивает каждому токену один из пяти следующих классов:

1. В неаспектный термин 2. Начало
соответствующего аспектного термина 3. Середина
соответствующего аспектного термина 4. Начало
другого аспектного термина 5. Середина другого
аспектного термина

Каждый токен классифицируется с помощью SVM с регуляризацией L2. Используемые функции кратко описаны ниже.

Мы используем систему Texterra (Turdakov et. al., 2014) в качестве решения общих задач NLP для токенизации текста, тегирования PoS и морфологического анализа. Также мы используем MaltParser (Nivre et al., 2007), обученный на корпусах SynTagRus[1] для синтаксического разбора.

2.2. Местные особенности

Локальные функции — это функции, которые вычисляются с использованием только предложения. Главный местный Особенностью, используемой в нашем методе, являются классификационные метки токенов в левом окне размера 2.

Отметим, что задача извлечения аспекта очень похожа на задачу распознавания именованных объектов (NERC). Итак, мы используем некоторые особенности, которые успешно используются в методе машинного обучения с учителем NERC (Zhang & Johnson, 2003). Используемые функции NERC описаны в разделе 2.2.1.

Поскольку в русском языке порядок слов свободный, мы решили использовать особенности синтаксической структуры предложения (см. раздел 2.2.2).

---

1 http://www.ruscorpora.ru/instruction-syntax.html

2.2.1. Особенности НКРЭ

Отметим, что задача извлечения аспекта очень похожа на задачу распознавания именованной сущности. Итак, в качестве основных признаков мы выбираем следующие признаки, описанные в (Zhang & Johnson, 2003).

Префиксы и суффиксы токенов длиной 1–4; токеновые словоформы, POS-теги, морфологические свойства, леммы в окне предложения размера 2; стоит ли токен в начале предложения; маска токена (все цифры в токене заменены на специальный символ) и некоторые особенности правописания токена в окне размером 2 (все символы в верхнем регистре / цифры или знаки препинания / не буквы / цифры или буквы; является ли любой символ цифрой; первый символ в верхнем регистре).

2.2.2. Синтаксические признаки Мы

используем следующие признаки, основанные на синтаксической структуре предложения. Расстояние в синтаксическом дереве предложения между текущим токеном и другими токенами в окне размера 3. Лемма, POS-тег и морфологические свойства токена для родительского токена (с точки зрения синтаксического дерева) и для каждого дочернего токена. Метки классификации, назначенные родительским и дочерним токенам в левом окне.

2.3. Глобальные особенности

Глобальные функции — это функции, которые вычисляются с использованием всего документа. Мы используем некоторые из признаков, используемых в методе NERC на основе машинного обучения с учителем (Ratinov & Roth, 2009): относительная частота классификационных меток для всех токенов, имеющих одинаковую словоформу с текущей в левом окне размером 1000; относительная частота появления первого символа в верхнем регистре для всех токенов, имеющих одинаковую с текущей словоформу в левом окне размера 200; относительная частота POS-тегов, морфологических свойств и лемм для всех токенов, имеющих одинаковую с текущей словоформу в левом окне размером 200.

2.4. Возможности на основе внешних ресурсов

2.4.1. Перчатка

Мы также используем встраивание слова в векторное пространство в качестве признаков. Для получения вложения в 50-мерное векторное пространство мы обучаем GloVe (Pennington, 2014) на русской Википедии. К сожалению, векторы, присвоенные словам, не поддаются интерпретации, но известно, что они подобны (с точки зрения евклидова расстояния) для подобных слов. Чтобы получить интерпретируемые признаки, мы обнаруживаем кластеры слов, используя подход нечеткой кластеризации — гауссовскую смешанную модель (GMM) с 200 кластерами — количество кластеров оптимизируется с помощью байесовского информационного критерия, который, как известно, является достаточной оценкой для GMM. (Редер и Вассерман, 1995). И, наконец, в качестве признаков используется апостериорное распределение кластеров, заданное для векторного вложения слова.

2.4.2. Тематическое

моделирование Тематическое моделирование — это метод нечеткой кластеризации, обычно используемый для кластеризации документов по темам. Была использована самая базовая тематическая модель — вероятностный латентный семантический анализ (Hofmann, 1999). Эта модель предполагает, что каждый документ был нарисован

из смеси полиномиальных распределений по словам. Компоненты смеси называются темами. Итак, в результате тематического моделирования мы получаем распределение слов по данной теме. Используя теорему Байеса, мы можем легко вычислить распределение тем по словам. Наконец, это распределение используется в качестве функции. Модель обучалась на большом немаркированном наборе отзывов пользователей. Использовалась реализация tm2.

2.4.3. Автоматическое распознавание терминов

Поскольку аспекты обычно выражаются терминами, специфичными для предметной области, мы проверяем, является ли конкретное слово-кандидат частью термина, специфичного для предметной области. Для этого мы применяем методы автоматического распознавания терминов. Большинство из них, в том числе и используемые нами, работают следующим образом: на вход принимается предметная текстовая коллекция; извлечение терминов-кандидатов (n-граммы, отфильтрованные по заранее заданной части речевых паттернов); вычислить характеристики (например, частоту появления термина или tf-idf); и, наконец, классифицировать или ранжировать кандидатов терминов на основе их векторов признаков. В этой работе мы пропускаем последний шаг, т. е. мы получаем вектор признаков для каждого термина-кандидата, а затем используем его следующим образом: во время обработки текста обзора мы жадно ищем термины-кандидаты среди последовательностей словесных токенов, так что выбирается самый длинный подходящий термин-кандидат., затем мы присоединяем соответствующий вектор признаков к каждому токену слова из совпадающей последовательности.

В частности, в качестве входной текстовой коллекции мы используем комбинацию наборов обучающих и тестовых данных, а также набор документов, просканированных из Интернета, а именно: 44567 документов (82,6 Мб) с сайта restoclub.ru для домена Ресторан и 7590 отзывов (28,5 Мб).) от otzovik.com для автомобильного домена.

Берутся следующие признаки: 3 общеизвестных признака: Частота; ТФ-ИДФ; C Value (Frantzi et al., 2000) в модификации, поддерживающей однословные термины (Lossio-Ventura et al., 2013); и 4 наших признака (Астраханцев, 2014): ExistsInKB — логический признак, указывающий, представлен ли термин-кандидат в Википедии; Link Probability — вероятность термина-кандидата быть гиперссылкой в Википедии; Связанность ключевых понятий — значение семантической связанности, вычисляемое по Википедии для автоматически найденных ключевых понятий; PUATR — результат вероятностного классификатора Positive-Unlabeled, обученного на 100 лучших терминах -кандидатах (найденных специальным методом на основе частот вложенных вхождений) как положительные и другие кандидаты как немаркированные со всеми ранее описанными признаками.

3. Оценка

3.1. Оценка параметра SVM

Для оценки параметра SVM мы проводим 10-кратную перекрестную проверку наименьших обучающих данных с параметром C от 0,001 до 0,2 с шагом 0,001 в двух настройках (см. рис. 1). Первые настройки — тестирование на обучающих данных (красная линия), вторые настройки — обычная перекрестная проверка (зеленая линия). Как видно, при C < 0,045 показатель F1 растет как для обучающих, так и для тестовых данных.

Для C > 0,45 показатель F1 для поезда растет, а для тестовых данных остается почти таким же, поэтому мы решили, что это граница между переобучением и недообучением. Таким образом, мы устанавливаем C равным 0,45

2 https://github.com/ispras/tm

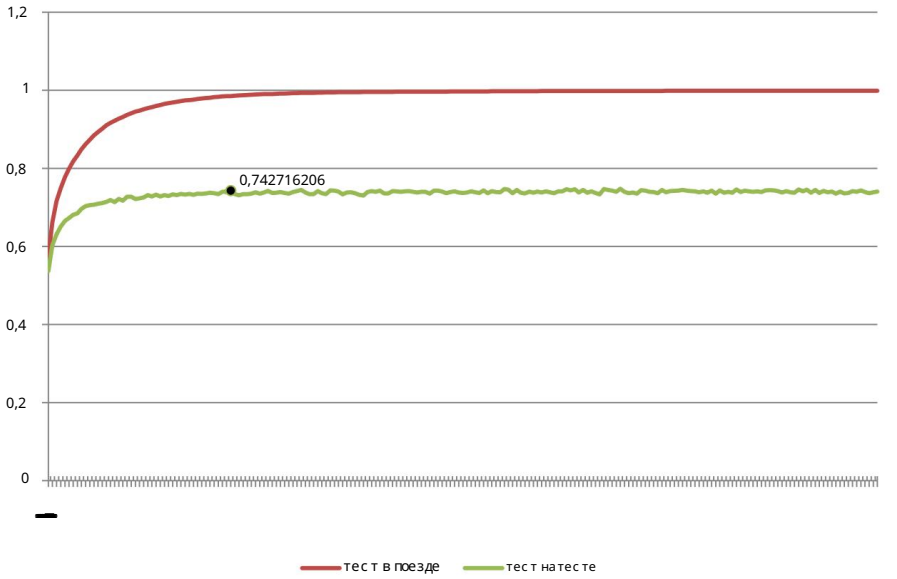

Рис. 1. Производительность метода с другим параметром SVM

3.2. Оценка влияния групп функций

Чтобы понять влияние каждой группы функций, мы последовательно удаляем каждую группу из нашего набора функций и измеряем качество метода для задачи A. Для измерения качества мы выполняем повторную 10-кратную 10-кратную перекрестную проверку и вычисляем 95% доверительный интервал для каждого качества. метрика. Результаты для автомобильного домена представлены в табл. 1. В табл. 2 представлены результаты для ресторанного домена.

Таблица 1. Результаты качества (95% доверительные интервалы) для различных наборов признаков для автомобильной области (задача A)

набор функций	точное соответствие			частичное совпадение		
	точный отзыв		f1	точный отзыв		f1
все	(0,7061; 0,7197)	(0,6500; 0,6618)	(0,6773; 0,6885)	(0,8080; 0,8200)	(0,6975; 0,7114)	(0,7493; 0,7604)
все—Перчатка	(0,7107; 0,7249)	(0,6467; 0,6584)	(0,6775; 0,6891)	(0,8139; 0,8257)	(0,6888; 0,7015)	(0,7467; 0,7573)
все — TM	(0,7031; 0,7166)	(0,6427; 0,6548)	(0,6720; 0,6832)	(0,8061; 0,8181)	(0,6882; 0,7016)	(0,7431; 0,7540)
все — ATR	(0,7032; 0,7165)	(0,6414; 0,6537)	(0,6713; 0,6826)	(0,8066; 0,8185)	(0,6915; 0,7059)	(0,7452; 0,7565)
все — глобальные	(0,7046; 0,7185)	(0,6509; 0,6633)	(0,6771; 0,6888)	(0,8068; 0,8190)	(0,6990; 0,7129)	(0,7496; 0,7609)

набор функций	точное соответствие			частичное совпадение		
	точность отзыва		f1	точный отзыв	(0,7069;	f1
все — синтаксические	(0,7132; 0,6706)	(0,6582; 0,6373;	(0,7276) (0,5120;	(0,6850; 0,6968)	(0,8155; 0,8268)	0,7203) (0,5812;	(0,7579; 0,7685)
все — НКРЭ	0,6535)	0,5253)		(0,5682; 0,5810)	(0,7655; 0,7798)	0,5968)	(0,6611; 0,6747)

Таблица 2. Результаты качества (95% доверительные интервалы) для различных наборов признаков для домена ресторана (задача А)

набор функций	точное соответствие			частичное совпадение		
	точный отзыв		f1	точный отзыв		f1
все	(0,7122; 0,7260)	(0,6546; 0,6692)	(0,6830; 0,6942)	(0,7894; 0,8024)	(0,7012; 0,7143)	(0,7439; 0,7530)
все—Перчатка	(0,7146; 0,7284)	(0,6529; 0,6672)	(0,6831; 0,6943)	(0,7956; 0,8080)	(0,6963; 0,7093)	(0,7438; 0,7528)
все — TM	(0,7140; 0,7281)	(0,6450; 0,6591)	(0,6786; 0,6896)	(0,7912; 0,8045)	(0,6884; 0,7017)	(0,7375; 0,7467)
все — ATR	(0,7106; 0,7247)	(0,6514; 0,6662)	(0,6805; 0,6920)	(0,7887; 0,8020)	(0,6972; 0,7106)	(0,7414; 0,7507)
все — глобальные	(0,7118; 0,7256)	(0,6551; 0,6696)	(0,6831; 0,6941)	(0,7893; 0,8017)	(0,7045; 0,7177)	(0,7458; 0,7545)
все — синтаксические	(0,7101; 0,7249)	(0,6570; 0,6713)	(0,6833; 0,6949)	(0,7947; 0,8076)	(0,7009; 0,7144)	(0,7461; 0,7554)
все — нерк	(0,6325; 0,6488)	(0,5109; 0,5265)	(0,5656; 0,5795)	(0,7426; 0,7571)	(0,5775; 0,5929)	(0,6504; 0,6627)

Как видно, только функции NERC вносят существенный вклад в метод. Другие группы признаков не столь значительны.

3.3. Производительность метода на тестовом наборе данных SentiRuEval

Качество предлагаемого метода, обученного на всех доступных обучающих данных со всеми описанными группами признаков, представлено в таблице 3 для задачи А и в таблице 4 для задачи В. Эти результаты получены организаторами SentiRuEval.

Таблица 3. Результаты эксперимента SentiRuEval Task A

Домен	точное соответствие			частичное совпадение		
	точный отзыв		f1	точный отзыв		f1
Автомобиль	0,760041	0,621793	0,676118	0,856055	0,655098	0,730366
Ресторан	0,723656	0,573800	0,631871	0,807759	0,616549	0,689096

Таблица 4. Результаты эксперимента SentiRuEval Task B

Домен	точное соответствие			частичное совпадение		
	точный	отзыв	f1	точный	отзыв	f1
Автомобиль	0,770100	0,553546	0,636623	0,866178	0,549210	0,659989
Ресторан	0,733599	0,513197	0,596179	0,814496	0,479988	0,590601

Заключение

Мы описали систему извлечения аспектных терминов, которая использует SVM с широким набором функций. Эта система работает с высокой точностью и хорошим показателем F1 на всех настройках и показала один из лучших результатов среди 21 прогона, полученных для задач SentiRuEval по извлечению аспектов.

Кроме того, мы провели оценку влияния различных групп признаков и обнаружили, что признаки, используемые для распознавания именованных объектов, также наиболее полезны для извлечения аспектов. Мы также обнаружили, что удаление некоторых функций может немного улучшить результаты перекрестной проверки. Одной из причин таких явлений является разреженность набора признаков. Поэтому можно предположить, что отбор признаков и уменьшение размерности могут улучшить качество предлагаемого метода. Кроме того, следует отметить, что из-за нехватки времени мы оценивали параметр SVM только на полном наборе признаков и использовали его для всех экспериментов. Однако оценка параметров SVM для каждой комбинации функций может повысить общую производительность системы. Это делает слот для будущего улучшения предлагаемого метода.

Рекомендации

1. Астраханцев Н., (2014), Автоматическое получение терминов из коллекции тематических текстов с использованием Википедии, Труды ИСП РАН, т. 1, с. 26, вып. 4, с. 7–20.

2. Fangtao L., Huang M., Zhu X. (2010), Анализ настроений с глобальными темами и локальной зависимостью, в материалах двадцать четвертой конференции AAAI по искусственному интеллекту (AAAI-2010).
3. Франти К., Ананиаду С., Мима Х. (2000), Автоматическое распознавание многословных терминов:. метод c-value/nc-value, International Journal on Digital Libraries, 3(2), 115–130.

4. Jin Wei, Hung Hay Ho, (2009), Новая лексикализированная структура обучения на основе HMM для сборамнений в Интернете, Proceedings of International Conference on Machine Learning (ICML-2009).
5. Хофманн Т. (1999), Вероятностное скрытое семантическое индексирование, в материалах 22-й ежегодной международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска (стр. 50–57). АКМ.

6. Лю Б. (2012), Анализ настроений и изучение мнений, Обобщающие лекции о Х у Языковые технологии человека, 5 (1), 1–167.
7. Lossio-Ventura JA, Jonquet C., Roche M., Teisseire M. (2013), Сочетание методов извлечения с-значения и ключевых слов для извлечения биомедицинских терминов, В LBM'2013: 5-й Международный симпозиум по языкам в биологии и медицине. (стр. 45–49).
8. Mei Qiaozhu, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai (2007), Смесь настроений по темам: моделирование граней и мнений в блогах , Proceedings of International Conference on World Wide Web (WWW-2007).
9. Никлас Дж., Гуревич И. (2010), Извлечение целевых мнений в одно- и междоменной обстановке с условными случайными полями, в материалах конференции по эмпирическим методам обработки естественного языка (EMNLP-2010).
10. Нивре Дж., Холл Дж., Нильсон Дж., Чанев А., Эригит Г., Кюблер С., Марси Э., (2007), MaltParser: независимая от языка система для анализа зависимостей, управляемых данными, Natural Языковая инженерия , 13 (02), 95–135.
11. Пеннингтон Дж., Сочер Р., Мэннинг К.Д., (2014), Перчатка: глобальные векторы для представления слов, Труды эмпирических методов обработки естественного языка (EMNLP 2014), 12.
12. Poria S., Cambria E., Ku LW, Gui C., Gelbukh A. (2014), Основанный на правилах подход к извлечению аспектов из обзоров продуктов, SocialNLP 2014, 28.
13. Попеску А.М., Этциони О. (2007), Извлечение характеристик продукта и мнений из обзоров, Обработка естественного языка и анализ текста (стр. 9–28), Springer London.

14. Qiu G., Liu B., Bu J., Chen C. (2011), Расширение слова мнения и извлечение цели посредством двойного распространения , Вычислительная лингвистика, 37(1), 9–27.
15. Ратинов Л., Рот Д. (2009) Проблемы проектирования и заблуждения при распознавании именованных объектов, Материалы Тринадцатой конференции по компьютерному изучению естественного языка / Ассоциация компьютерной лингвистики, стр. 147–155.

16. Родер К., Вассерман Л. (1997), Практическая байесовская оценка плотности с использованием смесей нормалей, Журнал Американской статистической ассоциации, 92(439), 894–902.

17. Scaffidi C., Bierhoff K., Chang E., Felker M., Ng H., Jin C. (2007), Red Opal: оценка характеристик продукта на основе обзоров, Материалы 8-й конференции ACM по электронной коммерции. , pp. 182–191 18. Турдаков Д., Астраханцев
Н., Недумов Ю, Сысоев А., Андрианов И., Майоров В., Федоренко Д., Коршунов А., Кузнецов С. (2014), Texterra: A Framework for Text Analysis, Труды ИСП РАН, том 26, вып. 1, с . 421–438.

19. Yejin C., Cardie C. (2010), Иерархическое последовательное обучение для извлечения мнений и их атрибутов, в Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010).
20. Чжан Т., Джонсон Д. (2003), Надежная система распознавания именованных объектов, основанная на минимизации рисков, Материалы седьмой конференции по изучению естественного языка в HLTNAACL, 2003 г., том 4 / Ассоциация вычислительной лингвистики, стр. 204–207. .