Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии

Н.А. Астраханцев <astrakhantsev@ispras.ru>
ИСП РАН, 109004, Россия, г. Москва, ул. А. Солженицына, дом 25

Аннотация. Автоматическое извлечение терминов является важной задачей во многих приложениях, связанных с обработкой текстов предметной области. В настоящее время существует множество методов извлечения терминов, однако они недостаточно полным образом используют внешние ресурсы, в частности – интернет-энциклопедию Википедия. Кроме того, существующие методы сильно зависят от языка и предметной области входной коллекции текстов. В данной работе предлагаются два новых признака: «Вероятность быть гиперссылкой» – нормализованная частота, с которой кандидат в термины является гиперссылкой в статьях Википедии; и «Близость к ключевым концептам» – среднее арифметическое значений семантической близости к ключевым понятиям заданной предметной области, определяемым автоматически на основе входной коллекции текстов предметной области. Также в данной работе предлагается новый автоматический метод извлечения терминов, основанный на алгоритме частичного обучения и не требующий размеченных данных. Схема метода состоит в извлечении лучших 100-300 кандидатов, присутствующих в Википедии, с помощью специального метода и последующем использовании этих кандидатов как положительных примеров для построения модели алгоритма обучения на основе положительных и неразмеченных примеров. Проведенное экспериментальное исследование на четырех предметных областях (настольные игры, биомедицина, информатика, сельское хозяйство) показывают значительное превосходство предложенного метода и его независимость от предметной области: средняя точность возросла на 5-17% по сравнению с лучшим из существующих методов для конкретного набора данных.

Ключевые слова: извлечение терминологии; Википедия; обучение на основе положительных и неразмеченных примеров.

1. Введение

Автоматическое извлечение терминов, то есть слов и словосочетаний, обозначающих определенные понятия заданной предметной области, является важным этапом во многих задачах, связанных с обработкой текстов предметной области. К таким задачам относятся, например, построение и

обогащение глоссариев, тезаурусов или онтологий [1], информационный поиск [2], извлечение информации [3], классификация и кластеризация документов [4] и т. п.

К настоящему времени разработано множество методов автоматического извлечения терминов, однако, как и во многих других задачах автоматической обработки текстов, большая часть методов существенно зависит от языка и предметной области входных текстов, что естественным образом сужает практическую применимость метода.

Кроме того, источником данных в большинстве методов является исключительно коллекция текстовых документов предметной области. Некоторые методы также используют внешние ресурсы, такие как корпуса текстов других предметных областей, поисковые машины или онтологии, созданные экспертами, однако все эти ресурсы обладает своими недостатками.

Так, внешние текстовые документы, в том числе найденные поисковыми машинами, лишены структуры и позволяют использовать только статистическую информацию о встречаемости слов и словосочетаний вне рассматриваемой предметной области; использование специализированных онтологий, например Генной онтологии [5], практически исключает возможность переноса метода на другие предметные области; универсальные онтологии, например WordNet [6] или РуТез [7], обладают малым объемом (порядка 100-150 тыс. терминов) и покрывают лишь самые общие понятия предметных областей.

Указанных недостатков лишена многоязычная интернет-энциклопедия Википедия[1]. Она содержит структурную информацию в виде соответствия статей понятиям реального мира, а также в виде гиперссылок между статьями; Википедия ежедневно пополняется сообществом пользователей, обладает очень большим размером (английская версия насчитывает более 4.5 млн. статей) и сочетает в себе универсальность и предметную специфичность.

За последние годы появилось несколько методов извлечения терминов на основе Википедии, однако в качестве возможных терминов они рассматривают только названия статей Википедии, что заведомо ограничивает полноту извлечения терминов (например, предметная область «Сельское хозяйство» покрывается Википедией только на 50% [8]).

В данной работе предлагается автоматический метод извлечения терминов из коллекции текстов с использованием Википедии, не ограничивающий набор возможных терминов и не зависящий от предметной области и языка.

Данная статья устроена следующим образом. Следующий раздел посвящен обзору существующих методов извлечения терминов. Раздел 3 представляет новый метод извлечения терминов на основе Википедии. В разделе 4 описывается проведенное экспериментальное исследование. В заключении подводятся итоги работы и предлагаются направления будущих исследований.

[1] http://wikipedia.org

2. Обзор существующих работ

Методы извлечения терминов, как правило, состоят из трех последовательных этапов:

- **Сбор кандидатов**: фильтрация слов и словосочетаний, извлеченных из коллекции документов, по статистическим и лингвистическим критериям.
- **Подсчет признаков**: перевод каждого кандидата в вектор признакового пространства.
- **Вывод на основе признаков**: классификация кандидатов на термины и не термины либо сортировка всех кандидатов по вероятности быть термином и взятие заранее определенного числа кандидатов.

2.1 Сбор кандидатов

Методы сбора кандидатов также состоят из нескольких шагов. На первом шаге применяются лингвистические фильтры, цель которых – оставить только существительные и именные группы, то есть словосочетания с существительным в роли главного слов. Для этого применяется поверхностный синтаксический разбор (shallow parsing, chunking) или фильтрация N-грамм по шаблонам частей речи.
На последующих шагах сбора кандидатов с целью снижения шума производится дополнительная фильтрация по частоте либо содержанию стоп-слов из заранее составленного списка.

2.2 Признаки для извлечения терминов

Подавляющее большинство признаков для извлечения терминов основано на частоте вхождения кандидатов в рассматриваемую коллекцию текстовых документов. К таковым относятся, например, TF-IDF [9], LexicalCohesion [10], CValue [11], DomainConsensus [12], статистические критерии Стьюдента (TTest) и логарифмического правдоподобия (Loglikelihood) [9], методы на основе тематических моделей [13]. Некоторые признаки также учитывают контекст вхождений, например NCValue [14] и Domain Model [15]. В других признаках используется частота вхождений во внешнюю коллекцию документов, не принадлежащую какой-либо предметной области: Weirdness [16], DomainRelevance [17], Relevance [18].

2.3 Вывод на основе признаков

На данном этапе, в случае использования нескольких признаков, возникает задача преобразования вектора признаков в число, показывающее уверенность метода в том, что данный кандидат является термином.

Наиболее простым способом является линейная комбинация, применяемая, например, в методе TermExtractor [17].

В работе [19] предлагается метод на основе алгоритма голосования:

$$V(t) = \sum_{i=1}^{n} \frac{1}{r(f_i(t))}$$

где t – кандидат в термины, n – количество признаков, $r(f_i(t))$ – позиция кандидата t среди всех кандидатов, отсортированных по значению признака $f_i(t)$. Данный метод не требует нормализации признаков и показывает в среднем лучшие результаты [19].

При наличии размеченных данных становится возможным применять алгоритмы машинного обучения с учителем, в частности AdaBoost [20], Ripper [21], машину опорных векторов (SVM) [22]. Как было показано в работе [23], классификаторы на основе машинного обучения достигают лучшей средней точности.

3. Метод извлечения терминов на основе Википедии

В данном разделе описываются новые признаки, основанные на использовании информации Википедии, а также новый метод, применяющий алгоритм обучения на основе положительных и неразмеченных примеров.

3.1 Признаки на основе Википедии

Признак *«Вероятность быть гиперссылкой»* (LinkProb) представляет собой нормализованную частоту, с которой кандидат в термины является гиперссылкой в статьях Википедии. Значение этого признака будет близко к нулю для слов и словосочетаний, являющихся частью общей лексики, то есть не принадлежащих какой-либо предметной области. Таким образом, мотивация использования этого признака заключается в фильтрации таких слов и словосочетаний, поскольку они, скорее всего, не принадлежат и к предметной области, для которой извлекаются термины. Стоит также отметить, что данный признак используется в методах разрешения лексической многозначности [24].

Например, словосочетание *Last card* (последняя карта) встречается 332 раза в статьях Википедии и всего лишь 4 раза в виде гиперссылки (на статью про карточную игру с одноименным названием). Таким образом, значение признака составит 0.012 и при прочих равных данное словосочетание не будет отнесено к терминам.

Признак «*Близость к ключевым концептам*» (KeyRel) основан на семантической близости к ключевым понятиям заданной предметной области и вычисляется по формуле:

$$Key\,\mathrm{Re}\,l(t) = \max_{c \in C(t)} \frac{1}{n} \sum_{i=1}^{n} sim(c, k_i)$$

где t – кандидат в термины; $C(t)$ – возможные концепты кандидата, то есть все статьи Википедии с названием, совпадающим с t с точностью до лемматизации; k_i – ключевые концепты предметной области, т.е. статьи Википедии, описывающие наиболее важные понятия для предметной области; $sim(c,k)$ – функция семантической близости, определенная для любой пары концептов и имеющая значения от 0 до 1: чем ближе значение функции к 1, тем больше общего между концептами.

Значение этого признака будет близко к нулю для слов и словосочетаний, обозначающих понятия, далекие по смыслу от ключевых понятий предметной области. Данный признак также используется в методах разрешения лексической многозначности [25].

Семантическая близость вычисляется с помощью системы Текстерра [24], в которой используется взвешенный алгоритм Дайса. Ключевые концепты для предметной области определяются на основе ключевых концептов, извлеченных для каждого документа из заданной коллекции: для каждого ключевого концепта подсчитывается, во скольких документах он был выбран ключевым, и отбирается 3 наиболее частых. Для извлечения ключевых концептов из документа используется метод KPMiner [26], реализация которого также взята из системы Текстерра.

Рассмотрим в качестве примера все то же словосочетание *Last card* применительно к предметной области «Настольные игры». Допустим, из коллекции текстов про настольные игры были извлечены следующие ключевые концепты:

- Board game (собственно, Настольная игра)
- Card game (Карточная игра)
- Hasbro Inc. (Компания, занимающаяся производством игрушек и настольных игр)

Как уже упоминалось выше, в Википедии существует статья *Last card* про одноименную игру; значения семантической близости этого концепта с указанными в списке составляют 0.001, 0.037 и 0, соответственно. Таким образом, значение признака составляет 0.0127. В то же время для термина Gene (Ген) семантическая близость к указанным концептам будет равна нулю и, таким образом, термин *Last card* является более вероятным термином предметной области «Настольные игры» с точки зрения признака «Близость к ключевым концептам».

3.2 Подход на основе частичного обучения

Идея метода основана на наблюдении, что небольшая верхняя часть списка кандидатов, полученных с помощью многих существующих методов извлечения терминов, представляет собой в основном действительно правильно определенные термины, которые могут рассматриваться в качестве обучающих данных для извлечения остальных терминов. Кроме того,

покрытие Википедии для большинства предметных областей позволяет с помощью простых методов найти небольшое число правильных терминов.

Более точно, схема метода состоит в определении лучших 100-300 кандидатов, присутствующих в Википедии, с помощью специального метода извлечения терминов и использовании этих кандидатов как положительных примеров для построения модели алгоритма обучения на основе положительных и неразмеченных примеров (Positive-unlabeled learning, PU-learning) – частного случая алгоритма частичного обучения (Semi-supervised learning). В данном случае неразмеченными примерами служат все остальные кандидаты в термины.

3.2.1 Метод извлечения положительных примеров

В данной работе используется собственный метод ModBasic в качестве метода извлечения кандидатов для последующего обучения. ModBasic представляет собой модификацию метода Basic, являющегося частью метода DomainModel [15] и, в свою очередь, модификацией широко распространенного метода CValue [11]. Метод Basic вычисляется по формуле:

$$b(t) = |t| \log f(t) + \alpha e_t$$

где t – кандидат в термины, $|t|$ - длина кандидата t, $f(t)$ – частота вхождений t в коллекции текстов, e_t – количество кандидатов, объемлющих кандидата t.

Метод Basic предназначен для извлечения «средне-специфичных» терминов, то есть терминов, описывающие распространенные понятия заданной предметной области, например, «колода карт» можно назвать средне-специфичным термином для предметной области «Настольные игры».

В методе ModBasic используется та же формула, однако вместо числа объемлющих кандидатов e_t обозначает количество кандидатов, содержащихся в данном кандидате t. Легко заметить, что данный метод предпочитает более специфичные термины, чем метод Basic, например термины «полная колода карт» или «сокращенная колода карт».

Кроме того, как было отмечено выше в описании подхода, в целях снижения шума производится фильтрация по наличию термина в названиях статей Википедии.

3.2.2 Алгоритм обучения на основе положительных и неразмеченных примеров

К настоящему времени разработано множество алгоритмов обучения на основе положительных и неразмеченных примеров; в данной работе рассмотрены следующие алгоритмы: Traditional PU learning [27], GradualReduction [28], Spy-EM [29] и PairwiseRanking SVM [30].

В качестве признаков для обучения использовались «Вероятность быть гиперссылкой», «Близость к ключевым концептам», Relevance и DomainModel.

4. Эксперименты

Извлечение терминов из коллекции текстов предметной области подразумевает следующую методологию оценки качества: для входной коллекции текстов формируется эталонное множество терминов, с которым и сравнивается результат работы системы извлечения терминов с помощью стандартных метрик полноты, точности и средней точности. Следует отметить, что на практике зачастую невозможно получить эталонное множество терминов, в точности соответствующее коллекции текстов. В таких случаях точность и полнота оцениваются приближенно.

Далее в этом разделе описываются используемые наборы данных, методика тестирования и результаты тестирования.

4.1 Наборы данных

К сожалению, в настоящее время нет общепринятого набора данных для тестирования методов извлечения терминов. Многие наборы данных и методы обладают спецификой, не позволяющей проводить тестирования в других работах; в некоторых случаях наборы данных не могут распространяться из-за нарушения интеллектуальных прав.

В работе [15] проводилось тестирование на трех открытых наборах данных: GENIA [31], Krapivin [32] и FAO [33].

GENIA можно назвать одним из наиболее популярных наборов данных; он представляет собой коллекцию из 2000 размеченных документов биомедицинской тематики.

FAO состоит из 780 размеченных вручную отчетов Продовольственной и сельскохозяйственной организации ООН (Food and Agriculture Organization).

Krapivin представляет собой 2304 научные статьи по информатике; в качестве эталонного множества терминов используются ключевые слова, выделенные авторами статей. При тестировании в данной работе к этому множеству был добавлен словарь предметной области «Вычислительная техника» (Computing), использованный в качестве эталона в системе Protodog [34].

Кроме того, для тестирования также использовался набор данных Board game [35] – коллекция из 1300 документов – описаний и рецензий настольных игр. 35 документов были размечены вручную, и для тестирования использовались только термины, имеющие хотя бы одно вхождение в размеченные документы.

4.2 Методика тестирования

Качество оценивалось с помощью средней точности:

$$AvP(N) = \sum_{i=1}^{N} P(i)(R(i) - R(i-1))$$

где *N* – общее количество оцениваемых кандидатов, или длина верхней части списка отсортированных кандидатов, *P(i)* – точность для *i* лучших кандидатов, *R(i)* – полнота для *i* лучших кандидатов. Данная метрика широко используется при оценке качества извлечения терминов, поскольку позволяет учитывать точность для всех срезов финального списка терминов.

Для набора данных Board game число N равнялось 500, для остальных наборов данных – 5000.

Кандидаты для всех оцениваемых методов представляли собой N-граммы (от 1 до 4) с фильтрацией по шаблонам частей речи и по частоте (пороговые значения для Board game и Genia – 2; для Krapivin и FAO – 3).

В методах на основе положительных и неразмеченных примеров число лучших кандидатов, используемых для обучения, равнялось 100 для Board game и 300 для остальных наборов данных.

В качестве программной реализации алгоритмов логистической регрессии, Наивного Байеса и машины опорных векторов, лежащих в основе методов Traditional, Gradual Reduction, Spy-EM и PairwiseRanking SVM, использовалась библиотека машинного обучения Weka [36]. Для реализации метода на основе тематических моделей [13] использовался открытый фреймворк[2] для построения многоязыковых регуляризованных робастных тематических моделей.

4.3 Результаты тестирования

Результаты представлены в табл. 1. Учитывая очень медленную работу и низкую эффективность алгоритма PairwiseRanking SVM на наборах данных Board game и Genia, этот алгоритм не тестировался на остальных наборах данных.

Табл. 1. Результаты тестирования

	Board game	GENIA	Krapivin	FAO
Существующие методы				
TermExtractor	0.35597	**0.79171**	0.35821	0.11126
CValue	0.36259	0.76916	0.41915	**0.29431**
Weirdness	0.30592	0.53960	0.29464	0.20923
DomainModel (DM)	0.36377	0.72815	**0.42182**	0.27956
Relevance	**0.38854**	0.55043	0.34864	0.27535
NovelTopicModel	0.34091	0.76585	0.10855	0.07801
Методы, предлагаемые в данной работе				

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

LinkProb	0.41065	0.81170	0.18491	0.02362
KeyRel	0.56113	0.79691	0.21254	0.08250
Voting (KeyRel+LinkProb+DM+Relevance)	0.46896	0.79963	0.33703	0.20577
Traditional PU	0.52806	**0.84843**	0.55675	**0.36508**
Spy-EM	0.49429	0.83089	0.47577	0.34436
PairwiseRanking SVM	0.30871	0.65465	-	-
GradualReduction	**0.56159**	**0.84843**	**0.55786**	0.34775

Как видно из таблицы, метод на основе положительных и неразмеченных примеров показывает лучшие результаты, из которых, в свою очередь, лучшим на 3 из 4 наборов данных является алгоритм GradualReduction.

Стоит отметить, что алгоритм DomainModel, продемонстрировавший одни из лучших показателей среди существующих методов, предназначен для извлечения средне-специфичных терминов, в то время как остальные методы, в том числе предложенный в настоящей работе, ставит целью извлечение всех терминов предметной области.

5. Заключение

В данной работе был представлен новый полностью автоматический метод извлечения терминов из коллекции текстов предметной области, использующий Википедию и алгоритм частичного обучения. В частности, были предложены два новых признака, использующих граф ссылок Википедии, а также специальный метод для извлечения лучших 100-200 кандидатов в термины, служащих в качестве положительных примеров для алгоритма обучения на основе положительных и неразмеченных примеров.

Экспериментальное исследование показало значительное превосходство предложенного метода по сравнению с существующими: средняя точность возросла на 5-16% (для разных наборов данных).

Среди основных направлений дальнейших исследований стоит выделить проведение более детальной оценки качества, в частности – отбора признаков и подбора параметров алгоритмов. Кроме того, представляет интерес разработка методов, повторяющих более одного раза описанную схему, то есть извлечение лучших кандидатов в термины и использование их как положительных примеров для последующего обучения.

Список литературы

[1]. Н.А. Астраханцев, Д.Ю. Турдаков. Методы автоматического построения и обогащения неформальных онтологий. Программирование, Т.39, №1, стр. 23-34, 2013.

[2]. Y. Lingpeng, J. Donghong, Z. Guodong, N. Yu. Improving retrieval effectiveness by using key terms in top retrieved documents. Advances in Information Retrieval, Springer, 2005, P. 169–184.
[3]. R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen. Automatic acquisition of domain knowledge for information extraction. Proceedings of the 18th conference on Computational linguistics-Volume 2, 2000, P. 940–946.
[4]. Д. Д. Голомазов. Методы и средства управления научной информацией с использованием онтологий. Диссертация на соискание ученой степени кандидата физико-математических наук. МГУ им. Ломоносова, 2012, 154 стр.
[5]. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet, 25(1), 2007.
[6]. G. A. Miller. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11. P. 39-41, 1995.
[7]. А.А. Алексеев, Б.В. Добров, Н.В. Лукашевич. Лингвистическая онтология тезаурус РуТез. Труды конференции Open Semantic Technologies for Intelligent Systems - OSTIS, P. 153–158, 2013.
[8]. D. Milne, O. Medelyan, I. H. Witten. Mining domain-specific thesauri from wikipedia: A case study. Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence, 2006, P. 442–448.
[9]. C.D. Manning and H. Schutze. Foundations of statistical natural language processing. MIT press, Cambridge, MA, USA. 1999. 680 p.
[10]. Y. Park, R.J. Byrd, and B.K. Boguraev. Automatic glossary extraction: beyond terminology identification. In Proceedings of the 19th international conference on Computational linguistics-Volume 1, P. 1–7. Association for Computational Linguistics, 2002.
[11]. K.T. Frantzi and S. Ananiadou. Extracting nested collocations. Proceedings of the 16th conference on Computational linguistics-Volume 1, P. 41–46. Association for Computational Linguistics, 1996.
[12]. R. Navigli and P. Velardi. Semantic interpretation of terminological strings. In Proc. 6th Intl Conf. Terminology and Knowledge Eng, 2002, P. 95-100.
[13]. S. Li, J. Li, T. Song, W. Li, B. Chang. A novel topic model for automatic term extraction. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013, P. 885-888.
[14]. K. Frantzi, S. Ananiadou, H. Mima. Automatic recognition of multi-word terms: the c-value/nc-value method. International Journal on Digital Libraries, vol. 3, no. 2, P. 115–130, 2000.
[15]. G. Bordea, P. Buitelaar, T. Polajnar. Domain-independent term extraction through domain modeling. 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), Paris, France, 2013.
[16]. K. Ahmad, L. Gillam, L. Tostevin, et al. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In The Eighth Text REtrieval Conference (TREC-8), 1999.
[17]. F. Sclano and P. Velardi. Termextractor: a web application to learn the shared terminology of emergent web communities. Enterprise Interoperability II, 2007, P. 287-290.
[18]. A. Penas, F. Verdejo, J. Gonzalo, et al. Corpus-based terminology extraction applied to information access. In Proceedings of Corpus Linguistics, volume 2001. Citeseer, 2001.

[19]. Z. Zhang, C. Brewster, F. Ciravegna. A comparative evaluation of term recognition algorithms. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco, 2008.
[20]. A. Patry and P. Langlais. Corpus-based terminology extraction. In Terminology and Content Development–Proceedings of 7th International Conference On Terminology and Knowledge Engineering, Litera, Copenhagen, 2005.
[21]. J. Foo, Term extraction using machine learning. Linkoping University, LINKOPING, 2009.
[22]. A. Judea, H. Schütze, S. Brügmann. Unsupervised Training Set Generation for Automatic Acquisition of Technical Terminology in Patents. The 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland, 2014, P. 290-300.
[23]. D. Fedorenko, N. Astrakhantsev, D. Turdakov. Automatic recognition of domain-specific terms: an experimental evaluation. Proceedings of SYRCoDIS 2013, 2013, P. 15-23.
[24]. Д. Турдаков, Н. Астраханцев, Я. Недумов, А. Сысоев, И. Андрианов, В. Майоров, Д. Федоренко, А. Коршунов, С. Кузнецов. Texterra: инфраструктура для анализа текстов. Труды Института системного программирования РАН, том 26, 2014 г. Выпуск 1. Стр. 421-438. DOI: 10.15514/ISPRAS-2014-26(1)-18.
[25]. Д. Ю. Турдаков, С. Д. Кузнецов. Автоматическое разрешение лексической многозначности терминов на основе сетей документов. Программирование, Том. 36, Номер 1, стр. 11-18, 2010.
[26]. S. R. El-Beltagy, A. Rafea. KP-Miner: A keyphrase extraction system for English and Arabic documents. Information Systems, 34(1), P. 132-144, 2009.
[27]. M. Montes, H. J. Escalante. Novel representations and methods in text classification. 7th Russian Summer School in Information Retrieval. Kazan, Russia, 2013.
[28]. M. Montes-y-Gómez, P. Rosso. Using PU-Learning to Detect Deceptive Opinion Spam. WASSA 2013, p. 38, 2013.
[29]. B. Liu, W. S. Lee, P. S. Yu, X. Li. Partially supervised classification of text documents. ICML, 2002, vol. 2, P. 387–394.
[30]. S. Sellamanickam, P. Garg, S. K. Selvaraj. A pairwise ranking based approach to learning with positive and unlabeled examples. Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, P. 663–672.
[31]. J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii. GENIA corpus--a semantically annotated corpus for bio-textmining. Bioinformatics, vol. 19, no. Suppl 1, P. 180–182, 2003.
[32]. M. Krapivin, A. Autaeu, M. Marchese. Large dataset for keyphrases extraction. 2009.
[33]. O. Medelyan, I. Witten. Domain-independent automatic keyphrase indexing with small training sets. Journal of the American Society for Information Science and Technology, 59.7, 2008, P. 1026-1040.
[34]. S. Faralli, R. Navigli. Growing Multi-Domain Glossaries from a Few Seeds using Probabilistic Topic Models. EMNLP, 2013, P. 170–181.
[35]. N. Astrakhantsev, D. Fedorenko, D. Turdakov. Automatic Enrichment of Informal Ontology by Analyzing a Domain-Specific Text Collection. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Issue 13, 2014, P. 29-42.
[36]. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009.

Automatic term acquisition from domain-specific text collection by using Wikipedia

N. Astrakhantsev <astrakhantsev@ispras.ru>
ISP RAS, 25 Alexander Solzhenitsyn Str., Moscow, 109004, Russian Federation

Abstract. Automatic term acquisition is an important task for many applications related to domain-specific texts processing. At present there are many methods for automatic term acquisition, but they are highly dependent on language and domain of input text collection. Also these methods, in general, use domain-specific text collection only, while many external resources are underutilized. We argue that one of the most promising external resources for automatic term acquisition is the online encyclopedia Wikipedia. In this paper we propose two new features: "Hyperlink probability" - normalized frequency showing how often the candidate terms is a hyperlink in Wikipedia articles; and "Semantic relatedness to the domain key concepts" - arithmetic mean of semantic relatedness to the key concepts of a given domain; those key concepts are determined automatically on the basis of input domain-specific text collection. In addition, we propose a new method for automatic term acquisition. It is based on semi-supervised machine learning algorithm, but it does not require labeled data. Outline of the method is to extract the best 100-300 candidates presented in Wikipedia by using a special method for term acquisition, and then to use these candidates as positive examples to construct a model for a classifier based on positive-unlabeled learning algorithm. An experimental evaluation conducted for the four domains (board games, biomedicine, computer science, agriculture) shows that the proposed method significantly outperforms existed one and is domain-independent: the average precision is higher by 5-17% than that of the best method for a particular data set.

Keywords: automatic term acquisition, automatic term recognition, Wikipedia, positive-unlabeled learning.

References

[1]. N.A. Astrakhantsev, D.Yu. Turdakov. Automatic construction and enrichment of informal ontologies: A survey, 2013, published in Programmirovanie, 2013, Vol. 39, No. 1.

[2]. Y. Lingpeng, J. Donghong, Z. Guodong, N. Yu. Improving retrieval effectiveness by using key terms in top retrieved documents. Advances in Information Retrieval, Springer, 2005, P. 169–184.

[3]. R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen. Automatic acquisition of domain knowledge for information extraction. Proceedings of the 18th conference on Computational linguistics-Volume 2, 2000, P. 940–946.

[4]. D. D. Golomazov. Metody i sredstva upravleniya nauchnoj informatsiej s ispol'zovaniem ontologij [Methods and tools for management of scientific information by using ontologies]. Ph.D. Thesis. Lomonosov MSU, 2012, 154 p.

[5]. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet, 25(1), 2007.
[6]. G. A. Miller. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11. P. 39-41, 1995.
[7]. А.А. Алексеев, Б.В. Добров, Н.В. Лукашевич. Лингвистическая онтология тезаурус РуТез. Труды конференции Open Semantic Technologies for Intelligent Systems - OSTIS, P. 153–158, 2013.
[8]. D. Milne, O. Medelyan, I. H. Witten. Mining domain-specific thesauri from wikipedia: A case study. Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence, 2006, P. 442–448.
[9]. C.D. Manning and H. Schutze. Foundations of statistical natural language processing. MIT press, Cambridge, MA, USA. 1999. 680 p.
[10]. Y. Park, R.J. Byrd, and B.K. Boguraev. Automatic glossary extraction: beyond terminology identification. In Proceedings of the 19th international conference on Computational linguistics-Volume 1, P. 1–7. Association for Computational Linguistics, 2002.
[11]. K.T. Frantzi and S. Ananiadou. Extracting nested collocations. Proceedings of the 16th conference on Computational linguistics-Volume 1, P. 41–46. Association for Computational Linguistics, 1996.
[12]. R. Navigli and P. Velardi. Semantic interpretation of terminological strings. In Proc. 6th Intl Conf. Terminology and Knowledge Eng, 2002, P. 95-100.
[13]. S. Li, J. Li, T. Song, W. Li, B. Chang. A novel topic model for automatic term extraction. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013, P. 885-888.
[14]. K. Frantzi, S. Ananiadou, H. Mima. Automatic recognition of multi-word terms: the c-value/nc-value method. International Journal on Digital Libraries, vol. 3, no. 2, P. 115–130, 2000.
[15]. G. Bordea, P. Buitelaar, T. Polajnar. Domain-independent term extraction through domain modeling. 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), Paris, France, 2013.
[16]. K. Ahmad, L. Gillam, L. Tostevin, et al. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In The Eighth Text REtrieval Conference (TREC-8), 1999.
[17]. F. Sclano and P. Velardi. Termextractor: a web application to learn the shared terminology of emergent web communities. Enterprise Interoperability II, 2007, P. 287-290.
[18]. A. Penas, F. Verdejo, J. Gonzalo, et al. Corpus-based terminology extraction applied to information access. In Proceedings of Corpus Linguistics, volume 2001. Citeseer, 2001.
[19]. Z. Zhang, C. Brewster, F. Ciravegna. A comparative evaluation of term recognition algorithms. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco, 2008.
[20]. A. Patry and P. Langlais. Corpus-based terminology extraction. In Terminology and Content Development–Proceedings of 7th International Conference On Terminology and Knowledge Engineering, Litera, Copenhagen, 2005.
[21]. J. Foo, Term extraction using machine learning. Linkoping University, LINKOPING, 2009.
[22]. A. Judea, H. Schütze, S. Brügmann. Unsupervised Training Set Generation for Automatic Acquisition of Technical Terminology in Patents. The 25th International

Conference on Computational Linguistics (COLING 2014), Dublin, Ireland, 2014, P. 290-300.
[23]. D. Fedorenko, N. Astrakhantsev, D. Turdakov. Automatic recognition of domain-specific terms: an experimental evaluation. Proceedings of SYRCoDIS 2013, 2013, P. 15-23.
[24]. D. Turdakov, N. Astrakhantsev, YA. Nedumov, A. Sysoev, I. Andrianov, V. Majorov, D. Fedorenko, A. Korshunov, S. Kuznetsov. Texterra: infrastruktura dlya analiza tekstov [Texterra: A Framework for Text Analysis]. Trudy ISP RAN [Proceedings of ISP RAS], 26(1), 2014. P. 421-438. DOI: 10.15514/ISPRAS-2014-26(1)-18.
[25]. D. Turdakov, S. Kuznetsov. Automatic word sense disambiguation based on document networks. Programming and Computer Software, Volume 36, Number 1, 11-18, 2010.
[26]. S. R. El-Beltagy, A. Rafea. KP-Miner: A keyphrase extraction system for English and Arabic documents. Information Systems, 34(1), P. 132-144, 2009.
[27]. M. Montes, H. J. Escalante. Novel representations and methods in text classification. 7th Russian Summer School in Information Retrieval. Kazan, Russia, 2013.
[28]. M. Montes-y-Gómez, P. Rosso. Using PU-Learning to Detect Deceptive Opinion Spam. WASSA 2013, p. 38, 2013.
[29]. B. Liu, W. S. Lee, P. S. Yu, X. Li. Partially supervised classification of text documents. ICML, 2002, vol. 2, P. 387–394.
[30]. S. Sellamanickam, P. Garg, S. K. Selvaraj. A pairwise ranking based approach to learning with positive and unlabeled examples. Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, P. 663–672.
[31]. J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii. GENIA corpus--a semantically annotated corpus for bio-textmining. Bioinformatics, vol. 19, no. Suppl 1, P. 180–182, 2003.
[32]. M. Krapivin, A. Autaeu, M. Marchese. Large dataset for keyphrases extraction. 2009.
[33]. O. Medelyan, I. Witten. Domain-independent automatic keyphrase indexing with small training sets. Journal of the American Society for Information Science and Technology, 59.7, 2008, P. 1026-1040.
[34]. S. Faralli, R. Navigli. Growing Multi-Domain Glossaries from a Few Seeds using Probabilistic Topic Models. EMNLP, 2013, P. 170–181.
[35]. N. Astrakhantsev, D. Fedorenko, D. Turdakov. Automatic Enrichment of Informal Ontology by Analyzing a Domain-Specific Text Collection. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Issue 13, 2014, P. 29-42.
[36]. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009.