Автоматическое извлечение новых концептов предметно-специфичных терминов

Д.Г. Федоренко, Н.А. Астраханцев
fedorenko@ispras.ru, astrakhantsev@ispras.ru

Аннотация. В статье описывается способ распознавания предметно-специфичных терминов, которые присутствуют в текущей базе знаний, но выражают отсутствующие в ней концепты. Разработанный метод может быть применен к неформальным базам знаний, поскольку требует только вычисления семантической близости между концептами и статистики встречаемости терминов в корпусе документов. Экспериментальная проверка показывает, что разработанный алгоритм превосходит существующие подходы, а также позволяет повысить точность разрешения лексической многозначности.

Ключевые слова: извлечение концептов; предметно-специфичные термины; обогащение баз знаний; обогащение онтологий; неформальная база знаний; неформальная онтология; разрешение лексической многозначности; семантический анализ.

1. Введение

Лексическая многозначность – неотъемлемая часть естественного языка. Слова и словосочетания могут нести различную смысловую нагрузку в зависимости от контекста, в котором они использовались. В компьютерной лингвистике задача определения значений (смысла) слов на основе контекста называется задачей *разрешения лексической многозначности*. Данная задача в настоящее время является одной из центральных и сложнейших [1] проблем обработки текстов.

Большинство современных методов разрешения лексической многозначности основаны на *базах знаний*, или *онтологиях* [2]. В контексте данной задачи, базой знаний обычно называют совокупность слов, их значений, или *концептов*, и связей между ними. Одна из проблем баз знаний заключается в их неполноте, т.е. в отсутствии подходящих значений слов, которые могут встретиться в некоторых текстах. Данная проблема особенно актуальна для

систем автоматического перевода и информационного поиска, т.к. на вход таких систем обычно подаются произвольные тексты, задаваемые пользователями. Дефицит значений слов отрицательно сказывается на качестве результатов систем, использующих механизм разрешения многозначности [3].

Наиболее часто проблема неполноты проявляется на текстах узких предметных областей, которые слабо покрываются базой знаний. Ключевую часть таких текстов составляют *специфичные термины* – текстовые представления понятий предметной области. Отсутствующие концепты специфичных терминов могут быть добавлены в базу знаний вручную, однако это потребует больших трудозатрат ввиду огромного числа существующих предметных областей.

Таким образом, возникает необходимость автоматического *обогащения базы знаний* – устранения неполноты за счет добавления в базу знаний новых концептов специфичных терминов.

Распознавание новых концептов — один из основных этапов процесса обогащения базы знаний [4]. Данный этап может происходить по-разному, в зависимости от масштаба базы знаний. В случае небольших баз знаний фиксированных предметных областей, обычно считают, что все специфичные термины несут в себе новые концепты [5]. Исключение составляют лишь термины с высоким уровнем специфичности, сам факт наличия которых в базе знаний означает существование для них подходящих концептов (явление однозначности или моносемии [6]). Однако в случае крупных баз знаний, покрывающих различные предметные области, возникает необходимость в более интеллектуальных методах распознавания концептов, поскольку для части специфичных терминов подходящие концепты уже существуют. Например, термин “hand” является специфичным для предметной области “Настольные игры” и обозначает концепт “набор карт, которые в рассматриваемый момент находятся в руках игрока”, однако для этого термина есть и другие концепты, начиная от самого распространенного — “часть тела” — и заканчивая названиями фильмов и фамилиями известных людей.

Также распознавание терминов, выражающих новые концепты, может существенно улучшить результаты алгоритмов разрешения лексической многозначности, поскольку большинство современных методов не позволяют определять случаи, когда выбор концепта невозможен по причине его отсутствия – алгоритм просто возвращает заведомо неверный концепт из доступных в текущей базе знаний [3].

В представленной работе описывается новый подход к распознаванию предметно-специфичных терминов, выражающих новые концепты. Данный подход основан на вычислении статистики встречаемости терминов в коллекции документов и *семантической близости* – функции двух концептов, показывающей, насколько концепты похожи по смыслу между собой. Важной

особенностью разработанного подхода является его применимость к *неформальным базам знаний*, характеризующимся относительно простой структурой, отсутствием формальных аксиом и детальной информации о типах связей между концептами.

Данная статья организована следующим образом. Сначала проводится обзор существующих методов распознавания специфичных терминов, выражающих новые концепты. В разделе 3 описывается разработанный метод. Далее, в разделе 4, представлено описание тестовых сценариев и результаты экспериментов. В заключении приводятся основные результаты работы и направления для дальнейших исследований.

2. Обзор методов распознавания специфичных терминов, выражающих новые концепты

В работе [7] был представлен простейший метод определения терминов, выражающих новые концепты. Данный метод основан на оценке весов концептов термина, присвоенных алгоритмом разрешения лексической многозначности. Так, в случае модели вида

$$P(S_i \mid C_1 \dots C_n) = \frac{P(S_i)P(C_1 \dots C_n \mid S_i)}{P(C_1 \dots C_n)}$$

где S_i — i-ый концепт, C_j — j-ый элемент контекста

весом концепта является число от 0 до 1, показывающее вероятность данного концепта. Предположение метода заключается в том, что если алгоритм возвращает низкий вес концепта, то такой концепт, скорее всего, не подходит и для данного термина на самом деле отсутствует подходящее значение. У данного метода есть два существенных недостатка. Первый заключается в применимости данного метода только для терминов, обладающих более чем одним концептом. Действительно, если у термина известен лишь один концепт, то вероятность выбора этого концепта алгоритмом разрешения многозначности равна единице и такой показатель веса невозможно оценить. Вторым недостатком подхода является то, что, как показали тесты в [3], при высоком пороговом значении веса сильно уменьшается показатель полноты, т.е. многие термины с верными концептами классифицируются как термины с отсутствующими концептами. Таким образом, данный подход представляет лишь теоретический интерес и может быть использован как “нижняя граница” (baseline) в решении поставленной задачи.

Более сложный подход был предложен в [3]. Автор данного подхода рассматривает термины с новыми концептами как “выбросы” из совокупности терминов, концепты которых формируют модель разрешения лексической многозначности. Для определения подобных “выбросов” все термины представляются как точки в n-мерном пространстве, в котором каждая координата является значением определенного признака. Из набора текстов,

называемого обучающей выборкой, извлекаются примеры терминов, для которых концепты известны. После чего на ранее неизвестном наборе текстов, тестовой выборке, алгоритм извлекает примеры терминов и пытается определить, относятся ли они к основному множеству примеров. Если пример достаточно сильно удален от обучающей выборки, то алгоритм полагает, что термин выражает в тексте новый концепт.

В данной статье автор в качестве признаков использует слова из контекста, части речи и именованные сущности. В качестве меры для измерения расстояния между точками используется мера Евклида. Правило, определяющее “выбросы” (новые концепты), заключается в сравнении расстояний от примера до основного множества и от ближайшей к примеру точки до оставшейся части основного множества:

$$p(x) = \frac{d_{xt}}{d_{tt}}$$

Если значение данного отношения больше заданного k (автор полагал k равным единице), то пример является выбросом и выражает новый концепт.

Результаты тестов показали, что данный подход способен находить 72% новых концептов с точностью равной 77%. Однако данный подход рассматривался лишь для одного заранее выбранного термина. Если обобщить его на случай многих терминов, то потребуется огромное количество обучающих данных, что сильно усложнит решение исходной задачи. Следовательно, данный подход также представляет лишь теоретический интерес.

Также исследователями были предложены методы, основанные на структуре и свойствах используемой базы знаний. Так, в работе [8] на основе текстового корпуса и базы знаний вычислялась специальная функция близости, позволяющая определять концепты, которые наименее близки к уже существующим.

В работе [9] задача поиска новых концептов рассматривалась как задача иерархической кластеризации с помощью нейронных сетей, строящихся на основе иерархии концептов существующей базы знаний.

На данный момент существует небольшое количество работ, затрагивающих задачу распознавания новых концептов. Самые простые методы показывают достаточно низкие результаты и поэтому могут быть использованы только в качестве “нижней границы” решения рассматриваемой задачи. Более сложные методы имеют ограниченное применение, т.к. зачастую сильно связаны со свойствами конкретных баз знаний. Обычно данные свойства заключаются в поддержке различных видов отношений между концептами [8] [10] и в возможности вычисления специальных функций близости.

3. Описание алгоритма

На вход алгоритма подается коллекция документов некоторой предметной области, а также предметно-специфичные термины, взятые из данной коллекции. Далее следует этап распознавания, результатом которого является список терминов, выражающих новые концепты.

Мы рассматриваем задачу распознавания специфичных терминов с новыми концептами как задачу классификации на два класса: “термин, выражающий существующий в текущей базе знаний концепт” и “термин, выражающий новый концепт”. Данная задача может быть решена с помощью методов машинного обучения, для применения которых необходимо перейти к признаковому описанию объектов.

3.1 Максимальная семантическая близость к ключевым концептам

Данный признак основан на предположении, что подходящий концепт специфичного термина, если он присутствует в текущей базе знаний, должен быть семантически близок к ключевым концептам предметной области. *Ключевые концепты* – это концепты, которые в совокупности формируют высокоуровневое описание документа или коллекции документов. Например, для предметной области “Настольные игры” ключевыми могут являться такие концепты, как “игровое поле”, “кость”, “карта”.

Для проверки данного предположения, для каждого известного концепта специфичного термина вычисляется семантическая близость к ключевым концептам предметной области, после чего величиной признака становится максимальная полученная близость:

$$relatedness(concepts, keyconcepts) = max_{c \in concepts} \sum similarity(c,k)$$

где *similarity(c, k)* – функция семантической близости между двумя концептами.

Ключевые концепты могут быть заданы вручную, либо определены автоматически с помощью одного из существующих алгоритмов [11] [12].

3.2 Отношение числа концептов к количеству вхождений термина

Специфичные термины, по предположению большинства методов построения баз знаний [5], должны выражать единственный концепт на протяжении всей коллекции документов. Например, слово “стек” в текстах о программировании всегда означает структуру данных и крайне редко может означать, например, трость.

Выполнимость данного предположения может быть проверена путем применения алгоритма разрешения лексической многозначности, основанного на рассматриваемой базе знаний, к коллекции документов и анализа его результатов для каждого вхождения термина. Для этого рассматривается отношение числа уникальных концептов, присвоенных термину, к общему количеству вхождений данного термина.

Чем меньше значение данной величины, тем реже алгоритм разрешения лексической многозначности присваивает термину разные концепты. В общем случае данный признак показывает, как часто меняются контексты термина и насколько сильно смена контекста влияет на выбор итогового концепта термина алгоритмом разрешения лексической многозначности.

3.3 Нормализованное число выбранных концептов

Предыдущий признак показывает, как часто алгоритм разрешения лексической многозначности меняет концепты термина, однако он не учитывает общее количество использованных концептов. Рассмотрим следующую ситуацию: алгоритм разрешения лексической многозначности определил 5 различных концептов термина, в то время как число вхождений термина равно 100. Тогда отношение числа концептов к количеству вхождений будет равно 0.05 и, следовательно, данный термин, вероятно, выражает существующий концепт. Однако данный признак перестает быть информативным, когда общее число концептов термина равно 5. Такие ситуации могут быть определены, если дополнительно рассматривать число выбранных алгоритмом разрешения многозначности концептов, поделенное на общее количество концептов, существующих в базе знаний для данного термина.

3.4 Специфичность термина

Данный признак показывает, насколько термин специфичен по отношению к внешнему корпусу общей тематики. Высокая специфичность может указывать на то, что термин обычно употребляется в единственном значении. Например, такие термины, как “синхрофазотрон” и “рибонуклеиновая кислота” обладают высоким уровнем специфичности и имеют единственное значение вне зависимости от контекста употребления. Таким образом, если специфичный термин уже существует в базе знаний, значит для данного термина существует и подходящий концепт.

В качестве признака специфичности используется *Релевантность домену (Domain Relevance)* [13], равная отношению числа вхождений термина в заданной коллекции документов к числу вхождений в коллекции общей тематики:

$$DR(t) = \frac{TF_{target}(t)}{TF_{target}(t) + TF_{reference}(t)}$$

4. Эксперименты

4.1 Тестовые данные

Для экспериментальной проверки предложенного алгоритма была использована база знаний системы Texterra [14], разрабатываемой в Институте Системного Программирования РАН. Данная база знаний состоит из терминов и концептов, полученных на основе англоязычной версии ресурса Википедия и содержит порядка четырех миллионов концептов различных предметных областей.

Одной из тематик, не полностью покрываемой базой знаний системы Texterra, является тематика “Настольные игры”. С целью проверки разработанного метода на данной предметной области, была собрана коллекция документов, состоящая из 1246 статей о настольных играх на английском языке. Для данной коллекции также был составлен список из 75 специфичных терминов, взятых из глоссария сайта Board Games Geek [1]. Каждый из этих терминов был вручную размечен либо как термин с существующим подходящим концептом, либо как термин, для которого в базе знаний системы Texterra нет подходящих концептов. В результате данной разметки было получено следующее соотношение терминов с новыми и существующими концептами: 32 и 43 соответственно.

Примеры терминов, выражающих новые концепты: movement, cell, tile-laying, chrome, crowning.

Примеры терминов, выражающих существующие концепты: dungeon crawl, rock-paper-scissors, replay value, dexterity game, deck.

4.2 Методика тестирования

В качестве методики тестирования использовался метод перекрестной проверки (*кросс-валидация*) со следующим соотношением тренировочного и тестового множеств: 90% и 10% терминов соответственно.

Для оценки качества алгоритмов вычислялись метрики точности и полноты. *Точность* — отношение числа правильно определенных терминов, выражающих новые концепты, к общему числу терминов, классифицированных алгоритмом как термины с новым концептами. *Полнота* — отношение числа правильно определенных терминов, выражающих новые концепты, к общему числу присутствующих в тестовой выборке терминов с новыми концептами.

При подсчете признаков, описанных в разделе 3, использовались алгоритмы разрешения лексической многозначности [14] и поиска ключевых концептов [11], реализованные в системе Texterra. В качестве корпуса общей тематики при подсчете признака специфичности использовался открытый корпус n-грамм современного английского языка [2].

В качестве алгоритмов машинного обучения использовались следующие методы: наивный Байесовский классификатор [15], случайный лес [16] и логистическая регрессия [17].

4.3 Результаты

4.3.1 Распознавание новых концептов

Целью данного теста является вычисление метрик качества разработанного метода. В табл. 1 представлены результаты тестирования различных алгоритмов машинного обучения, реализующих разработанный метод, и “нижней границы” – алгоритма на основе оценки веса концепта (см. раздел 2), при этом итоговым весом считался максимальный вес среди всех вхождений термина в заданной коллекции документов.

Алгоритм	Точность	Полнота	F-мера	Аккуратность (accuracy)
Наивный Байесовский классификатор	71.37%	**85.88%**	77.91%	78.26%
Случайный лес	74.13%	75.05%	74.58%	77.46%
Логистическая регрессия	**74.18%**	83.97%	**78.77%**	**80.16%**
Метод на основе оценки веса значения, threshold=0.7	63.5%	67.93%	65.64%	69.93%

Табл. 1 Результаты распознавания новых концептов.

[1] Адрес глоссария: http://boardgamegeek.com/wiki/page/Glossary

[2] Корпус доступен по адресу: http://www.ngrams.info

Как видно из таблицы, наилучшая точность на классе терминов с новыми концептами достигается с помощью методов логистической регрессии и случайного леса. В то же время наивный Байесовский классификатор показал наилучшую полноту результатов. По показателю аккуратности (точности по обоим классам) и F-меры (среднего гармонического точности и полноты) наилучшим оказался метод логистической регрессии.

Нужно также отметить, что все алгоритмы показали значительно более высокие результаты, чем метод на основе оценки веса концепта.

4.3.2 Разрешение лексической многозначности специфичных терминов

Целью данного теста является определение того, насколько повышается точность разрешения лексической многозначности за счет распознавания терминов, для которых невозможно выбрать правильные концепты ввиду их отсутствия в базе знаний.

В табл. 2 представлены результаты сравнения алгоритма разрешения многозначности системы Texterra и его комбинации с разработанным методом, при этом считалось, что для термина выбран правильный концепт, если он выбран хотя бы для одного вхождения данного термина.

	Точность	**Полнота**	**F-мера**
Texterra	52%	**52%**	52%
Texterra + разработанный метод	**78%**	44%	**56%**

Табл. 2 Результаты разрешения лексической многозначности специфичных терминов.

Как видно из таблицы, разработанный метод позволяет повысить точность разрешения многозначности специфичных терминов с 52% до 78%.

5. Заключение

В данной работе представлен метод распознавания предметно-специфичных терминов, которые присутствуют в текущей базе знаний, но выражают отсутствующие в ней концепты. Данный метод основан на вычислении статистики встречаемости терминов в корпусе документов и семантической близости между концептами. Важной особенностью разработанного подхода является его применимость к неформальным базам знаний, характеризующимся относительно простой структурой.

Разработанный метод был протестирован на основе базы знаний системы Texterra и превзошел существующие подходы. Также было выявлено, что разработанный метод позволяет существенно повысить точность разрешения лексической многозначности специфичных терминов.

Одним из направлений дальнейшей работы является кросс-доменное тестирование метода, то есть использование модели классификатора, обученного на одной предметной области, для распознавания специфичных терминов, извлеченных из коллекции отличной тематики.

Список литературы

[1]. Mallery J. C. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers // Master's thesis, MIT Political Science Department. 1988.

[2]. Turdakov D. Y. Word sense disambiguation methods // Programming and Computer Software. 2010. 36. No 6. P. 309–326.

[3]. Erk K. Unknown word sense detection as outlier detection // Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2006. P. 128–135.

[4]. Астраханцев Н., Турдаков Д. Методы автоматического построения и обогащения неформальных онтологий // Программирование. 2013. 39. No 1. С. 23–34.

[5]. Biemann C. Ontology learning from text: A survey of methods // LDV forum. 20. 2005. P. 75–93.

[6]. Agirre E., Edmonds P. G. Word sense disambiguation: Algorithms and applications. Springer Science+ Business Media. 2006. 33.

[7]. Erk K., Pado S. Shalmaneser–a toolchain for shallow semantic parsing // Proceedings of LREC. 6. 2006.

[8]. Faatz A., Steinmetz R. Ontology enrichment with texts from the www // Semantic Web Mining. 2002. P. 20.

[9]. Chifu E. T., Le Ia I. A. Text-based ontology enrichment using hierarchical self-organizing maps. 2008.

[10]. Georgiu M., Groza A. Ontology enrichment using semantic wikis and design patterns.

[11]. Grineva M., Grinev M., Lizorkin D. Effective extraction of thematically grouped key terms from text // Proc. of the AAAI 2009 Spring Symposium on Social Semantic Web. 2009. P. 39–44.

[12]. El-Beltagy S. R., Rafea A. KP-Miner: A keyphrase extraction system for English and Arabic documents //Information Systems. 2009. 34. No 1. P. 132-144.

[13]. Pazienza M. T., Pennacchiotti M., Zanzotto F. M. Terminology extraction: an analysis of linguistic and statistical approaches // Knowledge Mining. Springer. 2005. P. 255–279.

[14]. Ivannikov V., Turdakov D., Nedumov Y. Fast Text Annotation with Linked Data. Eighth International Conference on Computer Science and Information Technologies 26–30 September. 2011. Yerevan, Armenia.

[15]. Manning C. D., Schutze H. Foundations of statistical natural language processing. MIT press. 1999. P. 237.

[16]. Breiman L. Random forests // Machine learning. 2001. 45. No 1. P. 5–32.

[17]. Manning C. D., Schutze H. Foundations of statistical natural language processing. MIT press. 1999. P. 589–594.

Automatic Extraction of New Concepts from Domain-Specific Terms

D.G. Fedorenko, N.A. Astrakhantsev
ISP RAS, Moscow, Russia
fedorenko@ispras.ru, astrakhantsev@ispras.ru

Abstract. Most of the state-of-the-art approaches for word sense disambiguation (WSD) are based on knowledge bases, or ontologies — databases of terms, their concepts and relations between them. One of the standing problems of knowledge bases is their incompleteness, i.e. the lack of appropriate concepts for terms occurred in some contexts; the problem is mostly actual for domain-specific terms. The consequence is that systems produce incorrect results because existing WSD algorithms simply assign one of the a-priori incorrect concepts to the terms.

This paper describes a novel approach for recognition of domain-specific terms that exist in the knowledge base but represent new concepts. In contrast to previous approaches requiring formal ontologies with hierarchical structure and different relation types, our method can be applied to informal knowledge bases — it requires only semantic similarity between concepts and statistics of terms extracted from the domain-specific corpus.

We show that our method performs better than existing approaches and achieves 74% precision and 83% recall for the collection of domain-specific terms not fully covered by our knowledge base. Also our method improves precision of WSD from 52% to 78% for the considered terms.

Keywords: concept extraction; domain-specific terms; knowledge base enrichment; ontology enrichment; informal knowledge base; informal ontology; word sense disambiguation; semantic analysis.

References

[1]. Mallery J. C. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers. Master's thesis. MIT Political Science Department, 1988.
[2]. Turdakov D. Y. Word sense disambiguation methods. Programming and Computer Software, 2010. vol. 36. no 6. pp. 309–326. doi: 10.1134/S0361768810060010
[3]. Erk K. Unknown word sense detection as outlier detection. Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006. pp. 128–135. doi: 10.3115/1220835.1220852
[4]. Astrakhantsev N., Turdakov D. Automatic construction and enrichment of informal ontologies: A survey. Programming and Computer Software, 2013. vol. 39. no 1. pp. 23–34. doi: 10.1134/S0361768813010039
[5]. Biemann C. Ontology learning from text: A survey of methods. LDV forum. 20. 2005. pp. 75–93.
[6]. Agirre E., Edmonds P. G. Word sense disambiguation: Algorithms and applications. Springer Science+ Business Media. 2006. 33.
[7]. Erk K., Pado S. Shalmaneser–a toolchain for shallow semantic parsing. Proceedings of LREC. 6. 2006.

[8]. Faatz A., Steinmetz R. Ontology enrichment with texts from the www. Semantic Web Mining, 2002.
[9]. Chifu E. T., Le Ia I. A. Text-based ontology enrichment using hierarchical self-organizing maps. 2008.
[10]. Georgiu M., Groza A. Ontology enrichment using semantic wikis and design patterns.
[11]. Grineva M., Grinev M., Lizorkin D. Effective extraction of thematically grouped key terms from text. Proc. of the AAAI 2009 Spring Symposium on Social Semantic Web, 2009. pp. 39–44.
[12]. El-Beltagy S. R., Rafea A. KP-Miner: A keyphrase extraction system for English and Arabic documents. Information Systems, 2009. vol. 34, no 1. pp. 132-144. doi: 10.1016/j.is.2008.05.002
[13]. Pazienza M. T., Pennacchiotti M., Zanzotto F. M. Terminology extraction: an analysis of linguistic and statistical approaches. Knowledge Mining. Springer, 2005. pp. 255–279. doi: 10.1007/3-540-32394-5_20
[14]. Ivannikov V., Turdakov D., Nedumov Y. Fast Text Annotation with Linked Data. Eighth International Conference on Computer Science and Information Technologies 26–30 September. 2011. Yerevan, Armenia.
[15]. Manning C. D., Schutze H. Foundations of statistical natural language processing. MIT press, 1999. p. 237.
[16]. Breiman L. Random forests. Machine learning, 2001. vol. 45, no 1. pp. 5–32. doi: 10.1023/A:1010933404324
[17]. Manning C. D., Schutze H. Foundations of statistical natural language processing. MIT press, 1999. pp. 589–594.