Извлечение низкочастотных терминов из специализированных текстов

© К. К. Боярский
Университет ИТМО
Санкт-Петербург
boyarin9@yandex.ru

© Н. А. Арчакова © Е. А. Каневский
Экономико-математический институт РАН
Санкт-Петербург
Assoul@yandex.ru kanev@emi.nw.ru

Аннотация

Исследована возможность повышения качества выделения терминов предметной области в узкоспециализированных научных текстах. Для этого вначале с помощью семантико-синтаксического анализа и построения дерева зависимостей выделялись тематически значимые фрагменты текста. Затем производились кластеризация фрагментов и поиск терминов с использованием семантического классификатора. Показано, что данный метод позволяет с высокой вероятностью обнаруживать термины даже с единичной встречаемостью.

1 Введение

Во многих дисциплинах сейчас разрабатываются стандартные онтологии предметных областей, которые предназначены для совместного использования экспертами и автоматическими системами обработки информации. Процесс создания онтологии характеризуется высокой трудоемкостью, поскольку необходимо адекватно и максимально полно описать каждый концепт (термин), входящий в нее, с указанием всех возможных связей с другими концептами. В нашем исследовании анализируется начальный этап построения онтологии предметной области – автоматическое формирование списка терминов. В качестве исходного материала были взяты статьи Большого экономического словаря [7]. Представлены результаты кластеризации корпуса определений экономических понятий с последующим выделением терминов для каждого кластера. Объем корпуса был ограничен размерами, позволяющими вести ручную экспертную обработку для контроля качества работы автомата.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

2 Смежные исследования

Отбор терминов для помещения в онтологию обычно осуществляется с помощью лингвистического и статистического анализа [3]. В работе [2] использован гибридный метод: сначала отбираются все существительные, из которых потом по результатам статистического анализа с использованием информации о семантике слов формируется список возможных терминов.

Для повышения точности отбора в [1] применялся кластерный анализ англоязычного текста, использующий адаптивный алгоритм Леска «с помощью нахождения пересечений смыслов слов в WordNet». Результаты применения метода сравнивались с результатами, полученными при использовании частотного словаря, частотного семантического словаря и позиционного метода. Стоит отметить, что в [1] изначально показатели полноты и точности в определении ключевых слов по частотному словарю были довольно высокими. Схожая методика используется и в нашей работе. Однако целью работы [1] было извлечения ключевых слов, характеризующих каждый текст из корпуса, а цель нашего исследования – извлечение терминов, характеризующих конкретную область знаний.

В [11] рассматривалась задача онтологического анализа терминологических словарей методом семантического анализа дефиниций и построения OWL-описаний выделенных объектов. Однако выявление концептов для описания в основном базировалось на ручной обработке.

3 Специфика текста и предварительная обработка

Для анализа мы выбрали фрагмент Большого экономического словаря [7], состоящий из 1020 словарных статей, относящихся к банковской деятельности.

Для контроля качества работы системы предварительно вручную было выделено 462 однословных термина. Мы сосредоточились именно на однословных терминах, потому, что их автоматическое выделение представляет наибольшие трудности. Двухсловные термины в данных текстах имеют четко выраженную структуру: прилагательное + существительное, расположенные контактно, и могут быть найдены стандартными частотными методами.

Выбранный тип текста (словарь) имеет свои особенности. С одной стороны, все словарные статьи построены по одному шаблону: заголовок статьи и дефиниция, состоящая, как правило, из гиперонима и дополнительной информации. При этом термины предметной области могут встречаться в обеих частях словарной статьи. Как оказалось, в заголовочные части входит только 59% терминов. Такая структурированность облегчает обработку текста.

С другой стороны, как отмечалось в [6], для текстов научной направленности типична ситуация, при которой одни термины предметной области встречаются очень часто (БАНК – 826 вхождений, 3% от всех неслужебных слов корпуса), а другие – имеют только единичные вхождения (РЕТРАТТА, ХЕДЖЕР и др.). При стандартных методах отбора лексем (типа TF-IDF, LDA) для дальнейшей обработки, отбрасываются как те, так и другие. В то же время среди слов со средней частотностью большую часть составляют слова общей лексики. Например, термин ПОКУПАТЕЛЬ с абсолютной частотой вхождений 45 в анализируемом корпусе имеет по частоте встречаемости существительных 56-й ранг из 98. В первом столбце табл. 1, построенному по частотному словарю, приведены 19 существительных из окрестности слова ПОКУПАТЕЛЬ в интервале с 50 ранга (частота 51) по 62 ранг (частота 39). Оказывается, что к экономическим терминам (выделены жирным шрифтом) относятся только 42% лемм.

Таблица 1 Сравнение разных методов создания частотных списков

Методы	**Частотный словарь**	**TF-IDF**	**Предлагаемый метод**
Леммы	условия	институт	**заемщик**
	покупка	соглашение	**кредитор**
	депозит	условия	**аккредитив**
	прибыль	**актив**	**цена**
	средства	**владелец**	средства
	требование	**прибыль**	**оплата**
	использование	уровень	**вкладчик**
	производство	обращение	**сделка**
	вложение	орган	**затрата**
	ПОКУПАТЕЛЬ	**ПОКУПАТЕЛЬ**	**ПОКУПАТЕЛЬ**
	Расход	**договор**	**компания**
	соглашение	**залог**	**облигация**
	договор	требование	**залог**
	вкладчик	часть	**валюта**
	часть	**долг**	**чек**
	владелец	осуществление	**фонд**
	обращение	**погашение**	организация
	риск	**чек**	**документ**
	время	время	**рынок**
Процент терминов	42%	47%	89%

Аналогичные результаты получаются при анализе распределения терминов по TF-IDF (табл. 1, второй столбец). Здесь термин ПОКУПАТЕЛЬ имеет 38-й ранг из 80. В таком же диапазоне из 19 существительных (относящихся к 36–43 рангам) терминами являются 47% лемм.

Дополнительные проблемы создает использование модели bag-of-words, явно противоречащей структурности словарной статьи. При проведении данного исследования вместо bag-of-words использовался фрагмент дерева подчинения, построенного с помощью семантико-синтаксического парсера SemSin [8]. Этот фрагмент («краткое определение») включал в себя заголовочный термин и его гиперонимы с препозитивными определениями и зависимыми существительными в родительном падеже (рис. 1).

В качестве примера рассмотрим следующую словарную статью (слова, включенные в краткое определение, выделены жирным шрифтом):

*ОБМЕННЫЙ КУРС – **курс**, по которому одна валюта обменивается на другую, **цена денежной единицы** страны, выраженная в иностранной валюте <…>.*

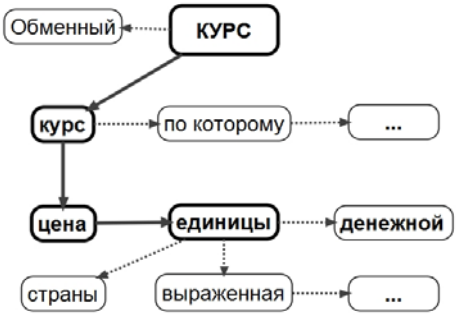

Рисунок 1 Построение краткого определения по фрагменту дерева подчинения

В правую часть кратких определений входит 16% выделенных вручную терминов, остальные – в полные определения. Такой подход позволил исключить из анализа большое количество слов

общей лексики и выявить «нелокальные» конструкции, состоящие из неконтактно расположенных слов

4. Кластеризация словарных статей

Выделение терминов происходило в три этапа:

1. Кластеризация текста.
2. Выделение наиболее частотных классов для каждого кластера.
3. Формирование списка терминов из слов, принадлежащих выделенным классам.

Для сравнения дефиниций была построена векторная модель текста. Каждому *i*-му абзацу, соответствующему одной словарной статье, был приписан нормализованный вектор $\{w_{n,i}\}$, где $w_{n,i}$ – частота токена *n* в *i*-ом абзаце. В качестве токена выступала либо лемма-существительное (сравнение «по леммам»), либо ее семантический класс (сравнение «по классам»), выявленный в ходе синтаксического анализа, выполняемого парсером SemSin. В последнем случае предполагалось, что слова, принадлежащие одному классу, эквивалентны: например, леммы *банкнота* и *валюта* относятся к одному классу «Купюры». Классы определялись в соответствии с семантическим классификатором, аналогичным описанному в [13]. В нем 192 тыс. лексем распределены по 1688 классам.

Существует много разных способов кластеризации данных. Выбор наиболее подходящего способа обусловлен особенностями анализируемых данных и целей исследования. Был применен стандартный иерархический агломеративный алгоритм кластеризации Уорда. Метод Уорда направлен на минимизацию суммы разностей квадратов расстояний внутри каждого кластера [5, 9].

Для кластеризации мы использовали пакет scikit-learn (Python) [4]. Данная реализация алгоритма накладывает ограничения на выбор способа измерения расстояния между векторами, поэтому была применена метрика Евклида.

Сначала все объекты являются отдельными кластерами. На каждой итерации алгоритма к текущему кластеру *s* добавляется один объект *t* так, что межкластерное расстояние между новым кластером $u = s \cap t$ и любым другим кластером меньше внутрикластерного расстояния.

Во многих отношениях метод Уорда является наиболее точным среди других иерархических методов [5]. В отличие от неиерархических методов, метод Уорда является устойчивым (не зависит от выбора начального приближения) и выделяет кластеры произвольной формы. Кроме того, как указано в [9], расстояние по Уорду обладает свойством растяжения. Это означает, что по мере роста кластера, расстояние от него до остальных кластеров увеличивается, что приводит к более чистому результату даже для «низкоконтрастных» текстов, в которых переход к новому разделу не означает смены лексикона. К недостаткам иерархического метода кластеризации относится то, что один из образуемых кластеров, как правило, значительно больше всех остальных.

Оптимальным для дальнейшей обработки оказалось разбиение текста на 35 кластеров, включающих от 4 до 282 словарных статей (среднее значение равно 29, медиана — 18).

Варианты классификации «по классам» и «по леммам» сравнивались с использованием данных о внутрикластерных и межкластерных расстояниях (Табл. 2). Чем больше разница между этими параметрами, тем точнее проведена кластеризация. Для используемой метрики расстояние лежит в интервале 0…1,41.

Таблица 2 Средние значения внутрикластерных и межкластерных расстояний

Вариант отбора лексики	**среднее внутрикластерное**	**среднее межкластерное**
«по классам»	0.41	1.04
«по леммам»	0.45	1.0

Отметим, что результаты кластеризации подтверждают вывод, сделанный в [6]: основной классифицирующей силой в русскоязычных текстах обладают существительные. Если при кластеризации по леммам-существительным отношение межкластерного и внутрикластерного расстояния равно 2,2 (табл. 2), то в контрольной кластеризации с учетом также прилагательных и глаголов оно составило только 1,08.

В целом, вариант отбора лексики «по классам» показывает более точные результаты. Например, термин ЦЕССИОНАРИЙ имеет три значения: лицо, становящееся кредитором (1); правопреемник (2); страховая компания (3). Во всех вариантах значение (1) верно определяется как относящееся к кластеру «Люди». При сравнении «по леммам», значения (2) и (3) объединяются в кластер, состоящий из 436 словарных статей. При сравнении «по классам» эта лемма в значении (2) попадает в кластер «Люди», а в значении (3) в кластер «Финансовые организации».

Для более точного исследования разницы результатов отбора лексики «по леммам» и «по классам», было сформировано два списка: список терминов-кандидатов «по классам» и список терминов-кандидатов «по словам» соответственно.

5 Автоматическое выделение терминов

Кластеризация, описанная в разделе 4, проводилась на основе словаря кратких определений, что позволило существенно понизить зашумленность данных. На следующих этапах алгоритма выделения

терминов мы использовали изначальные полные словарные статьи, поскольку, как было отмечено ранее, термины могут встречаться в любой части словарной статьи.

Сначала было выделено по три наиболее частых класса для каждого кластера. Стоит отметить, что полученный список классов, отличается от частотного списка классов, созданного до кластеризации из данных обо всем тексте в целом. Например, при отборе лексики «по классам» после кластеризации найдено 36 самых распространенных классов (Финансы, Деньги, Платежи, Учреждения…). Кроме того, обнаружено 7 существенных классов, которых не было в изначальном частотном списке классов (Документы, Торговля и сервис…). К ним относятся, например, термины РЫНОК, АУКЦИОН.

Затем, для каждого кластера отбирались существительные, принадлежащие наиболее частотным классам данного кластера. Эти существительные и вошли в итоговый список терминов-кандидатов.

6 Обсуждение результатов

6.1 Анализ списка терминов-кандидатов «по классам»

Из 311 отобранных слов «по классам» к терминам, выделенным экспертами относилось 249. Таким образом, точность отбора составила 0,8. В третьем столбце табл. 1 показана окрестность термина ПОКУПАТЕЛЬ (19 ранг) в частотном словаре итогового отбора. Если, как и выше, рассмотреть группу из 19 слов (с 13 по 24 ранги), то оказывается, что доля терминов в выборке увеличилась до 89% (по исходному частотному словарю – 42%). При этом из 17 терминов 13 «уникальны», т. е. встречаются в корпусе только один раз, например, ЗАЕМЩИК, ВАЛЮТА, ФОНД, РЫНОК. В то же время, в остальных столбцах присутствуют только по три уникальных термина.

Лексический анализ показал, что итоговый список терминов-кандидатов включает 95 из 147 терминов с единичной встречаемостью: ПРЕДОПЛАТА, ЛУИДОР, ФИДУЦИАРИЙ, КОМПАНИЯ-ХОЛДИНГ…. Из них 40 терминов встречались только в правой части словарной статьи, включая 25 терминов, не вошедших в краткие определения.

Для оценки качества выделения терминов был выбран стандартный способ оценки эффективности выделения информации, основанный на показателях точности и полноты. В нашем случае точность – доля правильных терминов среди слов-кандидатов, найденных автоматически, а полнота показывает, какую часть из терминов, выделенных экспертами, удалось обнаружить автоматически.

На рис. 2 и 3 представлены оценки точности и полноты соответственно. Предварительно список терминов-кандидатов был разбит на 10 примерно равных интервалов так, что термины-кандидаты с одинаковой частотой были в одном интервале. По оси абсцисс показана округленная средняя частотность терминов-кандидатов в данном интервале.

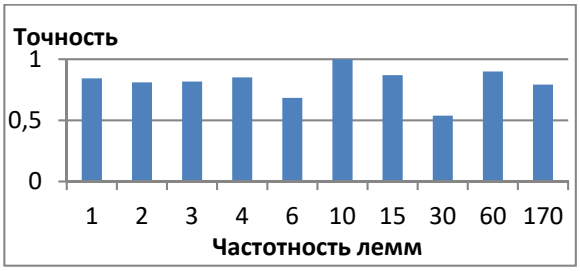

Рисунок 2 Зависимость точности выделения терминов от частотности их употребления в тексте

Как видно из рис. 2, точность не зависит от частоты встречаемости термина-кандидата. Между тем, чем чаще встречается термин, тем выше вероятность, что он будет обнаружен автоматически (рис. 3). Стоит заметить, что представленный метод извлекает термины, встречающиеся один или два раза, с вероятностью 40%. Данные из таблицы 3 показывают, что стандартные методы выделяют такие низкочастотные термины значительно хуже.

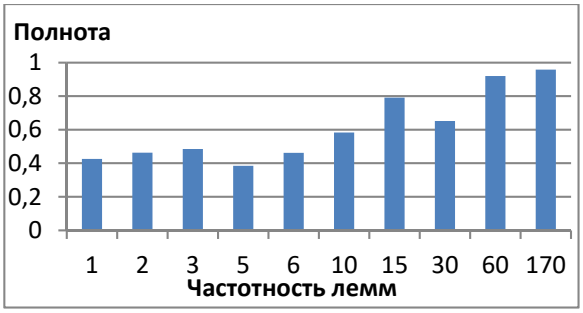

Рисунок 3 Зависимость полноты выделения терминов от частотности их употребления в тексте

В табл. 3 представлены средние оценки точности и полноты, подсчитанные с помощью 4-х разных методов:

1. по частотному словарю;
2. по списку слов, составленному по весам TF-IDF;
3. по частотному словарю слов, являющихся заголовками словарных статей;
4. по частотному списку терминов-кандидатов, созданному по описанному методу «по леммам» и «по классам».

Количество слов во всех списках одинаковое (311 по количеству терминов-кандидатов, выделенных автоматически), слова располагаются в порядке убывания частоты встречаемости. Например, частотность в первом списке изменяется от 826 до 10. Каждый список сравнивался со списком терминов, выделенных экспертами. По таблице 3 видно, что результаты предлагаемого метода значительно выше

остальных представленных способов выделения терминов.

Таблица 3 Точность и полнота, полученные по разным методам выделения терминов

	Точность	Полнота	F-мера
Частотный словарь	0.37	0.25	0.3
TF-IDF	0.27	0.19	0.22
Частотный словарь по заголовкам	0.63	0.42	0.5
Предлагаемый метод «по леммам»	0.64	0.48	0.55
Предлагаемый метод «по классам»	0.8	0.54	0.64

Был проведен анализ терминов-кандидатов, не отмеченных экспертами как термины. Можно выделить несколько причин избыточности списка терминов-кандидатов:

1. эти слова являются частью многословных терминов: ресурсы (финансовые ресурсы), карточка (пластиковая карточка), бумага (ценная бумага);
2. это гипероним, относящийся к общей лексики (организация, лицо);
3. эти слова описывают семантическую иерархию (сеть, множество, объединения, вид).

6.2. Сравнение списков терминов-кандидатов «по классам» и «по леммам»

Список терминов-кандидатов «по леммам» включает 349 лемм. Как видно из таблицы 3, результаты оценки качества выделения терминов немного выше результатов выделения по заголовкам, но значительно ниже результатов «по классам».

Списки терминов-кандидатов «по классам» и «по леммам» включают 192 общих терминов (42% от общего количества терминов): ВАЛЮТА, ДОЛЛАР, ЧЕКОДЕРЖАТЕЛЬ… и 54 слова общей лексики: БУМАГА, ГАРАНТИЯ, ОРГАНИЗАЦИЯ, ОСНОВАТЕЛЬ.… При этом 57 терминов встречаются только в списке терминов-кандидатов «по классам»: ФИНАНСЫ, БРОКЕР, ВАРРАНТ, ДЕБЕТ… В списке терминов-кандидатов «по леммам» таких слов всего 30.

Таким образом, сравнение двух способов отбора первичных данных показало, что предпочтительнее использовать способ отбора «по классам».

7 Заключение

В работе представлены результаты многоэтапного выделения однословных терминов из словаря предметной области с использованием классификатора. Показано, что традиционные методы определения терминов по частотному словарю или по весам TF-IDF не дают верной картины распределения терминов в узкоспециализированном тексте. Кроме того, несмотря на очевидность решения задачи — выборки терминов из заголовков статей словаря — оказалось, что уникальные термины могут встречаться также в любых частях дефиниций. Качество выделения терминов-кандидатов оценивалось с помощью списка терминов, найденных экспертами.

Метод, примененный в этой работе, увеличивает вероятность выделения терминов почти в два раза. Он зависит не от начального частотного распределения терминов в тексте, а от качества кластеризации. Благодаря этому, учитываются как термины с единичным вхождением, так и высокочастотные термины, равномерно распределенные по всему тексту. Средняя по частотности точность выделения терминов составила 0,8, полнота – 0,54, F-мера – 0,65.

Для проверки общности полученных результатов была проведена обработка текстов совершенно иной структуры и из другой предметной области, а именно, глав двух монографий, посвященных парусному вооружению судов [10, 12]. Несмотря на то, что лексика этих книг сильно различается (между ними полтора века), общие закономерности сохраняются. Вручную было выделено 244 термина, имеющих отношение к кораблям. Как и в экономической сфере имеется термин с очень высокой частотностью (ПАРУС – 238 вхождений, 4,4% от всех неслужебных слов). Наряду с этим 112 терминов (46%) встречаются только один раз.

При обработке по описанному выше алгоритму было найдено 158 лемм-кандидатов, из которых 152 термина. Таким образом, точность выделения терминов для судовой тематики составила 0,96, а полнота – 0,62. Этот результат показывает независимость предлагаемого метода выделения терминов от выбора конкретной предметной области.

Литература

[1] Haggag M. H., Abutabl A., Basil A. Keyword extraction using Clustering and Semantic Analysis. International Journal of Science and Research, Vol.3 #11 (November 2014). Pp 1128-1132.

[2] Lacasta J., Nogueras-Iso J., Zarazaga-Soria J. Terminological Ontologies: Design, Management and Practical Applications. Semantic Web and Beyond: Computing for Human Experience. Springer-Verlag, 2010.

[3] Pazienza M. T., Pennacchiotti M., Zanzotto F.M. Terminology extraction: an analysis of linguistic and statistical approaches. Knowledge Mining, Springer Verlag, 2005. Pp 255-281.

[4] SciPy [Электронный ресурс] URL: http://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarhie.linkage (дата обращения: 5.05.2016).

[5] Srinivasan R., Pepe A., Rodrigez V.F. A comparision between ethnographic and clustering-based semi-automatic technics for cultural ontologies. Journal of the American Society for Information Science and Technology, 2008, №5.
[6] Артемова Г.В., Боярский К.К., Гусарова Н.Ф., Добренко Н.В., Каневский Е.А. Категоризация текстов для структурирования массива исторических документов // Труды XVI Всероссийской научной конференции RCDL-2014 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Дубна, 2014. С. 159–164.
[7] Борисов А.Б. Большой экономический словарь. – М.: Книжный мир, 2003. 895 с.
[8] Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SemSin // Научно-технический вестник информационных технологий, механики и оптики. 2015. т. 15. № 5. С. 869–876.
[9] Воронцов К.В. Лекции по алгоритмам кластеризации и многомерного шкалирования. 2007. URL: http://www.ccas.ru/voron/download/Clustering.pdf (дата обращения: 5.05.2016.
[10] Курти О. Постройка моделей судов. Энциклопедия судомоделизма. Сокращенный пер. с итал. А.А. Чебана. Л.: Судостроение, 1977. 544 с.
[11] Лезин Г.В., Клименко Е.Н., Силина Е.Ф. Онтологическая интерпретация дефиниций терминологического словаря // Прикладная лингвистика в науке и образовании. Сборник трудов VII международной конференции. – СПб.: «Книжный дом», 2014. С. 50–54.
[12] Ромм Ш. Морское искусство или Главные начала и правила, научающие искусству строения, вооружения, правления и вождения кораблей. Часть 1. Пер. с франц. А.А. Шишков. Типография Морского шляхетского кадетского корпуса. Часть 1, 1793. 542 с. Часть 2, 1795. 355 с.
[13] Тузов В.А. Компьютерная семантика русского языка. СПб: Изд-во С.-Петерб. ун-та, 2004

Extraction of low-frequent terms from domain-specific texts

K. Boyarsky, N. Archakova, E. Kanevsky

We examined the way to improve the quality of low-frequent term recognition in scientific texts. Firstly, domain-relevant fragments were extracted from the text with the help of dependency tree. Then the fragments were clustered and candidate terms were defined using the semantic classifier. The studies suggest that this approach allows extracting unique terms as well.