АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ТЕРМИНОЛОГИИ С ИСПОЛЬЗОВАНИЕМ ПОИСКОВЫХ МАШИН ИНТЕРНЕТА

AUTOMATIC TERM EXTRACTION USING INTERNET SEARCH ENGINES

Браславский П.И. *(pb@imach.uran.ru),* ***Соколов Е.А.*** *(esokolov@list.ru)*
Институт машиноведения УрО РАН, Екатеринбург

В статье описываются методы автоматического извлечения двухсловных терминов из отдельного текста или корпуса текстов с использованием поисковых машин интернета. Рассмотрены пять различных вариантов подсчета «терминологичности» словосочетаний. Проведены эксперименты на трех наборах данных, относящихся к разным областям знаний. Предложена комбинированная методика оценки, приведены результаты сравнительной оценки методов.

Введение

Задача выделения ключевых слов и терминов из текста возникает в библиотечном деле, лексикографии и терминоведении, а также в информационном поиске. Объемы и динамика информации, которая подлежит обработке в этих областях в настоящее время, делают особенно актуальной задачу *автоматического выделения* терминов и ключевых слов. Выделенные таким образом слова и словосочетания могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, классификации.

В нашей предыдущей работы [1] мы исследовали четыре метода автоматического выделения двухсловных терминоподобных конструкций из текста. Все эти методы используют минимум исходной информации: 1) статистику встречаемости пар и отдельных слов в тексте и 2) морфологические шаблоны-фильтры. Мы использовали пять морфологических шаблонов и четыре метода подсчета «устойчивости» словосочетания: 1) прямой подсчет количества пар (**freq**); 2) **t-тест**; 3) **χ^2-тест**; 4) отношение функций правдоподобия (**LR**) (подробнее методы описаны в [1]).

Как показала оценка, методы **freq** и **t-тест** сравнимы по эффективности и могут быть использованы для составления списка терминов-кандидатов в задачах полуавтоматического формирования терминологических ресурсов. Основной тип ошибок обоих методов – выделение устойчивых общеупотребительных словосочетаний, удовлетворяющих шаблонам. Для фильтрации этих выражений мы предложили использовать дополнительный «контрастный» корпус. В данной работе рассматривается использование Веба в качестве такого корпуса, доступ к которому осуществляется с помощью поисковых машин интернета. Очевидны преимущества такого подхода: по актуальности, разнообразию и размеру с Вебом не могут конкурировать другие корпуса, а наличие интерфейсов к машинам поиска существенно облегчает реализацию метода. В качестве недостатков такого подхода можно назвать тематическую несбалансированность Веба (о чем пойдет речь ниже), а также различные артефакты – дубликаты документов, опечатки, спам и т.п.

Для отделения терминоподобных словосочетаний от общеупотребительных выражений мы используем два параметра: 1) частотность словосочетания и 2) совместная встречаемость словосочетаний. Первый параметр должен отражать то, что специальный термин имеет ограниченное употребление; а второй – системность терминологии (термины одной предметной области сильно связаны между собой и значительно слабее – с терминами из других предметных областей).

Эксперименты проводились для двусловий и трехсловий, в статье описаны результаты только для двусловий, так как экспертная оценка проводилась только для них.

Состояние исследований

В литературе можно найти описания близких подходов к извлечению терминов с использованием данных из Веба.

В работе [3] описан метод автоматического извлечения терминов предметной области «Технические сред-

ства обучения» на основе двух корпусов, полученных из Веба. Эксперименты проводились для испанского языка. Основной корпус был сформирован путем обхода двух тематических веб-сайтов и содержал около 1000 документов. В качестве контрастного был использован корпус газеты El Pais, содержавший более 7000 новостных сообщений. Термины-кандидаты извлекались из основного корпуса с помощью шаблонов. Формальный показатель терминологичности основывался на абсолютной частоте и документной частоте термина-кандидата в основном корпусе, а также частоте термин-кандидата в контрастном корпусе.

В работе [5] описывается подход к извлечению *согласованного* набора ключевых фраз из отдельного документа. Задача извлечения ключевых фраз решается как задача бинарной классификации методами машинного обучения. В качестве исходных признаков используются частотные характеристики фраз и их положение в тексте. После первого прохода для наиболее вероятных ста терминов вычисляются значения признаков *согласованности* (*coherence*). Значения признаков согласованности вычисляются на основе совместной встречаемости фраз-кандидатов с четверкой наиболее вероятных (по результатам той же первичной классификации) ключевых фраз в Вебе.

Работа [4] описывает метод формирования списка терминов, близких данному исходному термину (*seed term*), с использованием поисковых машин интернета. Эксперименты проводились с японскими терминами. Предлагаемый подход состоит из следующих шагов: 1) запросы, дополненные служебными словами, задаются машинам поиска; 2) скачиваются топ-100 документов из каждого результата поиска; 3) формируется корпус: из документов извлекаются предложения, содержащие исходный термин; 4) из корпуса извлекаются термины-кандидаты; 5) термины-кандидаты фильтруются на основе частоты встречаемости в интернете (термин не должен встречаться очень редко и очень часто); 6) термины-кандидаты проверяются на близость к исходному термину; близость вычисляется на основе лексических характеристик (вхождение элементов исходного термина) и совместной встречаемости терминов в интернете. Предлагается рекурсивно использовать метод для формирования списка терминов предметной области.

Система BootCaT, описанная в работе [2], решает несколько иную задачу – автоматического формирования тематического корпуса из Веба, – однако для ее решения используются близкие нашему исследованию методы. Процесс построения корпуса начинается на основе набора исходных терминов (*seed terms*). С помощью автоматических запросов к поисковой машине извлекаются документы, содержащие исходные термины; в свою очередь из этих документов извлекаются новые *однословные* термины (на основе сравнения частот в сформированном корпусе со «стандартным корпусом»), которые вновь можно использовать в качестве запросов, и т.д. Финальный корпус и список однословных терминов используется для итеративного извлечения многословных терминов на основе структурной информации и частотных характеристик.

Данные

Мы проводили эксперименты на трех наборах данных для того, чтобы проверить гипотезу о независимости методов от предметной области. Мы использовали электронные версии двух книг, а также корпус статьей научного журнала:

1. Олифер Н.А., Олифер В.Г. Сетевые операционные системы. СПб.: Питер, 2005. (**СОС**)
2. Щедровицкий Г.П. Философия. Наука. Методология. М.: ШКП, 1989. (**ФНМ**)
3. «Информационный вестник ВОГиС», http://www.bionet.nsc.ru/vogis. (**ВОГиС**)

Первая книга – это монография, описывающая достаточно узкую техническую область – сетевые операционные системы. Особенностью второй книги является то, что это не цельный текст, а сборник статей одного автора по обширной тематике. Границы предметной области здесь намного более расплывчаты, и сам текст менее насыщен специальными терминами. Эти два набора данных – СОС и ФНМ – использовались в экспериментах, описанных в [1]. Новый набор данных – это 100 статей разных авторов по генетике, селекции, а также смежным наукам, опубликованных в «Информационном вестнике ВОГиС» с 1997 по 2006 год. Были взяты все статьи журнала за этот период, за исключением редакционных статей, посвященных юбилеям ученых и памятным датам.

Книги СОС и ФНМ содержат предметный указатель (ПУ), который мы принимаем за список терминов, выделенных автором, и используем для формальной оценки извлеченных терминов-кандидатов. Журнал «Информационный вестник ВОГиС» не содержит предметного указателя. Поэтому для формальной оценки результатов его обработки мы используем русскую часть словаря терминов по молекулярной и клеточной биологии (http://www.mblogic.net/glossary/), предоставленную нам Анастасией Барышниковой. Словарь содержит 6 315 входов (всего 7 227 терминов). Интересно отметить, что словарь содержит достаточно много «терминов-метафор» (обычно употребляются в кавычках) – как однословных (*“аркан”, “булава”, “восьмерка”, “газон”* и др.), так и многословных (*“шитье назад”, “горячая точка”, “узлы-на-веревке”, “счастливые уроды”* и др.). Кроме того, словарь содержит много терминов специфической структуры, например, с цифрами (*1-метил-4-*

амино-6-оксипиримидин, 4-тиоуридин и др.), греческими и латинскими буквами (*α-гетерохроматин*, *β-талассемия*, *D-петли, F-эписома, HKG-бэндинг* и др.), а также сложные термины (например, *хронический остеомиелит длинных костей после огнестрельных повреждений*).

Тексты анализировались в формате *plain text*. По сравнению с экспериментами, описанными в [1], изменилась обработка слов с дефисом (такие слова всегда обрабатываются как одно слово, что немного изменило статистику). В экспериментах, описанных в данной работе, каждый из трех наборов обрабатывался как монолитный документ, без учета разбиения на отдельные статьи или главы. (Ср. пять наиболее частотных словосочетаний из корпуса ВОГиС, которые встретились более чем в 10 документах: *настоящее время, институт цитологии, точка зрения, последний год, высокий уровень*.) Некоторые характеристики текстов приведены в табл. 1.

	СОС	**ФНМ**	**ВОГиС**
Всего слов	99 337	180 048	256 255
Без стоп-слов	66 438	98 065	179 635
Пар	9 391	11 719	30 245
Уникальных пар, удовлетворяющих шаблонам, с частотой >1	2 653	2 947	7 656

Таблица 1. Характеристики обработанных текстов

Описание методов

Кандидаты в термины – это пары слов, которые извлекаются непосредственно из обрабатываемого текста. Отбираются все пары слов, не разделенные знаками препинания (кроме дефиса и кавычек) и стоп-словами. Далее кандидаты проходят морфологический фильтр. Мы используем четыре шаблона для двухсловных терминов (табл. 2). Морфологическая обработка осуществлялась с помощью программы *mystem* (http://company.yandex.ru/technology/products/mystem/mystem.xml); при неоднозначности морфологического разбора мы требуем совпадения хотя бы одного из возможных сочетаний с шаблоном. Прилагательные и причастия должны быть согласованы с существительным, пары с участием кратких прилагательных и причастий отсеиваются.

Шаблон	**Пример**
[Прил. + Сущ.]	*файловая система*
[Прич. + Сущ.]	*вытесняющая многозадачность*
[Сущ. + Сущ., Род.п.]	*менеджер памяти*
[Сущ. + Сущ., Твор.п.]	*управление ресурсами*

Таблица 2. Морфологические шаблоны для двухсловных терминов

В качестве базового метода мы используем метод **freq**, когда отфильтрованные пары упорядочиваются по убыванию частоты употребления в исходном тексте (см. [1]). Далее мы берем 200 (150 – для метода **coherence**) элементов из верхушки списка **freq** и пытаемся по-новому ранжировать этот список с учетом данных о встречаемости словосочетаний в Вебе. Эти данные мы получаем с помощью сервиса Яндекс.XML (http://xml.yandex.ru), который позволяет автоматически задавать поисковые запросы и получать отклик в формате XML. Для нахождения частоты встречаемости словосочетания в Вебе мы задаем запрос вида [$word_1$ /+1 $word_2$].

Комбинация Веб-данных с ранжированием на основе исходного текста дает нам дополнительные четыре списка терминов-кандидатов – **iFreq**, **TF*IDF**, **freq/iFreq**, **coherence**. Метод **iFreq** – это ранжирование исходного списка по увеличению частоты встречаемости термина в Интернете (верхушка списка – наиболее редкие термины). Список **freq/iFreq** получен исключением из списка **freq** 30 наиболее частотных словосочетаний по Вебу («хвост» списка **iFreq**). Значение порога (30) выбрано нами на основании анализа результатов предыдущих экспериментов (см. [1]). **TF*IDF** – это применение известного метода к двусловиям. Частота термина-кандидата (TF) считается по тексту/корпусу, а обратная документная частота (IDF) – по Вебу. Список ранжируется по убыванию TF*IDF. И, наконец, метод **coherence** должен отражать взаимосвязь терминов из списка, ранжируя выше термины-хабы (те, которые часто встречаются с другими терминами из списка). Простой показатель связности двух словосочетаний – это количество документов, в которых встречаются оба словосочетания. Для получения этого параметра мы задаем запросы вида: [($word_{11}$ /+1 $word_{12}$) && ($word_{21}$ /+1 $word_{22}$)]. Нормированная попарная связность двух терминов-кандидатов определяется следующим образом:

$$Coherence(term_1, term_2) = \frac{iFreq(term_1 \bigcap term_2)}{iFreq(term_1) \cdot iFreq(term_2)},$$

где

iFreq(term $_1$*)* и *iFreq(term* $_2$*)* – количество документов, содержащих $term_1$ и $term_2$ соответственно;

iFreq(term $_1 \cap$ *term* $_2$*)* – количество документов в интернете, содержащих одновременно оба термина.

Для вычисления общей связности словосочетания «со всем списком» мы для каждого суммируем его попарные связности (здесь мы неявно делаем предположение, что в списке больше «хороших» терминов, чем нетерминов или терминов других предметных областей, результаты [1] дают нам основания для этого). Конечно, такой подход достаточно затратный. Если у нас уже есть частоты *n* словосочетаний в Вебе, то дополнительно мы должны задать $(n^2 - n)/2$ запросов поисковой машине, т.е. для 150 кандидатов нужно послать всего (150 + 11 175) запросов. Список **coherence** мы получаем ранжированием ста пятидесяти словосочетаний по убыванию общей связности.

Итого, мы имеем пять списков терминов-кандидатов: **freq**, **iFreq**, **freq/iFreq**, **TF*IDF**, **coherence** для каждого набора данных.

Методика оценки

Для оценки полученных списков мы используем методику, описанную в [1]. Мы комбинируем ручную (экспертную) оценку и формальную оценку по «эталонному списку» (предметному указателю для СОС и ФНМ; словарю – для ВОГиС).

При формальной оценке мы подсчитываем три параметра: 1) *точные совпадения* выделенных терминов с терминами предметного указателя, 2) *включение* однословных терминов «эталонного списка» в выделенные словосочетания и 3) *вхождение* выделенного словосочетания в более сложные (три и более слова) термины «эталонного списка». Для проведения автоматической оценки список подвергается ручной нормализации (см. [1]).

В предыдущем эксперименте большинство терминов-кандидатов из списков **freq** для наборов данных СОС и ФНМ уже были оценены экспертами, мы использованием эти результаты повторно. Тем же экспертам было предложено дооценить списки извлеченных терминов-кандидатов. Для третьего набора данных мы проводим оценку с нуля.

Экспертная оценка организована следующим образом. Сначала эксперту предъявляется краткое описание предметной области, а также несколько положительных и отрицательных примеров терминов для данной области. После этого эксперт, используя простой интерфейс, последовательно для каждого элемента списка отвечает на вопрос: «Является ли данное словосочетание термином предметной области?» Варианты ответа эксперта: «да», «нет» и «затрудняюсь ответить». Список предъявляется эксперту «порциями» по 10 словосочетаний, порядок предъявления словосочетаний – случайный. Каждый термин-кандидат оценивается независимо двумя экспертами в данной предметной области. В случае *сильной оценки* термином считается словосочетание, которое оба эксперта признали термином; в случае *слабой оценки* только один из экспертов оценил словосочетание как термин.

В итоге для каждого набора данных у нас есть 200 терминов списка **freq**, оцененных экспертами. Оценки ранжирования других списков (топ-100) основываются на этой оценке.

Результаты и обсуждение

Пример верхушек двух списков для корпуса ВОГиС приведен табл. 3. Результаты оценки списков (каждый из 100 словосочетаний), соответствующих трем наборам данных и пяти методам, приведены в табл. 4–6. (Некоторые расхождения оценок метода **freq** для СОС и ФНМ с оценками, приведенными в [1], вызваны изменениями при обработке слов с дефисом).

freq	**coherence**
экспрессия генов	индекс Кроу
настоящее время	случайный инбридинг
естественный отбор	подразделенная популяция
наследственная болезнь	микросателлитный локус
генетическое разнообразие	специфичность расщепления
фосфодиэфирная связь	одноцепочечный участок
стволовая клетка	микросателлитная изменчивость

этническая группа	химические рибонуклеазы
точка зрения	искусственная рибонуклеаза
окружающая среда	генное разнообразие

*Таблица 3. Top-10 списков **freq** и **coherence** (ВОГиС)*

Оценка		freq	iFreq	freq/iFreq	TF*IDF	coherence
Экспертная	слабая	90	83	87	89	88
	строгая	50	38	48	49	44
Полуавтоматическая	Точно	27	24	26	27	27
	включение	27	30	28	27	28
	вхождение	10	10	10	10	13

Таблица 4. Результаты оценки top-100 (СОС)

Оценка		freq	iFreq	freq/iFreq	TF*IDF	coherence
Экспертная	слабая	90	83	87	89	87
	строгая	68	51	66	66	63
Полуавтоматическая	Точно	29	26	31	34	31
	включение	42	44	44	39	42
	вхождение	18	11	17	17	13

Таблица 5. Результаты оценки top-100 (ФНМ)

Оценка		freq	iFreq	freq/iFreq	TF*IDF	coherence
Экспертная	слабая	90	83	78	81	85
	строгая	53	65	61	65	71
Полуавтоматическая	Точно	25	20	28	25	28
	включение	36	59	40	47	52
	вхождение	5	6	7	6	5

Таблица 6. Результаты оценки top-100 (ВОГиС)

Оценка Freq iFreq freq/iFreq TF*IDF coherence Экспертная слабая 70 83 78 81 85 строгая 53 65 61 65 71 Полу-автоматическая точно 25 20 28 25 28 включение 36 59 40 47 52 вхождение 5 6 7 6 5

Согласие экспертов (доля совпадающих оценок, включая «затрудняюсь ответить») для разных наборов данных выглядит следующим образом: СОС – 62%, ФНМ – 62,5%, ВОГиС – 79,5%. В среднем эксперты тратили 4,5 с на оценку одного словосочетания.

Как видно из табл. 4–6, методы, основанные на использовании поисковых машин интернета, ухудшают «базовый» результат (по строгой экспертной оценке) для наборов данных СОС и ФНМ, и существенно улучшают для корпуса ВОГиС. Это можно объяснить как «происхождением» данных (книга vs. сборник статей), так и особенностями терминологии соответствующих предметных областей и тематической структуры Сети. Так, эксперты при оценке результатов обработки ФНМ сочли за термины словосочетания *точка зрения, решение задачи, постановка вопроса, новая проблема*, которые были вытеснены из новых списков на основе статистики по Вебу. Тематическая несбалансированность интернета как корпуса выражается в том, что многие компьютерные термины имеют в сети высокую частоту. Например, в список **iFreq** (верхушка списка **freq**, переранжированного по возрастанию частоты словосочетания в интернете) не попали термины *операционная система, файловая система, адресное пространство, оперативная память, рабочая станция, база данных, программное обеспечение, имя файла*. Полученные результаты говорят о том, что использование Веба в качестве «контрастного корпуса» при извлечении терминов не подходит для любой предметной области. Хорошие результаты в случае корпуса ВОГиС подсказывают, что метод скорее всего будет работать для областей со специфичной терминологией (той, которая по большей части отлична от общеупотребительных выражений, редко использует выражения для универсальных понятий), к тому же – не очень широко представленных в Сети. «Затратность» метода *coherence* в слу-

чае ВОГиС оправдывается: в первую сотню попал 71 термин из 75 (по строгой экспертной оценке), содержавшихся в списке **freq-150**.

Еще один важный результат – согласованность формальной и экспертной оценок даже в том случае, когда «эталонный список» и термины-кандидаты получены из различных источников («Словарь терминов по молекулярной и клеточной биологии» / ВОГиС). Формальный метод оценки в случае ВОГиС дает нам возможность грубо оценить общее количество терминов соответствующей структуры (см. табл. 2) в корпусе. Сравнив все термины-кандидаты из корпуса (7 656, см. табл. 1) со словарем (7 227 терминов), мы получили следующие результаты: точных совпадений – 245, включений – 559, вхождений – 155. На основании данных табл. 6 можно предположить, что верхняя оценка общего количества терминов в корпусе лежит в районе 700, но с учетом того, что предложенные методы довольно хорошо ранжируют список терминов-кандидатов (соотношение можно выразить так: в 1/70 части полного списка находится примерно 1/10 точных совпадений со словарем), более реалистичной оценкой представляется 300-400 терминов. Это согласуется с такими выкладками: одна научная статья в журнале обычно аннотируется 5-10 ключевыми словами; в нашем корпусе – 100 статей; с учетом пересечения наборов ключевых слов разных статей в одном специализированном журнале, а также того, что в данном эксперименте мы ограничивались терминами определенной структуры, мы получаем примерно ту же оценку. Таким образом, зная оценку точности методов, мы можем оценивать длину списка, необходимую для обеспечения заданной полноты.

Как показывает эксперимент, оценка списка терминов требует относительно небольших затрат времени экспертов (предположительно, существенно меньших, чем ручное извлечение терминов из оригинальных документов). Так, в среднем на оценку 200 словосочетаний уходит 15 мин. Мы надеемся, что при рациональной комбинации автоматических методов извлечения терминов-кандидатов, промежуточной ручной оценки и последующего использования поисковых машин интернета (по аналогии с методами, описанными в [2, 4]) можно добиться существенно более высокого качества извлечения при минимальных затратах ручного труда.

В дальнейшем мы планируем использовать Веб и механизм поисковых машин для задачи полуавтоматического выделения связей между терминами.

Благодарности

Мы благодарим компанию Яндекс (www.yandex.ru) за предоставленную программу морфологического анализа *mystem* и возможность интенсивно использовать службу Яндекс.XML; Анастасию Барышникову – за предоставленный словарь; а также всех экспертов, которые приняли участие в оценке.

Список литературы

1. Браславский П.И., Соколов Е.А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог'2006. М.: Изд-во РГГУ, 2006. – С. 88–94.
2. Baroni M., Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web // Proceedings of LREC 2004. Lisbon: ELDA, 2004. – P. 1313–1316.
3. Peñas A., Verdejo F., Gonzalo J. Corpus-Based Terminology Extraction Applied to Information Access // Proceedings of Corpus Linguistics 2001, Lancaster University, UK, 2001. – P. 458–465.
4. Sato S., Sasaki Y. Automatic Collection of Related Terms from the Web // The Companion Volume to the Proceedings of 41st Annual Meeting of the ACL, Sapporo, Japan, 2003. – P. 121–124.
5. Turney P.D. Coherent Keyphrase Extraction via Web Mining // Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, 2003. – P. 434–439.