СРАВНЕНИЕ ЧЕТЫРЕХ МЕТОДОВ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ДВУХСЛОВНЫХ ТЕРМИНОВ ИЗ ТЕКСТА

Павел Браславский, Евгений Соколов

Институт машиноведения УрО РАН, Екатеринбург

pb@imach.uran.ru, esokolov@list.ru

В статье рассматриваются четыре метода автоматического извлечения двухсловных терминов из текста на основе статистики встречаемости и морфологических шаблонов. Приведены результаты работы методов на двух текстах разных предметных областей. 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 ходе работ по созданию метапоисковой системы *ProThes* [5] быстро обнаружилось узкое место, сдерживающее развитие подхода, – ручное создание и поддержка тезаурусов предметной/научной области. Таким образом, мы обратились к задаче разработки «легких» инструментальных средств для полуавтоматического создания тезаурусов узкой научной/предметной области. Исходными данными для таких инструментов должны быть относительно небольшие тематические коллекции документов.

В данной работе мы описываем эксперименты, направленные на решение одной узкой задачи – автоматического выделения двухсловных терминоподобных конструкций. Близкой задачей является задача выделения устойчивых словосочетаний (*collocations*) [7].

На основе знакомства с литературой можно выделить два основных подхода к выделению терминов: 1) на основе шаблонов [1, 4] и 2) статистики встречаемости (см. обзор в [7]). Некоторые методы являются объединением этих подходов (например, [8]). Многие методы ориентируются на *пополнение* существующих терминологических ресурсов, т.е. исходят из наличия готового словаря, тезауруса или списка терминов [2, 6]. Некоторые методы автоматического построения тезаурусов решают одновременно задачи выделения терминов и связей между ними.

В нашей работе мы сравниваем четыре простых метода для выделения двухсловных терминов-кандидатов, которые используют минимум исходной информации: 1) статистику встречаемости пар и отдельных слов в тексте (коллекции) и 2) некоторые предположения о структуре двухсловных терминов.

Методы

В этой работе сравниваются четыре метода выделения терминов, которые являются модификацией методов автоматического выделения двусловий (*bigrams*), описанных в [7]:

1. прямой подсчет количества пар (**freq**);
2. **t-тест**;
3. **χ^2-тест**;
4. отношение функций правдоподобия (**LR**).

Первый из методов использует простейшую технику – двусловия упорядочиваются по убыванию их встречаемости в тексте (т.е. частоты встречаемости отдельных слов не учитываются). Последние три метода заключаются в проверке статистических гипотез, соответствующих случайной или неслучайной «встрече» слов в паре. Проверка основана на подсчете частоты отдельных слов и пар. На практике вычисленные статистики используются не для принятия/отвержения гипотез (иначе пришлось бы «принять» в качестве устойчивых словосочетаний большинство – так проявляется «неслучайная» природа речи), а для упорядочения словосочетаний-кандидатов.

Во втором подходе (t-тест) используется t-статистика Стьюдента для сравнения теоретического и выборочного среднего:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$, где

$\bar{x}$ – выборочное среднее;

μ – теоретическое среднее;

s^2 – выборочная дисперсия;

N – размер выборки.

В соответствии со схемой Бернулли, в качестве теоретического среднего (соответствует гипотезе о случайном образовании двусловия) берется произведение вероятностей появления отдельных слов, составляющих двусловие; в качестве выборочного

среднего – вероятность появления двусловия. Дисперсия распределения Бернулли $s^2 = p(1-p) \approx p$ (для малых значений p). Двусловия упорядочиваются по убыванию значения t.

В третьем методе используется χ^2-критерий Пирсона для анализа таблиц сопряженности 2x2. Четыре значения, формирующие таблицу, – это 1) частота данного двусловия, 2) частота двусловий с участием первого слова (но не второго), 3) частота двусловий с участием второго слова (но не первого), и 4) частота всех остальных двусловий. В качестве меры расхождения берется значение:

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i^* - n_i)^2}{n_i}, \text{ где}$$

n_i* – наблюдаемая частота;

n_i– ожидаемая частота (в соответствии с предположением о случайности сочетания слов).

В качестве ожидаемых значений берутся маргинальные частоты. Двусловия упорядочиваются по убыванию значения χ^2.

Наконец, в четвертом методе используется отношение функций правдоподобия, соответствующих двум гипотезам – о случайной и неслучайной природе двусловия. Логарифм отношения функций правдоподобия выглядит следующим образом:

$$\log \lambda = \log \frac{b(c_{12}, c_1, p) b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1) b(c_2 - c_{12}, N - c_1, p_2)}, \text{ где}$$

$b(k, n, x)$ – значение биномиального распределения для k успешных исходов в n независимых испытаниях при вероятности успешного исхода в каждом испытании, равном x;

c_1 – частота первого слова двусловия;

c_2 – частота второго слова двусловия;

c_{12} – частота двусловия;

N – длина текста;

$p = c_2/N$;

$p_1 = c_{12}/c_1$;

$p_2 = (c_2 - c_{12})/(N - c_1)$.

Двусловия упорядочиваются по возрастанию значения $\log \lambda$.

Обсуждение особенностей этих четырех методов (включая применимость к различным объемам данных и диапазонам вероятностей, а также предположения о свойствах выборочных распределений) можно найти в [7].

Необходимо дополнительно отметить, что, в отличие от примеров, приведенных в [7], мы учитывали разделители (знаки препинания и стоп-слова) при формировании списка пар слов, а также применяли методы к текстам значительно меньшего объема.

Морфологические шаблоны

Основная модификация методов заключается в предварительном использовании морфологических шаблонов-фильтров. Мы выделили пять шаблонов (Табл. 1), которые являлись фильтром для словосочетаний, подлежащих анализу. Морфологическая обработка осуществлялась с помощью программы *mystem*[*]; при неоднозначности морфологического разбора мы требовали совпадения хотя бы одного из возможных сочетаний с шаблоном.

Таблица 1. Морфологические шаблоны

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

Ясно, что, ограничиваясь двухсловными словосочетаниями определенного вида, мы не можем рассчитывать на очень высокую полноту: например, в [3] показано, что номинативность не является исключительной характеристикой терминов во многих предметных областях.

Данные

Набор методов был применен к электронным версиям двух книг:

1. Олифер Н.А., Олифер В.Г. Сетевые операционные системы. СПб.: Питер, 2005.
2. Щедровицкий Г.П. Философия. Наука. Методология. М.: ШКП, 1989.

Тексты относятся к разным областям знаний, что позволило проверить гипотезу о независимости методов от научной/предметной области.

Первая книга является монографией, описывающей достаточно узкую предметную 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.

Важно, что в обеих книгах есть предметный указатель (ПУ), который мы принимаем за список терминов, выделенных автором, и используем на этапе проверки методов.

Тексты анализировались в формате *plain text*. При анализе текста (1) было обработано 99337 отдельных слов (включая стоп-слова) и 9897 пар; при анализе текста (2) было обработано 180048 слов (включая стоп-слова) и 12694 пар.

Результаты эксперимента

Результатом эксперимента являются четыре списка словосочетаний, упорядоченных по убыванию параметра, отражающего их «устойчивость», для каждого из двух текстов.

Верхушки этих списков приведены в Табл. 2 и 3. Как видно из приведенных данных, топ-10, полу-

[*] См. http://corpora.narod.ru/mystem

ченных методами freq и t-тест, не отличаются совсем (Табл. 2) или незначительно отличаются ранжированием (Табл. 3). Наиболее «контрастный» набор – список, полученный с помощью метода χ^2. Характерно, что в верхушку списка χ^2 попали словосочетания, элементы которых не встречаются в других контекстах.

Табл. 4 и 5 дают более полное представление о похожести списков – в них указаны доли попарных пересечений в топ-100 соответствующих списков для текста (1) и (2) соответственно.

Таблица 2. Топ-10 терминов-кандидатов, «Сетевые операционные системы»

freq, t-тест	**LR**	**χ^2**
операционная система	операционная система	Карнеги Меллон
файловая система	файловая система	ввод-вывод
адресное пространство	адресное пространство	накладные расходы
ввод-вывод	ввод-вывод	грамматический разбор
оперативная память	рабочая станция	оранжевая книга
рабочая станция	оперативная память	доска объявлений
системный вызов	база данных	адресное пространство
база данных	системный вызов	рабочая станция
право доступа	критическая секция	Денис Ритчи
программное обеспечение	программное обеспечение	критическая секция

Таблица 3. Топ-10 терминов-кандидатов, «Философия. Наука. Методология»

freq	**t-тест**	**LR**	**χ^2**
процесс мышления	процесс мышления	процесс мышления	филиал ВНИИТЭ
процесс мысли	процесс мысли	суть дела	Миклухо-Маклай
знаковая форма	знаковая форма	знаковая форма	родимое пятно
суть дела	суть дела	сия пора	Павлик Морозов
научное мышление	картина мира	картина мира	категорический императив
картина мира	математическое отношение	математическое отношение	экологическая ниша
математическое отношение	научное мышление	целый ряд	древние греки
научный предмет	научный предмет	процесс мысли	бочка портвейна
методологическая работа	методологическая работа	онтологическая картина	конная армия
целый ряд	целый ряд	единая картина	уральский филиал

Выборочный анализ результатов показывает, что наряду с «хорошими» терминами-кандидатами в списках присутствуют, например, имена (*Денис Ритчи*), общеупотребительные устойчивые словосочетания (*суть дела*, *целый ряд*, *сия пора*), а также части более крупных терминов (*единая картина* → *единая картина мира*; *Карнеги Меллон* → *университет Карнеги Меллона*).

Таблица 4. Пересечение топ-100 списков, «Сетевые операционные системы»

	freq	**t-тест**	**χ^2**	**LR**
freq	1	0,93	0,25	0,73
t-тест	0,93	1	0,26	0,77
χ^2	0,25	0,26	1	0,39
LR	0,73	0,77	0,39	1

Таблица 5. Пересечение топ-100 списков, «Философия. Методология. Наука»

	freq	**t-тест**	**χ^2**	**LR**
freq	1	0,94	0,17	0,71
t-тест	0,94	1	0,19	0,75
χ^2	0,17	0,19	1	0,26
LR	0,71	0,75	0,26	1

Методика оценки

Важной составной частью эксперимента является методика оценки процедуры извлечения терминов. Мы предлагаем использовать методику, объединяющую 1) полуавтоматическую оценку и 2) экспертную оценку.

Для полуавтоматической оценки в качестве образца мы используем предметный указатель, помещаемый в конце книги. Мы подсчитываем три параметра: 1) *точные совпадения* выделенных терминов с терминами предметного указателя, 2) *включение* однословных терминов ПУ в выделенные словосочетания и 3) *вхождение* выделенного словосочетания в более сложные (три и более слова) термины ПУ.

Для экспертной оценки формируется список терминов, образованный слиянием верхушек списков, полученных разными методами, с добавлением двухсловных терминов из предметного указателя (так мы хотим дополнительно оценить терминологичность элементов предметного указателя с точки зрения эксперта для валидации полуавтоматической оценки). Из-за ограниченности ресурсов мы используем объединение топ-100 четырех списков для экспертной оценки. Эксперту предъявляется краткое

описание предметной области (абзац), а также положительные и отрицательные примеры терминов для данной области. После этого эксперт последовательно для каждого элемента списка отвечает на вопрос: «Является ли данное словосочетание термином предметной области?» Варианты ответа эксперта: «да», «нет» и «затрудняюсь ответить». Порядок предъявления словосочетаний из списка эксперту – случайный. Объединенный список (для каждого из текстов) оценивается минимум двумя экспертами.

Данные для оценки

Для полуавтоматической оценки методов выделения терминов необходимо было нормализовать термины предметных указателей двух книг. Частично такая нормализация включала принятие решения, является ли элемент предметного указателя термином. В бОльшей степени это касалось предметного указателя книги «Философия. Методология. Наука», который наряду со специальными терминами включает обозначения наиболее общих философских категорий (*время, наука* и т.п.), а также словосочетания, которые не являются терминами (*цель методологии, понятие металла, понятие объекта, связь логики с мышлением, связь логики с деятельностью, проблема объекта знания в логике* и др.). В качестве примера из книги «Сетевые операционные системы» можно привести элемент предметного указателя *эволюция операционных систем*, который отсылает к разделу книги, описывающему основные этапы развития операционных систем.

Очевидно, что исключение некоторых элементов предметного указателя в рамках нашей методики может только занизить оценки автоматических методов.

Фрагмент предметного указателя книги Г.П. Щедровицкого и соответствующий ему нормализованный список терминов представлены на Рис. 1, 2.

Фрагмент предметного указателя книги «Сетевые операционные системы» и соответствующий ему нормализованный список терминов представлены на Рис. 3, 4. Слово *система* не было внесено в список как общее слово, которое, к тому же, не является отсылкой к конкретной странице книги.

Топ-100 терминов-кандидатов каждого из методов автоматически сравнивались с полным нормализованным ПУ, как описано выше.

Деятельность
  как идеальный предмет изучения
  как объект изучения
  как структура
  воспроизводство
  носитель
  замещающая
  практическая и познавательная
  и рефлексия

Рис. 1. Фрагмент предметного указателя, «Философия. Методология. Наука»

деятельность
предмет изучения
объект изучения
структура
воспроизводство деятельности
носитель деятельности
замещающая деятельность
практическая деятельность
познавательная деятельность
рефлексия

Рис. 2. Нормализованное представление фрагмента, «Философия. Методология. Наука»

система
  аутентификации, 493
  дисковая, 357
  защиты данных, 226
  реального времени, 92
    жесткая, 119
    мягкая, 119
  удаленного ввода заданий, 16
  файловая, 15, 35, 357
  шифрования, 482

Рис. 3. Фрагмент предметного указателя, «Сетевые операционные системы»

система аутентификации
дисковая система
система защиты данных
система реального времени
жесткая система реального времени
мягкая система реального времени
система удаленного ввода заданий
файловая система
система шифрования

Рис. 4. Нормализованное представление фрагмента, «Сетевые операционные системы»

Списки для экспертной оценки были получены объединением топ-100 каждого из методов и ста двухсловных терминов из нормализованного ПУ, выбранных случайным образом. Список, соответствующий книге «Сетевые операционные системы», включал 272 элемент, книге «Философия. Методология. Наука» – 281. Каждый из списков оценивался двумя экспертами.

Результаты оценки

Результаты сравнения топ-100 каждого из методов с нормализованными предметными указателями приведены в Табл. 6, 7.

Результаты экспертной оценки приведены в Табл. 8, 9 («строгая оценка» соответствует случаям, когда оба эксперта давали положительную оценку, «слабая оценка» – хотя бы один из экспертов дал положительную оценку). Интересно отметить, что показатели согласия экспертов (доля совпадающих оценок) значительно различаются: 44% – для книги «Сетевые операционные системы» и 77% – для книги «Философия. Методология. Наука».

Таблица 6. Результаты формальной оценки с использованием предметного указателя, «Сетевые операционные системы»

	точное совпадение	включение	вхождение
freq	27	18	23
t-тест	27	19	28
χ^2	14	12	12
LR	27	15	27

Таблица 7. Результаты формальной оценки с использованием предметного указателя, «Философия. Методология. Наука»

	точное совпадение	включение	вхождение
freq	29	19	26
t-тест	29	20	26
χ^2	2	5	4
LR	21	17	20

Таблица 8. Результаты экспертной оценки, «Сетевые операционные системы»

	строгая оценка	слабая оценка
freq	38	83
t-тест	36	84
χ^2	14	57
LR	29	80
ПУ	35	85

Таблица 9. Результаты экспертной оценки, «Философия. Методология. Наука»

	строгая оценка	слабая оценка
freq	62	83
t-тест	58	80
χ^2	14	36
LR	47	72
ПУ	79	92

Заключение

Результаты эксперимента позволяют сделать вывод, что методы **freq** и **t-тест** сравнимы по эффективности и могут быть использованы для составления списка терминов-кандидатов в задачах полуавтоматического формирования терминологических ресурсов. Повышение качества этих методов может быть достигнуто за счет удаления устойчивых словосочетаний общей лексики. Эту задачу можно решить с помощью дополнительного «контрастного» корпуса (в качестве универсального корпуса можно использовать Веб).

Оценка методов с помощью предметных указателей демонстрирует, что для комплексного решения задачи выделения терминов из текста необходимо учитывать термины разной длины и структуры.

Результаты сравнения методов на основе формальной и экспертной оценок хорошо согласуются.

Сравниваемые методы доставляют схожие результаты для различных предметных областей.

Благодарности

Мы благодарим компанию Яндекс за предоставленный модуль морфологического анализа *mystem*.

Мы благодарим экспертов, которые приняли участие в оценке методов.

Литература

1. Васильева Н.Э. Шаблоны употреблений терминов и их использование при автоматической обработке научно-технических текстов// Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог’2004. («Верхневолжский», 2-7 июня 2004 г.). М., 2004. С. 96-101.
2. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических сочетаний по текстам предметной области // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды пятой Всероссийской научной конференции (С.-Петербург, 29-31 октября 2003 г.), 2003. С. 201–210.
3. Шелов С.Д. Терминоведение: семь вопросов и семь ответов по семантике термина // НТИ. Сер. 2. Информационные процессы и системы, 2001. №2. С. 1-11.
4. Bourigault D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases // Proc. of COLING-92, Nantes, France, August 23-28, 1992. P. 977-981.
5. Braslavski P., Shishkin A., Alshanski G. 3 in 1: Meta-Search, Thesaurus, and GUI for Focused Web Information Retrieval// Digital Libraries: Advanced Methods and Technologies, Digital Collections. Proceedings of the 6th National Russian Research Conference, September 29 - October 1, 2004, Pushchino. P. 135-140.
6. Jacquemin C. A Symbolic and Surgical Acquisition of Terms Through Variation // Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing. Heidelberg: Springer, 1996. P. 425-438.
7. Manning C., Schütze H. Collocations// Manning C., Schütze H. Foundations of Statistical Natural Language Processing, 2002. P. 151-189.
8. Smaja F. Retrieving Collocations from Text: Xtract // Computational Linguistics, 1993. № 19(1). P. 143-177.