УДК 004.853

КОМБИНИРОВАНИЕ ПРИЗНАКОВ ДЛЯ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ТЕРМИНОВ

Н. В. Лукашевич[1], **Ю. М. Логачев**[1]

В статье описывается эксперимент по извлечению двухсловных терминологических словосочетаний на основе комбинирования различных признаков этих словосочетаний. Признаки вычисляются на основе трех источников: статистики употребления слов в текстовой коллекции предметной области, выдачи глобальных поисковых машин и тезауруса предметной области. Для оценки качества извлечения терминов используются терминологические словосочетания из онтологии по естественным наукам и технологиям ОЕНТ. Показано, что использование совокупности признаков словосочетаний значительно улучшает извлечение терминов.

Ключевые слова: извлечение знаний из текстов, извлечение терминов, тезаурус, машинное обучение, поисковая система, Интернет.

1. Введение. Одним из актуальных направлений в области автоматической обработки текстов и информационного поиска являются исследования, посвященные методам автоматического извлечения терминов из текстов предметной области. Особую сложность представляет собой автоматическое извлечение многословных терминов — терминологических словосочетаний.

Как известно, под термином понимается слово (или сочетание слов), являющееся точным обозначением определенного понятия какой-либо специальной области науки, техники, искусства, общественной жизни и т.п. [9]. Понятно, что такое определение невозможно применить для автоматического извлечения терминов из текстов, поэтому на практике для отбора терминов применяется некоторый набор лингвистических и статистических характеристик словосочетаний.

Существующие методы автоматического извлечения устойчивых словосочетаний, терминов обычно используют некоторое сочетание следующих факторов [7, 12]:

— статистические характеристики употребления словосочетания и его компонентов (частотность по коллекции, взаимная ассоциация, вхождение в объемлющие словосочетания и т.п.);

— синтаксические ограничения: извлекаются словосочетания заданной синтаксической структуры (прилагательное + согласованное существительное, существительное + существительное в родительном падеже и др.);

— лексические фильтры (например, не извлекаются словосочетания, включающие в себя географические названия, эмоциональную лексику и др.) [10].

В свою очередь, статистические характеристики могут быть разделены на два подвида. Во-первых, используются характеристики, которые определяют единство употребления слов словосочетания, для чего используются такие статистические меры, как мера взаимной информации (mutual information), тест Стьюдента (t-test), функция логарифмического правдоподобия — log likelihood [3, 19].

Второе направление использования статистических мер для определения терминологичности выражений связано с автоматическим определением принадлежности слова или словосочетания к заданной предметной области. Для этого применяются характеристики, сопоставляющие частотности употребления словосочетания (слова) в исходной коллекции и в некотором контрастном корпусе. Для сопоставления частотностей используются такие статистические меры, как классическая мера информационного поиска — tf.idf [15], сопоставляющая частотность в заданной коллекции и встречаемость в документах контрастной коллекции, а также характеристика странности — weirdness [14], которая сопоставляет частотную долю употребления выражения в коллекции по сравнению с употреблением в контрастной коллекции.

В последнее время как дополнительные факторы используются статистические меры, основанные на запросах в Интернет [4], а также структура общеязыковых тезаурусов [17].

Существующие методы извлечения терминологических словосочетаний приводят, по большому счету, к одному и тому же результату. Обычно в результате работы программы отбора словосочетаний порождается список, упорядоченный по весу в соответствии с заложенной моделью. Верхняя часть такого списка

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, 119991, Москва; Н. В Лукашевич, ст. науч. сотр., e-mail: louk_nat@mail.ru; Ю. М. Логачев, студент ф-та ВМК МГУ, e-mail: yulogachev@gmail.com

наполнена общеизвестными терминами предметной области. Далее доля очевидных терминов резко снижается и большинство начинают составлять словосочетания, для которых очень трудно решить, являются ли они терминами — для этого требуется серьезный дополнительный анализ. Поэтому качество методов автоматического извлечения терминологических словосочетаний достаточно трудно оценивать [7]. Привлеченные эксперты часто дают очень противоречивые оценки [4].

При работе с большими предметными областями и корпусами даже лучшие методы извлечения терминологических словосочетаний показывают значительное падение процентного содержания терминов с 90% на первой сотне списка извлеченных терминологических словосочетаний до 60% на третьей тысяче [7]. Таким образом, актуальной является разработка новых методов по улучшению качества упорядочения списков словосочетаний с целью повышения процентной доли терминов в начале списка.

Одним из потенциальных направлений улучшения качества извлечения терминологических словосочетаний является комбинирование имеющихся статистических мер [16, 19]. Так, в работе [16] комбинирование различных статистических признаков словосочетаний, извлеченных на основе корпуса чешского языка, применяется для извлечения разных типов устойчивых словосочетаний. Авторы работы используют более 80 различных признаков и посредством их комбинации получают улучшение извлечения устойчивых словосочетаний более, чем на 20% по сравнению с лучшим отдельным признаком.

В настоящей статье мы опишем эксперимент по извлечению двухсловных терминологических словосочетаний на основе комбинирования трех видов признаков:

— признаки, построенные на основе текстовой коллекции предметной области;

— признаки, полученные на основе информации глобальной поисковой машины;

— признаки, полученные на основе заданного тезауруса предметной области; здесь мы моделируем ситуацию развития существующего тезауруса и хотим выяснить, насколько знания, описанные в текущей версии тезауруса, могут улучшить качество автоматического извлечения следующих терминов.

Эксперимент проводится в широкой области естественных наук. Для оценки качества извлеченных словосочетаний используется терминология онтологии по естественным наукам и технологиям, которая создается вручную [8].

2. Набор словосочетаний для извлечения терминов. В качестве базы для экспериментов по упорядочению терминологических словосочетаний мы используем словосочетания, которые были автоматически извлечены из публикаций по естественным наукам в начале работы над онтологией OEHT [8].

Онтология OEHT представляет собой так называемую лингвистическую онтологию, т.е. онтологию, понятия в которой основаны на значениях существующих языковых выражений — в данном случае естественнонаучных терминов. Одновременно онтология OEHT может рассматриваться как тезаурус, поскольку описывает формализованные отношения между терминами предметной области [13].

По структуре онтология OEHT — это иерархия понятий, которые связаны между собой отношениями. К каждому понятию приписан набор текстовых входов, т.е. терминов, посредством которых данное понятие может быть упомянуто в тексте. Набор отношений между понятиями, используемый в онтологии, совпадает с набором отношений, применяемых при создании тезауруса русского языка РуТез [11].

В настоящее время онтология OEHT включает в себя 56 тысяч понятий и 150 тысяч различных терминов математики, химии, физики, геологии, биологии. На первом шаге разработки онтологии OEHT (в 2004 г.) мы собрали текстовые коллекции по таким наукам, как математика, физика, химия, геология (от 3000 до 8000 документов, от 50 до 90 Мб по каждой из наук), извлекли из них терминологические словосочетания-кандидаты (более 600 тысяч словосочетаний), и наиболее частотные их них (60 тысяч словосочетаний) стали одним из источников терминологии онтологии OEHT.

Автоматически извлеченные слова и словосочетания из собранных коллекций текстов можно использовать для тестирования различных методов извлечения и упорядочения терминов-кандидатов. Входящие в текущий терминологический состав онтологии OEHT термины могут служить хорошей основой для оценки качества методов. Это позволяет получить достаточно объективные оценки качества каждого метода, что очень важно в связи с проблемой субъективности экспертных оценок [4].

Эксперименты проводились на 5 тысячах наиболее частотных двухсловных словосочетаний списка. Задачей применения методов является переупорядочение исходного списка словосочетаний (первоначально упорядоченного по мере снижения частотности) так, чтобы в начало списка попало как можно больше словосочетаний-терминов. Таким образом, наилучшее переупорядочение списка снизит трудозатраты эксперта по вводу терминов в онтологию — эксперт будет меньше просматривать словосочетания, не являющиеся терминами.

В табл. 1 приведены примеры словосочетаний из исходного списка. Словосочетания упорядочены по частотности их встречаемости в текстовой коллекции.

Начало списка в таблице содержит достаточно большое число общеизвестных терминов. Группа менее частотных словосочетаний содержит как явные нетермины (*пересечение окружности, современная химия, задача теории*), так и словосочетания, для которых сложно определить, являются ли они терминами (*перпендикулярная ось*).

Таблица 1

Примеры словосочетаний из исходного списка

Наиболее частотные словосочетания	Словосочетания средней частотности
программа образования	пересечение окружности
решение задачи	итерация метода
случайная величина	задача теории
земная кора	угловой размер
точные науки	абсолютный возраст
система координат	обязательный минимум
собственное значение	подъемная сила
горная порода	перпендикулярная ось
дифференциальное уравнение	современная химия

3. Наборы характеристик словосочетаний. Для извлеченных словосочетаний вычисляются признаки трех типов.

Во-первых, используются признаки, вычисленные на основе исходной текстовой коллекции предметной области. В данном эксперименте — это различные соотношения частот слов и словосочетаний, извлеченных из документов коллекции.

Во-вторых, мы используем признаки, полученные на основе выдачи глобальной поисковой машины по специально сформулированным запросам. В качестве базы для подсчета признаков используется информация о встречаемости слов в 100 первых сниппетах поисковой машины Яндекс. В терминологии информационного поиска сниппет — это краткий фрагмент текста, используемый поисковыми сервисами в качестве краткой аннотации найденного документа.

Кроме того, мы предполагаем, что термины нам нужны для пополнения тезауруса предметной области, и некоторая часть тезауруса у нас уже есть. Поэтому третий тип используемых нами признаков состоит в учете существующей части тезауруса для оценки терминологичности словосочетаний-кандидатов. Рассмотрим способы вычисления этих типов признаков подробнее.

3.1. Признаки, полученные по коллекции документов предметной области. Признак частотности словосочетаний в коллекции `Freq` часто используется для извлечения терминологических словосочетаний, поскольку известно, что в число наиболее частотных словосочетаний коллекции предметной области входит достаточно высокий процент терминологических словосочетаний.

Признак `взаимная информация слов MI` словосочетания обычно вычисляется по формуле

$$\mathtt{MI}(ab) = \log\left(\frac{N \cdot \mathtt{freq}(ab)}{\mathtt{freq}(a) \cdot \mathtt{freq}(b)}\right),$$

где `freq`( ) — частотность слов и словосочетаний в коллекции, N — число слов в коллекции. Признак показывает, насколько употребление слов в словосочетании отличается от их независимого употребления.

Признак $\mathtt{MI}_3$ является модификацией признака `MI` и вычисляется по формуле следующего вида:

$$\mathtt{MI}_3(ab) = \log\left(\frac{N \cdot \mathtt{freq}^3(ab)}{\mathtt{freq}(a) \cdot \mathtt{freq}(b)}\right).$$

В корпусных исследованиях данная модификация предложена как оптимально сочетающая частотность словосочетания и взаимную информацию слов [15].

Признак `усеченное словосочетание Inside` предназначен для выявления двухсловных словосочетаний, которые являются частью более длинного термина. Значение признака для словосочетания ab определяется следующим образом. Среди всех словосочетаний, извлеченных из коллекции документов, таких, что ab является частью этого (более длинного) словосочетания, выбирается словосочетание $*ab*$ с максимальной частотностью. Тогда $\mathtt{Inside}(ab) = \dfrac{\mathtt{freq}(*ab*)}{\mathtt{freq}(ab)}$. Так, например, частотность словосочетания *равенство треугольников* равна 273, при этом объемлющее словосочетание *признак равенства треугольников* имеет частотность, равную 217. Таким образом, значение `Inside` для словосочетания *равенство треугольников* равно $217/273 = 0.79$. Для словосочетания *геологические работы* в собранной коллекции текстов не нашлось ни одного более длинного словосочетания. Поэтому значение признака `Inside` равно 0.

3.2. Признаки, полученные по сниппетам глобальной поисковой машины. Обращение за дополнительной информацией к глобальным поисковым машинам важно по нескольким причинам.

Во-первых, коллекции документов широкой предметной области всегда недостаточно, поскольку множество достаточно значимых терминов предметной области может иметь относительно низкую частотность в данной коллекции. Привлечение Интернета помогает получить дополнительную информацию по таким словосочетаниям. Во-вторых, использование информации из Интернета позволяет выяснить, насколько употребление данного словосочетания жестко связано с заданной предметной областью.

Наконец, обращение в Интернет — это достаточно простой путь получения контекстов употребления словосочетания. В качестве таких контекстов мы используем сниппеты (аннотации документов в выдаче), получаемые от поисковой машины Яндекс через xml-интерфейс. Для получения сниппетов поисковой машине задавались запросы в виде самого словосочетания, а также запросы в виде его отдельных слов-компонентов. Например, при анализе словосочетания *инверсионная ось*, задаются поисковые запросы *инверсионная ось*, *инверсионная*, *ось*.

Таблица 2

Самые частотные леммы сниппетов для словосочетания *инверсионная ось*

По сниппету *инверсионная ось*		По сниппету *инверсионная*		По сниппету *ось*	
ось	98	инверсионный	176	ось	200
симметрия	95	применение	10	вращение	25
инверсионный	62	мочь	8	прямой	9
порядок	36	температура	7	можно	7
центр	34	иметь	6	тело	7
плоскость	24	основание	6	центр	7
инверсия	17	представлять	5	поворот	7
отражение	13	главный	5	угол	7
поворот	13	являться	4	мочь	6
класс	11	рис	3	называть	6
четвертый	11	ряд	3	плоскость	6

Для вычисления разных признаков использовалось по 100 сниппетов из выдачи. Сниппеты, получаемые по одному запросу, объединяются в один документ и обрабатываются программой морфологического анализа. В результате для каждого набора сниппетов может быть определена совокупность лемм (т.е. слов в словарной форме) и их частотность встречаемости в данном наборе сниппетов. Наборы сниппетов формируются для словосочетания в целом и для составляющих его слов. Обозначим S_{ab} — вектор лемм с частотностями, полученных из сниппетов словосочетания, S_a и S_b — векторы лемм из сниппетов слов-компонентов словосочетания. В векторы лемм не включаются служебные части речи. В качестве примера рассмотрим словосочетание *инверсионная ось*. Для него было получено три набора сниппетов. По каждому из наборов была посчитана статистика употребления отдельных лемм и получены векторы S_{ab}, S_a и S_b. В табл. 2 показаны наиболее частотные леммы каждого из этих трех векторов (с указанием частотности).

Далее рассмотрим подробнее признаки словосочетаний, вычисляемых по сниппетам.

3.2.1. Признаки векторного сравнения сниппетов. В стандартах по разработке информационно-поисковых тезаурусов считается, что одним из важных факторов является возможность внесения в тезаурус таких словосочетаний предметной области, значения которых не следуют из значений их компонентов [6]. Мы предполагаем, что такая семантическая особенность словосочетания может проявляться в контекстах употребления данного словосочетания.

Для сопоставления контекстов употребления словосочетания и составляющих его слов мы используем векторы S_{ab} лемм, полученных для словосочетания, и для его отдельных компонентов S_a и S_b. Сравнение векторов сниппетов производится с помощью вычисления скалярного произведения векторов и фиксируется в признаках $\texttt{Scalar}_1$ и $\texttt{Scalar}_2$:

$$\texttt{Scalar}_1 = \frac{(S_{ab}, S_a)}{\|S_{ab}\|\,\|S_a\|}, \quad \texttt{Scalar}_2 = \frac{(S_{ab}, S_b)}{\|S_{ab}\|\,\|S_b\|}.$$

При замене частотностей лемм в векторах S_{ab}, S_a и S_b на булевские признаки $\{0, 1\}$ в зависимости от присутствия или отсутствия леммы в сниппетах получаются булевские векторы и на их основе вычисляются соответствующие скалярные произведения — признаки $\texttt{BinarScalar}_1$ и $\texttt{BinarScalar}_2$. Признаки в виде бинарных скаляров показали высокую эффективность для извлечения устойчивых словосочетаний в [16].

3.2.2. Признаки максимально отличающегося контекста. Другим способом определения специфики употребления словосочетания является нахождение одного характерного слова, с которым чаще всего совместно встречается это словосочетание. Мы предполагаем, что если значение словосочетания не выводимо из значений его компонент, то это может проявиться в том, что это словосочетание употребляется в сниппетах рядом с такими словами, с которыми мало употребляются отдельные слова исходного словосочетания. Кроме того, мы считаем, что словосочетание имеет тем бóльшую семантическую

особенность, чем больше максимальная разница между частотностью употребления некоторой леммы в сниппетах словосочетания по сравнению с употреблением этой же леммы в сниппетах отдельных слов.

Пусть лемма L встречается f_{ab} раз в сниппетах словосочетания S_{ab}, f_a — в сниппетах первого слова словосочетания S_a, f_b — в сниппетах второго слова словосочетания S_b. Тогда коэффициент устойчивости $\texttt{SnipFreq}_0$ вычисляется по формуле

$$\texttt{SnipFreq}_0 = \max_L f_{ab-a-b} \log\left(\frac{N - d_{\text{Lcol}}}{d_{\text{Lcol}}}\right),$$

где $f_{ab-a-b} = \max(f_{ab} - f_a - f_b, 0)$. Лемма L не совпадает с леммами a и b и не является однокоренным словом для слов a и b, d_{Lcol} — это количество документов, в которых встречалась лемма L в некоторой объемлющей коллекции. Множитель $\log\left(\frac{N - d_{\text{Lcol}}}{d_{\text{Lcol}}}\right)$ представляет собой известный в информационном поиске множитель `idf`, который учитывает частотность употребления слова в документах объемлющей коллекции, N — размер коллекции. Этот коэффициент позволяет снизить вес частотных общеупотребительных слов.

Данные по количеству документов в множестве индексирования поисковиков постоянно меняются, количество документов, в которых употребляется то или иное слово, также точно не известно (число, которое выдается поисковиком на странице поиска, является лишь некоторым приближением). Поэтому в качестве объемлющей коллекции была взята коллекция Университетской информационной системы Россия (www.cir.ru), которая в настоящее время включает в себя более 2 млн. документов.

Наиболее высокие значения признака SnipFreq0 для рассматриваемой в эксперименте совокупности словосочетаний получили словосочетания *инверсионная ось* (наиболее характерное слово — *симметрия*), *капельная модель* (*ядро*), *характеристический корень* (*уравнение*).

В некоторых случаях двухсловное сочетание представляет собой несамостоятельный фрагмент более длинного словосочетания, и тогда такое двухсловное сочетание имеет наиболее высокую сочетаемость с остальными словами этого словосочетания. Для учета такой ситуации та же формула использовалась, чтобы посчитать наиболее характерное слово на расстоянии более чем 1 и 2 слова от исходного словосочетания (признаки $\texttt{SnipFreq}_1$, $\texttt{SnipFreq}_2$).

3.2.3. Частотность упоминания словосочетания в собственных сниппетах. Признак частотности упоминания словосочетания в собственных сниппетах `FreqbySnip` может отражать различные особенности словосочетания. Если значение этого признака значительно меньше 100, то это означает, что поисковая машина не находит такое словосочетание в Интернет и, таким образом, это словосочетание, возможно, ошибочно извлечено при обработке коллекции (например, за счет неправильной лемматизации или неточной обработки таблиц в исходных документах). Если же значение этого признака значительно больше 100 (иногда этот признак достигает величины 250–300 на 100 сниппетах), то это означает, что имеется множество контекстов, в которых это словосочетание подробно объясняется, является темой фрагмента, и, скорее всего, это словосочетание означает важное понятие или конкретную сущность.

3.2.4. Количество слов-определений в сниппетах. Смысл признака количество слов-определений в сниппетах `Neardefwords` заключается в том, что если в сниппетах рядом со словосочетанием встречаются слова, характерные для определения в терминологических словарях (*это, тип, вид, класс* и др.), то, скорее всего, это термин, для которого вводится определение. Признак `Neardefwords` равен количеству таких слов, появившихся непосредственно рядом (слева или справа) с исходным словосочетанием в сниппетах, полученных по запросу, совпадающему с исходным словосочетанием.

3.2.5. Количество слов-маркеров предметной области. Для словосочетаний-терминов существенным является принадлежность к предметной области. Простейший способ учесть фактор принадлежности к предметной области является задание списка маркеров предметной области, включающих в себя некую совокупность (от нескольких единиц до нескольких десятков) наиболее характерных слов предметной области. Признак `Markers` учитывает количество таких слов, встретившихся в сниппетах, полученных для словосочетания. В данном случае в качестве таких слов-маркеров мы используем существительные-названия основных наук и образованные от них прилагательные: *математика, математический, физика, физический* и др.

3.3. Признаки, полученные по тезаурусу. Признаки, полученные по тезаурусу, представляют собой необычный тип признаков, однако, на наш взгляд, использование этих признаков является чрезвычайно важным. Обычно в начале работ с терминологией предметной области несколько десятков наиболее существенных терминов являются очевидными, приводятся во всех терминологических словарях предметной области; часто достаточно понятно, как описать взаимоотношения между ними в тезаурусе. Чем больше производится работа над тезаурусом широкой предметной области, тем сложнее его пополнять.

Поэтому представляется важным, чтобы термины, которые уже отобраны экспертами в тезаурус, онтологию, терминологический словарь, помогали выявлять остальные термины данной предметной области. В нашем эксперименте в качестве тезауруса используется онтология ОЕНТ.

На основе терминологического состава онтологии ОЕНТ вычисляются признаки, которые должны помочь предсказать, относится ли к терминам данное словосочетание. Если словосочетание входит в состав терминов ОЕНТ, то, естественно, это словосочетание исключается из множества терминов, являющихся базой для порождения признаков. Был протестирован ряд признаков, основанных на структуре и составе тезауруса, однако в настоящее время удалось эффективно использовать только три следующих признака.

3.3.1. Синоним к термину. В текстах предметной области может встречаться много вариантов называния одного и того же термина [2], поэтому можно предположить, что если словосочетание похоже на словосочетание, которое уже считается термином (включено в тезаурус), то это словосочетание также является термином предметной области.

Пусть a и b являются словами-компонентами словосочетания ab, по поводу которого нужно принять решение. Мы будем считать, что словосочетание ab является синонимом словосочетания $a'b'$, если a совпадает или является синонимом a', а b совпадает или является синонимом b' (признак `Синоним к термину (SynTerm)`). Синонимия отдельных слов задана в исходном тезаурусе посредством того, что слова указаны как текстовые входы одного и того же понятия тезауруса. Так, например, если слова *объект* и *предмет* указаны в тезаурусе как текстовые входы одного и того же понятия, то словосочетания *учебный объект* и *учебный предмет* будут рассматриваться системой как потенциальные синонимы.

3.3.2. Синоним к нетермину. Если на текущем этапе работы обнаружилось определенная в предыдущем разделе синонимичность данного словосочетания к словосочетанию, не включенному в начальный тезаурус, то эта информация фиксируется в специальном признаке `Синоним к нетермину (SynNotTerm)`.

3.3.3. Полнота описания. Для извлеченного словосочетания ab может оказаться, что его слова-компоненты a и/или b уже включены в начальный тезаурус в качестве текстовых входов понятий и для соответствующих понятий уже описана некоторая совокупность отношений. Мы подсчитываем число всех отношений понятий, к которым, возможно, приписаны слова a и b, и фиксируем эту величину в виде характеристики `Completeness`. Предполагается, что чем больше отношений у соответствующих понятий, тем более они важны для предметной области, и это может также поднять значимость словосочетания ab.

Рассмотрим, например, словосочетание *собственный вектор*. Лемма *собственный* не соответствует ни одному понятию онтологии ОЕНТ, а лемма *вектор* соответствует одному понятию ВЕКТОР, у которого 56 отношений. Таким образом, значение признака `Completeness` для этого словосочетания равно 56. Если рассмотреть словосочетание *физическая химия*, то лемма *физический* относится к понятию ФИЗИКА, лемма *химия* — к понятию ХИМИЯ. Суммарное число отношений у этих понятий и соответственно значение признака `Completeness` равно 116.

Если у слов-компонентов словосочетания нет соответствующих понятий в тезаурусе, как, например, у словосочетания *последнее поступление*, то значение признака `Completeness`=0.

4. Метрика для оценки качества извлечения терминологических словосочетаний. Для оценки качества списка извлеченных терминологических словосочетаний используется мера, заимствованная из информационного поиска — так называемая средняя точность — `AvP` [1].

Характеристика средней точности определяется через характеристику точности.

Точность `PrecTerm` извлечения терминологических словосочетаний из некоторого числа N словосочетаний определяется как отношение числа терминов T к общему количеству словосочетаний в списке: $\texttt{PrecTerm} = \dfrac{T}{N}$. В упорядоченном списке может быть определена точность извлечения терминов на уровне i-го элемента $\texttt{PrecTerm}(i)$, которая представляет собой точность на множестве i словосочетаний от начала списка.

Характеристика средней точности `AvP` в задаче извлечения терминологических словосочетаний вычисляется следующим образом. Пусть в упорядоченном списке словосочетаний имеется k терминов и $\texttt{pos}(i)$ — позиция i-го термина от начала списка. Тогда точность на уровне i-го терминологического словосочетания $\texttt{PrecTerm}_i$ в упорядоченном списке равна $\texttt{PrecTerm}(\texttt{pos}(i))$.

Средняя точность для данного упорядочения списка словосочетаний равна среднему значению величины $\texttt{PrecTerm}_i$: $\texttt{AvP} = \dfrac{1}{k}\sum\limits_i \texttt{PrecTerm}_i$. Например, если в списке из трех словосочетаний термины встречаются на первом и на третьем месте, то $\texttt{AvP} = \dfrac{1+2/3}{2} = \dfrac{5}{6}$, или $\approx 83.3\%$.

5. Результаты эксперимента. Все эксперименты проводились с выборкой величиной 5 тысяч наи-

более частотных двухсловных словосочетаний исходного списка, для которых были обсчитаны все вышеперечисленные признаки. Для обучения наилучшей комбинации признаков исходная выборка делилась в соотношении 3 (обучающая выборка) к 1 (контрольная выборка). В качестве эталонного множества терминов использовались двухсловные термины, включенные в состав онтологии ОЕНТ.

Табл. 3 представляет меру средней точности AvP для отдельных признаков словосочетаний на контрольной выборке. Отметим, что в качестве базового уровня, в котором не было сделано реально никакого разумного упорядочения, для эксперимента можно взять простое упорядочение по алфавиту, для которого величина средней точности оказалась равной 57%.

Таблица 3

Значения AvP по отдельным характеристикам

Признак	AvP по контрольной выборке
Упорядочение по алфавиту	57%
Частотность (Freq)	66%
Взаимная информация (MI)	64%
MI_3	67%
Усеченное словосочетание (Inside)	75%
Частотность по сниппетам (FreqBySnip)	69%
Частотность слов-определений (NearDefWords)	73%
Scalar_1	61%
Scalar_2	60%
BinarScalar_1	64%
BinarScalar_2	62%
Snipfreq_0	66%
Snipfreq_1	67%
Snipfreq_2	67%
Markers	65%
Полнота описания (Completeness)	69%

Таким образом, по таблице видно, что среди признаков, рассчитанных по текстовой коллекции, наибольшую величину средней точности показал предложенный нами признак `Inside`.

Среди признаков, полученных на основе выдачи поисковой машины, наиболее высокими оказались предложенные нами признаки `Частотность по сниппетам FreqBySnip` и `Частотность слов определений NearDefWords`.

Два признака SynNotTerm и SynTerm являются бинарными, поэтому их некорректно оценивать с помощью оценки `AvP`. Для этих признаков оценивались информативность (т.е. качество информативности закономерности [5]) и точность классификации в целом. Данные, полученные на контрольной выборке, отражены в табл. 4.

Точность классификации бинарных признаков `Синоним к термину` и `Синоним к нетермину` невелика, не более 58%. Однако признак `Синоним к термину` очень хорошо “отделяет” (пользуясь терминологией логических алгоритмов классификации [5]) термины, т.е. имеет место высоко информативная закономерность: если `Синоним к термину`$(x) = 1$, то x — термин.

Таблица 4

Таблица точности признаков `Синоним к термину` и `Синоним к нетермину`

Название признака	Значение признака	Нетермин	Термин	Точность
Синоним к термину	0	415	406	50.5%
Синоним к термину	1	18	160	89.0%
Синоним к нетермину	0	387	538	58.0%
Синоним к нетермину	1	26	28	62.0%

Мы предполагаем, что вычисленные признаки могут отражать разные особенности терминологических словосочетаний, и поэтому имеет смысл найти наилучшую комбинацию этих признаков по мере средней точности.

Для подбора алгоритма комбинирования полученных признаков был использован программный пакет алгоритмов машинного обучения RapidMiner [18]. Наилучшим методом по величине средней точности `AvP` оказался метод логистической регрессии W-Logistic, на основе которого было достигнуто значение `AvP` $= 83\%$.

Логистическая регрессия реализует формулу $a(x) = \sigma(a_1x_1 + a_2x_2 + \ldots + a_nx_n)$, где $\sigma(z) = \dfrac{1}{1 - e^{-z}}$ —

сигмоидная функция, x_i — значения признаков объекта x, a_i — оптимизируемые параметры алгоритма. Значение функции $a(x)$ интерпретируется как вероятность принадлежности объекта x одному из двух классов.

Для практической обработки большого числа словосочетаний важным является выделение относительно небольшого числа признаков, на основе которых можно получить практически те же результаты упорядочения терминов. В результате произведенного отбора признаков было получено, что с использованием метода логистической регрессии удалось достичь уровня AvP 82% на следующем наборе признаков: `Inside, MI, FreqBySnip, Neardefwords, BinarScalar1, SynTerm, Completeness`. Как видим, среди отобранных признаков присутствуют признаки всех трех типов.

Таким образом, с помощью комбинации признаков удалось достичь качества упорядочения исходного списка словосочетаний, которое более чем на 10% лучше по сравнению с упорядочением по наилучшему признаку. Найден ряд признаков, которые показали значительно более высокий уровень средней точности, чем широко известные признаки.

6. Заключение. В данной работе мы предложили использовать для автоматического извлечения двухсловных терминологических словосочетаний три типа признаков различного происхождения и показали, что для определения терминологичности словосочетания полезно использовать все эти три типа признаков.

Впервые для определения терминологичности словосочетаний предложено использовать структуру разрабатываемого тезауруса предметной области и описанные в нем знания о предметной области. Эта информация улучшает качество определения терминологических словосочетаний и полезна в ситуации пополнения существующих тезаурусов предметной области.

Кроме того, мы показали, что при комбинировании разных признаков словосочетаний удается достичь намного более качественного упорядочения словосочетаний, т.е. повышения процентной доли терминов в начале упорядоченного списка словосочетаний. Эксперименты по автоматическому извлечению терминологических словосочетаний проводились на основе сопоставления результатов работы разных алгоритмов с вручную отобранными терминами в онтологию по естественным наукам и технологиям (ОЕНТ).

Качественное определение терминологичности словосочетаний может значительно сократить время работы экспертов для подготовки терминологических ресурсов, а также позволяет с высокой степенью точности автоматически извлекать “хорошие” словосочетания для использования их в процедурах визуализации выдачи информационно-поисковых систем, присваивания ключевых слов документам и т.п.

СПИСОК ЛИТЕРАТУРЫ

1. *Агеев М.С., Кураленок И.Е.* Официальные метрики РОМИП’2004 // Российский семинар по оценке методов информационного поиска. Пущино, 2004. 142–150.
2. *Большакова Е.И., Васильева Н.Э.* Терминологическая вариантность и ее учет при автоматической обработке текстов // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием. Т. 2. М.: ЛЕНАНД, 2008. 174–182.
3. *Браславский П.И., Соколов Е.А.* Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии. Тр. Международной конференции “Диалог 2006”. М.: Изд-во РГГУ, 2006. 88–94.
4. *Браславский П.И., Соколов Е.А.* Автоматическое извлечение терминологии с использованием поисковых машин Интернета // Компьютерная лингвистика и интеллектуальные технологии. Тр. Международной конференции “Диалог 2007”. М.: Изд-во РГГУ, 2007. 89–94.
5. *Воронцов К.В.* Лекции по логическим алгоритмам классификации. 2007 (http://www.ccas.ru/voron/download/LogicAlgs.pdf).
6. ГОСТ 7.25.-2001 Тезаурус информационно-поисковый одноязычный: Правила разработки: структура, состав и форма представления: Межгосударственный стандарт. Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2001.
7. *Добров Б.В., Лукашевич Н.В., Сыромятников С.В.* Формирование базы терминологических словосочетаний по текстам предметной области // Тр. 5-й Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” (RCDL-2003). СПб, 2003. 201–210.
8. *Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н.* Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска // Тр. 7-й Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” (RCDL-2005). Ярославль, 2005. 70–79.
9. Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцевой. М.: Советская энциклопедия, 1990.
10. *Лукашевич Н.В.* Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер. 2. 1995. № 3. 21–24.

11. *Лукашевич Н.В., Добров Б.В.* Отношения в онтологиях для решения задач информационного поиска в больших разнородных текстовых коллекциях // Девятая национальная конференция по искусственному интеллекту с международным участием (КИИ 2004). Т 2. М.: Физматлит, 2004. 544–551.
12. *Лукашевич Н.В., Добров Б.В., Чуйко Д.С.* Отбор словосочетаний для словаря системы автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии. Тр. Международной конференции "Диалог 2008". М.: Изд-во РГГУ, 2007. 339–344.
13. *Никитина С.Е.* Семантический анализ языка науки. М.: Наука, 1987.
14. *Ahmad K., Gillam L., Tostevin L.* University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval // Proc. of Eighth Text Retrieval Conference (Trec-8). Gaithersburg, 1999. 717–724.
15. *Daille B., Gaussier E., Lang J.M.* An evaluation of statistics scores for word association // Proc. of the Tbilisi Symposium on Logic, Language and Computation. Chicago: CSLI Publications. 1998. 177–188.
16. *Pecina P., Schlesinger P.* Combining association measures for collocation extraction // Annual Meeting of the Association for Computational Linguistics (ACL 2006). Sydney: ACM, 2006. 651–658.
17. *Pearce D.* Synonymy in collocation extraction // Proc. of the NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. Pittsburgh, 2001. 41–46.
18. RapidMiner (www.rapidminer.com).
19. *Zhang Z., Iria J., Brewster Ch., Ciravegna F.* A Comparative Evaluation of Term Recognition Algorithms // Proc. of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, 2008.

Поступила в редакцию
05.10.2010