КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 001.103:002

ФУНКЦИОНАЛЬНЫЙ ПОДХОД К ВЫДЕЛЕНИЮ КЛЮЧЕВЫХ СЛОВ: МЕТОДИКА И РЕАЛИЗАЦИЯ

И. Е. Воронина, А. А. Кретов., И. В. Попова, Л. В. Дудкина

Воронежский государственный университет

Поступила в редакцию 1.03.2009 г.

Аннотация. Рассматривается задача автоматизированного выделения ключевых слов с использованием алгоритма выделения тематически маркированной лексики. Представлены результаты вычислительного эксперимента на тексте Общей части УК РФ. Предлагается способ построения семантического пространства путем объединения алгоритма выделения тематически маркированной лексики и алгоритма Е.Л. Гинзбурга.

Ключевые слова. Компьютерная лингвистика, выделение ключевых слов, построение семантического пространства, вычислительный эксперимент, семантическое поле слова

Abstract. The article discusses the task of computer-aided key words selection. To solve the task the algorithm of thematically marked vocabulary selection was applied. Results of a computational experiment based on the text of the General Part of the Criminal Code of Russian Federation are presented. The authors suggest a way of semantic space modeling by union of the algorithm of thematically marked vocabulary selection and E.L.Ginzburg's algorithm.

Keywords: Computational linguistics, key words selection, semantic space modeling, computational experiment, semantic field of the word.

ВВЕДЕНИЕ

Современные процессы информатизации, позволяющие предоставлять информацию потребителю, давно стали важным фактором жизни общества, функционирующего в едином информационном пространстве. Поэтому трудно переоценить важность задачи предоставления нужной информации в нужном виде и в нужное время. Несмотря на то, что необходимая информация генерируется многочисленными информационными системами, аккумулируется в различных хранилищах, базах и банках данных, доля неструктурированной информации, содержащейся в текстах на естественном языке и извлекаемой из них, несоизмеримо выше. Отсюда актуальность задачи автоматизированного анализа текстов и проведения исследований по формализации естественного языка. Серьезной проблемой является семантический анализ текста, когда происходит выявление смысла предложений или их отдельных частей. Семантический анализ невозможен без использования результатов анализа отдельных слов, для чего в качестве общей основы для всех методов анализа может быть использован тезаурус языка. Кроме того, необходимо разработать методы выделения ключевых слов из текста, которые отражают специфику текста, позволяют выявить его тематику. Следующей, и весьма нетривиальной задачей, является задача определения слов, близких по значению заданному слову (получение семантического поля слова) [1].

Трудно переоценить актуальность задачи автоматизации выделения ключевых понятий и словосочетаний в текстовых источниках.

Применение компьютером не только ускорило создание и обработку текстовых документов, но и невероятно увеличило их количество и объем. Пользователи регулярно сталкиваются с потребностью быстро просматривать большой объем документов и выбирать из них только действительно нужные. Это происходит при работе с текстовыми базами данных, разборе электронной почты и при поиске в Интернет. Во всех указанных случаях полезна возмож-

ность автоматически составлять сжатые описания содержания документов — аннотации. Ключевые слова или фразы могут дать высокоуровневое описание содержания документа. Тематика документа определяется его словарным запасом. Выделение ключевых слов и словосочетаний необходимо также при решении задачи классификации документов по заданному набору тематик

В инженерии знаний группа текстологических методов объединяет методы извлечения знаний, основанных на изучении текстов, содержащих профессиональные знания. Для формирования макроструктуры текста в виде реферата или в форме графа выделяются ключевые слова и выражения, а затем определяются связи между ними. Далее на основании макроструктур строится поле знаний — условное, неформальное описание основных понятий и взаимосвязей между понятиями предметной области. Поле знаний — основа для создания формализованного представления знаний [ссылка на тезисы конференции].

1. ВЫДЕЛЕНИЕ ТЕМАТИЧЕСКИ МАРКИРОВАННОЙ ЛЕКСИКИ

Тексты на любом языке состоят из двух частей: того, что обязательно должно быть выражено по законам данного языка, и того, что отражает специфику тематики текста и стиля автора. Эти составляющие называются соответственно тематически нейтральной и тематически маркированной лексикой. Выделение обеих групп лексики — шаг на пути к определению содержательной отнесенности текста. Оно позволяет как проникнуть к содержанию текста, так и составить определенное мнение о своеобразии лексики автора и его языка.

В [2] был предложен метод формального, а, следовательно, — автоматизируемого выделения тематически маркированной лексики статистическим посредством «взвешивания» слов по функциональным параметрам.

Лексические единицы, встречающиеся в тексте, имеют свои частотные характеристики. Чем ярче количественные особенности некоторых словоформ по сравнению с остальными во всем тексте, тем выше их тематическая маркированность.

В контексте решаемой задачи словом считается непрерывная последовательность букв русского алфавита без учета регистра.

Выделение тематически маркированной и тематически нейтральной лексики производится методом системного взвешивания слов по двум функциональным параметрам: прямому (частотному — Q-параметр) и косвенному (длина слова — F-параметр).

Традиционным способом выявления тематически маркированной лексики является частота словоформ (или их множеств, относящихся к одному слову-лемме). При этом предполагается, что чем чаще употребляется слово в тексте, тем оно важнее для его содержания. Это справедливо лишь отчасти: служебные и дискурсивные слова обычно имеют большую частоту, но, тем не менее, специфики текста не отражают.

Необходимо учитывать и еще одно обстоятельство. Установлена зависимость, существующая между частотой слова (словоформы) и его длиной: чем чаще употребляется слово, тем оно короче и наоборот. Но для этого требуется, чтобы слова устойчиво и продолжительно были частотными. Следовательно, если короткое слово в тексте будет частотным, это будет характеризовать его как короткое слово, а если длинное слово в тексте будет обладать частотой, необычной для слов такой длины, то оно будет отражать специфику данного текста. Таким образом, для выделения тематически маркированной лексики в данном тексте недостаточно информации об их длине и частоте, нужно соотнести оба типа информации, а для этого сделать информацию о длине и частоте сопоставимой.

С этой целью единообразно вычисляются функциональные веса словоформ: прямой (частотный — Q-вес) и косвенный (длина слова в звуках — F-вес).

Q-параметр характеризует функционирование слова в данном тексте. Для его подсчета после построения частотного словаря исходного текста слова располагаются в порядке убывания абсолютной частоты. Назначение рангов осуществляется для группы слов с одинаковой частотой. Слова с наибольшей частотой получают ранг 1. При уменьшении частоты слова в тексте ранг увеличивается на 1. Таким образом, слова, имеющие одинаковую абсолютную частоту, получают одинаковый ранг.

Q-параметр вычисляется по формуле, предложенной в [3], и несколько модифицированной нами.

$$Q_{ri} = \frac{\sum r - R_{i-1}}{\sum r}, \qquad (1)$$

где $\sum r$ — сумма единиц всех рангов, R_{i-1} — сумма единиц от первого до данного ранга (исключительно).

При этом речь идёт о рангах частот, а не о рангах слов.

Длина слова в звуках — **F-параметр — характеризует** функционирование слова на продолжитетьном отрезке времени, настолько продолжительном, чтобы функционирование успело повлиять на длину слова. Для его подсчета после построения частотного словаря исходного текста слова располагаются в порядке убывания их длины в звуках. Назначение рангов осуществляется для группы слов с одинаковой длиной в звуках. Слова с наибольшей длиной получают ранг 1. При уменьшении длины слова ранг увеличивается на 1. Таким образом, слова, имеющие одинаковую длину в звуках, получают одинаковый ранг.

F-параметр вычисляется по формуле

$$F_{ri} = \frac{\sum r - R_{i-1}}{\sum r}, \qquad (2)$$

где $\sum r$ — сумма единиц всех рангов, R_{i-1} — сумма единиц от первого до данного ранга (исключительно).

Вводится Индекс тематической маркированности слова (ИнТеМ), вычисляемый по формуле

$$\text{ИнТеМ}=\text{Q-вес} - \text{F-вес}, \qquad (3)$$

где Q-вес — вес слова по частоте, F-вес — вес слова по длине.

Положительные значения ИнТеМа будут характеризовать тематически маркированную лексику, отрицательные значения — тематически нейтральную лексику. При этом значения ИнТеМа будут варьироваться в интервале от -1 до +1.

2. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Алгоритм выделения тематически маркированной лексики:

1. подсчитываем для каждого слова его абсолютную частоту в тексте и длину в звуках;

2. назначаем ранги частот и подсчитываем для каждого слова Q-вес;

3. назначаем ранги для длин слов в звуках и подсчитываем для каждого слова F-вес;

4. вычисляем для каждого слова ИнТеМ.

5. Если полученный ИнТеМ положительный, слово относится к тематически маркированной лексике, отрицательный — к тематически нейтральной лексике.

В соответствии с выбранным методом был разработан набор инструментальных и алгоритмических средств, позволяющих анализировать текст на русском языке и выявлять тематически маркированную лексику посредством вычисления ИнТеМ и ранжирования слов по нему.

В процессе разработки программы было сделано следующее:

— реализованы методы построения частотного словаря и работы с ним;

— разработаны алгоритмы и процедуры расчета функциональных весов на основе построенного частотного словаря;

— предоставлены возможности работы с полученными результатами (сортировки по различным параметрам и сохранение в удобном для пользователя виде);

— предоставлена возможность по выбору пользователя не включать в статистику слова, составляющие специфику естественного языка в целом, либо слова, не являющиеся значимыми в рамках конкретного исследования.

3. ПРОВЕДЕНИЕ ВЫЧИСЛИТЕЛЬНОГО ЭКСПЕРИМЕНТА

Полученные результаты сравнивались со списком ключевых элементов, полученным при ручной обработке. Тестирование проводилось на тексте Общей части УК РФ. Чтобы проверить эффективность работы алгоритма, были выявлены те категории и понятия Общей части УК РФ, чей вклад в смысловую нагрузку текста является определяющим.

Выделение ключевых понятий вручную должно осуществляться специалистом предметной области, поскольку для этого необходимо использовать специальные познания в области юриспруденции, в частности в области Уголовного права. Так как Уголовный Кодекс РФ состоит из 2-х частей — Общей и Особенной, то для получения более точного, полного и объективного списка ключевых понятий, отражающего специфику всей отрасли Уголовное право, целесообразным было обработать каждую часть в отдельности.

Процесс выделения ключевых понятий из Уголовного Кодекса можно условно разбить на 3 этапа:

— выделение ключевых понятий из Общей части УК;

— выделение ключевых понятий из Особенной части УК;

— формирование общего списка ключевых понятий УК

Рассмотрим каждый из этапов подробнее.

ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ ПОНЯТИЙ ИЗ ОБЩЕЙ ЧАСТИ УК

На данном этапе рассматривалась и анализировалась Общая часть Уголовного Кодекса.

В список ключевых понятий вошли понятия и категории, характеризующие:

— признаки уголовного права как отрасли права (то есть объединяющие его с другими отраслями права, некие закономерности права как категории знаний) — например, принципы права, законодательство, ответственность и т.п.;

— специфические особенности Уголовного права (отличающие УП от других отраслей права) — например, преступление, основание уголовной ответственности, преступное сообщество и т.п.

ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ ПОНЯТИЙ ИЗ ОСОБЕННОЙ ЧАСТИ УК

На данном этапе рассматривалась и анализировалась Особенная Часть УК.

В список ключевых понятий вошли понятия и категории, характеризующие:

— конкретные преступления (наименования конкретных составов), например, убийство, умышленное причинение тяжкого вреда здоровью, убийство матерью новорожденного ребенка и.т.п.;

— элементы составов преступлений (характеризующие каждый конкретный состав), например, корыстный мотив, месть, честь, достоинство, половая неприкосновенность и.т.п.

ФОРМИРОВАНИЕ ОБЩЕГО СПИСКА

На данном этапе анализировались списки, полученные на каждом из предыдущих этапов, из общего списка исключались дублирующиеся понятия (если таковые встречались), некоторые понятия объединялись в одно или происходило разделение одного понятия на несколько, исключались «лишние» и ошибочные понятия.

В итоге общий список составил порядка 700 понятий, причем среди них оказались как отдельные слова, так и словосочетания. Если анализировать соотношение количества ключевых понятий Общей и Особенной частей, то можно сделать вывод, что в Общей части встречается приблизительно 29% от общего количества ключевых понятий во всем кодексе.

В ходе выделения ключевых понятий возникли определенные трудности:

— неоднозначность в определении, все ли элементы словосочетания составляют одно ключевое понятие, или целесообразнее разбить его на несколько;

— неоднозначность в определении количества элементов в словосочетании, претендующем на роль ключевого понятия;

— большое количество элементов в некоторых ключевых понятиях-словосочетаниях;

— элементы некоторых ключевых понятий стоят не подряд (например, «охрана общественной безопасности, общественного порядка» дает 2 ключевых понятия – «охрана общественной безопасности», «охрана общественного порядка»).

Тем не менее, на основе анализа Уголовного кодекса был составлен список ключевых понятий, который можно считать своеобразным эталоном для сравнения результатов, полученных при использовании различных методов выделения ключевых понятий из текста.

Среди выявленных понятий термины, состоящие из одного слова, составляют около 15%. Остальная часть приходится на терминологические словосочетания. Обработка текста показала, что положительный индекс тематической маркированности имеет около 40% терминов, состоящих из одного слова. На результаты сильно повлияла специфика текста, взятого для тестирования. Выяснилось, что плохо поддаются выделению выбранным методом слова с низкой частотностью, термины, которые значимы и встречаются всего в нескольких (1—3) статьях кодекса. Хотя формально они входят в список ключевых, очень небольшая область употребления в тексте (несколько предложений) не обеспечивает частотности, нужной для их эффективного выделения алгоритмом. Возможный способ решения этой проблемы в дальнейшем – включение в метод распознавания структуры текста и соответствующая корректировка весов, например, повышение индекса значимости для слов в заголовках и т. п.

Остальные слова с положительным индексом тематической маркированности (до 100%) входят в состав выделенных вручную терминологических словосочетаний, что подтверждает эффективность применения выбранного метода.

ЗАКЛЮЧЕНИЕ

Результаты можно значительно улучшить (особенно в части определения ключевых словосочетаний), если объединить алгоритм выде-

ления тематически маркированной лексики с алгоритмом Е.Л. Гинзбурга [1, 4].

С использованием алгоритма Гинзбурга можно построить семантическое пространство следующим способом.

Строится граф, узлами которого являются ключевые слова; ребра графа – слова-спутники, притянутые ключевым словом по алгоритму Гинзбурга.

Для построения графа сначала формируется матрица ключевых слов, представленная на рис. (КС – ключевое слово).

На пересечении строк и столбцов указывается сила связи между ключевыми словами (при желании её можно интерпретировать и как величину обратную расстоянию между ключевыми словами в семантическом пространстве).

Простейший способ определения силы связи: подсчитывается сумма общих слов-спутников. Более точный способ: подсчитывается доля общих слов для меньшего из двух множеств. Третий способ (самый интересный): суммируются веса каждого из ключевых слов в обоих множествах и делятся на два. Полученные величины суммируются и помещаются на пересечении КС как показатель силы их связи. Заполненная таблица дает материал для визуализации данных и их представления в виде графа.

	КС1	КС2	КС3	КС4 …
КС1				
КС2				
КС3				
КС4 …				

Рис. 1. Матрица ключевых слов

Изложенные соображения лягут в основу дальнейших исследований по автоматизированному выделению ключевых слов и словосочетаний и построения семантического пространства текста.

СПИСОК ЛИТЕРАТУРЫ

1. *Воронина И. Е.* Программные средства выявления семантического поля слов / И. Е. Воронина, А.А. Кретов, О.С. Титова // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2008. – № 2. — С. 111—122.

2. *Кретов А.А.* Метод формального выделения тематически нейтральной лексики (на примере старославянских текстов) // Вестник Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2007. – № 1. С. 81—90.

3. *Титов В.Т.* Частная квантитативная лексикология романских языков: Монография / В. Т. Титов; Воронеж. гос. ун-т. – Воронеж: Изд-во Воронеж. гос. ун-та, 2004. –552 с.

4. *Гинзбург Е.Л.* Идиоглоссы: проблемы выявления и изучения контекста / Е. Л. Гинзбург // Семантика языковых единиц: Доклады VI Международной конференции. Т.I, М., 1998. — С. 26—28.

Воронина Ирина Евгеньевна – к.т.н., доцент кафедры программного обеспечения и администрирования информационных систем факультета ПММ, ВГУ. Тел. 208-337. E-mail: voronina@amm.vsu.ru

Кретов Алексей Александрович – д.фил.н., профессор, зав. каф. структурной и прикладной лингвистики факультета РГФ, ВГУ. Тел. 204-149. E-mail: a_a_kretov@rambler.ru

Попова Ирина Викторовна – студентка 4 курса факультета ПММ, специальность Математическое обеспечение и администрирование информационных систем, ВГУ.

Дудкина Людмила Владимировна – студентка 5 курса факультета ПММ, специальность Прикладная информатика в юриспруденции, ВГУ.

Voronina Irina Ye. - Candidate of Technical Sciences, Associate Professor, the dept. of the Software and Information System Administration, Voronezh State University. Tel. 208-337. E-mail: voronina@amm.vsu.ru.

Kretov Aleksey A. - Doctor of Philology Sciences, Professor, the Head of the dept. of Theoretical and Applied Lingvistics , Voronezh State University. Tel. 204-149. E-mail: a_a_kretov@rambler.ru

Popova Irina V. – 4th year student of Department of Applied Mathematics, Computer Science & Mechanics, Software & Information Systems Administering Program

Dudkina Lyudmila V. - 5th year student of Department of Applied Mathematics, Computer Science & Mechanics, Computer Science in Law Program