УДК 81'32

ПРОЦЕДУРА АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ТЕРМИНОПОДОБНЫХ СЛОВОСОЧЕТАНИЙ РУССКОГО ЯЗЫКА

Бурукина И.С.
научный руководитель канд.филол.наук Азарова И.В.
Санкт-Петербургский государственный университет

Цель – разработка алгоритма извлечения терминов и терминоподобных словосочетаний из корпусов текстов на русском языке. Несмотря на важность данной задачи, она остается нерешенной и нуждается в тщательной проработке. Были изучены существующие подходы к выделению терминов и устойчивых словосочетаний из текстов на английском, немецком и французском языках и рассмотрена возможность их применения к документам на русском языке.

Введение. Согласно S. Ananiadou, существуют три базовых подхода к выделению терминов и терминологических словосочетаний: 1) лингвистический подход, 2) статистический подход и 3) смешанный подход, основанный на применении и статистической, и лингвистической информации. Помимо данных методов, обычно применяемых для поиска терминов в тексте, разумно рассмотреть также подходы, выработанные для извлечения устойчивых словосочетаний (коллокаций). В настоящем исследовании мы обратились к теории составных наименований (multiword expression theory), разработанной в Стэнфордском университете, США. Составное наименование может быть описано как «уникальное толкование, стирающее границы слов». Так же, как терминоподобные словосочетания, составные наименования обладают значением не равным сумме значений их компонентов, и, что более важно, так же возникают «по договоренности»: мы говорим *компьютерная лингвистика*, а не **вычислительная филология*, поскольку так «исторически сложилось». Для извлечения составных наименований было предложено несколько статистических методов; наиболее интересным является так называемый критерий силы связи, используемый для определения силы зависимости между компонентами выражения, подобного составному наименованию.

База исследования. Поскольку наша задача заключается в разработке алгоритма, подходящего для извлечения терминоподобных выражений из корпусов текстов, был построен и вручную размечен корпус текстов тематики административное право (30.000 словоупотреблений). С одной стороны, юридическая терминология хорошо разработана; с другой стороны, многочисленные тексты данной тематики находятся в свободном доступе и удобны в качестве материалов для исследования. Для изучения современной терминологии области административного права был использован Словарь административного права под редакцией Бачило И.Л. (Москва, 1999). Также в ходе исследования было проведено сопоставления корпуса административного права с корпусом общелитературного русского языка Бокрёнок (21 млн словоупотреблений).

Формулировка гипотезы. Изучив современную терминологию административного права, мы сформулировали следующие гипотезы: 1) для достижения наиучшего результата при извлечении терминов необходимо использовать как статистическую, так и лингвистическую информацию, 2) методы, предложенные для выделения составных наименований, так же могут быть применимы к выделению терминоподобных словосочетаний русского языка.

Мы предполагаем, что поиск терминоподобных словосочетаний следует начинать с изучения контекста однословных терминов. Такие однословные термины мы

называем *ключевые слова*. Таким образом, выделение терминоподобных словосочетаний распадается на три этапа:

- выделение ключевых слов,
- анализ контекста ключевых слов и выделение терминоподобных словосочетаний,
- подтверждение терминологического статуса найденных словосочетаний.

Выделение ключевых слов. Эмпирически мы определили пороговую частоту встречаемости термина для сравнительно небольшого корпуса текстов – 200 IPM. В соответствии с данным пороговым значением были отобраны 602 лексические единицы (25%). Первым мы применили частеречный фильтр. Было решено уделить внимание наиболее частотному типу терминологических словосочетаний - имя существительное + имя прилагательное. Также были оставлены причастия из-за регулярной омонимии между вербоидами и именами прилагательными в русском языке. Далее был применен статистический фильтр. Частота встречаемости термина в корпусе специальных текстов выше, чем частота его встречаемости в корпусе общелитературного языка. Поэтому для каждой единицы был вычислен коэффициент относительной частоты. 270 единиц были оставлены для дальнейшего изучения по результатам применения обоих фильтров.

$$K = \frac{IPM_c}{IPM_b}$$

Критерий относительной частоты, где IPM_c – частота в корпусе специальных текстов, IPM_b – частота в корпусе общелитературного языка.

Выделение терминоподобных словосочетаний. Поскольку мы ограничились рассмотрением конструкций имя существительное + имя прилагательное, мы использовали синтаксическую модель Прилагательное(0) + Существительное(+1) для дальнейших поисков словосочетаний. Были оставлены 35 ключевых имен прилагательных. Далее вручную из корпуса специальных текстов были выделены 280 комбинаций слов. Применялись различные статистические оценки для подтверждения терминологического статуса найденных словосочетаний: коэффициенты взаимной информации (mutual information score), t-score, коэффициент Дайса.

$$MI(w_1, w_2) = \frac{f(w_1, w_2)}{f(w_1) \cdot f(w_2)}$$

Коэффициент взаимной информации 1, где w_1, w_2 – компоненты словосочетания.

$$MI_2(w_1, w_2) = \frac{f(w_1, w_2)}{f(w_1|w_2) \cdot f(w_2|w_1)}$$

Коэффициент взаимной информации 2.

Эмпирически мы также определили наилучшие пороговые значения для данных оценок. Таблица ниже показывает число выделенных терминоподобных словосочетаний и пороговые значения для каждой оценки.

	Коэффициент взаимной информации 1	Коэффициент взаимной информации 2	t-score	Коэффициент Дайса
Пороговое значение	0.01	0.01	0	0.1

Число выделенных словосочетаний	49	44	26	63

Для подтверждения смысловой связности словосочетаний мы также использовали критерий силы, проанализировав различные возможные контексты для каждого ключевого элемента.

$$k(w_i) = \frac{f(w_i) - \bar{f}}{\sigma}$$

Критерий силы, где w_1 – элемент из множества C (множество единиц, совместно употребленных в тексте).

$$\bar{f} = \frac{\sum_{w_i \in C} f(w_i)}{|C|}$$

$$\sigma = \sqrt{\frac{\sum_{w_i \in C} (f(w_i) - \bar{f})^2}{|C|}}$$

Средняя частота.

Были выделены 33 терминоподобных словосочетания, например: *основной акт, государственная власть, административное право, юридическое лицо.*

Оценка результатов. Для оценки полученных результатов мы использовали «золотой стандарт» и оценку экспертом. Во-первых, мы сравнили терминоподобные словосочетания, выделенные с вычислением различных статистических оценок, с материалом словаря («золотой стандарт»). Результаты сравнения представлены в таблице ниже (число совпавших словосочетаний).

Коэффициент взаимной информации 1		Коэффициент взаимной информации 2		t-score		Коэффициент Дайса		Критерий силы	
3	6.12%	1	2.27%	3	11.54%	8	12.70%	3	9.10%

Во-вторых, группе экспертов (студенты магистратуры юридических факультетов Санкт-Петербургского государственного университета и Пензенского государственного университета) было предложено оценить терминологичность найденных словосочетаний. Число словосочетаний, чей терминологический статус подтвердился, представлено в таблице ниже.

Коэффициент взаимной информации 1		Коэффициент взаимной информации 2		t-score		Коэффициент Дайса		Критерий силы	
8	14.3%	7	6.8%	9	30.8%	11	22.2%	7	27.3%

Заключение. Был предложен алгоритм выделения терминоподобных словосочетаний на русском языке. Первый этап – выделение так называемых ключевых слов (однословных терминов) с использованием частеречного и статистического фильтров. Второй этап – анализ контекста ключевых слов и выделение

терминоподобных словосочетаний. Здесь мы уделили основное внимание наиболее частотному типу конструкций имя существительное + имя прилагательное с прилагательным в качестве ключевого элемента. Третий этап – подтверждение терминологического статуса найденных словосочетаний с использованием различных статистических оценок. Наилучший результат был достигнут с применением t-score. Также удачно применение критерия силы, предложенного теорией составных наименований: была подтверждена терминологичность 27.2% всех найденных словосочетаний.