Рецензия на статью

Кочетковой Н.А., Ермакова П.Д. «МЕТОД ИЗВЛЕЧЕНИЯ ОДНОСЛОВНЫХ ТЕРМИНОВ НА ОСНОВЕ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ СЛОВ ВНУТРИ КОНТЕКСТА»

Статья Кочетковой Н.А., Ермакова П.Д. «МЕТОД ИЗВЛЕЧЕНИЯ ОДНОСЛОВНЫХ ТЕРМИНОВ НА ОСНОВЕ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ СЛОВ ВНУТРИ КОНТЕКСТА» посвящена актуальной проблеме теоретического и прикладного исследования терминологии. В статье представлен формализованный метод определения того, что считать терминологией (в рамках монотематической коллекции текстов). Решением прикладной задачи является предложенный авторами метод извлечения **однословных** терминов из рассматриваемых коллекций документов, основанный на статистическом распределении терминов в контексте. **Материалом** для экспериментального исследования Кочетковой Н.А., Ермакова П.Д. являются коллекции текстов журналов (7,2 млн. словоупотреблений) и "Вестник Томкого Государственного университета. Биология" (1,8 млн. словоупотреблений).

Представленное исследование обладает несомненной **научной новизной**, **актуальностью** и **практической значимостью** как в области теории языка (терминоведении), так и в прикладной области автоматической обработки текста, прежде всего, автоматической обработки научного текста (например, построения терминологических словарей, извлечения информации и разнообразных систем понимания научного текста).

Методику эксперимента авторы описывают следующим образом. «Из монотематической размеченной коллекции со снятой омонимией извлекаются все биграммы. Помимо этого, для коллекции составляется ее словарь нормальных форм. Для всех нормальных форм, встречаемость которых выше заданного порога проводится следующий алгоритм. Для всех биграмм, в которых встречается текущее выбранное слово, рассчитывается частота их встречаемости, после чего биграммы сортируются по убыванию этой частоты. Далее берется биграмма с максимальной частотой встречаемости P_1 и на ее основе рассчитывается распределение Ципфа с таким же количеством элементов: $Q_n=P_1/n$. Далее для этих двух распределений рассчитывается значение дивергенции Кулльбака-Лейблера: $D(P\|Q) = sum(P_i*log(P_i/Q_i))$. Слова с максимальным значением дивергенции считаются кандидатами на термины. Они должны быть просмотрены экспертами для окончательного принятия решения» (Кочетковой Н.А., Ермакова П.Д.).

Эти и предшествующие работы авторов подтвердили положение о том, что если из

текста после лемматизации извлечь все биграммы для целевого слова и расположить их по убыванию частоты встречаемости, то полученное распределение подчиняется закону Ципфа. В данной работе была верифицирована гипотеза о том, что для слов, не являющихся терминами данной предметной области, степенной показатель распределения Ципфа будет близок к 1 (или, возможно, к среднему значению по монотематической коллекции текстов). Для терминов же данный показатель должен существенно отличаться в большую сторону.

Показательно, что по результатам экспериментов Кочетковой Н.А., Ермакова П.Д. в топ по дивергенции вошли более специфичные термины, которых нет в топе по частоте.

Статья написана понятным языком и будет интересна представителям разных наук (от филологии до информатики, особенно прикладной информатики в области лингвистики).

Полагаю, что научная ценность статьи несомненна. Рекомендую статью Кочетковой Н.А., Ермакова П.Д. «МЕТОД ИЗВЛЕЧЕНИЯ ОДНОСЛОВНЫХ ТЕРМИНОВ НА ОСНОВЕ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ СЛОВ ВНУТРИ КОНТЕКСТА» для публикации в журнале «Вестник Пермского университета. Российская и зарубежная филология» или в журнале «Вестник Пермского университета. Серия: Математика, Механика, Информатика».

Доктор филологических наук
профессор, Кафедра информационных систем
в искусстве и гуманитарных науках СПбГУ Ягунова Е.В.

17.10.2016 г.

Специалист по кадрам Потапова Л.Ю.