Н. В. Зыгмантович

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ТЕРМИНОВ ИЗ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

Интенсивное развитие современных информационных технологий привело к значительным изменениям в процессах обработки научно-технических и публицистических текстов. Эти задачи неразрывно связаны с понятием *термин*. Существуют различные определения понятия *термин*. Например: в Большой советской энциклопедии: *термин* (от лат. *terminus* – предел, граница) – это слово (или сочетание слов), являющееся точным обозначением определенного понятия какой-либо специальной области науки, техники, искусства, общественной жизни.

Ученый В. М. Перерва в своей статье «О принципах и проблемах отбора терминов и составления словника терминологических словарей» дает следующее определение этого понятия: «Термин – это языковой знак, который может быть и отдельным словом, и словосочетанием. Он, как языковой знак, является носителем элементарной научной, технической, производственной и тому подобной информации в виде отдельного научного понятия, входящего в систему понятий определенной области знания или деятельности».

Необходимость решения проблемы автоматического извлечения из текстов терминов и терминологических словосочетаний связана с решением таких задач, как «автоматическая обработка текстов, «классификация документов», «кластеризации документов», «индексирование и реферирование текстов» и т.д.

Терминологический словарь – это не самый распространенный тип лингвистического словаря, но его значение трудно преувеличить. Благодаря определенной структуре и систематизации терминологического словаря поиск нужного термина значительно упрощается.

Прототипом терминологических словарей на Руси стали религиозные словари, появившиеся уже в XI веке. Первый словарь, систематизирующий научные термины, был составлен в 1780 году К. А. Кондратовичем – до этого существовали лишь краткие списки терминов и определений из различных областей знаний. В России первые работы по изучению терминов и терминологии начались в начале 30-х годов XX столетия, когда в 1931 году была опубликована статья ученого Д. С. Лотте «Очередные задачи научно-технической терминологии». Работа по созданию терминологических словарей активно продолжается и в наши дни, что описывается в статье А. В. Зубова «Способы автоматического извлечения терминов из текста».

Создание терминологических словарей – работа сложная, кропотливая. Для того, чтобы облегчить работу лингвиста по созданию терминологических словарей используются различные методы выделения из текстов терминов-слов и терминов-словосочетаний. Часто для этого используется статистический метод. Применение этого метода, который основан на частоте вхождения слова в рассматриваемую коллекцию текстовых документов, позволяет с большой достоверностью выделить термины-слова и термины-словосочетания в специальных текстах различных подъязыков. Использование статистического метода для автоматического выделения терминов-слов детально описано в работе Р. Г. Пиотровского «Статистическое опознание термина».

Одной из программ, которая использует статистический метод для извлечения возможных терминов из совокупности текстов по определенной предметной области, является программа «Менеджер терминологии Lite». Это программа входит в пакет «PROMT Professional», который предназначен для профессионального перевода документов различных форматов. Она позволяет работать с текстовыми документами на разных языках, которые поддерживает Promt (английский, немецкий, русский, французский, итальянский, испанский).

Программа «Менеджер терминологии Lite» предназначена для автоматизации поиска терминологии в текстах, ее извлечения и сохранения для дальнейшей обработки. Она упрощает работу по созданию терминологических словарей, которые широко используются при переводе текстов.

Возможности программы «Менеджер терминологии Lite»:

1) позволяет одновременно открывать для анализа множество файлов;

2) выполняет подсчет относительной частоты употребления слов и словосочетаний;

3) формирует список терминологических кандидатов – слов и словосочетаний, которые встретились в анализируемых текстах, причем удаляет из этого списка общеупотребительную и служебную лексику и сортирует сформированный список по убыванию относительной частоты употребления;

4) позволяет отобразить все контекстные примеры для любого термина;

5) сохраняет в виде отдельного файла список всех терминологических кандидатов, что удобно использовать при создании терминологических словарей, и список терминологических кандидатов с контекстом, т.е. позволяет для совокупности текстов создать конкордансы по всем или отдельно выделенным терминам.

Сегодня уже разработаны новые методы и программные средства с использованием алгоритмов машинного обучения, позволяющие извлекать термины из коллекции текстовых документов предметной области с использованием структуры гиперссылок Википедии. Эти методы описаны в диссертации Н. А. Астраханцева, что позволит значительно усовершенствовать возможности ПК для решения этих задач.