Панченко Александр Иванович

Построение информационно-поискового тезауруса из корпуса текстов предметной области

Информационно-поисковый тезаурус – составленный по определенным правилам словарь терминов и словосочетаний, для определенной предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области. Тезаурус предоставляет набор нормализованной лексики, состоящий из ключевых понятий (концептов) какой-либо предметной области для индексирования коллекции документов. Кроме этого, он задает множество семантических отношений между концептами, которые могут быть эффективно использованы для улучшения навигации по корпусу текстов, рубрикации текстовых документов или даже переформулирования поисковых запросов с помощью синонимов.

Использование подобных семантических ресурсов наиболее распространено в корпоративных системах поиска и в системах поиска ориентированных на какую-либо предметную область. На данный момент существует множество тезаурусов, описывающих различные предметные области, как свободно распространяемых, таких как EuroVOC и AgroVOC, так и платных, составленных с учетом специфики предметной области той или иной организации-заказчика.

Использование информационно-поисковых тезаурусов ограничено в силу того, что ручное составление тезауруса при помощи документалистов это дорогостоящий и длительный процесс. Поддержание в актуальном состоянии подобного словаря – еще более трудная задача.

Решаемая нами задача – автоматизация составления такого тезауруса с помощью методов обработки естественного языка. Мы работаем с корпусом политических текстов из 20.000 документов на французском языке и пытаемся “реконструировать” тезаурус, созданный вручную документалистами на основании анализа того же корпуса. Поставленная задача может быть разделена на два основных этапа:

1. Выделение ключевых слов и словосочетаний предметной области, с которой соотносится корпус текстов.
2. Выявление структуры семантических отношений между данным множеством ключевых понятий.

Для того, чтобы выделить из текста множество ключевых слов мы используем коэффициент TF-IDF. Выделение множества ключевых словосочетаний производится с помощью конечных автоматов и специализированных словарей словоформ французского языка. Найденные словосочетания группируются по признаку максимальной схожести формы (максимальной общей подстроки). Для эффективного решения этой подпроблемы были применены суффиксные деревья. Фильтрация словосочетаний производится опять же с помощью коэффициента TF-IDF.

Множество отношений между ключевыми понятиями строится с помощью метода дистрибутивно-статистического анализа, основанного на анализе совместной встречаемости синтаксических контекстов, в которых встречается то или иное ключевое понятие.

Для оценки качества построенной семантической сети мы сравниваем ее с “эталонным” тезаурусом, построенным вручную. В используемой нами методике, слово (словосочетание) считается релевантным, когда в образцовом тезаурусе находиться *приблизительно такое же* слово (словосочетание). Аналогично, мы считаем, что найденное семантическое отношение является верным, если можно найти связь, либо *короткий* путь между соответствующими концептами в эталонном тезаурусе.

Проведенные эксперименты показали, что порядка 35-40% найденных отношений являются корректными, в соответствии с приведенной методикой оценивания качества.

Из проделанной работы можно сделать вывод, что полностью автоматическое построение информационно-поискового тезауруса настолько же качественного как и продукт ручной работы – сложная задача. Принципиальной трудностью при автоматическом построении тезауруса является отсутствие явной информации о некоторых семантических связях в тексте, которая, однако, используется документалистом при ручном составлении семантического словаря. Это приводит к заключению о необходимости использования дополнительных словарей, таких как WordNet, для восполнения недостатка в информации о семантических связях *высокого уровня* при автоматизированном построении информационно-поискового тезауруса.

PS

Значительная часть данной научной работы была проведена в Центре обработки естественного языка при Католическом Университете Лёвена (Бельгия). Автор благодарит фонд WBI региона Валлонии за поддержку данного исследования.

Сведения об авторе

Александр Панченко — аспирант кафедры «Системы обработки информации и управления» МГТУ им. Н.Э.Баумана

Адрес электронной почты: panchenko.alexander@gmail.com