Бабина О.И.
Babina O.
г. Челябинск

ФОРМИРОВАНИЕ ЛЕКСИКОНА ЯЗЫКА ДЛЯ СПЕЦИАЛЬНЫХ ЦЕЛЕЙ ПОСРЕДСТВОМ ЧАСТОТНО-ДИСТРИБУТИВНОГО АНАЛИЗА ЛЕКСИКИ

BUILDING A LEXICON FOR AN LSP USING FREQUENCY-DISTRIBUTION ANALYSIS

The article centers on the problem of selecting vocabulary for teaching and lexicographic resources, as well as for the purpose of developing linguistic software, which is based on the functional approach. It describes the algorithm for automatic text processing and the detection of noun phrases particular for a restricted domain, modern computer technologies for composing frequency lists and analyzing syntagmatic relations among words in the discourse of the domain applied. The algorithm allows for both monolexemic and polylexemic terms. The performance of the algorithm enables significant curtailing the list of candidate terms of the domain for further manual analysis.

Современные исследования в области лингвистики уже давно перешли на экспериментальную базу, и не обходятся без наблюдений функционирования языковых единиц в речи посредством использования обширных корпусов текстов. Знание о парадигматических особенностях языковых единиц получается в результате исследований синтагматики и наоборот. При этом, очевидно, подъязыки определенных областей (языки для специальных целей), являются наиболее благодатной почвой для применения корпусных методов, так как, в силу ограниченности используемых языковых средств, описания и закономерности, наблюдаемые в подъязыке, в большей степени регулярны по сравнению с языком в целом, объединяю-

щем в себе несколько подъязыковых подсистем [Раскин 2008]. Последние могут в значительной степени отличаться, а – в отдельных аспектах – противоречить друг другу.

В нашем исследовании в качестве языка для специальных целей нами избрана предметная область «Программирование». Корпус текстов, положенный в основу изучения, составлен из текстов статей научных журналов (таких как The Computer Journal, Computer Languages, Systems and Structures, Artificial Intelligence, Data & Knowledge Engineering, Journal of Algorithms, Science of Computer Programming, Theoretical Computer Science, The Journal of Logic and Algebraic Programming, Journal of Parallel and Distributed Programming, Journal of Discrete Algorithms, Journal of Visual Languages and Computing) за 2008 год, отобранных методом сплошной выборки, и насчитывает около 475 тыс. словоупотреблений.

Рассматривая функциональный аспект лексики, начиная с исследований группы «Статистика речи» [Статистика 1980, Алексеев 2001], методология частотного исследования лексических единиц подъязыка лежит в основе практически любого изучения подъязыка. Статистический анализ дает возможность определить степень важности той или иной лексемы (и соответствующего ей понятия) для данной предметной области. Структурный аспект лексем может быть исследован с применением частотного анализа n-грам. Содержательный компонент исследования базируется на дефиниционном анализе, а также применении аппарата теории нечетких множеств.

Для рассмотрения всех аспектов лексического состава на основе собранного корпуса текстов нами применялся следующий алгоритм анализа, автоматизация шагов которого осуществлялась с применением комплекса программных средств FLAT, разработанного под руководством С.О. Шереметьевой, и созданного нами комплекса LingAssistant:

1) Автоматическое построение частотных списков 1-, 2-, 3- и 4-грам. При этом частотный список слов (1-грам) представляется ключевым, так как именно этот список определяет состав лексики предметной области. Слово определяется по формальным признакам как последовательность словообразующих символов от пробела до пробела. В качестве словообразующих символов рассматриваются латинские буквы, цифры, знаки «-» и «'», при этом в слове должна быть хотя бы одна буква.

2) Отбор слов, составляющих необходимый лексический минимум. Под таким минимум понимается набор лексики, необходимой для понимания содержания текстов исследуемой предметной области. В работах Н.Н. Петрушевской было показано, что для понимания содержания текста достаточно знать 85% составляющих его лексических единиц [Петрушевская 1981]. В связи с этим за необходимый минимум нами принят список наиболее частотных лексических единиц (1-грам), употребление которых составляет 85% корпуса. По данным спектрового распределения [Алексеев 2001:

63] лексических единиц, составленных на основе частотных списков, нами выявлена пороговая частота для единиц, включенных в список-минимум; остальные единицы в дальнейшем не рассматриваются. В корпусе по программированию общий список словоформ составляет 17947 единиц, из которых единицы с частотой 33 и выше покрывают 85,24% текста, что составляет 1761 словоформу. Лемматизация этого списка вручную выявила 1178 лексем, которые составляют необходимый минимум [Хомутова 2009].

3) Расширение лексического минимума посредством полилексемных термов из корпуса. Асимметрия плана выражения и плана содержания является нормой для любого подъязыка, поэтому очевиден тот факт, что понятия исследуемой предметной области могут быть представлены посредством более чем одной лексемы. Для отбора регулярных полилексемных термов предметной области полезными оказываются списки n-грам. При этом очевидно, что не каждая последовательность n словоформ является семантически цельным блоком, характеризующим одно понятие. Для отбора полилексемных единиц нами разработан подалгоритм, позволяющий на основе анализа частотных списков n-грам выявить дистрибутивные последовательности, претендующие на роль терма. Алгоритм направлен на отбор из списка n-грам последовательностей, являющихся именными группами.

Алгоритм частотно-дистрибутивного анализа. В качестве лингвистической базы знаний (ЛБЗ), лежащей в основе алгоритма, выступает автоматически сформированный список парадигматических форм 1178 лексем, покрывающих 85% словоупотреблений корпуса текстов, размеченный морфологическими «тэгами» с помощью автоматизированной процедуры парадигматической идентификации словоформ [Бабина 2007] на основании материала корпуса текстов. В список входят исходные лексемы, а также дополнительно 1 форма для существительных (с окончанием *-(e)s*) и 3 для глаголов (с окончаниями *-(e)s, -(e)d, -ing*). В силу незначительной степени синтетичности английского языка, генерация парадигматических форм лексем в незначительной степени увеличивает список, который в итоге насчитывает 2716 единиц.

Работа алгоритма основана на частично контроллируемой процедуре пошаговой фильтрации списков n-грам с использованием созданной ЛБЗ. Каждый из списков n-грам (2-х, 3-х и 4-х) последовательно проходит через следующие фильтры:

1. Фильтр отбора n-грам, содержащих в своем составе только словоформы из ЛБЗ. Процедура тривиальна, и заключается в сравнении составляющих n-грамы слов с ЛБЗ.

2. Фильтр, отбирающий последовательности, являющиеся именными группами. Для отбора этих единиц сформирована база шаблонов в форме регулярных выражений. При разработке базы шаблонов исключаются именные группы, начинающиеся с определителей (артиклей, местоимений и

т.д.), так как эти элементы не включаются в денотат понятий. Для прохождения через этот фильтр для каждого n-грама генерируются все возможные варианты его разметки на основе меток соответствующих лексических единиц в ЛБЗ. Фильтр работает через механизм сопоставления с шаблонами (pattern match) – каждый вариант разметки n-грама сравнивается с шаблоном из базы. Так, например, по шаблону ~Adj(+) ~N отбирается словосочетание *propositional normal modal logics*.

После прохождения фильтров в каждом словосочетании списков выделяется ядерное слово, которое автоматически приводится к начальной форме (единственному числу) с помощью алгоритма морфологического анализа и выделения основы. Далее все n-граммы, которые являлись парадигматическими формами одного и того же словосочетания (то есть такие пары, в которых ядерное слово встречалось в единственном и множественном числе соответственно) сводятся к одному нормализованному n-граму.

В результате работы указанного алгоритма на материале корпуса по программированию получены следующие количественные данные:

2-грамы: всего – 125255; n-грамы, содержащие только лексемы из ЛБЗ – 60218; n-грамы, являющиеся именными группами – 12976; нормализованные n-грамы – 11660.

3-грамы: всего – 226931; n-грамы, содержащие только лексемы из ЛБЗ – 117692; n-грамы, являющиеся именными группами – 3993; нормализованные n-грамы – 3866.

4-грамы: всего – 244355; n-грамы, содержащие только лексемы из ЛБЗ – 116941; n-грамы, являющиеся именными группами – 283; нормализованые n-грамы – 282.

Однако около двух третей полученных таким образом нормализованных n-грамов встречаются в корпусе лишь единожды. Такие сочетания рассматриваются как случайная комбинация слов и не включаются в набор регулярных дистрибутивных конструкций (кандидатов в термы предметной области). Отбросив n-грамы с частотой 1, мы сформировали списки, которые насчитывают: 2-грамы - 4813; 3-грамы - 783; 4-грамы – 32, итого 5628 словосочетаний.

Словосочетания из этих окончательных списков являются кандидатами для включения в лексикон языка для специальных целей в качестве термов. Очевидно, что автоматически отобранные n-грамы могут содержать неточности и должны быть проверены вручную. Однако представленный алгоритм в значительной степени сокращает список кандидатов в термы предметной области, которые необходимо подвергнуть мануальной проверке.

Таким образом, представленный алгоритм позволяет включать в лексикон языка для специальных целей как однословные, так и многословны термы. При этом разработанный комплекс фильтров и процедур с привле-

чением лингвистической базы знаний дает возможность минимизировать усилия лингвистов в процессе отбора термов для включения в лексикон.

Библиография

1. Раскин, В. К теории языковых подсистем. Изд. 2-е, доп. / В. Раскин. – М.: Изд-во ЛКИ, 2008. – 424 с.

2. Статистика речи и автоматический анализ текста / под ред. П.М. Алексеева, Р.Г. Пиотровского и др. – Л.: Наука, 1974. – 402 с.

3. Алексеев, П.М. Частотные словари: Учебное пособие / П.М. Алексеев. – СПб.: Изд-во С.-Петерб. ун-та, 2001. – 156 с.

4. Петрушевская, Н.Н. Формирование рецептивного и потенциального словарей в процессе обучения чтению на иностранном языке в неязыковом вузе (на материале англ. яз. свароч. пр-ва): Автореф. дис. ... канд. пед. наук: Специальность 13.00.02 / Н.Н. Петрушевская. – Л., 1981. – 22 с.

5. Хомутова, Т.Н. Научный текст: лингво-когнитивный анализ / Т.Н. Хомутова, О.И. Бабина // Вестник ЮУрГУ. Сер. Лингвистика. – Челябинск: Изд-во ЮУрГУ, 2009. – Вып. 9.

6. Бабина, О.И. Автоматическая первичная парадигматическая идентификация словоформ на основе анализа их морфологической квазиструктуры / О.И. Бабина // Актуальные проблемы теоретической и прикладной лингвистики: материалы международной научной конференции / отв. ред. О.А. Турбина (Челябинск, 10-13 декабря 2007 г.). – Челябинск: Изд-во ЮУрГУ, 2007. – Ч. 1. – С. 42-45.