// Machine Learning. − Springer Netherlands. − 1999. − №36. – P. 105–139. ***11.*** Hastie T., Tibshirani R., Friedman J. The Elements of Statistic Learning; Data Mining, Inference, and Predicting. – Springer-Verlag. –2001.

Поступила в редколлегию 26.11.07

УДК 007.681

О. Я. ЛАЗАРЕВА, канд. техн. наук

МЕТОДИКА АВТОМАТИЗАЦИИ ФОРМИРОВАНИЯ ТЕРМИНОЛОГИЧЕСКИХ СЛОВАРЕЙ

В статті пропонується один з підходів до створення термінологічних словників. Основою реєстру термінологічного словника може бути частотний список словосполучень певної граматичної структури, автоматично виділених з текстів деякої предметної галузі. Напевно, що такий словник буде надмірним і буде потребувати коригування експертом.

В статье предлагается один из подходов к созданию терминологических словарей. Основой словника терминологического словаря может служить частотный список словосочетаний определенной грамматической структуры, автоматически выделенных из текстов определенной предметной области. Естественно, что такой словарь будет избыточным и потребует корректирования экспертом.

The paper suggests an approach to creation of the vocabulary of terminological dictionaries. Word combinations of definite grammar structures being automatically extracted from the texts of some domain and arranged in a frequency list may constitute the basis for a terminological dictionary. The obtained vocabulary is sure to be excessive and should be edited by an expert.

Современное состояние «инфосферы», пронизывающей все виды человеческой деятельности, характеризуется лавинообразным увеличением потоков информации, все большим слиянием и взаимопроникновением отдельных областей знаний, что вызывает потребность в создании и обновлении словарей, обслуживающих данные предметные области. В связи с этим возникает необходимость пересмотра подходов к самому процессу создания или пополнения словарей. Традиционно специалисты-предметники, часто в сотрудничестве с лингвистами, составляли такие словари - и словари на бумажных носителях, и словари, использующиеся в автоматизированных информационных системах (классификаторы, рубрикаторы, информационно-поисковые тезаурусы) - на основе анализа достаточно большой коллекции документов по некоторой тематике. И совершенно очевидно, что сам процесс накопления массива лексических единиц был и остается чрезвычайно трудоемкой операцией. Кроме того, учитывая стремительное развитие практически всех отраслей науки и технологий, с одной стороны, и достаточно длительный процесс подготовки или модернизации словарей, с

другой, такие словари всегда будут отставать по составу лексики от современного состояния отрасли, то есть не отвечать требованиям полноты.

Эти проблемы и послужили причиной поиска методики, позволяющей автоматизировать процесс первичного накопления базы терминологический единиц некоторой предметной области, которая позволит существенно сократить время и трудозатраты на создание словарей.

Основополагающим принципом такой методики служит предположение, что необходимым и достаточным источником терминологической лексики являются специальные тексты, или, другими словами, достаточно большой массив специальных текстов может практически полностью покрывать лексический и терминологический состав некоторого терминологического подъязыка. Другим важным принципом методики является предположение, что потенциально любое именное словосочетание может быть термином или может содержать термин. Таким образом, для решения поставленной задачи необходимо разработать методы, которые позволяли бы автоматизировать процесс выделения из текстов таких именных словосочетаний, структура которых соответствует структуре термина. Сформированный таким образом словарь, естественно, будет избыточным и будет требовать постредактирования. Однако, трудозатраты, связанные с постредактированием, будут несоизмеримо меньшими по сравнению с трудозатратами на создание словарей традиционными методами. Точное определение корпуса текстов, необходимых для создания словаря, представляет отдельную задачу и не рассматривается в данной работе. Однако, приблизительную оценку можно сделать эмпирически – в процессе создания словаря следить за скоростью его роста. Если поступление новых слов или словосочетаний практически прекращается, то есть процент новых поступлений относительно, например, общего количества слов в тексте составляет меньше некоторого достаточно малого значения, то пополнение словаря можно прекратить.

Прежде всего, определимся, какие же словосочетания следует искать в текстах. Исследователи терминологии [6] выделяют достаточно много синтаксических структур терминов, которые включают, прежде всего, наиболее распространенную модель «единичное существительное», а также двух-, трех- или многокомпонентные словосочетания, содержащие прилагательные, причастия, наречия и даже глаголы. Однако, наиболее продуктивными [4] являются субстантивированные словосочетания с прилагательным или причастием (адъективный тип): *программное обеспечение, обратимый термодинамический процесс*, а также субстантивированные словосочетания с существительным (или их последовательностью) в родительном падеже (атрибутивный тип): *главный момент инерции*, *цепь переменного тока*. Оценки, приведенные в работе [5], показывают, что термины другой структуры, например, с предлогами или наречиями, составляют меньше одного процента от общего числа терминов.

Часто исследователи, работающие над аналогичными задачами [1], вводят ограничение на длину словосочетания, включаемого в словарь – не более 3 слов. С одной стороны, действительно – длина большинства терминов не превышает трех элементов. Однако существуют и намного более длинные термины, например: *многокомпонентная мелкодисперсная сухая смесь, линейное бесконечномерное евклидово полное сепарабельное пространство, область управления программными средствами, поворот плоскости поляризации световой волны* и пр., поэтому в предлагаемой методике мы снимаем все ограничения на длину выделяемого словосочетания.

Основой для автоматического выделения в тексте словосочетаний заданной структуры является, прежде всего, морфологический анализ текста, т.е. определение грамматических характеристик слов. Наиболее приемлемым для решения поставленной задачи считаем метод морфологического анализа с использованием словаря квазифлексий [2, 3]. Преимущество такого метода состоит в том, что он использует общие словоизменительные закономерности языка и тем самым обеспечивает достаточную точность, а также независимость применяемых методов автоматизации от тематики анализируемых текстов.

Схематично типы наиболее продуктивных терминологических конструкций можно представить следующим образом:

$$S_1:=[<\text{адъектив}>] \dots <\text{существительное}> \text{ или } [A_i] \dots N;$$

$$S_2:=S_1 \{S_1\}^p \dots ,$$

где квадратные скобки содержат необязательный элемент, а фигурные – обязательный.

Совершенно очевидно, что не все словосочетания такой структуры будут терминами: это и общеупотребительные слова, и свободные словосочетания, и словосочетания, в которых термин является их составной частью. Но что же можно считать критерием для включения или не включения выделенного словосочетания в словарь? Это, прежде всего, его принадлежность предметной области и еще целый перечень критериев, которые определяют свойства лексической единицы как элемента терминосистемы. А это уже экстралингвистическая информация, формализация которой представляется маловероятной сегодня. Поэтому мы ставим задачу выделения всех слов и словосочетания данной структуры по формальным признакам, а затем предлагаем предоставить специалисту возможность редактировать полученный список.

Если проанализировать расположение терминов в предложении с учетом их структуры, то оказывается, что термин (в частности структуры типа S_1 или S_2) могут быть как самостоятельной синтаксической единицей предложения, так и входить в состав словосочетания аналогичной или другой структуры. Поэтому предлагается выявленные в тексте начальные словосочетания

разделять на компоненты и формировать из них все возможные сочетания – производные дескрипторы, структура которых также должна соответствовать структуре термина.

Производные дескрипторы для конструкции S_1 формируются путем генерации всех возможных комбинация адъективов с существительным. При этом заведомо исключаются инвертированные цепочки типа $A_i+_kA_iN$. Это объясняется тем, что порядок следования адъективов носит не случайный характер, а обусловлен семантическими отношениями. Так, например, из цепочки *незатухающее гармоническое колебание* будет сформировано 3 производных дескриптора: *колебание; незатухающее колебание и гармоническое колебание*, которые вместе с начальным будут занесены в словарь.

Для атрибутивных конструкций (тип S_2) процесс генерации производных дескрипторов носит несколько иной характер, исходя из того, что отношения между компонентами этой конструкции линейные, то есть направление синтаксических связей направлено от предшествующего элемента к последующему без скачков: $S_1^1 \rightarrow S_1^2 \rightarrow ... \rightarrow S_1^3$. Генерация состоит в формировании всех допустимых линейных цепочек адъективных компонентов длиной от 1 до *n*, где *n* - число адъективных компонентов атрибутивной конструкции. Таким образом, из начального словосочетания этого вида, например, *закон сохранения энергии невзаимодействующих частиц*, будет сформировано 9 производных единиц: *закон; сохранения; энергии; невзаимодействующих частиц; закон сохранения; сохранения энергии; энергии невзаимодействующих частиц; закон сохранения энергии; сохранения энергии невзаимодействующих частиц*, и они также будут занесены в словарь. Кроме того, в соответствии с правилами обработки адъективных словосочетаний в словарь будут включены также элементы *частиц; энергии частиц; сохранения энергии частиц; закон сохранения энергии частиц*. Некоторые из производных дескрипторов (а, возможно, и многие) не будут являться терминами, а иногда и вообще не будут иметь смысла для данной предметной области. Тем не менее, подобная процедура позволит не пропустить термины, являющиеся составной частью свободного словосочетания.

Полученные в результате описанных преобразований словосочетания дополнительно подвергаются процедуре инвертирования, т.е. расположения элементов таким образом, чтобы главное слово – первое существительное – оказалось на первой позиции. Это дает возможность при сортировке полученного списка по алфавиту получить интересную картину лексико-семантических гнезд, объединенных общим главным словом.

Кроме того, во избежание повторения одних и тех же слов и словосочетаний, которые в текстах встречаются в различных падежно-числовых формах, дескрипторы необходимо нормализовать, т.е. привести к

начальной форме главное существительное и согласованные с ним прилагательные.

Важным критерием для включения полученных дескрипторов в тематический словарь является показатель частоты употребления этого слова или словосочетания. Очевидно, что вероятность того, что некоторый дескриптор действительно является термином данной предметной области, прямо пропорциональна его частоте и тем выше, чем больше слов входит в него. Таким образом, показатель

$$K = (F-1)*Q,$$

где К – вероятностный коэффициент, F - частота дескриптора, Q - количество слов в словосочетании, можно использовать как оценочный для принятия решения о включении данного слова или словосочетания в словарь. Сортировка полученного словаря по частоте, или по описанному выше показателю, несмотря на кажущуюся простоту такого подхода, дает неплохую вершину списка на достаточно больших объемах текстов. Конечно, наиболее частотными будут однословные дескрипторы, в особенности общеупотребительные и общенаучные слова типа *система, метод* и т.п. Однако мы уже отмечали, что данная методика не учитывает прагматику и семантику, а использует лишь формальные методы для отбора слов-претендентов на включение в словарь, а решающее слово всегда предоставляется эксперту.

Описанная методика автоматизации первичного отбора дескрипторов из текстов некоторой предметной области позволит существенно сократить время на разработку терминологических словарей и избавит эксперта от рутинной и чрезвычайно трудоемкой работы по обработке текстов.

Список литературы: ***1***. *Антонов А.В.* Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации // НТИ. Сер. 1. – 2001. – № 8. – С.12–21. ***2***. *Белоногов Г.Г., Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации. - М., 1979. - 253 с. ***3***. *Грязнухина Т.А., Дарчук Н.П., Клименко Н.Ф.* Использование ЭВМ в лингвистических исследованиях. - К., 1990. - 223 с. ***4***. *Лазарєва О.Я.* Про деякі продуктивні моделі науково-технічних термінів // Українська термінологія і сучасність: Зб.наук.праць. Вип. 4. – К: КНЕУ, 2001. – С.185-188. ***5***. *Лукашевич Н.В.* Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России//НТИ. Сер.2. - 1995. - №3. - С.21-24. ***6***. *Суперанская А.В., Подольская Н.В., Васильева Н.В.* Общая терминология. Вопросы теории. – М.: УРСС – 2003.

Поступила в редколлегию 27.11.07