

Авторы: С.М. Вороной, А.А. Егорова

Источник: <http://dspace.nbuv.gov.ua:8080/bitstream/handle/123456789/7799/08-Voronoy.pdf?sequence=1>

Словообразовательная база знаний экспертной обучающей системы

Для экспертной обучающей системы предложена логическая структура словообразовательной базы знаний и формальная модель узлов дерева, включающая описание методов словообразования с использованием функций выбора.

Введение

Проблема обработки естественной языковой информации остается актуальной на протяжении последних десятилетий. Системы информационного поиска, диалоговые системы, средства машинного перевода и автоматического реферирования, модули проверки правописания используют анализ текстов, написанных на естественном языке.

Использование словообразовательного компонента в информационно-поисковых системах предоставляет возможности для расширения полноты запроса, необходимость которого вызвана малым количеством обнаруженных ресурсов. Наличие модуля словообразования в системах обработки текстов приспособливает их к работе с неопознанными словами, которые образованы путем сложения основ, конверсией частей речи, с помощью аффиксов и т.д.

Словообразовательные процессы – это основной путь пополнения лексики языка, в связи с этим модуль словообразования является неотъемлемой частью современных интеллектуальных информационно-поисковых и обучающих систем с естественной языковым интерфейсом.

Актуальность разработки системы обучения словообразованию обусловлена тем, что знание словообразовательной системы способствует формированию и развитию у изучающих русский язык навыков грамотной речи: правильного употребления производных слов в структуре синтаксических единиц, соблюдения норм согласования и управления, умения пользоваться синонимическими равноуровневыми языковыми средствами и т.д. [1].

Постановка задачи

Обучение включает в себя больше, чем просто представление информации; необходима проверка действий обучаемого с динамичной обратной связью в процессе обучения для избежания ошибочных выводов, а также отложенная обратная связь для периодической оценки знаний обучаемого. Парадигма экспертной системы позволяет очень четко разделить знания и их обработку, увеличивая возможность многократного проведения такого процесса [2].

Одним из основных компонентов экспертной системы является база знаний (БЗ), предназначенная для хранения долгосрочных данных, описывающих словообразова-

тельную область (словари флективных классов, корней, аффиксов и окончаний) [3], а также правил словообразовательного синтеза и чередований.

Разработка БЗ на основе устроенного по семантическому принципу словообразовательного словаря Тихонова позволит применять при словообразовательном анализе и синтезе основные принципы объектно-ориентированного программирования, в первую очередь – наследование. Мотивационные и семантические отношения в словообразовании можно трактовать и использовать как связи множественного наследования признаков. Множественного наследования потому, что мотивированное слово наследует признаки как минимум от двух источников: от слова основы и от словообразующего форманта.

Целью статьи является разработка логической структуры БЗ для экспертной обучающей системы словообразованию русского языка.

Словообразовательная база знаний экспертной обучающей системы

Словообразовательная БЗ представляет собой лес, в качестве деревьев которого выступают словообразовательные гнезда словаря Тихонова.

Дерево – одна из наиболее распространенных структур, используемых для представления данных в ЭВМ. Подобные структуры широко применяются при организации банков данных, систем управления базами данных, в системах программного имитационного моделирования сложных комплексов и т.д. Особое значение сетевые структуры приобрели в системах искусственного интеллекта, в которых они адекватно отражают логику организации данных и сложные отношения, возникающие в таких системах между различными элементами данных. В этих системах деревья применяются для представления логических конструкций, необходимых для представления знаний, образования понятий и осуществления логических выводов.

Формально дерево (tree) представляет собой конечное множество T одного или более узлов со следующими свойствами:

- существует один выделенный узел, а именно корень (root) данного дерева T ;
- остальные узлы распределены среди $m \geq 0$ непересекающихся множеств T_1, \dots, T_m , и каждое из этих множеств в свою очередь является деревом, деревья T_1, \dots, T_m называются поддеревьями (subtrees) данного корня.

Выбор представления дерева зависит от решаемой задачи и способа ее решения.

Узлом дерева назовем структуру

$$A_{ij}^k = \langle U(R), z_a, S_a(x_{ij}) \rangle, \quad (1)$$

где $U(R)$ – объединение элементов множества формантов R , представляющее собой производящую основу;

z_a – часть речи слова, образующегося в узле A_{ij}^k ;

$S_a(x_{ij})$ – функция, задающая способ словообразования, с помощью которого образуется узел A_{ij}^k , x_{ij} – формант.

Корнями деревьев являются первые и обязательные ступени словообразования, которые являются производными.

На каждой ступени словообразования может быть образовано большое число производных слов. Порядок размещения узлов (производных слов) дерева на каждом ярусе учитывает их семантическую близость к родительскому узлу (производящему слову), а также лексико-грамматические и словообразовательные отношения. Например,

для имени существительного со значением лица наиболее семантически близкими являются уменьшительно-ласкательные и увеличительные существительные; за ними идут названия лиц женского пола, детей (при названии животных – самок и детенышей).

Самые близкие производные слова в лексико-грамматическом отношении для качественных прилагательных – это формы оценки. Для глаголов – это возвратные глаголы и существительные со значением процесса.

Наиболее широким является первый ярус дерева. Принцип размещения узлов (производных слов) следующий:

1) если родительский узел – имя существительное, то дочерние узлы размещены в таком порядке:

- а) формы оценки исходного существительного;
- б) остальные имена существительные;
- в) имена прилагательные;
- г) наречия;
- д) префиксальные и префиксально-суффиксальные имена существительные и имена прилагательные (в алфавитном порядке);
- е) глаголы.

В каждом из этих разрядов может быть один, несколько или множество узлов. Однако редки случаи, когда в ярусе встречаются все перечисленные разряды.

2) если родительский узел – имя прилагательное:

- а) субстантивные прилагательные;
- б) формы оценки исходного прилагательного;
- в) приставочные имена прилагательные;
- г) наречия;
- д) имена существительные;
- е) префиксальные прилагательные и наречия;
- ж) глаголы.

3) если родительский узел – имя числительное:

- а) собирательные числительные;
- б) существительные;
- в) количественные существительные;
- г) наречия;
- д) прилагательные.

4) если родительский узел – глагол:

- а) возвратный глагол;
- б) суффиксальная форма несовершенного вида;
- в) однократный глагол;
- г) многократный глагол;
- д) отглагольные существительные;
- е) причастия;
- ж) прилагательные;
- з) наречия;
- и) префиксальные и префиксально-суффиксальные глаголы.

Местоимения и наречия как производящие основы выступают редко, поэтому необходимость разработки принципа размещения производных не возникает.

Для учета описанного порядка размещения узлов в структуру, описывающую элемент узла дерева, вводится дополнительный элемент *K*, представляющий собой

бинарный массив, длина которого равна максимальному числу категорий перечисленных выше частей речи. Наибольшим числом категорий, равным девяти, обладает глагол

$$K = [k_1, k_2, \dots, k_9]. \quad (2)$$

Если у текущего родительского узла существуют потомки, обладающие i -м свойством, то $k_i = 1$, если же таких потомков нет или число свойств меньше i , то $k_i = 0$. То есть, например,

- для существительного $\forall k_i (i > 6 \rightarrow k_i = 0)$;
- для прилагательного $\forall k_i (i > 7 \rightarrow k_i = 0)$.

Таким образом, структура (1) будет иметь вид

$$A_{ij}^k = \langle U(R), z_a, S_a(x_{ij}), K \rangle. \quad (3)$$

Причем, под именами элементов массива будет подразумеваться название категории, свойственной части речи z_a . То есть, если узел – имя существительное, то выражение $k_l = 1$ будет означать наличие потомков, обозначающих формы оценки исходного существительного, однако если в качестве родительского узла выступает глагол, то выражение $k_l = 1$ будет означать наличие потомков, являющихся возвратными глаголами.

Рассмотрим первую ступень словообразования прилагательного бедный:

бедн(ый)

бедн-ейш-ий

бедн-оват-ый

бедн-еньк-ий

бедн-о

бедн-ость

бедн-от-а

бедн-як

без-бедн-ый

пре-бедн-ый

бедн-е-ть

при-бедн-ить-ся

Для данного примера элементы массива категорий будут иметь такие значения: $K = \{0, 1, 1, 1, 1, 1, 1, 0, 0\}$. Первые семь элементов обозначают наличие или отсутствие потомков, обладающих характеристиками, свойственными производным прилагательного. Последние два элемента равны нулю, так как категорий производных прилагательного существует только семь.

Аффиксы, хранящиеся в словарях, структура которых приведена в [3], будем трактовать как факты БЗ. А функции $S_a(x_{ji})$, задающие законы словообразования, – как правила БЗ.

Однако, следует отметить, что поведение одного и того же аффикса в разных ситуациях различно, то есть одному и тому же аффиксу могут соответствовать разные правила.

Например, рассмотрим суффикс *-щик* при образовании имен существительных.

Случай 1: мотивирующее слово (родительский узел) – глагол. В данной ситуации суффикс *-щик* вызывает следующие чередования на морфемном шве: $k - ч$, $z - ж$, $л - ль$ (размолоть – размольщик). К тому же конечная гласная производящей основы и финаль *-ива* не сохраняются (протирать – протирщик).

Случай 2: мотивирующее слово – имя прилагательное. В этом случае суффикс *-щик* не дает чередований вообще. Но финали *-н-* (после согласной) и *-ск-* основы

мотивирующего слова отсутствуют в образованном существительном (инструментальный – инструментальщик).

Случай 3: мотивирующее слово – имя существительное. В данной ситуации перед морфом *-щик* чередуются *л – ль* (факел – факельщик), *ск – щ* (сыск – сыщик), *ст – щ* (поместье – помещик). Финаль *-к- /-ок* мотивирующего слова в производном слове отсутствует (шарманка – шарманщик).

Таким образом, способ словообразования можно представить следующим выражением:

$$S_a = \{s_i \in S|y\}, \quad (4)$$

где s_i – способ словообразования объекта, выбираемый из множества S альтернативных способов, выбираемых по некоторому условию y .

Условие y представляет собой следующий кортеж:

$$y = \langle \pi, \varphi \rangle, \quad (5)$$

где π – совокупность сведений об объекте, а φ – множество правил (словообразования и чередования).

Сведения об объекте представляют собой множество информационных характеристик, таких, как код аффикса (x), часть речи родительского узла и текущего (c), финаль основы производящего слова (f)

$$\pi = \langle x, c, f \rangle. \quad (6)$$

Части речи родительского и текущего узлов предлагается представить в виде кодов, первый разряд которых соответствуют коду части речи родительского узла, а второй – текущего.

Заключение

Таким образом, в настоящей работе предложена логическая структура словообразовательной базы знаний и формальная модель узлов дерева, включающая описания методов словообразования с использованием функций выбора.

В дальнейшем планируется разработка эвристического алгоритма нахождения пути к узлу дерева, обладающему требуемой семантикой. Полученные результаты применяются при разработке экспертной обучающей системы словообразованию русского языка.

Литература

1. Потиха З.А. Современное русское словообразование. – М.: Просвещение, 1970.
2. Рыбина Г.В., Рыбин В.М. Опыт разработки и перспективы использования обучающих интегрированных экспертных систем в учебном процессе // Научная сессия МИФИ – 2007. Сб. научных трудов: В 17 т. – М.: МИФИ. – Т. 3. Интеллектуальные системы и технологии. – С. 37-39.
3. Егошина А.А. Об одном способе построения статического словаря морфологического процессора // Материалы Седьмой Международной научно-технической конференции «Искусственный интеллект. Интеллектуальные и многопроцессорные системы – 2006». – Таганрог: Изд-во ТРТУ. – 2006. – Т. 2. – 404 с.

С.М. Вороной, Г.А. Егошина

Словотворча база знань експертної навчальної системи

Для експертної навчальної системи запропонована логічна структура словотворчої бази знань і формальна модель вузлів дерева, що включає опис методів словотвору з використанням функцій вибору.

S. Voronoy, A. Yegoshina

Word Formation Knowledge Base for Expertise Learning System

A logical structure of word formation knowledge base and formal model of tree nodes, which includes a description of the methods using the word choice was offered for expertise learning system.

Статья поступила в редакцию 26.11.2008.