7В

УДК 004.89:004.4

С.М. Вороной, А.А. Егошина
Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина
postmaster@iai.donetsk.ua

Метод поиска ключевого узла словообразовательного дерева для экспертной обучающей системы

В статье рассматривается проблема обработки значительного объема производных слов, получаемых в результате словообразовательного синтеза в естественно-языковых системах, что приводит к затруднению и замедлению процесса поиска необходимого слова. Предлагается эвристический метод словообразования по заданной семантике, позволяющий снизить временные затраты на построение производного слова за счет уменьшения количества анализируемых вершин словообразовательного дерева.

Введение

В настоящее время существует значительное количество систем, исполняющих такие задачи, как компьютерная обработка больших массивов естественно-языковых текстов, естественно-языковое общение системы с пользователем, создание больших банков информации на основе естественного языка, разработка языка посредников в многоязыковой информационной среде. Неотъемлемой частью подобных систем является словообразовательный модуль, обеспечивающий пополнение словарного запаса новыми словами и словосочетаниями. Нерешенной остается задача обработки значительного объема производных слов, получаемых в результате словообразовательного синтеза, что приводит к затруднению и замедлению процесса поиска необходимого слова.

В работах [1], [2] проводится исследование зависимости семантики производного слова от ситуации, определяемой рядом формальных, семантических и грамматических свойств мотивирующего слова. На основании данной зависимости разработана формализация семантики словообразовательных формантов с помощью базисных лексических функций и предложено решение задачи словообразовательного синтеза на основе применения семантических свойств формантов. Работа [3] посвящена разработке логической структуры словообразовательной базы знаний и формальной модели узлов дерева словообразования для экспертной обучающей системы.

На основании предложенных моделей в данной работе рассматривается решение задачи поиска ключевого узла словообразовательного дерева. Под ключевым узлом понимается узел словообразовательного дерева, семантика которого соответствует семантическому описанию, заданному пользователем экспертной обучающей системы для построения необходимого производного слова.

Целью данной работы является разработка эвристического метода поиска ключевого узла словообразовательного дерева, согласно которому вершины-кандидаты словообразовательного дерева упорядочиваются по убыванию оценочной функции, в качестве которой предложено использовать меру семантической близости вершины-кандидата к заданной семантике.

Основные этапы словообразовательного синтеза

Задача словообразовательного синтеза может быть представлена в виде последовательности следующих этапов.

1. Морфологический анализ начального (производящего) слова. На данном этапе по алгоритму, предложенному авторами в работе [4], проводится разбиение исходного слова на составляющие его морфемы с последующим определением его грамматических характеристик, основной из которых является часть речи.

2. В словообразовательной базе знаний, представляющей собой лес (Forest), деревьями (trees) которого являются словообразовательные гнезда, выполняется поиск дерева (tree_cur), корень которого соответствует корню исходного слова, выделенного на этапе морфологического анализа.

3. В найденном на предыдущем шаге tree_cur выполняется поиск узла (node_cur), соответствующего базовому слову [5]. В результате выделяется поддерево (subtree_cur), корнем которого является ключевое слово.

4. В subtree_cur проводится поиск пути к узлу дерева, семантика которого соответствует заданной F0. Необходимо определить именно путь, представляющий собой цепочку правил, соответствующих законам словообразования, которые необходимо выполнить для построения слова заданной семантики.

Так как корни (узлы) деревьев представляют собой структуру, содержащую код производящей основы узла, часть речи слова, образующегося в узле, и функцию, задающую способ словообразования, с помощью которого образуется узел, то вначале выполняется поиск кода радикала (исходного корня), выделенного на этапе морфологического анализа, в словаре корней.

В связи с тем, что trees упорядочены в порядке возрастания кодов их корней, поиск дерева (tree_cur), корень которого соответствует корню исходного слова, не является сложной задачей и выполняется методом бинарного поиска.

Метод поиска ключевого узла словообразовательного дерева для экспертной обучающей системы

Поиск узла с заданной семантикой проводится в поддереве subtree_cur, выделенном на предыдущем шаге. В качестве основы алгоритма используется метод поиска в глубину.

Идея этого метода – идти вперед в неисследованную область, пока это возможно, если же вокруг все исследовано, отступить на шаг назад и искать новые возможности для продвижения вперед. Метод поиска в глубину известен под разными названиями, например, «бэктрекинг», «поиск с возвращением». Понятия новой, открытой, закрытой и активной вершин для поиска в глубину имеют такой же смысл, как и для поиска в ширину. Отметим, что всегда имеется не более чем одна активная вершина.

Обход начинается с посещения заданной стартовой вершины $A^0_{subtree}$, которая становится активной и единственной открытой вершиной. Затем выбирается инцидентное вершине $A^0_{subtree}$ ребро $(A^0_{subtree}, A^j_{ik})$ и посещается вершина A^j_{ik}. Она становится открытой и активной. Заметим, что при поиске в ширину вершина оставалась активной до тех пор, пока не были исследованы все инцидентные ей ребра. В дальнейшем, как и при поиске в ширину, каждый очередной шаг начинается с выбора активной вершины из множества открытых вершин. Если все ребра, инцидентные активной вершине, уже исследованы, она превращается в *закрытую*. В противном случае выбирается одно из неисследованных ребер, это ребро исследуется. Если вершина новая, то она посещается и превращается в *открытую*.

Главное отличие от поиска в ширину состоит в том, что при поиске в глубину в качестве активной выбирается та из открытых вершин, которая была посещена последней. Для реализации такого правила выбора наиболее удобной структурой хранения множества открытых вершин является стек: открываемые вершины складываются в стек в том порядке, в каком они открываются, а в качестве активной выбирается последняя вершина. Методы полного перебора (поиска в глубину) обеспечивают решение задачи поиска пути, но в случае поиска узла, обладающего заданной семантикой, данные методы невозможно использовать, поскольку при поиске необходимо раскрыть слишком много вершин, прежде чем требуемый путь будет найден. Так как всегда имеются практические ограничения на время вычисления и объем памяти, то нужны другие методы, более эффективные, чем методы слепого перебора.

В рассматриваемой задаче можно сформулировать правила, позволяющие уменьшить объем перебора. Такие правила зависят от особенностей семантики словообразования, база знаний которого представляется в виде леса. Описание особенностей словообразовательной семантики является эвристической информацией (помогающей найти решение), а использующие ее процедуры поиска – эвристическими методами поиска.

Один из путей уменьшения перебора состоит в том, чтобы вместо размещения вновь построенных вершин в произвольном порядке в начале списка *ОТКРЫТ* их можно расположить в нем некоторым определенным образом, зависящим от эвристической информации. Так, при переборе в глубину в первую очередь будет раскрываться та вершина, которая представляется наилучшей. Для того чтобы применить процедуру упорядочения, нам необходима мера, которая позволяла бы оценивать «перспективность» вершин. Такие меры называют оценочными функциями $\varphi(A^j_{ik})$. Оценочная функция должна обеспечивать возможность ранжирования вершин-кандидатов на раскрытие с тем, чтобы выделить ту вершину, которая с наибольшей вероятностью находится на лучшем пути к цели.

Значение производящего слова всегда в той или иной мере организует семантику производного. Как правило, лексическое значение производящего слова входит в лексическое значение производного в полном объеме, всеми своими семантическими компонентами. Поэтому в качестве меры оценки «перспективности» вершины $\varphi(A^j_{ik})$ используется мера семантической близости вершины-кандидата к заданной семантике:

$$\varphi(A^j_{ik}) = \delta(A^j_{ik}, F^0). \quad (1)$$

Мера семантической близости δ слов А и В определяется следующим образом:

$$\delta(A,B) = \frac{\tau(A,B)}{L_{\max}}, \quad (2)$$

где

$$\tau(A,B) = \sum_{i=1}^{L\min} \lambda_i, \lambda_i = \begin{cases} 1, a_i = b_i; \\ 0, a_i \neq b_i \end{cases} \quad (3)$$

$$L_{\max} = \max(L_a, L_b), L_{\min} = \min(L_a, L_b). \quad (4)$$

После принятия этих необходимых мер метод упорядоченного поиска может быть представлен такой последовательностью шагов:

1. Поместить начальную вершину $A^0_{subtree}$ в список, называемый *ОТКРЫТ*, и вычислить $\varphi(A^0_{subtree})$.

2. Если список *ОТКРЫТ* пуст, то на выход дается сигнал о неудаче; в противном случае – переходи к следующему этапу.

3. Взять из списка *ОТКРЫТ* ту вершину, для которой φ имеет наименьшее значение, и поместить ее в список *ЗАКРЫТ*. Дать этой вершине название A^n. (В слу-

чае совпадения значений выбирать вершину с минимальными φ произвольно, но всегда отдавая предпочтение целевой вершине) .

4. Если A^n есть целевая вершина ($f(A^n) = F^0$), то на выход выдать решающий путь, получаемый прослеживанием соответствующих указателей; в противном случае – переходить к следующему шагу.

5. Раскрыть вершину A^n, построив все непосредственно следующие за ней вершины. (Если таковых нет, переходить к шагу (2).) Для такой дочерней вершины A^n_i вычислить значение $\varphi(A^n_i)$.

6. Связать с теми из вершин A^n_i, которых еще нет в списках *ОТКРЫТ* или *ЗАКРЫТ*, только что прочитанные значения $\varphi(A^n_i)$. Поместить эти вершины в список *ОТКРЫТ* и провести от них к вершине A^n указатели.

7. Связать с теми из непосредственно следующих за A^n вершинами, которые уже были в списке *ОТКРЫТ* или *ЗАКРЫТ*, меньшие из прежних или только что вычисленных значений φ. Поместить в список *ОТКРЫТ* те из непосредственно следующих за A^n вершин, для которых новое значение φ оказалось ниже, и изменить направление указателей от всех вершин, для которых значение φ уменьшилось, направив их к A^n.

8. Перейти к (2).

Практической реализацией предложенных методов является экспертная система обучения словообразованию русского языка. Экспертная обучающая система взаимодействует с учащимся в двух основных режимах: в первом режиме подсистема демонстрирует и поясняет обучаемому ход решения задачи построения слова по заданной семантике; второй режим заключается в решении аналогичной задачи обучаемым под постоянным контролем со стороны системы, которая анализирует полученную от обучаемого информацию на каждом шаге решения задачи, корректирует его действия в случае неудачных шагов и может перейти к первому режиму обучения по просьбе обучаемого, т.е. продолжить решение задачи за него.

Исследование эффективности метода поиска ключевого узла словообразовательного дерева

С помощью разработанной экспертной обучающей системы проведено исследование алгоритмов морфологического анализа [4], словообразовательного синтеза и эффективности разработанной экспертной обучающей системы в целом. В качестве критериев оценки адекватности разработанной экспертной обучающей системы словообразованию русского языка были использованы следующие: эффективность; время отклика; надежность.

Для оценки эффективности и надежности результатов работы алгоритма морфологического анализа была проведена серия экспериментов с привлечением экспертов, которые определяли такие грамматические характеристики словоформ, как часть речи, род, число и падеж. Для этих же словоформ были получены аналогичные грамматические характеристики с помощью предложенного алгоритма. Результаты экспериментов показали, что в двух из двадцати проведенных экспериментов ответы эксперта и системы не совпали. В обоих случаях в дереве словоформ (окончаний) возникала омонимия. Так как анализируемые словоформы не обладали суффиксом, то дальнейший анализ дерева суффиксов, как предложено в алгоритме для устранения омонимии, был невозможен. Однако наиболее значимым из морфологических признаков для дальнейшей корректной работы алгоритма семантически-ориентированного словообразовательного синтеза является часть речи, которая была правильно идентифицирована во всех экспериментах. На основании анализа полученных результатов можно сделать заключение, что алгоритм является эффективным и надежным.

Эффективность метода поиска узла дерева, соответствующего ключевому слову, и метода поиска узла с заданной семантикой была оценена с помощью функции временной сложности *T(Km)*, где *Km* – количество морфем в слове. Было проведено 10 экспериментов на словообразовательных деревьях различной сложности, результаты которых показывают, что модифицированный алгоритм превосходит базовый по времени вычислений на простых деревьях в 1,2 раза, на сложных – в 2 раза, а разработанный эвристический алгоритм – в 1,6 и 2,3 раза соответственно.

Одним из важных критериев, характеризующих качество результатов словообразования, является точность, доля правильных результатов в общем числе полученных, которая была вычислена с помощью выражения (5):

$$\Pr = \frac{R_p}{A}, \qquad (5)$$

где R_p – количество производных слов, образованных системой, которые совпадают со словами, образованными экспертом;

A – множество всех результатов словообразовательного синтеза.

Было проведено 100 серий экспериментов программной реализации обобщенного алгоритма словообразовательного синтеза. В каждой серии синтезировались слова различных частей речи и с разным количеством морфем, составляющих ключевое слово. Количество морфем в наиболее распространенных словах варьируется от одной до пяти, причем наименьшее число морфем, составляющих имена прилагательные, составляет два форманта. На рис. 1 представлена зависимость точности словообразовательного синтеза (*Pr*) от количества морфем, составляющих слово (*Km*). Анализируя полученные результаты, можно отметить, что точность словообразовательного синтеза во всех случаях составляет не менее 93%, что превышает правильность использования словообразовательных моделей русскоговорящим человеком на 10 – 15%.

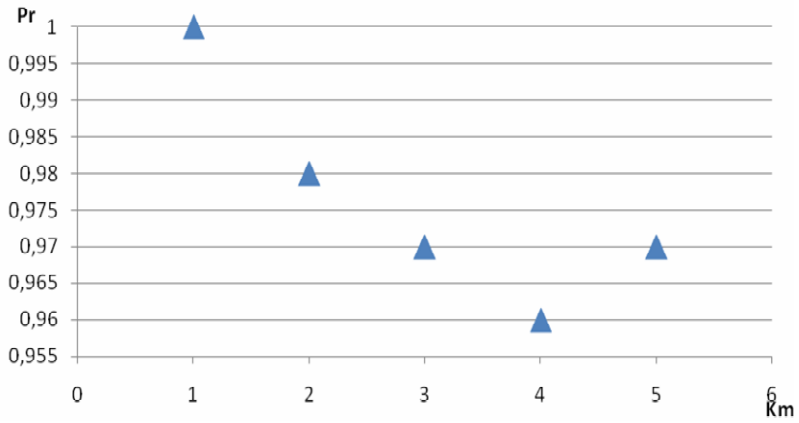

Рисунок 1 – Результаты словообразовательного синтеза имен существительных

Была определена семантическая близость (δ) между заданной пользователем семантикой производного слова (в формальном представлении) и семантикой слова, полученного в результате словообразовательного синтеза. На рис. 2 приведена зависимость семантической близости от количества морфем в слове.

На основании полученных результатов можно сделать следующие выводы: мера близости между заданной и полученной словообразовательной семантикой производного слова варьируется от 93,5% до 100%; наиболее часто употребляемые слова русского языка (слова, состоящие из двух или трех морфем) образуются в соответствии с заданной семантикой (мера семантической близости не опускается ниже 0,96).

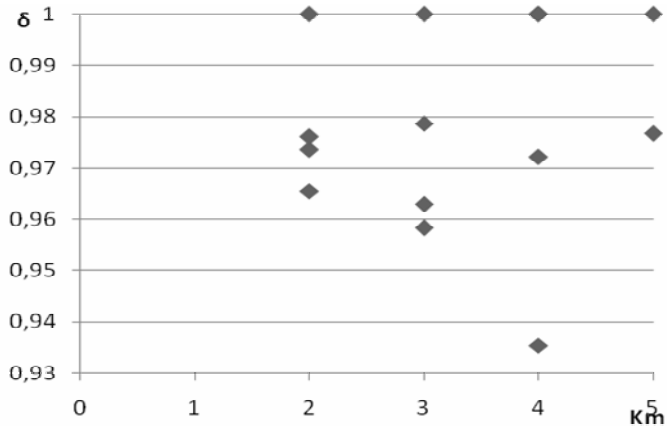

Рисунок 2 – Семантическая близость результатов словообразования к заданной пользователем семантике производного слова

Выводы

В работе предложен эвристический метод словообразования по заданной семантике, согласно которому вершины-кандидаты словообразовательного дерева упорядочиваются по убыванию оценочной функции, в качестве которой предложено использовать меру семантической близости вершины-кандидата к заданной семантике. Применение предложенной меры семантической близости позволяет снизить временные затраты на построение производного слова за счет уменьшения количества анализируемых вершин.

Литература

1. Вороной С.М. Формализация словообразовательного синтеза на основе семантических свойств формантов / С.М. Вороной, А.А. Егошина // VIII Международная конференция «Интеллектуальный анализ информации ИАИ-2008», (Киев, 14 – 17 мая 2008 г.) : сб. тр. / [Рос.ассоц.искусств.интеллекта и др. ; под ред. Т.А. Таран]. – К. : Просвіта, 2008. – 334 с.
2. Вороной С.М. Формализация семантических единиц при словообразовательном синтезе / С.М. Вороной, А.А. Егошина // IX Международная конференция «Интеллектуальный анализ информации ИАИ-2009», (Киев, 14 – 17 мая 2009 г.) : сб. тр. / [Рос.ассоц.искусств.интеллекта и др. ; под ред. Т.А. Таран]. – К. : Просвіта, 2009. – 334 с.
3. Вороной С.М. Словообразовательная база знаний экспертной обучающей системы / С.М. Вороной, А.А. Егошина // Искуственный интеллект. – 2009. – № 1. – С. 31-37.
4. Вороной С.М. Определение грамматических характеристик словоформы методом графов / С.М. Вороной, А.А. Егошина // Искусственный интеллект. – 2008. – № 1. – С. 80-85.
5. Вороной С.М. Метод поиска базового узла дерева словообразовательного синтеза для экспертной обучающей системы / С.М. Вороной, А.А. Егошина // X Международная конференция «Интеллектуальный анализ информации ИАИ-2010», (Киев, 17 – 21 мая 2010 г.) : сб. тр. / [Рос.ассоц. искусств. интеллекта и др. ; под ред. Т.А. Таран]. – К. : Просвіта, 2010. – 326 с.

S.M. Voronoy, A.A. Yegoshina

Search Method of a Tree Root of a Word-creating Tree for an Expert Learning System

The paper is devoted to the problem of processing large volume of derivatives of words in derivative synthesis in natural language systems which leads to complications and slow down the search for a word. It is proposed an heuristic method of derivation on a given semantics that provide to reduce time spent on the construction of the original words by reducing the number of vertices derivational trees that are analyzed.

Статья поступила в редакцию 09.07.2010.