УДК 81.322

Проблемы создания универсального морфосемантического словаря

С.В. Елкин[1], Э.С. Клышинский[2], С.Е. Стеклянников[3]

В работе описываются основы создания универсального морфологического словаря, обладающего возможностью хранить семантическую информацию о словах.

Введение

Традиционно в научной и инженерной мысли сложилась следующая последовательность этапов проведения такого анализа: графематический, морфологический и синтаксический. Неудовлетворительность обработки текстов статистическими методами вызвало к жизни большое количество работ направленых на разработку семантических методов поиска и анализа. Подобные системы работают как с глубинным, так и с поверхностным смыслом текста. В них дополнительно вводится этап семантического анализа. Каждый из данных этапов требует хранения собственной информации, в результате чего обычно создаются раздельные программные модули для хранения, пополнения и извлечения данных из словарей соответствующей направленности.

В наиболее простых случаях при создании систем для работы в узких предметных областях используют тезаурусный подход. Для этого частотным анализом из массива текстов выделяют список терминов и составляют словарь содержащий родовидовые отношения между указанными терминами. При необходимости работать с текстами политематическими или общелитературными составляют семантические сети, представляющие из себя неориентированные графы с огромным количеством вершин. Обработка текстовой информации с использованием такой сети-графа представляет существенную вычислительную проблему, а качество семантического анализа оставляет желать лучшего, так как количество рёбер идущих от одной вершины в таком графе весьма

[1] *115409, Москва, г. Москва Каширское шоссе д.31, МИФИ, elkin_serg@mail.ru*
[2] *109028, Москва, пер.Б. Вузовский, 3/12, МИЭМ, klyshinsky@mail.ru*
[3] *109028, Москва, пер.Б. Вузовский, 3/12, МИЭМ, steklakkinen@narod.ru*

ограничено и обычно не превышает 5-7. Попытки автоматической генерации семантических сетей и словарей не увенчались успехом, так как для определения соответствия некоторого слова к отвечающей ему семантической структуре необходима интеллектуальная система с развитой семантикой, которую необходимо составить «вручную». Как показала практика, владея базовой сематикой невозможно автоматически создать семантику более высокого уровня не выполняя контролирущих работ по устранению ошибок такой автоматической генерациии.

Семантическая информация может закрепляться за отдельными словами и словосочетаниями. В первом случае одно понятие может соответствовать нескольким словам, и одному слову может соответствовать несколько понятий. Во втором случае семантическая информация привязывается к словосочетанию (возможно разрывному).

В первом случае очевидно, что существует однозначное соответствие между словом и набором связанных с ним понятий. То есть определив какому слову из морфологического словаря соответствует данная строка, можно определить набор понятий, связанный с данным словом. При программной реализации последовательное определение нормальной формы слова по мофологическому словарю и извлечения из тезауруса понятий не является оптимальным как по производительности, так и по объему хранимой информации. Повысить производительность систем обработки текстов можно с помощью морфосемантического словаря, хранящего как морфологическую, так и семантическую информацию. В этом случае, найдя слово в морфологическом словаре, помимо информации о его парадигме мы получаем и семантическую информацию из тезауруса. При этом время на поиск в тезаурусе не тратится, а сами нормальные формы хранятся только один раз.

Введение в морфологический словарь семантики позволяет выделять семантическую информацию на более ранних этапах – синтаксического анализа и выявления глубинных падажей, что позволяет поысить качество и скорость перевода. Например, в фразе «Мы едем на поезде на юг на конференцию на две недели» без использования семантической информации нельзя определить роли обстоятельств, тогда как ее использование позволит разделить их и не порождать неверные варианты.

1. Программная реализация

Основной целью при создании новой версии словаря было повышение скорости поиска при сохранении умеренного объема занимаемой памяти. При этом словарь должен решать задачи синтеза и анализа слов. В первой на вход словаря поступает нормальная форма слова и парадигма, в соответствии с которой синтезируется требуемая форма этого слова. Во

второй на вход словаря поступает форма слова, на выходе же нужно получить его парадигму и нормальную форму.

Для повышения скорости поиска слов применяют следующие методы: поиск по бинарному дереву, цифровой поиск, хеширование. На первый взгляд, хеширование выглядит наиболее привлекательно, особенно с точки зрения быстродействия. В результате дальнейшего изучения вопроса и ознакомления с результатами других работ оказалось, что частота возникновения коллизий не так уж и мала, особенно при анализе слов не входящих в словарь. Так, например, поисковая система Яndex находит все формы слова “печати” в ответ на запрос на поиск слова “петила” (являющегося дворовым прозвищем реально существующей личности). Эта проблема может быть решена либо за счет увеличения хеш-таблицы (увеличение объема занимаемой памяти), либо за счет проверки проверяемого и найденного слов (увеличение времени поиска).

Традиционно словарь был разделен на две части – на словарь основ, в который вошли приставки и корни слов, и словарь постфиксов, состоящий из суффиксов и окончаний слов. При этом в словаре основы хранятся слева направо, а постфиксы – в инвертированном виде. Такая организация словарей позволяет за один проход по слову найти все встречающиеся в нем основы и постфиксы.

За основу при создании словаря был взят метод цифрового поиска, так как он обеспечивает наибольшую скорость. В этом методе ключ представляется в виде последовательности символов, принадлежащих рассматриваемому алфавиту, а структура хранения данных – это М-арное дерево, где М равно количеству символов в алфавите. Каждый узел уровня L является набором всех ключей, начинающихся с определенной последовательности L символов. Узел определяет разветвление на М путей в зависимости от (L+1)-го символа.

Основным достоинством этого метода является его высокая скорость – $\log_M N$, где N – количество ключей – основ или постфиксов (в зависимости от рассматриваемого словаря), М – размер вектора указателей. Его основным недостатком является огромный объем занимаемой памяти. В целях экономии памяти был сокращен размер алфавита а также была создана таблица перекодировки символов. В ней символы алфавита располагаются непрерывным блоком.

Дальнейшее сокращение объема памяти достигается за счет перехода от М-арного дерева к бинарному в случае, когда объем занимаемой памяти М-арного узла больше объема памяти занимаемого узлами бинарного дерева.

При построении бинарного дерева целесообразно учитывать частоты появления символов в позиции L+1 для каждого множества ключей

начинающихся с рассматриваемого префикса. Пусть p_i – вероятность того, что после рассмотренного префикса длиной L следует символ с кодом I, причем

$$\sum_{i=1}^{M} p_i = 1 .$$

Если учитывать вероятность того, что искомая строка S может не соответствовать ни одному из ключей, то ожидаемое количество сравнений при поиске составит

$$\sum_{j=1}^{K} p_j (уровень_узла(j) + 1) .$$

Это значение назовем ценой дерева. Дерево с минимальной ценой называется оптимальным. Однако стоит заметить, что цена дерева зависит и от того, каким образом перенумерованы символы алфавита.

Рассмотрим пример. Пусть код символа А меньше кода символа Б, а код символа Б меньше кода символа В. Пусть вероятности появления этих символов в позиции L+1 следующие: $p_А$=0,4; $p_Б$=0,3; $p_В$=0,3. Цена этого дерева будет: $1*p_А+2*p_Б+3*p_В=1,9$. После оптимизации цена будет равна: $2*p_А+1*p_Б+2*p_В=1,7$.

Это дерево будет оптимальным для данной нумерации символов алфавита. Теперь пусть В>А>Б и вероятности появления этих символов в позиции L+1 и порядок их поступления для построения бинарного дерева остались прежними. Тогда цена дерева равна $2*p_Б+1*p_А+2*p_В=1,6$. Из примера видно, что оптимизировать дерево поиска можно не только изменяя расположение узлов, но и модифицируя таблицу перекодировки.

Анализ построенного таким образом словаря показал, что только 1% от всех узлов дерева имеют структуру М-арного дерева, остальные принадлежат “бинарным” поддеревьям, причем среди последних преобладают деревья состоящие только из корневой вершины (почти 90%). Это свойство позволило резко сократить объем занимаемой памяти за счет объединения нескольких идущих друг за другом одноузельных поддеревьев.

2. Состояние разработки

В соответствии с вышеизложенным был разработан программный модуль морфосемантического словаря. Также были созданы морфосемантические базы русского, английского, испанского и турецкого языков.

Ранее в работах **[Бетин и др., 2001]** и **[Ёлкин и др., 2001]** описывалась разработка машинного семантического словаря для задачи автоматического перевода общелитературных текстов. Предварительно был

разработан иерархический ветвящийся рубрикатор с глубиной от 5 до 9 рубрик и содержащий около 3000 относительно самостоятельных понятий (сем). В дальнейшем каждое слово семантического словаря ставилось в соответствие набору сем от 5 до 30 (в виде графа), в зависимости от значимости слова. Это позволило ввести понятие семантической близости и соответствующую формулу для её расчета. В окончательном варианте словарь содержит 80 000 слов, причем омонимы в нём имеют различное семантическое представление. В словарь не входят предлоги, союзы и частицы, так как они анализируются на более ранних этапах машинной превода. Объем словаря, представление в нём каждого понятия через набор исходных сем и возможность математически определять близость между различными понятиями позволили говорить о нем как о семантическом пространстве. Словарь является открытой системой и его можно пополнять как в плане увеличения словарных входов, так и в плане увеличения и уточнения рубрик (сем). Необходимо так же заметить, что в словарь входят только нормальные формы слов, так как морфологический анализ предваряет семантический. В дальнейшем на основе семантического словаря был разработан семантический реферратор текстов, позволяющий извлекать из массива информацию связанную по смыслу.

Разработанный морфосемантический словарь в ходе тестирования показал скорость морфологического анализа на уровне 200000 слов в секунду.

Список литературы

[Бетин и др., 2001] Бетин В.Н. Ёлкин С.В. Хачукаев Э.М. Принципы построения семантического словаря для решения задачи устранения омонимии // Вестник ВИНИТИ НТИ. 2001. сер 2, N1. С.34

[Ёлкин и др., 2001] Ёлкин С.В., Бетин В.Н., Жигарев А.Е, Простаков О.В, Хачукаев Э.М. Разработка семантического анализа текстов при автореферировании // Вестник ВИНИТИ НТИ. 2001. сер 2, N12. С. 18-21