УДК 004.89:004.4

А.А. Егошина
Донецкий государственный институт искусственного интеллекта, Украина

Об одном способе построения статического словаря морфологического процессора

Предложены состав и структура статического словаря морфологического процессора полнотекстовых информационно-поисковых систем с естественно-языковым интерфейсом, для ускорения поиска часто используемые основы и соответствующие аффиксы помещаются в кэш-словарь.

В 60 – 70 гг. центральным звеном экспериментальных исследований в области машинного морфологического анализа являлось создание машинного словаря. Не существовало единого общепринятого формата и структуры такого словаря. В результате этого обстоятельства каждый алгоритм разрабатывался под определенный формат словаря и являлся словарнозависимым. В современных интеллектуальных информационно-поисковых системах морфологический процессор стал неотъемлемой частью.

Основная проблема при разработке лексического и алгоритмического обеспечения компонентов морфологического процессора состоит в хранении словаря большого объема и, соответственно, большого времени поиска в нем. Исследования в этой области направлены на минимизацию исходных данных. Работы, посвященные морфологии, можно условно разделить на две категории:

1) теоретические, в некоторых представлены описания морфологических законов и формальные модели русской морфологии;

2) прикладные, описание программно-реализованных систем с морфологическим модулем.

Теоретические работы посвящены построению многоуровневых формальных моделей морфологии и в большинстве своем предназначены для синтеза. Подобные модели морфологического синтеза подразумевают наличие больших словарей со сложной структурой. Они описывают широкий круг морфологических явлений (фонетическая реализация слова, акцентная парадигма, большое число словообразовательных аффиксов). Недостатком таких моделей является их сложность: несколько уровней представления морфологической информации, специальные грамматики для перехода с одного уровня на другой, избыточность грамматических признаков, часть из которых выделена в модели для описания частных случаев [1].

Модели, которые используют словарь, способны дать более полный анализ словоформы (т.е. оперировать большим числом грамматических признаков). Степень точности такого анализа выше по сравнению с моделями, которые не используют словаря. На пространстве реальных текстов системы, использующие словарь, могут часто давать сбои. Это обусловлено тем, что не существует полных словарей. Лексика языка непрерывно пополняется – появляются новые слова. Для каждой предметной области существует своя терминология, свое подмножество лексики языка, и включить в общий словарь всю существующую терминологию – невозможно.

Существует два базовых подхода к проектированию морфологических машинных словарей (лексиконов) для флективных языков. Первый копирует академическую лингвистическую модель описания, где выделяются основные парадигматические классы, соответствующие типу склонения и спряжения, и правила регулярных альтернаций (фонетических чередований), а нерегулярные формы (например, сильные глаголы в немецком и английском языках) задаются перечислением. Такого типа лексиконы для русского языка составляются на базе модели грамматического словаря А.Зализняка, разрабатывая 8 классов именного склонения и 16 глагольного спряжения, а чередования в основе и глагольной темы выносятся в отдельное множество пост-морфологических правил альтернаций. Второй подход рассматривает любого вида регулярное и нерегулярное чередование как часть расширенной псевдофлексии (в таком случае, основа словоформы ‘день’ – ‘д’, а флексия – ‘-ень’; для словоформы ‘песок’: ‘пес’ и ‘-ок’). В подобной модели описания число парадигматических классов для русского языка возрастает до 3 000, но рост числа классов при проектировании компенсируется однородностью лексикона и отсутствием как исключений, так и правил альтернаций.

Внутреннее устройство лексиконов первого и второго типов не влияет ни на процесс лемматизации – приведения словоформы к нормальной форме слова, репрезентирующей лексему, ни на морфологический анализ. Анализаторы, построенные на разных типах лексиконов, могут одинаково эффективно использоваться как для морфологического анализа, так и для синтеза [1].

Постановка задачи. Целью работы является разработка состава и структуры статического словаря морфологического процессора для полнотекстовых ИПС с естественно-языковым интерфейсом, а также определение возможности использования полученной модели для более полного и точного анализа словоформ. Целью и результатом морфологического анализа является определение морфологических характеристик слова и его основной словоформы. Перечень всех морфологических характеристик слов и допустимых значений каждой из них зависят от естественного языка. Тем не менее, ряд характеристик (например, название части речи) присутствует во многих языках. Результаты морфологического анализа слова неоднозначны, что можно проследить на множестве примеров.

В полнотекстовых ИПС с естественно-языковым интерфейсом обрабатываются запросы двух классов: запросы добавления новых документов и запросы пользователя на поиск документов в существующей базе документов. При обработке запросов каждого класса основными функциями, реализуемыми процессором морфоанализа, являются: получение всех словоформ слова, постановка слова в заданную форму и получение грамматических характеристик словоформы.

Для реализации этих функций морфологический процессор содержит основные модули, показанные на рис. 1.

Модуль разделения текста на составляющие

Принимает исходный текст документа или текст от компонентов пользовательского интерфейса. Анализируемое предложение попадает на вход модуля разделения текста в виде массива символов, содержащего прописные и строчные буквы русского алфавита, цифры, знаки пунктуации. Полученный массив преобразуется в массив лексических единиц. Для каждой лексической единицы формируется отдельная строка, в которую копируются все символы, принадлежащие данной лексической единице. При этом удаляются пробелы, символы переноса,

конца строки и незнакомые символы. В зависимости от результатов обработки полученная цепочка символов направляется в один из трех потоков данных:

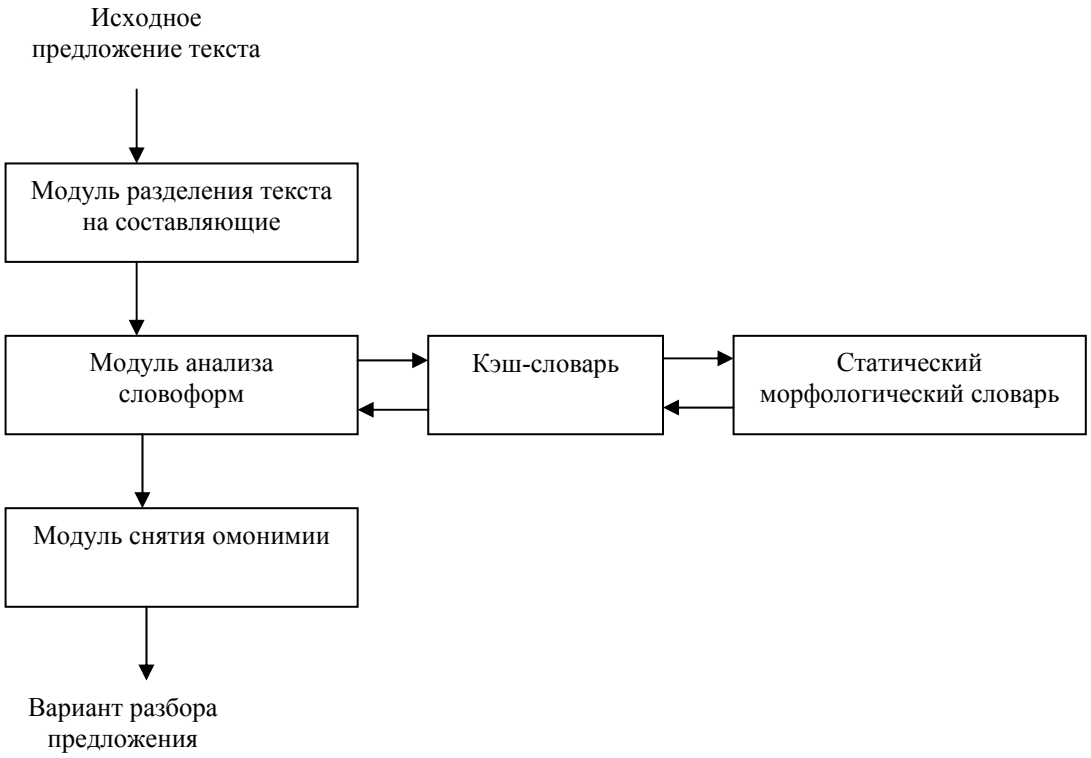

Рисунок 1 – Компоненты морфологического процессора

Основными компонентами процессора являются:

– цифровые или символьные сокращения (‘см’, ’15.06.06’);
– аббревиатуры (‘ДГИИИ’);
– полные словоформы.

Первые два потока данных считаются проиндексированными и не подвергаются дальнейшему морфологическому анализу.

Полные словоформы поступают на вход модуля анализа словоформ, цель которого разбить все множество словоформ на подмножества по признаку принадлежности к той или иной лексеме (множеству словоформ, отличающихся друг от друга только словоизменительными значениями [2]), привести все элементы каждого такого подмножества к уникальной основе, однозначно определить грамматические характеристики лексемы и проиндексировать тексты по встретившимся в них основам.

Модуль анализа словоформ

На вход поступает массив лексических единиц, выделенных из входного текста модулем разделения. Для каждой лексической единицы выполняется поиск аффиксов в статическом морфологическом словаре, состав которого представлен на рис. 2.

При этом вначале в наборе суффиксов подбирается соответствующий исходному слову, запоминается его вид (например, суффикс существительного). Затем из слова выделяется предшествующая суффиксу часть. Это может быть сочетание корня и

приставки (префикса) или только корень. На следующем шаге выполняется аналогичный поиск-проверка на наличие приставки, если она есть – отделяется. На последнем этапе выполняется поиск корня, т.е. оставшаяся после разбиения часть слова ищется в наборе корней.

Рисунок 2 – Состав статического морфологического словаря

Если в результате поиска не найдено ни одного успешного варианта, то проводится поиск среди исключений, также хранящихся в словаре. Разница между исключениями и обычными основами состоит в том, что словообразование исключений происходит нестандартным образом. В связи с этим в словаре хранятся не морфемы исключений, а слова целиком. Таким образом, при поиске среди исключений приходится просматривать весь набор слов. Это занимает много времени, поэтому поиск среди исключений проводится только в том случае, когда не найдено ни одного варианта среди обычных основ.

В случае, когда все этапы поиска дали отрицательный результат (не найдено ни одного варианта), пользователю выдается запрос на ввод новой основы в словарь. В случае его отказа это сделать выполнение морфологического анализа прекращается. Если же новое слово введено в словарь, то вся процедура поиска повторяется [3].

В результате работы модуля анализа словоформ получаем набор морфем, из которых состоит исходное слово, и его принадлежность к определенной части речи.

Статический морфологический и кэш- словари

Прообразом морфологического словаря является словарь А.Н. Тихонова [4]. Он содержит информацию об основах, аффиксах и исключениях. Все выделяющиеся в слове части имеют ту или иную семантику: непроизводная основа выражает основное лексическое значение слова, а остальные морфемы (их называют служебными морфемами или аффиксными) – дополнительное лексическое и грамматическое значение. Все множество морфем русского языка делится по разным основаниям на несколько классов. В классификации учитывается следующее: роль морфемы в слове, значение морфем, их место в слове, их происхождение. Выделяются корневые морфемы и аффиксальные. Основой для такого членения есть место и роль таких морфем в слове. Корневые морфемы – это обязательная часть слова. Без корня не существует слов. Аффиксальные морфемы – это факультативная часть слова. Аффиксы, входя в слово, относят его к какой-нибудь разновидности, к какому-нибудь классу предметов, признаков, процессов. В этом и заключается

принципиальное различие между аффиксальными и корневыми морфемами – обязательная повторяемость аффиксов в аналогично построенных и обладающих общим элементом значения словах и безразличие к этому свойству корней [5]. Семантический принцип организации подобного словаря позволяет использовать при анализе значений слов основной принцип объектно-ориентированного программирования – наследование.

Для ускорения поиска часто используемые основы и соответствующие аффиксы помещаются в кэш-словарь.

Метод морфологического разбора словоформы позволяет значительно уменьшить требования к памяти, так как требует хранения только основных частей словоформ и таблицы флективных частей (префиксов, суффиксов и инфиксов). Требования к оперативной памяти уменьшаются в 50 – 200 раз, но из-за использования более сложных алгоритмов сопоставления увеличиваются требования к производительности процессора.

Литература

1. Еськова Н.А., Бидер И.Г. Формальная модель русской морфологии.
2. Егошина А.А. Языковые и алгоритмические аспекты построения морфологических процессоров для интеллектуального поиска в полнотекстовых базах данных // VI Международная конференция «Интеллектуальный анализ информации ИАИ-2006». Киев, 16-19 мая 2006 г.: Сб. тр. / Рос. ассоц. искусств .интеллекта и др.; Под ред. Т.А. Таран. – К.:Просвіта. – 2006. – 334 с.
3. Смирнов Ю.М., Андреев А.М., Березкин Д.В., Брик А.В. Об одном способе построения синтаксического анализатора текстов на естественном языке // Изв. вузов. Приборостроение. – 1997. – Т. 40, № 5. – С. 34-42.
4. Тихонов А.Н. Словообразовательный словарь русского языка. – Русский язык, 1985.
5. Лефевр В.А., Земская Е.А. Современный русский язык и словообразование. – М.: Просвещение, 1973.
6. Мельчук И. Курс общей морфологии. – М., 1997. – Т 1.

Г.А. Єгошина

Про один засіб побудови статичного словника морфологічного процесора

Подана розробка складу та структури статичного словника морфологічного процесора повнотекстових інформаційно-пошукових систем з природно-мовним інтерфейсом, для прискорення пошуку часто основи, що використовуються, та належні афікси розміщуються у кеш-словнику.

A. Yegoshina

Mode of Construction Morphological Process Static-linc Dictionary

The propose composition and structure syntax dictionary morphological full-text information storage and retrieval system this natural language interface for acceleration search ofen used foundations and ansver the affix consist in cache memory.

Статья поступила в редакцию 11.07.2006.