УДК 004.89: 004.4

Языковые и алгоритмические аспекты построения морфологических процессоров для интеллектуального поиска в полнотекстовых базах данных

Егошина А.А., Донецкий государственный институт искусственного интеллекта, г. Донецк
postmaster@iai.donetsk.ua

В работе предложен вариант организации морфологического процессора, основанного на процедурных методах анализа и основном статическом словаре. Для сокращения времени поиска предлагается использовать кэш словарь. Выбран словообразовательный словарь для автоматизированного получения точной морфологической информации. Определен состав и связи основных компонентов морфологического процессора, предложена организация полнотекстовой поисковой системы на его основе. Разрабатываемая информационно-поисковая система будет использоваться как инструментальное средство обновления баз знаний интеллектуальных обучающих систем.

Введение

В настоящее время в глобальной сети Интернет накоплены и стали доступны огромные массивы информации. Ведется усиленная работа по оцифровке и размещению на серверах глобальной сети текстовой, графической и прочей информации накопленной человечеством. Это значит, что любая информация в принципе будет доступна. Основной проблемой становится поиск требуемой информации. Существующие технологии поиска информации в Интернете недостаточно эффективны при поиске в массивах информации такого объема. В результате поиска по запросу поисковая система обычно выдает огромное количество ссылок, большинство которых не отвечают запросу, и являются информационным мусором. И чем дальше развивается сеть Интернет, чем больше накапливается в ней информации, тем труднее найти то, что нужно. Можно говорить о кризисе поиска информации в Интернет.

Реальную возможность высокоэффективного поиска информации дает двухэтапный процесс поиска, где первый этап, это предварительный поиск и отбор информации в тематические базы

данных, а второй этап, это поиск нужной информации конечным пользователем в сетевых или локальных полнотекстовых базах [1]. Технические возможности современных компьютеров позволяют создавать полнотекстовые базы, включающие миллионы страниц тематически отобранных материалов: из Интернет, из баз данных на серверах локальных сетей предприятий, с тиражируемых носителей, других источников. Эффективность поиска в полнотекстовых базах данных, в предварительно отобранном по каким то признакам материале может быть значительно выше, чем поиск в массиве разнородной информации.

Реализация семантических подходов к автоматизации задач поиска в массовой персональной автоматизации, т. е. на рабочих местах отдельных пользователей или небольших рабочих групп затруднена из-за необходимости серьезной предварительной проработки соответствующей предметной области.

Средством, освобождающим пользователя от необходимости сложной предварительной структуризации предметной области и затратных процедур индексирования текстовых данных, являются полнотекстовые информационно-поисковые системы (ИПС) и полнотекстовые СУБД, появившиеся на рынке программных продуктов в конце 80-х годов [2,3]. Недостатком существующих полнотекстовых ИПС является их низкая гибкость из-за применения пользовательских интерфейсов на основе информационно-поисковых языков дескрипторного типа.

Постановка задачи. Целью работы является создание лексического и алгоритмического обеспечения компонентов морфологического разбора текстов документов и пользовательских запросов в ИПС с интеллектуальным интерфейсом. Интеллектуальные интерфейсы интенсивно разрабатываются в настоящее время [4] и по своему названию предполагают использование методов и технологий человеко-машинного взаимодействия и искусственного интеллекта. Интеллектуальный интерфейс дает возможность вести с пользователем диалог на естественном языке. “Естественность” языка общения состоит не в том, чтобы он позволял использовать весь словарь и весь арсенал синтаксиса естественного языка, а в том, чтобы он позволял вести взаимодействие с системой без какой бы то ни было предварительной подготовки пользователя. Язык общения должен обеспечивать пользователя простыми средствами однозначной формулировки задачи поиска необходимых документов.

В работе предложен вариант организации морфологического процессора ИПС, основанного на процедурных методах анализа и

статическом словаре. Для сокращения времени поиска предлагается использовать кэш словарь. Выбран словообразовательный словарь для автоматизированного получения точной морфологической информации. Определен состав и структура морфологического процессора, предложена организация полнотекстовой ИПС на его основе. Для улучшения качества работы морфологического процесора необходим учет контекста не только для отвержения неадекватных вариантов разбора слова, но и для порождения достоверных вариантов в тех случаях, когда информации о написании слова и алгоритмов морфоанализа не достаточно для определения лексико-семантического разряда.

Лингвистические основы построения морфологического процессора

Важной особенностью, оказывающей существенное влияние на эффективность полнотекстовых ИПС, является наличие либо отсутствие лингвистического процессора (ЛП).

Идеальная модель лингвистического процессора состоит из четырех основных процессоров-анализаторов: графематического (внешнее представление текста), морфологического, синтаксического и семантического. Реализация полнофункционального ЛП в ИПС системах является сложной задачей и возможна в ИПС для узких предметных областей. Рассмотрим лингвистические особенности реализации процессора – анализатора для морфологического разбора при обработке естественно языковых запросов пользователей и индексировании документов.

В лингвистических процессорах используется в основном декларативные и процедурные методы морфологического анализа. Методы декларативной ориентации используют полный словарь всех возможных словоформ для каждого слова. Из-за необходимости хранения словаря большого объема и, соответственно, большого времени поиска в нем, декларативные методы рассматривать не будем.

Процедурные методы основаны на особенностях словообразования и словоизменения в различных языках. Возможность анализа неизвестных слов – необходимое качество морфологического процессора. Выступая в языке в качестве его основной значимой единицы, слово предстает всегда как определенное структурное целое, так или иначе соотносительное по своему строению с другими словами. Подавляющее большинство полнозначных слов имеют богатую и разветвленную систему форм словоизменения и словообразования. Структурный характер какого-либо конкретного слова определяется не

только его морфемным составом, т. е. наличием в них тех или иных морфем, но и их словообразовательной спецификой, значением и соотношениями друг с другом.

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 значения. Значение, присущее морфемам, может быть свободным и связанным. Свободным значением называют такое, которое четко выявляется в морфеме. Оно имеет себе аналогию в прямом, номинативном значении слова. Так, например, «лес» в слове «перелесок», «гор» в слове «горный». Связанное значение является ясным лишь тогда, когда морфема рассматривается как определенная часть более сложного целого. Оно имеет себе аналогию во фразеологически связанном значении слова. Например, «ул» (ср. улица, переулок).

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. Основное различие между корневыми и аффиксальными морфемами состоит в степени абстрактности значения и частоте употребления. Так, корни могут быть единичные, то есть, могут встречаться только в одном слове, единичных же аффиксов не бывает. 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]

Все аффиксы делятся на словообразовательные и словоизменительные. Словообразовательные аффиксы русского языка могут располагаться перед корнем (префиксы) и после корня (суффиксы). Префиксов и суффиксов существует ограниченное количество. Различие между ними сводится не только к их местоположению в составе слова. Эти внешние структурные особенности лежат в основе целого ряда специфических черт суффиксов и префиксов. Присоединение приставки не изменяет

принадлежности слова к определенной части речи, а присоединение суффикса может оставлять в пределах той же самой части речи (лес - лесник), а может переводить производное слово в иную часть речи (белый - белок - белеть). В русском языке нет суффиксов, которые бы производили слова разных частей речи. Например, суффикс – ок-, производящий только существительные (бел**ок**, желт**ок**), -ыва – только глаголов (разбрас**ыва**ть, перепис**ыва**ть).

Для приставок не обязательна тесная связь со словами определенной части речи. В русском языке есть приставки, которые присоединяясь к словам различных частей речи, сохраняют свое значение. Например, приставки раз-: **раз**веселый, **рас**красавица, со-: **со**автор, **со**существовать, сверх-: **сверх**скорость, **сверх**прочный, анти-: **анти**вещество, **анти**революционный.

Присоединение к слову приставки обычно не меняет значения слова коренным образом, а лишь добавляет к нему некоторый оттенок значения. Например, глаголы с приставкой: убежать, выбежать, перебежать, добежать, пробежать, сбежать, обозначают то же действие, что и глагол бежать. Приставка лишь добавляет к их значению указание на направление движения. Поэтому приставки в русском языке соединяются преимущественно с глаголами, прилагательными и наречиями. Для существительных префиксальный способ словообразования не столь характерен. Число таких образований очень невелико. В существительные, так же как в прилагательные, наречия и глаголы присоединительные приставки вносят дополнительное значение, как-то: указание на меру, степень, временный характер.

Значение суффиксов иное: от широких и отличительных до очень конкретных. Так, например, суффиксы -ск- обозначает отношение к тому, что названо производящей основой (одес**ск**ий, мор**ск**ой, шофер**ск**ий), то есть очень отвлеченное значение. Суффиксы, образующие существительные, более конкретны. Так, суффикс -онок- обозначает детенышей животных: тигре**нок**, слоне**нок**, галчо**нок**. Суффиксы имен существительных в русском языке самые многочисленные и разнообразные. Они классифицируют предметы действительности: названия людей по профессии, по признаку, по действию, по месту жительства, названия.

Для автоматизированного получения точной морфологической информации целесообразно использовать словообразовательный словарь А.Н. Тихонова [6]. Семантический принцип организации данного словаря позволяет использовать при анализе значений слов основной принцип объектно-ориентированного программирования - наследование. Словообразовательные связи в подавляющем

большинстве случаев можно трактовать и использовать как связи множественного наследования признаков. Словарь ИПС должен содержать только основы слов вместе со ссылками на соответствующие строки в таблице возможных аффиксов. Основной критерий при разбиении слова на основу и аффикс — основа должна оставаться неизменной во всех возможных словоформах данного слова.

Структура ИПС с естественно-языковым интерфейсом

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.

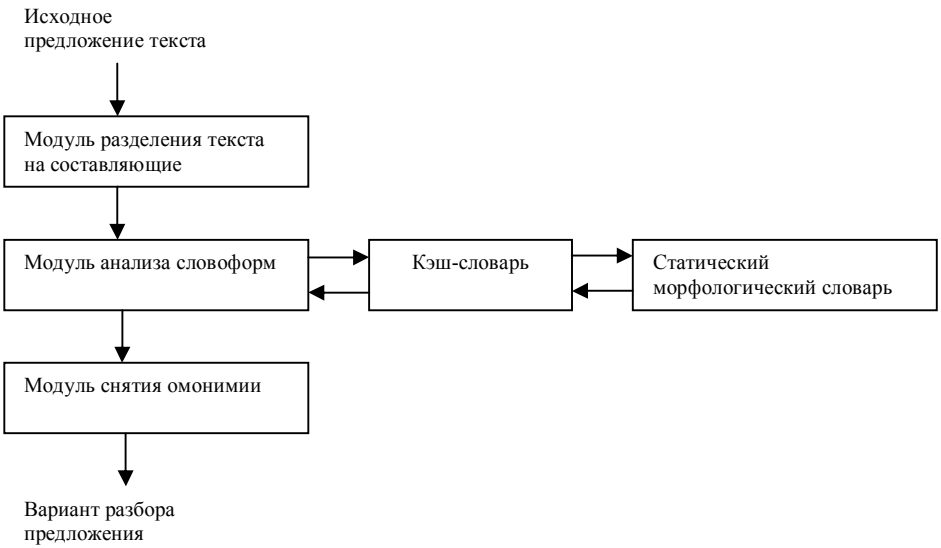

Рис.1 – Компоненты морфологического процессора

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.

На вход модуля анализа словоформ поступает массив лексических единиц, выделенных из входного текста модулем разделения. Для каждой лексической единицы выполняется поиска в словаре основ. При этом ищутся все основы, с которых может начинаться анализируемое слово. Если очередная основа удовлетворяет этому условию, то из словаря аффиксов извлекается строка, содержащая все возможные аффиксы для данной основы. Каждый аффикс из этой строки поочередно присоединяется к основе, и результат сравнивается с анализируемым словом. В случае их точного совпадения формируется очередная запись в список результатов поиска: по порядковому номеру аффикса в строке аффиксов определяются переменные морфологические параметры слова (например, для существительного - число и падеж), а по словарной информации данной основы - его постоянные параметры (для существительного — род и одушевленность). Для быстроты поиска при анализе все основы хранятся в виде дерева. Если в результате поиска не найдено ни одного успешного варианта, то проводится поиск среди исключений. Исключения присутствуют в словаре основ наряду с обычными основами. И те, и другие имеют в словаре информацию о постоянных морфологических признаках и о номере строки допустимых аффиксов.

Прообразом морфологического словаря является словарь А.Н. Тихонова. Он содержит информацию об основах, аффиксах и исключениях. Для ускорения поиска, часто используемые основы и соответствующие аффиксы помещаются в кэш словарь.

Структура полнотекстовой ИПС на базе морфологического процессора приведена на рис. 2.

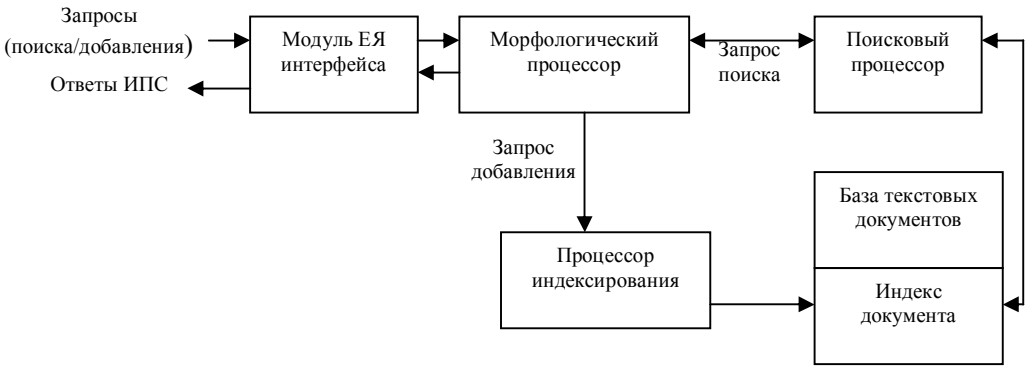

Рис.2 – Структура полнотекстовой ИПС на базе морфологического процессора

В состав ИПС входят:

- модуль естественно-языкового интерфейса;
- морфологический процессор;
- поисковый процессор;
- процессор индексирования;
- база текстовых документов с индексами.

База текстовых документов организуется как информационная структура с массивом индексов, обеспечивающим быстрый поиск документов. Для компактного хранения документы могут быть сжаты архиваторами. Процессор индексирования обрабатывает запросы добавления новых документов в базу путем построения индекса документа. Индекс документа строится на основе морфологического разбора текста документа. Поисковый процессор оперирует с индексами документов используя результаты морфологического разбора естественно-языкового запроса пользователя.

Заключение

На основе предложенного варианта организации морфологического процессора и полнотекстовой поисковой системы в настоящее время разрабатываются инструментальные средства для пополнения и обновления баз знаний интеллектуальных систем дистанционного обучения в Донецком государственном институте искусственного интеллекта.

Литература

1. *Карпова Г.Д. Пирогова Ю.К. Кобзарева Т.Ю. Микаэлян Е.В.*. Компьютерный синтаксический анализ: описание моделей и направлений разработок.// Итоги науки и техники (серия "Вычислительные науки").Т.6 – М.: ВИНИТИ. – 1991
2. *Попов Э.В*. Общение с ЭВМ на естественном языке. -М.: Наука, 1982.
3. *Кузин Е.С. Ройтман А.И. Фоминых И.Б. Хахалин Г.К.* Интеллектуализация ЭВМ. - М.: Высш. шк., 1989.
4. *Поспелов Д.А*. Системы общения и экспертные системы /Искусственный интеллект -М.: Наука, 1990.
5. . *Лефевр В.А Земская Е.А.*. Современный русский язык и словообразование. – М.: Просвещение. – 1973. – с.24
6. *Тихонов А.Н.* Словообразовательный словарь русского языка, Русский язык, 1985