5-я Юбилейная Международная Конференция "Крым 98"

Конференция проводится в рамках мероприятий ИФЛА 1998 г.

Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества

Материалы конференции

Том 2

Судак, Автономная Республика Крым, Украина
6-14 июня 1998

Морфологическая сегментация (частичная) словоформ русского языка: опыт эксплуатации, новые решения

Morphologic Segmentation of the Russian Language Wordforms: Experience and New Solutions

Морфологічна сегментація (часткова) словоформ російської мови: досвід експлуатації, нові рішення

Мазов Н.А.
Объединенный институт геологии, геофизики и минералогии Сибирского отделения академии наук, Новосибирск, Россия

Mazov N.A.
Joint Institute of Geology, Geophysics and Mineralogy of the Russian Academy of Sciences Siberian Branch, Novosibirsk, Russia

Мазов М.А.
Об"єднаний інститут геології, геофізики та мінералогії Сибірського відділення академії наук, Новосибірськ, Росія

Рассказывается о разработке аппарата автоматической морфологической сегментации слов естественного языка, обеспечивающего единообразие "запрос-база данных", высокую производительность и достоверность обработки словоформ, а также высокое качество функционирования информационной системы.

The paper covers the development of automated morphologic segmentation of natural language words ensuring the "inquiry-database" consistency, high efficiency and reliability of wordforms processing as well as the high quality of the information system operation.

В доповіді розглядаються проблеми розробки апарату автоматичної морфологічної сегментації слів природної мови, який міг би мати об"єктивний характер та забезпечував би одноманітність "запит-база даних", високу продуктивність та достовірність обробки словоформ, і, як наслідок, високу якість функціонування інформаційної системи.

Формулирование поискового запроса представляет собой сложный процесс "трансформирования" информационной потребности и перевода ее содержания с естественного языка на информационно-поисковый язык системы. В особенности это касается правильного отсечения флексий для словоформ естественного языка. Выходом из этой ситуации является разработка аппарата автоматической морфологической сегментации слов естественного языка, который носил бы объективный характер и обеспечивал единообразие "запрос-база данных", высокую производительность и достоверность обработки словоформ, и как следствие, высокое качество функционирования информационной системы.

Реальные условия функционирования крупных информационных систем (и в особенности в Internet) определяют актуальность решения целого ряда проблем, среди которых важное место занимают проблемы автоматизации процессов формулирования поисковых запросов к базам данных и морфологической сегментации слов естественного языка.

Необходимость решения первой проблемы вытекает из трудоемкости процедуры первоначального формулирования поискового запроса, ввиду специфичности используемых конкретных баз данных и информационно-поисковых языков систем, реализующих доступ к этим базам данных.

Решение второй проблемы частично способствует успешному решению первой, однако по значимости, она имеет важное самостоятельное значение.

В частности, при наличии достаточно простого и надежного аппарата морфологической сегментации слов естественного языка, появляется возможность более гибкого управления процессом поиска (на полноту или точность) в автоматическом режиме, а также не исключается возможность существенного сокращения словарей баз данных, используемых в информационных системах, ввиду значительной флективности русского языка.

В литературе описано достаточно большое число подходов к решению задачи автоматизации морфологической сегментации слов естественного языка. Так, большинство работ посвящены различным методам автоматического индексирования (в состав которых так или иначе входят методы морфологического анализа текста), основу которых составляют базовые наперед заданные словари основ, суффиксов, окончаний и грамматические правила морфологии. Эксплуатация систем, реализующих

такие подходы, требует наличия в информационных службах высококвалифицированных специалистов, профессионально владеющих русским языком для ведения эталонных словарей и значительных накладных расходов при эксплуатации. Однако, по ряду причин, иметь таких специалистов для информационных служб не всегда представляется возможным.

В работах [1, 2] подробно рассмотрен процесс формального выделения основ слов естественного языка. Такой подход требует в качестве опорной базы для построения алгоритма морфологической сегментации словоформ лишь массивов диагностических словоформ и конечных буквосочетаний словоформ.

Реализация изложенного алгоритма с уточненными массивами была успешно реализована в ГПНТБ России, о чем докладывалось на одной из предыдущих конференций [4]. Алгоритм показал свою практичность при эксплуатации различных баз данных, достаточно высокую степень достоверности отсечения окончаний словоформ.

По ряду причин (необходимость переноса алгоритма на другую платформу, сравнительно низкая скорость обработки словоформы, ввиду использования служебной базы данных) авторами была предпринята попытка пересмотра алгоритма и повторной реализации предложенного алгоритма [3] с точки зрения увеличения скорости обработки словоформы, поступающей на вход процедуры морфологической сегментации.

Как показывает опыт эксплуатации систем с большими объемами информации, не всегда бывает удобно иметь дело со словарями баз данных, в которых отсутствуют флексии словоформ. И как показывает практика, экономия на объемах словарей баз данных порой мифическая, особенно при нынешней стоимости одного мегабайта. В связи с этим необходимо было пересмотреть алгоритм морфологической сегментации в сторону увеличения скорости обработки словоформы, что позволило бы отображать словарь базы данных с учетом морфологической сегментации в реальном масштабе времени. С другой стороны такой подход на наш взгляд позволил бы однозначно трактовать поисковый запрос пользователя к базе данных в зависимости от поисковой ситуации, другими словами, за пользователем остается решение выполнения запроса - на полноту или точность.

Настоящая процедура реализована как обычная функция в языке программирования Delphi-2 и не требует никаких дополнительных установок или компонент, легко может быть встроена в динамическую библиотеку.

Аппробация реализованной процедуры проводилась на машиночитаемых информационных изданиях ВИНИТИ, а именно РЖ "Геология", РЖ "Геофизика", РЖ "Механика", РЖ "Химия", РЖ "Охрана окружающей среды" с общим объемом более 70 млн. словоформ (более 700 тыс. - уникальных). В процессе тестирования процедуры был получен ряд статистических характеристик для научно-технических текстов на русском языке.

В докладе будут приведены статистические характеристики по появлению буквенных окончаний в словоформах русского языка научно-технического текста, а также диагностические массивы, полученные в результате апробации алгоритма.

Литература:

1. Зализняк А.А. Русское именное словоизменение. М.: Наука, 1967.

2. Зализняк А.А. Грамматический словарь русского языка. М.: Русский язык, 1987.

3. Зайцева Е.М. Алгоритм отсечения окончаний словоформ, входящих в состав терминов. Научно-технические библиотеки. Сб. статей. М.: ГПНТБ СССР, 198?, С. 173-185

4. Универсальная технология формирования словаря баз данных CDS/ISIS с использованием основ терминов: 3-я Междунар. конф. "Крым 96", Форос, Ялта, Авт. Респ. Крым, Украина, 1-9 июня 1996 г.: Материалы конф.: [В 2 т.] // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества. - М. - С. 169-171.