Процессор автоматизированного морфологического анализа без словаря. Деревья и корреляция

И. Ножов

РГГУ

Igor@Diatech.ru

Сущность интеллекта состоит в способности принимать разумные решения в условиях отсутствия полноты данных и фактов [1]. Интеллектуальность системы повышается с уменьшением объема статической информации, используемой в процессе анализа данных. В нашем случае, речь идет об использовании лингвистической информации при морфологическом анализе в задачах автоиндексации текстовых баз данных (БД). Основная проблема в разработке машинно-ориентированного алгоритма для лингвистических процессоров состоит в объеме словарей, используемых программой, которые приходится составлять вручную. Исследования в этой области направлены на минимизацию исходных данных. На пространстве реальных текстов системы, использующие словарь, часто дают 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. Равно как невозможно и перечислить все существующие имена и фамилии, которые имеют регулярное склонение. Сейчас хотелось бы выделить основные критерии, отличающие большинство интеллектуальных систем, которых придерживается описываемый процессор автоиндексации текстов:

- Способность системы объяснить каждой шаг принятых решений. В процессе анализа не используются вероятностные и статистические методы.
- Использование правил и свойств, характеризующих данный предмет анализа. Для построения морфологических гипотез словоформ используется формальная грамматика и то свойство русского языка, что большая часть грамматических категорий в русском вычисляется из флексии.
- Модульность системы, которая обеспечивает эффективное изменение и пополнение правил и свойств, а также задает возможность настраивать анализатор на другие естественные языки с развитой морфологией.
- Множественность интерпретаций. Анализатор оставляет все омонимы значений словоизменительных категорий.
- Самообучаемость и механизм исправления принятых ранее неверных решений. Объем прочитанных текстов пополняет число словоформ, используемых в процессе анализа, тем самым повышая точность морфологического анализа и позволяя корректировать неправильно построенные основы и значения их грамматических категорий.
- Моделирование интеллектуального поведения человека. В данном случае, речь идет о попытке эмулировать размышления человека, изучающего иностранный язык, перед которым стоит задача классифицировать слова данного языка, в условиях, когда в его распоряжении находится большой массив текстов, некоторые знания о морфологии языка

и отсутствует словарь языка, на котором написаны тексты. Надо сказать, что при разработке алгоритмов не ставилось задачи опровергнуть мысленный эксперимент Джона Сёрля «Китайская комната» и разделить искусственный интеллект на «сильный» и «слабый» [2].

В данной статье будет рассмотрено общее описание процессора, взаимодействие его модулей и функциональная схема алгоритма морфологического анализа.

Схема процесса автоиндексации представлена на рис.1: на вход процесса автоиндексации поступает все множество текстов, хранящихся в базе данных, на выходе формируется словарь основ и таблица соответствий (текст ó основа), которая соответствует потоку индексированных текстов.

Блоки, которые осуществляют процесс автоиндексации, представлены на рис.2.

Процессы (рис.2):

1. Графематический анализ.
2. Морфологический анализ.

На рис.3 показана схема таблиц для хранения потоков данных, сформированных процессами графематического и морфологического анализа.

Потоки данных (рис.3):

1. Тексты;
2. Полные словоформы;
3. Аббревиатуры;
4. Цифровые и символьные комплексы;
5. Основы и значения их грамматических категорий;

Основная цель графематического блока получить выборку полных словоформ из массива текстов БД. Графематический анализ работает с внешним представлением текста (см. ст. «Прикладной морфологический анализ») и использует таблицу стоп-слов. В этой таблице хранятся цифры, спецсимволы и частотные слова языка, нерелевантные для поиска по текстам.

Графематический анализ выполняет три функции:

1. отсечение стоп-слов в тексте;
2. разбиение данных на три потока;
3. индексация каждого потока.

Единицей графематического анализа является цепочка символов, выделенная с двух сторон пробелами. Выделенная цепочка символов подвергается последовательной обработке эвристическими правилами: отсечь знаки пунктуации, проверить присутствие гласных внутри цепочки, чередование верхнего и нижнего регистров и т.д. В зависимости от

результатов обработки полученная цепочка символов направляется в один из трех потоков данных:

— цифровые и символьные комплексы (‘кг’, ‘ст.’, ’12.01.99’);

— аббревиатуры - названия государств, организаций, предприятий (‘СССР’, ‘ЮНЕСКО’, ‘ДорСтройСервис’);

— полные словоформы;

Каждой записи из любого потока ставятся в соответствие коды документов, в которых она встретилась. Первых два потока данных считаются проиндексированными, причем только аббревиатуры являются релевантным поисковым образом. Графематику можно считать лишь вспомогательным звеном для морфологического анализа. Графематический и морфологический процессы способны проиндексировать массивы текстов независимо от предметной области конкретной базы данных.

Полные словоформы поступают на вход морфологического анализа, цель которого разбить все множество словоформ на подмножества по признаку принадлежности к той или иной лексеме[1], привести все элементы каждого такого подмножества к уникальной основе, однозначно определить грамматические характеристики лексемы и проиндексировать тексты по встретившимся в них основам.

рис.1

рис.2

рис.3

Блок морфологического анализа использует минимальный объем исходной информации:

— таблицу предлогов;

— таблицу местоимений и числительных, имеющих нерегулярное склонение.

Блок морфологического анализа использует минимальный объем исходной информации:

— таблицу предлогов;

— таблицу местоимений и числительных, имеющих нерегулярное склонение.

На выходе морфологического анализа формируется словарь основ данной БД, уникальность записи в таком словаре задается тройкой значений [основа, часть речи, парадигматический класс]. Морфологический анализ состоит из трех модулей и соблюдает определенную последовательность действий.

Первый модуль содержит статический массив флексий и правила формальной грамматики русской морфологии, построенной на основе работ А.Зализняка [4]. Данный модуль может быть заменен формальной морфологией любого другого флективного языка. Методы, описанные в модулях два и три, являются универсальными, независящими от языка.

Второй модуль, используя правила формальной грамматики, позволяет строить морфологическое дерево словоформы, в узлах которого хранятся все возможные гипотезы об основах и значениях грамматических категорий словоформы. Традиционно в синтаксических и семантических теориях используется метод представления языковой структуры с помощью деревьев. В описываемой системе, пожалуй, впервые данный метод оправдано был применен к морфологии.

Третий модуль содержит метод подбора словоформ на одну лексему[2], то есть выбор коррелятов для дерева исходной словоформы. После того, как набраны корреляты, для каждой словоформы также строится морфологическое дерево всех возможных гипотез, в результате чего образуется «лес деревьев»[5]. Математический метод корреляции осуществляет сравнение морфологических деревьев внутри леса и унификацию гипотез. Корреляция проводится по гипотезам основ и значениям классифицирующих грамматических категорий, таких как часть речи, парадигматический класс, спряжение глаголов и род существительных. Значения словоизменительных категорий в корреляции не участвуют. Во время работы корреляции происходит удаление ложных гипотез: ветвей дерева или полного дерева коррелята. Этот модуль позволяет построить уникальную гипотезу об основе и значениях ее грамматических категорий для всех словоформ одной лексемы, найденных в текстах. Метод корреляции очищает лес от ложных коррелятов, оставляя, таким образом, только словоформы, принадлежащие одной лексеме. Уникальная основа, единая для всех словоформ, участвовавших в корреляции, значение части речи и парадигматического класса добавляются в словарь основ. По сути, основа в словаре репрезентирует лексему.

Для унификации гипотезы метод корреляции использует матрицы корреляций. Лесом называется множество деревьев словоформ F = {T_1,..,T_j,..,Tn}. Множество всех построенных гипотез об основе в F обозначим U = {s_1,..,s_i,..,s_m}. Параметром корреляции t называется значение грамматической категории. Матрицей корреляции A(t) = леса F с m гипотезами об основах и n деревьями словоформ называется ()-матрица, в которой , если заданный параметр корреляции t определен для s_i в T_j, и в противном случае.

В настоящей статье детальное описание формальной грамматики, построения деревьев гипотез и операций над матрицами корреляций опускается. Последовательность шагов морфологического анализа (Д1..Д13) представлена на рис.4.

Общая схема алгоритма.

рис.4

Д0. Выход из программы.

Д1. Выбрать из таблицы полных словоформ (рис.3) непроиндексированную словоформу, то есть словоформу, для которой еще не построена основа (ДА: словоформа выбрана; НЕТ: все словоформы в таблице проиндексированы).

Д2. Проверить, что данная словоформа не является предлогом или местоимением. Построить дерево всех возможных гипотез для данной словоформы. (ДА: не является; НЕТ: является)

Д3. Выбрать из таблицы полных словоформ (рис.3) словоформы на одну лексему. Создать список коррелятов.

(ДА: корреляты выбраны; НЕТ: список коррелятов пуст)

Д4. Если список коррелятов непустой, то построить деревья всех возможных гипотез для каждого коррелята.

Д5[3]. Провести корреляцию по гипотезам основ.

Д6. Провести корреляцию по значениям части речи.

Д7. Провести корреляцию по значениям спряжения глагола.

Д8. Провести корреляцию по значениям рода существительных.

Д9. Провести корреляцию по значениям парадигматического класса.

Д10. Проверить, что корреляция не привела к удалению полного дерева (дерева коррелята) из леса. (ДА: не привела; НЕТ: привела)

Д11. Удалить ложный коррелят из списка коррелятов.

Д12. Выбрать уникальную основу и ряд грамматических характеристик к данной основе. Проиндексировать тексты, то есть выбрать для построившейся тройки [основа, часть речи, парадигматический класс] коды текстов, в которых встретились словоформы, принадлежащие данной основе.

Д13. Применить метод распределения элементов пересеченных множеств коррелятов.

Литература

990. M. Boden. «Artificial intelligence and images of man» Perspectives From Artificial Intelligence, 1990.

Серл Д. «Мозг, сознание и программы». Аналитическая философия: Становление и развитие. М., 1998.

И.А.Мельчук «Курс общей морфологии» М., Т.№1, 1997 г.

А.А.Зализняк «Грамматический словарь русского языка» М.: Русский язык, 1980 г.

Ф.Харари «Теория графов» М., 1973 г.

[1] Лексема - это множество словоформ, отличающихся друг от друга только словоизменительными значениями [3].

[2] Словоформы, которые гипотетически принадлежат одной лексеме, для сокращения записи мы будем называть «словоформы на одну лексему» [прим. автора].

[3] Для Д5 - Д9 ДА: корреляция прошла успешно, то есть в деревьях словоформ были обнаружены ложные ветви и удалены; НЕТ: корреляция прошла неуспешно, то есть ложных ветвей не обнаружено. Цифровые индексы на стрелках задают маршрут продвижения по схеме, то есть индекс стрелки выхода из блока Д10 должен совпасть с индексом стрелки входа.