Компьютерная модель морфологического анализа словоформ французского языка

Alexandre Bolkhovityanov, Elena Egorova, Alexei Lavrentiev, Andrey Chepovskiy

Компьютерная модель морфологического анализа словоформ французского языка

А.В. Болховитянов[1], Е.Е. Егорова[2], А.М. Лаврентьев[3], А.М. Чеповский[2]

[1] Московский государственный технический университет им. Н.Э. Баумана
[2] Национальный исследовательский университет «Высшая школа экономики»
[3] ICAR Research Lab - CNRS, Université de Lyon, ASLAN Labex

Аннотация. В данной работе представлено описание модели морфологического анализа словоформ французского языка. Дано краткое описание алгоритма выделения псевдоосновы, главная идея которого заключается в построении структурных схем и соответствующих множеств суффиксов. Описана процедура их построения для случая французского языка. Предложена методика восстановления грамматических характеристик словоформ. Представлены результаты работы предложенных алгоритмов.

Ключевые слова: алгоритм выделения основ, алгоритм морфологического разбора, анализ текстов на естественном языке.

Computerized model of morphological analysis applied to French wordforms

A. Bolkhovityanov[1], E. Egorova[2], A. Lavrentev[3], A. Chepovskiy[2]

[1] Bauman Moscow State Technical University, Moscow, Russia
[2] National Research University Higher School of Economics, Moscow, Russia
[3] ICAR Research Lab - CNRS, Université de Lyon, ASLAN Labex

Abstract. In this paper a computerized model for morphological analysis of French wordforms is proposed. Firstly, a brief description of a stemming algorithm is given. The main idea is to create structural schemes and corresponding lists of suffixes, and the detailed procedure of this work is described in the second part of the paper.. The next part of this work concerns the technique of determining of grammatical characteristics for the French language. In the final part of the work, some results of execution of the proposed algorithms are provided.

Keywords: stemming algorithm, algorithms of morphological analysis, natural language processing.

1 Постановка задачи и методика решения

Данная работа посвящена компьютерной модели морфологического анализа словоформ французского языка, предназначенной для решения следующих задач:

- Разбиение слова на две основные части: основу или псевдооснову и окончание. Под окончанием здесь и далее подразумевается та часть слова, которая не является основой или псевдоосновой.
- Разбиение окончания на значимые составляющие — морфы.
- Восстановление ряда грамматических характеристик словоформ.

Метод структурных схем. Для перечисленных выше задач предлагается использовать метод структурных схем [1]. Рассмотрим

языки, у которых словообразование происходит преимущественно аффиксальным способом. В связи с этим мы можем определить конечный список различных типов аффиксов. Каждому типу соответствует множество морфов, иными словами множество представителей определенного типа. Будем исходить из предположения, что все возможные окончания могут быть представлены, как последовательность из морфов или, что тоже самое, как последовательность из типов аффиксов. Последовательности типов будем называть структурными схемами. Количество различных структурных схем конечно, поэтому задача сводится к перечислению схем, которые описывали бы все возможные окончания словоформ анализируемого языка. Алгоритм сопоставления слова с одной из допустимых структурных схем следующий: на первом шаге необходимо выделить окончание, затем разбить его на морфы, которые существуют в анализируемом языке, далее нужно определить какой последовательности типов морфов соответствует найденное разбиение, и на финальном шаге проверить является ли полученная схема допустимой или нет.

Разрабатываемая модель может быть применена при решении большого спектра задач компьютерной лингвистики: создании оптимального полнотекстового индекса, анализа смысловой составляющей текстов, разметки в задачах корпусной лингвистики.

2 Принципы построения структурных схем для французского языка

Опишем некоторые основные принципы того, как строить структурные схемы для словоформ французского языка. Основной акцент делается на морфологические и словообразовательные особенности французского языка, т.к. именно они позволяют производить корректный морфологический анализ словоформ.

Рассмотрим суффиксальный способ словообразования. Согласно [2], во французском языке выделяют 9 различных частей речи: существительное (nom), прилагательное (adjectif), глагол (verbe), наречие (adverbe), артикль (article), местоимение (pronom), предлог (préposition), союз (conjonction) и междометие (interjection). Суффиксальный способ словообразования чаще всего применяется для образования существительных, прилагательных, глаголов и наречий. Остановимся именно на этих четырех частях речи.

Суффиксы делятся на: словообразовательные и флективные. Для прилагательных, глаголов и некоторых существительных флективные суффиксы служат для изменения слова по роду (мужской и женский) и числу (единственное, множественное). Мы не будем выделять отдельный тип для таких суффиксов. При спряжении глагола изменяются лицо, число, время, род, наклонение и залог. Суффиксы, служащие для этого, также являются флективными. Их тип обозначим за v. Словообразовательные суффиксы служат для формирования новых слов. С их помощью слова могут менять свою часть речи либо обретать определенный оттенок значения (уменьшительность, многократность и т.п.).

Процесс построения структур окончаний строится на основе грамматики французского языка [2-5]. Для французского языка удобнее всего это делать анализируя каждую часть речи, поэтому при построении нашей модели анализировались и строились структурные схемы для четырех основных частей речи: глаголов, существительных, прилагательных и наречий. В силу ограниченности текста данной работы мы приводим далее описание структурных схем только для двух частей речи французского языка – глаголов и существительных.

3 Построение структурных схем на примере глаголов французского языка

Во французском языке при спряжении окончания глаголов изменяются в зависимости от лица, числа, времени, наклонения и залога. По типам спряжения глаголы во французском языке делятся на три группы, которые различаются по типам окончаний в инфинитиве. Глаголы первой и второй группы называются правильными, т.к. имеют определенные правила спряжения, т.е. для них однозначно определен вид окончания в зависимости от различных грамматических характеристик. Глаголы третьей группы являются «неправильными», т.к. при спряжении они меняют свою основу и не имеют единых правил изменения окончания. Все возможные окончания инфинитивов и их формы, которые они приобретают при спряжении, будут учитываться в суффиксах, соответсвующих типу v. Получаем первую возможную схему, описывающую окончания глаголов:

- основа $+ v$

К примеру, глагол первой группы marcher (ходить), его можно представить как march-er. При спряжении этот же глагол может принять форму march-ions (ходили), march-erai (буду ходить) и т.п.

Кроме суффиксов -er, -ir и тп. во французском также можно выделить более сложные суффиксы, например -ifier и -iser. Они служат для образования глаголов от существительных (нарицательных и имен собственных), прилагательных и аббревиатур. Сами эти суффиксы будем рассматривать как сложные, т.е. разбивать их на две части: -ifi-er и -is-er. Это объясняется тем, что при спряжении будет меняться только финальная часть (-er) и тем, что от таких глаголов также можно образовывать новые слова, частички -ifi и -is- будут входить в их состав, а -er нет. Также, во французском языке есть суффиксы, которые не меняют часть речи, а только придают определенный оттенок. К ним относятся суффиксы -aill-, -ass-, -ill-, -och-, -onn-, -ot-, ouill-, -ard-, -âtr-, -et-, -in-; их также будем понимать как сложные. Примеры: chant-er $\rightarrow$ chant-onn-er (петь $\rightarrow$ напевать), touss-er $\rightarrow$ touss-ot-er (кашлять $\rightarrow$ покашливать) и т.д. Таким образом, можно ввести еще один тип суффиксов c, который будет отвечать за одну из частей в сложных глагольных окончаниях. Теперь можем предложить еще одну схему, описывающую окончания глаголов:

- основа $+ c + v$

Во французском языке существуют случаи, когда слово, к которому будет прибавляться глагольный суффикс, также было образовано суффиксальным способом. Иными словами, когда производное слово становится производящим. Здесь стоит выделить два случая: образование глаголов на основе существительных и на основе прилагательных. В связи с этим, можно предложить еще две схемы, описывающие окончания глаголов:

- основа $+ n + v$, за n будем обозначать суффиксы производных существительных, которые стали производящими.

- основа $+ av + c + v$, за av будем обозначать суффиксы производных прилагательных, которые стали производящими.

4 Восстановление грамматических характеристик

Для восстановления грамматических характеристик словоформ вслучае французского языка будем рассматривать самый последний морф окончания.

Для глаголов стоит разделять два случая: сложную и простую форму глаголов. Под сложной формой глагола подразумевается наличие вспомогательного глагола, который используется для того, чтобы поставить глагол в форму определенного времени. Общая схема такова: $A+V_{pp}$, где A — это вспомогательный глагол (avoir или être), V_{pp} — смысловой глагол в форме participe passé. В данном случае, анализируя форму вспомогательного глагола, можно определить следующие грамматические характеристики: время, лицо, число. Под простой формой, соответственно, понимается такая форма глагола, которая не требует использования вспомогательного глагола. Здесь анализируется последний морф глагола. Грамматические характеристики, которые возможно определить: время, род, число.

В словоформах существительных и прилагательных анализируется самый последний морф в окончании. Грамматические характеристики для восстановления: род, число. Здесь возникает проблема множественности, которую без анализа контекста разрешить невозможно. Для наречий последний морф окончания может дать информацию только о части речи.

5 Реализация в программном комплексе

Представленные в предыдущих разделах алгоритмы (алгоритм выделения псевдоосновы, алгоритм разбиения окончания на морфы и алгоритм восстановления грамматических характеристик) реализованы на языке C++ и функционируют в рамках программного комплекса, предназначенного для автоматического анализа словоформ большого числа естественных языков. Такого рода анализ может быть использован в большом количестве прикладных задач: классификация, рубрицирование и т.д.

В основе работы рассматриваемых алгоритмов лежит структура данных типа «бор». В боре хранятся строки в кодировке Unicode (UTF-16), являющиеся морфами. В узлах бора, являющихся финальными, хранится указатель на данные, ассоциированные с данным морфом (например, грамматические характеристики, которые он несет). При

этом, для повышения быстродействия все грамматические характеристики задаются битовыми масками. Например, грамматические характеристики числа:

```
static const uint64 FRANCE_GC_SINGULAR = (uint64)1 << 4;
static const uint64 FRANCE_GC_PLURAL = (uint64)1 << 5;
```

Отдельно для каждого типа схемы хранятся ограничения на длину окончания/основы, которые затем используются на этапе проверки допустимости построенной структурной схемы окончания словоформы.

Так как в некоторых случаях для анализа словоформы необходимо знание контекста, то процедура анализа работает не с отдельными словоформами, а с предложениями. Таким образом, на вход процедуре анализа подается результат парсера, выделяющего в том числе предложения. При обработке слова оно проверяется в хэш-таблице на предмет наличия в списке вспомогательных глаголов. В соответствии с результатом проверки будет использован один из алгоритмов анализа: как одиночное слово или как пара глагол – вспомогательный глагол.

Сама процедура анализа (выделения окончания) одиночного слова представляет собой рекурсивную процедуру с посдеовательным отсечением суффиксов разной длины и проверкой их на наличие в боре морфов. В результате выполнения такой рекусивной процедуры проятся все возможные структурные схемы окончания анализируемой словоформы, представимые известными из грамматики французского языка морфами. После этого отдельно для каждой части речи (а каждая структурная схема окончания соответствует одной части речи) осуществляется проверка на длину получаемой псевдоосновы. Другими словами, для каждой структруной схемы известна минимальная длина псевдоосновы анализируемой словоформы, которая должна оставаться после отсечения окончания. В такой процедуре естественным образом учитывается случай так называемого «пустого» окончания, то есть когда анализируемая словоформа состоит только из основы.

Для восстановления грамматических характеристик по структурной схеме окончания используются сведения и грамматических характеристиках морфов, образующих окончание. Эти грамматические характеристики берутся из бора морфов за время, пропорциональное длине морфа. А так как длина обычно не превышает 6 символов, то время можно считать константным. Грамматические характеристики имеют особый вид в зависимости от части речи анализируемой словоформы, что позволяет легко идентифицировать некорректные варианты построенных рекурсивной процедурой структурных схем

окончания. Грамматические характеристики каждого из морфов пересекаются между собой, чтобы получить общую характеристику для всей словоформы. При этом часть речи восстанавливается по структурной схеме окончания.

В текущей реализации грамматические характеристики отдельных морфов и ограничения на длину псевдоосновы для структурных схем встроены в программный код программного комплекса. Однако, при необходимости, эта информация может быть вынесена в конфигурационные файлы.

Разработанный программный комплекс создан на основе кроссплатформенных технологий и предназначен для использования в операционных системах семества Linux и Windows в различных артитектурах: 32- и 64-бита.

6 Заключение

В данной работе была предложена компьютерная модель морфологического анализа словоформ французского языка, основная идея которой заключается в построении структурных схем окончаний. Для определения эффективности такого подхода на языке C++ были реализованы описанные выше алгоритмы (выделение псевдооснов, разбиение окончания на морфы и восстановление грамматических характеристик) и проведены некоторые эксперименты. В частности, была проанализирована корректность разбиения на структурные схемы и определение грамматических характеристик случайных наборов слов. Считалось, что корректное разбиение существует, если во множестве возможных разбиений присутствует такое, которое оставляет правильную основу и имеет корректное разбиение с точки зрения словообразовательных процессов. Также, считалось, что грамматические характеристики определяются правильно, если в списке возможных грамматических «профилей» есть верный. В итоге получили, что 95 из 100 слов имеют верное разбиение и 82 из 100 верные грамматические характеристики. Также был проведен ряд экспериментов для выявления зависимости между средним количеством подходящих структурных схем для словоформ разных частей речи от ограничения на количество букв в основе. Результаты показывают, что чем большее число букв в основе задается, тем меньше формируется возможных вариантов схем. К примеру, при увеличении количества букв в основе с 3 до 5 множественность падает для прилагательных на 22%, а для существительных на 29%.

Приведенные результаты иллюстрируют возможность применения предложенного подхода для решения задач информационного поиска.

Список литературы

1. Егорова Е.Е., Чеповский А.М. Морфологическая модель для анализа и индексирования текстов на индоевропейских языках // Труды Международной конференции по физико-технической информатике СРТ2013, 12-19 мая 2013 г., Ларнака, Республика Кипр, Изд.ИФТИ, Протвино-Москва, ISBN 978-5-88835-025-6, С.154-159
2. DuBois J.,Lagane R. Livres de bord : Grammaire. – 1995.
3. Grevisse M., Goosse A. Le bon usage. – 2007.
4. Катагощина Н. А. Как образуются слова во французском языке. – М.: URSS: КомКнига, 2006.
5. Huot H. La morphologie: forme et sens des mots du français. – Armand Colin, 2006.-249p.