К ВОПРОСУ ОБ АВТОМАТИЧЕСКОМ МОРФОЛОГИЧЕСКОМ АНАЛИЗЕ ФЛЕКТИВНЫХ ЯЗЫКОВ *

ON AUTOMATIC MORPHOLOGICAL ANALYSIS OF INFLECTIVE LANGUAGES

А. Ф. Гельбух, Г. О. Сидоров

Лаборатория обработки естественного языка,
Центр Компьютерных исследований (CIC),
Национальный Политехнический Институт (IPN),
г. Мехико, Мексика
gelbukh@gelbukh.com, sidorov@cic.ipn.mx; www.Gelbukh.com

Аннотация. В статье рассматривается вопрос о построении систем с морфологическим анализом с точки зрения количества затраченных на разработку усилий. Приводится анализ моделей, которые могут использоваться в разных подходах, и предлагается простой подход, основанный на «анализе через синтез», который позволяет использовать при анализе модели, ориентированные на синтез, что существенно уменьшает затраты на разработку. Описаны системы для русского и для испанского языков, разработка которых была основанные на этом подходе. Обе системы доступны для свободного использования и могут быть загружены с соответствующей Интернет-страницы.

Введение

Морфологическая система флективных языков конечна. Это значит, что любой подход к построению таких систем, основан ли он на правилах или на конечных автоматах, пользуется ли словарем основ или базой данных словоформ, в конце концов приводит к одному и тому же результату, если, конечно, используемые модели не редуцируются. Важными параметрами оценки систем с мофрологическим анализом обычно считаются размер словаря, скорость обработки текстов и возможность обработки незнакомых слов. Однако нам представляется, что в этот перечень необходимо добавить очень существенный параметр, относящийся у тому, сколько усилий затрачено на разработку системы — и связанный с ним вопрос, насколько морфологические модели, использованые для разработки алгоритмов, соответствуют лингвистической реальности и интуиции носителей языка. Заметим, что обычно эти соображения во внимание не принимаются (Коваль, 2003).

В данной статье мы предлагаем подход, позволяющий использовать практически без изменений морфлогические модели, описанные в традиционных грамматиках и словарях. Это позволяет сократить до минимума усилия при разработке систем морфологического анализа и сохранить интуитивную ясность моделей.

Предположим, что у нас имеется хороший традиционный словарь, описывающий словоизменение — для русского языка таким словарем является словарь Зализняка (1980) — то есть, что нет надобности создавать с нуля такой словарь. Как всякий традиционный словарь, такой словарь ориетирован на синтез, то есть на порождение словоформ, а не на анализ. Тогда имеется две возможности — в зависимости от материальных ресурсов, которыми мы располагаем.

Если имеется достаточно ресурсов — например, по опыту разработки систем морфлогического анализа для русского языка можно оценить как достаточную возможность задействовать порядка десяти квалифицированных разработчиков (программистов и лингвистов) в течении одного года — то вопросами о моделях можно не задавтся: моделируемая система конечна, ресурсов достаточно, а значит, можно сгенерировать все возможные уникальные классы словоизменения и приписать их словам, создав соответствующий алгоритм, который на базе классов слов, представленных в словаре для синтеза, строит классы слов для анализа. Для русского языка количество таких классов составляет порядка 1000 (Gel'bukh, 1992); правда, если мы будем учитывать различные схемы ударений (скажем, для последующей расстановки ударений), то, вероятно, количество классов будет порядка 2500 (Сокирко,

* Работа выполнена при частичной поддержке правительства Мексики (Конацит, СНИ) и Национального политехнического института (CGPI, COFAA). Мы благодарим И.А. Большакова за полезное обсуждение. Имена авторов даны в алфавитном порядке. Work was done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute, Mexico (CGPI, COFAA). We thank Prof. Igor A. Bolshakov for useful discussions. Author names are given in alphabetic order.

Для каждого известного системе окончания, имеющегося в данном языке,
Это окончание отделяется от входной словоформы, что дает потенциальную основу.
Если основа имеется в словаре, то
Синтезируются все ее формы соответствии с гипотезами для данного окончания
И сравниваются с входной словоформой.
Каждая сравнившаяся форма приводит к выдаче соответствующего набора грамматических характеристик, известного из процесса синтеза.

Рис 1. *Алгоритм анализа через синтез*

2004; однако неясно, связано ли это количество классов еще и с отдельно обрабатываемыми префиксами). Для чешского языка количество таких классов составляет около 1500 (Sedlacek and Smrz, 2001). Про том, что в исходной модели Зализняка для русского синтеза имеется порядка 40 словоизменительных классов, хотя каждый из них может иметь дополнительные пометы, отражающие различные особенности словоизменения, что и приводит к очень большому количеству словоизменительных классов для анализа. При этом полученные словоизменительные классы для использования в анализе никак не соотносятся с интуицией носителей языка, которая также ориентирована на синтез.

Если же ресурсы ограничены, но имеется необходимость в разработке системы морфологического анализа для какого-либо флективного языка, то можно применить подход, предлагаемый в данной статье, который позволяет использовать уже существующую морфологическую модель, ориентированную на синтез, практически без изменений. Более того, этот подход сохраняет соответствие используемых морфологических моделей с интуицией носителей языка. По нашему опыту, для русского языка хватило работы одного человека в течении года — и можно было сделать это в несколько раз быстрее, если бы подход уже был разработан; именно размышления над процессом разработки системы для русского языка послужили основой для разработки предлагаемого подхода (Сидоров, 1996). При разработке аналогичной системы для испанского языка (у которого, правда, более простая морфологическая система) для построения моделей и программирования алгоритмов потребовалось около двух месяцев работы одного студента (Velázquez *et al*., 2002).

Далее в статье описывается предлагаемый подход и дается описание его применения к русскому и испанскому языкам, а также обсуждается подход, основанный на хранении всех словоформ в словаре.

Подход «анализ через синтез»

Идея «анализа через синтез» была высказана достаточно давно в области искусственного интеллекта. Она состоит в том, что одни модули генерируют гипотезы, а другие модули эти гипотезы проверяют на соответствие с наблюдаемыми входными данными. Общий принцим, стоящий за этой идеей, состоит в том, что синтез всегда проще анализа, потому что не содержит необходимости проверять комбинаторику всех возможных вариантов (Мельчук 1974); достаточно сравнить, например, нынешнее состояние работ по синтезу речи и распознаванию речи.

Применение этой идеи к морфологическому анализу дает схему обрабоки слов, показанную на Рис. 1.

Аналогичная идея применяется к хранению основ в словаре. Можно обрабатывать основы динамически (Мальковский, 1985), то есть построить много правил построения «первой» основы из ее алломорфов. По оценке Хауссера (Hausser, 1999, p. 252) эта проблема имеет очень большую сложность: приходится разрабатывать много достаточно антиинтуитивных правил и выполянть излишний поиск в словаре, чтобы проверить, существуют ли построенные основы. Альтернативная возможность состоит в синтезировании заранее всех вариантов основ и хранении их в словаре с соответствующей пометой о номере основы (Gel'bukh, 1992) — т.е., сначала производится синтез, который затем используется в анализе.

Рассмотрим вопрос о том, какие типы лингвистичесих моделей должны использоваться при таком подходе.

Окончания → граммемы: модель анализа, ставящая в соответствие окончаниям все возможные наборы граммем, которые являются гипотезами. Эта модель в явном виде присутсвует в традиционных парадигмах, хотя в нашем случае её входом является окончание, а не граммемы. Например, для некоторых существительных в *Им. п., ед. ч.*, используется окончание *-а*, то есть по этому окончанию высказывается гипотеза о таком наборе граммем для данного грамматического класса. Очевидно, что гипотезы могут повторятся как для окончаний, так и для наборов граммем, то есть, окончанию *-а* могут соответствовать и другие наборы граммем (гипотезы), а набору граммем *Им. п., ед. ч* могут соответствовать и другие окончания, включая нулевое. Заметим, что информация о классе основы и об особенностях словоизменения, которая хранится в словаре, на

этом этапе не требуется, поскольку реальная совместимость будет проверена при синтезе.

Граммемы → окончания: модель синтеза, то есть, практически, традиционная парадигма, в которой уже принимаются во внимание все необходимые особенности словоизменения. В этой модели наборам граммем ставтся в соответствие окончания. Например, для *существительных женского рода, класса 1* в *Им.п., ед.ч* используется окончание *-а*.

Выбор алломорфа основы: практически это модель синтеза с точки зрения основ — какой алломорф основы соотвествует заданному набору граммем в данном классе. Например, для слов среднего рода с чередованием основ, для набора граммем *Род. п., мн. ч*. используется вторая основа, а в остальных случаях —первая (нумерация основ, очевидно, условна). Эта модель в явном виде не присутствует в традиционных описаниях, но легко выводится из приведенных там парадигм.

После построения этих трех моделей процесс разработки системы сводится к программированию простых правил вида «если—то» для синтеза и для генерации гипотез.

Подход «все словоформы в словаре»

Альтернативная возможность реализации автоматического морфологического анализатора сводится к предварительному синтезу всех возможных словоформ и их хранению в одной большой базе данных, с последующим поиском в ней. Такой подход является самым простым и дает максимальную скорость, потому что весь алгоритм сводится к поиску словоформ в словаре (т.е. в базе данных или в конечном автомате). Алгоритмы синтеза при этом, однако, все равно требуются — а как мы уже показали, эти алгоритмы составляют бóльшую часть системы анализа.

Имеются у такого подхода и другие недостатки, связанные с отсутствием структуры в анализе. Прежде всего это относится к возможностям обработки слов, отсутствующих в словаре. Кроме того, база данных получается излишне большой; в этом смысле можно рассматривать алгоритм анализа как способ сжатия базы данных словоформ. Возникают проблемы и при поддержке такой базы данных, потому что вести какую-либо ручную обрабоку ее практически невозможно из-за слишком большого числа словоформ, соответствующих каждой лемме.

Тем не менее, этот подход позволяет затратить еще меньше усилий на разработку системы анализа, потому что даже при подходе «анализ через синтез» приходится разработать примерно в два раза больше алгоритмов. Следовательно, при необходимости разработать систему морфологического анализа для какого-либо флективного языка при крайне ограниченных ресурсах надо решать, является ли размер базы данных чрезмерной платой за отсутствие необходимости разработки алгоритмов анализа. Если да, то можно использовать подход с «анализом через синтез»; если нет, то возможно хранение всех форм в базе данных (словаре словоформ).

Система морфологического анализа для русского языка

Опишем две разработанные нами системы, основанные на подходе «анализ через синтез» — для русского и для испанского языков.

В качестве основы для разработки моделей русского словоизменения были использованы модели из словаря Зализняка (1980). В соответствии с предлагаемым подходом было разработано три типа моделей: блок высказывания гипотез, блок синтеза для основ и блок синтеза для окончаний. Отдельно обрабатываются особые случаи, связанные с удалением морфемы *-ся* и с возможностью построения «дубликатных» форм, таких как второй родительный падеж.

Система доступна для свободного использования[1]. Ее можно загрузить (в варианте для Windows) с сайта www.cic.ipn.mx/~sidorov/rmorph или www.Gelbukh.com/rmorph.

Система проста в подключении. Она поставляется в двух версиях. Одна версия представляет собой исполняемый модуль (.exe), который обрабатывает входные файлы и строит выходной файл в следующем формате: *входная словоформа, лемма*1 *[информация*1*]* *лемма*2 *[информация*2*]* и т.д., например:

быстро быстрый [кратк.,ед.ч.,ср.р.,]
стол стол [вин.п.,ед.ч.,] стол [им.п.,ед.ч.,]
был быть [изъяв.,прош.вр.,ед.ч.,муж.р.,]

Другая версия является динамически загружаемой библиотекой (DLL), что позволяет легко включать ее в пользовательские программы.

В данной версии программ омонимия словоформ не разрешается. В случае необходимости рекомендуем пользоваться существующими тэггерами (Brill tagger, TnT и др.).

Система содержит весь словарь Зализняка — около 100,000 лемм. Ее быстродействие меньше чем, скажем, у системы Сокирко (2004) — около 40 КБ текста в секунду на процессоре Pentium IV, что, однако, вполне достаточно для большинства возможных приложений. Надо отметить, что мы не ставили своей целью достичь максимального быстродействия: это потребовало бы гораздо больших затрат на программирования, что сделало бы невозможным предоставление системы для свободного использования.

---

[1] В отличии, скажем, от системы Сокирко (2004), которая пока не доступна для свободного использования в варианте для Windows; на сайте укзано, что система должна стать бесплатной с декабря 2005 г.

Выдача программы	Приблизительный перевод слов
me (13) *yo* (*PP1CSR0) (0)	*-ся: я*
encontré (14) *encontrar* (*VMID1S0) (0)	*встретил: встретить*
a (15) *a* (*SPS00) (0)	вин. п.: вин. п.
toda (16) *todo* (*DI3FS00) (0) *todo* (*PI3FS00) (1)	*всю: весь всё*
la (17) *ella* (*PP3FSR0) (0) *la* (*TDFS0) (1)	артикль*: она* артикль
tripulación (18) *tripulación* (*NCFS000) (0)	*команду: команда*

Рис. 2. *Пример результатов морфологического анализа для испанского языка*

Система морфологического анализа для испанского языка

Основанная на тех же принципах система была разработана для морфологического анализа испанского языка. В ней использован тот же набор моделей, как и в русской версии. Поскольку морфологический строй испанского языка проще русского, то для построения моделей и программирования алгоритмов потребовалось около двух месяцев работы одного студента.

Так же как и система для русского языка, данная система реализована в двух версиях: исполняемый модуль (.exe) и динамически загружаемая библиотека (DLL). Скорость обработки составляет около 5 КБ в секунду, что меньше, чем у модулей в системе анализа русского языка. Это связано с использованием стандартной (медленной) базы данных для хранения словаря, однако мы ожидаем в ближайшее время улучшить в несколько раз этот показатель, заменив модуль поиска в словаре на модуль, аналогичный использованному в русской версии. В настоящий момент размер словаря составляет около 26,000 лемм, что не очень много, однако покрывает практически всю лексику неспециализированных текстов. Мы планируем пополнять словарь системы. В качестве схемы обозначений использована схема PAROLE — фактический стандарт для испанского языка. В этом стандарте словоформам приписывается код, первый символ которого обозначает часть речи, второй — опциональную лексическую характеристику, и т.д., например, V в VMID1S0 означает глагол. На Рис. 2 приводится пример выдачи системы (для фразы ‘*я встретил всю команду*’, с приблизительным переводом).

Система под Windows доступна для свободного использования. Ее можно загрузить с сайта www.cic.ipn.mx/~sidorov/agme или www.Gelbukh.com/agme.

Заключение

При необходимости разработки систем морфологического анализа для флективных языков в условиях отсутствия достаточных ресурсов можно воспользоваться предлагаемым подходом, сосотоящим в «анализе через синтез», который требует небольших затрат на разработку алгоритмов. Кроме того, предлагаемый подход позволяет сохранить практически без изменений морфлогические модели из традиционных грамматик и словарей, сответствующие интуиции носителей языка.

Список литературы

1) Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2. // М.: Наука, 1989. 296 стр.
2) Зализняк .А.А. Грамматический словарь русского языка. // М., Русский язык, 1980.
3) Коваль С.А.Сравнимости и эквивалентности компьютерных представлений морфологии. // Труды межд. конф. Диалог-2003, Москва, 2003.
4) Мальковский М.Г. Диалог с системой искусственного интеллекта. // МГУ, Москва, 1985. 213 стр.
5) Мельчук И.А. Опыт теории моделей смысл ⇔ текст. // Москва, 1974.
6) Сидоров Г.О. Разработка и реализация лингвистического обепечения систем с морфлогическим анализом/синтезом для русского языка. // Автореф. дисс. на соиск. уч. степ. канд. филол. наук. Москва, МГУ, 1995. 29 стр.
7) Сокирко А.В. Морфологические модули на сайте www.aot.ru. // Труды межд. конф. Диалог-2004, Москва, 2004.
8) Beesley K.B. and Karttunen L. Finite state morphology. // CSLI publications, Palo Alto, CA, 2003.
9) Gel’bukh A.F. Effective implementation of morphology model for an inflectional natural language. // Automatic Documentation and Mathematical Linguistics, Allerton Press, vol. 26, N 1, 1992, pp. 22–31.
10) Gelbukh A. and Sidorov G. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. // In: A.Gelbukh (ed.) Proc. of CICLing-2003 (4th International Conference on Intelligent Text Processing and Computational Linguistics), February 15–22, 2003, Mexico City. Lecture Notes in Computer Science, N 2588, 2003, Springer-Verlag, pp. 215–220.

11) Hausser R. Foundations of Computational linguistics. // Springer, 1999, 534 p.

12) Karttunen L. Computing with realizational morphology. // In: A.Gelbukh (ed.) Proc. of CICLing-2003 (4th International Conference on Intelligent Text Processing and Computational Linguistics), February 15–22, 2003, Mexico City. Lecture Notes in Computer Science N 2588, Springer-Verlag, pp. 203–214.

13) Koskenniemi K. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. // Tesis Doctoral. Universidad de Helsinki. 1983. 160 p.

14) Sedlacek R. and Smrz P. A new Czech morphological analyzer AJKA. // Proc. of TSD-2001. LNCS 2166, Springer, 2001, pp. 100–107.

15) Sproat R. Morphology and computation. // Cambridge, MA, MIT Press, 1992, 313 p.

16) Velásquez F., Gelbukh A. y Sidorov G. AGME: un sistema de análisis y generación de la morfología del español. // In: Proc. of Workshop "Multilingual information access and natural language processing" of IBERAMIA 2002 (8th Iberoamerican conference on Artificial Intelligence), Sevilla, Spain, November, 12, 2002, pp 1–6.