Построение морфологического анализатора неизвестных слов на основе словарей системы ЭТАП-3

Казенников А.О.
ИППИ РАН
kzn@iitp.ru

Аннотация

В настоящей работе представлен способ построения морфологического анализатора для неизвестных слов на основе словарей системы ЭТАП-3. Анализатор строится на базе конечного автомата. В качестве исходного материала для построения анализатора используются словари системы ЭТАП-3. При построении анализатора в автомат вносится изменяемая часть слова с приписанными морфологическими характеристиками. Представленный в работе алгоритм был экспериментально опробован на корпусе СинТагРус, где показал свою эффективность. Таким образом, алгоритм может использоваться как для непосредственной задачи анализа неизвестных слов, так и для задачи стемминга (лемматизации).

1. Введение

Морфологический анализ является первым шагом в традиционной схеме анализа текстов. Результаты этого шага являются важными, поскольку большинство задач более высокого уровня опирается на результаты морфологического анализа.

В системе ЭТАП-3 [1] исторически используется подход на основе знаний, которые формализуются экспертами-лингвистами. При таком подходе возникает проблема запаздывания. Новые явления или знания не сразу формализуются и заносятся в систему. В частности, к таким знаниям относится появление новых слов в языке. Создание морфологического анализатора незнакомых слов, или гессера, в значительной мере способствует решению этой проблемы или хотя бы смягчит ее остроту.

2. Постановка задачи

В настоящей работе предлагается несколько видоизмененное понятие морфологического анализа, отличное от того, которое используется в системе ЭТАП-3. В контексте системы ЭТАП-3 под морфологическим анализом понимается процесс преобразования цепочки символов в цепочку записей о словах, которые в свою очередь состоят из набора пар {лемма, морфологические характеристики}. В более узком смысле морфологический анализ — представление одного слововхождения в виде вероятны х пар {лемма, морфологические характеристики}, поскольку из-за омонимии слововхождение может быть разобрано несколькими способами. Лемма осуществляет привязку данного слововхождения к словарям системы ЭТАП. Очевидно, что для неизвестных слов такую привязку сделать невозможно, поэтому под морфологическим анализом неизвестных слов будет пониматься получение морфологических характеристик для заданного неизвестного слова. В системе ЭТАП такие слова связываются со словарными записями служебной группы FICT.

Другой особенностью архитектуры системы ЭТАП-3 является то, что морфологическая омонимия на этапе морфологического анализа может не разрешаться полностью. Это сделано для того, чтобы как можно в меньшем числе случаев отбросить правильную информацию. Таким образом, на выходе морфологического анализа получается не одна пара «лемма-морфологические характеристики», а несколько. В англоязычной литературе обычно также предполагается, что из нескольких таких пар выбирается одна. Однако, если в процессе выбора возникнет ошибка, она распространится на все остальные шаги анализа данного текста.

Задачу морфологического анализа неизвестных слов можно сформулировать следующим образом. Для заданного неизвестного слова необходимо получить список наборов морфологических характеристик, которыми потенциально может обладать данное слово. Анализатор должен выдавать как можно более короткий список, в котором бы присутствовал правильный с человеческой точки зрения набор морфологических характеристик.

Тем самым морфологический анализ неизвестных слов можно представить как задачу ранжирования наборов морфологических характеристик для конкретного слова. В такой постановке задачи изначально каждому рассматриваемому неизвестному слову приписываются все возможные наборы морфологических характеристик. Затем для этого слова невозможные наборы характеристик удаляются. Потом список оставшихся наборов морфологических характеристик сортируется в порядке убывания вероятности приписывания данного набора характеристик рассматриваемому слову. Последний шаг — отсечение маловероятных наборов. Это делается с целью уменьшения степени неоднозначности разборов.

В настоящей работе рассматривается построение анализатора неизвестных слов для русского языка. Хотя с точки зрения системы ЭТАП-3 важно располагать средствами обработки неизвестных слов для обоих основных языков, каковыми являются русский и английский, для английского языка существуют разработанные решения в виде алгоритмов частеречной разметки, которые в прямом виде слабо применимы для русского. Именно поэтому мы сосредоточимся на решении для русского языка.

3. Построение анализатора

Построение анализатора основано на двух предположениях.

Первое предположение состоит в том, что морфология языка в основном регулярна. В частности, было экспериментально установлено, что список всех словоформ русского языка с приписанными им морфологическими характеристики объемом около 4 млн. единиц описывается конечным автоматом, состоящим приблизительно из 130 тыс. состояний[2]. Таким образом, на одно состояние конечного автомата приходится в среднем 30 словоформ. Второе предположение заключается в том, что новые слова являются регулярными относительно достаточно большого морфологического словаря (например, словаря системы ЭТАП-3), поскольку естественно ожидать, что все нерегулярные слова давно укоренились в языке и уже присутствуют в словаре.

По этой причине в качестве базового материала для построения анализатора используется морфологический словарь системы ЭТАП-3.

Принципиальный алгоритм построения анализатора состоит в следующем:

1. Создать конечный автомат
2. Для каждой словоформы словаря
3. Выделить изменяемую часть — псевдоокончание

1. Выделить морфологические характеристики

5. Добавить в конечный автомат пару {псевдоокончание (в обратном порядке), морфологические характеристики}
6. Сохранить конечный автомат для дальнейшего использования

Разбор неизвестных слов производится следующим образом:

1. Загрузить конечный автомат, полученный при построении анализатора.
2. Определить границы неизвестного слова.
3. Начать обход конечного автомата по неизвестному слову в обратном порядке
4. В процессе обхода сохранять все встретившиеся наборы морфологических характеристик.
5. Отсортировать полученный список пар {псевдоокончание, набор морфологических характеристик}
6. Удалить из списка наименее вероятные варианты разбора

В результате разбора получается искомый список наборов потенциальных морфологических характеристик для заданного неизвестного слова. В представленных алгоритмах необходимо доопределить процедуры выделения изменяемой части слова и процедуру сортировки полученного списка наборов характеристик. Эти процедуры полностью определяют качество работы анализатора.

В такой формулировке задача морфологического анализа неизвестных слов родственна задаче лемматизации, или стемминга — отсечения изменяемой части слова для уменьшения суррогатного словаря. Такая задача встречается в области информационного поиска,

когда для сокращения размера индекса необходимо уменьшение списка всех встреченных словоформ.

Возможно несколько вариантов выделения изменяемой части словоформы. Простейшим способом является нахождение общей неизменяемой части слова на основании словаря. В таком варианте словарь рассматривается как список кортежей {словоформа, лемма, морфологические характеристики}. Из каждого кортежа на основании пары {словоформа, лемма} выделяется изменяемая часть слова, которой приписываются морфологические характеристики рассматриваемого кортежа. Этот способ является исключительно алгоритмическим и в дальнейшем будет использоваться в качестве базового метода при сравнении.

Другим способом является выделение изменяемой части на основе полной словарной информации. Морфологический словарь системы ЭТАП-3 можно представить двумя способами. С одной стороны, словарь является компактной записью всех возможных кортежей {словоформа, лемма, расширенная лемма, набор морфологических характеристик}, а с другой — описывает механизм построения такого списка. В частности, запись словаря описывает словоизменение заданной лексемы. Этот механизм состоит в том, что каждая словоформа собирается из морфем, а компактная запись содержит описание того, какие морфемы или списки морфем использовать. Сами морфемы бывают нескольких типов: приставка, основа, тема, суффикс, окончание, (возвратная) частица. Таким образом, словоформа является не только цепочкой знаков, но и цепочкой морфем. Поэтому изменяемую часть слова возможно тоже собирать из морфем или их частей. В частности интуитивно оправданным вариантом использования морфемной записи является формирование псевдоокончания на основе морфем окончания и частицы. Расширением этой модели является формирование псевдоокончания не только на основе окончания и частицы, но также и на основе суффикса и темы. Дальнейшим расширением модели является добавление нескольких букв основы слова. В этой модели возможные варианты разбора ранжируются по длине совпавшего псевдоокончания. В случае использования части основы ранжирование сначала производится по длине совпавшей основы, а затем по остальной части псевдоокончания.

Другим существенным изменением модели является ранжирование не только по длине совпавшей части псевдоокончания, но и по степени вероятности парадигмы, связанной с рассматриваемым псевдоокончанием. Это возможно при использовании большого неразмеченного корпуса текстов. Для этого из корпуса извлекаются все возможные словоформы. Далее, на их основе производится оценка покрытия потенциальной парадигмы ~~в корпусе~~. В настоящей работе используется наиболее простой способ оценки покрытия — отношение доли встреченных словоформ рассматриваемого слова к размеру парадигмы. Например, в интернет-жаргоне есть слово «мкадск» в значении «город Москва». При рассмотрении анализатором словоформы «мкадска» среди возможных вариантов разбора будет интерпретация данной словоформы, как существительного мужского рода на «-ск», по типу «Курск», «Смоленск», или как существительного женского рода на «-ска» типа «краска». Идея подхода состоит в том, чтобы оценить вероятности парадигм на основе встречаемости слов данной парадигмы в корпусе. Например, если в каком-либо тексте встретилось слово «мкадска», то наличие в корпусе словоформы «мкадском», как подсказывает интуиция, повышает вероятность того, что «мкадска» относится к парадигме существительного мужского рода. Однако если в корпусе будет только словоформа «мкадске», то на ее основе невозможно предпочесть один из двух представленных вариантов разбора. Для построения такой модели необходимо дополнительно не только извлекать из словаря псевдоокончания, но также и составлять группы таких псевдоокончаний — парадигмы.

7. Предшествующий опыт

Фактически в настоящей работе морфологический анализ сводится к идентификации окончания и приписыванию рассматриваемому слову набора признаков, связанных с этим окончанием.

В литературе представлены подходы к решению первой части задачи — идентификации окончания заданного слова. В частности, аналогичная задача решается при проведении стемминга. Задача стемминга состоит в том, чтобы отсечь от заданного слова изменяемую часть. Такая задача востребована в приложениях информационного поиска с целью сокращения

размера словаря слов для индексирования корпуса текстов. Наиболее широко известен алгоритм лемматизации Портера [3]. Первоначально он был создан для английского языка. Основная идея заключается в том, что словоформа рассматривается как последовательность гласных и согласных. На основе такого представления записываются правила отсечения изменяемых частей. Такое представление позволяет ограничиться небольшим набором правил для эффективной работы. Позже на основе этой идеи был разработан пакет Snowball, который позволяет записывать правила отсечения окончаний на специальном языке. В настоящее время существуют правила лемматизации для многих языков, в том числе и для русского.

Другим подходом является подход лемматизатора CST [4]. Здесь на основе некоторого аннотированного корпуса формируются правила преобразования символьных цепочек, где исходной цепочкой служит словоформа в тексте, а конечной — лемма этой словоформы. Основная идея заключается в том, что эти правила могут быть не только полностью определенными, но и являться схемой правил на основе регулярных выражений. Это достигается с помощью нахождения общих частей словоформы и леммы. В настоящей работе одним из подходов (он именуется алгоритмическим, см. ниже) является модификация этого алгоритма. Выделение изменяемой части тут производится исключительно алгоритмически.

Развитием подхода CST можно считать алгоритм, представленный в работе [5]. В нем не только производится выделение общей части, но и используется аппарат парадигм для проверки некоторого разбора слова.

Настоящая работа во многом опирается на [5], однако, есть и существенные различия. Во-первых, различается постановка задачи. Нам необходимо не столько получить из неизвестной словоформы лемму, сколько полный набор морфологических характеристик; кроме того, необходимо получать все множество высоковероятных возможных разборов. Следовательно, требуется процедура ранжирования этих наборов, а также процедура отсечения маловероятных наборов. В-третьих, в качестве исходных данных в настоящей работе использовался не только размеченный корпус текстов, но и морфологический словарь системы ЭТАП-3, из которого возможно извлекать не только словоформы, связанные с определенной леммой, но также и морфемы, из которых эти словоформы формируются.

8. Эксперименты

Оценка различных методов разбора неизвестных слов является нетривиальной задачей, поскольку готового проверочного эталона не существует. Для проверки было решено использовать корпус СинТагРус [6] в качестве морфологически размеченного материала. Поскольку СинТагРус использует расширенную и идеализированную модель как морфологии, так и синтаксиса, то для проведения экспериментов по проверке анализатора неизвестных слов необходима предварительная обработка корпуса. В частности, из проверочного множества были исключены некоторые классы слов, в том числе:

– различные цифровые и буквенно-цифровые последовательности, например «2001», «2-НДФЛ» и т.п.
– словоформы, для которых в морфологическом словаре системы ЭТАП-3 нет соответствующего набора морфологических характеристик (наборы имеют расширенные характеристики или являются неполными).

Алгоритм анализатора	Средний ранг
Базовый	55
Только окончания	32
Псевдоокончания	26
Псевдоокончания + 3 буквы основы	12
Использование парадигм	10

Таблица 1. Средний ранг правильного набора характеристик

Таким образом, было получено около 631 тыс. проверочных слов.

Для сравнения представленных алгоритмов построения анализатора в качестве базового используется метод выделения общей части слова.

Метод оценки анализатора состоит в оценке среднего ранга эталонного набора характеристик при анализе проверочных слов как неизвестных. Кроме того, необходимо оценить покрытие

анализатором неизвестных слов — хороший анализатор должен правильно анализировать наибольшее число неизвестных слов.

В таблице 1 представлены результаты оценки среднего ранга каждого из подходов. Из экспериментов следует, что алгоритмический подход неприменим для практических приложений, поскольку для правильной работы необходимо рассмотреть в среднем около 60 вариантов разбора. Если учитывать только словарное окончание, то число таких вариантов снижается до 32. Интересно, что добавление суффикса лишь немного снижает средний ранг – до 28. Гораздо более существенный эффект оказывает добавление к псевдоокончанию нескольких букв основы слова. Добавление информации о парадигме снижает средний ранг до 12.

Кроме того, была проведена серия более точных экспериментов. В ней оценивалось качество анализатора неизвестных слов только на множестве слововхождений, которые не были корректно проанализированы с помощью стандартного морфологического модуля системы ЭТАП-3. В корпусе СинТагРус таких слов обнаружилось около 6 тыс. Однако около 2 тыс. из них были записаны латинскими буквами и поэтому были непригодны для эксперимента (например «Auchan», «NASDAQ», «Post», «mortem» и т. д.). Таким образом, объем проверочного корпуса составил около 4 тыс. слововхождений. В таблице 2 представлены результаты оценки анализатора на этом подмножестве слововхождений.

Алгоритм анализатора	Средний ранг
Псевдоокончания	30
Псевдоокончания + 3 буквы основы	14
Использование парадигм	12

Таблица 2. Средний ранг для неизвестных слов корпуса СинТагРус

При сравнении результатов, представленных в таблицах 1 и 2, видно ухудшение при переходе к анализу только неизвестных слов. Это обусловлено тем, что в подмножестве неизвестных слов много практически неизменяемых фамилий, которые имеют одинаковое написание как в мужском, так и в женском роде (например «Джонс», «Шнирельман»). В таких ситуациях правильный набор характеристик может оказаться далеко от первой позиции. С другой стороны, довольно часто встречается ситуация, когда правильный набор характерик входит в первые три варианта разбора (например при разборе слововхождения «одновитковой» правильный набор характеристик находится на втором месте).

9. Выводы

В настоящей работе представлен способ построения морфологического анализатора неизвестных слов на основе словарей системы ЭТАП-3. Было установлено, что наиболее эффективной основой такого анализатора являются псевдоокончания, состоящие из нескольких букв основы и обычной изменяемой частью слов, состоящей из суффикса, окончания и частицы. В качестве алгоритма сортировки самым эффективным оказался алгоритм на основе оценки вероятности парадигм.

Представленный анализатор можно использовать в системе ЭТАП-3. Кроме того, возможно его использование в качестве процедуры стемминга.

10. Литература

1. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2. М., Наука, 1992
2. Казенников А.О. Использование конечных автоматов для морфологического анализа и синтеза на основе словарей системы ЭТАП. ИТиС’08. г. М., 2008. с. 201-205. ISBN 978-5-901158-08-01.
3. Porter M.F., An algorithm for suffix stripping, *Program*, 14(3), 1980, pp 130−137.
4. Jongejan B., Dalianis H., Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike, AFNLP'09, pp. 145-153.
5. Segalovich I., A fast morphological algorith with unknown word guessing induced by a dictionary for a web search engine. MLTMA'03, Las Vegas
6. Ю.Д. Апресян, И.М. Богуславский, Б.Л. Иомдин, Л.Л.Иомдин, А.В. Санников, В.З. Санников, В.Г. Сизов, Л.Л. Цинман. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–2005 г. (результаты и перспективы). М: «Индрик», 2005. С. 193-214