Распознавание словоформы в разных вариантах транскрипционной записи [1]

И.В.Жарков

igor@iz2803.spb.edu

На филологическом факультете Санкт-Петербургского государственного университета проводятся работы, направленные на создание системы автоматического распознавания речи. При этом авторы системы ставят перед собой задачу распознавания речи в различных произносительных вариантах нормы.

Одна из главных проблем, решение которых входит в компетенцию любой системы распознавания речи, вне зависимости от подхода и применяемых методов, — это задача детранскрипции, то есть получения орфографического представления последовательности транскрипционных знаков.

На вход детранскрипционного компонента системы распознавания речи подается транскрипция (или несколько возможных вариантов транскрипции) произвольной синтагмы. Программа осуществляет автоматическое членение последовательности транскрипционных знаков на фразы и синтагмы. Для выделения фраз и синтагм используются два основных признака: паузы и синтагматические ударения.

При наличии в транскрипционном тексте пауз между синтагмами выделение последних является задачей совершенно тривиальной. При их отсутствии, тем не менее, синтагмы могут быть выделены с опорой на синтагматическое ударение, которое в русском языке, как правило, характеризует последнее фонетическое слово синтагмы.

В задачу компонента входит отождествление транскрипционного представления синтагмы с одной или с несколькими цепочками орфографически правильных словоформ с возможностью последующего подключения компонента контекстного и/или синтаксического анализа, позволяющего выбрать один из орфографических вариантов, предложенных на этапе детранскрипции. На настоящий момент нами полностью реализована детранскрипция отдельных, изолированных словоформ. В данное время ведутся работы, направленные на создание компонента членения транскрибированной синтагмы на последовательность словоупотреблений. Выделение фонетических слов может производиться, например, следующим способом. Начальный фрагмент транскрипционной записи синтагмы, включающий в себя первый ударный квазислог (см. ниже), но не включающий в себя второй ударный квазислог, подается на вход функции детранскрипции словоформы. При удачном поиске словоформы в словаре словоформа считается выделенной; при неудачном поиске последовательность транскрипционных знаков каждый раз укорачивается на один символ, и поиск повторяется. После того, как первая словоформа идентифицирована, на вход функции детранскрипции словоформы подается следующий фрагмент траскрипционной записи синтагмы (включающий в себя второй ударный слог, но не включающий третьего), и т. д.

Отождествление транскрипционного и орфографического представления словоформ производится при помощи словаря. Словарный подход, вполне укладываясь в рамки общей тенденции развития технологий распознавания речи, состоящей в использовании словарей большого объема, обладает в то же время несомненной новизной в области распознавания русской речи: большинство систем распознавания русской речи работают с небольшими словарями (от нескольких десятков до 2–3 тысяч слов).

Чрезвычайно важный фактор, влияющий на эффективность детранскрипции, — это структура данных, в том числе логическая и физическая структура словарной базы данных. Представление словарных данных должно обеспечивать, с одной стороны, максимально быструю обработку транскрибированной синтагмы, а с другой стороны, реализацию детранскрипции изолированных словоформ с учетом особенностей произносительной нормы (московской, петербургской и др.).

Природа задачи распознавания речи определяет ряд требований к системам распознавания. Одним из самых существенных требований является скорость распознавания, обеспечивающая возможность работы системы в реальном времени. Организация словарной базы данных должна обеспечивать максимально высокую скорость доступа к данным; минимизация физического объема базы отходит здесь на второй план.

В качестве лексикографической основы словарной базы нами использован оригинальный акцентно-морфологический словарь (авторы: И. В. Жарков, О. А. Кузнецова и С. Л. Слободянюк) объемом около 130 000 лексем, что соответствует приблизительно 3,5 млн. форм слов.

Оказалось целесообразно при создании базы данных для поиска орфографических словоформ по их транскрипции преобразовать акцентно-морфологический словарь лексем в словарь словоформ, так как у большей части русских слов при словоизменении происходит модификация основы, вызываемая акцентными, морфонологическими и т. п. причинами. Таким образом, представить основу как цепочку фонем или аллофонов не представляется возможным, а это значит, что для получения некоторой формы заданного слова необходимо осуществить ряд трансформаций основы, что увеличивает время поиска словоформы в словаре приблизительно в три раза.

Каждой единице словаря словоформ поставлены в соответствие ее морфологические характеристики, как словоизменительные, так и классифицирующие. Например, для существительных это показатели рода и одушевленности, а также порядковый номер формы в классе словоизменения, позволяющий определить число и падеж.

Создание словарной базы данных происходило в несколько этапов:

1) Из акцентно-морфологического словаря был получен список орфографических форм слов с указанием места ударения и с информацией о частеречной принадлежности, классифицирующих и словоизменительных морфологических категориях;

2) Список словоформ был подан на вход программы автоматической транскрипции, и каждой словоформе была приписана ее транскрипция;

3) Каждая словоформа была разбита на открытые псевдослоги (последовательность транскрипционных знаков, заканчивающаяся на гласный либо завершающая словоформу);

4) Для каждой позиции слогов относительно ударного слога был построен индекс, ставящий в соответствие каждому слогу, отмеченному в словаре в данной позиции, множество словарных статей словаря словоформ.

5) Был создан двоичный файл, содержащий словарь словоформ, и индекс, обеспечивающий произвольный доступ к этому файлу по номеру словарной статьи.

Программа детранскрипции получает на входе транскрипцию произвольной словоформы, разбивает ее на открытые псевдослоги по тем же правилам, которые использовались при создании словарной базы (шаг 3), определяет для каждого слога его позицию относительно ударного, а затем обращается к индексам, описанным в шаге 4. Полученные множества номеров словарных статей подвергаются операции пересечения, результатом которой является множество номеров словарных статей, соответствующих словоформам, содержащим все псевдослоги, присутствующие в искомой словоформе, в тех же позициях. Из этого множества отбираются те словарные статьи, словоформы которых не содержат никаких дополнительных слогов. Полученное множество номеров словарных статей позволяет обратиться к словарю и получить список словоформ, транскрипция которых совпадает с искомой.

Примененный нами метод слоговой индексации следует считать оригинальным.

При работе с транскрипцией синтагмы, не разделенной на словоформы, потребуется также дополнительная структура данных, отражающая комбинаторные изменения фонем на стыке словоформ. Начальный и конечный псевдослоги словоформы, в зависимости от ее окружения, в общем случае могут иметь несколько вариантов реализации. Кроме того, несколько вариантов реализации произвольного слога возможны также при обработке реального потока речи, не вполне соответствующего произносительной норме, и могут использоваться также в качестве резерва для исправления ошибок в распознавании отдельных звуков. При этом возможно полное сохранение описанного представления данных при его расширении.

В случае, если в разных вариантах транскрипционной записи (скажем, в петербургской и в московской норме произношения) словоформа выглядит по-разному, соответствующая словарная статья делится на две или большее количество словарных статей, в зависимости от количества вариантов транскрипционной записи.

Описанный подход не лишен недостатков. Система детранскрипции оказывается довольно требовательной к ресурсам компьютера. Так, для хранения словаря, объем которого соответствует 60 тыс. лексем, или 2,5 млн. словоформ, требуется около 150 Мб дискового пространства. При учете в словаре пяти различных вариантов транскрипционной записи количество словоформ возрастает до 6,5 млн. Для хранения такой базы данных требуется около 320 Мб дискового пространства. Кроме того, время поиска всех орфографических вариантов одной словоформы в транскрипционной записи составляет около 0,7 секунды на компьютере с процессором Intel Celeron с тактовой частотой 333 МГц и с объемом оперативной памяти 64 Мб. Такую скорость вряд ли следует признать удовлетворительной.

В то же время, на компьютере указанной конфигурации поиск одной словоформы в транскрипционной записи в словаре объемом 3,5 млн. словоформ производится за 0,1 секунды, что вполне обеспечивает потребность системы детранскрипции текста, работающей в реальном времени. Следовательно, возможна реализация системы детранскрипции в реальном времени, рассчитанной на несколько вариантов произносительной нормы, в состав которой включено несколько отдельных словарей словоформ (по количеству вариантов нормы). Однако настройка на тот или иной вариант нормы (выбор соответствующего словаря) в этом случае производится вручную.

Вместе с тем, в ходе реализации программы детранскрипции выявлен ряд перспективных путей оптимизации организации словарной базы данных. В частности, ее физический объем, требования к ресурсам компьютера и время поиска могут быть сокращены в несколько раз, если в качестве единицы хранения — словарной статьи будут использованы не словоформы, но орфографические эквиваленты описанных выше квазислогов.

[1] Работа поддержана грантом РФФИ (№ 98-06-80431 — “Сегментация потока речи как модель взаимодействия языковых уровней ”)