УДК 519.688

О. Г. Хомицевич, С. В. Рыбин, И. М. Аничкин

ИСПОЛЬЗОВАНИЕ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ДЛЯ НОРМАЛИЗАЦИИ ТЕКСТА И СНЯТИЯ ОМОНИМИИ В СИСТЕМЕ СИНТЕЗА РУССКОЙ РЕЧИ

Исследована проблема разрешения неоднозначности прочтения различных элементов при работе системы синтеза русской речи по тексту VitalVoice. Описываются особенности использования морфологического и синтаксического анализа при расшифровке сокращений и специальных знаков, а также снятии омонимии (омографии). Данные экспериментов свидетельствуют о том, что выбранные методы позволяют правильно прочесть более 95 % сложных элементов естественного текста.

Ключевые слова: *синтез речи по тексту, синтаксический анализ, морфологический анализ, омонимия, омография, нормализация текста.*

Введение. Система автоматического синтеза речи преобразует текст, подающийся на ее вход, в звучащую речь. Необходимым этапом преобразования текста в речь является подготовка текста к тому, чтобы на его основании могла быть составлена фонетическая транскрипция, использующаяся далее для подбора необходимых звуковых элементов. Так, записи, которые не могут быть прочитаны непосредственно в том виде, в котором они встречаются в тексте (сокращения, цифры, небуквенные значки, элементы других алфавитов, например, латиницы в русском тексте), должны быть в итоге приведены к виду „полноценных“ слов русского языка. Например, в предложении *Порядка 22 % от объема полученных в 2011 г. средств — 172 тыс. долларов — были переданы 39 благотворительным организациям* элементы *22, %, 2011 г., 172, тыс, 39* не могут быть транскрибированы в исходном виде и долж-

ны быть преобразованы в слова *двадцати двух, процентов, две тысячи одиннадцатом году, сто семьдесят две, тысячи, тридцати девяти*. Кроме того, для русского языка при построении транскрипции необходима информация о месте ударения в слове, которое не обозначается на письме и должно быть определено отдельно для каждого слова в предложении.

Основной проблемой, возникающей при решении данных задач, является неоднозначность в прочтении элементов. Например, при расшифровке цифр возможен выбор количественного либо порядкового числительного; так, в вышеприведенном примере *22, 172, 39* — количественные числительные, *2011* — порядковое. Существенной проблемой для русского языка, обладающего богатой морфологией, является выбор правильной формы (падежа, рода, числа) при расшифровке элемента: *двадцать два* ***процента***, но *двадцать один* ***процент***, *двадцать* ***процентов***; ***тридцати девяти*** *благотворительным организациям*, но ***тридцатью девятью*** *благотворительными организациями*. При определении места ударения в слове источником неоднозначности является совпадение написания различных слов (омонимия, в более узком смысле — омография, т.е. совпадение написания слов, различающихся по звучанию).

Для решения проблемы выбора варианта расшифровки элемента, варианта ударения в слове и т.п. применяются различные методы анализа текста. Достаточно популярны статистические методы [1, 2], которые основаны на выявлении закономерностей в тексте путем обучения математических моделей. Недостатком таких методов является необходимость опираться на большие объемы соответствующим образом подготовленных текстов: для обучения автоматической программы расшифровки сокращений, цифр и т.п. требуется большой корпус текстов, где все такие элементы соотносятся с правильной расшифровкой, а для снятия омонимии (омографии) — корпус текстов, включающий разнообразные омографы с указанным правильным прочтением. Для русского языка получение такой текстовой базы является проблематичным.

Возможно также использовать в системе синтеза речи синтаксический и семантический анализ (парсинг) текста [3—5]. Однако полноценный разбор зачастую требует существенных вычислительных ресурсов, что нежелательно для коммерческих систем автоматического синтеза речи, которые должны работать в режиме реального времени или с опережением; к тому же именно неоднозначность многих словоформ языка вызывает наибольшие затруднения для многих синтаксических анализаторов [6].

Для расшифровки специальных обозначений и снятия омонимии при синтезе речи в системе VitalVoice используется частичный (локальный) лингвистический (морфологический и синтаксический) анализ текста, т.е. в процессе работы программы анализируется окружение конкретного слова (цифры, знака…). Дополнительное достоинство данного метода заключается в том, что алгоритм может быть сформулирован в виде контекстных правил с интуитивно понятным синтаксисом, которые содержатся в отдельных файлах, а не в программном коде, и могут оперативно редактироваться лингвистом для настройки работы системы. Экспериментальная проверка показывает, что этот метод позволяет корректно разрешить подавляющее большинство случаев неоднозначности чтения в естественном тексте на русском языке.

Расшифровка сокращений и специальных знаков. В текстах на русском языке, таких как газетные статьи, новостные сообщения, научно-популярная и художественная литература и др., встречаются различные типы специальных обозначений. В первую очередь, это сокращения и условные обозначения из различных элементов (буквы, цифры, небуквенные символы): *км, и.о., мск, Гб, м/с, м^2, C#,* а также специальные знаки: *%, °, $, №*. Помимо необходимости расшифровки трудности создает тот факт, что многие сокращения пишутся с точкой, а значит, их наличие должно быть дополнительно учтено в алгоритме деления текста на предложения.

Расшифровка сокращений и специальных знаков производится за счет анализа соседних, а также других слов предложения. Прежде всего нужно учесть семантическую неоднозначность: многие сокращения имеют разную расшифровку в зависимости от контекста, например, *м.* может обозначать „метр“ или „метро“; *ст.* — „станция“ или „статья“, или совпадать с несокращенными словами, например, *Кб* — „килобайт“ или аббревиатурой КБ („конструкторское бюро“), *им.* — „имени“ или личное местоимение. Для снятия подобной неоднозначности осуществляется поиск слова или другого элемента, ключевого для расшифровки: *2012 г.* („год“), *г. Псков* („город“), *ст. 105 УК РФ* („статья“), *ст. Москва-Сортировочная* („станция“) и т.п. Выбор правильной формы осуществляется при помощи анализа ближайшего контекста слова; основную роль играет наличие числительного слева (*1 км* „километр“*, 2 км* „километра“*, 12 км* „километров“*, 22 км* „километра“) и наличие предлога слева, в том числе перед числительным (*более 1 км* „километра“, *к 1 км* „километру“, *до ст. Бологое* „станции“).

Расшифровка цифровых записей включает в себя несколько этапов. В первую очередь выделяются специальные форматы записи, которые должны быть прочитаны определенным стандартизированным способом: телефон, дата, время, почтовый индекс и т.п. При этом анализируется вид записи (например, соответствует ли выражение стандартному виду записи типа XXX-XX-XX, XX:XX; входят ли цифры в возможный диапазон обозначения даты или времени, например, 13—30 или 60—65), а также наличие в предложении ключевых слов или словосочетаний (например, *телефон*, *мобильный*, *по московскому времени*…). Далее определяется разряд числительного (количественное или порядковое), прежде всего с помощью поиска ключевых слов, по преимуществу сочетающихся с порядковыми числительными (например, различные формы слова *год*). Следует заметить, что находящиеся в тексте римские цифры также должны быть расшифрованы как количественные или (чаще) порядковые числительные, например, *I квартал*, *Бенедикт XVI*.

Далее необходимо определить форму числительного, т.е. его падеж и (для числительных, обладающих данной категорией) род. При этом учитывается ближайший контекст числительного слева и справа: наличие предлога или другого управляющего слова слева (например, *к 23* „двадцати трем“, *до 23* „двадцати трех“, *владеет 23* „двадцатью тремя“ и т.п.), согласованного существительного или прилагательного справа (*10 пальцев* „десять“, *10 пальцами* „десятью“, *на 23 московских театральных площадках* „двадцати трех“, *10 этажа* „десятого“ и т.п.).

Снятие омонимии (омографии). Для синтеза речи наиболее важен анализ слов-омонимов, различающихся произношением (омографы), поскольку выбор между двумя омонимичными словоформами напрямую влияет на правильность синтезированного текста [5, 7]. Омографы в русском языке могут различаться ударением (например, *ст<u>о</u>ит—стои<u>т</u>*), а также наличием буквы „ё“, которая в современной орфографии чаще всего передается как *е* (*все—всё*), либо и тем и другим (*б<u>е</u>рег—бер<u>ё</u>г*).

Омонимичные словоформы могут иметь одинаковые грамматические признаки (например, *з<u>а</u>мок—зам<u>о</u>к*, *з<u>а</u>мка—замк<u>а</u>*…) либо различаться грамматическими характеристиками. В последнем случае омонимичными могут быть как различные словоформы внутри одной парадигмы (например, род.п. ед.ч.—им.п. мн.ч.: *облак<u>а</u>—<u>о</u>блака*, *стран<u>ы</u>—стр<u>а</u>ны*…), так и формы разных парадигм (например, существительное-инфинитив: *вест<u>и</u>—в<u>е</u>сти*, *проп<u>а</u>сть—пр<u>о</u>пасть*…). В случае с омонимами, одинаковыми по грамматическим характеристикам, разрешение неоднозначности может осуществляться только с помощью анализа лексического содержания предложения (ключевые слова, устойчивые выражения и т.п.). Если грамматические характеристики различаются, то можно использовать и анализ грамматического окружения слова для выбора омонима, подходящего по синтаксическому контексту. Усложняет проблему то, что омонимичные словоформы могут существенно различаться по частотности (на-

пример, *ух*а—*у*х*а*, *сорок*а—*сор*о*ка*, *кред*и*т*—*кр*е*дит*, *мо*ю—*м*о*ю*…). В таком случае зачастую становится продуктивным подход, при котором задаются специальные условия для нахождения низкочастотного омонимичного варианта, а в остальных случаях по умолчанию берется вариант с высокой частотностью.

Разрешение омонимии, как и расшифровка специальных обозначений, производится при помощи анализа контекста. На уровне индивидуальных слов-омонимов производится поиск в предложении ключевых слов или выражений. Этот этап включает анализ слов непосредственно рядом с текущим, как, например, в случае устойчивых выражений: *скрыто за семью замк*а*ми, в четырех стен*а*х*. Также анализируется состав предложения целиком, например: *Дверь была заперта на необычный зам*о*к* (ключевое слово *заперта*).

На уровне классов словоформ анализируется грамматическое окружение, т.е. выполняется поиск согласованных слов в предложении. Для формализации этого принципа были введены грамматические правила, увеличивающие условный „вес“ словоформы в зависимости от ее окружения. Правила хранятся в формализованном виде, позволяющем быстро оценивать и корректировать работу системы.

Результаты работы алгоритма лингвистического анализа. Для оценки качества лингвистического анализа в системе синтеза речи VitalVoice были проведены эксперименты по подсчету ошибок при обработке текстов. Для оценки правильности расшифровки нестандартных обозначений были взяты тексты с одного из новостных интернет-сайтов, поскольку данный тип текста содержит большое количество цифр, сокращений, специальных знаков и т.п. В ходе эксперимента был подсчитано число обозначений, неверно расшифрованных программой; результаты приведены ниже.

Расшифровка нестандартных обозначений

Слов в тексте, ед.	34235
Нестандартных обозначений в тексте, ед.	1066
Ошибок, ед	50
Ошибок, %	4,69
Правильно выполнено, %	95,31

Для оценки снятия омографии были взяты художественные тексты (произведения А.П.Чехова и Ю.В.Трифонова), поскольку они отличаются большим лексическим разнообразием. В ходе эксперимента был подсчитан процент слов-омографов, для которых было неверно определено место ударения; результаты приведены ниже.

Снятие омонимии (омографии)

Слов в тексте, ед.	37955
Омографов, ед.	2837
Ошибок, ед	113
Ошибок, %	3,98
Правильно выполнено, %	96,02

Обобщая результаты экспериментов, можно заметить, что лингвистический анализ, использующийся в системе VitalVoice, позволяет корректно разрешить неоднозначность чтения сложных элементов текста более чем в 95 % случаев. Основными источниками ошибок становятся сложные для анализа:

— случаи, когда для правильного прочтения элемента требуется анализ не только непосредственного контекста, но и дистанционных синтаксических связей. К примеру, во фрагменте: „*выбирать между 154 млрд кубометров по более низкой цене и 150 млрд по более высокой*“ второе числительное отделено несколькими другими членами предложения от относящегося к нему предлога;

— ошибочные или необщепринятые формы записи, например, *в 300-стах метрах* вместо *в 300-х метрах; 437 доллара* вместо *437 долларов*;

— формы записи, изначально не предназначенные для чтения вслух, такие как сложные цифровые записи, слова, полностью или частично замененные звездочками и т.п.

Развитие системы синтеза речи VitalVoice предполагает внедрение более глубокого синтаксического и семантического анализа текста, что позволит сократить количество ошибок, в особенности тех, которые связаны с недостаточно полным анализом предложения.

Заключение. Нормализация текста и определение правильного места ударения в слове — необходимый этап синтеза речи по тексту. Процедура морфологического и синтаксического анализа, реализованная в системе синтеза русской речи VitalVoice, позволяет выбрать корректный вариант прочтения таких элементов текста, как сокращения, цифры, специальные знаки, омографы и т.п. Как показывают эксперименты, проведенные на материале новостных и художественных текстов, точность правильного прочтения сложных элементов текста превышает 95 %.

СПИСОК ЛИТЕРАТУРЫ

1. *Taylor P.* Text to Speech Synthesis. Cambridge University Press, 2009.

2. *Sproat R.* et al. Normalization of Non-Standard Words // Computer Speech and Language. 2001. Vol. 15, N 3. P. 287—333.

3. *Allen J., Hunnicutt M. S., Klatt D*. From Text to Speech: The MI Talk system. Cambridge University Press, 1987.

4. *Liberman M. Y.* Text analysis and word pronunciation in text-to-speech synthesis // Advances in speech signal processing. 1992. P. 791—831.

5. *Иомдин Л. Л., Лобанов Б. М., Гецевич Ю. С.* Говорящий „ЭТАП“. Опыт использования синтаксического анализатора системы ЭТАП в русском речевом синтезе // Компьютерная лингвистика и интеллектуальные технологии: Матер. Междунар. конф. „Диалог“. М.: РГГУ, 2011. Вып. 10 (17). С. 669—679.

6. *Дружкин К. Ю., Цинман Л. Л.* Синтаксический анализатор лингвистического процессора ЭТАП-3: эксперименты по ранжированию синтаксических гипотез // Там же. М.: РГГУ, 2008. Вып. 7 (14). С. 147—153.

7. *Yarowsky D.* Homograph Disambiguation in Text-to-speech Synthesis // Progress in speech synthesis. 1996. P. 157—172.

Сведения об авторах

Ольга Гурьевна Хомицевич — PhD; ООО „ЦРТ“, Санкт-Петербург; старший научный сотрудник; E-mail: khomitsevich@speechpro.com

Сергей Витальевич Рыбин — канд. физ.-мат. наук; ООО „ЦРТ“, Санкт-Петербург; ведущий программист; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; доцент; E-mail: rybin@speechpro.com

Илья Михайлович Аничкин — ООО „ЦРТ“, Санкт-Петербург; старший программист; E-mail: anichkin@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.