УДК 629.73.05

РАЗРАБОТКА ИНТЕГРИРОВАННОГО ПРОГРАММНОГО КОМПЛЕКСА ОБРАБОТКИ И РАЗБОРЧИВОГО ОЗВУЧИВАНИЯ ПРОАНАЛИЗИРОВАННЫХ ТЕКСТОВЫХ ФРАГМЕНТОВ С ЕДИНИЦАМИ ИЗМЕРЕНИЯ ДЛЯ АРМ ПО ОБРАБОТКЕ ИНФОРМАЦИИ С ТЕЛЕМЕТРИЧЕСКОЙ И ЦЕЛЕВОЙ АППАРАТУРЫ КОСМИЧЕСКОГО АППАРАТА

А.Д. Карпенко, Ю.С. Гецевич, Е.С. Зеновко, Ю.С. Бородина
Объединенный институт проблем информатики НАН Беларуси, Минск

Описывается программный комплекс обработки и разборчивого озвучивания текстовых фрагментов с единицами измерения для АРМ по обработке информации с телеметрической и целевой аппаратуры космического аппарата, которая может являться составной частью программно-аппаратных средств для лабораторной отработки комплексов управления, функциональных модулей и узлов бортовой и обеспечивающей аппаратуры нано- и пикоспутников ДЗЗ. Данный комплекс наряду с управлением, приемом и обработкой данных может решать задачи отработки надежности, работоспособности и живучести нового оборудования и переподготовки специалистов аэрокосмической отрасли.

Введение

Тексты научно-технического тематического домена часто содержат количественные выражения с единицами измерения, которые представляют собой форму особо структурированной информации. Она выражается в сочетании количественного дескриптора, переданного на письме лингвистически (с помощью слов) либо математически (посредством цифр, знаков и символов), и обозначения единицы измерения. В качестве примеров приведем следующие буквенно-символьные выражения: 100 мА; 3 тыс. лет назад; +212 °F; 6,022 141; 1(37)•2013; моль-1 и т. д. Помимо сложноструктрированности, они характеризуются вариативностью форм содержания и выражения. Сразу записать правила локализации сложных выражений для всех случаев практично невозможно. Для упрощения этого процесса необходимо использовать приспособления, которые позволяют удобно корректировать уже разработанные правила и добавлять новые.

1. Общая схема процесса синтеза

Выражения с единицами измерения сложно идентифицировать и анализировать (выделить в их структуре число или числительное и название единицы) без хорошо подготовленных лингвистических словарей с описанием всех словоформ, сокращений и правил построения производных форм названий единиц. Это необходимо для правильной локализации единиц измерения. В письменной речи одна и та же единица измерения может в зависимости от использования разных шрифтов выглядеть поразному. Например, в латинской графике метр как единица измерения – m, а в кириллице – м. Поэтому для каждого языка следует создавать специальные уточнения при разработке алгоритмов идентификации.

На рис. 1 представлена общая схема процесса обработки и разборчивого озвучивания текстовых фрагментов с единицами измерения для автоматизированного рабочего места (АРМ) по обработке информации с телеметрической и целевой аппаратуры космического аппарата. Наглядно представлено, как поступающие от оператора данные в виде набора электронных текстовых символов проходят через четыре блока, прежде чем будут озвучены системой синтеза речи по тексту (СРТ).

Рис. 1. Общая схема процесса синтеза текстовых фрагментов
с единицами измерения в звуковой сигнал

2. Блок идентификации

В первом блоке (блоке идентификации) происходит поиск по всем поступившим текстовым данным именно количественных выражений с единицами измерения. Для этого необходимо реализовать соответствующие алгоритмы (рис. 2).

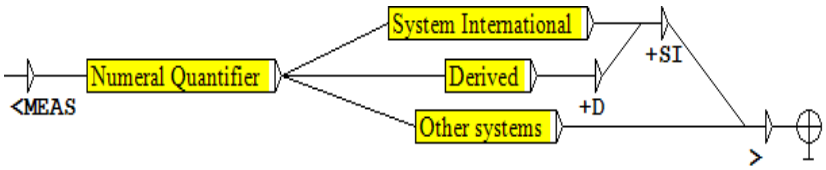

Рис. 2. Схема алгоритма для блока идентификации текстовых фрагментов
с единицами измерения

Согласно данной схеме любой текстовый набор первоначально будет проверяться на наличие количественного дескриптора (Numeral Quantifier), при этом стоит учитывать не только простые, десятичные числа и дроби в разных вариациях письменной записи, но и цифровые выражения с экспоненциальными частями. Если количественный дескриптор будет обнаружен, алгоритм начнет иные проверки через соответствующие линии-переходы на предмет наличия обозначения единицы измерения. Найденные единицы будут классифицированы согласно общей классификации измерений Международного бюро мер и весов СИ: системные (System International), производные (Derived) и внесистемные (Other systems) единицы. Также будет осуществлена соотнесенность единицы измерения с физической величиной.

3. Блок анализа

Задачей второго блока является анализ словообразовательных структур количественного дескриптора и единиц измерения. Принципиальным отличием в подходах к решению задачи идентификации и задачи анализа является необходимость примене-

ния во втором случае дополнительных морфологических компонентов, которые позволят проанализировать словообразовательную структуру.

На рис. 3. приведена схема решения задачи. Сначала введенный текст на любом языке обрабатывается посредством лингвистических ресурсов в виде словарей: каждый токен обозначается лексикографическими пометами (например, род, число, падеж, флективный класс), а далее – независимыми морфологическими (M1, M2, …, Mn) и синтаксическими (S1, S2, …, Sq) компонентами. Морфологические компоненты последовательно применяются к словоформам-токенам, чтобы проанализировать их строение и обозначить, по желанию разработчика, их особенности через специальные произвольные пользовательские маркеры. Синтаксические компоненты также позволяют применять маркеры, но уже для сочетаний слов, цифр, знаков пунктуации и других символов.

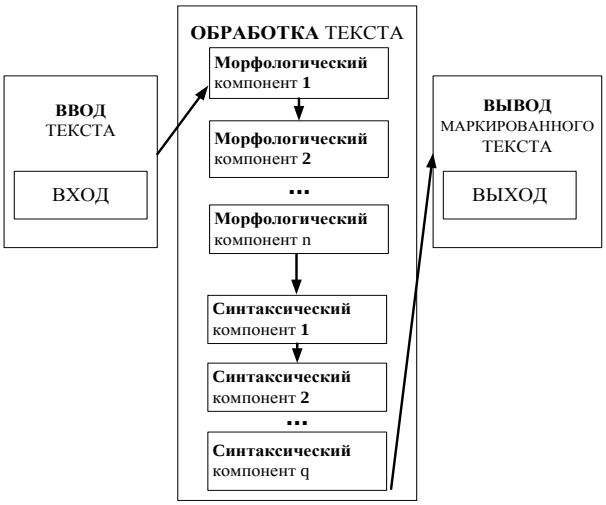

Рис. 3. Решение задачи анализа структуры элементов количественных выражений с единицами измерения

Теоретически возможны и ошибочные сращения метрологических приставок и корней едениц измерения: «микромегафарад» и т. п. Таким образом, необходимо создать именно такой блок анализа, который позволит получить не только точные, но и максимально полные результаты.

Согласно описанной выше классификации по словообразовательному принципу предвидятся четыре морфологических самостоятельных компонента, которые будут использовался в словаре и лингвистических ресурсах с префиксами: Fsubmultiple, Fmultiple, Ssubmultiple, Smultiple.

4. Блок обработки и разборчивого озвучивания количественных выражений с единицами измерения

Третий блок (см. рис. 1) нацелен на генерирование идентифицированных, классифицированных и проанализированных текстовых фрагментов с единицами измерения в орфографические последовательности слов. Особенности, которые необходимо учесть, заключаются в грамматическом устройстве беларуского и русского языков, а именно том факте, что лексические единицы, за которыми стоит количественный дескриптор, управляют родом, числом и падежом лексических единиц, которые выражают единицы измерения, не говоря уже о том, что и количественные дескрипторы также

могут подчиняться другим членам предложения. Например, для именительного падежа схематически он будет выглядеть так, как изображено на рис. 4.

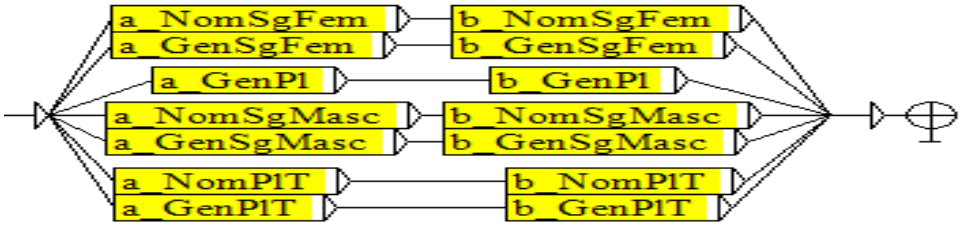

Рис. 4. Схема алгоритма обработки количественных выражений с единицами измерения

Генерирование будет происходить последовательно: от чисел (целесообразно ограничить диапазон до четырех триад, т. е. 999 999 999 999) к единицам измерения. Если число является единицей либо оканчивается на единицу, существительные в русском языке будут принимать форму именительного падежа единственного числа (например, двадцать один килограмм), кроме так называемых pluralia tantum (например, сутки), которые сохранят множественное. Таким образом, для указанной категории количественных дескрипторов и в зависимости от рода и числа единицы измерения количественное выражение попадет либо в первую ветвь (a_NomSgFem->b_NomSgFem), либо в четвертую (a_NomSgMasc->b_NomSgMasc), либо в шестую (a_NomPlT->b_NomPlT).

Заключение

После прохождения через блоки идентификации, анализа и обработки текстовые фрагменты с количественными выражениями приводятся к орфографическому представлению и подаются на вход системы синтеза речи по тексту для дальнейшего преобразования в звуковые сигналы. В то же время словарные базы системы синтеза речи по тексту должны быть соответствующим образом настроены на прием текстовых данных научно-технического тематического домена. Только в этом случае будет достигнуто разборчивое озвучивание количественных выражений с единицами измерения.

Список литературы

1. Carlson, R. A Text-to-Speech System Based Entirely on Rules / R. Carlson, B. Granstrom // Proc. of ICASSP 76. – Philadelhia, 1976. – P. 686–688.

2. Text-to-speech system wuth acoustic processor based on the instantaneous harmonic analysis / E. Azarov [et al.] // Speech and Computer : proc. of the 13-th Intern. Conf. SPECOM'2009. – SPb., 2009. – P. 414–418.

3. Апресян, Ю. Лингвистический процессор для сложных информационных систем / Ю. Апресян, И. Богуславский, Л. Иомдин. – М. : Наука, 1992. – С. 256.

4. Ахманова, О. Словарь лингвистических терминов / О. Ахманова. – КомКнига, 2007. – 576 с.

5. Информационный портал речевых технологий [Электронный ресурс]. – Режим доступа : http://www.sintezator.narod.ru. – Дата доступа : 20.12.2016.

6. Black, A. System documentation for Festival Version 1.4.0 / A. Black, P. Taylor, R. Caley // The Festival Speech Synthesis System [Electronic resource]. – Mode of access : http://www.cstr.ed.ac.uk/projects/festival/manual/. – Date of access : 02.06.2017.

7. Brailcom, o.p.s. Singing Computer. Free(b)soft [Electronic resource]. – Mode of access : http://www.freebsoft.org/singing-computer. – Date of access : 24.04.2017.