А.А. ТУРОВЕЦ
Intel Corporation

ФОРМИРОВАНИЕ БАЗОВОГО СЛОВАРЯ ДЛЯ СИСТЕМЫ РАСПОЗНАВАНИЯ СЛИТНОЙ РУССКОЙ РЕЧИ

Фонетический словарь является одним из основных языковых ресурсов, необходимых для построения системы распознавания слитной речи. При распознавании русской речи декодер должен различать несколько миллионов слов. Традиционные методы представления словаря в виде префиксного фонетического дерева приводят к недопустимому расходу памяти. Предложенный подход позволяет построить компактный русский фонетический словарь для декодера речи. Определены метрики для оценки сложности словаря. Проделана большая работа по верификации написания и произношения слов. Сейчас базовый словарь содержит более 4 000 000 словоформ.

Задача машинного распознавания речи привлекает внимание специалистов уже очень давно. Формально процесс распознавания речи можно описать следующим образом: аналоговый сигнал, генерируемый микрофоном, оцифровывается, и далее в речи выделяются фонемы, то есть атомарные объекты, из которых состоят все произносимые слова. Затем определяется, какое слово соответствует выделенному сочетанию фонем. При этом используется фонетический словарь, содержащий написания и произношения слов. Распознать слово – значит найти в этом словаре слово, наилучшим образом соответствующее произнесенному сочетанию фонем. Таким образом, разработка по возможности полного и непротиворечивого словаря для автоматического распознавателя речи является важной задачей. Поэтому, целью данной работы явилась разработка методов адаптации электронной версии книги «Грамматический словарь русского языка» А.А. Зализняка [1] для задач распознавания речи.

Работа разделена на следующие направления:

- исправление ошибок в электронной версии словаря,
- устранение избыточности в электронной версии словаря,
- разработка структуры словаря, позволяющая минимизировать размер префиксного дерева,
- сбор статистики о количестве лемм, словоформ, грамматических категорий.

Всего исходный словарь содержит более 100 000 лемм и более 4 800 000 словоформ.

В процессе работы выявлены следующие типы ошибок:

 ISBN 5-7262-0523-5. НАУЧНАЯ СЕССИЯ МИФИ-2004. Том 3

- неправильная часть речи (28 лемм)
- в винительном падеже неправильная одушевленность (44 830 лемм)
- неправильное время у причастий (509 лемм)
- два ударения в слове (125 лемм)
- неправильное ударение в слове (52 331 словоформа)
- неправильная лексика (347 словоформ)

Устранена избыточность, обусловленная следующими факторами:

- целиком одинаковые словоформы (с учетом лексики, транскрипции и морфологических признаков) (136 словоформ)
- целиком одинаковые леммы (леммы у которых все словоформы целиком совпадают) (11 лемм)
- частично одинаковые леммы (леммы у которых часть словоформ в лексике или в транскрипции отличаются друг от друга) (1053 леммы)

Минимизация размера префиксного дерева происходила в два этапа. Первый – сокращение максимального числа словоформ леммы с 369 – в исходной версии словаря до 58 в последней версии словаря путем разбиения леммы с большим числом словоформ на несколько меньших согласно значению морфологических признаков. Второй – для каждой леммы выделялась основа и поддерево окончаний, покрывающие все словоформы данной леммы. Далее, поддеревья с одинаковыми окончаниями и наборами морфологических признаков удалялись. Результирующий словарь состоит из словаря основ и словаря уникальных поддеревьев окончаний. Необходимый объем памяти для размещения словаря уменьшен более чем в 8 раз.

Список литературы

1. А.А. Зализняк. Грамматический словарь русского языка, Москва, Русский язык, 1980.

2. J.-L. Gauvain, L. Lamel. Large-Vocabulary Continuous Speech Recognition: Advances and Applications, Proceedings of the IEEE, Vol. 88, No. 8, pp 1181-1200, August 2000.

3. А.Б. Холоденко. "О построении статистических языковых моделей для систем распознавания русской речи", Интеллектуальные системы, т.6, вып.1-4, М.: МГУ.