Основные проблемы создания акустических корпусов

Пилипенко В.В.

Международный научно-учебный центр
Информационных технологий и систем
Киев
valery_pylypenko@mail.ru

Что такое корпус?

1) Это коллекция звуковых или текстовых файлов
2) Аннотация содержимого
3) Обслуживающий комплекс программ

Что это дает для распознавания и синтеза речи?

Немного истории

50-е годы

Задача: создать **устройство** распознавания речи.

Закончился созданием ЦЭВМ.

Новое качество — ПЭВМ, а теперь портативные утройства.

Теперь звуковой сигнал преобразовывается в цифру и нужен алгоритм

60-70-е годы

Задача: создать **алгоритм** распознавания речи.

Закончился применением алгоритма ДП

Новое качество — статистический подход на основе скрытых Марковских моделей

80-90-е годы

Задача: создать **программу** распознавания речи

Закончился открытым проектом HTK

Новое качество — Julius, Sphinx даже на портативных устройствах 20 тыс слов слитной речи в 2008 году

Почему тексты программ стали открытыми?

Сейчас

Задача: **научить** программу распознавания речи

Нужны **корпуса** речи и текстов

Размер речевого корпуса

100 часов речи = около 1 миллиона слов

Словарь в 100 тыс уникальных слов

Корпус Верховной рады

200 тыс секунд (55 часов) Словарь - 36 тыс уникальных слов

20 М отсчетов (100 отсчетов в секунду)

50 **фонем** X 3 состояния X 1000 смесей = 150 тыс состояний

→ 130 измерений на состояние

10 000 **трифонов** X 3 состояния X 64 смеси = 2 М состояний

→ 10 измерений на состояние

Размер текстового корпуса

Верховная рада за 19 лет наговорила около 14 М слов

Словарь - 150 тыс уникальных слов

С частотой встречаемости больше 50 - 14 тыс слов (95% текстов)

Биграммы $(14 \text{ тыс})^2 = 200$ М пар достаточно 1/10

Триграммы $(14 \text{ тыс})^3 = 3$ Т троек достаточно 1/30 или 100 Г

Основная проблема создания корпусов

Объем

Основные проблемы создания корпусов

- Организационные
- Технические
- Программистские
- Научные

Организационные

- Большой коллектив, необходимо разделение функций
- Разработка происходит в процессе создания корпуса
- Нужна автоматизация даже простейших действий
- Уровень ошибок на разных стадиях не должен превышать 1%, необходимо выявление и исправление ошибок

Технические

- 100 часов речи = 16Г, а промежуточных данных гораздо больше
- Это 200 тыс файлов — ОС не справляется
- Один цикл обучения около 1 месяца
- Распознавание КВ для одного комплекта параметров — 1 неделя

4-х ядерный компьютер, 2.4ГГц, 2Г Озу, 500Г диск

круглосуточно работает уже 3 года

Программистские

- Разработка происходит в процессе создания корпуса
- Необходимо осваивать и модифицировать новые продукты полезные для проекта (HTK, Transcriber, Julius и т.п.)
- Необходимо реализовывать и проверять известные подходы:

✓ идентификация и адаптация к диктору и каналу записи

✓ трифоны

✓ триграммы

✓ коартикуляция

Научные

Сейчас надежность - ВР 75%, ТВ 55%

- Собственные имена — 15-20% словаря
- При известном словаре + 15% надежности
- При известной лингвистической модели языка +10%
- Разброс надежности для разных дикторов 30%
- Экстралингвистические явления от 2 до 15% случаев

Наша цель

Система распознавания слитной речи многих дикторов для радио и телепередач

Надежность распознавания — не меньше 85%

Словарь — не меньше 100 тыс слов

Реальное время для 4-8 ядерного компьютера

Основные проблемы создания акустических корпусов

Спасибо за внимание !

Пилипенко В.В.

Международный научно-учебный центр
Информационных технологий и систем
Киев

valery_pylypenko@mail.ru