Створення корпусів поскладового розпізнавання реплік усного діалогу

Васильєва Ніна,
м.н.с., МНЦІТ та С

Створення НВ

Мета: створення НВ, як для розпізнавання так і для синтезу мовлення. НВ має містити якомога широкий спектр різних фонем-трифонів.

Початкові тексти для НВ

1. текстовий корпус.
2. словник УМІФ;
3. частотний словник української мови;

Текстовий корпус містить:

1. художні твори українських авторів;
2. публіцистичні твори;
3. новини;
4. історичні довідки.

Кроки виконання

1. попередня обробка текстів: видалення приміток, номерів розділів, заміна скорочень тощо;
2. перетворення орфографічного тексту в фонемний;
3. (для НВ з окремими словами – вилучення слів, які не містять голосної)
4. складання статистики по кожному із обраних початкових корпусів;
5. оброблення отриманого результату через “Жадібний алгоритм”;
6. запис отриманої навчальної вибірки.

Проблеми опрацювання

Людський фактор:

- помилково написані речення (написані правильно орфографічно, але позбавлені семантики);
- Транскиптор не враховує всіх особливостей різниці вимовляння та написання, таких як числівники, в тому числі і дати, ізольована літера і т.д.

Запис НВ:

- затрати часу;
- втома голосового тракту;
- специфіка вимови.

Статистичні дані за різними джерелами

	Загальна кількість речень/ слів до роботи “Жадібного“ алгоритму	Загальна кількість речень/ слів після роботи “Жадібного“ алгоритму	Загальна кількість реалізацій фонем-трифонів до роботи “Жадібного“ алгоритму	Загальна кількість реалізацій фонем-трифонів після роботи “Жадібного“ алгоритму	Кіль-кість фонем-трифо-нів
Текстовий корпус (709 файлів; 49 МБ)	**815 995**	**18 018**	**41 179 780**	**1 020 356**	**51 442**
Словник УМІФ	**1 874 744**	**13 706**	**23 734 396**	**120 194**	**27 772**
Частотний словник	**137 634**	**8 207**	**1 487 679**	**71 019**	**18 337**

Кількість фонем-трифонів, які зустрілися один раз

	Загальна кількість фонем-трифонів у кожній з НВ	Загальна кількість фонем-трифонів у кожній з КВ	Кількість фонем трифонів, які зустрічаються лише один раз		
			Початковий текст	НВ	КВ
Текстовий корпус	**51 442**	**9660**	**5072**	**23972**	**4264**
Словник УМІФ	**27772**	**-**	**1720**	**17372**	**-**
Частотний словник	**18337**	**-**	**2273**	**10798**	**-**

Статистика по словах

Тип вибірки	Кількість речень	Кількість слів, що зустрілися	Загальна кількість реалізацій слів
НВ	**18 018**	**47 621**	**184 910**
КВ	**2 988**	**3 227**	**8 985**

Статистика по складах

Тип складу	Кількість складів у НВ	Кількість складів у КВ	Кількість реалізацій складів у КВ
Відкритих складів	**18 134**	**2098**	**16 862**
Склади, що діляться за правилами складоподілу	**10 584**	**2 285**	**15 858**

Результати розпізнавання

Тип розпізнавання	Коректність, %	Надійність, %
Пофонемне	**69.28**	**64.05**
Поскладове, звичайний склад	**69.82**	**60.89**

Плани на майбутнє

- Записати КВ випадковим чином.
- Провести розпізнавання.
- Порівняти результати.

Дякую за увагу!