Автоматичне розбиття текстів на тематики для розпізнавання українського мовлення з телевізійних новин

Н.Б. Васильєва, В.В. Пилипенко, В.В.Яценко

Міжнародний науково-навчальний центр інформаційних технологій та систем
просп. Академіка Глушкова 40, Київ 03680
{n.vassilleva; valeriy.pylypenko; yatsenko.valya} @gmail.com

Сучасні системи розпізнавання потребують формування словника, від обсягу якого залежать показники системи, такі як точність розпізнавання та швидкодія. При збільшенні кількості слів в словнику зменшується кількість помилок за рахунок більшого охоплення мовлення. Але значно зростають вимоги до комп'ютерних ресурсів: потужності обчислень, пам'яті та інших. При розпізнаванні мовлення використовується статистична лінгвістична модель (ЛМ), яка описує породження мовлення. При збільшенні словника значно зростає обсяг потрібних текстів для побудови ЛМ.

Існує гіпотеза що, мовлення розділяється на окремі теми, які мають відносно обмежений словник та множину текстів, які належать до теми.

Описується інформаційна технологія розпізнавання мовлення за допомогою тематичних ЛМ на прикладі телевізійних новин. При побудові системи розпізнавання спочатку проводиться розділення текстів для ЛМ на теми та побудова окремих специфічних словників для кожної теми. При розпізнаванні використовується двопрохідний підхід, де на першому кроці шукаються ключові слова для всіх тем та визначається тема повідомлення. На другому проході відбувається розпізнавання мовлення з використанням визначеної тематичної ЛМ.

Для формування корпусів текстів використовуються ресурси Інтернет. Обсяг корпусу текстів склав близько 2ГБ. Розроблено алгоритм автоматичного віднесення текстів до теми, наводяться приклади роботи алгоритму. Усі тексти розподілялися на 12 тематичних кластерів та додатковий кластер (інше). Для кожного кластеру текстів формується словник, який складається з двох частин: словник специфічний для кластеру та словник загальновживаних слів Української мови.

Акустична модель будувалась по звуковим файлам записаним з каналу *NewsOne* тривалістю 60 годин.

Таблиця показує середню точність розпізнавання звукових файлів, які належать до теми ПОДІЇ, з використанням різних ЛМ. Для ЛМ ПОДІЇ досягнута найкраща точність розпізнавання.

Тема ЛМ	Послівна точність, %
досягнення	54.22
економіка	76.04
культура	75.47
київ	56.06
інше	79.95
події	81.44
політика	77.71
релігія	57.73
спорт	64.40
суд	64.01
тварини	44.01
явища	68.04
здоров’я	55.21

Розроблено та експериментально перевірено технологію автоматичного розбиття текстів на тематики для побудови специфічних тематичних лінгвістичних моделей. Для найбільш представлених тематик досягнута найкраща точність розпізнавання мовлення. Сумарний обсяг словників системи розпізнавання становив біля 300 тис. слів.