Кластеризація екстралінгвістичних явищ на прикладі мовленнєвих записів Верховної Ради України

Пилипенко Валерій Васильович, Гузієнко Ірина Віталіївна

Відділ розпізнавання та синтезу звукових образів
МННЦ Інформаційних технологій та систем
valeriy.pylypenko@gmail.com, irinaguzienko@mail.ru

Анотація

Дана стаття присвячена проблемі розпізнавання злитого мовлення зі словником, що містить екстралінгвістичні явища.

В статті наводиться опис математичних моделей, обгрунтування вибору та опис методу кластеризації екстралінгвістичних явищ для забезпечення ефективного розпізнавання злитого мовлення, наведено результати роботи програмного комплексу.

Система розпізнавання злитого мовлення базується на програмному комплексі НТК і виконує перетворення існуючих словників і граматик, вносячи в них мета символи, що відповідають заданим екстралінгвізмам.

1. Вступ

Варто відзначити, що в реальній комунікації переважає спонтанне мовлення, іншими словами, воно є основним, домінуючим в порівнянні з підготовленою усною промовою. Усвідомлення цього факту і призвело до виникнення дослідницького інтересу до спонтанної мови як особливого лінгвістичного феномену. Під спонтанним мовленням розуміється мова непідготовлена, здійснювана мовцем в постійно (іноді щохвилини) змінюваних комунікативних умовах. Стан мовця відображають характеристики його голосу, які стосуються екстралінгвістичних явищ.

Екстралінгвістичними явища - це включення в мову таких особливостей спілкування, як «е» -кання, «а» -кання, «н» -кання, «м» -кання та інших, а також різного роду психофізіологічних проявів людини: плач, кашель, сміх, вдих і т.д.

Загальновідомо, що мовні відхилення від мовної норми при розпізнаванні визначаються як помилки в розпізнаванні. Такі помилки, що виникають в випадку спонтанного мовлення, пов'язані з тимчасовим ослабленням уваги і контролю над промовою в процесі мовленнєвої діяльності. Вони не мають постійного систематичного характеру. Всі відхилення від норми в спонтанній мові, обумовлені поточною ситуацією і залежать від емоційно-психологічного стану мовця.

Відзначимо, що ефективність розпізнавання підготованого тексту (наприклад, новин), є достатньо високою (близько 95%) [1], але непідготований аудіо сигнал (тобто такий, що містить шуми) розпізнається на порядок гірше. Ще більше погіршує ефективність розпізнавання наявність у мовленні екстралінгвістичних явищ, тобто неінформативних звуків.

Метою нашої роботи є підвищення ефективності розпізнавання злитого мовлення за рахунок кластеризації екстралінгвістичних явищ, що здійснюється на основі аналізу попередніх результатів розпізнавання.

2. Опис роботи програмного комплексу

Робота системи полягає в модифікації її базових словників і граматик для поліпшення розпізнавання аудіосигналів, що містять екстралінгвістичні явища.

Першим етапом обробки є ручна модифікація словників і граматик з метою внесення до них даних про екстралінгвізми, а також створення файлів розмітки аудіо сигналу для навчання системи розпізнавання. Дані файли використовуються для навчання ПММ та отримання файлів результату. Загальна схема роботи програмного комплексу показана на рис 1.

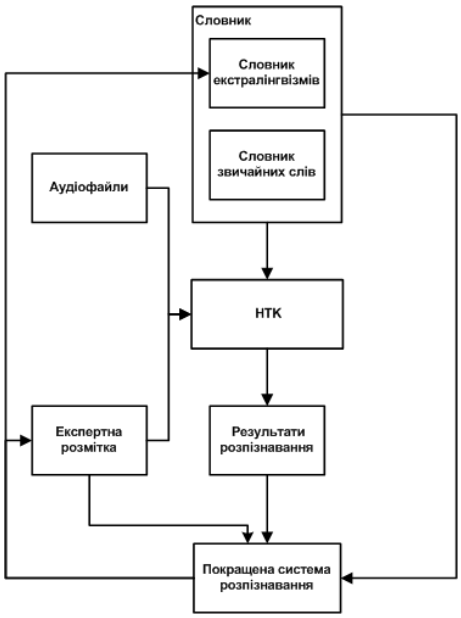

Рис. 1. Загальна схема роботи програмного комплексу

3. Моделі екстралінгвістичних явищ

Розпізнавання мовлення відбувається на основі побудови прихованих марковських моделей(ПММ). Особливістю роботи ПММ з екстралінгвізмами є те, що екстралінгвізм являє собою одну фонему, причому в загальному випадку не існує ніяких обмежень на їх розміщення. Екстралінгвістичні явища можуть розташовуватися в довільному місці промови. Вони можуть з'являтися всередині ланцюжків слів, що не дозволяє коректно їх розпізнавати.

Основний підхід для роботи з екстралінгвістичними явищами полягає в поданні їх як спеціальних слів зі словника (табл. 1).

Табл. 1. Словник екстралінгвістичних явищ

Слово/Фонема	Опис(пояснення)
е	*е*-кання
а	*а*-кання
пл	Плямкання
вд	вдих/видих
к	Кашель
н	*н*-кання
м	*м*-кання
…	

Модель цих явищ є простою, оскільки ми розглядаємо тільки ті екстралінгвізми, які зустрічаються між словами і фразами. Для даного словника (табл. 1.) ПММ матиме наступний вигляд (рис. 2.):

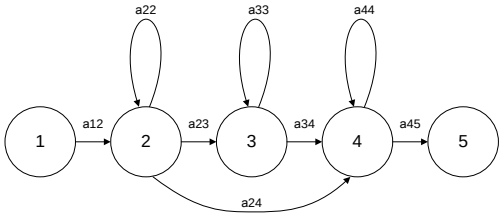

Рис. 2. ПММ для розпізнавання екстралінгвістичних явищ

Використання ПММ для розпізнавання цих явищ таке ж як і у випадку розпізнавання злитого мовлення, оскільки ми розглядаємо окремі екстралінгвістичні явища, як слова. Після цього відбувається:

- Перевірка наскільки правильно були розпізнані екстралінгвізми;
- Автоматичне визначення точності розпізнавання;
- Аналіз помилок розпізнавання екстралінгвізмів;
- Модифікація вхідних словників і граматик;
- Повторення процесу розпізнавання екстралінгвістичних явищ кілька разів.

4. Обгрунтування обраного методу

Як показують результати, екстралінгвізми є неточними фонемами і спроба відрізнити один екстралінгвізм від іншого може бути потенційним джерелом помилок і може вплинути на процес розпізнавання основного тексту. Тому з метою підвищення ефективності розпізнавання спонтанної мови доцільно скоротити кількість категорій екстралінгвістичних явищ, об'єднавши категорії схожих за звучанням екстралінгвізмов в одну. Таке об'єднання було вирішено здійснити за допомогою методів кластерного аналізу.

Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя [2]. Кластерний аналіз — це багатовимірна статистична процедура, яка виконує збір даних, що містять інформацію про вибірку

об'єктів і потім упорядковує об'єкти в порівняно однорідні групи — кластери.

Алгоритм кластеризації — це функція a:X→Y, яка будь-якому об'єкту x∈X ставить у відповідність номер кластера y∈Y. Множина Y в деяких випадках відома заздалегідь, проте частіше ставиться завдання визначити оптимальне число кластерів, з погляду деякого критерію якості кластеризації.

Об'єднання схожих об'єктів у групи може бути здійснене різними способами. Саме для цього етапу існує цілий ряд методів: метод просіювання, нейронна мережа Кохонена, метод К-середніх, нечітка кластеризація С-середніх, графові алгоритми кластеризації, ієрархічна кластеризація, та інші. В нашому випадку об'єднання категорій екстралінгвізмів було вирішено реалізувати таким методом як ієрархічна кластеризація.

Ієрархічна кластеризація виходить з того, що об'єкти з заданої множини характеризується певним ступенем близькості. Близькість задається функцією близькості, яка, в загальному випадку, є оберненою до функції відстані $\rho(x,x')$, тобто $\frac{1}{\rho(x,x')}$. Часто для зручності на основі функції близькості формують матрицю близькості.

Для кластеризації по матриці близькості використовується метод найближчого сусіда: об'єкти, що мають найближчу відстань, об'єднуються в один кластер. Для даного методу існує запропонована А.Н. Колмогоровим формула:

$$K_\eta([i,j],k) = \left[\frac{n_i K(i,k)^\eta + n_j K(j,k)^\eta}{n_i + n_j}\right]^{\frac{1}{\eta}}, -\infty \leq \eta \leq +\infty$$

де $[i,j]$ — група з двох об'єктів (кластерів) i та j;

k — об'єкт (кластер), з яким шукається схожість зазначеної групи;

n_i — кількість елементів у кластері i;

n_j — кількість елементів у кластері j.

5. Кластеризація екстралінгвізмів

Кластеризація відбувається на основі попередніх результатів аналізу розпізнавання мови. В ході аналізу частина екстралінгвізмів розпізнається неправильно - як інші екстралінгвізми.

Для початку кластеризації необхідно нормалізувати таблицю помилок, привівши всі значення в відсотковий вигляд. Для цього необхідно поділити кількості розпізнаних символів на загальну кількість таких символів у вхідному потоці. Отримана таблиця являє собою прообраз матриці близькості.

Основною проблемою в даній таблиці є її несиметричність, що, по суті, означає несиметричність відповідної функції ρ:

$$\rho(x,x') \neq \rho(x',x) \qquad (5.1)$$

Щоб виправити цей недолік, модифікуємо функцію:

$$\rho^*(x,x') = \max(\rho(x,x'),\rho(x',x)) \qquad (5.2)$$

В результаті отримаємо остаточну матрицю близькості.

Далі по цій матриці виконується кластеризація за методом найближчого сусіда.

При кластеризації необхідно контролювати кількість кластерів. Оскільки при старті алгоритму їх кінцева кількість невідома, слід задати порогові значення близькості, які не дозволятимуть об'єднувати недостатньо близькі екстралінгвізми. Даний поріг φ визначається емпірично для заданої вибірки. Тепер функція близькості модифікується наступним чином:

$$\rho^{**}(x,x') = \begin{cases} \rho^*(x,x'), \rho^*(x,x') \geq \varphi \\ 0 \qquad\quad , \rho^*(x,x') < \varphi \end{cases} \qquad (5.3)$$

Отже, проаналізувавши результати розпізнавання екстралінгвістичних явищ отримуємо матрицю близькості. На основі цієї матриці робимо кластеризацію екстралінгвізмів і, відповідно, об'єднання моделей екстралінгвістичних явищ. Така методика дозволить підвищити ефективність системи розпізнавання.

6. Опис алгоритму кластеризації

На першому кроці алгоритму кластеризації будується таблиця близькості. Вона буде отримана шляхом паралельного проходження по двом файлам розмітки і порівнянні часових міток. Якщо в обох файлах

з невеликим інтервалом часу (в районі 500 мс) знаходиться екстралінгвізм (не обов'язково один і той самий) то екстралінгвізм вважається розпізнаним (в тому розумінні, що система ідентифікувала його як неінформативний звук, а не як інформативне слово). Складається таблиця за наступним правилом: якщо у вхідному файлі в певній позиції стоїть і-й екстралінгвізм, а в вихідному файлі на цьому місці стоїть j-й, то елемент a_{ij} таблиці близькості збільшується на одиницю. Результуюча таблиця нормалізується за наступним правилом:

$$a^*_{i,j} = \frac{a_{i,j}}{\sum_k a_{i,k}}$$

Нормалізована таблиця приводиться до симетричної форми:

$$a^{**}_{i,j} = a^{**}_{j,i} = \max(a^*_{i,j}, a^*_{j,i})$$

До таблиці застосовується поріг φ:

$$a^{***}_{i,j} = \begin{cases} a^{**}_{i,j}, a^{**}_{i,j} \geq \varphi \\ 0 \quad , a^{**}_{i,j} < \varphi \end{cases}$$

Над отриманою таблицею $B = A^{***}$ виконується кластерний аналіз методом найближчого сусіда.

На кожному кроці визначаються два найближчі сусіди:

$$i, j, \text{такі що } b_{i,j} = \max_{t \neq k} b_{t,k}$$

Відповідні групи екстралінгвізмів вважаються схожими, і тому об'єднуються. Відповідні стовпчики і колонки видаляються з таблиці, і замість них вводиться один, який містить сумарну близькість двох об'єднаних категорій екстралінгвізмів.

Якщо в ході кластеризації виникла ситуація, що якась колонка чи стовпець містить лише одне число – на головній діагоналі, то це означає що відповідний кластер не містить екстралінгвізмів, що схожі на певні екстралінгвізми поза групою. Тоді даний кластер можна зразу виключити з процесу кластеризації і перенести в список готових кластерів.

Доцільним також є перед початком кластеризації перевірити всі існуючі екстралінгвізми. Якщо якісь з екстралінгвізмів не мають схожості з іншими, то їх можна виключити з таблиці ще до початку кластеризації. Для них створюються спеціальні кластери, що складаються з одного елементу. Ці кластери зразу заносяться в список готових кластерів, а відповідні колонки і строки таблиці схожості видаляються.

Після завершення кластеризації отримаємо завершений список готових кластерів. З нього видаляються всі одноелементні кластери, оскільки вони не внесуть ніяких змін в файл розмітки.

7. Результати експериментів

В результаті кластеризації отримуємо такі групи кластерів:

- Екстралінгвізми * е *, * а *, * і * об'єднуємо в кластер № 1;
- Екстралінгвізми * вд * і * пл * об'єднуємо в кластер № 2;
- Ексралінгвізми * м, * к *, * і *, * н, * ук * не ввійшли ні в один із кластерів, тому ми кожен з них залишаємо в окремому кластері з відповідною назвою.

В вихідному файлі розмітки категорії екстралінгвізмів замінюються сформованими кластерами, тобто кластер з екстралінгвізмів *е*, *а*, *и* замінюється мета символом *еаи*, а кластер з екстралінгвізмів *вд* і *пл* - *вд_пл*.

8. Висновки

Після повторного навчання НТК збільшився відсоток правильно розпізнаних екстралінгвізмів.

Алгоритм кластеризації дозволив зменшити кількість розрізнюваних екстралінгвізмів і збільшити точність розпізнавання мовлення вцілому.

9. Літературні джерела

[1] M. Gales and S. Young. "The Application of Hidden Markov Models in Speech Recognition." 2007, 1(3).

[2] Джордж Ф. Люггер. Искусственный интеллект. Стратегии и методы решения сложных проблем. Москва, Санкт-Петербург, Киев – 2003.