Декомпозиція N-грамної моделі мовленнєвого потоку для задачі розпізнавання зв’язного мовлення

Мета роботи

Аналіз N-грамної моделі мовленнєвого потоку для задачі розпізнавання зв’язного мовлення для віднайдення способу покращити результати розпізнавання

Розпізнавання зв'язного мовлення

Баєсівська інтерпретація задачі: визначення яка послідовність слів відповідає заданому мовленнєвому сигналу. Вона починається з розгляду усіх можливих класів, до яких ми можемо віднести вхідну звукову послідовність.

Баєсівська формула

$$\hat{w} = \arg\max_{w \in V} P(O \mid w) P(w)$$

$P(w)$ називають апріорною ймовірністю, а - $P(O \mid w)$ достовірністю

Мовна модель

При розпізнаванні зв'язного мовлення цей термін зазвичай використовується для позначення статистичної моделі словесних послідовностей.

Словоформа та Лема

Терміном **словоформа** позначають різні форми слова, на приклад, «кішка», «кішкою», «кішкам», «кішки» - це різні словоформи, що належать до одного абстрактного класу еквівалентності (у даному випадку, до класу {всі можливі форми слова «кішка»}) такий клас еквівалентності називатимемо **лемою** (lemma)

N-грама

N-грама – послідовність словоформ фіксованої довжини, з порядком елементів, що визначається порядком словоформ в якомусь тексті.

Порядок N-грами – її довжина в словоформах

Імовірність N-грами

Позначатимемо $w_1 \ldots w_n$ або w_1^n

Ланцюжок слів, що утворює N-граму тоді імовірність $P(w_1, w_2, \ldots, w_{n-1}, w_n)$

Можна виразити так:

$$P(w_1^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)\cdots P(w_n \mid w_1^{n-1}) = \prod_{k=1}^{n} P(w_k \mid w_1^{k-1})$$

Про навчальні та тестові множини

N-грамна модель має суттєвий недолік. Оскільки вона будується на основі випадкового процесу, то результуючі імовірності гарантовано містять шум. Їх необхідно згладити.

Розпізнавання зв'язного мовлення

Основною ідеєю розпізнавання є те, що побудована на корпусі та згладжена n-грамна модель дозволяє для акустичного входу визначити ланцюжок словоформ, що з найбільшою імовірністю відповідає вхідному сигналу.

Проблеми україномовного розпізнавання

- Українська мова належить до мов з переважно вільним порядком слів;
- Слова української мови мають багато словоформ.

Разом це перш за все призводить до побудови поганих моделей.

Ефективність алгоритмів

Сучасні алгоритми, як, наприклад, Вітербі або A^* є потужними, надійними та не допускають покращення без докорінної зміни методики побудови результатів розпізнавання зв’язного мовлення. Єдиним способом покращити результати є зміна способу побудови моделі мови.

Спосіб розв'язання проблеми

Пропонується декомпонувати модель на дві:

- модель лематичних n-грам;
- модель n-грам побудованих на граматичних ознаках.

Кожна з цих моделей матиме суттєво меншу кількість елементів, а відтак в середньому більшу частоту і достовірність кожного елементу.

Практична перевірка.

- Запущено декілька тестів, де кожна модель діяла як незалежний розпізнавач.
- Результати розпізнавачів порівнювались.
- Вибирався кращий.

HTK Results Analysis (Raw)

- Ref : word_2002_09_27.mlf
- Rec : res\answ_2002_09_27_.mlf
- SENT: %Correct=31.38 [H=322, S=704, N=1026]
- WORD: %Corr=65.36, Acc=48.81 [H=2551, D=134, S=1218, I=646, N=3903]
- Або 1993 неправильних слова.

HTK Results Analysis (Mod)

- Rec : res\answ_2002_09_27_.mlf
- SENT: %Correct=29.43 [H=302, S=724, N=1026]
- WORD: %Corr=63.31, Acc=46.45 [H=2471, D=141, S=1291, I=658, N=3903]
- Або 2086 неправильних слова.

Вибір кращого.

- Для базових N-грам:

1024 avg:1917.0 best:1917 worst:1917

Correct Sent =31.4453125

1917 неправильних слів.

- Для модифікованих N-грам:

1024 avg:1937.0 best:1937 worst:1937

Correct Sent =30.859375

1937 неправильних слів

Висновки

- В роботі отримано суперечливі результати, що з одного боку підтверджують тезу про ефективність застосування декомпозиції та рекомпозиції для побудови нових моделей мови.
- З іншого боку є певні незрозумілі моменти, що вимагають подальшого дослідження.

ДЯКУЮ ЗА УВАГУ.

Згладжування

$$c^* = \frac{(c+1)\frac{N_{c+1}}{N_c} - c\frac{(k+1)N_{k+1}}{N_1}}{1 - \frac{(k+1)N_{k+1}}{N_1}}, for \quad 1 \leq c \leq k$$

Гуда - Тьюринга

Де:

k – емпіричний коефіцієнт, звичайно < 6

N_i – кількість n-грам, що зустрілися i разів.

Згладжування

$$p_i^* = \begin{cases} \frac{T}{Z} \cdot \frac{N}{N+T}, \text{if } c_i = 0 \\ c_i \frac{N}{N+T}, \text{if } c_i > 0 \end{cases}$$

Віттена - Белла

Де:

T – кількість класів N-грам, що зустрілися у тексті;

N – кількість токенів;