УДК 004.93

Тарануха В.Ю., асистент

Модифікація n-грамної моделі засновані на класах для розпізнавання слов'янських мов

Стаття присвячена дослідженню можливої модифікації n-грамної моделі заснованої на класах для слов'янських мов в задачі розпізнавання мовлення

Ключові слова: n-грамна модель мови, модель на класах

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр-т. Глушкова 4д, e-mail: *taranukha@mail.ru*

V.Y. Taranukha, assistant

Modification of class-based n-gram model for slavic language

The article investigates class-based n-gram language model for slavic language recognition

Key words: n-gram language model, class-based model

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova av., 4d, e-mail:*taranukha@mail.ru*

Статтю представив чл.-кор. НАН України, д.ф.-м.н., проф. Анісімов А.В.

Завдяки розвитку ЕОМ та мережі Інтернет об'єм даних доступних у вигляді цифрових аудіо записів весь час зростає, і це викликає потребу в нових засобах обробки даних, включаючи переведення аудіо записів у текст. Широко вживається модель, на основі n-грам[1], за допомогою якої зручно оцінювати ймовірність появи ланцюжка з *n* слів у деякому тексті. Проте при застосуванні такої моделі до слов'янських мов, зокрема до української, проявляється ряд недоліків, для боротьби з якими пропонується внести зміни в метод побудови моделі.

Традиційний підхід до побудови моделей

Передбачається, що мова може бути описана за допомогою марківського ланцюжка, де станами будуть слова мови. Послідовність слів мови $w_1 \ldots w_n$ називається n-грамою довжини n. Для зручності її позначають w_1^n.

Тоді, послідовність слів можна представити як послідовність n-грам, а імовірність оцінити за формулою:

$$p(w_1^i) = p(w_i \mid w_1^{i-1}) p(w_{i-1} \mid w_1^{i-2}) \ldots p(w_1) \qquad (1)$$

Відповідні імовірності обчислюються за оцінкою максимальної правдоподібності на базі частот n-грам у корпусі текстів.

Позначимо $C()$ - частота відповідної n-грами(яка може бути рівна 0).

$$\hat{p}(w_i \mid w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i})}{C(w_{i-n+1}^{i-1})} \qquad (2)$$

При цьому виконується рівність:

$$C(w_{i-n+1}^{i-1}) = \sum_{j=0}^{|V|} C(w_{i-n+1}^{i-1} w_j), \qquad (3)$$

Оцінка (2) будується на припущенні, що в корпусі переставлені всі необхідні n-грами, а спостережувані в корпусі частоти являються хорошими наближеннями істинних частот n-грам у мові. Насправді, це не завжди так, і для розв'язання цієї проблеми використовуються методи для згладжування частот та імовірностей[1] відповідних n-грам.

Необхідно зазначити, що в слов'янських мовах характерною рисою є вільній порядок слів у реченнях. Для забезпечення цього в словоформі відповідного слова часто зберігається інформація, що вказує на потенційні синтаксичні зв'язки словоформи. Що в свою чергу призводить до збільшення числа словоформ відносно романо-германських мов.

На одне слово англійської мови припадає приблизно 1,7 словоформи, в той час, як на одне слово української мови може припасти від 5,5 до 19,9 залежно від вибраного словника, на одному і тому самому корпусі. Таким чином, при побудові таблиці n-грам, при *n*=2, розмір зростає

принаймні в 10,47 рази , а при n=3 – в понад 33 рази. Крім всього іншого це призводить до того, що значна кількість n-грам набуває малих значень частот, отже відповідно, оцінка імовірностей по формулі (2) стає значно чутливішою до викидів та шумів.

Зручним способом зменшити розмірність словника системи є застосування n-грам заснованих на класах[2]. Словоформи розбиваються на класи відповідно до того, в яких контекстах вони стоять.

Вводиться функція розбиття, що ставить у відповідність кожному слову w_i з словника V клас c_i. При цьому виконується:

$$P(w_i \mid w_1^{i-1}) = P(w_i \mid c_i)P(c_i \mid c_1^{i-1}), \forall i, 1 \le i \le n \quad (4)$$

Цей підхід суттєво зменшує розмірність словника системи, розмір таблиць і певною мірою мав би покращувати розпізнавання. Проте оцінка (3) виконується досить наближено, і таким чином при економії на обчислювальних видатках загальна якість розпізнавання не підвищується.

Альтернативний підхід до побудови частот n-грам для слов'янських мов

Пропонується попередньо побудувати дві окремі моделі n-грам, окремо на основі канонічних форм слів, окремо на основі граматичних класів слів. Тобто, одна словоформа дає інформацію в дві різні часткові моделі. Аналогічний підхід розглядався[3,4], проте пропонується розглянути суттєво відмінний метод оцінювання рядка на основі отриманих моделей порівнювано з [3], оскільки він передбачає побудову об'єднаної моделі, та не вживається фільтрація як у [4].

Для забезпечення достовірності об'єднаної моделі необхідно щоб сума частот канонічних форм та сума частот граматичних класів після перерозподілу лишалася незмінною.

Позначимо: $L(w_1^k)$ – сукупність послідовностей канонічних форм для послідовності слів w_1^k. $G(w_1^k)$ – сукупність послідовностей граматичних класів для послідовності слів w_1^k.

Позначимо $El(w_1^k)$ – сукупність послідовностей слів, що після приведення до канонічних форм мають однаковий запис, тобто сукупність $w_{i_1}^{i_k}$, таких що, $L(w_{i_1}^{i_k}) = L(w_1^k), \forall i$.

Позначимо $G(El(w_1^k))$ – сукупність $G(w_{i_1}^{i_k})$, таких що є послідовностями граматичних класів відповідних $w_{i_1}^{i_k} \in El(w_1^k)$

Тоді оцінка частоти w_1^k визначається:

$$C(w_1^k) = \frac{\mathrm{C}(L(w_1^k))\mathrm{C}(G(w_1^k))}{\sum\limits_{G_F \in G(El(w_1^k))} G_F} \quad (5)$$

На її основі за формулою (2) можна будувати імовірності, а отже і оцінювати імовірність тієї чи іншої послідовності слів звичайним чином. Це робіть формулу (5) формулою згладжування на основі зарані визначених граматичних класів.

При цьому виконання формули (3) є умовою коректності формули для реалізації моделі мови.

Розглянемо редуковану мову отриману з початкової шляхом злиття словоформ. Редукована мова має два граматичні класи та два класи канонічних форм.

Таблиця 1.

Повністю сумісні класи

	g_1	g_2
l_1	w_1	w_2
l_2	w_3	w_4

Лема 1. Для повністю сумісних класів після перерозподілу через формулу (5) виконується залежність (3)

Доведення. Розглянемо для $C(w_1)$

$$C(w_1) = \frac{C(L(w_1))C(G(w_1))}{\sum\limits_{G_F \in G(El(w_1))} C(G_F)} = \frac{C(l_1)C(g_1)}{C(g_1)+C(g_2)},$$

з іншого боку

$$C(w_1) = \sum_j C(w_1 w_j) = \frac{C(L(w_1 w_1))C(G(w_1 w_1))}{\sum\limits_{G_F \in G(El(w_1 w_1))} C(G_F)} +$$

$$\frac{C(L(w_1 w_2))C(G(w_1 w_2))}{\sum\limits_{G_F \in G(El(w_1 w_2))} C(G_F)} +$$

$$\frac{C(L(w_1 w_3))C(G(w_1 w_3))}{\sum\limits_{G_F \in G(El(w_1 w_3))} C(G_F)} +$$

$$\frac{C(L(w_1 w_4))C(G(w_1 w_{14}))}{\sum\limits_{G_F \in G(El(w_1 w_4))} C(G_F)};$$ врахуємо, що (3) тут і надалі виконується для граматичних класів та канонічних форм за замовчуванням

$$\sum_{G_F \in G(El(w_1 w_1))} C(G_F) = C(g_1) + C(g_2) =$$

$$= \sum_{G_F \in G(El(w_1 w4))} C(G_F) = \sum_{G_F \in G(El(w_1 w_3))} C(G_F) = \sum_{G_F \in G(El(w_1 w_2))} C(G_F);$$

тому

$$\sum_j C(w_1 w_j) = \frac{C(g_1 g_1)C(l_1 l_1) + C(g_1 g_2)C(l_1 l_1)}{C(g_1) + C(g_2)} +$$

$$\frac{C(g_1 g_1)C(l_1 l_2) + C(g_1 g_2)C(l_1 l_2)}{C(g_1) + C(g_2)} = \frac{C(l_1)C(g_1)}{C(g_1) + C(g_2)},$$

що і треба було довести.

Для $C(w_4)$ аналогічно, бо в таблиці можна поміняти елементи w_1 і w_4 місцями, перепозначити g_1 як g_2 і навпаки, також l_1 як l_2 і навпаки, при цьому вигляд таблиці не зміниться.

Розглянемо для $C(w_2)$

$$C(w_2) = \frac{C(L(w_2))C(G(w_2))}{\sum_{G_F \in G(El(w_2))} C(G_F)} = \frac{C(l_1)C(g_2)}{C(g_1) + C(g_2)},$$

з іншого боку

$$C(w_2) = \sum_j C(w_2 w_j) = \frac{C(L(w_2 w_1))C(G(w_2 w_1))}{\sum_{G_F \in G(El(w_2 w_1))} C(G_F)} +$$

$$\frac{C(L(w_2 w_2))C(G(w_2 w_2))}{\sum_{G_F \in G(El(w_2 w_2))} C(G_F)} + \frac{C(L(w_2 w_3))C(G(w_2 w_3))}{\sum_{G_F \in G(El(w_2 w_3))} C(G_F)} +$$

$\frac{C(L(w_2 w_4))C(G(w_2 w_4))}{\sum_{G_F \in G(El(w_2 w_4))} C(G_F)}$; аналогічно як для $C(w_1)$

$$\sum_{G_F \in G(El(w_1 w_2))} C(G_F) = C(g_1) + C(g_2) =$$

$$= \sum_{G_F \in G(El(w_4 w_2))} C(G_F) = \sum_{G_F \in G(El(w_3 w_2))} C(G_F) = \sum_{G_F \in G(El(w_2 w_2))} C(G_F);$$

тому

$$\sum_j C(w_2 w_j) = \frac{C(g_2 g_1)C(l_1 l_1) + C(g_2 g_2)C(l_1 l_1)}{C(g_1) + C(g_2)} +$$

$$\frac{C(g_2 g_1)C(l_1 l_2) + C(g_2 g_2)C(l_1 l_2)}{C(g_1) + C(g_2)} = \frac{C(l_1)C(g_2)}{C(g_1) + C(g_2)},$$

що і треба було довести.

Для $C(w_3)$ аналогічно, бо в таблиці можна поміняти елементи w_1 і w_4 місцями, перепозначити g_1 як g_2 і навпаки, також l_1 як l_2 і навпаки, при цьому вигляд таблиці не зміниться.

Таким чином, показано, що залежність (3) виконується для всіх елементів таблиці, при використанні перерозподілу (5).

Таблиця 2.

Повністю не сумісні класи

	g_1	g_2
l_1	w_1	
l_2		w_4

Лема 2. Для повністю не сумісних класів після перерозподілу через формулу (5) виконується залежність (3).

Доведення. Розглянемо для $C(w_1)$,

$$C(w_1) = \frac{C(L(w_1))C(G(w_1))}{\sum_{G_F \in G(El(w_1))} C(G_F)} = \frac{C(l_1)C(g_1)}{C(g_1)} = C(l_1),$$

з іншого боку

$$C(w_1) = \sum_j C(w_1 w_j) = \frac{C(L(w_1 w_1))C(G(w_1 w_1))}{\sum_{G_F \in G(El(w_1 w_1))} C(G_F)} +$$

$\frac{C(L(w_1 w_4))C(G(w_1 w_4))}{\sum_{G_F \in G(El(w_1 w_4))} C(G_F)}$; враховуючи, що

$$\sum_{G_F \in G(El(w_1 w_1))} C(G_F) = C(g_1 g_1), \quad \sum_{G_F \in G(El(w_1 w_4))} C(G_F) = C(g_1 g_2)$$

$\sum_j C(w_j w_1) = C(l_1)$, що і треба було довести.

Для $C(w_4)$ аналогічно, бо в таблиці можна поміняти відповідні елементи місцями і вигляд таблиці не зміниться.

Таким чином, показано, що залежність (3) виконується для всіх елементів таблиці, при використанні перерозподілу (5).

Теорема 1. При виконанні умов або повної сумісності, або повної несумісності множин словоформ за граматичними класами та канонічними формами залежність (3) виконується при використанні перерозподілу (5), для як завгодно великого словника.

Доведення. При побудові редукованої мови в одну умовну словоформу w_i можна внести як завгодно велику множину слів. При виконанні вимог лем 1 або 2 маємо, що виконання залежності (3) не залежить від істинної кількості та властивостей словоформ конвертованих в умовну словоформу w_i.

Вплив на можливі чисельні моделі

Таким чином, в задачах з реальними даними можна отримати оцінки навіть для тих елементів, що не спостерігалися у корпусі з якого збиралися дані, якщо доповнити словник словоформами для яких є канонічні форми та граматичні класи. Не обмежуючи загальності розглянемо приклад для української мови, для російської та інших все аналогічно.

Приклад. Для слів „корабель”, „корабля” „літак”, „літаком” „зелений”, „зеленого” „червоний”, „червоним” спостерігалися біграми „зелений корабель”, „зеленого корабля”, „червоний літак”, „червоним літаком”.

1) Поетапну згортку за канонічними формами („корабель” ,літак”) і („зелений” та „червоний”),

2) Поетапну згортку за граматичними класами „одн., чол. рід, наз. відм”, „одн., чол. рід, род. відм.”, „одн., чол. рід, орудн. відм.”

3) Після розгортки (застосування формули (5)) біграми „зеленим кораблем” та „червоного літака” отримають ненульові значення частот.

Для визначення того, наскільки насправді можна згортати словоформи треба робити чисельний експеримент для кожного нового корпуса.

Межі придатності моделі

Розглянемо випадок часткової сумісності. Так, наприклад, слово „небеса” має лише форми множини і не має форм однини. Виникає питання про допустимість використання спільних граматичних класів для таких іменників та іменників, що мають всі форми, як однини так і множини.

Таблиця 3.

Частково сумісні класи

	g_1	g_2
l_1	w_1	w_2
l_2		w_4

Теорема 2. Для частково сумісних класів після перерозподілу частоти канонічних форм та частоти граматичних класів залежність (3) не виконується.

Доведення. Розглянемо для $C(w_1)$

$$C(w_1)=\frac{C(L(w_1))C(G(w_1))}{\sum\limits_{G_F\in G(El(w_1))}C(G_F)}=\frac{C(l_1)C(g_1)}{C(g_1)+C(g_2)},$$

з іншого боку

$$C(w_1)=\sum_j C(w_1w_j)=\frac{C(L(w_1w_1))C(G(w_1w_1))}{\sum\limits_{G_F\in G(El(w_1w_1))}C(G_F)}+$$

$$\frac{C(L(w_1w_2))C(G(w_1w_2))}{\sum\limits_{G_F\in G(El(w_1w_2))}C(G_F)}+\frac{C(L(w_1w_4))C(G(w_1w_4))}{\sum\limits_{G_F\in G(El(w_1w_4))}C(G_F)};$$

враховуючи що:

$$\sum_{G_F\in G(El(w_1w_1))}C(G_F)=C(g_1)+C(g_2),$$

$$\sum_{G_F\in G(El(w_1w_2))}C(G_F)=C(g_1)+C(g_2),$$

$$\sum_{G_F\in G(El(w_1w_4))}C(G_F)=C(g_1g_2)+C(g_2g_2)$$

$$\sum_j C(w_1w_j)=\frac{C(l_1l_1)C(g_1)}{C(g_1)+C(g_2)}+\frac{C(l_1l_2)C(g_1g_2)}{C(g_1g_2)+C(g_2g_2)}$$

після зведення до спільного знаменника виникає вираз нетотожній $\frac{C(l_1)C(g_1)}{C(g_1)+C(g_2)}$, що і є демонстрацією суперечності.

Таким чином, щоб можна було застосовувати формулу перерозподілу (5) необхідно рознести слова з частково суміжними граматичними класами по групах, де після перевизначення граматичні класи будуть повністю несумісні. Це еквівалентно введенню додаткових ознак до множини граматичних. Наприклад, „множинний іменник”, щоб не використовувати комплекти роду, числа та відмінку від звичайних іменників для множинних.

Висновок

Запропоновано новий метод згладжування на основі граматичних класів, досліджено межі його застосування.

Подальше дослідження повинно бути спрямоване на визначення придатності для вживання в реальних задачах, особливо аспекти оптимального розбиття на граматичні класи та оцінка складності обчислення частот за формулою (5), оскільки при n-грамах довжини 3 та більше, що є сучасним стандартом, знаменник може виявитися незручним для обчислення.

Список використаних джерел

1. *S. F. Chen* An empirical study of smoothing techniques for language modeling. /*S. F. Chen, J. T. Goodman* // Computer Speech and Language 1999, 13, –PP. 448-453.

2. *Peter F. Brown* Class-based *n*-gram models of natural language / Peter F. Brown, Peter V. de Souza, Robert L. Mercer, et al. // Journal of Computational Linguistics –vol. 18, issue 4, – 1992. –PP. 467-479.

3. Babyn D.N. On the prospects of creating a system of automatic continuous speech recognition for spoken Russian. / D.N. Babyn, I.L. Mazurenko, A.B. Holodenko // Intelligent systems, – vol.8 , issue 1-4, –2004. –PP. 45-70 (in Russian).

4.*S. Ostrogonac* Language model reduction for practical implementation in LVCSR systems / *S. Ostrogonac, B. Popović, M.. Sečujski, et al.* // Infoteh-Jahorina Vol. 12, March 2013. – PP. 391-394.

Надійшла до редколегії 16.12.2013