УДК 004.93

Тарануха В. Ю.[1], асистент.

Згладжена n-грамна модель, заснована на класах, для розпізнавання слов'янських мов

[1] Київський національний університет імені Тараса Шевченка,83000, м. Київ, пр-т. Глушкова 4д,
e-mail: taranukha@ukr.net

V. Y. Taranukha [1], assistant.

Smoothed class-based n-gram model for recognition of the Slavic languages

[1] Taras Shevchenko National University of Kyiv, 83000, Kyiv, Glushkova st., 4d,
e-mail: taranukha@ukr.net

Стаття присвячена дослідженню n-грамної моделі для розпізнавання слов'янських мов та методу згладжування, призначеному для зменшення розміру моделі та підвищення якості розпізнавання.

Ключові слова: n-грамна модель мови, редукція моделі, модель на класах, згладжування.

The article investigates n-gram language model for Slavic language recognition. Syntactic links in Slavic languages are built mostly with changing forms of words and less with word order. It leads to creation of vast vocabularies with many wordforms matching single word (or lemma). Subsequently it increases transition matrix size and reduces frequencies of most elements in the matrixes. This makes models designed on Markov chains and n-grams less reliable for Slavic language comparing to Germanic and Romance languages. Method of smoothing aimed both for reduction of the model and decrease of recognition rate loss is developed. It is based on decomposition of word form n-grams into n-grams based on grammatical classes and lemma classes. Partial filtering is used for trigrams to reduce the model size. Trigrams composed of codes only are filtered. New smoothed pseudocounts are calculated based on decomposed model. Numerical tests have shown possible improvements in entropy of the model thus implying improvements in recognition rate.

Key Words: n-gram language model, model reduction, class-based model, smoothing.

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

Завдяки розвитку мережі Інтернет значно зріс об'єм даних формально доступних у електронній формі. Частина даних представлена у вигляді зображень документів та аудіо записів і це певною мірою ускладнює їх обробку. Відповідно виникає потреба у створенні засобів, що дозволять перетворити зображення та аудіо записи в текст. Розповсюдженою є модель, що спирається на харківські ланцюги, а відповідно і на n-грами[1]. За її допомогою можна зручно оцінювати ймовірність появи ланцюжка з *n* слів у деякому тексті. При застосуванні такої моделі до української мови і до слов'янських мов взагалі, проявляються проблеми, що роблять використання цієї моделі менш зручним, порівняно з романо-германськими мовами. Пропонується розробити модифіковану модель, щоби оцінювати імовірності кращим чином.

Звичайна модель

Послідовність слів мови $w_1 \ldots w_n$ називається n-грамою довжини n, позначимо її w_1^n. Тоді імовірність w_1^n можна оцінити за формулою:

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

Умовні імовірності визначаються як

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

де $C(w_{i-n+1}^{i})$ - частота відповідної n-грами.

В словянських мовах є відчутно вища

кількість словоформ з розрахунку на одну лему[2]. Це призводить до необхідності збільшувати словник. Через це значна кількість n-грам набуває малих значень частот, і оцінка імовірностей по формулі (2) стає значно чутливішою до збурень.

Для оцінювання якості моделі пропонується застосовувати оцінку на основі кросентропії:

$$H(p,m)=\lim_{n\to\infty}-\frac{1}{n}\log m(w_1w_2\dots w_n) \qquad (3)$$

де $m(w_1w_2\dots w_n)$ - модель для мовірності $p(w_1w_2\dots w_n)$. При цьому відомо, що

$$H(p)\le H(p,m) \qquad (4)$$

В якості методу вирішення вищезгаданої проблеми пропонувалося розділення моделі на дві: модель на лемах та модель на граматичних класах[1,2]. Також пропонувалася фільтрація на основі різних евристик[3]. Експерименти показали, що ці методи, взяті окремо або об'єднані в одній моделі практично не розв'язують поставлену задачу. Безпосередня фільтрація вимагає додаткових засобів та погіршує модель, розділення та оцінювання компонентних моделей незалежно, щоби потім додати оцінки, одразу підвищує ентропію.

Згладжена модель

В якості нового методу пропонується модифікувати модель засновану на класах.

Для розбиття на класи вводиться функція розбиття, що ставить у відповідність кожному слову w_i з словника системи V клас c_i. При цьому виконується:

$$P(w_i \mid w_1^{i-1})=P(w_i \mid c_i)P(c_i \mid c_1^{i-1}),\forall i,1\le i\le n \quad (5)$$

Висунуто припущення, що для слів, про які відомо, що вони мають однакову синтаксичну поведінку можна стверджувати, що у схожих контекстах вони повинні мати схожі імовірності зустрічання.

Нехай для слів „стілець", „стільця", „молоток", „молотком" „синій", „синього" „жовтий", „жовтим" в корпусі спостерігалися біграми „синій стілець", „синього стільця", „жовтий молоток", „жовтим молотком".

На базі знань про те, що ці іменники та прикметники мають схожу граматичну поведінку, а саме: мають однакові множини граматичних класів, можна побудувати припущення про імовірності появи їх в формах, що не спостерігалися в корпусі.

Враховуючи, омонімію в українській мові модель будується таким чином.

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

Тоді оцінка частоти w_1^k визначається:

$$C(w_1^k)=\frac{\mathrm{C}(L(w_1^k))\mathrm{C}(G(w_1^k))}{\sum\limits_{G_F\in G(El(w_1^k))}G_F} \qquad (6)$$

При цьому в ході експериментів встановлено, що вимога не виконується, щоби обчислені наново псевдо частоти були коректними:

$$C_{(w_{i-n+2}^{i})}=\sum_{j=0}^{|V|}C(w_jw_{i-n+2}^{i}) \qquad (7)$$

де $|V|$ - розмір словника.

Для забезпечення коректності моделі висувається вимоги: необхідно щоб сума частот канонічних форм та сума частот граматичних класів після перерозподілу лишалася незміною[4].

$$G(C(w_1^{k-1}))=\sum_{m=1}^{|V|}G(C(w_1^{k_m})) \qquad (8)$$

$$L(C(w_1^{k-1}))=\sum_{m=1}^{|V|}L(C(w_1^{k_m})) \qquad (9)$$

Якщо (8) та (9) виконується, то це дозволить суттєво оптимізувати обчислення (6), та забезпечити його коректність.

Чисельний експеримент показав, що попри покриття всіх існуючих в навчальній вибірці комбінацій лем та комбінаціїй граматичних класів, повне покриття всіх потенційно існуючих мовних явищ не відбувається. Тому виникає потреба у додатковій формулі згладжування, що дозволить обчислити оцінки псевдо частот для тих елементів, що мають значення частоти рівним 0 навіть після застосування (6). В якості тестової моделі згладжування було вибрано модель Віттена-Бела[5] в варіанті з поверненнями. Моделі на основі евристики Гуда[6] є непридатними, тому що неможливо коректно побудувати їх форму для нецілих значень псевдочастот.

Модель визначається так:

$$\hat{p}(w_i \mid w_{i-n+1} \dots w_{i-1}) =$$

$$\begin{cases} d(w_{i-n+1} \dots w_i),\ C(w_{i-n+1} \dots w_i) > 0 \\ \alpha_{w_{i-n+1} \dots w_{i-1}}\ \hat{p}(w_i \mid w_{i-n+2} \dots w_{i-1})\ інакше \end{cases} \quad (10)$$

де $d(w_{i-n+1} \dots w_i)$ - відповідним чином згладжене значення, $C(w_{i-n+1}^i)$, $\alpha_{w_{i-n+1} \dots w_{i-1}}$ - відповідний коефіцієнт, що визначає імовірнісну масу, перерозподілену для побудови імовірностей на n-грамах моделі меншого порядку.

$$\alpha_{w_{i-n+1} \dots w_{i-1}} = \frac{\beta_{w_{i-n+1} \dots w_{i-1}}}{\sum_{\{w_i : C(w_{i-n+1}^i) = 0\}} \hat{p}\ (w_i \mid w_{i-n+2}^{i-1})} \quad (11)$$

$$\beta_{w_{i-n+1} \dots w_{i-1}} = 1 - \sum_{\{w_i : C(w_{i-n+1}^i) > 0\}} d(w_{i-n+1}^i) \quad (12)$$

Для методу Віттена-Бела параметр *d* оцінюється так:

$$d_{WB}(w_i \mid w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^i)}{C(w_{i-n+1}^i) + T(w_{i-n+1}^i)} \quad (13)$$

де $T(w_{i-n+1}^i)$ - кількість типів n-грам, що передують слову w_i. При цьому, за замовчуванням, n-грами найвищого порядку з частотою 1 видаляються з моделі після побудови всіх таблиць.

Чисельний есперимент

Для чисельного експерименту використано модель на n-грамах розмірності ≤ 3, зібраних зі стенограм Верховної Ради України. Було сформовано корпус обсягом 112,5 МБ. Для цього відповідні стенограми було зібрано з сайту http://rada.gov.ua/meeting/stenogr . На корпусі було виділено словник системи з 10.000 словоформ, всі інші слова були замінені на стоп-слово "#". Словник було пропущено через систему морфолексичного аналізу, і сформовано словники канонічних форм та словники граматичних класів. При цьому множина граматичних класів однозначно визначає словник канонічних форм. За замовчуванням деякі слова не замінялися на леми та граматичні класи, наприклад: іменники родового відмінку, займенники, тощо.

Ентропію було обчислено для трьох різних варіантів реалізації згладжування. В Базовому варіанті застосовується лише формула(1). „Згладжування 1" передбачає використання у формулі (6) лише граматичних класів отриманих із морфологічного словника. Варіант „Згладжування 2" передбачає використання строгих граматичних підкласів класів, щоби забезпечити виконання умов (8),(9) як описано в вимогах[4]. Як виявилось, з наявною навчальною вибіркою неможливо забезпечити строге виконання (8) та (9) без внесення у словник великої кількості слів, для яких немає n-грам в навчальній частині корпуса. Експеримент показав, що за таких обставин варіант „Згладжування 2" показує поганий результат. Тому було вирішено комбінувати підходи, щоби отримати найкраще наближення без зайвої втрати якості. Було зроблено декілька ітерацій, словник обновлювався методом заміни словоформ (як на наявні у комплекті n-грам зібраних з навчальної частини корпуса таі і ні), при збереженні загальної кількості елементів. Застосовувалася фільтрація: відфільтровувалися n-грами частоти 1 з комплектів граматичних та лематичних n-грам при обчисленні формули (6). Результати наведено в Таблиці 1.

Таблиця 1

Результат експерименту

Метод	Ентропія
Базовий	$7{,}272 \pm 0{,}036$
Згладжування 1	$7{,}281 \pm 0{,}034$
Згладжування 2	$7{,}283 \pm 0{,}043$

На різних ітераціях значення ентропії мінялося, при чому на деяких ітераціях результат „Згладжування 2" був кращий за Базовий для отриманого словника.

Висновки

Показано, що формула згладжування на основі тематичної та граматичної інформації дозволяє ефективно згладжувати n-грамну модель мови. Застосування апріорної інформації про морфолексичні та граматичні характеристики слів дозволяє визначати структуру класів, для оптимізації моделей мови з метою покращення розпізнавання. Чисельні експерименти показали, що про безпосереднє застосування формули (6) не гарантує хорошого результату навіть з фільтрацією та використанням більш строгих класів. Відповідно, необхідно виконувати підбір елементів словника. При цьому крім прямої перевірки значення ентропії немає способу гарантувати підвищення якості.

Також, метод Віттена-Бела хоч і є показовим, проте в реальних задачах треба використовувати метод на зразок Кнесера-Нея[7], що в свою чергу ускладнює процес побудови остаточного комплекту таблиць та коефіцієнтів у формулах.

Подальші дослідження повинні бути

спрямовані на пошук можливості гарантувати зростання якості моделі для випадків, коли навчальний тестовий корпус має порівняно малий об'єм і не дає можливості гарантувати виконання вимог (8),(9) про сумісність та несумісність класів без втрати якості моделі.

Список використаних джерел

1. *Тарануха В.Ю.* Застосування класів основаних на канонічних формах слів та на граматичних класах в задачі редукції n-грамної моделі мови для розпізнавання української мови ./Тарануха В.Ю // Вісник Київського національного університету імені Тараса Шевченко Серія: фізико-математичні науки. –2013, – Спецвипуск. – С. 176-179
2. *Бабин Д.Н.* О перспективах создания системы автоматического распознавания слитной устной русской речи. /Бабин Д.Н, Мазуренко И.Л. , Холоденко А.Б.// Интеллектуальные системы. –2004. – Т.8, –Вып. 1-4,– С.45-70.
3. *S. Ostrogonac* Language model reduction for practical implementation in LVCSR systems / S. Ostrogonac, B. Popović, M.. Sečujski, et al. // Infoteh-Jahorina – 2013: proceedings of 12th International Scientific–Professional Symposium INFOTEH - JAHORINA, At Jahorina 20-22 march. –Vol. 12, March 2013. –PP. 391-394.
4. *Тарануха В.Ю.* Модифікація n-грамної моделі, заснованої на класах, для розпізнавання слов'янських мов / Тарануха В.Ю.// Вісник Київського національного університету імені Тараса Шевченко Серія: фізико-математичні науки. – 2014. – Вип 1. – С. 193-196
5. *I. H. Witten* The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression/ I. H. Witten and T. C. Bell // IEEE Transactions on Information Theory, – 1991. Vol. 37(4), –P. 1085–1094.
6. *I.J. Good*, The population frequencies of species and the estimation of population parameters //Biometrika, –1953. Vol. 40 (3–4),–P. 237–264.
7. *R. Kneser* Improved backing-off for m-gram language modeling / R. Kneser and H. Ney// Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, –1995. Vol. 1, –P. 181–184.

References

1. TARANUKHA, V. (2013) Applying classes based on canonical forms of words and the grammar classes in the problem of reduction of n-грамної language model for recognition of the Ukrainian language, *Bulletin of Taras Shevchenko National University of Kyiv Series Physics & Mathematics*, (Special issue), pp. 176-179.
2. BABIN, D., MAZURENKO, I., HOLODENKO, A. (2004) O perspektivah sozdanija sistemy avtomaticheskogo raspoznavanija slitnoj ustnoj russkoj rechi, *Intelligent sistems*, 8(1-4), pp. 45-70.
3. OSTROGONAC, S., POPOVIĆ, B., SEČUJSKI, M., et al. (2013) *Language model reduction for practical implementation in LVCSR systems*: *12th International Scientific–Professional Symposium Infoteh-Jahorina-2013*, (12) Jahorina 20-22 March 2013, pp. 391-394.
4. TARANUKHA, V. (2014) Modification of class-based n-gram model for slavic speech recognition, *Bulletin of Taras Shevchenko National University of Kyiv Series Physics & Mathematics*, (1), pp. 193-196.
5. WITTEN, I. and BELL, T. (1991) The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression, *IEEE Transactions on Information Theory*, 37(4), pp. 1085-1094.
6. GOOD, I. (1953) The population frequencies of species and the estimation of population parameters, *Biometrika*, 40(3–4), pp. 237–264.
7. KNESER, R. and NEY, H. (1995) Improved backing-off for m-gram language modeling,*Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1, pp. 181–184.

Надійшла до редколегії 22.04.14