

УДК 004.93

Тарануха В. Ю., асистент.

**Властивості згладженої n -грамної моделі
для слов'янських мов, заснованої на
класах**

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

V. Y. Taranukha, assistant.

**Properties of smoothed class-based n -gram
model for Slavic languages**

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 властивостям методу згладжування, проаналізовано різні способи формування моделі та їх наслідки для оцінок якості.

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

Статья посвящена исследованию n -грамной модели для распознавания славянских языков в свойствам метода сглаживания, проанализировано разные способы формирования модели и их результаты для оценок качества.

Ключевые слова: n -грамная модель языка, модель на классах, сглаживание.

The article investigates n -gram language model for Slavic language recognition. Idea of improvement is based on specific features of Slavic languages. Syntactic links in Slavic languages are built mostly with changing forms of words and much less with word order. It makes extended vocabularies with many wordforms for single meaningful word (or lemma). For the same number of meaningful words it expands transition matrix size. Given the same corpora size it reduces frequencies of most elements in the matrixes that represent model. 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 for improvement of recognition rate is investigated. It is based on decomposition of wordform n -grams into n -grams based on grammatical classes and lemma classes. Different methods are used to reduce the model size. After reduction new model with smoothed pseudocounts is calculated based on decomposed model. Numerical tests were performed for different setups and different secondary smoothing techniques. They have shown minor 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.

Швидкий розвиток комунікаційних технологій призвів до зростання доступних обсягів даних в електронній формі. Частина з цих даних представлена у вигляді зображень документів та аудіозаписів, а не у вигляді текстових документів. Тому значна увага приділяється засобам, що дозволяють перетворити зображення та аудіозаписи в текст. Розповсюдженою є модель, що спирається на марковські ланцюги, матриці яких зручно представляти у вигляді наборів n -грам[1]. За допомогою марковських

ланцюгів можна зручно оцінювати ймовірність появи ланцюжка з n слів у певному тексті. Проте при застосуванні такої моделі до слов'янських мов і, зокрема, до української виникають проблеми, що пов'язані з властивостями мови і вони роблять використання цієї моделі менш зручним, порівняно з використанням цієї ж моделі для романо-германських мов. Була розроблена спеціальна модель, щоби оцінювати ймовірності кращим чином спираючись на особливості саме української мови.

Звичайна модель

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

$$p(w_1^n) = p(w_1 | w_1^{i-1}) p(w_{i-1} | w_1^{i-2}) \dots p(w_1) \quad (1)$$

Умовні імовірності визначаються як

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

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

Така оцінка є оцінкою максимальної правдоподібності.

В українській мові суттєво більша кількість слів форм припадає на одну лему (канонічну форму)[2]. При однаковому наповненні моделі змістовними зв'язками це призводить до необхідності збільшувати словник слів форм.

Оскільки марковський ланцюг, а відповідно і n -грами будуються на слів формах, значна кількість n -грам набуває малих значень частот. Тому оцінка імовірностей по формулі (2) стає значно менш точною і сильніше залежить від корпусу на якому збиралися дані для моделі.

Ще одним загальним недоліком оцінки (2) є те, що в реальних корпусах не представлені всі можливі n -грами. Виникає потреба в застосуванні методу для згладжування частот та імовірностей відповідних n -грам. Раніше було запропоновано комбінувати відомі загальні методи згладжування з спеціалізованими[1].

Згладжена модель

Досліджується модифікована модель заснована на класах. Для розбиття на класи вводиться функція розбиття, що ставить у відповідність кожному слову w_i з словника системи V (тобто кожній слів формі, а не лемі) клас c_i . При цьому виконується:

$$P(w_i | w_1^{i-1}) = P(w_i | c_i) P(c_i | c_1^{i-1}), \forall i, 1 \leq i \leq n \quad (3)$$

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

Нехай для слів „каша”, „каші”, „страва”, „стравою”, „смачна”, „смачної”, „гаряча”, „гарячою” в корпусі спостерігалися біграми „смачна каша”, „смачної каші”, „гаряча страва”, „гарячою стравою”. Тоді спираючись на те, що ці іменники та прикметники мають схожу граматичну поведінку, можна побудувати припущення про імовірності появи їх в формах,

що не спостерігалися в корпусі, наприклад, для біграми „гарячої страви”.

Модель будується таким чином. $L(w_1^k)$ – сукупність послідовностей лем для послідовності слів форм w_1^k . $G(w_1^k)$ – сукупність послідовностей граматичних класів для послідовності слів форм w_1^k . $El(w_1^k)$ – сукупність послідовностей слів форм, що після приведення до лем 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 (псевдо частота) визначається так:

$$C(w_1^k) = \frac{C(L(w_1^k)) C(G(w_1^k))}{\sum_{G_F \in G(El(w_1^k))} G_F} \quad (4)$$

Щоби обчислені наново псевдо частоти були коректними висувається вимога:

$$C(w_{i-n+2}^i) = \sum_{j=0}^{|V|} C(w_j w_{i-n+2}^i) \quad (5)$$

де $|V|$ – розмір словника, а деякі $C(w_j w_{i-n+2}^i)$ можуть бути рівні 0.

Для забезпечення коректності моделі висуваються вимоги про незмінність суми частот лем та сума частот граматичних класів після перерозподілу[3].

$$C(G(w_{i-n+2}^i)) = \sum_{j=0}^{|V|} C(G(w_j w_{i-n+2}^i)) \quad (6)$$

$$C(L(w_{i-n+2}^i)) = \sum_{j=0}^{|V|} C(L(w_j w_{i-n+2}^i)) \quad (7)$$

Чисельний експеримент

Експеримент показав, що попри покриття всіх існуючих в навчальній вибірці комбінацій лем та комбінацій 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 (4). В якості основної тестової моделі згладжування було вибрано метод Віттена-Бела[4] в варіанті з поверненнями. Також використано метод згладжування Лідстона, для порівняння. Модель містить n -грами розмірності ≤ 3 .

Модель Віттена-Бела визначається так:

$$\hat{p}(w_i | w_{i-n+1} \dots w_{i-1}) = \begin{cases} d(w_{i-n+1} \dots w_i), C(w_{i-n+1} \dots w_i) > 0 \\ \alpha_{w_{i-n+1} \dots w_{i-1}} \hat{p}(w_i | w_{i-n+2} \dots w_{i-1}) \text{ інакше} \end{cases} \quad (8)$$

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

$$\alpha_{w_{i-n+1} \dots w_{i-1}} = \frac{\beta_{w_{i-n+1} \dots w_{i-1}}}{\sum_{\{w_i: C(w_{i-n+1}^i)=0\}} \hat{p}(w_i | w_{i-n+2}^{i-1})} \quad (9)$$

$$\beta_{w_{i-n+1} \dots w_{i-1}} = 1 - \sum_{\{w_i: C(w_{i-n+1}^i)>0\}} d(w_{i-n+1}^i) \quad (10)$$

Для методу Віттена-Бела параметр d оцінюється так:

$$d_{WB}(w_i | w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^i)}{C(w_{i-n+1}^i) + T(w_{i-n+1}^i)} \quad (11)$$

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.

Для згладжування Лідстона:

$$\hat{p}(w_i | w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^i) + \delta}{C(w_{i-n+1}^{i-1}) + \delta |V|} \quad (12)$$

де δ – параметр, що підбирається для отримання оптимальної оцінки.

Для оцінювання якості моделі застосовується оцінка на основі кросентропії:

$$H(p, m) = \lim_{n \rightarrow \infty} -\frac{1}{n} \log m(w_1 w_2 \dots w_n) \quad (13)$$

де $m(w_1 w_2 \dots w_n)$ – модель для імовірності $p(w_1 w_2 \dots w_n)$. При цьому відомо, що

$$H(p) \leq H(p, m) \quad (14)$$

Для чисельного експерименту було сформовано корпус обсягом 112,5 МБ з стенограм Верховної Ради України. Відповідні стенограми було зібрано з сайту <http://rada.gov.ua/meeting/stenogr>. На корпусі було виділено словник системи з 10.000 словоформ різними методами. Словник на якому більшість методів показали хорошу оцінку був вибраний в якості основного. Було побудовано дві допоміжні n -грамні моделі на основі граматичних класів та на основі лем.

Ентропію було обчислено для таких варіантів реалізації згладжування:

- 1) без перерозподілу, метод Віттена-Бела,
- 2) без перерозподілу, метод Лідстона,
- 3) з перерозподілом, з фільтрацією а потім методом Віттена-Бела,
- 4) з перерозподілом, фільтрацією а потім методом Лідстона,
- 5) з перерозподілом, без фільтрації, із врахуванням повної інформації, та методом Віттена-Бела.

В перших 4 варіантах всі непотрібні слова одразу замінювалися на стоп-слово „#”. Фільтрація передбачає викидання з триграм згладженої моделі тих, що мають відповідні низькі частоти в допоміжних моделях. Для урахування повної інформації спочатку виконується перерозподіл, а лише після того непотрібні слова замінюються на стоп-слово „#”. Це суттєво зменшує кількість триграм, якими поповнюється модель. Всі варіанти обчислювались з крос-валідацією.

Таким чином, ця серія експериментів, разом з раніше проведеними[5] повністю покриває всі змістовні методи формування моделей.

Результати експерименту наведені в таблиці

1. Як видно з наведених чисел, використання

Таблиця 1

Результат експерименту

Метод	Ентропія(середня)
1	6,92
2	7,11
3	6,91
4	7,17
5	7,63

методу 3 (ключового в усьому дослідженні) принаймні не погіршує оцінки.

Висновки

Показано, що метод згладжування на основі граматичної інформації дозволяє ефективно згладжувати n -грамну модель мови, і принаймні не погіршує оцінку. Чисельні експерименти показали, що про безпосереднє застосування формули (4) не гарантує хорошого результату навіть з фільтрацією та використанням більш строгих класів. Відповідно, необхідно виконувати підбір елементів словника, що крім іншого помітно впливає на абсолютне значення ентропії. При цьому крім прямої перевірки значення ентропії немає способу гарантувати підвищення якості.

Як показав аналіз, методи на основі евристики Гуда[6], або Кнесера-Нея[7] не дуже підходять, бо або не мають форми для псевдочастот, або вимагають підгонки

параметрів. При підгонці параметрів, як це було показано на прикладі метода Лідстона, перерозподілена модель втрачає свої переваги.

Також, добре видно, що спроба врахувати повну інформацію з корпусу призводить до погіршення якості моделі. Схоже, це пов'язано з тим, що n-грами з стоп-словом „#” можна

сприймати як скіп-грами – n-грами з пропущеними словами, які теж корисні, коли мають високі частоти. А от намагання врахувати всі словоформи призводить до накопичення низькочастотних елементів (шумів).

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 Шевченка 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. Тарануха В.Ю. Модифікація n-грамної моделі, заснованої на класах, для розпізнавання слов'янських мов / Тарануха В.Ю.// Вісник Київського національного університету імені Тараса Шевченка Серія: фізико-математичні науки. – 2014. – Вип 1. – С. 193-196
4. 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.
5. Тарануха В.Ю. Згладжена n-грамна модель, заснована на класах, для розпізнавання слов'янських мов / Тарануха В.Ю. // Вісник Київського національного університету імені Тараса Шевченка Серія: фізико-математичні науки. –2014. – Вип 2. – С. 202-205.
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-gram 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) О перспективах создания системы автоматического распознавания слитной устной русской речи, *Intelligent systems*, 8(1-4), pp. 45-70.
3. 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.
4. 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.
5. TARANUKHA, V. (2014) Smoothed class-based n-gram model for recognition of Slavic languages, *Bulletin of Taras Shevchenko National University of Kyiv Series Physics & Mathematics*, (2), pp. 202-205.
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 2.07.15