УДК 004.93

В.Ю. Тарануха

# Метод згладжування *n*-грамної моделі для розпізнавання мовлення, заснованої на класах, з використанням граматичної та лексичної інформації

Рассмотрен новый метод сглаживания, ориентированный на особенности славянских языков, в том числе на украинский. Показано, что он улучшает оценку качества моделей языка.

A new smoothing method focused on the features of the Slavic languages, including the Ukrainian is proposed. It is shown that the method improves the quality of the language models.

Розглянуто новий метод згладжування, орієнтований на особливості слов'янських мов, в тому числі на українську. Показано, що він покращує оцінку якості моделей мови.

**Вступ.** Значне зростання обсягів інформації у вигляді цифрових аудіозаписів та зображень текстів потребує ефективних засобів для переведення даних в текстову форму для подальшої обробки. Стандартом де-факто є використання статистичної моделі на основі *n*-грам [1] та алгоритмів, що з нею працюють. Така модель в цілому добре розроблена [2], проте при використанні для слов'янських мов, зокрема для української, виявляється низка недоліків, пов'язаних з властивостями слов'янських мов у порівнянні з романо-германськими. Пропонувалися різні підходи до вирішення цієї проблеми: перехід до *n*-грам з вільним порядком слів [3], використання даних синтаксичного аналізатора [4], фільтрація на основі евристик [5]. В даній статті розглянуто модифікацію класичної моделі, спираючись на лексичні та граматичні класи.

**Побудова та оцінювання моделей**

Ймовірнісна модель звичайно передбачає, що мова має властивості, які дозволяють описати її як марківський процес. Тоді ймовірність послідовності слів можна буде оцінити явно [1]. Послідовність слів мови $w_1 \ldots w_n$ називається *n*-грамою довжини *n*, її позначають $w_1^n$. Тоді послідовність слів можна представити як послідовність *n*-грам, а ймовірність оцінити за формулою $p(w_1^i) = p(w_i \mid w_1^{i-1}) p(w_{i-1} \mid w_1^{i-2}) \ldots . p(w_1)$.

При цьому можна побудувати оцінку ймовірностей, що спирається на частоти відповідних *n*-грам $\hat{p}(w_i \mid w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i})}{C(w_{i-n+1}^{i-1})}$, де $C(w_{i-n+1}^{i})$ – частота відповідної *n*-грами.

У слов'янських мовах характерною рисою є вільний порядок слів у реченнях. При цьому слово має більше словоформ, оскільки в словоформах зберігається інформація, що вказує на потенційні синтаксичні зв'язки слова.

Якщо на одне слово англійської мови припадає приблизно 1,7 словоформи, то на одне слово української мови, залежно від вибраного словника, може припасти від 5,5 до 19,9 на одному і тому ж корпусі. Отже, при побудові таблиці *n*-грам, при $n = 2$, розмір зростає принаймні в 10,47 рази, а при $n = 3$ – в понад 33 рази. Це призводить до того, що значна кількість *n*-грам набуває малих значень частот, і оцінка ймовірностей стає значно чутливішою до викидів та шумів. Це і складає головну проблему та заважає досягти таких самих високих показників розпізнавання, як для романо-германських мов.

Ще однією властивістю наведеного підходу є те, що в реальних корпусах не представлені всі можливі *n*-грами. Це створює потребу в застосуванні методу для згладжування частот та ймовірностей відповідних *n*-грам.

Для оцінки якості моделі без необхідності виконувати експеримент з розпізнаванням використовується ентропія

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x).$$

Це дозволяє оцінити якість марківського ланцюжка, хоча ігноруються певні аспекти реального розпізнавання. Наприклад, ігнорується схожість звучання слів, що може призвести до погіршення результатів у реальному експерименті.

Крім того, в ідеальному випадку необхідно обчислювати ентропію з розрахунку на слово, на потенційно нескінченному реченні, що опи-

сує мову. Проте в реальному експерименті доводиться обходитися вибіркою певного розміру, сподіваючись, що наближене значення буде близьким до теоретичного

$$H(L) = \lim_{n\to\infty} -\frac{1}{n}\log p(w_1 w_2 \dots w_n).$$

Для порівняння двох моделей зручно використовувати крос-ентропію. Нехай $m(w_1w_2\dots w_n)$ – модель для ймовірності $p(w_1w_2\dots w_n)$, тоді крос-ентропія з розрахунку на слово виражається так: $H(p,m) = \lim_{n\to\infty} -\frac{1}{n}\log m(w_1 w_2 \dots w_n)$, для неї відомо, що $H(p) \le H(p, m)$.

Також, при потребі можна застосувати більш детальну міру якості, що називається *перплексією*, і обчислюється так: $PPW = 2^{H(p,\,m)}$.

В реальному тесті може трапитися, що відповідна *n*-грама, яка трапилася в тесті, відсутня в корпусі, за яким будувалася модель. Для оцінки події, що не траплялася, використовують методи згладжування, які будуть описані далі.

**Аналіз відомих методів для підвищення якості моделі**

***Перехід до n-грам з вільним порядком слів*** [3]: $\hat{p}(w_i \mid w_{i-n+1}^{i-1}) = p(w_i \mid \{w_{i-n+1}....w_{i-1}\})$. Фігурні дужки означають, що слова, крім останнього, добираються за довільним порядком. Чисельний експеримент [3] показав непридатність цієї моделі для розпізнавання через високу ентропію.

***Використання даних синтаксичного аналізатора*** [4]. В комплект *n*-грам додаються *n*-грами, отримані як коректні словосполучення, отримані аналізом дерева синтаксичного розбору. Отримано гарантований ефект підвищення якості розпізнавання. Нажаль, цей метод передбачає необхідність використання відповідного синтаксичного аналізатора, що не завжди є можливим.

***Використання декомпозиції моделі на дві***: модель, засновану на граматичних класах, та модель, засновану на канонічних формах слів. Згідно наведених оцінок для перплексії [3], при одночасному використанні обох часткових моделей, результуюча перплексія набагато вища за перплексію моделі, створеної лише на словоформах. Це позбавляє сенсу безпосереднє використання двох моделей, хоча відповідно до запевнень авторів [3] залишає можливість винести в модель на граматичних класах частину інформації, необхідної для боротьби з акустичною схожістю різних форм одного слова.

***Використання оптимізації ентропії*** або перплексії моделі шляхом оцінки та вилучення шумів з моделі [5]. При цьому можна оцінювати як модель, зібрану на словоформах, так і модель, зібрану на канонічних формах слів. Метод показав дієвість такої оптимізації, проте він має відчутний недолік, оскільки фактично спирається в оптимізації на *n*-грами малої частоти, а отже, – на шуми та викиди.

**Новий метод**

Пропонуємо модифікацію класичної моделі, яка спирається на лексичні та граматичні класи. При побудові через лексичні та граматичні класи будуються дві окремі моделі *n*-грам окремо на основі канонічних форм слів (лексична) та окремо на основі граматичних класів слів. Тобто, одна послідовність слів дає інформацію в дві різні часткові моделі. Після того на основі двох моделей будується спільна модель, яка використовує інформацію з обох часткових моделей. При потребі допускається фільтрація частини отриманих *n*-грам, якщо це покращить результат.

Для розбиття на класи в загальному випадку вводиться функція розбиття, що ставить у відповідність кожному слову $w_i$ з словника $V$ клас $c_i$. При цьому виконується

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

Ідея полягає в тому, що для слів, про які відомо, що вони мають однакову синтаксичну поведінку можна зробити припущення про те, що у схожих контекстах вони повинні мати схожі ймовірності зустрічання.

Нехай для слів *автомобіль*, *автомобіля*, *вертоліт*, *вертольотом*, *синій*, *синього*, *жовтий*, *жовтим* у корпусі спостерігалися біграми *синій автомобіль*, *синього автомобіля*, *жовтий вертоліт*, *жовтим вертольотом*.

Тоді можна виконати поетапну згортку за канонічними формами (*автомобіль*, *вертоліт*) і (*синій* та *жовтий*), поетапну згортку за граматичними класами *одн.*, *чол. рід*, *наз. відм.*,

*одн.*, *чол. рід*, *род. відм.*, *одн.*, *чол. рід*, *орудн. відм.*

На базі знань про те, що ці іменники та прикметники поводяться схоже, тобто мають однакові множини граматичних класів, можна побудувати припущення про ймовірності появи їх у формах, що не спостерігалися в корпусі.

Враховуючи, що в українській мові спостерігається омонімія, 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 $L(w_{i_1}^{i_k}) = L(w_1^k), \forall i$.

Тоді оцінка частоти $w_1^k$ визначається:

$$C(w_1^k) = \frac{\mathrm{C}(L(w_1^k))\mathrm{C}(G(w_1^k))}{\sum\limits_{G_F \in G(El(w_1^k))} G_F} \quad . \qquad (1)$$

Для забезпечення достовірності об'єднаної моделі припустимо: необхідно, щоб сума частот канонічних форм та сума частот граматичних класів після перерозподілу лишалася незміною, тобто

$$G(C(w_1^{k-1})) = \sum_{m=1}^{V} G(C(w_1^{k_m})) \;\; \&$$

$$\& \; L(C(w_1^{k-1})) = \sum_{m=1}^{V} L(C(w_1^{k_m})), \qquad (2)$$

де $|V|$ – розмір словника, а відповідні частоти $C(w_1^k)$ можуть бути нульовими, якщо для них не існує $w_1^k$ в корпусі, на якому будується модель. Якщо (2) виконується, то це дозволить суттєво оптимізувати обчислення (1).

Повертаючись до прикладу, дана модель дозволяє за формулою (1) оцінити ймовірності для біграм *синім автомобілем* та *жовтого вертольота*.

Підбір параметрів, а саме множини граматичних класів та множини канонічних форм, було проведено підчас чисельного експерименту.

**Метод згладжування**

Для згладжування та заповнення пропусків використовуються різні методи [2], в даній статті розглядаються згладжування з поверненням Віттена–Белла, оскільки воно є простим і водночас включає в себе всі необхідні параметри.

$$\hat{p}(w_i \,|\, w_{i-n+1}^{i-1}) = \begin{cases} d(w_{i-n+1}^i) \;\; , \; C(w_{i-n+1}^i) > 0, \\ \alpha_{w_{i-n+1}\ldots w_{i-1}} \hat{p}(w_i \,|\, w_{i-n+2}^{i-1}) \;\; \text{інакше}, \end{cases} \qquad (3)$$

де $d(w_{i-n+1}^i)$ – відповідним чином згладжене значення $C(w_{i-n+1}^i)$, $\alpha_{w_{i-n+1}\ldots w_{i-1}}$ – відповідний коефіцієнт, що визначає ймовірнісну масу, перерозподілену для побудови ймовірностей на *n*-грамах моделі нижчого порядку.

$$\alpha_{w_{i-n+1}\ldots w_{i-1}} = \frac{\beta_{w_{i-n+1}\ldots w_{i-1}}}{\sum\limits_{\{w_i : C(w_{i-n+1}^i)=0\}} \hat{p}\;(w_i \,|\, w_{i-n+2}^{i-1})}, \qquad (4)$$

$$\beta_{w_{i-n+1}\ldots w_{i-1}} = 1 - \sum_{\{w_i : C(w_{i-n+1}^i)>0\}} d(w_{i-n+1}^i) \; . \qquad (5)$$

Для методу Віттена–Белла параметр *d* оцінюється так:

$$d_{WB}(w_i \,|\, w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^i)}{C(w_{i-n+1}^i) + T(w_{i-n+1}^i)}, \qquad (6)$$

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 одиниця видаляються з моделі.

Оскільки після застосування формули (1) *n*-грами отримають не частоти, а псевдочастоти, то згладжування за формулою Віттена–Белла є зручним, оскільки не потребує регулювання, від якого значення псевдочастоти необхідно відраховувати допустимі елементи у $T(w_{i-n+1}^i)$.

Метод Катца з поверненням не підходить як додатковий метод згладжування, оскільки спирається на евристику Гуда–Тьюринга [1], яка не має зрозумілого способу інтерпретації, якщо в неї подати псевдочастоти замість частот.

Відповідно до якості моделі, що визначається розміром ентропії, модифікований метод Кнесера–Нея [1] буде найкращим для даної задачі, проте потребуватиме додаткового навчання для

визначення діапазонів, що у випадку використання псевдочастот замінять фіксовані рівні відбору.

**Чисельні експерименти**

Експерименти проведено на *n*-грамах розмірності $\leq 3$, зібраних зі стенограм Верховної Ради України. Було сформовано корпус обсягом 112,5 Мб, для чого відповідні стенограми зібрано з сайту *http*://*rada.gov.ua*/ *meeting*/*stenogr*.

На корпусі було виділено словник системи з 10 тис. словоформ, всі інші слова замінені на стоп-слово «#». Словник пропущено через систему морфолексичного аналізу, і сформовано словники канонічних форм та словники граматичних класів. При цьому множина граматичних класів однозначно визначає словник канонічних форм.

Певні обмеження введені на всіх словниках: службові частини мови не розділяються на форми, іменники родового відмінку також не розділяються на форми. Отже, словники канонічних форм містять, крім власне форм, ще і додаткові слова, виділені як окремі умовні канонічні форми, які насправді є лише словоформами. Аналогічно з умовними граматичними класами.

Необхідно зауважити, що словникова система, використана в аналізі, не є повною і не описує всі можливі граматичні ознаки за українською граматикою, а лише частину. Тому для уникнення втрат інформації в деяких випадках розділення не виконувалося.

Перша пара словників визначається такими параметрами: словник канонічних форм обсягом 7012 одиниць, та словник граматичних класів обсягом 5409 одиниць. Для дієслів не виконується розділення на канонічні форми та граматичні класи. Це пов'язано з тим припущенням, що перехідні та неперехідні дієслова потребують підмета у різних відмінках. Перехідні дієслова означають дію, спрямовану (переходить) на певний предмет, названий іменником або займенником у знахідному відмінку без прийменника: виконати (що?) вправу; зустріти (кого?) друзів; прочитати (що?) книгу. Якщо присудок у реченні вживається з заперечною часткою *не*, іменник ставиться не в знахідному, а в родовому відмінку. Неперехідні дієслова означають дію або стан, які на інший предмет не переходять і не потребують знахідного відмінка без прийменника від іменника чи займенника.

Друга пара словників визначається такими параметрами: словник канонічних форм обсягом 6044 одиниці та словник граматичних класів обсягом 3770 одиниць. В цих словниках для всіх форм дієслів, крім інфінітива, виконано розділення на канонічні форми та граматичні класи.

Великий розмір другої пари часткових словників (понад 50 відсотків для словника канонічних форм, і понад 35 відсотків для словника граматичних класів) пояснюється тим, що, крім значної частки інфінітивів, у словник системи входить велика кількість слів з малим набором словоформ, що суттєво менше за можливий повний набір.

Також використано базовий словник системи, щоб дізнатися, чи покращує, чи погіршує метод роботу системи. Множина *n*-грам була фільтрована від три-грам частоти один, але після побудови таблиць $T(w_{i-n+1}^{i})$.

В усіх експериментах для побудови моделі використано 75 відсотків від корпуса, для обчислення ентропії та перплексії використано решту – 25 відсотків.

***Експеримент*** 1. Перевіряється припущення, описане формулою (2) для всіх *n*-грам моделей, для всіх способів згладжування як з фільтрацією,так і без неї. Як міру близькості вибрано добуток косинус кута між відповідними векторами частот канонічних форм та граматичних класів. Обчислення показали, що формула (2) не виконується навіть без фільтрації.

Було вирішено провести ще кілька експериментів, щоб перевірити, чи порушення умови (2) погіршує результат, і якщо так, то наскільки. В обчислені псевдочастот для заданої системи граматичних класів необхідно кожного разу будувати триграми на словоформах, і потім з них будувати біграми та уніграми, інакше буде некоректно обчислюватися формула (3).

***Експеримент*** 2. Порівнюються результати звичайної схеми Вітена–Белла. На основі отриманих комплектів словників за формулою (1) обчислюються відповідні триграми, а біграми та уніграми обчислюються на основі триграм.

**Таблиця 1.** Експеримент 2

| Умови | Ентропія | Перплексія |
|---|---|---|
| Базовий словник(фільтрація) | 7,623 | 197,129 |
| Набір 1 | 7,598 | 193,742 |
| Набір 2 | 7,674 | 204,222 |

Формула (1) дає незначне покращення порівняно з базовою моделлю, як вона описана раніше.

***Експеримент*** 3. Вводиться додаткова умова фільтрації. При обчисленні формули (1) ігноруються триграми з граматичних класів, якщо їх частота дорівнює одиниці. Аналогічно попередньому випадку, на основі отриманих комплектів словників за формулою (1) обчислюються відповідні триграми, а біграми та уніграми обчислюються на основі триграм.

**Таблиця 2**. Експеримент 3

| Умови | Ентропія | Перплексія |
|---|---|---|
| Базовий словник (фільтрація) | 7,623 | 197,129 |
| Набір 1 | 7,554 | 187,923 |
| Набір 2 | 7,511 | 182,404 |

Як видно з результатів експерименту, крім явного покращення, порівняно з базовим значенням, змінилося ранжування між першим та другим комплектом словників. Це пов'язано з суттєво іншим значенням $T(w_{i-n+1}^{i})$ з формули (6). Справді, при вказаній фільтрації кількість очікуваних типів триграм зменшується за рахунок найменш правдоподібних.

З огляду на результати першого та другого експериментів, можна стверджувати, що, якщо формула (1) порушує розподіл канонічних форм та граматичних класів, то, вибравши таку систему граматичних класів, щоб формула (2) виконувалася, можна буде покращити оцінку моделі, а отже і якість розпізнавання.

**Висновки.** Проаналізовано можливість застосування інформації про морфолексичні та граматичні характеристики слів для оптимізації моделей мови з метою покращення розпізнавання. Експерименти показали, що перерозподіл за запропонованою формулою (1) покращує оцінку ентропії, а отже потенційно покращує розпізнавання.

Показано, що спостережене порушення вимоги (2) не призводить до погіршення оцінки моделі, отже можна стверджувати, що коректно підібрана система граматичних класів дозволить значно покращити якість моделі, а отже і якість розпізнавання.


1. *Jurafsky D.*, *Martin J.H.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition // Prentice Hall PTR Upper Saddle River, NJ, 2000. – 934 p.
2. *Chen S.F.*, *Goodman J.T.* An empirical study of smoothing techniques for language modeling // Computer Speech and Language. – 1999. – N 13. – P. 448–453.
3. *Бабин Д.Н.*, *Мазуренко И.Л.*, *Холоденко А.Б.* О перспективах создания системы автоматического распознавания слитной устной русской речи // Интеллектуальные системы. – 2004. – **8**, **1**–**4**, – Р. 45–70.
4. *Кипяткова И.С.* Применение синтаксического анализа при создании *n*-граммной модели языка для систем распознавания русской речи / Пятый междисциплинарный семинар «Анализ разговорной русской речи» АР$^3$-2011. – СПб., 25–26 авг. 2011. – С 13–18.
5. *Language* model reduction for practical implementation in LVCSR systems / S. Ostrogonac, B. Popović, M. Sečujski et al. // Infoteh-Jahorina. – March 2013. – **12**. – P. 391–394.



Поступила 13.03.2014


Тел. для справок: +38 044 502-6319 (Киев)




В.Ю. Тарануха


## Метод сглаживания *n*-граммной модели для распознавания речи, основанной на классах, с использованием грамматической и лексической информации

**Введение.** Значительный рост объемов информации в виде цифровых аудиозаписей и изображений текстов требует эффективных средств, позволяющих переводить данные в текстовую форму для дальнейшей обработки. Стандартом де-факто есть использование статистической модели на основе *n*-грамм [1], и алгоритмов, с ней работающих. Такая модель в целом хорошо разработана [2], однако при использовании для славянских языков, в том числе для украинского, в сравнении с романо-германскими языками проявляется ряд недостатков, связанных с особенностями славянских языков. Предлагались различные подходы к решению этой проблемы: переход к *n*-граммам со свободным порядком слов [3], использование данных синтаксического анализатора [4], фильтрация на основе эвристик [5]. В настоящей статье предложена модификация классической модели с акцентом на лексические и грамматические классы.

**Построение и оценка моделей**

Обычно вероятностная модель полагает, что речь обладает свойствами, позволяющими описать ее как марковский процесс. Тогда вероятность последовательности слов можно будет оценить явно [1]. Последовательность слов языка $w_1 \ldots w_n$ называется *n*-граммой длины *n* и обозначается $w_1^n$. Последовательность слов можно предста-

вить как последовательность *n*-грамм, а вероятность оценить по формуле $p(w_1^i) = p(w_i \mid w_1^{i-1}) p(w_{i-1} \mid w_1^{i-2}) .... p(w_1)$. При этом можно построить оценку вероятностей, опирающихся на частоты соответствующих *n*-грамм, $\hat{p}(w_i \mid w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^i)}{C(w_{i-n+1}^{i-1})}$, где $C(w_{i-n+1}^i)$ – частота соответствующей *n*-граммы.

Характерной особенностью славянских языков есть свободный порядок слов в предложениях. При этом значимые слова имеют большое количество словоформ, поскольку в них хранится информация, указывающая на потенциальные синтаксические связи слова.

Если на одно слово английского языка приходится примерно 1,7 словоформы, то на одно слово украинского языка, в зависимости от выбранного словаря, может приходиться от 5,5 до 19,9 словоформ на одном и том же корпусе. Таким образом, при построении таблицы *n*-грамм, при $n = 2$, размер возрастает по крайней мере в 10,47 раза, а при $n = 3$ – в более чем 33 раза. Как следствие, значительное количество *n*-грамм приобретает малые значения частот, и оценка вероятностей становится значительно более чувствительной к выбросам и шумам, что и составляет серьезную проблему распознавания, препятствуя достижению таких же высоких показателей, как те, что получены для романо-германских языков. Еще одно свойство указанного подхода – то, что в реальных корпусах не представлены все возможные n-граммы. Это создает потребность в применении метода сглаживания частот и вероятностей для соответствующих *n*-грамм. Для оценки качества модели без необходимости проведения эксперимента с распознаванием используется энтропия

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x) .$$

Это позволяет оценить качество марковской цепочки, хотя определенные аспекты реального распознавания игнорируются. Например, игнорируется сходство звучания слов, что может привести к ухудшению результатов в реальном эксперименте.

Кроме того, в идеальном случае необходимо вычислять энтропию в расчете на слово на потенциально бесконечном предложении, описывающем язык. Однако в реальном эксперименте приходится обходиться выборкой определенного конечного размера, в надежде, что приближенное значение будет близким к теоретическому

$$H(L) = \lim_{n \to \infty} -\frac{1}{n} \log p(w_1 w_2 \ldots w_n) .$$

Для сравнения двух моделей удобно использовать кросс-энтропии. Пусть $m(w_1w_2\ldots w_n)$ – модель для вероятности $p(w_1w_2\ldots w_n)$, тогда кросс-энтропия в расчете на слово выражается так: $H(p,m) = \lim_{n \to \infty} -\frac{1}{n} \log m(w_1 w_2 \ldots w_n)$, когда известно, что $H(p) \le H(p, m)$.

Также при необходимости можно применить более подробную меру качества, называемую *перплексией* и рассчитываемую следующим образом: $PPW = 2^{H(p,m)}$.

В реальном тесте можно наблюдать *n*-грамму, отсутствующую в корпусе, по которому строилась модель. Для оценки события, которое не встречалось в корпусе, используют методы сглаживания, описываемые далее.

**Анализ известных методов для повышения качества модели**

***Переход к n-граммам со свободным порядком слов*** [3].: $\hat{p}(w_i \mid w_{i-n+1}^{i-1}) = p(w_i \mid \{w_{i-n+1} .... w_{i-1}\})$. Фигурные скобки обозначают, что слова, кроме последнего, подбираются в произвольном порядке. Численный эксперимент [3] показал непригодность этой модели для распознавания при высокой энтропии.

***Использование данных синтаксического анализатора*** [4]. В комплект *n*-грамм добавляются *n*-граммы, полученные как корректные словосочетания, полученные анализом дерева синтаксического разбора. Получен гарантированный эффект повышения качества распознавания. Этот метод, к сожалению, предполагает необходимость использования соответствующего синтаксического анализатора, что не всегда возможно.

***Использование декомпозиции модели на две***: модель, основанную на грамматических классах, и модель, основанную на канонических формах слов. Согласно приведенным оценкам для перплексии [3], при одновременном использовании обеих частичных моделей результирующая перплексия значительно выше, чем перплексия модели, созданной только на словоформах. Это лишает смысла непосредственное использование двух моделей, хотя, согласно мнению авторов [3], оставляет возможность вынести в модель на грамматических классах часть информации, необходимой для борьбы с акустическим сходством различных форм одного слова.

***Использование оптимизации энтропии*** или перплексии модели путем оценки и устранения шумов из модели [5]. При этом можно оценивать как модель, собранную на словоформах, так и модель, собранную на канонических формах слов. Метод показал действенность такой оптимизации, однако он имеет ощутимый недостаток, поскольку опирается в оптимизации на *n*-граммы малой частоты, а следовательно, на шумы и выбросы.

**Новый метод**

Предлагается модификация классической модели, построенной на лексических и грамматических классах. При таком построении создается две модели *n*-грамм – отдельно на основе канонических форм слов (лексическая) и отдельно на основе грамматических классов слов, т.е. одна последовательность слов дает информацию в две различные частичные модели. Затем на основе двух моделей строится общая модель, использующая информацию из обеих частичных моделей. При необходимости допускается фильтрация части полученных *n*- грамм, если это улучшит результат.

Для разбиения на классы в общем случае вводится функция разбиения, что приводит в соответствие каждому слову $w_i$ из словаря $V$ класс $c_i$. При этом выполняется $P(w_i \mid w_1^{i-1}) = P(w_i \mid c_i) P(c_i \mid c_1^{i-1}), \forall i, 1 \le i \le n$ .

Идея заключается в том, что для слов, о которых известно, что они имеют одинаковое синтаксическое поведение, можно сделать предположение о том, что и в похожих контекстах они должны иметь схожие вероятности встречаемости в контексте.

Пусть для слов *автомобиль*, *автомобиля*, *вертолет*, *вертолетом*, *синий*, *синего*, *желтый*, *желтым* в корпусе наблюдались биграммы *синий автомобиль*, *синего автомобиля*, *желтый вертолет*, *желтым вертолетом*. Тогда можно выполнить поэтапную свертку по каноническим формам (*автомобиль*, *вертолет*) и (*синий*, *желтый*), поэтапную свертку по грамматическим классам *ед.*, *м.род.*, *им. падеж*, *ед.*, *м.род.*, *род. падеж*, *ед.*, *м.род.*, *тв. падеж*. На базе знаний о том, что эти существительные и прилагательные ведут себя похожим образом, т.е. имеют одинаковые множества грамматических классов, можно построить предположение о вероятности появления их в формах, не наблюдавшихся в корпусе.

Учитывая, что в украинском языке наблюдается омонимия, обозначим: $L(w_1^k)$ – совокупность последовательностей канонических форм для последовательности слов $w_1^k$. $G(w_1^k)$ – совокупность последовательностей грамматических классов для последовательности слов $w_1^k$. Обозначим $El(w_1^k)$ – совокупность последовательностей слов, которые после приведения к каноническим формам имеют одинаковую запись, совокупность $w_{i_1}^{i_k}$, таких, что $L(w_{i_1}^{i_k}) = L(w_1^k), \forall i$.

Тогда оценка частоты $w_1^k$ определяется так:

$$C(w_1^k) = \frac{C(L(w_1^k))C(G(w_1^k))}{\sum\limits_{G_F \in G(El(w_1^k))} G_F} \quad . \qquad (1)$$

Для обеспечения достоверности объединенной модели выдвигается предположение: необходимо, чтобы сумма частот канонических форм и сумма частот грамматических классов после перераспределения оставалась неизменной:

$$G(C(w_1^{k-1})) = \sum_{m=1}^{V} G(C(w_1^{k_m})) \;\&\; L(C(w_1^{k-1})) = \sum_{m=1}^{V} L(C(w_1^{k_m})), \qquad (2)$$

где $|V|$ – размер словаря, а соответствующие частоты $C(w_1^k)$ могут быть нулевыми, если для них не существует $w_1^k$ в корпусе, на котором строится модель. Если (2) выполняется, то это позволит существенно оптимизировать вычисления (1).

Возвращаясь к ранее приведенному примеру: данная модель позволяет по формуле (1) оценить вероятности для биграмм *синим автомобилем* и *желтого вертолета*.

Подбор параметров, а именно множества грамматических классов и множества канонических форм, был проведен во время численного эксперимента.

**Метод сглаживания**

Для сглаживания и заполнения пропусков используются различные методы [2], в данной статье рассматриваются сглаживания с возвращением Витте–Белла, поскольку оно простое и одновременно включает в себя все необходимые параметры:

$$\hat{p}(w_i \mid w_{i-n+1}^{i-1}) = \begin{cases} d(w_{i-n+1}^i), C(w_{i-n+1}^i) > 0, \\ \alpha_{w_{i-n+1}\dots w_{i-1}} \hat{p}(w_i \mid w_{i-n+2}^{i-1}) \quad \text{иначе}, \end{cases} \qquad (3)$$

где $d(w_{i-n+1}^i)$ – соответственно сглаженное значение $C(w_{i-n+1}^i)$, $\alpha_{w_{i-n+1}\dots w_{i-1}}$ – соответствующий коэффициент, определяющий вероятностную массу, перераспределенную для построения вероятностей на *n*-граммах модели низшего порядка:

$$\alpha_{w_{i-n+1}\dots w_{i-1}} = \frac{\beta_{w_{i-n+1}\dots w_{i-1}}}{\sum\limits_{\{w_i : C(w_{i-n+1}^i)=0\}} \hat{p}(w_i \mid w_{i-n+2}^{i-1})}, \qquad (4)$$

$$\beta_{w_{i-n+1}\dots w_{i-1}} = 1 - \sum_{\{w_i : C(w_{i-n+1}^i)>0\}} d(w_{i-n+1}^i) \; . \qquad (5)$$

Для метода Виттена–Белла параметр *d* оценивается так:

$$d_{WB}(w_i \mid w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^i)}{C(w_{i-n+1}^i) + T(w_{i-n+1}^i)}, \qquad (6)$$

где $T(w_{i-n+1}^i)$ – количество типов *n*-грамм, предшествующих слову $w_i$.

При этом, по умолчанию, *n*-граммы высокого порядка с частотой единица удаляются из модели.

Поскольку после применения формулы (1) *n*-граммы получат не частоты, а псевдочастоты, то сглаживание по формуле Виттена–Белла удобно, поскольку не требует регулировать, от какого значения псевдочастоты необходимо отчислять допустимые элементы в $T(w_{i-n+1}^i)$.

Метод Катца с возвращением не подходит в качестве дополнительного метода сглаживания, поскольку опирается на эвристику Гуда–Тьюринга [1], которая не имеет понятного способа интерпретации, если вместо частот в нее подставить псевдочастоты.

В зависимости от качества модели, определяемой размером энтропии, для данной задачи целесообразно использовать модифицированный метод Кнесера–Нея [1], однако потребуется дополнительное обучение для определения диапазонов, которые в случае использования псевдочастот заменят фиксированные уровни отбора.

**Численные эксперименты**

Эксперименты проведены на *n*-граммах размерности $\leq 3$, собранных из стенограмм Верховной Рады Украины. Сформирован корпус объемом 112,5 Мб. Для этого соответствующие стенограммы были собраны с сайта *http://rada.gov.ua/meeting/stenogr*.

На корпусе был выделен словарь системы из 10 тыс. словоформ, все остальные слова заменены на стоп-слово «#». Словарь пропущен через систему морфолексического анализа, сформированы словари канонических форм и словари грамматических классов. При этом множество грамматических классов однозначно определяет словарь канонических форм.

Определенные ограничения были применены ко всем словарям: как служебные части речи, так и существительные родительного падежа не получают грамматические формы. Таким образом, словари канонических форм содержат, кроме собственно форм, еще и дополнительные

слова, выделенные как отдельные условные канонические формы, которые на самом деле есть лишь словоформами. По аналогии с условными грамматическими классами.

Отметим, что словарная система, использованная при анализе, не полная и не описывает все возможные грамматические признаки по украинской грамматике, а лишь их часть. Поэтому во избежание потерь информации в ряде случаев разделение не выполнялось.

Первая пара словарей определяется следующими параметрами: словарь канонических форм объемом 7012 единиц и словарь грамматических классов объемом 5409 единиц. Для глаголов не выполняется разделение на канонические формы и грамматические классы. Это связано с предположением, что переходные и непереходные глаголы требуют управления в разных падежах. Переходные глаголы обозначают действие, направленное (переходящее) на определенный предмет, названный существительным или местоимением в винительном падеже без предлога: выполнить (что?) упражнение; встретить (кого?) друзей; прочитать (что?) книгу. Если сказуемое в предложении употребляется с отрицательной частицей *не*, существительное ставится не в винительном, а в родительном падеже. Непереходные глаголы обозначают действие или состояние, которые на другой предмет не переходят и не требуют винительного падежа без предлога от существительного или местоимения.

Вторая пара словарей определяется следующими параметрами: словарь канонических форм объемом 6044 единицы и словарь грамматических классов объемом 3770 единиц. В этих словарях для всех форм глаголов, кроме инфинитива, выполнялось разделение на канонические формы и грамматические классы.

Большой размер второй пары частичных словарей (более 50 процентов для словаря канонических форм и более 35 процентов для словаря грамматических классов) объясняется тем, что, помимо значительной доли инфинитивов, в словарь системы входит большое количество слов с малым набором словоформ, существенно меньший, чем возможен полный набор.

Также использовался базовый словарь системы, чтобы узнать, улучшает или ухудшает метод работу системы. Множество *n*-грамм было профильтровано от триграмм частоты один, но после построения таблиц $T(w_{i-n+1}^{i})$.

Во всех экспериментах для построения модели использовано 75 процентов от корпуса, для вычисления энтропии и перплексии использованы остальные 25 процентов.

***Эксперимент* 1**. Проверяется предположение, описанное формулой (2) для всех *n*-грамм моделей, для всех способов сглаживания как с фильтрацией, так и без нее. В качестве меры близости выбрано произведение косинус угла между соответствующими векторами частот канонических форм и грамматических классов. Вычисления показали, что формула (2) не выполняется даже без фильтрации.

Было принято решение провести еще ряд экспериментов, чтобы проверить, ухудшают ли результат нарушения условия (2), и если да, то насколько. При вычислении псевдочастот для заданной системы грамматических классов необходимо каждый раз строить триграммы на словоформах, а затем из них строить биграммы и униграммы, иначе формула (3) не будет вычисляться корректно.

***Эксперимент* 2**. Сравниваются результаты обычной схемы Виттена–Белла. На основе полученных комплектов словарей по формуле (1) вычисляются соответствующие триграммы, а биграммы и униграммы вычисляются на основе триграмм.

**Таблица 1**. Эксперимент 2

| Условия | Энтропия | Перплексия |
|---|---|---|
| Базовый словарь(фильтрация) | 7,623 | 197,129 |
| Набор 1 | 7,598 | 193,742 |
| Набор 2 | 7,674 | 204,222 |

Формула (1) дает незначительное улучшение в сравнении с базовой моделью в том виде, в каком она описана ранее.

***Эксперимент* 3**. Вводится дополнительное условие фильтрации. При исчислении формулы (1) игнорируется триграмма из грамматических классов, если их частота равна единице. Аналогично предыдущему случаю, на основе полученных комплектов словарей по формуле (1) вычисляются соответствующие триграммы, а биграммы и униграммы исчисляются на основе триграмм.

**Таблица 2**. Эксперимент 3

| Условия | Энтропия | Перплексия |
|---|---|---|
| Базовый словарь (фильтрация) | 7,623 | 197,129 |
| Набор 1 | 7,554 | 187,923 |
| Набор 2 | 7,511 | 182,404 |

Как видно из результатов эксперимента, кроме явного улучшения в сравнении с базовым значением изменилось ранжирование между первым и вторым комплектом словарей. Это связано с существенно иным значением $T(w_{i-n+1}^{i})$ из формулы (6). Действительно, при указанной фильтрации количество ожидаемых типов триграмм уменьшается за счет наименее правдоподобных.

Учитывая результаты первого и второго экспериментов, можно утверждать, что, если формула (1) нарушает распределение канонических форм и грамматических классов, то, выбрав такую систему грамматических классов, чтобы формула (2) выполнялось, можно будет улучшить оценку модели, а значит, и качество распознавания.

**Заключение**. Проанализирована возможность применения информации о морфолексических и грамматических характеристиках слов для оптимизации моделей языка с целью улучшения распознавания. Численные эксперименты показали, что перераспределение по предложенной формуле (1) улучшает оценку энтропии, и, следовательно, потенциально улучшает распознавание.

Показано, что наблюдаемое нарушение требования (2) не приводит к ухудшению оценки модели, так что можно утверждать, что корректно подобранная система грамматических классов, позволит улучшить качество модели, а значит, и качество распознавания.