# АНАЛІЗ КОРПУСУ УКРАЇНСЬКОЇ МОВИ МЕТОДАМИ МОДЕЛІ ПРИХОВАНОГО ЛАНЦЮГА МАРКОВА

Є.М. ОКУНЄВ

В цьому дослідженні ми проаналізували фрагмент корпусу української мови методами, розробленими для аналізу моделі прихованого ланцюга Маркова, аналогічно до дослідження [1]. В межах моделі розвиваються два випадкові процеси: процес Маркова, за яким ми не можемо спостерігати та який, певним чином, впливає на еволюцію другого - спостережуваного процесу. Модель прихованого ланцюга Маркова формально визначається як впорядкована трійка $\lambda = (\pi, A, B)$, де $\pi$- - вектор розподілу початкового стану прихованого процесу Маркова, $A$ - матриця перехідних ймовірностей прихованого процесу Маркова та $B$ - матриця умовних ймовірностей спостережуваного процесу для різних станів прихованого процесу Маркова. Хоча у визначенні про це не йшлося, для подальшої роботи нам також знадобиться вектор із інформацією про перебіг спостережуваного процесу $O$.

Найцікавішим результатом теорії моделей прихованого ланцюга Маркова є алгоритм Баума-Велша. Це ітеративний алгоритм, який модифікує параметри моделі, щоб максимізувати ймовірність вектору спостережень $O$ [2]. Якщо розглядати модель прихованого ланцюга Маркова як нейронну мережу, то застосування алгоритму Баума-Велша відповідає процесу неконтрольованого навчання нейронної мережі. Ми застосували алгоритм Баума-Велша до частини корпусу української мови в якості вектора спостережень та розділили літери українського алфавіту на 2, 3 та 4 статистично значущі групи. Також ми провели порівняльний аналіз відомими розподілами літер на групи, що застосовуються в української філології: наприклад поділ літер українського алфавіту на дві статистично значущі групи в результаті дає поділ літер на голосні та приголосні.

Наведемо таблицю розподілу літер на дві групи: на перетині рядку $i$ та стовпця $j$ знаходяться наступні ймовірності:

$$P\left(\text{Спостерігаємо літеру } X_i | \text{Літера належить до групи } j\right)$$

Можна побачити, що більшість літер мають одну нульову та одну ненульову умовні ймовірності належності до певної групи. В дослідженні

ми вважаємо, що літера належить до тої групи, де умовна ймовірність - найбільш. В результаті розподілу літер на дві статистично значущі групи алгоритм Баума-Велша розділив літери на голосні та приголосні із високою точністю. Алгоритм зробив хибний висновок про літери ґ,є,ї,ь. Достеменно не можна сказати чим викликані ці помилки, але можна зробити обмірковані здогадки: ґ - дуже рідкісна літера і висновок стане правильним при застосуванні більшої частини корпусу; є,ї - дифтонги, які починаються на звук [й]; ь - літера, яка не має власного звуку і, відповідно, не відноситься до жодної з груп.

Філологічні відомості для інтерпретації результатів взяті з [3] .

Табл. 1. Розподіл літер на дві групи

| Літера | Група 1 | Група 2 | Літера | Група 1 | Група 2 |
|---|---|---|---|---|---|
| А | 0.21239 | 0 | Н | 0 | 0.09203 |
| Б | 0 | 0.03651 | О | 0.20561 | 0 |
| В | 0 | 0.09022 | П | 0.00010 | 0.042601 |
| Г | 0 | 0.02689 | Р | 0 | 0.06599 |
| ґ | 0.00003 | 0 | С | 0.00049 | 0.07362 |
| Д | 0 | 0.05632 | Т | 0 | 0.09845 |
| Е | 0.12115 | 0 | У | 0.08790 | 0 |
| Є | 0 | 0.01750 | Ф | 0 | 0.00021 |
| Ж | 0 | 0.02053 | Х | 0 | 0.02239 |
| З | 0.00166 | 0.03692 | Ц | 0 | 0.00786 |
| И | 0.15188 | 0 | Ч | 0 | 0.02564 |
| І | 0.10994 | 0.00047 | Ш | 0 | 0.01918 |
| Ї | 0.00290 | 0.00603 | Щ | 0 | 0.01180 |
| Й | 0 | 0.02461 | Ь | 0.04635 | 0 |
| К | 0 | 0.07672 | Ю | 0.00847 | 0.01119 |
| Л | 0 | 0.07247 | Я | 0.05072 | 0.00911 |
| М | 0 | 0.05460 | | | |

## Література

[1] R. L. Cave and L. P. Neuwirth, Hidden Markov models for English, Princeton, NJ, IDA-CRD, October 1980.

[2] Nejati, A., Unsworth, C. (2014). A Concise Information-Theoretic Derivation of the Baum-Welch algorithm. arXiv preprint arXiv:1406.7002.

[3] Наконечна, Л. Б. (2014). Сучасна українська мова. Фонетика. Орфоепія. Морфонологія. Графіка. Орфограгія (видання 2-е). Івано-Франківська,НАІР.

КПІ ім. Ігоря Сікорського, Київ, Україна
*Email address*: egorky96@gmail.com