Skip-bigrams reconstruct trigrams in 2-word languages

Shohei Hidaka[1]
[1]Japan Advanced Institute of Science and Technology
shhidaka@jaist.ac.jp

Abstract

In natural language processing, it has been empirically known that skip-grams, co-occurrence statistics of two words with some number of words in between them, is an effective source of data to learn semantic nature of the words. In this study, we propose a new theoretical account for why a set of skip-grams is effective at least for two-word languages, by giving a theorem that a set of trigram probabilities is representable with a set of skip bigrams. This representation theorem justifies the use of skip bigrams or so-called shiftgrams as a computationally efficient source to access higher order n-gram.

1 Effectiveness of skip-gram statistics

In natural language processing, it has been empirically known that semantic structure of words are represented by the word vector by learning the skip-grams [1, 2], which is co-occurrence statistics (X_t, X_{t+s}) of a pair of word at t and word $t+s$ with a skip length $s = 1, 2, \ldots$. There are previous studies that have tried to explain this empirical finding [3, 4, 5]. Most of such previous studies have hypothesized that the word vector gives a effective representation due to their special settings of the learning scheme of the word vector models (i.e., negative sampling) [3].

In this study, we take an apprroach distinct from these previous studies, and mathematically analyze the fundamental nature of language systems, represented by n-gram statistics. Our primary focus is how trigram statistics, the conditional probability $P(X_{t+2}|X_t, X_{t+1})$, can be represented by a set of s-skip bigrams, $P(X_{t+s}|X_t)$ for $s = 1, 2, \ldots$. There is a trivial relationship that s-skip bigrams $P(X_{t+s}|X_t)$ for each $s = 1, 2, \ldots$ is constructed by a given trigram $P(X_{t+2}|X_t, X_{t+1})$ of a Markov process. Our question is the converse – can we construct the trigram only from a set of s-skip bigrams? This is a focus special case, that may be generalized to the relationship between s-skip bigrams and a general n-grams. If such fundamental relationship between (n-1)-grams and n-grams is established, it would explain why skip-gram statistics is a good source of data to learn semantic nature of words or language in general – skip-gram gives a sufficient statistics of n-grams and is computable efficiently.

2 Skip bigram

In this study, we assume a language L has a set of k words $\mathbb{W}_k := \{0, 1, 2, \ldots, k-1\}$, and we call a Markov process over a series of $X_0, X_1, \ldots \in \mathbb{W}_k$ **language system**. In particular, a language system is called n-**grams** of L, if $P(X_t|X_{t-1}, X_{t-2}, \ldots, X_{t-n-s}) = P(X_t|X_{t-1}, X_{t-2}, \ldots, X_{t-n})$ for any t and $s = 0, 1, 2, \ldots$. So any n-gram language system with k words has $(k-1)k^{n-1}$ parameters, those are the conditional probabilities $P(X_t|X_{t-n+1}, \ldots, X_{t-2}, X_{t-1}) \geq 0$ with $\sum_{X_t \in \mathbb{W}_k} P(X_t|X_{t-n+1}, \ldots, X_{t-2}, X_{t-1}) = 1$. In this study, we assume any language system under analysis is ergodic, or equivalently it has a unique set of stationary probabilities.

To encode the joint random variables of m-series, without loss of generality, we fix the encoder map $h_{k,m} : \mathbb{W}_k^m \to C_{k,m} := \{1, 2, \ldots, k^m\}$ by

$$h_{k,m}(X_{t-m+1}, X_{t-m+2}, \ldots, X_t) := 1 + \sum_{j=1}^{m} \left(X_{t-m+j} - 1\right) k^{j-1}. \tag{1}$$

In this encoding of the joint random variables, the transition matrix $Q_2 \in \mathbb{R}^{k \times k}$ of any bigram language system is of the form

$$Q_2 := \begin{pmatrix} q_{0|0} & q_{0|1} & \cdots & q_{0|k-1} \\ q_{1|0} & q_{1|1} & \cdots & q_{1|k-1} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k-1|0} & q_{k-1|1} & \cdots & q_{k-1|k-1} \end{pmatrix}, \tag{2}$$

where $q_{i|j} := P(X_t = i|X_{t-1} = j)$ and $\sum_{i \in \mathbb{W}_k} q_{i|j} = 1$ for any $j \in \mathbb{W}_k$. Moreover, the transition matrix $Q_3 \in \mathbb{R}^{k^2 \times k^2}$

of any trigram language system is of the form

$$Q_3 := \sum_{i,j\in\mathbb{W}_k} e_{k,i} \otimes e_{k,j} e_{k,j}^\top \otimes r_{i,j}, \tag{3}$$

where $r_{i,j} = (q_{i|(0,j)}, \ldots, q_{i|(k-1,j)})$.

Let $\theta_2 = (\theta_0, \theta_1, \ldots, \theta_{k-1})^\top \in \mathbb{R}^k$ be the stationary probability vector of the bigram system such that $\theta_2 = Q_2\theta_2$, and $\theta_3 = (\theta_{(0,0)}, \theta_{(1,0)}, \ldots, \theta_{(k-1,k-1)})^\top \in \mathbb{R}^{k^2}$ be the stationary probability vector of the trigram system such that $\theta_3 = Q_3\theta_3$.

2.1 Tensor form and tensor product

The set of n-gram conditional probabilities $P(X_t|X_{t-1}, \ldots, X_{t-n+1})$ is naturally represented by a n^{th} order tensor (n-tensor in short). Real-valued n-tensor $\mathbb{R}^{k_1\times k_2\times\ldots\times k_n}$ is a vector space of real-valued maps $\{0, \ldots, k_1 - 1\} \times \ldots \times \{0, \ldots, k_n - 1\} \to \mathbb{R}$. Let us denote $\mathbb{R}^{k^n} := \mathbb{R}^{k_1\times k_2\times\ldots\times k_n}$ for $k = k_1 = k_2 = \ldots = k_n$.

We call a tensor product $\star : \mathbb{R}^{k^n} \times \mathbb{R}^{k^n} \to \mathbb{R}^{k^n}$ **convolution** defined by

$$P \star Q := \tag{4}$$
$$\sum_{X_{t-1}\in K} P(X_t, X_{t-1}, \ldots, X_{t-n+1}) Q(X_{t-1}, X_{t-2}, \ldots, X_{t-n}),$$

for $P, Q \in \mathbb{R}^{k^n}$. In particular, denote for $m \geq 0$

$$Q^m := \begin{cases} E_{n,k} & \text{if } m = 0, \\ Q^{m-1} \star Q & \text{otherwise} \end{cases}, \tag{5}$$

where $E_{n,k} \in \mathbb{R}^{k^n}$ is the left unit tensor satisfying $E_{n,k} \star Q = Q$ for any tensor $Q \in \mathbb{R}^{k^n}$. Specifically,

$$E_{n,k}(i_1, i_2, \ldots, i_n) = \begin{cases} 1 & \text{if } i_1 = i_2 \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

This convolution is useful to represent a time-shift operation in the following sense: If $Q \in \mathbb{R}^{k^n}$ is time-invariant n-gram conditional probability $Q(X_t, X_{t-1}, \ldots, X_{t-n+1}) = P(X_t|X_{t-1}, \ldots, X_{t-n+1})$, the convolution of k^{th} power represents **shift** in the time step of random variables:

$$Q^k(X_t, X_{t-k}, \ldots, X_{t-n-k+2}) \tag{7}$$
$$= P(X_t|X_{t-k}, \ldots, X_{t-n-k+2}). \tag{8}$$

Define reduction operator $r_{\theta,I} : \mathbb{R}^{k^n} \to \mathbb{R}^{k^{|I|}}$ for any $I \subseteq N = \{1, 2, \ldots, n\}$ for $Q \in \mathbb{R}^{k^n}$ by

$$r_I(Q) = \sum_{(i_j)_{\in\{1,\ldots,n\}\setminus I}\in K^{n-|I|}} Q(i_1, i_2, \ldots, i_n), \tag{9}$$

m-shifgram $S_m : \mathbb{R}^{k^n} \to \mathbb{R}^{k^2}$ is defined by

$$S_m(Q) := r_{\{1,2\}}\left(Q^m \Theta_Q\right), \tag{10}$$

where Θ_Q is the stationary tensor associated to Q.

The m-shiftgram of n-gram tensor $Q \in \mathbb{R}^{k^n}$ has the following properties. For $m' = 2, 3, \ldots, n$,

$$S_m(Q) = r_{\{1,m'\}}\left(Q^{m-m'+2}\Theta_Q\right). \tag{11}$$

3 Inverse problem

3.1 effective isomorphism between trigrams and shiftgrams

Suppose we have the series of all m-shiftgrams $S(Q) := \{S_m(Q)\}_{m=0,1,\ldots}$ of some unknown n-gram probability tensor $Q \in \mathbb{R}^{k^n}$. Then can we uniquely identify the original probability tensor Q that generates $S(Q)$? Let us focus on $n = 3$ in this paper. For each fixed $m' = 0, 1, \ldots$, we have $n - 1$ different m-shiftgrams $S_m(Q) = r_{\{1,(m+2-m')\}}\left(Q^{m'}\Theta_Q\right)$ for $m' \leq m \leq m' + n - 2$ due to the identity (11). For each m', Q is constrained by $(n - 1)$ matrices of m-shiftgrams $S_m(Q)$ and the sum $\sum_{i\in K} Q^{m'}(i, j, k) = 1$ is also constrained. So $Q^{m'}$ may have at most $(k - 1)^n$ polynomial equations, but only $(k - 1)^{n-1}$ equations are new constraints not expressed by (11) for $m' \geq n - 1$. Thus, there are at most $k^2 + k(k-1) + m'(k-1)^2$ polynomial equations for a series of $S_0(Q), S_1(Q), \ldots, S_{m'}(Q)$, and thus at least $m' \geq k$ to have the sufficient number k^3 of polynomial equations to identify $Q \in \mathbb{R}^{k^n}$.

3.2 Case with $k = 2$ and $n = 3$

To be specific, let us study $k = 2$ and $n = 3$ as a minimal example. In this case $S_1(Q), S_2(Q)$ is needed to have a sufficient number of equations. Let $\left(\begin{array}{c|c|c|c} Q_1 & Q_2 & \ldots & Q_m \end{array}\right)$ denote the third order tensor by series of matrices $Q(i, j, k) = Q_k(i, j)$ for $i, j, k \in K$. For $k = 2$ and $n = 3$, the trigram probability tensor is

$$Q = \left(\begin{array}{cc|cc} q_{00} & q_{10} & q_{01} & q_{11} \end{array}\right) \tag{12}$$
$$= \left(\begin{array}{cc|cc} q_{0|00} & q_{0|10} & q_{0|01} & q_{0|11} \\ q_{1|00} & q_{1|10} & q_{1|01} & q_{1|11} \end{array}\right), \tag{13}$$

where $q_{ij} = (q_{0|ij}, q_{1|ij})^\top \in \mathbb{R}^2$, and

$$Q^2 = \left(\begin{array}{cc|cc} (q_{00}, q_{10})\, q_{00} & (q_{01}, q_{11})\, q_{10} & (q_{00}, q_{10})\, q_{01} & (q_{01}, q_{11})\, q_{11} \end{array}\right). \tag{14}$$

Thus, with the stationary tensor $\Theta_Q(i,j,k) := \theta_{jk}$, the m-shiftgrams for $m = 1, 2, 3$ are

$$S_1(Q) = \sum_{j\in\{0,1\}} q_{ij}\theta_{ij}e_{n,i}^\top. \tag{15}$$

$$S_2(Q) = \sum_{i\in\{0,1\}} q_{ij}\theta_{ij}e_{n,j}^\top. \tag{16}$$

$$S_3(Q) = \sum_{i,j\in\{0,1\}} (q_{0i}, q_{1i})\, q_{ij}\theta_{ij}e_{n,j}^\top. \tag{17}$$

With (15), (16), and the sum-to-one constraint for $i, j \in K$

$$\mathbf{1}_k^\top q_{i,j} = 1, \tag{18}$$

7 independent linear equations are for $Q \in \mathbb{R}^{2^3}$ by fixing 1, 2-shiftgrams.

Lemma 1 For $k = 2$ and $n = 3$, there are at most two trigram probability tensors Q satisfy (18), (15), (16), and (16) for a given $S_1(Q)$, $S_2(Q)$, and $S_3(Q)$, if

$$\theta_{ij} = \mathrm{e}_{k,i}^\top S_1(Q)\mathrm{e}_{k,j} \tag{19}$$

$$\mathbf{1}_k^\top S_1(Q) = \mathbf{1}_k^\top S_2(Q) \tag{20}$$

$$S_1(Q)\mathbf{1}_k = S_2(Q)\mathbf{1}_k. \tag{21}$$

Otherwise, there is no Q satisfying the equations (18), (15), (16), and (17).

Proof Here we explicitly solve the equations (18), (15), (16), and (16) by letting the tensor Q as its variables. Specifically, the vectorized variables $\mathrm{vec}\,(Q) \in \mathbb{R}^{k^n}$ is required to be in the kernel of the matrix $C\mathrm{vec}\,(Q) = s \in \mathbb{R}^{3k^2}$ such that:

$$C := \mathrm{e}_{3,1} \otimes I_{k^2} \otimes \mathbf{1}_k^\top + \mathrm{e}_{3,2} \otimes \mathbf{1}_k^\top \otimes I_{k^2} + \mathrm{e}_{3,3} \otimes I_k \otimes \mathbf{1}_k^\top \otimes I_k \tag{22}$$

$$s := \mathrm{e}_{3,1} \otimes \mathrm{vec}\left(\Theta_Q\right) + \mathrm{e}_{3,2} \otimes \mathrm{vec}\,(S_1(Q)) + \mathrm{e}_{3,3} \otimes \mathrm{vec}\,(S_2(Q)) \tag{23}$$

This equation $C\mathrm{vec}\,(Q) = s$ gives a set of 7 independent linear equations, only if (19), (20), and (21) holds. Specifically, the solution is $q_{i|jk} = a_{ijk}x + b_{ijk}$ for each $i, j, k \in K$ with any $x \in \mathbb{R}$, where

$$a_{ij} = \theta_{ij}^{-1}(-1)^{i+j}(1,-1)^\top \tag{24}$$

$$b_{ij} = \theta_{ij}\mathrm{e}_{k,2} + \left(\delta_{1i}\delta_{0j}S_{00}^{(2)} + \delta_{0i}\delta 1 j S_{00}^{(1)} + \delta_{1i}\delta_{1j}\left(S_{01}^{(1)} - S_{00}^{(2)}\right)\right)(1,-1)^\top \tag{25}$$

Inserting $q_{i|jk} = a_{ijk}x + b_{ijk}$ to (17), it gives a quadratic equation

$$\alpha x^2 + \beta x + \gamma = 0, \tag{26}$$

where

$$\alpha = \sum_{i,j\in K} \theta_{ij}^{-1} \tag{27}$$

$$\beta = \theta_{00}(2a_{000}b_{000} + a_{010}b_{100} + b_{010}a_{100}) + \theta_{01}(a_{000}b_{001} + b_{000}a_{001} + a_{010}b_{101} + b_{010}a_{101}) \tag{28}$$

$$\gamma = (b_{0|00}, b_{0|10})\begin{pmatrix} b_{00} & b_{01}\end{pmatrix}(\theta_{00}, \theta_{01})^\top. \tag{29}$$

This quadratic equation has the leading coefficient $\alpha \neq 0$. Thus, it has at most two probability tensors Q satisfying the equations, unless the quadratic equation has a factor $(a_{ijk}x + b_{ijk} - q_{i|jk})$ for some $i, j, k \in \mathbb{K}$. □

4 Summary and Conjecture

Lemma 1 demonstrates a given set of m-shiftgrams is generally sufficient to reconstruct trigrams in two-word languages up to finite samples (there two possible trigram probability tensors Q). We expect that this special lemma can be probably extended to any general $k > 2$, and perhaps for $n > 3$ as well. This putative generalized theorem would fully explains why a set of m-shiftgrams or skip-bigrams approximates n-gram probabilities well. Also this generalized theorem would give mapping how higher n-grams are embedded into a series of m-grams, and the number of such maps will be bounded by the number of words k, which is much smaller than an exponential function of n. Thus, it may open up a theoretical explanation why n-grams, with an exponential number of combinations, can be learned efficiently.

To tackle further general cases with more words $k > 2$ and higher $n > 3$-grams, we need to understand how convolution $\star$ behaves over n-gram tensor and which algebra is suitable to understand such tensor operations.

Acknowledgements

This work was supported by JSPS KAKENHI JP23H0369, JST PRESTO JPMJPR20C9.

References

[1] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. 2013.

[2] A Vaswani. Attention is all you need. **Advances in Neural Information Processing Systems**, 2017.

[3] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In **Advances in Neural Information Processing Systems**, 2014.

[4] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 483–495, 2018.

[5] Takuma Torii, Akihiro Maeda, and Shohei Hidaka. Distributional hypothesis as isomorphism between word-word co-occurrence and analogical parallelograms. **PloS one**, Vol. 19, No. 10, p. e0312151, 2024.