Vol.2015-NL-221 No.14
Vol.2015-SLP-106 No.14
2015/5/26

対数的共起ベクトルの加法構成性

田 然[1,a)]　岡崎 直観[1,b)]　乾 健太郎[1,c)]

概要：この論文では、単語ベクトルの算術平均によって短いフレーズの意味を近似できる理由について初めての数学的解明を行う。具体的には、その近似による「誤差」に対する上界が理論的に与えられ、実験的に検証された。このような加法構成性が成り立つ必要条件として、対数関数と文脈のオーバーラップが重要であることや、低い共起頻度を Zipf 則に従って補完するのが有効であることなど、理論上予測される幾つかの性質も実験によって確かめられた。更に、加法構成性を考える上では、特異値分解による単語埋め込みは、最先端な埋め込み手法に匹敵する性能を達成できることを示す。

1. Introduction

Additive composition has been a commonly used baseline method since the advent of compositional distributional semantics, in which averages of individual word vectors are used to represent the meanings of longer linguistic sequences [5], [10]. Despite the considerable research that has been devoted to the exploration of more advanced composition frameworks [1], [2], [4], [17], [19], [21], [22], [25], additive composition remains a simple and effective way of handling phrase semantics. For example, [24] uses additive composition in a logic-based textual entailment recognition system, by scoring paraphrase candidates (e.g., "*blamed for death*" and "*cause loss of life*") using the cosine similarity between sums of word vectors (e.g., **blamed** + **death** and **cause** + **loss** + **life**).

However, the theoretical underpinnings of additive composition have so far been less clear. In this paper, we provide the first mathematical analysis of additive composition, and prove that the context vector of a bigram can be approximated by the average of the context vectors of its two words, given certain conditions and regarding a particular type of context vectors. More precisely, for a target $t \in T$ (i.e., a unigram or bigram), the context of t is derived from the event frequency $\mathrm{freq}(c,t)$ of a word $c \in C$ occurring within a window of t in a corpus (Table 1). In order to formulate the context vector $\mathbf{w}_t$, we sort the context lexicon C and use the i-th context word $c_i \in C$ to define the i-th entry of $\mathbf{w}_t$, as $s(c_i,t) := \ln \mathrm{freq}(c_i,t) - \alpha(c_i) - \beta(t)$. Therefore, $\mathbf{w}_t$ is formally defined as $\mathbf{w}_t := (s(c_i,t))_{i=1}^{|C|}$.

*These young women often **face difficulty** in acquiring needed resources . . .*

target	context
face_difficulty	*These, young, women, often, in, acquiring, needed, resources*
face	*These, young, women, often, difficulty, in, acquiring, needed, resources*
difficulty	*These, young, women, often, face, in, acquiring, needed, resources*

表 1　A context window of size 4 to each side for the bigram target "*face_difficulty*", and context windows of size 5 for the unigrams "*face*" and "*difficulty*".

The function $s(c_i,t) := \ln \mathrm{freq}(c_i,t) - \alpha(c_i) - \beta(t)$ represents the "strength" of c_i, occurring as a context of t. If c_i and t co-occur frequently, $\ln \mathrm{freq}(c_i,t)$ becomes relatively large, and so does $s(c_i,t)$. The terms $\alpha(c_i)$ and $\beta(t)$ are "shift" functions, to be specified later. This family of strength functions $s(c_i,t)$ contains special cases, such as the log-likelihood $\ln \Pr(c_i|t)$ (when $\alpha(c_i) = 0$ and $\beta(t) = \ln \mathrm{freq}(t)$), and the point-wise mutual information $\mathrm{PMI}(c_i,t)$ (when $\alpha(c_i) = \Pr(c_i)$ and $\beta(t) = \ln \mathrm{freq}(t)$). We also discuss low-dimensional reductions of $\mathbf{w}_t$ (i.e., matrix factorizations of $s(c_i,t)$), which include state-of-the-art word embeddings, such as the skip-gram model with negative sampling (SGNS) [16] and the GloVe model [20] (Section 3). Our theory provides insights into the performances of these models, regarding additive compositionality.

The main result of this paper (Section 2) is a theoretical upper bound for the Euclidean distance $\|\mathbf{w}_{t_1t_2} - \frac{1}{2}(\mathbf{w}_{t_1} + \mathbf{w}_{t_2})\|$, which represents the "error" in the approximation of the context vector $\mathbf{w}_{t_1t_2}$ of a bigram t_1t_2 by the average of the two vectors $\mathbf{w}_{t_1}$ and $\mathbf{w}_{t_2}$. We show

1　東北大学
a)　tianran@ecei.tohoku.ac.jp
b)　okazaki@ecei.tohoku.ac.jp
c)　inui@ecei.tohoku.ac.jp

that, as the bigram t_1t_2 occurs more often, the error $\|\mathbf{w}_{t_1t_2} - \frac{1}{2}(\mathbf{w}_{t_1} + \mathbf{w}_{t_2})\|$ has a smaller upper bound.

Furthermore, our analysis provides the following suggestions that have never been discussed from a theoretical viewpoint so far:

(A) We can generalize Zipf's law [26], an empirical law on word occurrences $\mathrm{freq}(c_i)$, to the co-occurrence frequencies $\mathrm{freq}(c_i, t)$ of any fixed target t. From this generalization of Zipf's law, we can derive the distribution of entries of the context vector $\mathbf{w}_t$, suggesting: **(A1)** the logarithmic function in $s(c_i, t)$ is important, in that a non-logarithmic strength, such as $s(c_i, t) := \Pr(c_i|t)$, may *not* yield similar upper bounds that guarantee additive compositionality (Section 2.1); **(A2)** for rarely seen (c_i, t) pairs, in particular when $\mathrm{freq}(c_i, t) = 0$ and $\ln \mathrm{freq}(c_i, t) = -\infty$, it is natural to complement co-occurrence frequencies according to the generalized Zipf's law (Section 2.1).

(B) The key observation to the proof of our main result is that when two unigrams t_1 and t_2 appear successively in a corpus (and if the context window size is not very small), the contexts of t_1 and t_2 have a large overlap (Table 1). Therefore: **(B1)** if the bigram t_1t_2 occurs often, then $\mathbf{w}_{t_1}$, $\mathbf{w}_{t_2}$, and $\frac{1}{2}(\mathbf{w}_{t_1} + \mathbf{w}_{t_2})$ are highly correlated (Section 2.2); **(B2)** during the addition $\mathbf{w}_{t_1} + \mathbf{w}_{t_2}$, components of $\mathbf{w}_{t_1}$ and $\mathbf{w}_{t_2}$ derived from the contexts where t_1 and t_2 appear independently tend to cancel each other out, whereas the component derived from bigram t_1t_2 reinforces itself. As a result, the average $\frac{1}{2}(\mathbf{w}_{t_1} + \mathbf{w}_{t_2})$ tends closer towards $\mathbf{w}_{t_1t_2}$ than both $\mathbf{w}_{t_1}$ and $\mathbf{w}_{t_2}$ (Section 2.2). In particular, this suggests that the overlap of contexts is important in deriving additive compositionality.

(C) It is better that shift term $\beta(t)$ is adjusted such that $\sum_{i=1}^{|C|} s(c_i, t) = 0$. Meanwhile, the shift term $\alpha(c_i)$ is not very relevant to additive compositionality. (Section 2.1)

(D) Low-dimensional reductions of $\mathbf{w}_t$ generally preserve additive compositionality. These include some state-of-the-art word embedding methods, such as SGNS and GloVe. However, the singular value decomposition (SVD) method is more compatible with our theory, which suggests that SVD could be at least as useful as other methods, regarding additive compositionality (Section 3).

By performing some experiments, we show that:

(E) The generalized Zipf's law actually holds in a real corpus (Section 4.1).

(F) Logarithmic context vectors in a real corpus fit with our theoretical upper bound, showing additive compositionality. In contrast, similar phenomena are not observed when using non-overlapping contexts, or taking non-logarithmic context vectors, such as $s(c_i, t) := \Pr(c_i|t)$. On the other hand, dimension reduction displays an effect of strengthening additive compositionality (Section 4.2).

(G) On a composition test set [18], we evaluated several SVD reductions of $\mathbf{w}_t$, shifted by different alpha terms. The results outperform SVD of non-overlapping contexts, and are competitive with SGNS and GloVe vectors. A constant performance gain is obtained by making $\mathbf{w}_t$ close to a PMI vector (Section 4.3).

(H) We also tested the SVD vectors on word analogy tasks [15]. The results outperform other state-of-the-art models, independent of alpha shift terms (Section 4.4).

2. Additive Compositionality

In this section, we derive our main result, and discuss some of the implications. Our goal is to bound the error $\|\mathbf{w}_{t_1t_2} - \frac{1}{2}(\mathbf{w}_{t_1} + \mathbf{w}_{t_2})\|$, where $\mathbf{w}_t$ is defined as $\mathbf{w}_t := (s(c_i, t))_{i=1}^{|C|}$, and $s(c_i, t) := \ln \mathrm{freq}(c_i, t) - \alpha(c_i) - \beta(t)$.

First, we consider a probabilistic trial, in which a word c is uniformly chosen from the context lexicon C at random. Then, for each target $t \in T$, we define a random variable S_t that outputs the value $s(c, t)$. Formally, we write $S_t := (s(c, t))_{c \sim C}$. The random variable S_t encodes the same information as the context vector $\mathbf{w}_t$, except that S_t does not depend on an explicit ordering of the lexicon C. The semantics of t are illustrated by the possible values $s(c, t)$ for each $c \sim C$ (e.g., for the target "*ice*", it is possible that $s(\mathrm{water}, \mathrm{ice}) = -3.7$ and $s(\mathrm{fashion}, \mathrm{ice}) = -5.4$), but we note that for the *distribution* of S_t there is much less information (e.g., 30% of context words c have a strength $s(c, \mathrm{ice}) \geq -3.5$). In the following subsection, we show that the distribution of S_t can be determined by a generalization of Zipf's law. Here, we convert our goal of bounding the error into the estimation of the second moment of a random variable:

$$\frac{1}{|C|}\|\mathbf{w}_{t_1t_2} - \frac{1}{2}(\mathbf{w}_{t_1} + \mathbf{w}_{t_2})\|^2 = E[(S_{t_1t_2} - \frac{1}{2}(S_{t_1} + S_{t_2}))^2], \quad (1)$$

where this equality is derived from the definition of $\mathbf{w}_t$ and S_t.

2.1 Generalized Zipf's Law

Zipf's law [26] states that the frequency $\mathrm{freq}(c)$ is inversely proportional to the rank of c in the frequency table, which in effect specifies a power law for the random variable $(\mathrm{freq}(c))_{c \sim C}$. We generalize this law to the random variable $(\mathrm{freq}(c,t))_{c\sim C}$, where t is any fixed target. To be precise, we assume the following distribution:

$$\Pr(\mathrm{freq}(c,t) \geq x) = \begin{cases} K \cdot m_t/\lceil x \rceil \ (m_t \leq x), \\ \text{unspecified} \ (x < m_t), \end{cases} \quad (2)$$

in which $m_t \in \mathbb{R}_{>0}$, and $\lceil x \rceil$ is the least integer $\geq x$. The constant K is chosen such that $K \cdot m_t/\lceil m_t \rceil = \#\{c|\,\mathrm{freq}(c,t) \geq m_t\}/|C|$, so that the number of context words with co-occurrence frequency $\geq m_t$ is exactly $\Pr(\mathrm{freq}(c,t) \geq m_t) \cdot |C|$. The parameter m_t represents the lower bound on the power law behavior, so that the distribution of frequencies $< m_t$ is unspecified. The derivation of (2) can be found in Appendix A.

(C) In order to estimate (1), we first note that the random variable $S_{t_1t_2} - \frac{1}{2}(S_{t_1} + S_{t_2})$ does not depend on the shift term $\alpha(c)$, because it is canceled out in this expression. Therefore, without loss of generality, we can that assume $\alpha(c) = 0$. Now, recall that the second moment of a random variable X can be written as $E[X^2] = V(X) + E[X]^2$, where $V(X)$ is the variance. Therefore, (1) becomes smaller when $E[S_{t_1t_2} - \frac{1}{2}(S_{t_1} + S_{t_2})] = 0$, which can be achieved by adjusting each $\beta(t)$ such that $E[S_t] = 0$. This is reasonable, because the strength $s(c,t)$ only makes sense when compared to some average level; its absolute magnitude does not directly represent the semantics of the target t. Hereon, we apply this setting, and assume that $E[S_t] = 0$.

Because $S_t = (\ln \mathrm{freq}(c,t) - \beta(t))_{c\sim C}$, and $\beta(t)$ is specified such that $E[S_t] = 0$, the distribution of S_t can be calculated from the distribution of $(\mathrm{freq}(c,t))_{c\sim C}$, which is given in (2). The following is proven in Appendix A.

Theorem 1. *If we assume the generalized Zipf's law* (2) *holds, then $S_t + 1$ has an approximately exponential distribution of rate parameter* 1.

(A1) From Theorem 1, we know that the random variable S_t has an exponential tail, which suggests that the logarithmic function in the definition of S_t is not arbitrary. Without the logarithm, the generalized Zipf's law (2) implies that $(\mathrm{freq}(c,t) - \beta(t))_{c\sim C}$ has a power law tail, which is very different from an exponential tail. For example, consider $S_t := (\Pr(c|t))_{c\sim C}$, a scalar multiplication of $(\mathrm{freq}(c,t))_{c\sim C}$. The generalized Zipf's law implies that $\Pr(c|t)$ is mostly very close to 0, yet has very large values for a significant portion of $c \in C$. Therefore, S_t is expected to yield an almost infinite second moment (in contrast to the logarithmic case, where $E[S_t^2] = 1$ by Theorem 1), which may exclude any nontrivial estimations for the second moment of $S_{t_1t_2} - \frac{1}{2}(S_{t_1} + S_{t_2})$. This prediction is verified by experiments (Section 4.2).

(A2) Noisy low-frequencies of rarely seen (c,t) pairs can be naturally complemented by the generalized Zipf's law (e.g., thinking of $\mathrm{freq}(c,t) = 1.6$, when the actually observed frequency is $\mathrm{freq}(c,t) = 1$). The idea is to extend the lower bound m_t to the power law behavior (2). That is, to extrapolate low frequencies $< m_t$ by assuming the unspecified part in (2) to be an exact, continuous power law as follows:

$$\Pr(\mathrm{freq}(c,t) \geq x) = \begin{cases} \tilde{m}_t/x \ (\tilde{m}_t \leq x < m_t), \\ 1 \qquad (x < \tilde{m}_t), \end{cases} \quad (3)$$

where $\tilde{m}_t = K \cdot m_t$. We will replace any frequency value $< m_t$ by a sample drawn from the above distribution (3), while preserving the frequency rank. Thus, the complemented frequency will be a real number $\geq \tilde{m}_t$, the new lower bound on this exact and continuous power law. From the proof of Theorem 1, we can deduce that $S_t + 1$ moves closer to the exponential distribution after complementing low-frequencies. We also need estimate m_t in order to implement this strategy; a method using [3] is described in Appendix B. Our experiments show that complementing low-frequencies can drastically improve the additive compositionality (Section 4.2).

2.2 Main Result

The observation that is key to our main result is the context overlap between two successively occurring unigrams (Table 1). In order to model this phenomenon, we assume that the contexts of any two unigrams t_1 and t_2 are generated by the following process. When an unordered pair $\{t_1, t_2\}$ appears successively (i.e., either t_1t_2 or t_2t_1) in a sentence, the contexts of t_1 and t_2 are *exactly the same* sample, drawn from a distribution $\Pr(c|t_1t_2)$. Meanwhile, all non-neighboring occurrences of t_1 and t_2 are assumed to be far from each other, so their contexts are independently drawn from $\Pr(c|t_1\backslash t_2)$ and $\Pr(c|t_2\backslash t_1)$, respectively. Formally,

$$\Pr(c|t_1) = \tau_1 \Pr(c|t_1 \backslash t_2) + (1-\tau_1) \Pr(c|t_1t_2),$$
$$\Pr(c|t_2) = \tau_2 \Pr(c|t_2 \backslash t_1) + (1-\tau_2) \Pr(c|t_1t_2),$$

where $\tau_1 = \Pr(t_1 \text{ not neighboring } t_2|t_1)$ is the proportion of t_1 occurrences *not* neighboring t_2. Therefore, τ_1 is small when $\{t_1, t_2\}$ occurs often. τ_2 is defined similarly. From this context model, we have

$$\begin{aligned}&\ln \Pr(c|t_1)\\&\quad=\ln\{\tau_1\Pr(c|t_1\setminus t_2)+(1-\tau_1)\Pr(c|t_1t_2)\}\\&\quad\risingdotseq\tau_1\ln\Pr(c|t_1\setminus t_2)+(1-\tau_1)\ln\Pr(c|t_1t_2),\end{aligned}$$

and a similar formula for $\ln\Pr(c|t_2)$ [*1]. Now, substitute $\ln\Pr(c|t_1)$ into $S_{t_1}=(\ln\operatorname{freq}(c,t_1)-\beta(t_1))_{c\sim C}=(\ln\Pr(c|t_1)-\widehat{\beta}(t_1))_{c\sim C}$, and note that $\widehat{\beta}(t_1)$ is specified such that $E[S_{t_1}]=0$, so we get

$$S_{t_1}\risingdotseq\tau_1 S_{t_1\setminus t_2}+(1-\tau_1)S_{t_1t_2},\tag{4}$$

and similarly

$$S_{t_2}\risingdotseq\tau_2 S_{t_2\setminus t_1}+(1-\tau_2)S_{t_1t_2}.\tag{5}$$

Using (4) and (5), we get

$$\begin{aligned}&S_{t_1t_2}-\frac{1}{2}(S_{t_1}+S_{t_2})\\&\qquad\risingdotseq\frac{1}{2}\{(\tau_1+\tau_2)S_{t_1t_2}-\tau_1 S_{t_1\setminus t_2}-\tau_2 S_{t_2\setminus t_1}\}.\end{aligned}$$

Hence, if $S_{t_1t_2}$, $S_{t_1\setminus t_2}$ and $S_{t_2\setminus t_1}$ are independent, we can calculate $E[(S_{t_1t_2}-\frac{1}{2}(S_{t_1}+S_{t_2}))^2]\risingdotseq\frac{1}{2}(\tau_1^2+\tau_2^2+\tau_1\tau_2)$. In practice, however, $S_{t_1t_2}$ almost always has a positive correlation with $S_{t_1\setminus t_2}$ and $S_{t_2\setminus t_1}$, because frequently used words are likely to be used in every context, regardless the target. As a consequence, the variance gets smaller, and we have the following estimation:

$$E[(S_{t_1t_2}-\frac{1}{2}(S_{t_1}+S_{t_2}))^2]\le\frac{1}{2}(\tau_1^2+\tau_2^2+\tau_1\tau_2).$$

Therefore, we obtain the main result:

$$\|\mathbf{w}_{t_1t_2}-\frac{1}{2}(\mathbf{w}_{t_1}+\mathbf{w}_{t_2})\|\le\sqrt{\frac{|C|}{2}(\tau_1^2+\tau_2^2+\tau_1\tau_2)}.$$

(B1) From (4), we show that S_{t_1} and $S_{t_1t_2}$ are linearly correlated. As $\{t_1,t_2\}$ occurs more often, τ_1 becomes smaller, and the correlation becomes higher. Similar behavior holds for S_{t_2} and $S_{t_1t_2}$. As manifested in the main result, this has the effect that when $\{t_1,t_2\}$ occurs often, the error of the approximation of $\mathbf{w}_{t_1t_2}$ by $\frac{1}{2}(\mathbf{w}_{t_1}+\mathbf{w}_{t_2})$ is small.

(B2) We could also deduce an upper bound simply from (4). Namely, that $\|\mathbf{w}_{t_1t_2}-\mathbf{w}_{t_1}\|\le\sqrt{2|C|}\tau_1$. From (5), we get that $\|\mathbf{w}_{t_1t_2}-\mathbf{w}_{t_2}\|\le\sqrt{2|C|}\tau_2$. However, we note that the upper bound given in the main result is tighter than the one derived from the triangular inequality: $\|\mathbf{w}_{t_1t_2}-\frac{1}{2}(\mathbf{w}_{t_1}+\mathbf{w}_{t_2})\|\le\frac{1}{2}(\|\mathbf{w}_{t_1t_2}-\mathbf{w}_{t_1}\|+\|\mathbf{w}_{t_1t_2}-\mathbf{w}_{t_2}\|)\le\sqrt{\frac{|C|}{2}}(\tau_1+\tau_2)$. This suggests that $\frac{1}{2}(\mathbf{w}_{t_1}+\mathbf{w}_{t_2})$ can get closer to $\mathbf{w}_{t_1t_2}$ than both $\mathbf{w}_{t_1}$ and $\mathbf{w}_{t_2}$. Intuitively, this is because when S_{t_1} and S_{t_2} add up, the two highly independent components $S_{t_1\setminus t_2}$ and $S_{t_2\setminus t_1}$ cancel each other out, whereas the common component $S_{t_1t_2}$ reinforces itself.

By performing experiments (Section 4.2), we verify the upper bound given by our main result, and we confirm that the overlap of contexts is important in deriving additive compositionality.

3. Dimension Reduction

In this section, we discuss low-dimensional reductions of the context vector $\mathbf{w}_t$. Given a dimension d, we want to use a d-dimensional vector $\mathbf{v}_t$ to approximate the $|C|$-dimensional vector $\mathbf{w}_t$. This can be formalized as the finding of a d-dimensional vector $\mathbf{v}_t$ for each $t\in T$, and a $(|C|\times d)$-matrix A, such that $\sum_{t\in T}L(A\mathbf{v}_t,\mathbf{w}_t)$ is minimized, where $L(\cdot,\cdot)$ is a given loss function.

(D) In general, dimension reductions preserve additive composition, as the argument below will show. First, by definition, $L(A\mathbf{v}_{t_1},\mathbf{w}_{t_1})$, $L(A\mathbf{v}_{t_2},\mathbf{w}_{t_2})$, and $L(A\mathbf{v}_{t_1t_2},\mathbf{w}_{t_1t_2})$ are small, which means that $A\mathbf{v}_{t_1}$, $A\mathbf{v}_{t_2}$, and $A\mathbf{v}_{t_1t_2}$ are close to $\mathbf{w}_{t_1}$, $\mathbf{w}_{t_2}$, and $\mathbf{w}_{t_1t_2}$, respectively. Therefore, $A\{\mathbf{v}_{t_1t_2}-\frac{1}{2}(\mathbf{v}_{t_1}+\mathbf{v}_{t_2})\}=A\mathbf{v}_{t_1t_2}-\frac{1}{2}(A\mathbf{v}_{t_1}+A\mathbf{v}_{t_2})$ is "near" to $\mathbf{w}_{t_1t_2}-\frac{1}{2}(\mathbf{w}_{t_1}+\mathbf{w}_{t_2})$. Second, $\|\mathbf{w}_{t_1t_2}-\frac{1}{2}(\mathbf{w}_{t_1}+\mathbf{w}_{t_2})\|$ is bounded by our main result, so we can bound $A\{\mathbf{v}_{t_1t_2}-\frac{1}{2}(\mathbf{v}_{t_1}+\mathbf{v}_{t_2})\}$ accordingly. Third, since A is bounded operator, we can obtain bounds for $\mathbf{v}_{t_1t_2}-\frac{1}{2}(\mathbf{v}_{t_1}+\mathbf{v}_{t_2})$ using the bounds for $A\{\mathbf{v}_{t_1t_2}-\frac{1}{2}(\mathbf{v}_{t_1}+\mathbf{v}_{t_2})\}$.

Some technical issues remain in the argument given above. First, the loss function L does not always satisfy a triangular inequality, meaning that $A\{\mathbf{v}_{t_1t_2}-\frac{1}{2}(\mathbf{v}_{t_1}+\mathbf{v}_{t_2})\}$ and $\mathbf{w}_{t_1t_2}-\frac{1}{2}(\mathbf{w}_{t_1}+\mathbf{w}_{t_2})$ may not always be close. Second, a bound for the Euclidean distance does not always imply a bound for the loss function L, or vice versa; so caution is required when applying the argument to a general loss. However, in the simplest case, where L is the L_2-loss, the above argument can be applied in a most compatible way. This suggests that the truncated SVD dimension reduction, which solves the L_2-loss minimization, is suitable for training additive compositional word vectors. In the following subsections, we compare SVD with two state-of-the-art methods, SGNS and GloVe. Empirical evaluations

[*1] This formula is valid, because $\Pr(c|t_1\setminus t_2)$ and $\Pr(c|t_1t_2)$ are very small (according to the generalized Zipf's law, the largest $\Pr(c|t)$ for a fixed t is approximately equal to $1/\sum_{r=1}^{n_t}\frac{1}{r}$, where $n_t:=\#\{c|\operatorname{freq}(c,t)>0\}$ is the number of distinct context words of t observed in the corpus. When the corpus size increases, $n_t\to+\infty$ and $\Pr(c|t)\to 0$). Therefore, for any x between $\Pr(c|t_1\setminus t_2)$ and $\Pr(c|t_1t_2)$, we can approximate $\ln(x)$ linearly.

Vol.2015-NL-221 No.14
Vol.2015-SLP-106 No.14
2015/5/26

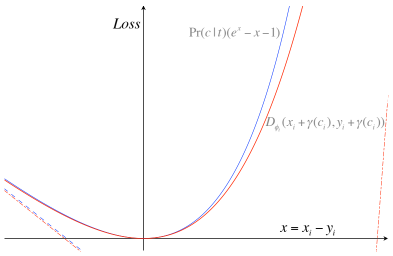

図 1 Graph of the SGNS loss function, which has two asymptotes (red). Its limit curve at $k \to +\infty$ has one asymptote (blue), and grows exponentially at $x \to +\infty$.

are conducted on a composition test set (Section 4.3) and word analogy (Section 4.4).

3.1 The Loss Function of SGNS

Recently, [13] have shown that the skip-gram model of negative sampling (SGNS) can be viewed as a factorization of the shifted-PMI matrix. More precisely, they showed that SGNS is a matrix factorization of $s(c,t) := \ln \Pr(c|t) - \ln(kP_{\text{noise}}(c))$, where k is an integer (the number of negative samples), and P_{noise} is a given noise distribution. This $s(c,t)$ is a special case of the strength functions we consider in this paper, so SGNS constitutes a dimension reduction of logarithmic context vectors. The difference between SGNS and the SVD reduction of the same $\mathbf{w}_t := (s(c_i,t))_{i=1}^{|C|}$ will be the loss function. In Appendix C, we prove the following theorem.

Theorem 2. *For the $|C|$-dimensional vectors $A\mathbf{v}_t$ and $\mathbf{w}_t$, SGNS uses the following loss function L_t:*

$$L_t(\mathbf{x},\mathbf{y}) = \Pr(t) \sum_{i=1}^{|C|} D_{\phi_i}(x_i + \gamma(c_i), y_i + \gamma(c_i)), \quad (6)$$

where $\gamma(c_i) := \ln(kP_{\text{noise}}(c_i))$, and $D_{\phi_i}(\cdot,\cdot)$ is the Bregman divergence associated with the convex function

$$\phi_i(x) = (\Pr(c_i|t) + e^{\gamma(c_i)}) \ln(e^x + e^{\gamma(c_i)}).$$

When $k \to +\infty$, the limit of D_{ϕ_i} is another Bregman divergence D_φ, associated with $\varphi(x) = e^x$.

A graph of $D_{\phi_i}(x_i+\gamma(c_i), y_i+\gamma(c_i))$, fixing $y_i = s(c_i,t)$ and varying $x = x_i - y_i$, is presented in Figure 1. D_{ϕ_i} becomes steeper as $\Pr(c|t)$ grows larger (note the $\Pr(c|t)$ coefficient in the equation of the limit curve), meaning that L_t puts more weight on frequent context words. In addition, the graph grows much faster at $x_i - y_i \to +\infty$ than at $x_i - y_i \to -\infty$ (Figure 1), so an x_i overestimating $y_i = s(c_i,t)$ is punished more than an underestimation. Therefore, the loss function (6) tends to enforce underestimations of $s(c,t)$ for a frequent context word c (since overestimating such $s(c,t)$ will be costly), and to compensate $s(c,t)$ for rarely seen contexts (i.e., overestimations on such c are affordable, so this will be done if necessary). This is a desirable property for a good generalization, and somewhat similar to the effect of complementing low-frequency data, as discussed in Section 2.1. However, the case of the SGNS loss function, where more weight is put on frequent context words, contrasts to the uniform L_2 loss in SVD. When too much weight is put on frequent contexts, the trained $A\mathbf{v}_t$ may fail to mimic the exponential distribution behavior of $\mathbf{w}_t$ on a large portion of relatively low-frequencies, which may hurt additive compositionality. This is because during the addition $\mathbf{w}_{t_1} + \mathbf{w}_{t_2}$, this portion should be the main area where the most cancellations occur, and the signal from $\mathbf{w}_{t_1t_2}$ reinforces itself. On the other hand, it seems reasonable to put more weight on frequent *targets*, much like the $\Pr(t)$ coefficient in (6) .

3.2 The GloVe Model

In the GloVe model [20], trained vectors $(\mathbf{v}_t, \tilde{\mathbf{v}}_c)$ are matrix factorizations of $\ln \mathrm{freq}(c,t) - b(t) - \tilde{b}(c)$, whereas the bias terms $b(t)$ and $\tilde{b}(c)$ are learned simultaneously, by minimizing a weighted L_2 loss as follows:

$$\sum_{c,t} f(c,t)(\mathbf{v}_t \cdot \tilde{\mathbf{v}}_c + b(t) + \tilde{b}(c) - \ln \mathrm{freq}(c,t))^2.$$

The weight $f(c,t) \to 0$ when $\mathrm{freq}(c,t) \to 0$. One notable difference between GloVe and the SVD approach discussed in this paper is the treatment of rarely seen (c,t) pairs. GloVe avoids the noisy low-frequencies and $\ln(0)$ by downgrading their weights in the loss function, which results in a sparse matrix and can be handled using the Stochastic Matrix Factorization (SMF) method [9]. In contrast, SVD should apply a uniform L_2 loss, which makes it mandatory to explicitly complement low-frequencies and unseen pairs. As a result, truncated SVD can be calculated using the extremely efficient random projection algorithm [7], which is usually faster and more precise than SMF. However, SVD needs to handle dense matrices, which becomes difficult (although it has been well studied) when scaling up to very large data.

4. Experiments

In this section, we test the assumptions and implications of our theory on practical data. We use the British National Corpus (BNC) [23], which contains about 100 million word tokens. We extract all sentences from texts (not including headings and captions) and utterances, and

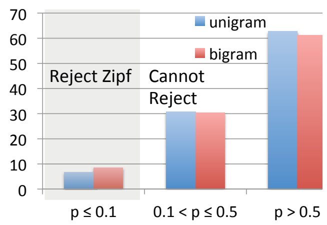

図 2 Aggregate of the p-value

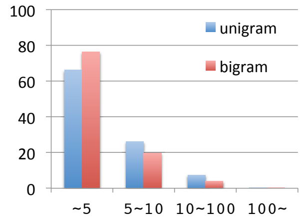

図 3 Aggregate of the estimated m_t

a sentence is regarded as a sequence of word tokens (punctuation not included). For context words, we take all words with a frequency ≥ 200, which results in a vocabulary of 22,000 words. For targets, we use unigrams with a frequency ≥ 200 (22,000 words, the same as the context vocabulary), as well as unordered bigrams of frequency ≥ 200 (47,000 word pairs). The window size used is five to each side for unigram targets, and four for bigram word pairs. Windows do not cross sentences.

4.1 Testing Generalized Zipf's Law

In this subsection, we test whether the generalized Zipf's law actually holds in a real corpus. For each target t (which is either a unigram target or an unordered bigram target), we compare the proposed distribution (2) to the distribution of $\mathrm{freq}(c, t)$ observed in data. I order to measure the goodness-of-fit, we run a Kolmogorov-Smirnov (KS) test, as described in [3], for each target. The KS test estimates the parameter m_t in (2) at the same time. For further details, see Appendix B.

The KS goodness-of-fit tests produce p-values, representing the plausibility of assuming that the generalized Zipf's law holds. A larger p-value indicates that the generalized Zipf's law fits the data well; and as pointed out in [3], it is a relatively conservative choice to reject Zipf's law when $p \leq 0.1$. The results of the KS tests are summarized in Figure 2 and Figure 3. According to the p-values (Figure 2), we should reject the generalized Zipf's law for below 10% of both unigram targets and unordered bigram targets. For the majority of targets ($> 60\%$), the generalized Zipf's law is very difficult to reject ($p > 0.5$).

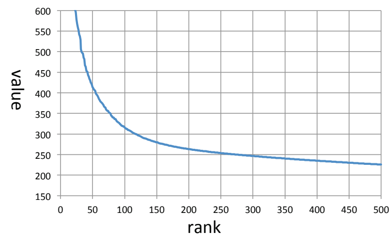

図 5 The top 500 singular values in SVD

As for the estimated m_t, in most cases this is less than 10 (Figure 3), which indicates that our complementing of low frequency context-target pairs does not substantially change the observed data.

4.2 Additive Compositionality in Practice

In this subsection, we verify our main result and confirm the implications, using some scatter plots that are constructed as follows. For each unordered bigram target $\{t_1, t_2\}$, we plot at $x = \frac{1}{2}(\tau_1^2 + \tau_2^2 + \tau_1\tau_2)$, and calculate y as the approximation error regarding additive compositionality, for different types of context vectors. In all settings, the shift term $\alpha(c_i)$ is set to zero, and the shift term $\beta(t)$ is always adjusted such that the entries of the vector sum up to zero. We omit this term for brevity.

First, as an alternative to complementing unseen pairs, we consider a naive setting where context words are restricted to a sub-lexicon $C' := \{c | \mathrm{freq}(c, t_1t_2) > 0\}$, whereas the context vectors $\mathbf{w}_{t_1t_2}$, $\mathbf{w}_{t_1}$ and $\mathbf{w}_{t_2}$ are restricted onto C'. Formally, $\mathbf{w}'_{t_1t_2} := (s(c_i, t_1t_2))_{c_i \in C'}$, $\mathbf{w}'_{t_1} := (s(c_i, t_1))_{c_i \in C'}$, and $\mathbf{w}'_{t_2} := (s(c_i, t_2))_{c_i \in C'}$. Then, we set $y = \frac{1}{|C'|}\|\mathbf{w}'_{t_1t_2} - \frac{1}{2}(\mathbf{w}'_{t_1} + \mathbf{w}'_{t_2})\|^2$. The plot is shown in Figure 4(ii). According to our main result, we would expect that all points lie under the theoretical bound of $y = x$ (solid red line). However, we note that a significant portion of points lie above this line.

Next, we complement low-frequencies as described in Section 2.1. The resulting context vectors are denoted as $\tilde{\mathbf{w}}_{t_1t_2}$, $\tilde{\mathbf{w}}_{t_1}$, and $\tilde{\mathbf{w}}_{t_2}$. We set $y = \frac{1}{|C|}\|\tilde{\mathbf{w}}_{t_1t_2} - \frac{1}{2}(\tilde{\mathbf{w}}_{t_1} + \tilde{\mathbf{w}}_{t_2})\|^2$. The plot is presented in Figure 4(iii). In contrast to Figure 4(ii), most points now lie under the solid red line, as predicted by our main result, showing the effect of low-frequency complementing. A dashed red line is drawn to show the level of average y of all points.

Next, we consider a setting in which contexts of neighboring unigrams do *not* overlap. This is achieved by labeling context words with relative positions. For example, in the sequence "`a b c d e`", the contexts of `c` are labeled words such as `b-1`, `a-2`, `d+1`, and `e+2`. We calculate context vectors in this setting and perform complementation,

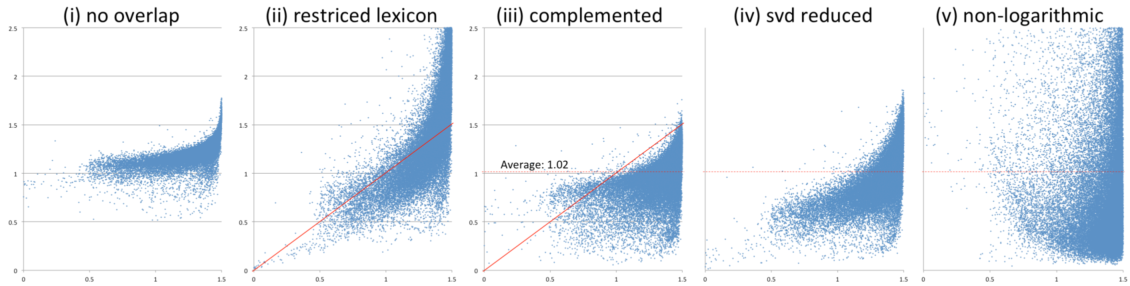

図 4 Additive compositionality in different settings

much in the same way as in the previous paragraph. We set $y = \frac{1}{|C|}\|\tilde{\mathbf{w}}_{t_1t_2} - \frac{1}{2}(\tilde{\mathbf{w}}_{t_1} + \tilde{\mathbf{w}}_{t_2})\|^2$. The plot is shown in Figure 4(i). We do not observe a tendency that the approximation error decreases as $\{t_1, t_2\}$ occurs more often.

Now, we consider the non-logarithmic setting where $s(c, t) = \Pr(c|t)$. The vector, no longer having $\ln(0)$-entries, does not need complementing. Therefore, we set $y = \frac{1}{|C|}\|\mathbf{w}_{t_1t_2} - \frac{1}{2}(\mathbf{w}_{t_1} + \mathbf{w}_{t_2})\|^2$. The plot is shown in Figure 4(v). Note that the absolute magnitudes of y for different types of vectors cannot be directly compared to each other, since the magnitude would change by multiplying all vectors by a constant scalar. Therefore, we do not draw a scale on the y-axis in Figure 4(v). Instead, we scale the y-axis such that the average level is the same as in Figure 4(iii). We see the variance in this plot is very large, and no obvious additive compositionality can be observed.

Finally, we plot the SVD reduction of complemented context vectors. The dimension of reduction is set to 200, which is selected by observing the top 500 singular values (Figure 5). At a dimension of 200, the singular values begin to decrease at a constant rate, which may suggest that there is not much information in dimensions ≥ 200. This setting will also produce better results in experiments described later. The reduced vectors are denoted as $\mathbf{v}_t$, and all reduced vectors are normalized. We set $y = \|\mathbf{v}_{t_1t_2} - \frac{1}{2}(\mathbf{v}_{t_1} + \mathbf{v}_{t_2})\|^2$. The plot is shown in Figure 4(iv). Compared with Figure 4(iii), the plot is neater and steeper, which suggests that some kind of "clustering" occurred, strengthening the tendency of additive compositionality.

4.3 Semantic Composition

To test if the vectors trained by SVD actually exhibit additive compositionality on linguistically meaningful phrases, we employ a data set[*2] created by [18], which consists of phrases extracted from BNC and annotated by humans on their semantic similarity.

Each instance in the dataset is a (*phrase1*, *phrase2*, *similarity*) triplet, and each phrase consists of two words. The *similarity* score is a value annotated by humans, ranging from 1 to 7, and indicating how similar the semantics of the two phrases are. For example, one participant annotated the similarity between *vast amount* and *large quantity* as 7 (the highest similarity), and the similarity between *hear word* and *remember name* as 1 (the lowest similarity). Phrases are divided into three categories: verb-noun, noun-noun, and adjective-noun. Each category has 108 phrase pairs, and is annotated by 18 human subjects (i.e., 1,944 instances in each category).

For each category, we compare the human ratings with computer outputs, which for each phrase pair are obtained by first adding up the two word vectors of each phrase, and then calculating the cosine similarity. The performance is measured by Spearman's ρ, which tells us how closely the computer outputs are related to the human ratings. We test several word vectors on each of the three categories.

The results are presented in Table 2. First, we tested the SVD reductions of the complemented context vectors, shifted by various alpha terms. For example, the '$x = 0.25$' row shows the results of the SVD reduction of the vector $\tilde{\mathbf{w}}_t := (\ln \tilde{\mathrm{freq}}(c_i, t) - 0.25 \ln \Pr(c_i))_{i=1}^{|C|}$, where $\tilde{\mathrm{freq}}(c_i, t)$ is the complemented frequency. We compare the results with the SVD reduction of non-overlapping

$\alpha(c_i) = x \ln \Pr(c_i)$	**VB-NN**	**NN-NN**	**JJ-NN**
$x = 0$	0.38	0.44	0.39
$x = 0.25$	0.38	0.44	0.39
$x = 0.5$	0.38	0.45	0.40
$x = 0.75$	0.40	0.45	0.41
$x = 1$	**0.40**	**0.46**	0.42
SVD-NoOverlap	0.34	0.43	0.36
GloVe	0.38	0.44	**0.45**
SGNS	0.36	0.43	**0.45**

表 2 Spearman's ρ on semantic composition

[*2] http://homepages.inf.ed.ac.uk/s0453356/

$\alpha(c_i) = x \ln \Pr(c_i)$	**Google**	**MSR**
$x = 0$	45.6	**58.2**
$x = 0.25$	45.0	57.6
$x = 0.5$	45.7	57.6
$x = 0.75$	**47.0**	57.3
$x = 1$	46.4	57.2
SVD-NoOverlap	31.8	53.7
GloVe	**45.8**	**57.4**
SGNS	39.9	50.4
3CosMul	40.3	43.6

表 3 Accuracy on analogy tasks

context vectors, as well as vectors produced by GloVe*3 and SGNS*4 toolkits, with dimension 200, window size 5 to each side, cutoff 10 for GloVe, subsampling 0 for SGNS, and other default settings.

First, we see that SVD-NoOverlap consistently performs worse than other vectors, indicating that composition may not be well captured by adding non-overlapping context vectors. Second, SVD reductions of complemented context vectors yield results that are competitive with the GloVe and SGNS vectors, outperforming the two on verb-noun and noun-noun categories. Finally, we note an intriguing tendency that the performance consistently improves as x changes from zero to one and $\mathbf{w}_t$ gets closer to the PMI vector. We believe that the reason for this is that, although the "degree" of additive compositionality is *not* altered by x, the composed vectors get closer to the PMI vectors of phrases as x increases, and the similarity of PMI vectors are closer to human intuitions on the semantic similarity.

4.4 Analogy Tasks

We also compared the performance of different word vectors and strategies on analogy tasks. We use the **MSR***5 [15] and **Google***6 [16] datasets, comprised of 4-tuples of words that are subject to "a is to b as c is to d". Tuples with out-of-vocabulary words are removed from data, which results in 4382 tuples in **MSR** and 8924 tuples in **Google***7.

A comparison of different strategies is presented in Table 3. The 3CosMul method was proposed in [12]; SVD-NoOverlap uses the SVD reduction of non-overlapping context vectors; GloVe and SGNS are vectors produced by the corresponding models*8. Among all of the compared methods, SVD reductions of complemented context vectors showed the best performance, although GloVe was almost the same. In addition, it is noteworthy that the performance only depended weakly on the shift term $\alpha(c_i)$.

Additive compositionality is thought to be related to analogy tasks, because additive compositionality enforces linearity. However, it is not known what exactly this relation is. In addition, we note that strategies not directly related to additive compositionality (e.g., 3CosMul and SVD-NoOverlap) can still achieve a high performance on analogy tasks.

5. Discussion

Computational linguistics is largely related to the application of general machine learning frameworks to different NLP tasks. However, natural language specific properties, such as the (generalized) Zipf's law, can have profound implications, which are not always trivial [8], [14]. We believe that there are more deep results still to be discovered in such "mathematical linguistics". In addition, we believe that our careful investigation on additive compositionality can lead to deeper insights, and find further applications to various tasks in NLP.

Appendices

A. Zipf's Law and Power Law

A.1 Zipf's Law as the Distribution of Word Occurrences

Zipf's law [26] states that the frequency of a word in a corpus is inversely proportional to its rank in the frequency table. Under the assumption that the frequency $\mathrm{freq}(w)$ of each word w is drawn i.i.d. from a probabilistic distribution, Zipf's law determines this distribution as follows.

Recall that the cumulative distribution function (CDF) defined as $F(x) := \Pr(\mathrm{freq}(w) \geq x)$ determines the probabilistic distribution. CDF should not be confused with the probabilistic density function (PDF), which is the derivative of CDF if $F(x)$ is differentialble. To calculate $F(x)$, we formally write the definition of rank as the following,

*3 http://nlp.stanford.edu/projects/glove/
*4 https://code.google.com/p/word2vec/
*5 http://research.microsoft.com/en-us/projects/rnn/
*6 https://code.google.com/p/word2vec/
*7 These are about half the size of the original datasets.
*8 In the default implementation, GloVe weights context words by the inverse of their distance to the target. Similar tricks also exist in the word2vec implementation of SGNS. These tricks are known to boost the performance on analogy tasks. However, regarding the context model we considered in this paper and for fair a comparison, we altered the implementations here to set equal weights to all context words.

$$\mathrm{rank}(w) := \#\{w' | \mathrm{freq}(w') \geq \mathrm{freq}(w)\} \quad (7)$$

which defines the frequency rank of a word w as the count of such word w' that occurs in a frquency higher than $\mathrm{freq}(w)$. Then, Zipf's law states that

$$\#\{w' | \mathrm{freq}(w') \geq \mathrm{freq}(w)\} = \mathrm{rank}(w) = \frac{E}{\mathrm{freq}(w)}, \quad (8)$$

where E is the proportionality constant. Now replace $\mathrm{freq}(w)$ by x in the above equation (8) , we get

$$\#\{w' | \mathrm{freq}(w') \geq x\} = \frac{E}{x}. \quad (9)$$

Hence, let the total number of words be N, we have

$$F(x) = \Pr(\mathrm{freq}(w) \geq x) = \frac{\#\{w' | \mathrm{freq}(w') \geq x\}}{N} = \frac{E}{\lceil x \rceil}, \quad (10)$$

where $\lceil x \rceil$ is the least integer greater than x, which is taken because originally the frequency $\mathrm{freq}(w)$ is always an integer.

In pactice, the above equation (10) cannot be everywhere true, for example $F(x) = \infty$ when $x = 0$, which is obviously absurd. As is usual in the analysis of a power law [3], we asume (10) holds for every $x \geq m$, where $m \in \mathbb{R}_{>0}$:

$$F(x) = \Pr(\mathrm{freq}(w) \geq x) = \begin{cases} K \cdot m/\lceil x \rceil & (x \geq m) \\ \text{unspecified} & (x < m) \end{cases} \quad (11)$$

Here the constant K is taken as the following, such that $F(m)$ is exactly the proportion of words which occur in frequencies $\geq m$.

$$F(m) = K \cdot m/\lceil m \rceil = \frac{\#\{w' | \mathrm{freq}(w') \geq m\}}{N}. \quad (12)$$

A.2 Proof of Theorem 1

Assume the frequency $\mathrm{freq}(w)$ follows Zipf's law. Let $S = \ln \mathrm{freq}(w) - \beta$, where β is chosen such that $E[S] = 0$. To calculate the distribution of S, we first prove that the distribution of $\ln \mathrm{freq}(w) - \ln Km$ is roughly an exponential distribution of rate parameter 1. Then, since $\ln \mathrm{freq}(w) - \ln Km = S + \text{Constant}$, by taking expected value of each side and noting $E[S] = 0$, we conclude that Constant $= 1$, so $S + 1$ is roughly an exponential distribution of rate parameter 1.

Now, the CDF of $\ln \mathrm{freq}(w) - \ln Km$ is calculated as follows.

$$\begin{aligned} &\Pr(\ln \mathrm{freq}(w) - \ln Km \geq x) \\ &= \Pr(\mathrm{freq}(x) \geq \exp(x + \ln Km)) \\ &= Km/\lceil \exp(x + \ln Km) \rceil \quad \text{(by (11) , when } x \geq -\ln K) \\ &\fallingdotseq \exp(-x) \end{aligned}$$

Hence, when $x \geq -\ln K$, the distribution of $\ln \mathrm{freq}(w) - \ln Km$ is roughly an exponential distribution of rate parameter 1. Theorem 1 is proven.

B. Estimating m and Testing Zipf's Law

B.1 Estimating the lower bound on power-law behavior

In Appendix A, we derived that the cumulative distribution function (CDF) of the distribution of $\mathrm{freq}(w)$ is of the form

$$F(x) = \Pr(\mathrm{freq}(w) \geq x) = \begin{cases} K \cdot m/\lceil x \rceil & (x \geq m) \\ \text{unspecified} & (x < m) \end{cases} \quad (13)$$

where the constant K is taken such that

$$F(m) = K \cdot m/\lceil m \rceil = \frac{\#\{w' | \mathrm{freq}(w') \geq m\}}{N}. \quad (14)$$

Hence, if we consider the sub-lexicon $C_x := \{w' | \mathrm{freq}(w') \geq x\}$ comprised of words of frequency $\geq x$, then we have the following power law restricted to the sub-lexicon C_m:

$$G_m(x) = \Pr(\mathrm{freq}(w) \geq x | w \in C_m) = \begin{cases} m/\lceil x \rceil & (x \geq m) \\ 1 & (x < m) \end{cases} \quad (15)$$

How to estimate this m from data? In this section, we give a brief introduction to the method described in [3].

The main idea is to consider the Kolmogorov-Smirnov (KS) statistic, which is a measure of how well an empirical sample can fit to a proposed distribution. In our case, the KS statistic (associated with m) is defined as

$$KS_m := \max_{x \geq m} \left| G_m(x) - \frac{\#C_x}{\#C_m} \right|, \quad (16)$$

in which, $G_m(x)$ is the theoretical probability of $\mathrm{freq}(w) \geq x$ proposed by the power law (15) , whereas $\#C_x/\#C_m$ is the probability observed in data. Hence, KS_m is smaller means G_m fits the data better. Therefore, we estimate m as

$$m^* := \operatorname*{arg\,min}_{m>0} KS_m = \operatorname*{arg\,min}_{m>0} \max_{x \geq m} \left| \frac{m}{\lceil x \rceil} - \frac{\#C_x}{\#C_m} \right|. \quad (17)$$

B.2 Testing Zipf's Law

The KS statistic can also be used to perform the Kolmogorov-Smirnov test, which estimates the plausibility of a proposed distribution. In our case, we want to test if the practical data actually follows Zipf's law (13) .

The procedure is as follows [3].

(1) Given a lexicon C and their frequencies $\mathrm{freq} : C \to \mathbb{N}$,

we firstly estimate m^* as described in Section B.1, and record the KS statistic KS_{m^*}.

(2) In order to find out if this KS_{m^*} is plausible, we compare it with KS statistics of synthesized samples drawn from the proposed distribution, which is (13) in our case.
 (a) We synthesize an artificial sample S comprised of $|C|$ sample points as follows. At probability $\#C_{m^*}/|C|$, the point is drawn from distribution (15) ; otherwise, we uniformly choose a $w \in C \setminus C_{m^*}$ at random, and use $\mathrm{freq}(w)$ as the sample point.
 (b) Estimate m_S^* for the sample S, and record the KS statistic $KS_{m_S^*}$.

(3) Repeat Step 2 for $\frac{1}{4}\epsilon^{-2}$ times, where ϵ is our required accuracy for the p-value. Then, the p-value is calculated as the fraction of the time the synthetic $KS_{m_S^*}$ is larger than KS_{m^*}. In our experiments, we use $\epsilon = 0.01$.

Hence, Zipf's law is more plausible when p-value is larger. As described in [3], it is relatively conservative to reject Zipf's law if $p \leq 0.1$.

C. The Loss Function of SGNS

In this appendix, we summarize the basics of the skip-gram model. The original explanation of the theory [16] was indeed cryptic, due to two missing links: (i) the link between the negative sampling objective (NEG) and the probability distribution it claims to model; and (ii) the link between NEG and the noise contrastive estimation (NCE) method. In the following, we will give a refined explanation, which shows that, though NEG was originally proposed as an adaptation of the NCE method, it is better understood as a special case within the NCE framework.

C.1 Noise Contrastive Estimation

NCE [6] is a relatively new method for solving an old problem: given a sample $(x_i)_{i=1}^N$ (wherein $x_i \in \mathcal{X}$) drawn from an *unknown* probability distribution P_{data}, and a function family $f(\cdot;\theta) : \mathcal{X} \to \mathbb{R}_{\geq 0}$ (parameterized by θ), we want to find the optimal θ^* such that $f(x;\theta^*)$ best approximates the distribution $P_{\mathrm{data}}(x)$. For example, recall the maximum likelihood estimation (MLE), in which θ^* is chosen as to maximize the log-likelihood of the sample $(x_i)_{i=1}^N$, with respect to the constraint that $f(\cdot;\theta^*)$ should be a probability:

$$\theta^*_{\mathrm{MLE}} = \arg\max_\theta \sum_{i=1}^N \ln f(x_i;\theta), \quad \text{s.t.} \sum_{x\in\mathcal{X}} f(x;\theta) = 1.$$

For MLE, the constraint $\sum_{x\in\mathcal{X}} f(x;\theta) = 1$ is important, because $f(x;\theta)$ can tend to arbitrarily large if we maximize the log-likelihood without constraint. NCE finds θ^* in a different way. It firstly mixes (x_i) with a noise sample drawn from a *known* distribution P_{noise}, each data point x_i mixed with k noise points $y_{i,1}, \ldots, y_{i,k} \sim P_{\mathrm{noise}}$. Hence

$$\Pr(x \text{ is data} \mid x) = \frac{P_{\mathrm{data}}(x)}{P_{\mathrm{data}}(x) + kP_{\mathrm{noise}}(x)}, \tag{18}$$

which calculates the probability of a given point $x \in \mathcal{X}$ being a data point. P_{data} is unknown in (18) , so we approximate $\Pr(x \text{ is data} \mid x)$ with $g(x;\theta)$:

$$g(x;\theta) = \frac{f(x;\theta)}{f(x;\theta) + kP_{\mathrm{noise}}(x)}. \tag{19}$$

Then, NCE maximizes the log-likelihood of "x_i being data and $y_{i,1}, \ldots, y_{i,k}$ being noise":

$$\theta^*_{\mathrm{NCE}} = \arg\max_\theta \sum_{i=1}^N \{\ln g(x_i;\theta) + \sum_{j=1}^k \ln(1 - g(y_{i,j};\theta))\}. \tag{20}$$

The most important point of NCE is that, $f(x;\theta)$ will *not* tend to infinity even we maximize (20) *without* the constraint $\sum_{x\in\mathcal{X}} f(x;\theta) = 1$. This is because making $f(x;\theta)$ large will accordingly make $1 - g(y_{i,j};\theta)$ small, which will *decrease* the likelihood of "$y_{i,1}, \ldots, y_{i,k}$ being noise". No longer necessary to repeatedly calculate $\sum_{x\in\mathcal{X}} f(x;\theta)$ during parameter update, NCE usually results in efficient training algorithms.

C.2 The Skip-gram Model

The skip-gram model learns the probability distribution $\Pr(c|t)$ from a corpus $\mathcal{C}$ comprised of target-context pairs [11]. SGNS approximates $\Pr(c|t)$ by the function family $\exp(\mathbf{u}_c \cdot \mathbf{v}_t + \ln kP_{\mathrm{noise}}(c))$, using NCE to optimize parameters. Here P_{noise} is a known noise distribution, and vectors $\mathbf{u}, \mathbf{v}$ are parameters to be learned from $\mathcal{C}$. Hence, if we put $\gamma(c) := \ln(kP_{\mathrm{noise}}(c))$ and $\theta(c,t) := \mathbf{u}_c \cdot \mathbf{v}_t + \gamma(c)$, the function family is defined as $f(c,t;\theta) := \exp(\theta(c,t))$. Substitute this $f(c,t;\theta)$ into (19) and substitute the obtained $g(c,t;\theta)$ into (20) , we get

$$g(c,t;\theta) = \frac{\exp(\theta(c,t))}{\exp(\theta(c,t)) + \exp(\gamma(x))} = \sigma(\mathbf{u}_c \cdot \mathbf{v}_t)$$

where $\sigma(x) = 1/\{1 + \exp(-x)\}$ is the sigmoid function, and the NCE objective (20) becomes

$$\arg\max_{\mathbf{u},\mathbf{v}} \sum_{(t,c)\in\mathcal{C}} \{\ln\sigma(\mathbf{u}_c\cdot\mathbf{v}_t) + \sum_{\substack{j=1\\ n_j\sim P_{\mathrm{noise}}}}^{k} \ln(1-\sigma(\mathbf{u}_{n_j}\cdot\mathbf{v}_t))\}, \tag{21}$$

which is exactly the NEG objective proposed in [16], now explained within the NCE framework.

C.3 Proof of Theorem 2

To prove Theorem 2, we consider $\frac{1}{\#\mathcal{C}}$ times the objective (21) :

$$O(\theta) := \frac{1}{\#\mathcal{C}} \sum_{(t',c')\in\mathcal{C}} \{\ln\sigma(\mathbf{u}_{c'}\cdot\mathbf{v}_{t'}) + \sum_{\substack{j=1\\ n_j\sim P_{\mathrm{noise}}}}^{k} \ln(1-\sigma(\mathbf{u}_{n_j}\cdot\mathbf{v}_{t'}))\}.$$

The above sum is taken across the corpus, in which the term $\ln\sigma(\mathbf{u}_c\cdot\mathbf{v}_t)$ appears $\Pr(c,t)$ times (i.e. we have a probability $\Pr(c,t)$ for the pair (c',t') to be equal to (c,t)), and the term $\ln(1-\sigma(\mathbf{u}_c\cdot\mathbf{v}_t)$ appears $kP_{\mathrm{noise}}(c)\Pr(t)$ times (i.e. we have a probability $P_{\mathrm{noise}}(c)$ for $n_j = c$, and a probability $\Pr(t)$ for $t' = t$). Hence,

$$O(\theta) = \sum_{c,t} \Pr(t)\{\Pr(c|t)\ln\sigma(\mathbf{u}_c\cdot\mathbf{v}_t) + kP_{\mathrm{noise}}(c)\ln(1-\sigma(\mathbf{u}_c\cdot\mathbf{v}_t))\}$$

We know the optimal of $O(\theta)$ is taken at $\mathbf{u}_c\cdot\mathbf{v}_t = s(c,t)$, so put

$$M := \sum_{c,t} \Pr(t)\{\Pr(c|t)\ln\sigma(s(c,t)) + kP_{\mathrm{noise}}(c)\ln(1-\sigma(s(c,t)))\}$$

Then, maximizing $O(\theta)$ is equivalent to minimizing $M - O(\theta)$, and by some calculation, we can find that

$$M - O(\theta) = \sum_{c,t} \Pr(t)\cdot D_\phi(\mathbf{u}_c\cdot\mathbf{v}_t + \gamma(c), s(c,t) + \gamma(c)),$$

where $D_\phi(p,q) := \phi(p) - \phi(q) - \phi'(q)(p-q)$ is the Bregman divergence associated with the convex function

$$\phi(x) = (\Pr(c|t) + e^{\gamma(c)})\ln(e^x + e^{\gamma(c)}).$$

The limit of D_ϕ at $k \to +\infty$ can be easily calculated.

参考文献

[1] Baroni, M. and Zamparelli, R.: Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space, *Proceedings of EMNLP* (2010).

[2] Blacoe, W. and Lapata, M.: A Comparison of Vector-based Representations for Semantic Composition, *Proceedings of EMNLP* (2012).

[3] Clauset, A., Shalizi, C. R. and Newman, M. E. J.: Power-Law Distributions in Empirical Data, *SIAM Rev.*, Vol. 51, No. 4 (2009).

[4] Coecke, B., Sadrzadeh, M. and Clark, S.: Mathematical foundations for a compositional distributional model of meaning, *Linguistic Analysis* (2010).

[5] Foltz, P. W., Kintsch, W. and Landauer, T. K.: The Measurement of Textual Coherence with Latent Semantic Analysis, *Discourse Process* (1998).

[6] Gutmann, M. U. and Hyvärinen, A.: Noise-contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics, *J. Mach. Learn. Res.*, Vol. 13, No. 1 (2012).

[7] Halko, N., Martinsson, P. G. and Tropp, J. A.: Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions, *SIAM Rev.*, Vol. 53, No. 2 (2011).

[8] Kobayashi, H.: Perplexity on Reduced Corpora, *Proceedings of ACL* (2014).

[9] Koren, Y., Bell, R. and Volinsky, C.: Matrix Factorization Techniques for Recommender Systems, *Computer*, Vol. 42, No. 8 (2009).

[10] Landauer, T. K. and Dutnais, S. T.: A solution to Plato ’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological review* (1997).

[11] Levy, O. and Goldberg, Y. .: Dependency-Based Word Embeddings, *Proceedings of ACL* (2014).

[12] Levy, O. and Goldberg, Y. .: Linguistic Regularities in Sparse and Explicit Word Representations, *Proceedings of CoNLL* (2014).

[13] Levy, O. and Goldberg, Y. .: Neural Word Embedding as Implicit Matrix Factorization, *Proceedings of NIPS* (2014).

[14] Li, W.: Random texts exhibit Zipf’s-law-like word frequency distribution, *IEEE Transactions on Information Theory* (1992).

[15] Mikolov, T., Wen-tau Yih and Zweig, G.: Linguistic Regularities in Continuous Space Word Representations, *Proceedings of NAACL-HLT* (2013).

[16] Mikolov, T., Ilya Sutskever, Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Proceedings of NIPS* (2013).

[17] Mitchell, J. and Lapata, M.: Vector-based Models of Semantic Composition, *Proceedings of ACL-HLT* (2008).

[18] Mitchell, J. and Lapata, M.: Composition in distributional models of semantics, *Cognitive Science*, Vol. 34, No. 8 (2010).

[19] Paperno, D., Pham, N. T. and Baroni, M.: A practical and linguistically-motivated approach to compositional distributional semantics, *Proceedings of ACL* (2014).

[20] Pennington, J., Socher, R. and Manning, C.: Glove: Global Vectors for Word Representation, *Proceedings of EMNLP* (2014).

[21] Socher, R., Huang, E. H., Pennin, J., Manning, C. D. and Ng, A. Y.: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, *Proceedings of NIPS* (2011).

[22] Socher, R., Huval, B., Manning, C. D. and Ng, A. Y.: Semantic Compositionality through Recursive Matrix-Vector Spaces, *Proceedings of EMNLP* (2012).

[23] The BNC Consortium: The British National Corpus, version 3 (BNC XML Edition), Distributed by Oxford University Computing Services (2007).
[24] Tian, R., Miyao, Y. and Matsuzaki, T.: Logical Inference on Dependency-based Compositional Semantics, *Proceedings of ACL* (2014).
[25] Zanzotto, F. M., Korkontzelos, I., Fallucchi, F. and Manandhar, S.: Estimating Linear Models for Compositional Distributional Semantics, *Proceedings of Coling* (2010).
[26] Zipf, G. K.: *The Psychobiology of Language: An Introduction to Dynamic Philology*, M.I.T. Press (1935).