Learning Features from Co-occurrences: A Theoretical Analysis

Yanpeng Li
IBM T. J. Watson Research Center
Yorktown Heights, New York 10598, USA
liyanpeng.lyp@gmail.com

Abstract

Representing a word by its co-occurrences with other words in context is an effective way to capture the meaning of the word. However, the theory behind remains a challenge. In this work, taking the example of a word classification task, we give a theoretical analysis of the approaches that represent a word X by a function $f(P(C|X))$, where C is a context feature, $P(C|X)$ is the conditional probability estimated from a text corpus, and the function f maps the co-occurrence measure to a prediction score. We investigate the impact of context feature C and the function f. We also explain the reasons why using the co-occurrences with multiple context features may be better than just using a single one. In addition, based on the analysis, we propose a hypothesis about the conditional probability on zero probability events.

1 Introduction

In natural language processing (NLP) and information retrieval (IR), representing a word by its co-occurrences with contexts is an effective way to learn the meaning of the word and lead to significant improvement in many tasks (Deerwester et al., 1990; Brown et al., 1992; Mikolov et al., 2013). The underlying intuition known as distributional hypothesis (Harris, 1954) can be summarized as: "a word is characterized by the company it keeps" (Firth, 1957). However, there is a lack of theory to justify why it works, even for a simple task such as word classification or clustering. For example, to determine if the word "phycoerythrin" is a gene name, we found it was effective to use the ratio of the count of "phycoerythrin gene" to the count of "phycoerythrin" as features (Li et al., 2009), denoted by

$$\hat{P}(\text{“}Xgene\text{”}|X = phycoerythrin) = \frac{Count(phycoerythrin\ gene)}{Count(phycoerythrin)}$$

The ratio is equivalent to the estimation of the conditional probability $P(\text{“}Xgene\text{”}|X = phycoerythrin)$ from a large text corpus. The pattern "$Xgene$" is a context feature that indicates if the word "gene" appears right next to the word X. Assuming $Y \in \{0, 1\}$ is the gold standard label such as "being a gene name" and $C \in \{0, 1\}$ is a binary context feature such as "$Xgene$". The nature of the approach is to predict $P(Y = 1|X)$ by $P(C = 1|X)$ based on the intuition that there could be correlation between these two functions of X. Also $P(Y = 1|X)$ tends to be more difficult to estimate than $P(C = 1|X)$, since the gold standard labels are more expensive to obtain than the context patterns. We need to know why and when $P(C = 1|X)$ is effective to predict $P(Y = 1|X)$. Obviously, this study is beyond computational linguistics only but of general interest of probability and statistical theory.

In this work, we first show that in a word classification task, simple co-occurrence with a single context feature can achieve perfect performance under some assumptions. Then we investigate the impact of context features and different co-occurrence measures without the assumptions. We also explain the reasons why using co-occurrences with multiple context features, e.g., vector representation can be better than just using a single one. In addition, we give a further analysis of the first theorem for the case of continuous random variables and discuss some hypothesis that may open the door to a new area in probability theory.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2846–2854
Santa Fe, New Mexico, USA, August 20-26, 2018.

2 Notions

We consider a word classification task that assigns each word in a text corpus one or more class labels by the co-occurrences of the word with one or more context patterns. By going through the occurrence of each word in the corpus, we define each example as a tuple $(word, context, labels)$ denoted by $(X, \mathbf{c}, \mathbf{y})$. The component *X* is a discrete random variable taking the values from the set $\{x_1, ..., x_m\}$, where each element refers to a different word. The component $\mathbf{c} = (C_1, ..., C_n)$ is a vector of features, where each component C_k(*k* is from 1 to *n*) is a random variable that takes the value 1 or 0, indicating if a word or pattern appears in the context. Similarly we define $\mathbf{y} = (Y_1, ..., Y_t)$ as a vector of binary labels such as concepts. We aim to investigate how well we can assign each example a set of labels **y** by the co-occurrence information of *X* and C_k with insufficient or without **y** at training stage.

Note that for a label variable Y, there are two cases. It may depend on both *X* and **c** or depend on *X* only. In the tasks such as word sense disambiguation or part-of-speech tagging, the same word can be assigned to different labels in different contexts. In some tasks that investigate the meaning of words independently of contexts, e.g., building WordNet (Miller, 1995) or calculating word-word similarity, the same word X can only be assigned to an unique label 0 or 1 for each *Y*. In other words, *Y* is function of *X*, that is, for each *X*, $P(Y|X)$ equals 0 or 1. We consider both cases in our analysis, but we find that we can get much simpler results for the second case. Moreover, the word classification task for the second case can be generalized to the task of annotating everything with words. For example, describing an image with several words or a sentence is closely related to the task of classifying words into the class of description for the given image.

In the following sections, when we aim to investigate the impact of a single context feature C_k or a single task Y_t, for simplicity, we use a random variable C or Y to denote C_k or Y_t respectively. So the tuple $(X, \mathbf{c}, \mathbf{y})$ can be simplified as (X, C, Y) in the setting of single context feature and single task. To avoid "dividing by zero", we assume that $P(X) \neq 0$, $P(Y) \neq 0$ and $P(C) \neq 0$ in all the sections except Section 6 where we discuss some open problems about the conditional probability on zero probability events.

3 Conditional independence assumption

We show that in a special case if a context feature C is conditionally independent with the word feature X on label Y, the Pearson correlation coefficient between $P(Y = 1|X)$ and $P(C = 1|X)$ equals 1 or -1.

Theorem 1. *Given random variables* $X \in \{x_1, ..., x_m\}$, $C \in \{0, 1\}$ *and* $Y \in \{0, 1\}$, *if*

1. $P(C = 1, Y = 1) \neq P(C = 1)P(Y = 1)$
2. $P(C = 1, X = x_i|Y = 1) = P(C = 1|Y = 1)P(X = x_i|Y = 1)$ *and* $P(C = 1, X = x_i|Y = 0) = P(C = 1|Y = 0)P(X = x_i|Y = 0)$ *for every i from 1 to m*

, we have:

$$Corr(P(Y = 1|X), P(C = 1|X))^2 = 1$$

Proof.

$$\begin{aligned}
P(C = 1|X) &= \frac{1}{P(X)}(P(C = 1, X|Y = 1)P(Y = 1) + P(C = 1, X|Y = 0)P(Y = 0)) \\
&= \frac{1}{P(X)}(P(C = 1|Y = 1)P(X|Y = 1)P(Y = 1) \\
&+ P(C = 1|Y = 0)P(X|Y = 0)P(Y = 0)) \\
&\qquad\qquad \text{(using the conditional independence assumption)} \\
&= \frac{1}{P(X)}(\frac{P(C = 1, Y = 1)P(X, Y = 1)}{P(Y = 1)} + \frac{P(C = 1, Y = 0)P(X, Y = 0)}{P(Y = 0)})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{P(X)}\left(\frac{P(C=1,Y=1)P(X,Y=1)}{P(Y=1)}\right. \\
&\left.+ \frac{(P(C=1)-P(C=1,Y=1))(P(X)-P(X,Y=1))}{P(Y=0)}\right) \\
&= \left(\frac{P(C=1,Y=1)}{P(Y=1)} - \frac{P(C=1)-P(C=1,Y=1))}{P(Y=0)}\right)P(Y=1|X) \\
&+ \frac{P(C=1)-P(C=1,Y=1)}{P(Y=0)} \\
&= \frac{P(C=1,Y=1)-P(C=1)P(Y=1)}{P(Y=1)P(Y=0)}P(Y=1|X) + P(C=1|Y=0)
\end{aligned}
$$

Since $P(C=1,Y=1) \neq P(C=1)P(Y=1)$, $P(C=1|X)$is a linear function of $P(Y=1|X)$, so we have $Corr(P(Y=1|X), P(C=1|X))^2 = 1$ □

This theorem indicates that a perfect context feature for a given task is not necessarily the class label itself. Therefore, even if we have little information about Y, we still have the chance to know the information of $P(Y|X)$ by $P(C|X)$. Some previous works also reported similar findings in different tasks (Blum and Mitchell, 1998; Abney, 2002; Veeramachaneni and Kondadadi, 2009; Li, 2013). However, in practice it is almost impossible to get the case with exact conditional independence in different tasks. Even if it exists, it is difficult to find it because in order to calculate the conditional dependence we still need to know the joint probability $P(X,C,Y)$. Therefore, we need a theory to describe the performance in the situation with certain degree of conditional dependence. We also need to know the impact of different functions that convert simple co-occurrence measures e.g., $P(C=1|X)$ to feature values or prediction scores. In the following sections, we will show the results for the cases.

4 Co-occurrences with a single context feature

In this section, we study the case of co-occurrence with a single context feature for a single task. Although it seems simple, there is still no systematic theoretical framework for it. Based on the theory about the function of discrete random variables, we analyze the impact of context features and the functions that convert co-occurrence measures to prediction scores. Note that the cases discussed in this section are not under the conditional independence assumption introduced in Section 3.

4.1 Function of a discrete random variable

Since the co-occurrence based learning actually converts the word variable $X \in \{x_1, ..., x_m\}$ to another variable $S \in \{s_1, ..., s_u\}$ by certain co-occurrence measure such as $P(X, C=1)$ or $P(C=1|X)$, or equivalently, we can say S is a function of X. Our goal is to find the function that can achieve the best performance. Therefore, we investigate some general principles about the function of a discrete variable. In this work, we use Pearson correlation coefficient of conditional probabilities as the measure of performance, because it tends to make the complicated analysis simpler.

Lemma 1. *Let g be a function that maps $X \in \{x_1, ..., x_m\}$ to another random variable $S \in \{s_1, ..., s_u\}$. For any binary random variable $Y \in \{0,1\}$ and any function f that assigns S a real number, we have the following results:*

(1) $Cov(P(Y=1|X), f(S)) = Cov(P(Y=1|S), f(S))$

(2) $Cov(P(Y=1|X), P(Y=1|S)) = Var(P(Y=1|S))$

(3) $Corr(P(Y=1|X), P(Y=1|S)) = \frac{\sqrt{Var(P(Y=1|S))}}{\sqrt{Var(P(Y=1|X))}}$

(4) $Corr(P(Y=1|X), f(S)) = Corr(P(Y=1|S), f(S))Corr(P(Y=1|X), P(Y=1|S))$

Proof. (1).

$$E(P(Y=1|X)) = \sum_{i=1}^{m} \frac{P(X=x_i)P(Y=1,X=x_i)}{P(X=x_i)} = \sum_{i=1}^{m} P(Y=1,X=x_i) = P(Y=1)$$

Similarly, we can prove $E(P(Y=1|S)) = P(Y=1) = E(P(Y=1|X))$

$$Cov(P(Y=1|X), f(S)) = Cov(P(Y=1|X), f(g(X))) \qquad \text{(the definition of } g\text{)}$$
$$= \sum_{i=1}^{m} P(X=x_i)(P(Y=1|X=x_i) - E(P(Y=1|X)))(f(g(x_i)) - E(f(g(X))))$$
$$= \sum_{j=1}^{u} (f(s_j) - E(f(S))) \sum_{i \in \{i|g(x_i)=s_j\}} P(X=x_i)(P(Y=1|X=x_i) - E(P(Y=1|X)))$$
$$= \sum_{j=1}^{u} (f(s_j) - E(f(S)))P(S=s_j)(P(Y=1|S=s_j) - E(P(Y=1|X)))$$
$$= \sum_{j=1}^{u} P(S=s_j)(f(s_j) - E(f(S)))(P(Y=1|S=s_j) - E(P(Y=1|S)))$$
$$\text{(by } E(P(Y=1|X) = E(P(Y=1|S))\text{)}$$
$$= Cov(P(Y=1|S), f(S))$$

(2). Since $P(Y=1|S)$ is a function of S, let $f(S)$ be $P(Y=1|S)$ and using (1), and we have $Cov(P(Y=1|X), P(Y=1|S)) = Cov(P(Y=1|S), P(Y=1|S)) = Var(P(Y=1|S))$

(3). Based on (2), we have

$$Corr(P(Y=1|X), P(Y=1|S)) = \frac{Cov(P(Y=1|X), P(Y=1|S)))}{\sqrt{Var(P(Y=1|X))}\sqrt{Var(P(Y=1|S))}}$$
$$= \frac{Var(P(Y=1|S)))}{\sqrt{Var(P(Y=1|X))}\sqrt{Var(P(Y=1|S))}} = \frac{\sqrt{Var(P(Y=1|S)))}}{\sqrt{Var(P(Y=1|X))}}$$

(4). By combining (1) and (3), we have

$$Corr(P(Y=1|X), f(S)) = \frac{Cov(P(Y=1|X), f(S)))}{\sqrt{Var(P(Y=1|X))}\sqrt{Var(f(S))}}$$
$$= \frac{Cov(P(Y=1|S), f(S)))}{\sqrt{Var(P(Y=1|X))}\sqrt{Var(f(S))}} = \frac{Cov(P(Y=1|S), f(S)))}{\sqrt{Var(P(Y=1|S))}\sqrt{Var(f(S))}} \frac{\sqrt{Var(P(Y=1|S))}}{\sqrt{Var(P(Y=1|X))}}$$
$$= Corr(P(Y=1|S), f(S))Corr(P(Y=1|X), P(Y=1|S))$$

□

The first equation plays a fundamental role, which indicates that if S is function of X, to calculate the covariance between $P(Y=1|X)$ and $f(S)$, we don't need the appearance of X but just S, as if X is safely hidden. It simplifies the analysis by avoiding explicit analysis of the divergence between $P(Y=1|X)$ and $f(S)$ such as $|P(Y=1|X) - f(S)|$ or $P(Y=1|X)f(S)$. Therefore, the correlation coefficient can also be written in a much simpler form. In the word classification task, the term $f(S)$ can be viewed as a classifier based on the new features S. The $Corr(P(Y=1|X), f(S))$ measures its correlation with the best possible classifier $P(Y=1|X)$ based on the word features X (single view) only. If Y is a function of X (the second case discussed in Section 2), we have $Corr(P(Y=1|X), f(S)) = Corr(Y, f(S))$, which measures almost exactly the performance of the word classification task. Based on these results, we are able to analyze the performance of the new features generated by co-occurrences.

4.2 An upper bound of any function

Theorem 2. *Given random variables $X \in \{x_1, ..., x_m\}$, $C \in \{0, 1\}$ and $Y \in \{0, 1\}$, for any function f that maps $P(C = 1|X)$ to a real number, there is:*

$$Corr(P(Y = 1|X), f(P(C = 1|X)))^2$$
$$= Corr(P(Y = 1|P(C = 1|X)), f(P(C = 1|X)))^2 Corr(P(Y = 1|X), P(Y = 1|P(C = 1|X)))^2$$

Proof. Since $P(C = 1|X)$ is a function of X, let S be the random variable that takes the value $P(C = 1|X)$ and based on Lemma 1 (4), we have:

$$Corr(P(Y = 1|X), f(S))^2 = Corr(P(Y = 1|S), f(S))^2 Corr(P(Y = 1|X), P(Y = 1|S))^2$$

Equivalently, we can write the equation as:

$$Corr(P(Y = 1|X), f(P(C = 1|X)))^2$$
$$= Corr(P(Y = 1|P(C = 1|X)), f(P(C = 1|X)))^2 Corr(P(Y = 1|X), P(Y = 1|P(C = 1|X)))^2$$

□

It shows that the performance of the classifier $f(P(C = 1|X))$ is determined by the product of two parts. The first part depends on both the context feature C and the function f. The second depends on the context feature only but not on f. Therefore, if we fix the context feature C to select the function f, using $Corr(P(Y = 1|X), f(P(C = 1|X)))^2$ (correlation with the gold standard $P(Y = 1|X)$) and $Corr(P(Y = 1|P(C = 1|X)), f(P(C = 1|X)))^2$ (correlation with the "silver" standard $P(Y = 1|P(C = 1|X))$) produce the same result. Similar to Lemma 1(1), we don't need the appearance of $P(Y = 1|X)$ for function selection. In addition, since any correlation coefficient ranges from -1 to 1, we have:

$$Corr(P(Y = 1|X), f(P(C = 1|X)))^2 \leq Corr(P(Y = 1|X), P(Y = 1|P(C = 1|X)))^2$$

It indicates that if we want to improve the performance by fixing a certain context feature C and changing different functions f , e.g., from $P(C = 1|X)$ to $log(P(C = 1, X)/(P(C = 1)P(X)))$ (the point wise mutual information), the performance cannot be arbitrarily high but upper bounded by the squared correlation coefficient between $P(Y = 1|X)$ and $P(Y = 1|P(C = 1|X))$. Given fixed word feature set X and label Y, the upper bound is determined only by the context feature C but not relevant to the function f. Actually, the upper bound is a special case of maximal correlation coefficient (Rényi, 1959; Hall, 1967). In order to build a high performing classifier $f(P(C = 1|X))$, we must find a way to improve the upper bound $Corr(P(Y = 1|X), P(Y = 1|P(C = 1|X)))^2$, in other words, to select a good context feature C. However, it is not easy to do it from this formula directly. Therefore, we relate the upper bound to the special case introduced in Section 3. Since $P(C = 1|X)$ is a function of itself, based on Theorem 2, we have:

$$Corr(P(Y = 1|X), P(C = 1|X))^2 \leq Corr(P(Y = 1|X), P(Y = 1|P(C = 1|X)))^2$$

The squared correlation coefficient $Corr(P(Y = 1|X), P(C = 1|X))^2$ not only reflects the performance of a special co-occurrence measure $P(C = 1|X)$, but also provide a way to improve the upper bound (for other co-occurrence measures). Based on Theorem 1, under the conditional independence assumption, $Corr(P(Y = 1|X), P(C = 1|X))^2$ equals 1, so that $Corr(P(Y = 1|X), P(Y = 1|P(C = 1|X)))^2$ equals 1 as well. As what we discussed in Section 3, it is important to know more about the cases out of the conditional independence assumptions, but it is difficult to analyze $Corr(P(Y = 1|X), P(C = 1|X))^2$ based on the conditional dependence directly without any other assumptions. In the following section, we show that we are able to obtain much simpler results under the assumption that the label variable $Y \in \{0, 1\}$ is a function of X.

4.3 Assuming that *Y* is a function of *X*

In Section 2, we have pointed out that there are many cases in practice that assume *Y* is a function of *X*. In addition, if we can explain everything about this simple task, we may naturally find the cues to move to more advanced tasks such as sentence classification.

Theorem 3. *If* $Y \in \{0,1\}$ *is a function of* $X \in \{x_1, ..., x_m\}$*, for a context feature* $C \in \{0,1\}$ *we have:*

$$Corr(P(Y=1|X), P(C=1|X))^2$$
$$= Corr(P(Y=1|Y), P(C=1|Y))^2 Corr(P(C=1|X), P(C=1|Y))^2$$

Proof. Since *Y* and *C* are both binary random variables, by Lemma 1 (4) we replace Y by C without loss of generality, and we have:

$$Corr(C=1|X), f(S))^2 = Corr(P(C=1|S), f(S))^2 Corr(P(C=1|X), P(C=1|S))^2$$

Since *Y* is a function of *X* and $P(Y=1|Y)$is a function of Y, in the above equation replace *S* by *Y* and $f(S)$ by $P(Y=1|Y)$, and we have:

$$Corr(C=1|X), P(Y=1|Y))^2 = Corr(P(C=1|Y), P(Y=1|Y))^2 Corr(P(C=1|X), P(C=1|Y))^2$$

Since *Y* is a function of *X*, for any X, there is $P(Y=1|X) = P(Y=1|Y) = Y$. So, we have:

$$Corr(C=1|X), P(Y=1|X))^2 = Corr(C=1|X), P(Y=1|Y))^2$$
$$= Corr(P(Y=1|Y), P(C=1|Y))^2 Corr(P(C=1|X), P(C=1|Y))^2$$

□

This result describes what a good context feature is and can give guidance for the selection of context features for a particular task. It is much simpler and more practical than what we can get by the direct analysis of the conditional dependence, since we just need to know the joint distribution of (X, C) and (C, Y) rather than (X, C, Y). This result can also be written as a simpler form as follows.

Corollary 1. *If* $Y \in \{0,1\}$ *is a function of* $X \in \{x_1, ..., x_m\}$*, for a context feature* $C \in \{0,1\}$ *we have:*

$$Corr(P(Y=1|X), P(C=1|X))^2 = \frac{(P(C=1, Y=1) - P(C=1)P(Y=1))^2}{P(Y=1)(P(Y=0))Var(P(C=1|X))}$$

Proof.

$$Corr(P(Y=1|X), P(C=1|X))^2$$
$$= Corr(P(Y=1|Y), P(C=1|Y))^2 Corr(P(C=1|X), P(C=1|Y))^2 \quad \text{(Theorem 3)}$$
$$= \frac{Cov(P(Y=1|Y), P(C=1|Y))^2}{Var(P(Y=1|Y))Var(P(C=1|Y))} \frac{Var(P(C=1|Y))}{Var(P(C=1|X)} \quad \text{(Lemma 1 (3))}$$
$$= \frac{(E(P(Y=1|Y)P(C=1|Y)) - E(P(Y=1|Y))E(P(C=1|Y)))^2}{Var(P(Y=1|Y))} \frac{1}{Var(P(C=1|X)}$$
$$= \frac{(P(C=1, Y=1) - P(C=1)P(Y=1))^2}{P(Y=1)(P(Y=0))Var(P(C=1|X))}$$

□

In the equation, the term $(P(C=1, Y=1) - P(C=1)P(Y=1))^2$ describes the dependency between context feature C and the class label Y. The term $Var(P(C=1|X))$ addresses the dependency between context feature C and each word feature X. We can see more clearly in another equivalent form:

$$Corr(P(Y=1|X), P(C=1|X))^2 = \frac{P(Y=1)}{P(Y=0)} \frac{(\frac{P(C=1,Y=1)}{P(C=1)P(Y=1)} - 1)^2}{E((\frac{P(C=1,X)}{P(C=1)P(X)} - 1)^2)}$$

From this mutual information style form, we can conclude that a good context feature for a particular task should be a trade-off between high dependence with label Y and low dependence with every word feature X. The conclusion is similar to our prevsious work (Li, 2013), but the result here is more general, under weaker assumptions, and more accurate (e.g., equations rather than inequalities). During the engineering work of context feature selection, intuitively, people tend to emphasize the numerator (tight connection with the label Y), but ignore the denominator (loose connection with each word X). However, the theory indicates that we should consider both factors to make a good co-occurrence. For example, some experiments showed that the co-occurrences with simply high frequency words such as "the" and "of" chould be used as good features (Li and Yu, 2014). Obviously, for these function words, in the equation both numerator and denominator tend to be smaller than the class-indicative words such as "$Xgene$" mentioned in Section 1, but the ratio of them may be larger. Therefore, it will be interesting to investigate what happen if both numerator and denominator approach to zero. Actually, the scope of candidates of context features is far beyond context words, since any binary pattern such as letters or sound that appears in the context can be a context feature.

5 Co-occurrences with multiple context features

So far, we have investigated the co-occurrences with a single context feature only. In practice, we usually find representing a word by a vector of co-occurrences with multiple context features tends to perform better. In the following theorem, we show part of the reason by analyzing the performance of the vector from multiple context features.

Theorem 4. *If we represent each X by a r-dimensional vector* $\mathbf{x} = (P(C_1 = 1|X), ..., P(C_r = 1|X))$*, where* $C_1, ..., C_r$ *are r context features (*$r \leq n$*), for the label* $Y \in \{0, 1\}$ *and any funcation f that maps each veoctor* $\mathbf{x}$ *to a real number, we have:*

(1) $Corr(P(Y=1|X), f(\mathbf{x}))^2 = Corr(P(Y=1|\mathbf{x}), f(\mathbf{x}))^2 Corr(P(Y=1|X), P(Y=1|\mathbf{x}))^2$

(2) $Corr(P(Y=1|X), P(Y=1|\mathbf{x}))^2 = \frac{Corr(P(Y=1|X), P(Y=1|P(C_k=1|X)))^2}{Corr(P(Y=1|\mathbf{x}), P(Y=1|P(C_k=1|X)))^2}$ *for every k from 1 to n*

Proof. (1) Since X is a discrete random variable and $\mathbf{x} = (P(C_1 = 1|X), ..., P(C_r = 1|X))$, we know that $\mathbf{x}$ is a discrete random variable and a function of X as well. Therefore, based on Lemma 1 (4), we come to the conclusion directly:

$$Corr(P(Y=1|X), f(\mathbf{x}))^2 = Corr(P(Y=1|\mathbf{x}), f(\mathbf{x}))^2 Corr(P(Y=1|X), P(Y=1|\mathbf{x}))^2$$

(2) Since for each $\mathbf{x} = (P(C_1 = 1|X), \ldots, P(C_r = 1|X))$ with the a different value, the value of $P(C_k = 1|X)$is unique, the discrete random variable $P(C_k = 1|X)$ is a function of $\mathbf{x}$. In Lemma 1 (3), replace X by $\mathbf{x}$, and S by $P(C_k = 1|X)$ without loss of generality, and we have

$$Corr(P(Y=1|\mathbf{x}), P(Y=1|P(C_k=1|X)))^2 = \frac{Var(P(Y=1|P(C_k=1|X)))}{Var(P(Y=1|\mathbf{x}))}$$

Since $\mathbf{x}$ is a function of X and $P(C_k = 1|X)$ is a function of X, according to Lemma 1 (3), we have

$$Corr(P(Y=1|X), P(Y=1|\mathbf{x}))^2 = \frac{Var(P(Y=1|\mathbf{x}))}{Var(P(Y=1|X))}$$

$$Corr(P(Y=1|X), P(Y=1|P(C_k=1|X)))^2 = \frac{Var(P(Y=1|P(C_k=1|X)))}{Var(P(Y=1|X))}$$

By combining the three equations above, we have:

$$Corr(P(Y=1|X), P(Y=1|\mathbf{x})) = \frac{Corr(P(Y=1|X), P(Y=1|P(C_k=1|X)))}{Corr(P(Y=1|\mathbf{x}), P(Y=1|P(C_k=1|X))^2}$$

□

Similar to Theorem 2, the performance of the vector-style representation can also be written as the products of two parts. The second part $Corr(P(Y=1|X), P(Y=1|\mathbf{x}))^2$ is an upper bound of the performance of any function, which is determined only by the set of context features $C_1, ..., C_r$. The result (2) shows that the upper bound from the vector is always better than the performance from any individual context feature, since we can prove $Corr(P(Y=1|X), P(Y=1|\mathbf{x})) \geq Corr(P(Y=1|X), P(Y=1|P(C_k|X)))^2$. Similar to the proof of (2), it can be proved that $Corr(P(Y=1|X), P(Y=1|\mathbf{x}))^2$ always increases when more context features are introduced. However, $Corr(P(Y=1|\mathbf{x}), f(\mathbf{x}))^2$ may decrease when the vector become longer, so its real performance $Corr(P(Y=1|X), f(\mathbf{x}))^2$ may increase or decrease depending on the first term. Therefore, we need to find the factors that determine the combined performance of the two terms in the future.

6 Assuming that the probability of *X* is zero

In the previous sections, we discussed the impact of context feature C. In addition, we should also consider the impact of the content feature X. We found an interesting case that satisfies exactly the conditional independence assumption in Theorem 1, although it is beyond the word classification task. If we assume that X is a continuous random variable such as $X \in \mathbb{R}$, we have $P(X=x)=0$, $P(X=x, C|Y)=0$, and $P(X=x|Y)=0$ for every real number $x \in \mathbb{R}$, which fits exactly the conditional independence assumption. The trouble is that we have the conditional probability $P(C=1|X)$ and $P(Y=1|X)$ in the format of $0/0$. It is well known that the conditional probability on an event with zero probability is undefined in the current probability theory, but there is no proof that it cannot be defined logically. If we assume that in some condition $P(C=1|X)$ and $P(Y=1|X)$ exist and are between 0 and 1, for example, as if we may define $f(x)=\frac{sin(x)}{x}$ when $x=0$ as $f(0)=\lim_{x\to 0}\frac{sin(x)}{x}$, the open question is: does the following equation (from Section 3) still hold?

$$P(C=1|X) = \frac{P(C=1,Y=1)-P(C=1)P(Y=1)}{P(Y=1)P(Y=0)}P(Y=1|X) + P(C=1|Y=0)$$

If the answer is "Yes", it suggests that given every specific point $X=x$, the possibility of appearance of any two non-independent events (e.g., $C=1$ and $Y=1$) has a linear relationship. One underlying philosophy is that two things with low dependence (e.g., low semantic similarity) could have strong correlation in some case, since $Corr(P(C=1|X), P(Y=1|X))$ is always 1 or -1 even if $|\frac{P(C=1,Y=1)-P(C=1)P(Y=1)}{P(Y=1)P(Y=0)}|$ is very small but not zero. Based on it we may have even more interesting findings. If the answer is "No", we would have a more complicated result, e.g., two or more different equations: one for $P(X)\neq 0$ and one or more for $P(X)=0$. An interesting fact is that for Bayes' theorem the answer of the similar question is "Yes", since if we assume $P(X)=0$ and $P(C=1|X)\in[0,1]$, we will have $P(X|C=1)=P(C=1|X)P(X)/P(C=1)=0$.

We also believe that such probability is more useful in practice because it gives to more exact knowledge and prediction. For example, assuming that $X=x$ is a specific time or position represented by a real number, $P(Y=1|X=x)$ (where $P(X=x)=0$) could mean the possibility of the occurrence of Y at the point of time or position x, which is more exact than the probability $P(Y=1|X\in[x1,x2])$, where $P(X\in[x1,x2])\neq 0$. Interestingly, such model is closely related to physics, and we will give a further analysis in the future.

7 Conclusions and future works

In this paper, we give a theoretical study of the approaches that learn the representation of word by the approaches like $f(P(C=1|X))$ or $f(P(C_1=1|X), ..., P(C_r=1|X))$. The theoretical framework is able to explain a set of approaches based on distributional semantics and give guidance for algorithm

design such as the selection of context feature and the co-occurrence metrics. In the next steps, we are going to give a deeper analysis of each formula that determines the performance, such as the diversity between multiple context features, find more principles that are constructive to algorithm design in practice and extend the theory to analyze other advanced tasks such as sentence classification and semantic similarity. Moreover, it will be interesting to see if we can verify the hypothesis in Section 6 and get inspiration from it.

References

[Deerwester et al.1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391.

[Brown et al.1992] Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D. and Lai, J. C. 1992. Class-based N-gram Models of Natural Language. *Computational linguistics*, 18(4), 467-479.

[Mikolov et al.2013] Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*, 1301.3781.

[Harris1954] Harris, Z. S. 1954. Distributional Structure. *Word*, 10(2-3), 146-162.

[Firth1957] Firth, J. R. 1957. A Synopsis of Linguistic Theory. *Studies in Linguistic Analysis*, 1930-1955.

[Li et al.2009] Li, Y., Lin, H. and Yang, Z. 2009. Incorporating Rich Background Knowledge for Gene Named Entity Classification and Recognition. *BMC Bioinformatics*, 10(1), 223

[Miller1995] Miller, G. A. 1954. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11), 39-41.

[Blum and Mitchell1998] Blum, A. and Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Co-training *Proceedings of the eleventh annual conference on Computational learning theory*, 92-100

[Abney2002] Abney, S. 2002. Bootstrapping. *Proceedings of the 40th Annual Meeting of the Association for Computa- tional Linguistics*, 360–367.

[Veeramachaneni and Kondadadi2009] Veeramachaneni, S. and Kondadadi, R. K. 2009. Surrogate Learning: from Feature Independence to Semi-supervised Classification *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, 10-18.

[Li2013] Li, Y. 2013. Reference Distance Estimator. *arXiv preprint*, 1308.3818.

[Rényi1959] Rényi, A. 1959. On Measures of Dependence. *Acta mathematica hungarica*, 10(3-4), 441-451.

[Hall1967] Hall, W. J. 1967. On Characterizing Dependence in Joint Distributions. University of North Carolina, Department of Statistics.

[Li and Yu2014] Li, Y. and Yu, H. 2014. A Robust Data-driven Approach for Gene Ontology Annotation. *Database*.