Online Computation of Mutual Information and Word Context Entropy

Wei-Hsuan Lin, Yi-Lun Wu, and Liang-Chih Yu

Abstract—Mutual information (MI) has been extensively used to measure the co-occurrence strength between two words in the field of natural language processing. Similarly, the word context entropy is also a useful measure to determine the distribution of words in contexts, and can be used to calculate word similarity. Calculating scores for both measures usually relies on a large text corpus to obtain a reliable estimation. However, calculation based on a static corpus may not reflect the dynamic nature of languages. In this paper, we consider the web documents as a text corpus, and develop an efficient online calculator for both mutual information and word context entropy. The major advantage of the online computation is that the web corpus not only is large enough to obtain a reliable estimation but also can reflect the dynamic nature of languages.

Index Terms—Mutual information, word context, entropy, natural language processing.

I. INTRODUCTION

Mutual information (MI) or pointwise mutual information (PMI) is a measure used to determine the co-occurrence strength between two words, and a high PMI score indicates a frequently co-occurred word pair. Knowing frequently co-occurred words is useful for many natural language applications such as lexical substitution [1], [2], feature selection [3], [4], and template matching [5]. Similarly, word context entropy is also a useful measure to determine the distribution of words in contexts, and a high entropy score indicates an even distribution, otherwise, a skewed distribution. By comparing the contextual distributions of two words, their similarity can be estimated [6], [7]. Knowing words with similar meanings is crucial for semantic-oriented applications such as (near-)duplicate detection for text summarization [8], concept mapping [9], [10], [11], computer-assisted language learning (CALL) [12], [13], [14], and query expansion in information retrieval (IR) [15], [16], [17], [18], [19]. Calculating scores for both measures usually relies on a large text corpus to obtain a reliable estimation. Researchers can use existing corpora to calculate both measures. However, such corpora are usually static knowledge resources because their contents are not updated with time, thus may not reflect the dynamic nature of languages. Furthermore, such large corpora may be unavailable for some application domains. Therefore, this study considers the web documents as a text corpus, and develops an efficient online calculator for both mutual information and word context entropy. The major advantage of the online computation is that the web corpus not only is large enough to obtain a reliable estimation but also can reflect the dynamic nature of languages. The aim of this work is summarized below.

1) **Online computation of pointwise mutual information:** To calculate the PMI score of two words, both the co-occurrence frequencies of the two words and frequencies of the individual words are obtained by querying Google.

2) **Online computation of word context entropy:** To calculate the context entropy of a word, the context distribution is estimated from the document titles containing the word returned by Google.

The rest of this work is organized as follows. Section 2 presents some related work. Section 3 describes the online computation procedure for pointwise mutual information and word context entropy. Conclusions are finally drawn in Section 4.

II. RELATED WORK

MI or PMI has been extensively used in the field of natural language processing. For the application of near-synonym substitution, PMI was used to examine whether a word matches the given contexts in so-called “fill-in-the-blank” (FITB) task [1]. Given a near-synonym set and a sentence containing one of the near-synonyms, the near-synonym was first removed from the sentence to form a lexical gap. The goal is to predict an answer (best near-synonym) that can fill the gap from the given near-synonym set. In this task, PMI was used to measures the co-occurrence strength between a near-synonym and the words in its context. A higher mutual information score indicates that the near-synonym fits well in the given context, and thus is more likely to be the correct answer. In feature selection, Doquire and Verleysen addressed the problem by adapting the MI criterion to handle missing data using a partial distance strategy [3]. Yu et al. acquired useful language patterns by incorporating the MI criterion into association rule mining to recursively discover frequent co-occurring words from a corpus of sentences [4]. Maciej et al. proposed the use of a mutual information-based template matching scheme to develop a computer-aided detection system for mammographic masses [5].

Estimating word context entropy usually relies on a vector representation of word contexts. For example, the Hyperspace Analog to Language (HAL) model constructed a high-dimensional context space to represent words [20]. Based on this representation, a word was represented as a vector of its context words where each dimension denotes the
weight of a context word. The word weights were then transformed into probabilistic representation. Each word (vector) thus can be viewed as a distribution of word contexts, and the context entropy of the word can then be estimated based on the context distribution in the vector. Furthermore, the similarity of the two words can be estimated by comparing the contextual distributions of two words (vectors) [6][7].

III. ONLINE COMPUTATION PROCEDURE

A. Pointwise Mutual Information

The pointwise mutual information [21] between two words \(x\) and \(y\) is defined as

\[
PMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)},
\]

(1)

where \(P(x, y) = C(x, y)/N\) denotes the probability that \(x\) and \(y\) co-occur; \(C(x, y)\) is the number of times \(x\) and \(y\) co-occur in the corpus, and \(N\) is the total number of words in the corpus. Similarly, \(P(x) = C(x)/N\), where \(C(x)\) is the number of times \(x\) occurs in the corpus, and \(P(y) = C(y)/N\), where \(C(y)\) is the number of times \(y\) occurs in the corpus. Therefore, (1) can be re-written as

\[
PMI(x, y) = \log_2 \frac{C(x, y) \cdot N}{C(x) \cdot C(y)}.
\]

(2)

All the frequency counts presented above are retrieved by querying Google. The value of \(N\) is usually unknown when using Google as the corpus. Therefore, we herein use \(N = 10^{12}\), the number of tokens in the Web 1T 5-gram corpus released by Linguistic Data Consortium (LDC). Fig. 1 shows an example of calculating the PMI score of two words natural and language. In this example, \(C(\text{natural, language}) = 16,900,000\), \(C(\text{natural}) = 2,420,000,000\), and \(C(\text{language}) = 3,890,000,000\), thus yielding a PMI score 1.80. Similarly, the PMI score of police and flower is 0.12, which is much smaller than that of natural and language, indicating that natural and language are more frequently co-occurred than police and flower.

B. Word Context Entropy

This measure is used to determine the context entropy of a word based on the distribution of its context words. To measure the distribution of word contexts, we use the document titles returned by Google as the corpus. Fig. 2 shows the search results for the keyword breaking.

Once the titles are obtained, the words occurring in the context of breaking can be extracted, and their frequency counts can also be retrieved from the corpus of document titles. Table 1 lists five most frequently occurred context words in the first 20 returned titles containing breaking. The proportion of each context word is defined as the frequency count of the word divided by the total frequency counts of all context words. According to the distribution of context words, the context entropy of a word can be defined as [22]

\[
H(w_i) = - \sum_{w_j \in \text{Context}(w_i)} P(w_j) \log_2 P(w_j),
\]

(3)

where \(H(w_i)\) denotes the entropy of \(w_i\), and \(w_j\) is a word occurring in the context of \(w_i\). For the sample word breaking, its entropy is 1.97 calculated based on the context distribution in Table 1. In our implementation, the online entropy calculator will output the entropy for both left and right contexts of a given word. In addition, the size of the title corpus is also adjustable by specifying the number of titles returned by Google (the maximum number is 999 per query).

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Example of calculating the PMI score by querying Google.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Sample search results for the keyword breaking.}
\end{figure}

\begin{table}
\centering
\begin{tabular}{|l|l|l|}
\hline
Context word & Frequency & Proportion \\
\hline
news & 11 & 0.50 \\
com & 4 & 0.18 \\
bad & 3 & 0.14 \\
dawn & 2 & 0.09 \\
free & 2 & 0.09 \\
\hline
\end{tabular}
\caption{Context distribution of the sample word `breaking`.}
\end{table}

IV. CONCLUSIONS

This work presents an online calculator for mutual
information and word context entropy. Both measures are calculated from the web by querying Google. The PMI is calculated using the frequency counts returned by Google, while the entropy is calculated from the returned document titles. Applications can benefit from such an online computation procedure to provide more reliable estimation due to the huge size of web corpora. The calculation results are also able to reflect the dynamic nature of languages. Future work will focus on incorporating the online computation module into real applications.

ACKNOWLEDGEMENTS
This work was supported by the National Science Council, Taiwan, ROC, under Grant No. NSC99-2221-E-155-036-MY3.

REFERENCES