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

Manuscript received April 04, 2012; revised May 8, 2012.
The authors are with Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C (Corresponding author: Tel.: + 886-3-463-8800, fax: + 886-3-435-2077, e-mail address: lcyu@saturn.yzu.edu.tw).

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

$$PMI(x,y)=\log_2\frac{P(x,y)}{P(x)P(y)}, \tag{1}$$

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

$$PMI(x,y)=\log_2\frac{C(x,y)\cdot N}{C(x)\cdot C(y)}. \tag{2}$$

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(natural, language) = 16,900,000, C(natural) = 2,420,000,000, and C(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*.

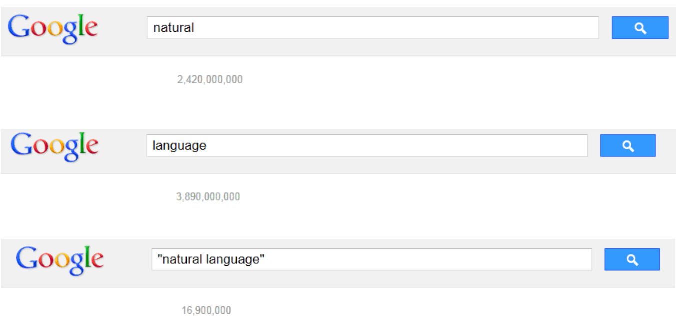

Fig. 1. Example of calculating the PMI score by querying Google.

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]

$$H(w_t)=-\sum_{w_i\in Context(w_t)}P(w_i)\log_2 P(w_i), \tag{3}$$

where $H(w_t)$ denotes the entropy of w_t, and w_i is a word occurring in the context of w_t. 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).

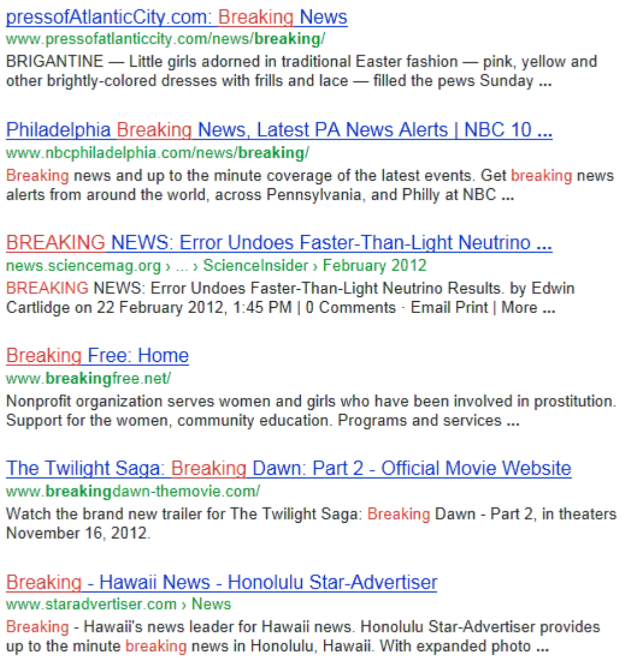

Fig. 2. Sample search results for *breaking*.

TABLE I: Context Distribution of the Sample Word Breaking.

Context word	Frequency	Proportion
news	11	0.50
com	4	0.18
bad	3	0.14
dawn	2	0.09
free	2	0.09

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

[1] D. Inkpen, “A Statistical Model of Near-synonym Choice,” *ACM Trans. Speech and Lang. Process.* 2007, vol. 4, no. 1, pp. 1-17.

[2] L. C. Yu, C. H. Wu, R. Y. Chang, C. H. Liu, and E. H. Hovy, “Annotation and Verification of Sense Pools in OntoNotes,” *Inf. Process. Manage*. 2010, vol. 46, no. 4, pp. 436-447.

[3] L. C. Yu, C. L. Chan, C .C. Lin, and I. C. Lin, “Mining Association Language Patterns Using a Distributional Semantic Model for Negative Life Event Classification,” *J. Biomed. Inform.* 2011, vol. 44, no. 4, pp. 509-518.

[4] G. Doquire and M. Verleysen, “Feature Selection with Missing Data Using Mutual Information Estimators,” *Neurocomputing*. 2012, vol. 90, no. 1, pp. 3-11.

[5] M. A. Mazurowski, J. Y. Lo, B. P. Harrawood, and G. D. Tourassi, “Mutual Information-based Template Matching Scheme for Detection of Breast Masses: From Mammography to Digital Breast Tomosynthesis,” *J. Biomed. Inform.* 2011, vol. 44, no. 5, pp. 815-823.

[6] J. F. Yeh, C. H. Wu, L. C. Yu, and Y. S. Lai, “Extended Probabilistic HAL with Close Temporal Association for Psychiatric Consultation Query Retrieval,” *ACM Trans. Inf. Syst.* 2008, vol. 27, no. 1, p. Article 4.

[7] L. C. Yu, C. H. Wu, J. F. Yeh, and F. L. Jang, “HAL-based Evolutionary Inference for Pattern Induction from Psychiatry Web Resources,” *IEEE Trans. Evol. Comput.* 2008, vol. 12, no. 2, pp. 160-170.

[8] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, “Beyond SumBasic: Task-focused Summarization with Sentence Simplification and Lexical Expansion,” *Inf. Process. Manage*. 2007, vol. 43, no. 6, pp. 1606-1618.

[9] C. H. Wu, L. C. Yu, and F. L. Jang, “Using Semantic Dependencies to Mine Depressive Symptoms from Consultation Records,” *IEEE Intell. Syst.* 2005, vol. 20, no. 6, pp. 50-58.

[10] M. Dong and L. Su, “Modeling a Configuration System of Product-Service System Based on Ontology Under Mass Customization,” *Adv. Sci. Lett.* 2011, vol. 4, no. 6-7, pp. 2256-2261.

[11] T. Merabti, L. F. Soualmia, J. Grosjean, O. Palombi, J. Müller, and S. J. Darmoni, “Translating the Foundational Model of Anatomy into French Using Knowledge-based and Lexical Method,” *BMC Med. Inform. Decis. Mak.* 2011, vol. 11, no. 65.

[12] C. C. Cheng, “Word-focused Extensive Reading with Guidance,” In: *Proc. of the 13th International Symposium on English Teaching*. pp. 24-32.

[13] C. H. Wu, C. H. Liu, H. Matthew, and L. C. Yu, “Sentence Correction Incorporating Relative Position and Parse Template Language Models,” *IEEE Trans. Audio Speech Lang. Process.* 2010, vol. 18, no. 6, pp. 1170-1181.

[14] J. Lee, S. Lee, H. Noh, K. Lee, and G. G. Lee, “Grammatical Error Simulation for Computer-Assisted Language Learning,” *Knowl-based. Syst.* 2011, vol. 24, no. 6, pp. 868-876.

[15] J. Bhogal, A. Macfarlane, and P. Smith, “A Review of Ontology Based Query Expansion,” *Inf. Process. Manage*. 2007, vol. 43, no. 4, pp. 866-886.

[16] L. C. Yu, C. H. Wu, and F. L. Jang, “Psychiatric Consultation Record Retrieval Using Scenario-based Representation and Multilevel Mixture Model,” *IEEE Trans. Inf. Technol. Biomed.* 2007, vol. 11, no. 4, pp. 415-427.

[17] L. C. Yu, C. H. Wu, and F. L. Jang, “Psychiatric Document Retrieval Using a Discourse-Aware Model,” *Artif. Intell.* 2009, vol. 173, no. 7-8, pp. 817-829.

[18] J. Zhai, C. Yuan, Y. Chen, and J. Li, “Knowledge Modeling and Semantic Retrieval of Product Data Based on Fuzzy Ontology and SPARQL,” *Adv. Sci. Lett.* 2011, vol. 4, no. 4-5, pp. 1855-1859.

[19] C. Crasto, D. Luo, F. Yu, A. Forero, and D. Chen, “GenDrux: A Biomedical Literature Search System to Identify Gene Expression-based Drug Sensitivity in Breast Cancer,” *BMC Med. Inform. Decis. Mak.* 2011, vol. 11, no.28.

[20] C. Burgess, K. Livesay, and K. Lund, “Explorations in Context Space: Words, Sentences, Discourse,” *Discl. Process.* 1998, vol. 25, no. 2-3, pp. 211-257.

[21] K. Church and P. Hanks, “Word Association Norms, Mutual Information and Lexicography,” *Comput. Linguist.* 1990, vol. 16, no. 1, pp. 22-29.

[22] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA, 1999.