Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 206-211

Extracting Recurrent Phrases and Terms from Texts Using a Purely Statistical Method

Zhao-Ming Gao* and Harold Somers**
Academia Sinica* and UMIST**

Most statistical measures for extracting interesting word pairs such as MI and t-score require a large corpus to work well. This paper evaluates some of the most widely used statistical measures and introduces a method that can identify significant bigrams in relatively small texts by adapting Fung and Church's (1994) K-vec algorithm, which was originally designed to extract word correspondences from unaligned parallel corpora. The proposed method captures the linguistic generalisation abou lexical patterning in texts and can identify recurrent co-occurring word sequences, which might be phrases, terms, or unknown words. In addition, it has the potential of identifying key phrases and terms that reveal topicality in a text.

1. Introduction

In recent years, there has been a growing interest in eliciting linguistic knowledge directly from corpora using statistical methods. Several quantitative measures have been proposed to identify significant lexical relations. These measures, however, are designed to work with large corpora with millions of words. Accordingly, they do not perform well in relatively small texts with a few thousand words. This paper presents a statistical method that is well-suited to extracting recurrent phrases and terms from relatively small texts. The method, a variant of Fung and Church's (1994) K-vec algorithm, is shown to be in line with linguistic generalisations about lexical cohesion in text structures.

2. Statistical Measures for Identifying Interesting Lexical Relations

Church and Hanks (1990) and Church et al. (1991) use mutual information (MI) to estimate associations between two words. Mutual information is defined as follows.

$$(1)\quad I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

MI compares the joint probability $P(x, y)$ (i.e. the probability of the co-occurrence of x and y) with $P(x)$ and $P(y)$, the independent probabilities of x and y (chance). If there is a strong association between word x and word y, then the joint probability $P(x, y)$ will be much larger than chance $P(x)P(y)$, and accordingly $I(x, y) > 0$. If no significant relation holds between x and y, $I(x, y)$ will approximate to zero. If x is in complementary distribution with y, $I(x, y)$ will be less than zero. Besides MI, Church and Hanks (1990) and Church et al. (1991) use t-score for testing the statistical significance of an co-occurrence. t-score can be approximated by (2).

$$(2).\quad t \approx \frac{f(x, y) - \frac{f(x)f(y)}{N}}{\sqrt{f(x, y)}}$$

where f(x), f(y), and f(x, y) are the number of occurrences of x, y, and x co-occurring with y, respectively; while N is the number of occurrences of all the tokens in the text.

*Chinese Knowledge Information Processing Group, Institute of Information Science, Academia Sinica, Nankang, Taipei 115, Taiwan.
E-mail: imgao@hp.iis.sinica.edu.tw

Several alternatives to MI and t-score have been proposed. These methods require the contingency table in (3).

(3).

a = k(A B)	b = k(~ A B)
c = k(A ~B)	d = k(~A ~B)

where *A* and *B* are the words in question, and *k* is the count of the bigrams. The ~ sign means not; so for example *c* is the count of the bigram where *A* is followed by a word other than *B*.

One of the alternatives to MI is the association measure IM, which is very similar to MI. IM is calculated as in (4) (cf. Daille (1996)).

(4). $IM = \log_2 \frac{a}{(a+b)(a+c)}$

In addition, Gale and Church (1991) introduce the Φ^2 coefficient using the formula in (5).

(5). $\Phi^2 = \frac{(a \times d - b \times c)^2}{(a+b)(a+c)(b+d)(c+d)}$

Dunning (1993) notes that MI is subject to overestimation when the counts are small and thus proposes using log likelihood ratio G^2 as a significance test for estimating surprise and coincidence of a rare event. G^2 is computed by the formula in (6).

(6). G^2 = a log a + b log b + c log c + d log d • (a+b)log(a+b) − (a+c)log(a+c) − (b+d)log(b+d) − (c+d)log(c+d) + (a+b+c+d)log(a+b+c+d)

We conducted experiments testing all the statistical measures described above with a Chinese text of 5155 words. The Chinese text was preprocessed by the Chinese word segmentation program reported in Chen and Liu (1992). The results of the tests are shown in Table 1. Bigrams with a t-score lower than 1.65 have been left out. As can be seen in Table 1, the performance of MI and t-score is not satisfactory, for many uninteresting bigrams containing pronouns and determiners are incorrectly extracted. It is obvious in Table 1 that Φ^2 and G^2 outperform MI and t-score. Nevertheless, Φ^2 gives a zero value for word pairs which always co-occur with each other, since b + d in (6) is zero if word pairs always co-occur. Therefore, bigrams consisting of proper nouns such as 小 虹 'Hsiao Hung', 賀 德 'Ho Te', 德 芬 'Te Fen' in Table 1 are given zero value, which is counterintuitive, because high values for rigid pairs are expected. Besides, Φ^2 and G^2 do not seem to be able to distinguish bigrams containing two content words from those containing one function word. For instance, G^2 gives a larger value to 一 位 *yi wei* 'one CLASSIFIER' than the more interesting proper names 小 虹 'Hsiao Hung' and 呂 安妮 'Lu Anni'. IM seems to outperform all the other statistical measures in small texts. By setting the threshold to –3, all the proper names together with some interesting terms such as 女性 主義 'feminism', 平 權 'equal right', 行政 人員 'administrative staff' can be extracted. However, IM has a serious defect: its threshold value is difficult to determine.

3. Modifying Fung and Church's (1994) K-vec Algorithm to Extract Recurrent Monolingual Terms

Fung and Church (1994) propose a simple algorithm to find word correspondences from unaligned parallel texts. The basic idea is that a true word pair should have similar distributions in terms of the position of its occurrence in the text. To estimate the similarity of co-occurrence, the parallel texts are split into the same number of segments (K) and the distributions of each word are represented in a 1...K binary vector. For instance, suppose the Chinese and English texts are divided into ten segments. Suppose further that the Chinese word 大學 *daxue* occurs ten times, with the first 3 occurrences in the fourth segment and the remaining 7 occurrences in the seventh segment and that the English word *university* appears twelve times, with the first 4 occurrences in the fourth segment and the remaining 8 occurrences in the

seventh segment. Using the K binary vectors, the distributions of both the Chinese and English words in question can be represented as <0,0,0,1,0,0,1,0,0,0>. Mutual information (MI) and t-score are then used to estimate the correlation of a proposed word correspondence. Mutual information and t-score are computed using the formulas in (8) and (9).

(7).

$$MI(V_C, V_e) = \log_2 \frac{P(V_C, V_e)}{P(V_C)P(V_e)}$$

$$P(V_C) = \frac{a+b}{K}$$

$$P(V_e) = \frac{a+c}{K}$$

$$P(V_C, V_e) = \frac{a}{K}$$

where a is the number of pieces of segments in which both the Chinese and the English word occur; b is the number of pieces of segment where only the Chinese word is found; c is the number of pieces of segment where only the English word is found.

(8). $$t(V_C, V_e) = \frac{P(V_C, V_e) - P(V_C)P(V_e)}{\sqrt{\frac{P(V_C, V_e)}{K}}}$$

The t-score in (8) is introduced to filter out word pairs with low frequency which happen to co-occur in the same segment by chance.[1] Fung and Church set the threshold value of MI to be 0 and t-score to be 1.65. Only word pairs with both MI and t-score higher than the predetermined threshold values and in the frequency range 3-10 are considered to be potential mutual translations.

The rationale behind the K-vec algorithm is that two words in parallel text associate strongly with each other if they co-occur more often than by chance in some text segments. The statistics of co-occurrence K-vec employs is actually grounded on a linguistic generalisation about lexical patterning in the text. Research by Halliday and Hasan (1976) and Hoey (1991) suggest that cohesion plays a very important role in the organisation of texts. They point out that the most straightforward form of cohesion is repetition. In addition, as each text has a topic, words or phrases closely related to the topic tend to recur in the text (cf. Salton and McGill (1983), Phillips (1985)).

K-vec can be easily applied to monolingual texts to identify recurrent noun phrases, collocations, or words not listed in the dictionary. The only necessary adaptation is that the source is the same as the target text. In addition, since sentences are the basic building blocks of a text, they are better units of a discourse segment than an ad hoc number of words as proposed by the original K-vec. As a result, a Word-Sentence Index (WSI) is required which records the position and the index of the sentence in which each word occurs. Based on WSI, adjacent word pairs that co-occur more often than by chance can be extracted.

Comparing Table 1 with Table 2 we can see that K-vec is better than MI, t-score, Φ^2, and G^2 in identifying collocations, recurrent proper names and phrases in a small text in terms of precision and recall. Like IM, K-vec can distinguish interesting bigrams from uninteresting ones. But unlike IM, the threshold value of K-value is predetermined (i.e. MI >=0 and t-score >= 1.65). K-value is thus more convenient than IM. In contrast with Smadja's (1993) Xtract, which was designed to extract collocations from large corpora, our proposed method is suitable for extracting recurrent rigid collocations in relatively small texts.

If two extracted bigrams are adjacent to each other, they are mostly likely to be phrases or proper nouns, as shown in Table 3. The proximity relation between two bigrams can be easily identified in the light of the WSI. It is interesting to note that many of the word pairs identified in Table 2 are key phrases that suggest topicality of the text, e.g. 呂安妮事件 'Lu Anni Incident', 女性主義 'feminism', 師生關係 'teacher-student relationship', 校務會議. 'Campus Affairs Committee Conference'.

[1]The approximation of t-score used by Fung and Church (1994) in (8) is slightly different from (2).

4. Conclusion

This paper reconfirms the importance of selecting an appropriate unit of text in lexical knowledge acquisition, as emphasized by Church et al. (1991). The proposed method, a simple variant of MI, t-score, and K-vec, has a higher precision than most current statistical algorithms in extracting recurrent word sequences from relatively small texts. The algorithm can be used to identify Chinese unknown words or key phrases in any language.

Acknowledgement

The first author would like to thank Prof. C.-R. Huang, Prof. K.-J. Chen, Dr. L.-F. Chien at Academia Sinica and anonymous PACLIC reviewers for their comments on an earlier draft of this paper.

References

Chen, K.-J. and Liu, S.-H. (1992) "Word Identification for Mandarin Chinese Sentences." In Proceedings of the International Conference on Computational Linguistics, pp. 101-107.

Church, K. and Hanks, P. (1990) "Word Association Norms, Mutual Information, and Lexicography." Computational Linguistics, Vol. 16, No. 1, pp. 22-29.

Church, K, W. Gale, P. Hanks, and D. Hindle. (1991) 'Using Statistics in Lexical Analysis,' in Zernik (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 115 - 164, Lawrence Erlbaum Associates Publishers.

Daille, B. (1996) "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology." In Klavans and Resnik (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, pp. 49-66.

Dunning, T. (1993) 'Accurate Methods for the Statistics of Surprise and Coincidences,' Computational Linguistics, Vol. 19, No. 1, pp. 61 - 74.

Fung, P. and Church, K. (1994) "K-vec: A New Approach for Aligning Parallel Texts." Proceedings of the International Conference of Computational Linguistics, pp. 1096-1102, Kyoto.

Gale, W. and Church, K. (1991) "Concordances for Parallel Texts." In Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, Using Corpora, pp. 40-62, Oxford.

Halliday, M. and Hasan, R. (1976) *Cohesion in English*. Longman Publishers.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.

Phillips, M. (1985) *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Elsevier Science Publishers.

Salton, G. and McGill, M. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill.

Smadja, F. (1993) 'Retrieving Collocations from Text: Xtract', Computational Linguistics, Vol. 19, No. 1, pp. 143 - 177.

Table 1. Output of Different Statistical Measures for Identifying Interesting Bigrams

C1	C2	MI	t-score	IM	Φ^2	G^2
一	位	6.23	3.95	-5.90	33.19	57.20
一	個	4.46	1.90	-7.67	1.45	8.03
也	有	3.33	1.80	-8.80	0.43	5.24
口試	教授	5.91	1.70	-6.22	4.01	9.21
大學	任教	7.18	1.72	-4.95	0.09	13.12
大學	校園	4.82	2.72	-7.31	3.85	18.24
女	學生	4.83	2.36	-7.30	1.25	16.04
女性	主義	9.81	1.73	-2.32	0.59	18.74
小	虹	9.14	2.82	-3.00	0.00	50.97
不	知道	6.41	1.97	-5.72	0.07	15.61
不	是	3.55	3.03	-8.59	1.01	16.49
不	喜歡	4.99	1.67	-7.14	1.67	7.14
不	願	6.41	1.71	-5.72	0.05	11.68
不少	教師	5.84	2.40	-6.29	6.13	19.05
之	一	4.97	1.67	-7.16	1.75	7.02
他	說	4.85	2.15	-7.29	2.97	11.22

平	權	9.81	1.99	-2.32	3609.00	25.72
各	種	7.91	1.72	-4.22	80.29	12.93
在	校園	3.36	2.21	-8.77	0.56	8.10
在	課堂	5.87	2.19	-6.26	4.86	16.37
安妮	事件	8.14	2.22	-4.00	175.80	23.29
成	了	7.18	1.72	-4.95	15.50	13.12
有	一	3.00	1.75	-9.13	0.29	4.48
有	些	6.58	2.96	-5.55	0.18	36.54
老	教授	6.64	2.61	-5.49	18.30	28.56
老師	的	1.19	1.68	-10.90	0.02	2.64
而	在	3.49	1.82	-8.64	0.47	5.65
行政	人員	9.33	1.72	-2.80	0.42	17.52
位	男	5.94	1.70	-6.19	7.22	8.77
呂	安妮	8.97	2.82	-3.16	0.88	48.24
男	老師	5.28	2.38	-6.85	3.65	16.06
男	教授	5.19	1.94	-6.95	2.91	9.99
系	上	6.11	1.70	-6.02	5.38	9.59
那	一	6.23	1.97	-5.90	6.58	13.80
並	不	5.15	2.17	-6.99	3.09	12.55
事實	上	6.85	1.98	-5.28	13.12	16.72
些	老師	5.31	2.17	-6.82	3.10	13.46
兩	年前	9.40	1.72	-2.73	676.60	16.79
官	俊	9.55	1.99	-2.58	0.66	24.57
東吳	大學	7.18	1.72	-4.95	15.50	13.12
林	錦	10.55	1.73	-1.58	0.00	21.67
的	主體	3.47	2.72	-8.67	0.34	15.50
的	事件	2.67	1.68	-9.46	0.07	4.33
的	師生	1.80	1.89	-10.30	0.04	4.00
的	課	3.00	1.75	-9.13	0.09	5.26
俊	榮	10.14	1.99	-2.00	0.00	27.89
指導	教授	6.64	1.980	-5.49	9.69	16.20
爲	了	5.60	1.69	-6.53	4.15	8.18
個	人	6.31	1.71	-5.82	11.19	9.51
師生	關係	6.40	3.42	-5.73	35.92	42.24
校	內	8.23	1.72	-3.90	112.60	13.77
校務	會議	8.09	3.30	-4.04	598.60	55.88
校園	中	6.01	3.11	-6.12	16.82	32.10
校園	的	1.45	1.67	-10.60	0.02	2.80
校園	裡	6.29	2.20	-5.84	10.27	17.10
珠	利	10.55	1.73	-1.58	0.00	21.67
高強	華	9.81	1.73	-2.32	1353.00	18.74
張	小	8.65	2.23	-3.48	402.30	25.67
教授	的	1.51	1.83	-10.60	0.03	3.42
現代	校園	5.36	1.69	-6.77	2.77	7.81
被	學生	2.80	1.71	-9.33	0.14	4.21
許多	老師	3.99	1.87	-8.14	0.78	6.88
許多	教師	4.67	2.35	-7.46	2.21	13.12
這	位	5.49	1.69	-6.64	4.51	7.86
這	種	7.01	1.71	-5.12	23.98	11.01
這樣	的	3.73	2.06	-8.40	0.22	10.04
賀	德	9.33	2.64	-2.80	0.00	45.41
越來越	多	8.23	1.72	-3.90	0.19	15.15
會議	上	5.85	2.60	-6.28	10.15	21.13

說	過去	5.06	1.68	-7.08	2.58	7.05
劉	珠	10.14	1.73	-2.00	0.74	19.71
德	芬	9.33	2.64	-2.80	0.00	45.41
課堂	上	6.59	2.21	-5.54	13.62	18.61
學生	代表	5.05	1.67	-7.08	0.02	9.23
學生	與	3.14	1.77	-8.99	0.20	5.04
整	個	7.63	1.72	-4.50	39.62	12.84
錦	川	10.55	1.73	-1.58	0.00	21.67

Table 2. Chinese Bigrams Extracted Using a Variant of K-vec

A	B	MI	t-score
女性	主義	9.76	1.73
小	虹	9.08	2.82
安妮	事件	8.08	2.23
呂	安妮	8.91	2.82
官	俊	9.49	2.00
林	錦	10.49	1.73
俊	榮	10.07	2.00
師生	關係	6.71	3.84
校務	會議	7.98	2.99
珠	利	10.49	1.73
高強	華	9.76	1.73
賀	德	9.27	2.45
劉	珠	10.07	1.73
德	芬	9.27	2.64
錦	川	10.49	1.73

Table 3. Proper Names Extracted On the Basis of Table 2 and Word-Sentence Index

呂	安妮	事件
賀	德	芬
劉	珠	利
官	俊	榮