Improving Collocation Extraction for High Frequency Words

David Wible[*], Chin-Hwa Kuo[**] and Nai-Lung Tsao[**]

[*]English Department and [**]Department of Computer Science and Information Engineering, Tamkang University
151 Ying-chuan Road Tamsui,Taipei County Taiwan 251, Republic of China
dwible@mail.tku.edu.tw

Abstract

The purpose of this paper is to introduce an alternative word association measure aimed at addressing the under-extraction collocations that contain high frequency words. While measures such as MI provide the important contribution of filtering out sheer high frequency of words in the detection of collocations in large corpora, one side effect of this filtering is that it becomes correspondingly difficult for such measures to detect true collocations involving high frequency words. As an alternative, we propose normalizing the MI measure by dividing the frequency of a candidate lexeme by the number of senses of that lexeme. We premise this alternative approach on the one sense per collocation assumption of Yarowsky (1992; 1995). Ten verb-noun collocations involving three high frequency verbs (*make, take, run*) are used to compare the extraction results of traditional MI and the proposed normalized MI. Results show the ranking of these high-frequency verbs as candidate collocates with the target focal nouns is raised by normalizing MI as proposed. Side effects of these improved rankings are discussed, such as increase in false positives resulting from higher recall. It is found that overall rank precision remains quite stable even with the increased recall of normalized MI.

Introduction

Computational lexicography has contributed fundamental improvements to the detection of collocations from large corpora by exploiting statistical word association measures, especially since the contributions of Church et al. (Church & Hanks 1989; Church, Hanks, Hindle & Gale 1991). One of the widely acknowledged reasons that word association measures such as mutual information (MI) and hypothesis testing methods provide an advantage over simple frequency counts is that sheer high frequency of cooccurence of two expressions does not necessarily constitute an interesting word association. Such high frequency of cooccurence may simply be the result of the raw high frequency of one or the other of the participating words in the expression (or both), and these word association measures are able to essentially filter out this noise. A strong word association score or, say, collocability score is achieved when the frequency of cooccurence is not merely high, but significantly higher than the rate that would be expected by chance given the frequency of the individual expressions in the corpus relative to the size of the corpus.

Despite their advantages, however, one persistent limitation of such word association measures is that, partially as a consequence of reducing the influence of the raw frequency of an expression in measuring collocability, these measures make it relatively hard for very high frequency lexemes to register high word association scores with other words. In this respect, these measures work almost too well. More specifically, while high frequency words threaten to introduce noise into attempts to extract collocations, still, there are true collocations which contain high frequency words, and these collocations are often undetectable by word association measures. The challenge is how to find a measure that registers these as collocations while still avoiding the true noise posed by raw high frequency lexemes. We propose a novel measure intended to help achieve this.

An Alternative Approach

To reiterate, the problem we seek to address is the under-extraction of high frequency words in the automatic extraction of collocations from large corpora. Our approach takes into account the semantics of the lexemes. Specifically, the high frequency words at the center of this problem are generally also highly polysemous. According to WordNet 1.6, the verb run, for example, has 42 senses, take has 41 senses, and make has 48 senses. Our approach is premised on the 'one sense per collocation' assumption of Yarowsky (1992; 1995). We exploit this assumption as follows. While a high frequency word is generally highly polysemous, only one of its many senses is used in a particular collocation. For example, while *run* has over forty senses, only one of these is relevant whenever *run* appears in the collocation '*run a risk*'. On the basis of this assumption, we find a plausible way to reduce the frequency count of high frequency lexemes and render them susceptible to collocation extraction with the more traditional word association measures. Specifically, we normalize the frequency count of these words according to the total number of senses of each of them. That is, we calculate the MI not according to the raw frequency and the size of the corpus but according to the frequency normalized according to the number of senses WordNet attributes to the word. There are a variety of ways this approach could be implemented. Due to limitations of space, we choose the most straightforward implementation for comparison of its results with the results obtained from traditional 'non-normalized' MI.

The formulation of so-called traditional MI, compares the probability of observing word x and word y together (the joint probability) with the probabilities of observing x and y independently (chance), which we take for our benchmark is as follows:

$$tradMI(x, y) = \log_2 \frac{P(x, y)}{P(x)\,P(y)}$$

In contrast, for the normalized MI that we propose, rather than using raw frequency scores to calculate the MI score, we divide the raw frequency of each lexeme by the number of senses of the lexeme, using WordNet to determine the latter. This gives the mean frequency per sense, and it is this ‘normalized’ frequency which is used for the frequency value in calculating the MI score. We formulate this as follows.

$$normMI(x, y) = \log_2 \frac{P(x, y)}{\left(P(x) \big/ sn(x) \right) \cdot \left(P(y) \big/ sn(y) \right)}$$

,where *sn* means sense number of the specific word. To compare the results we choose ten verb-noun collocations involving three high frequency verbs (*make*, *take*, and *run*), listed in Table 1.

Verb	Noun
make	change
make	decision
make	effort
run	business
run	risk
take	bath
take	effort
take	medicine
take	risk
take	time

Table 1: Verb-noun pairs

In these verb-noun collocations, it is the noun that serves as the focal word and the verb as the collocate (See Manning and Schutze (1999) on the distinction between focal word and collocate). Thus, our approach is to use the focal noun and seek to extract all collocating verbs preceding it, using the two formulations of the MI measure (traditional and normalized) to do this. The results for each focal noun, then, consist of a ranked list candidate collocate verbs for that noun. The focus of comparison of our proposed normalized MI to the traditional MI is to see whether the two measures result in different rankings of collocating verbs for the same noun (a example of focal noun *time* shown in Table 2). The most extreme contrast between the two measures would be cases where a high frequency verb does not even appear on the list of candidate collocate verbs extracted by one of the measures but appears highly ranked by the other measure. The less extreme difference would be simply a difference in the candidate collocate verbs generated by the two measures, with the high frequency verb being ranked higher by one measure than the other.

Trad MI	Norm MI
bide(1)	waste(10)
waste(10)	bide(1)
idle(2)	take(41)
spend(3)	give(45)
magnify(3)	idle(2)
clock(1)	spend(3)
devote(2)	magnify(3)
flower(1)	break(63)
spare(4)	save(10)
multiply(4)	fall(32)
repeat(6)	pass(25)
sow(3)	mark(15)
date(5)	repeat(6)
postpone(1)	play(29)
save(10)	run(42)
coincide(3)	get(37)
invest(5)	occupy(8)
occupy(8)	date(5)
wake(2)	clear(24)
dive(3)	spare(4)
	reduce(19)
	make(48)
	multiply(4)
	devote(2)
	come(23)
	hold(36)
	meet(14)
	beat(21)
	go(30)

Table 2: The sorted collocate verb candidates of focal noun *time.* The number after each word means WordNet sense number.

Results and Discussion

First we give a simple comparison of the rankings of the targeted high frequency collocate verbs for the two measures, traditional and normalized MI. The numbers on the Table 3 indicates where the verb ranks on the list of candidate collocate verbs generated by that particular MI measure. This is shown in Table 3.

	Trad MI ranking of the verb	Norm MI ranking of the verb
make (change)	--	2
make (decision)	9	1
make (effort)	6	2
run (business)	6	1
run (risk)	9	1
take (bath)	8	3
take (effort)	--	5
take (medicine)	4	2
take (risk)	12	5
take (time)	--	3

Table 3

A cursory look at Table 3 shows that in each case the high frequency verb is ranked higher by the normalized MI than by traditional MI, essentially the main effect we were hoping for. In fact in three cases, the normalized MI extracts a verb that the traditional MI leaves completely undetected (*take time, take effort* and *make (a) change*). The success of a word association measure, however, cannot be judged simply on whether it improves the ranking of a narrowly specified type of word (in our case, high frequency collocates). It is worth considering other effects of the measure, and to do this we look more closely at the rankings of candidate collocates provided for particular focal words listed in Table 1. Some basic questions to consider are whether the normalized MI wreaked havoc on the other aspects of extraction and ranking of candidate collocates. That is, did it introduce false positives as well, and did it lose true positives that the traditional MI originally succeeded in detecting?

First, it is worth noting that the normalized MI creates consistently higher recall than traditional MI. The average number of candidates extracted for traditional MI over the ten collocations tested here is 17, whereas the average number extracted by the normalized MI is 29. Higher recall does not entail precision. Recall can introduce unwanted candidates. It appears, however, that other than the candidates we would want to pull up in the rankings, the other additions to the candidate list introduced by normalized MI essentially appear near the bottom of the list, generally leaving rank order precision relatively unaffected. There are notable exceptions. For example, the focal noun *change,* in addition to the welcome effect of ranking the verb *make* as second among candidate collocates though it did not appear at all on the candidate list of traditional MI, also raises the verb *mark* to fourth on the candidate list from its rank as 21st on the traditional MI candidate list. One reason is that the verb *mark*, while not an extremely frequent verb, is surprisingly polysemous, with 15 senses. Similarly, for the focal noun *time*(shown in Table 2), along with the impressive result that the verb *take* is ranked third by the normalized MI even though it is not detected at all by traditional MI, the verb *give*, which traditional MI rightly skips over in its list of 20 candidate collocates, is ranked 4th by normalized MI. The reason for this effect on *give* becomes clear once we note that the verb has 45 senses.

This points to a limitation of the approach of normalizing the MI measure by number of senses: normalized MI raises the tide for all highly polysemous verbs that co-occur with the target focal noun, whether or not that highly polysemous verb constitutes a collocation with that noun. Hence, the same tide that happily raises the otherwise undetected *take time* to 3rd place on the collocate list also unfortunately raises *give time* right behind it to 4th place.

Conclusion

The preliminary results suggest that normalizing the MI measure according to the number of senses is a worthwhile direction to pursue for improving the extraction of high frequency words in collocation extraction from large corpora. Some surprisingly dramatic differences in the ranking of high frequency verbs demonstrate that normalizing MI in this way can detect collocate verbs which fly below the radar of traditional MI. The limitation of the approach lies in the fact that it improves the ranking of highly polysemous words regardless of their collocability with the target focal noun. It appears that this effect is surprisingly mild, however, in the upper end of the candidate lists. A more thorough investigation of a wider range of collocations and high frequency words would help to assess the overall potential of the proposed approach.

References

Church, K. and Hanks, P.(1989). Word Association Norms, Mutual Information, and Lexicography. In Proceedings of the 27rd Annual Meeting of the Association for Computational Linguistics (pp. 76-83).

Church, K., Hanks, P., Hindle, D., Gale, W. (1991). Using Statistics in Lexical Analysis. In Zernik (ed), Lexical Acquisition: Using On-line Resources to Build a Lexicon (pp. 115-164). Lawrence Erlbaum.

Manning, Christopher and Schutze, Hinrich (1999) Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

Yarowsky, David (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In Proceedings of the 15th International Conference on Computational Linguistics (pp. 454-460).

Yarowsky, David (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In

Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (pp. 189-196).