Measures of collocational strength and flexibility for the identification of MWEs.

Patrick Hanks
RIILP, University of Wolverhampton
WV1 1NA, England
Patrick.W.Hanks@gmail.com

Ismail El-Maarouf
RIILP, University of Wolverhampton
WV1 1NA, England
I.El-Maarouf@wlv.ac.uk

Michael Oakes
RIILP, University of Wolverhampton
WV1 1NA, England
Michael.Oakes@wlv.ac.uk

Abstract

This poster describes our most recent work in the automatic detection and characterisation of Multi-Word Expressions (MWE). A large number of statistical measures designed to measure collocational strength have been used in the past to detect MWEs – stronger collocates are more likely to be meaningful MWEs. We also look at the flexibility of collocations, which is a function of how many alternative surface forms they can take, using the standard deviation of the length of the span of words between them, and Shannon's diversity index. A range of measures of collocational strength as well as the two measures of collocational flexibility were used as features for a Support Vector Machine, designed to act as a combined classifier which works better than any of the individual features in isolation.

1 Introduction

This poster describes our most recent work in the automatic detection and characterisation of Multi-Word Expressions (MWE) using statistical information from a subset of the BNC corpus consisting of 2 million words. In the past, a large number of word association measures have been derived to estimate the collocational strength of MWEs, which often appear as idiomatic expressions. A tool for measuring word associations, using the mutual information statistic, was first described by Church and Hanks (1989), and subsequently elaborated by Kilgarriff et al. (2004). A survey of word association measures is given by Pecina (2006).

Smadja (1993) recommends that collocations should not only be measured by their strength, such as by using the z-score, but also by their flexibility. This can be done by finding the mean of the relative distances between two words, and the *spread* of each collocation, which is the standard deviation of the relative distances between the two words. High spread would indicate a flexible or semantic, rather than a rigid, lexical collocation. In a study of David Wyllie's English translation of Kafka's *Metamorphosis*, Oakes (2012) found that *stuck fast* and *office assistant* had mean inter-word distances of 1 with a standard deviation of 0. This showed that in this particular text, they were completely fixed collocations where the first word was always immediately followed by the second. Conversely, *collection* and *samples* had a mean distance of 2.5 with a standard deviation of 0.25. This collocation was a little more flexible, occurring both as *collection of samples* and *collection of textile samples*. *Mr. Samsa* had a mean distance of 1.17 and a standard deviation of 0.32. This is because it usually appeared as *Mr. Samsa* with no intervening words, but sometimes as *Mr. and Mrs. Samsa*.

Another way of looking at the flexibility of a collocation is by measuring the diversity of surface forms found for that collocation. A rigid collocation, where all found examples are identical in form and length, has very low diversity, while a collocation which has many surface forms has much higher diversity. One measure of diversity, popular in ecological studies, is Shannon's diversity index, which is equivalent to entropy in information theory, and given by the formula:

$$E = -\sum_{i=1}^{N} p_{i\ log_2} p_i$$

E is entropy, N is the number of different surface forms found for the collocation, i refers to each surface form in turn, and p_i is the proportion of all surface forms made up of the surface form currently under consideration. The choice of logarithms to the base 2 ensures that the units of diversity are bits. The minimum value of diversity (when all the examples of a phrase or idiom are identical) is 0, while the maximum value (when all the examples occur in different forms) is the logarithm to the base 2 of the number of examples found. For standard deviation, the minimum value when all the examples are identical length is 0, and there is no theoretical upper limit.

2 Results for idioms containing "bite"

In this section we describe corpus-driven results for idioms containing the verb "bite". The phrase "bite the bullet", as in "I was going to bite the bullet and buy a real computer", occurs 9 times in our corpus. Thus the mean length was 3, the standard deviation of the length was 0 (indicating a maximally rigid collocation), and Shannon's index of diversity was also 0 (its minimum possible value, also indicating a very rigid collocation). Another example of a rigid collocation was "bite back", which occurred 3 times in exactly that form, so both the standard deviation and the Shannon index of diversity were 0. The phrase "bite … head(s) off", as in "If I ask my boss for a raise he'll sack me or bite my head off" was found 6 times in the corpus, twice in the form "bite their heads off", twice in the form "his head bitten off", once as "bite my head off" and once as "biting your head off". The lengths of these variants was 4 in every case, so the standard deviation of their lengths was 0, indicating maximal uniformity of length. The Shannon index of variety was 1.057, indicating an intermediate degree of diversity of surface forms for this idiom. Two examples of highly flexible phrases were found. One was "bitten by the … bug" (as in "But he had been bitten by the newspaper bug"), which also occurred as "bitten by the flying bug", "bitten by the London bug" and the structurally contrasting "bitten by the bug of the ocean floor". Due to the greater length of this last example, the standard deviation of the lengths was 1.5 and since the surface forms were all different, the Shannon index of diversity took its maximum value for four examples, namely $\log_2(4) = 2$. The other flexible phrase appeared in various forms of "bite the hand that feeds you", which took that form only in the example "It is hard to bite the hand that feeds you". In the other five examples the forms were "bite the hand that holds them", "bite the hand that feeds", "bite the hand that feeds them", "bite the hand that strokes them" and "bite the hand that strokes it". Although the lengths of these forms were similar, yielding a standard deviation of the lengths of just 0.408, the fact that all the forms were different produced its maximum value for a set of 6 examples, $\log_2(6) = 2.585$.

3 Conclusion

To complement the variety of measures which measure the strength of word association, we propose two measures of the flexibility of MWEs, standard deviation of MWE length and Shannon's diversity index. Both types of measures were used as features in a machine learning classifier which can discriminate between word sequences which are MWEs and those which are not.

References

Kenneth W. Church and Patrick Hanks. Word Association Norms, Mutual Information and Lexicography. 1989. *27th ACL:* 78-83.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz and David Tugwell. The Sketch Engine. 2004 *Proc. Euralex.* Lorient, France, July: 105-116.

Michael P. Oakes.. Describing a Translational Corpus. In: Oakes, M. P. and Ji, M., Quantitative Methods in Corpus-Based Translation Studies, Amsterdam: John Benjamins, 2012: 115-148.

P. Pecina. Lexical Association Measures: Collocation Extraction. PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic. 2008.

Frank Smadja. Retrieving collocations from text: Xtract. 1993. Computational Linguistics 19: 143-177.

Wikipedia. Diversity Index. http://en.wikipedia.org/wiki/Diversity_index