社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－167 (2)
2005－SLP－ 56 (2)
2005／5／26

Automatic extraction of fixed multiword expressions

Campbell HORE Masayuki ASAHARA Yūji MATSUMOTO
Graduate School of Infomation Science, Nara Institute Science and Technology
{campbe-h, masayu-a, matsu}@is.naist.jp

Abstract : Fixed multiword expressions are strings of words which together behave like a single word. This research establishes a method for the automatic extraction of such expressions. Our method involves three stages. In the first, statistical measures are used to extract candidate bigrams. In the second, we use this list to select occurrences of candidate expressions in a corpus, together with their surrounding contexts. These examples are used as training data for supervised machine learning, resulting in a classifier which can identify true multiword expressions. The final stage is the estimation of the parts of speech of the extracted expressions. Evaluation demonstrated that collocation measures alone are not effective in identifying target expressions. However, when trained on one million examples, the classifier identified true multiword expressions with precision greater than 90%. Part of speech estimation had precision and recall of over 95% for the part of speech types measured.

Keywords : multiword expressions, automatic extraction, support vector machines, part of speech

1 Introduction

1.1 Multiword Expressions

For natural language processing purposes, a naive definition of a word in English is '*a sequence of letters delimited by spaces*'. By this definition, the expression *ad hoc*, which originally came from Latin, consists of two "words", *ad* and *hoc*. However, in isolation *hoc* is not a meaningful English word. It is always preceded by *ad*. This suggests that treating these two words as if they together form a single "word with spaces" more closely models their behaviour in text. A sequence of words which for one reason or another is more sensibly treated as a single lexical item, rather than as individual words, is known as a multiword expression (MWE). In other words, an MWE is a sequence of words which together behave as though they were a single word.

MWEs are not limited to imported foreign phrases such as *ad hoc*. They cover a large range of expression types including proper nouns such as *New York*, verb-particle constructions such as *to call up* (i.e. to telephone someone), and light verbs such as *to make a mistake*. The justification for treating such expressions as MWEs is that their linguistic properties are odd in some way as compared to "normal" expressions: either their part of speech or their meaning is unpredictable based on an analysis of their constituent words.

1.2 Fixed Multiword Expressions

By *fixed*, we mean that this particular type of MWE consists of a contiguous sequence of words. Other MWE types can consist of discontinuous word sequences. For example, the verb-particle construction *to call up* takes an indirect object, the person who receives the telephone call. This person can appear after the verb-particle construction (e.g. "I called up *Mohammad*") but it can also appear in the *middle* of the verb-particle construction, (e.g. "I called *Mohammad* up"). In contrast, fixed MWEs consist of contiguous word sequences. For example, *by and large* cannot be modified by insertion of other words (e.g. *by and *very* large).

1.3 Multiword Expressions in Parsing

The aim of our research is the development of a method for the automatic extraction and part of speech estimation of fixed MWEs. The ability to identify this type of MWE in texts is of potential use in a wide variety of natural language processing tasks because it should enable an improvement in the precision of sentence parsing. Sentence parsing is frequently the first step in more sophisticated language processing tasks, so an increase in parsing precision should improve results in a large number of natural language processing applications.

A parser generally takes a sentence as input, together with the parts of speech of the tokens[1] in the

[1]We use the term *tokens* here rather than *words* because texts actually contain many space delimited character

sentence. The parser then attempts to estimate the most probable syntactic structure for the sentence. Some kinds of MWE have the potential to disrupt this process because the part of speech of the MWE as a whole, cannot be predicted on the basis of the parts of speech of its constituent words. For example, the part of speech sequence for the expression *by and large* is Preposition + Conjunction + Adjective, which is a sequence almost certainly unique to this expression. A parser which is not explicitly informed about *by and large* will therefore struggle to cope with this part of speech sequence, and may incorrectly try to group one or more parts of the expression with words in the surrounding context.

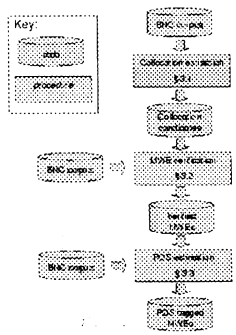
The solution to this problem of syntactically unpredictable MWEs is to add them to the dictionary used by the parser. When the parser comes across a sequence of words that matches an MWE in its dictionary, it can use the MWE's part of speech to parse the sentence as if the MWE were a single lexical item.

In reality, some sequences of words are an MWE only in specific contexts. For example, *in the main* is an MWE when it means "overall" or "mostly" as seen in the sentence "*In the main, the biggest improvements have been in child health*". In contrast, in a sentence such as "*Village hotels ought to be in the main square, not at the outskirts of a village*" the word sequence *in the main* is not an MWE, and can therefore be treated normally by the parser as separate words.

1.4 Target Problems

We approach the task of extracting fixed MWEs by decomposing it into three sub-problems, as illustrated in Figure 1. The first (§3.1, described in section 3.1) is simple collocation extraction. We use a standard collocation measure to extract as many candidate MWEs from the corpus as possible. The second problem (§3.2) is refinement of the list of candidate MWEs. Many of the candidates are not target multiword expressions. Distinguishing between word sequences that are MWEs, and those that never are, represents one sub-task. Hereafter, we refer to word sequences which are never MWEs as *non-MWEs*. Some word sequences have dual identities. In one context a word sequence may be an MWE, but in another, it may be just a normal, literal word sequence. For example, a child forced to help wash the family car, and acting petulantly, might be scolded "Don't kick the bucket!". In this case *kick the bucket* is a normal, compositional phrase; its meaning can be understood based on the literal meaning of its constituent words. When *kick the bucket* is used as a euphemism for "to die" however, its meaning is non-compositional, and thus in this context it *is* an MWE. In this paper we refer to occurrences of literal word sequences which have the *appearance* of being an MWE as *pseudo-MWEs*. We deal with the two sub-tasks simultaneously, using supervised machine learning. The third problem we tackle (§3.3) is estimation of the part of speech of MWEs. This problem is also solved using supervised machine learning. In this research we limit ourselves to MWEs containing only two words (i.e. bigrams). In the future, we plan to generalize the method so that it works with MWEs of arbitrary length.

Figure 1: Flowchart of Processing



strings which are not normally thought of as words, such as numbers and punctuation. In addition, some purely alphabetic character strings are not words in their own right, as is the case for *hoc*, mentioned above.

2 Related Work

Collocation extraction has been covered extensively in the literature. One of the earliest attempts to automatically extract collocations from a corpus was Church and Hanks work [2]. The statistical measure they used to identify collocations was based on mutual information. Smadja (see [7]) developed a tool called Xtract for the extraction of collocations. His definition of a collocation differed slightly from that of Church and Hanks because he claimed expressions such as *doctors and nurses* are not real collocations, just words related by virtue of their shared domain or semantics. Thanopoulos, Fakotakis and Kokkinakis in [8] reviewed the statisti-

cal measures most frequently used for collocation extraction, and evaluated them, comparing them with two new measures of their own. Their first novel measure, Mutual Dependency (MD) is pointwise mutual information minus self-information. The second measure attempts to introduce a slight frequency bias by combining the t-score with mutual dependency. Although frequency alone is not sufficient evidence of collocational status, they argue that candidate collocations that have a high frequency are more likely to be valid than those that are very rare.

While collocations have received attention over a number of years, MWEs have only relatively recently emerged as a research topic within natural language processing. In consequence, there are few articles specifically about MWEs. Sag et al. in [6] gave a linguistic categorisation of the different types of MWE, and described ways of representing them efficiently within a computationally tractable framework[2]. Although MWEs as a whole have yet to receive widespread investigation, attention has been paid to specific types of MWE. For example, verb-particle constructions have been the subject of several studies (for example, see [1] and [9]).

3 Method

In order to extract information about fixed MWEs from a corpus, we use a three stage process. In the first stage we identify a list of candidate MWEs based on the statistical behaviour of the tokens in the corpus. Two words whose probability of appearing together is greater than what would be expected based on their individual frequencies, are considered to constitute a potential MWE and are extracted for later processing. In other words, stage one is collocation extraction.

In the second stage, we use this list of candidate MWEs as the basis for extracting from the corpus examples of candidates together with their contexts. These examples are then used as training data for supervised machine learning resulting in a classifier capable of distinguishing between, on the one hand, true MWEs, and on the other, non-MWEs and pseudo-MWEs.

In the final stage we use supervised machine learning to train a classifier to perform MWE part of speech assignment. By examining the context surrounding an MWE, it is possible for the classifier to determine the most likely part of speech for the MWE.

[2]Head-driven Phrase Structure Grammar (HPSG)

3.1 Collocation Extraction

Collocation extraction was performed using each of the statistical measures discussed in [8]. These are: frequency, χ-square [5], log-likelihood ratio [4], t-score [5], mutual information (MI) [2], mutual dependency (MD – mutual information minus self-information) [8], and log-frequency biased mutual dependency (LFMD – a combination of the t-score and mutual dependency) [8]. The equations of these measures are shown in Figure 2.

We compared the resulting ranked lists of bigrams with a list of target MWEs extracted from the British National Corpus (BNC)[3]. The target list was produced by starting with a list of all MWEs tagged as such in the BNC, and removing MWEs with a frequency of less than five, and MWEs with a part of speech of noun, or adjective. This reduction was performed for two reasons. Firstly, in the candidate bigrams list there were many low frequency bigrams, many of which represented noise such as spelling variants, and features of spoken language. Secondly, a collocation consisting entirely of a combination of one or more nouns and adjectives is almost certainly a noun phrase, or part of a noun phrase. Noun phrases tend to be easily identifiable as such by parsers, and so are not relevant to this research. By removing the above MWEs we were able to reduce computational costs in the later stages of processing.

3.2 Verification

Verification was performed with the aim of extracting a much higher quality list of candidate MWEs from the list of candidate MWEs produced in the collocation stage.

Features

In order to train a classifier, a decision must be made about which features to include in the training data. We decided to use the tokens and their part of speech from a context window of three tokens to the left and right of each candidate MWE. We also used the words in the candidate MWE and their parts of speech. The cutoff value of three is somewhat arbitrary, but most lexical dependencies can be assumed to be relatively local. A cutoff of three can therefore be assumed to be large enough to capture the most useful information available from the context.

Contexts were not allowed to cross sentence boundaries. In cases where the available context

[3]http://www.natcorp.ox.ac.uk/

Figure 2: Collocation Measures

t-score

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Where $\overline{x}$ is the sample mean, s^2 is the sample variance, N is the sample size and μ is the distribution's mean.

χ-square

$$\chi^2 = \frac{(f_{w_1 w_2} - f_{w_1} f_{w_2})^2}{f_{w_1} f_{w_2}} + \frac{(f_{w_1 \overline{w_2}} - f_{w_1} f_{\overline{w_2}})^2}{f_{w_1} f_{\overline{w_2}}}$$

$$+ \frac{(f_{\overline{w_1} w_2} - f_{\overline{w_1}} f_{w_2})^2}{f_{\overline{w_1}} f_{w_2}} + \frac{(f_{\overline{w_1 w_2}} - f_{\overline{w_1}} f_{\overline{w_2}})^2}{f_{\overline{w_1}} f_{\overline{w_2}}}$$

Where f is the frequency of an event, $w_1 w_2$ is the sequence of events (in this case words) w_1 then w_2, and $\overline{w_1}$ is the negation of the event w_1.

log-likelihood ratio

$$-2 \log \lambda = 2 \cdot \log \frac{L(H_1)}{L(H_0)}$$

Where $L(H)$ is the likelihood of hypothesis H assuming a binomial distribution.

pointwise mutual information (PMI)

$$I(w_1, w_2) = log_2 \frac{P(w_1 w_2)}{P(w_1) \cdot P(w_2)}$$

Where $P(w)$ is the probability of a given word.

mutual dependency (MD)

$$D(w_1, w_2) = log_2 \frac{P^2(w_1 w_2)}{P(w_1) \cdot P(w_2)}$$

Where $P(w)$ is the probability of a given word.

log-frequency biased mutual dependency (LFMD)

$$D_{LF}(w_1, w_2) = D(w_1, w_2) + log_2 P(w_1 w_2)$$

Where $P(w)$ is the probability of a given word.

surrounding a candidate MWE was shorter than the three token window, we inserted the appropriate number of dummy tokens and dummy parts of speech to fill up the deficit: "BOS" (Beginning Of Sentence) tokens in the left context, and "EOS" (End Of Sentence) tokens in the right context.

Part of Speech Tagging

The part of speech information was provided by tagging the corpus using the part of speech tagger TnT[4]. The BNC is a part of speech tagged corpus, but retagging was necessary because although MWEs in the BNC are tagged with parts of speech, their constituent tokens are not. The tokens which make up each MWE must therefore be tagged with individual parts of speech. It might be argued that combining the original BNC tagging with the TnT tagging of the words in the MWEs would have produced more accurate training data, but in a real world application, the part of speech information in the classifier's input data will be produced entirely using a tagger such as TnT. By using the same part of speech tagger at both the training, and application stages, any systematic tagging mistakes will hopefully (at least in part) be learnt and compensated for by the classifier. The tagger was trained on a subset of BNC files containing just under 5.5 million tokens. It was then tested on a different set of files, containing approximately 5.3 million tokens. The tagger's precision when tested on this data was 94.7%.

Training

The corpus used for training the classifier was a sub-corpus of the BNC containing approximately ninety million tokens covering all domains in the corpus. Examples used for training were the occurrences in the training corpus of the top 10,000 bigrams identified using the t-score and LFMD collocation measures. Most of the bigrams in the corpus were negative examples, either non-MWEs, or pseudo-MWEs.

We used TinySVM[5] to create a binary classifier. Training was performed (using the software's default settings) on the training corpus. Several training runs were performed using different amounts of data in order to investigate the relationship between volume of training data and the resulting model's performance.

[4] http://www.coli.uni-sb.de/~thorsten/tnt/

[5] http://chasen.org/~taku/software/TinySVM/

Testing

Another sub-corpus of the BNC, independent of that used in training, was used for testing the classifier. This testing corpus contained approximately six million tokens and included texts from each domain in the corpus.

3.3 Part of Speech Estimation of Multiword Expressions

We treated the estimation of the overall part of speech of a given MWE as a classification task using the same approach as we used for the classification of true and false MWEs. We trained a separate classifier for each target part of speech. A positive training example was an occurrence of an MWE with the target part of speech. A negative example was an occurrence of any MWE with a non-target part of speech. The features used were the token and part of speech of three tokens to the left and right of the target MWE, as well as the tokens and parts of speech of the words in the MWE itself. The training and testing corpora were the same as used at the verification stage described above (section 3.2). This was acceptable because the two tasks are independent of each other.

We chose the target parts of speech (adverbs, prepositions and conjunctions) because these relatively closed class, high frequency types are expected to be most useful in applications like parsing. We also experimented with an open class type (nouns) for comparison. Verbs could not be tested because there were insufficient numbers of them in the testing corpus. There are few fixed MWE verbs, so a scarcity of data was not surprising.

4 Results and Discussion

4.1 Collocation Extraction

Results for collocation extraction (Table 1) show that standard collocation measures perform poorly in the task of extracting the target MWEs. Even when calculated based on the top 10,000 candidate collocations, recall is only 26% (using the t-score).

A limitation in our approach to measuring collocation extraction may be partly to blame for the poor results in this task. Our target list consisted of *all* target MWEs, irrespective of their length. Since the collocations extracted were limited to bigrams, some of these may in fact be only *part* of a larger MWE in our target list. A fairer measure might therefore have been a comparison with a list containing two word MWEs only.

Table 1: Precision and recall for top 100, 1,000 and 10,000 candidate multiword expressions extracted using different collocation measures

Measure	Cutoff	Prec.	Rec.	F
freq	10,000	0.009	0.251	0.017
	1,000	0.032	0.091	0.047
	100	0.010	0.003	0.004
t-score	10,000	0.009	**0.257**	0.017
	1,000	0.037	0.106	0.055
	100	0.030	0.009	0.013
χ^2	10,000	0.006	0.169	0.011
	1,000	0.013	0.037	0.019
	100	0.020	0.006	0.009
log-like	10,000	0.004	0.117	0.008
	1,000	0.008	0.023	0.012
	100	0.000	0.000	0.000
MI	10,000	0.003	0.083	0.006
	1,000	0.002	0.006	0.003
	100	0.000	0.000	0.000
MD	10,000	0.003	0.091	0.006
	1,000	0.003	0.009	0.004
	100	0.000	0.000	0.000
LFMD	10,000	0.008	**0.229**	0.015
	1,000	0.017	0.049	0.025
	100	0.080	0.023	0.036

Nevertheless, it may be that collocation measures are relatively ineffective at extracting fixed MWEs. Collocation measures are most effective when applied to expressions such as noun compounds. Many of the target MWEs contain high frequency function words such as prepositions, and thus are atypical of the types of expressions for which collocation measures were originally developed.

4.2 Verification

Verification of candidate MWEs produced better results (Table 2). For example, a classifier trained using one million examples, had precision of 96.56%, and recall of 89.11% giving an F-measure of 92.69.

Initial review of classification results suggests a number of sources of error. Tagging errors seem to cause many of the false negative results. Proper nouns tend to be tagged as "unclassified words" which is intelligent in as much as it bundles all unusual words the tagger is unsure of together, but it results in incorrect tagging which prevents the

Table 2: Performance of verifier using models trained on different quantities of data

Measure	Examples	Prec.	Rec.	F
t-score	1,000,000	91.52	89.11	90.30
	100,000	86.90	79.87	83.24
	10,000	76.03	56.04	64.52
	1,000	79.62	11.56	20.19
LFMD	1,000,000	96.56	89.11	92.69
	100,000	96.35	79.87	87.34
	10,000	93.66	56.04	70.12
	1,000	92.83	11.56	20.56

classifier identifying true MWEs. Similarly, capitalisation of words in titles results in incorrect tagging of ordinary words as proper nouns. One title in particular *Sport in Short* occurs multiple times in the corpus, resulting in numerous errors.

False positives seem to be caused by proper nouns (e.g. *Kuala Lumpur*) and foreign words (e.g. *Vive L'Empereur*). Both false positives and false negatives seem to occur often in the context of punctuation, suggesting that this presents a particular difficulty for the classifier.

Interestingly, some false positives are in fact substrings of longer MWEs. Because we focused on bigrams in this research, MWEs of longer than two tokens were ignored when assessing whether a candidate MWE was a true or false MWE. However, some of these candidate MWEs were in fact substrings of a longer MWE. The classifier may therefore be recognising that a given substring occurs in a context typical of MWEs, and is identifying the MWE substring as being a MWE in its own right. A fuller implementation which extracts MWEs longer than two tokens might therefore be expected to eliminate this source of error.

In spite of the occasional error, applying a classifier to the context surrounding a candidate MWE seems to offer an effective means of distinguishing true MWEs from non-MWEs and pseudo-MWEs.

4.3 Part of Speech Estimation

Evaluation of the part of speech classifiers shows them to be an effective means of estimating an MWE's part of speech based on its context of occurrence (Table 3). As we might expect, the part of speech for nouns performed best, with near perfect recall and high precision. The conjunctions classifier performed least well with recall in particular being lower than that achieved for other parts of speech. This may reflect a greater variability in the contexts surrounding conjunctive MWEs. Conjunctions often play a discursive role in sentences, so evidence of an expression being a conjunction or not might be found at a higher level of linguistic analysis than the immediate lexical context used in our experiment.

Table 3: Part of speech estimation results

Part of speech	Prec.	Rec.	F
Prepositions	98.06	98.40	98.23
Conjunctions	97.10	95.37	96.23
Adverbs	98.73	98.72	98.72
Nouns	98.88	99.25	99.07

5 Future Work

In this work we have focused on bigrams. We hope to generalise our approach, so that MWEs of length greater than two can be extracted and assigned a part of speech.

We plan to evaluate the performance of the BNC models described above on another corpus to determine their flexibility. Specifically, we plan to use a corpus of North American English such as the Penn Treebank, in the hope of demonstrating the models' ability to handle American as well as British English.

We also plan to check the effect of the extracted expressions on parsing accuracy, by using them in the input to a parser such as Collins' [3] or Yamada's [10].

6 Conclusion

In this research we aimed to identify a method for the automatic extraction and part of speech estimation of fixed MWEs. Knowledge about fixed MWEs has the potential to improve the accuracy of numerous natural language processing applications. Generating such a list therefore represents an important natural language processing task.

Our method uses a collocation measure to produce a list of candidate bigrams. These candidates are then used to select training data for a classifier. The trained classifier was successfully able to distinguish between contexts containing a true MWE, from contexts containing a pseudo-MWE or no MWE at all. The classifier trained on one million examples of candidates identified using the t-score had precision of 91.5%, and recall of 89.1%, giving an F-measure of 90.3%. Part of speech classifiers were then trained and tested. The classifiers

were able to identify the correct part of speech for an MWE with a precision and recall of over 95%.

These results show that the local context surrounding an MWE contains sufficient information to identify its presence, and estimate its part of speech. If this information is detailed enough, we may be able to perform additional processing steps. For example, it may be possible to distinguish between specific sub-types of fixed MWE. The present method needs to be generalised so it can deal with MWEs of any length, not just bigrams. We plan to explore these issues in future research.

References

[1] Timothy Baldwin and Aline Villavicencio. Extracting the unextractable: A case study on verb-particles. In Dan Roth and Antal van den Bosch, editors, *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 98–104, Taipei, Taiwan, August-September 2002.

[2] Ken Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, March 1990.

[3] Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

[4] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. In *Computational Linguistics*, pages 61–74, 1993.

[5] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[6] Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico, 2002. CICLING.

[7] Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993. *Special Issue on Using Large Corpora: I.

[8] Aristomenis Thanopoulos, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of collocation extraction metrics. In *International Conference on Language Resources and Evaluation (LREC-2002)*, pages 620–625, 2002.

[9] Aline Villavicencio. Verb-particle constructions and lexical resources. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 57–64. ACL, 2003.

[10] Hiroyasu Yamada and Yūji Matsumoto. Statistical dependency analysis with support vector machines. In *IWPT 2003: 8th International Workshop on Parsing Technologies*, pages 195–206, 2003.