Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool

Scott S.L. Piao[1] s.piao@lancaster.ac.uk

Guangfan Sun[2] morgan2001_sun@sohu.com

Paul Rayson[1] paul@comp.lancs.ac.uk

Qi Yuan[2] yq@trans.ccidnet.com

[1]UCREL
Computing Department
Lancaster University
Lancaster, UK

[2]CIPOL
China Centre for Information Industry Development (CCID)
Beijing, China

Abstract

In this paper, we report on our experiment to extract Chinese multiword expressions from corpus resources as part of a larger research effort to improve a machine translation (MT) system. For existing MT systems, the issue of multiword expression (MWE) identification and accurate interpretation from source to target language remains an unsolved problem. Our initial test on the Chinese-to-English translation functions of Systran and CCID's Huan-Yu-Tong MT systems reveal that, where MWEs are involved, MT tools suffer in terms of both comprehensibility and adequacy of the translated texts. For MT systems to become of further practical use, they need to be enhanced with MWE processing capability. As part of our study towards this goal, we test and evaluate a statistical tool, which was developed for English, for identifying and extracting Chinese MWEs. In our evaluation, the tool achieved precisions ranging from 61.16% to 93.96% for different types of MWEs. Such results demonstrate that it is feasible to automatically identify many Chinese MWEs using our tool, although it needs further improvement.

1 Introduction

In real-life human communication, meaning is often conveyed by word groups, or meaning groups, rather than by single words. Very often, it is difficult to interpret human speech word by word. Consequently, for an MT system, it is important to identify and interpret accurate meaning of such word groups, or multiword expressions (MWE hereafter), in a source language and interpret them accurately in a target language. However, accurate identification and interpretation of MWEs still remains an unsolved problem in MT research.

In this paper, we present our experiment on identifying Chinese MWEs using a statistical tool for MT purposes. Here, by multiword expressions, we refer to word groups whose constituent words have strong collocational relations and which can be translated in the target language into stable translation equivalents, either single words or MWEs, e.g. noun phrases, prepositional phrases etc. They may include technical terminology in specific domains as well as more general fixed expressions and idioms. Our observations found that existing Chinese-English MT systems cannot satisfactorily translate MWEs, although some may employ a machine-readable bilingual dictionary of idioms. Whereas highly compositional MWEs may pose a trivial challenge to human speakers for interpretation, they present a tough challenge for fully automatic MT systems to produce even remotely fluent translations. Therefore, in our context, we expand the concept of MWE to include those compositional ones which have relatively stable identifiable patterns of translations in the target language.

By way of illustration of the challenge, we experimented with simple Chinese sentences containing some commonly-used MWEs in SYSTRAN (http://www.systransoft.com/) and Huan-Yu-Tong (HYT henceforth) of CCID (China Centre for Information Industry Development) (Sun, 2004). The former is one of the most efficient MT systems today, claiming to be "the leading provider of the world's most scalable and modular translation architecture", while the latter is one of the most successful MT systems in China. Table 1 shows the result, where SL and TL denote source and target languages respectively.. As shown by the samples, such highly sophisticated MT tools still struggle to produce adequate English sentences..

Chinese Sentences	English (Systran)	English (HYT)
今天下午会练球吗？ 我*希望不会*。	This afternoon can practice a ball game? I hope not to be able.	Can practise a ball game this afternoon? I hope can not.
你不可以那样做，让我们*各付各*的。	You may not such do, let us pay respectively each.	You cannot do like that, and let us make it Dutch.
恐怕没办法让你们坐*同桌*,你们介不介意分开坐呢？	Perhaps does not have the means to let you sit shares a table, did you mind sits separately?	Perhaps no way out(ly) let you sit with table, are you situated between not mind to separate to sit?
来点冰镇的*奶咖啡*。	Selects the milk coffee which ices.	Ice breasts coffee take is selected.
好的，我要啤酒，*再来点*咖啡。	Good, I want the beer, again comes to select the coffee.	Alright, I want beer, and take the coffee of ordering again.

Table 1: Samples of Chinese-to-English translations of Systran and HYT.

Ignoring the eccentric English syntactic structures these tools produced, we focus on the translations of Chinese MWEs (see the italic characters in the Table 1) which have straightforward expression equivalents in English. For example, in this context, *希望不会* can be translated into "hope not", *各付各* into "go Dutch", *同桌* into "together" or "at the same table", *奶咖啡* into "white coffee" or "coffee with milk", *再来点* into "want some more (in addition to something already ordered)". While these Chinese MWEs are highly compositional ones, when they are translated word by word, we see verbose and awkward translations (for correct translations, see the appendix).

To solve such problems, we need algorithms and tools for identifying MWEs in the source language (Chinese in this case) and to accurately map them to their adequate translation equivalents in the target language (English in our case) that are appropriate for given contexts. In the previous examples, an MT tool should be able to identify the Chinese MWE 各付各 and either provide the literal translation of "pay for each" or map it to the more idomatic expressions of "go Dutch".

Obviously, it would involve a wide range of issues and techniques for a satisfactory solution to this problem. In this paper, we focus on the sub-issue of automatically recognising and extracting Chinese MWEs. Specifically, we test and evaluate a statistical tool for automatic MWE extraction in Chinese corpus data. As the results of our experiment demonstrate, the tool is capable of identifying many MWEs with little language-specific knowledge. Coupled with an MT system, such a tool could be useful for addressing the MWE issue.

2 Related Work

The issue of MWE processing has attracted much attention from the Natural Language Processing (NLP) community, including Smadja, 1993; Dagan and Church, 1994; Daille, 1995; 1995; McEnery *et al.*, 1997; Wu, 1997; Michiels and Dufour, 1998; Maynard and Ananiadou, 2000; Merkel and Andersson, 2000; Piao and McEnery, 2001; Sag *et al.*, 2001; Tanaka and Baldwin, 2003; Dias, 2003; Baldwin et al., 2003; Nivre and Nilsson, 2004 Pereira et al,. 2004; Piao et al., 2005. Study in this area covers a wide range of sub-issues, including MWE identification and extraction from monolingual and multilingual corpora, classification of MWEs according to a variety of viewpoints such as types, compositionality and alignment of MWEs across different languages. However studies in this area on Chinese language are limited.

A number of approaches have been suggested, including rule-based and statistical approaches, and have achieved success to various extents. Despite this research, however, MWE processing still presents a tough challenge, and it has been receiving increasing attention, as exemplified by recent MWE-related ACL workshops.

Directly related to our work is the development of a statistical MWE tool at Lancaster for searching and identifying English MWEs in running text (Piao et al., 2003, 2005). Trained on corpus data in a given domain or genre, this tool can automatically identify MWEs in running text or extract MWEs from corpus data from the similar domain/genre (see further information about this tool in section 3.1). It has been tested and compared with an English semantic tagger (Rayson et al., 2004) and was found to be efficient in identifying domain-specific MWEs in English corpora, and complementary to the semantic tagger which relies on a large manually compiled lexicon.

Other directly related work includes the development of the HYT MT system at CCID in Beijing, China. It has been under development since 1991 (Sun, 2004) and it is one of the most successful MT systems in China. However, being a mainly rule-based system, its performance degrades when processing texts from domains previously unknown to its knowledge database. Recently a corpus-based approach has been adopted for its improvement, and efforts are being made to improve its capability of processing MWEs.

Our main interest in this study is in the application of a MWE identification tool to the improvement of MT system. As far as we know, there has not been a satisfactory solution to the efficient handling of Chinese MWEs in MT systems, and our experiment contributes to a deeper understanding of this problem.

3 Automatic Identification and extraction of Chinese MWEs

In order to test the feasibility of automatic identification and extraction of Chinese MWEs on a large scale, we used an existing statistical tool built for English and a Chinese corpus built at CCID. A CCID tool is used for tokenizing and POS-tagging the Chinese corpus. The result was thoroughly manually checked by Chinese experts at CCID. In this paper, we aim to evaluate this existing tool from two perspectives a) its performance on MWE extraction, and b) its performance on a language other than English. In the following sections, we describe our experiment in detail and discuss main issues that arose during the course of our experiment.

3.1 MWE extraction tool

The tool we used for the experiment exploits statistical collocational information between near-context words (Piao et al., 2005). It first collects collocates within a given scanning window, and then searches for MWEs using the collocational information as a statistical dictionary. As the collocational information can be extracted on the fly from the corpus to be processed for a reasonably large corpus, this process is fully automatic. To search for MWEs in a small corpus, such as a few sentences, the tool needs to be trained on other corpus data in advance.

With regards to the statistical measure of collocation, the option of several formulae are available, including mutual information and log likelihood, etc. Our past experience shows that log-likelihood provides an efficient metric for corpus data of moderate sizes. Therefore it is used in our experiment. It is calculated as follows (Scott, 2001).

For a given pair of words *X* and *Y* and a search window *W*, let *a* be the number of windows in which *X* and *Y* co-occur, let *b* be the number of windows in which only *X* occurs, let *c* be the number of windows in which only *Y* occurs, and let *d* be the number of windows in which none of them occurs, then

$$\begin{aligned}G_2 = 2\,(a\mathrm{ln}a + b\mathit{ln}b + c\mathit{ln}c + d\mathit{ln}d - (a+b)\mathit{ln}(a+b)\\ - (a+c)\mathit{ln}(a+c) - (b+d)\mathit{ln}(b+d)\\ - (c+d)\mathit{ln}(c+d)) + (a+b+c+d)\mathit{ln}(a+b+c+d))\end{aligned}$$

In addition to the log-likelihood, the *t-score* is used to filter out insignificant co-occurrence word pairs (Fung and Church, 1994), which is calculated as follows:

$$t = \frac{prob(W_a, W_b) - prob(W_a)prob(W_b)}{\sqrt{\frac{1}{M} prob(W_a, W_b)}}$$

In order to filter out weak collocates, a threshold is often used, i.e. in the stage of collocation extraction, any pairs of items producing word affinity scores lower than a given threshold are excluded from the MWE searching process. Furthermore, in order to avoid the noise caused by functional words and some extremely frequent words, a stop word list is used to filter such words out from the process.

If the corpus data is POS-tagged, some simple POS patterns can be used to filter certain syntactic patterns from the candidates. It can either be implemented as an internal part of the process, or as a post-process. In our case, such pattern filters are mostly applied to the output of the MWE searching tool in order to allow the tool to be language-independent as much as possible.

Consequently, for our experiment, the major adjustment to the tool was to add a Chinese stop word list. Because the tool is based on Unicode, the stop words of different languages can be kept in a single file, avoiding any need for adjusting the program itself. Unless different languages involved happen to share words with the same form, this practice is safe and reliable. In our particular case, because we are dealing with English and Chinese, which use widely different characters, such a practice performs well.

Another language-specific adjustment needed was to use a Chinese POS-pattern filter for selecting various patterns of the candidate MWEs (see Table 6). As pointed out previously, it was implemented as a simple pattern-matching program that is separate from the MWE tool itself, hence minimizing the modification needed for porting the tool from English to Chinese language.

A major advantage of this tool is its capability of identifying MWEs of various lengths which are generally representative of the given topic or domain. Furthermore, for English it was found effective in extracting domain-specific multiword terms and expressions which are not included in manually compiled lexicons and dictionaries. Indeed, due to the open-ended nature of such MWEs, any manually compiled lexicons, however large they may be, are unlikely to cover them exhaustively. It is also efficient in finding newly emerging MWEs, particularly technical terms, that reflect the changes in the real world.

3.2 Experiment

In this experiment, our main aim was to examine the feasibility of practical application of the MWE tool as a component of an MT system, therefore we used test data from some domains in which translation services are in strong demand. We selected Chinese corpus data of approximately 696,000 tokenised words (including punctuation marks) which cover the topics of food, transportation, tourism, sports (including the Olympics) and business.

In our experiment, we processed the texts from different topics together. These topics are related to each other under the themes of entertainment and business. Therefore we assume, by mixing the data together, we could examine the performance of the MWE tool in processing data from a broad range of related domains. We expect that the different features of texts from different domains will have a certain impact on the result, but the examination of such impact is beyond the scope of this paper.

As mentioned earlier, the Chinese word tokeniser and POS tagger used in our experiment has been developed at CCID. It is an efficient tool running with accuracy of 98% for word tokenisation and 95% for POS annotation. It employs a part-of-speech tagset of 15 categories shown in Table 2. Although it is not a finely grained tagset, it meets the need for creating POS pattern filters for MWE extraction.

N	Name
V	Verb
A	Adjective
F	Adverb
R	Pronoun
I	Preposition
J	Conjunction
U	Number
S	classifier (measure word)
G	Auxiliary verb
E	Accessory word
L	directional noun
P	Punctuation
H	Onomatopoeia
X	Subject-predicate phrase

Table 2: CCID Chinese tagset

Since function words are found to cause noise in the process of MWE identification, a Chinese stop list was collected. First, a word frequency list was extracted. Next, the top items were considered and we selected 70 closed class words for the stop word list. When the program searches for MWEs, such words are ignored.

The threshold of word affinity strength is another issue to be addressed. In this experiment, we used log-likelihood to measure the strength of collocation between word pairs. Generally the log-likelihood score of 6.6 ($p < 0.01$ or 99% confidence) is recommended as the threshold (Rayson *et al.*, 2004), but it was found to produce too many false candidates in our case. Based on our initial trials, we used a higher threshold of 30, i.e. any word pairs producing log-likelihood score less than this value are ignored in the MWE searching process. Furthermore, for the sake of the reliability of the statistical score, when extracting collocates, a frequency threshold of five was used to filter out low-frequency words, i.e. word pairs with frequencies less than five were ignored.

An interesting issue for us in this experiment is the impact of the length of collocation searching window on the MWE identification. For this purpose, we tested two search window lengths 2 and 3, and compared the results obtained by using them. Our initial hypothesis was that the shorter window length may produce higher precision while the longer window length may sacrifice precision but boost the MWE coverage.

The output of the tool was manually checked by Chinese experts at CCID, including cross checking to guarantee the reliability of the results. There were some MWE candidates on which disagreements arose. In such cases, the candidate was counted as false. Furthermore, in order to estimate the recall, experts manually identified MWEs in the whole test corpus, so that the output of the automatic tool could be compared against it. In the following section, we present a detailed report on our evaluation of the MWE tool.

3.3 Evaluation

We first evaluated the overall precision of the tool. A total of 7,142 MWE candidates (types) were obtained for window lengths of 2, of which 4,915 were accepted as true MWEs, resulting in a precision of 68.82%. On the other hand, a total of 8,123 MWE candidates (types) were obtained for window lengths of 3, of which 4,968 were accepted as true MWEs, resulting in a precision of 61.16%. This result is in agreement with our hypothesis that shorter search window length tends to produce higher precision.

Next, we estimated the recall based on the manually analysed data. When we compared the accepted MWEs from the automatic result against the manually collected ones, we found that the experts tend to mark longer MWEs, which often contain the items identified by the automatic tool. For example, the manually marked MWE 网球 运动 发展 计划 (development plan for the tennis sport) contains shorter MWEs 网球 运动 (tennis sport) and 发展 计划 (development plan) which were identified by the tool separately. So we decided to take the partial matches into account when we estimate the recall. We found that a total 14,045 MWEs were manually identified and, when the search window length was set to two and three, 1,988 and 2,044 of them match the automatic output, producing recalls of 14.15% and 14.55% respectively. It should be noted that many of the manually accepted MWEs from the automatic output were not found in the manual MWE collection. This discrepancy was likely caused by the manual analysis being carried out independently of the automatic tool, resulting in a lower recall than expected. Table 3 lists the precisions and recalls.

Window length = 2		Window length = 3	
Precision	Recall	Precision	Recall
68.82%	14.15%	61.16%	14.55%

Table 3: Overall precisions and recalls

Furthermore, we evaluated the performance of the MWE tool from two aspects: frequency and MWE pattern.

Generally speaking, statistical algorithms work better on items of higher frequency as it depends on the collocational information. However, our tool does not select MWEs directly from the collocates. Rather, it uses the collocational information as a statistical dictionary and searches for word sequences whose constituent words have significantly strong collocational bonds between them. As a result, it is capable of identifying many low-frequency MWEs. Table 4 lists the breakdown of the precision for five frequency bands (window length = 2).

Freq	Candidates	True MWEs	Precision
>= 100	17	9	52.94%
10 ~ 99	846	646	76.36%
3 ~ 9	2,873	2,178	75.81%
2	949	608	64.07%
1	2,457	1,474	59.99%
Total	7,142	4,915	68.82%

Table 4: Breakdown of precision for frequencies (window length = 2).

As shown in the table above, the highest precisions were obtained for the frequency range between 3 and 99. However, 2,082 of the accepted MWEs have frequencies of one or two, accounting for 42.36% of the total accepted MWEs. Such a result demonstrates again that our tool is capable of identifying low-frequency items. An interesting result is for the top frequency band (greater than 100). Against our general assumption that higher frequency brings higher precision, we saw the lowest precision in the table for this band. Our manual examination reveals this was caused by the high frequency numbers, such as "one" or "two" in the expressions "一个" (a/one) and "一种" (a kind of). This type of expression were classified as uninteresting candidates in the manual checking, resulting in higher error rates for the high frequency band.

When we carry out a parallel evaluation for the case of searching window length of 3, we see a similar distribution of precision across the frequency bands except that the lowest frequency band has the lowest precision, as shown by Table 5. When we compare this table against Table 4, we can see, for all of the frequency bands except the top one, that the precision drops as the search window increases. This further supports our earlier assumption that wider searching window tends to reduce the precision.

Freq	candidates	true MWEs	Precision
>= 100	17	9	52.94%
10 ~ 99	831	597	71.84%
3 ~ 9	3,093	2,221	71.81%
2	1,157	669	57.82%
1	3,025	1,472	48.66%
Total	8,123	4,968	61.16%

Table 5: Breakdown of precision for frequencies (window length = 3).

In fact, not only the top frequency band, much of the errors of the total output were found to be caused by the numbers that frequently occur in the test data, e.g. 一_U 个_S (one), 两_U 个_S (two) etc. When a POS filter was used to filter them out, for the window length 2, we obtained a total 5,660 candidates, of which 4,386 were accepted as true MWEs, producing a precision of 77.49%. Similarly for the window length 3, a total of 6,526 candidates were extracted in this way and 4,685 of them were accepted as true MWEs, yielding a precision of 71.79%.

Another factor affecting the performance of the tool is the type of MWEs. In order to examine the potential impact of MWE types to the performance of the tool, we used filters to select MWEs of the following three patterns:

1) AN: Adjective + noun structure;
2) NN: Noun + noun Structure;
3) FV: Adverb + Verb.

Table 6 lists the precision for each of the MWE types and for search window lengths of 2 and 3.

Search window length = 2			
Pattern	Candidate	True MWEs	Precision
A+N	236	221	93.64%
N+N	644	589	91.46%
F+V	345	321	93.04%
total	1,225	1,131	92.33%
Search window length = 3			
Pattern	Candidate	True MWEs	Precision
A+N	259	233	89.96%
N+N	712	635	89.19%
F+V	381	358	93.96%
Total	1,352	1,226	90.68%

Table 6: Precisions for three types of MWEs

As shown in the table, the MWE tool achieved high precisions above 91% when we use a search window of two words. Even when the search window expands to three words, the tool still obtained precision around 90%. In particular, the tool is efficient for the verb phrase type. Such a result demonstrates that, when we constrain the search algorithm to some specific types of MWEs, we can obtain higher precisions. While one may argue that rule-based parser can do the same work, it must be noted that we are not interested in all grammatical phrases, but those which reflect the features of the given domain. This is achieved by combining statistical word collocation measures, a searching strategy and simple POS pattern filters.

Another interesting finding in our experiment is that our tool extracted clauses, such as *想喝些什么* (What would you like to drink?) and *先喝点什么？* (Would you like a drink first?). The clauses occur only once or twice in the entire test data, but were recognized by the tool because of the strong collocational bond between their constituent words. The significance of such performance is that such clauses are typical expressions which are frequently used in real-life conversation in the contexts of the canteen, tourism etc. Such a function of our tool may have practical usage in automatically collecting longer typical expressions for the given domains.

4 Discussion

As our experiment demonstrates, our tool provides a practical means of identifying and extracting domain specific MWEs with a minimum amount of linguistic knowledge. This becomes important in multilingual tasks in which it can be costly and time consuming to build comprehensive rules for several languages. In particular, it is capable of detecting MWEs of various lengths, sometimes whole clauses, which are often typical of the given domains of the corpus data. For example, in our experiment, the tool successfully identified several daily used long expressions in the domain of food and tourism. MT systems often suffer when translating conversation. An efficient MWE tool can potentially alleviate the problem by extracting typical clauses used in daily life and mapping them to adequate translations in the target language.

Despite the flexibility of the statistical tool, however, there is a limit to its performance in terms of precision. While it is quite efficient in providing MWE candidates, its output has to be either verified by human or refined by using linguistic rules. In our particular case, we improved the precision of our tool by employing simple POS pattern filters. Another limitation of this tool is that currently it can only recognise continuous MWEs. A more flexible searching algorithm is needed to identify discontinuous MWEs, which are important for NLP tasks.

Besides the technical problem, a major unresolved issue we face is what constitutes MWEs. Despite agreement on the core MWE types, such as idioms and highly idiosyncratic expressions, like 成语 (Cheng-Yu) in Chinese, it is difficult to reach agreement on less fixed expressions.

We contend that MWEs may have different definitions for different research purposes. For example, for dictionary compilation, lexicographers tend to constrain MWEs to highly non-compositional expressions (Moon, 1998: 18). This is because monolingual dictionary users can easily understand compositional MWEs and there is no need to include them in a dictionary for native speakers. For lexicon compilation aimed at practical NLP tasks, however, we may apply a looser definition of MWEs. For example, in the Lancaster semantic lexicon (Rayson et al., 2004), compositional word groups such as "youth club" are considered as MWEs alongside non-compositional expressions such as "food for thought" as they depict single semantic units or concepts. Furthermore, for the MT research community whose primary concern is cross-language interpretation, any multiword units that have stable translation equivalent(s) in a target language can be of interest.

As we discussed earlier, a highly idiomatic expression in a language can be translated into a highly compositional expression in another language, and vice versa. In such situations, it can be more practically useful to identify and map translation equivalents between the source and target languages regardless of their level of compositionality.

Finally, the long Chinese clauses identified by the tool can potentially be useful for the improvement of MT systems. In fact, most of them are colloquial expressions in daily conversation, and many such Chinese expressions are difficult to parse syntactically. It may be more feasible to identify such expressions and map them as a whole to English equivalent expressions. The same may apply to technical terms, jargon and slang. In our experiment, our tool demonstrated its capability of detecting such expressions, and will prove useful in this regard.

5 Conclusion

In this paper, we have reported on our experiment of automatic extraction of Chinese MWEs using a statistical tool originally developed for English. Our statistical tool produced encouraging results, although further improvement is needed to become practically applicable for MT system in terms of recall. Indeed, for some constrained types of MWEs, high precisions above 90% have been achieved. This shows, enhanced with some linguistic filters, it can provide a practically useful tool for identifying and extracting MWEs. Furthermore, in our experiment, our tool demonstrated its capability of multilingual processing. With only minor adjustment, it can be ported to other languages. Meanwhile, further study is needed for a fuller understanding of the factors affecting the performance of statistical tools, including the text styles and topic/domains of the texts, etc.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (grant no. 60520130297) and the British Academy (grant no. SG-42140).

References

Biber, D., Conrad, S., Cortes, V., 2003. Lexical bundles in speech and writing: an initial taxonomy. In: Wilson, A., Rayson P., McEnery, T. (Eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Peter Lang, Frankfurt. pp. 71-92.

Baldwin, T., Bannard, C., Tanaka, T. and Widdows, D. 2003 An Empirical Model of Multiword Expression Decomposability, In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 89–96.

Dagan, I., Church, K., 1994. Termight: identifying and translating technical terminology. In: *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart, German. pp. 34-40.

Daille, B., 1995. *Combined approach for terminology extraction: lexical statistics and linguistic filtering.* Technical paper 5, UCREL, Lancaster University.

Dias, G., 2003. Multiword unit hybrid extraction. In: *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL'03*, Sapporo, Japan. pp. 41-48.

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1), 61-74.

Fung, P., Church, K., 1994. K-vec: a new approach for aligning parallel texts. In: *Proceedings of COLING '94*, Kyoto, Japan. pp. 1996-2001.

Maynard, D., Ananiadou, S., 2000. Trucks: a model for automatic multiword term recognition. *Journal of Natural Language Processing* 8 (1), 101-126.

McEnery, T., Lange, J. M., Oakes, M., Vernonis, J.., 1997. The exploitation of multilingual annotated corpora for term extraction. In: Garside, R., Leech, G., McEnery, A. (Eds.), *Corpus Annotation --- Linguistic Information from Computer Text Corpora*. Longman, London & New York. pp 220-230.

Merkel, M., Andersson, M., 2000. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In: *Proceedings of 2000 Conference User-Oriented Content-Based Text and Image Handling (RIAO'00)*, Paris, France. pp. 737-746.

Michiels, A., Dufour, N., 1998. DEFI, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In: *Proceedings of the First International Conference on Language Resources & Evaluation*, Granada, Spain. pp. 1179-1186.

Moon, R. 1998. *Fixed expressions and idioms in English: a corpus-based approach*. Clarendon Press: Oxford.

Nivre, J., Nilsson, J., 2004. Multiword units in syntactic parsing. In: *Proceedings of LREC-04 Workshop on Methodologies & Evaluation of Multiword Units in Real-world Applications*, Lisbon, Portugal. pp. 37-46.

Pereira, R., Crocker, P., Dias, G., 2004. A parallel multikey quicksort algorithm for mining multiword units. In: *Proceedings of LREC-04 Workshop on Methodologies & Evaluation of Multiword Units in Real-world Applications*, Lisbon, Portugal. pp. 17-23.

Piao, S. L., Rayson, P., Archer, D. and McEnery, T. 2005. Comparing and Combining A Semantic Tagger and A Statistical Tool for MWE Extraction. *Computer Speech & Language* Volume 19, Issue 4, pp. 378-397.

Piao, S.L , Rayson, P., Archer, D., Wilson, A. and McEnery, T. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, at ACL'03, Sapporo, Japan, pp. 49-56.

Piao, S., McEnery, T., 2001. Multi-word unit alignment in English-Chinese parallel corpora. In: *Proceedings of the Corpus Linguistics 2001*, Lancaster, UK. pp. 466-475.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. 2004. The UCREL semantic analysis system. In *proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with LREC 2004*, Lisbon, Portugal, pp. 7-12.

Rayson, P., Berridge, D. and Francis, B. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. In Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium. pp. 926-936.

Sag, I., Baldwin, T., Bond, F., Copestake, A., Dan, F., 2001. Multiword expressions: a pain in the neck for NLP. *LinGO Working Paper No. 2001-03*, Stanford University, CA.

Scott, M., 2001. Mapping key words to problem and solution. In: Scott, M., Thompson, G. (Eds.), *Patterns of Text: in Honour of Michael Hoey*. Benjamins, Amsterdam. pp. 109 – 127.

Smadja, F., 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19 (1), 143-177.

Sun, G. 2004. Design of an Interlingua-Based Chinese-English Machine Translation System. In *Proceedings of the 5th China-Korea Joint Symposium on Oriental Language Processing and Pattern Recognition*, Qingdao, China. pp. 129-134.

Tanaka, T., Baldwin, T., 2003. Noun-noun compound machine translation: a feasibility study on shallow processing. In: *Proceedings of the ACL-03 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan. pp. 17-24.

Wu, D., 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23 (3), 377-401.

Appendix: English translations of the sample Chinese sentences

1. 今天下午会练球吗？我*希望不会*。

Tran: Do we have (football) training this afternoon? I *hope not*.

2. 你不可以那样做，让我们*各付各*的。

Tran: You can’t do that. Let’s go Dutch.

3. 恐怕没办法让你们坐*同桌*，你们介不介意分开坐呢？

Tran: I am afraid I can’t arrange for you to sit at the same table. Would you mind if you sit separately?

4. 来点冰镇的*奶咖啡*。

Tran: I’d like iced white coffee (please).

5. 好的，我要啤酒，*再来点*咖啡。

Tran: OK, I want beer and some coffee (please).