Multi-word Unit Alignment in English-Chinese Parallel Corpora

Scott Songlin Piao
Department of Computer Science
University of Sheffield
Email: s.piao@dcs.shef.ac.uk

Tony McEnery
Department of Linguistics and MEL
Lancaster University
Email: mcenery@comp.lancs.ac.uk

1. Introduction

Multi-word unit (MWU) alignment in bilingual/multilingual parallel corpora is an important goal for natural language engineering. An efficient algorithm for aligning MWUs in different languages could be of use in several practical applications, including machine translation, lexicon construction and cross-language information retrieval. A number of algorithms have been proposed and tested for this purpose, including collocation and association-strength testing statistics (Dagan *et al.*, 1994; Smadja *et al.*, 1996), n-gram, approximate string matching techniques (ASMT), finite state automata, (McEnery *et al.*, 1997), bilingual parsing matching (Wu, 1997), and a hybrid connectionist framework (Wermter *et al.*, 1997). Despite the work undertaken to date, however, reliable and robust MWU alignment remains an elusive goal.

In this paper, we describe an algorithm which combines the n-gram approach, linguistic filters and co-occurrence statistical metrics to extract and align English and Chinese nominal MWUs in English-Chinese parallel corpora. We decided to focus on nominal MWUs as they are stable and form the largest group of MWUs in natural language[1]. This algorithm has been evaluated on a sentence aligned English-Chinese Parallel corpus (Piao, 2000). It obtained precision rates of 92.17% and 87.37% for English and Chinese nominal MWU extraction respectively while it obtained a precision rate of 80.63% for English-Chinese MWU alignment.

2. Related works

As noted, a number of techniques have already been applied to the problem of MWU extraction and alignment, including Smadja *et al.*' (1996) collocation translation system, McEnery *et al.*'s (1997) approximate string matching techniques (ASMT) and finite state automata, and Wu's (1997) stochastic inversion transduction grammars used to align Chinese-English phrases.

Dagan and Church (1994) also describe a semi-automatic tool, *Termight*, which extracts technical terms and translations from parallel corpus data. This tool first extracts candidate multi-word noun phrases and single words from a POS tagged corpus using a syntactic pattern filter. Then it groups terms by head words, and sorts terms within each group in reverse word order. Finally, concordance lines are produced so that human experts can distinguish true terms from the candidate terms. Then, the *Termight* aligns bilingual MWUs using aligned words. For each source term, the tool identifies a candidate translation by selecting a sequence of target words whose first and last word are aligned with any of the words in the source term. Bilingual concordances are produced for selected term pairs to allow terminologists to verify true translations. Dagan *et al*, (1994) tested Termight on 192 terms found in English and German versions of a technical manual. They found that in 40% and 7% of the cases the first and second target language candidates respectively were the correct translations. In other cases, the correct translation was always somewhere in the concordances.

The Champollion system of Smadja *et al.*' (1996) can produce translations of the source collocations in the target language. It is based on a MWU extraction system, Xtract, which was also developed by Smadja *et al*. (1993). They first extracted source language (English) collocations with Xtract. After that, for each source collocation, they extracted its translationd in the target language (French) by testing Dice-score and collocation lists. Champollion was tested on the Hansard corpus and an accuracy of between 65%-78% was reported.

[1] Biber *et al.* (1999: 231) observe: "In the news report, nominal elements make up about 80 per cent of the text (measured in terms of words). The corresponding figures for the other text samples are approximately: academic prose 75 per cent, fiction 70 percent, and conversation 55 per cent. In other words, nominal elements make up between a half and four-fifth of the text."

McEnery *et al*. (1997) tested extracting multi-word cognates from parallel corpora using ASMT and finite state automata. In this algorithm, all n-grams from source text are compared against all potential m-grams in an aligned region of the parallel text. Dice's similarity coefficient is calculated for each pair (n, m). The (n, m) pair which gains the highest score is selected as the best potential MWU cognate. For 3,142 windows matched in a million-words English-Spanish parallel corpus, an overall precision of 96.5% was reported. They also developed finite state automata for typical English and Spanish compound noun constructions to extract alignment candidates, then filtered the candidates with a similarity score. This effectiveness of the technique was found to be directly linked to the strength of the similarity scores computed for candidate terms.

Wu (1997) approached Chinese-English phrase alignment by bilingual parsing with stochastic inversion transduction grammars. Wu reports that his approach produced phrase alignments as long as 12 English tokens and 15 Chinese characters. After pruning, a precision rate of 81.5% (based on random samples drawn from about 2,800 phrasal alignments) was reported.

3. A hybrid algorithm for aligning English-Chinese nominal MWUs

In this paper, we describe an algorithm of nominal MWU alignment in which a number of techniques outlined in the previous section are combined. Our approach uses n-grams, POS filters and co-occurrence metrics in concert. The n-gram approach is used to extract candidate MWUs. POS filters are then used to extract a shortlist of candidate nominal MWUs. After this, co-occurrence metrics are used to extract true nominal MWUs and their alignments (explained in detail later in this paper). To show the effectiveness of this approach, an English-Chinese parallel corpus is used as a testbed. This corpus contains 61,534 English words and 98,537 Chinese characters. The English data in the corpus was POS tagged with the Lancaster CLAWS tagger (Garside *et. al*. 1987) and the Chinese data was tagged with Zhang *et al*'s (2000) Chinese tagger. Additionally, the corpora have been sentence aligned using Piao's (2000) program.

One assumption underlying the approach taken to nominal MWU alignment in this paper is that nominal MWUs in the source language are *generally* translated into a nominal MWU (hereafter, MWU in this paper refers to nominal MWU) in the target language, and hence their occurrences in the parallel translation texts are correlated. Based on this assumption, we propose that nominal MWU alignment can be approached through a) first extracting significant MWUs from each language in the parallel corpus, then b) aligning them based on their co-occurrence affinity. An algorithm was designed to implement this approach. Two main stages are involved in this algorithm: a) English and Chinese MWU extraction, b) MWU alignment.

3.1. Candidate Nominal MWU extraction

The first stage of our algorithm is to extract candidate nominal English and Chinese MWUs from the English-Chinese parallel corpus. We assume that most MWUs are stable continuous strings of words. Therefore we adopted an n-gram approach to extracting candidate MWUs from the corpus texts. In order to remove irrelevant candidates from the process, simple POS filters are used to filter out n-grams whose POS structures are unlikely to constitute nominal MUWs.

Firstly, candidate MWUs are extracted from the corpus with the following algorithm:

(1) Extract English and Chinese n-grams ($2 <= n <= 6$)[2] from the English and Chinese section of the corpus respectively. Considering the unreliability of statistical scores for low frequency items, those n-grams whose frequency is lower than three were ignored.
(2) An English and Chinese POS filter is used to filter out those n-grams whose POS patterns are unlikely nominal MWUs .

In the first step of this algorithm, any one of the n-grams extracted, from bi-grams to 6-grams, can be candidate MWUs. However, as noted, many of these n-grams have POS sequences that make them

[2] Because only twelve 6-grams with frequencies equal or greater than three were found after POS filtering in the tesbed corpus, we assumed that the nominal MWUs longer than six words are non-existent in this corpus.

unlikely candidate nominal MWUs. These false MWUs cause "noise" for the algorithm, and hence it is desirable to filter them out before proceeding.

In filtering the candidate n-grams, the POS sequence associated with each n-gram is matched against simple POS filters. The English POS filter[3] is a simple rule system that excludes any of the candidates if the any of the following conditions are met:

1) For the initial word of a n-gram, if
   a) the initial code of its POS tag is any one of "A", 'C', 'D', 'G', 'I', 'M', 'P', 'R' or 'T', or
   b) the initial code of its POS tag is 'V' and the last code is not 'G' or 'N', or
   c) it has the POS tag of "AT".
2) For the last word of a n-gram, if its initial POS code is not 'N'.

Table 1 shows the English POS categories denoted by the POS tags or tag-initials used in the English filter (for English C7 tagset see appendixIII in).

Code	Category
*	Any codes
A*	Possessive pronouns and articles
C*	Conjunction
D*	Determiner
G*	Genitive markers ' and 's
I*	Preposition
M*	Numeral
N*	Noun/Proper noun
P*	Pronoun
R*	Adverbs
T*	Infinitive marker "to"
V*	Verb
V*G	*-ing* verb
V*N	Past participle

Table 1: English POS categories denoted by the codes in the English POS filter

The Chinese POS filter[4] works by excluding candidate Chinese n-grams if any of the following conditions are met:

1) For the initial word,
   a) the initial code of its POS tag is any of 'C', 'D', 'M', 'P' and 'V',
   c) its POS tag is "AUX", "AV0" or "NMW".
2) For the last word, the initial code of tag is not 'N'.

Table 2 shows the Chinese POS categories denoted by the codes in the Chinese filter.

Code	Category
AUX	Auxiliary word
AV0	Adverb
C*	Conjunction
D*	Markers "的", "地" and "得"
M*	Numeral
P*	Pronoun and determiner
V*	Verb

[3] For C7 tagset, see appendix III in Roger Garside, Geoffrey Leech, Anthony McEnery (eds) (1997) *Corpus annotation, London & New York, Longman.*
[4] The Chinese tagset used by Zhang *et al*'s tagger was modified. For details, see appendix.

Table 2: Chinese POS categories denoted by the codes in the Chinese POS filter

A notable feature of the filter is that it does not thoroughly examine n-grams in the sense that it only considers the initial and last words of the n-grams. In spite of the simplicity of this filtering mechanism, it generally performs well, particularly on bi-grams and tri-grams. For example, in an experiment, these filters filtered out 6,009/4853 irrelevant English/Chinese bigrams from 6767/5599 English/Chinese bi-grams. Thus while the filter is effective, it does not seem to exclude good candidates while demonstrating a fair degree of success at filtering out bad candidates.

3.2. Seed-gram extraction

After the POS filter has eliminated a proportion of the bad candidate n-grams, a further filtering process, to identify the true nominal MWUs from the candidate list, is necessary. In order to achieve this, we identify *seed-grams.* Seed-grams are short MWUs, including bi-grams and tri-grams, in which the element-tokens are strongly associated. Such seed-grams are identified by testing their co-occurrence associations (for bi-seed-grams) and some specific POS patterns (for tri-seed-grams). It is assumed that a good nominal MWU must contain one or more *seed-grams*. Therefore, seed grams can be used to find longer significant nominal MWUs, i.e. a *seed-gram* can "grow" longer.

Bi-seed-grams are extracted as follows. For each bi-gram, the mutual information (*MI*)[5] and *t*-score is calculated. These scores reflect the co-occurrence affinity between the two tokens of the bi-gram. These two scores are calculated by the following formulas:

Mutual information

$$MI^2 = \log_2 \frac{a^2}{(a+b)(a+c)},$$

t-score

$$t = \frac{prob(W_a, W_b) - prob(W_a)\,prob(W_b)}{\sqrt{\frac{1}{M}\,prob(W_a, W_b)}} = \sqrt{a} - \frac{(a+b)(a+c)}{\sqrt{a}(a+b+c+d)}.$$

where, *a*, *b*, *c* and *d* are elements of a contingency table. For example, given a bi-gram containing tokens *x* and *y*,
a = number of bi-grams in which both *x* and *y* occur;
b = number of bi-grams in which only *x* occurs;
c = number of bi-grams in which only *y* occurs;
d = number of bi-grams in which neither *x* nor *y* occurs.

For the purposes of seed-extraction, thresholds of 1.65 for MI and –3 for t-score are used[6]. If a given bi-gram produces both an MI and t-score greater than the thresholds, it is accepted as a *seed-gram*. This process does not vary in the sense that the same algorithm is applied to both English and Chinese.

When applied to the testbed corpus, this algorithm extracted 414 English *seed-grams* and 315 Chinese *seed-grams* out of the 758 and 746 English and Chinese candidates respectively. All of the extracted seed-grams were found meaningful and accepted to be true *seed-grams*, giving the technique a precision of 100%. However, the recall for the technique for both English and Chinese seedgram extraction is significantly lower at 52.74% and 42.22% respectively. Figures 1 and 2 show examples of English and Chinese seed grams.

5 In formula 1, numerator a is squared, for previous study shows that this modified formula performs better than the original one in which the exponent of a is one (see Piao, 2000).
[6] These parameters were established empirically for the corpus we tested the algorithm on. We accept that these parameters may vary depending upon the corpus being exploited.

MI	t-score	f	Seed gram	POS pattern
4.717	6.527	43	Education Commission	NN1 NNJ
4.373	4.990	25	Nobel Prize	NP1 NN1
4.334	4.683	22	Lianhe Zaobao	NP1 NP1
4.272	5.087	26	hanyu pinyin	NN1 NN1
4.164	4.682	22	Hong Kong	NP1 NP1
4.005	6.864	48	higher learning	JJR NN1
3.816	4.236	18	United States	NP1 NP1
3.700	3.603	13	Chen Qing-shan	NP1 NP1

Figure 1: A sample of top English seed grams

MI	t-score	f	Seed gram	POS pattern
4.431	9.497	98	高等 学校	AJ0 NN0
4.130	5.627	32	汉语 拼音	NN0 NN0
4.104	7.539	59	少数 民族	AJ0 NN0
3.286	3.990	16	股 价	NN0 NN0
3.225	8.569	85	特殊 教育	AJ0 NN0
3.126	5.919	36	国家 教委	NN0 NN0
2.831	3.311	11	荔子 情	NN0 NN0
2.645	5.405	30	残疾 儿童	NN0 NN0

Figure 2: A sample of top Chinese seed grams

For trigrams, POS matching patterns are used for extracting *seed-grams*, as shown below. Trigrams are matched against POS patterns and deemed to be seed grams if they match the following patterns for English and Chinese:

(1) English POS patterns:
[Noun/Proper_noun + of/genitive_marker + Noun/Proper_noun]
(2) Chinese POS patterns:
[Noun/Adjective/Proper_noun + *de*(的) + Noun/Proper_noun]

These heuristic patterns were developed on the basis of the grammatical properties of English and Chinese noun phrases. This proved to be an effective technique, as when tested on the corpus, all of the tri-grams extracted with these POS patterns proved to be significant nominal MWUs. Fig. 1 shows sample *tri-seed-grams* extracted in this way.

f	Eng. tri-seed-gram	POS pattern	f	Chi. tri-seed-gram	POS pattern
33	People 's Republic	NN GE NN1	6	中国 的 教育	NP0 DE1 NN0
30	Republic of China	NN1 IO NP1	5	教育 的 权利	NN0 DE1 NN0
8	institutions of China	NN2 IO NP1	5	中国 的 成人	NP0 DE1 NN0
6	China 's education	NP1 GE NN1	5	中国 的 研究生	NP0 DE1 NN0
6	Ministry of Education	NN1 IO NN1	4	多 的 资料	AJ0 DE1 NN0
6	number of people	NN1 IO NN	4	好 的 教师	AJ0 DE1 NN0
6	number of women	NN1 IO NN2	4	文化 的 社会	NN0 DE1 NN0
6	women 's education	NN2 GE NN1	4	新 的 经济	AJ0 DE1 NN0

Fig. 3: A sample of English and Chinese seed tri-grams

As shown previously, the n-gram approach, simple POS filters and co-occurrence metrics combined provide an efficient algorithm for extracting significant short MWUs, or *seed-grams*. The process of

'growing' these seed-grams to identify nominal MWUs of length greater than 3 is described in the following section.

3.3. Extracting longer MWUs based on2 and 3 length seed-grams

In natural languages, a nominal MWU can clearly be longer than three words. As discussed previously, our supposition was that if an MWU is significant, it is likely to contain one or more *seed-grams*; conversely, if an MWU contains one or more *seed-grams*, it is likely to be significant. Based on this assumption, we used seed-grams to identify true MWUs of a length greater than 3.

To determine the usefulness of this approach to extracting nominal MWUs, we applied the the hypothesis to n-grams of length 3 – 6[7]. The POS filters used to filter out candidate nominal MWUs work well on short n-grams, but work less efficiently on longer n-grams. Hence we applied a further filter to candidate n-grams (3 <= n <=6) as follows:

a) English n-grams containing tags "VV*" (verbs) or "APPGE" (pre-nominal possessive pronoun) are filtered,
b) Chinese n-grams containing "VV0" (verbs), "VM" (modal verbs) are filtered,

where the asterisk denotes any letter(s).

The candidates survived the pruning are taken as candidate nominal MWUs. Each of them is matched against the seed-grams. Those which contain one or more seed-grams are accepted as nominal MWUs. In the experiment, This approach extracted 626 English nominal MWUs and 467 Chinese nominal MWUs from the corpus. Figures 4 and 5 show samples of the extracted English and Chinese noun MWUs respectively.

f	MWU	POS
6	Goh Chok Tong	NP1 NP1 NP1
3	Gross Domestic Product	JJ JJ NN1
7	HIV infection	NP1 NN1
3	Hanyu pinyin	NN1 NN1
5	Harvard University	NP1 NN1
3	Health Information	NN1 NN1
3	Health Publications	NN1 NN2
3	Health Publications Unit	NN1 NN2 NN1
4	Health Service	NN1 NN1
3	Heywood Stores	NP1 NN2

Figure 4: A sample of automatically extracted English nominal MWUs

f	Nominal MWU	POS Tags
4	高层 次 专门 人才	NN0 NMW PND NN0
30	高等 教育	AJ0 NN0
3	高等 教育 体系	AJ0 NN0 NN0
98	高等 学校	AJ0 NN0
3	高等 学校 和 科研 机构	AJ0 NN0 CJ0 NN0 NN0
10	高等 学校 科学	AJ0 NN0 NN0
14	高等 学校 科学 技术	AJ0 NN0 AJ0 NN0
3	高粱 舅	NP0 NN0
6	高中 阶段	NN0 NN0
3	高 的 华人	AJ0 DE1 NN0

Figure 5: A sample of automatically extracted Chinese nominal MWUs

[7] Note we still consider n-grams of length 3 at this stage, as these may be missed by the POS pattern matcher, but contain significant seed-grams of length 2.

A manual examination of the results showed precision rates of 92.17% and 87.37% for English and Chinese respectively. In a further analysis, it was found that 59 Chinese bad MWUs, or 20.34% of the mistakes, were caused by errors in the Chinese POS tagging. This means that, if a more accurate Chinese POS tagger is available, a higher success rate could reasonably be expected for Chinese. Given the considerably high precision yielded throughout the various stages of the technique developed for monolingual nominal MWU extraction, we use the output from the English and Chinese algorithms as the basis upon which alignment of MWUs between the two languages is attempted.

3.4. Aligning English and Chinese MWUs

With the English and Chinese MWUs extracted the next step is to align them. As assumed previously, the English and Chinese translation equivalents are generally expected to co-occur in corresponding sentence translations. Since the test-bed corpus is aligned at sentence level, it is possible to test co-occurrence affinity between the candidate MWUs at sentence level.

Again, the mutual information (MI) and *t*-score (see formulae 1 and 2) are used for testing co-occurrence correlation. For each English MWU x_i (i = 1, 2, …, m) every Chinese MWU y_j (j = 1, 2, …, n) is considered as a potential translation (n and m denote the numbers of English and Chinese MWUs). For a given x_i, a contingency table is extracted against every y_j. The contingency table contains elements, a, b, c and d which are defined as follows:

(1) a denotes the number of aligned English-Chinese sentence pairs in which both x_i and y_j occur;
(2) b denotes the number of aligned English-Chinese sentence pairs in which only y_j occurs;
(3) c denotes the number of aligned English-Chinese sentence pairs in which only x_i occurs;
(4) d denotes the number of aligned English-Chinese sentence pairs in which none of x_i and y_j occur.

It was found that letter case distinctions in English caused considerable "noise" by dispersing frequencies. For instance, "GREAT BRITAIN" and "Great Britain" are indexed as different items despite the fact that they are variants of a single MWU, dividing their common frequency between them. In order to avoid this problem, all lowercase was taken as the canonical form of English MWUs.

The MWU translations are identified as follows:

1) For each English MWU, all of the Chinese candidate MWUs are collected;
2) The Chinese candidates with *t*-scores lower than 1.65[8] are filtered out;
3) The candidates with MI lower than –0.2 are removed;
4) Finally, the surviving candidates are sorted by MI into descendent order.
5) The top one accepted as true Chinese translation of the given English MWU.
6) If no Chinese candidate survives the filtering, the given English MWU is ignored.

In the experiment, out of the 626 English nominal MWUs and 467 Chinese nominal MWUs extracted by the monolingual extraction process, the alignment algorithm extracted 191 potential English-Chinese MWU alignments. Figure 6 shows a sample of aligned MWU pairs. In the sample, the first set of square brackets encloses the frequency of the English MWU and the second set of square brackets contains the MI-score, co-occurrence frequency and frequency of the Chinese MWU.

```
1) [48] chinese_JJ culture_NN1:
   (1) [2.3339 ; 22; 44] 中华_NP0 文化_NN0
----------------------------
2) [8] chinese_JJ intellectual_NN1:
   (1) [0.9475 ; 6; 14] 文化_NN0 精英_NN0
   (2) [-0.5670 ; 3; 5] 华文_NN0 文化_NN0 精英_NN0
----------------------------
3) [7] chinese_JJ intellectual_NN1 and_CC cultural_JJ elite_NN1:
   (1) [1.1402 ; 6; 14] 文化_NN0 精英_NN0
   (2) [-0.3744 ; 3; 5] 华文_NN0 文化_NN0 精英_NN0
----------------------------
4) [10] chinese_JJ language_NN1 teachers_NN2:
   (1) [0.3455 ; 6; 17] 华文_NN0 教师_NN0
```

[8] A *t*-score threshold of 1.65 indicates the significance level of MI is above 95%.

```
----------------------------
5) [16] chinese_JJ singaporeans_NN2:
   (1) [0.9475 ; 6; 7] 华族_NN0 新加坡_NP0 人_NN0
   (2) [-1.5850 ; 4; 12] 新加坡_NP0 华人_NN0
   (3) [-1.7370 ; 6; 45] 新加坡_NP0 人_NN0
   (4) [-1.8301 ; 3; 6] 杨荣文_NP0 准将_NN0
----------------------------
6) [12] chinese_JJ studies_NN2:
   (1) [3.3339 ; 11; 11] 中文_NN0 学院_NN0
   (2) [0.3808 ; 5; 8] 南洋_NP0 理工_NN0
   (3) [0.3808 ; 5; 8] 南洋_NP0 理工_NN0 大学_NN0
   (4) [0.0589 ; 5; 10] 理工_NN0 大学_NN0
----------------------------
7) [4] cigarette_NN1 rolling_JJ tobacco_NN1 box_NN1:
   (1) [-0.5850 ; 2; 3] 卷烟_NN0 盒_NN0
----------------------------
8) [3] coffee_NN1 shops_NN2:
   (1) [0.8480 ; 3; 5] 小贩_NN0 中心_NN0
   (2) [-0.1699 ; 2; 3] 咖啡_NN0 店_NN0
----------------------------
9) [8] cold_JJ weather_NN1:
   (1) [0.1699 ; 3; 3] 极其_AV0 寒冷_AJ0 的_DE1 天气_NN0
   (2) [0.1699 ; 3; 3] 寒冷_AJ0 的_DE1 天气_NN0
   (3) [-1.5850 ; 2; 3] 寒冷_AJ0 天气_NN0
----------------------------
```

Figure 6: A sample of aligned English-Chinese nominal MWUs

As shown in figure 6, for most of the English MWUs more than one candidate survives the filtering. But three of them, "Chinese culture", "Chinese language teachers" and "cigarette rolling tobacco box" are precisely matched with unique candidates "中华 文化", "化文 教师" and "卷烟 盒"[9].

Due to multiple translations, some English MWUs may have more than one true translation in Chinese. For example in figure 6, "Chinese Singaporean" is translated as both "华族 新加坡 人" and "新加坡 华人". In this particular case, the English MWU and the two Chinese translations occurred for 16 , 7, and 12 times in the corpus. Of the Chinese candidates, "华族 新加坡 人" co-occurred with the English MWU 6 times (in the same aligned English-Chinese sentence pairs) while "新加坡 华人" co-occurred with the English MWU only four times. This shows that, in the corpus "Chinese Singaporean" is translated as "华族 新加坡 人" more often than "新加坡 华人". Their MI-scores, 0.9475 and -1.5850 reflect the situation accurately.

Considering that for a given nominal MWU more than one true alignment may exist, the process of identifying true and false alignments is somewhat complex. To represent this complexity, in the evaluation of the MWU alignment algorithm, two categories are used for 'correct' alignments, precise match and partial match. Precise match refers to an exact match between the source MWU (English in this case) and the top target MWU (Chinese in this case) in the candidate list. Partial match refers to cases in which alignments are approximate matches or where the correct translation is the second ranked item from the candidate list. For example in figure 6, Pair (1) is judged to be a pair of precise matches while pairs (2), (3) and (8) are judged to be partial matches. A manual examination revealed 99 precise matches, 55 partial matches and 37 mismatches among the total 191 potential MWU alignments. If precise matches only are taken into account, the precision of the technique is 51.83%; if both precise and partial matches are considered, the precision score increases to 80.63%. In both cases recall is significantly lower; recall is calculated by dividing the number of English MWUs with the number of resultant MWU alignments, giving a recall (100% × 191/626=) 30.51%.

4. Conclusion

Bilingual/multilingual MWU alignment is a challenging but worthwhile task. An efficient and effective system for MWU alignment will be of use to a number of areas including bilingual/multilingual

[9] All of these MWUs reflect major topics in the corpus.

contrastive studies, machine translation and multilingual lexicon building. Although much effort has been made in this area, no satisfactory solution has been found yet.

In this paper, we described a hybrid algorithm of English-Chinese MWU alignment which combines the n-gram approach, POS filters and co-occurrence coefficients. Given a sentence aligned English-Chinese parallel corpus, this algorithm automatically identifies and aligns nominal English and Chinese MWUs with high precision, but relatively low recall. As the result shows, it is a practical algorithm for extracting MWU alignments. So while a limited success has been achieved, it provides an inexpensive but practical tool for aligning nominal English-Chinese MWUs with a high degree of precision. Also, although not tested, the possibility exists that this algorithm could be ported to other language pairs by modifying the POS filters.

References:

Garside, Roger, Leech, Geoffrey and Sampson, Geoffrey 1987 *The Computational Analysis of English*, London, Longman.

McEnery Tony, Langé Jean-Marc, Oakes Michael, Véronis Jean 1997 The exploitation of multilingual annotated corpora for term extraction. In Garside Roger, Leech Geoffrey, McEnery Anthony (eds), *Corpus annotation --- linguistic information from computer text corpora*, London & New York, Longman, pp 220-230.

Piao Scott Songlin 2000 Sentence and word alignment between Chinese and English. Ph.D. thesis, Lancaster University.

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

Smadja Frank, McKeown Kathleen R., Hatzivassiloglou Vasileios 1996 Translating collocations for bilingual lexicons: a statistical approach. In *Computational Linguistics* 22(1): 1-38.

Wermter Stefan, Joseph Chen 1997 Cautious steps towards hybrid connectionist bilingual phrase alignment. In *Proceedings of International Conference Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, pp 364-368.

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

Zhang Min, Li Sheng 1997 Tagging Chinese corpus using statistics techniques and rule techniques. In *Proceedings of 1997 International Conference on Computer Processing of Oriental Languages (ICCPOL'97)*, Hong Kong, pp 503-506.

Appendix: Modified Zhang et al.'s Chinese tagset

No.	Part-of-Speech Classification (Eng)	Part-of-Speech Classification (Chi)	Original Tags	Modified Tags
1	Punctuation mark	标点符号	w	XX0
2	Idiom	成语	I	IDM
3	Positional Noun	处所词	s	NNC
4	Pronoun	代词	r	PN0
5	Verb	动词	v	VV0
6	Directional	方位词	f	NND
7	Non-linguistic Code	非语素词	x	ZZ0
8	Adverb	副词	d	AV0
9	Suffix	後接成分	k	SUF
10	Acronym	简称略语	j	ACN
11	Preposition	介词	p	PRP
12	Conjunction	连词	c	CJ0
13	Classifier	量词	q	NMW
14	Noun	名词	n	NN0
15	Prefix	前接成分	h	PRF
16	Determiner	区别词	b	PND
17	Temporal Noun	时间词	t	TIM
18	Numeral	数词	m	MC
19	Exclamatory word	叹词	e	ITJ
20	Onomatopoeic word	象声词	o	OP0
21	Idiomatic Expression	习用语	l	FIX
22	Adjective	形容词	a	AJ0
23	Auxiliary of Mood	语气词	y	AS0
24	Morpheme	语素	g	ELM
25	State word	*状态词 → 形容词	z	AJ0
26	Auxiliary Word	助词	u	AUX
Added POS Categories				
27	Proper Noun	专有名词		NP0
28	Modal Verb	情态动词		VM
29	Attribute Marker *de*	结构助词 "的"		DE1
30	Adverbial Marker *de*	结构助词 "地"		DE2
31	Complement Marker *de*	结构助词 "得"		DE3
32	Auxiliary *suo*	结构助词 "所"		SUO