International Journal of Knowledge and Language Processing
Volume **2**, Number 2, April **2011**

www.ijklp.org
AKLP International ©2011 ISSN 2191-2734
pp. **14-22**

Extracting Multiword Translation Equivalents Using Hybrid Strategy

Shasha Zhu[1], Jingsong Yu[1]
[1] School of Software and Microelectronics
Peking University
Beijing, 100871, China
zssjane@126.com; yjs@ss.pku.edu.cn

Received April 2011; revised April 2011

ABSTRACT. *Automatic extraction of multiword translation equivalents is an important task in the domain of natural language processing. This paper proposes a hybrid strategy for extracting multiword translation equivalents in parallel corpus. The strategy consists of using log likelihood and point-wise mutual information technique to generate the bigram/trigram model, extending multiword coverage by integrating the dependency parsing technique into the similarity approach and mining multiword translation equivalents based on Google Translate engine. Experimental results show that the precision and recall of this hybrid strategy are improved.*
Keywords: Multiword translation equivalent, hybrid strategy, log likelihood, point-wise mutual information, semantic similarity

1. **Introduction.** Automatic extraction of multiword translation equivalents is an important task in the domain of natural language processing. Previously, most researchers lay their emphasis on extracting monolingual multiword expressions instead of bilingual ones. Smadja claimed that multiword expressions were "recurrent combinations of words that co-occur more often than expected by chance" [1]. Sag et al. defined the multiword expressions as "idiosyncratic interpretations that cross word boundaries (or spaces)" [7]. Besides, Biber described multiword expressions as "lexical bundles" [2]. Different from monolingual multiword expressions, multiword translation equivalents have both source multiword expressions and target multiword expressions in two languages. Between them exists a form of strong bounded translation-ship. Rayson regarded that multiword expressions play a critical role in terminology extraction, machine translation, text summarization and so on [9, 10]. Likely, automatic extraction of multiword translation equivalents is of great importance in the corpus research, language teaching, translation and other applications. So far, however, efficient extraction of multiword translation equivalents still remains an unsolved issue. To resolve this problem, this paper proposes a hybrid strategy for extracting multiword translation equivalents, which incorporates the statistic tool, similarity-based extension technique into aligned translated equivalents.

The remainder of this paper is organized as follows. Section 2 describes the related work

in extracting multiword translation equivalents. Section 3 introduces our hybrid strategy. Section 4 presents experiments of this strategy. We conclude in section 5.

2. **Related Work.** Piao and McEnery employed mutual information and t-test approaches to extract multiword translation equivalents from parallel corpus [6, 9, 10]. They started from extracting multiword expressions from Chinese and English corpus separately and then align the Chinese multiword expressions with the English ones. Their multiword expressions cover 2-6 English words. Additionally, they tried to seek the seed bigram and trigram and obtain more multiword expressions with these seeds. In the alignment phase, they aligned and ranked the Chinese-English translation equivalents and reserved those having a high ranking score. Helena et al. presented an alignment-based approach for extracting multiword translation equivalents [4]. In the English-Portugal parallel corpus, they measured the affinity between target language words by means of point-wise mutual information and mutual information techniques. Our research shares a little similarity with theirs, but we differ from them in that we use log likelihood and point-wise mutual information to compute word affinity, integrate dependency parse technique into similarity approach to enlarge multiword coverage and use translate engine to get the multiword translation equivalents for the extended multiword expressions.

Tanaka and Baldwin took dictionary- and template-based translation technique to extract noun-noun translation equivalents [12]. They extracted translation equivalents from bilingual dictionary. They called this method as memory based machine translation because this method required the ALTDIC and EDICT dictionaries. Their template-based approach first extracted multiword expressions from the source language and then used the translation templates to convert the source multiword expression into the target one. Chang extracted translation equivalents by using various techniques, including log likelihood, point-wise mutual information, DICE to calculate affinities between translation equivalents [1]. Du and Chen proposed average affinity and normalized affinity to mine bilingual translation equivalents [5].

This paper presents a departure from the previous researches. We proposed a hybrid strategy which employs log likelihood and point-wise mutual information techniques to measure word affinities, integrates dependency parsing technique with the similarity measure to extend multiword coverage and finally translates these extended multiword expressions by means of Google engine. Experimental results show that this hybrid strategy can extract translation equivalents with high precision and recall, including both high-frequency and low-frequency ones.

3. **Hybrid Strategy.**

3.1. Generate bigram/trigram model using log likelihood and PMI. Prior to generating bigram/trigram model, we use the GIZA++ tool to get candidate translation equivalents [3]. Hereafter, we take the log likelihood and point-wise mutual information techniques to measure the affinity between words in the bigram/trigram and then filter the candidate translation equivalents according to the word affinities, thus improving the accuracy of

bigram/trigram translation equivalents. Take bigram "chinese government" for example, its contingency table is as follows:

TABLE 1. Bigram contingency table

	government	~government	
chinese	C_{11}	C_{12}	C_{1p}
~chinese	C_{21}	C_{22}	C_{2p}
	C_{p1}	C_{p2}	C_{pp}

In table 1, C_{11} is the number of times words Chinese and government co-occur, C_{12} denotes the number of times word Chinese occurs without government being the second word, C_{21} is the number of times word government occurs without word Chinese, and C_{22} refers to the number of times words Chinese and government do not occur, then we calculate the expectation of C_{11} and log likelihood using the formula described by Banerjee and Pedersen [8]:

$$E_{11} = \frac{C_{p1} * C_{1p}}{C_{pp}} \tag{1}$$

Where $C_{p1} = C_{11} + C_{21}$, $C_{1p} = C_{11} + C_{12}$, and $C_{pp} = C_{11} + C_{12} + C_{21} + C_{22}$.

With these expectations, we can calculate log likelihood and PMI by the following formula:

$$LL = 2\left[C_{11} * \log\left(C_{11} / E_{11} \right) + C_{12} * \log\left(C_{12} / E_{12} \right) + C_{21} * \log\left(C_{21} / E_{21} \right) + C_{22} * \log\left(C_{22} / E_{22} \right) \right] \tag{2}$$

$$PMI = \log \frac{C_{11}}{E_{11}} \tag{3}$$

Equations 2 and 3 will calculate word affinities in bigram. Higher log likelihood and PMI values produce high-quality multiword expressions.

As for the trigram human right cause, its contingency table can be:

TABLE 2. Trigram contingency table

		cause	~cause
human	right	C_{111}	C_{112}
human	~right	C_{121}	C_{122}
~human	right	C_{211}	C_{212}
~human	~right	C_{221}	C_{222}

The log likelihood of this trigram is:

$$LL = 2\left[\begin{array}{l} C_{111} * \log\left(C_{111} / E_{111} \right) + C_{112} * \log\left(C_{122} / E_{122} \right) + C_{221} * \log\left(C_{221} / E_{221} \right) + C_{222} * \log\left(C_{222} / E_{222} \right) + \\ C_{211} * \log\left(C_{121} / E_{211} \right) + C_{212} * \log\left(C_{212} / E_{212} \right) + C_{221} * \log\left(C_{221} / E_{221} \right) + C_{222} * \log\left(C_{222} / E_{222} \right) \end{array} \right] \tag{4}$$

where E is the expectation value and calculated in the same way as for the bigram.
Point-wise mutual information is calculated as follows;

$$PMI = \log \frac{C_{111}}{E_{111}} \tag{5}$$

With these formulas, we will get the statistical information between words in bigram and trigram and filter the candidate bilingual translation equivalents using statistical information. However, statistical tool has a better performance in high-frequency multiword expressions than in low-frequency ones. To remedy this, we extend multiword coverage using similarity-based approach.

3.2 Extend multiword coverage based on similarity. To resolve the low-frequency multiword expressions, this paper integrates the dependency parsing technique into similarity approach. We use the multiword expressions extracted in section 3.1 as seeds, utilize the Stanford dependency parser to generate the dependency pairs and extracts the head words in these dependency pairs. Stanford dependency parser[1] is employed to get the head words of the seeds and those of candidate multiword expressions. Afterwards, similarities between head words of the seeds and those of candidates are calculated. For example, seed multiword expression like "human right system" has the word "system" as its head, and candidate multiword expressions like "spring bud program, medical insurance system, harmonious socialist society, human right theory" have words "program, system, society and theory" as their head. Then we measure similarities between seed head and candidate heads by means of JC[2] algorithm.

TABLE 3. Similarity between head words

Candidate Multiword	Seed Multiword
	human right system
spring bud program	system : program (0.52495)
medical insurance system	system : system (5.15)
harmonious socialist society	system : society (0.1069)
human right theory	system : theory (0.13719)

By doing so, multiword coverage can be enlarged based on the similarities between head words, thus resolving the problem of statistical tool's inedibility in finding low-frequency multiword expressions.

3.3 Extract translation equivalents using online translation. Multiword coverage is extended in section 3.2. Its corresponding Chinese translations, however, are not provided in the extended set. Hence, we take the extended English multiword expressions as translation sources, and send them to Google translation engine for Chinese translations. We then compare the difference between Chinese translations returned by Google

[1] http://nlp.stanford.edu/software/lex-parser.shtml
[2] http://www.stanford.edu/class/cs224u/lec/224u.10.lec2.pdf

translation engine and those in the answer set by means of the edit distance. Table 4 gives the Google translation results and compares them with the standard answers using edit distance.

TABLE 4. Google translation and edit distance

English Multiword	Google Translation	Standard Answer	Edit Distance
agricultural economy	农业经济	农业经济	0
educational system	教育系统	教育制度	2
guarantee system	保证体系	保障体系	1
human right issue	人权问题	人权问题	0
national academic body	全国性学术团体	全国性学术团体	0

Table 4 shows that if Google translation is found in the standard answer set, then the edit distance between them is zero. Experiments show that most of the translation results are covered by the answer set, thus contributing to the performance of our hybrid strategy.

4. **Experiments.** The Chinese-English parallel corpus of Chinese government white papers from 1991 to 2010 is used as our dataset. We randomly select 500 Chinese English sentence pairs. We then segment Chinese sentences tokenize and lemmatize English sentences. Precision, recall and F-score are taken to evaluate the performance of our algorithm. We initially extract candidate bigram and trigram translation equivalents using GIZA++ tool. For example,

international community ||| 国际 社会 ||| 0-0 1-0 1-1
Hong Kong ||| 香港 ||| 0-0 1-0
the Chinese Government ||| 中国 政府 ||| 0-0 1-0 1-1 2-1
political consultative conference ||| 政治 协调 会议 ||| 0-0 1-1 2-2

In the above example, 0-0 1-0 1-1 denotes the word indexes in the source and target language. The bigram/trigram translation equivalents extracted by the GIZA++ tool is used as the baseline for comparison. Our hybrid strategy bases itself on the GIZA++ extraction results. We then use the log likelihood and point-wise mutual information approaches to measure word affinities of English multiword expressions, integrate dependency parsing technique into similarity approach for extending multiword coverage and finally extract Chinese translation equivalents for the extended multiword expressions.

Our experiment begins with using GIZA++ tool, continues to add log likelihood measure, similarity-based extension and Google translate technique. Table 5 shows the performance of our system for bigram translation equivalents.

TABLE 5. Performance for bigram translation equivalent (1)

Bigram	Precision	Recall	F-Score
GIZA++	35.24%	33.94%	34.58%
+ Log likelihood	43.21%	32.11%	36.84%
+ Similarity	57.41%	56.88%	57.14%
+ Google Translate	56.81%	55.50%	56.15%

Table 5 shows that the performance is improved by incorporating more techniques. When the log likelihood method is employed, our system's recall exhibits a marginal decrease but its precision increases nearly by 8%. A big improvement is achieved when using similarity extension approach which is based on the dependency parsing results. After Google Translate technique is added, our system removes those translation equivalents that do not exist in the parallel corpus. Therefore, some English multiword expressions are in the answer set, but their corresponding Chinese translations are missing in the answer set. This causes the decrease in both precision and recall.

Hereafter, we continue our experiment by replacing log likelihood with point-wise mutual information for bigram translation equivalents. Table 6 shows the experiment results:

TABLE 6. Performance for bigram translation equivalent (2)

Bigram	Precision	Recall	F-Score
GIZA++	35.24%	33.94%	34.58%
+ PMI	43.48%	32.11%	36.94%
+ Similarity	57.67%	56.88%	57.27%
+ Google Translate	57.08%	55.50%	56.28%

Table 6 shows that PMI performs in the same way as the log likelihood does. To simplify the comparison, we call GIZA++, log likelihood, similarity plus Google translation engine as hybrid strategy one and GIZA++, point-wise mutual information, similarity plus Google translation engine as hybrid strategy two. Figure 1 shows its comparison with the baseline system for bigram translation equivalents.

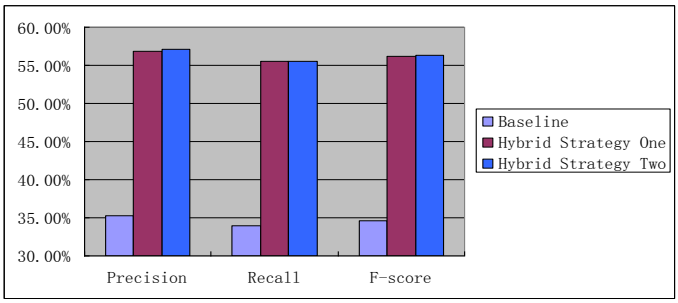

FIGURE 1. Comparison with baseline for bigram translation equivalent

For trigram translation equivalents, we experiment with adding more techniques and

obtain the results as follows:

TABLE 7. Performance for trigram translation equivalent

Trigram	Precision	Recall	F-Score
GIZA++	31.58%	16.51%	21.69%
+ Log likelihood	36.73%	16.51%	22.78%
+ Similarity	42.59%	21.10%	28.22%
+ Google Translate	41.51%	20.18%	27.16%

Trigram	Precision	Recall	F-Score
GIZA++	31.58%	16.51%	21.69%
+ PMI	39.02%	14.68%	21.33%
+ Similarity	44.44%	18.35%	25.97%
+ Google Translate	43.18%	17.43%	24.84%

Table 7 demonstrates that the two hybrid strategies show a good performance in trigram translation equivalents as well. We compare them with the baseline in figure 2:

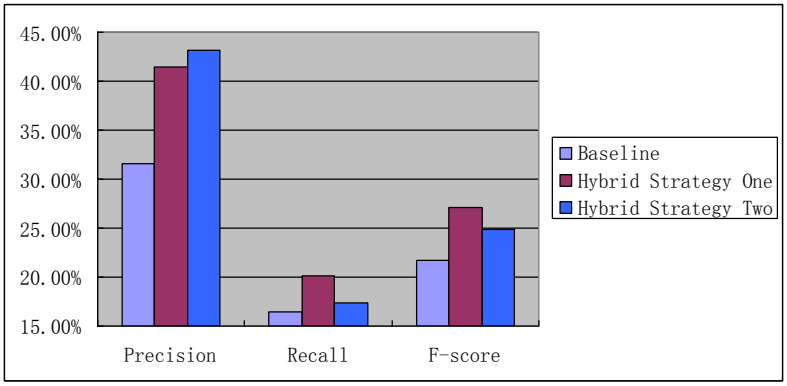

FIGURE 2. Comparison with baseline for trigram translation equivalent

We also compared our results with previous researches. Because of the differences in datasets, we can only compare our algorithm with log likelihood and chi-square mostly taken by previous researchers. These approaches first extracts multiword expressions of each language respectively and align them afterwards. Different from them, we first align bilingual translation equivalents with GIZA++ tool and then conduct a hybrid strategy to extract qualified Chinese-English translation equivalents. Figure 3 and figure 4 show the comparison of our algorithm with other approaches for bigram and trigram translation equivalents.

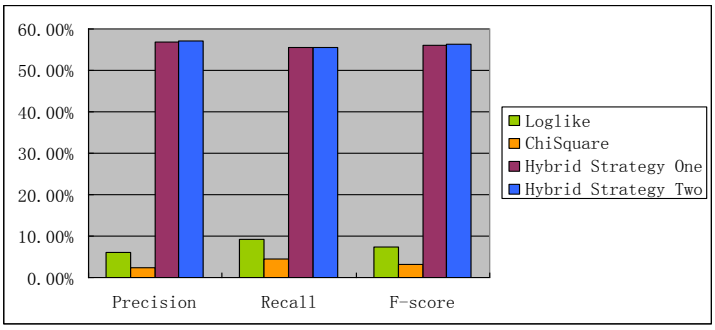

FIGURE 3. Comparison with other approaches for bigram translation equivalent

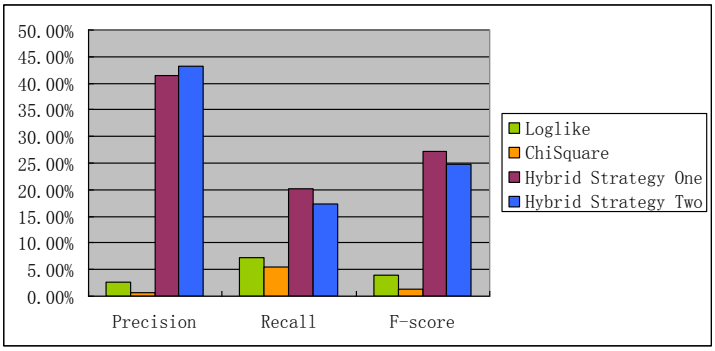

FIGURE 4. Comparison with other approaches for trigram translation equivalent

Figure 3 and figure 4 show that our hybrid strategies perform better than other approaches both in precision and recall.

5. **Conclusions.** We presented a hybrid strategy for accurately extracting multiword translation equivalents from the parallel corpus. It achieves high precision and recall compared with other approaches taken by most of the researchers. It makes use of GIZA++ capability in high precision of extracting candidate Chinese-English translation equivalents, and employs statistic approaches to find tightly collocated words, thus reducing the noises in the candidate translation equivalents. Moreover, this strategy utilizes the similarity-based approach to extend multiword coverage. This complements statistic tool's inability in extracting low-frequency multiword expression. From the experiments, we find that it is important to have well-aligned candidate translation equivalents. However, this still poses a big challenge and we will try to optimize it with some other techniques.

Acknowledgment. We would like to thank Prof. Chang and Yu for their continuous supports and everlasting encouragements. We also would like to thank our anonymous reviewers for their constructive suggestions which helped improve our paper.

REFERENCES

[1] Baobao Chang, Pernilla Danielsson, Wolfgang Teubert. "Extraction of Translation Unit from Chinese-English Parallel Corpora." *Proceedings of the first SIGHAN Workshop on Chinese Language Processing*. 2002. 1–5.

[2] Biber, Douglas, Susan Conrad and Viviana Cortes. "Lexical Bundles in Speech and Writing: an Initial Taxonomy." *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*. Frankfurt: 2003. 71-92.

[3] Franz Josef Och, Hermann Ney. "Improved Statistical Alignment Models." *Proceedings of the 38th Annual Meeting of the ACL*. Hong Kong: 2000. 440–447.

[4] Helena de Medeiros Caseli, Aline Villavicencio, Andr´e Machado, Maria Jos´e Finatto. "Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains." *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications*. 2009. 1-8.

[5] Limin Du, Boxing Cheng. "Automatic Extraction of Multiword Translation Equivalents from Bilingual Corpus." Intellectual Property Press: Beijing. 2005.

[6] Piao, S., McEnery, T. “Multi-word Unit Alignment in English‑Chinese Parallel Corpora.” *Proceedings of the Corpus Linguistics*. 2001. 466‑475.

[7] Sag, I., Baldwin, T., Bond, F., Copestake, A., Dan, F. “Multiword Expressions: a Pain in the Neck for NLP.” LinGO Working Paper No. 2001-03, Stanford University, CA.

[8] Satanjeev Banerjee, Ted Pedersen. “The Design, Implementation and Use of the Ngram Statistics Package.” *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. 2003. 370–381.

[9] Scott S. L. Piao, Guangfan Sun, Paul Rayson, Qi Yuan. “Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool.” *Proceedings of the Workshop on Multi-word expressions in a Multilingual Context*. 2006. 17–24.

[10] Scott Songlin Piao, Paul Rayson, Dawn Archer, Tony McEnery. “Comparing and Combining a Semantic Tagger and a Statistical Tool for MWE Extraction.” *Computer Speech and Language*. 19.4(2005): 378‑397.

[11] Smadja, F. “Retrieving Collocations from Text: Xtract.” *Computational Linguistics*. 19.1(1993): 143-177.

[12] Takaaki Tanaka, Timothy Baldwin. “Noun-Noun Compound Machine Translation: a Feasibility Study on Shallow Processing.” *Proceedings of the ACL 2003 workshop on Multiword expressions*. 2003. 17–24.