Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions

Zhixiang Ren[1] **Yajuan Lü**[1] **Jie Cao**[1] **Qun Liu**[1] **Yun Huang**[2]

[1]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{renzhixiang,lvyajuan
caojie,liuqun}@ict.ac.cn

[2]Department of Computer Science
School of Computing
National University of Singapore
Computing 1, Law Link, Singapore 117590
huangyun@comp.nus.edu.sg

Abstract

Multiword expressions (MWEs) have been proved useful for many natural language processing tasks. However, how to use them to improve performance of statistical machine translation (SMT) is not well studied. This paper presents a simple yet effective strategy to extract domain bilingual multiword expressions. In addition, we implement three methods to integrate bilingual MWEs to Moses, the state-of-the-art phrase-based machine translation system. Experiments show that bilingual MWEs could improve translation performance significantly.

1 Introduction

Phrase-based machine translation model has been proved a great improvement over the initial word-based approaches (Brown et al., 1993). Recent syntax-based models perform even better than phrase-based models. However, when syntax-based models are applied to new domain with few syntax-annotated corpus, the translation performance would decrease. To utilize the robustness of phrases and make up the lack of syntax or semantic information in phrase-based model for domain translation, we study domain bilingual multiword expressions and integrate them to the existing phrase-based model.

A *multiword expression (MWE)* can be considered as word sequence with relatively fixed structure representing special meanings. There is no uniform definition of MWE, and many researchers give different properties of MWE. Sag et al. (2002) roughly defined MWE as "idiosyncratic interpretations that cross word boundaries (or spaces)". Cruys and Moirón (2007) focused on the non-compositional property of MWE, i.e. the property that whole expression cannot be derived from their component words. Stanford university launched a MWE project[1], in which different qualities of MWE were presented. For *bilingual multiword expression (BiMWE)*, we define a bilingual phrase as a bilingual MWE if (1) the source phrase is a MWE in source language; (2) the source phrase and the target phrase must be translated to each other exactly, i.e. there is no additional (boundary) word in target phrase which cannot find the corresponding word in source phrase, and vice versa. In recent years, many useful methods have been proposed to extract MWEs or BiMWEs automatically (Piao et al., 2005; Bannard, 2007; Fazly and Stevenson, 2006). Since MWE usually constrains possible senses of a polysemous word in context, they can be used in many NLP applications such as information retrieval, question answering, word sense disambiguation and so on.

For machine translation, Piao et al. (2005) have noted that the issue of MWE identification and accurate interpretation from source to target language remained an unsolved problem for existing MT systems. This problem is more severe when MT systems are used to translate domain-specific texts, since they may include technical terminology as well as more general fixed expressions and idioms. Although some MT systems may employ a machine-readable bilingual dictionary of MWE, it is time-consuming and inefficient to obtain this resource manually. Therefore, some researchers have tried to use automatically extracted bilingual MWEs in SMT. Tanaka and Baldwin (2003) described an approach of noun-noun compound machine translation, but no significant comparison was presented. Lambert and Banchs (2005) presented a method in which bilingual MWEs were used to modify the word alignment so as to improve the SMT quality. In their work, a bilingual MWE in training corpus was grouped as

[1]http://mwe.stanford.edu/

Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pages 47–54,
Suntec, Singapore, 6 August 2009.

one unique token before training alignment models. They reported that both alignment quality and translation accuracy were improved on a small corpus. However, in their further study, they reported even lower BLEU scores after grouping MWEs according to part-of-speech on a large corpus (Lambert and Banchs, 2006). Nonetheless, since MWE represents liguistic knowledge, the role and usefulness of MWE in full-scale SMT is intuitively positive. The difficulty lies in how to integrate bilingual MWEs into existing SMT system to improve SMT performance, especially when translating domain texts.

In this paper, we implement three methods that integrate domain bilingual MWEs into a phrase-based SMT system, and show that these approaches improve translation quality significantly. The main difference between our methods and Lambert and Banchs' work is that we directly aim at improving the SMT performance rather than improving the word alignment quality. In detail, differences are listed as follows:

- Instead of using the bilingual n-gram translation model, we choose the phrase-based SMT system, Moses[2], which achieves significantly better translation performance than many other SMT systems and is a state-of-the-art SMT system.

- Instead of improving translation indirectly by improving the word alignment quality, we directly target at the quality of translation. Some researchers have argued that large gains of alignment performance under many metrics only led to small gains in translation performance (Ayan and Dorr, 2006; Fraser and Marcu, 2007).

Besides the above differences, there are some advantages of our approaches:

- In our method, automatically extracted MWEs are used as additional resources rather than as phrase-table filter. Since bilingual MWEs are extracted according to noisy automatic word alignment, errors in word alignment would further propagate to the SMT and hurt SMT performance.

- We conduct experiments on domain-specific corpus. For one thing, domain-specific corpus potentially includes a large number of technical terminologies as well as more general fixed expressions and idioms, i.e. domain-specific corpus has high MWE coverage. For another, after the investigation, current SMT system could not effectively deal with these domain-specific MWEs especially for Chinese, since these MWEs are more flexible and concise. Take the Chinese term "*软 坚 散 结*" for example. The meaning of this term is "soften hard mass and dispel pathogenic accumulation". Every word of this term represents a special meaning and cannot be understood literally or without this context. These terms are difficult to be translated even for humans, let alone machine translation. So, treating these terms as MWEs and applying them in SMT system have practical significance.

- In our approach, no additional corpus is introduced. We attempt to extract useful MWEs from the training corpus and adopt suitable methods to apply them. Thus, it benefits for the full exploitation of available resources without increasing great time and space complexities of SMT system.

The remainder of the paper is organized as follows. Section 2 describes the bilingual MWE extraction technique. Section 3 proposes three methods to apply bilingual MWEs in SMT system. Section 4 presents the experimental results. Section 5 draws conclusions and describes the future work. Since this paper mainly focuses on the application of BiMWE in SMT, we only give a brief introduction on monolingual and bilingual MWE extraction.

2 Bilingual Multiword Expression Extraction

In this section we describe our approach of bilingual MWE extraction. In the first step, we obtain monolingual MWEs from the Chinese part of parallel corpus. After that, we look for the translation of the extracted MWEs from parallel corpus.

2.1 Automatic Extraction of MWEs

In the past two decades, many different approaches on automatic MWE identification were reported. In general, those approaches can be classified into three main trends: (1) statistical approaches (Pantel and Lin, 2001; Piao et

[2]http://www.statmt.org/moses/

al., 2005), (2) syntactic approaches (Fazly and Stevenson, 2006; Bannard, 2007), and (3) semantic approaches (Baldwin et al., 2003; Cruys and Moirón, 2007). Syntax-based and semantic-based methods achieve high precision, but syntax or semantic analysis has to be introduced as preparing step, so it is difficult to apply them to domains with few syntactical or semantic annotation. Statistical approaches only consider frequency information, so they can be used to obtain MWEs from bilingual corpora without deeper syntactic or semantic analysis. Most statistical measures only take two words into account, so it not easy to extract MWEs containing three or more than three words.

Log Likelihood Ratio (LLR) has been proved a good statistical measurement of the association of two random variables (Chang et al., 2002). We adopt the idea of statistical approaches, and propose a new algorithm named LLR-based Hierarchical Reducing Algorithm (HRA for short) to extract MWEs with arbitrary lengths. To illustrate our algorithm, firstly we define some useful items. In the following definitions, we assume the given sentence is "*A B C D E*".

Definition 1 *Unit:* A unit is any sub-string of the given sentence. For example, "*A B*", "*C*", "*C D E*" are all units, but "*A B D*" is not a unit.

Definition 2 *List:* A list is an ordered sequence of units which exactly cover the given sentence. For example, {"*A*","*B C D*","*E*"} forms a list.

Definition 3 *Score:* The score function only defines on two adjacent units and return the LLR between the last word of first unit and the first word of the second unit[3]. For example, the score of adjacent unit "*B C*" and "*D E*" is defined as LLR("*C*","*D*").

Definition 4 *Select:* The selecting operator is to find the two adjacent units with maximum score in a list.

Definition 5 *Reduce:* The reducing operator is to remove two specific adjacent units, concatenate them, and put back the result unit to the removed position. For example, if we want to reduce unit "*B C*" and unit "*D*" in list {"*A*","*B C*","*D*","*E*"}, we will get the list {"*A*","*B C D*","*E*"}.

Initially, every word in the sentence is considered as one unit and all these units form a initial list L. If the sentence is of length N, then the list contains N units, of course. The final set of MWEs, S, is initialized to empty set. After initialization, the algorithm will enter an iterating loop with two steps: (1) select the two adjacent units with maximum score in L, naming U_1 and U_2; and (2) reduce U_1 and U_2 in L, and insert the reducing result into the final set S. Our algorithm terminates on two conditions: (1) if the maximum score after selection is less than a given threshold; or (2) if L contains only one unit.

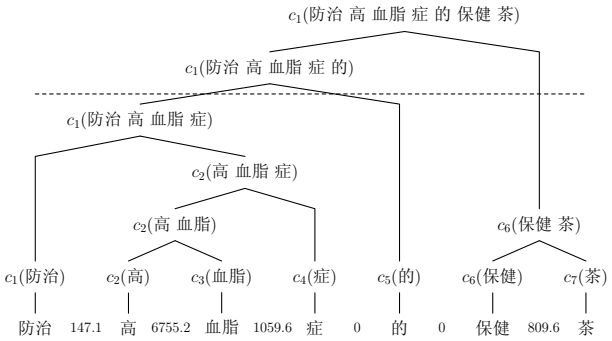

Figure 1: Example of Hierarchical Reducing Algorithm

Let us make the algorithm clearer with an example. Assume the threshold of score is 20, the given sentence is "*防治 高 血脂 症 的 保健 茶*"[4]. Figure 1 shows the hierarchical structure of given sentence (based on LLR of adjacent words). In this example, four MWEs ("*高 血脂*", "*高 血脂 症*", "*保健 茶*", "*防治 高 血脂 症*") are extracted in the order, and sub-strings over dotted line in figure 1 are not extracted.

From the above example, we can see that the extracted MWEs correspond to human intuition. In general, the basic idea of HRA is to reflect the hierarchical structure pattern of natural language. Furthermore, in the HRA, MWEs are measured with the minimum LLR of adjacent words in them, which gives lexical confidence of extracted MWEs. Finally, suppose given sentence has length N, HRA would definitely terminate within $N-1$ iterations, which is very efficient.

However, HRA has a problem that it would extract substrings before extracting the whole string, even if the substrings only appear in the particular whole string, which we consider useless. To solve this problem, we use contextual features,

[3]we use a stoplist to eliminate the units containing function words by setting their score to 0

[4]The whole sentence means "healthy tea for preventing hyperlipidemia", and we give the meaning for each Chinese word: 防治(preventing), 高(hyper-), 血脂(-lipid-), 症(-emia), 的(for), 保健(healthy), 茶(tea).

contextual entropy (Luo and Sun, 2003) and C-value (Frantzi and Ananiadou, 1996), to filter out those substrings which exist only in few MWEs.

2.2 Automatic Extraction of MWE's Translation

In subsection 2.1, we described the algorithm to obtain MWEs, and we would like to introduce the procedure to find their translations from parallel corpus in this subsection.

For mining the English translations of Chinese MWEs, we first obtain the candidate translations of a given MWE from the parallel corpus. Steps are listed as follows:

1. Run GIZA++[5] to align words in the training parallel corpus.

2. For a given MWE, find the bilingual sentence pairs where the source language sentences include the MWE.

3. Extract the candidate translations of the MWE from the above sentence pairs according to the algorithm described by Och (2002).

After the above procedure, we have already extracted all possible candidate translations of a given MWE. The next step is to distinguish right candidates from wrong candidates. We construct perceptron-based classification model (Collins, 2002) to solve the problem. We design two groups of features: translation features, which describe the mutual translating chance between source phrase and target phrase, and the language features, which refer to how well a candidate is a reasonable translation. The translation features include: (1) the logarithm of source-target translation probability; (2) the logarithm of target-source translation probability; (3) the logarithm of source-target lexical weighting; (4) the logarithm of target-source lexical weighting; and (5) the logarithm of the phrase pair's LLR (Dunning, 1993). The first four features are exactly the same as the four translation probabilities used in traditional phrase-based system (Koehn et al., 2003). The language features include: (1) the left entropy of the target phrase (Luo and Sun, 2003); (2) the right entropy of the target phrase; (3) the first word of the target phrase; (4) the last word of the target phrase; and (5) all words in the target phrase.

[5]http://www.fjoch.com/GIZA++.html

We select and annotate 33000 phrase pairs randomly, of which 30000 pairs are used as training set and 3000 pairs are used as test set. We use the perceptron training algorithm to train the model. As the experiments reveal, the classification precision of this model is 91.67%.

3 Application of Bilingual MWEs

Intuitively, bilingual MWE is useful to improve the performance of SMT. However, as we described in section 1, it still needs further research on how to integrate bilingual MWEs into SMT system. In this section, we propose three methods to utilize bilingual MWEs, and we will compare their performance in section 4.

3.1 Model Retraining with Bilingual MWEs

Bilingual phrase table is very important for phrase-based MT system. However, due to the errors in automatic word alignment and unaligned word extension in phrase extraction (Och, 2002), many meaningless phrases would be extracted, which results in inaccuracy of phrase probability estimation. To alleviate this problem, we take the automatically extracted bilingual MWEs as parallel sentence pairs, add them into the training corpus, and retrain the model using GIZA++. By increasing the occurrences of bilingual MWEs, which are good phrases, we expect that the alignment would be modified and the probability estimation would be more reasonable. Wu et al. (2008) also used this method to perform domain adaption for SMT. Different from their approach, in which bilingual MWEs are extracted from additional corpus, we extract bilingual MWEs from the original training set. The fact that additional resources can improve the domain-specific SMT performance was proved by many researchers (Wu et al., 2008; Eck et al., 2004). However, our method shows that making better use of the resources in hand could also enhance the quality of SMT system. We use "Baseline+BiMWE" to represent this method.

3.2 New Feature for Bilingual MWEs

Lopez and Resnik (2006) once pointed out that better feature mining can lead to substantial gain in translation quality. Inspired by this idea, we append one feature into bilingual phrase table to indicate that whether a bilingual phrase contains bilingual MWEs. In other words, if the source language phrase contains a MWE (as substring) and

the target language phrase contains the translation of the MWE (as substring), the feature value is 1, otherwise the feature value is set to 0. Due to the high reliability of bilingual MWEs, we expect that this feature could help SMT system to select better and reasonable phrase pairs during translation. We use "Baseline+Feat" to represent this method.

3.3 Additional Phrase Table of bilingual MWEs

Wu et al. (2008) proposed a method to construct a phrase table by a manually-made translation dictionary. Instead of manually constructing translation dictionary, we construct an additional phrase table containing automatically extracted bilingual MWEs. As to probability assignment, we just assign 1 to the four translation probabilities for simplicity. Since Moses supports multiple bilingual phrase tables, we combine the original phrase table and new constructed bilingual MWE table. For each phrase in input sentence during translation, the decoder would search all candidate translation phrases in both phrase tables. We use "Baseline+NewBP" to represent this method.

4 Experiments

4.1 Data

We run experiments on two domain-specific patent corpora: one is for traditional medicine domain, and the other is for chemical industry domain. Our translation tasks are Chinese-to-English.

In the traditional medicine domain, table 1 shows the data statistics. For language model, we use SRI Language Modeling Toolkit[6] to train a trigram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998) on the target side of training corpus. Using our bilingual MWE extracting algorithm, 80287 bilingual MWEs are extracted from the training set.

	Chinese	English
Training Sentences	120,355	
Words	4,688,873	4,737,843
Dev Sentences	1,000	
Words	31,722	32,390
Test Sentences	1,000	
Words	41,643	40,551

Table 1: Traditional medicine corpus

[6]http://www.speech.sri.com/projects/srilm/

In the chemical industry domain, table 2 gives the detail information of the data. In this experiment, 59466 bilingual MWEs are extracted.

	Chinese	English
Training Sentences	120,856	
Words	4,532,503	4,311,682
Dev Sentences	1,099	
Words	42,122	40,521
Test Sentences	1,099	
Words	41,069	39,210

Table 2: Chemical industry corpus

We test translation quality on test set and use the open source tool mteval-vllb.pl[7] to calculate case-sensitive BLEU 4 score (Papineni et al., 2002) as our evaluation criteria. For this evaluation, there is only one reference per test sentence. We also perform statistical significant test between two translation results (Collins et al., 2005). The mean of all scores and relative standard deviation are calculated with a 99% confidence interval of the mean.

4.2 MT Systems

We use the state-of-the-art phrase-based SMT system, Moses, as our baseline system. The features used in baseline system include: (1) four translation probability features; (2) one language model feature; (3) distance-based and lexicalized distortion model feature; (4) word penalty; (5) phrase penalty. For "Baseline+BiMWE" method, bilingual MWEs are added into training corpus, as a result, new alignment and new phrase table are obtained. For "Baseline+Feat" method, one additional 0/1 feature are introduced to each entry in phrase table. For "Baseline+NewBP", additional phrase table constructed by bilingual MWEs is used.

Features are combined in the log-linear model. To obtain the best translation $\hat{e}$ of the source sentence f, log-linear model uses following equation:

$$\begin{aligned} \hat{e} &= \arg\max_{e} p(e|f) \\ &= \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e,f) \end{aligned} \quad (1)$$

in which h_m and λ_m denote the mth feature and weight. The weights are automatically turned by minimum error rate training (Och, 2002) on development set.

[7]http://www.nist.gov/speech/tests/mt/resources/scoring.htm

4.3 Results

Methods	BLEU
Baseline	0.2658
Baseline+BiMWE	0.2661
Baseline+Feat	0.2675
Baseline+NewBP	0.2719

Table 3: Translation results of using bilingual MWEs in traditional medicine domain

Table 3 gives our experiment results. From this table, we can see that, bilingual MWEs improve translation quality in all cases. The Baseline+NewBP method achieves the most improvement of 0.61% BLEU score compared with the baseline system. The Baseline+Feat method comes next with 0.17% BLEU score improvement. And the Baseline+BiMWE achieves slightly higher translation quality than the baseline system.

To our disappointment, however, none of these improvements are statistical significant. We manually examine the extracted bilingual MWEs which are labeled positive by perceptron algorithm and find that although the classification precision is high (91.67%), the proportion of positive example is relatively lower (76.69%). The low positive proportion means that many negative instances have been wrongly classified to positive, which introduce noises. To remove noisy bilingual MWEs, we use the length ratio x of the source phrase over the target phrase to rank the bilingual MWEs labeled positive. Assume x follows Gaussian distributions, then the ranking score of phrase pair (s,t) is defined as the following formula:

$$Score(s,t) = \log(LLR(s,t)) \times \frac{1}{\sqrt{2\pi}\sigma} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Here the mean μ and variance σ^2 are estimated from the training set. After ranking by score, we select the top 50000, 60000 and 70000 bilingual MWEs to perform the three methods mentioned in section 3. The results are showed in table 4.

From this table, we can conclude that: (1) All the three methods on all settings improve BLEU score; (2) Except the Baseline+BiMWE method, the other two methods obtain significant improvement of BLEU score (0.2728, 0.2734, 0.2724) over baseline system (0.2658); (3) When the scale of bilingual MWEs is relatively small (50000, 60000), the Baseline+Feat method performs better than others; (4) As the number of bilingual MWEs increasing, the Baseline+NewBP method outperforms the Baseline+Feat method; (5) Comparing table 4 and 3, we can see it is not true that the more bilingual MWEs, the better performance of phrase-based SMT. This conclusion is the same as (Lambert and Banchs, 2005).

Methods	50000	60000	70000
Baseline		0.2658	
Baseline+BiMWE	0.2671	0.2686	0.2715
Baseline+Feat	**0.2728**	**0.2734**	0.2712
Baseline+NewBP	0.2662	0.2706	**0.2724**

Table 4: Translation results of using bilingual MWEs in traditional medicine domain

To verify the assumption that bilingual MWEs do indeed improve the SMT performance not only on particular domain, we also perform some experiments on chemical industry domain. Table 5 shows the results. From this table, we can see that these three methods can improve the translation performance on chemical industry domain as well as on the traditional medicine domain.

Methods	BLEU
Baseline	0.1882
Baseline+BiMWE	0.1928
Baseline+Feat	0.1917
Baseline+Newbp	0.1914

Table 5: Translation results of using bilingual MWEs in chemical industry domain

4.4 Discussion

In order to know in what respects our methods improve performance of translation, we manually analyze some test sentences and gives some examples in this subsection.

(1) For the first example in table 6, "*通 脉*" is aligned to other words and not correctly translated in baseline system, while it is aligned to correct target phrase "*dredging meridians*" in Baseline+BiMWE, since the bilingual MWE ("*通 脉*", "*dredging meridians*") has been added into training corpus and then aligned by GIZA++.

(2) For the second example in table 6, "*药 茶*" has two candidate translation in phrase table: "*tea*" and "*medicated tea*". The baseline system chooses the "*tea*" as the translation of "*药 茶*", while the Baseline+Feat system chooses the "*med-*

Src	该 食品 具有 补血 、 逐寒 、 通 脉 、 生 津 、 利 水 、 安神 等 滋补 功效 , 可 达 到 健身 营养 的 目的 。
Ref	the obtained product is effective in tonifying blood , expelling cold , **dredging meridians , promoting production of body fluid , promoting urination** , and tranquilizing mind ; and can be used for supplementing nutrition and protecting health .
Baseline	the food has effects in tonifying blood , dispelling cold , **promoting salivation and water** , and tranquilizing , and tonic effects , and making nutritious health .
+Bimwe	the food has effects in tonifying blood , dispelling cold , **dredging meridians , promoting salivation , promoting urination** , and tranquilizing tonic , nutritious pulverizing .
Src	还 可 制成 片剂 、 丸剂 、 散剂 、 药 茶 、 注射 剂 。
Ref	the product can also be made into tablet , pill , powder , **medicated tea** , or injection .
Baseline	may also be made into tablet , pill , powder , **tea** , or injection .
+Feat	may also be made into tablet , pill , powder , **medicated tea** , or injection .

Table 6: Translation example

icated tea" because the additional feature gives high probability of the correct translation "*medicated tea*".

5 Conclusion and Future Works

This paper presents the LLR-based hierarchical reducing algorithm to automatically extract bilingual MWEs and investigates the performance of three different application strategies in applying bilingual MWEs for SMT system. The translation results show that using an additional feature to represent whether a bilingual phrase contains bilingual MWEs performs the best in most cases. The other two strategies can also improve the quality of SMT system, although not as much as the first one. These results are encouraging and motivated to do further research in this area.

The strategies of bilingual MWE application is roughly simply and coarse in this paper. Complicated approaches should be taken into account during applying bilingual MWEs. For example, we may consider other features of the bilingual MWEs and examine their effect on the SMT performance. Besides application in phrase-based SMT system, bilingual MWEs may also be integrated into other MT models such as hierarchical phrase-based models or syntax-based translation models. We will do further studies on improving statistical machine translation using domain bilingual MWEs.

Acknowledgments

This work is supported by National Natural Science Foundation of China, Contracts 60603095 and 60873167. We would like to thank the anonymous reviewers for their insightful comments on an earlier draft of this paper.

References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. Going beyond aer: an extensive analysis of word alignments and their impact on mt. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisiton and Treatment*, pages 89–96.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions*, pages 1–8.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Baobao Chang, Pernilla Danielsson, and Wolfgang Teubert. 2002. Extraction of translation unit from chinese-english parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing*, pages 1–5.

Stanley F. Chen and Joshua Goodman. 1998. Am empirical study of smoothing techniques for language modeling. Technical report.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual*

Meeting on Association for Computational Linguistics, pages 531–540.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 1–8.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32.

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

Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the unified medical language system. In *Proceedings of the 20th international conference on Computational Linguistics table of contents*, pages 792–798.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the EACL*, pages 337–344.

Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting nested collocations. In *Proceedings of the 16th conference on Computational linguistics*, pages 41–46.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Patrik Lambert and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 396–403.

Patrik Lambert and Rafael Banchs. 2006. Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*, pages 9–16.

Adam Lopez and Philip Resnik. 2006. Word-based alignment, phrase-based translation: What's the link? In *proceedings of the 7th conference of the association for machine translation in the Americas: visions for the future of machine translation*, pages 90–99.

Shengfen Luo and Maosong Sun. 2003. Two-character chinese word extraction based on hybrid of internal and contextual measures. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 24–30.

Franz Josef Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.d. thesis, Computer Science Department, RWTH Aachen, Germany.

Patrick Pantel and Dekang Lin. 2001. A statistical corpus based term extractor. In *AI '01: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, pages 36–46.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics*, pages 311–318.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech and Language*, 19(4):378–397.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3th International Conference on Intelligent Text Processing and Computational Linguistics(CICLing-2002)*, pages 1–15.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of Conference on Computational Linguistics (COLING)*, pages 993–1000.