Improving Word Alignment Models Using Structured Monolingual Corpora

Wei Wang *
New York University
719 Broadway Room 716
New York, NY
10003-6806 USA
wei@cs.nyu.edu

Ming Zhou
Microsoft Research Asia
5/F, Beijing Sigma Center
No.49, Zhichun Road, Haidian District
Beijing 100080 China
mingzhou@microsoft.com

Abstract

We propose a new method to improve the performance of word alignment algorithms, in particular the recall, using structured monolingual corpora as sources to estimate cross language word similarities. Normally, cross language word similarities, i.e., the similarity between a source language word and a target language word, can be estimated with a bilingual corpus of enough size. We use a method to estimate them from two structured monolingual corpora based on the dependency correspondence assumption justified on large and balanced bilingual corpora. We selected three typical word alignment models ranging over statistical-based ones and heuristic-based ones, to test whether cross language similarities can improve the performance of word alignment models. The crosslingual word similarities are simply interpolated into these models. The experiments show that crosslingual word similarities estimated from structured monolingual corpora can effectively improve the performance of word alignment models, in particular the recall.

1 Introduction

Word alignment algorithms, e.g., (P. Brown *et al.*, 1993), accept a bilingual sentence pair as input, and output the links between words across sentences of the pair. Such links are very useful knowledge for machine translation.

Many of previous works implicitly assume that bilingual resources (e.g., bilingual corpora, bilingual dictionaries) on a large scale are available. Better performance of word alignment algorithms can be achieved if the training is conducted on larger parallel corpora. The increasing requirement for bilingual resources thus becomes a bottleneck of constructing practical word aligners for a general domain. To alleviate the problem, some works, like (Resnik & Smith, 2003; J. Nie *et al.*, 1999), try to obtain parallel corpora by automatically scrawling them from the web.

Instead of heavily relying on bilingual corpora, some works try to solve the bottleneck problem in a different way: to mine bilingual knowledge from monolingual corpora, which can be more easily obtained in a large volume.

Koehn & Knight (2000) present an approach to estimating word translation probabilities by using unrelated monolingual corpora based on the EM algorithm. Promising results of their method are exhibited in selecting the right translation among several options provided by a bilingual dictionary. However, their method focuses on how to stochastize a bilingual dictionary instead of how to extend it by adding new translations. They discussed that better language modeling, e.g., lexical dependencies, may be useful for further improvement.

Fung & Lee (1998) use an IR approach to inducing new word translations from comparable corpora. Their method is able to induce new translations, thus to extend a bilingual dictionary. As discussed in their paper, their method suffers form low recall. They gave some suggestions in achieving better results, including introducing PoS tagging, improving word segmentation accuracy. Using a better language model will certainly do these.

Zhou *et al.* (2001) compute crosslingual word similarities from two dependency triple databases, each of which is obtained by parsing a large monolingual corpus with a dependency parser. Since bilexical dependency correspondences are considered in the computation, the resulting similarities are expected to be more accurate. Better language modeling imposes more constraints on the confidences/weights of translation candidates. Encouraging results are shown in the application of the similarities to

* This work was done while the author was visiting Microsoft Research Asia.

translation selection.

In this paper, we are interested in the question how and to what extent the word alignment task can benefit from the work in monolingual language processing, e.g., monolingual parsers and monolingual corpora, so that the requirement for large bilingual corpora could be minimized. Specifically, we are concerned with how to "convert" structured monolingual corpora into bilingual word similarities (or translation probabilities) and how to use the similarities to enhance the existing word alignment models.

To do this, we first empirically justify the assumption exploiting the structural correspondence between different languages.[1] Then crosslingual word similarities are computed from structured corpora using the method in (Zhou *et al.*, 2001) based on this assumption. After that, the crosslingual word similarities are normalized and incorporated into word alignment models. Experiments are conducted on different types of alignment models ranging from statistical-based ones and heuristic-based ones. Experimental results show that word alignment models can be consistently improved with the crosslingual similarities estimated from structured monolingual corpora.

The remainder of this paper is organized as follows. Section 2 lists the related research. Section 3 justifies the dependency correspondence assumption. Section 4 presents the method to estimate crosslingual word similarities from structured monolingual corpora. Sections 5 and 6 describe how to incorporate estimated similarities into word alignment methods. Experiments are reported in Section 7. The last section makes conclusions and points out future works.

2 Related Research

One of the related topics is word alignment. Word alignment models can generally be classified into two categories (Och & Ney, 2003): statistical alignment models and heuristic ones. A statistical alignment model $p(\mathbf{c}, \mathbf{a}|\mathbf{e})$ describes the relationship (e.g., word alignment $\mathbf{a}$) between a source language string $\mathbf{e}$ and a target language string $\mathbf{c}$. Different decompositions of $p(\mathbf{c}, \mathbf{a}|\mathbf{e})$ result in different variants of word alignment models. Heuristic based word alignment approaches use similarity functions of two languages. Readers might want to refer to (Och & Ney, 2003) for a comprehensive examination of word alignment models.

Another related research topic is the automatic estimation of crosslingual word similarities (or probabilities) from monolingual corpora. For example, the works (Koehn & Knight, 2000; Fung & Lee, 1998; Zhou *et al.*, 2001) that we have mentioned in Section 1.

Methods to estimate crosslingual word similarities are sometimes related to methods of monolingual word clustering. For instance, the method used in (Zhou *et al.*, 2001) is motivated by the work in (Lin, 1998). Lin (1998) presents a method to cluster monolingual words based on similarities between words in the same language. The word similarities are estimated from a parsed monolingual corpus.

The third related topic is the justification of Direct Correspondence Assumption (DCA), which underlies the models/applications exploiting high level linguistic structures. For example, tree-tree alignment, e.g., (Matsumoto, 1993), in example-based machine translation; synchronous grammars, e.g., (Wu, 1997), for statistical machine translation (SMT). R. Hwa *et al.* (2002) formulate the DCA, and evaluate it in terms of the accuracies (precision and recall) of the Chinese syntactic parses projected from the corresponding English parse trees, which are the output of an English parser. Their Experiments are done on small newswire corpora. They conclude that DCA is useful with some principled transformation of syntactic structures.

3 Dependency Correspondence Assumption

Methods, e.g., (Zhou *et al.*, 2001), to estimate crosslingual word similarities from structured monolingual corpora also take advantage of this assumption. The justification of this assumption using a large scale and balanced data is thus necessary.

The Dependency Correspondence Assumption can be formally expressed as follows. Let triple $\langle w_1, R, w_2 \rangle$ be a syntactic dependency consisting of two words w_1 and w_2 and a dependency relation R between them. Given a pair of sentences E and C that are translation of each other with syntactic structures $Tree_E$ and $Tree_C$, if $Tree_E$ contains a dependency triple $\langle e_1, R, e_2 \rangle$, and e_1 and e_2 are aligned to words

[1] Although the language pair used in our paper is English and Chinese, the basic idea of this paper, however, can be generalized to other language pairs.

Dependency	Correct	Incorrect	Map Ratio
VO(E-to-C)	9,991	2,088	82.71%
VO(C-to-E)	8,823	1,697	83.87%

Table 1: Verb-Object dependency correspondence between English and Chinese.

c_1 and c_2 in C, respectively, then there is a dependency triple $\langle c_1, R, c_2 \rangle$ in $Tree_C$.

The experiment that we have done differentiates from that in (R. Hwa *et al.*, 2002) in the following aspects.

First, we use a much larger and balanced corpora consisting of 10,000 English-Chinese sentence pairs,[2] coming from newswire, novels, general bilingual dictionaries, and software product manuals.

Second, instead of examining all the types of dependency relations, we examine only the Verb-Object (VO) dependency type since we think that VO is one of the dependency types that are often preserved across languages, and DCA will thus mostly hold among these dependency types.

Third, we evaluate the DCA in terms of *mapping ratio*: the ratio of the number of correct crosslingual dependency mappings versus the number of overall mappings. To do this, we first manually add the word alignment information to the 10,000 sentence pairs. Then, we run the English parser MiniPar (Lin, 1993) and the Chinese dependency parser BlockParser (Zhou, 2000) on both sides of the corpora, respectively. Next, we extract all the mappings from English dependency triples to Chinese dependency triples (and vice versa) based on the word alignment and parsing results.

Table 1 lists the mapping results. The first column shows the dependency type and the mapping directions. The correct mapping ratio reaches 82.71% from English to Chinese, and 83.87% from Chinese to English. Note that we treated the dependency type Verb-Object-Prep as VO.

The intent of this experiment is not to compare with the method and results in (R. Hwa *et al.*, 2002), but to evaluate the dependency correspondence assumption for a certain type of dependency concerned by us.

The numbers in Table 1 are very encouraging because it indicates the feasibility of DCA. This also suggests that translating the sentences in the way of keeping DCA on some key dependency types can normally get understandable translation results.

[2]Created by Microsoft Research Asia, and not publicly available.

4 Crosslingual Word Similarities

We now briefly describe the method that we use to estimate crosslingual words similarities from structured monolingual corpora. The method was proposed in (Zhou *et al.*, 2001). We shall use slightly different notations.

The information of a dependency triple $\tau = \langle w_1, R, w_2 \rangle$ is defined as:

$$I(\tau) = \log_2 \frac{c(w_1, R, w_2)c(\cdot, R, \cdot)}{c(w_1, R, \cdot)c(\cdot, R, w_2)} \quad (1)$$

where $c(...)$ is the counting function. τ is called a *supportive dependency* of w_1 (or w_2) if $I(\tau) > 0$. Let $D(e)$ be the set of e's supportive dependencies collected from a structured corpus in language L_1. Let $D(c)$ be the set of c's supportive dependencies collected from a structured corpus in language L_2. Let $\delta(\tau)$ be a function returning 1 if dependency $\tau \in D(e)$ has a corresponding dependency in $D(c)$, and e corresponds to c;[3] and returning 0 otherwise. Note that we need a bilingual dictionary to "bridge" corresponding dependencies in different languages, and this bilingual dictionary is the only bilingual resource used. The *common information* between c and e is then defined as:

$$I_c(e, c) = \sum_{\tau \in D(e)} \delta(\tau) I(\tau) + \sum_{\tau \in D(c)} \delta(\tau) I(\tau) \quad (2)$$

The *overall information* between e and c is defined as:

$$I_o(e, c) = \sum_{\tau \in D(e)} I(\tau) + \sum_{\tau \in D(c)} I(\tau) \quad (3)$$

The similarity $sim(c, e)$ between a pair of crosslingual words c in language L_1 and e in language L_2 is defined as the ratio of $I_c(c, e)$ to $I_o(c, e)$:

$$sim(c, e) = \frac{I_c(c, e)}{I_o(c, e)} \quad (4)$$

Readers might want to refer to (Zhou *et al.*, 2001) for detailed derivations.

[3]Or in the other direction, returning 1 if $\tau \in D(c)$ has a corresponding dependency in $D(e)$.

It is worth mentioning that this method has the ability of inducing new translations. In principle, the similarity of any pair of crosslingual words can be computed using Formula 4. Since the information used in the computation is distributedly encoded in the relevant dependency triples in the entire treebank, this method is also very robust.

Although the value of function $sim(c, e)$ computed by their method ranges over [0,1], it, however, is not a probability distribution because of the fact that $\sum_{c,e} sim(c, e) \neq 1$. We thus use the following normalization so that $sim(c, e)$ can be incorporated into a statistical translation model:

$$q(c|e) = \frac{sim(c, e)}{\sum_{c'} sim(c', e)} \quad (5)$$

In the following two sections, we are going to present methods to integrate $q(c|e)$ into word alignment models for performance improvements. Word alignment models that will be involved range from statistical-based ones to heuristic-based ones.

5 Improving Statistical-Based Word Alignment Models

IBM Model 2 is used to test the usefulness of $q(c|e)$ to statistical word alignment algorithms. The reason why we have not used more complex models is that our objective is not the comparison of different word alignment models. It is reasonable to conclude that, if Model 2 can be improved by integrating $q(c|e)$, more complex models can be improved, too.[4] For brevity, we shall call the statistical word alignment model **STATS** hereafter.

We use a simple interpolated model to combine $q(c|e)$ with the the word-to-word translation probabilities $p(c|e)$ estimated from bilingual corpora:

$$p_{stats-interp}(c|e) = \lambda\, q(c|e) + (1 - \lambda)\, p(c|e) \quad (6)$$

where $0 \leq \lambda \leq 1$. We have made λ a constant for all $\langle c, e\rangle$ pairs to get around the data sparseness problem in estimation.

[4]Of course, word alignment models (or translation models) can always be improved by exploiting more information, e.g., the structural information; but this does not conflict with out objective because a more complex model is usually composed of more parameters, and thus requires larger size of bilingual corpora for training. How to extract bilingual knowledge from monolingual corpora and how to combine them into word alignment models is exactly our objective.

Like the estimation of interpolated monolingual language models, the optimal interpolation coefficient λ can be estimated via the EM algorithm from held-out bilingual corpora X such that

$$\lambda^* = \arg\max_{\lambda} p_{stats-interp}(X) \quad (7)$$

We shall refer to the interpolated statistical model as **STATS-interp** hereafter.

6 Improving Heuristic-Based Word Alignment Models

We consider the usefulness of $q(e|c)$ to the heuristic-based word alignment models. We use two types of heuristic-based word alignment methods: the dictionary-based method and the class-based method. They exploit different types of bilingual knowledge.

6.1 Dictionary-Based Models

A way to examine the degree to which the coverage of a translation dictionary can be improved by $q(c|e)$ is to show the word alignment accuracy (e.g., precision and recall). The baseline for comparison is the dictionary-based word alignment method (**DICT**) in (Ker & Zhang, 1997), which is briefly described in Figure 1. The input of DICT is a pair $\langle S, T\rangle$ of sentences and a bilingual dictionary. The following method can be used to incorporate $q(c|e)$ into DICT.

$$p_{dict-interp}(c|e) = \lambda_{(c,e)} q(e|c) + [1 - \lambda_{(c,e)}] sign((c, e) \in \text{BD}) \quad (8)$$

where BD stands for the bilingual dictionary. $\lambda_{(c,e)}$ is chosen such that if $q(c|e)$ is larger than a threshold, it is a non-zero value (e.g., empirically chosen as 0.3), and 0 otherwise. When $\lambda_{(c,e)}$ is 0, the bilingual dictionary will take full control — *sign* functor returns 1 if (c, e) is in the bilingual dictionary, and 0, otherwise. We shall refer to the interpolated DICT as **DICT-interp**.

6.2 Class-Based Models

Class-based word alignment method (Ker & Zhang, 1997) attempts to broaden the word alignment coverage/recall using *crosslingual concept similarities*. Concepts are classes defined in a monolingual thesaurus. Crosslingual concept similarities are estimated from bilingual corpora by generalizing the words into their classes.

1. Enumerate all words W_S in S and words W_T in T.
2. Foreach s in W_s, find the set of translations DT_s of s based on a bilingual dictionary.
3. For $d \in DT_s$ and $t \in W_T$, calculate the dictionary
based similarity ($DTSim(s,t)$)
4. Foreach word s, if $DTSim(s,t)$ is maximized over $t \in W_T$, produce a connection (s,t).
5. Compile the list ALN of word alignments.

Figure 1: Dictionary-based word alignment method (DICT) (Ker & Zhang, 1997)

The basic idea of the class-based word alignment algorithm is as follows. Taking a bilingual sentence pair as input, the algorithm first conducts DICT (see Figure 1), resulting in a list ALN of word alignments. Words that are not in ALN are are aligned using the following model:

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} conceptsim(c_j|e_i)d(i,j) \quad (9)$$

where $conceptsim(c_j|e_i)$ is the concept similarity between word c_j in one language and e_i in the other language. $d(i,j)$ is the distortion model. We shall use **CLASS** to refer to the class-based method.

As with the integration of $q(c|e)$ into statistical word alignment models in Section 5, $q(c|e)$ can play a role in the improvement on class-based word alignment models by interpolating $q(c|e)$ with $conceptsim(c|e)$ as follows.

$$\begin{aligned} & p_{class-interp}(c|e) \\ =& \lambda conceptsim(c|e) + (1-\lambda) q(c|e) \end{aligned} \quad (10)$$

where $0 \leq \lambda \leq 1$. The optimal interpolated coefficient λ^* can be computed via the EM algorithm with held-out bilingual data. The class-based model interpolating with $q(c|e)$ will be referred to as **CLASS-interp**.

7 Experiments

The experiments include the estimation of crosslingual word similarities from structured monolingual corpora (Section 4), and the application of the estimated similarities to word alignment algorithms.

7.1 Estimation of Crosslingual Word Similarities

To obtain two sets of dependency triples, each set per language, the English dependency parser MiniPar (Lin, 1993) is applied to 750M bytes of English corpora of Wall Street Journal (1980-1990), resulting in 1.9×10^7 English dependency triples. The Chinese dependency parser BlockParser (Zhou, 2000) is applied to 1,200M bytes of Chinese corpora of People's Daily (1980-1998), resulting in 3.3×10^7 Chinese dependency triples.

The HIT English-Chinese bilingual dictionary[5] consisting of 66,248 Chinese words, 73,693 English words, and 164,794 translation links is used.

7.2 Improving Word Alignment Models

Evaluation metrics

Let A denote the set of word alignments from a word alignment method for a pair of bilingual sentences, and let R denote the set of word alignments of the same sentence pair in the reference corpora, let $|\cdot|$ denote the size of set $\cdot$, then we use

$$\text{Precision} = \frac{|A \cap R|}{|A|} \quad (11)$$

$$\text{Recall} = \frac{|A \cap R|}{|R|} \quad (12)$$

$$\text{F–measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

where $\cap$ is a functor of two sets of word alignments, returning the number of *matched* alignments.

Experimental settings

215,347 pairs of English and Chinese sentences are used to train the statistical word alignment algorithm. Since the interpolation coefficients are manually assigned in our experiments, no held-out data were reserved.

The HIT English-Chinese bilingual dictionary is used in the dictionary-based word alignment methods — DICT and DICT-interp.

The same 215,347 bilingual sentences as used in the training of statistical-based models are used to train the crosslingual concept similarities $conceptsim(e|c)$ and the distortion model $d(i,j)$. Monolingual thesauri are WordNet (G. Miller, 1990) with 45,784 classes for English and Xian Dai Han Yu Tong Yi Ci Dian (Mei, 2002) with 3,724 classes (40,289 words) for Chinese.

A single test set is used for all the word alignment methods. It consists of 1,000 sentence pairs that are disjoint with the training data.

[5] Created by Harbin Institute of Technology (HIT).

Algorithms	Prec. (%)	Rec. (%)	F
STATS	73.69	67.31	70.36
STATS-interp	**74.54**	**68.63**	**71.46**

Table 2: Word alignment methods STATS vs. STATS-interp

Algorithms	Prec. (%)	Rec. (%)	F
DICT	95.87	44.60	60.88
DICT-interp	**95.28**	**46.84**	**62.80**

Table 3: Word alignment methods DICT vs. DICT-interp

Experimental results

The performance comparison between STATS and STATS-interp are shown in Table 2. All metrics including the precision, recall and thus F-measure have been improved. The interpolation of crosslingual word similarities estimated from monolingual corpora into the statistical word alignment method STATS makes the word-to-word translation probability more reliable than otherwise. Although the training bilingual corpus is relatively large, there still exists the data sparseness problem. Probabilities of rare word translations are often inaccurate. These probabilities are "adjusted" by being interpolated with crosslingual word similarities from a completely different knowledge source — monolingual copora. The (optimized) interpolated coefficient λ decides how the knowledge from difference sources (monolingual corpora or bilingual corpora) are weighted.

The performance comparison between DICT and DICT-interp are shown in Table 3. We see that the recall and F-measure metrics of DICT have been improved by DICT-interp. It is worth mentioning that the recall is even improved by more than 2% with only a slight drop of precsion. This implies that the crosslingual word similarities provide more chances for words to be aligned when the bilingual dictionary fails. Table 4 shows an example. The words and alignments in bold fonts are those where the crosslingual word similarities help, and the bilingual dictionary fails. Chinese are written in pinyin.

Table 5 lists the top 8 Chinese words (in pinyin) mined from the monolingual corpora that are the possible translations of the English word "salary". The rightmost column shows the

E:	20000/1 is/2 a/3 very/4 **respectable/5 salary/6** ./7
C:	20000/1 Ying1Bang4/2 De0/3 **Xin1Jin1/4** Shi4/5 Fei1Chang2/6 **Ke3Guan1/7** De0/8 ./9
Align:	[1:1] [4:6] **[5:7]** **[6:4]** [7:9]

Table 4: An example of word alignment output from DICT-interp.

Gong1Zi1 [salary]	0.178
Nian2 [year]	0.150
Xin1Jin1 [salary]	0.150
Xin1Shui3 [salary]	0.112
Yue4 [month]	0.060
Shou1Yi4 [benefit]	0.059
Jin1Nian2 [this year]	0.044
Jia1Xin1 [raise]	0.004

Table 5: Crosslingual word similarities.

$q(c|e)$. Words in brackets are the real translations of the corresponding pinyin (in the same rows).

The performance comparisons between CLASS and CLASS-interp are shown in Table 6. There are two reasons why CLASS is improved: First, although the generalization of words into their classes alleviates the data sparseness problem in CLASS, it gives rise to the overgeneralization problem.[6] The combination with crosslingual word similarities makes the crosslingual concept similarities more informative.

Second, the thesauri used are not large enough for a general domain. For example, the Chinese thesaurus provides classes only for 40,289 Chinese words. These words even cannot cover all the Chinese words in the bilingual dictionary we used. Furthermore, the definition of Chinese words is not consistent, e.g, between the Chinese thesaurus and the Chinese word segmentor. In the case where thesauri fails, our crosslingual word similarities takes control, raising the recall.

7.3 Discussions

One of essential reasons why the performance of different types (statistical-based and heuristic-based) of alignment algorithms can be improved

[6]Like in monolingual language modeling, too small number of classes will not guarantee the reduction of perplexity.

Algorithms	Prec. (%)	Rec. (%)	F
CLASS	86.84	54.96	67.32
CLASS-interp	**86.03**	**56.53**	**68.23**

Table 6: Word alignment methods CLASS vs. CLASS-interp

is that the crosslingual word similarities estimated from monolingual corpora are able to broaden the coverage of, or adjust the knowledge learned from bilingual corpora. The coverage is broadened because we can induce from structured monolingual corpora reliable new translations that do not co-occur in the same sentence pair in bilingual corpora. The knowledge in bilingual corpora is adjusted in a way that probabilities of translations with low frequencies can be smoothed by simply being interpolated with the normalized word similarities estimated from structured monolingual corpora.

It is worth emphasizing that the improvements that we have achieved are valuable for the word alignment task in a general domain. With the large training corpus, and a large bilingual dictionary, the word alignment methods we have used in experiments have set a high baseline for comparisons. In a general domain, a slight improvement on the recall metric of the word alignment result usually requires a large increment of the size of training bilingual corpora. By utilizing the crosslingual word similarities that are estimated from monolingual corpora, we have got around the problem of how to obtain bilingual corpora on a large scale.

The limitation of our experiments is that we used two independent monolingual corpora for the estimation of crosslingual word similarities. The corpora are independent in the following ways: first, they are collected independently; second, they are collected during different dates; third, their major topics are different, e.g., political versus business. Better results could be expected if we use comparable corpora, because, in comparable copora, crosslingual phrases with similar structures will provide more information for the estimation of crosslingual word similarities.

Our approach to estimating crosslingual word similarities assumes that two monolingual parsers are available. Although the construction of monolingual parsers is expensive, there are free parsers available on the web.

8 Conclusions

We have been concerned with how to "convert" structured monolingual corpora into bilingual word similarities (or translation probabilities) and how to incorporate them into word alignment models for performance improvement. Our aim is to take advantage of monolingual resources, e.g., corpora, parsers, treebank, for bilingual tasks, e.g., word alignment, so that the requirement for large amounts of training bilingual corpora could be alleviated.

To do this, we first empirically justified the dependency correspondence assumption between different languages using large and balanced corpora. Then, we computed crosslingual word similarities using the method in (Zhou *et al.*, 2001) based on this assumption. After that, we normalized the crosslingual word similarities and integrated them into word alignment algorithms using interpolated models.

Experiments are conducted on word alignment algorithms ranging from statistical-based ones and heuristic-based ones. Experimental results show that word alignment models have been consistently improved, in particular the recall metric.

The main contribution of this paper is that we have presented an approach to combining the knowledge mined from structured monolingual corpora with the knowledge mined from bilingual corpora, and showed the usefulness of the approach to the improvement on word alignment performance, and thus showed the usefulness of monolingual resources to bilingual tasks. Moreover, we also justified the dependency correspondence assumption based on a large and balanced corpora. This assumption underlies the method that we have used to estimated crosslingual word similarities from structured monolingual corpora.

One of the interesting topics deserving study in the future could be to divide the crosslingual word pairs into clusters and estimate the optimal interpolation coefficient λ for each of them using the minimum error rate (MER) as the optimization goal, instead of maximum likelihood. Results in machine translation (Och, 2003) and speech recognition have shown the advantage of discriminative training.

Another topic will be to compute the (normalized) word similarities $q(e|c)$ in an bootstrapping manner, for instance, using the EM algorithm. In each iteration, the (normalized) word similarities output from the previous it-

eration are used as weights of two corresponding dependency triples in two languages. These weights are involved in the computation of the common information (Section 4) between two crosslingual words. We also desire to improve statistical models of the state of art with these re-estimated similarities.

Acknowledgments

We would like to thank the reviewers for their valuable comments.

References

Brown, P.F., Della Pietra, S.A., Della Pietra V.J., and Mercer, R.L. (1993) "The mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, 19(2), pp.263-311.

P. Fung and Y. Lee (1998) "Translating Unknown Words Using Nonparallel, Comparable Texts," In Proceedings of *COLING-ACL98*, Montreal, Canada: Aug. 1998, 414-420.

R. Hwa, P. Resnik, A. Weinberg, and O. Kolak (2002) "Evaluating Translational Correspondence using Annotation Projection" In proceedings of *ACL-02*, Philadelphia, July, 2002.

S. Ker and J. Zhang (1997) "A class-based approach to word alignment," *Computational Linguistics*, Vol. 23, No. 2, pp 313-343.

P. Koehn and K. Knight (2000) "Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm," *AAAI 2000.*

D. Lin (1993) *Principle-Based Parsing without Overgeneration.* Proceedings of ACL-93, pp 112-120, Columbus, Ohio.

D. Lin (1998) "Automatic retrieval and clustering of similar words," *COLING-ACL98*, Mentreal, Canada.

Y. Matsumoto (1993) "Structural Matching of Parallel Texts," In *Proceedings of ACL 93.*

J. Mei (2002) "Xiandai Hanyu TongYi CiDian," *The Commercial Press LTD. of China.*

G. Miller (1990). "WordNet: An on-line lexical database," *International Journal of Lexicography*, 3(4):235-312, 1990.

J. Nie, M. Simard, P. Isabelle, R. Durand (1999) "Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web", *22nd ACM-SIGIR*, pp. 74-81.

F. Och and H. Ney (2003) "A Systematic Comparison of Various Statistical Alignment Models," In *Computational Linguistics.*

F. Och (2003) "Minimum Error Rate Training in Statistical Machine Translation", in *ACL 2003.*

P. Resnik and N. Smith (2003) "The Web as a Parallel Corpus," *Computational Linguistics*, Vol. 29, Issue 3, pp 349-380.

D. Wu (1997) "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Computational Linguistics 23(3)*:377-404, September.

M. Zhou (2000) "A Block-Based Robust Dependency Parser for Unrestricted Chinese Text," *2nd Workshop on Chinese language processing*, Hong Kong.

M. Zhou, Y. Ding, and C. Huang (2001) "Improving Translation Selection with a New Translation Model Trained by Independent Monolingual Corpora," *Computational Linguistics and Chinese Language Processing*, Vol.6, No.1, Feb. 2001, pp. 1-26.