Collocation Translation Acquisition Using Monolingual Corpora

Yajuan LÜ
Microsoft Research Asia
5F Sigma Center,
No. 49 Zhichun Road, Haidian District,
Beijing, China, 100080
t-yjlv@microsoft.com

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

Abstract

Collocation translation is important for machine translation and many other NLP tasks. Unlike previous methods using bilingual parallel corpora, this paper presents a new method for acquiring collocation translations by making use of monolingual corpora and linguistic knowledge. First, dependency triples are extracted from Chinese and English corpora with dependency parsers. Then, a dependency triple translation model is estimated using the EM algorithm based on a dependency correspondence assumption. The generated triple translation model is used to extract collocation translations from two monolingual corpora. Experiments show that our approach outperforms the existing monolingual corpus based methods in dependency triple translation and achieves promising results in collocation translation extraction.

1 Introduction

A collocation is an arbitrary and recurrent word combination (Benson, 1990). Previous work in collocation acquisition varies in the kinds of collocations they detect. These range from two-word to multi-word, with or without syntactic structure (Smadja 1993; Lin, 1998; Pearce, 2001; Seretan et al. 2003). In this paper, a collocation refers to a recurrent word pair linked with a certain syntactic relation. For instance, <*solve*, *verb-object*, *problem*> is a collocation with a syntactic relation verb-object.

Translation of collocations is difficult for non-native speakers. Many collocation translations are idiosyncratic in the sense that they are unpredictable by syntactic or semantic features. Consider Chinese to English translation. The translations of "解决" can be "*solve*" or "*resolve*". The translations of "问题" can be "*problem*" or "*issue*". However, translations of the collocation "解决 ~ 问题" as "*solve~problem*" or "*resolve~issue*" is preferred over "*solve~issue*" or "*resolve ~problem*". Automatically acquiring these collocation translations will be very useful for machine translation, cross language information retrieval, second language learning and many other NLP applications. (Smadja et al., 1996; Gao et al., 2002; Wu and Zhou, 2003).

Some studies have been done for acquiring collocation translations using parallel corpora (Smadja et al, 1996; Kupiec, 1993; Echizen-ya et al., 2003). These works implicitly assume that a bilingual corpus on a large scale can be obtained easily. However, despite efforts in compiling parallel corpora, sufficient amounts of such corpora are still unavailable. Instead of heavily relying on bilingual corpora, this paper aims to solve the bottleneck in a different way: to mine bilingual knowledge from structured monolingual corpora, which can be more easily obtained in a large volume.

Our method is based on the observation that despite the great differences between Chinese and English, the main dependency relations tend to have a strong direct correspondence (Zhou et al., 2001). Based on this assumption, a new translation model based on dependency triples is proposed. The translation probabilities are estimated from two monolingual corpora using the EM algorithm with the help of a bilingual translation dictionary. Experimental results show that the proposed triple translation model outperforms the other three models in comparison. The obtained triple translation model is also used for collocation translation extraction. Evaluation results demonstrate the effectiveness of our method.

The remainder of this paper is organized as follows. Section 2 provides a brief description on the related work. Section 3 describes our triple translation model and training algorithm. Section 4 extracts collocation translations from two independent monolingual corpora. Section 5 evaluates the proposed method, and the last section draws conclusions and presents the future work.

2 Related work

There has been much previous work done on monolingual collocation extraction. They can in

general be classified into two types: window-based and syntax-based methods. The former extracts collocations within a fixed window (Church and Hanks 1990; Smadja, 1993). The latter extracts collocations which have a syntactic relationship (Lin, 1998; Seretan et al., 2003). The syntax-based method becomes more favorable with recent significant increases in parsing efficiency and accuracy. Several metrics have been adopted to measure the association strength in collocation extraction. Thanopoulos et al. (2002) give comparative evaluations on these metrics.

Most previous research in translation knowledge acquisition is based on parallel corpora (Brown et al., 1993). As for collocation translation, Smadja et al. (1996) implement a system to extract collocation translations from a parallel English-French corpus. English collocations are first extracted using the Xtract system, then corresponding French translations are sought based on the Dice coefficient. Echizen-ya et al. (2003) propose a method to extract bilingual collocations using recursive chain-link-type learning. In addition to collocation translation, there is also some related work in acquiring phrase or term translations from parallel corpus (Kupiec, 1993; Yamamoto and Matsumoto 2000).

Since large aligned bilingual corpora are hard to obtain, some research has been conducted to exploit translation knowledge from non-parallel corpora. Their work is mainly on word level. Koehn and Knight (2000) presents an approach to estimating word translation probabilities using unrelated monolingual corpora with the EM algorithm. The method exhibits promising results in selecting the right translation among several options provided by bilingual dictionary. Zhou et al.(2001) proposes a method to simulate translation probability with a cross language similarity score, which is estimated from monolingual corpora based on mutual information. The method achieves good results in word translation selection. In addition, (Dagan and Itai, 1994) and (Li, 2002) propose using two monolingual corpora for word sense disambiguation. (Fung, 1998) uses an IR approach to induce new word translations from comparable corpora. (Rapp, 1999) and (Koehn and Knight, 2002) extract new word translations from non-parallel corpus. (Cao and Li, 2002) acquire noun phrase translations by making use of web data. (Wu and Zhou, 2003) also make full use of large scale monolingual corpora and limited bilingual corpora for synonymous collocation extraction.

3 Training a triple translation model from monolingual corpora

In this section, we first describe the dependency correspondence assumption underlying our approach. Then a dependency triple translation model and the monolingual corpus based training algorithm are proposed. The obtained triple translation model will be used for collocation translation extraction in next section.

3.1 Dependency correspondence between Chinese and English

A dependency triple consists of a head, a dependant, and a dependency relation. Using a dependency parser, a sentence can be analyzed into dependency triples. We represent a triple as (w_1,r,w_2), where w_1 and w_2 are words and r is the dependency relation. It means that w_2 has a dependency relation r with w_1. For example, a triple (*overcome*, *verb-object*, *difficulty*) means that "*difficulty*" is the object of the verb "*overcome*".

Among all the dependency relations, we only consider the following three key types that we think, are the most important in text analysis and machine translation: verb-object (VO), noun-adj(AN), and verb- adv(AV).

It is our observation that there is a strong correspondence in major dependency relations in the translation between English and Chinese. For example, an object-verb relation in Chinese (e.g.(克服, VO, 困难)) is usually translated into the same verb-object relation in English(e.g. (*overcome*, VO, *difficulty*)).

This assumption has been experimentally justified based on a large and balanced bilingual corpus in our previous work (Zhou et al., 2001). We come to the conclusion that more than 80% of the above dependency relations have a one-one mapping between Chinese and English. We can conclude that there is indeed a very strong correspondence between Chinese and English in the three considered dependency relations. This fact will be used to estimate triple translation model using two monolingual corpora.

3.2 Triple translation model

According to Bayes's theorem, given a Chinese triple $c_{tri} = (c_1, r_c, c_2)$, and the set of its candidate English triple translations $e_{tri} = (e_1, r_e, e_2)$, the best English triple $\hat{e}_{tri} = (\hat{e}_1, r_e, \hat{e}_2)$ is the one that maximizes the Equation (1):

$$\begin{aligned}\hat{e}_{tri} &= \arg\max_{e_{tri}} p(e_{tri} \mid c_{tri}) \\ &= \arg\max_{e_{tri}} p(e_{tri})\, p(c_{tri} \mid e_{tri}) / p(c_{tri}) \\ &= \arg\max_{e_{tri}} p(e_{tri})\, p(c_{tri} \mid e_{tri})\end{aligned} \quad (1)$$

where $p(e_{tri})$ is usually called the language model and $p(c_{tri} \mid e_{tri})$ is usually called the translation model.

Language Model

The language model $p(e_{tri})$ is calculated with English triples database. In order to tackle with the data sparseness problem, we smooth the language model with an interpolation method, as described below.

When the given English triple occurs in the corpus, we can calculate it as in Equation (2).

$$p(e_{tri}) = \frac{freq(e_1, r_e, e_2)}{N} \quad (2)$$

where $freq(e_1, r_e, e_2)$ represents the frequency of triple e_{tri}. N represents the total counts of all the English triples in the training corpus.

For an English triple $e_{tri} = (e_1, r_e, e_2)$, if we assume that two words e_1 and e_2 are conditionally independent given the relation r_e, Equation (2) can be rewritten as in (3)(Lin, 1998).

$$p(e_{tri}) = p(r_e)\, p(e_1 \mid r_e)\, p(e_2 \mid r_e) \quad (3)$$

where $p(r_e) = \frac{freq(*, r_e, *)}{N}$,

$$p(e_1 \mid r_e) = \frac{freq(e_1, r_e, *)}{freq(*, r_e, *)},$$

$$p(e_2 \mid r_2) = \frac{freq(*, r_e, e_2)}{freq(*, r_e, *)}.$$

The wildcard symbol * means it can be any word or relation. With Equations (2) and (3), we get the interpolated language model as shown in (4).

$$p(e_{tri}) = \lambda \frac{freq(e_{tri})}{N} + (1-\lambda)\, p(r_e)\, p(e_1 \mid r_e)\, p(e_2 \mid r_e) \quad (4)$$

where $0 < \lambda < 1$. λ is calculated as below:

$$\lambda = 1 - \frac{1}{1 + freq(e_{tri})} \quad (5)$$

Translation Model

We simplify the translation model according the following two assumptions.

Assumption 1: Given an English triple e_{tri}, and the corresponding Chinese dependency relation r_c, c_1 and c_2 are conditionally independent. We have:

$$\begin{aligned}p(c_{tri} \mid e_{tri}) &= p(c_1, r_c, c_2 \mid e_{tri}) \\ &= p(c_1 \mid r_c, e_{tri})\, p(c_2 \mid r_c, e_{tri})\, p(r_c \mid e_{tri})\end{aligned} \quad (6)$$

Assumption 2: For an English triple e_{tri}, assume that c_i only depends on e_i ($i \in \{1,2\}$), and r_c only depends on r_e. Equation (6) is rewritten as:

$$\begin{aligned}p(c_{tri} \mid e_{tri}) &= p(c_1 \mid r_c, e_{tri})\, p(c_2 \mid r_c, e_{tri})\, p(r_e \mid e_{tri}) \\ &= p(c_1 \mid e_1)\, p(c_2 \mid e_2)\, p(r_c \mid r_e)\end{aligned} \quad (7)$$

Notice that $p(c_1 \mid e_1)$ and $p(c_2 \mid e_2)$ are translation probabilities within triples, they are different from the unrestricted probabilities such as the ones in IBM models (Brown et al., 1993). We distinguish translation probability between head ($p(c_1 \mid e_1)$) and dependant ($p(c_2 \mid e_2)$). In the rest of the paper, we use $p_{head}(c \mid e)$ and $p_{dep}(c \mid e)$ to denote the head translation probability and dependant translation probability respectively.

As the correspondence between the same dependency relation across English and Chinese is strong, we simply assume $p(r_c \mid r_e) = 1$ for the corresponding r_e and r_c, and $p(r_c \mid r_e) = 0$ for the other cases.

$p_{head}(c_1 \mid e_1)$ and $p_{dep}(c_2 \mid e_2)$ cannot be estimated directly because there is no triple-aligned corpus available. Here, we present an approach to estimating these probabilities from two monolingual corpora based on the EM algorithm.

3.3 Estimation of word translation probability using the EM algorithm

Chinese and English corpora are first parsed using a dependency parser, and two dependency triple databases are generated. The candidate English translation set of Chinese triples is generated through a bilingual dictionary and the assumption of strong correspondence of dependency relations. There is a risk that unrelated triples in Chinese and English can be connected with this method. However, as the conditions that are used to make the connection are quite strong (i.e. possible word translations in the same triple structure), we believe that this risk, is not very severe. Then, the expectation maximization (EM) algorithm is introduced to iteratively strengthen the correct connections and weaken the incorrect connections.

EM Algorithm

According to section 3.2, the translation probabilities from a Chinese triple c_{tri} to an English triple e_{tri} can be computed using the English triple language model $p(e_{tri})$ and a translation model from English to Chinese $p(c_{tri} \mid e_{tri})$. The English language model can be

estimated using Equation (4) and the translation model can be calculated using Equation (7). The translation probabilities $p_{head}(c \mid e)$ and $p_{dep}(c \mid e)$ are initially set to a uniform distribution as follows:

$$p_{head}(c \mid e) = p_{dep}(c \mid e) = \begin{cases} \frac{1}{|\Gamma_e|}, & if\ (c \in \Gamma_e) \\ 0, & otherwise \end{cases} \quad (8)$$

Where Γ_e represents the translation set of the English word *e*.

Then, the word translation probabilities are estimated iteratively using the EM algorithm. Figure 1 gives a formal description of the EM algorithm.

Train language model for English triple $p(e_{tri})$ *;*
Initialize word translation probabilities $p_{head}(c \mid e)$ *and* $p_{dep}(c \mid e)$ *uniformly as in Equation (8);*
Iterate
Set $score_{head}(c \mid e)$ *and* $score_{dep}(c \mid e)$ *to 0 for all dictionary entries (c,e);*
for *all Chinese triples* $c_{tri} = (c_1, r_c, c_2)$
for *all candidate English triple translations* $e_{tri} = (e_1, r_e, e_2)$
compute triple translation probability $p(e_{tri} \mid c_{tri})$ *by* $p(e_{tri}) p_{head}(c_1 \mid e_1) p_{dep}(c_2 \mid e_2) p(r_c \mid r_e)$
end for
normalize $p(e_{tri} \mid c_{tri})$ *, so that their sum is 1;*
for *all triple translation* $e_{tri} = (e_1, r_e, e_2)$
add $p(e_{tri} \mid c_{tri})$ *to* $score_{head}(c_1 \mid e_1)$
add $p(e_{tri} \mid c_{tri})$ *to* $score_{dep}(c_2 \mid e_2)$
endfor
endfor
for *all translation pairs (c,e)*
set $p_{head}(c \mid e)$ *to normalized* $score_{head}(c \mid e)$ *;*
set $p_{dep}(c \mid e)$ *to normalized* $score_{dep}(c \mid e)$ *;*
endfor
enditerate

Figure 1: EM algorithm

The basic idea is that under the restriction of the English triple language model $p(e_{tri})$ and translation dictionary, we wish to estimate the translation probabilities $p_{head}(c \mid e)$ and $p_{dep}(c \mid e)$ that best explain the Chinese triple database as a translation from the English triple database. In each iteration, the normalized triple translation probabilities are used to update the word translation probabilities. Intuitively, after finding the most probable translation of the Chinese triple, we can collect counts for the word translation it contains. Since the English triple language model provides context information for the disambiguation of the Chinese words, only the appropriate occurrences are counted.

Now, with the language model estimated using Equation (4) and the translation probabilities estimated using EM algorithm, we can compute the best triple translation for a given Chinese triple using Equations (1) and (7).

4 Collocation translation extraction from two monolingual corpora

This section describes how to extract collocation translation from independent monolingual corpora. First, collocations are extracted from a monolingual triples database. Then, collocation translations are acquired using the triple translation model obtained in section 3.

4.1 Monolingual collocation extraction

As introduced in section 2, much work has been done to extract collocations. Among all the measure metrics, log likelihood ratio (LLR) has proved to give better results (Duning, 1993; Thanopoulos et al., 2002). In this paper, we take LLR as the metric to extract collocations from a dependency triple database.

For a given Chinese triple $c_{tri} = (c_1, r_c, c_2)$, the LLR score is calculated as follows:

$$\begin{aligned} Logl = & a \log a + b \log b + c \log c + d \log d \\ & - (a+b) \log(a+b) - (a+c) \log(a+c) \\ & - (b+d) \log(b+d) - (c+d) \log(c+d) \\ & + N \log N \end{aligned} \quad (9)$$

where,

$a = freq(c_1, r_c, c_2)$,
$b = freq(c_1, r_c, *) - freq(c_1, r_c, c_2)$,
$c = freq(*, r_c, c_2) - freq(c_1, r_c, c_2)$,
$d = N - a - b - c$.
N is the total counts of all Chinese triples.

Those triples whose LLR values are larger than a given threshold are taken as a collocation. This syntax-based collocation has the advantage that it can represent both adjacent and long distance word association. Here, we only extract the three main types of collocation that have been mentioned in section 3.1.

4.2 Collocation translation extraction

For the acquired collocations, we try to extract their translations from the other monolingual

corpus using the triple translation model trained with the method proposed in section 3.

Our objective is to acquire collocation translations as translation knowledge for a machine translation system, so only highly reliable collocation translations are extracted. Figure 2 describes the algorithm for Chinese-English collocation translation extraction. It can be seen that the best English triple candidate is extracted as the translation of the given Chinese collocation only if the Chinese collocation is also the best translation candidate of the English triple. But the English triple is not necessarily a collocation. English collocation translations can be extracted in a similar way.

For each Chinese collocation c_{col}:

a. *Acquire the best English triple translation* $\hat{e}_{tri}$ *using C-E triple translation model*:
$$\hat{e}_{tri} = \arg\max_{e_{tri}} p(e_{tri})\, p(c_{tri} \mid e_{tri})$$

b. *For the acquired* $\hat{e}_{tri}$, *calculate the best Chinese triple translation* $\hat{c}_{tri}$ *using E-C triple translation model*:
$$\hat{c}_{tri} = \arg\max_{c_{tri}} p(c_{tri})\, p(\hat{e}_{tri} \mid c_{tri})$$

c. *If* $c_{col} = \hat{c}_{tri}$, add $c_{col} \Leftrightarrow \hat{e}_{tri}$ *to collocation translation database.*

Figure 2: Collocation translation extraction

4.3 Implementation of our approach

Our English corpus is from Wall Street Journal (1987-1992) and Associated Press (1988-1990), and the Chinese corpus is from People's Daily (1980-1998). The two corpora are parsed using the NLPWin parser[1] (Heidorn, 2000). The statistics for three main types of dependency triples are shown in tables 1 and 2. *Token* refers to the total number of triple occurrences and *Type* refers to the number of unique triples in the corpus. Statistic for the extracted Chinese collocations and the collocation translations is shown in Table 3.

Class	#Type	#Token
VO	1,579,783	19,168,229
AN	311,560	5,383,200
AV	546,054	9,467,103

Table 1: Chinese dependency triples

[1] The NLPWin parser is a rule-based parser developed at Microsoft research, which parses several languages including Chinese and English. Its output can be a phrase structure parse tree or a logical form which is represented with dependency triples.

Class	#Type	#Token
VO	1,526,747	8,943,903
AN	1,163,440	6,386,097
AV	215,110	1,034,410

Table 2: English dependency triples

Class	#Type	#Translated
VO	99,609	28,841
AN	35,951	12,615
AV	46,515	6,176

Table 3: Extracted Chinese collocations and E-C translation pairs

The translation dictionaries we used in training and translation are combined from two dictionaries: HITDic and NLPWinDic [2]. The final E-C dictionary contains 126,135 entries, and C-E dictionary contains 91,275 entries.

5 Experiments and evaluation

To evaluate the effectiveness of our methods, two experiments have been conducted. The first one compares our method with three other monolingual corpus based methods in triple translation. The second one evaluates the accuracy of the acquired collocation translation.

5.1 Dependency triple translation

Triple translation experiments are conducted from Chinese to English. We randomly selected 2000 Chinese triples (whose frequency is larger than 2) from the dependency triple database. The standard translation answer sets were built manually by three linguistic experts. For each Chinese triple, its English translation set contain English triples provided by anyone of the three linguists. Among 2000 candidate triples, there are 101 triples that can't be translated into English triples with same relation. For example, the Chinese triple (讲, VO, 价钱) should be translated into "*bargain*". The two words in triple cannot be translated separately. We call this kind of collocation translation no-compositional translations. Our current model cannot deal with this kind of translation. In addition, there are also 157 error dependency triples, which result from parsing mistakes. We filtered out these two kinds of triples and got a standard test set with 1,742 Chinese triples and 4,645 translations in total.

We compare our triple translation model with three other models on the same standard test set with the same translation dictionary. As the

[2] These two dictionaries are built by Harbin Institute of Technology and Microsoft Research respectively.

baseline experiment, Model A selects the highest-frequency translation for each word in triple; Model B selects translation with the maximal target triple probability, as proposed in (Dagan 1994); Model C selects translation using both language model and translation model, but the translation probability is simulated by a similarity score which is estimated from monolingual corpus using mutual information measure (Zhou et al., 2001). And our triple translation model is model D.

Suppose $c_{tri} = (c_1, r_c, c_2)$ is the Chinese triple to be translated. The four compared models can be formally expressed as follows:

Model A:

$$e_{\max} = (\arg\max_{e_1 \in Trans(c_1)}(freq(e_1)), r_e, \arg\max_{e_2 \in Trans(c_2)}(freq(e_2))$$

Model B:

$$e_{\max} = \arg\max_{e_{tri}} p(e_{tri}) = \arg\max_{\substack{e_1 \in Trans(c_1) \\ e_2 \in Trans(c_2)}} p(e_1, r_e, e_2)$$

Model C:

$$e_{\max} = \arg\max_{e_{tri}}(p(e_{tri}) \times \text{likelyhood}(c_{tri} \mid e_{tri}))$$

$$= \arg\max_{\substack{e_1 \in Trans(c_1) \\ e_2 \in Trans(c_2)}}(p(e_{tri}) \times \text{Sim}(e_1, c_1) \times \text{Sim}(e_2, c_2))$$

where, $\text{Sim}(e, c)$ is similarity score between e and c (Zhou et al., 2001).

Model D (our model):

$$e_{\max} = \arg\max_{e_{tri}}(p(e_{tri})p(c_{tri} \mid e_{tri}))$$

$$= \arg\max_{\substack{e_1 \in Trans(c_1) \\ e_2 \in Trans(c_2)}}(p(e_{tri})p_{head}(c_1 \mid e_1)p_{dep}(c_2 \mid e_2)p(r_c \mid r_e))$$

	Cove-Rage(%)	Accuracy(%)		Oracle (%)
		Top 1	Top 3	
Model A	83.98	17.21	----	66.30
Model B		33.56	53.79	
Model C		35.88	57.74	
Model D		**36.91**	**58.58**	

Table 4: Translation results comparison

The evaluation results on the standard test set are shown in Table 4, where coverage is the percentages of triples which can be translated. Some triples can't be translated by Model B, C and D because of the lack of dictionary translations or data sparseness in triples. In fact, the coverage of Model A is 100%. It was set to the same as others in order to compare accuracy using the same test set. The oracle score is the upper bound accuracy under the conditions of current translation dictionary and standard test set. Top *N* accuracy is defined as the percentage of triples whose selected top *N* translations include correct translations.

We can see that both Model C and Model D achieve better results than Model B. This shows that the translation model trained from monolingual corpora really helps to improve the performance of translation. Our model also outperforms Model C, which demonstrates the probabilities trained by our EM algorithm achieve better performance than heuristic similarity scores.

In fact, our evaluation method is very rigorous. To avoid bias in evaluation, we take human translation results as standard. The real translation accuracy is reasonably better than the evaluation results. But as we can see, compared to the oracle score, the current models still have much room for improvement. And coverage is also not high due to the limitations of the translation dictionary and the sparse data problem.

5.2 Collocation translation extraction

47,632 Chinese collocation translations are extracted with the method proposed in section 4. We randomly selected 1000 translations for evaluation. Three linguistic experts tag the acceptability of the translation. Those translations that are tagged as acceptable by at least two experts are evaluated as correct. The evaluation results are shown in Table 5.

	Total	Acceptance	Accuracy (%)
VO	590	373	63.22
AN	292	199	68.15
AV	118	60	50.85
All	1000	632	63.20
ColTrans	334	241	72.16

Table 5: Extracted collocation translation results

We can see that the extracted collocation translations achieve a much better result than triple translation. The average accuracy is 63.20% and the collocations with relation AN achieve the highest accuracy of 68.15%. If we only consider those Chinese collocations whose translations are also English collocations, we obtain an even better accuracy of 72.16% as shown in the last row of Table 5. The results justify our idea that we can acquire reliable translation for collocation by making use of triple translation model in two directions.

These acquired collocation translations are very valuable for translation knowledge building. Manually crafting collocation translations can be time-consuming and cannot ensure high quality in a consistent way. Our work will certainly improve the quality and efficiency of collocation translation acquisition.

5.3 Discussion

Although our approach achieves promising results, it still has some limitations to be remedied in future work.

(1) Translation dictionary extension

Due to the limited coverage of the dictionary, a correct translation may not be stored in the dictionary. This naturally limits the coverage of triple translations. Some research has been done to expand translation dictionary using a non-parallel corpus (Rapp, 1999; Keohn and Knight, 2002). It can be used to improve our work.

(2) Noise filtering of parsers

Since we use parsers to generate dependency triple databases, this inevitably introduces some parsing mistakes. From our triple translation test data, we can see that 7.85% (157/2000) types of triples are error triples. These errors will certainly influence the translation probability estimation in the training process. We need to find an effective way to filter out mistakes and perform necessary automatic correction.

(3) Non-compositional collocation translation.

Our model is based on the dependency correspondence assumption, which assumes that a triple's translation is also a triple. But there are still some collocations that can't be translated word by word. For example, the Chinese triple (富有, VO, 成效) usually be translated into "*be effective*"; the English triple (*take*, VO, *place*) usually be translated into "发生". The two words in triple cannot be translated separately. Our current model cannot deal with this kind of non-compositional collocation translation. Melamed (1997) and Lin (1999) have done some research on non-compositional phrases discovery. We will consider taking their work as a complement to our model.

6 Conclusion and future work

This paper proposes a novel method to train a triple translation model and extract collocation translations from two independent monolingual corpora. Evaluation results show that it outperforms the existing monolingual corpus based methods in triple translation, mainly due to the employment of EM algorithm in cross language translation probability estimation. By making use of the acquired triple translation model in two directions, promising results are achieved in collocation translation extraction.

Our work also demonstrates the possibility of making full use of monolingual resources, such as corpora and parsers for bilingual tasks. This can help overcome the bottleneck of the lack of a large-scale bilingual corpus. This approach is also applicable to comparable corpora, which are also easier to access than bilingual corpora.

In future work, we are interested in extending our method to solving the problem of non-compositional collocation translation. We are also interested in incorporating our triple translation model for sentence level translation.

7 Acknowledgements

The authors would like to thank John Chen, Jianfeng Gao and Yunbo Cao for their valuable suggestions and comments on a preliminary draft of this paper.

References

Morton Benson. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*. 3(1):23–35

Yunbo Cao, Hang Li. 2002. Base noun phrase translation using Web data and the EM algorithm. *The 19th International Conference on Computational Linguistics*. pp.127-133

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutural information, and lexicography. *Computational Linguistics*, 16(1):22-29

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563-596

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

Hiroshi Echizen-ya, Kenji Araki, Yoshi Momouchi, Koji Tochinai. 2003. Effectiveness of automatic extraction of bilingual collocations using recursive chain-link-type learning. *The 9th Machine Translation Summit*. pp.102-109

Pascale Fung, and Yee Lo Yuen. 1998. An IR approach for translating new words from nonparallel, comparable Texts. *The 36th annual conference of the Association for Computational Linguistics*. pp. 414-420

Jianfeng Gao, Jianyun Nie, Hongzhao He, Weijun Chen, Ming Zhou. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations**.** *The 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*l. pp.183 - 190

G. Heidorn. 2000. Intelligent writing assistant. In R. Dale, H. Moisl, and H. Somers, editors, A

Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text. Marcel Dekker.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. *National Conference on Artificial Intelligence.* pp.711-715

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. *Unsupervised Lexical Acquisition: Workshop of the ACL Special Interest Group on the Lexicon*. pp. 9-16

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. *The 31st Annual Meeting of the Association for Computational Linguistics*, pp. 23-30

Cong Li, Hang Li. 2002. Word translation disambiguation using bilingual bootstrapping. *The 40th annual conference of the Association for Computational Linguistics*. pp: 343-351

Dekang Lin. 1998. Extracting collocation from Text corpora. *First Workshop on Computational Terminology*. pp. 57-63

Dekang Lin 1999. Automatic identification of non-compositional phrases. *The 37th Annual Meeting of the Association for Computational Linguistics*. pp.317--324

Ilya Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. *The 2nd Conference on Empirical Methods in Natural Language Processing*. pp. 97~108

Brown P.F., Pietra, S.A.D., Pietra, V. J. D., and Mercer R. L. 1993. The mathematics of machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-313

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. *The 37th annual conference of the Association for Computational Linguistics.* pp. 519-526

Violeta Seretan, Luka Nerima, Eric Wehrli. 2003. Extraction of Multi-Word collocations using syntactic bigram composition. *International Conference on Recent Advances in NLP*. pp. 424-431

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

Frank Smadja, Kathleen R. Mckeown, Vasileios Hatzivassiloglou. 1996. Translation collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22:1-38

Aristomenis Thanopoulos, Nikos Fakotakis, George Kokkinakis. 2002. Comparative evaluation of collocation extraction metrics. *The* 3rd *International Conference on Language Resource and Evaluation*. pp.620-625

Hua Wu, Ming Zhou. 2003. Synonymous collocation extraction using translation Information. *The 41th annual conference of the Association for Computational Linguistics*. pp. 120-127

Kaoru Yamamoto, Yuji Matsumoto. 2000. Acquisition of phrase-level bilingual correspondence using dependency structure. *The 18th International Conference on Computational Linguistics*. pp. 933-939

Ming Zhou, Ding Yuan and Changning Huang. 2001. Improving translation selection with a new translation model trained by independent monolingual corpora. *Computaional Linguistics & Chinese Language Processing*. 6(1): 1-26