A Graph-based Bilingual Corpus Selection Approach for SMT *

Wenhan Chao[a] and Zhoujun Li[a]

[a] School of Computer Science & Engineering, BeiHang University,
37# Xueyuan Rd, Haidian District, Beijing 100191, China
{chaowenhan, lizj}@buaa.edu.cn

Abstract. In statistical machine translation, the number of sentence pairs in the bilingual corpus is very important to the quality of translation. However, when the quantity reaches some extent, enlarging the corpus has less effect on the translation quality; whereas increasing greatly the time and space complexity to train the translation model, which hinders the development of statistical machine translation. In this paper, we propose a graph-based bilingual corpus selection approach, which makes use of the structural information of corpus to measure and update the importance of each sentence pair, and then selects a sentence pair with the highest importance each time. Our experiments in a Chinese-English translation task show that, selecting only 50% of the whole corpus by the graph-based selection approach as training set, we can obtain the near translation result with the one using the whole corpus.

Keywords: Statistical Machine Translation, Corpus Selection, Graph

1 Introduction

In statistical machine translation, large scale of bilingual corpus is very important. In order to improve the quality of translation, there are two viewpoints about the use of corpus.

One way is to collect more and more bilingual corpus to improve the quality of translation model, such as extracting the sentence pairs from the comparable corpus (Smith et al., 2010; Uszkoreit et al., 2010). However, some researchers found that, after the quantity of the sentence pairs (Han et al., 2009) in the corpus reaches some extent, adding more sentence pairs will not improve the quality of translation significantly. On the other hand, larger and larger corpus will consume more and more resources, which hinders the research progress of machine translation in some degree.

This type of approaches assumes the sentence pairs in the corpus are independent each other, not considering the relationship between sentence pairs and their effect on the translation.

The other view is to mine the potential of training corpus through corpus selection and optimization to improve the quality of the translation model. And it also includes three ways: the first one is to select and optimize the training corpus to adapt to the test set (Lu et al., 2007) or the domain (Yasuda et al,. 2008); the second one is to select the sentence pairs with high quality as training corpus (Chen et al., 2006; Han et al., 2009), in which the quality is measured through the features of the sentence pair itself, such as the number of words that can be translated each other in the sentence pair; the third one is to measure and sort the sentence pairs based on the number of unknown n-grams in the sentences, and then select the sentence pair with the highest scores each time (Eck et al. 2005).

This type of approaches considers the quality difference between the sentence pairs in the corpus. However, it still views the sentence pairs as independent.

* The work reported in this paper was supported by National Natural Science Foundation of China, Contract No. 61003111; and supported by Research Fund for the Doctoral Program of Higher Education of China (New teacher Fund), Contract No. 20101102120016

In this paper, we assume the quality of the translation model is related to the coverage and quality of the selected corpus, and expect to select the sentence pairs with high quality as possible when maximizing the coverage of the selected corpus. And we propose a graph-based bilingual corpus selection approach, which makes use of the structural information of corpus to measure and update the importance of each sentence pair, and then selects a sentence pair with the highest importance each time. The underlying principle is that we should select a sentence pair each time to maximize the coverage and quality of the selected sentence pairs.

In the rest of this paper, we first introduce how to measure the importance of each sentence pair based on the bilingual graph in Section 2, and then describe the framework of graph-based bilingual corpus selection approach in Section 3, emphasizing on corpus selection algorithm. Section 4 shows the results of the experiments, and we conclude in Section 5 and 6.

2 Graph-based Sentence Pair Ranking

2.1 Terminology and Notation

Monolingual Sub-Graph

Assume that the bilingual corpus ***BC*** is composed of a collection of sentence pairs <*f*,*c*>, which consists of two sentences that come from two languages *F* and *C* respectively and can be translated each other.

The monolingual sentences of each language in the collection of sentence pairs will construct an undirected graph, called *Monolingual Sub-Graph*. The two graphs are represented as G_f=<V_f,E_f> and G_c=<V_c,E_c> respectively.

Where G_f=<V_f,E_f> represents the undirected graph constructed by the sentences of language *F* in the corpus, and a node $f \in V_f$ represents a sentence of language *F*, and if the similarity between two sentences f_1 and f_2 are greater than the threshold σ_f, i.e. $sim(e_1, e_2) \geq \sigma_f$, then there exists an edge $(e_1, e_2) \in E_f$.

Similarly, we use G_c=<V_c,E_c> represents the undirected graph constructed by the sentences of language *C* in the corpus, and a nodes $c \in V_c$ represents a sentence of language *C*, and an edge $(c_1, c_2) \in E_c$ represents that the similarity between the two sentences c_1 and c_2 are greater than the threshold σ_c i.e. $sim(c_1, c_2) \geq \sigma_c$.

Bilingual Graph

Bilingual graph is an undirected graph constructed by the sentence pairs in the bilingual corpus ***BC***, represented as G=<$V_{f,c}$, $E_{f,c}$,>, where $v \in V_{f,c}$ represents a sentence pair in the corpus. For each sentence pair <*f*,*c*> $\in V_{f,c}$, it will be $f \in V_f$ and $c \in V_c$.

An edge $(v_1, v_2) \in E_{f,c}$ represents the similarity between two sentence pairs v_1 and v_2 is greater than the threshold $\sigma_{f,c}$, i.e. $sim(v_1, v_2) \geq \sigma_{f,c}$, and the similarity between two sentence pairs can be calculated based on the similarities between the monolingual sentences.

If the node *v* does not connected to any other node, then we call the node *v* as an isolated sentence pair.

Quantity of Information (QI)

Given the set of selected sentence pairs ***S***, the *quantity of information* of a sentence pair is the quantity of the novel information it can provide, i.e. it represents the novelty of the sentence pair.

In the beginning, $S = \phi$, and the quantity of information for each sentence pair will be the largest value. And as the *S* increases, the quantity of information for each unselected sentence

pair will be updated dynamically, removing the redundancy information between *S* and the sentence pair.

The quantity of information for the whole corpus will be the sum of quantity of information for all sentences pairs in the corpus, and it represents the coverage of the selected corpus.

Coverage of Sentence Pair (CSP)

For each unselected sentence pair, the *coverage* is the quantity of redundancy information between the sentence pair and all of the other unselected sentence pairs in the bilingual corpus. And it is the sum of redundancy information between the sentence pair and each unselected sentence pair in the corpus.

In the bilingual graph, the coverage for each sentence pair only considers the sentence pairs that connect to it.

The underlying principle is that the more the number of the similar sentence pairs with high quality is, the better of the quality of the sentence pair is. Thus, CSP represents the quality of the sentence pair.

Importance of Sentence Pair (ISP)

The *importance* of sentence pair consists of two parts: the *quantity of information* (*QI*) and the *coverage* (*CSP*) of the sentence pair.

The underlying assumption is that if the quantity of information and the coverage for the sentence pair is high, then importance of the sentence pair is high.

2.2 Importance Equation

After the bilingual graph has been constructed, our goal is to compute the importance for each sentence pair via the structural information of the corpus, and then make the sentence selection.

The importance of sentence pair (*ISP*) consists of two parts: the quantity of information *QI* and the coverage *CSP*. Given the bilingual corpus *BC* and the set of selected sentence pair *S*, the *QI* for a sentence pair v_a is equal to the initial quantity of information of v_a, represented as QI_0, subtracting the redundancy information contained in the *S* according to v_a:

$$QI(v_a, S) = QI_0(v_a) - RI(S, v_a) \tag{1}$$

Where $QI(v_a, S)$ represents the quantity of information for v_a when given *S*; $QI_0(v_a)$ is the initial *QI* for v_a, i.e. the *QI* for $S = \phi$.

$RI(S, v_a)$ represents the redundancy information contained in *S* according to v_a.

The importance of sentence pair *ISP* is the sum of the quantity of information *QI* and the coverage *CSP*.

$$ISP(v_a, S) = QI(v_a, S) + SRI(v_a, S, BC) \tag{2}$$

$$SRI(v_a, S, BC) = \sum_{\substack{v_b \in BC \\ v_a \neq v_b}} RI(v_a, v_b, S) = \sum_{\substack{v_b \in BC \\ v_a \neq v_b}} sim(v_a, v_b) \cdot QI(v_b, S) \tag{3}$$

Where $ISP(v_a, S)$ represents the importance of v_a when given *S*, $SRI(v_a, S, BC)$ represents the coverage of v_a in corpus *BC*, when given *S* and *BC*. $RI(v_a, v_b, S)$ represents the redundancy information contained in v_a according to v_b, when given *S*; $sim(v_a, v_b)$ represents the similarity between v_a and v_b; and $RI(v_a, v_b, S)$ equals the multiple of the $sim(v_a, v_b)$ and the *QI* of v_b.

Given the bilingual graph *G*, we assume that the redundancy information contained in *S* for v_a only be relevant to the sentence pairs that connect to v_a in *S*.

So, we rewrite the $RI(S, v_a)$ as follows:

$$RI_G(S, v_a) = sim(\bigcup_{v_b \in S} v_b, v_a) \cdot QI_0(v_a) \quad (4)$$

Where $sim(\bigcup_{v_b \in S} v_b, v_a)$ represents the similarity between the union of all of the similar sentence pairs of v_a in S and v_a.

Similarly, we can rewrite $RI(v_a, v_b, S)$ and $SRI(v_a, S, BC)$ as follows:

$$SRI_G(v_a, S, BC) = \sum_{(v_a, v_b) \in E_G} RI_G(v_a, v_b, S) = \sum_{(v_a, v_b) \in E_G} sim(v_b, v_a) \cdot QI_G(v_b, S) \quad (5)$$

Thus, our importance equation will be:

$$QI_G(v_a, S) = QI_0(v_a)[1 - sim(\bigcup_{v_b \in S} v_b, v_a)] \quad (6)$$

$$ISP_G(v_a, S) = QI_G(v_a, S) + \sum_{(v_a, v_b) \in E_G} sim(v_b, v_a) \cdot QI_G(v_b, S) \quad (7)$$

The underlying principle is: given the set of selected sentence pair *S*, the importance of sentence pair will be the sum of the quantity of information itself and the redundancy information between the sentence pair and all the unselected sentence pairs it connects to.

3 Graph-based Bilingual Corpus Selection

3.1 Framework

In order to implement the graph-based bilingual corpus selection, the graph-based bilingual corpus selection consists of three steps as shown in Figure 1:

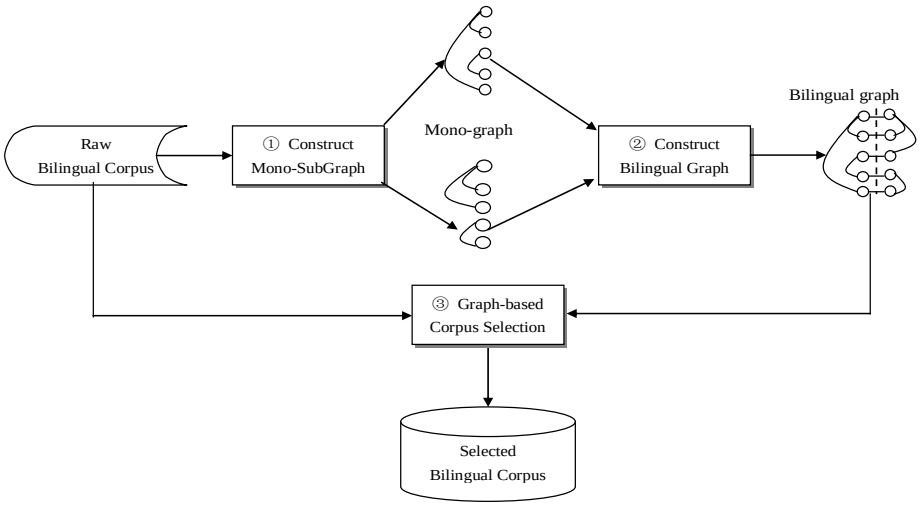

Figure 1: The graph-based corpus selection framework.

3.2 Construct Monolingual Sub-Graph

Measure the Similarity between Sentences

In order to construct the monolingual graph, we need to measure the similarities between the sentences. Considering the corpus is very large, we will take a very simple approach, which counts the number of co-occur words and obtain the similarity as follows:

$$sim(c_1, c_2) = \frac{2 \times |c_1 \cap c_2|}{|c_1| + |c_2|} \quad (8)$$

Where $|c_1 \cap c_2|$ represents the number of the co-occur words between the sentences c_1 and c_2, $|c_1|$ and $|c_2|$ represent the number of words in sentences c_1 and c_2 respectively.

Select the similarity threshold σ_f and σ_c

The similarity threshold will affect the structure of the monolingual graph, and then affect the computing of the importance of the sentence pair. If σ_f or σ_c is too large, it will decrease the edges in the graph, and make the isolated sentence pair increasing; on the other hand, if σ_f or σ_c is too small, there will be too many edges, weakening the ability of distinguishing the importance of sentence pair using the structural information.

Thus, we should try to find a balance between avoiding too many isolated sentence pairs and avoiding too many connected sentence pairs.

3.3 Construct Bilingual Graph

After constructing the two monolingual graphs respectively, we can construct the bilingual graph. We obtain the connection between two sentence pairs in the following way:

- For any two sentence pairs $v_a =< f_a, c_a >$ and $v_b =< f_b, c_b >$, if f_a connects to f_b and c_a connects to c_b in the two monolingual graph respectively, we say v_a connects to v_b, i.e. $(v_a, v_b) \in E_{f,c}$;
- The similarity between two sentence pairs is the average of the two similarities between the monolingual sentences:

$$sim(v_a, v_b) = \frac{1}{2} \times [sim(f_a, f_b) + sim(c_a, c_b)] \tag{9}$$

3.4 Graph-based Bilingual Corpus Selection

After constructing the bilingual graph, we can now make the corpus selection as follows:

1. **Initialize** the $QI_0(v_a) = 1$;
2. **Update** the importance of each unselected sentence pair by Equ.(6) and (7);
3. **Sort** the sentence pairs by the importance;
4. **Select** the sentence pair with the highest importance, and add it to the list S;
5. **Repeat** 2-4, until all of the sentence pairs in the corpus have been selected.
6. **Output** the sentence pairs in the list *S* in order.

When selecting a sentence pair each time, it need updated all the other sentence pairs that the selected sentence pair can reach through a connected path, and then sort all of the sentence pairs again. So the complexity of the selecting algorithm is very high.

We note when selecting a sentence pair, only the sentence pairs that the selected sentence connects will change their quantities of information, so we divide the updating of the importance into two steps:

Step 1: update the quantity of information using the following iterative way:

When t=k=0, $QI_G(v_a, \phi) = 1$, i.e. $QI_0(v_a) = 1$ for all of the v_a;

When t=k+1, assuming the new selected sentence pair is s_n, then only updating the quantity of information of each sentence pair v_a that s_n connects to:

$$QI_G^{k+1}(v_a, S \cup s_n) = QI_G^k(v_a, S) \times [1 - sim(v_a, s_n)] \tag{10}$$

The above equation will compute the quantity of information of v_a approximately.

Note that we set the initial quantity of information of each sentence pair QI_0 as 1 here, that is, all of the sentence pairs have the same *QI* in the beginning.

Step 2: in the selecting step, re-calculate the importance of the sentence pair with the highest importance by equation (7). If the importance has been changed, then sort it and test the next sentence pair with the highest importance; otherwise select the sentence pair, execute Step 1.

Taking the above two steps, it avoids updating and sorting the importance of each sentence pair that the selected sentence pair can reach through a connected path.

Thus the pseudo code of the final algorithm is shown as follows:

```
ALGORITHM:  Graph_Based_BiCorpus_Selection
INPUT:      Bilingual Graph G of Corpus BC
OUTPUT:     Selected Sentence List S
1: S = <> ;
2. FOR each Va in G DO
3.   Calculate the initial ISP_Va by equation (6) and (7);
4.   Insert into the list L in descending order;
5. WHILE L is not empty DO
6:    Va = L. RemoveHead();
7:    ISP_new = CalcISP(Va);    //by equation (7)
8:    IF  ISP_new < ISP_Va
9:       ISP_Va = ISP_new
10:      Insert into L in descending order
11:   ELSE
12:      S = S ∪ {Va}
13:      FOR each Vb that Va connects to DO
14:         Update QT_Vb by equation(10)
15: RETURN S
```

Figure 2: The pseudo code of graph-based corpus selection algorithm.

4 Evaluation

4.1 Data Set

We choose the training set in the Chinese-English news translation task in CWMT2009 as our bilingual corpus. After removing the sentence pairs in which one of the lengths of the monolingual sentences is greater than 50, we obtain a bilingual corpus containing about 2M sentence pairs, represented as *BC*.

In order to tune the translation model, we take one of the development sets in CWMT2009 as our development set, which is the test set in the Chinese-English news translation task in SSMT2007.

We also choose another develop set in CWMT2009 as our test set, the statistics of our data sets are shown in Table 1.

Table 1: The statistics of the data sets.

		Chinese	English
Train. corpus	Sentences	2,378,944	
	Words	34,362,755	34,921,267
	Vocabulary	193309	307095
Dev. Set	Sentences	1002	
	Words	26,285	
Test Set	Sentences	1006	
	Words	27,477	

4.2 Experiments

In order to analyze the efficiency of our graph-based corpus selection algorithm, we implement several different corpus selection algorithms and compare them with our selection algorithm.

In each experiment, we select a subset of the whole bilingual corpus by specifying the ratio via each selection algorithm, and take it as training set to train the translation model, and then compare the translation quality.

We take the state-of-the-art statistical translation system Moses[1] as the decoder, and BLEU as the evaluation metric.

Our experiments are designed as follows:

Method 1: Baseline I (Random Selection)

- Take the whole of the ***BC*** as training corpus to train the translation model;
- Random select specific ratios (10%, 30%, 50%, 60%, 70%, 80%) of the sentence pairs in the BC as training corpora to train the translation models respectively;

Method 2: Baseline II (Unseen gram-based Selection)

- Select the sentence pair with the highest weight each time, and we calculate the weight of the sentence using the $weight_{1,2}$ (Eck et al. 2005), which considered the length of the sentence and bi-grams, and generated the best results as reported in (Eck et al. 2005).
- Select specific ratios (10%, 30%, 50%, 60%, 70%, 80%) of the sentence pairs in the BC as training corpora to train the translation models respectively;

Method 3: Graph-based Corpus Selection (Considering the *QI* Only)

- Take the graph-based Corpus Selection algorithm, but it only consider the quantity of information of the sentence pair, i.e. the importance is equal to the quantity of information. So, the algorithm need not update the coverage.
- Select specific ratios (10%, 30%, 50%, 60%, 70%, 80%) of the sentence pairs in the BC as training corpora to train the translation models respectively;

Method 4: Graph-based Corpus Selection (Considering the *QI* and *CSP*)

- Take the graph-based Corpus Selection algorithm, here the importance is the sum of the quantity of information and the coverage;
- Select specific ratios (10%, 30%, 50%, 60%, 70%, 80%) of the sentence pairs in the BC as training corpora to train the translation models respectively;

Methods 3 and 4 need to construct the monolingual graphs and bilingual graph. We set the similarity threshold as 0.4, and the statistics of the graphs are shown in Table 2.

In table 2, the column 1 represents the three graphs (two monolingual graphs and a bilingual graph), the column 2 is the amount of edges in each graph, the column 3 is the average edge for each sentence or sentence pair in each graph, and the column 4 is the amount of the isolated nodes, which have no similar sentences or sentence pairs in the graph. Note the amount of the points is 2378944.

Table 2: The statistics of the graphs (N=2378944).

	Amount of Edges	Avg. Edge	Amount of Isolated Nodes
Mono. Graph (Chinese)	77,135,825	64.8	445,684 (18.7%)
Mono. Graph (English)	208,614,318	175.4	366,690 (15.4%)
Bilingual Graph	19,731,976	16.6	864,281 (36.3%)

From the table we can see that about 36.3% of sentence pairs in the bilingual graph are isolated sentence pairs. So we should adjust the thresholds to avoid so many isolated sentence pairs in the future.

After selecting the subsets of the corpus with specific ratios, the statistics of them are shown as Table 3.

[1] http://www.statmt.org

Table 3: The statistics of the subsets of the corpus.

Ratios	Method 1		Method 2		Method 3		Method 4	
	Avg.	OOV	Avg.	OOV	Avg.	OOV	Avg.	OOV
10%	14.4	389	11.9	191	15.2	361	13.5	359
30%	14.4	228	14.0	150	15.6	189	16.2	261
50%	14.4	186	14.8	148	15.8	158	16.3	156
60%	14.4	176	15.4	148	15.7	151	15.6	151
70%	14.4	165	14.6	148	15.1	150	15.2	149
80%	14.4	165	14.2	148	14.6	148	14.8	148
100%	14.4	148						

In table 3, the column 1 represents the specific ratios, the columns 2 to 5 represents the four selection methods, and each of them consists of two sub-columns, average length of the source sentences (Avg.) and the number of out of vocabulary words (OOV).

From the table, we can see that when using the random selection, the average lengths of the sentences are all near to the average sentence lengths of the whole corpus.

Both in method 2, 3 and 4, the numbers of OOV words decrease very quickly, and it can reflect the coverage of the selected corpus.

Finally, we obtained the translation results shown in Table 4.

Table 4: The test results for different selection methods(=BLEU%).

Ratios	Method 1	Method 2	Method 3	Method 4
10%	18.84	**20.03**	19.32	19.51
30%	19.91	20.68	**20.78**	20.30
50%	20.76	21.08	21.10	**21.25**
60%	20.96	21.00	21.00	**21.34**
70%	21.14	21.26	**21.54**	21.27
80%	21.25	21.26	21.30	**21.58**
100%	21.51			

In table 4, the column 1 represents the specific ratios, the columns 2 to 5 represents the BLEU% scores for four selection methods.

The results from the table 4 show, given the specific ratios of the training corpus, using unseen gram-based selection (Method 2) and graph-based corpus selection methods (Method 3 and 4) will obtain better results than using the random selection method.

The results obtained by Method 2 and Method 3 are similar, since both of them consider the *QI* when given the set of selected sentence pairs *S* only.

However, when comparing the Method 2 and 3 with Method 4, we find Method 2 and 3 obtained better results when selecting only 10% and 30% of the corpus, and after increasing the ratios, method 4 obtains the better results, especially it obtains the best results when selecting only 80% of the corpus.

We conclude that the quality of the translation model depends on both of the quality and coverage of the bilingual corpus. In the beginning, Method 2 and 3 obtain better coverage than Method 4 (see the number of the OOV in table 3); however, as enlarging the corpus, Method 4 can get similar coverage with Method 2 and 3, but it can obtain better sentence pairs, since it selects the sentence pair with best importance each time. Thus, it generates better results later.

And we can also see from the table 4 that, when using the method 4 to make corpus selection, selecting only 50% of the bilingual corpus will generate near result with selecting 60%~100% of the whole corpus. That is, when using more than 50% of the whole corpus, it does not obtain

significant improvement. This shows the efficiency of our graph-based corpus selection approach.

Especially, when selecting 80% of the corpus, it will get the best results, overcoming the result using the whole corpus. This shows that there may be noisy data in the corpus, which decreases the quality of the translation, and Method 4 could filter the noisy sentence pairs.

5 Related Work

In statistical machine translation, there are three ways to make effective use of the bilingual corpus. One is assuming the sentence pair has different effects on the translation, so it need estimate the quality of each sentence pair, and sort them.

Chen et al. (2006) provided a quality sorting model for the sentence pair, which estimated the quality of each sentence pair via many features, such as language model, sentence length, word alignment etc. Their experiments showed that, when using the same number of sentence pairs as training corpus, selecting the sentence pairs with high quality would improve the quality of translation.

Han et al. (2009) provided another approach. They divided the sentence pairs in the corpus into two types: literal translation and free translation, the first was low-level word-word translation, and the latter was high-level translation. They assumed that SMT could be viewed as low-level translation system, which should be supervised by the sentence pairs with literal translation. So they provided word-match metric and grammar-match metric to find the sentence pairs with literal translation, and selected them as training corpus. Their experiments showed that, when taking the sentence pairs with literal translation as baseline, adding the sentence pairs with free translation would not improve the translation quality all the while.

These approaches considered the difference between the qualities of sentence pairs. However, they only used the features of each sentence pair itself, and the quality would not be updated as the selection process.

Our approach measures the importance of each sentence pair, which only uses the structural information, i.e. the relationship between sentence pairs, and the importance will be updated dynamically during the selection.

The other way is to select and optimize the corpus according to the test set. Lu et al. (2007) proposed the corpus selection and optimization approaches based on the information retrieval methods. The first one retrieved the similar sentence pairs in the corpus according to the test set, and took them as the training corpus; the latter increased the occur number for each similar sentence pair, so that the importance of the similar sentence pair in the translation model will be enlarged. These approaches made the translation model more adaptive to the test set. Their experiments showed the improvement in the translation quality.

The graph-based selection approach in this paper does not consider the test set at all; however, it just considers the corpus itself, especially the structural information within the corpus.

Eck et al. (2005) provided a simple way to sort and select the sentence pairs based on the number of unseen n-grams in the selected data set. The approach considered only the quantity of information between the selected data set and the unselected sentence pair.

6 Conclusion

In this paper, we proposed a graph-based bilingual corpus selection framework, which measures and updates the importance of each sentence pair based on the structural information of the bilingual corpus, and then selects the sentence pair with the highest importance each time, until it obtains the subset of the corpus with specific ratio.

Experiments showed that, through selecting only 50% of the corpus, we can obtain near translation quality with the whole corpus using the graph-based selection approach. We can

even obtain better results than the whole corpus when selecting 80% of the corpus, which suggests that the corpus may contain noisy data and decrease the quality of the translation.

Thus, through the graph-based corpus selection approach, we can select only a part of corpus to train the translation model, which will reduce the time and space complexity largely when building machine translation system, while not decreasing the translation quality significantly.

However, since our approach is just a basic framework, we will improve the following issues in the future:

- Considering more effective approaches to build bilingual graph;
- Improving the graph-based selection algorithm, such as considering the difference between the sentence pairs' QI_0;
- Combining the other features with the structural feature to measure the importance.

References

Chen, Y.D., X.D. Shi and C.L. Zhou. 2006. Research on Filtering Parallel Corpus: A Ranking Model. *Journal of Chinese Information Processing*, Vol.20 Supplement, pp.66-70, 2006.

Eck, M., S. Vogel and A. Waibel. 2005. Low cost portability for statistical machine translation based on n-gram coverage. *Conference Proceedings: the tenth Machine Translation Summit (MT Summit X)* pp.227-234, *Phuket, Thailand, September 13-15,*2005.

Han, X.W., H.Z. Li and T.J. Zhao. 2009. Train the machine with what it can learn – corpus selection for SMT. *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora*, Suntec, Singapore, 6 August 2009; pp.27-33, 2009.

Lü, Y.J., J. Huang and Q. Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. *EMNLP-CoNLL-2007: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June 28-30, 2007, Prague, Czech Republic; pp. 343-350, 2007.

Moore, R.C. and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. *Proceedings of the ACL 2010 Conference Short Papers*, pp. 220–224, Uppsala, Sweden, 11-16 July 2010.

Smith, J.R., C. Quirk and K. Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 403–411, Los Angeles, California, June 2010.

Yasuda. K.J., R.Q. Zhang, H. Yamamoto and E. Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. *IJCNLP 2008: Third International Joint Conference on Natural Language Processing*, January 7-12, 2008, Hyderabad, India; pp.655-660.

Yasuda. K.J., H. Yamamoto and E. Sumita. 2007. Method of selecting training sets to build compact and efficient language model. *MT Summit XI Workshop: Using corpora for natural language generation: language generation and machine translation*, 11 September 2007, Copenhagen, Denmark; pp.31-37, 2007.

Uszkoreit, J., J.M. Ponte, A.C. Popat, and M. Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1101–1109, Beijing, August 2010.