An Efficient Framework to Extract Parallel Units from Comparable Data

Lu Xiang, Yu Zhou, and Chengqing Zong

NLPR, Institute of Automation Chinese Academy of Sciences, Beijing, China
{lu.xiang,yzhou,cqzong}@nlpr.ia.ac.cn

Abstract. Since the quality of statistical machine translation (SMT) is heavily dependent upon the size and quality of training data, many approaches have been proposed for automatically mining bilingual text from comparable corpora. However, the existing solutions are restricted to extract either bilingual sentences or sub-sentential fragments. Instead, we present an efficient framework to extract both sentential and sub-sentential units. At sentential level, we consider the parallel sentence identification as a classification problem and extract more representative and effective features. At sub-sentential level, we refer to the idea of phrase table's acquisition in SMT to extract parallel fragments. A novel word alignment model is specially designed for comparable sentence pairs and parallel fragments can be extracted based on such word alignment. We integrate the two levels' extraction task into a united framework. Experimental results on SMT show that the baseline SMT system can achieve significant improvement by adding those extra-mined knowledge.

Keywords: statistical machine translation, comparable corpora, two-level parallel units extraction, parallel sentences, parallel sub-sentential fragments.

1 Introduction

Parallel corpus is an important resource in many natural language processing tasks, such as statistical machine translation (SMT) and cross-lingual information retrieval. Especially for SMT system, the size and quality of the training data has a vital impact on its performance. However, parallel corpora are always very limited in size, domain, and language pairs. Moreover, it is impractical to build such parallel corpora manually for it will take enormous human material and financial resources. Hence, we have to shift our attention to the large amount of available resources from the Internet and try to extract useful information automatically.

While parallel data are very scarce, comparable corpora are much more available and diverse. To alleviate the lack of parallel data, many methods have been proposed to extract parallel text from comparable resources. These works include identifying parallel sentences[3,8,15,18,19,20] and finding parallel fragments[9,12,13].

For the task of identifying parallel sentences, most of the previous work adopts the classification method. However, the performance of various features has not been investigated in-depth, which cannot achieve a good balance on classification accuracy

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 151–163, 2013.

and speed. For the extraction of parallel fragments, most of the previous work has not found an appropriate method to build an alignment model for the comparable corpus. However, this alignment model is essential for the performance of the subsequent parallel fragments extraction. Furthermore, the previous solutions are restricted to extract either sentential or sub-sentential fragments. In reality, it's very common that both of them do coexist. In this case, existing approaches fail to extract both sentences and fragments at the same time, which will lead lots of useful resources unexploited.

Therefore, in this paper, we propose a two-level parallel text extraction framework which can extract both parallel sentential and sub-sentential fragments at the same time. At sentence level, a classifier is used to identify whether the bilingual sentence pair is parallel or not. We investigate the impact of different groups of features on the performance of classification and choose one set of features that can give us better performance. Moreover, at sub-sentential level, a novel word alignment model for comparable sentence pairs is presented, and a new approach is proposed to extract sub-sentential fragments from comparable sentences using our word alignment model. We applied our framework to the extraction of Chinese-English parallel units from Wikipedia. Experiments show that our framework can extract all parallel units at both sentential and sub-sentential level which can help to improve the translation quality significantly when adding the extracted data to large-scale Chinese-English training data.

The remainder of this paper is organized as follows: Section 2 introduces the related work. Section 3 gives our two-level parallel text extraction framework. Section 4 presents the experiments on data extraction and SMT results. Finally, we conclude the paper in Section 5.

2 Related Work

Much research work has been done on the task of mining parallel units from comparable corpus. Comparable corpora may contain parallel documents, parallel sentences, or parallel sub-sentential fragments. In existing approaches, the parallel sentence extraction process is often divided into two steps: (1) identify document level alignment, and (2) detect parallel sentences within the identified document pairs. [8] uses cross-lingual information retrieval methods to get more precious article pairs and a maximum entropy classifier is used to extract parallel sentences from the article pairs. [15] exploits "inter-wiki" links in Wikipedia to align documents. [18] implements hash-based algorithms to find cross-lingual article pairs efficiently. [19] extends parallel sentence alignment algorithm to align comparable news corpora. [3] calculates pair-wise cosine similarities to identify parallel sentence pairs. [17] extends the classification approach and adopts a beam-search algorithm to abandon target sentences early during classification.

Furthermore, not all comparable documents contain parallel sentence pairs but they could still have plenty of parallel sub-sentential fragments. Typical phrase extraction method used in SMT only works on parallel sentences and it only collects phrase pairs that are consistent with the word alignment [4, 10]. However, this typical method is not suitable for comparable sentence pairs since it will produce lots of non-parallel phrases due to the un-aligned parts. [9] firstly attempts to detect parallel sub-sentential fragments in comparable sentences and they use a word-based signal filter method to extract sub-sentential fragments. Since the source and target signals are filtered separately, there is no guarantee that the extracted sub-sentential fragments are translations of each other. [12] proposes two generative models for segment extraction from comparable sentences, but they don't show any improvements when applied to in-domain test data for MT. [13] uses a hierarchical alignment model and its derivation trees to detect parallel fragments. The experiments show their method can obtain good fragments but their method need some extra data like gold-standard alignments and parse trees for source and target sentences to train the alignment model.

3 Two-Level Parallel Units Extraction Framework

Our two-level parallel text extraction framework is shown as Fig. 1. Since our work mainly focuses on parallel sentences identifying and parallel sub-sentential fragments detecting, we suppose the given documents have been already aligned before.

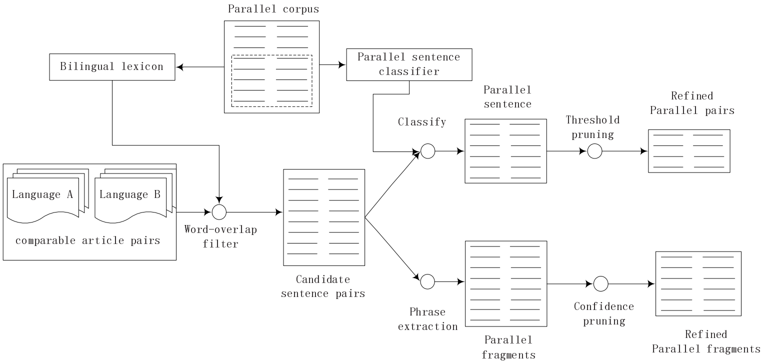

Fig. 1. Two-level parallel unit extraction framework

The resource our framework needed is only a small amount of parallel data used to train the classifier. The bilingual dictionary used for the word-overlap filter process is automatically learned from the small amount of data. First, we obtain a GIZA-lexicon by running GIZA++[1] from two directions on the corpus. Considering the GIZA-lexicon contains too much noise, we adopt log-likelihood-ratios[2,6,7] (LLR) statistic in two directions to get LLR-lexicon. This statistic can estimate the correlation of word pairs and can be calculated using the formula below:

$$2\log\left[\frac{P(y\mid x)^{C(x,y)}\cdot P(y\mid\neg x)^{C(\neg x,y)}\cdot P(\neg y\mid x)^{C(x,\neg y)}\cdot P(\neg y\mid\neg x)^{C(\neg x,\neg y)}}{P(y)^{C(y)}\cdot P(\neg y)^{C(\neg y)}}\right] \tag{1}$$

[1] http://www.statmt.org/moses/giza/GIZA++.html

In the formula, x and y represent two words for which we wish to estimate the strength of association. $C(y)$ and $C(\neg y)$ are the observed frequencies of y occurring or not occurring in the corpus; $C(x, y), \ldots, C(\neg x, \neg y)$ are the joint frequencies of the different possible combinations of x and y occurring and not occurring; and $p(y), p(\neg y), p(y|x), \ldots, p(\neg y|\neg x)$ are the maximum likelihood estimates of the corresponding marginal and conditional probabilities[7].

3.1 Candidate Sentence Pairs' Selection

We use a method based on word-overlap to filter the candidate sentence pairs. For each article pairs, we take all possible sentence pairs and pass them through the word-overlap filter. It checks that the percentage of words that have a translation in the other sentence must be above a threshold according to LLR-lexicon. Any sentence pairs that don't meet the condition will be discarded. This step can help to remove most unrelated sentence pairs and the remaining sentences will be passed on to the parallel sentences identifying and parallel fragments detecting stage.

3.2 Parallel Sentences' Identification

We use a binary classifier to identify whether a sentence-pair is parallel or not. Considering that the maximum entropy model (ME) proves to be effective in [8], we adopt the maximum entropy model in our classifier.

3.2.1 Establishment of Training Data

As Fig. 1 shows, the training data of the classifier is from the seed parallel sentence corpus. We randomly choose 5,800 parallel sentence pairs[2] as positive instances and then generate non-parallel sentence pairs from those sentence pairs. This will generate (5800^2-5800) non-parallel sentence pairs. Since we only apply classifier on the sentence pairs passing the word-overlap filter, we randomly select 5,800 negative instances under the same condition. After training data is prepared, we use the features we proposed to train parallel sentence classifier.

3.2.2 Selection of Features

Features are quite important to the performance of classifier. Besides the features used in [8], we believe that there must be some other useful features. After comparing parallel and non-parallel sentence pairs, we find the following three features should be useful: strong translation sentinels[20] , log probability of word alignment and pair-wise sentences cosine similarity.

Log probability is word alignment score and pair-wise sentences cosine similarity can tell us the similarity between two sentences. Intuitively, if two sentences are translations of each other, the translated words should distribute in the whole sentence. Strong translation sentinels come from that intuition. In order to extract this feature conveniently, we only take the content words near the beginning and end

[2] We choose 5800 parallel sentence pairs in our experiment. This can be adjusted.

of the sentences into consideration (the first and last two content words) and count the number of these four words that have a translation on the other side. These three features are light-weight features and can be computed efficiently.

To sum up, we will use the following features[3] to train our classifier:

- The features in [8]:
  1) General features: Length feature; Translation coverage feature;
  2) Features derived from IBM Model-1 alignments.
- Our new features: 1) Log probability of the alignment; 2) Pair-wise sentences cosine similarity; 3) Strong translation sentinels.

We will investigate the impact of all above features to the classification performance in our experiment. By analyzing the impact, we can choose a group of features that can best meet our needs and train a better classifier for the parallel sentence extraction task.

After passing the candidate sentence pairs through the classifier, we extract those pairs that are classified as positive and finally we only keep the ones with classification confidence higher than a predefined threshold.

3.3 Parallel Sub-sentential Fragments' Extraction

3.3.1 Alignment Model

Since the comparable data is huge, we need a simple but efficient word alignment model to get the word alignment. IBM Model-1[1] is one such model. But it has some shortcomings like it often produces vast false word correspondences in high frequency words. Inspired by YAWA[16], we propose a modified IBM Model-1 that uses bilingual translation lexicons in two stages to obtain a better alignment.

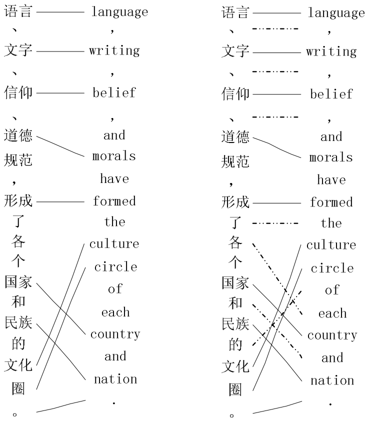

Fig. 2. An example of our modified IBM Model-1. The solid lines in the left are the links of content words obtained in the first step. The dotted lines are the new added links of function words in the second step.

[3] We use all the features with numerical value in ME classifier.

Phase 1: Content Words Alignment. Content words can represent the skeleton of one sentence and the links between content words are much more reliable than those between function words[4]. First, we compute the content words alignment by greedily linking each content word in source sentence with its best translation candidate which has the highest log-likelihood (LL) association score from target sentence according to LLR-lexicon. If a pair of words doesn't exist in the LLR-lexicon or some words appear more than once in the target sentence, we just leave such words and align them in the second stage since the existing links in the first stage can give contextual information for the unaligned words. If the end punctuations are translations, we also link them. The left in Fig. 2 exemplifies the links created in end of the first phase.

Phase 2: Function Words Alignment. The existing links in stage 1 can be treated as phrase boundaries. In this stage, we will heuristically match the blocks and add links to the unaligned words. The algorithm is given as follows: align two blocks if the surrounding words are already aligned and then align words in the two blocks. We illustrate this situation in Fig. 2. For example, we have (了(le), 各(ge), 个(ge)) unaligned in the Chinese side after phase 1. Its surrounding words are "形成(xingcheng)" and "国家(guojia)" which are aligned with "formed" and "country" respectively. Thus, the chunk (了, 各, 个) is aligned with block (the, culture, circle, of, each) and then we can align words in these two blocks. We can align the rest unaligned words in the same way. The new links are shown in the right side of Fig. 2 by the dotted lines.

3.3.2 Parallel Sub-sentential Fragments' Detection

Fig. 3 gives an example of a comparable sentence pair that contains two parallel sub-sentential fragments. In this part, we are concerned about finding such fragments.

以 " 一 国 两 制 " 的 办法 解决 台湾 问题 ， 才 能 最大 限度 地 寻求 两 岸 利益 的 公分母 。
we have hoped for many years to use the formula of " one country , two systems " to peacefully resolve the taiwan issue .

Fig. 3. An example of comparable sentences containing parallel sub-sentential fragments

To detect the sub-sentential fragments, we generate word alignments using our modified IBM Model-1 from bi-directions for each comparable sentence pair and then use intersection [4] method to obtain better alignments. After that, we traverse the word alignment and extract phrases that satisfy the following constraints:

(a) The length of the source and target phrase is no less than 3;
(b) Words inside the source phrase can only be aligned with words inside the target phrase and the same for the words outside the phrase boundaries;
(c) The phrase span can contain a small percentage of content words without any connection;
(d) The unaligned boundary word of the phrase can only be function words.

[4] We count words appearing in the text we will experiment with and take the 100 most frequent words on each side as function words. The remaining words are treated as content words.

When traversing the word alignment, it will often produce short fragments which are not really parallel. And the fragment is more confident when it is longer. Constraint (a) helps us to exclude phrases with less than 3 words and extract long fragments. Constraints (b) and (c) are the content constraints. Constraint (b) limits us from extracting phrases that contain words aligned outside the phrases. Because of the coverage of the LLR-lexicon, not every word can find its translation in the other side. Constraint (c) allows us to extract fragments with a small percentage of words unaligned and such constraint can help us obtain some new lexicon entries. Constraint (d) decides whether a phrase can be extended or not. This makes some function words like “the” and “of” in English and “的(de)” and “了(le)” in Chinese can be included in the phrases.

After sub-sentential fragments extraction, we also need to do some pruning to get better fragments. We filter those low-frequent fragment pairs.

4 Experiments

We evaluate our two-level extraction framework following the steps blow: (1) investigating the impact of different features used in our classifier on the extraction of parallel sentence pairs, (2) describing the extraction of sub-sentential fragments, and (3) evaluating the impact of the extracted data on SMT performance.

4.1 Experiments Setup

We perform our two-level extraction method on Chinese and English Wikipedia articles. We download the dump files from Wikimedia dump[5]. For the seed corpora, we use a large bilingual training data from LDC corpus[6]. Since Wikipedia is a different domain from our seed corpora, we extract article titles from Wikipedia documents as the additional baseline training data. Thus, the initial parallel data consists of large bilingual data from LDC corpus and Wikipedia article title pairs and we denote it as LDC&WikiTitle. Table 1 shows the relevant statistics of LDC&WikiTitle.

Table 1. Statistics of our initial parallel corpus

	Parallel initial corpus	sentences	tokens
Corpus from LDC	Chinese	2,085,331	27,640,151
	English	2,085,331	31,826,668
Wikipedia title pairs	Chinese	236,565	639,813
	English	236,565	642,743
LDC&WikiTitle	Chinese	2,321,896	28,279,964
	English	2,321,896	32,469,411

[5] `http://dumps.wikimedia.org/backup-index.html`

[6] LDC category number: LDC2000T50, LDC2002E18, LDC2003E07, LDC2004T07, LDC2005T06, LDC2002L27, LDC2005T10 and LDC2005T34.

To obtain LLR-lexicon, we run GIZA++ with default setting on LDC&WikiTitle from two directions, use "grow-dial-final-and" strategy to combine the alignments and then keep lexicon with LLR score above 10. Finally, we obtain 4,631,477 GIZA-lexicon and 325,001 LLR positive lexicons.

4.2 Experimental Results on Parallel Sentences Extraction

We create training data from the large bilingual data of LDC as description in Subsection 3.2.1. A maximum entropy binary classifier using the Maximum Entropy Modeling Toolkit[7] is used as the classifier model.

4.2.1 Evaluation of Various of Features

In order to investigate the influence of features on the performance of ME classifier, we use the same method to build a test set of 1,000 positive instances and 1,000 negative instances from CWMT'2011 data[8]. We first test the features in [8] by incrementally adding features from F1 to F8. Table 2 reports the performance evaluation.

Table 2. Evaluation on different features. F2-F5 use only bidirectional alignments. F6-F8 add three combinations (intersection, union, refined) of the bi-directional alignments respectively.

Features	Precision	Recall	F-score
F1:General features	0.7021	0.608	0.6516
F2: +no connection	0.7414	0.889	0.8085
F3: +fertility	0.7395	0.9	0.8119
F4: +contiguous connected span	0.7300	0.933	0.8191
F5: +unconnected substring	0.7288	0.941	0.8214
F6: +intersection	0.7344	0.91	0.8128
F7: +union	0.7435	0.928	0.8256
F8: +refined	0.7606	0.896	0.8227

From Table 2, the classifier evaluation results show that the most useful features are F1 and F2. The other features are relatively difficult to calculate and don't help to improve the performance a lot. Due to the large amount of comparable data, we give up these features. Instead, we use three other features which are very easy to compute and the evaluation scores are given in Table 3. In order to compare the real performance of our features with those features in [8], we give the detailed scores both on speed and accuracy presented in Table 4.

[7] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[8] The 7th China Workshop on Machine Translation Evaluation.

Table 3. Evaluation on our added features

Features	Precision	Recall	F-score
F1:General features	0.7021	0.608	0.6516
F2:+no connection	0.7414	0.889	0.8085
F9:+log probability	0.7429	0.893	0.8110
F10:+cosine similarity	0.7439	0.895	0.8125
F11:+sentinels	0.7466	0.899	0.8157

Table 4. Performance comparison under different features

Features	F-score	Time (s)
F1-F5	0.8214	2259.26
F1-F8	0.8227	4097.63
F1+F2+F9+F10+F11	0.8157	403.45

Table 4 presents the time it costs to extract different set of features from 100,000 sentence pairs. Using F1-F8 cost nearly twice the time of using F1-F5, but the F-score improvements are slight. Our features have achieved the comparable performance to the features in [8], and it is much faster to extract these features. Thus we finally use the features in Table 3 to train a ME classifier and we can affirm that our classifier will be much efficient than Munteanu and Marcu's.

4.2.2 Parallel Sentences' Extraction

We use "inter-wiki" links to align Wikipedia article pairs and finally obtain 199,984 article pairs. Then we use word-overlap filter process to obtain the candidate sentence pairs based on LLR-lexicon. Here sentence pairs with translation coverage above 0.3 are maintained. Then we apply our ME classifier to the candidate sentence pairs and extract sentence pairs with confidence higher than 0.75. The amount of candidate sentence pairs and extracted parallel sentence pairs are shown in Table 5.

Table 5. Size of extracted parallel text

	Chinese-English
Candidate sentence pairs	2,101,770
Parallel sentence pairs	201,588
Parallel fragments	7,708,424

4.3 Experimental Results on Parallel Sub-sentential Fragments Extraction

We employ the method described in Sub-Section 3.3 to extract parallel sub-sentential fragments. We use the sentence pair shown in Fig. 3 to illustrate the extraction procedure. The word alignment result is shown in Fig. 4 and part of the extracted fragments is presented in Table 6.

From Table 6, we can see that our method can find the parallel sub-sentential fragments in comparable sentence pairs. Our method is quite simple and easy to be implemented. We apply this method to the candidate sentence pairs and the results are shown in Table 5. The number of the extracted sentence pairs is 201,588 and it makes up about 10% of the candidate sentence pairs. The amount of the fragments is 7,708,424 and it is nearly 40 times larger than the parallel sentences. This also shows the importance of the exploration of the parallel fragment resource.

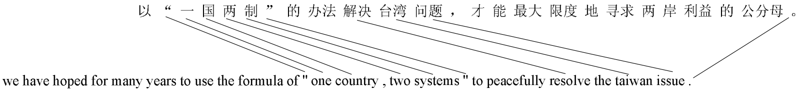
以 “ 一 国 两 制 ” 的 办法 解决 台湾 问题 ， 才 能 最大 限度 地 寻求 两 岸 利益 的 公分母 。
we have hoped for many years to use the formula of " one country , two systems " to peacefully resolve the taiwan issue .

Fig. 4. Word alignment result using intersection strategy to combine bi-directional alignment

Table 6. Examples of the extracted fragments

Chinese	English
一 国 两 制	one country , two systems
一 国 两 制 ”	one country , two systems "
一 国 两 制 ” 的 办法 解决 台湾 问题	one country , two systems " to peacefully resolve the taiwan issue
两 制 ” 的 办法 解决 台湾 问题	two systems " to peacefully resolve the taiwan issue
制 ” 的 办法 解决 台湾 问题	systems " to peacefully resolve the taiwan issue
解决 台湾 问题	resolve the taiwan issue

4.4 Experimental Results on SMT Performance

We use LDC&WikiTitle as our baseline SMT training data. We adopt GIZA++ and grow-diag-final-and for word alignment, Moses toolkit[5] with default settings to train the SMT systems. Language model is trained on the English part of LDC&WikiTitle plus Wikipedia data with SRILM[14]. We have two development and testing sets: one is the 790 parallel sentence pairs manually selected from Wikipedia and half of them as development set (Dev A) and the rest as test set (Test A); the other is NIST MT 2003 evaluation data as the development set (Dev B) and NIST MT05 as test set (Test B). Translation performance is evaluated using BLEU metric[11].

In order to show the advantage of our two-level parallel text extraction method, we conduct the following evaluation respectively: (1) only adding the extracted parallel sentences, (2) only adding the extracted parallel sub-sentential fragments, and (3) adding both the extracted parallel sentences and sub-sentential fragments to the original corpora and then evaluate the impact to an end-to-end SMT system.

As Table 7 shows, the translation performance on Test A is much lower than that on MT05 under the baseline system. This is because the LDC&WikiTitle corpus is mainly consisted of news data and it can only learn less knowledge about Wikipedia. When adding the extracted parallel text, the SMT performance on Test A has been improved significantly (+16 BLEU points for extracted sentences and +20 points for extracted fragments). It is due to the extra data can provide much translation knowledge about Wikipedia. This means that our parallel sentences and fragments extraction method is useful for machine translation.

Table 7. SMT evaluation results

	Test A	**Test B**
baseline	24.49	29.96
baseline +extracted sentence (201,588 sentence pairs)	41.31	30.84
baseline+ extracted fragment (7,708,424 fragments)	45.20	30.21
baseline + sentence + fragment	50.52	30.23

However, the extracted fragments don't seem to be much too useful for MT05 by only slightly improvement for MT05. Intuitively, this could come down to the following reasons: (1) MT05 is in news domain, which is a different domain from Wikipedia. However, domain adaptation is a very difficult research task. (2) The fragments are much more than the baseline corpus which may be drowned by the new added data. We also conduct the evaluation on MT06. The evaluation results show 0.24 improvements when adding the extracted sentences and 0.27 improvements when adding both extracted sentences and fragments compared to the baseline.

In order to find why the effect is not so obvious on Test B, we give the statistical information of LLR positive lexicon size with different training corpus in Table 8. From Table 8 we can see that the new lexicon entries don't increase too much by the extracted sentences (19,067, 5.87%) but increase dramatically by the extracted fragments (364,613, 112.2%). This illustrates that the extracted sentences are very close to the baseline training data LDC&WikiTitle, thus the translation result is improved relatively significantly by 0.88 BLEU score. However, for the extracted fragments, it has two times of lexicon entries compared to the LDC&WikiTitle, so those two corpora have much difference in domain and the original data may be drowned by the new added fragments. Consequently the SMT performance did not meet our expectations on MT05 and MT06. In this respect, it is not wise to add all fragments directly to the LDC&WikiTitle to train translation model. We have to do further experiments to inspect and verify how to add our resource appropriately to the baseline SMT model as a beneficial supplement. However, this can also show our new framework is able to help mine much more useful information for SMT from another aspect.

Table 8. LLR positive lexicon size

	LLR positive lexicon size
LDC&WikiTitle corpus	325,001
LDC&WikiTitle corpus + extracted sentence	344,068
LDC&WikiTitle corpus+ extracted fragment	678,652
LDC&WikiTitle corpus + sentence + fragment	689,614

5 Conclusion

In this paper, we propose a simple and effective two-level parallel unit extraction method for extracting both sentential and sub-sentential parallel text from comparable corpora. At sentential level, we treat the task of identifying parallel sentences as a classification problem and investigate the impact of different features in detail to find the best group of features. For sub-sentential fragment, we developed a novel word alignment model for comparable sentence pairs and describe how to extract parallel sub-sentential fragments based on such word alignment. Our new framework can help us extract much more useful information with a good trade-off performance at accuracy and speed. We applied our framework to the extraction of Chinese-English parallel units from Wikipedia used for Chinese-to-English SMT. Experimental results show that it can improve the translation quality significantly by adding the extracted data to large-scale Chinese-English training data.

In the next step, we will study on the method of how to evaluate the confidence of the extracted sentences and fragments and how to use those mined parallel fragments appropriately for a SMT system with a given domain.

Acknowledgement. The research work has been partially funded by the Natural Science Foundation of China under Grant No. 61003160 and Hi-Tech Research and Development Program (“863” Program) of China under Grant No. 2011AA01A207, and also supported by the Key Project of Knowledge Innovation Program of Chinese Academy of Sciences under Grant No. KGZD-EW-501.

References

1. Brown Peter, F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of machine translation: Parameter estimation. Computational Linguistics 19(2), 263–311 (1993)
2. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1), 61–74 (1993)
3. Fung, P., Cheung, P.: Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and EM. In: EMNLP 2004, pp. 57–63 (2004a)
4. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase based translation. In: Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL (2003)
5. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R.-C., Dyer, C., Bojar, O.: Moses: Open source toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 177–180 (2007)
6. Moore, R.C.: Improving IBM word alignment model 1. In: ACL 2004, pp. 519–526 (2004a)
7. Moore, R.C.: On log-likelihood-ratios and the significance of rare events. In: EMNLP 2004, pp. 333–340 (2004b)
8. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics 31(4), 477–504 (2005)
9. Munteanu, D.S., Marcu, D.: Extracting parallel sub-sentential fragments from nonparallel corpora. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, pp. 81–88 (2006)

10. Och, F.J., Tillmann, C., Ney, H.: Improved alignment models for statistical machine translation. In: Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 20–28 (1999)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of ACL, Philadelpha, Pennsylvania, USA, pp. 311–318 (2002)
12. Quirk, C., Udupa, R.U., Menezes, A.: Generative models of noisy translations with applications to parallel fragment extraction. In: Proceedings of the Machine Translation Summit XI, Copenhagen, Denmark, pp. 377–384 (2007)
13. Riesa, J., Marcu, D.: Automatic parallel fragment extraction from noisy data. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 538–542. Association for Computational Linguistics (2012)
14. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proceedings of ICSLP, vol. 2, pp. 901–904 (2002)
15. Smith, J.R., Quirk, C., Toutanova, K.: Extracting parallel sentences from comparable corpora using document level alignment. In: Proceedings of the Human Language Technologies/North American Association for Computational Linguistics, pp. 403–411 (2010)
16. Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D.: Improved Lexical Alignment by Combining Multiple Reified Alignments. In: Proceedings of EACL 2006, Trento, Italy, pp. 153–160 (2006)
17. Tillmann, C.: A Beam-Search extraction algorithm for comparable data. In: Proceedings of ACL, pp. 225–228 (2009)
18. Ture, F., Lin, J.: Why not grab a free lunch? Mining large corpora for parallel sentences to improve translation modeling. In: HLT-NAACL, pp. 626–630 (2012)
19. Zhao, B., Vogel, S.: Adaptive parallel sentences mining from web bilingual news collection. In: IEEE International Conference on Data Mining, Maebashi City, Japan, pp. 745–748 (2002)
20. Ştefănescu, D., Ion, R., Hunsicker, S.: Hybrid parallel sentence mining from comparable corpora. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), Trento, Italy (2012)