Exploiting Non-Parallel Corpora for Statistical Machine Translation

Hoang Cuong, Le Anh Cuong
University of Engineering and Technology
Vietnam National University, Hanoi
Vietnam

Ho Tu Bao
School of Knowledge Science
Japan Advanced Institute of Science and Technology (JAIST)
Japan

***Abstract*—Constructing a corpus of parallel sentence pairs is an important work in building a Statistical Machine Translation system. It impacts deeply how the quality of a Statistical Machine Translation could achieve. The more parallel sentence pairs we use to train the system, the better translation's quality it is. Nowadays, comparable non-parallel corpora become important resources to alleviate scarcity of parallel corpora. The problem here is how to extract parallel sentence pairs automatically but accurately from comparable non-parallel corpora which are usually very "noisy". This paper presents how we can apply the reinforcement learning scheme with our new proposed algorithm for detecting parallel sentence pairs. We specify that from an initial set of parallel sentences in a domain, the proposed model can extract a large number of new parallel sentence pairs from non-parallel corpora resources in different domains, concurrently increasing the system's translation ability by time.**

I. Introduction

Statistical Machine Translation (SMT) is a machine translation approach which depends on creating a parameter probabilistic model by analyzing parallel sentence pairs in a bilingual corpus. With the "de-facto" Moses SMT Engine [5], the effort in building an acceptable translation system quality is reduced by releasing a lot of works. Nowadays, extracting a large number of parallel sentence pairs is one of the most consuming-time and important work for building a good SMT. Unfortunately, parallel corpora have been "limited in size, language coverage, and language register" [9]. Comparable non-parallel corpora are much more available from various resources in different domains, such as from Wikipedia, News websites, etc. However, these resources are very "noisy" environments for the work.

This paper focuses on the problem: from an initial training corpus which usually contain a small number of parallel sentence pairs, how do we can expand this training corpus, specially to a new knowledge domain. By this way the we can improve the SMT system repeatedly. In addition, the SMT system can also enhance its translation's ability in new knowledge domains.

In a general framework of extracting parallel corpus we firstly derive parallel sentence pair candidates, and then determine whether a pair is parallel or not based on the similarity measurement between the two sentences in the pair. Gale-Church [4] measured the rate of lengths between two bilingual sentences. This method is suitable in very "clean" environments (purely parallel corpora); Michel Simard [12] submitted mixing length scheme from Gale-Church [4] and "cognate" criterion to a unify criterion to obtain better results in the environments which are a little "noise". But these two criteria are not powerful, robust, "noise-protecting" or accurate enough for extracting parallel sentence pairs from extremely "noisy" environments [7].

Some other studies such those of Utsuro [13] and Zhao [17] used a statistical translation lexicon model. However, these works suffer much from the ambiguity of word translations, that cause the error rate problem (i.e. the high error rate of extracted parallel sentence pairs). Building a statistical translation lexicon model from a set of available parallel sentence pairs for constructing a dictionary is an error-prone challenge, because its error rate measurement is extremely hard to control[10], [15], [8], [17].

Some methods tried to reduce the error rate by limiting top translations for each source language word, for example, Munteanu [9] used "top five translations of each of its words" to reduce as much as possible the error rate of the parallel corpus extraction. However, it was not a steady solving, and thus there was still a risky error-prone challenge and it almost rejected correct word-by-word translations in dictionary.

The risk of the error rate is also bounded the power of other methods which are based on the statistical translation lexicon model, such as in ([8], [14], [9], [1]). They purely used the word translation model as the main criterion. Under our observation [1] this model will not be sufficient enough to satisfy the requirement of extracting parallel sentences in a noisy environment.

In this paper we improve the similarity measurement by proposing a new algorithm in which we combine Length-based filtering, Cognate condition, and content-based similarity measurement with some appropriate modifications. Specially we improve the algorithm of measuring content-based similarity by using phrasal overlapping calculation denoted in [11] with the help of a complete phrase-based SMT system. We also isolate the translation phrases (also called segments) and use a constraint rule to prevent the unexpected overlapping phrases. The proposed algorithm will extend a large number of parallel sentence pairs and concurrently reduce considerably error rate,

[1] This observation will be confirmed in the section Experiment.

in comparison with previous studies. In addition, by using the scheme of reinforcement learning, we will show that the SMT system's translation quality is deeply increased by time.

The rest of this paper is organized as follows: Section II presents the reinforcement learning idea and its application to our task. Section III shows our method for detecting parallel sentence pairs. Section IV describes in detail the algorithm for measuring the content-based similarity between the two sentences. Section V presents our experimental evaluations to clarify our contributions. Finally, conclusion is derived in section VI.

II. Reinforcement Model for expanding parallel corpora

The basic reinforcement learning model typical consists of:

- a set of environment states S;
- a set of actions A;
- a set R of scalar immediate rewards.

A reinforcement learning agent interacts with its environment in discrete time steps. At each specific time t with state $s_t \in$ S and the set of actions available $A(s_t)$. It elects an action in the action set $A(s_t)$ and the environment walks to a new state s_{t+1} with a *reward* r_{t+1}. Concerning to the environments which have a terminated state, the goal of the agent in reinforcement learning is to try to establish a plan π: S $\rightarrow$ A which collects as much *reward* as possible: $R = r_0 + r_1 + ... + r_n$.

The prior target of parallel sentence extracting at each specific time t (corresponding to the *system's translation ability* at that time C_t) could not extract all the parallel sentence pairs from a comparable non-parallel corpora resource. Our scheme is that the most highest priority of finding job at each specific time is extracting all of possible candidates based on the system's translation ability at that time and re-extracting it latter to get the lack of parallel sentence pairs which could not achieve previously or extracting new non-parallel corpora in different domains due to the increase of the system's translation ability.

Applying the simplest case of reinforcement learning model, where each state is represented as the system's translation ability at each specific time t, we have at each state only one available action (corresponding to extracting all of parallel sentence pairs as much as possible and we get the number of parallel sentence pairs reward achieving). After finding all possible parallel sentence pairs at each time t, we then retrain the SMT system and go to the new state s_{t+1}. Fig. 1 shows the architecture of our reinforcement scheme that deals with both tasks: expanding training corpus of parallel sentence pairs and improving the corresponding SMT system.

III. Method for Detecting Parallel Sentence Pairs

The proposed method for detecting parallel sentence pairs includes three steps as follows:

- *Step 1* - Filtering candidates based on the ratio of lengths of the two bilingual sentences in each candidate.

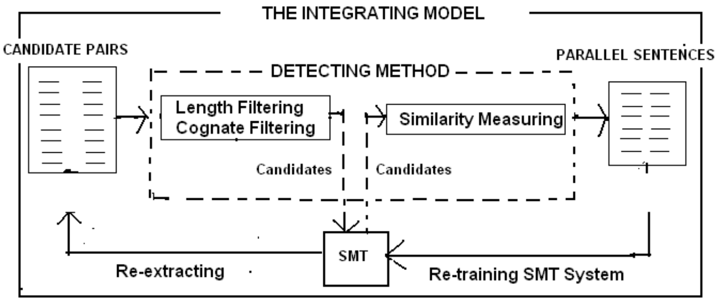


Fig. 1. Architecture of the proposed Model

- *Step 2* - Filtering candidates based on the similarity of cognates in the two bilingual sentences.
- *Step 3* - Measuring content-based similarity described in Figure 2, and then determine whether a candidate is parallel or not.

In step 1, Gale-Church [4] indicated that the length measurement could be ”used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences”. We use this measurement as a length filtering criterion by checking the ratio of candidate pair's length. From all parallel sentence pairs of the training corpus we calculate the mean and variance of all the length ratios. If a candidate has its ratio staying in the circle of this mean and variance, it passes the step 1 and goes to step 2.

In step 2, we adjust Simard [12] idea by comparing the orders of cognate sequences between two sentences in a bilingual sentence candidate (the cognates are just non-translation symbols, such as question mark, bracket, parentheses, etc.). For example, if the *structure* of source language sentence is:

A, B, C ” D ”.

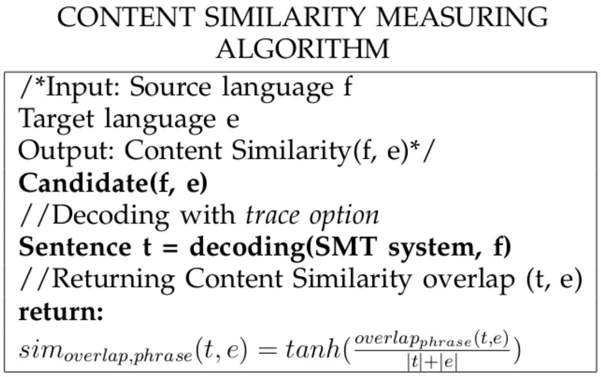
and target language sentence's structure is:

E, F, G ” H ” .

In this step we compare the orders of cognate sequences between the two sentences in a pair and also checking the length filtering condition in its sub-parts. This condition is satisfied if the cognate sequences are the same and its sub-parts pass the length-filtering condition (as in the step 1), we then go to step 3.

In step 3, we estimate the content based similarity between the two sentences in the candidate. This similarity is assigned a score, and if this score is greater or equal a threshold called λ we obtain a new parallel sentence pair.

Figure 1 specifies the architecture of our proposed model, which applies our new proposed method of detecting parallel sentence pairs to the reinforcement learning scheme. In detail, the detecting method consists of two parts. The first part will remove candidates based on the conditions of length and cognate. The second one uses a content-based similarity measurement for filtering candidates. Firstly the candidates go through the first part and if they are satisfied its conditions then they are moved the second part through the SMT. That is because in the algorithm of measuring content-based similarity



```
CONTENT SIMILARITY MEASURING
ALGORITHM
/*Input: Source language f
Target language e
Output: Content Similarity(f, e)*/
Candidate(f, e)
//Decoding with trace option
Sentence t = decoding(SMT system, f)
//Returning Content Similarity overlap (t, e)
return:
$sim_{overlap,phrase}(t,e) = tanh(\frac{overlap_{phrase}(t,e)}{|t|+|e|})$
```

Fig. 2. The content-based similarity measuring algorithm

we will use the translations of source sentences. Finally, if these candidates can pass the conditions of content-based similarity measurement we will obtained new parallel sentence pairs. Each time when we complete expanding the training parallel corpus, we can re-train the SMT system and repeat the processing of expanding parallel corpus as well as improving the SMT system.

IV. The Algorithm for Measuring Content based Similarity

The content-based similarity measurement is an exciting problem. In our specific problem, we have to find out the similarity value between two sentences, which one is a normal human sentence and another one is a machine-generated sentence. The machine-generated sentence is yielded by a complete phrase-based SMT system, which its generation is based on the language model and the translation model [6]. So we can't treat it like a normal human sentence. And we can't apply "semantic" metric or "grammar" metric on measuring the similarity between two normal human sentences. The word overlapping metric seems to be appropriate for this problem.

Banerjee and Pedersencite [3] introduced the lexical overlap measurement based on *Zipf's law* between the *length of phrases* and their *frequencies* in a text collection, which is called the "multi-word phrases overlap" measuring method. Ponzetto and Strube [11] used the sum of sentence lengths and apply the hyperbolic tangent function to minimize the effect of the outliers. In fact, the traditional lexical overlap measurements treat sentences as a bag of words and does not regard highly the differences between *single words overlap* and *multi-word phrases overlap*. The comparison of the content-based similarity measurements is credited by [2] and they pointed out that the "multi-word phrases overlap" measurement is the best measurement in the lexical overlap metric.

For the work of determining parallel sentences in the two languages, we first use a complete phrase-based SMT system to translate a sentence (called translation sentence) to the target language (language of the second sentence). And then apply the phrasal overlap measurement for the two sentences in the same language. Different from other studies in which they measured the overlap of words/phrases between the two sentences, we here can utilize more information from outputs of the complete phrase-based SMT system (based on the MOSES Framework[5]). That will help to reduce "noisy" phrases, avoid inaccurate results and fix flaw as mentioned in [2].

The MOSES Framework[5] trains parallel corpus by extracting all *legal parallel phrases* consists with the *word alignment* - words in one phrase are only aligned to words in another phrase of a pair. The following criterion proposed by Zens[16] defines the set of bilingual phrases BP $BP(f_1^J, e_1^I, A)$ of sentence pair (f_1^J, e_1^I) consists with word alignment matrix A generated from IBM Model:

$$\{(f_j^{j+m}, e_i^{i+n}) : \forall (i', j') \in A\} : \\ j \leq j' \leq j+m \Leftrightarrow i \leq i' \leq i+n$$

The basic formula for finding e_{best} (decoding step) in statistical phrase-based model (mixing several components which contribute to the overall score: the *phrase translation probability* ϕ, *reordering model* d and the *language model* p_LM), which gives all i = 1, ... , I input phrases fi and output phrase ei and their positions $start_i$ and end_i:

$$argmax_e \prod_{i=1}^{I} \phi(\overline{f_i}|\overline{e_i}) d(start_i - end_{i-1} - 1) p_{LM}(e)$$

For almost parallel sentence candidates, we have a very long number of words of each sentence in a pair. The number of words are usually more than 10-15 words per a sentence. By the way, surprisingly, Koehn[6] pointed out that limiting the length to a maximum of "only three words" per a phrase already gains top performance. Using longer phrases does not yield much improvement, and occasionally leads to worse results. In general, we also usually use trigram as the default language model parameter of a phrase-based SMT system[6].

We see that P_{LM} is actually is *not clue enough* to be an relationship between all phrases e_i. This means that we can assume the output of the complete phrase-based SMT system decoding step, all of each phrase element e_i is *independent with other elements*, or there's *no (or rarely) relationship between the elements*.

From the results of the MOSES's decoding process we can split the translation sentence into separate segments. For example, a translation_sentence_with_trace[2] has format sequence of segments like the following:

$$t = |w_1 w_2 ... w_k||w_{k+1} w_{k+2} ... w_n|...$$

So that if we treat these segments independently we can avoid measuring the overlap on the phrases such as $w_k w_{k+1}, w_{k-1} w_k w_{k+1}, \ldots$. It means that we will not take the phrases in which their words appear in different translation segments. Note that in a "noisy" environment this phenomenon may cause many wrong results.

Note that for computing phrase overlapping measurement between the sentence *t* and *e* (i.e. the content-based similarity) we use the formulate denoted in [3] and [11], as follows:

[2] Running the MOSES decoder with the segmentation trace switch using -t option

$$sim_{overlap,phrase}(t,e) = tanh(\frac{overlap_{phrase}(t,e)}{|t|+|e|})$$

where

$$overlap(t,e) = \sum_n^N \sum_m n^2$$

here m is a number of n-word phrases that appear in both sentences.

From our observation, long overlapping phrases take a large proportion in the score of overlapping measurement between the two sentences. Therefore, the appearance of overlapping phrases in non-parallel sentences may cause much mis-detection of parallel sentence pairs. In a very ”noisy” environment there easily exist overlapping phrases randomly. It is worth to notice that this phenomenon hasn’t been mentioned in previous studies.

To overcome this drawback, we added a constraint rule to the algorithm of measuring content-based similarity, that is: an overlapping phrase with *N* words (called N-word overlapping phrase) will be counted if there are at least *N* overlapping phrases/words which have their lengths shorter than *N* and doesn’t appear in the fragment of the N-word overlapping phrase.

With this constraint-improving, our detecting model is *extremely ”strong” noise-filtering* and better than previous studies. The superior quality of our detecting model based on constraint-improving will show more detail in the section Experiment.

V. Experiment

This experiment is deployed on an English - Vietnamese phrase-based SMT Project, using Moses framework [5]. We implement three evaluations to clarify major contributions of the proposed model.

- The first evaluation estimates the threshold λ, concurrently shows the improvement of the proposed algorithm of measuring content-based similarity.
- The second evaluation compares our parallel sentence pair detecting method with a previous dictionary-based method which is considered as the baseline.
- The third evaluation shows the ability of expanding parallel corpus, consequently improving the SMT system via measuring its BLEU score.

A. Data preparation

For the initial parallel corpus, we extract from a Subtitle resource and obtain 50,000 parallel sentence pairs. To do that, we first use the length-based filtering method [4] and then check again by hand.

The Wikipedia resource is chosen for expanding the parallel corpus. It is worth to emphasize that the knowledge domain of Wikipedia is far different from Subtitle domain. We found Wikipedia resource a clue for bilingual connection, like that: a Wikipedia page (in source language) will connect to (if exists) another wiki page (in target language) via Wikipedia’s hyperlink structure. By this evidence, we can collect a set of bilingual pages (in English and Vietnames). And then, from a pair of bilingual pages, denoted as page A (containing *m* sentences) and page B (containing *n* sentences) we have $n \times m$ candidates of parallel sentence pairs.

B. Evaluation 1

This experiment comes from a development set with about 34,000 parallel sentence candidates getting from Wikipedia resource (they are both satisfied the conditions on length and cognation) and to be going through the content-based similarity measuring algorithm. With the obtained parallel sentence pairs, we will check by hand that each pair is true or wrong (i.e. the sentences in a pair are parallel or not). Table 1 shows the results when using 1-gram overlapping measurement; Table 2 shows the results when implementing the improved algorithm with phrasal overlapping, independent fragments, and the constraint rule.

λ	Total	True	Wrong	Error(%)
0.45	778	582	196	25.19
0.5	474	389	85	17.93
0.55	266	231	35	13.15
0.6	156	140	16	10.25

Table 1 - Error rate of detecting parallel sentence pairs using 1-gram overlapping

λ	Total	True	Wrong	Error(%)
0.35	595	586	9	1.51
0.4	404	401	3	0.74
0.45	272	272	0	0.00
0.5	172	172	0	0.00
0.55	108	108	0	0.00
0.6	83	83	0	0.00

Table 2 - Error rate of detecting parallel sentence pairs using phrasal overlapping, independent fragment, and the constraint rule.

The obtained results in Table 1 and Table 2 have shown that the proposed algorithm for detecting parallel sentences are much better than the normal algorithm. They shows that the error rate of using the normal overlapping method is very high, meanwhile with the improved method we can achieve very low error rate even there are no errors when λ is greater than 0.4. It is interesting that, the improved algorithm even bring a larger number of true parallel sentence pairs (586 pairs in comparison with 582 pairs). This evaluation also determines that $\lambda = 0.35$ is a suitable threshold.

C. Evaluation 2

This evaluation was done in an exact condition with the Evaluation 1 to compare how better our proposed method is in comparison with previous studies [13], [9] which used Length, Cognate, and Dictionary for measuring content-based similarity. For this evaluation, we use ”top five translations of each of its words” as described in [9].

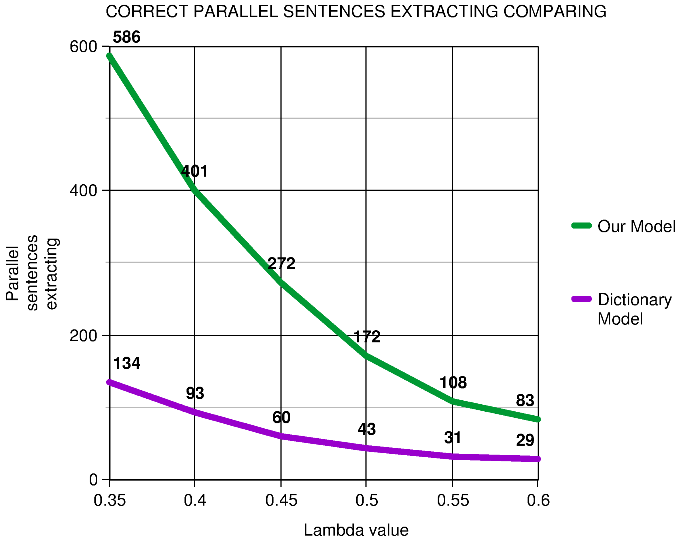


Fig. 3. Comparing the proposed algorithm and the baseline algorithm

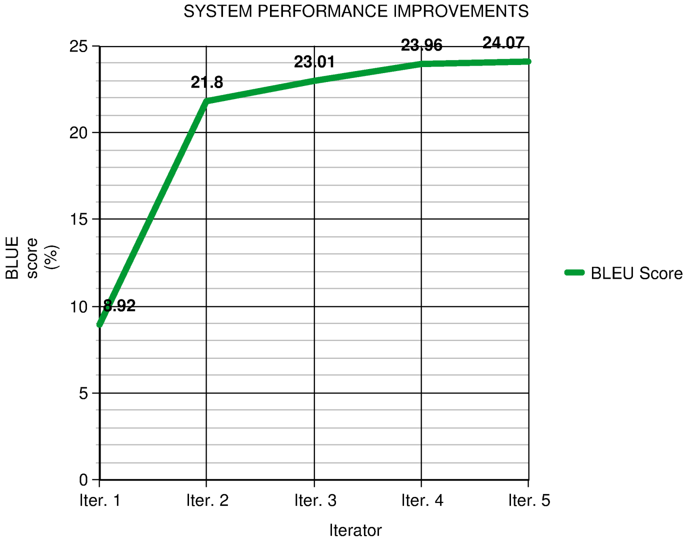


Fig. 4. The system performance improvements gradually

λ	Total	True	Wrong	Error(%)
0.35	149	134	15	10.06
0.4	105	93	12	11.43
0.45	64	60	4	6.25
0.5	45	43	2	4.65
0.55	31	31	0	0.00
0.6	29	29	0	0.00

Table 3 - Error rate of detecting parallel sentence pairs using Dictionary.

Table 3 shows the obtained results of error rate when detecting parallel sentence pairs using Dictionary. We can see that Table 3 denotes that a dictionary created from lexical translation probability output is extremely noisy. In addition, Figure 3 points out that when the word-based translation models limit the number of possible word-translations to balance the error rate controlling, it will reduce intensely the number of extracted parallel sentences. In contrast, although our proposed method satisfies deeply the error rate controlling, meanwhile it could also extract more parallel sentence pairs.

D. Evaluation 3

This evaluation tests the re-training with the re-extracting capacity of our proposed model. We extract 9,998 parallel links from Wikipedia (via Wikipedia's hyperlink structure) and use this resource for evaluating the scalability our method. Note that by using step 1 and step 2 of the parallel sentence detecting algorithm, we remove a large number of parallel sentence pair candidates from the whole of candidates (tens of million of parallel sentence pair candidates). Consequently, there are only about 958,800 candidates to be used for the step 3.

Starting up with a set of 50,000 available parallel sentence pairs collected from Subtitle resources we will test the capacity of extending parallel sentence pairs. This is done by applying the reinforcement scheme and using the method for detecting parallel sentences. Each time when the training is extended we will retrain the SMT system, and then apply it to the candidates again to find out new results. The Table 4 shows the experimental results:

Iterator	Training	BLEU(%)	Extract
Iter. 1	50,000	8.92	22,835
Iter. 2	72,835	21.80	16,505
Iter. 3	89,340	23.01	4,742
Iter. 4	94,082	23.96	1,130
Iter. 5	95,212	24.07	0

Table 4 - The Results of Integrating Reinforcement Learning with Our Detecting method

We use a test set of parallel sentence pairs of 10,000 parallel sentence pairs handling [3] from Wikipedia resource to test the improvement of the BLEU score of our phrase-based SMT system.

At first the BLEU score is 8.92% obtained by using the initial training set. And then, at the first iteration of the process of extending parallel corpus, we achieve 22,835 new parallel sentence pairs. Retraining a with the new training set containing 72,835 parallel sentence pairs the BLEU score is up to 21.80%. And then the SMT system continue extracting more 16,505 new parallel sentence pairs which could not extract at the previous iterations. This result denotes that the SMT system is now upgrading its translation ability. At the end, we can extract in the total of 45,212 new parallel sentence pairs and the BLEU score reaches to 24.07% which is far from the beginning sytem.

[3] We first extract automatically a set of parallel sentence pair candidates based on the proposed algorithm, and then check them again by hand for obtaining exactly parallel sentence pairs.

Figure 4 shows out more details the increasing of BLEU Score value gradually. Another interesting thing is when we try to extract other resources, at a suitable time, we could turn back and could still extract more parallel sentence pairs from this resource. That's one of the most valuable things of our proposed model - applying reinforcement learning scheme integrating with our detecting method.

VI. Conclusion

This paper has proposed a model that integrates our new method of detecting parallel sentences into a reinforcement learning scheme for the purpose of extending a parallel corpus, and consequently improving statistical machine translation. Various experiments have been conducted and the obtained results have shown that the proposed algorithm of detecting parallel sentences could extract a larger number of parallel sentence pairs and the error rate has been much reduced in comparison with the previous results. This new algorithm is then applied into a reinforcement scheme, which allows the SMT system can be upgraded by time with its improved ability of translation, specially for covering new knowledge domains.

References

[1] Sadaf AbduI-Rauf and Holger Schwenk. On the use of comparable corpora to improve smt performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 16–23, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[2] Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. The evaluation of sentence similarity measures. In *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, DaWaK '08, pages 305–316, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 805–810, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[4] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19:75–102, March 1993.

[5] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[6] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[7] I. Dan Melamed. Bitext maps and alignment via pattern recognition. *Comput. Linguist.*, 25:107–130, March 1999.

[8] Robert C. Moore. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK, 2002. Springer-Verlag.

[9] Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31:477–504, December 2005.

[10] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[11] Simone Paolo Ponzetto and Michael Strube. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Int. Res.*, 30:181–212, October 2007.

[12] Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2*, CASCON '93, pages 1071–1082. IBM Press, 1993.

[13] Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. Bilingual text, matching using bilingual dictionary and statistics. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, COLING '94, pages 1076–1082, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[14] Dekai Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 80–87, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[15] Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 144–151, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[16] Richard Zens, Franz Josef Och, and Hermann Ney. Phrase-based statistical machine translation. In *Proceedings of the 25th Annual German Conference on AI: Advances in Artificial Intelligence*, KI '02, pages 18–32, London, UK, 2002. Springer-Verlag.

[17] Bing Zhao and Stephan Vogel. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–, Washington, DC, USA, 2002. IEEE Computer Society.