A Dependency based Large-scale Web Parallel Corpus Filtering Method*
一种基于依存句法分析的互联网海量双语平行语料过滤方法

王明轩，熊浩，刘群
（中国科学院计算技术研究所，北京市 100190）

Abstract

Large-scale web parallel corpus could potentially improve the performance of statistical machine translation models. However, the quality of those web parallel corpus is not reliable since a certain number of them stem from machine translation. In this paper, we propose a dependency based method to filter those corpus that do harm to translation models. We perform dependency parsing on both source and target sides, and filter the corpus that obtain lower score in respect of the similarity between source and target dependency trees. Large scale experiments on Chinese-English spoken translations show that our models filter 21% corpus while still significantly improve the performance 0.9 in term of BLEU.

摘要:对于统计机器翻译来说 ,互联网中存在的海量双语平行语料能够潜在的提升翻译系统的性能。然而当前互联网中包含很多翻译质量较差、意译、甚至是机器翻译结果的双语语料 ,利用这部分语料训练的翻译模型显然将影响最后的翻译效果。本文提出一种基于依存句法分析的语料过滤方法 ,通过对双语之间进行快速依存句法分析 ,计算源语言和目标语言之间依存句法树的相似度 ,过滤掉一些相似度较低的平行句对。在大规模汉英口语翻译实验中 ,使用该方法对互联网语料进行过滤后 ,在过滤掉 21%语料的同时 ,翻译性能显著的提高了0.9个BLEU值。

1 Introduction

Statistical machine translation (SMT) relies heavily on the parallel training corpus. No natter the quantity or the quality of the corpus contributes to the performance of SMT. In recent years, network data mining has provided adequate parallel corpus for SMT. However the quality of those web parallel corpus is not reliable. Through the analysis of a large number of web bilingual sentence pairs, we have found that mainly three types of bilingual sentence pairs decrease the value of the corpus and do harm to the translation model.

• **Machine-translated sentences** often demonstrate better word correspondence than human translated sentences and are easier to align, but a certain num -ber of those sentences are translated disordered and the rules extracted from them are likely to mislead the translation model because the language model trained from them may be unnatural.

• **Paraphrase sentences pairs** express the same meaning, but the alignment is really bad and some key words from source sentences may be aligned to empty. The rules extracted from them will obviously reduce the translation model's performance.

• Further, there are also many other mistakes (e.g. spelling mistakes and unmatched sentences) in those web parallel corpus. These sentences pairs obviously reduce the quality of the corpus and are likely to reduce the translation model's performance. Therefore, to improve the performance of the SMT model, these bad sentences must be filtered.

Most previous researches on SMT training data are focused on improving the scale of the corpus. Some researches try to collect parallel sentences from

*本文受 863 重大项目课题 2011AA01A207 对本文研究工作的资助。
*作者简介： 王明轩：男，1989 年生；硕士研究生，研究方向是自然语言处理、机器学习。
熊皓：男，1985 年生，助理研究员，研究方向是自然语言处理、机器学习。
刘群：男，1966 年生，研究员，研究方向是自然语言处理、机器翻译、信息提取。

website(Nie et al. 1999; Resnik and Smith 2003; Chen et al. 2004). Others try to extract parallel sentences from comparable corpus(Munteanu and Marcu 2005, 2006). These work aims to enlarge the corpus, while our work aims to improve the quality of the corpus by filtering out the bad sentences pairs.

There are few work on filtering corpus for translation model training. Most successful and recent study was that of (Lu et al., 2007; Keiji et al., 2009). Their work both focus on selecting translation pairs as the training set from a training parallel corpus by using a small in-domain parallel corpus. Although we all aims to improve the value of the corpus, they filter the out-domain sentence pairs while our task is to filter bad sentences pairs.

A simple method to filter out the corpus is to use the alignment probability, but there are a certain number of bad sentence pairs are well-aligned and this method is less effective. Further we proved this in the experiment. In this paper, we propose a method of filtering out bad sentences pairs from parallel corpus to improve the quality of the corpus. This method enables a certain number of bad sentences to be filtered out by comparing the matched-degree of the dependency tree of the pairs. Therefore the size of training corpus will be reduced and we can acquire an efficient translation model. As a result, training the translation model will become faster. Further, because the bad sentences pairs which may mislead the translation model have been filtered out, the performance will improved significantly. Our method take full advantage of syntactic information and the logical relationships between words, unnatural machine-translated and paraphrase sentence pairs can be accurately identified.

Section 2 describes the dependency parsing. Section 3 details the method to filter the corpus based on dependency parsing. Section 4 describes another method only use lexical information as a comparison. Section 5 details the experiment result for filtering the training set and compares the results of the dependency-based method with lexical-based method. Section 6 concludes the paper.

2 Background

2.1 Dependency parsing

Dependency structure, as the first step towards semantics, represents the grammatical relations that hold between words in a sentence. It encodes semantic relations directly, and has the best interlingual phrasal properties(Fox, 2002). Those attractive characteristics make it possible to identify some well-aligned but disordered or unnatural sentences.

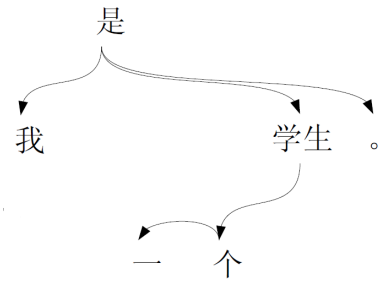
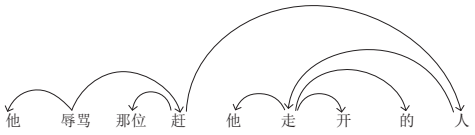


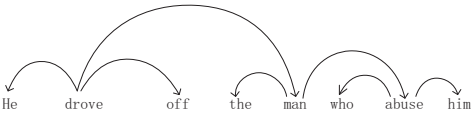
Figure 1: a dependency structure

The dependency structure for a sentence is a directed acyclic graph with words as nodes and modification as edges. Each edge directs from a head to a dependent. Figure 1 shows a dependency structure for a Chinese sentence. In this figure, “是” is head of “学生” and “学生” is the head of “个”. Therefore the relationship between “我” and “学生” is closer than that of “我” and “个”.

3 Filtering method



(a) dependency structure for Chinese sentences



(b) dependency structure for English sentence

Figure 2: disorder-translation sentence pair

If only based on the lexical information, the three types of bad sentence pairs mentioned in section1

can not be filtered out easily. There are some reasons for that.

- Some disordered but well-aligned sentence pairs will mislead the translation model. These sentence pairs may stem form machine-translations or some bad human-translations. As we can see from fig 2, the two sentences are really well-aligned, but the sentence in fig 2b is not the correct translation of sentence in fig 2a, because there are some order mistakes in the target sentence. It is truly difficult to filiter out this kind of sentences. We must make use of some semantic information to solve this problem.

- Some paraphrase sentences are hard to identify, because they are not true sentences translation errors but really do harm to the SMT model, such as the sentences pair "把 你 从 大西洋 城 来的 吗" 和 "From Atlantic city, isn't it". These sentences have the same meaning as a whole, but the phrase extracted from them may not comparable or some key words from source sentences may aligned to empty. We must filter out these well-translate but paraphrase sentences.

Consider dependency structure represents the semantic relationship that hold between two words in a sentence, we use the *match-degree* of the dependency structure between the source sentence and the target sentence to estimate whether this sentence pair will reduce the value of the corpus.

In this section, we will introduce a method to estimate the match-degree of the dependency structure. This method is based on the conclusion that for a well-translate sentence pair if two words have tightly relationship in source dependency structure, their aligned-words in the target dependency structure must also have closely relationship.Roughly speaking, if a certain number of word pairs have closely relationship in source dependency structure/ but their aligned-word in target dependency structure doesn't connect closely, we believe that this sentence pair is a bad sentence pair and may do harm to the translation model. Taking fig 2 as an example, in source dependency structure" 他$_1$" is the dependent of"辱骂$_2$", while in the target dependency structure, the aligned-word "He$_1$" is far away from the aligned-word "abuse$_7$". This sentence pair is well-aligned but doesn't match in semantic lever. Using dependency information we can easily filter it out.

3.1 Model description

Following convention，we will assume throughout this paper that the task is to estimate the *match-degree* of a sentence pair. In general, we will use f to refer to a sentence in "source" language: f is a sentence of words $f_1, f_2 \dots f_m$ where m is the length of the sentence and f_i for $i \in \{1 \dots m\}$ is the j'th word in the sentence. We will use e to refer to an "target" sentence: e is equal to e_1,$e_2 \dots e_l$ where l is the length of the sentence.

We also need to introduce additional alignment variables in this paper. We will have alignment variables $A = \{a_1, \dots, a_m\}$—that is, one alignment variable for each "source" word in the sentence—where each alignment variable can be represented as a vector which can take any value in $\{0, 1, \dots, l\}$. We now describe the alignment variables in detail. We assume that each a_i for $i \in \{1, \dots, m\}$ is equal to a vector $< a_i^1, \dots, a_i^{k_i} >$ which means that a "source" word will be aligned to k_i "target" word. Each alignment variable a_i^j specifies that the "source" word f_i is aligned to the "target" word $e_{a_i^j}$ and $e_{a_i^j}$ is the j'th alignment to f_i.

Further, the dependency structure we used should be introduced. Formally, $W = w_1 w_2 \dots w_n$ is a sentence, and $\Psi = (V, E)$ is the dependency structure for this sentence where:

- $V = \{root\} \cup \{W\}$ which means that the vertex set of the dependency structure contains all the words in W and an artifical node *root*.

- $E = \{w_h, r, w_d\}$ where $w_h \in V$, $w_d \in V$, $r \in R = \{dep\}$ that is, each edge of the edge set is a link from head word to dependent word. In our model, the link relation is simple and we neglect the label of link. The reason is that the relation in Chinese dependency sturcture is somewhat diffrent from English dependency structure, and the simple relation is enough to estimate the match-degree between these two structures.

- $\forall(w_h, r, w_d) \in E \Rightarrow w_d \neq root$,that is, artifical node has no head.
- $\forall w \in V \wedge w \neq root, \exists w_h(w_h, r, w) \in E$, that is, each node in V expect $root$ has the head node.
- $\forall w(w_h, r, w) \in E \wedge (w'_h, r, w) \in E \Rightarrow w_h = w'_h$, that is, the node in V only has one head node.
- E doesn't contain subset of arcs $\{(w_h, r, w_i), (w_i, r, w_k), \ldots, (w_j, r, w_h)\}$, that is, there is no cycles in the graph.

In this paper we use Ψ_f to represent the dependency structure for f and Ψ_e for e. Each edge in the dependency structure denote the relationship between the word pair connected by the edge. Consider that word "A" relates to word "B" in source sentence, we believe that the aligned-word of "A" must relate to aligned-word of "B". Hence we can use the word pair relation between the two dependency structure to estimate the match-degree.

We use Γ to present the match-degree between Ψ_f and Ψ_e. Formally, we have:

$$\Gamma(\Psi_f, \Psi_e) = \frac{\sum_{w_i^f, w_j^f} sim(w_i^f, w_j^f, w_{a_i}^e, w_{a_j}^e)}{\sum_{w_i^f, w_j^f} Q(w_i^f, w_j^f)} \tag{1}$$

Where $sim(w_i^f, w_j^f, w_{a_i}^e, w_{a_j}^e)$ represents the comparable relation between the edges (w_i^f, w_j^f) and $(w_{a_i}^e, w_{a_j^{k_q}}^e)$. $Q(w_i^f, w_j^f)$ is the relation between between w_i^f and w_j^f, formally we have

$$Q(w_i, w_j) = \begin{cases} 0 & \text{if } w_i \text{ is the head of } w_j \\ 1 & \text{if } w_i \text{ is not the head of } w_j \end{cases} \tag{2}$$

Consider w_i or w_j may have more than one aligned-word, we use the average relation to estimate $sim(w_i^f, w_j^f, w_{a_i}^e, w_{a_j}^e)$. Further, we should punish the word wich aligned to empty. If w_i or w_j is aligned to empty, we employ 0 to $sim(w_i^f, w_j^f, w_{a_i}^e, w_{a_j}^e)$.

$$sim(w_i^f, w_j^f, w_{a_i}^e, w_{a_j}^e) = \frac{\sum_{k_p, k_q} R(w_i^f, w_j^f, w_{a_i^{k_p}}, w_{a_j^{k_q}})}{\#a_i \times \#a_j} \tag{3}$$

Where $R(w_i^f, w_j^f, w_{a_i^{k_p}}^e, w_{a_j^{k_q}}^e)$ donates to the relation score between word pair (w_i^f, w_i^f) and $(w_{a_i^{k_p}}^e, w_{a_j^{k_q}}^e)$. Typically, the distance between w_i and w_j in the dependency structure relates to the relation between the word pair. Far distance implies far relation. We use $dist(w_i, w_j, \Psi)$ to represent the distance between w_i and w_j in Ψ. Through a great quantity of bilingual training corpus, we have found that $dist(w_i, w_j, \Psi_f)$ is always similar to $dist(w_{a_i}, w_{a_j}, \Psi_e)$ in good quality sentence pairs. Hence we use the difference between $dist(w_i, w_j, \Psi_f)$ and $dist(w_{a_i}, w_{a_j}, \Psi_e)$ to estimate the relation score. Briefly, far relation means low relation score and we use a simple function to measure this relation:

$$R(w_i^f, w_j^f, w_{a_i^{k_p}}^e, w_{a_j^{k_q}}^e) = \frac{1}{|dist(w_i, w_j, \Psi_f) - dist(w_{a_i^{k_p}}, w_{a_j^{k_q}}, \Psi_e)| + 1} \tag{4}$$

Following convention, we assume that the distance between w_i and w_j is the shortest path from w_i to w_j in the dependency structure. For convenience, we use $trace(w, \Phi)$ to present the path from the artificial node *root* to w, and $comm(w_i, w_j.\Phi)$ to present the common path of $trace(w_i, \Phi)$ and $trace(w_j, \Phi)$. "#path" is employed to present the length of this "path". The distance between a word pair can be defined as follows:

$$dist(w_i, w_j, \Phi) = \#trace(w_i, \Phi) + \#trace(w_j, \Phi) - 2 \times \#comm(w_i, w_j, \Phi) \tag{5}$$

Through Eq. 1 we can calculate the match-degree between the two dependency tree, however we must go over all the word-pairs in source dependency tree. It will take $|V|^2$ time. Further most of the word pairs are not related and this may mislead our model, hence we just consider the word-pair which has close relation in source dependency structure. This will reduce the complexity of our model.

$$\Gamma(\Psi_f, \Psi_e) = \frac{\sum_{(w_i^f, w_j^f) \in E} sim(w_i^f, w_j^f, w_{a_i}^e, w_{a_j}^e)}{|E_f|} \tag{6}$$

According to the Eq. 6, we can calculate the final dependency match-degree, and it just costs |E| time. Further the model will be more reliable, because we neglect most unrelated word-pairs from source dependency structure.

3.2 An instance

Take Figure 2 as an example, this sentence pair are aligned well but really are not matched. Briefly we just analysis one word pair to describe our method and shows its efficiency. Take word-pair (他$_1$,辱骂$_2$) from Figure 2a and their aligned-words are as follows:

$$a_1 =< 1 >, a_2 =< 7 >$$

That is, He$_1$ is aligned to 他$_1$ and abuse$_7$ is aligned to 辱骂$_2$. Then we can get the distance of these word pairs through Eq. 5.

$$dist(w_1, w_2, \Psi_f) = 1 + 0 - 2 \times 0 = 1$$

$$dist(w_1, w_7, \Psi_e) = 1 + 2 - 2 \times 0 = 3$$

We can find that the relation between w_1^f and w_2^e is close, but the relation between w_1^e and w_7^e is far. Aligned word pairs should have the same dependency relation, so this bad aligned-pairs has low comparable relation score as we can estimate through Eq. 4.

$$R(w_1^f, w_2^f, w_1^e, w_7^e) = \frac{1}{3 - 1 + 1} = \frac{1}{3}$$

Because both w_1^f and w_2^f have only one aligned-word, we have

$$sim(w_1^f, w_2^f, w_{a_1}^e, wa_{2_e}) = R(w_1^e, w_2^e, w_1^f, w_7^f)$$

Same as this word pair, we can go over all the edges and finally based on Eq. 6 to estimate the match-degree of the sentence pair.

4 The method based on lexical information

A simple method is to employ the lexical information and filter the sentence pairs with low alignment probability. We define this probability as $P(f|e)$, that is, the probability of e given f, where e is the target-language sentence and f is the source-language sentence. To estimate this probability we have:

$$P(f|e) = \sum_A P(f, A|e) \tag{7}$$

Where $A = a_1 a_2 \dots a_m$ is alignment information. Generally, based on *chain rules* we have:

$$\begin{aligned} P(f, A|e) =& P(m|e) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, s_1^{j-1}, m, e) \\ & \cdot P(s_j|a_i^j, s_i^{j-1}, m, T) \end{aligned} \tag{8}$$

Finally, the probability can be calculated by IBM model(P.F.Brown et al, 1993). And in this paper we adopt the Open source program GIZA++[2] to get the probability. Then we filter out the pairs with low alignment probability and use the corpus left to train the system. But this method is less effective because well-aligned pairs may also do harm to our translation model and the accurate rate of this method isn't credible.

5 Experiments

5.1 Framework of SMT

We conduct our experiment on Chinese-to-English translation tasks. The baseline system is the hierarchical phrase-based SMT system which is implemented by using log-liner translation model (He et al, 2006). This model express the probability of a target-language word sequence e given a source-language word sequence f.

$$p(e|f) = \frac{exp(\sum_{i=1}^{M} \lambda_i f_i(e, f))}{\sum_{e'} exp(\sum_{i=1}^{M} \lambda_i f_i(e', f))} \tag{9}$$

Where $h(f, e)$ is the feature function and λ_i is the weight of the feature. M is the number of feature. The translation results $\hat{e}$ can be obtained by

$$\hat{e} = \text{argmax}_e \sum_{i=1}^{M} \lambda_i h_i(f, e) \tag{10}$$

5.2 Experiment setting

All the experiments is on the baseline system and the only difference is the training data.In training

[2]http://www.fjoch.com/GIZA++.html

process we use GIZA++[3] to align the sentence pairs on both direction. Then we employ "grow-diag-final" method to refine it (Konhn et al.,2003). To extract the rules we use the method described in (Chiang et al. 2007). For the log-linear model training, we adopt minimum-error-rate training method as described in (Och, 2003). And the translation quality is evaluated by BLEU metric(Papineni et al, 2002) as calculated by mteval-v11b.pl[4]with case-sensitive matching of n-grams. To conduct the dependency parsing on both sides of the corpus, we adopt stanford parser[5].

We dug about 10 million bilingual sentence pairs from Hujiang[6]and Renren[7] website and randomly extract 4,000 pairs to be translated by human. Then we select 2,000 pairs from this human-translation pairs as development set and the other 2,000 pairs as the test set. The experiment employs two method to filter the corpus. The first method is based on dependency structure as described in Section 3 and the second one only use the lexical information in Section 4.

5.3 Experiment results

5.3.1 Dependency based method

In Section 3, we filter bad sentence pairs through the match-degree of their dependency structure. A threshold λ should be set up to filter. When the match-degree is smaller than λ, the pair should be filtered out.

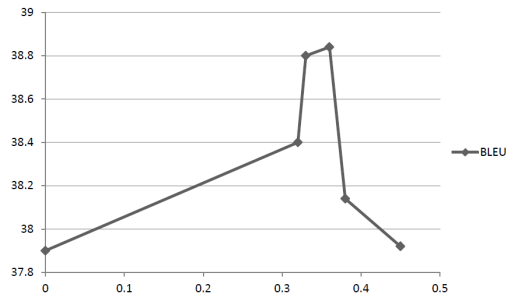


Figure 3: the relation between λ and BLEU

Figure 3 plots the result of the experiments. The horizontal axis represents the threshold to filter the corpus and the vertical axis represents the automatic metric of translation. $\lambda = 0$ in the figure indicates automatic scores of a baseline system. This baseline system is trained by the original corpus without filtering. Here the filtering was carried out by using Eq. 6. In Figure 3, the best translation quality is obtained where the threshold is equal to 0.36. The best translation significantly improve the performance 0.9 over the baseline in term of BLEU. The translation performance is equal to the baseline where the threshold is equal to 0.45.

λ	0	0.33	0.34	0.36	0.38	0.45
Filter-size	0	1.44	1.57	2.10	2.54	4.10

Table 1: The relationship between λ and filter-size(million)

Table 1 shows the filter-size changes with λ. Where λ equals to 0.36, 2.1 million pairs was filtered out and here the best performance we can get. Further where λ equals to 0.45, almost half of the corpus was filtered out but the performance of the translation model keeps the same as the baseline. As the table indicates, our method not only improved the quality of the corpus but also significantly reduced the size of the translation model. This reduction had a positive effect on the com- putational load of decoding.

5.3.2 Comparison with lexical based method

We also adopt another method to filter the corpus as an comparison. The method only use the lexical information and based on alignment probability described in Eq. 7. We estimated the alignment probability by using GIZA++. Then we sorted the sentence pairs by the probability and filtered the pairs with low probability respectively. We used the corpus left to train our model. Table 2 shows that only using lexical information (L-result) is less effective in improving the corpus quality. The best performance only improve 0.35 in term of BLEU where 2.1 million pairs was filtered out. Further when 4.1 million pairs was filtered, the performance even reduce 1.3 in term of BLEU compared to the baseline, while using the dependency information here the BLEU keeps the same(37.9).

[3]http://www.fjoch.com/GIZA++.html
[4]http://www.nist.gov/speech/tests/mt/resources/scoring.htm
[5]http://nlp.stanford.edu/software/lex-parser.shtml
[6]http://www.hujiang.com
[7]http://www.yyets.com/subtitle

F-size	0	1.44	1.57	2.10	2.54	4.10
D-result	37.9	38.8	38.35	**38.84**	38.14	37.9
L-result	37.9	38.1	**38.25**	37.71	37.1	36.62

Table 2: The comparison with lexical based method. D-result is obtained by employing the method described in Section 3 and L-result is got by using the method in Section 4. F-size is the number of pairs to be filtered and unit for F-size is million.

5.4 Analysis

Because a certain number of web parallel corpus stem from machine translation or have some mistakes. Lexical-based method cannnot capture the semantic mistakes which may mislead the translation model. Further lexical-based method isn't sensitive to some mistakes. If only few key words are aligned to empty, the sentence pairs may also acquire high alignment probability. This is not uncommon especially in paraphrase pairs. Our dependency-based method can solve this problem. If some key word are spelled wrong or aligned empty, the dependency structure are sensitive to find these mistakes. This method can also capture the semantic mistakes because dependency parsing is the first step towards semantic and represents the grammatical relations that hold between words in a sentence.

Table 3 shows the sentence pairs to be filtered by our dependency-based method. Sentence 1 and 5 has spell mistakes (abo and had) and it's difficulty to filter this sentence by lexical-based method. However, the mistake will mislead the dependency parsing and the match-degree will be reduced. Hence dependency-matched method is sensitive to find this type of mistakes. Sentence 2,6,7 are paraphrase pairs, and their dependency structure are obviously different. Our method can easily identified this type of mistakes. Sentence 3,4,8 have some key words aligned to empty. This type of mistakes may lead us to extract some rules that the key words are aligned to empty. Roughly, key words are some notional words such as the verbs and nouns which play an important role in a sentence. In spoken translation, a large number of corpus lost the key words. Our method punish the word aligned to empty, and if a word are aligned empty, all the edges link to the word will acquire 0 score. That is, if an aligned-empty word has more dependent node, we will punish more. The reasons generally if a word has more dependent node, it is more important. Hence our dependency-based method can easily tackle with this type of mistakes.

6 Conclusion and Future work

In this paper, we propose a dependency based Large-scale web parallel corpus filtering method to improve the performance of the translation model. This method makes use of the semantic information and assumes that if a word pair has close relation in source-language side, their aligned-word pairs should has the same relation in target-language side. If a pair is not meet this assumption, we will punish it. Then we will estimate the match-degree and filtered the corpus based on this assumption. Further we conduct another method based on lexical information as comparison. This method calculates the alignment probability of the sentence pairs using IBM model. Experimental results showed that dependency based method is far more effective than this lexical based method. The method not only significantly improved the performance of translation model but also reduced the scale of corpus. Our work also demonstrated that well-aligned sentence pairs not always improve the translation model.

In future work, we will improve our method in several aspects. Currently, the dependency based method and the similarity between word pairs are simple. It might work better by trying other sophisticated similarity measure models or using the label of dependency structure. Further we assume that each edge in the dependency structure has the same weight and introducing more optimization weight may improve our system.

References

Franz Josef Och 2003. *Minimum Error Rate Training in Statical machine Translation* ACL-2003:160:167

Jian-Yun Nie, Michel Simard, Pierre Isabelle, Richard Durand 1999. *Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web*. SIGIR-1999:74-81

Philip Resnik and Noah A. Smith 2003. *The Web as a Parallel Corpus.* Computational Linguistics

1	因为你从不告诉我们你有什么事.	It's not like you ever talk **abo** happend.
2	不具备**减震**功能的鞋子	…footwear which does not **absorb the impact of the foot striking the ground**.
3	Dubois 几乎丧失了三分之一在那条天堑里.	Almost a thier of Dubois's **brigade** fell into that abyss.
4	针刺与肌肉张力平衡促通法，**治疗，中风，偏瘫**，康复理念.	Acupuncture and the myo-tensitu pass to pass the law treament stroke hemiparalysus's recovery idea balanced.
5	艾德森把手压在口袋, 沮丧地走在回家路上.	Adelson presses **had** on his pocket , going home in low spirit.
6	把你从大西洋城**带到**这来的吗?	From Atlantic city , is n't it?
7	黄河远上白云间，一片孤城万仞山.	Where a yellow river climbs to the white clods. Near the one city -all among ten-thousand-foot miuntains.
8	本文综述了壁面机器人的现状及其优缺点，**提出了新的避面清洗机器人的设计方案**.	The theme summarizes the advantage and dis advantage of a various types of Wall-climbing robot today.
9	牧师爬上一个**柱子**，魔王就能爬上十个**柱子**.	While the priest climbs one, the devil climbs ten.

Table 3: Sentence pairs to be filtered. There are mainly three types of mistakes:1) Spelling mistakes, such as sentences 1,5. 2) Paraphrase sentences, such as sentences 2,6,7. 3) Aligned empty sentences, such as sentences 3,4,8,9.

29(3):349-380

Jisong Chen, Rowena Chau, Chung-Hsing Yeh 2003. *Discovering Parallel Text from the World Wide Web*. ACSW Frontiers 2004: 157-161

Stanley F. Chen and Joshua Goodman. 1998 *An Empirical Study of Smoothing Techniques for Language Modeling*. Technical Report TR-10-98, Harvard University Center for Research in Computing Technology.

Zhongjun He, Yang Liu, Deyi Xiong, Hongxu Hou, and Qun Liu 2006 *ICT System Description for the 2006 TC-STAR Run#2 SLT Evaluation*. Proceedings of TCSTAR Workshop on Speech-to-Speech Translation: 63-68

Philipp Koehn, Franz J. Och, and Daniel Marcu 2003 *Statistical phrase-based translation*. Proceedings of HLT-NAACL 2003: 127–133.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation* . ACL-2002: 311–318

Heidi J. Fox 2002. *Phrasal cohesion and statistical machine translation*. In Proceedings of EMNLP 2002, pages 304–311.

Dragos Stefan Munteanu and Daniel Marcu 2005. *Improving Machine Translation Performance by Exploiting Comparable Corpora* . Computational Lin- guistics, 31 (4): 477-504