Confidence Measure for Word Alignment

Fei Huang

IBM T.J.Watson Research Center
Yorktown Heights, NY 10598, USA
huangfe@us.ibm.com

Abstract

In this paper we present a confidence measure for word alignment based on the posterior probability of alignment links. We introduce sentence alignment confidence measure and alignment link confidence measure. Based on these measures, we improve the alignment quality by selecting high confidence sentence alignments and alignment links from multiple word alignments of the same sentence pair. Additionally, we remove low confidence alignment links from the word alignment of a bilingual training corpus, which increases the alignment F-score, improves Chinese-English and Arabic-English translation quality and significantly reduces the phrase translation table size.

1 Introduction

Data-driven approaches have been quite active in recent machine translation (MT) research. Many MT systems, such as statistical phrase-based and syntax-based systems, learn phrase translation pairs or translation rules from large amount of bilingual data with word alignment. The quality of the parallel data and the word alignment have significant impacts on the learned translation models and ultimately the quality of translation output. Due to the high cost of commissioned translation, many parallel sentences are automatically extracted from comparable corpora, which inevitably introduce many ”noises”, i.e., inaccurate or non-literal translations. Given the huge amount of bilingual training data, word alignments are automatically generated using various algorithms ((Brown et al., 1994), (Vogel et al., 1996) and (Ittycheriah and Roukos, 2005)), which also introduce many word alignment errors.

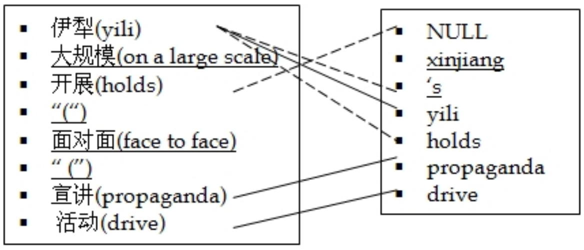

Figure 1: An example of inaccurate translation and word alignment.

The example in Figure 1 shows the word alignment of the given Chinese and English sentence pair, where the English words following each Chinese word is its literal translation. We find untranslated Chinese and English words (marked with underlines). These spurious words cause significant word alignment errors (as shown with dash lines), which in turn directly affect the quality of phrase translation tables or translation rules that are learned based on word alignment.

In this paper we introduce a confidence measure for word alignment, which is robust to extra or missing words in the bilingual sentence pairs, as well as word alignment errors. We propose a sentence alignment confidence measure based on the alignment's posterior probability, and extend it to the alignment link confidence measure. We illustrate the correlation between the alignment confidence measure and the alignment quality on the sentence level, and present several approaches to improve alignment accuracy based on the proposed confidence measure: sentence alignment selection, alignment link combination and alignment link filtering. Finally we demonstrate

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 932–940,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

the improved alignments also lead to better MT quality.

The paper is organized as follows: In section 2 we introduce the sentence and alignment link confidence measures. In section 3 we demonstrate two approaches to improve alignment accuracy through alignment combination. In section 4 we show how to improve a MaxEnt word alignment quality by removing low confidence alignment links, which also leads to improved translation quality as shown in section 5.

2 Sentence Alignment Confidence Measure

2.1 Definition

Given a bilingual sentence pair (S,T) where S=$\{s_1,\ldots, s_I\}$ is the source sentence and T=$\{t_1, \ldots,t_J\}$ is the target sentence. Let $A = \{a_{ij}\}$ be the alignment between S and T. The alignment confidence measure $C(A|S,T)$ is defined as the geometric mean of the alignment posterior probabilities calculated in both directions:

$$C(A|S,T) = \sqrt{P_{s2t}(A|S,T)P_{t2s}(A|T,S)}, \quad (1)$$

where

$$P_{s2t}(A|S,T) = \frac{P(A,T|S)}{\sum_{A'} P(A',T|S)}. \quad (2)$$

When computing the source-to-target alignment posterior probability, the numerator is the sentence translation probability calculated according to the given alignment A:

$$P(A,T|S) = \prod_{j=1}^{J} p(t_j|s_i, a_{ij} \in A). \quad (3)$$

It is the product of lexical translation probabilities for the aligned word pairs. For unaligned target word t_j, consider $s_i = NULL$. The source-to-target lexical translation model $p(t|s)$ and target-to-source model $p(s|t)$ can be obtained through IBM Model-1 or HMM training. The denominator is the sentence translation probability summing over all possible alignments, which can be calculated similar to IBM Model 1 in (Brown et al., 1994):

$$\sum_{A'} P(A',T|S) = \prod_{j=1}^{J} \sum_{i=1}^{I} p(t_j|s_i). \quad (4)$$

Aligner	F-score	Cor. Coeff.
HMM	54.72	-0.710
BM	62.53	-0.699
MaxEnt	69.26	-0.699

Table 1: Correlation coefficients of multiple alignments.

Note that here only the word-based lexicon model is used to compute the confidence measure. More complex models such as alignment models, fertility models and distortion models as described in (Brown et al., 1994) could estimate the probability of a given alignment more accurately. However the summation over all possible alignments is very complicated, even intractable, with the richer models. For the efficient computation of the denominator, we use the lexical translation model.

Similarly,

$$P_{t2s}(A|T,S) = \frac{P(A,S|T)}{\sum_{A'} P(A',S|T)}, \quad (5)$$

and

$$P(A,S|T) = \prod_{i=1}^{I} p(s_i|t_j, a_{ij} \in A). \quad (6)$$

$$\sum_{A'} P(A',S|T) = \prod_{i=1}^{I} \sum_{j=1}^{J} p(s_i|t_j). \quad (7)$$

We randomly selected 512 Chinese-English (C-E) sentence pairs and generated word alignment using the MaxEnt aligner (Ittycheriah and Roukos, 2005). We evaluate per sentence alignment F-scores by comparing the system output with a reference alignment. For each sentence pair, we also calculate the sentence alignment confidence score $-\log C(A|S,T)$. We compute the correlation coefficients between the alignment confidence measure and the alignment F-scores. The results in Figure 2 shows strong correlation between the confidence measure and the alignment F-score, with the correlation coefficients equals to -0.69. Such strong correlation is also observed on an HMM alignment (Ge, 2004) and a Block Model (BM) alignment (Zhao et al., 2005) with varying alignment accuracies, as seen in Table1.

2.2 Sentence Alignment Selection Based on Confidence Measure

The strong correlation between the sentence alignment confidence measure and the alignment F-

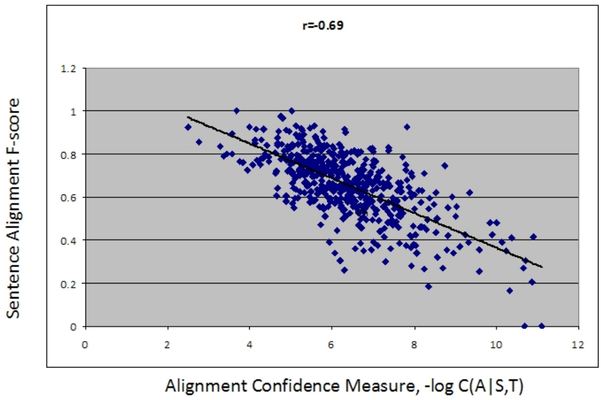

Figure 2: Correlation between sentence alignment confidence measure and F-score.

measure suggests the possibility of selecting the alignment with the highest confidence score to obtain better alignments. For each sentence pair in the C-E test set, we calculate the confidence scores of the HMM alignment, the Block Model alignment and the MaxEnt alignment, then select the alignment with the highest confidence score. As a result, 82% of selected alignments have higher F-scores, and the F-measure of the combined alignments is increased over the best aligner (the MaxEnt aligner) by 0.8. This relatively small improvement is mainly due to the selection of the *whole* sentence alignment: for many sentences the best alignment still contains alignment errors, some of which could be fixed by other aligners. Therefore, it is desirable to combine alignment links from different alignments.

3 Alignment Link Confidence Measure

3.1 Definition

Similar to the sentence alignment confidence measure, the confidence of an alignment link a_{ij} in the sentence pair (S, T) is defined as

$$c(a_{ij}|S,T) = \sqrt{q_{s2t}(a_{ij}|S,T)q_{t2s}(a_{ij}|T,S)} \tag{8}$$

where the source-to-target link posterior probability

$$q_{s2t}(a_{ij}|S,T) \quad = \quad \frac{p(t_j|s_i)}{\sum_{j'=1}^{J} p(t_{j'}|s_i)}, \tag{9}$$

which is defined as the word translation probability of the aligned word pair divided by the sum of the translation probabilities over all the target words in the sentence. The higher $p(t_j|s_i)$ is, the higher confidence the link has. Similarly, the target-to-source link posterior probability is defined as:

$$q_{t2s}(a_{ij}|T,S) \quad = \quad \frac{p(s_i|t_j)}{\sum_{i'=1}^{I} p(s_{i'}|t_j)}. \tag{10}$$

Intuitively, the above link confidence definition compares the lexical translation probability of the aligned word pair with the translation probabilities of all the target words given the source word. If a word t occurs N times in the target sentence, for any $i \in \{1, ..., I\}$,

$$\sum_{j'=1}^{J} p(t_{j'}|s_i) \geq Np(t|s_i),$$

thus for any $t_j = t$,

$$q_{s2t}(a_{ij}) \leq \frac{1}{N}.$$

This indicates that the confidence score of any link connecting t_j to any source word is at most $1/N$. On the one hand this is expected because multiple occurrences of the same word does increase the confusion for word alignment and reduce the link confidence. On the other hand, additional information (such as the distance of the word pair, the alignment of neighbor words) could indicate higher likelihood for the alignment link. We will introduce a context-dependent link confidence measure in section 4.

3.2 Alignment Link Selection

From multiple alignments of the same sentence pair, we select high confidence links from different alignments based on their link confidence scores and alignment agreement ratio.

Typically, links appearing in multiple alignments are more likely correct alignments. The alignment agreement ratio measures the *popularity* of a link. Suppose the sentence pair (S, T) have alignments $A_1, \ldots, A_D$, the agreement ratio of a link a_{ij} is defined as

$$r(a_{ij}|S,T) = \frac{\sum_d C(A_d|S,T\colon a_{ij} \in A_d)}{\sum_{d'} C(A_{d'}|S,T)}, \tag{11}$$

where $C(A)$ is the confidence score of the alignment A as defined in formula 1. This formula computes the sum of the alignment confidence scores for the alignments containing a_{ij}, which is

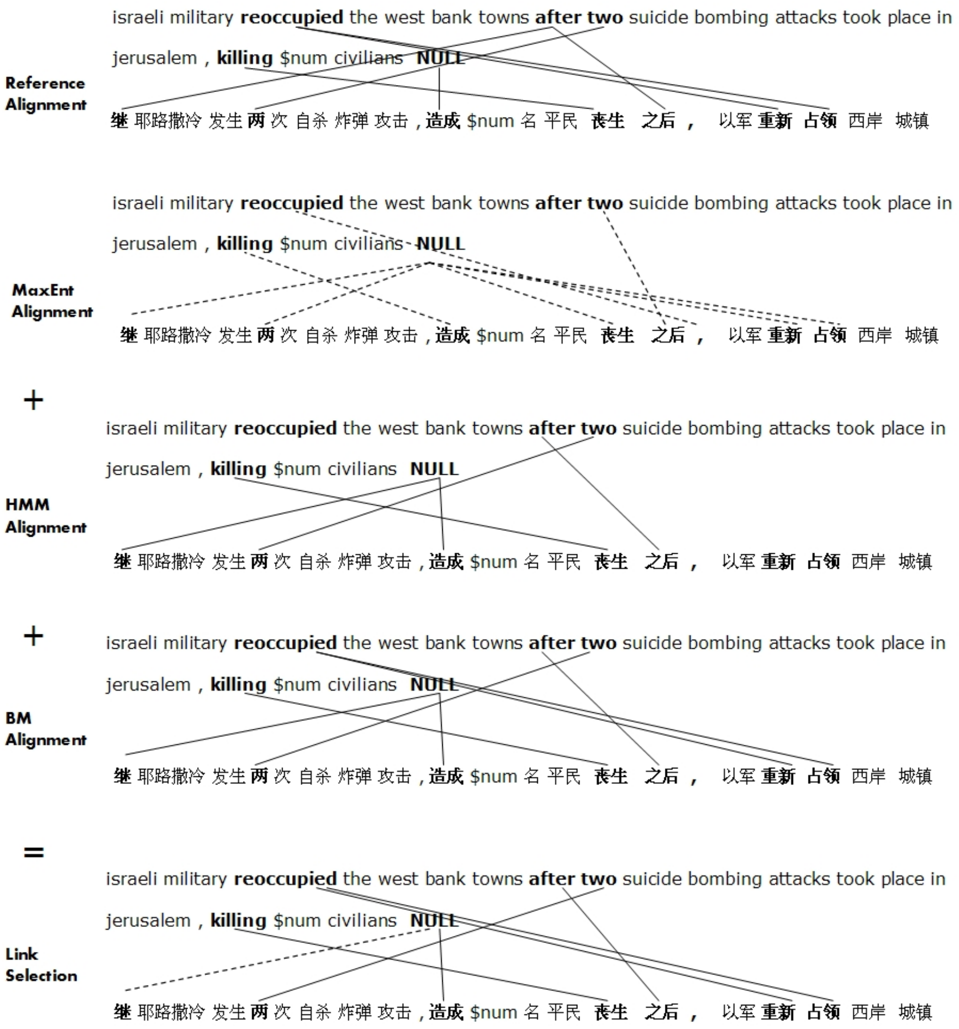
Reference Alignment
israeli military reoccupied the west bank towns after two suicide bombing attacks took place in
jerusalem , killing $num civilians NULL
继 耶路撒冷 发生 两 次 自杀 炸弹 攻击 , 造成 $num 名 平民 丧生 之后 , 以军 重新 占领 西岸 城镇
MaxEnt Alignment
israeli military reoccupied the west bank towns after two suicide bombing attacks took place in
jerusalem , killing $num civilians NULL
继 耶路撒冷 发生 两 次 自杀 炸弹 攻击 , 造成 $num 名 平民 丧生 之后 , 以军 重新 占领 西岸 城镇
+
HMM Alignment
israeli military reoccupied the west bank towns after two suicide bombing attacks took place in
jerusalem , killing $num civilians NULL
继 耶路撒冷 发生 两 次 自杀 炸弹 攻击 , 造成 $num 名 平民 丧生 之后 , 以军 重新 占领 西岸 城镇
+
BM Alignment
israeli military reoccupied the west bank towns after two suicide bombing attacks took place in
jerusalem , killing $num civilians NULL
继 耶路撒冷 发生 两 次 自杀 炸弹 攻击 , 造成 $num 名 平民 丧生 之后 , 以军 重新 占领 西岸 城镇
=
Link Selection
israeli military reoccupied the west bank towns after two suicide bombing attacks took place in
jerusalem , killing $num civilians NULL
继 耶路撒冷 发生 两 次 自杀 炸弹 攻击 , 造成 $num 名 平民 丧生 之后 , 以军 重新 占领 西岸 城镇

Figure 3: Example of alignment link selection by combining MaxEnt, HMM and BM alignments.

normalized by the sum of all alignments' confidence scores.

We collect all the links from all the alignments. For each link we calculate the link confidence score $c(a_{ij})$ and the alignment agreement ratio $r(a_{ij})$. We link the word pair $(s_i,\ t_j)$ if either $c(a_{ij}) > h_1$ or $r(a_{ij}) > r_1$, where h_1 and r_1 are empirically chosen thresholds.

We combine the HMM alignment, the BM alignment and the MaxEnt alignment (ME) using the above link selection algorithm. Figure 3 shows such an example, where alignment errors in the MaxEnt alignment are shown with dotted lines. As some of the links are correctly aligned in the HMM and BM alignments (shown with solid lines), the combined alignment corrects some alignment errors while still contains common incorrect alignment links.

Table 2 shows the precision, recall and F-score of individual alignments and the combined alignment. F-content and F-function are the F-scores for content words and function words, respectively. The link selection algorithm improves the recall over the best aligner (the ME alignment) by 7 points (from 65.4 to 72.5) while decreasing the precision by 4.4 points (from 73.6 to 69.2). Overall it improves the F-score by 1.5 points (from 69.3 to 70.8), 1.8 point improvement for content words and 1.0 point for function words. It also significantly outperforms the traditionally used heuristics, ”intersection-union-refine” (Och and Ney, 2003) by 6 points.

4 Improved MaxEnt Aligner with Confidence-based Link Filtering

In addition to the alignment combination, we also improve the performance of the MaxEnt aligner through confidence-based alignment link filtering. Here we select the MaxEnt aligner because it has

	Precision	Recall	F-score	F-content	F-function
HMM	62.65	48.57	54.72	62.10	34.39
BM	72.76	54.82	62.53	68.64	43.93
ME	72.66	66.17	69.26	72.52	61.41
Link-Select	69.19	72.49	70.81	74.31	60.26
Intersection-Union-Refine	63.34	66.07	64.68	70.15	49.72

Table 2: Link Selection and Combination Results

the highest F-measure among the three aligners, although the algorithm described below can be applied to any aligner.

It is often observed that words within a constituent (such as NP, PP) are typically translated together, and their alignments are close. As a result the confidence measure of an alignment link a_{ij} can be boosted given the alignment of its context words. From the initial sentence alignment we first identify an anchor link a_{mn}, the high confidence alignment link closest to a_{ij}. The anchor link is considered as the most reliable connection between the source and target context. The context is then defined as a window centering at a_{mn} with window width proportional to the distance between a_{ij} and a_{mn}. When computing the context-dependent link confidence, we only consider words within the context window. The context-dependent alignment link confidence is calculated in the following steps:

1. Calculate the *context-independent* link confidence measure $c(a_{ij})$ according to formula (8).

2. Sort all links based on their link confidence measures in decreasing order.

3. Select links whose confidence scores are higher than an empirically chosen threshold H as anchor links [1].

4. Walking along the remaining sorted links. For each link $\{a_{ij} : c(a_{ij}) < H\}$,

   (a) Find the closest anchor link a_{mn}[2],

   (b) Define the context window width $w = |m - i| + |n - j|$.

   (c) Compute the link posterior probabilities within the context window:

   $$q_{s2t}(a_{ij}|a_{mn}) = \frac{p(t_j|s_i)}{\sum_{j'=j-w}^{j+w} p(t_{j'}|s_i)},$$
   $$q_{t2s}(a_{ij}|a_{mn}) = \frac{p(s_i|t_j)}{\sum_{i'=i-w}^{i+w} p(s_{i'}|t_j)}.$$

   (d) Compute the context-dependent link confidence score $c(a_{ij}|a_{mn})$ =

   $$\sqrt{q_{s2t}(a_{ij}|a_{mn})q_{t2s}(a_{ij}|a_{mn})}.$$

   If $c(a_{ij}|a_{mn}) > H$, add a_{ij} into the set of anchor links.

5. Only keep anchor links and remove all the remaining links with low confidence scores.

The above link filtering algorithm is designed to remove incorrect links. Furthermore, it is possible to create new links by relinking unaligned source and target word pairs within the context window if their context-dependent link posterior probability is high.

Figure 4 shows context-independent link confidence scores for the given sentence alignment. The subscript following each word indicates the word's position. Incorrect alignment links are shown with dashed lines, which have low confidence scores ($a_{5,7}$, $a_{7,3}$, $a_{8,2}$, $a_{11,9}$) and will be removed through filtering. When the anchor link $a_{4,11}$ is selected, the context-dependent link confidence of $a_{6,12}$ is increased from 0.12 to 0.51. Also note that a new link $a_{7,12}$ (shown as a dotted line) is created because within the context window, the link confidence score is as high as 0.96. This example shows that the context-dependent link filtering not only removes incorrect links, but also create new links based on updated confidence scores.

We applied the confidence-based link filtering on Chinese-English and Arabic-English word alignment. The C-E alignment test set is the same

[1] H is selected to maximize the F-score on an alignment devset.

[2] When two equally close alignment links have the same confidence score), we randomly select one of the tied links as the anchor link.

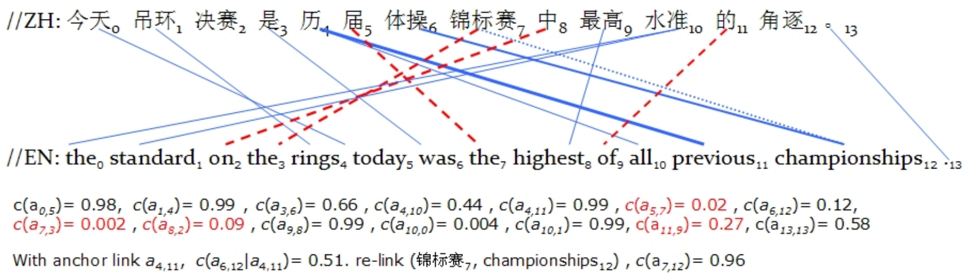
//ZH: 今天$_0$ 吊环$_1$ 决赛$_2$ 是$_3$ 历$_4$ 届$_5$ 体操$_6$ 锦标赛$_7$ 中$_8$ 最高$_9$ 水准$_{10}$ 的$_{11}$ 角逐$_{12}$ 。$_{13}$
//EN: the$_0$ standard$_1$ on$_2$ the$_3$ rings$_4$ today$_5$ was$_6$ the$_7$ highest$_8$ of$_9$ all$_{10}$ previous$_{11}$ championships$_{12}$.$_{13}$
$c(a_{0,5})= 0.98$, $c(a_{1,4})= 0.99$, $c(a_{3,6})= 0.66$, $c(a_{4,10})= 0.44$, $c(a_{4,11})= 0.99$, $c(a_{5,7})= 0.02$, $c(a_{6,12})= 0.12$,
$c(a_{7,3})= 0.002$, $c(a_{8,2})= 0.09$, $c(a_{9,8})= 0.99$, $c(a_{10,0})= 0.004$, $c(a_{10,1})= 0.99$, $c(a_{11,9})= 0.27$, $c(a_{13,13})= 0.58$
With anchor link $a_{4,11}$, $c(a_{6,12}|a_{4,11})= 0.51$. re-link (锦标赛$_7$, championships$_{12}$) , $c(a_{7,12})= 0.96$

Figure 4: Alignment link filtering based on context-independent link confidence.

	Precision	Recall	F-score
Baseline	72.66	66.17	69.26
+ALF	78.14	64.36	70.59

Table 3: Confidence-based Alignment Link Filtering on C-E Alignment

	Precision	Recall	F-score
Baseline	84.43	83.64	84.04
+ALF	88.29	83.14	85.64

Table 4: Confidence-based Alignment Link Filtering on A-E Alignment

512 sentence pairs, and the A-E alignment test set is the 200 Arabic-English sentence pairs from NIST MT03 test set.

Tables 3 and 4 show the improvement of C-E and A-E alignment F-measures with the confidence-based alignment link filtering (ALF). For C-E alignment, removing low confidence alignment links increased alignment precision by 5.5 point, while decreased recall by 1.8 point, and the overall alignment F-measure is increased by 1.3 point. When looking into the alignment links which are removed during the alignment link filtering process, we found that 80% of the removed links (1320 out of 1661 links) are incorrect alignments, For A-E alignment, it increased the precision by 3 points while reducing recall by 0.5 points, and the alignment F-measure is increased by about 1.5 points absolute, a 10% relative alignment error rate reduction. Similarly, 90% of the removed links are incorrect alignments.

5 Translation

We evaluate the improved alignment on several Chinese-English and Arabic-English machine translation tasks. The documents to be translated are from difference genres: newswire (NW) and web-blog (WB). The MT system is a phrase-based SMT system as described in (Al-Onaizan and Papineni, 2006). The training data are bilingual sentence pairs with word alignment, from which we obtained phrase translation pairs. We extract phrase translation tables from the baseline MaxEnt word alignment as well as the alignment with confidence-based link filtering, then translate the test set with each phrase translation table. We measure the translation quality with automatic metrics including BLEU (Papineni et al., 2001) and TER (Snover et al., 2006). The higher the BLEU score is, or the lower the TER score is, the better the translation quality is. We combine the two metrics into (TER-BLEU)/2 and try to minimize it. In addition to the whole test set's scores, we also measure the scores of the ”tail” documents, whose (TER-BLEU)/2 scores are at the bottom 10 percentile (for A-E translation) and 20 percentile (for C-E translation) and are considered the most difficult documents to translate.

In the Chinese-English MT experiment, we selected 40 NW documents, 41 WB documents as the test set, which includes 623 sentences with 16667 words. The training data includes 333 thousand C-E sentence pairs subsampled from 10 million sentence pairs according to the test data. Tables 5 and 6 show the newswire and web-blog translation scores as well as the number of phrase translation pairs obtained from each alignment. Because the alignment link filtering removes many incorrect alignment links, the number of phrase translation pairs is reduced by 15%. For newswire, the translation quality is improved by 0.44 on the whole test set and 1.1 on the tail documents, as measured by (TER-BLEU)/2. For web-blog, we observed 0.2 improvement on the whole test set and 0.5 on the tail documents. The tail documents typically have lower phrase coverage, thus incorrect phrase translation pairs derived from incorrect

	# phrase pairs	Average			Tail		
		TER	BLEU	**(TER-BLEU)/2**	TER	BLEU	**(TER-BLEU)/2**
Baseline	934206	60.74	28.05	**16.35**	69.02	17.83	**25.60**
ALF	797685	60.33	28.52	**15.91**	68.31	19.27	**24.52**

Table 5: Improved Chinese-English Newswire Translation with Alignment Link Filtering

	# phrase pairs	Average			Tail		
		TER	BLEU	**(TER-BLEU)/2**	TER	BLEU	**(TER-BLEU)/2**
Baseline	934206	62.87	25.08	**18.89**	66.55	18.80	**23.88**
ALF	797685	62.30	24.89	**18.70**	65.97	19.25	**23.36**

Table 6: Improved Chinese-English Web-Blog Translation with Alignment Link Filtering

alignment links are more likely to be selected. The removal of incorrect alignment links and cleaner phrase translation pairs brought more gains on the tail documents.

In the Arabic-English MT, we selected 80 NW documents and 55 WB documents. The NW training data includes 319 thousand A-E sentence pairs subsampled from 7.2 million sentence pairs with word alignments. The WB training data includes 240 thousand subsampled sentence pairs. Tables 7 and 8 show the corresponding translation results. Similarly, the phrase table size is significantly reduced by 35%, while the gains on the tail documents range from 0.6 to 1.4. On the whole test set the difference is smaller, 0.07 for the newswire translation and 0.58 for the web-blog translation.

6 Related Work

In the machine translation area, most research on confidence measure focus on the confidence of MT output: how accurate a translated sentence is. (Gandrabur and Foster, 2003) used neural-net to improve the confidence estimate for text predictions in a machine-assisted translation tool. (Ueffing et al., 2003) presented several word-level confidence measures for machine translation based on word posterior probabilities. (Blatz et al., 2004) conducted extensive study incorporating various sentence-level and word-level features thru multilayer perceptron and naive Bayes algorithms for sentence and word confidence estimation. (Quirk, 2004) trained a sentence level confidence measure using a human annotated corpus. (Bach et al., 2008) used the sentence-pair confidence scores estimated with source and target language models to weight phrase translation pairs. However, there has been little research focusing on confidence measure for word alignment. This work is the first attempt to address the alignment confidence problem.

Regarding word alignment combination, in addition to the commonly used ”intersection-union-refine” approach (Och and Ney, 2003), (Ayan and Dorr, 2006b) and (Ayan et al., 2005) combined alignment links from multiple word alignment based on a set of linguistic and alignment features within the MaxEnt framework or a neural net model. While in this paper, the alignment links are combined based on their confidence scores and alignment agreement ratios.

(Fraser and Marcu, 2007) discussed the impact of word alignment’s precision and recall on MT quality. Here removing low confidence links results in higher precision and slightly lower recall for the alignment. In our phrase extraction, we allow extracting phrase translation pairs with unaligned *functional* words at the boundary. This is similar to the ”loose phrases” described in (Ayan and Dorr, 2006a), which increased the number of correct phrase translations and improved the translation quality. On the other hand, removing incorrect *content word* links produced cleaner phrase translation tables. When translating documents with lower phrase coverage (typically the “tail” documents), high quality phrase translations are particularly important because a bad phrase translation can be picked up more easily due to limited phrase translation pairs available.

7 Conclusion

In this paper we presented two alignment confidence measures for word alignment. The first is the sentence alignment confidence measure, based on which the best whole sentence alignment is se-

	# phrase pairs	Average			Tail		
		TER	BLEU	**(TER-BLEU)/2**	TER	BLEU	**(TER-BLEU)/2**
Baseline	939911	43.53	50.51	**-3.49**	53.14	40.60	**6.27**
ALF	618179	43.11	50.24	**-3.56**	51.75	42.05	**4.85**

Table 7: Improved Arabic-English Newswire Translation with Alignment Link Filtering

	# phrase pairs	Average			Tail		
		TER	BLEU	**(TER-BLEU)/2**	TER	BLEU	**(TER-BLEU)/2**
Baseline	598721	49.91	39.90	**5.00**	57.30	30.98	**13.16**
ALF	383561	48.94	40.00	**4.42**	55.99	31.92	**12.04**

Table 8: Improved Arabic-English Web-Blog Translation with Alignment Link Filtering

lected among multiple alignments and it obtained 0.8 F-measure improvement over the single best Chinese-English aligner. The second is the alignment link confidence measure, which selects the most reliable links from multiple alignments and obtained 1.5 F-measure improvement. When we removed low confidence links from the MaxEnt aligner, we reduced the Chinese-English alignment error by 5% and the Arabic-English alignment error by 10%. The cleaned alignment significantly reduced the size of phrase translation tables by 15-35%. It furthermore led to better translation scores for Chinese and Arabic documents with different genres. In particular, it improved the translation scores of the tail documents by 0.5-1.4 points measured by the combined metric of (TER-BLEU)/2.

For future work we would like to explore richer models to estimate alignment posterior probability. In most cases, exact calculation by summing over all possible alignments is impossible, and approximation using N-best alignments is needed.

Acknowledgments

We are grateful to Abraham Ittycheriah, Yaser Al-Onaizan, Niyu Ge and Salim Roukos and anonymous reviewers for their constructive comments. This work was supported in part by the DARPA GALE project, contract No. HR0011-08-C-0110.

References

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion Models for Statistical Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 529–536, Sydney, Australia, July. Association for Computational Linguistics.

Necip Fazil Ayan and Bonnie J. Dorr. 2006a. Going beyond aer: An extensive analysis of word alignments and their impact on mt. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Sydney, Australia, July. Association for Computational Linguistics.

Necip Fazil Ayan and Bonnie J. Dorr. 2006b. A maximum entropy approach to combining word alignments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 96–103, New York City, USA, June. Association for Computational Linguistics.

Necip Fazil Ayan, Bonnie J. Dorr, and Christof Monz. 2005. Neuralign: Combining word alignments using neural networks. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 65–72, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Nguyen Bach, Qin Gao, and Stephan Vogel. 2008. Improving word alignment with language model based confidence scores. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 151–154, Columbus, Ohio, June. Association for Computational Linguistics.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 315, Morristown, NJ, USA. Association for Computational Linguistics.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1994. The Mathematic of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Comput. Linguist.*, 33(3):293–303.

Simona Gandrabur and George Foster. 2003. Confidence estimation for translation prediction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 95–102, Morristown, NJ, USA. Association for Computational Linguistics.

Niyu Ge. 2004. Max-posterior hmm alignment for machine translation. In *Presentation given at DARPA/TIDES NIST MT Evaluation workshop*.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 89–96, Morristown, NJ, USA. Association for Computational Linguistics.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Chris Quirk. 2004. Training a sentence-level machine translation confidence measure. In *In Proc. LREC 2004*, pages 825–828, Lisbon, Portual. Springer-Verlag.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence measures for statistical machine translation. In *In Proc. MT Summit IX*, pages 394–401. Springer-Verlag.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841, Morristown, NJ, USA. Association for Computational Linguistics.

Bing Zhao, Niyu Ge, and Kishore Papineni. 2005. Inner-outer bracket models for word alignment using hidden blocks. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 177–184, Morristown, NJ, USA. Association for Computational Linguistics.