Improving Word Alignment with Bridge Languages

Shankar Kumar and **Franz Och** and **Wolfgang Macherey**
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043, U.S.A.
{shankarkumar,och,wmach}@google.com

Abstract

We describe an approach to improve Statistical Machine Translation (SMT) performance using multi-lingual, parallel, sentence-aligned corpora in several bridge languages. Our approach consists of a simple method for utilizing a bridge language to create a word alignment system and a procedure for combining word alignment systems from multiple bridge languages. The final translation is obtained by consensus decoding that combines hypotheses obtained using all bridge language word alignments. We present experiments showing that multilingual, parallel text in Spanish, French, Russian, and Chinese can be utilized in this framework to improve translation performance on an Arabic-to-English task.

1 Introduction

Word Alignment of parallel texts forms a crucial component of phrase-based statistical machine translation systems. High quality word alignments can yield more accurate phrase-pairs which improve quality of a phrase-based SMT system (Och and Ney, 2003; Fraser and Marcu, 2006b).

Much of the recent work in word alignment has focussed on improving the word alignment quality through better modeling (Och and Ney, 2003; Deng and Byrne, 2005; Martin et al., 2005) or alternative approaches to training (Fraser and Marcu, 2006b; Moore, 2005; Ittycheriah and Roukos, 2005). In this paper we explore a complementary approach to improve word alignments using multi-lingual, parallel (or multi-parallel) corpora. Two works in the literature are very relevant to our approach. Borin (2000) describes a non-statistical approach where a pivot alignment is used to combine direct translation and indirect translation via a third language. Filali and Bilmes (2005) present a multi-lingual extension to the IBM/HMM models. Our current approach differs from this latter work in that we propose a simple framework to combine word alignments from any underlying statistical alignment model without the need for changing the structure of the model. While both of the above papers focus on improving word alignment quality, we demonstrate that our approach can yield improvements in translation performance. In particular, we aim to improve an Arabic-to-English (Ar-En) system using multi-parallel data from Spanish (Es), French (Fr), Russian (Ru) and Chinese (Zh). The parallel data in these languages $X \in \{Es, Fr, Ru, Zh\}$ is used to generate word alignments between Arabic-X and X-English. These alignments are then combined to obtain multiple word alignments for Arabic-English and the final translation systems.

The motivation for this approach is two-fold. First, we believe that parallel corpora available in several languages provide a better training material for SMT systems relative to bilingual corpora. Such multi-lingual parallel corpora are becoming widely available; examples include proceedings of the United Nations in six languages (UN, 2006), European Parliament (EU, 2005; Koehn, 2003), JRC Acquis corpus (EU, 2007) and religious texts (Resnik et al., 1997). Word alignment systems

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 42–50, Prague, June 2007.

trained on different language-pairs (e.g. French-English versus Russian-English) make errors which are somewhat orthogonal. In such cases, incorrect alignment links between a sentence-pair can be corrected when a translation in a third language is available. Thus it can help resolve errors in word alignment. We combine word alignments using several bridge languages with the aim of correcting some of the alignment errors. The second advantage of this approach is that the word alignment from each bridge language can be utilized to build a phrase-based SMT system. This provides a diverse collection of translation hypotheses for MT system combination (Bangalore et al., 2002; Sim et al., 2007; Matusov et al., 2006; Macherey and Och, 2007). Finally, a side benefit of this paper is that it provides a study that compares alignment qualities and BLEU scores for models in different languages trained on parallel text which is held identical across all languages.

We show that parallel corpora in multiple languages can be exploited to improve the translation performance of a phrase-based translation system. This paper gives specific recipes for using a bridge language to construct a word alignment and for combining word alignments produced by multiple statistical alignment models.

The rest of this paper is organized as follows: Section 2 gives an overview of our framework for generating word alignments in a single language-pair. In Section 3, we describe how a bridge language may be used for producing word alignments. In Section 4, we describe a scheme to combine word alignments from several bridge languages. Section 5 describes our experimental setup and reports the alignment and translation performance. A final discussion is presented in Section 6.

2 Word Alignment Framework

A statistical translation model (Brown et al., 1993; Och and Ney, 2003) describes the relationship between a pair of sentences in the source and target languages ($\mathbf{f} = f_1^J, \mathbf{e} = e_1^I$) using a translation probability $P(\mathbf{f}|\mathbf{e})$. Alignment models introduce a hidden alignment variable $\mathbf{a} = a_1^J$ to specify a mapping between source and target words; $a_j = i$ indicates that the j^{th} source word is linked to the i^{th} target word. Alignment models assign a probability $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ to the source sentence and alignment conditioned on the target sentence. The translation probability is related to the alignment model as: $P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P_\theta(\mathbf{f}, \mathbf{a}|e)$, where θ is a set of parameters.

Given a sentence-pair $(\mathbf{f}, \mathbf{e})$, the most likely (Viterbi) word alignment is found as (Brown et al., 1993): $\hat{\mathbf{a}} = \mathrm{argmax}_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$. An alternate criterion is the Maximum A-Posteriori (MAP) framework (Ge, 2004; Matusov et al., 2004). We use a refinement of this technique.

Given any word alignment model, posterior probabilities can be computed as (Brown et al., 1993)

$$P(a_j = i|\mathbf{e}, \mathbf{f}) = \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{f}, \mathbf{e})\delta(i, a_j), \qquad (1)$$

where $i \in \{0, 1, ..., I\}$. The assignment $a_j = 0$ corresponds to the NULL (empty) alignment. These posterior probabilities form a matrix of size $(I+1) \times J$, where entries along each column sum to one.

The MAP alignment for each source position $j \in \{1, 2, ..., J\}$ is then computed as

$$a_{MAP}(j) = \underset{i}{\mathrm{argmax}}\, P(a_j = i|\mathbf{e}, \mathbf{f}). \qquad (2)$$

We note that these posterior probabilities can be computed efficiently for some alignment models such as the HMM (Vogel et al., 1996; Och and Ney, 2003), Models 1 and 2 (Brown et al., 1993).

In the next two sections, we describe how posterior probabilities can be used to a) construct alignment systems from a bridge language, and b) merge several alignment systems.

3 Constructing Word Alignment Using a Bridge Language

We assume here that we have triples of sentences that are translations of each other in languages F, E, and the bridge language G: $\mathbf{f} = f_1^J, \mathbf{e} = e_1^I, \mathbf{g} = g_1^K$. Our goal is to obtain posterior probability estimates for the sentence-pair in FE: $(\mathbf{f}, \mathbf{e})$ using the posterior probability estimates for the sentence pairs in FG: $(\mathbf{f}, \mathbf{g})$ and GE: $(\mathbf{g}, \mathbf{e})$. The word alignments between the above sentence-pairs are referred to as $\mathbf{a}^{FE}$, $\mathbf{a}^{FG}$, and $\mathbf{a}^{GE}$ respectively; the notation $\mathbf{a}^{FE}$ indicates that the alignment maps a position in F to a position in E.

We first express the posterior probability as a sum over all possible translations g in G and hidden alignments a^{FG}.

$$
\begin{aligned}
& P(a_j^{FE}=i|\mathbf{e},\mathbf{f}) \\
= & \sum_{\mathbf{g}} P(a_j^{FE}=i,\mathbf{g}|\mathbf{e},\mathbf{f}) \\
= & \sum_{\mathbf{g},k} P(a_j^{FE}=i,\mathbf{g},a_j^{FG}=k|\mathbf{e},\mathbf{f}) \\
= & \sum_{\mathbf{g},k} \Big\{ P(\mathbf{g}|\mathbf{e},\mathbf{f}) P(a_j^{FG}=k|\mathbf{g},\mathbf{e},\mathbf{f}) \\
& \times P(a_j^{FE}=i|a_j^{FG}=k,\mathbf{g},\mathbf{e},\mathbf{f}) \Big\} \qquad (3)
\end{aligned}
$$

We now make some assumptions to simplify the above expression. First, there is exactly one translation g in bridge language G corresponding to the sentence-pair $\mathbf{f},\mathbf{e}$. Since $a^{GE}_{a_j^{FG}} = i = a_j^{FE}$, we can express
$P(a_j^{FE}=i|a_j^{FG}=k,\mathbf{g},\mathbf{f},\mathbf{e}) = P(a_k^{GE}=i|\mathbf{g},\mathbf{e})$.
Finally, alignments in FG do not depend on E.

Under these assumptions, we arrive at the final expression for the posterior probability FE in terms of posterior probabilities for GF and EG

$$
\begin{aligned}
& P(a_j^{FE}=i|\mathbf{e},\mathbf{f}) = \qquad (4) \\
& \sum_{k=0}^{K} P(a_j^{FG}=k|\mathbf{g},\mathbf{f}) P(a_k^{GE}=i|\mathbf{g},\mathbf{e})
\end{aligned}
$$

The above expression states that the posterior probability matrix for FE can be obtained using a *simple matrix multiplication* of posterior probability matrices for GE and FG. In this multiplication, we prepend a column to the GE matrix corresponding to $k=0$. This probability $P(a_k^{GE}=i)$ when $k=0$ is not assigned by the alignment model; we set it as follows

$$
P(a_k^{GE}=i|k=0) = \begin{cases} \epsilon & i=0 \\ \frac{1-\epsilon}{I} & i \in \{1,2,...,I\} \end{cases}
$$

The parameter ϵ controls the number of empty alignments; a higher value favors more empty alignments and vice versa. In our experiments, we set $\epsilon = 0.5$.

4 Word Alignment Combination Using Posterior Probabilities

We next show how Word Alignment Posterior Probabilities can be used for combining multiple word alignment systems. In our context, we use this procedure to combine word alignments produced using multiple bridge languages.

Suppose we have translations in bridge languages $G_1, G_2, ..., G_N$, we can generate a posterior probability matrix for FE using each of the bridge languages. In addition, we can always generate a posterior probability matrix for FE with the FE alignment model directly without using any bridge language. These $N+1$ posterior matrices can be combined as follows. Here, the variable B indicates the bridge language. $B \in \{G_0, G_1, ..., G_N\}$; G_0 indicates the case when no bridge language is used.

$$
\begin{aligned}
& P(a_j^{FE}=i|\mathbf{e},\mathbf{f}) \qquad (5) \\
= & \sum_{l=0}^{N} P(B=G_l, a_j^{FE}=i|\mathbf{e},\mathbf{f}) \\
= & \sum_{l=0}^{N} P(B=G_l) P(a_j^{FE}=i|G_l,\mathbf{e},\mathbf{f}),
\end{aligned}
$$

where $P(a_j^{FE}=i|G_l,j,\mathbf{e},\mathbf{f})$ is the posterior probability when bridge language $B=G_l$. The probabilities $P(B=G_l)$ sum to one over $l \in \{0,1,2,...,N\}$ and represent the prior probability of bridge language l. In our experiments, we use a uniform prior $P(B=G_l) = \frac{1}{N+1}$. Equation 5 provides us a way to combine word alignment posterior probabilites from multiple bridge languages. In our alignment framework (Section 2), we first interpolate the posterior probability matrices (Equation 5) and then extract the MAP word alignment (Equation 2) from the resulting matrix.

5 Experiments

We now present experiments to demonstrate the advantages of using bridge languages. Our experiments are performed in the open data track of the NIST Arabic-to-English (A-E) machine translation task [1].

5.1 Training and Test Data

Our approach to word alignment (Section 3) requires aligned sentences in multiple languages. For training alignment models, we use the ODS United Na-

[1] `http://www.nist.gov/speech/tests/mt/`

Set	# of Ar words (K)	# of sentences
dev1	48.6	2007
dev2	11.4	498
test	37.8	1610
blind	36.5	1797

Table 1: Statistics for the test data.

tions parallel data (UN, 2006) which contains parliamentary documents from 1993 onwards in all six official languages of the UN: Arabic (Ar), Chinese (Zh), English (En), French (Fr), Russian (Ru), and Spanish (Es).

We merge the NIST 2001-2005 Arabic-English evaluation sets into a pool and randomly sample this collection to create two development sets (dev1,dev2) and a test set (test) with 2007, 498, and 1610 sentences respectively. Our blind test (blind) set is the NIST part of the NIST 06 evaluation set consisting of 1797 sentences. The GALE portion of the 06 evaluation set is not used in this paper. We report results on the test and blind sets. Some statistics computed on the test data are shown in Table 1.

5.2 Alignment Model Training

For training Arabic-English alignment models, we use Chinese, French, Russian and Spanish as bridge languages. We train a model for Ar-En and 4 models each for Ar-X and X-En, where X is the bridge language. To obtain aligned sentences in these language pairs, we train 9 sentence aligners. We then train alignment models for all 9 language-pairs using a recipe consisting of 6 Model-1 iterations and 6 HMM iterations. Finally, Word Alignment Posterior Probabilities are generated over the bitext. In Table 2, we report the perplexities of the alignment models for the translation directions where either Arabic or English is predicted. There are 55M Arabic tokens and 58M English tokens. We observe that the alignment model using Spanish achieves the lowest perplexity; this value is even lower than the perplexity of the direct Arabic-English model. Perplexity is related to the hardness of the word alignment; the results suggest that bridge languages such as Spanish make alignment task easier while others do not. We stress that perplexity is not related to the alignment or the translation performance.

Bridge Lang	Perplexity $\rightarrow$ Ar	$\rightarrow$En
None	113.8	26.1
Es	99.0	22.9
Fr	138.6	30.2
Ru	128.3	27.5
Zh	126.1	34.6

Table 2: Perplexities of the alignment models.

5.3 Bridge Language Word Alignments

Each of the 4 bridge languages is utilized for constructing a word alignment for Arabic-English. Using each bridge language X, we obtain Arabic-English word alignments in both translation directions (AE and EA). The posterior matrix for AE is obtained using AX and XE matrices while the EA matrix is obtained from EX and XA matrices (Equation 4). The AE (EA) matrices from the bridge languages are then interpolated with the AE (EA) matrix obtained from the alignment model trained directly on Arabic-English (Section 4). The MAP word alignment for AE (EA) direction is computed from the AE (EA) matrix. We next outline how these word alignments are utilized in building a phrase-based SMT system.

5.4 Phrase-based SMT system

Our phrase-based SMT system is similar to the alignment template system described in Och and Ney (2004). We first extract an inventory of phrase-pairs up to length 7 from the union of AE and EA word alignments. Various feature functions (Och and Ney, 2004) are then computed over the entries in the phrase table. 5-gram word language models in English are trained on a variety of monolingual corpora (Brants et al., 2007). Minimum Error Rate Training (MERT) (Och, 2003) under BLEU criterion is used to estimate 20 feature function weights over the larger development set (dev1).

Translation is performed using a standard dynamic programming beam-search decoder (Och and Ney, 2004). Decoding is done in two passes. An initial list of 1000-best hypotheses is generated by the decoder. This list is then rescored using Minimum Bayes-Risk (MBR) decoding (Kumar and Byrne, 2004). The MBR scaling parameter is tuned on the smaller development set (dev2).

Bridge Language	Metrics(%)					
	AE			EA		
	Prec	Rec	AER	Prec	Rec	AER
None	74.1	73.9	26.0	67.3	57.7	37.9
Es	61.7	56.3	41.1	50.0	40.2	55.4
Fr	52.9	48.0	49.7	42.3	33.6	62.5
Ru	57.4	50.8	46.1	40.2	31.6	64.6
Zh	44.3	39.3	58.3	39.7	29.9	65.9
AC1	70.0	65.0	32.6	56.8	46.4	48.9

Table 3: Alignment Performance with Bridge Languages

5.5 Alignment Results

We first report alignment performance (Table 3) of the alignment models obtained using the bridge languages. Alignment results are reported in terms of Precision (Prec), Recall (Rec) and Alignment Error Rate (AER). We report these numbers on a 94-sentence test set with translations in all six languages and human word alignments in Arabic-English. Our human word alignments do not distinguish between *Sure* and *Probable* links (Och and Ney, 2003).

In these experiments, we first identify the common subset of sentences which have translations in all six languages. Each of the 9 alignment models is then trained on this subset. We report Alignment performance in both translation directions: Arabic-to-English (AE) and English-to-Arabic (EA). The first row (None) gives the results when no bridge language is used.

Among the bridge languages, Spanish gives the best alignment for Arabic-English while Chinese results in the worst. This might be related to how different the bridge language is relative to either English or Arabic. The last row (AC1) shows the performance of the alignment obtained by combining None/Es/Fr/Ru/Zh alignments. This alignment outperforms all bridge alignments but is weaker than the alignment without any bridge language. Our hypothesis is that a good choice of interpolation weights (Equation 5) would reduce AER of the AC1 combination. However, we did not investigate these choices in this paper. We report alignment error rates here to give the readers an idea of the vastly different alignment performance using each of the bridge languages.

5.6 Translation Results

We now report translation performance of our techniques. We measure performance using the NIST implementation of case sensitive BLEU-4 on true-cased translations. We observed in experiments not reported here that results are almost identical with/without Minimum Error Rate Training ; we therefore report the results without the training. We note that the blind set is the NIST subset of the 2006 NIST evaluation set. The systems reported here are for the Unlimited Data Track in Arabic-to-English and obtain competitive performance relative to the results reported on the NIST official results page [2]

We present three sets of experiments. In Table 4, we describe the first set where all 9 alignment models are trained on nearly the same set of sentences (1.9M sentences, 57.5M words in English). This makes the alignment models in all bridge languages comparable. In the first row marked None, we do not use a bridge language. Instead, an Ar-En alignment model is trained directly on the set of sentence pairs. The next four rows give the performance of alignment models trained using the bridge languages Es, Fr, Ru and Zh respectively. For each language, we use the procedure (Section 3) to obtain the posterior probability matrix for Arabic-English from Arabic-X and X-English matrices. The row AC1 refers to alignment combination using interpolation of posterior probabilities described in Section 4. We combine posterior probability matrices from the systems in the first four rows: None, Es, Ru and Zh. We exclude the Zh system from the AC1 combination because it is found to degrade the translation performance by 0.2 points on the test set.

In the final six rows of Table 4, we show the performance of a consensus decoding technique that produces a single output hypothesis by combining translation hypotheses from multiple systems; this is an MBR-like candidate selection procedure based on BLEU correlation matrices and is described in Macherey and Och (2007). We first report performance of the consensus output by combining None systems with/without MERT. Each of the following rows provides the results from consensus decoding for adding an extra system both with/without MERT. Thus, the final row (TC1) combines transla-

[2] http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html

tions from 12 systems: None, Es, Fr, Ru, Zh, AC1 with/without MERT. All entries marked with an asterisk are better than the None baseline with 95% statistical significance computed using paired bootstrap resampling (Koehn, 2004).

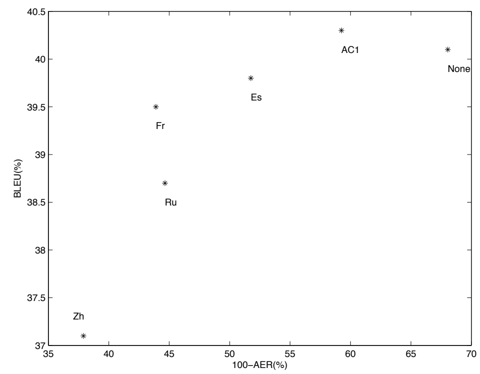

Figure 1: 100-AER (%) vs. BLEU(%) on the blind set for 6 systems from Table 3.

Figure 1 shows the plot between 100-AER% (average of EA/AE directions) and BLEU for the six systems in Table 3. We observe that AER is loosely correlated to BLEU ($\rho = 0.81$) though the relation is weak, as observed earlier by Fraser and Marcu (2006a). Among the bridge languages, Spanish gives the lowest AER/highest BLEU while Chinese results in highest AER/lowest BLEU. We can conclude that Spanish is closest to Arabic/English while Chinese is the farthest. All the bridge languages yield lower BLEU/higher AER relative to the No-Bridge baseline. Therefore, our estimate of the posterior probability (Equation 4) is always worse than the posterior probability obtained using a direct model. The alignment combination (AC1) behaves differently from other bridge systems in that it gives a higher AER and a higher BLEU relative to None baseline. We hypothesize that AC1 is different from the bridge language systems since it arises from a different process: interpolation with the direct model (None).

Both system combination techniques give improvements relative to None baseline: alignment combination AC1 gives a small gain (0.2 points) while the consensus translation TC1 results in a larger improvement (0.8 points). The last 4 rows of the table show that the performance of the hypothesis consensus steadily increases as systems get added to the None baseline. This shows that while bridge language systems are weaker than the direct model, they can provide complementary sources of evidence. To further validate this hypothesis, we compute inter-system BLEU scores between None/es and all the systems in Table 5. We observe that the baseline (None) is very dissimilar from the rest of the systems. We hypothesize that the baseline system has an alignment derived from a real alignment model while the rest of the bridge systems are derived using matrix multiplication. The low inter-system BLEU scores show that the bridge systems provide diverse hypotheses relative to the baseline and therefore contribute to gains in consensus decoding.

Bridge Lang	# Msents	BLEU (%)	
		test	blind
None	1.9	52.1	40.1
Es	1.9	51.7	39.8
Fr	1.9	51.2	39.5
Ru	1.9	50.4	38.7
Zh	1.9	48.4	37.1
AC1	1.9	52.1	40.3
Hypothesis Consensus			
None	1.9	51.9	39.8
+Es	1.9	52.2	40.0
+Fr	1.9	52.4*	40.5*
+Ru	1.9	52.8*	40.7*
+Zh	1.9	52.6*	40.6*
+AC1 = TC1	1.9	**53.0***	**40.9***

Table 4: Translation Experiments for Set 1; Results are reported on the test and blind set: (NIST portion of 2006 NIST eval set).

Ref	None	es	fr	ru	zh	AC1
None	100.0	60.0	59.8	59.7	59.5	58.7
es	59.6	100.0	79.9	69.3	67.4	70.5

Table 5: Inter-system BLEU scores (%) between None/es and all systems in Table 3.

To gain some insight about how the bridge systems help in Table 4, we present an example in Table 6. The example shows the consensus Translations and the 12 input translations for the consensus decoding. The example suggests that the inputs to the consensus decoding exhibit diversity.

Table 7 reports the second and third sets of experiments. For both sets, we first train each bridge language system X using all aligned sentences avail-

System	MERT	Hypothesis
None	N	The President of the National Conference Visit Iraqi Kurdistan Iraqi
None	Y	President of the Iraqi National Conference of Iraqi Kurdistan Visit
Es	N	President of the Iraqi National Congress to Visit Iraqi Kurdistan
Es	Y	President of the Iraqi National Congress to Visit Iraqi Kurdistan
Fr	N	President of the Iraqi National Conference Visits Iraqi Kurdistan
Fr	Y	Chairman of the Iraqi National Conference Visits Iraqi Kurdistan
Ru	N	The Chairman of the Iraqi National Conference Visits Iraqi Kurdistan
Ru	Y	Chairman of the Iraqi National Conference Visit the Iraqi Kurdistan
Zh	N	The Chairman of the Iraqi National Conference Visits Iraqi Kurdistan
Zh	Y	The Chairman of the Iraqi National Conference Visit Iraqi Kurdistan
AC1	N	President of the Iraqi National Congress to Visit Iraqi Kurdistan
AC1	Y	Chairman of the Iraqi National Congress to Visit Iraqi Kurdistan
TC1	-	The Chairman of the Iraqi National Conference Visits Iraqi Kurdistan
Ref	-	Head of Iraqi National Congress Visits Iraqi Kurdistan

Table 6: An example showing the Consensus Translation (TC1) and the 12 inputs for consensus decoding. The final row shows the reference translation.

able in Ar, En and X. In Set 2, the first row (Union) is an alignment model trained on all sentence-pairs in Ar-En which are available in at least one bridge language X. AC2 refers to alignment combination using bridge languages Es/Fr/Ru and Union. TC2 refers to the translation combination from 12 systems: Es/Fr/Ru/Zh/Union/AC2 with/without Minimum Error Rate training. Finally, the goal in Set 3 (last 3 rows) is to improve the best Arabic-English system that can be built using all available sentence pairs from the UN corpus. The first row (Direct) gives the performance of this Ar-En system; AC3 refers to alignment combination using Es/Fr/Ru and Direct. TC3 merges translations from Es/Fr/Ru/Zh/Direct/AC3. All entries marked with an asterisk (plus) are better than the Union (Direct) baseline with 95% statistical significance computed using paired bootstrap resampling (Koehn, 2004).

The motivation behind Sets 2 and 3 is to train all bridge language systems on as much bitext as possible. As a consequence, these systems give better results than the corresponding systems in Table 4. The Union system outperforms None by 1.7/1.4 BLEU points and provides a better baseline. We show under this scenario that system combination techniques AC2 and TC2 can still give smaller improvements (0.3/0.5 and 1.0/0.7 points) relative to this baseline.

As mentioned earlier, our approach requires sentence-aligned corpora. In our experiments, we use a single sentence aligner for each language pair (total of 9 aligners). Since these aligners make independent decisions on sentence boundaries, we end up with a smaller pool of sentences (1.9M) that is common across all language pairs. In contrast, a sentence aligner that makes simultaneous decisions in multiple languages would result in a larger set of common sentence pairs (close to 7M sentence pairs). Simard (1999) describes a sentence aligner of this type that improves alignment on a trilingual parallel text. Since we do not currently have access to such an aligner, we simulate that situation with Sets 2 and 3: AC2/AC3 do not insist that a sentence-pair be present in all input word alignments. We note that Set 2 is a data scenario that falls between Sets 1 and 3.

Set 3 provides the best baseline for Arabic-English based on the UN data by training on all parallel sentence-pairs. In this situation, system combination with bridge languages (AC3/TC3) gives reasonable improvements in BLEU on the test set (0.4/1.0 points) but only modest improvements (0.1/0.4 points) on the blind set. However, this does show that the bridge systems continue to provide orthogonal evidence at different operating points.

6 Discussion

We have described a simple approach to improve word alignments using bridge languages. This includes two components: a matrix multiplication to assemble a posterior probability matrix for the desired language-pair FE using a pair of posterior probability matrices FG and GE relative to a bridge language G. The second component is a recipe for combining word alignment systems by linearly in-

Bridge Lang	# Msents	BLEU (%) test	BLEU (%) blind
Es	4.7	53.7	40.9
Fr	4.7	53.2	40.7
Ru	4.5	52.4	39.9
Zh	3.4	49.7	37.9
Set 2			
Union	7.2	53.8	41.5
AC2	7.2	54.1	42.0*
TC2	-	**54.8***	**42.2***
Set 3			
Direct	7.0	53.9	42.2
AC3	9.0	54.3^{+}	42.3
TC3	-	**54.9**$^{+}$	**42.6**$^{+}$

Table 7: Translation performance for Sets 2 and 3 on test and blind:NIST portion of 2006 NIST eval set.

terpolating posterior probability matrices from different sources. In our case, these sources are multiple bridge languages. However, this method is more generally applicable for combining posterior matrices from different alignment models such as HMM and Model-4. Such an approach contrasts with the log-linear HMM/Model-4 combination proposed by Och and Ney (2003).

There has been recent work by Ayan and Dorr (2006) on combining word alignments from different alignment systems; this paper describes a maximum entropy framework for this combination. Their approach operates at the level of the alignment links and uses maximum entropy to decide whether or not to include an alignment link in the final output. In contrast, we use posterior probabilities as the interface between different alignment models. Another difference is that this maxent framework requires human word aligned data for training feature weights. We do not require any human word aligned data to train our combiner.

Another advantage of our approach is that it is based on word alignment posterior probability matrices that can be generated by any underlying alignment model. Therefore, this method can be used to combine word alignments generated by fairly dissimilar word alignment systems as long as the systems can produce posterior probabilities.

Bridge languages have been used by NLP researchers as a means to induce translation lexicons between distant languages without the need for parallel corpora (Schafer and Yarowsky, 2002; Mann and Yarowsky, 2001). Our current approach differs from these efforts in that we use bridge languages to improve word alignment quality between sentence pairs. Furthermore, we do not use linguistic insight to identify bridge languages. In our framework, a good bridge language is one that provides the best translation performance using the posterior matrix multiplication. Our experiments show that Spanish is a better bridge language relative to Chinese for Arabic-to-English translation. We speculate that if our approach was carried out on a data set with hundreds of languages, we might be able to automatically identify language families.

A downside of our approach is the requirement for exact sentence-aligned parallel data. Except for a few corpora such as UN, European Parliament etc, such a resource is hard to find. One solution is to create such parallel data by automatic translation and then retaining reliable translations by using confidence metrics (Ueffing and Ney, 2005).

Our approach to using bridge languages is extremely simple. Despite its simplicity, the system combination gives improvements in alignment and translation performance. In future work, we will consider several extensions to this framework that lead to more powerful system combination strategies using multiple bridge languages. We recall that the present approach trains bridge systems (e.g. Arabic-to-French, French-to-English) until the alignment stage and then uses these for constructing Arabic-to-English word alignment. An alternate scenario would be to build phrase-based SMT systems for Arabic-to-Spanish and Spanish-to-English, and then obtain Arabic-to-English translation by first translating from Arabic into Spanish and then Spanish into English. Such end-to-end bridge systems may lead to an even more diverse pool of hypotheses that could further improve system combination.

References

N. Ayan and B. Dorr. 2006. A maximum entropy approach to combining word alignments. In *HLT-NAACL*, New York, New York.

S. Bangalore, V. Murdock, and G. Riccardi. 2002. Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *COLING*, Taipei, Taiwan.

L. Borin. 2000. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *COLING*, pages 97–103, Saarbrucken, Germany.

T. Brants, A. Popat, P. Xu, F. Och, and J. Dean. 2007. Large language models in machine translation. In *EMNLP*, Prague, Czech Republic.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Y. Deng and W. Byrne. 2005. HMM word and phrase alignment for statistical machine translation. In *EMNLP*, Vancouver, Canada.

EU, 2005. *European Parliament Proceedings*. http://www.europarl.europa.eu.

EU, 2007. *JRC Acquis Corpus*. http://langtech.jrc.it/JRC-Acquis.html.

K. Filali and J. Bilmes. 2005. Leveraging multiple languages to improve statistical mt word alignments. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico.

A. Fraser and D. Marcu. 2006a. Measuring word alignment quality for statistical machine translation. Technical Report ISI-TR-616, ISI/University of Southern California.

A. Fraser and D. Marcu. 2006b. Semi-supervised training for statistical word alignment. In *ACL*, pages 769–776, Sydney, Australia.

N. Ge. 2004. Improvements in word alignments. In *Presentation given at DARPA/TIDES workshop*.

A. Ittycheriah and S. Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *EMNLP*, Vancouver, Canada.

P. Koehn, 2003. *European Parliament Proceedings, Sentence Aligned*. http://people.csail.mit.edu/koehn/publications/europarl/.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, Barcelona, Spain.

S. Kumar and W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-NAACL*, pages 169–176, Boston, MA, USA.

W. Macherey and F. Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *EMNLP*, Prague, Czech Republic.

G. Mann and D. Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *NAACL*, Pittsburgh, PA, USA.

J. Martin, R. Mihalcea, and T. Pedersen. 2005. Word alignment for languages with scarce resources. In *ACL Workshop on Building and Using Parallel Texts*, pages 65–74, Ann Arbor, MI, USA.

E. Matusov, R. Zens, and H. Ney. 2004. Symmetric word alignments for statistical machine translation. In *COLING*, Geneva, Switzerland.

E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *EACL*, Trento, Italy.

R. C. Moore. 2005. A discriminative framework for bilingual word alignment. In *EMNLP*, Vancouver, Canada.

F. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19 – 51.

F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417 – 449.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, Sapporo, Japan.

P. Resnik, M. Olsen, and M. Diab. 1997. Creating a parallel corpus from the book of 2000 tongues. In *Text Encoding Initiative 10th Anniversary User Conference*, Providence, RI, USA.

C. Schafer and D. Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*, Taipei, Taiwan.

K. C. Sim, W. J. Byrne, M. J. F. Gales, H. Sahbi, and P. C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA.

M. Simard. 1999. Text translation alignment: Three languages are better than two. In *EMNLP-VLC*, College Park, MD, USA.

N. Ueffing and H. Ney. 2005. Word-level confidence estimation for machine translation using phrase-based translation models. In *EMNLP*, pages 763 – 770, Vancouver, Canada.

UN, 2006. *ODS UN Parallel Corpus*. http://ods.un.org/.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM based word alignment in statistical translation. In *COLING*, pages 836–841, Copenhagen, Denmark.