Multi-Source Translation Methods

Lane Schwartz
University of Minnesota
Minneapolis, MN 55455, USA
lane@cs.umn.edu

Abstract

Multi-parallel corpora provide a potentially rich resource for machine translation. This paper surveys existing methods for utilizing such resources, including hypothesis ranking and system combination techniques. We find that despite significant research into system combination, relatively little is know about how best to translate when multiple parallel source languages are available. We provide results to show that the MAX multilingual multi-source hypothesis ranking method presented by Och and Ney (2001) does not reliably improve translation quality when a broad range of language pairs are considered. We also show that the PROD multilingual multi-source hypothesis ranking method of Och and Ney (2001) cannot be used with standard phrase-based translation engines, due to a high number of unreachable hypotheses. Finally, we present an oracle experiment which shows that current hypothesis ranking methods fall far short of the best results reachable via sentence-level ranking.

1 Introduction

To date, the vast majority of research in machine translation has focused on the task of translating from a single source language into a single target language. Yet governments, companies, and other international organizations commonly translate documents into many languages. In general, documents translated into more than one language will likely be translated into many more languages (Kay, 2000).

The recent development of large *multi-parallel* corpora has made research into multilingual translation practical. A multi-parallel corpus contains the same texts in more than two languages. The Europarl (Koehn, 2005), Acquis Communautaire (Steinberger et al., 2006), and News Commentary (Callison-Burch et al., 2007) corpora are freely available multi-parallel corpora that together include most European languages. Other multi-parallel corpora available in machine-readable format include many United Nations documents (UN, 1994), the Bible (Resnik et al., 1999), and George Orwell's novel *1984* (Erjavec, 2004).

Multi-parallel texts provide a rich source of information which could be exploited to reduce ambiguity and improve translation choices. This paper surveys the current state of the art in techniques to exploit multi-parallel corpora and techniques for using multiple source languages in statistical machine translation and presents experiments which show the limitations of existing hypothesis ranking methods.

The main contributions of this paper are as follows. We show that significant gains in translation quality are reachable by simply selecting the best output hypotheses from a list of system output hypotheses, without performing any word-level or phrase-level system combination. We show that the MAX ranking technique of Och and Ney (2001) does not reliably improve translation quality, in contradiction to the results initially reported for this technique. Finally, we show that the PROD ranking technique is impractical to use with current phrase-based translation, primarily due to problems regarding unreachable hypotheses.

The remainder of this paper is structured as follows. Section 2 reviews related translation techniques that exploit multilingual resources and multiparallel corpora. Section 3 presents oracle experiments which illustrate the gains possible by sentence-level hypothesis ranking.

Section 4 examines existing techniques for selecting the best hypothesis from multiple translation engines, with particular attention paid to the MAX and PROD ranking techniques of Och and Ney (2001). We present new experiments that show the limits of the MAX ranking method for many language pairs, and show that the PROD ranking method of Och and Ney (2001) cannot be directly applied when using a standard phrase-based decoder. The results in this section show that the MAX and PROD ranking methods fall far short of the best results reachable via system output ranking.

Finally, section 5 explores additional methods concerning how multiple sources could be incorporated during translation.

2 Related Work

While relatively little research has examined how multiple source languages can explicitly be used to find a higher quality target translation, numerous existing techniques use multilingual resources to enhance bilingual translation resources.

Simard (1999) presents techniques for aligning sentences in a multi-parallel corpus. Kumar et al. (2007) describe a technique for word alignment in a multi-parallel sentence-aligned corpus and show that this technique can be used to obtain higher quality bilingual word alignments than traditional bilingual word alignment techniques.

Multilingual resources can also be used to directly improve the quality of a bilingual translation phrase table. Callison-Burch (2002) presents a technique in which the best output of several translation systems is used as additional training data, leading to an improvement in translation quality.

Eisele (2006) proposes that existing bilingual translation systems which share one or more common *pivot* languages can be coupled to build translation systems for language pairs for which no parallel corpus exists; using this approach, for example, existing Arabic-English, Arabic-Spanish, Spanish-Chinese, and English-Chinese systems could together be used to effect an Arabic-Chinese translation system. Wu and Wang (2007) report positive results using a similar technique with a single pivot language in conjunction with a small bilingual training corpus. Utiyama and Isahara (2007) show that in addition to sentence-based pivot methods, phrase translation tables can be built directly from phrase tables that share a pivot language.

Cohn and Lapata (2007) present another pivot approach centered on phrase tables, which they call triangulation. This technique maintains separate phrase tables for each language pair; during decoding, source phrases are translated into multiple intermediate language phrases, which are finally translated into target language phrases.

In contrast to pivot-based techniques, consensus network decoding (Mangu et al., 2000) attempts to improve translation quality by finding a novel, higher quality hypothesis based on the hypotheses produced by multiple translation systems. Much recent research (Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007) has explored consensus decoding where all systems translate the same language pair. Matusov et al. (2006) adapts this approach to a multilingual setting, performing consensus decoding when translating Japanese and Chinese into English; gains of 4.8 BLEU higher than the single best system are reported. Callison-Burch et al. (2008) report preliminary results that indicate promising results when applying system combination techniques on the multi-source News Commentary corpus.

Alternatively, hypothesis ranking techniques attempt to select the single best hypothesis from a list of output hypotheses produced by different translation systems. Several techniques designed for bilingual sentence-level system combination could be applied with no changes to the multi-source task. Kaki et al. (1999) and Callison-Burch and Flourney (2001) use only the target language model to rank the hypotheses. This approach follows the intuition that the hypothesis with the highest language model score will be the most fluent. Nomoto (2004) take this one step further by using multiple language models which vote on candidate hypotheses. Och and Ney (2001) present two techniques, called MAX and PROD, designed specifically for multi-source translation.

3 Oracle Experiments

The two techniques that have been used successfully for multi-source translation are sentence-level hypothesis ranking (Och and Ney, 2001) and consensus decoding (Matusov et al., 2006). In this work we are interested in determining whether the techniques presented in Och and Ney (2001) can be replicated using current multi-parallel corpora with long sentences and modern phrase-based decoders, and measuring the translation quality of these techniques against current metrics.

In order to provide a context for the possible gains from the hypothesis ranking methods of Och and Ney (2001), it is worth examining the maximum possible gains in translation quality which these methods can achieve.

languages	BLEU	TER	METEOR
da-en	28.4	57.5	52.9
de-en	27.3	58.9	52.4
el-en	29.3	56.4	53.6
es-en	**32.5**	**52.8**	**56.3**
fi-en	24.6	62.1	50.4
fr-en	31.9	53.1	55.8
it-en	29.2	57.1	53.7
nl-en	25.7	62.7	50.4
pt-en	31.8	53.7	56.0
sv-en	**32.7**	**52.3**	**56.6**

Table 1: Results of ten bilingual phrase based decoders into English. All systems were trained on Europarl v3. Test set is Europarl test05. Best results are bold.

To begin, ten bilingual translation systems were trained on the Europarl corpus. The standard phrase-based Moses decoder (Koehn et al., 2007) was used for all ten systems. The system parameters were tuned using minimum error rate training (Och, 2003) to optimize BLEU (Papineni et al., 2001) on the dev2006 development set. The target language for all systems was English. Table 1 shows results for these ten systems on the in-domain Europarl test05 data. Scores are listed for the TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005) metrics in addition to BLEU. We observe that the Swedish and Spanish systems perform the best according to all three metrics. The Dutch and Finnish systems perform the worst. In all experiments, each test sentence had only one reference translation.

Two oracle experiments were conducted to estimate the maximum possible gains in translation quality achievable by hypothesis ranking techniques. All hypothesis ranking methods by definition simply choose one target sentence from a list of two or more possible hypotheses. The best such method is one that always chooses the target sentence which represents the best translation from the available options. In each experiment, an oracle selected the best target sentence from the available hypotheses by selecting the one with the lowest word error rate (WER) when compared with the reference.

languages	BLEU	TER	METEOR
oracle-all	40.8	40.5	62.5

Table 2: Scores after combining results of ten bilingual phrase based decoders into English, using a WER-based oracle to choose which system output to select.

The first oracle experiment examined the possible gains when all ten bilingual systems are combined using sentence-level hypothesis ranking. For each test sentence, the oracle selected the hypothesis from the list of system output hypotheses with the lowest WER against the reference. Table 2 lists the results. The oracle BLEU score achieved here is 8.3 BLEU higher than the best individual system. This indicates that the combined translation systems together provide substantial additional information to positively influence translation quality.

system	% selected
da-en	14.1
de-en	9.6
el-en	10.3
es-en	14.0
fi-en	4.0
fr-en	12.9
it-en	7.2
nl-en	5.5
pt-en	9.8
sv-en	12.9

Table 3: Percentage of time that sentences from each system were selected in an All-English oracle WER experiment. Score for overall oracle output was 43.8 WER and 40.8 BLEU.

	da	de	el	es	fi	fr	it	nl	pt	sv
da	—	3.2	**3.7**	2.4	1.9	2.6	**4.0**	2.4	2.4	1.7
de		—	2.7	2.0	1.9	2.0	3.3	2.7	2.1	1.6
el			—	2.1	1.8	2.3	**3.7**	2.6	2.5	2.5
es				—	1.2	3.1	2.4	1.7	3.1	**3.7**
fi					—	1.0	1.9	2.7	1.1	0.6
fr						—	2.4	1.6	3.5	**3.7**
it							—	2.4	2.5	2.7
nl								—	1.8	1.3
pt									—	3.5
sv										—

Table 4: Absolute change in BLEU after combining two languages using oracle compared with the best BLEU of either language individually. The largest increases come from combining da & el, el & it, es & sv, fr & sv (each +3.7) and da & it (+4.0). The smallest increases come from combining fr & fi (+1.0) and fi & sv (+0.6). Best results in bold.

This oracle experiment also tracked for each system the number of times its hypothesis was selected as the best overall hypothesis. Table 3 lists these percentages. This distribution is flatter than we anticipated. It is not surprising that the systems which performed the best individually (sv-en, fr-en, and es-en) were chosen a large percentage of the time. However, the da-en system, which ranked seventh individually, was chosen by the oracle more often than any other. Even fi-en and nl-en, systems which performed substantially worse than the others individually, were selected a reasonable number of times. This data suggests that additional research is warranted to investigate the types of sentences for which bilingual systems with different source languages systems perform well.

The second experiment calculated oracle hypotheses for each pair of systems. In this experiment each of the 45 pairs of systems were combined by the WER oracle to simulate ideal sentence-level hypothesis ranking. Table 4 lists the absolute increase in BLEU score achieved by the oracle on the test set compared with the best BLEU score achieved by either system individually.

The difference between the best BLEU score for each pair and the oracle score was substantial for nearly all pairs of systems. The lowest absolute increase in BLEU scores (0.6) is seen when combining the worst individual system, Finnish, with the best individual system, Swedish. The mean and median increase in BLEU score for the 46 system pairs is 2.4 BLEU. The average difference between oracle BLEU and the score achieved by the MAX method for the same system pair is 3.3 BLEU (median 3.4 BLEU). This data shows that substantial gains are achievable from sentence-level hypothesis ranking methods, even when only two systems are combined.

4 Multi-Source Translation as a Hypothesis Ranking Problem

Given that significant gains in translation quality are possible through sentence-level hypothesis ranking, this section considers the MAX and PROD ranking methods proposed by Och and Ney (2001).

The original work is limited in the scope of its experiments. At the time, no large multi-parallel corpus was available. The authors assembled a training corpus from the *Bulletin of the European Union* with 117k-139k sentences per language for eleven European languages. Their test set was restricted to sentences 10 to 14 words in length. The metrics in common use today, including BLEU, METEOR, and TER, had not been developed. Results were reported in terms of word error rate (WER) and position-independent word error rate (PER). Their decoder used the alignment template system (Och et al., 1999).

In the following sections, we attempt to reproduce the techniques and results presented in Och and Ney (2001). We do this to answer three important questions:

- Can the techniques for multi-source translation presented in Och and Ney (2001) be replicated using current phrase-based decoders?

- Can the positive results they report be replicated on a larger data set which includes longer sentences?

- And finally, do the results presented for WER correlate with current automatic evaluation metrics?

4.1 Ranking Method MAX

Och and Ney (2001) propose that the best output translation from distinct bilingual translation systems can be chosen by taking the hypothesis with the highest score according to a noisy channel model. Given n source languages, the best translation $\hat{\mathbf{e}}$ is defined using both the language model and a translation model as

$$\begin{aligned} \hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \{ p(\mathbf{e}) \cdot \max_{n} p(\mathbf{f}_n|\mathbf{e}) \} \quad (1) \\ &= \arg\max_{\mathbf{e},n} \{ p(\mathbf{e}) \cdot p(\mathbf{f}_n|\mathbf{e}) \} \quad (2) \end{aligned}$$

This method is straightforward, and has the advantage that no modifications to the bilingual decoders are needed. The decoders must simply be capable of reporting language model and translation model probabilities along with each hypothesis. We note that the translation model probability reported by the decoder is in fact an approximation of $p(\mathbf{f}|\mathbf{e})$ as $p(\mathbf{f}|\mathbf{e}, \mathbf{d})$ where $\mathbf{d}$ is the derivation selected by the decoder.

Because the translation model probabilities from various systems are not necessarily comparable, it might be valuable to train weights for each system. Och and Ney (2001) report that such weights did not diverge much from one in their experiments. Due to time constraints, we did not perform system weighting.

4.2 Experiments using MAX

To examine how well MAX performs, we can take the output hypotheses produced by the bilingual translation systems, and apply the method using the translation model probabilities and language model probabilities reported by the decoder. We begin by selecting the translation system which produced the highest BLEU score, then added the system which gave the highest incremental gain. Table 5 reports results for the MAX method when applied with an increasing number of source languages. Och and Ney (2001) report the highest gains from MAX by combining three languages, French, then Swedish, then Spanish. We see the same three languages in our results, but with Swedish placed before French.

Because Och and Ney (2001) report only WER and PER, we report those metrics in addition to current metrics so that our results can be more directly compared with theirs. We see that our best results using MAX (58.8 WER for sv+es+fr) are significantly worse than their best result (52.0 WER for fr+sv+es). Most of this discrepancy is likely due to our use of a different corpus with longer sentences.

We also performed MAX ranking on all 45 system pairs, using every pair of foreign languages to translate into English. Table 6 shows the absolute change in BLEU score for the MAX ranking compared with the best BLEU score for either input system. For comparison with the results in Och and Ney (2001), table 7 presents this data in terms of WER. Our experiment reports results for all available languages, including German and Finnish; results for those two languages were not included in Och and Ney (2001).

Och and Ney (2001) show positive results using the MAX method for all 21 language pairs on which they report results. They report positive results showing absolute decreases in WER ranging from -0.5 (fr & it) to -4.3 (da & nl). Even if Finnish and German are excluded, we observe a wide range of mostly negative results (+5.0 for fr & nl, -2.0 for da & it). In total, 56% of combinations (25 out of 44) resulted in higher word error rate. Fully 80% of combinations (35 out of 44) resulted in an decrease in BLEU.

The results above show that the simple MAX approach simply does not improve translation quality for the majority of language pairs. In addition, Akiba et al. (2002) report (in a bilingual setting) that the hypothesis chosen by MAX ranking often differs from the hypothesis chosen by a human performing manual ranking.

For many MAX combinations, an improvement in WER was matched by an improvement in BLEU, and an increase in WER was matched with a lower

languages	BLEU	WER	PER	TER	METEOR
sv	32.7	60.2	53.6	52.3	56.6
sv+es	**33.1**	59.2	52.6	51.2	**56.9**
sv+es+fr	33.0	**58.8**	**52.1**	**50.9**	56.8
sv+es+fr+el	32.6	58.9	52.3	51.0	56.3

Table 5: Combination using Max ranking method.

	da	de	el	es	fi	fr	it	nl	pt	sv
da	—	**0.4**	**0.1**	-0.8	-1.3	-1.3	**0.3**	-1.4	-0.7	-1.6
de		—	-0.2	-0.6	-0.8	-2.0	-0.1	-0.8	-0.8	-1.1
el			—	-0.2	-1.8	-1.0	**0.6**	-1.9	-0.3	-0.5
es				—	-1.5	**0.5**	-0.9	-2.6	**0.1**	**0.3**
fi					—	-2.9	-1.3	-0.3	-1.9	-2.3
fr						—	-1.6	-3.7	**0.2**	**0.2**
it							—	-1.5	-1.0	-1.0
nl								—	-2.4	-2.9
pt									—	-0.1
sv										—

Table 6: Absolute change in BLEU after combining two languages using Max ranking method compared with the best BLEU of either language individually. Best results come from combining es & sv (+0.4), es & fr (+0.5), and el & it (+0.6). Worst results come from combining fi & fr (-2.9), nl & sv (-2.9), and fr & nl (-3.7). Only 20% of Max pairwise combinations led to an improvement in BLEU. Results which indicate an improvement in BLEU are bold.

	da	de	el	es	fi	fr	it	nl	pt	sv
da	—	**-0.8**	**-1.1**	**-0.1**	2.1	0.6	**-2.3**	1.6	**-0.7**	1.1
de		—	0.6	0.0	0.7	3.1	**-1.0**	0.1	0.2	1.2
el			—	**-1.0**	3.4	0.2	**-1.5**	2.9	**-0.9**	**-0.2**
es				—	2.0	**-2.0**	**-0.2**	2.6	**-1.6**	**-1.0**
fi					—	5.1	1.2	**-2.3**	2.4	3.5
fr						—	1.5	5.0	**-1.0**	**-0.9**
it							—	0.9	**-0.2**	0.3
nl								—	2.4	3.6
pt									—	**-0.6**
sv										—

Table 7: Absolute change in WER after combining two languages using Max ranking method compared with the best WER of either language individually. Best results come from combining da & it (-2.0) or from fi & nl (-2.0). Worst results come from combining fr with nl (+5.0) or with fi (+5.1). Only 44% of Max pairwise combinations led to an improvement in WER. Results which indicate an improvement in WER are bold.

	da	de	el	es	fi	fr	it	nl	pt	sv
da	—	**0.9**	-0.4	-2.6	-1.8	-2.8	**0.3**	-1.5	-2.0	-1.4
de		—	-1.4	-3.1	-1.3	-3.5	**0.3**	-1.1	-1.9	-0.5
el			—	0.0	-2.9	-1.0	**0.2**	-3.1	-0.1	-0.2
es				—	-4.4	**0.3**	-2.8	-5.7	-0.2	**0.4**
fi					—	-4.3	-0.8	-0.1	-4.1	-4.0
fr						—	-2.3	-5.4	**0.2**	**0.1**
it							—	-0.8	-2.2	-2.1
nl								—	-4.7	-3.9
pt									—	**0.4**
sv										—

Table 8: Absolute change in BLEU after combining two languages using MAXLL ranking method compared with the best BLEU of either language individually.

BLEU score. However, 26% of language pairs showed an improvement in WER but a decline in BLEU.

4.3 Extending MAX to a Log-linear Framework

Given that most current statistical translation systems are based on a log-linear combination of features rather than a noisy channel model, the question immediately arrives whether MAX might work if the log-linear score of each sentence is used in the argmax calculation (equation 2) instead of the noisy channel product. We define MAXLL as follows:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}}\{\max_{n} exp(\sum_{i} \lambda_i h_i(e, f_n))\} \tag{3}$$

$$= \arg\max_{\mathbf{e},n}\{exp(\sum_{i} \lambda_i h_i(e, f_n))\} \tag{4}$$

To test this method, we combine all 45 system pairs (as in section 4.2) but use MAXLL in place of MAX. This method performs quite badly. The number of systems for which BLEU increases is the same as for MAX (9 out of 45 pairs); however, for pairs in which MAXLL does poorly, it performs worse than MAX. The worst performance comes from the es & nl pair, where MAXLL scores -5.7 BLEU worse than the best of either system individually. This poor performance should not be surprising, as the scores returned by each system are not comparable.

4.4 Ranking Method PROD

The MAX method provides a simple method to choose the best output from among two or three bilingual translation systems. However, it fails to effectively make use of larger numbers of source languages; in fact, translation quality degrades when additional source languages are incorporated. The PROD method presented in Och and Ney (2001) addresses this shortcoming in MAX. Kay (1997) observed that if multiple translation engines independently produce the same hypothesis, that is strong evidence that the hypothesis is a good one. The PROD method follows this insight.

The PROD method attempts to approximate a true multi-source decoding algorithm by incorporating probabilities associated with each bilingual translation model. In this approach, given n source languages, the best translation $\hat{\mathbf{e}}$ is defined as

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}}\{p(\mathbf{e}) \cdot \prod_{n=1}^{N} p(\mathbf{f}_n|\mathbf{e})\} \tag{5}$$

Paul et al. (2005) explore a related technique in the bilingual setting, where a hypotheses is selected if its average translation model times language model score is significantly higher than competing hypotheses.

4.5 Constraint Decoding

The PROD method requires that for each target hypothesis e, a translation probability $p(\mathbf{f}_n|\mathbf{e})$ must be calculated for each source language sentence f_n.

Each target hypothesis e is produced by one of the bilingual decoders described earlier in section 4. Each sentence f_n is given as a source sentence.

The standard phrase-based decoder produces a 1-best or n-best list of hypotheses when given a source sentence. There is no guarantee that a particular target sentence e will appear in a decoder's n-best list of hypotheses. So, in order to calculate $p(\mathbf{f}_n|\mathbf{e})$ for every source language n, we modified the Moses decoder to permit *constraint decoding*. When constraint decoding is used, the phrase-based search is constrained so that only hypotheses which are consistent with the desired target output are considered.

The standard phrase-based decoding model (Koehn, 2004) creates "stacks" of translation options that cover contiguous phrases in the input sentence. Each translation option stores a target language phrase. The decoder attempts to trace a path through the translation options, creating partial hypotheses as it proceeds, in such a way that all source words are covered by a translation option, resulting in a complete hypothesis.

Each partial hypothesis represents a partial translation into the target language. Constraint decoding is defined by restricting the creation of partial hypotheses. Whenever a partial hypothesis would be constructed, the partial translation for that partial hypothesis is examined. If the partial translation is compatible with the desired target sentence, the partial hypothesis is constructed. If the partial translation is not compatible, meaning it is not a prefix to the desired target sentence, the partial hypothesis is pruned.

In this way, a desired target sentence can be provided to the decoder as a constraint; the desired target sentence will be produced as the result as long as the decoder's model and parameters are capable of reaching the desired target sentence.

4.6 Experiments using PROD

The decoder provides feature values, including $p(\mathbf{f}|\mathbf{e})$, for each sentence that it successfully translates. By using constraint decoding, and providing the decoder with both source sentence and desired target sentence, the translation model probabilities required for PROD can be obtained.

Unfortunately, the vast majority of target sentences presented to the constraint decoder resulted in failure. In these cases, the decoder was not able to reach the desired target sentence from the provided source sentence given the translation model, the language model, and the feature parameters. Table 9 presents a sample of results that illustrate this problem.

	da-en	de-en	es-en	fr-en
% reachable	10.5	9.8	11.5	10.6

Table 9: Percentage of sentences reachable by the Swedish-English system when constrained by the output of the listed systems.

Attempts to increase the number of reachable constraint sentences by turning off all pruning during constraint decoding did not lead to a substantial increase in the number of reachable sentences. Even worse, the particular sentences reachable in the test set was not the same across the various translation systems. As as result, the decoder was unable to provide $p(\mathbf{f}_n|\mathbf{e})$ for all required source languages n in the vast majority of test sentences.

It is reasonable to ask whether the PROD method could be applied for those minority of sentences which are reachable. What conditions are necessary in order to apply the PROD method when just two source languages are used? In other words, are we able to use equation 5 to determine ê? Equation 5 requires that we know $p(\mathbf{f}_n|\mathbf{e})$. Consider the concrete example where Spanish-English and French-English outputs are to be ranked using the PROD method. The Spanish system translates the Spanish input into English hypothesis e_{es} and provides $p(\mathbf{f}_{es}|\mathbf{e}_{es})$. Likewise the French system translates the French input into English hypothesis e_{fr} and provides $p(\mathbf{f}_{fr}|\mathbf{e}_{fr})$. If the Spanish system is able to successfully translate the Spanish source sentence f_{es} into e_{fr} using constraint decoding, $p(\mathbf{f}_{es}|\mathbf{e}_{fr})$ becomes available. Likewise $p(\mathbf{f}_{fr}|\mathbf{e}_{es})$ becomes available if the French system is able to translate f_{fr} into e_{es}. Only a very small number of test sentences fulfill these conditions for each pair of systems. For three or more systems, the problem is even worse.

4.7 Discussion on PROD

Och and Ney (2001) do not discuss the problem of unreachable sentences when calculating $p(\mathbf{f}_n|\mathbf{e})$ for PROD. We now briefly examine why our experiments found so few sentences to be reachable by a constrained phrase based decoder, while no such problem was reported by Och and Ney (2001).

The first possible factor that presents itself is the data used in the experiments. The test corpus used by Och and Ney (2001) was extracted from the *Bulletin of the European Union* and was restricted so that all reference sentences in the test set were 10 to 14 words long. By contrast, the Europarl test05 test set includes reference sentences up to 135 words long. In the experiments in section 4.6, the sentences reachable during constraint decoding have an average length of 14.2 words. We note that this is longer than the *maximum* sentence length used by Och and Ney (2001). The average reference sentence length in our complete test set is 29.0 words.

If we restrict our test set to only those sentences where the reference has 10-14 words, the percentage of reachable sentences increases from approximately 10% of the entire test corpus to approximately 25% of the subset of 10-14 word sentences for a given pair of systems. In other words, even when only short sentences are considered, a large majority cannot be reached during constraint decoding.

The second possible factor is the choice of translation search algorithm. The experiments presented in this paper used the standard phrase based Moses decoder, modified to allow constraint decoding[1].

Phrase based translation allows a single word to be translated as a phrase only if the word and phrase were aligned during training. Consider the case where the decoder needs to translate source word a, and adjacent target words xyz still need to be generated; no phrase $a \rightarrow xyz$ exists in the phrase table, but entries for $a \rightarrow x$, $a \rightarrow y$, and $a \rightarrow z$ all do exist. A word based decoder might be able to deal with this by assigning a a fertility of three, then translating the three words individually, but a phrase-based system cannot.

Och and Ney (2001) use an alignment template decoder which implements an early phrase-based translation model. We hypothesize that the use of word classes in the alignment template approach may have allowed more hypotheses to be reachable during constraint decoding, which could allow $p(\mathbf{f}_n|\mathbf{e})$ to be obtained for PROD calculations.

Given the positive results initially reported for PROD, we believe that there is still value in replicating this technique, at minimum for use as a baseline as more advanced techniques are developed.

5 Conclusion & Future Work

We have shown that significant gains in translation quality are possible using hypothesis ranking, but that the MAX technique is not a reliable method for hypothesis ranking. We have also shown that limitations in current decoding techniques prevent the use of PROD with phrase-based systems. Our findings show a limit to the claim by Och and Ney (2001) that this method for combining multiple source languages is independent of translation models. In particular, PROD is useful only insofar as $p(\mathbf{f}_n|\mathbf{e})$ can be reasonably approximated for an arbitrary source language sentence f_n when constrained by an arbitrary target language sentence e.

As part of our research into multi-source translation, we are looking at methods for finding the closest reachable hypothesis to a specified target sentence; this should make a more thorough examination of PROD possible. We are currently undertaking a more thorough examination of consensus decoding in multi-source translation than has been previously published. We also believe that significant gains can be found by integrating multiple source into the actual decoding process. One promising area where we are currently working is the use of lattice inputs (Dyer et al., 2008) where multiple source language inputs are encoded in the input lattice.

Acknowledgements

Thanks to Josh Schroeder, Chris Callison-Burch, Adam Lopez, Chris Dyer, Phil Blumson, and Martin Kay for invaluable advice and guidance, and Hieu Huang for assistance with the Moses code. This work was partially supported by the DARPA GALE program (Contract No HR0011-06-2-0001).

References

Yasuhiro Akiba, Taro Watanabe, and Eiichiro Sumita. 2002. Using language and translation models to select the best among outputs from multiple MT systems. In *Proc. COLING*.

[1] Moses decoder, subversion revision 1857

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL*.

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proc. ASRU*.

Chris Callison-Burch and Raymond Flourney. 2001. A program for automatically selecting the best output from multiple machine translation engines. In *Proc. MT Summit VIII*, pages 63–66.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proc. WMT*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc. WMT*, pages 70–106.

Chris Callison-Burch. 2002. Co-training for statistical machine translation. Master's thesis, U. Edinburgh.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proc. ACL*, pages 728–735.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proc. ACL*, pages 1012–1020.

Andreas Eisele. 2006. Parallel corpora and phrase-based statistical machine translation for new language pairs via multiple intermediaries. In *Proc. LREC*.

Tomaž Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. LREC*, pages 1535–1538.

Robert Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *Proc. ANLP*.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. EAMT*.

Satoshi Kaki, Setsuo Yamada, and Eiichiro Sumita. 1999. Scoring multiple translations using character n-gram. In *Proc. NLPRS*, pages 298–302.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12:3–23. First appeared as a Xerox PARC working paper in 1980.

Martin Kay. 2000. Triangulation in translation. Keynote at the MT 2000 Conference, University of Exeter.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL Demonstration Session*.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proc. AMTA*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*.

Shankar Kumar, Franz Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proc. EMNLP-CoNLL*, pages 42–50.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400, Oct.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypothesis alignment. In *Proc. EACL*, pages 33–40.

Tadashi Nomoto. 2004. Multi-engine machine translation with voted language model. In *Proc. ACL*.

Franz Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proc. MT Summit VIII*.

Franz Och, Christoph Tillman, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. SIGDAT-EMNLP*.

Franz Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Michael Paul, Takao Doi, Youngsook Hwang, Kenji Imamura, Hideo Okuma, and Eiichiro Sumita. 2005. Nobody is perfect: ATR's hybrid approach to spoken language translation. In *Proc. IWSLT*, pages 55–62.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1–2):129–153.

Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007. Combining outputs from multiple machine translation systems. In *Proc. NAACL-HLT*, pages 228–235.

Michel Simard. 1999. Text-translation alignment: Three languages are better than two. In *Proc. EMNLP*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. AMTA*, pages 223–231.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. LREC*, pages 2142–2147.

UN. 1994. United Nations parallel text. LDC Catalog No.: LDC94T4A.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proc. NAACL/HLT*.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proc. ACL*, pages 856–863.