Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora

Trevor Cohn and **Mirella Lapata**
Human Computer Research Centre, School of Informatics
University of Edinburgh
{tcohn,mlap}@inf.ed.ac.uk

Abstract

Current phrase-based SMT systems perform poorly when using small training sets. This is a consequence of unreliable translation estimates and low coverage over source and target phrases. This paper presents a method which alleviates this problem by exploiting multiple translations of the same source phrase. Central to our approach is *triangulation*, the process of translating from a source to a target language via an intermediate third language. This allows the use of a much wider range of parallel corpora for training, and can be combined with a standard phrase-table using conventional smoothing methods. Experimental results demonstrate BLEU improvements for triangulated models over a standard phrase-based system.

1 Introduction

Statistical machine translation (Brown et al., 1993) has seen many improvements in recent years, most notably the transition from word- to phrase-based models (Koehn et al., 2003). Modern SMT systems are capable of producing high quality translations when provided with large quantities of training data. With only a small training sample, the translation output is often inferior to the output from using larger corpora because the translation algorithm must rely on more sparse estimates of phrase frequencies and must also 'back-off' to smaller sized phrases. This often leads to poor choices of target phrases and reduces the coherence of the output. Unfortunately, parallel corpora are not readily available in large quantities, except for a small subset of the world's languages (see Resnik and Smith (2003) for discussion), therefore limiting the potential use of current SMT systems.

In this paper we provide a means for obtaining more reliable translation frequency estimates from small datasets. We make use of *multi-parallel* corpora (sentence aligned parallel texts over three or more languages). Such corpora are often created by international organisations, the United Nations (UN) being a prime example. They present a challenge for current SMT systems due to their relatively moderate size and domain variability (examples of UN texts include policy documents, proceedings of meetings, letters, *etc.*). Our method translates each target phrase, $\mathbf{t}$, first to an intermediate language, $\mathbf{i}$, and then into the source language, $\mathbf{s}$. We call this two-stage translation process *triangulation* (Kay, 1997). We present a probabilistic formulation through which we can estimate the desired phrase translation distribution (phrase-table) by marginalisation, $p(\mathbf{s}|\mathbf{t}) = \sum_{\mathbf{i}} p(\mathbf{s}, \mathbf{i}|\mathbf{t})$.

As with conventional smoothing methods (Koehn et al., 2003; Foster et al., 2006), triangulation increases the robustness of phrase translation estimates. In contrast to smoothing, our method alleviates data sparseness by exploring additional multiparallel data rather than adjusting the probabilities of existing data. Importantly, triangulation provides us with separately estimated phrase-tables which could be further smoothed to provide more reliable distributions. Moreover, the triangulated phrase-tables can be easily combined with the standard source-target phrase-table, thereby improving the coverage over unseen source phrases.

As an example, consider Figure 1 which shows the coverage of unigrams and larger n-gram phrases when using a standard source target phrase-table, a triangulated phrase-table with one (*it*) and nine languages (*all*), and a combination of standard and triangulated phrase-tables (*all+standard*). The phrases were harvested from a small French-English bitext

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 728–735,
Prague, Czech Republic, June 2007.

and evaluated against a test set. Although very few small phrases are unknown, the majority of larger phrases are unseen. The *Italian* and *all* results show that triangulation alone can provide similar or improved coverage compared to the standard source-target model; further improvement is achieved by combining the triangulated and standard models (*all+standard*). These models and datasets will be described in detail in Section 3.

We also demonstrate that triangulation can be used on its own, that is *without a source-target distribution*, and still yield acceptable translation output. This is particularly heartening, as it provides a means of translating between the many "low density" language pairs for which we don't yet have a source-target bitext. This allows SMT to be applied to a much larger set of language pairs than was previously possible.

In the following section we provide an overview of related work. Section 3 introduces a generative formulation of triangulation. We present our evaluation framework in Section 4 and results in Section 5.

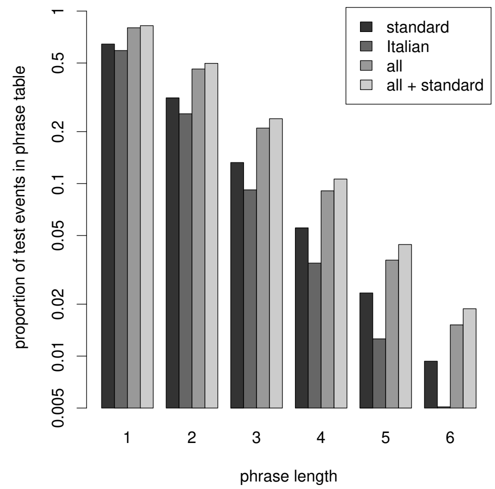

Figure 1: Coverage of *fr* → *en* test phrases using a 10,000 sentence bitext. The standard model is shown alongside triangulated models using one (Italian) or nine other languages (all).

2 Related Work

The idea of using multiple source languages for improving the translation quality of the target language dates back at least to Kay (1997), who observed that ambiguities in translating from one language onto another may be resolved if a translation into some third language is available. Systems which have used this notion of triangulation typically create several candidate sentential target translations for source sentences via different languages. A single translation is then selected by finding the candidate that yields the best overall score (Och and Ney, 2001; Utiyama and Isahara, 2007) or by co-training (Callison-Burch and Osborne, 2003). This ties in with recent work on ensemble combinations of SMT systems, which have used alignment techniques (Matusov et al., 2006) or simple heuristics (Eisele, 2005) to guide target sentence selection and generation. Beyond SMT, the use of an intermediate language as a translation aid has also found application in cross-lingual information retrieval (Gollins and Sanderson, 2001).

Callison-Burch et al. (2006) propose the use of paraphrases as a means of dealing with unseen source phrases. Their method acquires paraphrases by identifying candidate phrases in the source language, translating them into multiple target languages, and then back to the source. Unknown source phrases are substituted by the back-translated paraphrases and translation proceeds on the paraphrases.

In line with previous work, we exploit multiple source corpora to alleviate data sparseness and increase translation coverage. However, we differ in several important respects. Our method operates over phrases rather than sentences. We propose a generative formulation which treats triangulation not as a post-processing step but as part of the translation model itself. The induced phrase-table entries are fed directly into the decoder, thus avoiding the additional inefficiencies of merging the output of several translation systems.

Although related to Callison-Burch et al. (2006) our method is conceptually simpler and more general. Phrase-table entries are created via multiple source languages without the intermediate step of paraphrase extraction, thereby reducing the exposure to compounding errors. Our phrase-tables may well contain paraphrases but these are naturally induced as part of our model, without extra processing effort. Furthermore, we improve the translation estimates for both seen and unseen phrase-table entries, whereas Callison-Burch et al. concentrate solely on unknown phrases. In contrast to Utiyama and Isahara (2007), we employ a large number of intermediate languages and demonstrate how triangulated phrase-tables can be combined with standard phrase-tables to improve translation output.

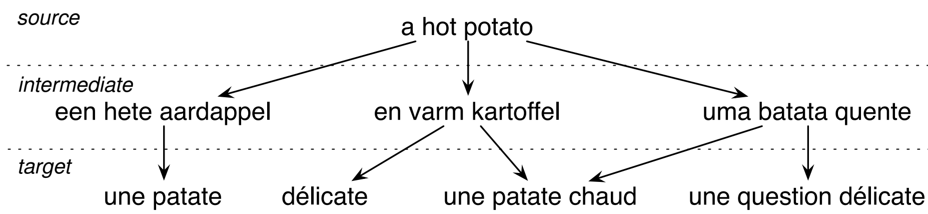

Figure 2: Triangulation between English (source) and French (target), showing three phrases in Dutch, Danish and Portuguese, respectively. Arrows denote phrases aligned in a language pair and also the generative translation process.

3 Triangulation

We start with a motivating example before formalising the mechanics of triangulation. Consider translating the English phrase *a hot potato*[1] into French, as shown in Figure 2. In our corpus this English phrase occurs only three times. Due to errors in the word alignment the phrase was not included in the English-French phrase-table. Triangulation first translates *hot potato* into a set of intermediate languages (Dutch, Danish and Portuguese are shown in the figure), and then these phrases are further translated into the target language (French). In the example, four different target phrases are obtained, all of which are useful phrase-table entries. We argue that the redundancy introduced by a large suite of other languages can correct for errors in the word alignments and also provide greater generalisation, since the translation distribution is estimated from a richer set of data-points. For example, instances of the Danish *en varm kartoffel* may be used to translate several English phrases, not only *a hot potato*.

In general we expect that a wider range of possible translations are found for any source phrase, simply due to the extra layer of indirection. So, if a source phrase tends to align with two different target phrases, then we would also expect it to align with two phrases in the 'intermediate' language. These intermediate phrases should then each align with two target phrases, yielding up to four target phrases. Consequently, triangulation will often produce more varied translation distributions than the standard source-target approach.

3.1 Formalisation

We now formalise triangulation as a generative probabilistic process operating independently on phrase pairs. We start with the conditional distribution over three languages, $p(\mathbf{s}, \mathbf{i}|\mathbf{t})$, where the arguments denote phrases in the source, intermediate and target language, respectively. From this distribution, we can find the desired conditional over the source-target pair by marginalising out the intermediate phrases:[2]

$$\begin{aligned} p(\mathbf{s}|\mathbf{t}) &= \sum_{\mathbf{i}} p(\mathbf{s}|\mathbf{i},\mathbf{t}) p(\mathbf{i}|\mathbf{t}) \\ &\approx \sum_{\mathbf{i}} p(\mathbf{s}|\mathbf{i}) p(\mathbf{i}|\mathbf{t}) \end{aligned} \tag{1}$$

where (1) imposes a simplifying conditional independence assumption: the intermediate phrase fully represents the information (semantics, syntax, *etc.*) in the source phrase, rendering the target phrase redundant in $p(\mathbf{s}|\mathbf{i}, \mathbf{t})$.

Equation (1) requires that all phrases in the intermediate-target bitext must also be found in the source-intermediate bitext, such that $p(\mathbf{s}|\mathbf{i})$ is defined. Clearly this will often not be the case. In these situations we could back-off to another distribution (by discarding part, or all, of the conditioning context), however we take a more pragmatic approach and ignore the missing phrases. This problem of missing contexts is uncommon in multi-parallel corpora, but is more common when the two bitexts are drawn from different sources.

While triangulation is intuitively appealing, it may suffer from a few problems. Firstly, as with any SMT approach, the translation estimates are based on noisy automatic word alignments. This leads to many errors and omissions in the phrase-table. With a standard source-target phrase-table these errors are only encountered once, however with triangulation they are encountered twice, and therefore the errors will compound. This leads to more noisy estimates than in the source-target phrase-table.

Secondly, the increased exposure to noise means that triangulation will omit a greater proportion of large or rare phrases than the standard method. An

[1] An idiom meaning a situation for which no one wants to claim responsibility.

[2] Equation (1) is used with the source and target arguments reversed to give $p(\mathbf{t}|\mathbf{s})$.

alignment error in either of the source-intermediate or intermediate-target bitexts can prevent the extraction of a source-target phrase pair. This effect can be seen in Figure 1, where the coverage of the Italian triangulated phrase-table is worse than the standard source-target model, despite the two models using the same sized bitexts. As we explain in the next section, these problems can be ameliorated by using the triangulated phrase-table in conjunction with a standard phrase-table.

Finally, another potential problem stems from the independence assumption in (1), which may be an oversimplification and lead to a loss of information. The experiments in Section 5 show that this effect is only mild.

3.2 Merging the phrase-tables

Once induced, the triangulated phrase-table can be usefully combined with the standard source-target phrase-table. The simplest approach is to use linear interpolation to combine the two (or more) distributions, as follows:

$$p(\mathbf{s}, \mathbf{t}) = \sum_j \lambda_j p_j(\mathbf{s}, \mathbf{t}) \tag{2}$$

where each joint distribution, p_j, has a non-negative weight, λ_j, and the weights sum to one. The joint distribution for triangulated phrase-tables is defined in an analogous way to Equation (1). We expect that the standard phrase-table should be allocated a higher weight than triangulated phrase-tables, as it will be less noisy. The joint distribution is now conditionalised to yield $p(\mathbf{s}|\mathbf{t})$ and $p(\mathbf{t}|\mathbf{s})$, which are both used as features in the decoder. Note that the resulting conditional distribution will be drawn solely from one input distribution when the conditioning context is unseen in the remaining distributions. This may lead to an over-reliance on unreliable distributions, which can be ameliorated by smoothing (e.g., Foster et al. (2006)).

As an alternative to linear interpolation, we also employ a weighted product for phrase-table combination:

$$p(\mathbf{s}|\mathbf{t}) \propto \prod_j p_j(\mathbf{s}|\mathbf{t})^{\lambda_j} \tag{3}$$

This has the same form used for log-linear training of SMT decoders (Och, 2003), which allows us to treat each distribution as a feature, and learn the mixing weights automatically. Note that we must individually smooth the component distributions in (3) to stop zeros from propagating. For this we use Simple Good-Turing smoothing (Gale and Sampson, 1995) for each distribution, which provides estimates for zero count events.

4 Experimental Design

Corpora We used the Europarl corpus (Koehn, 2005) for experimentation. This corpus consists of about 700,000 sentences of parliamentary proceedings from the European Union in eleven European languages. We present results on the full corpus for a range of language pairs. In addition, we have created smaller parallel corpora by sub-sampling 10,000 sentence bitexts for each language pair. These corpora are likely to have minimal overlap — about 1.5% of the sentences will be shared between each pair. However, the phrasal overlap is much greater (10 to 20%), which allows for triangulation using these common phrases. This training setting was chosen to simulate translating to or from a "low density" language, where only a few small independently sourced parallel corpora are available. These bitexts were used for direct translation and triangulation. All experimental results were evaluated on the ACL/WMT 2005[3] set of 2,000 sentences, and are reported in BLEU percentage-points.

Decoding Pharaoh (Koehn, 2003), a beam-search decoder, was used to maximise:

$$\mathbf{T}^* = \arg\max_{\mathbf{T}} \prod_j f_j(\mathbf{T}, \mathbf{S})^{\lambda_j} \tag{4}$$

where $\mathbf{T}$ and $\mathbf{S}$ denote a target and source sentence respectively. The parameters, λ_j, were trained using minimum error rate training (Och, 2003) to maximise the BLEU score (Papineni et al., 2002) on a 150 sentence development set. We used a standard set of features, comprising a 4-gram language model, distance based distortion model, forward and backward translation probabilities, forward and backward lexical translation scores and the phrase- and word-counts. The translation models and lexical scores were estimated on the training corpus which was automatically aligned using Giza++ (Och et al., 1999) in both directions between source and target and symmetrised using the growing heuristic (Koehn et al., 2003).

[3] For details see `http://www.statmt.org/wpt05/mt-shared-task`.

Lexical weights The lexical translation score is used for smoothing the phrase-table translation estimate. This represents the translation probability of a phrase when it is decomposed into a series of independent word-for-word translation steps (Koehn et al., 2003), and has proven a very effective feature (Zens and Ney, 2004; Foster et al., 2006). Pharaoh's lexical weights require access to word-alignments; calculating these alignments between the source and target words in a phrase would prove difficult for a triangulated model. Therefore we use a modified lexical score, corresponding to the maximum IBM model 1 score for the phrase pair:

$$lex(\mathbf{t}|\mathbf{s}) = \frac{1}{Z} \max_{\mathbf{a}} \prod_{k} p(t_k|s_{a_k}) \tag{5}$$

where the maximisation[4] ranges over all one-to-many alignments and Z normalises the score by the number of possible alignments.

The lexical probability is obtained by interpolating a relative frequency estimate on the source-target bitext with estimates from triangulation, in the same manner used for phrase translations in (1) and (2). The addition of the lexical probability feature yielded a substantial gain of up to two BLEU points over a basic feature set.

5 Experimental Results

The evaluation of our method was motivated by three questions: (1) How do different training requirements affect the performance of the triangulated models presented in this paper? We expect performance gains with triangulation on small and moderate datasets. (2) Is machine translation output influenced by the choice of the intermediate language/s? Here, we would like to evaluate whether the number and choice of intermediate languages matters. (3) What is the quality of the triangulated phrase-table? In particular, we are interested in the resulting distribution and whether it is sufficiently distinct from the standard phrase-table.

5.1 Training requirements

Before reporting our results, we briefly discuss the specific choice of model for our experiments. As mentioned in Section 3, our method combines the triangulated phrase-table with the standard source-target one. This is desired in order to compensate for the noise incurred by the triangulation process. We used two combination methods, namely linear interpolation (see (2)) and a weighted geometric mean (see (3)).

	standard	interp	+indic	separate
en → *de*	12.03	12.66	12.95	12.25
fr → *en*	23.02	24.63	23.86	23.43

Table 1: Different feature sets used with the 10K training corpora, using a single language (*es*) for triangulation. The columns refer to standard, uniform interpolation, interpolation with 0-1 indicator features, and separate phrase-tables, respectively.

Table 1 reports the results for two translation tasks when triangulating with a single language (*es*) using three different feature sets, each with different translation features. The *interpolation* model uses uniform linear interpolation to merge the standard and triangulated phrase-tables. Non-uniform mixtures did not provide consistent gains, although, as expected, biasing towards the standard phrase-table was more effective than against. The *indicator* model uses the same interpolated distribution along with a series of 0-1 indicator features to identify the source of each event, *i.e.*, if each $(\mathbf{s}, \mathbf{t})$ pair is present in phrase-table j. We also tried per-context features with similar results. The *separate* model has a separate feature for each phrase-table.

All three feature sets improve over the standard source-target system, while the interpolated features provided the best overall performance. The relatively poorer performance of the separate model is perhaps surprising, as it is able to differentially weight the component distributions; this is probably due to MERT not properly handling the larger feature sets. In all subsequent experiments we report results using linear interpolation.

As a proof of concept, we first assessed the effect of triangulation on corpora consisting of 10,000 sentence bitexts. We expect triangulation to deliver performance gains on small corpora, since a large number of phrase-table entries will be unseen. In Table 2 each entry shows the BLEU score when using the standard phrase-table and the absolute improvement when using triangulation. Here we have used three languages for triangulation (*it* $\cup$ $\{de, en, es, fr\} \backslash \{s, t\}$). The source-target languages were chosen so as to mirror the evaluation setup of NAACL/WMT. The translation tasks range

[4]The maximisation in (5) can be replaced with a sum with similar experimental results.

s ↓ t →	*de*	*en*	*es*	*fr*
de	-	17.58	16.84	18.06
	-	+1.20	+1.99	+1.94
en	12.45	-	23.83	24.05
	+1.22	-	+1.04	+1.48
es	12.31	23.83	-	32.69
	+2.24	+1.35	-	+0.85
fr	11.76	23.02	31.22	-
	+2.41	+2.24	+1.30	-

Table 2: BLEU improvements over the standard phrase-table (top) when interpolating with three triangulated phrase-tables (bottom) on the small training sample.

from easy (*es* → *fr*) to very hard (*de* → *en*). In all cases triangulation resulted in an improvement in translation quality, with the highest gains observed for the most difficult tasks (to and from German). For these tasks the standard systems have poor coverage (due in part to the sizeable vocabulary of German phrases) and therefore the gain can be largely explained by the additional coverage afforded by the triangulated phrase-tables.

To test whether triangulation can also improve performance of larger corpora we ran six separate translation tasks on the full Europarl corpus. The results are presented in Table 3, for a single triangulation language used alone (*triang*) or uniformly interpolated with the standard phrase-table (*interp*). These results show that triangulation can produce high quality translations on its own, which is noteworthy, as it allows for SMT between a much larger set of language pairs. Using triangulation in conjunction with the standard phrase-table improved over the standard system in most instances, and only degraded performance once. The improvement is largest for the German tasks which can be explained by triangulation providing better robustness to noisy alignments (which are often quite poor for German) and better estimates of low-count events. The difficulty of aligning German with the other languages is apparent from the Giza++ perplexity: the final Model 4 perplexities for German are quite high, as much as double the perplexity for more easily aligned language pairs (*e.g.*, Spanish-French).

Figure 3 shows the effect of triangulation on different sized corpora for the language pair *fr* → *en*. It presents learning curves for the standard system and a triangulated system using one language (*es*). As can be seen, gains from triangulation only diminish slightly for larger training corpora, and that

task	standard	interm	triang	interp
de → *en*	23.85	*es*	23.48	**24.36**
en → *de*	17.24	*es*	16.28	**17.42**
es → *en*	30.48	*fr*	29.06	**30.52**
en → *es*	**29.09**	*fr*	28.19	**29.09**
fr → *en*	29.66	*es*	29.59	**30.36**
en → *fr*	**30.07**	*es*	28.94	29.62

Table 3: Results on the full training set showing triangulation with a single language, both alone (triang) and alongside a standard model (interp).

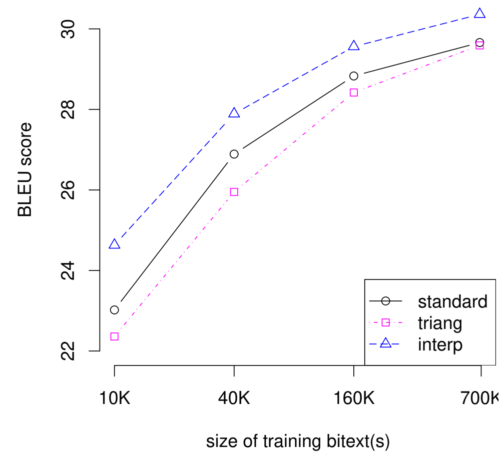

Figure 3: Learning curve for *fr* → *en* translation for the standard source-target model and a triangulated model using Spanish as an intermediate language.

the purely triangulated models have very competitive performance. The gain from interpolation with a triangulated model is roughly equivalent to having twice as much training data.

Finally, notice that triangulation may benefit when the sentences in each bitext are drawn from the same source, in that there are no unseen 'intermediate' phrases, and therefore (1) can be easily evaluated. We investigate this by examining the robustness of our method in the face of disjoint bitexts. The concepts contained in each bitext will be more varied, potentially leading to better coverage of the target language. In lieu of a study on different domain bitexts which we plan for the future, we bisected the Europarl corpus for *fr* → *en*, triangulating with Spanish. The triangulated models were presented with *fr-es* and *es-en* bitexts drawn from either the same half of the corpus or from different halves, resulting in scores of 28.37 and 28.13, respectively.[5] These results indicate that triangulation is effective

[5]The baseline source-target system on one half has a score of 28.85.

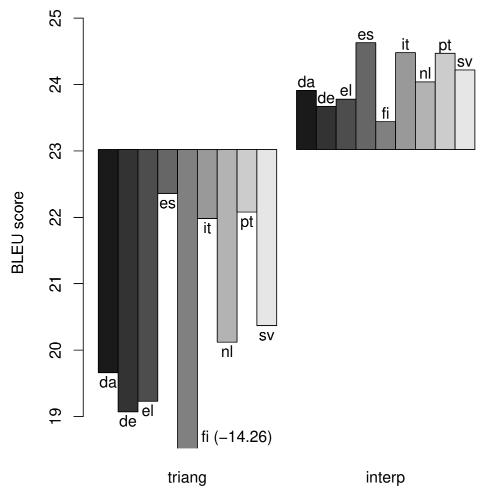

Figure 4: Comparison of different triangulation languages for *fr* → *en* translation, relative to the standard model (10K training sample). The bar for *fi* has been truncated to fit on the graph.

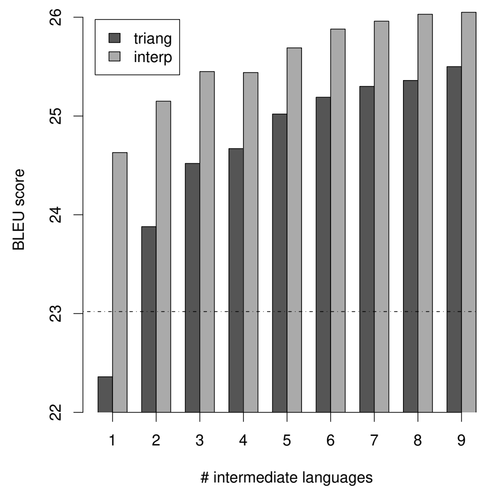

Figure 5: Increasing the number of intermediate languages used for triangulation increases performance for *fr* → *en* (10K training sample). The dashed line shows the BLEU score for the standard phrase-table.

for disjoint bitexts, although ideally we would test this with independently sourced parallel texts.

5.2 The choice of intermediate languages

The previous experiments used an ad-hoc choice of 'intermediate' language/s for triangulation, and we now examine which languages are most effective. Figure 4 shows the efficacy of the remaining nine languages when translating *fr* → *en*. Minimum error-rate training was not used for this experiment, or the next shown in Figure 5, in order to highlight the effect of the changing translation estimates. Romance languages (*es, it, pt*) give the best results, both on their own and when used together with the standard phrase-table (using uniform interpolation); Germanic languages (*de, nl, da, sv*) are a distant second, with the less related Greek and Finnish the least useful. Interpolation yields an improvement for all 'intermediate' languages, even Finnish, which has a very low score when used alone.

The same experiment was repeated for *en* → *de* translation with similar trends, except that the Germanic languages out-scored the Romance languages. These findings suggest that 'intermediate' languages which exhibit a high degree of similarity with the source or target language are desirable. We conjecture that this is a consequence of better automatic word alignments and a generally easier translation task, as well as a better preservation of information between aligned phrases.

Using a single language for triangulation clearly improves performance, but can we realise further improvements by using additional languages? Figure 5 shows the performance profile for *fr* → *en* when adding languages in a fixed order. The languages were ordered by family, with Romance before Germanic before Greek and Finnish. Each addition results in an increase in performance, even for the final languages, from which we expect little information. The purely triangulated (triang) and interpolated scores (interp) are converging, suggesting that the source-target bitext is redundant given sufficient triangulated data. We obtained similar results for *en* → *de*.

5.3 Evaluating the quality of the phrase-table

Our experimental results so far have shown that triangulation is not a mere approximation of the source-target phrase-table, but that it extracts additional useful translation information. We now assess the phrase-table quality more directly. Comparative statistics of a standard and a triangulated phrase-table are given in Table 4. The coverage over source and target phrases is much higher in the standard table than the triangulated tables, which reflects the reduced ability of triangulation to extract large phrases — despite the large increase in the number of events. The table also shows the overlapping probability mass which measures the sum of probability in one table for which the events are present in the other. This shows that the majority of mass is shared by both tables (as joint distributions), although there are significant differences. The Jensen-Shannon divergence is perhaps more appropriate for the comparison, giving a relatively high divergence

	standard	triang
source phrases (M)	8	2.5
target phrases (M)	7	2.5
events (M)	12	70
overlapping mass	0.646	0.750

Table 4: Comparative statistics of the standard triangulated table on *fr* → *en* using the full training set and Spanish as an intermediate language.

of 0.3937. This augurs well for the combination of standard and triangulated phrase-tables, where diversity is valued. The decoding results (shown in Table 3 for $fr \rightarrow en$) indicate that the two methods have similar efficacy, and that their interpolated combination provides the best overall performance.

6 Conclusion

In this paper we have presented a novel method for obtaining more reliable translation estimates from small datasets. The key premise of our work is that multi-parallel data can be usefully exploited for improving the coverage and quality of phrase-based SMT. Our triangulation method translates from a source to a target via one or many intermediate languages. We present a generative formulation of this process and show how it can be used together with the entries of a standard source-target phrase-table.

We observe large performance gains when translating with triangulated models trained on small datasets. Furthermore, when combined with a standard phrase-table, our models also yield performance improvements on larger datasets. Our experiments revealed that triangulation benefits from a large set of intermediate languages and that performance is increased when languages of the same family to the source or target are used as intermediates.

We have just scratched the surface of the possibilities for the framework discussed here. Important future directions lie in combining triangulation with richer means of conventional smoothing and using triangulation to translate between low-density language pairs.

Acknowledgements The authors acknowledge the support of EPSRC (grants GR/T04540/01 and GR/T04557/01). Special thanks to Markus Becker, Chris Callison-Burch, David Talbot and Miles Osborne for their helpful comments.

References

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

C. Callison-Burch, M. Osborne. 2003. Bootstrapping parallel corpora. In *Proceedings of the NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada.

C. Callison-Burch, P. Koehn, M. Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the HLT/NAACL*, 17–24, New York, NY.

A. Eisele. 2005. First steps towards multi-engine machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 155–158, Ann Arbor, MI.

G. Foster, R. Kuhn, H. Johnson. 2006. Phrase-table smoothing for statistical machine translation. In *Proceedings of the EMNLP*, 53–61, Sydney, Australia.

W. A. Gale, G. Sampson. 1995. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

T. Gollins, M. Sanderson. 2001. Improving cross language retrieval with triangulated translation. In *Proceedings of the SIGIR*, 90–95, New Orleans, LA.

M. Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(1–2):3–23.

P. Koehn, F. J. Och, D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the HLT/NAACL*, 48–54, Edomonton, Canada.

P. Koehn. 2003. *Noun Phrase Translation*. Ph.D. thesis, University of Southern California, Los Angeles, California.

P. Koehn. 2005. Europarl: A parallel corpus for evaluation of machine translation. In *Proceedings of MT Summit*, Phuket, Thailand.

E. Matusov, N. Ueffing, H. Ney. 2006. Computing consesus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of the EACL*, 33–40, Trento, Italy.

F. J. Och, H. Ney. 2001. Statistical multi-source translation. In *Proceedings of the MT Summit*, 253–258, Santiago de Compostela, Spain.

F. J. Och, C. Tillmann, H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the EMNLP and VLC*, 20–28, University of Maryland, College Park, MD.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the ACL*, 160–167, Sapporo, Japan.

K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the ACL*, 311–318, Philadelphia, PA.

P. Resnik, N. A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

M. Utiyama, H. Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of the HLT/NAACL*, 484–491, Rochester, NY.

R. Zens, H. Ney. 2004. Improvements in phrase-based statistical machine translation. In D. M. Susan Dumais, S. Roukos, eds., *Proceedings of the HLT/NAACL*, 257–264, Boston, MA.