Intersecting multilingual data for faster and better statistical translations

Yu Chen[1,2], **Martin Kay**[1,3], **Andreas Eisele**[1,2]
1: Universität des Saarlandes, Saarbrücken, Germany
2: Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Saarbrücken, Germany
3: Stanford University, CA, USA
{yuchen,kay,eisele}@coli.uni-saarland.de

Abstract

In current phrase-based SMT systems, more training data is generally better than less. However, a larger data set eventually introduces a larger model that enlarges the search space for the translation problem, and consequently requires more time and more resources to translate. We argue redundant information in a SMT system may not only delay the computations but also affect the quality of the outputs. This paper proposes an approach to reduce the model size by filtering out the less probable entries based on compatible data in an intermediate language, a novel use of *triangulation*, without sacrificing the translation quality. Comprehensive experiments were conducted on standard data sets. We achieved significant quality improvements (up to 2.3 BLEU points) while translating with reduced models. In addition, we demonstrate a straightforward combination method for more progressive filtering. The reduction of the model size can be up to 94% with the translation quality being preserved.

1 Introduction

Statistical machine translation (SMT) applies machine learning techniques to a bilingual corpus to produce a translation system entirely automatically. Such a scheme has many potential advantages over earlier systems which relied on carefully crafted rules. The most obvious is that it at dramatically reduces cost in human labor and it is able to reach many critical translation rules that are easily overlooked by human being.

SMT systems generally assemble translations by selecting phrases from a large candidate set. Unsupervised learning often introduces a considerable amount of noise into this set as a result of which the selection process becomes more longer and less effective. This paper provides one approach to these problems.

Various filtering techniques, such as (Johnson et al., 2007) and (Chen et al., 2008), have been applied to eliminate a large portion of the translation rules that were judged unlikely to be of value for the current translation. However, these approaches were only able to improve the translation quality slightly. In this paper, we describe a triangulation approach (Kay, 1997) that incorporates multilingual data to improve system efficiency and translation quality at the same time. Most of the previous triangulation approaches (Kumar et al., 2007; Cohn and Lapata, 2007; Filali and Bilmes, 2005; Simard, 1999; Och and Ney, 2001) add information obtained from a third language. In other words, they work with the union of the data from the different languages. In contrast, we work with the intersection of information acquired through a third language. The hope is that the intersection will be more precise and more compact than the union, so that a better result will be obtained more efficiently.

2 Noise in a phrase-based SMT system

The phrases in a translation model are extracted heuristically from a word alignment between the parallel texts in two languages using machine learning techniques. The translation model feature values are stored in the form of a so-called *phrase-table*,

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 128–136,
Boulder, Colorado, June 2009.

while the distortion model is in the *reordering-table*. As we have said models built in this way tend to contain a contains a considerable amount of noise. The phrase-table entries are far less reliable than the lexicons and grammar rules handcrafted for rule-based systems.

The main source of noise in the phrase table is errors from the word alignment process. For example, many function words occur so frequently that they are incorrectly mapped to translations of many function words in the other language to which they are, in fact, unrelated. On the other hand, many words remain unaligned on account of their very low frequency. Another source noise comes from the phrase extraction algorithm itself. The unaligned words are usually attached to aligned sequences In order to achieve longer phrase pairs.

The final selection of entries from the phrase table is based not only on the values assigned to them there, but also to values coming from the language and reordering models, so that entries that receive an initially high value may end up not being preferred.

(1) Sie lieben ihre Kinder nicht.
they love their children not
They don't love their children.

The frequently occurring German negative "*nicht*" in (1). is sometimes difficult for SMT systems to translate into English because it may appear in many positions of a sentence. For instance, it occurs at the end of the sentence in (1). The phrase pairs "*ihre kinder nicht → their children are not*" and "*ihre kinder nicht → their children*" are both likely also to appear in the phrase table and the former has greater estimated probability. However, the language model would preferred the latter in this example because the sentence "*They love their children are not.*" is unlikely to be attested. Accordingly, SMT system may therefore produce the misleading translation in (2).

(2) They love their children.

The system would not produce translations with the opposite meanings if the noisy entries like "*ihre kinder nicht → their children*" were excluded from the translation candidates. Eliminating the noise should help to improve the system's performance, for both efficiency and translation quality.

3 Triangulated filtering

While direct translation and pivot translation through a bridge language presumably introduce noise, in substantially similar amounts, there is no reason to expect the noise in the two systems to correlate strongly. In fact, the noise from such different sources, tends to be quite distinct, whereas the more useful information is often retained. This encourages us to hope that information gathered from various sources will be more reliable overall.

Our plan is to ameliorate the noise problem by constructing a smaller phrase-table by taking the *i*ntersection of a number of sources. We reason that a target phrase is will appear as a candidate translation of a given source phrase, only if it also appears as a candidate translation for some word or phrase in the bridge language mapping to the source phrase. We refer to this triangulation approach as *triangulated phrase-table filtering*.

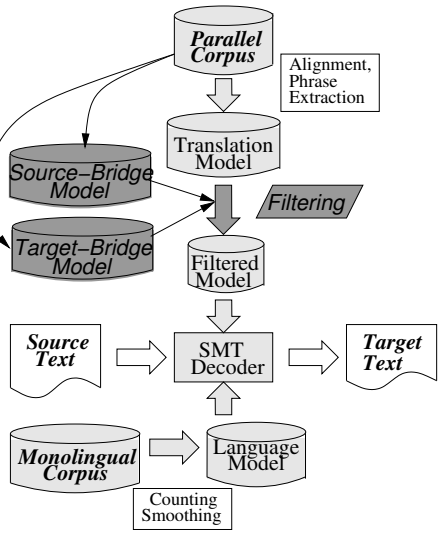

Figure 1: Triangulated filtering in SMT systems

Figure 1 illustrates our triangulation approach. Two bridge models are first constructed: one from the source language to the bridge language, and another from the target language to the bridge language. Then, we use these two models to filter the original source-target model. For each phrase pair in the original table, we try to find a common link in these bridge models to connect both phrases. If such links do not exist, we remove the entry from the table. The probability values in the table remain

unchanged. The reduced table can be used in place of the original one in the SMT system.

There are various forms of links that can be used as our evidence for the filtering process. One obvious form is complete phrases in the bridge language, which means, for each phrase pair in the model to be filtered, we should look for a third phrase in the bridge language that can relate the two phrases in the pair.

This approach to filtering examines each phrase pair presented in the phrase-table one by one. For each phrase pair, we collect the corresponding translations using the models for translation into a third language. If both phrases can be mapped to some phrases in the bridge language, but to different ones, we should remove it from the model. It is also possible that neither of the phrases appear in corresponding bridge models. In this case, we consider the bridge models insufficient for making the filtering decision and prefer to keep the pair in the table.

The way a decoder constructs translation hypotheses is directly related to the weights for different model features in a SMT system, which are usually optimized for a given set of models with minimum error rate training (MERT) (Och, 2003) to achieve better translation performance. In other words, the weights obtained for a model do not necessarily apply to another model. Since the triangulated filtering method removes a part of the model, it is important to readjust the feature weights for the reduced phrase-table.

4 Experimental design

All the text data used in our experiments are from Release v3 of "European Parliament Proceedings Parallel Corpus 1996-2006" (Europarl) corpus (Koehn, 2005). We mainly investigated translations from Spanish to English. There are enough structural differences in these two language to introduce some noise in the phrase table. French, Portuguese, Danish, German and Finnish were used as bridge languages. Portuguese is very similar to Spanish and French somewhat less so. Finnish is unrelated and fairly different typologically with Danish and German occupying the middle ground. In addition, we also present briefly the results on German-English translations with Dutch, Spanish and Danish as bridges.

For the Spanish-English pair, three translation models were constructed over the same parallel corpora. We acquired comparable data sets by drawing several subsets from the same corpus according to various maximal sentence lengths. The subsets

Model	Sentences	Tokens	
		Spanish	English
EP-20	410,487	5,220,142	5,181,452
EP-40	964,687	20,820,067	20,229,833
EP-50	1,100,813	26,731,269	25,867,370
Europarl	1,304,116	37,870,751	36,429,274

Table 1: Europarl subsets for building the Spanish-English SMT system

we used in the experiments are presented by "EP-20", "EP-40" and "EP-50", in which the numbers indicate the maximal sentence length in respective Europarl subsets. Table 1 lists the characteristics of the Spanish-English subsets. Although the maximal sentence length in these sets is far less than that of the whole corpus (880 tokens), EP-50 already includes nearly 85% of Spanish-English sentence pairs from Europarl.

The translations models, both the models to be filtered and the bridge models, were generated from compatible Europarl subsets using the Moses toolkit (Koehn et al., 2007) with the most basic configurations. The feature weights for the Spanish-English translation models were optimized over a development set of 500 sentences using MERT to maximize BLEU (Papineni et al., 2001).

The triangulated filtering algorithm was applied to each combination of a translation model and a third language. The reordering models were also filtered according to the phrase-table. Only those phrase pairs that appeared in the phrase-table remained in the reordering table. We rerun the MERT process solely based on the remaining entries in the filtered tables. Each table is used to translate a set of 2,000 sentences of test data (from the shared task of the third Workshop on Statistical Machine Translation, 2008 [1]). Both the test set and the development data set have been excluded from the training data.

We evaluated the proposed phrase-table filtering

[1] For details, see
http://www.statmt.org/wmt08/shared-task.html

method mainly from two points of view: the *efficiency* of systems with filtered tables and the *quality* of output translations produced by the systems.

5 Results

5.1 System efficiency

Often the question of machine translation is not only how to produce a good translation, but also how to produce it quickly. To evaluate the system efficiency, we measured both storage space and time consumption. For recording the computation time, we run an identical of installation of the decoder with different models and then measure the average execution time for the given translation task.

In Table 2, we give the number of entries in each phrase table (N), and the physical file size of the phrase table (S_{PT}) and the reordering table (S_{RT}) (without any compression or binarization), T_l, the time for the program to load phrase tables and T_t the time to translate the complete test set. We also highlighted the largest and the smallest reduction from each group.

All filtered models showed significant reductions in size. The greatest reduction of model sizes, taking both phrase-table and reordering table into account, is nearly 11 gigabytes for filtering the largest model (EP-50) with a Finnish bridge, which leads to the maximal time saving of 939 seconds, or almost 16 minutes, for translating two thousand sentences.

The reduction rates from two larger models are very close to each other whereas the filtered table scaled down the most significantly on the smallest model (EP-20), which was in fact constructed over a much smaller subset of Europarl corpus, consisting of less than half of the sentences pairs in the other two Europarl subsets. Compared to the larger Europarl subsets, the small data set is expected to produce more errors through training as there is much less relevant data for the machine learning algorithm to correctly extract useful information from. Consequently, there are more noisy entries in this small model, and therefore more entries to be removed. In addition, the filtering is done by exact matching of complete phrases, which presumably happens much less frequently even for correctly paired phrase pairs in the very limited data supplied by the smallest training set. For the same reason, the distinction between different bridge languages was less clear for this small model.

Due to hardware limitation, we are not able to fit the unfiltered phrase tables completely into the memory. Every table was filtered based on the given input so only a small portion of each table was loaded into memory. This may diminish the difference between the original and the filtered table to a certain degree. The relative time consumptionnevertheless agrees with the reduction in size: phrase tables from the smallest model showed the most reduction for both loading the models and processing the translations.

For loading time, we count the time it takes to start and to load the bilingual phrase-tables plus reordering tables and the monolingual language model into the memory. The majority of the loading time for the smallest model, even before filtering, has been used for loading language models and other start-up processes, could not be reduced as much as the reduction on table size.

5.2 Translation quality

Bridge	EP-20	EP-40	EP-50
—	26.62	31.43	31.68
pt	28.40	**32.90**	**33.93**
fr	28.28	32.69	33.47
da	**28.48**	32.47	33.88
de	28.05	32.65	33.13
fi	**28.02**	**31.91**	**33.04**

Table 3: BLEU scores of translations using filtered phrase tables

Efficiency aside, a translation system should be able to produce useful translation. It is important to verify that the filtering approach does not affect the translation quality of the system. Table 3 show the BLEU scores of each translation acquired in the experiments.

Between translation models of different sizes, there are obvious performance gaps. Different bridge languages can cause different effects on performance. However, the translation qualities from a single model are fairly close to each other. We therefore take it that the effect of the triangulation approach is rather *robust* across translation models of different sizes.

	Time		Table Size		
Model+Bridge	T_l (s)	T_t (s)	N	S_{PT} (byte)	S_{RT} (byte)
EP-20+ —	55	3529	7,599,271	953M	717M
EP-20+ pt	**53**	**2826**	**1,712,508** (22.54%)	**198M**	**149M**
EP-20+ fr	48	2702	1,536,056 (20.21%)	172M	131M
EP-20+ da	52	2786	1,659,067 (21.83%)	186M	141M
EP-20+ de	**43**	2732	**1,260,524** (16.59%)	**132M**	**101M**
EP-20+ fi	47	**2670**	1,331,323 (17.52%)	147M	111M
EP-40+ —	65	3673	19,199,807	2.5G	1.9G
EP-40+ pt	**50**	**3091**	8,378,517 (43.64%)	1.1G	1.8G
EP-40+ fr	46	3129	**8,599,708** (44.79%)	**1.1G**	**741M**
EP-40+ da	42	3050	6,716,304 (34.98%)	842M	568M
EP-40+ de	46	3069	6,113,769 (31.84%)	725M	492M
EP-40+ fi	**40**	**2889**	**4,473,483** (23.30%)	**533M**	**353M**
EP-50+ —	140	4130	54,382,715	7.1G	5.4G
EP-50+ pt	78	3410	13,225,654 (24.32%)	1.6G	1.3G
EP-50+ fr	**97**	**3616**	**24,057,849** (44.24%)	**3.0G**	**2.3G**
EP-50+ da	81	3418	12,547,839 (23.07%)	1.5G	1.2G
EP-50+ de	95	3488	15,938,151 (29.31%)	1.9G	1.5G
EP-50+ fi	**71**	**3191**	**7,691,904** (17.75%)	**895M**	**677M**

Table 2: System efficiency: time consumption and phrase-table size

It is obvious that the best systems are usually NOT from the filtered tables that preserved the most entries from the original. All the filtered models showed some improvement in quality with updated model weights. Mostly around 1.5 BLEU points, the increases ranged from 0.36 to 2.25. Table 4 gives a set of translations from the experiments. The unfiltered baseline system inserted the negative by mistake while all the filtered systems are able to avoid this. It indicates that there are indeed noisy entries affecting translation quality in the original table. We were able to achieve better translations by eliminating noisy entries.

The filtering methods indeed tend to remove entries composed of long phrases. Table 5 lists the average length of phrases in several models. Both source phrases and target phrases are taken into account. The best models have shortest phrases on average. Discarding such entries seems to be necessary. This is consistent with the findings in (Koehn, 2003) that phrases longer than three words improve performance little for training corpora of up to 20 million words.

Quality gains appeared to converge in the results across different bridge languages while the original models became larger. Translations generated using large models filtered with different bridge languages are less diverse. Meanwhile, the degradation is less for a larger model. It is reasonable to expect improvements for extremely large models with arbitrary bridge languages. For relatively small models, the selection of bridge languages would be critical for the effect of our approach.

Bridge	EP-20	EP-40	EP-50
—	3.776	4.242	4.335
pt	3.195	3.943	3.740
fr	3.003	3.809	3.947
da	3.005	3.74	3.453
de	2.535	3.501	3.617
fi	2.893	3.521	3.262

Table 5: Average phrase length

5.3 Language clustering

To further understand how the triangulated filtering approach worked and why it could work as it did, we examined a randomly selected phrase table fragment through the experiments. The segment included 10 potential English translations of the same Spanish word "*fabricantes*", the plural form of the word "*fabricante*" (manufacturer).

Table 6 shows the filtering results on a randomly selected segment from the original "EP-40" model, including 10 English translations of the same source

source	Así, se van modificando poco a poco los principios habituales del Estado de derecho por influencia de una concepcin extremista de la lucha con tra las discriminaciones..
ref	thus , the usual principles of the rule of law are being gradually altered under the influence of an extremist approach to combating discrimination.
baseline	we are not changing the usual principles of the rule of law from the influence of an extremist approach in the fight against discrimination.
pt	so , are gradually changing normal principles of the rule of law by influence of an extremist conception of the fight against discrimination.
fr	so , we are gradually changing the usual principles of the rule of law by influence of an extremist conception of the fight against discrimination.
da	so , are gradually changing the usual principles of the rule of law by influence of an extremist conception of the fight against discrimination.
de	thus , we are gradually altering the usual principles of the rule of law by influence of an extremist conception of the fight against discrimination.
fi	so , are gradually changing normal principles of the rule of law by influence of an extremist conception of the fight against discrimination.

Table 4: Examples

fabricantes	pt	fr	da	de	fi	
a manufacturer	✓	✓	✓		✓	4
battalions	✓	✓	✓			3
car manufacturers have						0
car manufacturers	✓	✓	✓	✓	✓	5
makers	✓	✓			✓	3
manufacturer	✓	✓	✓	✓	✓	5
manufacturers	✓	✓	✓	✓	✓	5
producers are		✓	✓	✓		3
producers need						0
producers	✓	✓	✓	✓	✓	5

Table 6: Phrase-table entries before and after filtering a model with different bridges

word "*fabricantes*". ✓ indicates that the corresponding English phrase remained in the table after triangulated filtering with the corresponding bridge language. We also counted the number of tables that included each phrase pair.

Regardless of the bridge language, the triangulated filtering approach had removed those entries that are clearly noise. Meanwhile, entries which are surely correct were always preserved in the filtered tables. The results of using different bridge languages turned out to be consistent on these extreme cases. The 5 filtering processes agreed on six out of ten pairs.

As for the other 4 pairs, the decisions were different using different bridge languages. The remaining entries were always different when the bridge was changed. None of the languages led to the identical eliminations. None of the cases excludes all errors. Apparently, the selection of bridge languages had immediate effects on the filtering results.

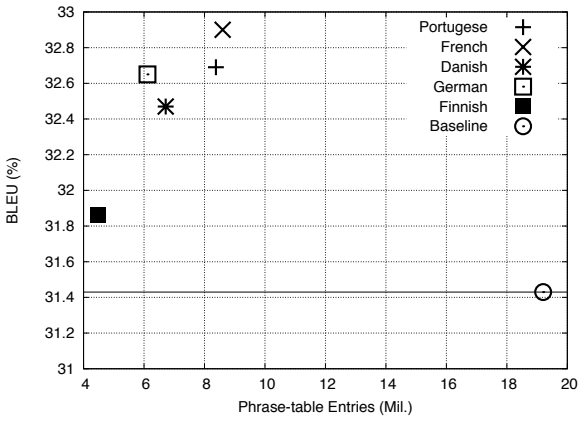

Figure 2: Clustering of bridge languages

We compared two factors of these filtered tables: their sizes and the corresponding BLEU scores. Figure 2 shows interesting signs of language similarity/dissimilarity. There are apparently two groups of languages having extremely close performance, which happen to fall in two language groups: Germanic (German and Danish) and Romance (French and Portuguese). The Romance group was associated with larger filtered tables that produced slightly better translations. The filtered tables created with Germanic bridge languages contained ap-

proximately 2 million entries less than Romance groups. The translation quality difference between these two groups was within 1 point of BLEU.

Observed from this figure, it seems that the translation quality was connected to the similarity between the bridge language and the source language. The closer the bridge is to the source language, the better translations it may produce. For instance, Portuguese led to a filtered table that produced the best translations. On the other hand, the more different the bridge languages compared to the source, the larger portion of the phrase-table the filtering algorithm will remove. The table filtered with German was the smallest in the four cases.

Finnish, a language that is unrelated to others, was associated with distinctive results. The size of the table filtered with Finnish is only 23% of the original, almost half of the table generated with Portuguese. Finnish has extremely rich morphology, hence a great many word-forms, which would make exact matching in bridge models less likely to happen. Many more phrase pairs in the original table were removed for this reason even though some of these entries were beneficial for translations. Even though the improvement on translation quality due to the Finnish bridge was less significant than the others, it is clear that triangulated filtering retained the useful information from the original model.

5.4 Further filtering

The filtering decision with a bridge language on a particular phrase pair is fixed: either to keep the entry or to discard it. It is difficult to adjust the system to work differently. However, as the triangulated filtering procedure does not consider probability distributions in the models, it is possible to further filter the tables according to the probabilities.

The phrase pairs are associated with values computed from the given set of feature weights and sorted, so that we can remove any portions of the remain entries based on the values. Each generated table is used to translate the test set again. Figure 3 shows BLEU scores of the translation outputs produced with tables derived from the "EP-50" model with respect to their sizes. We also included the curve of probability-based filtering alone as the baseline.

The difference between filtered tables at the same

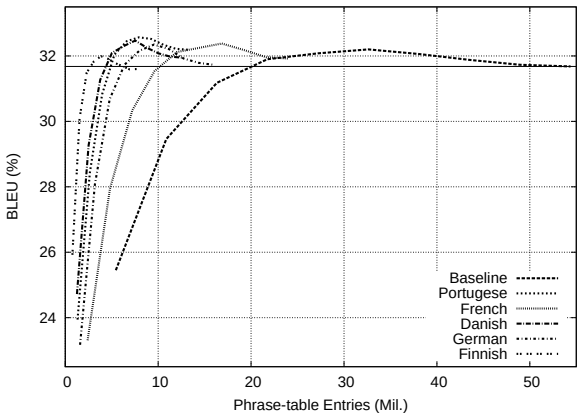

Figure 3: Combining probability-based filtering

size can be over 6 BLEU points, which is a remarkable advantage for the triangulated filtering approach always producing better translations. The curves of the triangulated filtered models are clearly much steeper than that of the naive pruned ones. Data in these filtered models are more compact than the original model before any filtering. The triangulated filtered phrase-tables contain more useful information than a normal phrase-table of the same size. The curves representing the triangulated filtering performance are always on the left of the original curves.

We are able to use less than 6% of the original phrase table (40% of the table filtered with Finnish) to obtain translations with the same quality as the original. The extreme case, using only 1.4% of the original table, leads to a reasonable BLEU score, indicating that most of the output sentences should still be understandable. In this case, the overall size of the phrase table and the reordering table was less than 100 megabytes, potentially feasible for mobile devices, whereas the original models took nearly 12.5 gigabytes of disk space.

5.5 Different source language

Bridge	EP-40		EP-50	
—	5.1G	26.92	6.5G	27.23
Dutch	562M	27.11	1.3G	28.14
Spanish	3.0G	27.28	3.6G	28.09
Danish	505M	28.04	780M	28.21

Table 7: Filtered German-English systems (Size and BLEU)

In addition to Spanish-English translation, we also conducted experiments on German-English translation. The results, shown in Table 7, appear consistent with the results of Spanish-English translation. Translations in most cases have performance close to the original unfiltered models, whereas the reduction in phrase-table size ranged from 40% to 85%. Meanwhile, translation speed has been increased up to 17%.

Due to German's rich morphology, the unfiltered German-English models contain many more entries than the Spanish-English ones constructed from similar data sets. Unlike the Spanish-English models, the difference between "EP-40" and "EP-50" was not significant. Neither was the difference between the impacts of the filtering in terms of translation quality. In addition, German and English are so dissimilar that none of the three bridge languages we chose turned out to be significantly superior.

6 Conclusions

We highlighted one problem of the state-of-the-art SMT systems that was generally neglected: the noise in the translation models. Accordingly, we proposed triangulated filtering methods to deal with this problem. We used data in a third language as evidence to locate the less probable items in the translation models so as to obtain the **intersection** of information extracted from multilingual data. Only the occurrences of complete phrases were taken into account. The probability distributions of the phrases have not been considered so far.

Although the approach was fairly naive, our experiments showed it to be effective. The approaches were applied to SMT systems built with the Moses toolkit. The translation quality was improved at least 1 BLEU for all 15 cases (filtering 3 different models with 5 bridge languages). The improvement can be as much as 2.25 BLEU. It is also clear that the best translations were not linked to the largest translation models. We also sketched a simple extension to the triangulated filtering approach to further reduce the model size, which allows us to generate reasonable results with only 1.4% of the entries from the original table.

The results varied for different bridge languages as well as different models. For translation from Spanish to English, Finnish, the most distinctive bridge language, appeared to be a more effective intermediate language which could remove more phrase pair entries while still improving the translation quality. Portuguese, the most close to the source language, always leads to a filtered model that produces the best translations. The selection of bridge languages has more obvious impact on the performance of our approach when the size of the model to filter was larger.

The work gave one instance of the general approach described in Section 3. There are several potential directions for continuing this work. The most straightforward one is to use our approaches with more different languages, such as Chinese and Arabic, and incompatible corpora, for example, different segments of Europarl. The main focus of such experiments should be verifying the conclusions we had in this paper.

Acknowledgments

This work was supported by European Community through the EuroMatrix project funded under the Sixth Framework Programme and the EuroMatrix Plus project funded under the Seventh Framework Programme for Research and Technological Development.

References

Yu Chen, Andreas Eisele, and Martin Kay. 2008. Improving Statistical Machine Translation Efficiency by Triangulation. In *the 6th International Conference on Language Resources and Evaluation (LREC '08)*, May.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech, June.

Karim Filali and Jeff Bilmes. 2005. Leveraging Multiple Languages to Improve Statistical MT Word Alignments. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*, Cancun, Mexico, November.

J. Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural*

Language Learning (EMNLP-CoNLL), Prague, Czech Republic, June.

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

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 *the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, June.

Philipp Koehn. 2003. *Noun Phrase Translation*. Ph.D. thesis, University of Southern California.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*.

Shankar Kumar, Franz Josef Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *MT Summit VIII*, Santiago de Compostela, Spain.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Michel Simard. 1999. Text-translation alignment: Three languages are better than two. In *EMNLP/VLC-99*, College Park, MD, June.