Improving Statistical Machine Translation Efficiency by Triangulation

Yu Chen, Andreas Eisele, Martin Kay

Saarland University
Saarbrücken, Germany
{yuchen, eisele, kay}@coli.uni-sb.de

Abstract

In current phrase-based Statistical Machine Translation 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 decoder, and consequently requires more time and more resources to translate. This paper describes an attempt to reduce the model size by filtering out the less probable entries based on testing correlation using additional training data in an intermediate third language. The central idea behind the approach is *triangulation*, the process of incorporating multilingual knowledge in a single system, which eventually utilizes parallel corpora available in more than two languages. We conducted experiments using Europarl corpus to evaluate our approach. The reduction of the model size can be up to 70% while the translation quality is being preserved.

1. Introduction

Statistical machine translation (SMT) is now generally taken to be an enterprise in which machine learning techniques are applied to a bilingual corpus to produce a translation system entirely automatically. Such a scheme has many potential advantages over earlier systems which relied on large bodies of carefully crafted rules. The most obvious is that it becomes thinkable to construct a system for a new pair of language in a matter of days, if not hours, rather than years, and at dramatically reduced cost in human labor. It can also be claimed, though perhaps more controversially, that a system based directly on authentic data will be more likely to take into account phenomena that a human might overlook or consider unimportant. Whereas a system built in the more traditional way might sometimes be unable to produce any translation at all, however unsatisfactory, for some inputs, this is fare less likely to happen in an automatically generated system.

On the other hand, the new approach also has potential drawbacks. In particular, the data on which it is based rarely contain more than one translation of each sentence, chosen essentially at random from an often large set of equally good candidates. The information on which the system is based therefore contains a great deal of noise. This paper describes a preliminary attempt to reduce that noise and, at the same time to enhance the efficiency of the resulting system in certain cases. We propose to take advantage of a special, but increasingly common, circumstance, namely that in which translations of a single original are required in more than one other language. Considerable amounts of data of the kind required to explore this proposal already exist in the form of parallel corpora are already available in more than two languages, such as Europarl (Philipp Koehn, 2005), JRC-Aquis (Steinberger et al., 2006), UN parallel text corpus (Graff, 1994), etc. The opportunity that such resources presents is that of incorporating multilingual knowledge in in a single statistical MT systems. Related proposals have sometimes been referred to as *triangulation* (Kay, 1997).

2. Phrase-based SMT

The dominant paradigm in SMT is referred to as *phrase-based* machine translation. The term refers to a sequence of words characterized by its statistical rather then any grammatical properties. At the center of the process is a *phrase table*, a list of phrases identified in a source sentence together with potential translations. Phrases in the source sentence may overlap and also may have several translations, so that a subset of the entries in the table must, in general, be selected to make a translation of the sentence. The members of the selected subset must then be arranged in a specific order to give a translation. These operations are determined by statistical properties of the target language enshrined in the so-called *language model*. Several current SMT systems work in this way and, most notably for our purposes, the freely available, open-source Moses Toolkit (Philipp Koehn et al., 2007) on which our work is based.

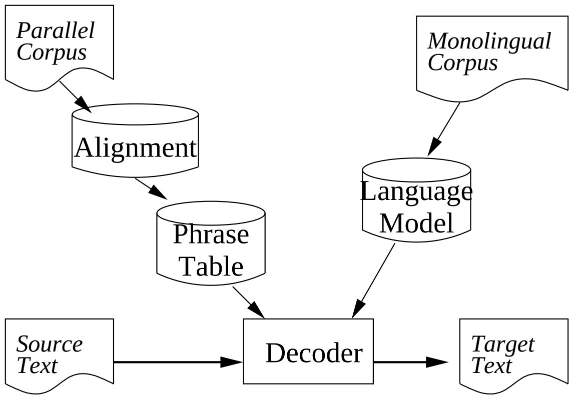

Figure 1: Phrase-based SMT system

Whereas the use of the phrase table is determined by the language model, the constitution of the table is determined by the *translation model* which captures the supposedly relevant statistical properties of a corpus consisting of paired source and target sentences. Very generally speaking, the faithfulness, or accuracy of a translation depends more on the translation model, and it fluency on the language model.

The particular kinds of translation model most generally used at present are constructed using another open-souce program, called GIZA++ (Och and Ney, 2003). This program identifies phrases and potential translations and associates a number of statistics with each. The results can therefore be used to construct a variety of different phrase tables and use them in a number of different ways.

As we pointed out earlier, there is generally much noise in the data on which SMT systems are based. Consequently, a phrase table can be expected to contain several phrases that are paired with putative translations to which they are not related in any interesting way. This is because the pair has been observed in corresponding pairs of sentences, but not as translations of one another. The strength of the association between a phrase and a good translation may be reduced simply because there are many good ways of translating that phrase. The more noise of this kind there is in the system, the greater the chance that unacceptable translations will be produced, and also, the greater the space of possibilities that must be explored in order to extract a translation from the table. In other words, both the quality and the efficiency of the system are affected. A translation model can easily occupy 30 gigabytes, and can take hours to load into the machine. Various filtering techniques can be used to remove parts of the model that can be predicted to be of no interest for the current translation. The proposals we make here are in this spirit.

3. Trianguation in SMT

Suppose that E_1 and E_2 being considered as possible translations of a Spanish sentence S. Suppose, furthermore, that a French translation, F, of the Spanish text is also available, and that candidate English translations of the corresponding French sentence were E_2 and E_3. We could reasonably take this as evidence that E_2 was likely to be a better translation of the original Spanish (or the French). There are two obvious bases for such an argument. One is that the French version of the text contains information, not available from the original, that can be used to reduce the space of possible English translations. The other is, in a sense, the converse of this, namely that the translation process has identified misleading information, or noise, in the Spanish that has no counterpart in the French and that was responsible for the inclusion or E_1 in the set. The first argument has merit only if there is some possible source of additional information in the French. If it has been produced by a method that had been shown to be superior, perhaps even involving human editing, then there is such a possible source. But if it was produced in essentially the same way as the translation from the Spanish, then it is hard to find where the additional information might have come from other than the language model used in that translation. This is presumably as weak source of information, but one that should not be entirely discounted. The second argument, however, remains in tact because, while both translations presumably introduce noise, in substantially similar amounts, there is no reason to expect the noise in the two systems to correlate strongly. E_2 should represent the nearest approximation to the information in S and F and this is likely to be the information that we want to preserve. The point E_2 is thus identified by triangulation from the points S and F.

The situation we have just sketched is unrealistic in that the translations of a pair of mutual translations, S and F would have an empty intersection in a large proportion of cases [1] But this is presumably not true of the phrase tables from which the translations would be produced in a SMT of the kind we have sketched. The tables used in translating the English and the French would be expected to have entries in common, and these should be less noisy than those that are found in only one of the tables. This is the primary observation on which our work is based.

The idea of triangulation in machine translation is not new. (Simard, 1999) describes experiments showing that a system based on trillingual texts can yield what amount to better bilingual sentence phrase tables, while retaining the same computational complexity as the common bilingual approach. (Kumar et al., 2007) incorporate third languages to construct word alignments, that is, essentially, single-word phrase tables. (Cohn and Lapata, 2007) creates a larger phrase table by incorporating one obtained by translating into a third language. In this way, lexical some gaps in the original training data are filled by training data from the third language.

These earlier approaches have all had the aim of improving the quality of translations and, as we have said, success in this enterprise rests on the assumption that a second translation introduces new information about the original text into the system. Such advantages as they have brought have generally been at considerable cost to the efficiency of the system as a whole because the technique has involved, effectively, adding new entries to the phrase table. With an eye more towards the efficiency of the process as a whole, we aim to *remove* from the table that are not whose presence there is not supported by the other language. While previous approaches effectively took the union of a pair of phrase tables, we aim to work with their intersection and thus crucially to reduce the search space.

4. Phrase table filtering with triangulation

Reducing the size of the phrase table should clearly improve the efficiency and a quick inspection of a phrase talble generally suggests that there should be many opportunities to do this. Two phrases often appear as one pair in a model with high probabilities only because they occur once and in a single sentence pair.

On the other hand, careless pruning may harm the translation quality as the process may eliminate good information that was originally in the model. Our plan is to eliminate a considerable amount of noise by constructing a new phrase table from those arrived at in translating both from the original and the third language. Essentially, we retain a pair in the original table only if a pair with the same output string appears in the table coming from the third language. We argue that, if a target phrase cannot be translated from any of the bridge translations, this phrase is also unlikely to be the translation of original source phrase. We refer to this as *phrase-table filtering*.

[1] However, an experiment along these lines, reported in (Och and Ney, 2001), showed no significant improvement in word error rate.

In this section, we introduce two specific techniques for phrase-table filtering with help from resources in a third language. The first method (Method 1) looks for phrases in the bridge language that can connect phrase pairs in the phrase table in question. It requires strict matches of complete phrases. These constraints are relaxed in the second method (Method 2) by scoring over vocabulary overlap.

4.1. Method 1

If phrases in two languages are true translations of each other, then they presumably have at least one meaning in common, regardless of other ones they each may have that do not correspond. We therefore hope to be able to find one phrase in the third language that also shares this meaning. Provided these facts are reflected in the training data, this will give rise to a reliable translation pair. On the other hand, when translating two phrases paired together by mistake into the third language, we are likely to have distinct sets of candidates for the phrases.

This approach to phrase table filtering examines each phrase pair presented in the phrase table one by one. For each phrase pair, we collect the corresponding translations using the model for translation to a third language. Even if both phrases can be mapped to some phrases in the additional language, but to different ones, it is reasonable to assume the pair is less useful and we should remove it from the model.

Suppose that a translation is to be made from language Spanish(S) to English(E), using French(F) as the bridge language. Let $P_{S\to E}$, $P_{S\to F}$, $P_{E\to F}$ be the phrase tables involved in the Spanish-to-English, Spanish-to-French, and English-to-French translations. We construct a new table $P'_{S\to E}$ containing just those entries (a, e) such that, for some f, $(a, f) \in P_{S\to F}$ and $(e, f) \in P_{E\to F}$.

This clearly leaves many implausible entries remain in the phrase table because, exact string matching is doubtless too strict. Errors in the models involving the bridge languages can lead to mismatches when examing a useful phrase pair.

4.2. Method 2

Our second method differs from the first in two important ways:

- We no longer look for an exact phrase that can be mapped to both phrases in a given pair. Instead, all words having appeared in one or more translations of a phrase to the bridge language are collected into a vocabulary set assigned to this phrase. Then, the question turns to whether there are any overlaps between the two sets extracted independently for the two original phrases in a pair. Even when various types of errors occur in the translation models, it is still reasonable to suggest there must be a common vocabulary in the bridge language being used to translate two phrases carrying the same meanings. In return, from a vocabulary established for the source phrase, if there is no way to form any phrases which can be translated into the target phrase, then we can conclude the target phrase is unlikely to be a translation of the source phrase.
- Aiming at more fine-grained filtering, we introduce a correlation measure based on external data. It is obvious that the overlap of two vocabularies has somewhat connection to the correlation of the two phrases. Therefore, we compare the size of the overlap to the size of two vocabularies to show the degree of correlation. If a complete match is found, the corresponding pair tends to be the most probable ones. On the contrary, the less common words found in the vocabularies, the less correlations between two phrases. Note, we also penalize the pair for which overlap does not exist in respective non-empty translation vocabularies so as to distinguish them with the phrase pairs from which one or two phrases cannot be found in given external data.

For language X and Y, the set of words that have appeared in any potential translation of a phrase x in language X to language Y is:

$$W_{X\to Y}(x) = \{w|\exists y,\ \text{s.t. } w \in y \text{ and } (x, y) \in P_{X\to Y}\}$$

We also define the set of words that appear in any phrases in language X that can be translated into the phrase y in language Y:

$$W'_{X\to Y}(y) = \{w|\exists x,\ \text{s.t. } w \in x \text{ and } (x, y) \in P_{X\to Y}\}$$

Using the same languages as above, we now get three sets of phrase pairs: $P_{S\to E}$, $P_{S\to F}$ and $P_{F\to E}$. Given a phrase pair $(s, e) \in P_{S\to E}$, we can infer $W_{S\to F}(s)$ and $W'_{F\to E}(e)$. An overlap score is assigned to the phrase pair according to Equation 1

$$O_{(s,e)} = \frac{|W_{S\to F}(s) \cap W'_{F\to E}(e)|}{\min(|W_{S\to F}(s)|, |W'_{F\to E}(e)|)}, \tag{1}$$

if $W_{S\to F}(s)$ and $W'_{F\to E}(e)$ contain any words in common. Otherwise, ie. $W_{S\to F}(s) \cap W'_{F\to E}(t) = \emptyset$, the language pair is penalized following Equation 2:

$$O_{(s,e)} = \begin{cases} 0 & \text{if } W_{S\to F}(s) = \emptyset \text{ and } W'_{F\to E}(t) = \emptyset \\ -1 & \text{if } W_{S\to F}(s) = \emptyset \text{ and } W'_{F\to E}(t) \neq \emptyset \\ -1 & \text{if } W_{S\to F}(s) \neq \emptyset \text{ and } W'_{F\to E}(t) = \emptyset \\ -2 & \text{if } W_{S\to F}(s) \neq \emptyset \text{ and } W'_{F\to E}(t) \neq \emptyset \end{cases} \tag{2}$$

Phrase tables can be filtered according to different thresholds. The phrase pairs in the second group are assigned with non-positive numbers. These scores only indicate the corresponding categories rather than the correlation degrees. Overlapping words are merely one of many potential indications of correlations. It is possible to introduce more factors into this scoring scheme to show more distinction between pairs. This is in particular important for studying the phrase pairs currently with negative scores.

5. Experiments

We construct two subsets of Europarl corpus in English, French, German and Spanish: one consists of sentences with a maximal length of 40 tokens and the other are made up of sentences of less than 50 tokens, respectively presented by Europarl-40 and Europarl-50. There are about

950,000 sentences for each language in the first subset and around 1,100,000 sentences in the second subset. Our approaches are evaluated on 2,000 sentences of test data from the shared task of the third Workshop on Statistical Machine Translation, 2008 [2].

The experiments are performed on translations from Spanish to English. We choose French and German as bridge languages used separately. Phrase-based models are then built for the following translation directions: Spanish-English, Spanish-French, Spanish-German, French-English, German-English, English-French and English-German. The Spanish-English phrase tables are filtered through either bridge language using other phrase tables. For each direction, two models are trained from both Europarl subsets. The Spanish-English models are considered as baselines. Neither of our approaches modify feature values in a phrase table. Only a part of entries are removed from the table. The probability distributions of the features remain the same as before filtering. We only train the feature weights for these two baseline translation models using minimum error rate training (Och, 2003) to maximize the BLEU scores on a set of 500 sentences from a development data provided by the mentioned shared task. The other systems simply adapt the weights obtained for respective baseline, namely the baseline trained on the same subset of Europarl. Moreover, all experiments use the same English 5-gram language model trained from Europarl corpus.

We apply both approaches to every combination of two original phrase tables and two bridge languages, that is, 4 setups in total. Reordering tables, which keep the distortion models, are also filtered following the same scheme when phrase tables are being filtered. As for method 2, the threshold is set to zero, namely only phrase pairs with negative scores are excluded.

6. Results

The results of these experiments are showed in Table 1 through 5. In the tables, the translation models are denoted by names consisting of two parts: the number in the first part indicates the corresponding approach and the language in the second part is used as the bridge.

Model	Phrase pairs	PT (Byte)	RT (Byte)
Baseline	19,199,807	2.5G	1.9G
1:French	8,599,708	1.1G	741M
2:French	14,877,456	1.9G	1.3G
1:German	6,113,769	725M	492M
2:German	13,600,633	1.8G	1.2G

Table 1: Size of the models on Europarl-40

Table 1-2 present the sizes of various models used in the experiments. In these two tables, PT stands for the phrase tables and RT refers to the reordering tables. Both physical sizes of the files and number of entries in them are listed. Table 3 shows the excluded portion for each model. The size of a phrase table can be reduced to less than 30% of the original. Obviously, Method 1 filters out more phrase pairs than Method 2 because exact phrase matches in Method 1 occur much less frequent than word overlaps between translation vocabularies in Method 2 given the same translation models. Both methods remove slightly more entries when the phrase table to filter is larger. The reduction appears to be more significant using German as the third language. French is so close to Spanish, the source language, that the data in French cannot provide as much information as German data.

Model	Phrase pairs	PT (Byte)	RT (Byte)
Baseline	54,382,715	7.1G	5.4G
1:French	24,057,849	3.0G	2.3G
2:French	41,821,489	5.5G	4.2G
1:German	15,938,151	1.9G	1.5G
2:German	37,841,524	5G	3.8G

Table 2: Size of the models on Europarl-50

	Europarl-40	Europarl-50
1:French	55.21%	55.77%
2:French	23.52%	24.10%
1:German	69.16%	70.70%
2:German	29.16%	30.42%

Table 3: Removed portions through filtering

	Bridge language		
Training set	None	French	German
Europarl-40	31.43	28.27	31.58
Europarl-50	31.65	31.73	31.92

Table 4: BLEU scores using models filtered with Method 1

	Bridge language		
Training set	None	French	German
Europarl-40	31.43	28.20	31.38
Europarl-50	31.65	31.69	31.75

Table 5: BLEU scores using models filtered with Method 2

The approaches are evaluated by means of BLEU score (Papineni et al., 2001). The results are listed in Table 4 and Table 5. It becomes clear that filtering would not reduce BLEU in most cases. In general, using German data results in more improvement. However, the performance of models filtered through French and German actually converges while the original phrase table becomes larger. The distinction between two approaches also decreases with larger models. The scores increase in larger models based on larger corpora. This is plausible since a larger models usually contains more noise to be excluded. In addition, larger corpus in the third language can also introduce more constraints for filtering. Taking the model sizes into account, we can infer that smaller models do not necessarily produce worse results. The best BLEU score was not produced by the largest phrase table we have produced in the experiments.

Our approaches reduce the sizes of models used for SMT

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

Source	Como ha señalado el Sr. de Soto, no esperamos que el progreso sea tarea fácil, y el éxito del proceso de las Naciones Unidas no está ni mucho menos garantizado.	
Reference	As Mr de Soto noted, we do not expect progress to be easy, and the success of the UN process is far from assured.	
Europarl-40	Baseline	As has been pointed out by Mr de Soto, we hope that progress is not an easy task, and the success of the UN process is far from guaranteed.
	1:French	As pointed out by the sr. of Soto , we hope that progress is not an easy task , and the success of the process of the United Nations is far from guaranteed .
	2:French	As pointed out by the sr. of Soto, we hope that progress is not an easy task, and the success of the process of the United Nations is far from guaranteed.
	1:German	As Mr de Soto, we expect that progress is not an easy task, and the success of the UN process is far from guaranteed.
	2:German	As has been pointed out by Mr de Soto, we hope that progress is not an easy task, and the success of the UN process is far from guaranteed.
Europarl-50	Baseline	As has been pointed out by Mr de Soto , we hope that progress is not an easy task, and the success of the UN process is far from guaranteed.
	1:French	As Mr de Soto , we do not expect that progress is easy , and the success of the UN process is far from guaranteed .
	2:French	As Mr de Soto , we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
	1:German	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
	2:German	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.

Table 6: Example translations

and thereby reduce the time and space costs required for translation tasks. Meanwhile, translation quality is improved after the models being filtered. Table 6 illustrates an example on how the translations can be better after filtering. The translations from both baselines have selected the word "*hope*" for "*esperamos*", which leads to a wrong meaning. This error disappeared in the results of 5 filtered models out of 8. Since we made no changes to probability distributions in the models, the only explanation for this is that the filtering procedure has removed the noisy entry/entries causing the problem.

7. Discussion

In this paper, we have presented two novel approaches to phrase tables filtering for more efficient translations. Multi-parallel data can be useful for reduce the computation costs without harming the translation quality of phrase-based SMT. Our triangulation methods verify a translation model via an intermediate language. The performance of the filtered models greatly relates to the choice of the bridge language.

There are several potential directions to continue this work. Selection of the intermediate language needs to be studied more systematically. According to the experimental results, there is a vague idea about what languages can be more helpful as the additional language for filtering, however more experiments, including using more than one intermediate languages for filtering, are important before we can draw any general conclusions.

Another potential work is refinement of the correlation measure introduced in Method 2. Current design of the scoring scheme is still ad hoc. We can only assign certain labels to the phrase pairs under consideration instead of numerical values that indeed reflect the degrees of correlation. The approaches in this paper are working with phrases on the model level. If the methods can be scaled up to hypotheses level, at which we work with complete sentences rather than small units, we can easily imagine resources in the third language can help to eliminate implausible translation candidates.

8. Acknowledgement

This work is supported by the EuroMatrix project funded by the European Commission (6^{th} Framework Programme). We also want to thank anonymous reviewers for their helpful comments.

9. References

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.

David Graff. 1994. UN Parallel Text (Complete). Linguistic Data Consortium, Philadelphia.

J. Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 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, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

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, June.

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

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

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 *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

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

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

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

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *the 5th International Conference on Language Resources and Evaluation (LREC'2006).*, Genoa, Italy, May.