Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages

Preslav Nakov
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nakov@comp.nus.edu.sg

Hwee Tou Ng
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nght@comp.nus.edu.sg

Abstract

We propose a novel language-independent approach for improving statistical machine translation for resource-poor languages by exploiting their similarity to resource-rich ones. More precisely, we improve the translation from a resource-poor source language X_1 into a resource-rich language Y given a bi-text containing a limited number of parallel sentences for X_1-Y and a larger bi-text for X_2-Y for some resource-rich language X_2 that is closely related to X_1. The evaluation for Indonesian→English (using Malay) and Spanish→English (using Portuguese and pretending Spanish is resource-poor) shows an absolute gain of up to 1.35 and 3.37 Bleu points, respectively, which is an improvement over the rivaling approaches, while using much less additional data.

1 Introduction

Recent developments in statistical machine translation (SMT), *e.g.*, the availability of efficient implementations of integrated open-source toolkits like Moses (Koehn et al., 2007), have made it possible to build a prototype system with decent translation quality for any language pair in a few days or even hours. In theory. In practice, doing so requires having a large set of parallel sentence-aligned bi-lingual texts (a *bi-text*) for that language pair, which is often unavailable. Large high-quality bi-texts are rare; except for Arabic, Chinese, and some official languages of the European Union (EU), most of the 6,500+ world languages remain resource-poor from an SMT viewpoint.

While manually creating a small bi-text could be relatively easy, building a *large* one is hard, *e.g.*, because of copyright. Most bi-texts for SMT come from parliament debates and legislation of multi-lingual countries (*e.g.*, French-English from Canada, and Chinese-English from Hong Kong), or from international organizations like the United Nations and the European Union. For example, the *Europarl* corpus of parliament proceedings consists of about 1.3M parallel sentences (up to 44M words) per language for 11 languages (Koehn, 2005), and the *JRC-Acquis* corpus provides a comparable amount of European legislation in 22 languages (Steinberger et al., 2006).

The official languages of the EU are especially lucky in that respect; while this includes such "classic SMT languages" like English and French, and some important international ones like Spanish and Portuguese, most of the rest have a limited number of speakers and were resource-poor until recently; this is changing quickly because of the increasing volume of EU parliament debates and the ever-growing European legislation. Thus, becoming an official language of the EU has turned out to be an easy recipe for getting resource-rich in bi-texts quickly. Of course, not all languages are that 'lucky', but many can still benefit.

In this paper, we propose using bi-texts for resource-rich language pairs to build better SMT systems for resource-poor ones by exploiting the similarity between a resource-poor language and a resource-rich one.

The proposed method allows non-EU languages to benefit from being closely related to one or more official languages of the EU, the most obvious candidates being Norwegian (related to Swedish), Moldavian[1] (related to Romanian), and Macedonian[2] (related to Bulgarian). After Croatia joins the EU, Serbian, Bosnian and Montenegrin will be able to benefit from Croatian gradually turning resource-rich (all four split from Serbo-Croatian after the breakup of Yugoslavia). The newly-made EU-official (and thus not as resource-

[1] Not recognized by Romania.
[2] Not recognized by Bulgaria and Greece.

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1358–1367,
Singapore, 6-7 August 2009.

rich) Czech and Slovak are another possible pair of candidates. As we will see below, even such resource-rich languages like Spanish and Portuguese can benefit from the proposed method. Of course, many pairs of closely related languages can be also found outside of Europe, Malay and Indonesian being just one such example we will experiment with.

The remainder of the present paper is organized as follows: Section 2 presents our method, Section 3 describes the experiments, and Section 4 discusses the results and the general applicability of the approach. Section 5 provides an overview of the related work. Finally, Section 6 concludes and suggests possible directions for future work.

2 Method

We propose a novel language-independent approach for improving statistical machine translation for resource-poor languages by exploiting their similarity to resource-rich ones. More precisely, we improve the translation from a resource-poor source language X_1 into a resource-rich target language Y given a bi-text containing a limited number of parallel sentences for X_1-Y and a much larger bi-text for X_2-Y for some resource-rich language X_2 that is closely related to X_1.

Our method exploits the similarity between related languages with respect to word order, syntax, and, most importantly, vocabulary overlap – related languages share a large number of cognates.

Before we present the method, we will describe two simple strategies for integrating the bi-text for X_2-Y into a phrase-based SMT system for X_1-Y.

2.1 Merging Bi-texts

We can simply concatenate the bi-texts for X_1-Y and X_2-Y into one large bi-text and use it to train an SMT system.

This offers several advantages. First, it can yield improved word alignments for the sentences that came from the X_1-Y bi-text, *e.g.*, since the additional sentences can provide new contexts for the rare words in that bi-text; rare words are hard to align, which could have a disastrous effect on the subsequent phrase extraction stage. Second, it can provide new source-language side translation options, thus increasing the lexical coverage and reducing the number of unknown words at translation time; it can also provide new useful non-compositional phrases on the source-language side, thus yielding more fluent translation output. Third, it can offer new target-language side phrases for known source phrases, which could improve fluency by providing more translation options for the language model (LM) to choose from. Fourth, bad phrases including words from X_2 that do not exist in X_1 will be effectively ignored at translation time since they could never possibly match the input, while bad new target-language translations still have the chance to be filtered out by the language model.

However, simple concatenation can be problematic. First, when concatenating the small bi-text for X_1-Y with the much larger one for X_2-Y, the latter will dominate during word alignment and phrase extraction, thus hugely influencing both lexical and phrase translation probabilities, which can yield poor performance. This can be counteracted by repeating the small bi-text several times so that the large one does not dominate. Second, since the bi-texts are merged mechanically, there is no way to distinguish between phrases extracted from the bi-text for X_1-Y (which should be good), from those coming from the bi-text for X_2-Y (whose quality might be questionable).

2.2 Combining Phrase Tables

An alternative way of making use of the additional bi-text for X_2-Y to train an improved SMT system for $X_1 \rightarrow Y$ is to build separate phrase tables from X_1-Y and X_2-Y, which can then be (a) used together, *e.g.*, as alternative decoding paths, (b) merged, *e.g.*, using one or more extra features to indicate the bi-text each phrase came from, or (c) interpolated, *e.g.*, using simple linear interpolation.

Building two separate phrase tables offers several advantages. First, the good phrases from the bi-text for X_1-Y are clearly distinguished from (or given a higher weight in the linear interpolation compared to) the potentially bad ones from the X_2-Y bi-text. Second, the lexical and the phrase translation probabilities are combined in a principled manner. Third, using an X_2-Y bi-text that is much larger than that for X_1-Y is not problematic any more. Fourth, as with bi-text merging, there are many additional source- and target-language phrases, which offer new translation options.

On the negative side, the opportunity is lost to obtain improved word alignments for the sentences in the X_1-Y bi-text.

2.3 Proposed Method

Taking into account the potential advantages and disadvantages of the above strategies, we propose a method that tries to get the best of both: (i) increased lexical coverage by using additional phrase pairs independently extracted from X_2-Y, and (ii) improved word alignments for X_1-Y by biasing the word alignment process with additional sentence pairs from X_2-Y (possibly also repeating X_1-Y several times). A detailed description of the method follows:

1. Build a bi-text B_{cat} that is a concatenation of the bi-texts for X_1-Y and X_2-Y. Generate word alignments for B_{cat}, extract phrases, and build a phrase table T_{cat}.

2. Build a bi-text B_{rep} from the X_1-Y bi-text repeated k times followed by one copy of the X_2-Y bi-text. Generate word alignments for B_{rep}, then truncate them, only keeping word alignments for one copy of the X_1-Y bi-text. Use these word alignments to extract phrases, and build a phrase table T_{rep_trunc}.

3. Produce a phrase table T_{merged} by combining T_{cat} and T_{rep_trunc}, giving priority to the latter, and use it in an $X_1 \rightarrow Y$ SMT system.

2.4 Transliteration

As we mentioned above, our method relies on the existence of a large number of *cognates* between related languages. While linguists define cognates as words derived from a common root[3] (Bickford and Tuggy, 2002), *computational* linguists typically ignore origin, defining them as words in different languages that are mutual translations and have a similar orthography (Bergsma and Kondrak, 2007; Mann and Yarowsky, 2001; Melamed, 1999). In this paper, we adopt the latter definition.

Cognates between related languages often exhibit minor spelling variations, which can be simply due to different rules of orthography, (*e.g.*, *senhor vs. señor* in Portuguese and Spanish), but often stem from real phonological differences. For example, the Portuguese suffix *-ção* corresponds to the Spanish suffix *-ción*, *e.g.*, *evolução* vs. *evolución*. Such correspondences can be quite frequent and thus easy to learn automatically[4]. Even more frequent can be the inflectional variations. For example, in Portuguese and Spanish respectively, verb endings like *-ou* vs. *-ó* (for 3rd person singular, simple past tense), *e.g.*, *visitou* vs. *visitó*, or *-ei* vs. *-é* (for 1st person singular, simple past tense), *e.g.*, *visitei* vs. *visité*.

If such systematic differences exist between the languages X_1 and X_2, it might be useful to learn and to use them as a pre-processing step in order to transliterate the X_2 side of the X_2-Y bi-text and thus increase its vocabulary overlap with the source language X_1.

We will describe our approach to automatic transliteration in more detail in Section 3.4 below.

3 Experiments

3.1 Language Pairs

We experimented with two language pairs: the closely related Malay and Indonesian and the more dissimilar Spanish and Portuguese.

Malay and Indonesian are mutually intelligible, but differ in pronunciation and vocabulary. An example follows[5]:

- **Malay:** *Semua manusia dilahirkan bebas dan samarata dari segi kemuliaan dan hak-hak.*

- **Indonesian:** *Semua orang dilahirkan merdeka dan mempunyai martabat dan hak-hak yang sama.*

Spanish and Portuguese also exhibit a noticeable degree of mutual intelligibility, but differ in pronunciation, spelling, and vocabulary. Unlike Malay and Indonesian, however, they also differ syntactically and have a high degree of spelling differences as demonstrated by the following examples[6]:

- **Spanish:** *Señora Presidenta, estimados colegas, lo que está sucediendo en Oriente Medio es una tragedia.*

- **Portuguese:** *Senhora Presidente, caros colegas, o que está a acontecer no Medio Oriente é uma tragédia.*

[3] E.g., Latin *tu*, Old English *thou*, Spanish *tú*, Greek *sú* and German *du* are all cognates meaning '2^{nd} person singular'.

[4] Not all such differences are systematic; many apply to a particular word only, *e.g.*, *kerana vs. kareña* in Malay and Indonesian, or *dizer vs. decir* in Portuguese and Spanish.

[5] In English: *All human beings are born free and equal in dignity and rights.* (from Article 1 of the Universal Declaration of Human Rights)

[6] In English: *Madam President, ladies and gentlemen, the events in the Middle East are a real tragedy.*

3.2 Datasets

In our experiments, we used the following number of training sentence pairs (number of words shown in parentheses) for English (en), Indonesian (in), Malay (ml), Portuguese(pt), and Spanish (es):

- **Indonesian-English (*in-en*):**
  - 28,383 pairs (0.8M, 0.9M words);
  - monolingual English en_{in}: 5.1M words.
- **Malay-English (*ml-en*):**
  - 190,503 pairs (5.4M, 5.8M words);
  - monoling. English en_{ml}: 27.9M words.
- **Spanish-English (*es-en*):**
  - 1,240,518 pairs (35.7M, 34.6M words);
  - monolingual English $en_{es:pt}$: 45.3M words (the same as for *pt-en*).
- **Portuguese-English (*pt-en*):**
  - 1,230,038 pairs (35.9M, 34.6M words).
  - monolingual English $en_{es:pt}$: 45.3M words (the same as for *es-en*).

All of the above datasets contain sentences with up to 100 tokens. In addition, for each of the four language pairs, we have a development and a testing bi-text, each with 2,000 parallel sentence pairs. We made sure the development and the testing bi-texts shared no sentences with the training bi-texts; we further excluded from the monolingual English data all sentences from the English sides of the training and the development bi-texts.

The training bi-text datasets for *es-en* and *pt-en* were built from release v.3 of the *Europarl* corpus, excluding the Q4/2000 portion out of which we created our testing and development datasets.

We built the *in-en* bi-texts from texts that we downloaded from the Web. We translated the Indonesian texts to English using *Google Translate*, and we matched[7] them against the English texts using a cosine similarity measure and heuristic constraints based on document length in words and in sentences, overlap of numbers, words in uppercase, and words in the title. Next, we extracted pairs of sentences from the matched document pairs using *competitive linking* (Melamed, 2000), and we retained the ones whose similarity was above a pre-specified threshold. The *ml-en* was built in a similar manner.

[7]Note that the automatic translations were used for matching only; the final bi-text contained no automatic translations.

3.3 Baseline SMT System

In the baseline, we used the following setup: We first tokenized and lowercased both sides of the training bi-text. We then built separate directed word alignments for English$\rightarrow X$ and $X\rightarrow$English ($X\in\{$Indonesian, Spanish$\}$) using IBM model 4 (Brown et al., 1993), combined them using the *intersect+grow heuristic* (Och and Ney, 2003), and extracted phrase-level translation pairs of maximum length seven using the *alignment template approach* (Och and Ney, 2004). We thus obtained a phrase table where each pair is associated with five parameters: forward and reverse phrase translation probabilities, forward and reverse lexical translation probabilities, and phrase penalty.

We then trained a log-linear model using standard SMT feature functions: trigram language model probability, word penalty, distance-based[8] distortion cost, and the parameters from the phrase table. We set all weights by optimizing Bleu (Papineni et al., 2002) using minimum error rate training (MERT) (Och, 2003) on a separate development set of 2,000 sentences (Indonesian or Spanish), and we used them in a beam search decoder (Koehn et al., 2007) to translate 2,000 test sentences (Indonesian or Spanish) into English. Finally, we detokenized the output, and we evaluated it against a lowercased gold standard using Bleu[9].

3.4 Transliteration

As was mentioned in Section 2, transliteration can be helpful for languages with regular spelling differences. Thus, we built a system for transliteration from Portuguese into Spanish that was trained on a list of automatically extracted likely cognates. The system was applied on the Portuguese side of the *pt-en* training bi-text.

Classic approaches to automatic cognate extraction look for non-stopwords with similar spelling that appear in parallel sentences in a bi-text (Kondrak et al., 2003). In our case, however, we need to extract cognates between Spanish and Portuguese given *pt-en* and *es-en* bi-texts only, *i.e.*, without having a *pt-es* bi-text. Although it is easy to construct a *pt-es* bi-text from the Europarl corpus, we chose not to do so since, in general, synthe-

[8]We also tried lexicalized reordering (Koehn et al., 2005). While it yielded higher absolute Bleu scores, the relative improvement for a sample of our experiments was very similar to that achieved with distance-based re-ordering.

[9]We used version 11b of the NIST scoring tool: `http://www.nist.gov/speech/tools/`

sizing a bi-text for X_1-X_2 would be impossible: *e.g.*, it cannot be done for *ml-in* given our training datasets for *in-en* and *ml-en* since the English sides of these bi-texts have no sentences in common.

Thus, we extracted the list of likely cognates between Portuguese and Spanish from the training *pt-en* and *es-en* bi-texts using English as a pivot as follows: We started with IBM model 4 word alignments, from which we extracted four conditional lexical translation probabilities: $\Pr(p_j|e_i)$ and $\Pr(e_i|p_j)$ for Portuguese-English, and $\Pr(s_k|e_i)$ and $\Pr(e_i|s_k)$ for Spanish-English, where p_j, e_i and s_k stand for a Portuguese, an English and a Spanish word respectively. Following Wu and Wang (2007), we then induced conditional lexical translation probabilities $\Pr(p_j|s_k)$ and $\Pr(s_k|p_j)$ for Portuguese-Spanish as follows:

$$\Pr(p_j|s_k) = \sum_i \Pr(p_j|e_i, s_k)\Pr(e_i|s_k)$$

Assuming p_j is conditionally independent of s_k given e_i, we can simplify the above expression:

$$\Pr(p_j|s_k) = \sum_i \Pr(p_j|e_i)\Pr(e_i|s_k)$$

Similarly, for $\Pr(s_k|p_j)$, we obtain

$$\Pr(s_k|p_j) = \sum_i \Pr(s_k|e_i)\Pr(e_i|p_j)$$

We excluded all stopwords, words of length less than three, and those containing digits. We further calculated $\mathrm{Prod}(p_j, s_k) = \Pr(p_j|s_k)\Pr(s_k|p_j)$, and we excluded all Portuguese-Spanish word pairs (p_j, s_k) for which $\mathrm{Prod}(p_j, s_k) < 0.01$. From the remaining pairs, we extracted likely cognates based on $\mathrm{Prod}(p_j, s_k)$ and on the orthographic similarity between p_j and s_k.

Following Melamed (1995), we measured the orthographic similarity using the *longest common subsequence ratio* (LCSR), defined as follows:

$$\mathrm{LCSR}(s_1, s_2) = \frac{|\mathrm{LCS}(s_1,s_2)|}{\max(|s_1|,|s_2|)}$$

where $\mathrm{LCS}(s_1, s_2)$ is the *longest common subsequence* of s_1 and s_2, and $|s|$ is the length of s.

We retained as likely cognates all pairs for which LCSR was 0.58 or higher; that value was found by Kondrak et al. (2003) to be optimal for a number of language pairs in the *Europarl* corpus.

Finally, we performed *competitive linking* (Melamed, 2000), assuming that each Portuguese wordform had at most one Spanish best cognate match. Thus, using the values of $\mathrm{Prod}(p_j, s_k)$, we induced a fully-connected weighted bipartite graph. Then, we performed a greedy approximation to the maximum weighted bipartite matching in that graph (*i.e.*, competitive linking) as follows: First, we accepted as cognates the cross-lingual pair (p_j, s_k) with the highest $\mathrm{Prod}(p_j, s_k)$ in the graph, and we discarded p_j and s_k from further consideration. Then, we accepted the next highest-scored pair, and we discarded the involved wordforms and so forth. The process was repeated until there were no matchable pairs left.

As a result of the above procedure, we ended up with 28,725 Portuguese-Spanish cognate pairs, 9,201 (or 32%) of which had spelling differences. For each pair in the list of cognate pairs, we added spaces between any two adjacent letters for both wordforms, and we further appended the start and the end characters ^ and $. For example, the cognate pair *evolução – evolución* became

^ e v o l u ç ã o $ — ^ e v o l u c i ó n $

We randomly split the resulting list into a training (26,725 pairs) and a development dataset (2,000 pairs), and trained and tuned a character-level phrase-based monotone SMT system similar to (Finch and Sumita, 2008) to transliterate a Portuguese wordform into a Spanish wordform. We used a Spanish language model trained on 14M word tokens (obtained from the above-mentioned 45.3M-token monolingual English corpus after excluding punctuation, stopwords, words of length less than three, and those containing digits): one per line and character-separated with added start and end characters as in the above example. We set both the maximum phrase length and the language model order to ten; this value was found by tuning on the development dataset. The system was tuned using MERT, and the feature weights were saved. The tuning Bleu was 95.22%, while the baseline Bleu, for leaving the Portuguese words intact, was 87.63%. Finally, the training and the tuning datasets were merged, and a new training round was performed. The resulting system was used with the saved feature weights to transliterate the Portuguese side of the training *pt-en* bi-text, which yielded a new pt_{es}-*en* training bi-text.

We did the same for Malay into Indonesian. We extracted 5,847 cognate pairs, 844 (or 14.4%) of which had spelling differences, and we trained a transliteration system. The highest tuning Bleu was 95.18% (for maximum phrase size and LM order of 10), but the baseline was 93.15%. We then re-trained the system on the combination of the training and the development datasets, and we transliterated the Malay side of the training *ml-en* bi-text, obtaining a new ml_{in}-*en* training bi-text.

#	Train	LM	Dev	Test	10K	20K	40K	80K	160K	320K	640K	1230K
1	ml-en	en_{ml}	ml-en	ml-en	44.93	46.98	47.15	48.04	49.01	–	–	–
2	ml_{in}-en	en_{ml}	ml-en	ml-en	38.99	40.96	41.02	41.88	42.81	–	–	–
3	ml-en	en_{ml}	ml-en	**in-en**	13.69	14.58	14.76	15.12	15.84	–	–	–
4	ml-en	en_{ml}	**in-en**	**in-en**	13.98	14.75	14.91	15.51	16.27	–	–	–
5	ml-en	**en**$_{in}$	**in-en**	**in-en**	15.56	16.38	16.52	17.04	17.90	–	–	–
6	ml_{in}-en	**en**$_{in}$	**in-en**	**in-en**	16.44	17.36	17.62	18.14	19.15	–	–	–
7	pt-en	$\text{en}_{es:pt}$	pt-en	pt-en	21.28	23.11	24.43	25.72	26.43	27.10	27.78	27.96
8	pt_{es}-en	$\text{en}_{es:pt}$	pt-en	pt-en	10.91	11.56	12.16	12.50	12.83	13.27	13.48	13.71
9	pt-en	**en**$_{es:pt}$	pt-en	**es-en**	4.40	4.77	4.57	5.02	4.99	5.32	5.08	5.34
10	pt-en	**en**$_{es:pt}$	**es-en**	**es-en**	4.91	5.12	5.64	5.82	6.35	6.87	6.44	7.10
11	pt_{es}-en	**en**$_{es:pt}$	**es-en**	**es-en**	8.18	9.03	9.97	10.66	11.35	12.26	12.69	13.79

Table 1: **Cross-lingual SMT experiments (shown in bold).** Columns 2-5 present the bi-texts used for training, development and testing, and the monolingual data used to train the English language model. The following columns show the resulting Bleu (in %s) for different numbers of training sentence pairs.

3.5 Cross-lingual Translation

In this subsection, we study the similarity between the original and the additional source languages.

First, we measured the vocabulary overlap between Spanish and Portuguese, which was feasible since our training *pt-en* and *es-en* bi-texts are from the same time span in the *Europarl* corpus and their English sides largely overlap. We found 110,053 Portuguese and 121,444 Spanish word types, and 44,461 (or 36.6%) of them were identical. Unfortunately, we could not do the same for Malay and Indonesian since the English sides of the *in-en* and *ml-en* bi-texts do not overlap.

Second, following the setup of the baseline system, we performed cross-lingual experiments. The results are shown in Table 1. As we can see, this yielded a huge decrease in Bleu compared to the baseline – three to five times – even for very large training datasets, and even when a proper English LM and development dataset were used: compare line 1 to lines 3-6, and line 7 to lines 9-11.

Third, we tried transliteration. Bleu doubled for Spanish (see lines 10-11), but improved far less for Indonesian (lines 5-6). Training on the transliterated data and testing on Malay and Portuguese yielded about 10-12% relative decrease for Malay (lines 1-2) but 50% for Portuguese (lines 7-8).[10] Thus, unlike Spanish and Portuguese, there were far less systematic spelling variations between Malay and Indonesian. A closer inspection confirmed this: many extracted likely Malay-Indonesian cognate pairs with spelling differences were in fact forms of a word existing in both languages, *e.g.*, *kata* and *berkata* ('*to say*').

[10]However, as lines 8 and 11 show, a system trained on 1.23M pt_{es}*-en* sentence pairs, performs equally well when translating Portuguese and Spanish text: 13.71% *vs.* 13.79%.

3.6 Using an Additional Language

We performed various experiments combining the original and an additional training bi-text:

Two-tables: We built two separate phrase tables for the two bi-texts, and we used them in the alternative decoding path model of Birch et al. (2007).

Interpolation: We built two separate phrase tables for the original and for the additional bi-text, and we used linear interpolation to combine the corresponding conditional probabilities: $\Pr(e|s) = \alpha \Pr_{orig}(e|s) + (1-\alpha)\Pr_{extra}(e|s)$. We optimized the value of α on the development dataset, trying .5, .6, .7, .8 and .9; we used the same α for all four conditional probabilities.

Merge: We built separate phrase tables, T_{orig} and T_{extra}, for the original and for the additional training bi-text. We then concatenated them giving priority to T_{orig}: We kept all phrase pairs from T_{orig}, adding to them those ones from T_{extra} that were not present in T_{orig}. For each phrase pair added, we retained its associated conditional probabilities and the phrase penalty. We further added three additional features to each entry in the new table: F_1, F_2 and F_3. The value of F_1 was 1 if the phrase pair came from T_{orig}, and 0.5 otherwise. Similarly, F_2=1 if the phrase pair came from T_{extra}, and F_2=0.5 otherwise. The value of F_3 was 1 if the pair came from both T_{orig} and T_{extra}, and 0.5 otherwise. We experimented using (1) F_1 only, (2) F_1 and F_2, (3) F_1, F_2, and F_3. We set all feature weights using MERT, and we optimized the number of features on the development set.[11]

[11]In theory, we should have also re-normalized the probabilities since they may not sum to one. In practice, this was not that important since the log-linear SMT model does not require that the features be probabilities at all (*e.g.*, the phrase penalty), and we had extra features whose impact was bigger.

Concat$\times k$: We concatenated k copies of the original and one copy of the additional training bi-text; we then trained and tuned an SMT system as for the baseline. The value for k was optimized on the development dataset.

Concat$\times k$**:align**: We concatenated k copies of the original and one copy of the additional training bi-text. We then generated IBM model 4 word alignments, and we truncated them, only keeping them for one copy of the original training bi-text. Next, we extracted phrase pairs, thus buildng a phrase table, and we tuned an SMT system as for the baseline.

Our Method: Our method was described in Section 2. We used **merge** to combine the phrase tables for **concat**$\times k$**:align** and **concat**$\times$**1**, considering the former as T_{orig} and the latter as T_{extra}. We had two parameters to tune: k and the number of extra features in the merged phrase table.

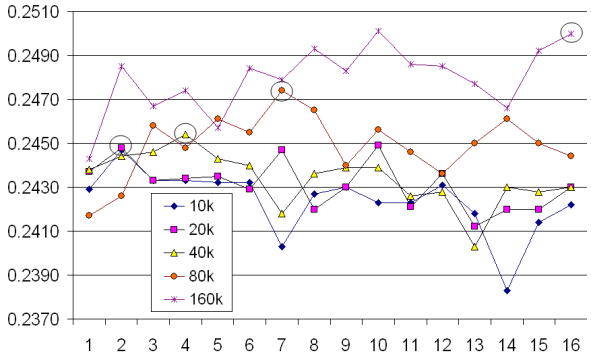

Figure 1: **Impact of k on Bleu for *concat*$\times k$ for different number of extra *ml-en* sentence pairs in Indonesian→English SMT.**

4 Results and Discussion

First, we studied the impact of k on **concat**$\times k$ for Indonesian→English SMT using Malay as an additional language. We tried all values of k such that $1 \leq k \leq 16$ with $10000n$ extra *ml-en* sentence pairs, $n \in \{1,2,4,8,16\}$. As we can see in Figure 1, the highest Bleu scores are achieved for $(n;k) \in \{(1;2),(2;2),(4;4),(8;7),(16;16)\}$, *i.e.*, when $k \approx n$. In order to limit the search space, we used this relationship between k and n in our experiments (also for Portuguese and Spanish).

Table 2 shows the results for experiments on improving Indonesian→English SMT using 10K, 20K, ..., 160K additional *ml-en* pairs of parallel sentences. Several observations can be made. First, using more additional sentences yields better results. Second, with one exception, all experiments yield improvements over the baseline. Third, the improvements are always statistically significant for **our method**, according to (Collins et al., 2005)'s sign test. Overall, among the different bi-text combination strategies, **our method** performs best, followed by **concat**$\times k$, **merge**, and **interpolate**, which are very close in performance; these three strategies are the only ones to consistently yield higher Bleu as the number of additional *ml-en* sentence pairs grows. Methods like **concat**$\times$1, **concat**$\times k$**:align** and **two-tables** are somewhat inconsistent in that respect. The latter method performs worst and is the only one to go below the baseline (for 10K *ml-en* pairs).

Table 3 shows the results when using *pt-en* data to improve Spanish→English SMT. Overall, the results and the conclusions that can be made are consistent with those for Table 2. We can further observe that, as the size of the original bi-text increases, the gain in Bleu decreases, which is to be expected. Note also that here transliteration is very important: it doubles the absolute gain in Bleu.

Finally, Table 4 shows a comparison to the pivoting technique of Callison-Burch et al. (2006). for English→Spanish SMT. Despite using just Portuguese, we achieve an improvement that is, in five out of six cases, much better than what they achieve with *eight* pivot languages (which include not only Portuguese, but also two other Romance languages, French and Italian, which are closely related to Spanish). Moreover, our method yields improvements for very large original datasets – 1.2M pairs, while theirs stops improving at 160K. However, our improvements are only statistically significant for 160K original pairs or less. Finally, note that our translation direction is reversed.

Based on the experimental results, we can make several conclusions. First, we have shown that using bi-text data from related languages improves SMT: we achieved up to 1.35 and 3.37 improvement in Bleu for *in-en* (+*ml-en*) and *es-en* (+*pt-en*) respectively. Second, while simple concatenation can help, it is problematic when the additional sentences out-number the ones from the original bi-text. Third, concatenation can work very well if the original bi-text is repeated enough times so that the additional bi-text does not dominate. Fourth, merging phrase tables giving priority to the original bi-text and using additional fea-

in-en	*ml-en*	Baseline	Two tables	Interpol.	Merge	concat×1	concat×k	concat×k:align	Our method
28.4K	10K	23.80$^{<}$	$^{\leq}$23.79$^{<}$	23.89$^{<}_{(.9)}$	23.97$^{<}_{(3)}$	24.29$^{<}$	24.29$^{<}_{(1)}$	24.01$^{<}_{(1)}$	$^{<}$**24.51**$_{(2;1)}$ (+0.72)
28.4K	20K	23.80$^{<}$	24.24$^{<}$	24.22$^{<}_{(.8)}$	$^{\leq}$24.46$^{<}_{(3)}$	24.37$^{<}$	$^{\leq}$24.48$_{(2)}$	$^{<}$24.35$^{<}_{(2)}$	$^{<}$**24.70**$_{(2;2)}$ (+0.90)
28.4K	40K	23.80$^{<}$	24.27$^{<}$	24.27$^{<}_{(.8)}$	24.43$^{\leq}_{(3)}$	24.38$^{\leq}$	$^{\leq}$24.54$_{(4)}$	$^{<}$24.39$^{<}_{(4)}$	$^{<}$**24.73**$_{(4;2)}$ (+0.93)
28.4K	80K	23.80$^{<}$	24.11$^{<}$	$^{\leq}$24.46$^{<}_{(.8)}$	$^{<}$24.67$_{(3)}$	24.17$^{<}$	$^{\leq}$24.65$^{<}_{(8)}$	24.18$^{<}_{(8)}$	$^{<}$**24.97**$_{(8;3)}$ (+1.17)
28.4K	160K	23.80$^{<}$	$^{<}$24.58$^{<}$	$^{<}$24.58$^{<}_{(.8)}$	$^{<}$24.79$^{\leq}_{(3)}$	$^{\leq}$24.43$^{<}$	$^{<}$25.00$_{(16)}$	$^{\leq}$24.27$^{<}_{(16)}$	$^{<}$**25.15**$_{(16;3)}$ (+1.35)

Table 2: **Improving Indonesian→English SMT using *ml-en* data.** Shown are the Bleu scores (in %s) for different methods. A subscript shows the best parameter value(s) found on the development set and used on the test set to produce the given result. Bleu scores that are statistically significantly better than the baseline/our method are marked on the left/right side by $^{<}$ (for $p < 0.01$) or $^{\leq}$ (for $p < 0.05$).

es-en	*pt-en*	Transl.	Baseline	Two tables	Interpol.	Merge	concat×1	concat×k	concat×k:align	Our method
10K	160K	no	22.87$^{<}$	$^{<}$23.81	$^{<}$23.73$_{(.5)}$	$^{<}$23.60$_{(2)}$	$^{<}$23.54$^{<}$	$^{<}$23.83$^{<}_{(16)}$	22.93$^{<}_{(16)}$	$^{<}$**23.98**$_{(16;3)}$ (+1.11)
		yes	22.87$^{<}$	$^{<}$25.29$^{\leq}$	$^{<}$25.22$^{<}_{(.5)}$	$^{<}$25.16$^{<}_{(2)}$	$^{<}$25.26	$^{<}$25.42$_{(16)}$	$^{<}$23.31$^{<}_{(16)}$	$^{<}$**25.73**$_{(16;3)}$ (+2.86)
20K	160K	no	24.71$^{<}$	$^{<}$25.22	$^{\leq}$25.02$^{<}_{(.5)}$	$^{<}$25.32$^{\leq}_{(3)}$	$^{<}$25.19$^{<}$	$^{<}$25.29$^{<}_{(8)}$	24.91$^{<}_{(8)}$	$^{<}$**25.65**$_{(8;2)}$ (+0.94)
		yes	24.71$^{<}$	$^{<}$26.07$^{\leq}$	$^{<}$26.07$_{(.7)}$	$^{<}$26.04$^{<}_{(3)}$	$^{<}$26.16$^{\leq}$	$^{<}$26.18$^{\leq}_{(8)}$	24.88$^{<}_{(8)}$	$^{<}$**26.36**$_{(8;3)}$ (+1.65)
40K	160K	no	25.80$^{<}$	25.96$^{<}$	26.15$^{<}_{(.6)}$	25.99$^{<}_{(3)}$	26.24$^{<}$	25.92$^{<}_{(4)}$	25.99$^{<}_{(4)}$	$^{<}$**26.49**$_{(4;2)}$ (+0.69)
		yes	25.80$^{<}$	$^{<}$26.68	$^{<}$26.43$_{(.7)}$	$^{<}$26.64$_{(3)}$	$^{<}$26.78	$^{<}$26.93$_{(4)}$	25.88$^{<}_{(4)}$	$^{<}$**26.95**$_{(4;3)}$ (+1.15)
80K	160K	no	27.08$^{\leq}$	$^{\geq}$26.89$^{<}$	27.04$^{<}_{(.8)}$	27.02$^{<}_{(3)}$	27.23	27.09$^{<}_{(2)}$	27.01$^{<}_{(2)}$	$^{\leq}$**27.30**$_{(2;2)}$ (+0.22)
		yes	27.08$^{<}$	27.20$^{<}$	27.42$_{(.5)}$	27.29$^{\leq}_{(3)}$	27.26$^{<}$	$^{\leq}$27.53$_{(2)}$	27.09$^{<}_{(2)}$	$^{<}$**27.49**$_{(2;3)}$ (+0.41)
160K	160K	no	27.90	27.99	27.72$_{(.5)}$	27.95$_{(2)}$	27.83$^{<}$	27.83$^{<}_{(1)}$	27.94$_{(1)}$	**28.05**$_{(1;3)}$ (+0.15)
		yes	27.90	28.11	$^{\leq}$28.13$_{(.6)}$	$^{\leq}$28.17$_{(2)}$	$^{\leq}$28.14	$^{\leq}$28.14$_{(1)}$	28.06$_{(1)}$	**28.16**$_{(1;2)}$ (+0.26)

Table 3: **Improving Spanish→English SMT using 160K additional *pt-en* sentence pairs.** Column three shows whether transliteration was used; the following columns list the Bleu scores (in %s) for different methods. A small subscript shows the best parameter value(s) found on the development set and used on the test set to produce the given result. Bleu scores that are statistically significantly better than the baseline/our method are marked on the left/right side by $^{<}$ (for $p < 0.01$) or $^{\leq}$ (for $p < 0.05$).

tures is a good strategy. Fifth, part of the improvement when combining bi-texts is due to increased vocabulary coverage because of cognates, but another part comes from improved word alignments. Sixth, the best results are achieved when the latter two sources are first isolated and then combined (our method). Finally, transliteration can help a lot in case of systematic spelling variations between the original and the additional source languages.

5 Related Work

In this section, we describe two general lines of related previous research: using cognates between the source and the target language, and source-language side paraphrasing with a pivot language.

5.1 Cognates

Many researchers have used likely cognates to obtain improved word alignments and thus build better SMT systems. Al-Onaizan et al. (1999) extracted such likely cognates for Czech-English using one of the variations of LCSR (Melamed, 1995) described in (Tiedemann, 1999) as a similarity measure. They used these cognates to improve word alignments with IBM models 1-4 in three different ways: (1) by seeding the parameters of IBM model 1, (2) by constraining the word co-occurrences when training IBM models 1-4, and (3) by adding the cognate pairs to the bi-text as additional "sentence pairs". The last approach performed best and was later used by Kondrak et al. (2003) who demonstrated improved SMT for nine European languages.

Unlike these approaches, which extract cognates between the source and the target language, we use cognates between the source and some other related language that is different from the target. Moreover, we only implicitly rely on the existence of such cognates; we do not try to extract them at all, and we leave them in their original sentence contexts.[12]

[12]However, in some of our experiments, we extract cognates for training a transliteration system from the resource-rich source language X_2 into the resource-poor one X_1.

Direction	System	10K	20K	40K	80K	160K	320K	1,230K
en→es	*baseline*	22.6	25.0	26.5	26.5	28.7	30.0	–
	pivoting (+8 languages × ∼1.3M pairs)	23.3	26.0	27.2	28.0	29.0	30.0	–
	improvement	**+0.7**	**+1.0**	**+0.7**	**+1.5**	**+0.3**	**+0.0**	–
es→en	baseline	22.87	24.71	25.80	27.08	27.90	28.46	29.90
	our method (+1 language × **160K** pairs)	23.98*	25.65*	26.49*	27.30$^\diamond$	28.05	28.52	29.87
	improvement	**+1.11***	**+0.94***	**+0.69***	**+0.22**$^\diamond$	**+0.15**	**+0.06**	**-0.03**
	our method (**translit.**, +1 lang. × **160K**)	25.73*	26.36*	26.95*	27.49*	28.16	28.43	29.94
	improvement	**+2.86***	**+1.65***	**+1.15***	**+0.41***	**+0.26**	**-0.03**	**+0.04**
	our method (+1 language × **1.23M** pairs)	24.23*	25.70*	26.78*	27.49	28.22$^\diamond$	28.58	29.84
	improvement	**+1.36***	**+0.99***	**+0.98***	**+0.41**	**+0.32**$^\diamond$	**+0.12**	**-0.06**
	our method (**translit.**, +1 lang. × **1.23M**)	26.24*	26.82*	27.47*	27.85*	28.50*	28.70	29.99
	improvement	**+3.37***	**+2.11***	**+1.67***	**+0.77***	**+0.60***	**+0.24**	**+0.09**

Table 4: **Comparison to the pivoting technique of Callison-Burch et al. (2006) for English→Spanish.** Shown are Bleu scores (in %s) and absolute improvement over the baseline for training bi-texts with different numbers of parallel sentences (10K, 20K, . . ., 1230K) and fixed amount of additional data: (1) about 1.3M sentence pairs for each of eight additional languages in Callison-Burch et al. (2006)'s pivoting, and (2) 160K and 1,230K pairs for one language (Portuguese) for our method. Statistically significant improvements over the baseline are marked with a * (for $p < 0.01$) and with a $^\diamond$ (for $p < 0.05$).

5.2 Paraphrasing with a Pivot-Language

Another relevant line of research is on using multilingual parallel corpora to improve SMT using additional languages as pivots.

Callison-Burch et al. (2006) improved English→Spanish and English→French SMT using source-language paraphrases extracted with the pivoting technique of Bannard and Callison-Burch (2005) and eight additional languages from the *Europarl corpus* (Koehn, 2005). For example, using German as a pivot, they extracted English paraphrases from a parallel English-German corpus by looking for English phrases that were aligned to the same German phrase: *e.g.*, if *under control* and *in check* were aligned to *unter controlle*, they were hypothesized to be paraphrases with some probability. Such (English) paraphrases were added as additional entries in the phrase table of an English→Spanish/English→French phrase-based SMT system and paired with the foreign (Spanish/French) translation of the original (English) phrase. The system was then tuned with MERT using an extra feature penalizing low-probability paraphrases; this yielded up to 1.8% absolute improvement in Bleu.

Other important publications about pivoting approaches for machine translation include (Wu and Wang, 2007), (Utiyama and Isahara, 2007), (Hajič et al., 2000) and (Habash and Hu, 2009).

Unlike pivoting, which can only improve source-language lexical coverage, we augment both the source- and the target-language sides. Second, while pivoting ignores context when extracting paraphrases, we take it into account. Third, by using as an additional language one that is related to the source, we are able to get increase in Bleu that is comparable and even better than what pivoting achieves with eight pivot languages. On the negative side, our approach is limited in that it requires that X_2 be related to X_1, while the pivoting language Z does not need to be related to X_1 nor to Y. However, we only need one additional parallel corpus (for X_2-Y), while pivoting needs two: one for X_1-Z and one for Z-Y. Finally, note that our approach is orthogonal to pivoting, and thus the two can be combined.

6 Conclusion and Future Work

We have proposed a novel method for improving SMT for resource-poor languages by exploiting their similarity to resource-rich ones.

In future work, we would like to extend that approach in several interesting directions. First, we want to make better use of multi-lingual parallel corpora, *e.g.*, while we had access to a Spanish-Portuguese-English corpus, we used it as two separate bi-texts Spanish-English and Portuguese-English. Second, we would like to exploit multiple auxiliary resource-rich languages the resource-poor source language is related to. Third, we could also experiment with using auxiliary languages that are related to the *target* language.

Acknowledgments

This research was supported by research grant POD0713875.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz Joseph Och, David Purdy, Noah Smith, and David Yarowsky. 1999. Statistical machine translation. Technical report, CLSP, Johns Hopkins University, Baltimore, MD.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL'05*, pages 597–604.

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of ACL'07*, pages 656–663.

Albert Bickford and David Tuggy. 2002. Electronic glossary of linguistic terms. http://www.sil.org/mexico/ling/glosario/E005ai-Glossary.htm.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG supertags in factored statistical machine translation. In *Proceedings of WMT'2007*, pages 9–16.

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL'06*, pages 17–24.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL'05*, pages 531–540.

Andrew Finch and Eiichiro Sumita. 2008. Phrase-based machine transliteration. In *Proceedings of WTCAST'08*, pages 13–18.

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proceedings of the WMT'09*, pages 173–181.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of ANLP'00*, pages 7–12.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of IWSLT'05*.

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 *Proceedings of ACL'07. Demonstration session*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for evaluation of machine translation. In *Proceedings of MT Summit*, pages 79–86.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of NAACL'03*, pages 46–48.

Gideon Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL'01*, pages 1–8.

Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *Proceedings of WVLC'95*, pages 184–198.

Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

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

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL'03*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL'02*, pages 311–318.

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 *Proceedings of LREC'2006*, pages 2142–2147.

Jorg Tiedemann. 1999. Automatic construction of weighted string similarity measures. In *Proceedings of EMNLP-VLC'99*, pages 213–219.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL-HLT'07*, pages 484–491.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.