Source Language Adaptation for Resource-Poor Machine Translation

Pidong Wang
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
wangpd@comp.nus.edu.sg

Preslav Nakov
QCRI
Qatar Foundation
Tornado Tower, P.O. 5825
Doha, Qatar
pnakov@qf.org.qa

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 machine translation from a resource-poor language to X by *adapting* a large bi-text for a related resource-rich language and X (the same target language). We assume a small bi-text for the resource-poor language to X pair, which we use to learn word-level and phrase-level paraphrases and cross-lingual morphological variants between the resource-rich and the resource-poor language; we then adapt the former to get closer to the latter. Our experiments for Indonesian/Malay–English translation show that using the large adapted resource-rich bi-text yields 6.7 BLEU points of improvement over the unadapted one and 2.6 BLEU points over the original small bi-text. Moreover, combining the small bi-text with the adapted bi-text outperforms the corresponding combinations with the unadapted bi-text by 1.5–3 BLEU points. We also demonstrate applicability to other languages and domains.

1 Introduction

Statistical machine translation (SMT) systems learn how to translate from large sentence-aligned bilingual corpora of human-generated translations, called *bi-texts*. Unfortunately, collecting sufficiently large, high-quality bi-texts is hard, and thus most of the 6,500+ world languages remain resource-poor. Fortunately, many of these resource-poor languages are related to some resource-rich language, with whom they overlap in vocabulary and share cognates, which offers opportunities for bi-text reuse.

Example pairs of such resource rich–poor languages include Spanish–Catalan, Finnish–Estonian, Swedish–Norwegian, Russian–Ukrainian, Irish–Gaelic Scottish, Standard German–Swiss German, Modern Standard Arabic–Dialectical Arabic (e.g., Gulf, Egyptian), Turkish–Azerbaijani, etc.

Previous work has already demonstrated the benefits of using a bi-text for a related resource-rich language to X (e.g., X=English) to improve machine translation from a resource-poor language to X (Nakov and Ng, 2009; Nakov and Ng, 2012). Here we take a different, orthogonal approach: we *adapt* the resource-rich language to get closer to the resource-poor one.

We assume a small bi-text for the resource-poor language, which we use to learn word-level and phrase-level paraphrases and cross-lingual morphological variants between the two languages. Assuming translation into the same target language X, we adapt (the source side of) a large training bi-text for a related resource-rich language and X.

Training on the adapted large bi-text yields very significant improvements in translation quality compared to both (a) training on the unadapted version, and (b) training on the small bi-text for the resource-poor language. We further achieve very sizable improvements when combining the small bi-text with the large adapted bi-text, compared to combining the former with the unadapted bi-text.

While we focus on adapting Malay to look like Indonesian in our experiments, we also demonstrate the applicability of our approach to another language pair, Bulgarian–Macedonian, which is also from a different domain.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 286–296, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

2 Related Work

One relevant line of research is on machine translation between closely related languages, which is arguably simpler than general SMT, and thus can be handled using word-for-word translation, manual language-specific rules that take care of the necessary morphological and syntactic transformations, or character-level translation/transliteration. This has been tried for a number of language pairs including Czech–Slovak (Hajič et al., 2000), Turkish–Crimean Tatar (Altintas and Cicekli, 2002), Irish–Scottish Gaelic (Scannell, 2006), and Bulgarian–Macedonian (Nakov and Tiedemann, 2012). In contrast, we have a different objective – we do not carry out full translation but rather adaptation since our ultimate goal is to translate into a third language X.

A special case of this same line of research is the translation between dialects of the same language, e.g., between Cantonese and Mandarin (Zhang, 1998), or between a dialect of a language and a standard version of that language, e.g., between some Arabic dialect (e.g., Egyptian) and Modern Standard Arabic (Bakr et al., 2008; Sawaf, 2010; Salloum and Habash, 2011). Here again, manual rules and/or language-specific tools are typically used. In the case of Arabic dialects, a further complication arises by the informal status of the dialects, which are not standardized and not used in formal contexts but rather only in informal online communities[1] such as social networks, chats, Twitter and SMS messages. This causes further mismatch in domain and genre.

Thus, translating from Arabic dialects to Modern Standard Arabic requires, among other things, normalizing informal text to a formal form. In fact, this is a more general problem, which arises with informal sources like SMS messages and Tweets for just any language (Aw et al., 2006; Han and Baldwin, 2011). Here the main focus is on coping with spelling errors, abbreviations, and slang, which are typically addressed using string edit distance, while also taking pronunciation into account. This is different from our task, where we try to adapt good, formal text from one language into another.

A second relevant line of research is on language adaptation and normalization, when done specifically for improving SMT into another language.

For example, Marujo et al. (2011) described a rule-based system for adapting Brazilian Portuguese (BP) to European Portuguese (EP), which they used to adapt BP–English bi-texts to EP–English. They report small improvements in BLEU for EP–English translation when training on the adapted "EP"–En bi-text compared to using the unadapted BP–En (38.55 vs. 38.29), or when an EP–English bi-text is used in addition to the adapted/unadapted one (41.07 vs. 40.91 BLEU). Unlike this work, which heavily relied on language-specific rules, our approach is statistical, and largely language-independent; moreover, our improvements are much more sizable.

A third relevant line of research is on reusing bi-texts between related languages without or with very little adaptation, which works well for very closely related languages. For example, our previous work (Nakov and Ng, 2009; Nakov and Ng, 2012) experimented with various techniques for combining a small bi-text for a resource-poor language (Indonesian or Spanish, pretending that Spanish is resource-poor) with a much larger bi-text for a related resource-rich language (Malay or Portuguese); the target language of all bi-texts was English. However, our previous work did not attempt language adaptation, except for very simple transliteration for Portuguese–Spanish that ignored context entirely; since it could not substitute one word for a completely different word, it did not help much for Malay–Indonesian, which use unified spelling. Still, once we have language-adapted the large bi-text, it makes sense to try to combine it further with the small bi-text; thus, below we will directly compare and combine these two approaches.

Another alternative, which we do not explore in this work, is to use cascaded translation using a pivot language (Utiyama and Isahara, 2007; Cohn and Lapata, 2007; Wu and Wang, 2009). Unfortunately, using the resource-rich language as a pivot (poor→rich→X) would require an additional parallel poor–rich bi-text, which we do not have. Pivoting over the target X (rich→X→poor) for the purpose of language adaptation, on the other hand, would miss the opportunity to exploit the relationship between the resource-poor and the resource-rich language; this would also be circular since the first step would ask an SMT system to translate its own training data (we only have one rich–X bi-text).

[1] The Egyptian Wikipedia is one notable exception.

3 Malay and Indonesian

Malay and Indonesian are closely related, mutually intelligible Austronesian languages with 180 million speakers combined. They have a unified spelling, with occasional differences, e.g., *kerana* vs. *karena* ('because'), *Inggeris* vs. *Inggris* ('English'), and *wang* vs. *uang* ('money').

They differ more substantially in vocabulary, mostly because of loan words, where Malay typically follows the English pronunciation, while Indonesian tends to follow Dutch, e.g., *televisyen* vs. *televisi*, *Julai* vs. *Juli*, and *Jordan* vs. *Yordania*.

While there are many cognates between the two languages, there are also a lot of false friends, e.g., *polisi* means *policy* in Malay but *police* in Indonesian. There are also many partial cognates, e.g., *nanti* means both *will* (future tense marker) and *later* in Malay but only *later* in Indonesian.

Thus, fluent Malay and fluent Indonesian can differ substantially. Consider, for example, Article 1 of the *Universal Declaration of Human Rights*:[2]

- *Semua manusia dilahirkan bebas dan samarata dari segi kemuliaan dan hak-hak. Mereka mempunyai pemikiran dan perasaan hati dan hendaklah bertindak di antara satu sama lain dengan semangat persaudaraan. (**Malay**)*
- *Semua orang dilahirkan merdeka dan mempunyai martabat dan hak-hak yang sama. Mereka dikaruniai akal dan hati nurani dan hendaknya bergaul satu sama lain dalam semangat persaudaraan. (**Indonesian**)*

There is only 50% overlap at the word level, but the actual vocabulary overlap is much higher, e.g., there is only one word in the Malay text that does not exist in Indonesian: *samarata* ('equal'). Other differences are due to the use of different morphological forms, e.g., *hendaklah* vs. *hendaknya* ('conscience'), derivational variants of *hendak* ('want').

Of course, word choice in translation is often a matter of taste. Thus, we asked a native speaker of Indonesian to adapt the Malay version to Indonesian while preserving as many words as possible:

- *Semua manusia dilahirkan bebas dan mempunyai martabat dan hak-hak yang sama. Mereka mempunyai pemikiran dan perasaan dan hendaklah bergaul satu sama lain dalam semangat persaudaraan.* (**Indonesian**)

Obtaining this latter version from the original Malay text requires three word-level operations: (1) deletion of *dari*, *segi*, (2) insertion of *yang*, *sama*, and (3) substitution of *samarata* with *mempunyai*.

Unfortunately, we do not have parallel Malay-Indonesian text, which complicates the process of learning when to apply these operations. Thus, below we restrict our attention to the simplest and most common operation of word substitution only, leaving the other two[3] operations for future work.

Note that word substitution is enough in many cases, e.g., it is all that is needed for the following Malay-Indonesian sentence pair:[4]

- *KDNK Malaysia dijangka cecah 8 peratus pada tahun 2010.*
- *PDB Malaysia akan mencapai 8 persen pada tahun 2010.*

4 Method

We improve machine translation from a resource-poor language (Indonesian) to English by *adapting* a bi-text for a related resource-rich language (Malay) and English, using *word-level* and *phrase-level* paraphrases and cross-lingual morphological variants.

4.1 Word-Level Paraphrasing

Given a Malay sentence, we generate a confusion network containing multiple Indonesian word-level paraphrase options for each Malay word. Each such Indonesian option is associated with a corresponding weight in the network, which is defined as the probability of this option being a translation of the original Malay word (see Eq. 1 below). We decode this confusion network using a large Indonesian language model, thus generating a ranked list of n corresponding adapted "Indonesian" sentences.

Then, we pair each such adapted "Indonesian" sentence with the English counter-part for the Malay sentence it was derived from, thus obtaining a synthetic "Indonesian"–English bi-text. Finally, we combine this synthetic bi-text with the original Indonesian–English one to train the final Indonesian–English SMT system.

Below we first describe how we generate word-level Indonesian options and corresponding weights for the Malay words. Then, we explain how we build, decode, and improve the confusion network.

[2] English: *All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.*

[3] There are other potentially useful operations, e.g., a correct translation for the Malay *samarata* can be obtained by splitting it into the Indonesian sequence *sama rata*.

[4] *Malaysia's GDP is expected to reach 8 percent in 2010.*

4.1.1 Inducing Word-Level Paraphrases

We use pivoting over English to induce potential Indonesian translations for a given Malay word.

First, we generate separate word-level alignments for the Indonesian–English and the Malay–English bi-texts. Then, we induce Indonesian-Malay word translation pairs assuming that if an Indonesian word i and a Malay word m are aligned to the same English word e, they could be mutual translations. Each translation pair is associated with a conditional probability, estimated by pivoting over English:

$$\Pr(i|m) \quad = \quad \sum_{e} \Pr(i|e)\Pr(e|m) \qquad (1)$$

$\Pr(i|e)$ and $\Pr(e|m)$ are estimated using maximum likelihood from the word alignments. Following (Callison-Burch et al., 2006), we further assume that i is conditionally independent of m given e.

4.1.2 Confusion Network Construction

Given a Malay sentence, we construct an Indonesian confusion network, where each Malay word is augmented with a set of network transitions: possible Indonesian word translations. The weight of such a transition is the conditional Indonesian-Malay translation probability as calculated by Eq. 1; the original Malay word is assigned a weight of 1.

Note that we paraphrase *each* word in the input Malay sentence as opposed to only those Malay words that we believe not to exist in Indonesian, e.g., because they do not appear in our Indonesian monolingual text. This is necessary because of the large number of false friends and partial cognates between Malay and Indonesian (see Section 3).

Finally, we decode the confusion network for a Malay sentence using a large Indonesian language model, and we extract an n-best list.[5] Table 1 shows the 10-best adapted "Indonesian" sentences[6] we generated for the confusion network in Figure 1.

4.1.3 Further Refinements

Many of our paraphrases are bad: some have very low probabilities, while others involve rare words for which the probability estimates are unreliable.

Moreover, the options we propose for a Malay word are inherently restricted to the small Indonesian vocabulary of the Indonesian–English bi-text. Below we describe how we address these issues.

Score-based filtering. We filter out translation pairs whose probabilities (Eq. 1) are lower than some threshold (tuned on the dev dataset), e.g., 0.01.

Improved estimations for $\Pr(i|e)$. We concatenate k copies of the Indonesian–English bi-text and one copy of the Malay–English bi-text, where the value of k is selected so that we have roughly the same number of Indonesian and Malay sentences. Then, we generate word-level alignments for the resulting bi-text. Finally, we truncate these alignments keeping them for one copy of the original Indonesian–English bi-text only. Thus, we end up with improved word alignments for the Indonesian–English bi-text, and with better estimations for Eq. 1. Since Malay and Indonesian share many cognates, this improves word alignments for Indonesian words that occur rarely in the small Indonesian–English bi-text but are relatively frequent in the larger Malay–English one; it also helps for some frequent words.

Cross-lingual morphological variants. We increase the Indonesian options for a Malay word using morphology. Since the set of Indonesian options for a Malay word in pivoting is restricted to the Indonesian vocabulary of the small Indonesian–English bi-text, this is a severe limitation of pivoting. Thus, assuming a large monolingual Indonesian text, we first build a lexicon of the words in the text. Then, we lemmatize these words using two different lemmatizers: the Malay lemmatizer of Baldwin and Awab (2006), and a similar Indonesian lemmatizer. Since these two analyzers have different strengths and weaknesses, we combine their outputs to increase recall. Next, we group all Indonesian words that share the same lemma, e.g., for *minum*, we obtain {*diminum, diminumkan, diminumnya, makan-minum, makananminuman, meminum, meminumkan, meminumnya, meminum-minuman, minum, minum-minum, minum-minuman, minuman, minumanku, minumannya, peminum, peminumnya, perminum, terminum*}. Since Malay and Indonesian are subject to the same morphological processes and share many lemmata, we use such groups to propose Indonesian translation options for a Malay word. We first lemmatize the target Malay word, and then we find all groups of Indonesian words the Malay lemmata belong to.

[5] For balance, in case of less than n adaptations for a Malay sentence, we randomly repeat some of the available ones.

[6] According to a native Indonesian speaker, options 1 and 3 in Table 1 are perfect adaptations, options 2 and 5 have a wrong word order, and the rest are grammatical though not perfect.

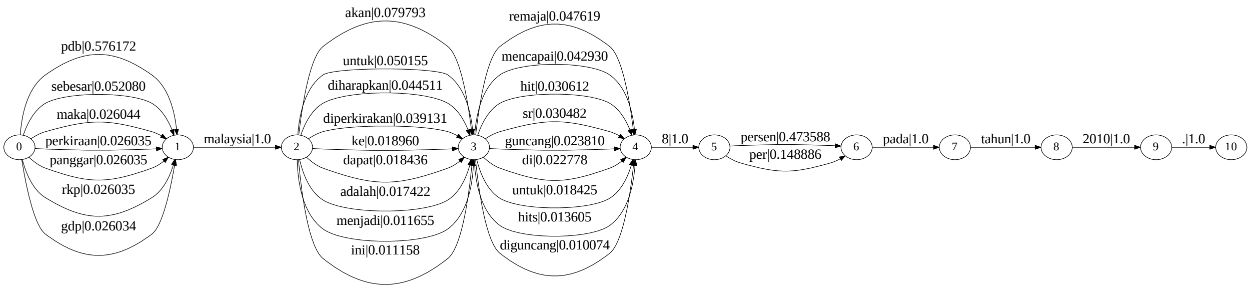

Figure 1: Indonesian confusion network for the Malay sentence "*KDNK Malaysia dijangka cecah 8 peratus pada tahun 2010.*" Arcs with scores below 0.01 are omitted, and words that exist in Indonesian are not paraphrased (for better readability).

Rank	**"Indonesian" Sentence**			
1	pdb	malaysia	akan	mencapai 8 persen pada tahun 2010 .
2	pdb	malaysia	untuk	mencapai 8 persen pada tahun 2010 .
3	pdb	malaysia	diperkirakan	mencapai 8 persen pada tahun 2010 .
4	maka	malaysia	akan	mencapai 8 persen pada tahun 2010 .
5	maka	malaysia	untuk	mencapai 8 persen pada tahun 2010 .
6	pdb	malaysia	dapat	mencapai 8 persen pada tahun 2010 .
7	maka	malaysia	diperkirakan	mencapai 8 persen pada tahun 2010 .
8	sebesar	malaysia	akan	mencapai 8 persen pada tahun 2010 .
9	pdb	malaysia	diharapkan	mencapai 8 persen pada tahun 2010 .
10	pdb	malaysia	ini	mencapai 8 persen pada tahun 2010 .

Table 1: The 10-best "Indonesian" sentences extracted from the confusion network in Figure 1.

The union of these groups is the set of morphological variants that we will add to the confusion network as additional options for the Malay word.[7] For example, given *seperminuman* ('drinking') in the Malay input, we first find its stem *minum*, and then we get the above example set of Indonesian words, which contains some reasonable substitutes such as *minuman* ('drink'). In the confusion network, the weight of the original Malay word is set to 1, while the weight of a morphological option is one minus the minimum edit distance ratio (Ristad and Yianilos, 1998) between it and the Malay word, multiplied by the highest probability for all pivoting variants for the Malay word.

4.2 Phrase-Level Paraphrasing

Word-level paraphrasing ignores context when generating Indonesian variants, relying on the Indonesian language model to make the right contextual choice. We also try to model context more directly by generating adaptation options at the *phrase level*.

Phrase-level paraphrase induction. We use standard phrase-based SMT techniques to build separate phrase tables for the Indonesian–English and the Malay–English bi-texts, where we have four conditional probabilities: forward/reverse phrase translation probability, and forward/reverse lexicalized phrase translation probability. We pivot over English to generate Indonesian-Malay phrase pairs, whose probabilities are derived from the corresponding ones in the two phrase tables using Eq. 1.

Cross-lingual morphological variants. While phrase-level paraphrasing models context better, it remains limited in the size of its Indonesian vocabulary by the small Indonesian–English bi-text, just like word-level paraphrasing was. We address this by transforming the sentences in the *development* and the *test* Indonesian–English bi-texts into confusion networks, where we add Malay morphological variants for the Indonesian words, weighting them as before. Note that we do not alter the training bi-text.

4.3 Combining Bi-texts

We combine the Indonesian–English and the synthetic "Indonesian"–English bi-texts as follows:

Simple concatenation. Assuming the two bi-texts are of comparable quality, we simply train an SMT system on their concatenation.

[7] While the different morphological forms typically have different meanings, e.g., *minum* ('drink') vs. *peminum* ('drinker'), in some cases the forms could have the same translation in English, e.g., *minum* ('drink', verb) vs. *minuman* ('drink', noun). This is our motivation for trying morphological variants, even though they are almost exclusively derivational, and thus quite risky as translational variants; see also (Nakov and Ng, 2011).

Balanced concatenation with repetitions. However, the two bi-texts are not directly comparable and are clearly not equally good as a source of training data for an Indonesian-English SMT system. For one thing, the "Indonesian"–English bi-text is obtained from n-best lists, i.e., it has exactly n very similar variants for each Malay sentence. Moreover, the original Malay–English bi-text is much larger in size than the Indonesian–English one, and now it has been further expanded n times in order to become an "Indonesian"–English bi-text, which means that it will dominate the concatenation due to its size. In order to counter-balance this, we repeat the smaller Indonesian–English bi-text enough times so that we can make the number of sentences it contains roughly the same as for the "Indonesian"–English bi-text; then we concatenate the two bi-texts and we train an SMT system on the resulting bi-text.

Sophisticated phrase table combination. Finally, we experiment with a method for combining phrase tables proposed in (Nakov and Ng, 2009; Nakov and Ng, 2012). The first phrase table is extracted from word alignments for the balanced concatenation with repetitions, which are then truncated so that they are kept for only one copy of the Indonesian–English bi-text. The second table is built from the simple concatenation. The two tables are then merged as follows: all phrase pairs from the first one are retained, and to them are added those phrase pairs from the second one that are not present in the first one. Each phrase pair retains its original scores, which are further augmented with 1–3 additional feature scores indicating its origin: the first/second/third feature is 1 if the pair came from the first/second/both table(s), and 0 otherwise. We experiment using all three, the first two, or the first feature only; we also try setting the features to 0.5 instead of 0. This makes the following six combinations (0, 00, 000, .5, .5.5, .5.5.5); on testing, we use the one that achieves the highest BLEU score on the development set.

Other possibilities for combining the phrase tables include using alternative decoding paths (Birch et al., 2007), simple linear interpolation, and direct phrase table merging with extra features (Callison-Burch et al., 2006); they were previously found inferior to the last two approaches above (Nakov and Ng, 2009; Nakov and Ng, 2012).

5 Experiments

We run two kinds of experiments: (a) *isolated*, where we train on the synthetic "Indonesian"–English bi-text only, and (b) *combined*, where we combine it with the Indonesian–English bi-text.

5.1 Datasets

In our experiments, we use the following datasets, normally required for Indonesian–English SMT:

- **Indonesian–English train bi-text (*IN2EN*):** 28,383 sentence pairs; 915,192 English tokens; 796,787 Indonesian tokens;
- **Indon.–English dev bi-text (*IN2EN-dev*):** 2,000 sentence pairs; 36,584 English tokens; 35,708 Indonesian tokens;
- **Indon.–English test bi-text (*IN2EN-test*):** 2,018 sentence pairs; 37,101 English tokens; 35,509 Indonesian tokens;
- **Monolingual English text (*EN-LM*):** 174,443 sentences; 5,071,988 English tokens.

We also use a Malay–English set (to be turned into "Indonesian"–English), and monolingual Indonesian text (for decoding the confusion network):

- **Malay–English train bi-text (*ML2EN*):** 290,000 sentence pairs; 8,638,780 English tokens; 8,061,729 Malay tokens;
- **Monolingual Indonesian text (*IN-LM*):** 1,132,082 sentences; 20,452,064 Indonesian tokens.

5.2 Baseline Systems

We build five baseline systems – two using a single bi-text, *ML2EN* or *IN2EN*, and three combining *ML2EN* and *IN2EN*, using simple concatenation, balanced concatenation, and sophisticated phrase table combination. The last combination is a very strong baseline and the most relevant one we need to improve upon.

5.3 Isolated Experiments

The isolated experiments only use the adapted "Indonesian"–English bi-text, which allows for a direct comparison to using *ML2EN* / *IN2EN* only.

5.3.1 Word-Level Paraphrasing

In our word-level paraphrasing experiments, we adapt Malay to Indonesian using three kinds of confusion networks (see Section 4.1.3 for details):

- ***CN:pivot*** – using word-level pivoting only;
- ***CN:pivot′*** – using word-level pivoting, with probabilities from word alignments for *IN2EN* that were improved using *ML2EN*;
- ***CN:pivot′+morph*** – *CN:pivot′* augmented with cross-lingual morphological variants.

There are two parameter values to be tuned on *IN2EN-dev* for the above confusion networks: (1) the minimum pivoting probability threshold for the Malay-Indonesian word-level paraphrases, and (2) the number of n-best Indonesian-adapted sentences that are to be generated for each input Malay sentence. We try $\{0.001, 0.005, 0.01, 0.05\}$ for the threshold and $\{1, 5, 10\}$ for n.

5.3.2 Phrase-Level Paraphrasing

In our phrase-level paraphrasing experiments, we use pivoted phrase tables (PPT) with the following features for each phrase table entry (in addition to the phrase penalty; see Section 4.2 for more details):

- ***PPT:1*** – only uses the forward conditional translation probability;
- ***PPT:4*** – uses all four conditional probabilities;
- ***PPT:4::CN:morph*** – *PPT:4* but used with a cross-lingual morphological confusion network for the dev/test Indonesian sentences.

Here we tune one parameter only: the number of n-best Indonesian-adapted sentences to be generated for each input Malay sentence; we try $\{1, 5, 10\}$.

5.4 Combined Experiments

These experiments assess the impact of our adaptation approach when combined with the original Indonesian–English bi-text *IN2EN* as opposed to combining *ML2EN* with *IN2EN* (as was in the last three baselines). We experiment with the same three combinations: simple concatenation, balanced concatenation, and sophisticated phrase table combination. We tune the parameters as before; for the last combination, we further tune the six extra feature combinations (see Section 4.3 for details).

6 Results and Discussion

For all tables, statistically significant improvements ($p < 0.01$), according to Collins et al. (2005)'s sign test, over the baseline are in **bold**; in case of two baselines, underline is used for the second baseline.

System	**BLEU**
ML2EN	14.50
IN2EN	18.67
Simple concatenation	**18.49**
Balanced concatenation	$\underline{\mathbf{19.79}}$
Sophisticated phrase table combination	$\underline{\mathbf{20.10}}_{(.5.5)}$

Table 2: **The five baselines.** The subscript indicates the parameters found on *IN2EN-dev* and used for *IN2EN-test*. The scores that are statistically significantly better than *ML2EN* and *IN2EN* ($p < 0.01$, Collins' sign test) are shown in **bold** and are underlined, respectively.

6.1 Baseline Experiments

The results for the baseline systems are shown in Table 2. We can see that training on *ML2EN* instead of *IN2EN* yields over 4 points absolute drop in BLEU (Papineni et al., 2002) score, even though *ML2EN* is about 10 times larger than *IN2EN* and both bi-texts are from the same domain. This confirms the existence of important differences between Malay and Indonesian. While simple concatenation does not help, balanced concatenation with repetitions improves by 1.12 BLEU points over *IN2EN*, which shows the importance of giving *IN2EN* a proper weight in the combined bi-text. This is further reconfirmed by the sophisticated phrase table combination, which yields an additional absolute gain of 0.31 BLEU points.

6.2 Isolated Experiments

Table 3 shows the results for the isolated experiments. We can see that word-level paraphrasing improves by up to 5.56 and 1.39 BLEU points over the two baselines (both statistically significant). Compared to *ML2EN*, *CN:pivot* yields an absolute improvement of 4.41 BLEU points, *CN:pivot′* adds another 0.59, and *CN:pivot′+morph* adds 0.56 more. The scores for TER (v. 0.7.25) and METEOR (v. 1.3) are on par with those for BLEU (NIST v. 13).

Table 3 further shows that the optimal parameters for the word-level SMT systems (*CN:**) involve a very low probability cutoff, and a high number of n-best sentences. This shows that they are robust to noise, probably because bad source-side phrases are unlikely to match the test-time input. Note also the effect of repetitions: good word choices are shared by many n-best sentences, and thus they would have higher probabilities compared to bad word choices.

		n-gram precision						
	System	**1-gr.**	**2-gr.**	**3-gr.**	**4-gr.**	**BLEU**	**TER**	**METEOR**
	ML2EN (baseline)	48.34	19.22	9.54	4.98	14.50	67.14	43.28
	IN2EN (baseline)	55.04	23.90	12.87	7.18	18.67	61.99	54.34
	CN:pivot	54.50	24.41	13.09	7.35	$\mathbf{18.91}^{(+4.41,+0.24)}_{(0.005,10best)}$	61.94	51.07
	CN:pivot′	55.05	25.09	13.60	7.69	$\underline{\mathbf{19.50}}^{(+5.00,+0.83)}_{(0.001,10best)}$	61.25	51.97
(i)	*CN:pivot′+morph*	55.97	25.73	14.06	7.99	$\underline{\mathbf{20.06}}^{(+5.56,+1.39)}_{(0.005,10best)}$	60.31	55.65
	PPT:1	55.11	25.04	13.66	7.80	$\underline{\mathbf{19.58}}^{(+5.08,+0.91)}_{(10best)}$	60.92	51.93
	PPT:4	56.64	26.20	14.53	8.40	$\underline{\mathbf{20.63}}^{(+6.13,+1.96)}_{(10best)}$	59.33	54.23
(ii)	*PPT:4::CN:morph*	56.91	26.53	14.76	8.55	$\underline{\mathbf{20.89}}^{(+6.39,+2.22)}_{(10best)}$	59.30	57.19
	System combination: (i) + (ii)	57.73	27.00	15.03	8.71	$\underline{\mathbf{21.24}}^{(+6.74,+2.57)}$	58.19	54.63

Table 3: **Isolated experiments.** The subscript shows the best tuning parameters, and the superscript shows the absolute test improvement over the *ML2EN* and the *IN2EN* baselines. The last line shows system combination results.

		Combining *IN2EN* with an adapted version of *ML2EN*		
	Combination with	**Simple Concatenation**	**Balanced Concatenation**	**Sophisticated Combination**
(i)	+ *ML2EN* (unadapted; baseline)	18.49	19.79	$20.10_{(.5.5)}$
	+ *CN:pivot*	$\mathbf{19.99}^{(+1.50)}_{(0.001,1best)}$	$20.16^{(+0.37)}_{(0.001,10best)}$	$20.32^{(+0.22)}_{(0.01,10best,.5.5)}$
	+ *CN:pivot′*	$\mathbf{20.03}^{(+1.54)}_{(0.05,1best)}$	$\mathbf{20.80}^{(+1.01)}_{(0.05,10best)}$	$\mathbf{20.55}^{(+0.45)}_{(0.05,10best,.5.5)}$
(ii)	+ *CN:pivot′+morph*	$\mathbf{20.60}^{(+2.11)}_{(0.01,10best)}$	$\mathbf{21.15}^{(+1.36)}_{(0.01,10best)}$	$\mathbf{21.05}^{(+0.95)}_{(0.01,5best,00)}$
	+ *PPT:1*	$\mathbf{20.61}^{(+2.12)}_{(1best)}$	$\mathbf{20.71}^{(+0.92)}_{(10best)}$	$20.32^{(+0.22)}_{(1best,000)}$
	+ *PPT:4*	$\mathbf{20.75}^{(+2.26)}_{(1best)}$	$\mathbf{21.08}^{(+1.29)}_{(5best)}$	$\mathbf{20.76}^{(+0.66)}_{(10best,.5.5.5)}$
(iii)	+ *PPT:4::CN:morph*	$\mathbf{21.01}^{(+2.52)}_{(1best)}$	$\mathbf{21.31}^{(+1.52)}_{(5best)}$	$\mathbf{20.98}^{(+0.88)}_{(10best,.5)}$
	System combination: (i) + (ii) + (iii)	$\mathbf{21.55}^{(+3.06)}$	$\mathbf{21.64}^{(+1.85)}$	$\mathbf{21.62}^{(+1.52)}$

Table 4: **Combined experiments: BLEU.** The best tuning parameter values are in subscript, and the absolute test improvement over the corresponding baseline (on top of each column) is in superscript.

The gap between *ML2EN* and *IN2EN* for unigram precision could be explained by vocabulary differences between Malay and Indonesian. Compared to *IN2EN*, all *CN:** models have higher 2/3/4-gram precision. However, *CN:pivot* has lower unigram precision, which could be due to bad word alignments, as the results for *CN:pivot′* show.

When morphological variants are further added, the unigram precision improves by almost 1% absolute over *CN:pivot′*. This shows the importance of morphology for overcoming the limitations of the small Indonesian vocabulary of the *IN2EN* bi-text.

The lower part of Table 3 shows that phrase-level paraphrasing performs a bit better. This confirms the importance of modeling context for closely-related languages like Malay and Indonesian, which are rich in false friends and partial cognates. We further see that using more scores in the phrase table is better. Extending the Indonesian vocabulary with cross-lingual morphological variants is still helpful, though not as much as at the word-level.

Finally, the combination of the output of the best PPT and the best CN systems using MEMT (Heafield and Lavie, 2010) yields even further improvements, which shows that the two kinds of paraphrases are complementary. The best overall BLEU score for our isolated experiments is 21.24, which is better than the results for all five baselines in Table 2, including the three bi-text combination baselines, which only achieve up to 20.10 BLEU.

6.3 Combined Experiments

Table 4 shows the performance of the three bi-text combination strategies (see Section 4.3 for additional details) when applied to combine *IN2EN* (1) with the original *ML2EN* and (2) with various adapted versions of it.

We can see that for the word-level paraphrasing experiments (*CN:**), all combinations except for *CN:pivot* perform significantly better than their corresponding baselines, but the improvements are most sizeable for the simple concatenation.

Note that while there is a difference of 0.31 BLEU points between the balanced concatenation and the sophisticated combination for the original *ML2EN*, they differ little for the adapted versions. This is probably due to the sophisticated combination assuming that the second bi-text is worse than the first one, which is not really the case for the adapted versions: as Table 3 shows, they all outperform *IN2EN*.

Overall, phrase-level paraphrasing performs a bit better than word-level paraphrasing, and system combination with MEMT improves even further. This is consistent with the isolated experiments.

7 Further Analysis

Paraphrasing non-Indonesian words only. In *CN:** above, we paraphrased *each* word in the Malay input, because of false friends like *polisi* and partial cognates like *nanti*. This risks proposing worse alternatives, e.g., changing *beliau* ('he', respectful) to *ia* ('he', casual), which confusion network weights and LM would not always handle. Thus, we tried paraphrasing non-Indonesian words only, i.e., those not in *IN-LM*. Since *IN-LM* occasionally contains some Malay-specific words, we also tried paraphrasing words that occur at most t times in *IN-LM*. Table 5 shows that this hurts by up to 1 BLEU point for $t = 0; 10$, and a bit less for $t = 20; 40$.

System	BLEU
CN:pivot, $t = 0$	$17.88_{(0.01,5best)}$
CN:pivot, $t = 10$	$17.88_{(0.05,10best)}$
CN:pivot, $t = 20$	$18.14_{(0.01,5best)}$
CN:pivot, $t = 40$	$18.34_{(0.01,5best)}$
CN:pivot (i.e., paraphrase all)	$18.91_{(0.005,10best)}$

Table 5: **Paraphrasing non-Indonesian words only:** those appearing at most t times in *IN-LM*.

Manual evaluation. We asked a native Indonesian speaker who does not speak Malay to judge whether our "Indonesian" adaptations are more understandable to him than the original Malay input for 100 random sentences. We presented him with two extreme systems: (a) the conservative *CN:pivot,t*=0 vs. (b) *CN:pivot′+morph*. Since the latter is noisy, the top 3 choices were judged for it. Table 6 shows that *CN:pivot,t*=0 is better/equal to the original 53%/31% of the time. In contrast, *CN:pivot′+morph* is typically worse than the original; even compared to the best in top 3, the better:worse ratio is 45%:43%.

Still, this latter model works better, which means that phrase-based SMT systems are robust to noise and prefer more variety. Note also that the judgments were at the sentence level, while phrases are sub-sentential, i.e., there can be many good phrases in a "bad" sentence.

System	Better	Equal	Worse
CN:pivot, $t = 0_{(Rank1)}$	53%	31%	16%
CN:pivot′+morph$_{(Rank1)}$	38%	8%	54%
CN:pivot′+morph$_{(Rank2)}$	41%	9%	50%
CN:pivot′+morph$_{(Rank3)}$	32%	11%	57%
CN:pivot′+morph$_{(Ranks:1-3)}$	45%	12%	43%

Table 6: **Human judgments: Malay vs. "Indonesian".** The parameter values are those from Tables 3 and 5.

Reversed Adaptation. In all experiments above, we were adapting the Malay sentences to look like Indonesian. Here we try to reverse the direction of adaptation, i.e., to adapt Indonesian to Malay: we thus build a "Malay" confusion network for each dev/test Indonesian sentence to be used as an input to a Malay–English SMT system trained on the *ML2EN* dataset. We tried two variations of this idea:

- ***lattice***: Use Indonesian-to-Malay confusion networks directly as input to the *ML2EN* SMT system, i.e., tune a log-linear model using confusion networks for the source side of the *IN2EN-dev* dataset, and then evaluate the tuned system using confusion networks for the source side of the *IN2EN-test* dataset.
- ***1-best***: Use the 1-best output from the Indonesian-to-Malay confusion network for each sentence of *IN2EN-dev* and *IN2EN-test*. Then pair each 1-best output with the corresponding English sentence. Finally, get an adapted "Malay"–English development set and an adapted "Malay"–English test set, and use them to tune and evaluate the *ML2EN* SMT system.

Table 7 shows that both variations perform worse than *CN:pivot*. We believe this is because *lattice* encodes many options, but does not use a Malay LM, while *1-best* uses a Malay LM, but has to commit to 1-best. In contrast, *CN:pivot* uses both n-best outputs and an Indonesian LM; designing a similar setup for reversed adaptation is a research direction we would like to pursue in future work.

System	BLEU
CN:pivot (Malay→Indonesian)	$18.91_{(0.005,10best)}$
CN:pivot (Indonesian→Malay) – lattice	$17.22_{(0.05)}$
CN:pivot (Indonesian→Malay) – 1-best	$17.77_{(0.001)}$

Table 7: **Reversed adaptation: Indonesian to Malay.**

Adapting Macedonian to Bulgarian. We experimented with another pair of closely-related languages,[8] Macedonian (*MK*) and Bulgarian (*BG*), using data from a different, non-newswire domain: the OPUS corpus of movie subtitles (Tiedemann, 2009). We used datasets of sizes that are comparable to those in the previous experiments: 160K *MK2EN* and 1.5M *BG2EN* sentence pairs (1.2M and 11.5M *EN* words). Since the sentences were short, we used 10K *MK2EN* sentence pairs for tuning and testing (77K and 72K English words). For the LM, we used 9.2M Macedonian and 433M English words.

Table 8 shows that both *CN:** and *PPT:** yield statistically significant improvements over balanced concatenation with unadapted *BG2EN*; system combination with MEMT improves even further. This indicates that our approach can work for other pairs of related languages and even for other domains.

We should note though that the improvements here are less sizeable than for Indonesian/Malay. This may be due to our monolingual *MK* dataset being smaller (10M *MK* vs. 20M *IN* words), and too noisy, containing many OCR errors, typos, concatenated words, and even some Bulgarian text. Moreover, Macedonian and Bulgarian are arguably somewhat more dissimilar than Malay and Indonesian.

System	BLEU	TER	METEOR
BG2EN (baseline)	24.57	57.64	41.60
MK2EN (baseline)	26.46	54.55	46.15
Balanced concatenation of *MK2EN* with an adapted *BG2EN*			
+ *BG2EN* (unadapted)	**27.33**	54.61	48.16
+ *CN:pivot′+morph*	**27.97**$^{(+0.64,+1.51)}$	54.08	49.65
+ *PPT:4::CN:morph*	**<u>28.38</u>**$^{(+1.05,+1.92)}$	53.35	48.21
Combining last three	**<u>29.05</u>**$^{(+1.72,+2.59)}$	52.31	50.96

Table 8: **Improving Macedonian–English SMT by adapting Bulgarian to Macedonian.**

[8] There is a heated political and linguistic debate about whether Macedonian represents a separate language or is a regional literary form of Bulgarian. Since there are no clear criteria for distinguishing a dialect from a language, linguists are divided on this issue. Politically, the Macedonian remains unrecognized as a language by Bulgaria and Greece.

8 Conclusion and Future Work

We have presented a novel approach for improving machine translation for a resource-poor language by *adapting* a bi-text for a related resource-rich language, using confusion networks, word/phrase-level paraphrasing, and morphological analysis.

We have achieved very significant improvements over several baselines (6.7 BLEU points over an unadapted version of *ML2EN*, 2.6 BLEU points over *IN2EN*, and 1.5–3 BLEU points over three bi-text combinations of *ML2EN* and *IN2EN*), thus proving the potential of the idea. We have further demonstrated the applicability of the general approach to other languages and domains.

In future work, we would like to add word deletion, insertion, splitting, and concatenation as allowed editing operations. We further want to explore tighter integration of word-based and phrase-based paraphrasing. Finally, we plan experiments with other language pairs and application to other linguistic problems.

Acknowledgments

We would like to give special thanks to Harta Wijaya and Aldrian Obaja Muis, native speakers of Indonesian, for their help in the linguistic analysis of the input and output of our system. We would also like to thank the anonymous reviewers for their constructive comments and suggestions, which have helped us improve the quality of this paper.

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

Kemal Altintas and Ilyas Cicekli. 2002. A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences*, ISCIS '02, pages 192–196.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, ACL-COLING '06.

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *Proceedings of the 6th International Conference on Informatics and Systems*, INFOS '08.

Timothy Baldwin and Su'ad Awab. 2006. Open source corpus analysis tools for Malay. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC '06, pages 2212–2215.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 9–16.

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

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multiparallel corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL '07, pages 728–735.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 531–540.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLP '00, pages 7–12.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL-HLT '11, pages 368–378.

Kenneth Heafield and Alon Lavie. 2010. Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme. *The Prague Bulletin of Mathematical Linguistics*, 93(1):27–36.

Luís Marujo, Nuno Grazina, Tiago Luís, Wang Ling, Luísa Coheur, and Isabel Trancoso. 2011. BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, EAMT '11, pages 129–136.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 1358–1367.

Preslav Nakov and Hwee Tou Ng. 2011. Translating from morphologically complex languages: A paraphrase-based approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL-HLT '11, pages 1298–1307.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 44:179–222.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL-Short '12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318.

Eric Ristad and Peter Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proc. of the Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*, AMTA '09.

Kevin P. Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the LREC 2006 Workshop on Strategies for Developing Machine Translation for Minority Languages*.

Jörg Tiedemann. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248.

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

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, ACL '09, pages 154–162.

Xiaoheng Zhang. 1998. Dialect MT: a case study between Cantonese and Mandarin. In *Proceedings of the 17th International Conference on Computational Linguistics*, COLING '98, pages 1460–1464.