Translating from Morphologically Complex Languages: A Paraphrase-Based Approach

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 approach to translating from a morphologically complex language. Unlike previous research, which has targeted word inflections and concatenations, we focus on the pairwise relationship between morphologically related words, which we treat as *potential paraphrases* and handle using paraphrasing techniques at the word, phrase, and sentence level. An important advantage of this framework is that it can cope with derivational morphology, which has so far remained largely beyond the capabilities of statistical machine translation systems. Our experiments translating from Malay, whose morphology is mostly derivational, into English show significant improvements over rivaling approaches based on five automatic evaluation measures (for 320,000 sentence pairs; 9.5 million English word tokens).

1 Introduction

Traditionally, statistical machine translation (SMT) models have assumed that the *word* should be the basic token-unit of translation, thus ignoring any word-internal morphological structure. This assumption can be traced back to the first word-based models of IBM (Brown et al., 1993), which were initially proposed for two languages with limited morphology: French and English. While several significantly improved models have been developed since then, including phrase-based (Koehn et al., 2003), hierarchical (Chiang, 2005), treelet (Quirk et al., 2005), and syntactic (Galley et al., 2004) models, they all preserved the assumption that words should be atomic.

Ignoring morphology was fine as long as the main research interest remained focused on languages with limited (e.g., English, French, Spanish) or minimal (e.g., Chinese) morphology. Since the attention shifted to languages like Arabic, however, the importance of morphology became obvious and several approaches to handle it have been proposed. Depending on the particular language of interest, researchers have paid attention to *word inflections* and *clitics*, e.g., for Arabic, Finnish, and Turkish, or to *noun compounds*, e.g., for German. However, *derivational morphology* has not been specifically targeted so far.

In this paper, we propose a paraphrase-based approach to translating from a morphologically complex language. Unlike previous research, we focus on the pairwise relationship between morphologically related wordforms, which we treat as *potential paraphrases*, and which we handle using paraphrasing techniques at various levels: word, phrase, and sentence level. An important advantage of this framework is that it can cope with various kinds of morphological wordforms, including derivational ones. We demonstrate its potential on Malay, whose morphology is mostly derivational.

The remainder of the paper is organized as follows: Section 2 gives an overview of Malay morphology, Section 3 introduces our paraphrase-based approach to translating from morphologically complex languages, Section 4 describes our dataset and our experimental setup, Section 5 presents and analyses the results, and Section 6 compares our work to previous research. Finally, Section 7 concludes the paper and suggests directions for future work.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1298–1307,
Portland, Oregon, June 19-24, 2011.

2 Malay Morphology and SMT

Malay is an Astronesian language, spoken by about 180 million people. It is official in Malaysia, Indonesia, Singapore, and Brunei, and has two major dialects, sometimes regarded as separate languages, which are mutually intelligible, but occasionally differ in orthography/pronunciation and vocabulary: Bahasa Malaysia (*lit.* 'language of Malaysia') and Bahasa Indonesia (*lit.* 'language of Indonesia').

Malay is an agglutinative language with very rich morphology. Unlike other agglutinative languages such as Finnish, Hungarian, and Turkish, which are rich in both inflectional and derivational forms, Malay morphology is mostly derivational. Inflectionally,[1] Malay is very similar to Chinese: there is no grammatical gender, number, or tense, verbs are not marked for person, etc.

In Malay, new words can be formed by the following three morphological processes:

- **Affixation**, i.e., attaching affixes, which are not words themselves, to a word. These can be prefixes (e.g., *ajar*/'teach' → ***pel****ajar*/'student'), suffixes (e.g., *ajar* → *ajar**an***/'teachings'), circumfixes (e.g., *ajar* → ***peng****ajar**an***/'lesson'), and infixes (e.g., *gigi*/'teeth' → *g**er**igi*/'toothed blade'). Infixes only apply to a small number of words and are not productive.

- **Compounding**, i.e., forming a new word by putting two or more existing words together. For example, *kereta*/'car' + *api*/'fire' make *kereta api* and *keretapi* in Bahasa Indonesia and Bahasa Malaysia, respectively, both meaning 'train'. As in English, Malay compounds are written separately, but some stable ones like *kerjasama*/'collaboration' (from *kerja*/'work' and *sama*/'same') are concatenated. Concatenation is also required when a circumfix is applied to a compound, e.g., *ambil alih*/'take over' (*ambil*/'take' + *alih*/'move') is concatenated to form ***peng****ambilalih**an***/'takeover' when targeted by the circumfix *peng-. . .-an*.

- **Reduplication**, i.e., word repetition. In Malay, reduplication requires using a dash. It can be full (e.g., *pelajar-pelajar*/'students'), partial (e.g., *adik-beradik*/'siblings', from *adik*/'younger brother/sister'), and rhythmic (e.g., *gunung-ganang*/'mountains', from the word *gunung*/'mountain').

Malay has very little inflectional morphology, It also has some **clitics**[2], which are not very frequent and are typically spelled concatenated to the preceding word. For example, the politeness marker *lah* can be added to the command *duduk*/'sit down' to yield *duduk**lah***/'please, sit down', and the pronoun *nya* can attach to *kereta* to form *kereta**nya***/'his car'. Note that clitics are not affixes, and clitic attachment is not a word derivation or a word inflection process.

Taken together, affixation, compounding, reduplication, and clitic attachment yield a rich variety of wordforms, which cause data sparseness issues. Moreover, the predominantly derivational nature of Malay morphology limits the applicability of standard techniques such as (1) removing some/all of the source-language inflections, (2) segmenting affixes from the root, and (3) clustering words with the same target translation. For example, if ***pel****ajar*/'student' is an unknown word and lemmatization/stemming reduces it to *ajar*/'teach', would this enable a good translation? Similarly, would segmenting[3] ***pel****ajar* as *peN+ ajar*, i.e., as 'person doing the action' + 'teach', make it possible to generate 'student' (e.g., as opposed to 'teacher')? Finally, if affixes tend to change semantics so much, how likely are we to find morphologically related wordforms that share the same translation? Still, there are many good reasons to believe that morphological processing should help SMT for Malay.

Consider *affixation*, which can yield words with similar semantics that can use each other's translation options, e.g., ***di****ajar*/'be taught (intransitive)' and ***di****ajarkan*/'be taught (transitive)'. However, this cannot be predicted from the affix, e.g., compare *minum*/'drink (verb)' – *minum**an***/'drink (noun)' and *makan*/'eat' – *makan**an***/'food'.

[1] *Inflection* is variation in the form of a word that is obligatory in some given grammatical context. For example, *plays, playing, played* are all inflected forms of the verb *play*. It does not yield a new word and cannot change the part of speech.

[2] A *clitic* is a morpheme that has the syntactic characteristics of a word, but is phonologically bound to another word. For example, *'s* is a clitic in *The Queen of England's crown*.

[3] The prefix *peN* suffers a nasal replacement of the archiphoneme *N* to become *pel* in *pelajar*.

Looking at *compounding*, it is often the case that the semantics of a compound is a specialization of the semantics of its head, and thus the target language translations available for the head could be usable to translate the whole compound, e.g., compare *kerjasama*/'collaboration' and *kerja*/'work'. Alternatively, it might be useful to consider a segmented version of the compound, e.g., *kerja sama*.

Reduplication, among other functions, expresses plural, e.g., *pelajar-pelajar*/'students'. Note, however, that it is not used when a quantity or a number word is present, e.g., *dua pelajar*/'two students' and *banyak pelajar*/'many students'. Thus, if we do not know how to translate *pelajar-pelajar*, it would be reasonable to consider the translation options for *pelajar* since it could potentially contain among its translation options the plural 'students'.

Finally, consider *clitics*. In some cases, a clitic could express a fine-grained distinction such as politeness, which might not be expressible in the target language; thus, it might be feasible to simply remove it. In other cases, e.g., when it is a pronoun, it might be better to segment it out as a separate word.

3 Method

We propose a *paraphrase-based approach* to Malay morphology, where we use paraphrases at three different levels: word, phrase, and sentence level.

First, we transform each development/testing Malay sentence into a *word lattice*, where we add simplified *word-level paraphrasing* alternatives for each morphologically complex word. In the lattice, each alternative w' of an original word w is assigned the weight of $\Pr(w'|w)$, which is estimated using pivoting over the English side of the training bi-text. Then, we generate *sentence-level paraphrases* of the training Malay sentences, in which exactly one morphologically complex word is substituted by a simpler alternative. Finally, we extract additional Malay phrases from these sentences, which we use to augment the phrase table with additional translation options to match the alternative wordforms in the lattice. We assign each such additional phrase p' a probability $\max_p \Pr(p'|p)$, where p is a Malay phrase that is found in the original training Malay text. The probability is calculated using *phrase-level pivoting* over the English side of the training bi-text.

3.1 Morphological Analysis

Given a Malay word, we build a list of morphologically simpler words that could be derived from it; we also generate alternative word segmentations:

(a) words obtainable by affix stripping
e.g., *pelajaran* → *pelajar*, *ajaran*, *ajar*

(b) words that are part of a compound word
e.g., *kerjasama* → *kerja*

(c) words appearing on either side of a dash
e.g., *adik-beradik* → *adik*, *beradik*

(d) words without clitics
e.g., *keretanya* → *kereta*

(e) clitic-segmented word sequences
e.g., *keretanya* → *kereta nya*

(f) dash-segmented wordforms
e.g., *aceh-nias* → *aceh - nias*

(g) combinations of the above.

The list is built by reversing the basic morphological processes in Malay: (a) addresses affixation, (b) handles compounding, (c) takes care of reduplication, and (d) and (e) deal with clitics. Strictly speaking, (f) does not necessarily model a morphological process: it proposes an alternative tokenization, but this could make morphological sense too.

Note that (g) could cause potential problems when interacting with (f), e.g., *adik-beradik* would become *adik - beradik* and then by (a) it would turn into *adik - adik*, which could cause the SMT system to generate two separate translations for the two instances of *adik*. To prevent this, we forbid the application of (f) to reduplications. Taking into account that reduplications can be partial, we only allow (f) if $\frac{|LCS(l,r)|}{\min(|l|,|r|)} < 0.5$, where l and r are the strings to the left and to the right of the dash, respectively, $LCS(x, y)$ is the longest common character subsequence, not necessarily consecutive, of the strings x and y, and $|x|$ is the length of the string x. For example, LCS(*adik*,*beradik*)=*adik*, and thus, the ratio is 1 (≥ 0.5) for *adik-beradik*. Similarly, LCS(*gunung*,*ganang*)=*gnng*, and thus, the ratio is 4/6=0.67 (≥ 0.5) for *gunung-ganang*. However, for *aceh-nias*, it is 1/4=0.25, and thus (f) is applicable.

As an illustration, here are the wordforms we generate for *adik-beradiknya*/'his siblings': *adik*, *adik-beradiknya*, *adik-beradik nya*, *adik-beradik*, *beradiknya*, *beradik nya*, *adik nya*, and *beradik*. And for *berpelajaran*/'is educated', we build the list: *berpelajaran*, *pelajaran*, *pelajar*, *ajaran*, and *ajar*. Note that the lists do include the original word.

To generate the above wordforms, we used two morphological analyzers: a freely available Malay lemmatizer (Baldwin and Awab, 2006), and an in-house re-implementation of the Indonesian stemmer described in (Adriani et al., 2007). Note that these tools' objective is to return a single lemma/stem, e.g., they would return *adik* for *adik-beradiknya*, and *ajar* for *berpelajaran*. However, it was straightforward to modify them to also output the above intermediary wordforms, which the tools were generating internally anyway when looking for the final lemma/stem. Finally, since the two modified analyzers had different strengths and weaknesses, we combined their outputs to increase recall.

3.2 Word-Level Paraphrasing

We perform word-level paraphrasing of the Malay sides of the development and the testing bi-texts.

First, for each Malay word, we generate the above-described list of morphologically simpler words and alternative word segmentations; we think of the words in this list as *word-level paraphrases*. Then, for each development/testing Malay sentence, we generate a lattice encoding all possible paraphrasing options for each individual word.

We further specify a weight for each arc. We assign 1 to the original Malay word w, and $\Pr(w'|w)$ to each paraphrase w' of w, where $\Pr(w'|w)$ is the probability that w' is a *good paraphrase* of w. Note that multi-word paraphrases, e.g., resulting from clitic segmentation, are encoded using a sequence of arcs; in such cases, we assign $\Pr(w'|w)$ to the first arc, and 1 to each subsequent arc.

We calculate the probability $\Pr(w'|w)$ using the training Malay-English bi-text, which we align at the word level using IBM model 4 (Brown et al., 1993), and we observe which English words w and w' are aligned to. More precisely, we use *pivoting* to estimate the probability $\Pr(w'|w)$ as follows:

$$\Pr(w'|w) = \sum_i \Pr(w'|w, e_i)\Pr(e_i|w)$$

Then, following (Callison-Burch et al., 2006; Wu and Wang, 2007), we make the simplifying assumption that w' is conditionally independent of w given e_i, thus obtaining the following expression:

$$\Pr(w'|w) = \sum_i \Pr(w'|e_i)\Pr(e_i|w)$$

We estimate the probability $\Pr(e_i|w)$ directly from the word-aligned training bi-text as follows:

$$\Pr(e_i|w) = \frac{\#(w,e_i)}{\sum_j \#(w,e_j)}$$

where $\#(x, e)$ is the number of times the Malay word x is aligned to the English word e.

Estimating $\Pr(w'|e_i)$ cannot be done directly since w' might not be present on the Malay side of the training bi-text, e.g., because it is a multi-token sequence generated by clitic segmentation. Thus, we think of w' as a pseudoword that stands for the union of all Malay words in the training bi-text that are reducible to w' by our morphological analysis procedure. So, we estimate $\Pr(w'|e_i)$ as follows:

$$\Pr(w'|e_i) = \Pr(\{v : w' \in forms(v)\}|e_i)$$

where $forms(x)$ is the set of the word-level paraphrases[4] for the Malay word x.

Since the training bi-text occurrences of the words that are reducible to w' are distinct, we can rewrite the above as follows:

$$\Pr(w'|e_i) = \sum_{v:w' \in forms(v)} \Pr(v|e_i)$$

Finally, the probability $\Pr(v|e_i)$ can be estimated using maximum likelihood:

$$\Pr(v|e_i) = \frac{\#(v,e_i)}{\sum_u \#(u,e_i)}$$

3.3 Sentence-Level Paraphrasing

In order for the word-level paraphrases to work, there should be phrases in the phrase table that could potentially match them. For some of the words, e.g., the lemmata, there could already be such phrases, but for other transformations, e.g., clitic segmentation, this is unlikely. Thus, we need to augment the phrase table with additional translation options.

One approach would be to modify the phrase table directly, e.g., by adding additional entries, where one or more Malay words are replaced by their paraphrases. This would be problematic since the phrase translation probabilities associated with these new

[4] Note that our paraphrasing process is directed: the paraphrases are morphologically simpler than the original word.

entries would be hard to estimate. For example, the clitics, and even many of the intermediate morphological forms, would not exist as individual words in the training bi-text, which means that there would be no word alignments or lexical probabilities available for them.

Another option would be to generate separate word alignments for the original training bi-text and for a version of it where the source (Malay) side has been paraphrased. Then, the two bi-texts and their word alignments would be concatenated and used to build a phrase table (Dyer, 2007; Dyer et al., 2008; Dyer, 2009). This would solve the problems with the word alignments and the phrase pair probabilities estimations in a principled manner, but it would require choosing for each word only one of the paraphrases available to it, while we would prefer to have a way to allow all options. Moreover, the paraphrased and the original versions of the corpus would be given equal weights, which might not be desirable. Finally, since the two versions of the bi-text would be word-aligned separately, there would be no interaction between them, which might lead to missed opportunities for improved alignments in both parts of the bi-text (Nakov and Ng, 2009).

We avoid the above issues by adopting a sentence-level paraphrasing approach. Following the general framework proposed in (Nakov, 2008), we first create multiple paraphrased versions of the source-side sentences of the training bi-text. Then, each paraphrased source sentence is paired with its original translation. This augmented bi-text is word-aligned and a phrase table T' is built from it, which is merged with a phrase table T for the original bi-text. The merged table contains all phrase entries from T, and the entries for the phrase pairs from T' that are not in T. Following Nakov and Ng (2009), we add up to three additional indicator features (taking the values 0.5 and 1) to each entry in the merged phrase table, showing whether the entry came from (1) T only, (2) T' only, or (3) both T and T'. We also try using the first one or two features only. We set all feature weights using minimum error rate training (Och, 2003), and we optimize their number (one, two, or three) on the development dataset.[5]

Each of our paraphrased sentences differs from its original sentence by a single word, which prevents combinatorial explosions: on average, we generate 14 paraphrased versions per input sentence. It further ensures that the paraphrased parts of the sentences will not dominate the word alignments or the phrase pairs, and that there would be sufficient interaction at word alignment time between the original sentences and their paraphrased versions.

3.4 Phrase-Level Paraphrasing

While our sentence-level paraphrasing informs the decoder about the origin of each phrase pair (original or paraphrased bi-text), it provides no indication about how good the phrase pairs from the paraphrased bi-text are likely to be.

Following Callison-Burch et al. (2006), we further augment the phrase table with one additional feature whose value is 1 for the phrase pairs coming from the original bi-text, and $\max_p \Pr(p'|p)$ for the phrase pairs extracted from the paraphrased bi-text. Here p is a Malay phrase from T, and p' is a Malay phrase from T' that does not exist in T but is obtainable from p by substituting one or more words in p with their derivationally related forms generated by morphological analysis. The probability $\Pr(p'|p)$ is calculated using phrase-level pivoting through English in the original phrase table T as follows (unlike word-level pivoting, here e_i is an English *phrase*):

$$\Pr(p'|p) = \sum_i \Pr(p'|e_i)\Pr(e_i|p)$$

We estimate the probabilities $\Pr(e_i|p)$ and $\Pr(p'|e_i)$ as we did for word-level pivoting, except that this time we use the list of the phrase pairs extracted from the original training bi-text, while before we used IBM model 4 word alignments. When calculating $\Pr(p'|e_i)$, we think of p' as the set of all possible Malay phrases q in T that are reducible to p' by morphological analysis of the words they contain. This can be rewritten as follows:

$$\Pr(p'|e_i) = \sum_{q:p' \in par(q)} \Pr(q|e_i)$$

where $par(q)$ is the set of all possible phrase-level paraphrases for the Malay phrase q.

The probability $\Pr(q|e_i)$ is estimated using maximum likelihood from the list of phrase pairs. There is no combinatorial explosion here, since the phrases are short and contain very few paraphrasable words.

[5] In theory, we should re-normalize the probabilities; in practice, this is not strictly required by the log-linear SMT model.

Number of sentence pairs	**1K**	**2K**	**5K**	**10K**	**20K**	**40K**	**80K**	**160K**	**320K**
Number of English words	30K	60K	151K	301K	602K	1.2M	2.4M	4.7M	9.5M
baseline	23.81	27.43	31.53	33.69	36.68	38.49	40.53	41.80	43.02
lemmatize all	22.67	26.20	29.68	31.53	33.91	35.64	37.17	38.58	39.68
	-1.14	-1.23	-1.85	-2.16	-2.77	-2.85	-3.36	-3.22	-3.34
‘noisier’ channel model (Dyer, 2007)	23.27	**28.42**	**32.66**	33.69	37.16	38.14	39.79	41.76	42.77
	-0.54	**+0.99**	**+1.13**	+0.00	**+0.48**	-0.35	-0.74	-0.04	-0.25
lattice + sent-par (orig+lemma)	**24.71**	**28.65**	**32.42**	**34.95**	**37.32**	38.40	39.82	41.97	43.36
	+0.90	**+1.22**	**+0.89**	**+1.26**	**+0.64**	-0.09	-0.71	+0.17	+0.34
lattice + sent-par	**24.97**	**29.11**	**33.03**	**35.12**	**37.39**	38.73	**41.04**	42.24	**43.52**
	+1.16	**+1.68**	**+1.50**	**+1.43**	**+0.71**	+0.24	**+0.51**	**+0.44**	**+0.50**
lattice + sent-par + word-par	**25.14**	**29.17**	**33.00**	**35.09**	**37.39**	38.76	*40.75*	**42.23**	**43.58**
	+1.33	**+1.74**	**+1.47**	**+1.40**	**+0.71**	+0.27	*+0.22*	**+0.43**	**+0.56**
lattice + sent-par + word-par + phrase-par	**25.27**	**29.19**	**33.35**	**35.23**	**37.46**	**39.00**	**40.95**	**42.30**	**43.73**
	+1.46	**+1.76**	**+1.82**	**+1.54**	**+0.78**	**+0.51**	**+0.42**	**+0.50**	**+0.71**

Table 1: **Evaluation results.** Shown are BLEU scores and improvements over the baseline (in %) for different numbers of training sentences. Statistically significant improvements are in **bold** for $p < 0.01$ and in *italic* for $p < 0.05$.

4 Experiments

4.1 Data

We created our Malay-English training and development datasets from data that we downloaded from the Web and then sentence-aligned using various heuristics. Thus, we ended up with 350,003 *training* sentence pairs, including 10.4M English and 9.7M Malay word tokens. We further downloaded 49.8M word tokens of monolingual English text, which we used for *language modeling*.

For *testing*, we used 1,420 sentences with 28.8K Malay word tokens, which were translated by three human translators, yielding translations of 32.8K, 32.4K, and 32.9K English word tokens, respectively. For *development*, we used 2,000 sentence pairs of 63.4K English and 58.5K Malay word tokens.

4.2 General Experimental Setup

First, we tokenized and lowercased all datasets: training, development, and testing. We then built directed word-level alignments for the training bitext for English→Malay and for Malay→English using IBM model 4 (Brown et al., 1993), which we symmetrized using the intersect+grow heuristic (Och and Ney, 2003). Next, we extracted phrase-level translation pairs of maximum length seven, which we scored and used to build a phrase table where each phrase pair is associated with the following five standard feature functions: forward and reverse phrase translation probabilities, forward and reverse lexicalized phrase translation probabilities, and phrase penalty.

We trained a log-linear model using the following standard SMT feature functions: trigram language model probability, word penalty, distance-based distortion cost, and the five feature functions from the phrase table. We set all weights on the development dataset by optimizing BLEU (Papineni et al., 2002) using minimum error rate training (Och, 2003), and we plugged them in a beam search decoder (Koehn et al., 2007) to translate the Malay test sentences to English. Finally, we detokenized the output, and we evaluated it against the three reference translations.

4.3 Systems

Using the above general experimental setup, we implemented the following baseline systems:

- **baseline**. This is the default system, which uses no morphological processing.

- **lemmatize all**. This is the second baseline that uses lemmatized versions of the Malay side of the training, development and testing datasets.

- **‘noisier’ channel model**.[6] This is the model of Dyer (2007). It uses 0-1 weights in the lattice and only allows lemmata as alternative wordforms; it uses no sentence-level or phrase-level paraphrases.

[6] We also tried the word segmentation model of Dyer (2009) as implemented in the *cdec* decoder (Dyer et al., 2010), which learns word segmentation lattices from raw text in an unsupervised manner. Unfortunately, it could not learn meaningful word segmentations for Malay, and thus we do not compare against it. We believe this may be due to its focus on word segmentation, which is of limited use for Malay.

sent.	system	1-gram	2-gram	3-gram	4-gram
1k	baseline	59.78	29.60	17.36	10.46
	paraphrases	62.23	31.19	18.53	11.35
2k	baseline	64.20	33.46	20.41	12.92
	paraphrases	66.38	35.42	21.97	14.06
5k	baseline	68.12	38.12	24.20	15.72
	paraphrases	70.41	40.13	25.71	17.02
10k	baseline	70.13	40.67	26.15	17.27
	paraphrases	72.04	42.28	27.55	18.36
20k	baseline	73.19	44.12	29.14	19.50
	paraphrases	73.28	44.43	29.77	20.31
40k	baseline	74.66	45.97	30.70	20.83
	paraphrases	75.47	46.54	31.09	21.17
80k	baseline	75.72	48.08	32.80	22.59
	paraphrases	76.03	48.47	33.20	23.00
160k	baseline	76.55	49.21	34.09	23.78
	paraphrases	77.14	49.89	34.57	24.06
320k	baseline	77.72	50.54	35.19	24.78
	paraphrases	78.03	51.24	35.99	25.42

Table 2: **Detailed BLEU n-gram precision scores:** in %, for different numbers of training sentence pairs, for *baseline* and *lattice* + *sent-par* + *word-par* + *phrase-par*.

Sent.	System	BLEU	NIST	TER	METEOR	TESLA
1k	baseline	23.81	6.7013	64.50	49.26	1.6794
	paraphrases	25.27	6.9974	63.03	52.32	1.7579
2k	baseline	27.43	7.3790	61.03	54.29	1.8718
	paraphrases	29.19	7.7306	59.37	57.32	2.0031
5k	baseline	31.53	8.0992	57.12	59.09	2.1172
	paraphrases	33.35	8.4127	55.41	61.67	2.2240
10k	baseline	33.69	8.5314	55.24	62.26	2.2656
	paraphrases	35.23	8.7564	53.60	63.97	2.3634
20k	baseline	36.68	8.9604	52.56	64.67	2.3961
	paraphrases	37.46	9.0941	52.16	66.42	2.4621
40k	baseline	38.49	9.3016	51.20	66.68	2.5166
	paraphrases	39.00	9.4184	50.68	67.60	2.5604
80k	baseline	40.53	9.6047	49.88	68.77	2.6331
	paraphrases	40.95	9.6289	49.09	69.10	2.6628
160k	baseline	41.80	9.7479	48.97	69.59	2.6887
	paraphrases	42.30	9.8062	48.29	69.62	2.7049
320k	baseline	43.02	9.8974	47.44	70.23	2.7398
	paraphrases	43.73	9.9945	47.07	70.87	2.7856

Table 3: **Results for different evaluation measures:** for *baseline* and *lattice* + *sent-par* + *word-par* + *phrase-par* (in % for all measures except for NIST).

Our full morphological paraphrasing system is **lattice + sent-par + word-par + phrase-par**. We also experimented with some of its components turned off. **lattice + sent-par + word-par** excludes the additional feature from phrase-level paraphrasing. **lattice + sent-par** has all the morphologically simpler derived forms in the lattice during decoding, but their weights are uniformly set to 0 rather than obtained using pivoting from word alignments. Finally, in order to compare closely to the 'noisier' channel model, we further limited the morphological variants of **lattice + sent-par** in the lattice to lemmata only in **lattice + sent-par (orig+lemma)**.

5 Results and Discussion

The experimental results are shown in Table 1.

First, we can see that *lemmatize all* has a consistently disastrous effect on BLEU, which shows that Malay morphology does indeed contain information that is important when translating to English.

Second, Dyer (2007)'s *'noisier' channel model* helps for small datasets only. It performs worse than *lattice + sent-par (orig+lemma)*, from which it differs in the phrase table only; this confirms the importance of our sentence-level paraphrasing.

Moving down to *lattice + sent-par*, we can see that using multiple morphological wordforms instead of just lemmata has a consistently positive impact on BLEU for datasets of all sizes.

Adding weights obtained using word-level pivoting in *lattice* + *sent-par* + *word-par* helps a bit more, and also using phrase-level paraphrasing weights yields even bigger further improvements for *lattice* + *sent-par* + *word-par* + *phrase-par*.

Overall, our morphological paraphrases yield statistically significant improvements ($p < 0.01$) in BLEU, according to Collins et al. (2005)'s sign test, for bi-texts as large as 320,000 sentence pairs.

A closer look at BLEU. Table 2 shows detailed n-gram BLEU precision scores for n=1,2,3,4. Our system outperforms the baseline on all precision scores and for all numbers of training sentences.

Other evaluation measures. Table 3 reports the results for five evaluation measures: BLEU and NIST 11b, TER 0.7.25 (Snover et al., 2006), METEOR 1.0 (Lavie and Denkowski, 2009), and TESLA (Liu et al., 2010). Our system consistently outperforms the baseline for all measures.

Example translations. Table 4 shows two translation examples. In the first example, the reduplication *bekalan-bekalan* ('supplies') is an unknown word, and was left untranslated by the baseline system. It was not a problem for our system though, which first paraphrased it as *bekalan* and then translated it as *supply*. Even though this is still wrong (we need the plural *supplies*), it is arguably preferable to passing the word untranslated; it also allowed for a better translation of the surrounding context.

`src :` Mercy Relief telah menghantar 17 khemah khas bernilai $5,000 setiap satu yang boleh menampung kelas seramai 30 pelajar, selain **bekalan-bekalan lain seperti 500 khemah biasa**, barang makanan dan ubat-ubatan untuk mangsa gempa Sichuan.

`ref1:` Mercy Relief has sent 17 special tents valued at $5,000 each, that can accommodate a class of 30 students, including **other aid supplies such as 500 normal tents**, food and medicine for the victims of Sichuan quake.

`base:` mercy relief has sent 17 special tents worth $5,000 each class could accommodate a total of 30 students, besides ***other bekalan-bekalan 500 tents as usual***, foodstuff and medicines for sichuan quake relief.

`para:` mercy relief has sent 17 special tents worth $5,000 each class could accommodate a total of 30 students, besides **other supply such as 500 tents**, food and medicines for sichuan quake relief.

`src :` Walaupun hidup susah, kami tetap berusaha untuk **menjalani kehidupan** seperti biasa.

`ref1:` Even though life is difficult, we are still trying to **go through life** as usual.

`base:` despite the hard life, we will always strive to ***undergo training*** as usual.

`para:` despite the hard life, we will always strive to **live** normal.

Table 4: **Example translations**. For each example, we show a source sentence (`src`), one of the three reference translations (`ref1`), and the outputs of *baseline* (`base`) and of *lattice + sent-par + word-par + phrase-par* (`para`).

In the second example, the baseline system translated *menjalani kehidupan* (lit. ‘go through life’) as *undergo training*, because of a bad phrase pair, which was extracted from wrong word alignments. Note that the words *menjalani* (‘go through’) and *kehidupan* (‘life/existence’) are derivational forms of *jalan* (‘go’) and *hidup* (‘life/living’), respectively. Thus, in the paraphrasing system, they were involved in sentence-level paraphrasing, where the alignments were improved. While the wrong phrase pair was still available, the system chose a better one from the paraphrased training bi-text.

6 Related Work

Most research in SMT for a morphologically rich **source** language has focused on inflected *forms of the same* word. The assumption is that they would have similar semantics and thus could have the same translation. Researchers have used *stemming* (Yang and Kirchhoff, 2006), *lemmatization* (Al-Onaizan et al., 1999; Goldwater and McClosky, 2005; Dyer, 2007), or *direct clustering* (Talbot and Osborne, 2006) to identify such groups of words and use them as *equivalence classes* or as possible *alternatives* in translation. Frameworks for the simultaneous use of different word-level representations have been proposed as well (Koehn and Hoang, 2007).

A second important line of research has focused on *word segmentation*, which is useful for languages like German, which are rich in *compound words* that are spelled concatenated (Koehn and Knight, 2003; Yang and Kirchhoff, 2006), or like Arabic, Turkish, Finnish, and, to a lesser extent, Spanish and Italian, where *clitics* often attach to the preceding word (Habash and Sadat, 2006). For languages with more or less regular inflectional morphology like Arabic or Turkish, another good idea is to segment words into *morpheme sequences*, e.g., prefix(es)-stem-suffix(es), which can be used instead of the original words (Lee, 2004) or in addition to them. This can be achieved using a lattice input to the translation system (Dyer et al., 2008; Dyer, 2009).

Unfortunately, none of these general lines of research suits Malay well, whose compounds are rarely concatenated, clitics are not so frequent, and morphology is mostly derivational, and thus likely to generate words whose semantics substantially differs from the semantics of the original word. Therefore, we cannot expect the existence of equivalence classes: it is only occasionally that two derivationally related wordforms would share the same target language translation. Thus, instead of looking for equivalence classes, we have focused on the pairwise relationship between derivationally related wordforms, which we treat as *potential paraphrases*.

Our approach is an extension of the *‘noisier’ channel model* of Dyer (2007). He starts by generating separate word alignments for the original training bi-text and for a version of it where the source side has been lemmatized. Then, the two bi-texts and their word alignments are concatenated and used to build a phrase table. Finally, the source sides of the development and the test datasets are converted into confusion networks where additional arcs are added for word lemmata. The arc weights are set to 1 for the original wordforms and to 0 for the lemmata. In contrast, we provide *multiple* paraphrasing alternatives for each morphologically complex word, including derivational forms that occupy intermediary positions between the original wordform and its lemma. Note that some of those paraphrasing alternatives are *multi-word*, and thus we use a *lattice* instead of a confusion network. Moreover, we give *different weights* to the different alternatives rather then assigning them all 0.

Second, our work is related to that of Dyer et al. (2008), who use a lattice to add a single alternative clitic-segmented version of the original word for Arabic. However, we provide *multiple* alternatives. We also include *derivational forms* in addition to clitic-segmented ones, and we give *different weights* to the different alternatives (instead of 0).

Third, our work is also related to that of Dyer (2009), who uses a lattice to add multiple alternative segmented versions of the original word for German, Hungarian, and Turkish. However, we focus on *derivational morphology* rather than on clitics and inflections, add *derivational forms* in addition to clitic-segmented ones, and use *cross-lingual word pivoting* to estimate paraphrase probabilities.

Finally, our work is related to that of Callison-Burch et al. (2006), who use cross-lingual pivoting to generate phrase-level paraphrases with corresponding probabilities. However, our paraphrases are derived through *morphological analysis*; thus, we do not need corpora in additional languages.

7 Conclusion and Future Work

We have presented a novel approach to translating from a morphologically complex language, which uses paraphrases and paraphrasing techniques at three different levels of translation: word-level, phrase-level, and sentence-level. Our experiments translating from Malay, whose morphology is mostly derivational, into English have shown significant improvements over rivaling approaches based on several automatic evaluation measures.

In future work, we want to improve the probability estimations for our paraphrasing models. We also want to experiment with other morphologically complex languages and other SMT models.

Acknowledgments

This work was supported by research grant POD0713875. We would like to thank the anonymous reviewers for their detailed and constructive comments, which have helped us improve the paper.

References

Mirna Adriani, Jelita Asian, Bobby Nazief, S. M.M. Tahaghoghi, and Hugh E. Williams. 2007. Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing*, 6:1–33.

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation. Technical report, JHU Summer Workshop.

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.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. 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 the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL '06, pages 17–24.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 263–270.

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.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL '08, pages 1012–1020.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, ACL '10, pages 7–12.

Christopher Dyer. 2007. The 'noisier channel': translation from morphologically complex languages. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 207–211.

Chris Dyer. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of Human Language Technologies: The 2009 Annual*

Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, pages 406–414.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL '04, pages 273–280.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP '05, pages 676–683.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, HLT-NAACL '06, pages 49–52.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 868–876.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '03, pages 187–193.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 48–54.

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 the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume on Demo and Poster Sessions*, ACL '07, pages 177–180.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL '04, pages 57–60.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. TESLA: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 354–359.

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

Preslav Nakov. 2008. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of the 18th European Conference on Artificial Intelligence*, ECAI '08, pages 338–342.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 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 the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 271–279.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, AMTA '06, pages 223–231.

David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, COLING-ACL '06, pages 969–976.

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

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 41–48.