THE NATIONAL UNIVERSITY
of SINGAPORE

School of Computing
Computing 1, 13 Computing Drive, Singapore 117417

TR22/10

Morphological Analysis for Resource-Poor Machine Translation

Ming-Feng Tsai, Preslav Nakov and Hwee Tou Ng

December 2010

T e c h n i c a l R e p o r t

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

OOI Beng Chin
Dean of School

Morphological Analysis for Resource-Poor Machine Translation

Ming-Feng Tsai
Preslav Nakov
Hwee Tou Ng
Department of Computer Science
National University of Singapore

In statistical machine translation, word-to-word probabilities are usually difficult to estimate because of the problem of data sparseness, especially for resource-poor languages. Furthermore, this problem would become more serious for translation from morphologically complex languages such as Malay or Indonesian to morphologically simple ones such as English, since we need to be able to translate word forms in many different morphological variants. This paper conducts a morphological analysis for such resource-poor and morphologically rich machine translation: one is Malay-English machine translation; another is Indonesian-English. Specifically, we use morphological analysis to modify the unknown words of morphologically complex languages, and explore the effect of using the modified input on translation quality with varying number of training sentences. In our experiments, a number of trials were carried out to assess the performance of the proposed approach. The experimental results show that our proposed method can improve translation quality when the rate of unknown words is higher than 20%, and the improvement gradually increases as the unknown word rate increases.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Machine Translation

General Terms: Machine Translation

Additional Key Words and Phrases: Resource-Poor Languages

1. INTRODUCTION

Statistical machine translation has improved translation quality substantially in recent years. However, state-of-the-art approaches are essentially based on lexicons. That is, in these approaches, every pair of surface word or phrase in the source language and its corresponding translation in the target language is considered an independent entity. This word-based or phrase-based approach causes sensitivity to data sparseness. The data sparseness problem is more severe for resource-poor languages when parallel corpora are scarce. The problem also becomes particularly

serious for morphologically complex languages, where words are used in many different surface forms, requiring a larger amount of parallel texts as training data in order to give good translation performance.

Below we briefly present the principle of statistical machine translation, and then introduce our proposed approach.

In statistical machine translation, the goal is to find the most likely translation of some foreign language text f into a desired language e. In order to attain this goal, a machine translation system attempts to maximize the probability $P(e|f)$. Instead of maximizing $P(e|f)$ directly, the standard generative model turns to using Bayes rules to divide the problem into two separate parts:

$$\arg\max_e P(e|f) = \arg\max_e P(e)P(f|e) \tag{1}$$

where $P(e)$ is the *language model* and $P(f|e)$ is the *translation model*. According to [Weaver 1955], we can think of the language model $P(e)$ as a stochastic model that generates target language sentences, and the translation model $P(f|e)$ as a second stochastic process that "corrupts" the target language to produce source language sentences.

Based on the above framework, the quality of machine translation would heavily depend on that of the translation model. Parallel corpora needed for training the translation model are usually not as readily available as monolingual resources needed for training the language model. Because of the difficulty in obtaining large parallel corpora, sparse data becomes a serious issue in estimating the parameters of the translation model. The problem worsens when one or both of the source and target language are morphologically complex languages.

In order to investigate the problem, we conducted a series of experiments and found that morphological analysis can be employed to reduce data sparseness and can increase the similarity between languages, thereby enhancing the quality of machine translation for morphologically complex languages. Specifically, we lemmatize those unknown words found in the development and test data sets, and then replace the unknown words with their lemmatized forms as input for translation. Our experiments consist of two parts: one is on Malay-English translation, the other is Indonesian-English translation. Both parts are aimed at translating a morphologically more complex language (i.e., Malay or Indonesian) into a morphologically simpler language (i.e., English). In addition to our investigation in morphology, this work is also related to the study of resource-poor machine translation because both Malay and Indonesian at present are without plentiful parallel corpora.

This paper is organized as follows: In Section 2, we review previous work on using morphological analysis for statistical machine translation. Section 3 describes our

proposed methods of utilizing morphological information for machine translation. Section 4 presents the experimental setup and results, and offers some observations and discussions on the results. We conclude the paper in Section 5.

2. RELATED WORK

Most modern statistical machine translation (SMT) models rely on the assumption that the *word* is the basic token unit of translation, thus ignoring any word-internal morphological structure. This tradition 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. More complex models have been introduced later, e.g., phrase-based [Koehn et al. 2003], hierarchical phrase-based [Chiang 2005], treelet [Quirk et al. 2005], and syntactic [Galley et al. 2004], but they all preserved the assumption that words should be atomic.

Ignoring morphology worked fine as long as the main research interest was focused on languages with limited (e.g., English, French, Spanish) or non-existent (e.g., Chinese) morphology. After attention shifted to agglutinative languages like Arabic, however, it became obvious that morphology simply could not be ignored any more and various proposals have been made of morphology-aware SMT models.

Most morphology-aware SMT models are extensions of existing models: word-based, phrase-based, hierarchical phrase-based, and syntactic. Often, an assumption is further made that the translation is between a morphologically rich and a morphologically poor language, and different approaches have been proposed depending on whether the source or the target is the morphologically rich one.

2.1 Morphologically Rich *Source* Language

This is the easy translation direction. Typically, the model needs to learn that many source wordforms have the same translation in the target language, e.g., all forms of the French adjective *petit* (namely *petit*, *petite*, *petits*, *petites*) could be translated as *little* in English. Learning such a many-to-one mapping[1] is easy and often can be done automatically by the translation model; this is why it is typically not handled in any special way for French-English. Still, there is a problem: having many possible morphological variants increases the vocabulary size of the SMT system, thus creating data sparseness issues.

One simple way to reduce data sparseness is to **remove all or most of the source language morphology**, e.g., by lemmatizing or stemming the input. In-

[1]We do not claim that these mappings are always many-to-one. For example, all four forms of *petit* could also be translated as *small*, i.e., the mapping is actually many-to-many but with more word forms on the source-language side (in general).

deed, [Al-Onaizan et al. 1999] have shown that lemmatization can yield improvements for a word-based Czech-English SMT system. Of course, removing all inflectional Czech morphology might be too radical (some of it might translate into corresponding English morphology, e.g., the plural markers for nouns), and thus they also tried using modified Czech lemmata where some morphological markers were added in order to increase the correspondence to English lemmata; this yielded some further improvements. Unfortunately, the experiments were performed on a small corpus with just 50K sentences and the evaluation was based on subjective scores.[2]

Using the same language pair, the same dataset, and the same kind of word-based SMT system, [Goldwater and McClosky 2005] compared four **pre-processing approaches** to cope with data sparseness: (1) lemmatizing the input (as before); (2) using modified Czech lemmata (similar, but a bit different than above); (3) introducing **pseudowords corresponding to some important inflections** that could help translation, e.g., `TEN_F` for future tense[3], and `PER_1` for first person singular[4]; and (4) using a **morpheme-aware word alignment** model. By combining all four approaches, they improved the Bleu score from 27% to 33.3%.

[Habash and Sadat 2006] performed similar experiments for Arabic-English translation using the following four pre-processing approaches: (1) simple tokenization; (2) minimizing the differences with English; (3) three types of rule-based decliticization; (4) breaking up words into stem and affixival morphemes. Experiments on a corpus of 5M words show statistically significant improvements over the baseline: from 34.6% to 37.1% Bleu when testing on NIST MT04, and from 37.8% to 38.7% Bleu when testing on NIST MT05.

[Lee 2004] segmented words as sequences of prefix(es)-stem-suffix(es) and used POS tagging, thus trying to impose morphological and syntactic symmetry between the source and the target languages. They further identified morphemes that needed to be merged or deleted in the process. The evaluation on Arabic-English translation when training on 3.3M sentence pairs showed an improvement from 18% to 32% Bleu for an IBM Model 1 SMT model, and from 36% to 39% Bleu when using a phrase-based SMT model.

Note that in the above work, approach (2) is based on manual decisions about which morphemes should be discarded and which ones should be preserved, which

[2]Bleu [Papineni et al. 2002] was invented three years later.
[3]Future tense is typically expressed by a prefix in Czech.
[4]Czech is a PRO-drop language, but, normally, the subject can be recovered from the verb inflection.

requires deep linguistic knowledge. [Talbot and Osborne 2006] proposed an alternative language-independent *clustering approach* that uses the training bi-text to discover automatically which morphological distinctions in the source language should be preserved and which ones should be considered redundant. They learned word clusters and used them to improve word alignments and to smooth phrase pair probabilities in a phrase-based SMT system, achieving statistically significant improvements in Bleu when translating from Czech, French, and Welsh to English for training corpora of up to 250K sentences.

[Yang and Kirchhoff 2006] proposed a **back-off model** for phrase-based SMT that reduced the unseen wordforms in the source text to morphologically simpler forms using stemming and compound splitting. The reduction was performed sequentially so that wordforms that were closer to the original wordform were preferred. The phrase table entries for words sharing the same reduced wordform were then modified by replacing the respective words with their reduced wordforms and re-estimating all phrase table probabilities using a discounted back-off model. The approach yielded statistically significant improvements in Bleu for tiny training corpora of up to 5K sentence pairs when translating from German or Finnish to English: from 15.3% to 16.3%, and from 12.9% to 14.0%, respectively.

All above-described approaches make their decisions about whether to use morphological information in a pre-processing and/or a post-processing step; more importantly, they need to commit to one possible analysis in case of ambiguities.

In contrast, [Dyer 2007] proposed what he called the **'noisier channel' model**[5], where a **confusion network** (CN) is constructed for each sentence on the source side of the development and the test datasets. In this CN, a single back-off form is provided at each position in the sentence where a lemmatizer yields a different wordform than the actual observed surface form. The back-off form is assigned a cost of 1 and the surface form a cost of 0. The evaluation for Czech-English translation using a hierarchical SMT model for a small training corpus of 57.8K sentence pairs (1.2M tokens) of News Commentary data shows that the proposed model outperforms a back-off model similar[6] to that of [Yang and Kirchhoff 2006] (25.01% vs. 23.94% Bleu), which in turn outperforms the baseline of using the surface wordforms only (22.74% Bleu). All pairwise differences between the three

[5]The idea has already been popular in spoken language translation, where confusion networks were commonly used to represent a lattice of alternative output hypotheses from the automatic speech recognition system, e.g., [Ney 1999; Casacuberta et al. 2004; Saleem et al. 2004; Matusov et al. 2005].

[6]The main difference is that back-off forms were generated for every possible surface string, not just the unknown words; the word alignments were improved as well.

models are statistically significant.

Note that the above approach works with words and lemmata only, which does not change the number of input tokens; it just creates alternatives, which can be handled with confusion networks.

[Dyer et al. 2008] took a step further. They proposed segmenting the words in the input sentences into morphemes and then combining the surface wordforms with the segmented forms. Note that since segmenting a word into morphemes changes the number of tokens in the sentence, their approach required moving from a confusion network to a **word segmentation lattice**. In their experiments, they used the Buckwalter morphological analyzer [Buckwalter 2004] and disambiguated the analysis using a simple unigram model trained on the Penn Arabic Treebank. Using a phrase-based SMT model trained on the entire NIST MT08 training data for Arabic-English and testing on the NIST MT05 test data, their method achieved 52.25% Bleu, which is a statistically significant improvement over using surface wordforms (46.82% Bleu) or morphologically segmented input (50.87% Bleu). Similarly, testing on the NIST MT06 test data yielded 40.08% Bleu, which is statistically significantly better than using surface wordforms (35.12% Bleu) or morphologically segmented input (38.41% Bleu). Similar statistically significant improvements were observed using a hierarchical model on the same training/testing datasets: for MT05, from 52.53% for surface wordforms and 53.77% for the morphologically segmented input to 54.53% for the word segmentation lattices, and for MT06, from 39.91% for surface wordforms and 41.80% for the morphologically segmented input to 42.87% for the word segmentation lattices. These are huge improvements given the large size of the training corpus.

Note that morphological analysis can be ambiguous. For example, the Buckwalter morphological analyzer produced many possible segmentations and [Dyer et al. 2008] had to perform a subsequent disambiguation step. In contrast, [Dyer 2009] proposed using a **segmentation lattice encoding many alternative segmentations for each input word** without the need for disambiguation. The lattices were produced by a maximum entropy model that was trained on monolingual text in an unsupervised manner (it was further tuned on manually prepared lattices for German); they were then used in a hierarchical SMT model. The evaluation showed statistically significant improvements in Bleu when translating from German, Hungarian, and Turkish to English: from 21.0% to 21.6%, from 11.0% to 12.3%, and from 26.9% to 28.7%, respectively. These results are impressive since they were achieved for very large training corpora: 1.4M German-English and 1.5M

Hungarian-English sentence pairs, respectively.[7]

2.2 Morphologically Rich *Target* Language

This is the hard translation direction since the system has to generate many more wordforms than are present in the source side.

[Nießen and Ney 2004] modeled the interdependencies between related inflected wordforms on the morphologically-rich target-language side by means of a **hierarchical lexicon** model, which represents words as combinations of full forms, base forms, and part-of-speech tags, and allowed the word alignment training procedure to interpolate counts based on these different levels of representation. They further used sentence-level restructuring, which aimed at assimilating the word order in related sentences. The evaluation for English-German using a word-based IBM Model-4 SMT system and training on 58K sentence pairs showed a statistically significant improvement from 53.7% to 57.1% Bleu.

[Toutanova et al. 2008] used an **inflection generation model trained independently of the SMT system** and predicted wordforms from their stems using extensive morphological and syntactic information from both the source and target languages. The generation model itself was based on [Minkov et al. 2007]. The evaluation on English-Russian translation showed an improvement from 29.24% to 31.80% Bleu using a treelet SMT system, and from 36.00% to 36.72% Bleu using a phrase-based SMT system. For English-Arabic translation, the improvement was from 35.54% to 37.41% Bleu using a treelet SMT system. These improvements are impressive since they were achieved for very large training corpora: 1.46M English-Russian and 0.46M English-Arabic sentence pairs (24M and 5.2M English words, respectively). They were further confirmed by human judgments.

Research efforts in improving translation into a morphologically rich language have been undertaken for some other languages such as *Greek* [Avramidis and Koehn 2008], *Hungarian* [Novák 2009; Koehn and Haddow 2009], and *Turkish* [Oflazer and El-Kahlout 2007]. These works, however, either only helped for small datasets [Oflazer and El-Kahlout 2007], or yielded very modest improvements when applied on large corpora, e.g., [Avramidis and Koehn 2008] improved 0.15% and 0.72% Bleu absolute over an 18.05% Bleu phrase-based English-Greek baseline SMT system.

[7] The results for Turkish were achieved on a much smaller training corpus: 45.7K Turkish-English sentence pairs.

2.3 Other Models

[Koehn and Hoang 2007] proposed the **factored translation model** as a general framework that allows for a principled integration of word-level annotations in a phrase-based SMT model. This is achieved by transforming each word into a vector of factors (e.g., surface form, lemma, part-of-speech, semantic class, morphological tag). The authors described some particular instantiations of the model. They experimented with translating from English into a morphologically-rich language, achieving the following improvements over the baseline from 18.15% to 18.22% Bleu for English-German (trained on 751K sentence pairs), from 23.41% to 24.66% Bleu for English-Spanish (trained on 40K sentence pairs), and from 25.82% to 27.62% Bleu for English-Czech (trained on 20K sentence pairs). They further experimented with translating from a morphologically-rich language into English, improving from 18.19% to 19.47% Bleu for German-English. The framework has been made part of the Moses toolkit [Koehn et al. 2007]. While it is theoretically very neat, it has three important drawbacks: (1) it leads to a combinatorial explosion in the search space, which makes it very time-consuming for large training corpora, (2) it yields large improvements for small training corpora only, (3) the word remains somewhat atomic in the sense that it cannot be represented as a sequence of morphemes, i.e., the number of tokens remains the same, it is just the token representation that changes into a vector of factors.

3. OUR PROPOSED METHOD

An approach to using morphological information to modify input data is to replace word forms with their associated lemmas. According to previous work in [Goldwater and McClosky 2005; Toutanova et al. 2008; Cartoni 2009], morphological information could bring in some improvement on translation quality because of the reduction of data sparseness. However, since lemmatization may also remove useful information from the source language, we suggest lemmatizing those words that cannot be found in the training set only. That is, we lemmatize the unknown words in the development and testing data sets only. Such a strategy not only keeps useful information intact, but also avoids introducing noise. The reason we lemmatize unknown words only is that they are usually problematic for any NLP system. According to the studies in [Ren and Perrault 1992; Cartoni 2009], about 5 to 10% of words of a text written in a "standard" language are unknown to lexical resources. For a machine translation system, unknown words results in incorrect translation.

In order to evaluate the effect of lemmatizing unknown words on translation

Original:	**ketiga-tiga** kumpulan utama – makanan dan minuman bukan alkohol , **perumahan** , air , elektrik , gas dan bahan api lain , dan pengangkutan ...
Tan lemmas:	**tiga** kumpulan utama – makanan dan minuman bukan alkohol , perumahan , air , elektrik , gas dan bahan api lain , dan pengangkutan ...
Baldwin lemmas:	**ketiga** kumpulan utama – makanan dan minuman bukan alkohol , **rumah** , air , elektrik , gas dan bahan api lain , dan pengangkutan ...
Original:	" itu sebab **mengapa** kami sangat kukuh di negara ini kerana ia menyediakan infrastruktur sempurna untuk **pengilang** yang mengeluarkan produk ...
Tan lemmas:	" itu sebab **kapa** kami sangat kukuh di negara ini kerana ia menyediakan infrastruktur sempurna untuk **kilang** yang mengeluarkan produk ...
Baldwin lemmas:	" itu sebab **apa** kami sangat kukuh di negara ini kerana ia menyediakan infrastruktur sempurna untuk pengilang yang mengeluarkan produk ...

Table I. **Examples of Malay sentences after the lemmatization of unknown words.**

quality, we use two lemmatizers to find the associated lemma for a Malay unknown word. These two Malay lemmatizers include one from Derry Tanti Wijaya developed in the present project at NUS (abbreviated as "Tan"), and one from [Baldwin et al. 2006][8] (abbreviated as "Baldwin"). Table I lists two examples of sentences after the lemmatization of unknown words by using these two lemmatizers. From the table, we observe that the results of the two lemmatizers are different. For example, the word *mengapa* is lemmatized as *kapa* by Tan's lemmatizer, and *apa* by Baldwin's. This variety enables us to investigate the effect of using different lemmatizers on translation quality. In addition, we also observe that word reduplication in Malay can be handled by both lemmatizers. For instance, the word *ketiga-tiga* is stemmed as *tiga* by Tan's lemmatizer, and *ketiga* by Baldwin's.

Table II lists twenty examples of unknown words with different stemmed forms, including ten single words and ten reduplicated words. Overall, judging from the table, we note that:

—Baldwin's lemmatizer is more aggressive than Tan's.

—Baldwin's lemmatizer generates some wrong lemmas. Take *pengguna-pengguna* as an example. The lemmatized word by Baldwin's (i.e., *una*) is non-existent in Malay. The same problem also occurs for the word *kenderaan-kenderaan*.

—Tan's lemmatizer misses some opportunities for further lemmatization.

For Indonesian-English machine translation, we use Aldrian Obaja Muis's Indonesian lemmatizer developed in the present project at NUS, which is modified from Tan's lemmatizer. Table III lists two examples of sentences lemmatized by

[8]http://code.google.com/p/malay-toklem/

Original word	Lemmatized by Tan's lemmatizer	Lemmatized by Baldwin's lemmatizer
penyampaian	sampai	nyampai
penyelenggaraan	selenggara	nyelenggara
kebajikan	kebaji	bajik
berwajaran	wajar	wajaran
menumpukan	menumpu	tumpu
penyenggaraan	senggara	nyenggara
mempengerusikan	kerusi	pengerusi
selebihnya	lebih	selebih
dinamisnya	dinamis	namis
memaksimumkan	maksimum	aksimum
ketiga-tiga	tiga	ketiga
kenderaan-kenderaan	kenderaan	ndera
pemandu-pemandu	mandu	pemandu
pengguna-pengguna	guna	una
keluaran-keluaran	keluar	luar
pengeluar-pengeluarnya	keluar	luar
pindaan-pindaan	pindaan	pinda
besar-besarannya	besar-besaran	besarannya
tele-pemasaran	tele-pasar	tele-pemasar
perdagangan-pelaburan	perdagangan-labur	dagangan-pelabur

Table II. **Examples of Malay lemmatized words by using different lemmatizers.**

Original sentence:	**klaim-klaim** itu belum bisa segera **dibuktikan kebenarannya** secara independen ...
Lemmatized sentence:	**klaim** itu belum bisa segera **bukti benar** secara independen ...
Original sentence:	indonesia , australia **tingkatkan pengamanan** perbatasan ...
Lemmatized sentence:	indonesia , australia **tingkat aman** perbatasan ...

Table III. **Examples of Indonesian sentences after the lemmatization of unknown words.**

the Indonesian lemmatizer. As listed in the table, the lemmatizer can also handle reduplicated words.

4. EXPERIMENTS

In this section, we first describe the experimental set up, which includes two datasets: one for Malay-English translation, another for Indonesian-English translation. We next describe the settings for the experiments. We finally present the experimental results and offer some observations and discussions.

	Malay-English		Indonesian-English	
Sentence pairs	296,230		32,518	
Language	Malay	English	Indonesian	English
Total # words	8,241,807	8,833,767	901,612	1,040,345
Vocabulary size	72,907	77,729	36,262	35,966

Table IV. **Statistics of the parallel texts.**

Dataset	Sentence pairs	Total # words		Vocabulary size	
Malay-English					
		Malay	English	Malay	English
Train	292,230	8,126,695	8,709,025	72,570	77,429
Dev	2,000	58,494	63,384	4,163	4,707
Test	2,000	56,618	61,358	4,757	5,327

Table V. **Statistics of the Malay-English datasets.**

4.1 Datasets

In our experiments, datasets of both language pairs (i.e., Malay-English and Indonesian-English) are used for the evaluation of our proposed method. The bi-texts of the two datasets are built from texts downloaded from the Internet. We first use *Google Translate* to translate the source language (i.e., Malay or Indonesian) texts into English. Next we match[9] the target language texts against the translated English texts using the measure of cosine similarity and several heuristic constraints based on document length in words and in sentences, overlap of numbers, words in uppercase, and words in the title. Finally, we extract pairs of sentences from the matched document pairs using *competitive linking* [Melamed 2000], and we keep the pairs whose similarity exceeds a pre-defined threshold. Table IV lists the statistics of the two constructed language pairs.

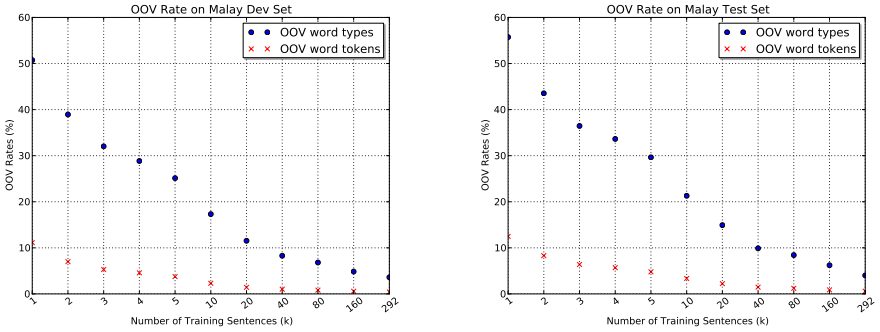
In addition, for each of the two language pairs, we separately take 2,000 parallel sentences as development and testing datasets, and the remaining is used as the training dataset. There is no overlap of parallel sentences between training, development, and testing datasets.

Table V lists the statistics of the datasets for the development of Malay-English machine translation, and Table VI lists those for Indonesian-English. Note that for Indonesian-English, we also filter out training sentences with length less than

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

Dataset	Sentence pairs	Total # words		Vocabulary size	
Indonesian-English					
		Indonesian	English	Indonesian	English
Train*	22,227	679,500	772,153	30,563	30,339
Dev	2,000	61,998	61,998	8,120	8,142
Test	2,000	56,541	63,437	8,215	8,474

Table VI. **Statistics of the Indonesian-English datasets.** *In order to remove URLs and advertisements, we filter out the sentences with length less than 15 or more than 60 tokens.



(a) OOV Rate on Malay Dev Set (b) OOV Rate on Malay Test Set

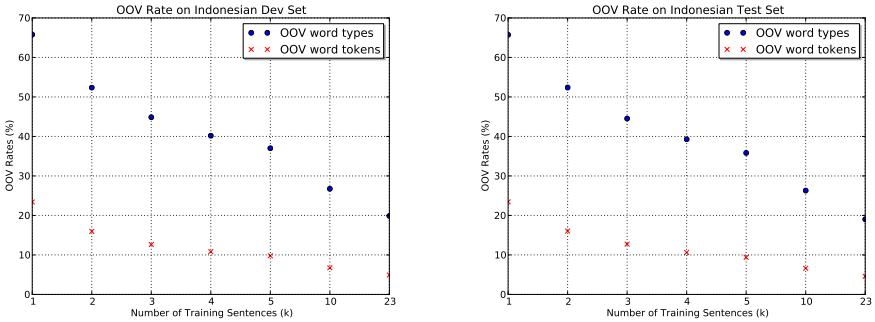
Fig. 1. **OOV Rate on Malay Datasets.**

15 or more than 60 tokens, in order to remove unsuitable texts such as URLs and advertisements. This results in 22,227 sentence pairs for training.

From these two tables, we expect that Indonesian-English translation will be more difficult than Malay-English translation. This is because the vocabulary size of development and testing sets in Indonesian-English is about two times that in Malay-English, while the vocabulary size of the Indonesian-English training set is about half that of Malay-English.

4.2 Experimental Settings

In order to explore the effect of the rate of out-of-vocabulary (OOV) words on translation quality, we further divide the training datasets into several different smaller sizes. Figure 1 shows the OOV rate of the Malay development and testing datasets, in terms of word types and word tokens. As observed from the figure, the OOV rate increases as the number of training sentences decreases. That is, the problem of data sparseness becomes more severe with fewer training sentences. Given fewer training sentences, we expect that our proposed method of lemmatizing unknown



(a) OOV Rate on Indonesian Dev Set (b) OOV Rate on Introduction Test Set

Fig. 2. **OOV Rate on Indonesian Datasets.**

words will be able to obtain more improvements. Note that with 1,000 training sentences, the OOV rate in word types exceeds 50% for the Malay development and test datasets, which means that half of the word types in the development and test sets are not present in the training set.

Figure 2 shows the OOV rate of the Indonesian development and testing datasets, in terms of word types and word tokens. Due to more diverse topics within the Indonesian dataset, the OOV rate is higher than that of the Malay dataset. With 1,000 training sentences, the OOV rate in terms of word types exceeds 60%. Hence, the problem of data sparseness in the Indonesian dataset is worse than that of Malay.

In addition, we also propose a hybrid approach that uses two copies of training sentences: one is the original sentence, and the other is the lemmatized one. Take the case of 1,000 training sentences as an example. The hybrid approach uses the original 1,000 sentences along with their lemmatized sentences as training data. In the following experimental results, for each trial of using a different number of training sentences, we report results of both the single-copy and hybrid approaches.

4.3 Experimental Results

In the following experiments, all results are reported in terms of the BLEU score, for which we use the default settings of mteval-v13a script[10]. Figure 3 plots the results of lemmatizing unknown words for Malay-English translation with varying sizes of training sentences. As observed from the figure, our proposed method begins to get some improvements over the corresponding baselines when the size of training data is 10,000 sentences or less, corresponding to an OOV rate of 20% or higher as shown in Figure 1. However, when the size of training data is at least 20,000 sentences,

[10] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl

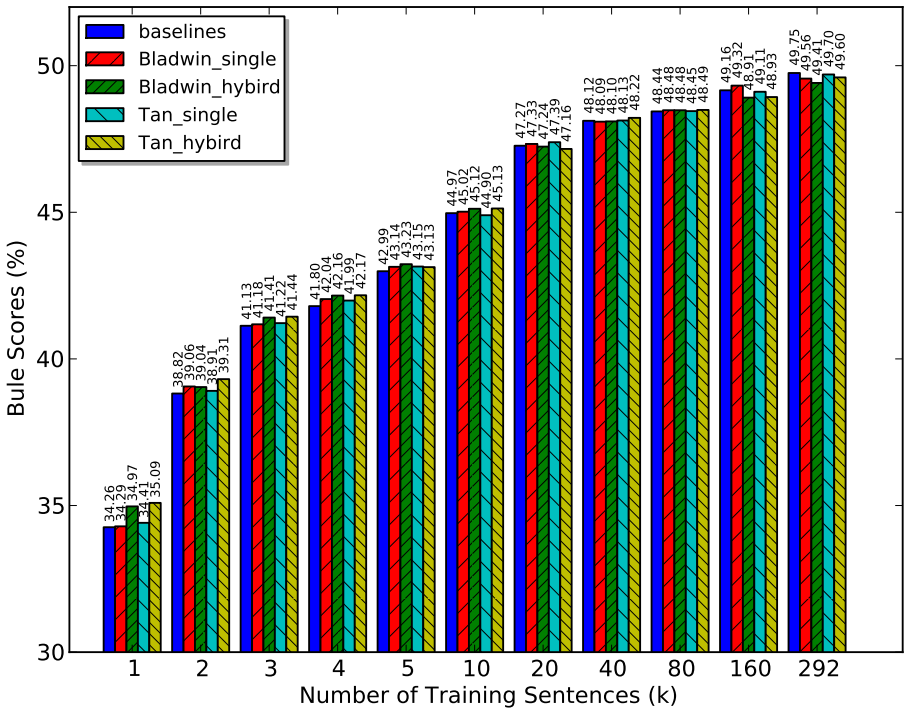


Fig. 3. **Experimental Results of Malay-English Translation.**

our method barely surpasses the corresponding baselines. A similar situation also occurs in the Indonesian-English experiments, as shown in Figure 4. Note that, as also shown in Figure 4, the trial with 5,000 training sentences unexpectedly fails to get improvement over the corresponding baseline. This exception may be due to the fact that the vocabulary within the Indonesian-English dataset is so diverse that it causes some uncertainties.

Furthermore, comparing the single approach with the hybrid approach, we also notice that in most cases, the improvement obtained from using the hybrid method is more than that of using the single method. This phenomenon appears in the two different language pairs, which suggests that the hybrid method alleviates the problem of data sparseness more effectively than the single method. However, for the cases with lower OOV rate, neither the single nor the hybrid method gets improvement.

In addition to the superiority of the hybrid method over the single method,

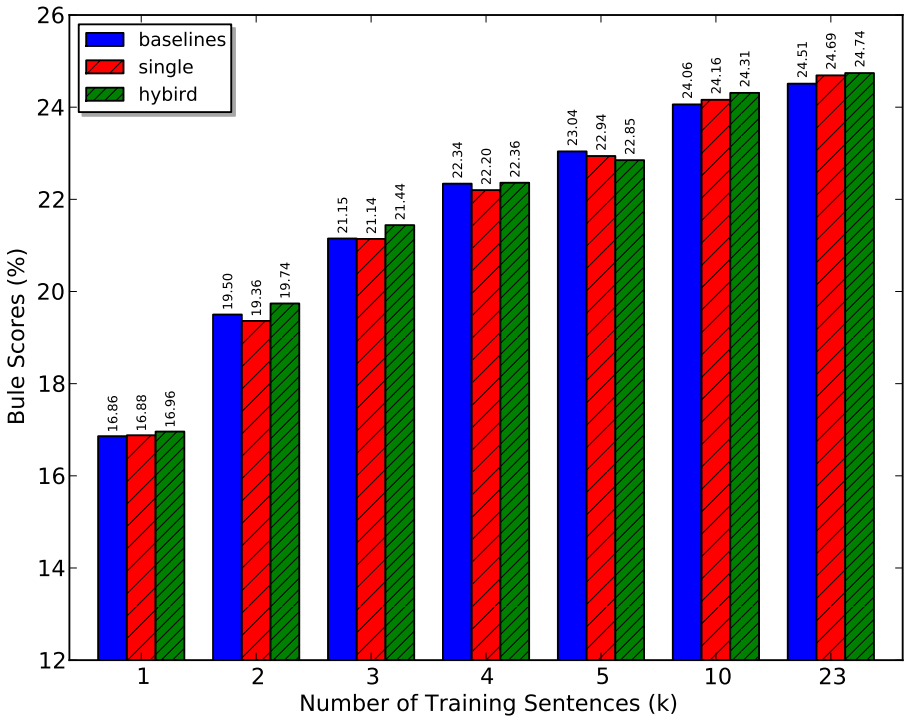


Fig. 4. **Experimental Results of Indonesian-English Translation.**

# Training sentences	Trials	p-value
5k	Baldwin_hybrid vs. Baseline	0.295
4k	Tan_hybrid vs. Baseline	0.217
3k	Tan_hybrid vs. Baseline	0.158
2k	Tan_hybrid vs. Baseline	0.123
1k	Tan_hybrid vs. Baseline	**0.025**

Table VII. **Results of Significance Tests on Malay-English Experiments.**

we also observe that for Malay-English translation, using Tan's lemmatizer for unknown words gets better translation quality than using Baldwin's. As shown in Figure 3, there are 9 out of 11 trials in which the performance of using Tan's lemmatizer is better than that of using Baldwin's. This phenomenon is due to the fact that, as mentioned in Section 3, Baldwin's lemmatizer is more aggressive such that it can produce wrong or non-existent lemmas.

# Training sentences	Trials	p-value
23k	Hybrid vs. Baseline	0.264
10k	Hybrid vs. Baseline	0.203
3k	Hybrid vs. Baseline	0.175
2k	Hybrid vs. Baseline	0.227
1k	Hybrid vs. Baseline	0.394

Table VIII. **Results of Significance Tests on Indonesian-English Experiments.**

In order to determine if the improvements are statistically significant, we conduct significance tests using the bootstrap resampling method [Koehn 2004]. Table VII lists the results of the tests and the corresponding p-values on five trials in the Malay-English experiments. As listed in the table, in the trial of using 1,000 training sentences, our hybrid method with Tan's lemmatizer outperforms the corresponding baseline significantly with a p-value of 0.025. This demonstrates that when data sparseness is severe and the OOV rate is high, our proposed method of lemmatizing unknown words for machine translation can give a statistically significant improvement. However, if the vocabulary size is too large and diverse, our proposed method still does not give significant improvement, as listed in Table VIII, in which all the five possible trials in the Indonesian-English experiments fail to give significant improvements.

5. CONCLUSION

This paper presents a study of utilizing morphological information for machine translation. The contribution of this work includes the proposal of lemmatizing unknown words for machine translation. In our experiments, two language pairs and three lemmatizers are used to determine the effectiveness of the proposed method. According to the experimental results, when the OOV rate is higher than 20%, our proposed method can relieve the problem of data sparseness, thereby enhancing the translation quality. One of our experimental trials even shows that the improvement is statistically significant, and the corresponding p-value is 0.025. Several research directions remain for future work:

—Considering that our translation model is rather simple, we will attempt to use more sophisticated translation models such as a back-off model, hierarchical model, confusion network, or word segmentation lattice in the future.

—According to our experimental results, the translation quality is also affected by the performance of the lemmatizer. So in future, we would also like to improve the

accuracy of the lemmatizer, and make it more suitable for machine translation.

—Another direction is to investigate how to better integrate the translation model and the lemmatizer. For instance, a lattice model can be employed to combine different lemmas of an unknown word for improving translation quality.

—Furthermore, we would also like to conduct an analysis on the Malay and Indonesian datasets, in order to determine what kinds of words can easily cause problem for translation. The analysis can help us better understand the problem.

REFERENCES

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. 1999. Statistical machine translation. Tech. rep., Final Report, JHU Summer Workshop.

Avramidis, E. and Koehn, P. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*. Columbus, Ohio, 763–770.

Baldwin, T., Awab, S., et al. 2006. Open source corpus analysis tools for Malay. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Citeseer, 2212–5.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics 19,* 2, 263–311.

Buckwalter, T. 2004. Buckwalter Arabic morphological analyzer version 2.0. Linguistic Data Consortium, Philadelphia.

Cartoni, B. 2009. Lexical morphology in machine translation: A feasibility study. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Association for Computational Linguistics, Athens, Greece, 130–138.

Casacuberta, F., Ney, H., Och, F. J., Vidal, E., Vilar, J. M., Barrachina, S., Garcia-Varea, I., Llorens, D., Martinez, C., and Molau, S. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 25–47.

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

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

Dyer, C. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 406–414.

Dyer, C., Muresan, S., and Resnik, P. 2008. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*. Columbus, Ohio, 1012–1020.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. 2004. What's in a translation rule? In *Proceedings of HLT-NAACL'04*. 273–280.

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

Habash, N. and Sadat, F. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York City, USA, 49–52.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP '05: Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Koehn, P. and Haddow, B. 2009. Edinburgh's submission to all tracks of the WMT2009 shared task with reordering and speed improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, 160–164.

Koehn, P. and Hoang, H. 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)*. 868–876.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic, 177–180.

Koehn, P., Och, F. J., and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Edmonton, Canada, 48–54.

Lee, Y.-S. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL 2004: Short Papers*, D. M. Susan Dumais and S. Roukos, Eds. Boston, Massachusetts, USA, 57–60.

Matusov, E., Kanthak, S., and Ney, H. 2005. On the integration of speech recognition and statistical machine translation. In *Proc. European Conference on Speech Communication and Technology*. 467–474.

Melamed, I. D. 2000. Models of translational equivalence among words. *Computational Linguistics 26,* 2, 221–249.

Minkov, E., Toutanova, K., and Suzuki, H. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, 128–135.

Ney, H. 1999. Speech translation: coupling of recognition and translation. In *ICASSP '99: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE Computer Society, Washington, DC, USA, 517–520.

Niessen, S. and Ney, H. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics 30,* 2, 181–204.

Novák, A. 2009. MorphoLogic's submission for the WMT 2009 shared task. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, 155–159.

Oflazer, K. and El-Kahlout, I. D. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*. 25–32.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.

Quirk, C., Menezes, A., and Cherry, C. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan, 271–279.

Ren, X. and Perrault, F. 1992. The typology of unknown words: an experimental study of two corpora. In *Proceedings of the 14th Conference on Computational Linguistics*. Association for Computational Linguistics, 408–414.

Saleem, S., chen Jou, S., Vogel, S., and Schultz, T. 2004. Using word lattice information for a tighter coupling in speech translation systems. In *Proceedings of the International Conference of Spoken Language Processing (ICSLP-2004)*. Jeju Island, Korea, 41–44.

Talbot, D. and Osborne, M. 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*. 969–976.

Toutanova, K., Suzuki, H., and Ruopp, A. 2008. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*. Columbus, Ohio, 514–522.

Weaver, W. 1955. Translation. *Machine Translation of Languages 14*, 15–23.

Yang, M. and Kirchhoff, K. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the European Chapter of the ACL*. 41–48.