Rule-based Machine Translation between Indonesian and Malaysian

Raymond Hendy Susanto[1] *Septina Dian Larasati*[2] *Francis M. Tyers*[3]
(1) Department of Computer Science, National University of Singapore
(2) Institute of Formal and Applied Linguistics, MFF, Charles University in Prague
(3) Dept. Lleng. i Sist. Inform., Universitat d'Alacant
raymondhs@nus.edu.sg, larasati@ufal.mff.cuni.cz, ftyers@dlsi.ua.es

ABSTRACT
We describe the development of a bidirectional rule-based machine translation system between Indonesian and Malaysian (id-ms), two closely related Austronesian languages natively spoken by approximately 35 million people. The system is based on the re-use of free and publicly available resources, such as the Apertium machine translation platform and Wikipedia articles. We also present our approaches to overcome the data scarcity problems in both languages by exploiting the morphology similarities between the two.

KEYWORDS: machine translation, Malay languages, morphology.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 191–200,
COLING 2012, Mumbai, December 2012.

1 Introduction

In this paper we describe the development of `apertium-id-ms`, a bidirectional Indonesian and Malaysian machine translation system based on the Apertium platform. The paper is laid out as follows: Section 2 gives a brief description of the two languages; Section 3 gives a short review of the previous work in Indonesian-Malaysian language pair; Section 4 describes the system and the creation of the resources; Section 5 presents an evaluation of the system, and finally we describes future work that could be done and some concluding remarks.

2 Languages

Indonesian (*Bahasa Indonesia*) and Malaysian (*Bahasa Malaysia*) are standards of the Malay language and both belong to the Austronesian family. Indonesian is spoken by approximately 35 million people[1], mostly from Indonesia, but also widespread in the Netherlands, the Philippines, Saudi Arabia, Singapore, and the United States. Malaysian has 10 million speakers across the Peninsular Malaysia, and other speakers coming from parts of Sarawak, Indonesia (Sumatra), Singapore, and United States.

Indonesian and Malaysian are closely related; both languages are mutually intelligible to a great extent. The morphology of the two languages are the same, where agglutination is used extensively (by means of affixation, reduplication, and compounding), although some affixes are more frequently used in one language over the other. For instance, the prefix *juru-* is often used in Malaysian to indicate an actor characterized by the stem it is attached to. While this affixation also happens in Indonesian, it is not as frequently found as in Malaysian.

The main difference between Indonesian and Malaysian languages lies in their vocabulary; Indonesian is largely influenced by Dutch and Javanese, whereas Malaysian has many words borrowed from English, e.g. *dokter* vs. *doktor* ('doctor') and *tas* vs. *beg* ('bag'). Moreover, there are frequent minor spelling differences, e.g. *kabar* vs. *khabar* ('news') and *mau* vs. *mahu* ('want').

3 Motivation

The development of this system is motivated by an early prototype Apertium system developed in Larasati and Kubon (2010). Since then, most of the components in the prototype were completely redesigned, and the direction of the development was based on the following rationale:

- Morphological analyser: We wanted to create a morphological analyser which is not only robust, but also does not overgenerate either. Both Indonesian and Malaysian morphology are characterized by rich derivational morphology, but poor inflectional morphology, unlike some other morphologically-complex languages such as Arabic and Turkish. Moreover, each root undergoes an idiosyncratic subset of these derivational processes (i.e. we cannot simply apply the derivational affixes to new words).
- Bidirectional: A new translation direction from Malaysian to Indonesian was added, and used as the starting point instead of the opposite direction, since translating from Indonesian to Malaysian appears to be more ambiguous.
- Evaluation: Finally, we performed a quality evaluation of our system, which has not been done before.

[1] `http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=89` Last accessed: October 2012

We have chosen the rule-based approach, instead of the ubiquitous corpus-based statistical approach, due to the dearth of parallel corpora for the two languages. Moreover, the closeness between the two languages makes the rule-based approach favourable. Recent development for closely related languages with the rule-based approach have shown competitive performance with respect to the statistical approach, e.g. the rule-based Swedish→Danish system in Tyers and Nordfalk (2009) and the Italian→Catalan system in Toral et al. (2011), where both systems outperform a rivaling statistical-based system.

4 System

The system is based on the Apertium machine translation platform (Forcada et al., 2011).[2] The platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other, more distantly related language pairs. The whole platform, both programs and data, are licensed under the Free Software Foundation's General Public Licence[3] (GPL) and all the software and data for the 33 supported language pairs (and the other pairs being worked on) is available for download from the project website.

4.1 Architecture of the system

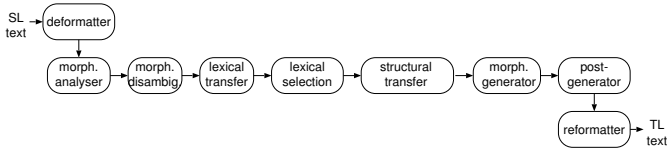

Figure 1: The pipeline architecture of the Apertium system.

The Apertium translation engine consists of a Unix-style *pipeline* or *assembly line* with the following modules (see Fig. 1):

- A *deformatter* which encapsulates the format information in the input as *superblanks* that will then be seen as blanks between words by the other modules.
- A *morphological analyser* which segments the text in surface forms (SF) (*words*, or, where detected, multi-word lexical units or MWLUs) and for each, delivers one or more *lexical forms* (LF) consisting of *lemma*, *lexical category* and morphological information.
- A *morphological disambiguator* (constraint grammar) which chooses, using linguistic rules the most adequate sequence of morphological analyses for an ambiguous sentence.
- A *lexical transfer* module which reads each SL LF and delivers the corresponding target-language (TL) LF by looking it up in a bilingual dictionary encoded as an FST compiled from the corresponding XML file. The lexical transfer module may return more than one TL LF for a single SL LF.
- A *lexical selection* module which chooses, based on context rules the most adequate translation of ambiguous source language LFs.

[2] `http://www.apertium.org/`
[3] `http://www.fsf.org/licensing/licenses/gpl.html`

- A *structural transfer* module which performs local syntactic operations, is compiled from XML files containing rules that associate an *action* to each defined LF *pattern*. Patterns are applied left-to-right, and the longest matching pattern is always selected.

- A *morphological generator* which delivers a TL SF for each TL LF, by suitably inflecting it.

- A *reformatter* which de-encapsulates any format information.

4.2 Morphological transducers

There is one publicly available morphological tool for Indonesian, MorphInd (Larasati et al., 2011). However, MorphInd is only designed for analysis, while we wanted a tool for both morphological analysis and generation. Moreover, there are rarely linguistic resources and tools available for Malaysian. Baldwin (2006) has developed the free/open-source lemmatiser for Malay, but this does not meet our need either since we wanted to include more robust morphological information into our Malaysian transducer, such as part-of-speech and affixes. Thus, we decided to build the morphological transducers from scratch.

Similar to most Apertium language pairs, the morphological transducers for both Indonesian and Malaysian are constructed using *lttoolbox*, a toolbox for morphological analysis and generation that is available under a free/open-source licence. The monolingual dictionary for each language is provided as XML-formatted entries, which is then compiled into a finite state transducers using *lttoolbox*.

4.2.1 Indonesian morphological transducer

A lexicon list was created semi-automatically to the Indonesian morphological analyser based on a words frequency list, with the most frequent words being added first. The frequency list was taken from a database dump of the Indonesian Wikipedia. For each word in the frequency list, we obtained its lemma and part-of-speech information from Kateglo[4], an online Indonesian dictionary with over 70,000 entries licensed under CC BY-SA 3.0.[5]. Since we also wanted to include affix information in our analyser, we wrote a rule-based morpheme segmentor to decompose a given Indonesian surface form into their constituent morphemes, also by making use of the lemma information from Kateglo. Moreover, closed word classes (e.g. pronouns, conjunctions) were added by hand.

4.2.2 Malaysian morphological transducer

A frequency list for Malaysian was also created based on a database dump of the Malaysian Wikipedia. Unlike Indonesian, we did not find a comprehensive Malaysian dictionary with adequate morphological information, such as lemma and part-of-speech. Hence, the Malaysian analyser was built using the two strategies below.

First, Malaysian words that also exist as an Indonesian word were assumed to share the same morphological information (i.e. the same lemma and part-of-speech), and were added automatically to the analyser. Although this method may introduce a number of false friends (e.g. *polisi* means 'policy' in Malay but 'police' in Indonesian), the benefit outweighs the risk

[4] http://kateglo.bahtera.org/
[5] http://creativecommonsorg/licenses/by-sa/3.0/

since there is huge overlap in the lexicons of the two languages. Moreover, most of these false friends usually belong to the same part-of-speech.

Second, we also added Malaysian words which appear in our bilingual dictionary. Since every entry in a bilingual dictionary is a pair of words with the same meaning, we can assume that these words also belong to the same part-of-speech most of the time. Our approaches to building the bilingual dictionary are presented in the following section.

4.3 Bilingual dictionary

There is no freely available bilingual dictionary between Indonesian and Malaysian, so we had to build the dictionary from scratch. At the moment, the bilingual dictionary contains 12,142 entries, which was developed in several ways described below.

First, most of the entries were added using automatic word alignments. We created an Indonesian-Malaysian parallel corpus by translating many articles taken from Malaysian Wikipedia. The translation process is mostly automatic, with the help of existing Malaysian-Indonesian machine translation systems such as Google Translate.[6]

Next the Wikipedia corpus is tagged using our morphological analyser, and word alignments were created by running `GIZA++` (Och and Ney, 2003) on the tagged corpus. We fed the probabilistic dictionary into the `ReTraTos` toolbox (Caseli et al., 2006), which extracts both phrases and single-word translations from alignments, and converts them into Apertium translation entries. The `ReTraTos` method gave us about 12,000 translation entries, but also required a manual check due the amount of noise in the resulting data.

Finally, some entries were added manually, which included closed word classes and words that frequently appeared in Wikipedia but were not yet added to the bilingual dictionary.

4.4 Disambiguation

The output from the morphological analysis is disambiguated using Apertium's statistical disambiguator module. The module implements a bigram part-of-speech tagger based on hidden Markov models (HMM). To improve the accuracy of our disambiguator, a Constraint Grammar (Karlsson, 1990) could be used as a pre-disambiguator module before feeding the input to the HMM, which is left for future work.

4.5 Lexical selection rules

Given the closeness of the languages, lexical selection is not a large problem between Indonesian and Malaysian. However, a number of rules can be written for ambiguous words; for example, the Malaysian preposition *daripada* 'from (to explain the origin of something), than (comparison)' can be translated into Indonesian as either *dari* 'from' or *daripada* 'than (comparison)', depending on the surrounding context.

Another example is the copulas *adalah* and *ialah* (both meaning 'be'), which exist in both Indonesian and Malaysian, but have a slightly different usage in each language. In Malaysian, *adalah* is used before an adjective phrase or a prepositional phrase, and *ialah* is used only before a noun phrase. In comparison to Indonesian, there are no strict rules governing the use of the two words, and their usage is more interchangeable.

[6] `http://translate.google.com/`

4.6 Transfer rules

There are barely differences between the grammar of Indonesian and Malaysian, in that the structure of words, phrases, clauses, and sentences are almost exactly the same. That said, the lexical transfer between the two languages works by simple word substitution in most cases.

(Malaysian) Input	Cuaca kelmarin amatlah sejuk.
Mor. analysis	^Cuaca/Cuaca<n><sg>$ ^kelmarin/kelmarin<adv>$ ^amatlah/amatlah<adv>$ ^sejuk/sejuk<adj>$ ^./.<sent>$
Mor. disambiguation	^Cuaca<n><sg>$ ^kelmarin<adv>$ ^amatlah<adv>$ ^sejuk<adj>$^.<sent>$
Transfer	^Cuaca<n><sg>$ ^kemarin<adv>$ ^amatlah<adv>$ ^dingin<adj>$^.<sent>$
Mor. generation	Cuaca kemarin amatlah dingin.

Table 1: Translation process for the sentence *Cuaca kelmarin amatlah sejuk.* 'The weather yesterday is very cold'.

5 Evaluation

The system was evaluated in three ways. The first was the coverage[7] of the system. The second was the word error rate (WER) of the translation output for our test data set, together with the error analysis of the translations. Lastly, we did a comparative evaluation with an existing system.

5.1 Coverage

Lexical coverage of the system is calculated over the Indonesian and Malaysian Wikipedia articles, as shown in Table 2. The database dump of the Indonesian Wikipedia[8] was from the 29th April 2012, and that of Malaysian Wikipedia[9] from the 28th April 2012. Both database dumps were stripped of formatting.

Corpus	**Tokens**	**Coverage**
Indonesian Wikipedia	19,021,087	80.70%
Malaysian Wikipedia	12,613,364	80.10%

Table 2: Naïve vocabulary coverage over Wikipedia articles.

5.2 Quantitative and Qualitative

We tested our system on a 2,084 word text taken from various articles in Malaysian Wikipedia. The translation quality was measured using Word Error Rate (WER), a metric based on the

[7] Here coverage is defined as *naïve coverage*, that is for any given surface form at least one analysis is returned by our monolingual dictionaries

[8] `http://id.wikipedia.org/`; `idwiki-20120429-pages-articles.xml.bz2`

[9] `http://ms.wikipedia.org/`; `mswiki-20120428-pages-articles.xml.bz2`

Levenshtein distance (Levenshtein, 1966). We calculated the WER for each sentence using the `apertium-eval-translator`[10] tool. The WER metric was preferred to other MT metrics such as BLEU (Papineni et al., 2002) since we want to evaluate the system for the *postedition* task. That is, we want to assess the amount of manual labour needed to improve the machine-generated translation.[11]

For the Malaysian to Indonesian direction, the sentences were translated by the system, and then postedited by a native Indonesian speaker. For the Indonesian to Malaysian direction, we used the reference translation, as postedited by the native speaker and used it as a source of Indonesian to be translated to Malaysian, then the original Malaysian sentence was used as the reference translation.

Corpus	Direction	Tokens	Unknown	WER
Malaysian Wikipedia	id→ms	2,079	211	14.43% (83.89%)
	ms→id	2,084	256	7.58% (69.53%)

Table 3: Word error rate over the Malaysian Wikipedia test data. Number in parentheses gives percentage of unknown words which were free rides.

We consider the WER of our system, as depicted in Table 3 is quite acceptable for postediting. In our system, unknown words are left unprocessed. Nonetheless, many of these unknown words are *free rides*[12], which will not affect the final quality of the translation.

As a direction for future improvement, we did an error analysis by reviewing the translation outputs from our system. Most of the translation errors were due to the mistakes and gaps in our analyser. Specifically, many of the stems do not have the complete set of its derived forms. As a result, it cannot provide analysis for a unknown derived wordform even if the stem already exists in the analyser. Moreover, the analyser cannot handle clitics attached to unknown words (e.g. possessive enclitic *-nya*). These errors can be fixed by a more thorough revision of the morphological analyser. Lastly, the system is often not capable of choosing the most suitable translation given a particular context. A lexical selection module can be written to alleviate this problem.

Dir.	System	WER
id→ms	Apertium	14.43%
	Google	13.90%
ms→id	Apertium	7.58%
	Google	4.07%

Table 4: Accuracy comparison between the two systems.

[10] `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-eval-translator/`

[11] BLEU is not used in our evaluation because we are using a reference translation which is a postedition of the machine-translated text, and the normal use of BLEU is for evaluating against a non-postedited reference. If we used BLEU it would give artifically high scores.

[12] That is, the word is unknown to the system, but the same in Indonesian and Malaysian. Typical free-rides include names and special terminology.

5.3 Comparative

We compared our system to another MT system for Indonesian to Malaysian and Malaysian to Indonesian, Google Translate, a web-based statistical machine translation system. The evaluation was performed the same way: the test data was translated with Google Translate, then postedited.

We notice from Table 4 that Google outperforms the Apertium system in both translation directions. For Malaysian to Indonesian, the error rate is reduced by almost a half. It is also interesting that both systems perform almost as worse for Indonesian to Malaysian, perhaps due to the fact that translating from Indonesian to Malaysian is more ambiguous than translating to Indonesian. Google seems to have a greater vocabulary coverage than Apertium, as exemplified in the first example in Table 5. The Malaysian word *mencuba* ('try') is unknown to the Apertium system (denoted by the asterisk), whereas it is correctly translated by Google as *mencoba*.

Moreover, in many cases, Google seems to be performing better in picking the most natural translations. In the second example, it translates the noun phrase *kuasa ketenteraan* ('military power') to *kekuatan militer*, which sounds much more natural than *kuasa ketentaraan*.

Source	Para saintis **mencuba** menjawab pertanyaan tersebut.
Apertium	Para ilmuwan ***mencuba** menjawab pertanyaan tersebut.
Google	Para ilmuwan **mencuba** menjawab pertanyaan tersebut.
Reference	Para ilmuwan **mencoba** menjawab pertanyaan tersebut.
	'Scientists are trying to answer that question.'
Source	Sebuah **kuasa ketenteraan** yang disegani.
Apertium	Sebuah **kuasa ketentaraan** yang disegani.
Google	Sebuah **kekuatan militer** yang disegani.
Reference	Sebuah **kekuatan militer** yang disegani.
	'A respectable military power.'

Table 5: Example translations from Malaysian to Indonesian.

Conclusion and future work

We have presented a bidirectional rule-based machine translation system between Indonesian and Malaysian, two closely-related Malay languages. The system is available as free/open-source software under the GNU GPL and the whole system may be downloaded from SVN.[13]

The resulting system provides comparable results with a leading corpus-based machine translation system, and we are looking forward to improving the translation quality of our system in the future. The long-term plan is to integrate the data created to make transfer systems with more distantly related languages such as Indonesian-English and Malaysian-English.

Acknowledgments

Development of the system was funded as part of the Google Summer of Code[14], an annual program sponsored by Google, Inc. to promote students' participation in open-source software projects.

[13] `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-id-ms`
[14] `http://code.google.com/soc/`

References

Baldwin, T. (2006). Open source corpus analysis tools for Malay. In *In Proc. of the 5th International Conference on Language Resources and Evaluation*.

Caseli, H. M., Nunes, M. D., and Forcada, M. L. (2006). Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20(4):227–245.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING '90, pages 168–173.

Larasati, S. and Kubon, V. (2010). A study of Indonesian-to-Malaysian MT system. In *Proceedings of the 4th International MALINDO Workshop, Depok*, pages 16–22.

Larasati, S., Kuboň, V., and Zeman, D. (2011). Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. *Systems and Frameworks for Computational Morphology*, pages 119–129.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics—Doklady 10, 707–710. Translated from Doklady Akademii Nauk SSSR*, pages 845–848.

Lewis, M. P. e. (2009). *Ethnologue: Languages of the World, Sixteenth edition*. Dallas, Tex.: SIL International.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Toral, A., Ginestí, M., and Tyers, F. M. (2011). An Italian to Catalan RBMT system reusing data from existing language pairs. In *Proceedings of the Second Workshop on Free/Open-Source Rule-Based Machine Translation*.

Tyers, F. M. and Nordfalk, J. (2009). Shallow-transfer rule-based machine translation for Swedish to Danish. In *Proceedings of the First Workshop on Free/Open-Source Rule-Based Machine Translation*.