Rule-Based Machine Translation between Indonesian and Malaysian

Raymond Hendy Susanto, Septina Dian Larasati, Francis M. Tyers

raymondhs@nus.edu.sg, larasati@ufal.mff.cuni.cz, ftyers@dlsi.ua.es

1 Introduction

Indonesian (Bahasa Indonesia) and Malaysian (Bahasa Malaysia) are standards of the Malay language, a major language of the Austronesian family.

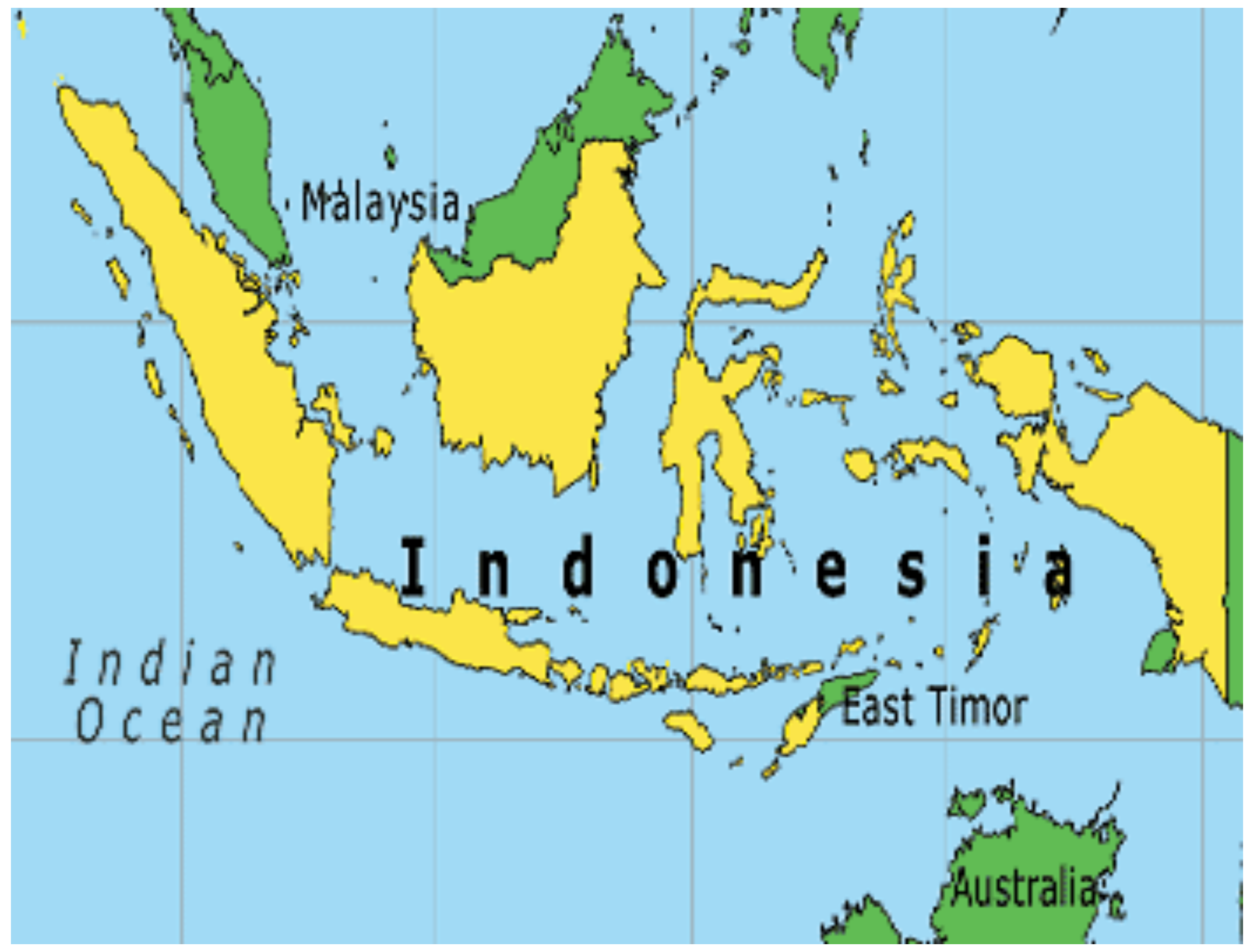

Figure 1: Map of Indonesian and Malaysian

Indonesian is natively spoken by about 35 million people; Malaysian is natively spoken by about 10 million people.

1.1 Language characteristics

Indonesian and Malaysian are characterized by:

- Mutually intelligible;
- Similar grammar;
- 50% overlap in vocabulary;
- Rich morphology

1.2 Why rule-based approach?

Rule-based approach is preferred to the statistical approach, since:

- No parallel corpora;
- Both languages are closely related;
- Simple word substitution works most of the time

2 System

The system is based on **Apertium** (http://www.apertium.org/), a free/open-source rule-based machine translation platform.

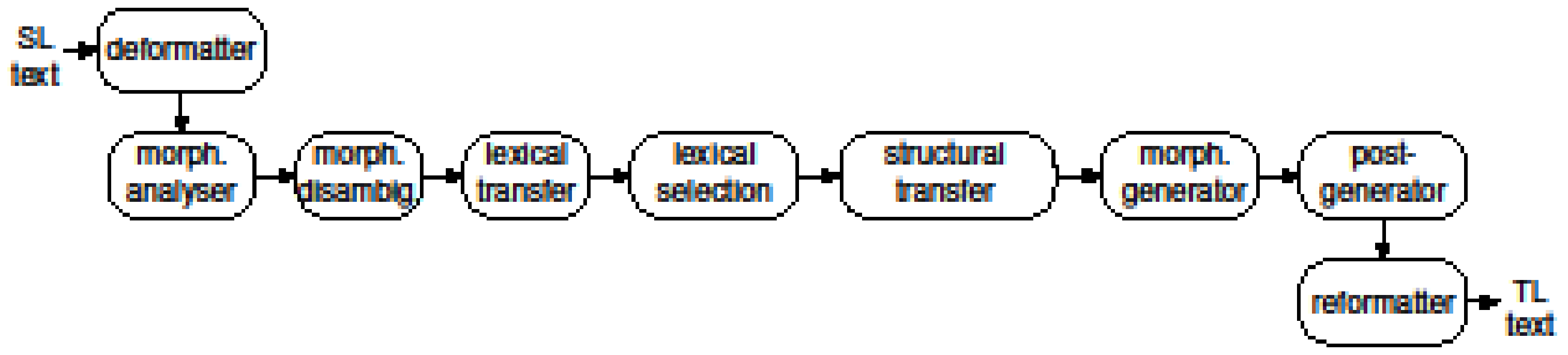

Figure 2: Modules of the Apertium translation system

2.1 Morphological analysis

Our morphological analyser returns, for every Indonesian/Malay word, the possible lexical forms (analyses) of the word. For the Indonesian sentence *Yakobus dan Maria sedang berada di kebun itu.* ('James and Mary are in that garden.')

```
^Yakobus/Yakobus<np><m><sg>$
^dan/dan<cnjcoo>$
^Maria/Maria<np><f><sg>$
^sedang/sedang<adv>$
^berada/ada<vblex><ber>$
^di/di<pr>$
^kebun/kebun<n><sg>$
^itu/itu<det><dem>/itu<prn><dem>$
^./.<sent>$
```

Figure 3: Output of morphological analysis with a finite-state transducer

The analysers have a coverage over 80% for the Wikipedia corpora.

2.2 Part-of-speech tagging

The part-of-speech tagging module for the system is based on a bigram HMM-based part-of-speech tagger. It can be trained on a database dump of the Indonesian Wikipedia (http://id.wikipedia.org/) and the Malaysian Wikipedia (http://ms.wikipedia.org/)..

2.3 Bilingual dictionary

The bilingual dictionary, or transfer lexicon contains mappings between lemmas, parts-of-speech and other tags.

```
<e><p><l>mobil<s n="n"/></l>
      <r>kereta<s n="n"/></r></p></e>
<e><p><l>mobilitas<s n="n"/></l>
      <r>mobiliti<s n="n"/></r></p></e>
<e><p><l>sepeda<s n="n"/></l>
      <r>basikal<s n="n"/></r></p></e>
<e><p><l>modern<s n="adj"/></l>
      <r>moden<s n="adj"/></r></p></e>
<e><p><l>karena<s n="cnjsub"/></l>
      <r>kerana<s n="cnjsub"/></r></p></e>
```

Figure 4: Extract from bilingual dictionary

3 Evaluation

We evaluated our system using word error rate (WER) on a corpus of 2,000 tokens from Malaysian Wikipedia.

Direction	WER
Indonesian-to-Malaysian	14.43%
Malaysian-to-Indonesian	7.58%

Table 1: Word error rate of the system

4 Future Work

- Improving coverage;
- Lexical selection;
- More transfer rules

Acknowledgments

This work was made during the Google Summer of Code 2012, a program for introducing students to free/open-source software development.