MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian

Hiroki Nomoto* Hannah Choi° David Moeljadi° Francis Bond°

*Tokyo University of Foreign Studies, °Nanyang Technological University

7 May 2018
The 13th Workshop on Asian Language Resources (ALR13)
Morphological dictionaries in NLP

- Lemmatization is an important task for morphological analysis.
- A good dictionary with wide coverage is crucial to the success of a robust morphological analysis, which in turn becomes the basis for higher-level tasks such as syntactic parsing.
- Open dictionaries for Japanese
  - NAIST Japanese Dictionary (IPAL)
  - UniDic
- Nothing comparable exists for Malay/Indonesian.
- So we created a morphological dictionary for Malay/Indonesian: MALINDO Morph
Organization

1. Malay and Indonesian
   - Their relationship
   - Morphology
2. Existing tools and their problems
3. MALINDO Morph and its creation
4. Ways of using MALINDO Morph
5. Future work
Malay and Indonesian

- The “Malay” language (msa\textsuperscript{1}): official language of four countries in the Malay Archipelago.
- Two regional varieties:
  - Malay in the narrow sense (zsm\textsuperscript{1}), used in Malaysia, Brunei and Singapore
  - Indonesian (ind\textsuperscript{1}), used in Indonesia
- Many tools and resources have been independently developed in each region.
- But the languages are mutually intelligible (about 10\% lexical difference (Asmah, 2001)) and share the same set of affixes.
  \[ \Rightarrow \text{A common morphological dictionary can be developed.} \]\n
\[ ^{1}\text{ISO693-3} \]
Malay/Indonesian morphology involves the use of

- Affixation
- Reduplication
- Cliticization
Affixation

- Productive: Prefixes, suffixes and circumfixes
- Non-productive: Infixes

(1)  

a. **Prefix**
   
   batas ‘limit’ + ter- $\rightarrow$ **terbatas** ‘limited’

b. **Suffix**
   
   batas ‘limit’ + -an $\rightarrow$ batasan ‘limitation’

c. **Circumfix**
   
   batas ‘limit’ + peN- -an $\rightarrow$ **pembatasan** ‘delimiting’
Reduplication

- Productive: Full reduplication
- Semi-productive: Partial and rhythmic reduplication

(2)  

a. Full reduplication  
   *kucing* ‘cat’ → *kucing-kucing* ‘cats’

b. Rhythmic reduplication  
   (vowel and/or consonant alternation)  
   *gunung* ‘mountain’ → *gunug-ganang* ‘mountain range’

c. Partial reduplication  
   (base-initial consonant + e + base)  
   *mula* ‘to start’ → *memula* ‘at first’  
   (Malay)
Cliticization

- Proclitics
- Enclitics

(3)  
a. **Proclitic (before the base)**  
*terima* ‘to receive’ + *ku* = ‘I’  
→ **kuterima** ‘I receive’

b. **Enclitics (after the base)**  
*buku* ‘book’ + =*ku* ‘me/my’  
→ **bukuku** ‘my book’
Interaction of different morphological processes

\[ \begin{align*}
  \text{batas} & \quad \text{‘limit’} \\
  \downarrow & \quad +\text{affixation: } \text{ter-} \\
  \text{terbatas} & \quad \text{‘limited’} \\
  \downarrow & \quad +\text{affixation: } \text{ke- -an} \\
  \text{keterbatasan} & \quad \text{‘limitation’} \\
  \downarrow & \quad +\text{reduplication} \\
  \text{keterbatasan-keterbatasan} & \quad \text{‘limitations’}
\end{align*} \]
Existing morphological dictionaries

- No large dictionary file is publicly available in an accessible format.
- One can create a larger dictionary by using the data from online dictionaries.
- However, no existing dictionary contains all the kinds of morphological information that MALINDO Morph offers: affixes, clitics and reduplication types.
Existing morphological analysers

**Stemmers/lemmatizers**
- Identify the stem/lemma.
- Much work has been done (Baldwin and Su’ad, 2006; Adriani et al., 2007; Larasati et al., 2011; Mohamad Nizam et al., 2016).

**Morphological analysers**
- Also analyse the non-stem/lemma strings.
- MorphInd (Larasati et al., 2011) seems to be the most sophisticated morphological analyser.
MorphInd (Larasati et al., 2011)

MorphInd identifies morpheme boundaries and assigns two POS tags to a token:

1. ‘Lemma tag’ (POS tag for the lemma)
2. ‘Morphological tag’ (POS tag for the entire token)

(4) a. Input: mengirim ‘to deliver’
   b. Output: meN+kirim<_VSA
      _VSA: lemma tag for verbs
      _VSA: morphological tag indicating that the entire token
      is a singular active verb
A common misunderstanding among NLP researchers: Circumfix ≡ prefix + suffix

- Circumfixes are incorrectly thought of as a combination of a prefix and a suffix.
- MorphIndo does not specify whether the non-lemma strings are a prefix, suffix or circumfix.

(5)  
a. Input: **pengiriman** ‘delivery’ (= *kirim* + circumfix *peN*-an)  
b. Output: **peN+kirim<v>+an**_NSD_  
- Not obvious whether **peN** and **an** are a combination of two morphemes (prefix *peN*- and suffix -an) or a single morpheme (circumfix *peN*- -an)...
Circumfix or “prefix + suffix”?

- The correct identification of circumfixes presents a major challenge to morphological analysis in Malay/Indonesian.
- A correct circumfix cannot be identified by just looking at the two strings at the left and right edges of a token.

(6) \textit{berakhir}an ‘suffixed’

\textbf{NOT} \textit{akhir} + circumfix \textit{ber-} \textit{-an}

\textbf{BUT} [\textit{akhir} + suffix \textit{-an}] + prefix \textit{ber-}
MALINDO Morph and its format

- Available at https://github.com/matbahasa/MALINDO_Morph
- Licensed under a CC BY 4.0 license.
- Version 20180418 has 232,516 lines (case-sensitive).
- Each line is made up of:
  - ID
  - Root
  - Surface form
  - Prefix(es), proclitic
  - Suffix(es), enclitic(s)
  - Circumfix(es)
  - Reduplication type
- Also include the analyser: morph_analyzer.py
Example: *perlu* ‘necessary’ and its derivatives

<table>
<thead>
<tr>
<th>Root</th>
<th>Surface form</th>
<th>Prefix</th>
<th>Suffix</th>
<th>Circumfix</th>
<th>Reduplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>perlu</td>
<td>perlu</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>perlu</td>
<td>seperlunya</td>
<td>0</td>
<td>0</td>
<td>se- -nya</td>
<td>0</td>
</tr>
<tr>
<td>perlu</td>
<td>memerlukan</td>
<td>meN-</td>
<td>-kan</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>perlu</td>
<td>perlu- memerlukan</td>
<td>meN-</td>
<td>-kan</td>
<td>0</td>
<td>R-full</td>
</tr>
<tr>
<td>perlu</td>
<td>keperluan</td>
<td>0</td>
<td>0</td>
<td>ke- -an</td>
<td>0</td>
</tr>
</tbody>
</table>
Two steps in building MALINDO Morph

1. **Core dictionary**
   Entries from the authoritative dictionaries in Malaysia and Indonesia (*Kamus Dewan*⁴ (KD) and *Kamus Besar Bahasa Indonesia*⁵ (KBBI))

   we would like to thank them for their cooperation

2. **Expanded dictionary**
   Other tokens found in the reclassified version of the Leipzig Corpora Collection for Malay and Indonesian (LCC; Goldhahn et al., 2012; Nomoto et al., under review)
Sizes of the MALINDO Morph dictionaries (unit: line)

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Checked</th>
<th>Unchecked</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>84,404</td>
<td>0</td>
<td>84,404</td>
</tr>
<tr>
<td>Expanded</td>
<td>47,400</td>
<td>100,712</td>
<td>148,112</td>
</tr>
<tr>
<td>Total</td>
<td>131,804</td>
<td>100,712</td>
<td>232,516</td>
</tr>
</tbody>
</table>
The morphological analysis of the core dictionary

- The morphological analyses were conducted using Microsoft Excel functions.
- The results were manually checked by Japanese undergraduate students of Malay/Indonesian, Indonesian research students and the first and second authors of the present paper.
- When the analyses provided by KD and KBBI differed from each other or were not precise as linguistic analyses, we adopted our own analyses.
- Hence, our core dictionary is not identical to either KD or KBBI.
Expanded dictionary

- Tokens that are not in the core dictionary were taken from the reclassified version of LCC.
- 300K (= 300K sentences) subset files × 16 (Malay 3, Indonesian 13)
- 1,005,007 word types (case-sensitive)
- Genuine Malay/Indonesian words, proper names, abbreviations, spelling variants/errors, foreign words and non-alphabets.
- Only tokens with frequency greater than ten in one of the sixteen subset files were further processed.
Frequent words in LCC

Total: 282,186 words
- English words: 57,633 → not included in MALINDO Morph
- Non-alphabets: 76,638 → not included in MALINDO Morph
- The others: 147,915 → analysed using the morphological analyser and checked by hand (ongoing)
Other items in the expanded dictionary

- Words in the core dictionary that can also be analysed as involving an enclitic.
- Handled manually → added to the “checked” category of the expanded dictionary.

(7) **penanya**

a. Core dictionary
   
   \[\text{penanya} = \text{Root } tanya \text{ ‘ask’ } + \text{prefix } peN- \text{ (‘questioner’)}\]

b. Expanded dictionary
   
   \[\text{penanya} = \text{Root } pena \text{ ‘pen’ } + \text{enclitic } =nya \text{ ‘his/her’ (‘his/her pen’)}\]
Limitations

- MALINDO Morph only targets productive native affixes and reduplication, but not borrowed affixes (with a few exceptions).
- No distinction is made between the suffix `-nya` and the enclitic `=nya`.
Morphological analyser: Preparation

1. **rootlist**: A list of roots in the core dictionary (core-dic).

2. **hyp-dic**: A hypothetical dictionary consisting of the basic and *di-* passive forms corresponding to the *meN-* verbs in core-dic.

- The forms in hyp-dic were created automatically and are merely hypothetical.
- They were added to the expanded dictionary (exp-dic) only if they were found to actually be used in the corpus.
Morphological analyser: The algorithm I

- Input $W$
- An ‘analysis’ is a list of the format
  \[ \langle \text{affix candidate, root, remaining string before root, remaining string after root, reduplication} \rangle. \]

1. Handle non-alphabets.
2. Handle English words.
3. Handle words present in *core-dic/hyp-dic*.
4. Strip $W/w$ of clitic strings. ($w$: $W$ in lower case)
5. Generate candidate sets $Cand_c$, $Cand_p$ and $Cand_s$, where $Cand_a$ is a set of candidate analyses for token $w$ based on affix/clitic type $a$
   \[ \in \{ c(ircumfix), p(refix/proclitic), s(uffix/enclitic) \}. \]
Morphological analyser: The algorithm II

6. Search $C_{andc} \times C_{andp} \times C_{ands}$ for members whose elements are mutually compatible.

7. Return $\langle root_c, w, p-, -s, c_1- -c_2, red_c \rangle$ for every such member.
Example: *sedianya* ‘actually’ I

Suppose the word were not in core-dic.

**Step 5: Candidate generation**

\[
\begin{align*}
C_{and_c} &= \left\{ \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset \rangle, \\
&\quad \langle \text{se}-\text{-nya}, \text{dia}, \emptyset, \emptyset, \emptyset \rangle \right\} \\
C_{and_p} &= \left\{ \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset \rangle, \\
&\quad \langle \text{se}, \text{dia}, \emptyset, \text{nya}, \emptyset \rangle \right\} \\
C_{and_s} &= \left\{ \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset \rangle, \\
&\quad \langle \text{-nya}, \text{sedia}, \emptyset, \emptyset, \emptyset \rangle, \langle \text{-nya}, \text{dia}, \text{se}, \emptyset, \emptyset \rangle \right\}
\end{align*}
\]
Example: *sedianya* ‘actually’ II

Step 6: Search $\text{Cand}_c \times \text{Cand}_p \times \text{Cand}_s$ for mutually compatible members

1. \[(\langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \text{-nya}, \text{sedia}, \emptyset, \emptyset, \emptyset \rangle)]

2. \[(\langle \text{se- -nya}, \text{dia}, \emptyset, \emptyset, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se, nya}, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se, nya}, \emptyset \rangle)]
Example: *sedianya* ‘actually’ III

**Step 7: Output**

1. $\langle \text{sedia, sedianya, } \emptyset, -\text{nya, } \emptyset \rangle$
2. $\langle \text{dia, sedianya, } \emptyset, \emptyset, \text{se- -nya, } \emptyset \rangle$

(The second output will be rejected by human checking.)
Conclusions

- With MALINDO Morph, stemming/lemmatizing frequent words in Malay/Indonesian will become a simple dictionary lookup with an additional disambiguation process for morphologically ambiguous words.
- The development of stemmers, lemmatizers and root identifiers should then focus on infrequent words.
- MALINDO Morph provides useful information for other tasks. E.g., POSs can be partly predicted from the outermost affix of a word:
  - `meN-` → verb (active)
  - `per- -an` → noun
  - `se- -nya` → adverb, ...

Nomoto, Choi, Moeljadi, Bond
Future work

In the future, the MALINDO Morph dictionary can be enriched by adding more linguistic information.

- Distinction between the suffix *-nya* (forming adverbials, nominalizing verbs and adjectives, occurring in exclamatives) and the enclitic *=nya* (3rd person pronoun, definite marker)
- Information about the variety, i.e. Malay, Indonesian and their dialects
- POSs
- Frequency of forms and derivations
References I


Nomoto, Hiroki, Shiro Akasegawa, and Asako Shiohara. under review. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian.