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**[1]): official language of four countries in the Malay Archipelago.
- Two regional varieties:
  - Malay in the narrow sense (**zsm**[1]), used in Malaysia, Brunei and Singapore
  - Indonesian (**ind**[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.

$\Rightarrow$ A common morphological dictionary can be developed.

[1] ISO693-3

Malay/Indonesian Morphology

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-* → *terbatas* 'limited'

b. Suffix
batas 'limit' + *-an* → *batasan* 'limitation'

c. Circumfix
batas 'limit' + *peN- -an*
→ *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

batas	‘limit’
↓	+affixation: *ter-*
terbatas	‘limited’
↓	+affixation: *ke- -an*
keterbatasan	‘limitation’
↓	+reduplication
keterbatasan-keterbatasan	‘limitations’

Existing morphological dictionaries

- No large dictionary file is publicly available in an accessible format.
- Baldwin and Su'ad's (2006) Malay tokenizer/lemmatizer: Word-lemma-POS triples for 2,499 words.
- 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<v>_VSA`

`<v>`: 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) *berakhiran* ‘suffixed’

NOT *akhir* + circumfix *ber- -an*

BUT [*akhir* + suffix *-an*] + prefix *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

Root	Surface form	Prefix	Suffix	Circumfix	Reduplication
perlu	perlu	0	0	0	0
perlu	seperlunya	0	0	se- -nya	0
perlu	memerlukan	meN-	-kan	0	0
perlu	perlu-memerlukan	meN-	-kan	0	R-full
perlu	keperluan	0	0	ke- -an	0

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)

Dictionary	Checked	Unchecked	Total
Core	84,404	0	84,404
Expanded	47,400	100,712	148,112
Total	131,804	100,712	232,516

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 $\times$ 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
penanya = Root *tanya* 'ask' + prefix *peN-*
('questioner')

b. Expanded dictionary
penanya = Root *pena* 'pen' + enclitic *=nya* '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
  $\langle$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 $Cand_c \times Cand_p \times Cand_s$ for members whose elements are mutually compatible.
7. Return $\langle root_c, w, p\text{-}, \text{-}s, c_1\text{- -}c_2, red_c \rangle$ for every such member.

Example: *sedianya* 'actually' I

Suppose the word were not in core-dic.

Step 5: Candidate generation

$$Cand_c = \left\{ \begin{array}{l} \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset \rangle, \\ \langle \text{se- -nya}, \text{dia}, \emptyset, \emptyset, \emptyset \rangle \end{array} \right\}$$

$$Cand_p = \left\{ \begin{array}{l} \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset \rangle, \\ \langle \text{se-}, \text{dia}, \emptyset, \text{nya}, \emptyset \rangle \end{array} \right\}$$

$$Cand_s = \left\{ \begin{array}{l} \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset \rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset \rangle, \\ \langle \text{-nya}, \text{sedia}, \emptyset, \emptyset, \emptyset \rangle, \langle \text{-nya}, \text{dia}, \text{se}, \emptyset, \emptyset \rangle \end{array} \right\}$$

Example: *sedianya* 'actually' II

Step 6: Search $Cand_c \times Cand_p \times Cand_s$ for mutually compatible members

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

Example: *sedianya* 'actually' III

Step 7: Output

1. ⟨sedia, sedianya, ∅, -nya, ∅⟩
2. ⟨dia, sedianya, ∅, ∅, se- -nya, ∅⟩

(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, ...

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

KD[4]. 2005. *Kamus Dewan*. Kuala Lumpur: Dewan Bahasa dan Pustaka, 4th edition.

KBBI[5]. 2016. *Kamus Besar Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa, 5th edition.

Adriani, Mirna, 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 (TALIP)* 6:1–33.

Asmah Haji Omar. 2001. The Malay language in Malaysia and Indonesia: From lingua franca to national language. *The Aseanists ASIA* II.

Baldwin, Timothy, and Su'ad Awab. 2006. Open source corpus analysis tools for Malay. In *Proceedings, the 5th International Conference on Language Resources and Evaluation (LREC2006)*, 2212–2215.

References II

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Larasati, Septina Dian, Vladislav Kuboň, and Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Systems and Frameworks for Computational Morphology*, ed. Cerstin Mahlow and Michael Piotrowski, 119–129. Verlag: Springer.

Mohamad Nizam Kassim, Mohd Aizaini Maarof, Anazida Zainal, and Amirudin Abdul Wahab. 2016. Word stemming challenges in Malay texts: A literature review. In *2016 4th International Conference on Information and Communication Technology (ICoICT)*, 1–6.

References III

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