Towards genuine stemming and lemmatization in Malay/Indonesian

Hiroki Nomoto
Tokyo University of Foreign Studies
nomoto@tufs.ac.jp

Abstract

Existing stemmers and lemmatizers for Malay/Indonesian identify stems/lemmas with roots and hence do not provide stems/lemmas in the normal sense. This paper discusses what actually count as stems/lemmas in Malay/ Indonesian and reports how stem/lemma information was added to MALINDO Morph, a morphological dictionary consisting of more than 230K surface forms.

1 Introduction

Stemming and lemmatization are among the most basic operations in computational processing of language data. They both turn a morphologically complex surface form of a word into more basic units: stem and lemma. By so doing, they enable dealing with different inflected forms as a group. Thus, a stemmer will group together {*cat*, *cats*} as *cat* and {*big*, *bigger*, *biggest*} as *big*. Similarly, a lemmatizer will group together {*take*, *takes*, *took*, *taking*, *taken*} as *take*. Grouping enabled by stemming and lemmatization is quintessential in areas involving word meanings, such as information retrieval.

A number of tools have been developed to handle these operations in various languages. However, not all such tools are designed based on a common understanding of the notions of 'stem' and 'lemma'. Different definitions exist for these notions. Moreover, the same stem can be represented differently. For instance, the stem for *taking*, namely [teɪk], may be represented by *take* (the standard orthography) or *tak* (the strings corresponding to the stem). Similarly, the choice of a specific lemma representation is in fact artibrary, following the convention in a specific language. Thus, although a verb lemma is represented by its infinitival form in English (e.g. *be* for *am/is/are*, *was/were*, etc.), a different langauge may choose, say, a present tense third person singular form as a representation.

Therefore, stemmers and lemmatizers cannot be used adequately without understanding the definition and representation of stems and lemmas assumed by their developers. Meanwhile, developers of these tools need to make decisions about how stems and lemmas are defined and represented in their tools, unless a well-established convention already exists in the language. In the present paper, I report the decisions I made when adding stem and lemma information to the Malay/Indonesian[1] morphological dictionary MALINDO Morph (Nomoto et al. 2018)[2].

The rest of the paper is organized as follows. Section 2 presents the definitions of stems and lemmas, and points out that the notions have not been understood in the normal sense in the context of stemmer and lemmatizer development for Malay/Indonesian. Section 3 describes how stems and lemmas are identified in MALINDO Morph. The section also serves as a brief guideline for stem and lemma identification in Malay/Indonesian. Section 4 concludes the paper.

2 Stems, lemmas and lexemes

2.1 Definitions

I adopt the following definitions:

Stem: the base targeted by inflectional morphology
Lemma: a concrete form representing a lexeme
Lexeme: an abstract lexical unit of a group of inflectionally related word forms

Importantly, the three notions are concerned only with inflectional as opposed to derivational morphology. Dictionaries normally use lemmas as the headword of an entry. Fig. 1 depicts the relationship between various concepts using English words. The 'root' level is also included to emphazie its distinctness from either lexeme/lemma or stem. Roots connect all morphologically related forms, regardless of inflection and derivation.

2.2 Confusion in existing tools

All existing stemmers and lemmatizers for Malay/Indonesian that I know confuse roots with stems or lemmas. The Sastrawi stemmer[3] does not return stems but roots, even

[1] Malay (ISO-639: zsm) and Indonesian (ISO-639: ind) are the two standard varieties of the macrolanguage Malay (ISO-639: msa). The two languages share the core grammar, though there are considerable lexical differences.

[2] `https://github.com/matbahasa/MALINDO_Morph`

[3] `https://github.com/sastrawi/sastrawi`

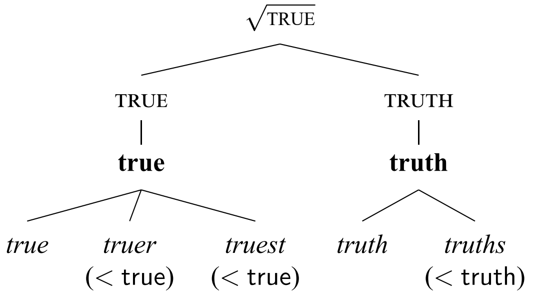

Figure 1: Stem, **lemma**, LEXEME and $\sqrt{\text{ROOT}}$

though it is called a “stemmer.” Malaya, a natural language toolkit for Malay powered by Deep Learning Tensorflow (Husein 2018), is impressive as a whole, but it inherits Sastrawi’s problem. The morphological analyser MorphInd (Larasati et al. 2011), which is currently most widely used in computational processing of Indonesian, regards roots as lemmas.

Therefore, no existing tools (but MALINDO Morph) provide stems and lemmas as defined above. Fig. 2 shows the stems and lemmas for several derived forms of the root *kirim* ‘send’. The existing tools will yield the root form *kirim* as the stems/lemmas for all these words.

This problem happens because the distinction between inflectional and derivational morphology is ignored when defining stems and lemmas. Consequently, stemming and lemmatization are simply taken as removing all sorts of affixes and reduplication. In fact, Knowles and Zuraidah (2006:71) advocate for the inclusion of derivational morphology in the definitions for some languages, including Malay. I do not think this is a wise move. Like many other languages in the world, Malay/Indonesian does have inflection paradigms, as shown in Fig. 2 and discussed in detail in section 3.3. Transitive verbs inflect according to voice: {V (active/passive), *meN*-V (active), *di-* (passive)}. Count nouns inflect for number, with plurality indicated by reduplication. These facts cannot be captured unless stems and lemmas are defined in terms of inflectional morphology alone.

It should be noted that the developers of stemmers and lemmatizers are not to blame. Malay/Indonesian grammars seldom present inflection paradigms as such, and stems, lemmas and roots are often not clearly distinguished. Malay/Indonesian dictionaries are normally organized in such a way that derivationally related words are listed under their common root as its subentries. This common root is referred to variously as *kata dasar* (*lit.* ‘base word’), *kata akar* (*lit.* ‘root word’)[4] or *akar kata* (*lit.* ‘word’s root’). The Sastrawi stemmer defines stemming using the first term: *Stemming adalah proses mengubah kata berimbuhan menjadi kata dasar* (Stemming is a process of changing affixed words into “kata dasar”). While *kata dasar* is used as an equivalent of *stem*, it is more commonly used as referring to roots.

3 Stems and lemmas in MALINDO Morph

3.1 MALINDO Morph

MALINDO Morph is a morphological dictionary that I developed with my colleagues (Nomoto et al. 2018). It is the first and only morphological dictionary for Malay/Indonesian. When it was first released, MALINDO Morph had a total of 232,546 lines, with each line containing an analysis for one (case-sensitive) token. An analysis consisted of root, surface form, prefix/proclitic, suffix/enclitic, circumfix and reduplication type.

Subsequently, the information about the source of each token was added. The latest release (ver. 20190923) saw a major upgrade with the addition of stems and lemmas. The total line number increased to 234,274, of which 134,101 (those with IDs starting with ‘cc’ or ‘ec’) have been manually checked. A sample line is given in Fig. 3.

3.2 Special signs

+: Token boundary. Some words contain clitics, which are spelt together with the host in the standard Malay/Indonesian orthography. Ideally, splitting clitics is the job of tokenizers. However, existing tools either ignore them by simply removing them as if they were affixes or attempt to split them, but not very accurately. MALNIDO Morph thus also includes words with clitics inside them and indicates the boundary with the ‘+’ sign. Moreover, multiple tokens that should be spelt separately are sometimes spelt together in casual writing. These cases also involve the ‘+’ sign. Some examples are given below, where clitics are underlined:

	Surface Form	**Stem**	**Meaning**
1.	kupikir	aku+pikir	‘I think’
2.	kuasaku	kuasa+aku	‘my power’
3.	yakah	ya+kah	‘yes Q’
4.	diatas	di+atas	‘at the top’
5.	itupun	itu+pun	‘that too’

In these examples, morphological analysis is unambiguous. However, ambiguity arises for two morphemes, namely *nya* and *se*. The morpheme *nya* is ambiguous between the third person enclitic pronoun (=*nya*)[5] and the suffix (-*nya*) of various functions such as forming exclamatives (in Malay), nominalizing adjectives and intransitive/passive verbs, forming adverbs, etc. The former should be tokenized, as it functions similarly to its free morpheme counterparts *dia* ‘he/she’ and *ia* ‘it’. In Fig. 2, this is indicated by the ‘+’ sign as in `menarik+dia` ‘his/her/its interesting (N)’; `tarik+dia` ‘to pull him/her/it’. The suffix -*nya*, on the other hand, should not be tokenized, hence no use of ‘+’: `menariknya` ‘how interesting; being interesting’.

[4] This term is problematic, as not all roots are words.

[5] The definite marker *nya* often found in Indonesian is also analysed as a clitic in MALINDO Morph. This is because it is syntactically a determiner, as are enclitic pronouns.

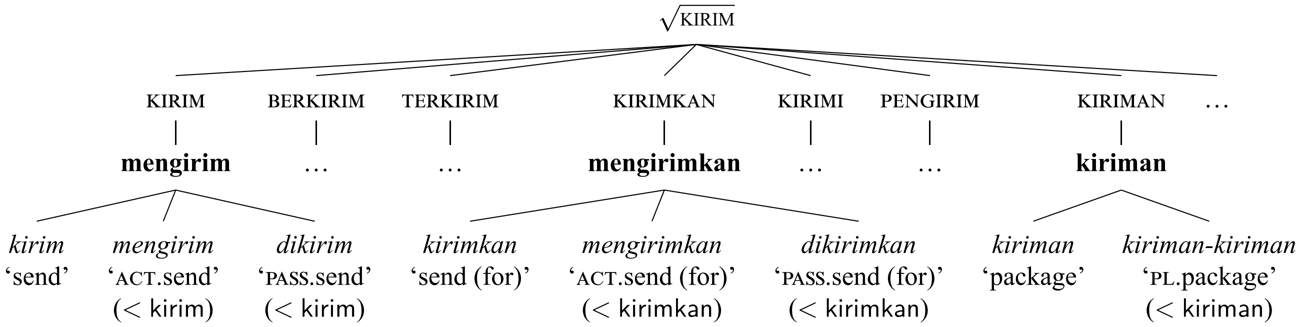

Figure 2: Hierarchical structure of derived forms of *kirim* 'send'

ID	Root	Surface form	Prefix	Suffix	Circumfix	Redup.	Source	Stem	Lemma
ec-42593	tarik	Me-nariknya	meN-	-nya	0	0	Leipzig	menariknya@menarik+dia@tarik+dia	menariknya@menarik+dia

Figure 3: Analysis for *Me-nariknya* 'his/her/its interesting (N); to pull him/her/it; how interesting, being interesting'

The prefix *se-* is also ambiguous. It should be tokenized when it is a prefixal form of the numeral *satu* 'one', but not otherwise. Thus, *sebuah* 'one CLF' involves two tokens: `satu+buah`. However, non-numeral *se-* such as in *sekampung* 'same village' (< *kampung* 'village'), *sebaik* 'as good as' (< *baik* 'good') (cf. Fig. 4), *sesudah* 'after' (< *sudah* 'already') is part of a single token.

@: Disjunction. As a consequence of such ambiguous morphological analyses, some forms have more than one stem and/or lemma. This is indicated by the '@' sign, a common shorthand for *atau* 'or'. See Fig. 3 for examples.

Multiple stems/lemmas also arise due to multiple POS possibilities. For example, the form *aku* is a first person singular pronoun 'I', but it is also used as the bare form of the verb *mengaku* 'to admit'. As seen in Fig. 2–3 and discussed further below, the lemma of a verb is the *meN-* form if its inflection paradigm contains one. Thus, the lemma column for the surface form *aku* is `aku@mengaku`.

3.3 Inflection paradigms

MALINDO Morph assumes three inflectional paradigms.

Transitive verbs. A number of transitive verbs have three inflectionally related forms: bare, *meN-* and *di-*. The bare form is used in the bare active voice (1a) and the bare passive voice (1b), the *meN-* form in the morphological active voice (1c) and the *di-* form in the morphological passive voice (1d) (Nomoto 2013).

(1) a. Mereka sudah *baca* buku itu.
3PL already read book that
b. Buku itu sudah mereka *baca*.
book that already 3PL read
c. Mereka sudah *mem-baca* buku itu.
3PL already ACT-read book that
d. Buku itu sudah *di-baca* oleh mereka.
book that already PASS-read by 3PL
'They already read the book./
The book was already read by them.'

It is obvious that the stem is identical to the bare form. The morphologically basic bare form is also basic syntactically in that it occurs in both active and passive voices. Given this, the bare form should also be chosen as the lemma. However, as stated in section 1, the choice of lemma is arbitrary. It is in fact the *meN-* form that was chosen by the communities, as used in dictionaries and grammar books. MALINDO Morph follows this convention. In short, the stem and lemma for verbs with the {bare, *meN-*V, *di-*V} paradigm are the bare form and the *meN-* form, respectively. See the lexemes KIRIM and KIRIMKAN in Fig. 2 for examples.

Note that not all transitive verbs involve this paradigm. In particular, the so-called *ter-* and *kena* passives involve no overt passive morphology in Malay and to some extent in Indonesian (see Nomoto 2013 and references cited therein). *Ter-* and *kena* are are a derivational prefix and a modal verb, respectively. Therefore, the stem and lemma for *ter-* verbs are the *ter-* form, but neither the bare nor the *meN-* form. For example, the stem and lemma for *termakan* 'to eat/be eaten accidentally' are both *termakan*.

Note also that intransitive *meN-* verbs and adjectives with *meN-* do not involve the relevant paradigm either. Hence, their stems and lemmas are the *meN-* form. For example, the stems/lemmas for *meningkat* 'to increase' and *menarik* 'interesting' are, respectively, *meningkat* and *menarik*, but not *tingkat* and *tarik*.

Count nouns. Count nouns inflect for number. The bare form is number-neutral, whereas the plural involves full reduplication. The stem and lemma are the bare form. See the lexeme KIRIMAN in Fig. 2 for an example.

Not all fully reduplicated forms are plural nouns. Full reduplication is also available for other POSs, in which case reduplication is not inflectional but derivational. Full reduplication of adjectives adds an additional shade of meaning or turns an adjective into an adverb. Moreover, some reduplicated nouns such as *kanak-kanak* 'child' are inherently reduplicated, with no non-reduplicated counterpart. The stems/lemmas for non-count noun reduplicated forms are the reduplicated forms themselves.

Gradable adjectives. Gradable adjectives inflect for degrees. The equative and the superlative are indicated by the prefixes *se-* and *ter-*, respectively. Their stems and lemmas are both the bare form. Fig. 4 shows the inflection paradigm of the adjective *baik* 'good'.[6]

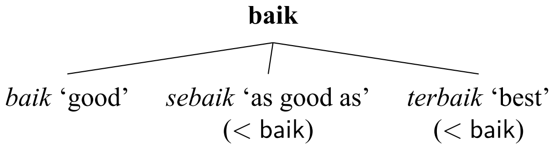

Figure 4: The inflection paradigm of *baik* 'good'

The superlative *ter-* should not be confused with the verbal prefix *ter-* mentioned above. The latter is derivational because it adds a non-volitionality meaning.

4 Conclusion

The stem and lemma information made available by MALINDO Morph will improve stemming and lemmatization in Malay/Indonesian. What was thought of as stemming and lemmatization in other existing tools is in fact 'root'-ing, that is, undoing all morphological processes to get a root. With such an understanding, stemming and lemmatization in Malay/Indonesian are not difficult to implement. However, as seen in the previous section, stemming and lemmatization in Malay/Indonesian is quite complicated, involving various kinds of ambiguities.

Calling 'root'-ing stemming/lemmatization is not only inadequate linguistically but will also cause confusion among the users of stemmers/lemmatizers. It is now not uncommon for one to analyse a language about which s/he has no knowledge, using tools developed by others. S/he will simply assume that what a Malay/Indonesian "stemmer" yields is a stem in the same sense as is used for English and continue his/her analysis or application development. The same problem will also happen if someome knows the language but does not understand what is actually produced by a "stemmer."

The stem and lemma information in MALINDO Morph can be used to develop a genuine stemmer and lemmatizer. It can also be used for language resource development. For example, some entries of Wordnet Bahasa (Bond et al. 2014), the wordnet for Malay/Indonesian, list more than one form for a lexeme. Synset 01437888-v ('send via the postal service') thus contains not only *mengirim* (lemma) but also *kirim* (non-lemma) (cf. Fig. 2). Removing the latter will make the entry less messy and more useful.

In the future, I would like to make a relational database based on MALINDO Morph, following what the National Institute of Japanese Language and Linguistics (NINJAL) did with their UniDic morphological dictionary[7] (Ogiso and Nakamura 2011). Given the neat hierarchical structure of the Malay/Indonesian lexicon (cf. Fig. 2 and 4), it should be possible and useful to apply the basic design of the UniDic database to Malay/Indonesian. The resulting database can in turn be used to annotate corpora, again as done by NINJAL.

Acknowledgements This study was supported by the JSPS grant 'Program for Fostering Globally Talented Researchers' and the JSPS grant #JP18K00568.

References

Bond, Francis, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined Wordnet Bahasa. *NUSA* 57: 83–100.

Husein Zolkepli. 2018. Malaya. GitHub repository. URL `https://github.com/huseinzol05/malaya`.

Knowles, Gerald O., and Zuraidah Mohd Don. 2006. *Word Class in Malay: A Corpus-Based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

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.

Nomoto, Hiroki. 2013. On the optionality of grammatical markers: A case study of voice marking in Malay/Indonesian. *NUSA* 54: 121–143.

Nomoto, Hiroki, Hannah Choi, David Moeljadi, and Francis Bond. 2018. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. In *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"*, ed. Kiyoaki Shirai, 36–43.

Ogiso, Toshinobu, and Takenori Nakamura. 2011. *"Gendai Nihongo Kakikotoba Kinkou Koopasu" Keitairon Jouhou Deetabeesu no Sekkei to Jissou Kaiteiban. [The Architecture and Development of the Morphological Information Database for "The Balanced Corpus of Contemporary Japanese" Revised Edition.]* Technical report, National Institute of Japanese Language and Linguistics.

[6] The inflection paradigms for many adjectives are not as complete as *baik*. This is because the superlative meaning is also expressed by the word *paling* 'most', which is more productive than *ter-*.

[7] `https://unidic.ninjal.ac.jp/`