Building a Tokenizer for Indonesian

David **Moeljadi** and Hannah **Choi**

Division of Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore

The 21st International Symposium on Malay/Indonesian Linguistics (ISMIL 21),
Langkawi Research Center

4 May 2017

Outline

1. Tokenization

2. Wordnet Bahasa

3. Our proposal

- There is no good tokenizer for Indonesian
  → we are building a good one (early stage)
- Many benefits we can get, esp. for natural language processing, corpora etc.
- We will propose our guidelines
  → open to comments and suggestions

Tokenization

Tokenization or **word segmentation** is the task of separating out (tokenizing) words or other meaningful elements (tokens) from running text; the segmentation of text [3]

Tokens:

- words,
- numbers,
- punctuation marks,
- parentheses,
- quotation marks,
- and similar entities

An Example in English

> “Most customers don’t want to sit in a turboprop for 2 1/2 to three hours,” Mr. Lowe said.

Wall Street Journal corpus

Tokenization result:

<S> “ Most customers do n’t
want to sit in a turboprop
for 2 1/2 to three hours ,
” Mr. Lowe said . </S>

Corpus linguistics: an international handbook, volume 1

An Example in Indonesian

> ...salah satu relawannya Ahok bilang ‘Kita kumpul di sana jam 19.00 WIB’. ...

KOMPAS.com “Merespons Pembakaran Bunga, Relawan Ahok-Djarot Nyalakan Lilin”

Tokenization result:

<S> salah satu relawan nya Ahok bilang
‘ Kita kumpul di sana jam
19.00 WIB ’ . </S>

Purpose of Tokenization

Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis.
The list of tokens becomes input for further processing such as parsing (taking an input and producing some sort of linguistic structure for it) or text mining (the process of deriving high-quality information from text).

Text → tokenization → part-of-speech (POS) tagging → lemmatization → sense/semantic tagging → semantic disambiguation → machine translation, information retrieval, sentiment analysis
→ syntactic parsing → treebank building, corpus query, lexicography identification of collocations, determining verb frames, information extraction, term extraction, ...

Current situation

- NLTK tokenizer (`http://text-processing.com/demo/tokenize/`)
- morphInd (`http://septinalarasati.com/work/morphind/`)
- `http://morphadorner.northwestern.edu/morphadorner/wordtokenizer/example/`
- ...

Tokenization problems

- **Multiword expressions**
  e.g. New York, *rumah sakit* "hospital", *memberi tahu* "tell, inform", *dan lain-lain* "et cetera", ...
  Problems: *orang tua* "parent/old person", *kamar kecil* "toilet/small room", *kambing hitam* "scapegoat/black goat", ...
- **Clitics**
  e.g. isn't, he's, we'll, *kukejar* "chased by me", *kaukejar* "chased by you", *dikejarnya* "chased by him/her", *mengejarmu* "chase you", *bukunya* "the/his/her book", ...
  Problems: *kucek* "rub,scrub/checked by me", *rumah bekuku* "Gilt-head bream's house/my frozen house", *keramu* "keramu tree/your monkey", *penanya* "questioner/his/her/the pen", ...
- **Affixes**
  e.g. se-Indonesia "whole/entire Indonesia", seekor "one CL", ...
- ...

Wordnet

- an open-source, free semantic lexicon
- a resource for the study of lexical semantics
- `http://wordnet.princeton.edu`
- synset (synonym set): a group of words with closely related meanings e.g. the noun “car” has 5 different meanings (senses), thus belongs to multiple synsets. One synset for “car” consists of many members.

[2]

Wordnet Bahasa

- http://wn-msa.sourceforge.net
- open source
- The Combined Wordnet Bahasa [1]:
  1. Malay Wordnet (Lim & Hussein, 2006)
  2. Indonesian Wordnet (Riza, Budiono & Hakim, 2010)
  3. Open Wordnet Bahasa (Nurril Hirfana, Suerya & Bond, 2011)
- Indonesian: 48,689 synsets and 58,541 words
  Malay: 38,736 synsets and 45,664 words
- has been used for sense tagging NTU Multilingual Corpus (NTU-MC) of English, Chinese, Japanese and Indonesian, ...

Our proposal

General rules:

1. Do not tokenize **multiword expressions** into words if they are in Wordnet
   e.g. *orang tua* "parent/old person" → *orang tua* "parent"
   (*orang*, *tua*, and *orang tua* are in Wordnet)
2. Split **clitics** from the bases
   e.g. *penanya* "questioner/my pen" → *pena* *nya*
   (both *pena* and *nya* are in Wordnet)
3. Split **affixes** from the stems if the affixes have consistent, predictable meanings
   e.g. seekor "one CL" → *se* *ekor*
   (both *se* and *ekor* are in Wordnet)

References

Francis Bond et al. “The combined Wordnet Bahasa”. In: *NUSA: Linguistic studies of languages in and around Indonesia* 57 (2014), pp. 83–100.

Christiane Fellbaum. *WordNet: an electronic lexical database*. Cambridge: MIT Press, 1998. URL: http://wordnet.princeton.edu/man/wninput.5WN.html (visited on 11/24/2014).

Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. 2nd ed. New Jersey: Pearson Education, Inc., 2009.