



Indonesian Part-of-Speech Tag

Fam Rashel, Andry Luthfi, Arawinda Dinakaramani, and Ruli Manurung

Faculty of Computer Science, Universitas Indonesia


Email: fam.rashel@ui.ac.id, andry.luthfi@ui.ac.id, ard51@ui.ac.id, maruli@cs.ui.ac.id

Workshop on Wordnet Bahasa


Nanyang Technological University, Singapore, October 2014

Overview

- Tagset and Manually Tagged Corpus
 - Analysis and Design of Initial Part of Speech Tagset
 - Data Description
 - Testing and Revisions of Tagset
 - Result
- Rule-Based Tagger
 - Language Resources
 - Rule-Based Tagging
 - Summary



Tagset and Manually Tagged Corpus



Analysis and Design of Initial Part of Speech Tagset

Analysis and Design of Initial Part of Speech Tagset

- Analyzed and compared POS tagsets from various previous works.
- Consulted authoritative Indonesian grammar references.
- Our guiding principle in designing a tagset:
 - Maintain useful linguistic distinctions.
 - Reducing the manual effort that would be required by the annotators.

Analysis and Design of Initial Part of Speech Tagset (cont.)

Adriani et al. [2]	Larasati et al. [6]
CC (coordinate conjunction)	H (coordinating conjunction)
CD (cardinal numerals)	C (numeral)
	B (determiner)
FW (foreign words)	F (foreign word)
IN (prepositions)	R (preposition)
JJ (adjectives)	A (adjective)
MD (modal or auxiliaries verbs)	M (modal)
NEG (negations)	G (negation)
NN (common nouns)	N (noun)
NNP (proper nouns)	
PR (common pronouns)	
PRP (personal pronouns)	P (personal pronoun)
RB (adverbs)	D (adverb)
	T (particle)
SC (subordinate conjunction)	S (subordinating conjunction)
SYM (symbols)	
	I (interjection)

Analysis and Design of Initial Part of Speech Tagset (cont.)

Adriani et al. [2]	Larasati et al. [6]
VB (verbs)	V (verb)
WDT (<i>wh</i> -determiners)	
WH (WH)	W (question)
. (sentence terminator)	
, (comma)	
; (colon or ellipsis)	
((opening parenthesis)	
) (closing parenthesis)	
" (opening quotation mark)	
" (closing quotation mark)	
-- (dash)	
	O (copula)
	X (unknown)
	Z (punctuation)



Data Description

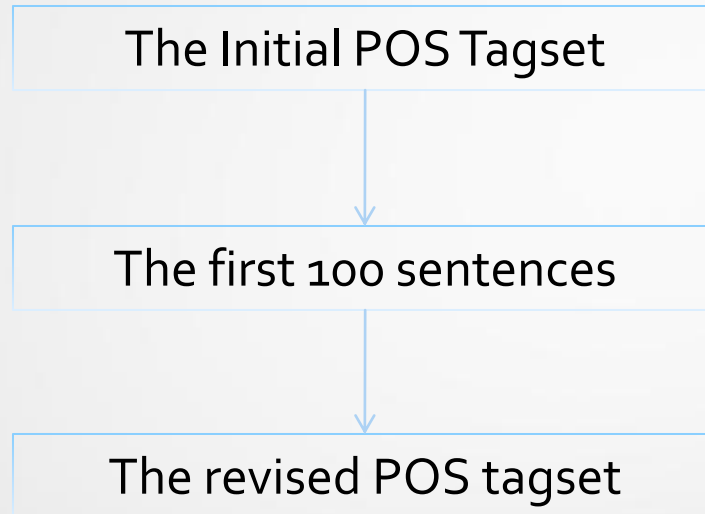
The IDENTIC Parallel Corpus

- The Penn Treebank corpus that were translated into Indonesian.
- Newspaper articles in economy, international news, science, and sports from the PAN Localization project output.
- Movie subtitles.

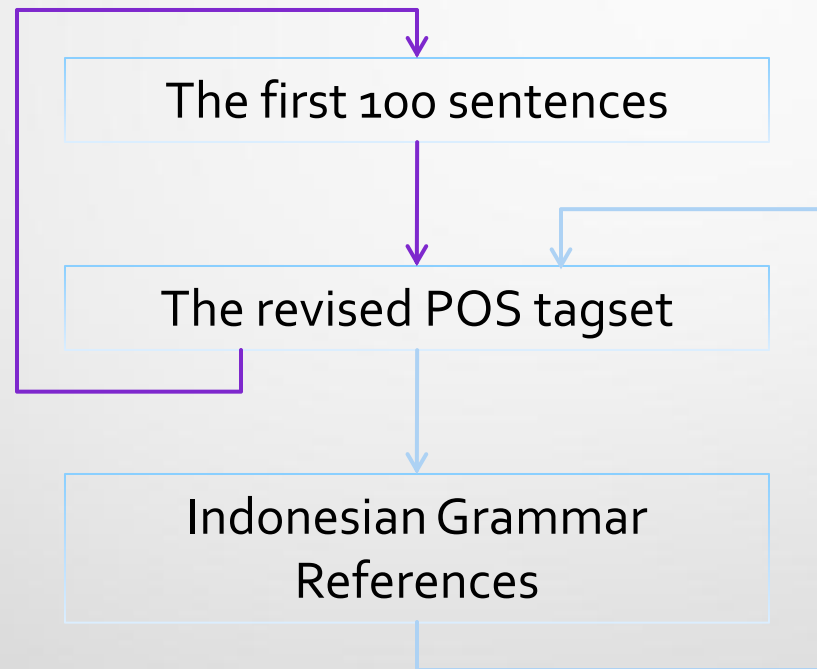


Testing and Revisions of Tagset

Testing and Revisions of Tagset



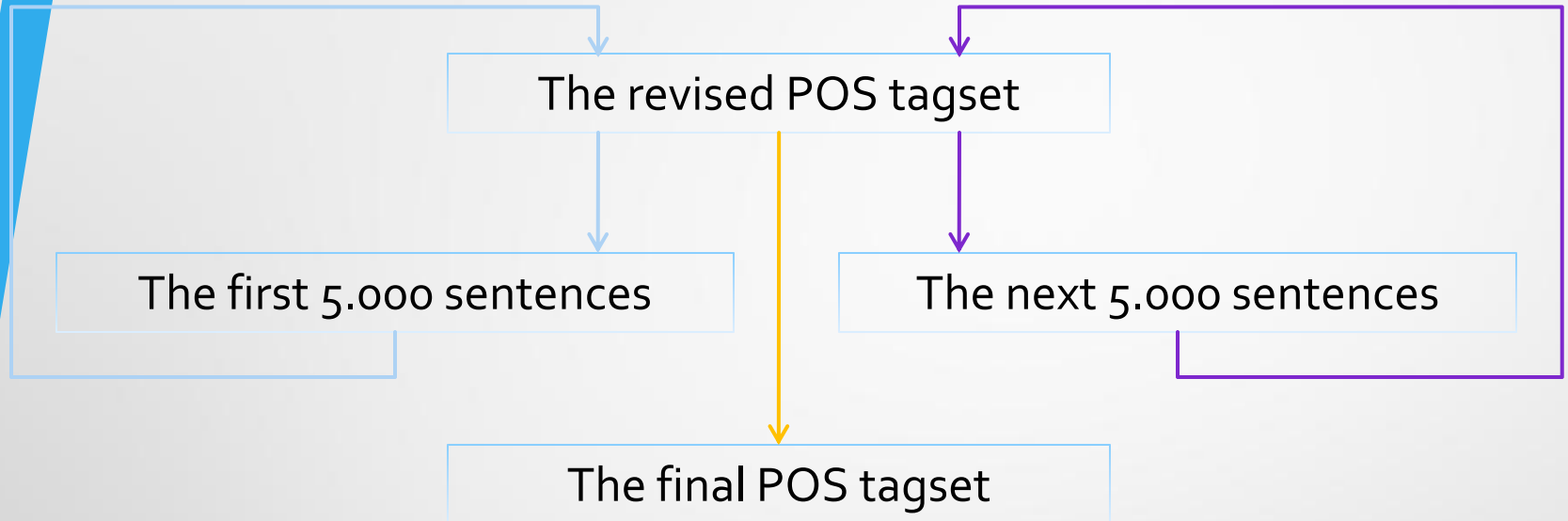
Testing and Revisions of Tagset (cont.)



→ 2nd step

→ 3rd step


Testing and Revisions of Tagset (cont.)



→ 5th step

→ 6th step

→ 7th step



Result

Output 1: Part of Speech Tagset

Tag	Description	Example
CC	Coordinating conjunction	dan 'and'
CD	Cardinal number	enam 'six'
OD	Ordinal number	pertama 'first'
DT	Determiner / Article	Sang 'The'
FW	Foreign word	change
IN	Preposition	dengan 'with'
JJ	Adjective	bersih 'clean'
MD	Modal and auxiliary verb	harus 'must'
NEG	Negation	tidak 'no'
NN	Noun	monyet 'monkey'
NNP	Proper noun	India
NND	Classifier, partitive, and measurement noun	ton

Output 1: Part of Speech Tagset (cont.)

Tag	Description	Example
PR	Demonstrative pronoun	ini 'this'
PRP	Personal pronoun	saya 'I/me'
RB	Adverb	sangat 'very'
RP	Particle	pun
SC	Subordinating conjunction	jika 'if'
SYM	Symbol	%
UH	Interjection	aduh 'auch'
VB	Verb	pergi 'go'
WH	Question	siapa 'who'
X	Unknown	statemen
Z	Punctuation	,

Output 2: Manually Tagged Indonesian Corpus

- The first 10.000 Indonesian sentences of the IDENTIC corpus.
- The corpus is made freely available online under a Creative Commons license.

<http://bahasa.cs.ui.ac.id/postag/corpus>

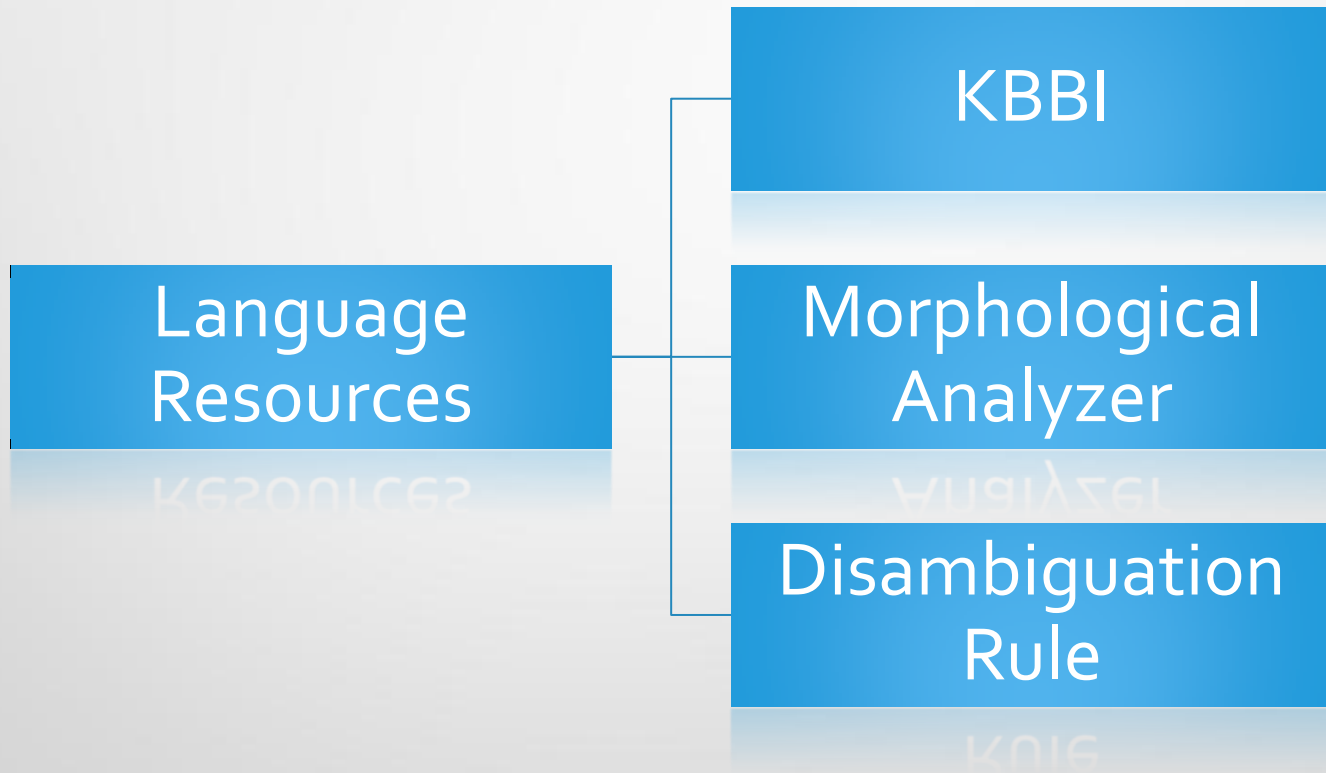


Rule-Based Tagger



Language Resources

Language Resources



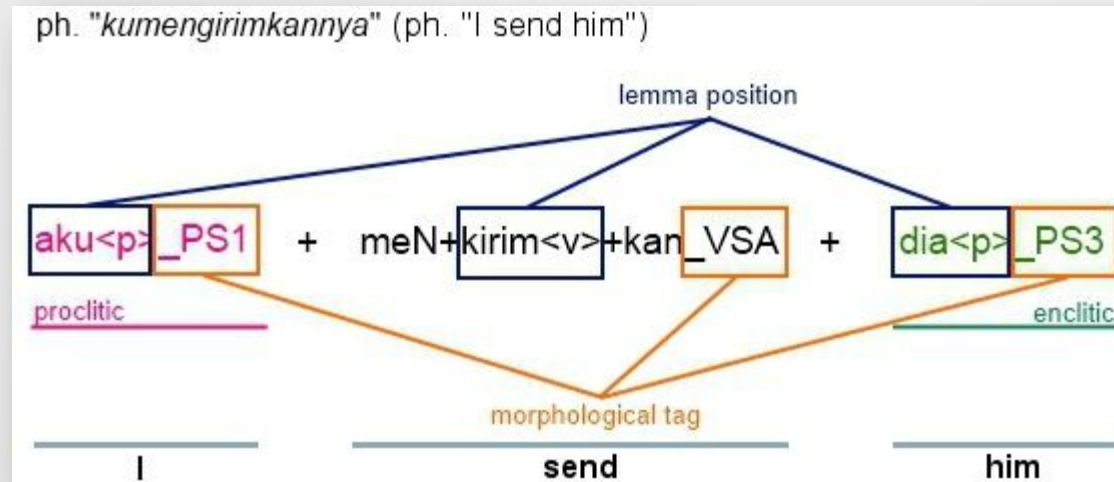
KBBI

We use Kamus Besar Bahasa Indonesia (KBBI) version 3 to extract the required information. From KBBI we managed to build closed-class tagging dictionary and multi-word expressions dictionary.

Closed-Class Words	Part-of-speech tag
dia	PRP
belum	NEG
atau	CC

Multi-word Expressions	Part-of-speech tag
rumah sakit jiwa	NN
balas dendam	VB
haru biru	NN

Morphological Analyzer



The system employs MorphInd to annotate noun, verb and adjective tag (open-class words) [2].

Disambiguation Rules

BISA

Modal? Noun?

The system provides a disambiguation feature by employing 15 disambiguation rules.

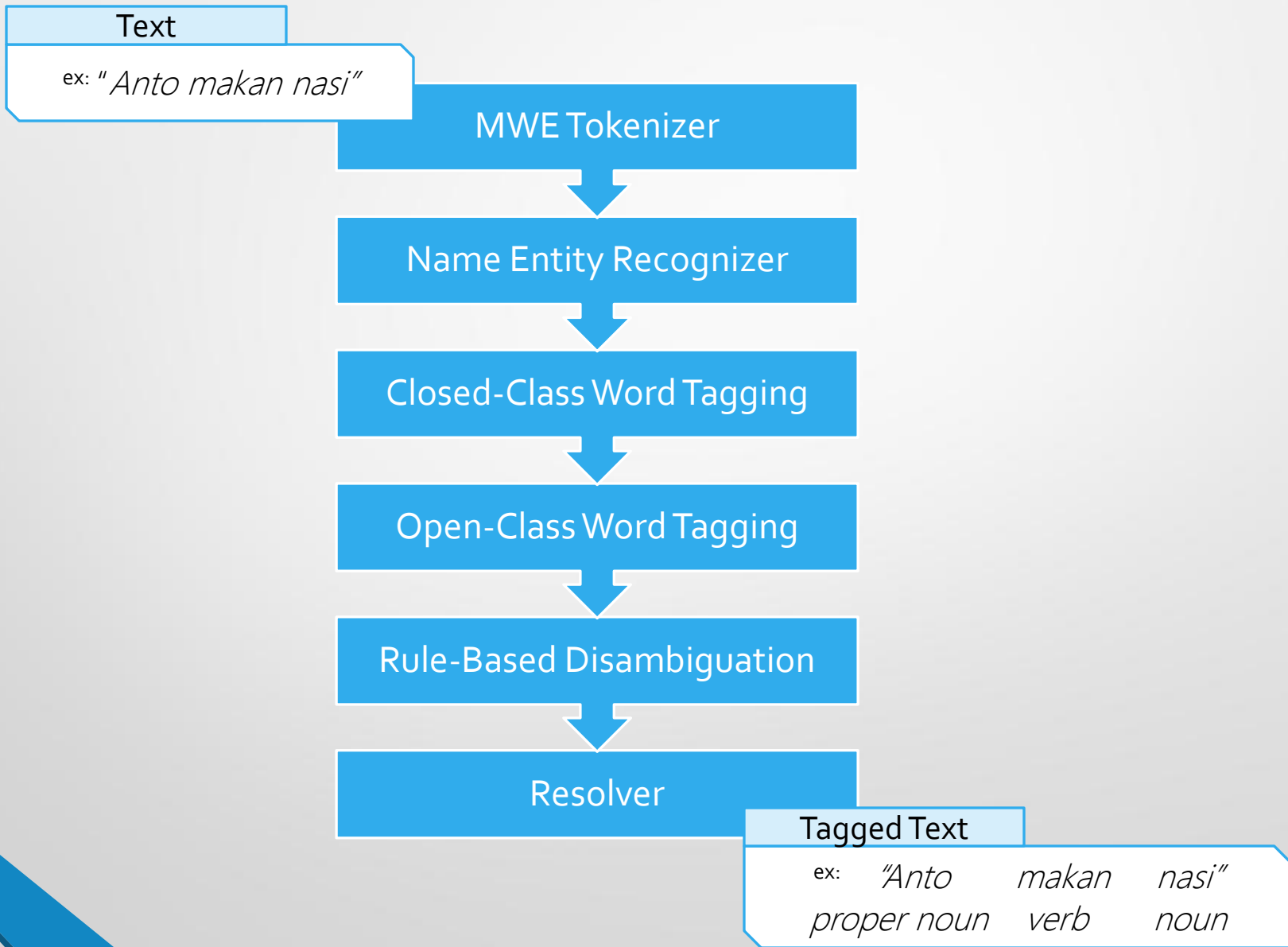
```
<rule id="rule-11" tags="MD/NN">  
  <premise grammar="+1:NN" output="NN"/>  
  <premise grammar="+1:VB" output="MD"/>  
  <premise grammar="+1:JJ" output="MD"/>  
  <premise grammar="-1:IN and +1:NN" output="NN"/>  
  <premise output="MD,NN"/>  
</rule>
```

The rules disambiguate a token by performing “lookup” for the neighboring token’s tag.



Rule-Based Tagging

Rule-Based Tagging



MWE Tokenizer

Kera untuk amankan pesta olahraga

Pemerintah kota Delhi mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena pesta olahraga Persemakmuran ...

**Multi-word Expressions
Dictionary
(KBBI)**

20677 token

Kera
untuk
amankan
pesta olahraga

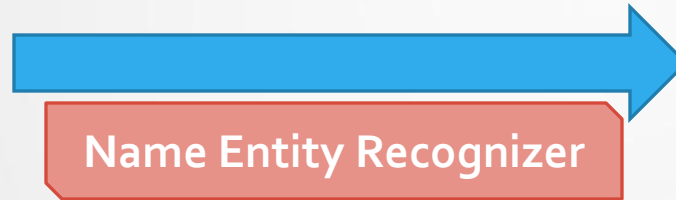
NN

Pemerintah
kota
Delhi
mengerahkan
monyet
untuk
mengusir
monyet-monyet
lain
yang
...

Name Entity Recognizer

Kera
untuk
amankan
pesta olahraga NN

Pemerintah
kota
Delhi
mengerahkan
monyet
untuk
mengusir
monyet-monyet
lain
yang
...



Kera
untuk
amankan
pesta olahraga NN

Pemerintah NNP
kota NNP
Delhi NNP
mengerahkan
monyet
untuk
mengusir
monyet-monyet
lain
yang
...

Closed-Class Words & Open-Class Words Tagging

Kera	
untuk	
amankan	
pesta olahraga	NN
Pemerintah	NNP
kota	NNP
Delhi	NNP
mengerahkan	
monyet	
untuk	
mengusir	
monyet-monyet	
lain	
yang	
...	

MorphInd

Closed-Class Word
Dictionary

word *postag*
word *postag*
word *postag*
...

Kera	NN
untuk	SC,IN
amankan	VB
pesta olahraga	NN
Pemerintah	NNP
kota	NNP
Delhi	NNP
mengerahkan	VB
monyet	NN
untuk	SC,IN
mengusir	VB
monyet-monyet	NN
lain	NN
yang	SC
...	

Rule-Based Disambiguation

Disambiguation Rules

Rule 1
Rule 2
Rule 3
...
Rule 15

Kera	NN
untuk	SC,IN
amankan	VB
pesta olahraga	NN

Pemerintah	NNP
kota	NNP
Delhi	NNP
mengerahkan	VB
monyet	NN
untuk	SC,IN
mengusir	VB
monyet-monyet	NN
lain	NN
yang	SC
...	

Kera	NN
untuk	SC
amankan	VB
pesta olahraga	NN

Pemerintah	NNP
kota	NNP
Delhi	NNP
mengerahkan	VB
monyet	NN
untuk	SC
mengusir	VB
monyet-monyet	NN
lain	NN
yang	SC
...	

Resolver

What if there is a token which does not have any POS tag?

The system would give a special "X" tag for the respective token as a meaning of unknown token. This special "X" tag indicates that the system does not know the right part-of-speech tag for that token. We believe that better for the system to tell that it does not know rather than giving unreliable answer.

Input-Output

Input

Anto bisa makan apa saja?

Output

Token	POSTag	Ambiguous Tag	Rule Applied
Anto	NNP		
bisa	MD	MD, NN	rule-11
makan	VB		
apa saja	PR		
?	Z		



Summary

Indonesian Rule-Based POS Tagger

Demo POS-Tagger

DEMO API KONTAK

Silahkan berikan sebuah masukan. Lalu mesin kami akan memberikan Jawaban

Anto bisa makan apa?

PROSES

Contoh Keluaran POS-Tagger

Kata	Tag	Tag Ambigu	Aturan Disambiguasi
Anto	NNP		
bisa	MD	MD,NN	rule-11
makan	VB		
apa	WH	SC,WH	rule-1
?	Z		

Future Work

- Develop better disambiguation rules,
- Foreign language detector,
- Expand the language resources, and
- Improve the tokenizer.



Thank You for Your Attention