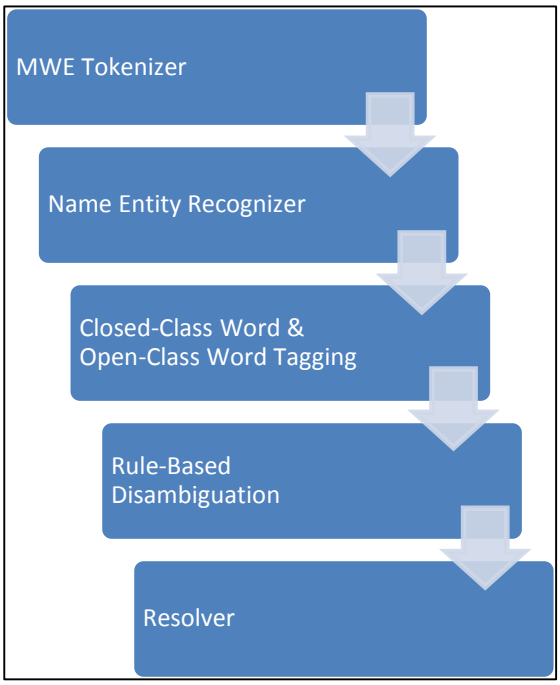
Indonesian Rule-Based Part-of-Speech Tagger

Intro

This is a technical documentation of Indonesian Rule-Based Part-of-Speech Tagger at bahasa.cs.ui.ac.id/postag. This POS Tagger was developed by employing rule-based approach and implemented in Java & Perl programming language. If you are using our POS Tagger please cite our publication.

Methodology

In general, our POS Tagger functionality can be divided into 6 main modules:



1. Multi-word Expression Tokenizer
   During this stage, tokenization is performed against the input document. The document is split using whitespaces then each token is considered as a multi-word expression (MWE) before finalizing the token. The system also annotates the part-of-speech tag for every token that is recognized as a multi-word expression.
2. Name Entity Recognizer
   The system employs Name Entity Recognizer as a reliable information source to recognize named entities as proper nouns, such as persons, places, organizations, etc.
3. Closed-Class Word Tagging
   The system scans through the document and iterates every token in it. The system then consults the token with the closed-class tagging dictionary.
4. Open-Class Word Tagging
   The system passes the document to MorphInd to analyze every token and get the corresponding part-of-speech tag for every token. The system filters the outputs from MorphInd for only noun, verb and adjective part-of-speech.
5. Rule-Based Disambiguation
   The system will try to disambiguate ambiguous tokens by employing a rule-based approach. The system would search through the defined disambiguation rules to find the suitable rule for the ambiguous token.
6. Resolver
   The system would give a special "X" tag for the untagged token as a meaning of unknown token.

System Requirements

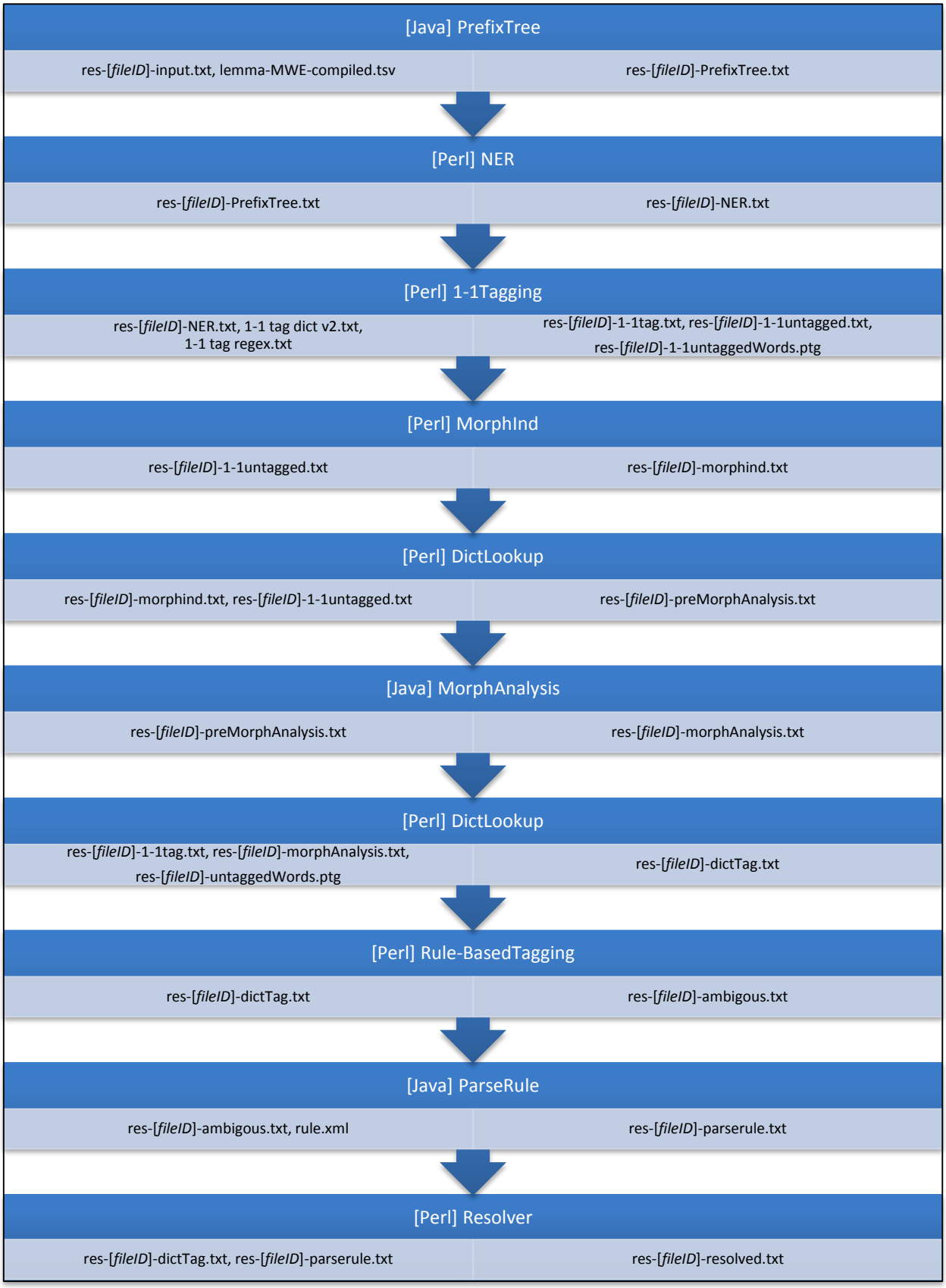
Below are the system requirements to run the Indonesian Rule-Based POS Tagger.

- Linux Operating System
- Morphind (http://septinalarasati.com/work/morphind/)
- Foma (https://code.google.com/p/foma/)
- Zlib ("*sudo apt-get install --reinstall zlibc zlib1g zlib1g-dev*")
- Readline (*"sudo apt-get install libreadline6 libreadline6-dev"*)
- Curl (*"sudo apt-get install curl"*)
- JDK & JRE 1.6
- Perl

Workflow

Below is the workflow of how code interact each other in overall tagging process. The blue boxes indicate source code title including the programming language. Below of each blue box, there are two boxes. The left one indicates input file(s) needed by the program, while the right one indicates the output file(s) produced by the program.

To run the program: $ perl NER.pl -f=*fileID*
Please reassure you have your input written in res-[*fileID*]-input.txt



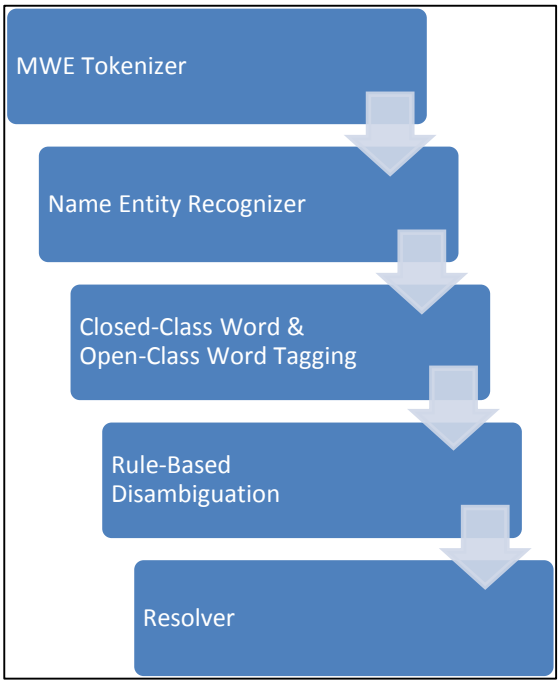
Rule-Based Part-of-Speech Tagger untuk Bahasa Indonesia

Intro

Ini adalah dokumen teknis yang berisi penjelasan mengenai *part-of-speech tagger* (POS Tagger) untuk bahasa Indonesia yang terdapat pada bahasa.cs.ui.ac.id/postag. POS Tagger dikembangkan dengan menggunakan pendekatan *rule-based* dan diimplementasikan dalam bahasa pemrograman Java dan Perl. Jika anda menggunakan POS Tagger kami, harap melakukan sitasi terhadap publikasi kami.

Metodologi

Secara garis besar, fungsionalitas POS Tagger dapat terbagi menjadi 6 modul besar, antara lain:



1. Multi-word Expression Tokenizer
   Merupakan modul tokenisasi yang memperhatikan ekspresi frase yang terdiri lebih dari satu kata. Dokumen diproses oleh tokenizer untuk menghasilkan token-token untuk diberikan part-of-speech yang sesuai kemudian.
2. Name Entity Recognizer
   Merupakan modul yang secara khusus menangani token-token entitas (*proper noun*), seperti nama orang, nama tempat, nama organisasi, dll.
3. Closed-Class Word Tagging
   Merupakan modul yang berguna untuk menangani token-token yang termasuk ke dalam kategori *closed-class word* menggunakan sebuah kamus.
4. Open-Class Word Tagging
   Merupakan modul yang berguna untuk menangani token-token yang termasuk ke dalam kategori *open-class word* dengan memanfaatkan MorphInd.
5. Rule-Based Tagging
   Merupakan modul yang berguna untuk menyelesaikan token-token ambigu berdasarkan aturan-aturan yang telah didefinisikan sebelumnya.
6. Resolver
   Merupakan modul yang berguna untuk memberikan tag kepada token-token yang tidak diketahui.

Kebutuhan sistem

Berikut ini adalah syarat-syarat yang perlu dipenuhi untuk dapat menjalankan program POS Tagger ini.

- OS: Linux
- Morphind (http://septinalarasati.com/work/morphind/)
- Foma (https://code.google.com/p/foma/)
- Zlib (*"sudo apt-get install --reinstall zlibc zlib1g zlib1g-dev"*)
- Readline (*"sudo apt-get install libreadline6 libreadline6-dev"*)
- Curl (*"sudo apt-get install curl"*)
- JDK & JRE 1.6
- Perl

Alur kode program

Berikut ini adalah alur kerja kode program yang digunakan. Kotak berwarna biru menunjukan nama kode program beserta bahasa pemrograman yang digunakan. Kotak di bawahnya terbagi menjadi dua. Sebelah kiri adalah berkas apa saja yang dibaca/dibutuhkan oleh kode program tersebut, sedangkan sebelah kanan adalah berkas yang dihasilkan oleh kode program tersebut.

Untuk menjalankan program: $ perl NER.pl -f=*fileID*
Pastikan ada teks input dalam file res-[*fileID*]-input.txt