Natural Language Processing: Programming Project III Partial Parsing

Reykjavik University – School of Computer Science

November 2009

1 Description

In this project you will develop an *incremental finite-state parser* for English text. Use JFlex (or a comparable tool) to program each transducer. The output of one transducer is used as the input to the next transducer in the sequence. The input to the first transducer is POS tagged text[1] and the output of the last transducer is the original text annotated with various labels denoting chunks. These labels and the corresponding transducers are described in the following sections.

1.1 Verb chunks

This transducer, *Chunk_ Verb.flex*, labels specific verb chunks with [VP ... VP].

- A past tense or present tense verb. Example:

  ```
  [VP found VBD VP]
  [VP finds VBZ VP]
  ```

- The triples <past tense verb, adverb (optional), past participle verb> or <modal verb, verb base form, past participle verb> Example:

  ```
  [VP was VBD required VBN VP]
  [VP was VBD not RB known VBN VP]
  [VP can MD be VB transmitted VBN VP]
  ```

[1] Using the Penn Treebank tagset.

- The pairs <infinitive marker, verb baseform> or <modal verb, verb base form> Example:

  ```
  [VP to TO work VB VP]
  [VP should MD buy VB VP]
  ```

1.2 Noun chunks

This transducer, *Chunk_Noun.flex*, labels specific noun chunks with [NP . . . NP].

- A single pronoun (tagged as PRP) Example:

  ```
  [NP We PRP NP]
  ```

- The quadruple <a determiner (optional), an adverb (optional) an adjective (optional), a noun>. Example:

  ```
  [NP EU NNP NP]
  [NP scientific JJ study NN NP]
  [NP any DT danger NN NP]
  [NP a DT highly RB specific JJ move NN NP]
  ```

- Sequences of two or more nouns in a row Example:

  ```
  [NP cow NN disease NN NP]
  [NP EU NNP Farm NNP Commissioner NNP Franz NNP Fischler NNP NP]
  ```

1.3 Preposition chunks

This transducer, *Chunk_Prep.flex*, labels preposition chunks with [PP . . . PP].

- The sequence a preposition and a noun chunk. Example:

  ```
  [PP from IN [NP the DT human NN NP] PP]
  ```

1.4 Cleaning

This transducer, *Chunk_Clean.flex*, removes multiple occurrences of white space from the output generated by any of the preceeding transducers.

1.5 Testing

You need to put together a test file of 30 sentences by selecting sentences from the file eng.sent (`http://www.ru.is/faculty/hrafn/Data/eng.zip`). Make sure that your test includes at least one occurence of each pattern desribed above. Name your test file *ChunkTest.txt.*

1.6 Everything put together

You need to write a shell with the name *chunk.bat* or *chunk.sh* which accepts one command line parameter, i.e. the file which should be parsed.

Your shell should call the transducers in the exact order as described above and return a file with the same name as the input file name but with ".out" appended. The output file thus contains the original input sentences annotated with chunks. Example:

```
chunk ChunkTest.txt
...
```

generates the file ChunkTest.txt.out

2 What to hand in

All program code (all .flex files and shell files), along with your test file *ChunkTest.txt* and the output file *ChunkTest.txt.out.*