Natural Language Processing – Assignment II

Reykjavik University – School of Computer Science

October 2009

1 Exercise I – 30%

The Xerox company has produced a number of NLP tools using finite-state technology, see `http://epine.xerox.fr/competencies/content-analysis/fsnlp/`. In this exercise, you will experiment with two of their tools: a morphological analyser and a part-of-speech tagger.

1. Describe briefly (2-3 sentences) what technique is used in the morphological analysers developed at Xerox.

2. Go to the demo web page `http://epine.xerox.fr/competencies/content-analysis/demos/english`. Use the tool to perform morphological analysis on the English sentence "That round table might collapse"

   - Why do you get back more than one analysis for each word?
   - How many possible analysis of the above sentence do you get back?
   - Disambiguate this sentence by hand, i.e. state the one single correct reading for each word. Moreover, for each correct reading for a word describe in words what the associated morphological features denote.

3. Describe briefly (2-3 sentences) what technique is used in the part-of-speech tagging developed at Xerox.

4. On the same demo web page, use the tool to perform part-of-speeech tagging on the same sentence "That round table might collapse". Is the tagging consistent with what you did by hand? Explain what the tags mean.

2 Exercise II – 30%

Perform PoS tagging by hand for the following Icelandic sentences by using tags from the Icelandic tagset (description of which is available on the course web page). If you are a non-Icelandic spekar then this can obviously be very difficult for you, but just try to do your best. For examle, try to derive some of the individual features for each tag. Note that the Database of Icelandic Inflections, `http://bin.arnastofnun.is`, can be of help.

1. ég stappaði fótunum á svarta gúmmimottu.

   ```
   ég stappaði fótunum  á  svarta gúmmimottu.
   I  stamped  feet-the on black  rubbermat.
   ```

2. hún tíndi hausana upp af gólfinu og setti þá í pokann.

   ```
   hún tíndi  hausana   upp af   gólfinu   og  setti þá    í  pokann.
   she picked heads-the up  from floor-the and put   them in bag-the.
   ```

3. hún reyndi að leyna kvíða sínum.

   ```
   hún reyndi að leyna kvíða   sínum.
   she tried  to hide  anxiety her.
   ```

4. hún hreyfði hvorki legg né lið.

   ```
   hún hreyfði hvorki  legg né  lið.
   she moved   neither leg  nor joint.
   ```

5. þau vissu bæði að þetta gengi ekki upp.

   ```
   þau  vissu bæði að   þetta         gengi ekki upp.
   they knew  both that this  (would) work  not  up.
   ```

Now use *IceTagger* (on `http://nlp.ru.is/icenlp.htm`) to tag the same sentences (don't cheat by using it when you tag the sentences by hand :-)).

- Point to the errors made by *IceTagger* – it makes **five** errors in total (again, a very difficult task for the non-Icelandic speakers).
- What is the accuracy of *IceTagger* for these test sentences?

3 Exercise III – 40%

In this exercise you need to train *TriTagger*, the trigram tagger which is part of the *IceNLP* toolkit, and then use it to tag test sentences.

1. Download *IceNLP* from `http://www.ru.is/faculty/hrafn/Software/IceNLP-1.2.zip`, and extract to a directory of your choice.

2. Read the section on *TriTagger* in the document *IceNLP.pdf* (available in the /doc directory) to become familiar with how to train and run the tagger.

3. Remove the first 46 lines (4 sentences) from the *eng.train* corpus (from `http://www.ru.is/faculty/hrafn/Data/eng.zip`) and copy these lines into a new file, *eng.test*. Make sure that the *eng.test* file only contains the words, not the PoS tags. Note that at this point, you have a training corpus and a test corpus for which the test sentences do not appear in the training corpus.

4. Build a training model using the modified *eng.train* corpus. **Show the command you use for training**.

5. Now use *TriTagger* to tag the file *eng.test* and make it write the output to the file *eng.out*. **Show the command you use for tagging (testing) and the result from the eng.out file.**

6. **What is the accuracy of *TriTagger* for these test sentences?**