Methods in Computational Linguistics
Master of Science *Computational Linguistics*
Exercise: Part-of-Speech Tagging

apl. Prof. Dr. Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

January 22, 2020

1 Part-of-Speech Tagging – Do it Yourself

1. Select two sentences of your choice (with a sentence length ≥ 8) from the play corpus *wiki-en-flower.txt*. Label the words in the sentences with the Penn Treebank part-of-speech tags provided in the handout of the lecture. Document decisions that you found difficult.

2. Suggest a non-lexicalised transformation based on one of the rule templates of the lecture handout, to correct the part-of-speech error in the following word/tag sequence:

   a/DT very/RB difficult/RB question/NN

3. Explain step-by-step and in your own words why we use the simplified way to calculate

   $$\hat{t}_1^n \approx \arg\max_{t_1^n} \prod_{i=1}^n P(w_i|t_i)\ P(t_i|t_{i-1})$$

   in order to estimate

   $$\hat{t}_1^n \approx \arg\max_{t_1^n} P(t_1^n|w_1^n)$$

   (cf. lecture handout).

4. Calculate the probabilities of the two part-of-speech sequences

   PPSS VB TO VB and PPSS VB TO NN

   for the sentence *I want to race* according to the probabilities in the lecture handout. Which part-of-speech sequence is more probable?

2 Part-of-Speech Tagging with the Tree Tagger

Apply Helmut Schmid's *Tree Tagger* to annotate our English play corpus *wiki-en-flower.txt* with part-of-speech tags.

2.1 Download and Installation of the Tree Tagger

1. You can either download the Tree Tagger from Helmut Schmid's website (`http://www.cis.uni-muenchen.de/~schmid/`) and install it yourself, or you can use an installation provided in `/mount/studenten/MethodsCL/2019/Tree-Tagger/`.

   If you download the tagger yourself, make sure that you also download the English parameter files.

2. Get familiar with the directory structure.

3. How many abbreviations are available for English?

2.2 Applying the Tree Tagger

1. Test the Tree Tagger.

   ```
   echo 'Hello world!'  | Tree-Tagger-path/cmd/tree-tagger-english
   ```

2. Apply the Tree Tagger to our play corpus *wiki-en-flower.txt*.

3. Describe the output of the Tree Tagger. What information is provided, and what is the output format?

4. How many different part-of-speech tags did the Tree Tagger assign to your corpus words? Which are the two most frequent part-of-speech types?

5. How many unknown words are in the corpus (types and tokens)? Can you think of an alternative way to deal with unknown words, in comparison to what the Tree Tagger does?

References

[1] Helmut Schmid. Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*, 1994.