Natural Language Programming Methods and Tools
Assignment 3, May 2016

Stochastic POS Tagging

Introduction

This assignment involves the construction and evaluation of a stochastic POS tagger. This is similar in function to the POS taggers that are available in NLTK whose input is a sentence (an ordered list of words) and whose output is a tagged sentence, e.g.

```
>>> tag('Bill saw that man yesterday')
['Bill/NP','saw/VB','that/DT','man/NN','yesterday/ADV']
```

where the tags assigned represent the best possible tagging sequence in comparison with a gold standard dataset.

Tasks

1. **Data preparation: (10%)**
   - Minimally, you need a dataset comprising at least 100K words.
   - You may use good quality tagged data available from NLTK or from any other source.
   - Use NLTK universal tagset
   - You should prepare test (10%) and training (90%) datasets.
   - Bonus marks for doing N-fold cross validation

2. **Training algorithms.(10%)**
   - Training data. You may need to prepare training datasets of different sizes
   - Output: Word dictionary consisting of words paired with an ordered list of possible POS tags together with their probabilities
   - Tag bigram dictionary pairing each tag bigram with its probability

3. **Tagging algorithms (40%)**

   - Take account of tag bigram probabilities and word tag probabilities. This is in lecture notes and also in Jurafsky and Martin's chapter on the Viterbi algorithm (download)

4. **Evaluation (20%)**

   - This should be in terms of accuracy and should compare different sizes of training data.

5. **Report (20%)**

   This should be in workshop format[1].

   - Introduction. Explain the problem and approach
   - Implementation - describe how your implementation(s) work(s).
   - Evaluation - describe how well your system works. The marks here concern the clarity with which results are presented.
   - Conclusion - limitations, possible improvements to your system
   - Bibliography (any other papers or materials looked at)

Templates for the report are provided on the VLE in word and latex. Max length: 5 pages

Marking

NB. This assignment in all is worth **15% of the mark** for the study unit. Marks will be apportioned according the percentages above

Submission

- Submission Deadline: **Wednesday 29th June**
- Submission Format: zip file containing report, programs + data + any instructions needed to run it + scan of signed declaration. Please submit through VLE

[1] Templates are provided at http://staff.um.edu.mt/mros1/COMMON/workshop_format_latex.zip and http://staff.um.edu.mt/mros1/COMMON/workshop_format_word.docx

Resources

- NLTK corpora
- NLTK book ch 5 [link]
- Jurafsky and Martin (2nd edition) Chapter 5
- Jurafsky and Martin (1999 edition) Chapter 8 [download]

Contact

Please send email (mike.rosner@um.edu.mt) in case of problems.