Assignment: A Non-parametric Bayesian Model for Word Segmentation

ULL 2015-16

In this exercise, we will re-implement the Bayesian word segmentation model of Goldwater et al. [1] and study its performance on the original dataset (a post-processed version of a portion of the CHILDES corpus of child-directed speech). Results of your study should be described in a report, requirements for the report are listed on the class website. The length of the report should be 7-8 pages (e.g., Times 11 pt).

1 The Segmenter

Corpus: The dataset is split into two segments: training and test sets. The data is provided already in a segmented form, however, your learning methods are not supposed to use the segmentation information (i.e. our methods are unsupervised). You can use performance on the training set to debug your model and perform parameter selection. You are not supposed to use the test set for any model selection (ideally you should run your method only once on this dataset).

Model: You will implement only the unigram DP model from Goldwater et al. The model has the following two free parameters: the concentration parameter of the Chinese Restaurant Process and the termination probability for a word. Additionally, the base distribution over characters can also be adjusted (e.g., taken as the uniform distribution over all characters or set to the empirical frequency estimates). You can experiment with these choices. However, also see which choices have been made in the paper and if these choices work best for you.

Evaluation: You can use all metrics from Goldwater et al. At least use the standard Precision, Recall and F-score measures. Report the scores

for different choices of parameters on the training set, and also include results for the best model on the final test set. Please also discuss some examples of produced segmentation, including both successful segmentations and typical mistakes. As discussed in the paper, the unigram model is likely to undersegment, see if this holds for your implementation.

Programming language: We suggest Python as the programming language. However, Java is also acceptable. Please discuss other choices with teaching assistants before starting with the project.

2 Submission

Submit via Blackboard, the deadline is **April 17, 23:59**. This is a **strict** deadline. If you have only a partial solution, still submit. Late submissions will not be accepted.

- The program (including source files) and a README file describing how to run the program. The program should accept hyperparamters, the number of sampling epochs, and input files from the command line. It should output the scores and save the induced segmentation.
- Test corpus segmented by your model
- The report

References

[1] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, 2009.