Applied Language Technology

Lab Assignment 2

Deadline 11:59pm, Monday, October 9, 2017

For questions, please contact the teaching assistant
Marzieh Fadaee: `m.fadaee@uva.nl` (English only!)

Preface

- You can work in groups of 2 or 3 students. Please state clearly in your accompanying document who was part of your team (full names and student ID numbers)
- Data sets for the assignment can be found on
  `https://staff.fnwi.uva.nl/m.fadaee/ALT/file.en`
  `https://staff.fnwi.uva.nl/m.fadaee/ALT/file.de`
  `https://staff.fnwi.uva.nl/m.fadaee/ALT/file.aligned`
- The bitext contains 50,000 lines of tokenized and lowercased sentences (`file.en` and `file.de`) and a (refined) word alignment `file.aligned`.
- Submit the code (code only, not the files!) and the report together as a gzipped tarball via blackboard.
- Place the files on a file sharing website of your choice and submit the path. Make sure the path is accessible and readable by others.

Exercise 1: Extraction of Reordering Estimates

Extract phrase-pair orientations ($l \rightarrow r$ and $r \rightarrow l$). Result: program and file with extracted phrases orientations of the form

`f ||| e ||| p1 p2 p3 p4 p5 p6 p7 p8`

where

- `p1`=$p_{l \rightarrow r}(m|(f,e))$
- `p2`=$p_{l \rightarrow r}(s|(f,e))$
- `p3`=$p_{l \rightarrow r}(d_l|(f,e))$
- `p4`=$p_{l \rightarrow r}(d_r|(f,e))$
- `p5`=$p_{r \rightarrow l}(m|(f,e))$
- `p6`=$p_{r \rightarrow l}(s|(f,e))$
- `p7`=$p_{r \rightarrow l}(d_l|(f,e))$

- p8=$p_{r\to l}(d_r|(f,e))$

Here d_l is discontinuous to the left and d_r is discontinuous to the right.

Estimation should use both orientation estimation approaches: word based and phrase based. Phrases f and e should be of maximum length 7.

Report: Describe and motivate your choices of which reordering events you have considered (this applies to the phrase-based estimation only).

Exercise 2: Empirical Analysis of Reordering

Given the observed orientations in the previous exercise provide a empirical analysis of the data. You slice-and-dice the data as you see fit. Here are some (not exclusive) suggestions:

- Histogram analysis of counts of orientations
- Histogram analysis of counts of actual distances
- Comparison of word-alignment reordering distances (not involving phrases) and phrase-based reordering counts
- Comparison of phrase-length or phrase-pair-frequency and observed reorderings
- Average Variance or standard deviation of reorderings
- Distribution of reorderings that are captured by phrases internally

You don't have to carry out an analysis for all of the above, nor do you have to stick to the suggestions above. This is just to give you some ideas. Also come up with your own analysis criteria.

Submit your analysis code in a tarball together with the report.

Report: Describe how you have analyzed the data and motivate your choice (what did you expect to see?). Describe your findings. Are there any peculiar observations? Are there any correlations? Are the distributions skewed, if so, how? ...

There is no fixed minimum nor maximum length for your report, something in the order of 3-4 pages (including figures).