COMP34412

Coursework 1

Task 1 – POS tagging

- Use a POS-tagger of your choice (e.g. NLTK tagger, Stanford, TreeTagger etc) to tag corpus A (see Blackboard).

(1a) Corpus POS tagged: 0 / 0.25 / 0.5 points

Task 1 – POS tagging

b) Let's assume that the POS tagging you have generated for the corpus is a gold standard. Use the annotated corpus to estimate word likelihood and tag transition probabilities you would need to be able to disambiguate which of the following two POS tagging results is more likely

(1) Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NN

(2) Secretariat/NNP is/VBZ expected/VBN to/TO race/NN tomorrow/NN

Explain what you have done and comment/explain the results in the report (½ page).

Task 1 – POS tagging

1b) Word likelihood probabilities calculated?

0 / 0.5 / 1 point

(1b) Tag transition probabilities calculated?

0 / 0.5 / 1 point

(1b) Comment on the result - is the tag 'correct'?

0 / 0.25 / 0.5 point

Task 2 – distributional semantics

a) *Implement* a program to cluster a given list of target words into n groups based on their distributional patterns. You may first want to construct a word-by-word matrix that captures co-occurrence patterns of the given target words using a given corpus. . . Your program should take as input a list of words to cluster and a number of clusters.

(2a) Word-by-word matrix created: 0 / 0.25 / 0.5 / 1 point

(2a) Target words clustered: 0 / 0.25 / 0.5/ 1 point

Task 2 – distributional semantics

b) Use corpus B and target list D (with 50 words) to evaluate the results of your clustering. Use the following pseudoword disambiguation approach: for each target word, randomly substitute half of its occurrences in the corpus with its reverse (e.g., "procedure" will be transformed into "erudecorp"). Now, apply your clustering algorithm to the list of 100 target words, which contains original words and their reverses, producing 50 clusters. If you generate 50 clusters, how many of them will contain correct pairs (i.e., a word and its reverse)? Repeat this process 5 times and give the average accuracy.

Task 2 – distributional semantics

(2b) Evaluation: corpus 'randomised' by target words?

0 / 0.25 / 0.5 / 1 point

(2b) Evaluation: average accuracy calculated

0 / 0.15 / 0.25 / 0.5 points

Task 2 – distributional semantics

c) Analyse the impact of (1) the *size of context*, (2) *type of features* and (3) *training data* on the quality of generated clusters. To analyse the contribution of contextual representation, consider different ways of constructing a word-by-word matrix (i.e. vary the dimensions of the context window) and experiment with different definitions of context (stems vs. words). To analyse the impact of training data, in addition to corpus B, use also corpus C and train your system on each corpora separately, and also on their combination. Comment the results and report any difference. What other type(s) of feature you may consider using?

Task 2 – distributional semantics

(2c) Analysis: window size

0 / 0.25 / 0.5 / 1 point

(2c) Analysis: words vs. stems

0 / 0.25 / 0.5 / 1 point

(2c) Analysis: corpora

0 / 0.25 / 0.5 / 1 point

(2c) Other ideas for features

0 / 0.15 / 0.25 / 0.5 points

General comments

- Don’t forget to submit code!
- Please add your name to the report and make it look professional