Language Technology (2019)

L5X: Synonym prediction

Marco Kuhlmann

Goal	Use state-of-the-art NLP libraries to train word space models of word meaning and evaluate them on a standard task.
Preparations	Read and understand this instruction in its entirety.
Report	Solve the problems (such as **P01**) and describe your solutions in a short report. Some problems require you to write Python code. Submit the code together with your report.

01 In the synonym part of the Test of English as a Foreign Language (TOEFL), the task of the examinee is to identify which of four alternatives is the correct synonym of a given target word. For example, given the target word and candidate word list

enormously: appropriately, uniquely, tremendously, decidedly

the examinee would have to select *tremendously*. The average performance of non-native English speakers on this test is around 64%. In this assignment you will implement and evaluate a system that beats this baseline using word space models.

02 You are only given data files. All of them can be found in

`/courses/729G17/labs/l5x/data/`

03 The file `toefl.txt` contains the test data, a set of 80 questions from the TOEFL synonym test. Each question is represented by one row with six columns: column 1 contains the question word, columns 3–6 contain the four alternatives, and column 2 contains the position of the correct alternative.

04 The file `oanc.txt.bz2` contains 11.4 million words from the written section of the Open American National Corpus, a balanced corpus with texts from various genres, such as fiction, non-fiction, journals and travel guides. The data was downloaded from

`http://www.anc.org/data/oanc/`

and preprocessed using the following Python script:

```
for line in sys.stdin:
   for token in re.findall(r'[a-zA-Z]+', line):
      print(" " + token.lower(), end='')
```

05 The file `enwik9.txt.bz2` contains 124.3 million words from the (full) English Wikipedia dump on 2006-03-03. The data was downloaded from

`http://mattmahoney.net/dc/textdata.html`

and preprocessed using the Perl script at the bottom of that page.

Problem set

P01 Have a look at the test data. How good are *you* at the test?

P02 Think about how you can use the word space model to solve the synonym test.

P03 Download and install Google's word2vec tool:

`https://github.com/tmikolov/word2vec`

Use the tool to train a word space model on the OANC data. Train using the command line options `-cbow 0` (which will train a so-called skip-gram model) and `-binary 1` (which will save the resulting model in binary format). Use default settings for all other options. How long does it take to train the model?

P04 Write Python code to load the trained model and evaluate it on the TOEFL data. You can use the `gensim` library that was used in lab L5:

`https://radimrehurek.com/gensim/`

Note that you will have to activate the course's virtual environment in order to work with `gensim`. What accuracy do you get?

P05 Try to train a model on the Wikipedia data. Note that this may take quite some time, especially if you are doing it on the lab machines. What accuracy do you get? Compare the improvement in accuracy to the increased data size and training time.