ITCS 4111/5111 Introduction to Natural Language Processing

Assignment 5 **Due: November 10th, 2017 11:59 pm**

Problem 1 (80 points):

Similar to language models, embeddings are trained directly from a large collection of natural language text without being tied to a specific NLP application or subtask. How can we measure the quality of learned embeddings? Some of the commonly accepted extrinsic evaluation methods are based on word similarity tasks, word analogies (e.g., "man is to woman as king is to queenIn this assignment, you will explore an analogy task.

The analogy prediction task is defined as follows. Given a pair of words <*a, b*> and a third word *c*, choose a fourth word *d* so that the analogy "*a* is to *b* as *c* is to *d*" holds. In other words, the relationship between *c* and *d* should be as close as possible to that between *a* and *b*.

Note: the system need not characterize this relationship or give it a name! The relationship is taken to be implicit in the pair <*a, b*>.

Mikolov et al. [2013a] proposed that simple algebraic operations could be applied to embeddings to find an analogy prediction.

Let v_a be the vector for *a*, v_b the vector for *b*, and so on. For the *d* such that the analogy holds, we expect

$$\mathbf{v}_b - \mathbf{v}_a \approx \mathbf{v}_d - \mathbf{v}_c. \tag{1}$$

We therefore seek

$$d = \arg \max_{d \in \mathcal{V} \setminus \{a,b,c\}} \cos\left(\mathbf{v}_d, \mathbf{v}_b - \mathbf{v}_a + \mathbf{v}_c\right). \tag{2}$$

Analogy Dataset Mikolov's analogy dataset [Mikolov et al., 2013a] includes four semantic relations and four syntactic relations. In the test files, each line represents one analogy question, in the form of four words <*a, b, c, d*>.

For example: "Bangkok Thailand Cairo Egypt"

A question is counted as correctly answered only if the predicted word is the same as the given word. For example, given the first three words "Bangkok Thailand Cairo", the task is to predict "Egypt".

The full set of analogy questions can be found in the file *word-test.v1.txt.*
Available here: http://www.fit.vutbr.cz/~imikolov/rnnlm/word-test.v1.txt

The groups of relations are delimited by lines starting with a colon (:) and you should only work with these groups: capital-world, currency, city-in-state, family, gram1-adjective-to-adverb, gram2-opposite, gram3-comparative, and gram6-nationality-adjective.

Word Embeddings "Pretrained" word embeddings are word embeddings that are already constructed in advance of your NLP project (whether your project is a neural language model, a text classifier, or a class assignment). The advantage of pretraining is that it simplifies learning your model, because the embedding parameters are fixed in advance. The disadvantage is that, if your embeddings happen to be bad for your task, you're stuck with them.

For this assignment, you are free to use any pretrained word embeddings you can find, as long as you cite their papers in the writeup.

Here are some popular choices:
• Word2vec [Mikolov et al., 2013a,b]: https://code.google.com/archive/p/word2vec/
• GloVe [Pennington et al., 2014]: http://nlp.stanford.edu/projects/glove/
• Dependency-based embeddings [Levy and Goldberg, 2014]: https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

Note that embeddings vary not only in the method of construction from data, but also in dimensionality, amount and type of text used, and more. When you report on your choices, be precise!

Additional notes. It is okay to use existing tools to efficiently find the most similar vector; as always, cite the tools you used. You may have to settle for lower-dimensional word embeddings or files that contain smaller vocabularies of word embeddings.

Problem 1 (60 points): Run the analogy test. Pick any **two** sets of pretrained embeddings, implement the analogy prediction method described in Equation 2, and compare their accuracies on the eight analogy tasks listed above. Make sure to mention the details of your selection in writing.

Problem 2 (20 points): One known problem with word embeddings is that antonyms (words with meanings considered to be opposites) often have similar embeddings. You can verify this by searching for the top 10 most similar words to a few verbs like increase or enter that have clear antonyms (e.g., decrease and exit, respectively) using the cosine similarity. Discuss why embeddings might have this tendency.

Problem 3 (20 points):** Design two new types of analogy tests that are not part of Mikolov's analogy dataset. You will create your own test questions (3 questions for each type, so in total 6 new questions). Report how well the two sets of embeddings perform on your test questions. You're encouraged to be adversarial so that the embeddings might get an accuracy of zero! Discuss any interesting observations you have made in the process.

What to submit

Code and write-up. Be precise and concise in your write-up.

Note: ITCS 5111 students will be graded out of a 100 points total for all problems in this assignment. ITCS 4111 students will be graded out of 80 points total for all problems in this assignment, except for those marked with **.