GloVe : Global Vectors for Word Representation

(Group Members : Vaishnavi Jha, Astha, Monalisa)

Q1. i). a) What is the role of $f\left(X_{ij}\right)$ in the cost function of GloVe. (2 + 1 + 2)

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 ,$$

Ans: $f\left(X_{ij}\right)$ is a weighting function that makes sure frequent co-occurrences are not over-weighted. If the cost function doesn't have a weighting function, it will treat frequent words such as 'the', 'a', with more weight, which may not be contributing much to the model.

b) Can we keep $f(x) = e^{-x}$?

Ans: No, f(x) must be a non decreasing function, if f(x) is kept as a decreasing function, it will give unreasonably high weight to rare co-occurrences.

c) Out of $\boldsymbol{w}$ and $\tilde{W}$ What should be used as the final word embedding?
Ans: If the co-occurrence matrix is symmetric, w and w~ will be similar, but can differ because of random initialization.

So, we can use either of them, or to capture both vectors we can use their sum (w + w~) or average. (w + w~) / 2

Q2. i) What is the main disadvantage of Matrix Factorization methods? (2+1+1+1)

Ans: a) The main problem with Matrix Factorization methods (like HAL and LSA) is that the most common words in the corpus such as (*the, and*) contribute disproportionate similarity between words .i.e. the number of times two words

co-occur with *the* or *and*, will have a large effect on their similarity despite having relatively little semantic relatedness.

b) It is primarily used to capture word similarity and can't capture complex patterns beyond word similarity well.

ii) Name the techniques that address this shortcoming of HAL.

Ans: COALS(Correlated Occurrence Analogue to Lexical Semantic), PPMI(Positive Pointwise Mutual Information), HPCA(Hellinger PCA).

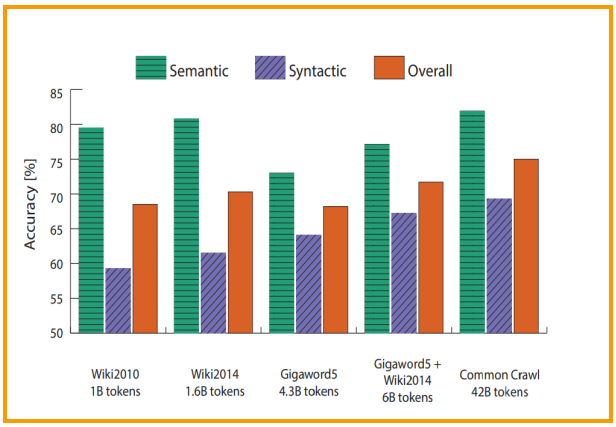
iii) Explain the main difference between Word2vec and Glove.

Ans: Word2vec embeddings are based on training a shallow feedforward neural network while Glove embeddings are learnt based on matrix factorization techniques. The two models differ in the way that they are trained, and hence lead to word vectors with subtly different properties. Glove model is based on global word to word co-occurance counts leveraging the entire corpus while Word2vec on leverages the co-occurance within local context (i.e neighbouring words).

iv) What are the parameters used by Word2Vec?

Ans: Vector size, window size, minimum word count, the number of iterations etc.

Q3. i) For syntactic subtask, there is monotonous increase in performance as the corpus size increases(see figure). But the same trend is not true for semantic subtask. Why?(2+2)



Ans: The performance not only depends on the size of the data but also depends on what you're training your word vectors on.There are large number of city and country based analogies in the analogy dataset.Wikipedia is really great as it has very comprehensive articles and good descriptions on all the capitals in the world. Gigaword is news repository.Abuja/Ashgabat might not be mentioned in the news very often.
Wikipedia has less Spelling errors than general Internet texts.Vectors for such words won't capture semantics really well and will perform worse.

ii) Explain the purpose of adding negative samples during model analysis.

Ans: This was done to draw a quantitative comparison between GloVe and Word2Vec but for GloVe the relevant parameter is the number of Training iterations but for Word2Vec, it is number of training epochs. And the code is currently designed for a single epoch i.e it specifies a learning schedule specific to a single pass through the data. Adding negative samples increases the number of training words as seen by the model and in a way it is similar to extra epochs.