# Assignment 5: Word Embeddings

# 1 Brown Cluster

1. 1) Named- entity recognition;

2) Semantic dependency parsing;

3) Question answering systems.

1. Word pairs with lowest mutual information would be merged together, as they are found in similar context with similar meaning. (i.e. have similar distribution)
2. I would use a greedy algorithm to construct word classes to gradually merging words to form classes by keeping track of the objective of maximizing the likelihood.
3. Brown cluster helps to find a sub-optimal mapping of words to clusters by producing derive lexical representations for unlabeled text. And supervised learning methods can utilize these generated features to learn other linguistic structure.
4. Low generality will result in more clusters that are tighter at each level of the cluster tree, while high generality will produce fewer clusters.

If the word in your dataset has ambiguity problem, then specific clusters are needed. On the contrary, it is better to use more general clusters when word has multiple meanings.

1. For a bigram model,

To measure the quality of a partition, we change it to a more convenient format:

According to the definition that ,

Then above formula:

We can see from the above formula that the first part is entropy of the word distribution, and the second part is mutual information of adjacent clusters.

Therefore, to maximize the log likelihood of bigram class model is same to find the maximum mutual information of adjacent class pairs in brown cluster.

# 2 Word Embeddings (Theory)

1. The distributional hypothesis is that “words that occur in similar contexts tend to have similar meanings”.

Given a corpus, Skip-Gram model loops on the words of each sentence and uses the current word to predict its neighbors (its context), while CBOW will use each of these contexts to predict the current word.

1. The actual data are pairs of word and context, denoting by the probability that comes from the corpus data. These pairs can be extracted from the training data set.

The noise are some pairs with low , i.e. pairs which are not in the data. These pairs can be generated randomly, assuming they are all incorrect.

1. Using subsampling, it will decrease the importance of frequent words with high discarded probability while increase the importance of infrequent words with low discard probability.

Subsampling will aggressively subsample words whose frequency is greater then while preserving the ranking of the frequencies.

1. As this subsampling is done before creating the windows, the context windows will get larger as they can now reach words that were not in their original -sized windows.

# 3 Word Embeddings (Practice)

1. The three nearest neighbors of ‘cat’ are ‘dog’ (0.99991), ‘mouse’ (0.80691) and ‘data’ (0.660421)
2. The three nearest neighbors of ‘computer’ are ‘data’ (0.99967), ‘mouse’ (0.816148) and ‘also’ (0.66427)
3. The three nearest neighbors of ‘the’ are ‘for’ (0.99909), ‘also’ (0.99932) and ‘it’ (0.99885)
4. Cluster 1 (animal): cat, dog, mouse

Cluster 2 (computer related): computer, data, mouse, it

Cluster 3 (preposition): the, for, also, it

Yes, it corresponds with the one obtained from the vector space as each sematic cluster I create here is subset of corresponding cluster formed by unsupervised clustering method.

1. Yes, because a hard clustering algorithm performs a yes/no decision on object membership and cannot model word ambiguity.