John Mathena

LING499J – Naomi Feldman

Final Project Write-up

**Summary**

In this project, I made a change to a model that used Vectors to represent word meaning by using a corpus of words and calculating how often each word appeared within a pre-determined distance of another word. The original model did not use parts of speech (POS) in generating the co-occurrence counts, which I believe to be a shortcoming of said model. The new model retains the POS tags in the corpus so that it can better capture words that have different parts of speech and/or meanings while their written form remains the same. I evaluated the old model without POS tags to the new model with POS tags by comparing the similarity judgements each model produced to actual human similarity judgements in order to determine which model produces judgments that are closer to humans.

**Motivation for change**

One of the assumptions of the original model was that context will tell you what you need to know about the meaning of a word, adopting the distributional hypothesis of semantics. I agree that context does play a large role in semantics but we are sacrificing a significant amount of precision by ignoring the fact that words can have different parts-of-speech and different meanings while their orthographic form remains the same. A Vector Space Model (VSM) that does not account for POS variation ignores key distinctions such as the fact that when humans are asked to judge the similarity of “rain” and “storm”, they are most likely considering “rain” to be a noun and not a verb. However, the original model cannot account for such differences because it has no way to differentiate between the verb “rain” and the noun “rain” if their distributions are similar.

A Turney & Pantel (2010) paper from class stated that Information Retrieval system performance was improved by the addition of syntactic annotation, such as POS tags. Additionally, the same paper explained the usefulness of annotation in measuring the similarity of words, which was the goal of the original model. Due to the support of the paper, I believe that the new model will have better selective capabilities as it has more information to access when making judgments on word pairs. The new model can distinguish between pairs like “dog” and “bark” verb and “dog” and “bark” noun. Whereas, the old model would have both distinct distributions represented in one pair, which would in turn lower the correlation with an actual human similarity judgment given that people are most likely not thinking of tree bark when they judge “dog” and “bark” to be similar and/or related.

Finally, I believe this new model to be an example of linguistic sophistication because it does incorporate a facet of linguistic theory into an existing model of human language. That new facet is parts-of-speech and their effects on the precision of model-produced word similarity ratings. As I stated earlier, POS tags have the potential to increase the precision with which VSMs make similarity judgments.

**Evaluation and baseline model**

Naturally, in order to assess how close the VSM cosine similarities are to human judgments, I am going to compare the cosine values of the different vectors produced by co-occurrence values to actual human similarity judgments. This is the same method used in Problem set 4. However, the human similarity dataset used differs from Problem set 4 in that it contains words that have POS tags. Additionally, the word pairs in the human set are all pairs that are orthographically the same but have different POSs and meanings. I quantified the comparisons between the cosine values and new dataset judgments by using spearman’s rho correlations. The judgments are made based on the similarity and/or relatedness of the pairs, much like the MEN dataset. The baseline model is the model from Problem set 4. I ran the old model on the new dataset and then compared the spearman’s rho correlations between the two models.

**Results**

Following the procedure of Problem set 4, I calculated spearman’s rho correlations between model based similarities and the human similarity judgements at 1000 and 400 dimensions and window sizes of 1 word, 2 words, and 5 words. The final correlation coefficients are represented in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **cooc\_1\_1000** | **cooc\_2\_1000** | **cooc\_5\_1000** | **cooc\_1\_400** | **cooc\_2\_400** | **cooc\_5\_400** |
| Baseline | 0.0143 | 0.025 | 0.0678 | 0.1498 | -0.0348 | -0.0178 |
| POS Model | -0.0769 | -0.0524 | 0.2975 | -0.052 | 0.0453 | 0.2549 |

Overall, the POS model had the highest correlation at 0.2975 with a window size of 5 with 1000 dimensions. The POS model also had the next highest correlation at 0.2549 with a window size of 5 and 400 dimensions. The baseline model had its highest correlations at 0.0678 and 0.1498 at a window size of 5 and 1000 dimensions and a window size of 1 with 400 dimensions, respectively. The POS model performed better as window size increased regardless of the number of dimensions while it performed better overall with more dimensions. The baseline model appeared to perform better as the window size increased when there were 1000 dimensions. However, it begun to perform worse as window size increased when there were 400 dimensions.

**Discussion**

The results demonstrate that using POS annotations can allow a computer model to better align with words that have the same characters but different meanings. However, the important question to ask now is how important POS annotation is to the performance of the model, which is something that can only be answered once the POS model is compared to the baseline model on a human dataset that includes all types of word pairs and not just the ones of interest to my research question. However, before we get into the limitations and future directions of the experiment, I will explain the implications of the results.

As the results show, the POS model-based similarities performed better than the baseline model overall, which is in line with what I predicted. More precisely, the POS model performed best with more dimensions and a larger window size. These findings are in line with the findings of Problem set 4 in which we saw higher window sizes and higher dimension counts leading to higher correlations. Therefore, future models should use dimension sizes closer to 1000 than 400 to achieve optimal performance. This would make sense when using POS tags as those annotations add more row vectors for the same number of words, which would translate into more dimensions needed in order to capture the extra information contained in the POS tags.

I have already eluded to one limitation to my model pertaining to the human similarity data set that I used to evaluate the model and the baseline. The dataset was generated using some of the words in the original Problem set 4 corpus with the POS tags still intact (as Problem set 4 did not make use of them). More specifically, words that had multiple POS tags while remaining orthographically the same. I selected pairs for the human dataset by hand. Therefore, the data was skewed in favor of the POS model. The results still do show that POS tags perform better when the data involves words that have multiple meanings across different POSs. The important question is if that performance boost will make a difference if the POS model is compared to the baseline model on a dataset that mixes in word pairs that do not vary across POSs. Any future experiment comparing POS annotated models and models such as the baseline should use human datasets that include POS tags but also include normal word pairs, or words that do not vary across parts-of-speech.

**Methods**

I used the Brown corpus provided to me in Problem set 2 as my vocab to train the model to get my co-occurrence counts using a function called coocmat.mat. This function takes a vocab and a number (window size) as its input and outputs a co-occurrence matrix for all the words in vocab as well as a map indexing every vocab word in the corpus. However, when training the model, I did not strip off the POS tags as was done for the baseline model. I created a co-occurrence matrix for window sizes of 1, 2, and 5. ‘

Then, I used a function called process\_matrix to convert the co-occurrence counts to positive pointwise mutual information over a set number of columns, which are determined by which of the columns in the raw matrix have the highest variance. I created reduced vectors with dimensions of 1000 or 400 and with window sizes of 1, 2, or 5 so six reduced vectors overall.

I then wrote a function cossim.m that takes two vectors from a reduced matrix and outputs the cosine of the angle between them. I used a function evaluate\_sim, which takes one of the reduced vectors, the wordlist output by coocmat\_mat, and a human similarity dataset as inputs and outputs a spearman rank correlation coefficient between the model’s similarity judgments and human dataset. The function calls cossim.m to calculate the model’s similarity judgement before comparing them to the human judgements. In order to create the dataset for this experiment, I manually looked through the vocab vector generated by coocmat.mat for words that had multiple POS tags while remaining the same orthographically. I put those words, with their POS tags, into a spreadsheet and paired them with other words I found to be similar. I then obtained similarity judgements on roughly twenty pairs. The judgments were on a scale from one to ten. Then, I imported the spreadsheet into MATLAB and used it as input in evaluate\_sim. Evaluate\_sim also had to be modified so it would recognize the POS tags in the new dataset. Pictured below is the dataset I used.



Works Cited

Turney, D. P., & Pantel. P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37.