Code presentation: Make-A-Thon

Dhiganth Rao

Problem Statement: Overview

One of the interesting problems in NLP for healthcare is in handling medical coded term associations and spelling errors. So we are going to mimic the problem with hangman / dump charades like NLP word games. Your ML model should predict the correct word from an intentionally obscured word and its description (Hint). Our evaluation set will have incomplete words and their descriptions.

Input masked word = DEM_G_A_HY and

Description = is the statistical study of populations, especially human beings.

Model prediction/output = DEMOGRAPHY

*You are free to choose / generate your own training dataset that would suit the problem

Possible Solutions:

Since the problem included a test dataset with masked words and their meanings, the meanings could be used somehow to represent that word. Given a dataset containing a lot of words and their meanings, we can convert each meaning into a vector representation, compare these vector representations with the vector representations of the masked word, and pick out the most similar vector representation. Hopefully, that should be the word we're looking for.

Please find my GitHub repository containing my solution.

Possible Solution: Using Vector Representations

Using pre-trained word vector representations such as GloVe, we can simplify this process a lot, as GloVe is trained on a very large corpus of documents (Wikipedia, to be exact). Here, we use the weights of the GloVe model as the vector representations of the particular word. The representations are in the form of a dictionary with each word mapped to its particular vector representation. GloVe representations are available in 50, 100, 200 and 300 dimensions. Here, the 50 dimension GloVe representations are used.

Dataset Overview

The dataset was obtained from this link.

This dataset is a collection of over 13,000 English words, their respective meanings as well as 5-10 examples of their usage in a sentence. The sentences were dropped from this dataset due to redundancy.

Code Overview

- The datasets were loaded and preprocessed.
- They were then converted into word-meaning dictionaries.
- The words in the meaning were converted into their vector representations using GloVe.
- The overall meaning was represented by the average of all the vector representations of the words.

Code overview

- After the vector representations of the meanings present in both the train dataset and the test dataset had been obtained, an iterative loop was created where one meaning from the test dataset was compared with every meaning from the train dataset to compute the similarity.
- The similarity was computed using Cosine Similarity.
- After comparing, 5 similar vectors from the train dataset were appended into a list based on the threshold of Cosine Similarity (if Cosine Similarity > 0.9, the vectors were considered.)

Results

A sample of the results obtained using the method suggested above are shown below.

```
 'C o _ _ n t h': ['Abattoir', 'Abnormal', 'Aboard', 'Abode '],
 '_ e c e i _ e': ['A bed of roses', 'A Priori', 'Abandon', 'Abase'],
 '_ o l l a g _': ['A bed of roses', 'A Priori', 'Abduct', 'Abject'],
 't _ _ s h': ['A bed of roses', 'A Priori', 'Abase', 'Abdicate'],
 'D e _ a w _ r _': ['Abdominal', 'Aberrant', 'Abnormal', 'Aboard'],
 'p a l e _ t _': ['Abbess', 'Abnormal', 'Abnormality', 'Aboriginal'],
 'i c _ _ a _ e _ r o n': [],
 'B a _ _ a c': ['Hasidic', 'Rococo'],
 '_ a l a _ s i _': ['Abnormal', 'Abode ', 'Abolish', 'Abolition'],
 's _ _ i _ o s i s': ['Cholesterol', 'Cyanosis', 'Hypertonic',
'Hypotrophy'],
 's _ b _ _ _ s i b l e': ['A bed of roses', 'A Priori', 'Abduct',
'Abet'],
 'f _ o t p _ d': ['Adolescence', 'Bayonet', 'Bonnet', 'Casualty'],
 'l a n g _ _': ['Buzzard', 'Cilia', 'Claw', 'Hare'],
 '_ _ s c a n': ['Abbess', 'Abnormal', 'Abode ', 'Aboriginal']
```

Results

The results obtained using the method suggested above are very poor. It is observed that a lot of common words are present in every element of the dictionary; implying these set of words are similar to many of the words present in the test dictionary.

Drawbacks and Suggestions

The poor performance of the method suggested above can be attributed to the following reasons:

- The train dataset; as not enough data was present to create enough vector representations out of.
- It was also observed that there were many flaws in the train dataset; many words had meanings that were “NaN”, which are non-string values and hence can’t be used.
- Creating sentence vectors by averaging the word vectors blurs information, thus leading to a drop in performance.
- One possible way to improve performance is to use higher dimensional word embeddings; as that increases the depth of representation of the word.

Conclusion

Thus, a possible method to ‘unmask’ a word based on its meaning was discussed. The results were poor, but this can mainly be attributed to the quality and quantity of the train dataset. It is also to be noted that no deep learning model was used, thus reducing computation time. However, some deep learning models that could be used for this particular problem are RNN’s (Recurrent Neural Networks), or Siamese networks.

Thank you for providing this interesting problem statement! It was fun to work on and try to provide a working solution.

References:

GitHub link to the repository:

https://github.com/dhiganthrao/WordUnMasker

GloVe: Global Vectors for Word Representation:
https://nlp.stanford.edu/projects/glove/

Training Dataset used: https://data.world/idrismunir/english-word-meaning-and-usage-examples

Cosine Similarity: https://www.sciencedirect.com/topics/computer-science/cosine-similarity