CW2 NLP RESEARCH QUESTION

Word embeddings have had a great effect for natural language processing (NLP), specially benefiting sequence-to-sequence (Seq2Seq) models by enriching their understanding of language semantics. They represent words as continuous numerical vectors in a multi-dimensional space, word embeddings capture semantic meaning based on the context of the words, providing Seq2Seq models with a better understanding of the language (Kak & Bouman 2024).

In contrast to traditional one-hot vectors, word embeddings offer dense representations that encode semantic information efficiently and alleviates the burden on Seq2Seq models and enables them to learn from limited training data more effectively. Our experimentation with GloVe embeddings within a Bidirectional GRU architecture validated these findings. We observed an increase in model performance, particularly in tasks like sentiment analysis. The Bidirectional GRU architecture, by considering both past and future contexts of words, further augmented the model's comprehension and sentiment analysis capabilities (Asudani et al., 2023).

Furthermore, pre-trained word embeddings like GloVe facilitate better generalization and handling of out-of-vocabulary (OOV) words, which are common challenges in NLP tasks. By integrating GloVe embeddings into both the Encoder and Decoder of Seq2Seq models, we observed improved semantic understanding and coherence in generated sequences.

In our experimentation with our dataset, we used GloVe (Global Vectors for Word Representation) embeddings in a Bidirectional GRU (Gated Recurrent Unit) and found that there was an increase in model’s performance upon using these word embeddings. The model gained a deeper understanding of word meanings which improved comprehension and allowed the model to better grasp the sentiment of the dataset. Bidirectional GRU architecture helped the model to consider past and the future contexts of words which further enhanced its abilities. We concluded that leveraging word embeddings improved sequence-to-sequence model’s effectiveness in the sentiment analysis task.

We also employed GloVe within Logistic Regression and SVM classifier and discovered that it increased the performance of our models. Despite the straightforwardness of Logistic Regression, the results showed and accuracy of 76.57% and the SVM showed the accuracy of 81.08%, additionally the TF-IDF vectorizer when combined with Logistic Regression achieved an accuracy of 80.68%.

In summary, our experimentation underscores the significant impact of word embeddings, particularly GloVe, on enhancing the effectiveness and performance of Seq2Seq models and traditional classifiers in NLP tasks. By capturing semantic meaning and mitigating data sparsity challenges, word embeddings serve as indispensable tools for advancing the capabilities of NLP systems.

References:

1. *Kak , A. and Bouman, C. (2024) Word embeddings and sequence-to-sequence learning, engineering.purdue.edu. Available at: https://engineering.purdue.edu/DeepLearn/pdf-kak/Seq2Seq.pdf (Accessed: 27 March 2024).*
2. *Asudani, D.S., Nagwani, N.K. and Singh, P. (2023) 'Impact of word embedding models on text analytics in Deep Learning Environment: A Review', Artificial Intelligence Review, 56(9), pp. 10345–10425. doi:10.1007/s10462-023-10419-1.*