Evidence Retrieval

Enlong Bo and Wei Xiang Wong
University of Manchester

Abstract

This project explores two different approaches to handle evidence retrieval as a classification task. The poster will provide a side-by-side comparison of these two models, providing insight into their strengths and limitation and some of the different methods that were tried along the way.

Introduction

In Natural Language Understanding, evidence retrieval is a well-known pairwise task that tries at supporting or refute a given claim based on given evidence. This project aims to compare two distinct approaches:

- Long Short-Term Memory (LSTM) networks: Modelling word dependencies, allowing contextual text understanding which is critical when understanding evidence and claim.

- Logistic Regression: Statistic approach using Maximum Likelihood Estimation to find set of parameters resulting in largest sum likelihood over training set.

Theoretically, LSTM should perform better then Logistic Regression as it has access to word dependencies providing better context when classifying.

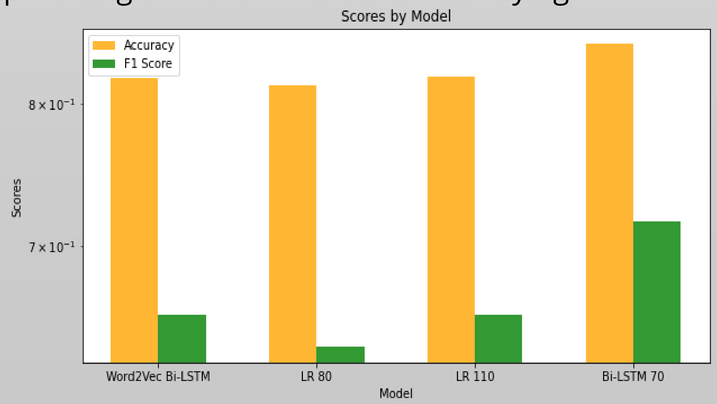
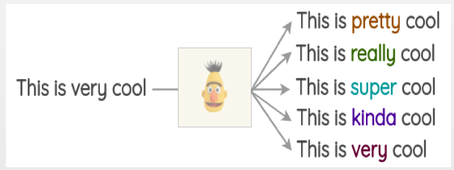


Figure 1: Accuracy and F1 Score of 4 different models

Methodology

- **Preprocess:**
  - Removing [REF] tags from Evidence column.
  - Apply Data Augmentation[1] to the train data (Synonym Replacement).
- **Generate Word embeddings**
  - Concatenate 'Claim' and 'Evidence' with a '[SEP]' tag.
  - Tokenize, generate embeddings using DistilBERT transformer[2].
- **Train models**:
  - **Logistic Regression:**
    - 1000 max iterations
  - **LSTM Model**:
    - 64 units Bi-LSTM.
    - 2 sets of 64 units Dense Layer with 0.3 dropout



Results

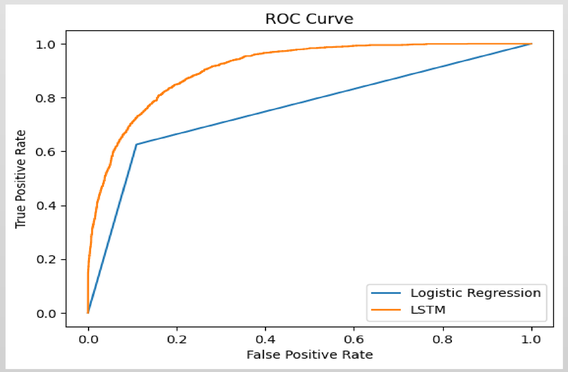


Figure 2: Roc Curve of the 2 best models

- Figure 1 demonstrates an increase in performance with longer maximum sequence length, supporting the fact that more context leads to better classification.
- DistilBERT's contextualized embeddings outperforms Word2Vec.
- LSTM showcase better performance over Logistic Regression in terms of both F1 score and accuracy even though it has a shorter maximum sequence length, showing the ability to capture sequential data and long-term dependencies within the text.

Conclusion

To conclude, here is what we found:

- Higher maximum sequential length helps with performance
- Data augmentation allow the model to generalise better[1]
- DistilBert[2] embeddings are more meaningful compared to more classic approaches like Word2Vec
- Deep learning methods performs better overall

Practical implication of these findings could influence development of different AI-driven tools like sentiment analysis and content moderation and hopefully help inform best practices for embedding and model selection.

What are some potential future steps?

- **Memory**: when creating embedding for large dataset it is easy to run out of memory hence it was important to have the right balance of maximum sequential length. Potentially, it may be better to explore with better hardware.
- **Ensemble**: Explore ensemble[3] networks combining the strengths of LSTM and Logistic Regression could yield even better performance.

Reference

[1] (Li, B., Hou, Y. and Che, W. (2022) 'Data augmentation approaches in natural language processing: A survey', AI Open, 3, pp. 71–90. Available at: https://doi.org/10.1016/j.aiopen.2022.03.001.

[2]Adoma, Acheampong Francisca, Henry, N.-M. and Chen, W. (2020) 'Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition', pp. 117–121. Available at: https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379.

[3]Dietterich, T.G. (2000) 'Ensemble Methods in Machine Learning', Multiple Classifier Systems, 1857, pp. 1–15. Available at: https://doi.org/10.1007/3-540-45014-9_1.