Analyzing Errors of Neural Models in Named Entity Recognition

By Dwija Parikh and Dr. Thamar Solorio (faculty mentor)
Department of Computer Science

UNIVERSITY of HOUSTON

Introduction & Background

Despite stellar performance on many NLP tasks, the behavior of neural models like BERT is not properly understood. We attempt to analyze the behavior and recognize patterns in errors for the NER task by BERT. Specifically, we try to answer:

- Does the model unintentionally memorize?
- Are there patterns in the errors generated?
- Does the model have true generalization ability?

Dataset and Model

We used the industry standard CoNLL 2003 dataset. Then we used a pretrained BERT for NER and obtained the following scores on the CoNLL dataset:

Metrics	
Classes	F1-Score
LOC	93.011
MISC	81.753
ORG	88.426
PER	95.885
macro avg	91.086
micro avg	91.056

Table: F1 Scores for BERT

The model performs well across all classes but there are still a considerable number of errors.
Some of these errors can be classified as annotation errors by humans are have not been considered.

Analysis

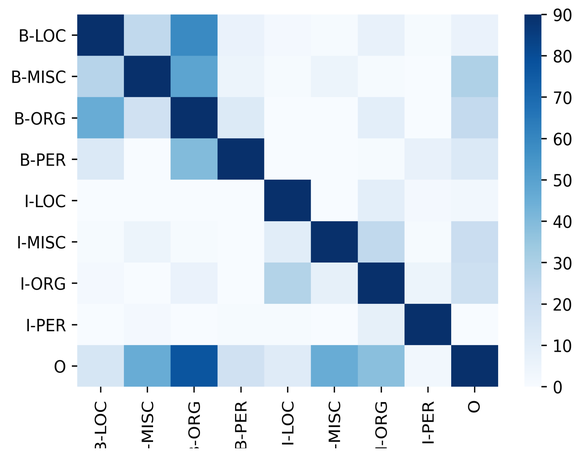

Figure: Confusion matrix for entities

- We begin by collecting errors by cross referencing gold standard with the model's predictions.
- Then we categorized these errors into broad types: **classification errors** and **sequential errors**.
- By analyzing the errors, we found a pattern of memorization. We calculated frequencies of entities appearing in both train and test sets.

Future Work

Our research shows that there are a few detectable patterns in the errors. The model also exhibits overfitting to an extent. However, there many cases of **unintentional memorization.** For example, there is evidence that model predictions are biased towards entities previously encountered in the train set with the same label. In cases where entities appear in the train set with multiple labels and no clear majority, the model behaves unpredictably.
Our next step is to further analyze trends in memorization. We plan to explore the effects of context on model performance. Another scope of research is fixing the errors made by the model and fine tuning it.

Acknowledgments

Special thanks to Dr Thamar Solorio for the invaluable mentorship this summer and to Gustavo Aguilar for his help.

References

- Carlini, Nicholas, et al. "The secret sharer: Evaluating and testing unintended memorization in neural networks." *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 2019.
- Helali, Mossad, Thomas Kleinbauer, and Dietrich Klakow. "Assessing Unintended Memorization in Neural Discriminative Sequence Models." 23rd International Conference on Text, Speech and Dialogue. 2020.
- Fu, Jinlan, et al. "Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study." AAAI. 2020.