Fine-tuning a Hierarchical Attention Based BiLSTM-CRF Model Using Minimal Labeled Data

Keerthana Komatineni*, MS[1], Rohan Goli*, MS[1], Shailesh Alluri, MS[1], Nina Hubig, PhD[1], Hua Min, PhD[2], Yang Gong, PhD[3], Dean F. Sittig, PhD[3], David Robinson, MD[4], Paul Biondich, MD[5], Adam Wright, PhD[6], Christian Nøhr, PhD[7], Timothy Law, DO[8], Arild Faxvaag, PhD[9], Ronald Gimbel, PhD[1], Lior Rennert, PhD[1], Xia Jing, PhD[1]

[1]Clemson University, Clemson, SC, USA; [2]George Mason University, Fairfax, VA, USA; [3]University of Texas Health Sciences Center at Houston, Houston, TX, USA; [4]Independent Consultant, Cumbria, UK; [5]Indiana University School of Medicine, Indianapolis, IN, USA; [6]Vanderbilt University Medical Center, Nashville, TN, USA; [7]Aalborg University, Aalborg, Denmark; [8]Ohio Musculoskeletal and Neurologic Institute, Ohio University, Athens, OH, USA; [9]Norwegian University of Science and Technology, Trondheim, Norway; *Coauthors contributed equally

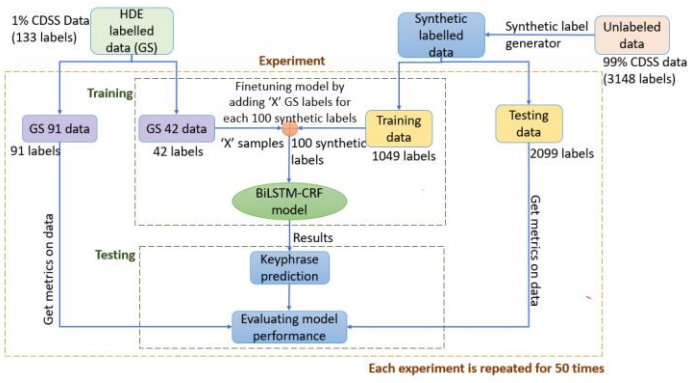
Introduction

Interoperability is a persistent challenge in healthcare. One specific challenge is sharing information between institutions due to lack of interoperability. Having interoperable Clinical Decision Support System (CDSS) rules can be achieved by building CDSS ontology. Ontology is usually constructed by human domain experts (HDE); however, this presents limitations in scalability and updating. An automatic approach can facilitate such manual efforts, e.g., by identifying keyphrases (KPs) automatically from literature. Our team harnesses natural language processing (NLP) and deep learning (DL) techniques to expedite KP identification to facilitate the building and updating of a CDSS ontology[1]. Although classic NLP algorithms often require human-labeled data, which is considered the gold standard (GS), our approach aims to minimize the use of GS and leverage unsupervised algorithms and synthetic labeling/similar to PU-Learning. This approach addresses the challenges of clinical NLP by complementing the arduous manual data labeling with lightweight DL models which facilitate CDSS ontology construction by identifying relevant KPs in the CDSS domain. The ontology empowered interoperable CDSS rules can provide a feasible pathway to achieve interoperable patient records, while also saving repeated efforts to develop and maintain individual CDSS rules institution by institution. Our effort aims to advance the foundational research in biomedical informatics.

Existing Model

The model we developed in [1] uses a multi-step machine learning process that includes generating synthetic labels for unlabeled data, pre-training, and constructing a hierarchical-attention-based Bidirectional Long Short-Term Memory network with Conditional Random Field (BiLSTM-CRF) model. The architecture of the model involves generating word embedding and introducing hierarchical-attention mechanisms. This method uses a combination of word- and sentence-level attention to recognize sentence and document structures, thereby facilitating KPs identification. In this study, we introduce the experiments used to fine-tune the model (Figure 1) and compare its performance with that of a minimally labeled dataset.

Fine-tuning Model



Experiment: The experiment involves iteratively fine-tuning the BiLSTM-CRF model by introducing 'X' GS labels for every 100 Synthetic labels during training using GS 42 and the Train dataset. Then the model performance is evaluated on Test and GS 91 dataset, where 'X' takes values such as 0, 2, 4, 6, 8, 10, 12. Hence, there will be a total of 7 experiments and each experiment is repeated for 50 times.

Figure 1. Process flow of experiments with a combination of 'X' GS labels for every 100 documents with synthetic labels.

To explore the model's performance when adding GS labels, we ran experiments on the hierarchical-attention-based BiLSTM-CRF model[1]. Our experiments involve combining GS labels with synthetic labels in varying proportions, such as 0, 2, 4, 6, 8, 10, and 12 GS-labeled documents, within each batch of 100 synthetic labeled documents, with 50 independent repetitions for each combination. This methodology allowed us to compare the learning performance more accurately with human

input. The flow of experiments with 50 repetitions is shown in Figure 1. We evaluated the model's performance using average and variance of Precision, Recall, F1-score, and Accuracy measures over 50 repetitions, with a focus on the F1-score as the primary measure for its balanced consideration of precision and recall, particularly in sequence-to-sequence multi-classification tasks.

Results

Figure 2 shows the results of the experiment. The model performance does not follow a linear trajectory over all repetitions. This is expected because we trained the model independently for each repetition and experiment. Additionally, it typically took 10 to 15 repetitions for the model to reach an average level of performance. This indicates that, although the model's performance has ups and downs across repetitions, its performance is close to the average after approximately 10 to 15 repetitions. When considering the best combination of GS labels (HDE efforts) and synthetic labels, we noted the following.

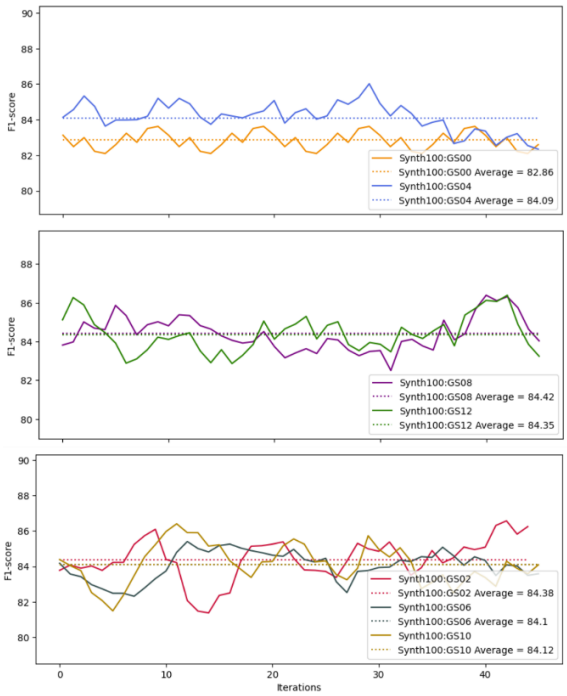


Figure 2. BiLSTM-CRF model performance: a comparison of F1-scores for different combinations of GS and Synthetic labels.

1.An Increase in model performance after adding GS labels: A 2% enhancement was achieved by including two GS labels for every batch of 100 synthetically labeled documents, which indicates the model learns more effectively when GS labels are integrated into the training process. If incorporating randomly selected two GS labels can enhance the performance of a model trained on nontruth labels, an even greater performance may be achieved by incorporating strategically chosen intelligent GS labels during training through Active Learning[2].

2.No upward trend of plot: Continuously increasing the number of ground-truth GS labels during training did not yield a substantial upward trend in the performance of the model. Although there was a slight improvement when 12 GS labels were added for every 100 documents with synthetic labels, the improvement remained relatively modest. Our initial objective was to achieve "learning more with less labeled data." A comparison of the performance of Synth 100: GS 12 and Synth 100: GS 02 suggests that the Synth 100: GS 02 combination is a better choice.

3.Variance of the F1-score metric: Lower variance of the model indicates the model is more robust. In Figure 2, the model performance is similar for Synth 100: GS 12 and Synth 100: GS 02, but the variance of the model is less at Synth 100: GS 02, making this combination more robust.

Discussion

Our experiments showed that the performance of the model did not exhibit a linear trend or consistent patterns across repetitions. Our goal was to find the best performance using the least HDE labeled data. The combination of two documents with GS labels for every 100 documents with synthetic labels significantly improved the model's performance. These experiments suggest that good results do not necessarily require a large amount of labeled data. We are currently investigating the strategic selection and incorporation of intelligent GS labels during training through active learning, as it may achieve even more significant improvements in the model. To the best of our knowledge, our work is the first operational framework within the CDSS sub-domain designed to identify KPs.

Acknowledgment

The work is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM138589.

References

1. Goli R, Hubig N, Min H, Gong Y, Sittig DF, Rennert L, Robinson D, Biondich P, Wright A, Nøhr C, Law T, Faxvaag A, Weaver A, Gimbel R, Jing X. Keyphrase Identification Using Minimal Labeled Data with Hierarchical Context and Transfer Learning. medRxiv [Preprint]. 2023 May 26:2023.01.26.23285060. doi: 10.1101/2023.01.26.23285060. PMID: 37292830; PMCID: PMC10246160.
2. Monarch RM. Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. Shelter Island, NY: Manning Publications, 2021.