Keyphrase Identification with a Limited Labeled Dataset Using Deep Active Learning and Domain Adaptation

Rohan Goli, 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

Introduction

Interoperability is a well-recognized barrier in the health information technology field. To facilitate the interoperability of clinical decision support systems' (CDSS) rules, we propose using Semantic Web technologies to build an ontology for CDSS. To iteratively improve the identification of concepts from unseen corpora and produce results comparable to a human domain expert (HDE) annotator, we are building a keyphrase (KP) identification model by using available CDSS text resources with minimal human feedback. This model will provide candidate phrases for HDE to review before adding to the CDSS ontology.

Methods

Machine learning (ML) models dealing with identifying entities in a text sequence (viz., KP Identification) would be considered sequence labeling tasks. In natural language processing (NLP), cutting-edge language models like GPT & BERT have been quite popular in accomplishing such a task using the context information with attention. Additionally, unsupervised algorithms[1] have a prominent role in similar tasks through grouping KP by similarity using statistical[1] features and embeddings[1]. These models are either computationally costly to train, require a massive amount of labeled data (L-Data), or cannot incorporate human expertise – do not satisfy all of the required constraints. Being neither explainable nor interpretable and unable to work with human feedback does not align with our goals.

Although a supervised learning model with a labeled text corpus and one or more features, such as statistical[1], embedding[1], linguistic[1] (e.g., Part-Of-Speech Tag), or context[1] (e.g., relative position, previous/following token information) as provided by Zhang et al., 2021[2] looks convincing to use, it requires enormous L-Data to train. The biggest challenge involves generating an L-Data for the CDSS from PubMed, as HDE annotation is expensive. So, the model has to learn the patterns with limited L-Data [1.2%] and extensive unlabeled data (U-Data) [98.8%].

While the high-performant prior NLP models and algorithms require expensive and enormous HDE annotated L-Data for model training, we propose a hybrid approach to solve the challenges by harnessing the potential of iterative or continuous human feedback on the model's predictions trained with minimal L-Data and supervision.

To create such a system, firstly, we generate synthetic biomedical labels for U-Data and combine them with actual CDSS labels for L-Data. Secondly, for domain adaptation[3] with an imbalanced dataset, we use deep active learning[4] (AL), and transfer learning[3] (TL) approaches to train a model that identifies relevant entities such as KP in any text. Here, the special semi-supervised learning approach AL[4] incorporates active human feedback on the least confident prediction results for diverse CDSS concepts.

In this regard, we develop an ML pipeline that works on titles and abstracts of research papers in the CDSS domain as part of iterative improvement. The experiment selects the best subset of input representations (i.e., statistical[1], linguistic[1], context[1], word/document/sentence/character-level embeddings[1]) and sentence-level attention that contribute to performance in identifying a KP. The components of the pipeline include: (1) a Pre-trained model for synthetic label generation; (2) Input Context Encoder[2] for combining various text features into an input representation vector; (3) Sequence Labeling Bidirectional Long-Short Term Memory (Bi-LSTM) Model[2] for identifying each token of a given text; (4) Output Tag Decoder[2] for predicting the best combination of KP annotations in each text (BIO Format); and (5) Active Learner[4], select most diverse and independent unlabeled text for human review.

Results

To overcome the limited L-Data, we generated synthetic labels using the scispaCy BERT model pre-trained on Biomedical entities. This model outperforms several unsupervised ranking algorithms (Multi-partite, Position & Topic) by 13% on F1 scores to closely match HDE annotations. Figure 1 shows the performance metrics on a 1% sample data compared to HDE annotations. Figure 2 shows our Bi-LSTM-CRF[2] model, which predicts the best combination of KP annotations from a given text sequence.

Accuracy	69.1%
Precision	37.1%
Recall	81.38%
Specificity	66.12%
F1-Score	50.97%

Figure 1: Performance metrics of synthetic label generation

Also, we want to use a pre-trained Language Model and pre-trained word embeddings on CDSS articles so that the model can understand the probability distribution of the words (or N-gram sequences) in our domain. To further enlighten the Bi-LSTM-CRF[2] model with the context of the research article while working on a single sentence, we calculate global neural attention over the other sentences of the same article. It helps the model to match human context understanding while identifying the current KP closely. By applying domain adaptation[3] from the Biomedical-BERT model to a CDSS-specific Bi-LSTM-CRF[2] model with various text features, context embeddings, pre-trained language model, sentence-level attention, and active learning sampling techniques[4] (Uncertainty or Diversity based test sample selection for iterative human feedback), we anticipate the pipeline results will be similar to those from domain experts with continuous feedback and improvement.

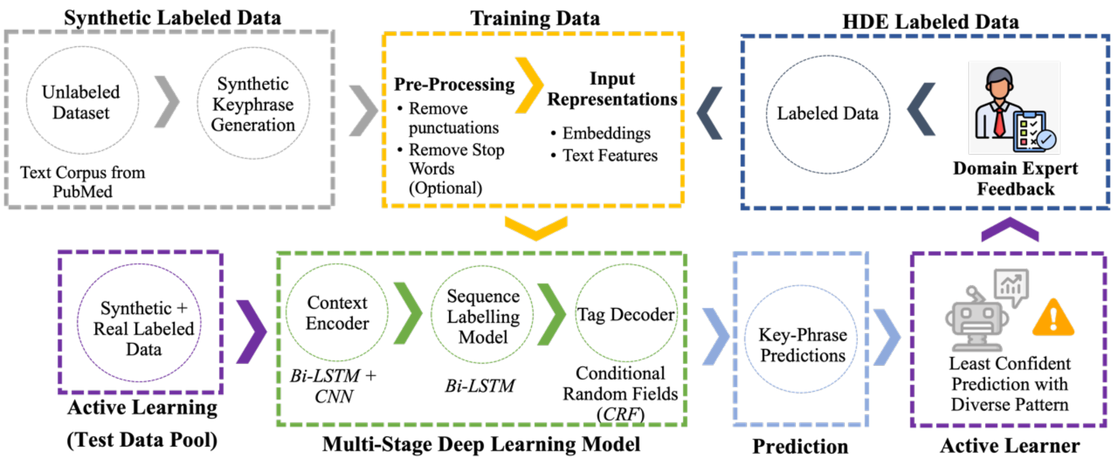

Figure 2: Architecture of Keyphrase Annotation Identification

Discussion

We are developing an ML pipeline and will share the codes when the model performs satisfactorily. With real-time annotators' feedback, we believe the performance of the pipeline will improve over time. It also arises new challenges like identifying the qualified HDEs for building crowdsourced annotations, consensus between HDEs for the actual labels, and selection bias over the research papers for human feedback. All the challenges need to be explored and addressed further. The output of this work will contribute to the human review process while constructing and maintaining the CDSS ontology, and the methodology will contribute to NLP. Knowledge graphs can be built over the entities and their context in the future to facilitate explainable and interpretable predictions.

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. Eirini Papagiannopoulou, Grigorios Tsoumakas. A Review of Keyphrase Extraction. 10.10.1002/widm.1339
2. Chengzhi Zhang, Lei Zhao, Mengyuan Zhao, Yingyi Zhang. Enhancing Keyphrase Extraction from Academic Articles with their Reference Information. 10.1007/s11192-021-04230-4
3. Guoliang Guan, Min Zhu. 2019. New Research on Transfer Learning Model of Named Entity Recognition. 10.1088/1742-6596/1267/1/012017
4. Fumimaro Odakura, Koga Kobayashi, Kei Wakabayashi. Active Learning for Extracting Technical Terms Covering Multiword Phrases. 10.1145/3487664.2487706