CS598 DL4HC Reproducibility Project Proposal

Michael Miller and Kurt Tuohy
{msmille3, ktuohy}@illinois.edu

Presentation link: n/a
Code link: https://github.com/mich1eal/cs598_dl4hc

1 Paper ID 41: Symptom Similarity Analysis

1.1 General Problem

The first paper we propose reproducing is "Disease Prediction and Early Intervention System Based on Symptom Similarity Analysis" by Zhang, Huang, and Li.

The primary goal of this paper is to rate the similarity of pairs of sentences. The idea is to aid prediction of disease by automatically identifying similarities in descriptions of patient symptoms. However, the paper does not use healthcare-related text.

1.2 Specific Novel Approach

Zhang et al. (2019) preprocess each sentence by extracting its "trunk" – the subject, predicate and object. The researchers found that these trunks help identify sentence similarity more reliably.

1.3 Specific Hypotheses to Verify

We propose reproducing experiment 1. This creates and tests a classifier using a dataset of human-labeled sentence similarities. Experiment 1 compares the authors' method to a baseline plus the methods in four related papers.

1.4 Additional Ablations

For the project, we are considering performing one/several of the following ablations:

- Incorporate synonyms, antonyms, and other word relationships into the model to gauge how this affects ratings of sentence similarities. The paper doesn't use healthcare data, so normal English synonyms could be a proxy for gauging whether the model's performance on healthcare data would improve by incorporating medical ontologies. One potential source of synonyms is the WordNet synonyms dataset on Kaggle.

- Add an attention mechanism to learn which portions of sentences contribute most to similarity scores.

- Reduce the size of the training set to learn how sensitive the outcome is to the training corpus size.

- Use a dependency parser instead of a constituency parser (no change expected for sub-verb-obj).

- Incorporate additional sentence parts of the sentence into the model (e.g. one modifier each for sub-verb-obj).

1.5 Data Access

The training and testing datasets are public: the Microsoft Research Paraphrase Corpus (MSRP) and the SemEval Semantic Text Similarity (STS) datasets.

1.6 Computational Feasibility

The training dataset is small: 4,076 pairs of sentences. The authors use the Stanford Parser to extract sentence trunks, and this parser is publicly available. Finally, the authors' model uses only a small CNN, with one convolution layer and one pooling layer. We do not anticipate challenges in terms of computation power.

The authors evaluate their novel model plus a number of models defined in other papers. Evaluating this large number of models may be computationally expensive. To avoid this issue, only the paper's model and a baseline model may be evaluated.

1.7 Code Reuse

To date, the authors have not responded to our request for code. We plan to implement these methods ourselves.

If useful, we may take advantage of Python modules such as Spacy or the Natural Language Tool Kit for word tokenization and dependency parsing. Additional modules such as PyTorch and scikit-learn may be used for data loading, model construction, and model evaluation.

2 Paper ID 68: MuLan: Multilevel Language-Based Representation Learning

2.1 General Problem

Our second paper is "MuLan: Multilevel Language-based Representation Learning for Disease Progression Modeling" by Sohn, Park, and Chi. This paper seeks to:

- Summarize EHRs as a chronology of medical events
- Interpret and visualize disease progression
- Provide early detection of dangerous conditions such as septic shock

2.2 Specific Novel Approach

Unlike with Med2Vec, the authors model EHR hierarchies by using supervised learning for the final stage. This learns the relationship between patient visits and medical events in terms of a target condition such as septic shock.

Before the final stage, the authors model medical events as a sequence of states and state changes, such as blood pressure changing from low to high. They then use these state representations to identify relationships between events. This makes a difference for identifying septic shock, which has no single biological marker.

2.3 Specific Hypotheses to Verify

We would verify the prediction of septic shock with a 40-hour hold-off window. We would compare four variations of MuLan with two baseline models: one based on Med2Vec and one based on (relatively) raw EHR features.

2.4 Additional Ablations

For the project, we are considering performing one/several of the following ablations:

- Before training the models, we would remove medical events that may represent target leakage, such as administration of vasopressors.
- We could vary the size of the training dataset to learn the model's sensitivity to the amount of data available.

2.5 Data Access

This paper uses data from the Christiana Care Health System (CCHS). This data is not publicly available. The authors of this paper were contacted to determine if the data can be retrieved. At the time of writing, no response has been received from the authors.

Without access to the authors' dataset, it may be infeasible to recreate this paper's results. If we desired to proceed, we could take the following course of action:

1. Substitute MIMIC-III data for the original dataset
2. Select a study population using the same criteria as the researchers
3. Label the presence of septic shock using the criteria listed in the paper

2.6 Computational Feasibility

Zhang et al.'s (2019) dataset is relatively large. The final set contains 5,928 visits with 795,314 events. Input vectors are feature-rich, with 10+ elements per timestep. Models are trained in multiple steps including pretraining. The models themselves are relatively complicated with numerous layers and channels. These factors suggest that computation power will need to be considered when determining whether or not to proceed with reproducing this paper's results.

Computational power requirements will be reassessed if/when the dataset becomes available.

2.7 Code Reuse

Without an author response, we would write our own code. Modules such as PyTorch and scikit-learn may be used for data loading, model construction, and model evaluation.

3 Paper ID 117: NLP for Cognitive Therapy

3.1 General Problem

Our final paper is "Natural language processing for cognitive therapy: Extracting schemas

from thought records" by Burger, Neerincx, and Brinkman. This paper seeks to classify a human subject's *maladaptive schema* based on their unstructured text input.

3.2 Specific Novel Approach

The task attempted by Burger et al. (2021) is particularly challenging. Inputs are unstructured natural language utterances from a variety of authors. In addition to the usual challenges of natural language, the classification target (psychological mental state) is highly subjective and likely varies person to person. This task is challenging for humans, requiring trained professionals who often do not reach consensus. Due to these obstacles, this task will likely be more challenging than typical natural language processing applications.

Not surprisingly, the authors found only limited existing use of NLP techniques as psychological assessment tools for interpreting free text. Essentially, the field is wide open. The authors performed their study as a benchmark for future research to improve upon. For example, the researchers used plain-vanilla word embeddings based on a Wikipedia-trained language model, and they used relatively simple LSTMs. More state-of-the-art approaches might improve on the researchers' results.

3.3 Specific Hypotheses to Verify

We will attempt to verify the paper's hypothesis H1. This hypothesis asserts that mental health schema can be predicted given an unstructured utterance.

3.4 Additional Ablations

For the project, we are considering performing one/several of the following ablations:

- Use a state of the art model like BERT and compare results to the original paper
- Use a hierarchical model to take advantage of structure in downward arrow data
- Add attention mechanism to identify what parts of utterances predict their schemas best

3.5 Data Access

All data used in this paper is publicly available on the 4TU.ResearchData repository. It has been successfully retrieved.

3.6 Computational Feasibility

This study relies on a relatively small dataset (5747 utterances). The models used are not especially complicated. However, Burger et al. (2021) trained hundreds of RNN models to decrease the role of chance in the final results, and training took about 18 hours. If computational time becomes prohibitive, we plan to reduce the number of models trained while still confirming key results.

3.7 Code Reuse

The code from this study is available on 4TU.ResearchData. The code will be reviewed, however new code will be written for this project.

If useful, we may take advantage of Python modules such as Spacy or the Natural Language Tool Kit for word tokenization and dependency parsing. Additional modules such as PyTorch and scikit-learn may be used for data loading, model construction, and model evaluation.

References

Franziska Burger, Mark A. Neerincx, and Willem-Paul Brinkman. 2021. Natural language processing for cognitive therapy: Extracting schemas from thought records. *PLOS ONE*, 16(10):e0257832.

Hyunwoo Sohn, Kyungjin Park, and Min Chi. 2020. MuLan: Multilevel Language-based Representation Learning for Disease Progression Modeling. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1246–1255.

Peiying Zhang, Xingzhe Huang, and Maozhen Li. 2019. Disease Prediction and Early Intervention System Based on Symptom Similarity Analysis. *IEEE Access*, 7:176484–176494.