lign 167 project proposal

Richard Li, Brian Chu, Julie Hang, Kyle Batalla

what task are you trying to accomplish?

The task we intend on accomplishing is from the Microsoft BLURB leaderboard / benchmark for biomedical natural language processing (https://microsoft.github.io/BLURB/tasks.html#dataset_ddi). The specific task is relation extraction with drug-drug interactions, with the goal of being able to predict adverse drug effects. We will be provided with entity names (i.e., the drugs), from which our goal will be to create a form of encoding to serve as embeddings for their interactions.

The natural language processing aspect of this task lies in the source of the dataset. The dataset is derived from named entity recognition on a large variety of biomedical papers. From there, the task centers around relation extraction of those named entities, such that we can generate a means of **predicting how multiple drugs will interact in relation with each other, without prior interactions.**

This is notable due to significant amounts of prior research from which we can learn others' approaches. The drug-drug interaction relation extraction benchmark (DDI) was used as a competition task in 2013, and thus has plenty of background information surrounding it.

The advantages of pursuing this task center around the existence of annotated datasets and reference solutions from which we can benchmark our accuracy, and thereby evaluate our success in this task.

---

how are you planning to do it?

Our approach centers on studying previously successful approaches to this task, and aiming to replicating the methodology implemented by others. Notably, many researchers have utilized knowledge graph generation to encode the relation extraction between the named drug entities, then used graph neural networks to predict the outcomes.

There are two sources of previous research into this specific task that we find fascinating -

i. **KGNN** (Lin et al.)
   - https://github.com/xzenglab/KGNN
   - https://www.ijcai.org/proceedings/2020/380

   We found this to accurately demonstrate the exact same approach we intend on utilizing. It requires no sources of additional knowledge to provide contextual information to improve model performance, and creates a novel methodology of automatically generating a knowledge graph from a matrix of every drug's interaction.

ii. **SumGNN** (Yu et al.)
   - https://github.com/yueyu1030/SumGNN
   - https://arxiv.org/abs/2010.01450

Similarly, SumGNN uses graph neural networks to power its prediction, but differs from KGNN because it focuses on extracting subgraphs from the primary knowledge graph, in order to generate a reasoning path for the predictions.

Further, we did research into how knowledge graphs can be generated, in order to suit our needs. Notably, we came across **MAMA** (Wang, et al.; https://arxiv.org/abs/2010.11967) as a means of creating knowledge graphs with unsupervised construction.

https://github.com/theblackcat102/language-models-are-knowledge-graphs-pytorch

Between our research into how we can create knowledge graphs, and how to apply them to create a knowledge graph, we believe that this approach will permit us to align our approach with previous tested methodologies, and demonstrate our understanding of unsupervised machine learning and graph neural networks. Lastly, we aim to find ways of improving upon these methodologies with variations - a hypothetical would be an implementation of KGNN with additional information - what would be the impact?

---

how are you planning to evaluate whether you have succeeded?

The benefits of committing to this task lie in the pre-existing annotation and named entity recognition of these datasets, as well as provided reference solutions from the competition task. As a result, we can use these benchmarks as a means of evaluating our efficacy.

We define success not in terms of F1-score or other benchmarking, but personally feel as if we will attain success in this endeavor provided that our model is *somewhat* effective. We understand the constraints of time and our experience in embarking on this project, and aim to create a model that demonstrates our understanding of the subject matter, although we aspire to reach efficacy standards set by the prior research referenced above.