Scientific Relation Extraction: Experimental Protocol

Cole Winstanley
colew@stanford.edu

Ethan Shen
ezshen@stanford.edu

Eli Pugh
epugh@stanford.edu

1 Task

Textbooks contain a vast wealth of information, but it is often difficult to find the answer to a specific question in the dense text. We adapt the natural language relation extraction task to solving this textbook query problem. In particular, we are building a system that takes in preprocessed sentences from a textbook and builds a knowledge base that can be queried with specific questions. In addition, this knowledge base will be able to point out the specific part of the text that corresponds to the answer to a question.

Our model will be trained on a textbook's raw text. The resulting model will can be queried using pairs of entities, and will return the relationship between those entities, if there is one. For example, it might give the following:

```
(cell wall, cell membrane) => "encloses"
```

The task of parsing a natural language question into such a query is not within the scope of this project.

2 Hypothesis

- A simple bidirectional LSTM model [6] with BERT word embeddings will perform well on the LIFE dataset as it has on the TACRED dataset.
- Using SciBERT embeddings, potentially along with regularization tweaks, will beat BERT embeddings on LIFE dataset because BERT lacks scientific vocabulary. This will allow us to perform useful relation extraction for knowledge inference on LIFE dataset.

3 Data

We will use two datasets when training and testing our models. The first is Text Analysis Conference Relation Extraction Dataset (TACRED) [7], and the second is the Intelligent Life biology textbook [4].

3.1 TACRED

TACRED is a widely-used relation extraction dataset compiled by Stanford researchers. It's built from the TAC Knowledge Base Population (TAC KBP) corpus which includes newswire and web text. TACRED has 106,264 examples spread over 41 relation types as well as the special no_relation type. [7]

The TACRED examples include the context sentence, subject and object spans, Stanford named entity recognition (NER) types, and the truth relation. [7] The examples are well-annotated and the NER types give extra information not found in raw text that is often useful for predicting relation type.

TACRED was built to more closely mirror real-world situations than existing datasets. The average sentence length is 36.4, which is makes predicting relations more challenging when compared to datasets with shorter sentence length. In addition, 79.5

Another advantage of the TACRED dataset is that Stanford researchers have also included benchmark performances of many standard models, such as logistic regression, patterns, CNNs, LSTMs, LSTMs with position aware attention, and other neural systems. In addition, there is a repository linked that uses PyTorch to implement a very performant position-aware attention model. [7]

3.2 Intelligent Life

Intelligent Life [4] is a biology textbook that is currently building a knowledge base (KB) for question answering as well as other tools like visualizing concept graphs. Currently only the first 10 chapters of the knowledge base have been human-verified. This includes 150,000 relation instances from around 20 relation types. While only the first 10 chapters have human-verified relations, the entire text is included, giving around 30,000 sentences of raw text. [4]

Since Intelligent Life is a biology textbook, it has a very domain-specific vocabulary. This poses a challenge, since many models use word embedding methods that will have a large number of lookup misses. This means that any model that wishes to perform well on this dataset will need to use character-level embedding or be trained on a scientific corpus.

Another challenge that Intelligent Life presents is that only 10 chapters are human labelled. This means that distant supervision will almost certainly improve accuracy, since there is very limited training data in only the first 10 chapters. This makes learning relations for the remaining chapters tricky when compared to a more simple supervised training approach used with TACRED.

4 Metrics

We intend to use mostly standard relation extraction metrics to evaluate how well our models perform the task. One challenge with relation extraction as a task is that recall is often difficult to define due to the distant supervision aspect; relationships that are not present in a knowledgebase or evaluation dataset are not necessarily actual negatives. So that we can address these issues, we will begin our evaluation of our model on our target dataset (the LIFE dataset) based on a confusion matrix, where "no relation" is a category in addition to the approximately 20 relations in the dataset. Our primary target will be the $F_{0.5}$ macro average over the 20 defined relations. This gives precision more weight than recall.

The focus on precision recognizes specific features of the task of building a question and answer system for a textbook. For a student trying to learn from the textbook, it is much more harmful to give a wrong answer to a question than giving no answer at all. Giving a wrong answer could deceive the student and therefore be deleterious on the effectiveness of the textbook. Using the macro average instead of the micro average recognizes that in general, it is more important for students to learn the relations themselves, instead of the instances of them. (e.g. it is more important to understand what a symbiosis is than the fact that ants and acacia are an example of it.) Therefore, the macro average, which prioritizes performance on the relation level independent of the number of instances, correlates best to benefit for student users.

4.1 Baselines

A difficulty of working with a new dataset is that baselines are difficult to formulate for a task that has not been attempted before. For this reason, we intend to bring in both outside datasets and outside models to create baselines to evaluate our system against. We discuss the choice of baselines in the General Reasoning section below.

When evaluating on the baseline TACRED dataset, it will be appropriate to analyze F_1 macro average, since that dataset has a more complete set of ground-truth relations defined, so we can be more confident that if our model predicts a relation for a triple marked as no relation, that this is an

actual false positive (i.e. recall is more meaningful). Ideally, good recall on TACRED compared with state-of-the-art models for TACRED will also correlate to good recall on the LIFE dataset, even though we don't have a way of directly evaluating it.

4.2 Error Analysis

We intend to perform extensive error analysis on our final model. We will follow some of the guidelines of Schneider et al.[5] in organizing errors into categories based on what part of the model failed: the featurizer, parser, or the extraction model. We will also analyze cases that impacted our final $F_{0.5}$ macro score, but do not actually indicate a problem with the model (rather, a problem with the metric or the evaluation dataset). We will also evaluate cases of missed extraction, where a relation failed to be extracted from a sentence. Characterizing the errors in a useful way will lead to understanding of where the best targets for improvement lie.

5 Models

We plan to use different models described in the TACRED dataset release, as well as propose an update to their highest performing model. TACRED compares results from many models including patterns, logistic regression, CNN, LSTM, and LSTM with position-aware attention. [7] We hope to test their LSTM and LSTM with position-aware attention models that are publicly released.

The most performant model from the TACRED dataset tests is described in their paper, "Position-aware Attention and Supervised Data Improve Slot Filling". [7] This model, shown in Figure 1, uses GloVe word embeddings [3] as inputs $\mathbf{x}_1, \cdots, \mathbf{x}_n$. It then passes these through an LSTM model where the hidden states are $\mathbf{h}_1, \cdots, \mathbf{h}_n$, and the final state $\mathbf{q}$ is $\mathbf{h}_n$. The positional encodings $\mathbf{p}_i^s$ are

$$\mathbf{p}_i^s = \begin{cases} i - s_1, & i < s_1 \\ 0, & s_1 \leq i \leq s_2 \\ i - s_2, & i > s_2 \end{cases}$$

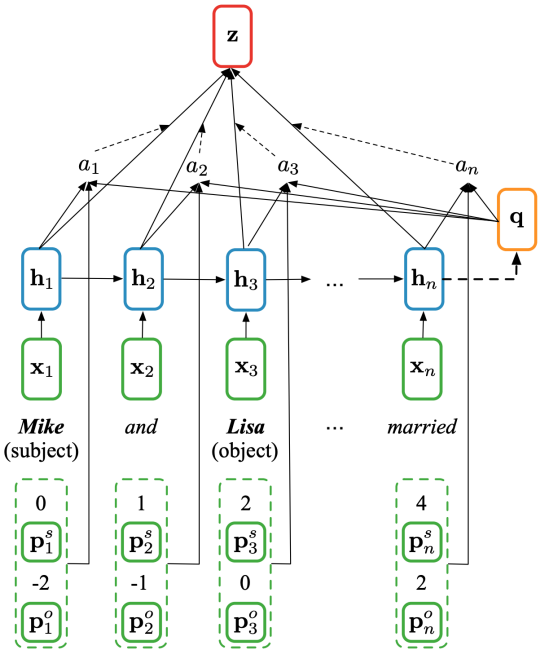
The attention weights a_i are calculated as

$$u_i = \mathbf{v}^T \tanh(\mathbf{W}_h \mathbf{h}_i + \mathbf{W}_q \mathbf{q} + \mathbf{W}_s \mathbf{p}_i^s + \mathbf{W}_o \mathbf{p}_i^o)$$

$$a_i = \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)}$$

where the $\mathbf{W}$s and $\mathbf{v}$ are trainable parameters. The final sentence representation $\mathbf{z}$ is calculated as

$$\mathbf{z} = \sum_{i=1}^{n} a_i \mathbf{h}_i,$$



LSTM with position-aware attention

which is then fed into a fully-connected softmax classifier to predict which of the 41 relations (or no_relation) the example exhibits.

In order to make an extension to this already impressive model, we plan to utilize BERT [2] and SciBERT [1] contextual embeddings as inputs $\mathbf{x}_1, \cdots, \mathbf{x}_n$ rather than GloVe embeddings. This idea stems from the work of Shi and Lin [6], where they use BERT embeddings and a bidirectional LSTM to feed into a mult-layer perceptron classifier. The bidirectionality could be a good addition to Zhang et al.'s position-aware attention model, and the BERT embeddings achieve SOTA results on almost all NLP tasks. While we realize that one potential advantage of using BERT embeddings is their built-in positional encoding, we hope that our model will still see some performance gains

with the position-aware attention before feeding to the output $\mathbf{z}$. If this is not the case, we may do ablative analysis to see what changes in performance BERT adds when compared to the original position-aware attention with GloVe.

6 General Reasoning

Here, we introduce and explain three planned baseline experiments for relation extraction using the models and datasets detailed above: BERT on TACRED, SciBERT on TACRED, and BERT on LIFE.

Our initial baseline experiment will be to use BERT embeddings for relation extraction as described in the above section on the TACRED dataset. We do this to validate that our model described above is indeed functional, and can match published results on an established dataset. We don't intend to do extensive hyperparameter tuning, or expect metric scores to exactly match or surpass these benchmarks, our purpose will be satisfied with scores in the general ballpark. In the same general spirit, we will replace the BERT backbone in our model with SciBERT, and train on the TACRED dataset to ensure that the model will train smoothly. This serves a similar purpose as the previous baseline experiment, and we expect that scores be around the same, or lower because it is biased towards scientific vocabulary. Our final baseline will be to train our BERT relation extraction model on the LIFE dataset. Here, we expect that our model will do poorly because the pretrained embeddings were trained on general text corpora such as news articles and Wikipedia, while the LIFE vocabulary will include specific scientific words/phrases. This issue is symptomatic of many applications in specialized fields, which restricts the effectiveness of NLP models in these cases.

Here, we present a few potential experiments to alleviate this issue, including entity masking. As a potential lead, Alt. et al. present entity masking as an important regularizer which increases model generalization for a relation extraction task on the TACRED dataset. These strategies may include replacing entity mentions with UNK tokens, named entity types (NE), or grammatical roles (GR). While BERT may already account for this using the way it breaks up words before embedding, Alt. et al. show that a combination of NE and GR gives the highest scores, suggesting that NE and GR entity masking on scientific words that are not in the dataset may give a boost to BERT relation extraction scores on the LIFE dataset.

Another approach is to use embeddings pre-trained on large corpora of scientific data, which was published by Beltagy et al. as SciBERT. The vocabulary in LIFE has a much higher overlap with the SciBERT vocabulary because of the scientific training data, which will perhaps alleviate the problems detailed above. This approach shows promise due to Beltagy et al's preliminary experiments in scientific domain-specific tasks such as sequence tagging, parsing, and text classification, showing SOTA results. By comparing our SciBERT results on the LIFE relation extraction task with our entity masking approach, as well as with our baselines, we hope to explore our central question on whether vocabulary specificity or masking is more effective for model generalization.

7 Progress Summary

We have collected the LIFE and TACRED datasets and prepared preprocessing scripts to load the data from the raw files into PyTorch. We have run a few initial checks to characterize the datasets and check for inconsistencies. We have also verified some results of [7] on our local machines with full TACRED data and GLoVE embeddings, without training for the full number of epochs because of the computational work necessary.

The remaining work involves adapting the model to use BERT/SciBERT embeddings, and to use the LIFE dataset. Additionally, the error analysis step after running the experiment will involve some new infrastructure to extract the error examples.

References

[1] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[4] David E. Sadava, David M. Hillis, H Craig Heller, and Sally D. Hacker. *Life: The Science of Biology*, volume 11. 2017.

[5] Rudolf Schneider, Tom Oberhauser, Tobias Klatt, Felix A. Gers, and Alexander Löser. Analysing errors of open information extraction systems. *CoRR*, abs/1707.07499, 2017.

[6] Peng Shi and Jimmy Lin. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255, 2019.

[7] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017.