MesEval: Counts and Measurements

SemEval 2021 Task 8

Team Name - CS60075_team22

Team Members –

Name	Roll no.	Contribution
Rishiraj Guha Ray	18CH10044	**Subtask 1** – Preprocessing, Model selection and training, hyper-parameter tuning, score evaluation of model **Subtask 3**-Preprocessing,Model selection and training, hyper-parameter tuning, score evaluation of model **Report writing**
Debapriyo Mukherjee	18IE10035	**Subtask 1** – Preprocessing, Model selection and training, score evaluation of model **Subtask 3**-Preprocessing, Model selection and training, score evaluation of model
K Rishith Reddy	16CS30016	**Subtask 1**-Preprocessing, Model Evaluation on validation /trial set and finding the final predictions. Fine-Tuning the model. **Subtask 3**-Preprocessing, Model Evaluation on validation /trial set and finding the final predictions. Fine-Tuning the model.
Venkata Vamsi	18CS10012	**NIL**
Akhilesh Hessa	14CS30003	**NIL**

Github Link - https://github.com/rishirajguharay/CS60075-Team-22-Task-8

APPROACH

The task requires identification of quantity spans and the corresponding entity spans and property spans. The task of identifying the quantity spans falls under the domain of entity extraction whereas the task of identifying the Measured-Property and Measured-Entity is a domain of Information Extraction, or more precisely Relation Extraction.

Entity Extraction can be done in various ways. We tried our hand with tools available in the spacy and NLTK library. We also tried using Conditional Random Fields but they were not giving high scores (F1-score). So we approached the problem in a deep learning way. We could have used LSTMs, Bi-LSTMs, BERT, etc. but because BERT usually outperforms a normal Bi-LSTM model we used BERT. Since the text is derived from scientific documents it would be more efficient if we use SciBERT (for tokenization purposes) which is a pre-trained BERT model on scientific documents. The steps are briefly described below :-

1. We tokenized the paragraph into sentences using spacy. The individual sentences were further tokenized into individual words using BertTokenizer pre trained on Scientific Corpus i.e SciBERT . The token were then converted to their respective ids. The maximum length is restricted to 256.
2. Input sentences were tokenized using SciBERT tokenizer from Huggingface implementation. The Quantity span was transformed into BIO/IOB format and used as true labels for training the model. The tokenized sentence is passed through

SciBERT. Tanh activation function is applied over the final hidden state of SciBERT.

3. CRF is a probabilistic model that makes it possible to extract structural dependencies using the BIO tags. We trained the model using CRF loss and Stochastic Gradient Descent optimizer.
4. We insert the special symbol “$” at the beginning and end of the quantity span. The modified sentences are tokenized using a SciBERT tokenizer. The span of the MeasuredEntity related to Quantity enclosed in the “$” symbol is transformed into BIO / IOB format and used as the true-label for training the model. The formatted data is used to train a model similar to the Quantity Extraction (SciBERT + CRF Model). The above model extracts the MeasuredEntity associated with the Quantity enclosed in “$”.
5. To extract MeasuredProperty we used a similar approach as used for MeasuredEntity. We enclosed the Quantity span in “$” symbol and the MeasuredEntity span in “#” symbol. The modified sentences are passed through the SciBERT tokenizer. The span of MeasuredProperty related to MeasuredEntity, Quantity pair is transformed into BIO / IOB format and used as the true-label for training the model. The formatted data is used to train a model similar to the Quantity Extraction (SciBERT + CRF Model). The model trained is used to extract MeasuredProperty linked with the MeasuredEntity, Quantity pair.
6. Once the predictions from all the models are available, we need to transform the predicted BIO/ IOB format into entity span format. We initially map each token’s span in the tokenized sentence and use it to determine the predicted entity’s span. While finding the span of the MeasuredEntity, MeasuredProperty, or Qualifier, if our model predicts multiple entities, then we predict the one which is closest to the

Quantity span. After that, we convert the sentence span of each entity extracted to the paragraph span.

7. The official metrics used by the SemEval organizer are F1-measure, F1-overlap, and Exact Match. Exact Match is a binary value of 0 or 1, while F1- measure is a token level overlap ratio of submission to true spans, where tokenization is done using simple white space delimiters. F1-overlap is a SQuAD (Rajpurkar et al., 2016) style Overlap score based on F1-measure, which penalizes the negative submissions more strictly. The final evaluation is based on a global F1-overlap score averaged across all subtasks.

EXPERIMENTS

We tried models like CRF, Bi-LSTM, stacked-LSTM but the models were not performing at par the SciBERT model either due to low score or a high degree of overfitting. We also tried to use Bi-LSTM on top of SciBERT but it overfitted. BERT-base, BERT-medium were also tried. BERT-large was avoided due to expensive computations. So we finally used the SciBERT + CRF model.

RESULTS

The quantity extraction task gave the following classification report on the validation set

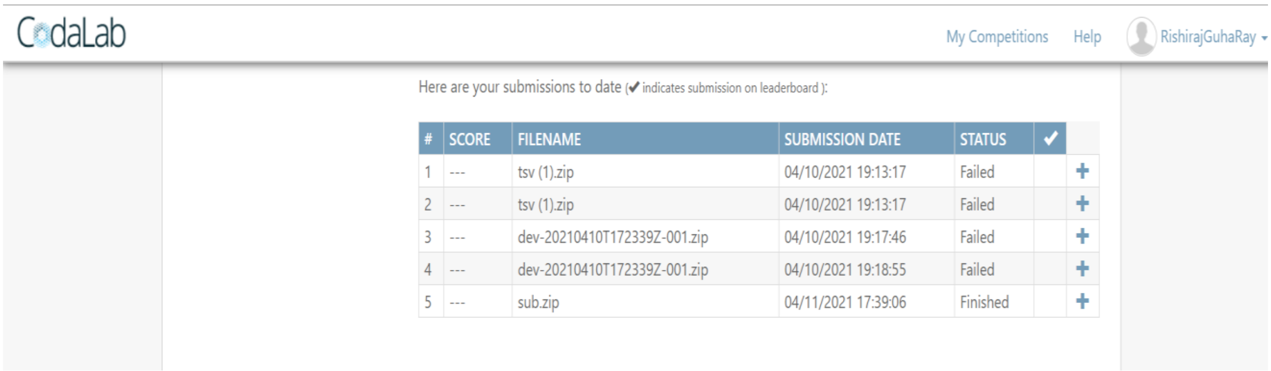
```
Entity recognition modified accuracy:-0.9017199017199017
--------NER RESULTS--------
Accuracy:-0.99875
Modified Accuracy:-0.9017199017199017
Precision:-0.9362244897959183
Recall:-0.9607329842931938
F1 score:-0.9483204134366926
```

The MeasuredEntity extraction task gave the following classification report in the validation set:-

```
Entity recognition modified accuracy:-0.35174418604651164
--------NER RESULTS--------
Accuracy:-0.9905315896739131
Modified Accuracy:-0.35174418604651164
Precision:-0.5707547169811321
Recall:-0.4782608695652174
F1 score:-0.5204301075268818
```

REFERENCES

1. Measurement Context Extraction from Text: Discovering Opportunities and Gaps in Earth Science: by Kyle Hundman and Chris A Mattmann
2. SciBERT: A Pretrained Language Model for Scientific Text: by Iz Beltagy Kyle Lo Arman Cohan Allen Institute for Artificial Intelligence, Seattle, WA, USA
3. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: by Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
4. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing: by Mark Neumann, Daniel King, Iz Beltagy, Waleed Ammar Allen Institute for Artificial Intelligence, Seattle, WA, USA
5. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data: by John Lafferty,Fernando Pereira and Andrew Mccallum



Here are your submissions to date (✓ indicates submission on leaderboard):

#	SCORE	FILENAME	SUBMISSION DATE	STATUS	✓	
1	---	tsv (1).zip	04/10/2021 19:13:17	Failed		+
2	---	tsv (1).zip	04/10/2021 19:13:17	Failed		+
3	---	dev-20210410T172339Z-001.zip	04/10/2021 19:17:46	Failed		+
4	---	dev-20210410T172339Z-001.zip	04/10/2021 19:18:55	Failed		+
5	---	sub.zip	04/11/2021 17:39:06	Finished		+