CS4248 G17 Final Project

Kan Yip Keng, Lin Mei An, Yang Zi Yun, Yew Kai Zhe

Introduction

- SemEval 2021 Task 11: NLPContributionGraph (NCG)
- Extract information from NLP scholarly articles
- Interconnect scholarly articles using the NCG

Problem Definition

Subtask 1 (Binary Classification)
Input: a set of sentences (NLP article)
Output: a set of sentences (all contributing sentences)

Subtask 2 (Named Entity Recognition)
Input: one sentence (contributing sentence)
Output: a set of phrases (scientific terms & predicates)

Dataset

- 237 NLP scholarly articles, 55201 total sentences
- 5096 contributing sentences, 51696 scientific phrases
- 1:10 contributing to non-contributing sentence ratio
- IAA score: 67.92% for subtask 1, 41.82% for subtask 2

Subtask 1

- Extraction of Contributing sentences
- Models used:
  - Naive-Bayes
  - SciBERT
  - Sentence--BERT
  - SciBERT + BiLSTM

Naive Bayes

- Uses TF-IDF word embeddings
- value increases proportionally to number of times a word appears
- F1 score: 0.326

SciBERT

- BERT based model trained on scientific texts
- inputs consists of sentence-label pair
  - label (1, 0) indicates a contributing sentence and non-contributing sentence
- correction of class imbalance with oversampling
- token limit = 100
- 2 epochs, lr 2e-5, batch size 32
- F1 score: 0.437

SciBERT + BiLSTM

- extension of SciBert
- Stack BiLTSM model on top of SciBERT
  - helps encode hidden semantics and long-distance dependencies
- dropout layer applied to avoid overfitting
- final layer outputs scores for the labels '1' and '0'
- 2 epochs, lr 2e-5, batch size 32
- F1 score: 0.422

Sentence-BERT

- Siamese BERT architecture
  - Contains 2 or more identical subnetworks
- Apply mean pooling separately for sentence and document embeddings
  - Learn a shared distance metric embedding space
- Not pretrained on domain specific dataset
- Experimented with all-mpnet-base-v2 and gsarti/scibert-nli
- 1 epoch, lr 1e-5, batch size 16
- F1 Score: 0.364
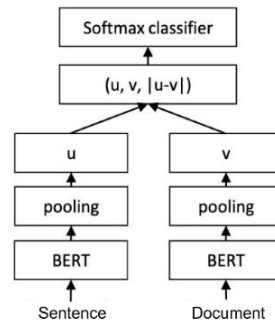
Subtask 2

Input: one sentence (contributing sentence)

- Eg. We used the BERT-BASE model pre-trained on English Wikipedia and BooksCorpus for 1M steps.

Output: a set of phrases (scientific terms & predicates)

- Eg. (used, BERT-BASE model, pre-trained on, English Wikipedia, BooksCorpus, for, 1M steps)

Subtask 2: BIO tagging scheme

B = start token of phrase
I = continuation tokens of phrase
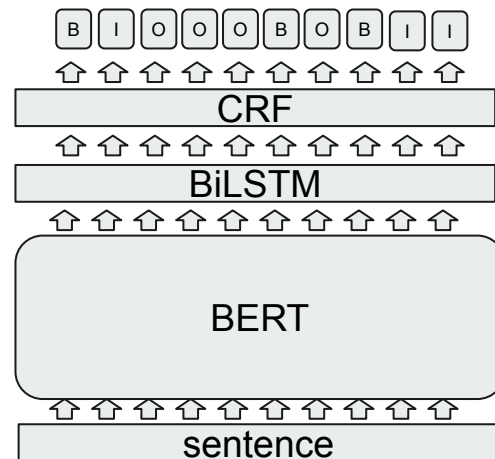O = non-phrase tokens

Sample input:

We used the BERT-BASE model pre-trained on English Wikipedia and BooksCorpus for 1M steps.
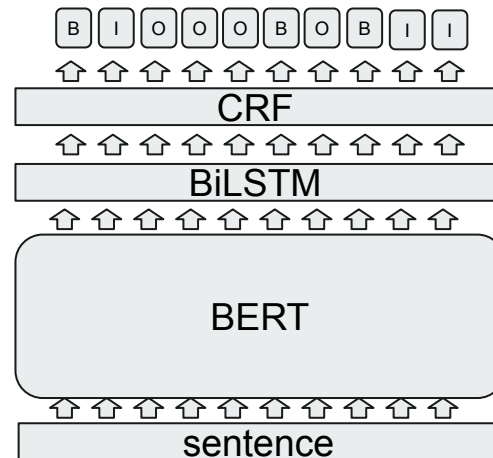
⇩

O B O B I B I B I O B B B I

Subtask 2: BERT + BiLSTM + CRF

- Sequence labelling task
- Similar to Named Entity Recognition (NER) task

B	I	O	O	O	B	O	B	I	I

CRF

BiLSTM

BERT

sentence

Subtask 2: Experiment

- BERT vs SciBERT
- BiLSTM vs without BiLSTM

Subtask 2: Results

- BERT vs **SciBERT**
- **BiLSTM** vs without BiLSTM

Approach	F1-Score
BERT + CRF	0.730
BERT + BiLSTM + CRF	0.734
SciBERT + CRF	0.748
SciBERT + BiLSTM + CRF	**0.751**