CS4248 G17 Project Proposal

Team members

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

Task background (Information Extraction)

The SemEval-2021 Shared Task NLP CONTRIBUTION GRAPH (a.k.a. 'the NCG task') tasks participants to develop automated systems that structure contributions from NLP scholarly articles in English.
Input: a set of research articles in plaintext format
Output: (1) a set of contributing sentences and
(2) a set of scientific knowledge terms and predicates from the contributing sentences

Approach

We will breakdown the NCG task into 2 subtasks:

(1) Extract a set of contributing sentences from a research article

The approach is similar to a classification task where every sentence in the article will be classified in 13 classes - 12 information units and not a contribution sentence. The model will follow an embedding - -hidden - hidden - output layer architecture where the headers and sentence position (where the sentence is found in the paper) are incorporated into the sentence representations (e.g. the header "related work" indicates that the sentence most likely discusses prior research). The best team used the BERT model to encode the sentence and its titles separately before concatenating this textual representation together with the positional feature. Since a primary challenge is dealing with a small, imbalanced dataset, we also aim to explore various strategies to deal with sparse data. These include data sampling and preprocessing strategies as well as training strategies (few-shot learning, adversarial training) to improve the generalization ability of the model.

(2) Extract a set of scientific knowledge terms and predicates from a set of contributing sentences

This task is a modified named entity recognition task that requires extracting scientific words and phrases belonging to one of twelve categories (Information Units). The best performing teams typically adopted one of the different variants of pre-trained Bidirectional Encoder Representations from Transformers (BERT) models, including BERT, RoBERTa, and sciBERT, combined with sequence modelling algorithms such as Hidden Markov Models or Conditional Random Fields. However, different parameter representations and optimisations made a comparison between models somewhat difficult. We aim to evaluate these models using the same system architecture as well as attempt to improve model F1 scores using an ensemble of these approaches.

Prior work

These are the 6 papers published by the participating teams of the NCG task which uses different approaches to tackle this shared task. Our team will be studying the merits and limitations of each approach, aiming to implement a better solution to resolve the common pitfalls of all prior work.

https://aclanthology.org/2021.semeval-1.59.pdf
https://arxiv.org/pdf/2105.05435.pdf
https://arxiv.org/pdf/2104.01619.pdf
https://aclanthology.org/2021.semeval-1.185.pdf
https://aclanthology.org/2021.semeval-1.61.pdf
https://aclanthology.org/2021.semeval-1.60.pdf

Dataset

We will be using the training dataset provided by the NCG task and evaluate using the CodaLab competition portal.

Breakdown of work

Kan Yip Keng: End-to-end pipeline, management, assisting in subtask 1&2
Lin Mei An, Yew Kai Zhe: Subtask (1) - Extract a set of contributing sentences
Yan Boshen, Yang Zi Yun: Subtask (2) - Extract a set of scientific knowledge terms and predicates

Timeline

Week 10 (18/10 - 24/10): Project consultation, research on prior work
Week 11 (25/10 - 31/10): Complete end-to-end pipeline, implement models
Week 12 (1/11 - 7/11): Evaluation 1, discussion & improvement
Week 13 (8/11 - 14/11): Evaluation 2, conclusion & documentation
Week 14 (15/11): Submission