Group#17: NLPContributionGraph

Siddharth Poonia[1], **Ankan Das**[2], **Aayush Mani**[3]
[1]16917687, [2]170120, [3]170007
[1]ECO, [2]EE, [3]EE
{spoonia, ankand, aayushm}@iitk.ac.in

Abstract

The NCG (NLPContributionGraph) task aims to provide structured contributions defined on NLP scholarly articles that can be integrable with the ORKG (Open Research Knowledge Graph). We plan to use advanced NLP/DL/ML techniques to solve the problem and contribute to the ORKG project. Techniques and algorithms from Transformers, BERT and BERT-KG are to be implemented on the dataset of NLP Scholarly papers and articles. Group 17 is participating in SemEval Task 11 to propose a solution to this problem. We have 288 papers as dataset to be used for training and testing. We are on our way to get first results out soon. The evaluations will start in January 2021. We are confident we will finish our work by December 2020.

1 Introduction

Knowledge Graph: A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge. In simple words, it's a programming solution to model a knowledge domain with the help of Machine Learning techniques.
The aim is to make searching for related papers and articles easily and fast. So that a researcher might spend less time going through articles, as that is a time consuming and cognitively demanding task. This knowledge graph solution aims to fasten up scientific research.

This task is being hosted for the first time in SemEval history. There hasn't been much done in SemEval regarding this problem. The ORKG framework was proposed in 2019 by Sören Auer et al (2019), and is hosted by TIB, the German National Library of Science and Technology.

A BERT based KG was proposed by Liang Yao et al(2019) to model triples from scholarly articles. Our approach is fairly straight-forward as of now. We plan to use sci-BERT as pretrained model for embeddings and further use BERT transformers to extract further information. Key points:

2 Problem Defintion

The OR-Knowledge-Graph has stated to provide structured contribution annotations as: (1) Contribution sentences: a set of sentences about the contribution in the article; (2) Scientific terms and relations: a set of scientific terms and relational cue phrases extracted from the contribution sentences; and (3) Triples: semantic statements that pair scientific terms with a relation, modeled toward subject-predicate-object RDF statements for KG building. The Triples are organized under three (mandatory) or more information units (viz., Research Problem, Approach, Model, Code, Dataset, Experimental Setup, Hyperparameters, Baselines, Results, Tasks, Experiments, and Ablation Analysis).
Thus the extraction task is defined in terms of these three dataset annotation elements where the extracting data element 3 relies on having extracted data element 2 which in turn depends on extracting 1.
This problem requires subject-verb-predicate analysis, fine-tuned for scientific vocabulary to solve it further.
The ORKG Framework, the motivation behind this task states to: (1) produce a semantic representation based on existing work, that can be well motivated as an annotation scheme for the application domain of NLP-ML scholarly articles (2) The annotated scholarly contributions based on NLPContributions should be integrable in the Open Research Knowledge Graph (ORKG) the

state-of-the-art content-based knowledge capturing platform of scholarly articles' contributions. (3) TheNLPContributions model should be useful to produce data for the development of machine learning models in the form of machine readers [18] of scholarly contributions.Such trained models can serve to automatically extract such structured information for downstream applications, either completely automated or semi-automated workflows recommenders. (4) TheNLPContributions model should be amenable to feed-back via a consensus approval or content annotation change suggestions from a large group of authors toward their schol-arly article contribution descriptions (an experiment that is beyond the scope of the present work and planned as following work)

3 Related Work

The approach by Liang Yao et al. (2019) takes entity and relation descriptions of a triple as input and computes scoring function of the triple with the KG-BERT language model. It represents entities and relations as their name/description textual sequences, and turn knowledge graph completion problem into a sequence classification problem. They intend to implement other pre-trained models like XLnet for further improvement.

4 Corpus/Data Description

Our dataset has been provided by the organisers of the task. It includes around 288 scholarly articles that are further classified into their research domains. The text parsing has been done by GROBID and further sentence tokenization has been performed by using Stanza library.

For the first task, the sentences from the [paper-name]-Stanza-out.txt has been stored in a dataframe and corresponding to each sentence, a value of either '1' or '0' has been assigned indicating whether this sentence is relevant or not, as given in sentences.txt.

5 Proposed Approach

WE use Sci-BERT for a vocabulary. We identify the contributory sentences based on the title and/or abstract. For that we create embeddings and use those embeddings for training our ML model (Logistic Regression as our first model). Here, the input is the CLS token, that can be considered as the sentence embedding using all the words present in that sentence. We have a binary class output of whether the sentence is relevant of not.

For the second task, we do NER(Named Entity Recognition) and for now, make a ruled based model. Then we do the same for the text and list down the matching sentences numbers for first subtask. Further work is going.

6 Experiments and Results

As of now, results have been produced on the test set from the text corpus we are working on. Hoping for something solid in the coming days.

7 Future Work

We plan to use BERT-SUM for summerization. In the first task, we find the embeddings using Sci-BERT. We will fine-tune the SciBERT model. Then fine-tune the BERT-SUM model so as to get most desirable ouput. We are planning to do little changes in these models as well. Also, we will be using the method given in the paper CUTS along with these models to get the maximum accuracy.

8 Conclusion

Baseline for our first task is complete and we have an accuracy of 0.912 as of now. Although, this basically reflects the underlying data distribution.

The work is a bit behind, but we are gaining pace. That's all.

References

1. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, Sören Auer
2. Relation Extraction with CNNs, Ji Young Lee et al.
3. Identifying and Labelling Keyphrases with Conditional Random Fields
4. Scientific Information Extraction with Semi-supervised Neural Tagging
5. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers