Tracking Dataset use across Conference Papers

Dataset mention extraction and clustering to construct a bipartite knowledge graph

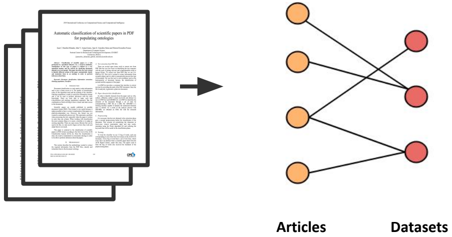
Pim Meerdink, Maarten Marx

Abstract

We build a system to extract the information necessary for the construction of a bipartite article-dataset graph. In this graph dataset node X and article node Y share an edge iff dataset X is used in article Y. We divide the task into two sub-tasks: dataset mention extraction and entity clustering. Dataset mention extraction entails identifying dataset referential phrases in scientific text. We train Googles sciBERT to extract dataset referential phrases on a dataset constructed for this project. SciBERT attains an F1 of 0.84 on our zero-shot test set. Entity clustering entails the clustering of our extracted dataset mentions so that dataset mentions referring to the same real-life dataset are assigned to the same cluster. We develop a task-specific graph-based algorithm that clusters based on lexical, semantic and document level features. The algorithm is able to attain a B-cubed f1 score of 0.86 on a self-constructed golden standard.

Introduction

- In the last few years, the scientific community has experienced considerable growth in the number of articles published.
- As such, it has become more difficult to find relevant articles quickly and efficiently.
- This projects aims to identify the datasets used in large corpora of scientific articles.
- This information will be used to construct a bipartite graph of dataset and article nodes
- Dataset X and article Y will have be connected through an edge if and only if dataset X was used in article Y.


Automatic classification of scientific papers in PDF for populating ontologies
Articles
Datasets

Approach

The task was divided into two sub-tasks:

Dataset mention extraction
Dataset mention extraction entails identifying the phrases in the text that refer to a dataset. This task is an example of Named Entity Recognition (NER). This task was performed using sciBERT.

Entity clustering
Entity clustering entails partitioning the identified dataset mentions so that partition contains all the dataset mentions corresponding to one real world dataset. This task is an example of cross-docume

nt coreference resolution. A task specific algorithm was developed for this subtask

Dataset Mention Extraction

- Allenai's sciBERT was used for the named entity recognition task, sciBERT is a BERT model pre-trained on scientific text [1].
- A dataset of sentences containing dataset mentions was constructed using 15,000 scientific articles taken from NIPS, SIGIR, VISION and SDM. The final dataset contained 6000 BIO-labeled sentences, 2864 of these sentences had a dataset mention.
- The model was evaluated on a zero-shot test set, this entails that all of the datasets in this set (e.g. CIFAR-10) are not in the training data.
- Extended SEMEVAL metrics were used for evaluation of the dataset mention extraction task [2]

	Exact			Begin		
	Prec.	Recall	F1	Prec.	Recall	F1
Train set	0.92	0.95	0.93	0.95	0.97	0.96
Eval set	0.90	0.86	0.88	0.96	0.90	0.93
Test set (Zero-shot)	0.82	0.88	0.84	0.88	0.93	0.90

Table 1: Precision, F1 and recall scores for the best performing model on the evaluation set. The scores for both the Exact and Beginning measure are reported

Entity Clustering

- For the entity clustering a task-specific algorithm was developed based loosely on [3]
- The choice to divert from established practice and implement a custom solution was made in large part due to the specific nature of the entities to be clustered (i.e. they all describe datasets).
- The distance based algorithm uses (dis)similarities in three different spaces
  - The **lexical space** is expressed in character level n-gram tf-idf vectors
  - The **semantic space** is expressed through pooling of sciBERTs last hidden layer [4]
  - The **document space** uses document embeddings from gensims doc2vec model [5]
- Linear interpolation of the distances in these spaces constitute the final distances used by our clustering algorithm
- The algorithm attained a b-cubed f1 of 0.86

Conclusion

This paper explores the subtasks of dataset mention extraction and entity clustering, working towards the development of a system that from some large corpuse of scientific articles, construct a bipartite knowledge graph

Works still needs to be performed before the developed system is ready to be deployed, in particular :

- While summations and ellipses are extracted properly, no specific steps are taken to parse and split them before the clustering steps, this is necessary for optimal performance
- Computational complexity remains an issue due to the distance based approach of the algorithm, the algorithm must be adapted to allow for larger amounts of data
- End to end evaluation must be performed to understand the systems actual, real life performance

References

1. Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In EMNLP , 2019.
2. Isabel Segura Bedmar. Extraction of drug-drug interactions from biomedical texts, 2013. URL https://www.cs.york.ac.uk/semeval-2013/task9/
3. Sourav Dutta and Gerhard Weikum. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment.TACL, 3:15–28, 12 2015. doi: 10.1162/tacl_a_00119.
4. Han Xiao. bert-as-service.https://github.com/hanxiao/bert-as-service
5. 2018.Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks , pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en