Tracking Dataset use across Conference Papers

Pim Meerdink[12]
Supervised by Maarten Marx[1,3]

[1] University of Amsterdam
[2] pim.meerdink@student.uva.nl
[3] maartenmarx@uva.nl

Keywords: Named Entity Recognition · Coreference Resolution · Conference Papers

The enormous growth in the amount of papers published that the scientific community has experienced demands preliminary information extraction from scientific articles; due to time constraints researchers cannot read and understand all published scientific articles within their domain. Constructing knowledge bases containing information corresponding to these published articles has become an important task in streamlining scientific research. We work towards an end to end system that, given some large corpus of scientific articles, builds a bipartite graph containing dataset nodes on one side, and articles on the other. Dataset X and article Y will have an edge between them if and only if dataset X was used in article Y. This knowledge graph can be applied to, for example, extending a scientific literature search engine with a feature that allows users to explore datasets. The full paper can be found at [3].

The task at hand can essentially be divided into two sub-tasks. First there is **dataset mention extraction**, this entails identifying the phrases in the text that refer to a dataset. This task is an example of Named Entity Recognition (NER). Second, **entity clustering**, this 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-document coreference resolution.

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. The network obtained a tight fit of the training data, with an 'exact' f1 of 0.93. This fit reflected well onto both the evaluation and test set, where the F1 scores are 0.88 and 0.84, respectively.

When observing the performance of our model it becomes apparent that the model has little added difficulty correctly classifying dataset mentions that occur in sentences with more than 4 positively labelled instances. This means that the network is also able to understand and interpret ellipses and summations, these more complex rules and structures are not harder for the network to identify than simple, one- or two-word dataset mentions. These structures and patterns are difficult even for human annotators to consistently parse and classify correctly, making the networks ability to understand the nuances of the labelling task significant.

For the entity clustering a task-specific algorithm was developed based loosely on [2]. 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). This aspect of the problem allowed for assumptions and steps that improve performance significantly. Examples are assumptions that can be made with respect to the lexical structure of the entities, in particular the important role that numbers play in denoting datasets (CIFAR10 vs CIFAR100). In short, the developed algorithm first normalises the entities within the input data, and performs intra-document clustering: entities within each document are clustered using lexical similarity. Afterwards, a linear interpolation of similarities in a lexical, semantic and document level space are used to construct a graph G where each node represents the groups of dataset mentions found within an article. Lexical similarity was expressed in character level n-gram tf-idf cosine similarity, semantic similarity using sciBERT sentence embeddings cosine similarity and the document similarity was expressed using gensims doc2vec model trained on our corpus of 15 000 scientific articles. The edges in G express similarity, and all edges below a certain value are dropped. Each component in G now corresponds to an equivalence class of intra document coreferring entities.

The algorithm attained a B-cubed F1 score of 0.86. When performing grid search of the linear interpolation parameters of the lexical, semantic and document similarities it was found that the algorithm relied heavily on the lexical distance, while also using document level information. The sciBERT sentence embeddings expressing semantic similarity did not add much to the models ability to correctly cluster dataset mentions, and the top performing set of parameters did not use them at all.

Several steps must be taken before the developed system is ready to be deployed and utilized in a practical setting. First, the entity clustering algorithm must be expanded to parse ellipses and summations separately, and split them into their separate elements. Further, the computational complexity of the entity clustering remains an issue, due to the distance based nature of the algorithm it complexity scales quadratically with the input size. Finally, end-to-end evaluation should be performed of the system. While the systems' performance for each of the two subtasks was thoroughly evaluated, the overall, end to end system was not evaluated properly. This is, of course, an essential step in the development and deployment of the system.

References

1. Beltagy, I., Cohan, A., Lo, K.: Scibert: Pretrained contextualized embeddings for scientific text. CoRR **abs/1903.10676** (2019), `http://arxiv.org/abs/1903.10676`

2. Dutta, S., Weikum, G.: Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. TACL **3**, 15–28 (12 2015). https://doi.org/10.1162/tacl_a_00119
3. Meerdink, P.: Tracking dataset use across conference papers (2020), https://scripties.uba.uva.nl/search?id=715259