Preface to the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at JCDL 2020

Chengzhi Zhang[1], Philipp Mayr[2], Wei Lu[3], Yi Zhang[4]

1. Nanjing University of Science and Technology, Nanjing, China, zhangcz@njust.edu.cn
2. GESIS-Leibniz-Institute for the Social Sciences, Cologne, Germany, philipp.mayr@gesis.org
3. Wuhan University, Wuhan, China, weilu@whu.edu.cn
4. University of Technology Sydney, Sydney, Australia, Yi.Zhang@uts.edu.au

1. Introduction

The 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2020) was launched at the ACM/IEEE Joint Conference on Digital Libraries (JCDL) on August 1, 2020. The goal of this workshop is to engage the related communities in open problems in the *extraction and evaluation of knowledge entities* from scientific documents. Participants are encouraged to identify knowledge entities, explore feature of various entities, analyze the relationship between entities, and construct the extraction platform or knowledge base. Results of this workshop are expected to provide scholars, especially early career researchers, with knowledge recommendations and other knowledge entity-based services [1].

2. Overview of the papers

This year, 14 papers (including 3 long papers, 6 short papers, 4 posters and 1 demo) were accepted for presentation and inclusion in the proceedings. In addition, the workshop featured two keynote talks in the different EEKE-related fields. All workshop contributions are documented in the workshop website[1]. The following section briefly lists the various contributions.

2.1 Keynotes

Two keynotes were presented in EEKE2020.

The first one was given by Ming Song: ***Entitymetrics 2.0: Measuring the Impact of Entities and Relations Extracted from Scientific Documents.*** The concept of entitymetrics was first introduced in 2013, entitymetrics [2] has been applied to measure the impact of entities as well as to gauge the knowledge usage and transfer anchored on entities for knowledge discovery. In this talk, the previous studies employing entitiymetrics are summarized and the limitations of the current approaches are discussed. In addition, the future directions of entitymetrics are suggested.

[1] https://eeke2020.github.io/

The second keynote was given by Markus Stocker: ***Building Scholarly Knowledge Bases with Crowdsourcing and Text Mining.*** Building on the Open Research Knowledge Graph (http://orkg.org) as a concrete research infrastructure, in this talk Dr. Stocker presented how using crowdsourcing and text mining humans and machines can collaboratively build scholarly knowledge bases. He discussed some key challenges that human and technical infrastructures face as well as the possibilities scholarly knowledge bases enable.

2.2 Research papers, Posters and Demo

The following papers were presented in 5 sessions.

Session 1: Knowledge Entity Extraction and Application

-Jennifer D'Souza and Sören Auer

NLPContributions: An Annotation Scheme for Machine Reading of Scholarly Contributions in Natural Language Processing Literature

This paper describes an annotation initiative to capture the scholarly contributions in natural language processing (NLP) articles, particularly, for the articles that discuss machine learning (ML) approaches for various information extraction tasks. They attempted to find a systematic set of patterns of subject-predicate-object statements for the semantic structuring of scholarly contributions, and to apply the discovered patterns in the creation of a larger annotated dataset for training machine readers of research contributions.

-Mengjia Wu and Yi Zhang

Intelligent Bibliometrics for Discovering the Associations between Genes and Diseases: Methodology and Case study

This paper proposes an adaptable and transferable methodology to extract biomedical entities including diseases, chemicals, genes and genetic variations from literature data. A heterogeneous co-occurrence network is constructed and a semantic adjacency matrix is generated to identify key genes and genetic variants, and capture the emerging disease-gene associations via a link prediction approach.

Session 2: Entity Extraction from Scientific Documents

-Liangping Ding, Zhixiong Zhang, Huan Liu, Jie Li and Gaihong Yu

Automatic Keyphrase Extraction from Scientific Chinese Medical Abstracts Based on Character-Level Sequence Labeling

In this paper, authors regard automatic keyphrase extraction from Chinese text as a character-level sequence labeling task. Unsupervised keyphrase extraction methods including term frequency (TF), TF-IDF, TextRank and supervised machine learning methods including Conditional Random Field (CRF), Bi-directional Long Short Term Memory Network (BiLSTM) and BiLSTM-CRF are used to extract keyphrases from academic papers in medical domain. The character-level sequence labeling model based on BERT obtains the best result.

-Jin Mao, Shiyun Wang and Xianli Shang

Investigating interdisciplinary knowledge flow through citances

This study attempts to investigate the content of knowledge flow towards an interdisciplinary field by analyzing the citation sentences (i.e., citances) in the articles of eHealth field. The associated knowledge phrases between citances and the references are identified and categorized to analyze the content and categories of knowledge spread from the source disciplines to the field. In general, this study contributes to the understanding of content characteristics about interdisciplinary knowledge integration.

-Yu Li, Tao Yue (Speaker) and Wu Zhenxin

IEKM-MD: An Intelligent Platform for Information Extraction and Knowledge Mining in Multi-Domains

This paper constructs a platform for information extraction and knowledge mining, namely IEKMMD. Two innovative technologies are proposed: Firstly, a phrase-level scientific entity extraction model combining neural network and active learning is designed to reduce the model's dependence on large-scale corpus. Secondly, a translation-based relation prediction model is provided, which improves the relation embedding by optimizing loss function. In addition, the platform integrates the advanced entity recognition model and the keyword extraction mode, and provides abundant services for fine-grained and multi-dimensional knowledge.

-Liang Chen, Shuo Xu, Weijiao Shang, Zheng Wang, Chao Wei and Haiyun Xu

What is Special about Patent Information Extraction?

This article aims at exploring the particularity in patent information extraction, thus to point out the direction for further research. To be more specific, they discuss: (1) what is the special about labeled patent dataset? (2) What is special about word embedding in patent information extraction? (3) What kind of method is more suitable for patent information extraction?

Sesson 3: Interactive demos

-Zi Xiong, Yue Qi, Wei Lu and Qikai Cheng

Design and Implementation of an Academic Search System Based on a General Query Language and Automatic Question Answering

This research designs and implements an academic search system with two major innovations: 1) proposing a general query language SSL to describe the academic search intention in a unified and standardized way, 2) proposing a user intention recognition method to help improve traditional automatic question-answering systems. The SSL language and intention recognition method is applied to a QA-oriented academic system which is innovative compared with traditional query-based systems.

Session 4: Entity Relation Extraction and Application

-Xin An, Jinghong Li, Shuo Xu, Liang Chen and Sainan Pi

A Novel Approach for Patent Similarity Measurement Based on Sequence Alignment

In order to measure the similarity among different patents, this study proposes a novel approach on the basis of sequence alignment. The method takes semantic direction of each sequence structure and the word order information of each component into consideration; an algorithm for calculating the global importance of each sequence structure is put forward. Extensive experimental results show that the proposed approach is significantly more accurate and is not sensitive to several core parameters.

-Fang Tan, Siting Yang, Xiaoyan Wu and Jian Xu

Exploring the Relation between Biomedical Entities and Government Funding

In order to analyze the effect of government funding on the promotion of scientific research, and to help the government manage research funds more rationally, this study proposes a framework for analyzing the relationship between entities in the field of medicine and funds. The results reveal that the field of genetic research is in a period of rapid development and disease research catch NIH's continuous attention. However, the stimulating effect of government funding on the research popularity is decreasing.

-Sahand Vahidnia, Alireza Abbasi and Hussein A. Abbass

Document Clustering and Labeling for Research Trend Extraction and Evolution Mapping

In this study, a method is proposed to extract research trends and their temporal evolution, throughout discrete time periods. Adapting contextualized word embedding techniques, the method utilizes published academic documents as knowledge units and clusters them into groups. Various labeling techniques are explored to evaluate the quality of clusters and explore their explain ability. The results show that utilization of neural embedding in conjunction with paragraph-term weights would provide simple and reliable paragraph embedding that can be used for clustering of the textual data.

Session 5: Poster/ Greeting Notes of EEKE2020

-Qikai Liu, Pengcheng Li, Wei Lu and Qikai Cheng

Long-tail dataset entity recognition based on Data Augmentation

Datasets play an important role in data-driven scientific research. It is important to recognize dataset entities correctly, especially when it comes to unusual long-tail dataset entities. However, it is very difficult to obtain high quality training corpus in named entity recognition. This paper obtained the data based on a distant supervision method along with two data augmentation methods. A BERT-BiLSTM-CRF model is used to predict long-tail dataset entity.

-Xiaole Li, Yuzhuo Wang

Assessing Impact of Method Entities in a Special Task

Methods play an important role in the research. Identifying and analyzing entities about research methods can help scholars understand methods used in their field and accelerate the efficiency of scientific research. This paper takes named entity recognition (NER) as an

example and evaluates the impact of method entities in this domain. They found that conditional random field (CRF) is the most influential algorithms in NER. Deep learning algorithms have developed rapidly in the past 5 years. F-measure, precision and recall are the most widely used indices and measurements. Scholars do not pay enough attention to use tools and they prefer to use classic datasets.

-Chong Chen, Jingying Zhang, Xiaoyu Chu and Jinglin Zheng

Study on the Difference between Summary Peer Reviews and Abstracts of Scientific Papers

This article proposes primary measurement to compare Summary peer reviews with abstracts from readability and semantic function types. The results show that summary peer reviews highlight some distinct function types, and the terminology in peer reviews is not as dense as in abstracts. Summary peer reviews can be complement to abstracts in literature searches, and can help readers understanding papers more thoroughly.

-Wei Shao, Hua Bolin, Qiang Ma, Jiaying Liu, Hongwei He, Keqi Chen

An Unsupervised Method for Terminology Extraction from Scientific Text

Finding new terminology is a kind of named entity recognition (NER) problem. However, many high performance methods need labeled data. This paper proposes an unsupervised method based on sentence pattern and part of speech. They initialize a few patterns to extract terminologies in certain sentences, and then try to find the same POS sequences in sentences not matched by initial patterns with obtained terminologies' POS sequences. The new patterns and more terminologies are obtained after several iterations.

3. Outlook and further reading

Currently the EEKE2020 organizers edit the following two Special issues:

-Special Issue on "Extraction and Evaluation of Knowledge Entities from Scientific Documents" in Journal of Data and Information Science (https://mc03.manuscriptcentral.com/jdis).

-Special Issue on "Scientific Documents Mining and Applications" in Data and Information Management (https://www.editorialmanager.com/dim/default.aspx).

References

[1] Chengzhi Zhang, Philipp Mayr, Wei Lu, Yi Zhang. (2020). Extraction and Evaluation of Knowledge Entities from Scientific Documents: EEKE2020. In: Proceedings of the 20th ACM/IEEE Joint Conference on Digital Libraries (JCDL2020), Wuhan, China, 2020. https://doi.org/10.1145/3383583.3398504

[2] Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, Tamy Chambers. Entitymetrics: measuring the impact of entities. Plos One, 8(8), e71416.