Preface to the 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at JCDL 2021

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 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2021) was co-located with the ACM/IEEE Joint Conference on Digital Libraries (JCDL) on September 30, 2021. 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].

2. Overview of the papers

This year 15 papers (including 3 long papers, 6 short papers, 6 posters) were accepted for presentation and 14 paper were included 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 at EEKE2021.

The first one was given by Heiko Paulheim: ***From Wikis to Knowledge Graphs: Approaches and Challenges beyond DBpedia and YAGO.***

Wikipedia was among the first sources to be identified for automatic knowledge graph construction. DBpedia and YAGO, two of the most widely used public knowledge graphs, perform knowledge extraction from Wikipedia by following the "one entity per Wiki page" paradigm. Thus, the resulting graphs are naturally limited by the coverage of Wikipedia, and they inherit many biases from it. In my talk, I will introduce recent alternatives for creating knowledge graphs from Wikis, in particular DBkWik and CaLiGraph, which use different approaches for identifying entities, and I will point out several challenges that exist off the beaten path of the "one entity per Wiki page" approaches.

The second keynote was given by Gong Cheng: ***Entity Summarization: Where We Are and What Lies Ahead?***

Semantic data such as knowledge graphs, describing entities with property values, are increasingly available on the Web. A large number of property values describing an entity may overload users with excessive amounts of information. One solution is to generate a summary (e.g., a small subset of key property values) for entity descriptions to satisfy users' information needs efficiently and effectively. This research topic, termed Entity Summarization, has received considerable attention in the past decade. In this talk, I will review existing

[1] https://eeke-workshop.github.io/2021/

methods and evaluation efforts on entity summarization. I will categorize existing methods by presenting a hierarchy of technical features that have been incorporated, including generic, domain-specific, and task-specific features. I will show various frameworks for combining multiple features to assemble a full entity summarizer, including graph-based models, grouping, re-ranking, and combinatorial optimization. I will particularly highlight some pioneering deep learning based methods. Finally, I will discuss limitations of existing methods and, based on that, I will suggest several directions for future research.

2.2 Research papers and posters

The following papers were presented in 4 sessions.

Session 1: Entity Extraction and Application

-Anastasia Zhukova, Felix Hamborg and Bela Gipp

ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts

This paper proposes ANEA, an automated (named) entity annotator to assist human annotators in creating domain-specific NER corpora for German text collections when given a set of domain-specific texts. In our evaluation, this paper finds that ANEA automatically identifies terms that best represent the texts' content, identifies groups of coherent terms, and extracts and assigns descriptive labels to these groups, i.e., annotates text datasets into the domain (named) entities.

-Santosh Tokala Yaswanth Sri Sai, Prantika Chakraborty, Sudakshina Dutta, Debarshi Kumar Sanyal and Partha Pratim Das

Joint Entity and Relation Extraction from Scientific Documents: Role of Linguistic Information and Entity Types

This paper aims to automatically extract entities and relations from a scientific abstract using a deep neural model. Given an input sentence, the authors use a pretrained transformer to produce contextual embeddings of the tokens which are then enriched with embeddings of their part-of speech (POS) tags. A sequence of enriched token representations forms a span, and entities and relations are jointly learned over spans. Entity logits predicted by the entity classifier are used as features in the relation classifier. The proposed model improves upon competitive baselines in the literature for entity and relation extraction on SciERC and ADE datasets.

-Masaya Tsunokake and Shigeki Matsubara

Classification of URLs Citing Research Artifacts in Scholarly Documents based on Distributed Representations

This paper describes methods for classifying URLs referring to research artifacts in scholarly papers, and examines their classification performance. The methods discriminate whether a URL refers to a research artifact or not and classify the identified URL into "tool" or "data." The methods use distributed representations obtained from citation contexts of the URL. Each component of a URL can be regarded as a word, and the meaning of the entire URL can be generated by synthesizing the distributed representation of each component using compositional functions. Experiments with using URLs in international conference papers showed the effectiveness of our proposed compositional functions.

Session 2: Keyword Exaction and Application

-Liangping Ding, Zhixiong Zhang, Huan Liu and Yang Zhao

Design and Implementation of Keyphrase Extraction Engine for Chinese Scientific Literature

This paper constructs a keyphrase extraction engine for Chinese scientific literature to assist researchers in improving the efficiency of scientific research. There are four key technical problems in the process of building the engine: how to select a keyphrase extraction algorithm, how to build a large-scale training set to achieve application-level performance, how to adjust and optimize the model to achieve better application results, and how to be conveniently invoked by researchers. Aiming at the above problems, we propose corresponding solutions. The engine is able to automatically recommend four to five keyphrases for the Chinese scientific abstracts given by the user, and the response speed is generally within 3 seconds. The keyphrase extraction engine for Chinese scientific literature is developed based on advanced deep learning algorithms, large-scale training set, and high-performance computing capacity, which might be an effective tool for researchers and publishers to quickly capture the key stating points of scientific text.

- Aofei Chang, Bolin Hua and Dahai Yu

Keyword Extraction and Technology Entity Extraction for Disruptive Technology Policy Texts

This article first crawls the texts of disruptive technologies from the science and technology policy websites of major countries. Then, the text is segmented by Spacy, the segment result is filtered by a word list to construct an applicable TF*IDF matrix, and finally the matrix weights are optimized with manually collected domain core words and important words. After these, extraction and statistics of

technical entity are performed according to a specified word list. Through comprehensive analysis, it can be found that the keyword hotspots of the experimental texts are focused on artificial intelligence, information security, new energy, etc. The key areas of specific disruptive technologies are artificial intelligence, air and space, and new generation communication technologies. The result reflects the current situation and policy focus of disruptive technology development in these countries.

-Jiabin Peng, Jing Chen and Guo Chen

Extracting Domain Entities from Scientific Papers Leveraging Author Keywords

This paper proposes a two-stage methodology that can make good use of existed author keywords of the given domain to solve this problem. Firstly, the author keyword set was used to mark the boundary of candidate entities, and then their features are integrated to classify their entity type. In the experiment on artificial intelligence (AI) documents from WOS, our approach obtains an F1 value of 0.753 without manual annotation, which is slightly lower than the BERT-BiLSTM-CRF baseline model (F_1=0.772) trained on manual annotation corpus, showing the usability of our approach in practice.

Sesson 3: Knowledge Graph and Application

-Johannes Stegmüller, Fabian Bauer-Marquart, Norman Meuschke, Terry Ruas, Moritz Schubotz and Bela Gipp

Detecting Cross-Language Plagiarism using Open Knowledge Graphs

This paper introduces the new multilingual retrieval model Cross-Language Ontology-Based Similarity Analysis (CL-OSA) for this task. CL-OSA represents documents as entity vectors obtained from the open knowledge graph Wikidata. Opposed to other methods, CL-OSA does not require computationally expensive machine translation, nor pre-training using comparable or parallel corpora. It reliably disambiguates homonyms and scales to allow its application to Web-scale document collections. We show that CL-OSA outperforms state-of-the-art methods for retrieving candidate documents from five large, topically diverse test corpora that include distant language pairs like Japanese-English. For identifying cross-language plagiarism at the character level, CL-OSA primarily improves the detection of sense-for-sense translations. For these challenging cases, CL-OSA's performance in terms of the well-established PlagDet score exceeds that of the best competitor by more than factor two.

-Yongmei Bai, Huage Sun and Jian Du

A PICO-based Knowledge Graph for Representing Clinical Evidence

In this paper, the clinical trial information is extracted from the semi-structured records on ClinicalTrials.gov to construct a PICO-based knowledge graph for representing clinical evidence. The knowledge graph is expected to give a whole picture on the research protocol and reported results of clinical trials. It can be quickly searched, visualized, and exported in batches and on-demand. The authors collect 6279 registered clinical trials on COVID-19 in ClinicalTrials.gov, among which 71 trials had reported results. Information extraction and term standardization are carried out in a semi-automated manner. The knowledge graph is constructed using neo4j.

Session 4: Poster

-Shiyun Wang, Jin Mao and Yaxue Ma

The correlation between content novelty and scientific impact

This paper proposes two indicators to measure the content novelty of a paper based on the knowledge entities it contains, and explored the relationship between content novelty and scienti0ic impact of papers. It is found that content novelty is negatively correlated with citation impact in the dataset. The findings of this paper suggest that science policy in favor of citation count based impact may be biased against novel research.

-Tohida Rehman, Debarshi Kumar Sanyal, Samiran Chattopadhyay, Plaban Kumar Bhowmick and Partha Pratim Das

Automatic Generation of Research Highlights from Scientific Abstracts

The huge growth in scientific publications makes it difficult for researchers to keep track of new research even in narrow sub-fields. While an abstract is a traditional way to present a high level view of the paper, recently it is getting supplemented with research highlights that explicitly identify the important findings in the paper. In this poster, the authors aim to automatically construct research highlights given the abstract of a paper. We use deep neural network-based models for this purpose and achieve high ROUGE and METEOR scores on a large corpus of computer science papers.

-Fang Tan, Tongyang Zhang and Jian Xu

Differential Analysis on Performance of Scientific Research Teams based on Analysis of the Popularity Evolution of Entities

In order to investigate the impact of research topic selection time on output performance of scientific collaborations, the aim of this paper is to develop a differential analysis framework of scientific collaboration performance at different stages of entity popularity. The framework consists of three main sections: (1) data acquisition and processing; (2) stage division of entity popularity; (3) differential analysis on performance of scientific collaborations at different stages of entities popularity. Our findings show that the popularity stage that research topics are going through can play a role in the collaboration output performance.

-Litao Lin, Dongbo Wang and Si Shen

Research on extraction of thesis research conclusion sentences in academic literature

The extraction of sentences with specific meaning in academic literature is an important work in academic full-text bibliometrics. This paper attempts to establish a practical model of extracting conclusion sentences from academic literature. In this research, SVM and SciBERT models were trained and tested using academic papers published in JASIST from 2017 to 2020. The experimental results show that SciBERT is more suitable for extracting thesis conclusion sentences and th***e*** optimal F1-value is 77.51%.

-Jinzhu Zhang and Linqi Jiang

Topic Evolution Path and Semantic Relationship Discovery Based on Patent Entity Relationship

This paper uses representation learning method to get the semantic representation of each entity/word, and computes the semantic similarity among them to find out pairs of words which are different but with the same meaning in a special context. Moreover, we define multiple semantic relationships among topics, and design a method to use patent entity relationships to obtain the semantic relationships among topics. Experiments in the technical field of UAV transportation have confirmed that the method in this paper can effectively identify the evolutionary relationship between topics and the semantic relationship between topic, Make the evolutionary relationship between topics more abundant and Interpretable. And provide a reference for further enriching and improving the topic evolution analysis method.

-Wang Zheng and Xu Shuo

Bureau for Rapid Annotation Tool: Collaboration can do more over Variety-oriented Annotations

This paper develops a novel workbench such that collaboration can do more over variety-oriented annotation. The workbench is named as Bureau for Rapid Annotation Tool (Brat for short). Main functionalities include enhanced semantic constraint system, Vim-like shortcut keys, annotation filter and graph-visualizing annotation browser. Until now, over 500,000 mentions have been annotated with our Brat workbench.

3. Outlook and further reading

Currently the EEKE2021 organizers edit the following Special issues:

Special Issue on "Extracting and Evaluating of Knowledge Entities" in Aslib Journal of Information Management (https://www.emeraldgrouppublishing.com/journal/ajim).

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] Chengzhi Zhang, Philipp Mayr, Wei Lu, Yi Zhang. (2020). Preface to the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at JCDL 2020. In: Proceedings of the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents co-located with the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL 2020), Wuhan, China, 2020.