Scientific Research Publishing

Open Journal of Social Sciences, 2020, 8, 315-325
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

A Framework Using Active Learning to Rapidly Perform Named Entity Extraction and Relation Recognition for Science and Technology Knowledge Graph

Ying Wang[1], Jing Dong[2*], Peng Ren[2*], Ye Wang[1], Jingjing Cao[1]

[1]Center for Science & Technology Talents, MoST, Beijing, China
[2]BNRist, DCST, RIIT, Tsinghua University, Beijing, China
Email: wangy@sttc.net.cn, *infantainmaple@gmail.com, *launcher.ix5@gmail.com

How to cite this paper: Wang, Y., Dong, J., Ren, P., Wang, Y., & Cao, J. J. (2020). A Framework Using Active Learning to Rapidly Perform Named Entity Extraction and Relation Recognition for Science and Technology Knowledge Graph. *Open Journal of Social Sciences, 8,* 315-325.
https://doi.org/10.4236/jss.2020.89025

Received: August 31, 2020
Accepted: September 24, 2020
Published: September 27, 2020

Abstract

Construct a knowledge graph is time-consuming and the knowledge graph in the scientific domain requires extremely high labor costs due to it requires high prior knowledge to extract knowledge from resources. To build a scientific research knowledge graph, the most of input are papers, patent, the description of their project and some national program (such as National High Technology Research and Development Program of China, Major State Basic Research Development Program of China, General Program, Key Program and Major Program) which all of them are unstructured data, that make human participation are mostly necessary to measure the quality. In this paper, we design and proposed a framework using active learning; this framework can be used to extract entity and relation from unstructured science and technology research data. This framework combines the human and machine learning approach together, which is active learning, to help user extract entity from those unstructured data with less time cost. By using those data to construct a CKG as annotation label, it further implements active learning tools and helps the expert to rapidly annotate the data with high accuracy. Those knowledge graph constructed by this framework can be used to finding similar research area, finding similar researchers, finding popular research areas and so on.

Keywords

Knowledge Graph, Human-in-the-Loop, Framework, Science and Technology

1. Introduction

In the scientific domain, knowledge graph can be used in many ways. For example, it can use to recognize deviant researchers who do not have enough research contribution. It can also be used as a tool to cluster similar researchers and help the organization to manage them better. Moreover, it can find popular research areas. Knowledge graph collects a massive amount of interrelated facts that connect different concepts and instances, and can be transformed into practical knowledge (Pujara, Miao, Getoor, & Cohen, 2013). These linked data triples can be queried by users (Verborgh, Vander Sande, & Hartig et al., 2016), Researchers have paid a great effort into the realms of constructing knowledge graph. There are already several developed science and technology research knowledge graphs available, such as Aminer.

In knowledge graph, RDF triples are stored to represent the knowledge, and there are three types of information stored as nodes: entity, event and concept. Some knowledge graphs only contain concept nodes, and are generally called Ontology. We redefine it as Concept Knowledge Graph (CKG). Correspondingly, we define the knowledge graph with entity nodes and event nodes as Instance Knowledge Graph (IKG). Based on that, we describe the knowledge graph, including both CKG and IKG as Factual Knowledge Graph (FKG) (Sheng, Shao, Zhang, Li, Xing, Zhang, Wang, & Gao, 2019; Sheng, Wang, Zhang, Li, Li, Xing, Li, Shao, & Zhang, 2019). In the scientific research domain, IKG contains instance data such as the title of the papers, content of research projects and so on. But these data sources for construction are generally unstructured data, in which the knowledge needs to be extracted manually with dramatic labor cost.

To reduce the labor cost, automatic named entity recognition and relation extraction are adopted. Machine learning method can be used to extract those unstructured data automatically. But the mechanical method to do so still requires preprocessed data and a lot of time in model training (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016). Besides, without the experts to provide useful prior knowledge and measure the process, the quality of the automatic result is relatively unreliable (Giorgi, Bader, & Wren, 2020).

To solve the labor cost problem, we implement active learning to reduce human participation workloads during the scientific unstructured data annotation process, and it's further combined with "expert-in-the-loop" methodology to maintain the quality of entity annotation and relation extraction result.

This paper is organized as follows. In Section 2, we introduce the related work in the relevant field. In Section 3, we present the detailed framework and workflow for the framework. In Section 4, we show the details of the modules used in this framework, in the end, we summarize the paper and propose future work in Section 5.

2. Related Work

There are already a lot of works that have been done on the field of extract useful

information in the science and technology domain from unstructured data., Wan et al. (Wan, Zhang, Zhang, & Tang, 2019) mined the social relations between academic researchers through unstructured data. Yuan et al. (Yuan, Shao, Liang, Tang, Hall, Liu, & Zhang, 2020) specifically extract the same information and relation in Artificial Intelligence domain. Some previous research not only focus on the instance information such as social relations between humans, but also focus on extraction of the concept in the related domain, such as Jibing's previous research (Gong, Wang, Wang, Feng, Peng, Tang, & Yu, 2020). They use novel method to extract useful information in the domain of science and technology.

We compared several unstructured data annotation frameworks currently used on entity recognition and relation extraction, to show what has been achieved in the related field, and briefly discuss what can be improved by our research. Those related works are Doccano, BRAT, Prodigy, YEDDA (Yang, Zhang, Li, & Li, 2018), DeepDive: Mindtagger, Anafora (Chen & Styler, 2013), WebAnno (Eckart de Castilho, Mújdricza-Maydt, Yimam, Hartmann, Gurevych, Frank, and Biemann, 2016), MAE and INCEpTION (Klie, Bugert, Boullosa, Eckart de Castilho, and Gurevych, 2018). Those frameworks that are discussed in this section are chosen based on their popularity in practice.

We explore those named entity recognition and relation extraction framework. They are compared to human participation method and labor cost level. The result shows only a few frameworks combine both machine and human effort to accelerate the annotation process with a reliable result. Among all the frameworks we explored, only the WebAnno provide full auto annotation but it is only available for project manager and administrators. Most of the named entity recognition and relation extraction frameworks are purely manual.

To reduce the labor cost level, our framework implements active learning method, which makes the extraction and recognition process become semi-auto at the beginning of an annotation task. With the model trained by active learning getting more and more accurate, the labor cost level of our framework will get lower through the annotation process.

3. Framework and Workflow

In this section, we introduce the framework and the general workflow. There are two parts in this framework: an interface used for extract meta concept knowledge graph for annotation standard; an active learning toolset implements active learning method and interactive with annotator, used for reducing the labor cost of annotation.

3.1. The Framework for Extract Science and Technology Research Data

The framework contains following parts:

1)The data source of provides unstructured data to be annotated; In this part, experts also need to manually annotated from high quality teaching material,

paper abstract, and summarization papers for the basic concept meta data extraction. Construct a concept knowledge graph for use as annotation standard.

2) The active learning toolset.

3) The output of this framework is the high quality annotated scientific research material and can be further used to construct high-quality IKG.

How to construct the meta concept knowledge graph and how the active learning module works to reduce the labor cost is explained in detail in section 4.

3.2. The Workflow

As shown in Figure 1, in the workflow of this framework, the unstructured data such as papers and project descriptions are taken as input into the active learning toolset, experts who are assigned as annotators are asked to annotate the abstract and summarization paper first, those step will generate standardized concept, which will be further used as label while the active learning loop takes part in. Those meta concepts will also be directly sent into machine learning loop to train the initial model. After initializing a learning model and start the loop of active learning, the algorithm will perform auto-extraction on the unstructured data, periodically returns the unconfident auto annotation result to the annotator, asks them to correct. The data predicted through active learning model with high confident will be combined with the correction result generated by human, and further alignment into a Science and Technology research IKG with decent accurate.

Through this process, with the machine learning model keeping convergence, it becomes more and more accurate while predicting the extraction and recognition result. Meanwhile, the framework gets less requirement for a human to participate in correction annotation. The measurement module is supervised during the entire process. With the active learning model evolving and convergence, the

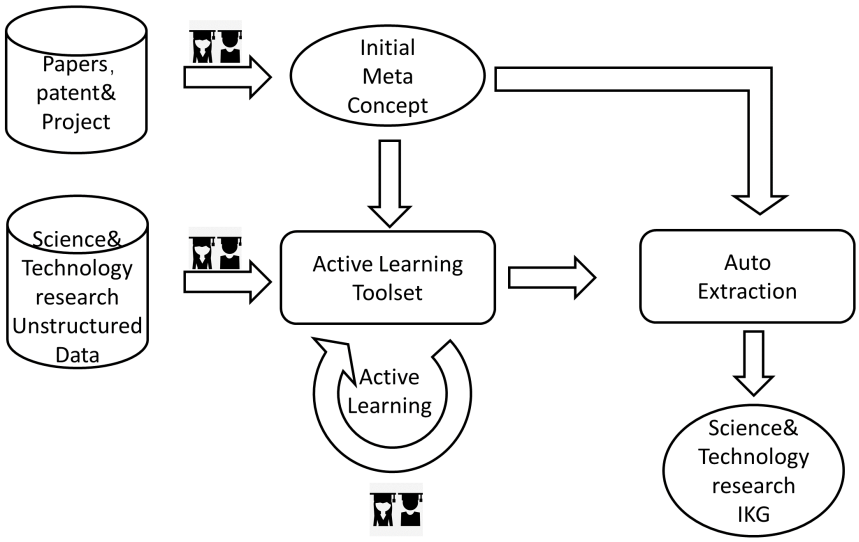

Figure 1. The detailed workflow to annotate science and technology research unstructured data.

labor cost of named entity recognition and relation extraction assignment continuously decreases. The fine trained model can be further implemented to automatically extract information from academic papers, patent and research projects, generate high-quality IKG with low labor cost.

4. Toolsets and Modules

To reduce the labor cost for name entity recognition and relation extraction, an active learning toolset is involved to help user perform annotation quickly and accurately. In the framework, we first extract basic concept to generated standard for annotator to use before assigning active learning loops. Then we use active learning to quick extract entities from scientific research materials. In this section, we explain how those two functions combine together and describe in detail about the active learning process.

4.1. Human-in-the-Loop Active Learning Toolset

Algorithms that involve humans' communication can be defined as "human-in-the-loop" (Holzinger, 2016). Human-in-the-loop has actually been applied to many aspects of artificial intelligence like named entity recognition (Coelho da Silva & Magalhães et al., 2019) and rules learning (Yang, Kandogan, Li, Sen, & Lasecki, 2019) to improve the performance. Active learning is a machine learning method that involves the human-in-the-loop methodology.

In this framework, an active learning toolset using deep active learning method has been developed to reduce the labor cost.

We use other's work (Shen, Yun, Lipton, Kronrod, & Anandkumar, 2017) to implement an active learning model to fulfill the function in this science and technology research extraction framework. When the active learning model is compared with other algorithms, deep learning needs a large amount of labelled data to perform well, but when it comes to small datasets, the advantage is less obvious. Meanwhile, expecting better performance with less manual labelling work, active learning methods seek to select a subset of examples that can critically improve the model before asking the annotators to label them.

The deep learning method we used in our experiment implemented a CNN-CNN-LSTM architecture including character-level encoder, word-level encoder and tag decoder. The input unstructured data with the low rank will be chosen for active learning use sequence tagging.

We managed to get 65% of accuracy as human manually annotation with around 1300 data samples, whereas the standard learning strategy takes far more numbers of papers and patent records to get the same accuracy. As the number of samples increases, the performance of the model still remains stable. Experiments on a number of datasets show that with as little as 25% of the training instances, it is possible to obtain similar or superior performance compared to that of the complete datasets. In other words, our active learning query strategies can not only reduce annotation costs but also result in better quality predictors

(Beck, Specia, & Cohn, 2013).

4.2. Named Entity Recognition and Relation Extraction Process

During the learning process, active learning algorithm iteratively queries the most informative instances to manual verification and revision. The appropriate selection of instances in each epoch ensures the cost of manual work to be limited in a relatively low level. The workflow of an annotation assignment using active learning is shown in Figure 2 and Figure 3.

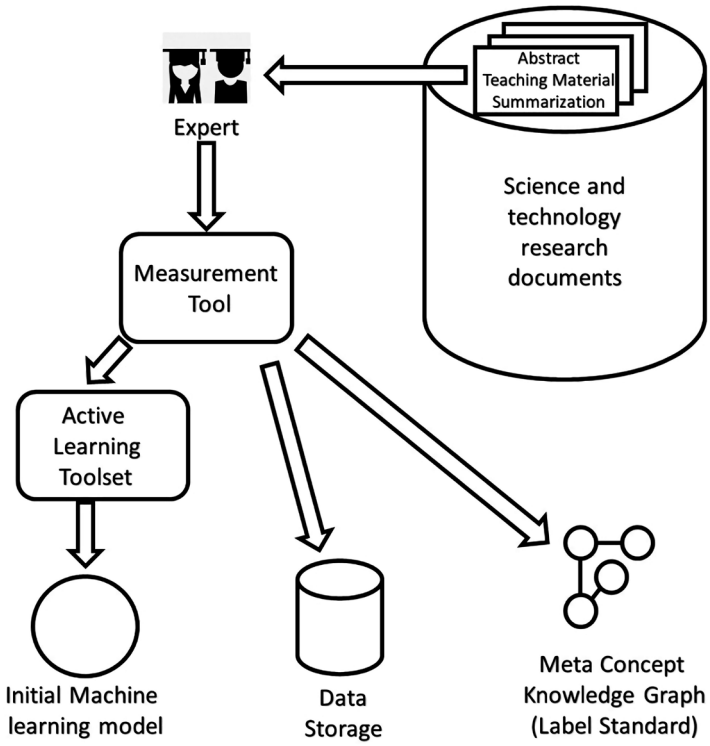

Figure 2. The workflow of the start-up for the framework.

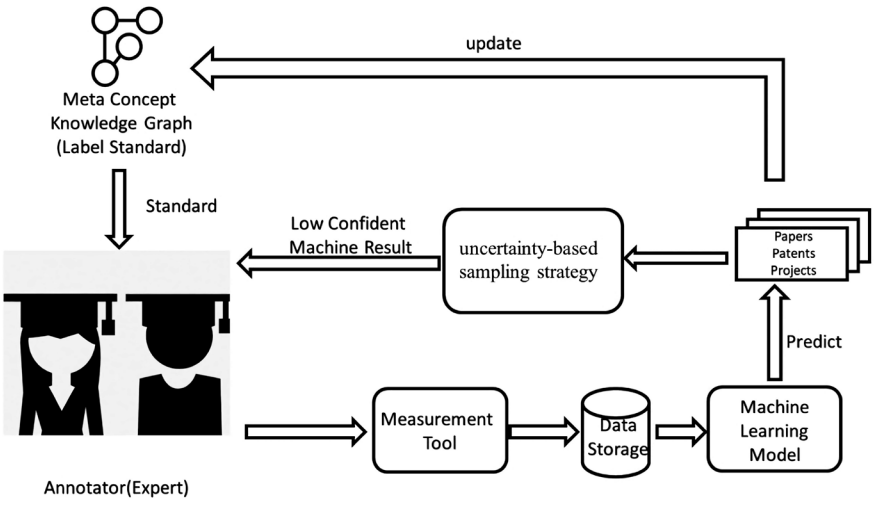

Figure 3. The workflow of the loop for the framework.

4.2.1. Start-Up Procedure for Active Learning Process

Before the start-up procedure of this framework, experts need to construct a concept set to provide standard labels for the annotation process.

This concept set will be generated by let experts annotate the data with more meta level information, such as paper abstract, patent abstract, summarizations and high-quality teaching materials. We'll also put the papers and researches that start up a field into consideration in this process. The annotated data will be stored in triples as RDF, and will be used as the initial data for the training of the startup active learning loop.

At the start-up of an annotation assignment, manager initializes it, and determines the research field of this assignment; chooses the range of target research documents that need to be annotated; the CKG construct using the meta concept extraction interface is used for annotation standard, and assigned to the experts. Then, the framework pushes part of randomly selected science and technology research documents to expert, and let them label the data. The labeled data is sent into the measurement module before being transferred to a "storage of training data". If the management determines to use the measurement tool, the training data will be passed to the acceleration tool for training. The initially trained model is be generated before the startup procedure using concept annotated by the expert at the beginning.

As we mentioned above, a high-quality meta concept knowledge graph will be used as the annotation standard. In this procedure, the labels that experts use will be provided by that CKG. Some frequently used concepts in the relevant field have already been displayed on the interface at the beginning of the start-up procedure. While the expert finds out an entity to be annotated, he or she should choose from those concepts to label the corpus. Sometimes, if the target label is not in the recommended label list, the expert needs to use the search function of the framework, the CKG will provide a list of the most relevant concepts in the result. And once the expert confirms one of those concepts is the label he or she wants, that concept and the concepts queried as nearest nodes in CKG will be added to the label queue, due to those concepts have closer relations with the chosen concept, which means they have higher chance to be needed in the same corpus.

In rare cases, the CKG may not contain the label that the expert need. We will discuss this situation in detail in Section 4.3.

By applying this function, the high-quality CKG standard prevents the problem caused by synonym and non-standardization. In the meantime, it also saves the time cost by the expert to self-define the label, which results in further labor cost saving.

4.2.2. Loop Procedure for Active Learning Process

After the startup, there is a loop of the active learning. As shown in **Figure 3**, the trained machine learning model tries to automatically perform NER on those unstructured science and technology research data not in the training storage,

resulting in a machine labelled records. Next, those records are applied for uncertainty-based sampling strategy and calculating the confidence to every machine labelled data. A certain amount of data with the lowest confidence is passed on to the experts for the annotation. The expert can choose to accept those annotation results labelled by model or re-annotate them again. Experts annotation results are transmitted to the training storage if they can pass the quality control tool.

After that, the machine learning model updates based on the new training storage. Finally, the framework starts the next cycle of a loop by applying the trained model to the unstructured research data out of the storage.

During the loop procedure, with the machine learning model starting to convergence for each time that the experts provide manual labeled data for training, the number of data that the model has low confident and requires the expert to manually labelled will become lower. For this advantage brought by active learning, the labor cost can be dramatically saved during the process. This is also the main labor cost saving function provided by our framework.

4.2.3. Termination Procedure for Active Learning Process

The loop terminates once the management demands that the performance of the model is good enough. The data in training storage and the rest of the machine labelled data is moved to the result storage and becomes the final result of this annotation assignment.

In this framework, those manually labelled instances can be directly transfer into science and technology research instance knowledge graph. Due to those annotation results come from the expert's prior knowledge and have been applied on measurement module, we regard that the knowledge graph has a relatively reliable quality. In the termination process, the converged model is applied to the remaining unstructured data in the data storage, generates the high accuracy auto extracted relations and entities, further automatically combined with the instance knowledge graph generated in the process to output the final product. With that converged active learning model, this procedure takes no labor cost and can still result in final product with good quality.

Due to the help of the active learning, with the growing of dataset to annotate, only a few of data need to be manually processed. Therefore, the labor cost is reduced using this toolset.

4.3. Quality Control

During the annotation process, a set of tools to evaluate the data is needed to help us measure the quality, which is essential and critical. To measure the quality of the generated data, we involve two measurement functions. One focuses on avoiding mistakes from the algorithm used in the process, while the other focuses on avoiding mistakes coming from the experts who participate in the annotation process. There is an additional mechanism to maintain the meta concept standard CKG by updating or modifying them. This mechanism can al-

so measure the quality of the FKG generated using this framework.

4.3.1. Measurement Methodology

In this framework, the method of machine learning is replaceable as long as the accuracy of the algorithm is assured. To assure the quality of framework, the measurement standards declared as follows need to be applied before implementing the machine learning model. We use an already annotated data from this framework and randomly divide part of it as a test set to apply to the algorithm. Then we evaluate by comparing the result with the dataset we use, generate a percentage as feedback to experts, and let the experts decide whether the error coming from this algorithm is acceptable or not.

To measure the quality of annotated data, the mistakes from the annotator should be minimized. Therefore, the framework needs an inner annotator agreement measurement system in order to alleviate the problem. Cohen's Kappa, has been proved can be used as a very effective agreement measurement evaluation method (Vieira, Kaymak, & Sousa, 2010). We apply Cohen's Kappa evaluation between the examiners who measure the labelled result from annotator, and only send the examined data which pass the evaluation score threshold to the active learning toolset. The manager should define the threshold at the start of the annotation assignment. In our framework, the examiner only needs to check the output from the annotator. With the active learning applied, only a few data will be labelled by annotator, as we mentioned before. This results in only a few data to be examined during the process. By using the active learning method, the framework not only saves the labor cost of annotator but also saves the labor cost in the process of quality control. All the participants and user of this framework should have solid research field knowledge, or otherwise the quality of final product cannot be measure.

4.3.2. Updateable CKG

This framework will construct reliable CKG to provide standards. During the annotation process, the experts are asked to choose from CKG standard to annotate on the target corpus rather than self-defined one.

During the annotation, the framework suggests labels chosen from CKG, which helps experts to annotate the science and technology research unstructured data and produce reliable and standardized annotation results. However, sometimes the CKG may have a defect. For example, in the initialization process, the experts make a mistake on annotated the meta concept, or through time, the concept has renewed and changed. If experts continue to annotate based on that CKG, the quality of final labelling result will be damaged. Therefore, we develop a CKG update mechanism. Any update to the CKG is sent to the inner annotator agreement measurement system. Since CKG should be recognized as a reliable source, aligned with the two former measurements functions, the CKG modification will only be accepted if the inspectors fully agree with it.

This mechanism not only maintains the quality of the annotation result but

also measures the quality of CKG at the same time. By applying direct mapping between annotation result and model prediction result with the updated CKG, the high-quality FKG will be generated and ready to use for providing further help.

5. Conclusion

In this paper, we designed and proposed a framework using active learning; this framework can be used to extract entity and relation from unstructured science and technology research data, such as papers, patents, and research project descriptions. This framework first asks experts to annotate the concept from more critical and standardized data, such as summarization and abstract, as well as teaching material. By using those data to construct a CKG as annotation label, it further implements active learning tools and helps the expert to rapidly annotate the data with high accuracy. The quality control has also been taking part in consider during this framework. Eventually, this framework will generate accurate science and technology knowledge graph with fast speech. Those knowledge graph constructed by this framework can be used to finding similar research area, finding similar researchers, finding popular research areas and so on.

In the future, we are going to improve this framework by developing a lower labor cost method on concept extraction part at the initial of this framework; we will also build serval science and technology research knowledge graph and use them in real-world situations.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Beck, D., Specia, L., & Cohn, T. (2013). Reducing Annotation Effort for Quality Estimation via Active Learning. In *Association for Computational Linguistics Conference* (pp. 543-548). Sofia: Association for Computational Linguistics.

Chen, W., & Styler, W. (2013). *Anafora: A Web-Based General Purpose Annotation Tool*. NAACL HLT Demonstration Session, 14-19.

Coelho da Silva, T. L., Magalhães, R. P. et al. (2019). Improving Named Entity Recognition Using Deep Learning with Human in the Loop. In *Proceedings of the 22nd International Conference on Extending Database Technology* (pp. 594-597). Lisbon: OpenProceedings.org.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A Web-Based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *LT4DH Workshop* (pp. 76-84). Osaka: The COLING 2016 Organizing Committee.

Giorgi, J. M., Bader, G. D., & Wren, J. (2020). Towards Reliable Named Entity Recognition in the Bio-Medical Domain. *Bioinformatics, 36,* 280-286. https://doi.org/10.1093/bioinformatics/btz504

Gong, J. B., Wang, S., Wang, J. L., Feng, W. Z., Peng, H., Tang, J., & Yu, P. S. (2020). Attentional Graph Convolutional Networks for Knowledge. Concept Recommendation in

MOOCs in a Heterogeneous View. In *SIGIR* (pp. 79-88). Virtual Event: ACM. https://doi.org/10.1145/3397271.3401057

Holzinger, A. (2016). Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop? *Brain Informatics, 3,* 119-131. https://doi.org/10.1007/s40708-016-0042-6

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The Inception Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics* (pp. 5-9). Santa Fe, NM: Association for Computational Linguistics.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 260-270). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1030

Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge Graph Identification. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 542-557). Berlin: Springer.

Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., & Anandkumar, A. (2017). *Deep Active Learning for Named Entity Recognition*. In *Proceedings of the 2nd Workshop on Representation Learning for NLP* (pp. 252-256). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2630

Sheng, M., Shao, Y., Zhang, Y., Li, C., Xing, C., Zhang, H., Wang, J., & Gao, F. (2019). DEKGB: An Extensible Framework for Health Knowledge Graph. In *ICSH* (pp. 27-38). Shenzhen: Springer. https://doi.org/10.1007/978-3-030-34482-5_3

Sheng, M., Wang, J., Zhang, Y., Li, X., Li, C., Xing, C., Li, Q., Shao, Y., & Zhang, H. (2019). DocKG: A Knowledge Graph Framework for Health with Doctor-in-the-Loop. In *HIS* (pp. 3-14). Xi'an: Springer. https://doi.org/10.1007/978-3-030-32962-4_1

Verborgh, R., Vander Sande, M., Hartig, O. et al. (2016). Triple Pattern Fragments: A Low-Cost Knowledge Graph Interface for the Web. *Journal of Web Semantics, 37,* 184-206. https://doi.org/10.1016/j.websem.2016.03.003

Vieira, S. M., Kaymak, U., & Sousa, J. M. C. (2010). Cohen's Kappa Coefficient as a Performance Measure for Feature Selection. In *WCCI 2010* (pp. 1-8). Barcelona: IEEE. https://doi.org/10.1109/FUZZY.2010.5584447

Wan, H. Y., Zhang, Y. T., Zhang, J., & Tang, J. (2019). AMiner: Search and Mining of Academic Social Networks. *Data Intelligence, 1,* 58-76. https://doi.org/10.1162/dint_a_00006

Yang, J., Zhang, Y., Li, L. W., & Li, X. X. (2018). YEDDA: A Lightweight Collaborative Text Span Annotation Tool. In *ACL 2018* (pp. 31-36). Melbourne: Association for Computational Linguistics.

Yang, Y., Kandogan, E., Li, Y., Sen, P., & Lasecki, W. S. (2019). A Study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics. In *CEUR Workshop* (Vol. 2327). Los Angeles: CEUR-WS.org.

Yuan, S., Shao, Z., Liang, Y. X., Tang, J., Hall, W., Liu, G., & Zhang, Y. T. (2020). International Scientific Collaboration in Artificial Intelligence an Analysis Based on Web Data. In *12th ACM Conference on Web Science* (pp. 69-75). Southampton: ACM. https://doi.org/10.1145/3394231.3397896