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

Yu Li†
National Science Library, Chinese Academy of Sciences
Beijing, China
yul@ mail.las.ac.cn

Tao Yue
National Science Library, Chinese Academy of Sciences
Beijing, China
taoyue@ mail.las.ac.cn

Wu Zhenxin
National Science Library, Chinese Academy of Sciences
Beijing, China
wuzx@ mail.las.ac.cn

ABSTRACT

The terminologies in different disciplines vary greatly, and the annotated corpora are scarce, which have limited the portability of information extraction models. The content of scientific articles is still underutilized. This paper constructs an intelligent platform for information extraction and knowledge mining, namely IEKM-MD. Two innovative technologies are proposed: Firstly, a phrase-level scientific entity extraction model combining neural network and active learning is designed, which can reduce the model's dependence on large-scale corpus. Secondly, a translation-based relation prediction model is provided, which improves the relation embeddings by optimizing loss function. In addition, the platform integrates the advanced entity recognition model (spaCy.NER) and the keyword extraction model (RAKE). It provides abundant services for fine-grained and multi-dimensional knowledge, including problem discovery, method recognition, relation representation and hot spot detection. We carried out the experiments in three different domains: Artificial Intelligence, Nanotechnology and Genetic Engineering. The average accuracies of scientific entity extraction respectively are 0.97, 0.69 and 0.78.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence → Natural language processing →Information extraction

KEYWORDS

Information extraction, Relation prediction, Active learning, Translation embedding, Neural network

ACM Reference format:

Yu Li, Tao Yue and Wu Zhenxin. 2020. IEKM-MD: An Intelligent Platform for information Extraction and Knowledge Mining in Multi-Domains. In *Proceedings of the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2020), Wuhan, Hubei, China, 4 pages.* https://doi.org/

1 Introduction

With the progress of science and technology, there are more and more fields and scientific articles. Information extraction and knowledge mining in the specific field enable scholars to quickly grasp the overall outline of information, and track the development of fine-grained knowledge. There are many mature models to extract information from texts, such as BiLSTM-CNN [1], CNN-BiLSTM-CRF [2], LM-LSTM-CRF [3], which have achieved high scores in various tasks of natural language processing. In fact, these supervised learning models inevitably consume large amounts of high-quality annotated corpus in order to fully learn the characteristics of natural language representation. In most case, however, the annotated corpus in one specific field is constructed manually by several experts, which is time-consuming and laborious. Therefore, it is hard to directly use a well-trained model to other domains.

How to extract information without massive annotated corpus is a big challenge. Active Learning (AL) [4] has been proved to be an effective way to solve the problem of corpus scarcity when dealing with the classification tasks [5, 6]. However, it has not been validated on the sequence labelling task, which is more difficult to find the optimal result because its complexity increases exponentially [7]. In this paper, we introduce multiple active learning strategies into information extraction for the first time, so as to explore a cheap and efficient solution for recognizing the fined-grain entities in multiple domains.

Relation predication is another basic technology for knowledge organization. Translation models see relation as a process of translating the head entity to the tail entity, which have been wildly used to predict relations. There are some classic translation models proposed from different perspectives: TransE [8] is the first translation embedding model with fewer parameters. TransH [9] is presented to solve the problem of complex relation representation. TransR [10] distinguishes the semantic embedding for different types of relations, which wined a better F-score. TransD [11] simplifies the projection process of TransR and improves the computing efficiency.

This paper aims to construct an intelligent platform for information extraction and knowledge mining, which can be used in multiple domains without much human intervention. The main contributions are as follows: 1). with the limited annotated corpus, an effective method combining neural network with active learning recognizes scientific entities in multiple domains; 2). By optimizing the loss function, an improved translation model represents the semantic vectors more accurately and reaches the convergence state faster with a small loss score compared with the original model.

2 Intelligent Platform: IEKM-MD

The technology framework of our platform is shown in Figure 1. This platform includes two innovative technologies: 1) the model combining neural network with active learning extracts "problem" and "method" entities, 2) the improved translation model predicts relations between "problem" and "method" entities. At the same time, the platform integrates two excellent tools (spaCy.NER[1] and RAKE[2]) to recognize the named entities and keywords. Finally, this platform provides a variety of knowledge services for researchers, including problem discovery, method recognition, relation representation and hot spot detection. Besides, the analyzers can perform richer downstream tasks based on our platform, such as discipling analysis, trend explosion, new technology detection, and so on.

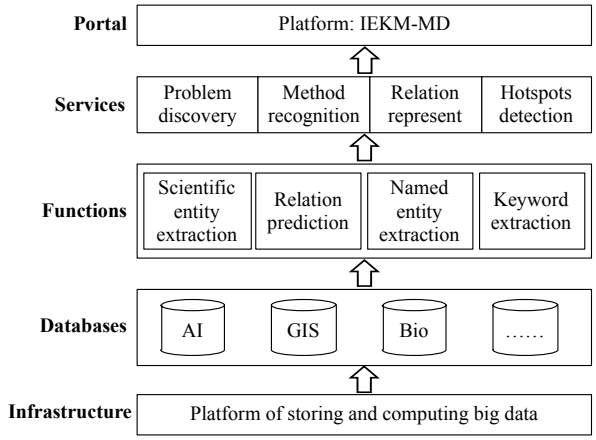
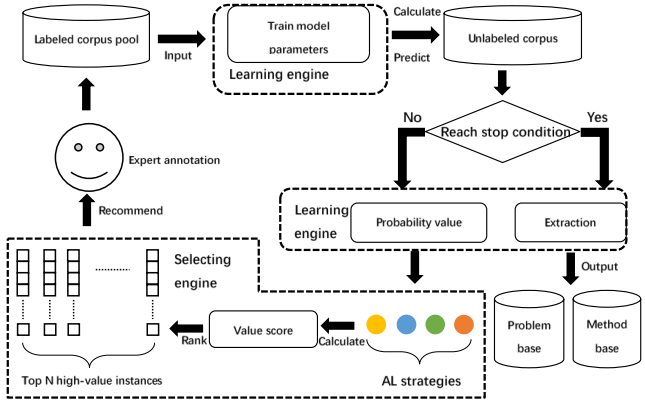


Figure 1: Technology framework of IEKM-MD

2.1 Scientific Entity Recognition

Scientific entity recognition contributes to extract phrases from scientific articles. These phrases consist of several words which describe the focus of article or the method proposed by author. In order to reduce the dependence on annotated corpus, this paper provides a semi-supervised learning model combining neural network with active learning.



[1] https://spacy.io/
[2] https://github.com/aneesha/RAKE

Figure 2: Information extraction model combining neural network with active learning

The framework of the information extraction model is shown in Figure 2. Firstly, the learning engine trains the parameters of neural network by using a small number of annotated samples. Then, the trained neural network predicts the labels of unannotated samples and inputs the predicted probabilities to the selecting engine. Secondly, according to the active learning strategies, the selecting engine decides which samples are valuable and should be annotated manually. Thirdly, the manually annotated samples are added into the training set to re-train the neural network, in order to improve the performance of label prediction. The whole process runs repeatedly until the performance of model has no significant optimization. Finally, the trained model predicts the "problems" and "methods" for all unlabeled articles.

Here we choose CNN-BiLSTM-CRF [12] as the learning engine. CNN focuses on the morphology features that are the prefix and suffix of word. BiLSTM learns the dependency relationship between words with a long distance by using two groups of long-short term memory networks in opposite directions. CRF decides the most optimal labeling sequence with a rational linguistic logic.

In addition, we propose a voting approach for the selecting engine. Firstly, the annotation values of the unlabeled samples are computed in a cumulative way by four types of active learning strategies: margin [13], N-best sequence entropy [14], maximum normalized log-probability [15] and label weighted probability. Secondly, the annotation values are listed in descending order, only the top 10% of samples are selected to be annotated manually in each iteration.

2.2 Entity Relation Prediction

Relation prediction decides whether a "problem" and a "method" is related or not. That means if a "problem" is related to a "method", the method can be used to solve this problem.

Translation model sees the relation in the triple (head entity, relation, tail entity) as a translational between two entities. There is a series of translation models. TransE [8] has few parameters and is low in complexity, but cannot distinguish two tail entities with the same relation. TransH [9] uses different vectors to represent one entity with various relations, which solves the problem of complex relation representation (1-N, N-1, N-N). TransR [10] supposes that different relations are in different semantic spaces. Thus, this model projects entities into their relation spaces at first, then builds the translation process. However, it greatly increases the time cost because of too many parameters. TransD [11] creates the projection matrix respectively for head entity and tail entity. It not only combines the effects of both entities and relations on projection, but also improves the computing efficiency.

After comparing the performance of various translation models, we choose TransH to predict relations, which keeps balance between accuracy and efficiency. To solve the problems of one-to-many, many-to-one, many-to-many relations, TransH generates

the relation-specific translation vector d_r in the relation-specific hyperplane w_r rather than in the same space of entity embeddings.

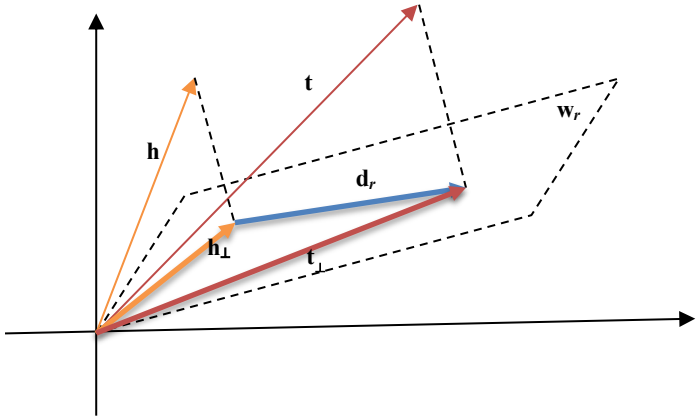


Figure 3: TransH projection [9]

As shown in Figure 3, the relation r in its hyperplane w_r has a translation vector d_r, the head embedding h and the tail embedding t in w_r have their projection vectors $h_\perp$ and $t_\perp$. The defined score function is: $||h_\perp + d_r - t_\perp||_2^2$.

However, the original TransH model does not match our goal exactly. We achieved three improvements.

1) TransH constructs the negative samples by replacing the head or tail entity with others in the positive samples. However, the replaced one may also be correct because of synonyms, which introduced many false negative labels into training. Considering that there are only two types of relationships, we simply construct the negative samples by modifying the correct relationship into its antonym. By this change, it is more convenient to construct a balanced annotated corpus. Moreover, the score function $f_r(h,t)$ is re-defined as Equation (1), which aims to move the attention from entity to relation.

$$f_r(h,t) = ||abs(h_\perp - t_\perp) - d_r||_2^2 \quad (1)$$

2) Comparing with the original model that initializes the entities with the random vectors, we use the word2vec model to generate the semantic representation of all head and tail entities.

3) To improve the ability of feature learning for the unknown entities, we add one hidden layer of linear transformation respectively for the head entities and tail entities.

2.3 Named Entity Recognition and Keyword Extraction

We use an enterprise open source toolkit spaCy.NER to recognize the named entities. spaCy.NER implements a very fast and efficient system based on the statistical machine learning algorithms, which can recognize 18 entity types, such as Person, Organization, Location, Geopolitics entity.

Furthermore, keyword extraction is achieved by the open source toolkit RAKE (Rapid Automatic Keyword Extraction). RAKE is an automatic keyword extraction technique. Based on the statistical method, RAKE outperformed TextRank and other supervised learning models, which obtained a high F value [16] and is more efficient.

3 Platform Evaluation and Display

We choose three different domains to verify the practical application effect of IEKM-MD. They are Artificial Intelligence (AI), Nanotechnology (Nano) and Genetic Engineering (GE). The abstracts are collected from NSTL database[3], the number of which are respectively 46680, 434400 and 7214 in total from 1991-2020.

3.1 Scientific Entity Recognition

We manually checked the top 30 results to calculate the recognition accuracy (ACC) as shown in Table 1. The results reflect that AI achieved the best performance with 0.97 accuracy score. The average accuracy of three fields reveals that problem extraction has a better score than method extraction. The first reason is that the total mentions of problem are smaller than methods, and they are usually described in the noun phrases, which contribute to an easier pattern to be caught by model. The second reason is that one article may contain multiple methods, which are modified by multiple attributives or adverbials, making it more challenging to recognize the complete methods.

However, our platform performed worst in the field of Nano. This may because that the articles of Nano include many complex and specialized terms in the subjects of biology, physics, chemistry, electronics, and metrology. Our platform still lacks the professional knowledge to learn the specific features.

Table 1: Multi-domain specific entity recognition results

Field	Type	ACC
AI	problem	0.97
	method	0.97
Nano	problem	0.70
	method	0.67
GE	problem	0.83
	method	0.73

The extracted top 10 problems of three fields are shown in Table 2, which reveal that AI focuses on the classification, prediction and recognition problems of data and images in the subject of Computer Science. Nano covers a wide range, including physics, biology, chemistry, and so on, which focuses on the applications on the basic disciplines. Therefore, the extracted problems involve detection, analysis and prediction of energy, atom and medicine. The scope of GE is relatively narrow and is related to drug development, disease treatment, and biological manufacturing in the biomedical field.

[3] https://www.las.ac.cn

Table 3: Multi-domain problem recognition

Top	AI	Nano	GE
1	Classification	Detection	Drug discovery
2	Prediction	Optimization	Identification
3	Pattern recognition	Energy storage chemical prediction	Disease resistance
4	Feature selection	Sensitive detection	Crop protection
5	Optimization	Remote sensing	Drug delivery
6	Datum mining	UV detection	Genetic engineering
7	Binary classification	Hydrothermal clinical diagnosis	Biodiesel production
8	Computer vision	Determination	Cancer immunotherapy
9	Feature extraction	Excitation limit of detection	Biofuel production
10	Image classification	Atomic layer deposition	Biomedical

Table 3 shows the extracted top 10 methods. In the field of AI, they are mostly based on machine learning models and neural networks. The technologies in Nano are specific instruments, such as microscope, spectrograph and ray. For GE, gene editing, manipulation and recombination are the three main techniques.

Table 2: Multi-domain method recognition

Top	AI	Nano	GE
1	Machine learning	X-Ray diffraction (XRD)	Polymerase Chain Reaction (PCR)
2	Support vector machine	Transmission electron microscopy (TEM)	Genetic engineering strategy
3	Classification	Scanning electron microscopy (SEM)	Gene therapy
4	Random forest	Raman spectroscopy	Southern blot analysis
5	Neural network	Fourier transform infrared spectroscopy (FTIR)	Biotechnology
6	Deep learning	Atomic force microscopy (AFM)	Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)
7	Decision tree	High Performance Liquid Chromatography (HPLC)	enzyme-linked immunosorbent assay (ELISA)
8	Feature selection	Elemental analysis	Genetic transformation
9	Datum mining	X-ray photoelectron spectroscopy (XPS)	Genetic manipulation
10	Artificial neural network	Hydrothermal atomic force microscopy	Recombinant DNA

3.2 Entity Relation Prediction

By predicting the relations between problem and method, we construct the method-problem networks for different domains. As shown in Figure 4, the methods and problems which were separate in the articles of AI are linked by relation prediction. The red dots refer to methods, and the blue dots refer to problems.

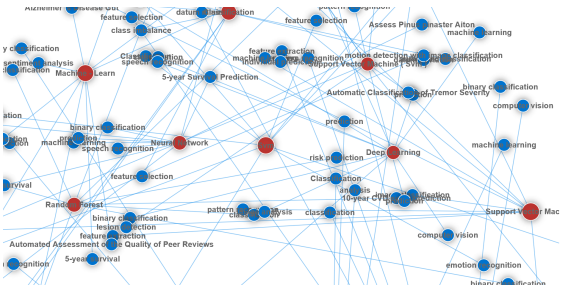


Figure 4: Problem-method relation network in AI

Specifically, we can get more details from the above-mentioned network. By setting the method X-Ray Diffraction (XRD) as a center, Figure 5 reveals that what problems are solved by XRD. They are Assisted Synthesis, Biomedical Application, Biosynthesis of Silver Nanoparticles and so on.

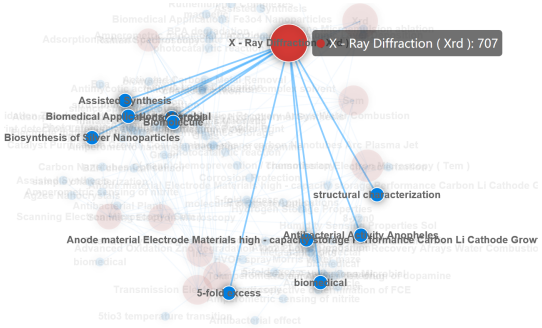


Figure 5: The problems solved by XRD method in Nano

3.3 Hotspot Detection

Hotspots are the most popular research topics. We use the extracted keywords to pick out the hotspots in multiple domains. As a hotspot, the total number occurring in articles should be increased year by year or keeps a steady top order in last three years. According by this rule, Figure 6 shows the hotspots in the field of AI. They are distinct from the scientific entities recognized in section 4.1, which have no semantic type but reflects the popularity degree of terms.

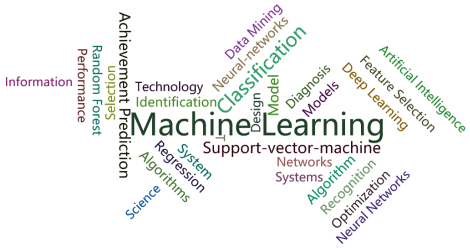


Figure 6: Hotspots in AI

4 Conclusion

This paper introduced an innovative and intelligent platform IEKM-MD to extract information and mine knowledge from scientific articles in multiple domains. One contribution is providing a hybrid active learning strategy to solve the problem of annotated corpus scarcity in supervised learning model. Another contribution is designing an improved Translation embedding approach based on TransH model to optimize the performance of relation prediction. Three datasets in AI, Nano and GE show that our platform is enable to achieve various knowledge services with a high accuracy in multiple domains.

ACKNOWLEDGMENTS

This work is supported by the project “Annotation and evaluation of the semantic relationship between geographical entities in Chinese web texts” (Grant No. 41801320) from the National natural science foundation of China youth science foundation.

REFERENCES

[1] Chiu Jason, Nichols Eric. 2015. Named entity recognition with bidirectional LSTM-SNNs. Transactions of the Association for Computational Linguist 6(Nov. 2015). DOI: https://doi.org/10.1162/tacl_a_00104.

[2] Ma Xuezhe, Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv:1603.01354. Retrieved from https://arxiv.org/abs/1603.01354.

[3] Liyuan Liu, Jingbo Shang, Frank F. Xu, Xiang Ren, Huan Gui, Jian Peng, Jiawei Han. 2017. Empower sequence labeling with task-aware neural language model. arXiv:1709.04109. Retrieved from https://arxiv.org/abs/1709.

[4] Kulkarni, Sanjeev and Mitter, Sanjoy and Tsitsiklis, John and Systems, Massachusetts. 1993. Active Learning Using Arbitrary Binary Valued Queries. Machine Learning 11, 1 (Apr. 1993), 23-35. DOI: https://doi.org/11.10.1023/A:1022627018023.

[5] Vijayanarasimhan Sudheendra, Grauman Kristen.2012. Active frame selection for label propagation in videos. In Proceedings of the 12th. European Conference on Computer Vision (ECCV’12), Florence, Italy. Springer-Verlag. Heidelberg, Berlin, 496-509. https://doi.org/10.1007/978-3-642-33715-4_36.

[6] Deng Yue, Dai Qionghai, Liu Risheng, Zhang Zengke, Hu Sanqing. 2013. Low-rank structure learning via non-convex heuristic recovery. IEEE Transactions on Neural Networks and Learning Systems, 24(3): 383–396. DOI: https://doi.org/10.1109/TNNLS.2012.2235082.

[7] Deng Yue, Chen Kawai, Shen Yilin, Jin Hongxia. 2018. Adversarial active learning for sequences labeling and generation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, July, 2018, Stockholm, Sweden. IJCAI-18. California, 4012-4018. https://doi.org/10.24963/ijcai.2018/558.

[8] Bordes Antonie, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In Proceedings of NIPS. MIT Press. Cambridge, MA, 2787-2795.

[9] Zhen Wang, Jianwen Zhang, Jianlin Feng, Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the 28th. AAAI Conference on Artificial Intelligence (AAAI’14), June, 2014. AAAI Press. Menlo Park, CA, 1112-1119. https://doi.org/10.5555/2893873.2894046.

[10] He Shizhu, Liu Kang, Ji Guoliang, Zhao Jun. 2015. Learning to represent knowledge graphs with Gaussian embedding. In Proceedings of CIKM. ACM. New York, 623-632. https://doi.org/10.1145/2806416.2806502.

[11] Ji Guoliang, He Shizhu, Xu Liheng, Liu Kang, Zhao Jun. 2015. Knowledge graph embedding via dynamic mapping matrix. In Proceedings of ACL. ACL. Stroudsburg, PA, 687-696. https://doi.org/10.3115/v1/P15-1067.

[12] Xuezhe Ma, Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [OL]. arXiv: 1603.01354. Retrieved from https://arxiv.org/abs/1603.01354.

[13] Yanyao Shen, Hyokun Yun, Zachary C. 2017. Lipton, Yakov Kronrod, Animashree Anandkumar. Deep active learning for named entity recognition. arXiv:1707.05928. Retrieved from https://arxiv.org/abs/1707.05928.

[14] Seokhwan Kim, Yu Song, Kyungduk Kim, Jeong-Won Cha, Gary Geunbae Lee. 2006. MMR-based active machine learning for bio entities. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, June, 2006, New York., New York, USA, 69–72.

[15] Balcan Maria-Florina, Broder Andrei, Zhang Tong. 2007. Margin based active learning. In Proceedings of the 20th. Annual Conference on Learning Theory (COLT’07), 2007, San Diego, CA, USA. Springer-Verlag., Berlin, Heidelberg, 35–50. https://doi.org/10.5555/1768841.1768848

[16] Stuart Rose, Dave Engel, Nick Cramer, Wendy Cowley. 2010. Automatic keyword extraction from individual documents. Text Mining: Applications and Theory 20, 1 (Mar. 2010), 1-20. DOI: https://doi.org/10.1002/9780470689646.ch1.