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 0.91, 0.52 and 0.76.

CCS CONCEPTS

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

KEYWORDS

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

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 widely 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.

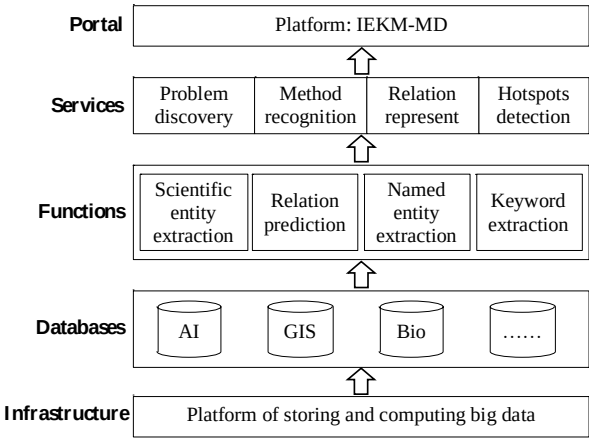

Figure 1: Technology framework of IEKM-MD

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.

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 (dozens of abstracts with semantic labels). Then, the trained neural network predicts the labels of unannotated samples and inputs the predicted scores to the selecting engine. Secondly, according to the active learning strategies, the selecting engine decides which samples are valuable and should be annotated manually. Only the top 10% most valuable samples are labelled by experts. 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 the unlabeled articles. More details about parameter setting will be discussed in Section 3.1.

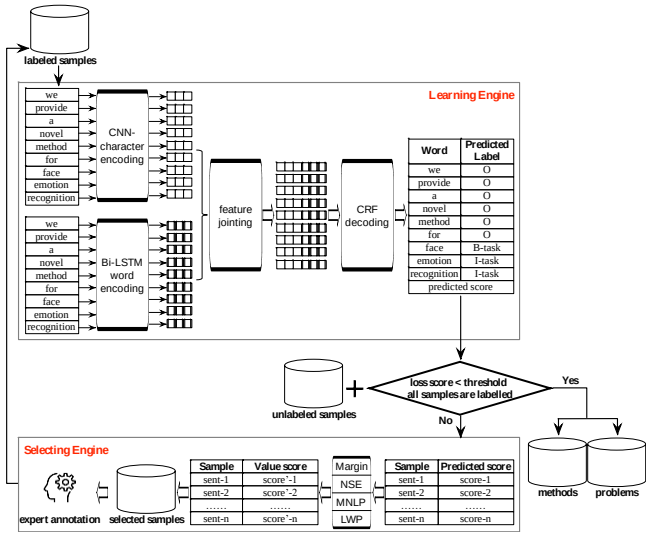

Figure 2: Information extraction model combining neural network with active learning

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 hybrid approach for the selecting engine. Firstly, the value score of each unlabeled sample is respectively computed by four different types of active learning strategies, and the sum of them is set as the final value score. Secondly, the value scores are listed in descending order, only the top 10% most valuable samples are selected to be annotated manually in each iteration.

This paper picked out three classical strategies from the uncertain sampling methods: margin [13], N-best sequence entropy [14] and maximum normalized log-probability [15]. Additionally, we propose a novel strategy, namely label weighted probability, which enhances on the importance of the number of labels. The more labels of problems or methods there are in a sentence, the more valuable the sentence is.

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

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

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.

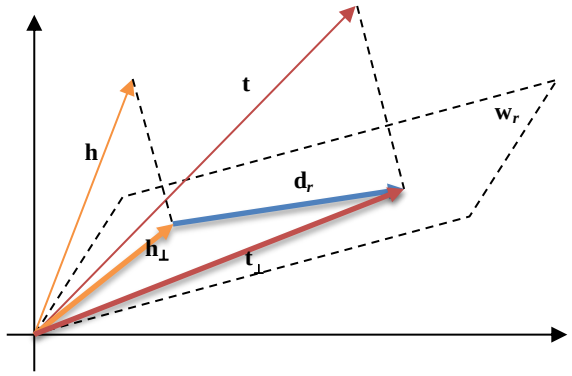

Figure 3: TransH projection [9]

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 evaluate the performance of information extraction of IEKM-MD in the field of Artificial Intelligence (AI). There are two datasets be used.

1) The top 100 AI conferences were picked out by the domain experts, and their abstracts were acquired from NSTL database[3], total in 9753 sentences. Next, we built the truth datasets. Each sentence is annotated synchronously by two students in the corresponding subjects (task, method or other). The annotation results are checked by one expert. The annotation format is shown as Figure 4. The AI annotated corpus contains 26,0000 tokens.

We	use	active	learning	to	extract	information
|	|	|	|	|	|	|
O	*O*	*B-method*	*I-method*	*O*	*B-task*	*I-task*

Figure 4: An example of annotation format

2) FTD datasets[4] shared by Stanford University in the field of Computational Linguistics. It comes from the Conference of the Association for Computational Linguistics and ranges from 1965 to 2009, which containing four types of labels: focus, technique, domain and other, in total 2628 sentences.

In addition, we show the effect of knowledge mining in three different kinds of domains. We choose three popular keywords (Neural Networks, Nano Structure and Genetic Engineering) that respectively respect the subjects of Computer Science, Material and Medicine to acquire abstracts from NSTL database. 200

[3] https://www.las.ac.cn
[4] https://nlp.stanford.edu/pubs/FTDDataset_v1.txt

abstracts of each subject are randomly selected from SCI journals and are used to verify the practical application effect of IEKM-MD.

3.1 Scientific Entity Recognition

We set the baselines only using the CNN-BiLSTM-CRF (CBC) model trained on all annotated samples. For each dataset (AI or FTD), the best performance is as the baseline, so as to detect whether active learning helps reduce the scale of annotated corpus for supervised learning models. The scale of training sets and the best F1 scores of CBC model are shown in Table 1.

Table 1: Best F1 of three datasets trained by CBC model

Metric	AI		FTD		
	Problem	Method	Focus	Technique	domain
Instances in training set	5763	12041	1740	1986	1652
Best F1 score	73.70%	71.24%	55.33%	51.33%	57.73%

In the model of IEKM-MD, initially only 0.01% annotated samples are used to carry out the cold starting process, then the highest valuable samples (10%) are added into the training sets in each iteration. Only if the F1 score of IEKM-MD reaches the baseline, can the learning process be stopped. The label scales and F1 scores of AI and FTD datasets in each iteration are show in Table 2.

Table 2: Learning effect of IEKM-MD in each iteration

Step	Metric	AI		FTD		
		Problem	Method	Focus	Technique	Domain
Initial	Labels	31	67	6	8	6
Iteration-1	Labels	694	1303	272	284	371
	F1	64.20%	60.23%	42.81%	42.33%	47.59%
Iteration-2	Labels	1232	2713	428	403	452
	F1	68.18%	66.43%	46.27%	49.02%	53.20%
Itereation-3	Labels	1729	3866	564	618	573
	F1	**75.87%**	**72.57%**	**57.41%**	50.70%	**58.00%**
Iteration-4	Labels	-	-	-	821	-
	F1	-	-	-	**52.23%**	-

Table 1 and 2 reveal that after combing supervised learning model with active learning strategies, the annotated samples can be cut down 60%-70%.

After IEKM-MD achieves the best performance as that CBC model did, the model extracts problems and methods from Neural Networks, Nano Structure and Genetic Engineering datasets. We manually checked the top 30 problems and methods and evaluated their accuracies as shown in Table 3.

The results reflect that Neural Networks achieved the best performance with 0.93 accuracy of problem extraction and 0.89 accuracy of method extraction. 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.

Table 3: Accuracies of scientific entity recognition

Metric	AI		Nano Structure		Genetic Engineering	
	Problem	Method	Problem	Method	Problem	Method
Accuracy	0.93	0.89	0.61	0.42	0.77	0.75

However, our platform performed worst in the field of Nano Structure. This may because that the articles of Nano Structure 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.

The extracted top 10 problems of three fields are shown in Table 4, which reveal that Neural Networks focuses on the classification, prediction and recognition problems of data and images in the subject of Computer Science. Nano Structure 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 Genetic Engineering is relatively narrow and is related to drug development, disease treatment, and biological manufacturing in the biomedical field.

Table 4: Problem recognition in multiple domains

Top	Neural Network	Nano Structure	Genetic Engineering
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 5 shows the extracted top 10 methods. In the field of Neural Networks, they are mostly based on machine learning models, such as support vector machine, random forest, deep learning. The technologies in Nano Structure are specific instruments, such as microscope, spectrograph and ray. For Genetic Engineering, gene editing, manipulation and recombination are the three main techniques.

Table 5: Method recognition in multiple domains

Top	Neural Network	Nano Structure	Genetic Engineering
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 5, the methods and problems which were separate in the articles of Neural Network are linked by relation prediction. The red dots refer to methods, and the blue dots refer to problems.

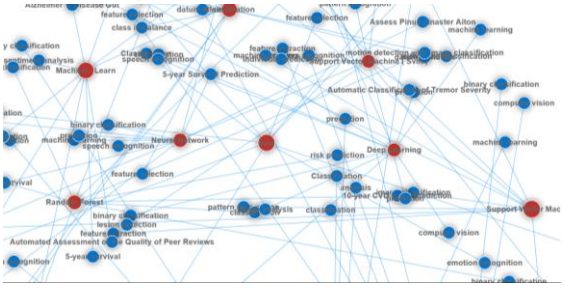

Figure 5: Problem-method relation network in Neural Networks

Specifically, we can get more details from the above-mentioned network. By setting the method X-Ray Diffraction (XRD) as a center, Figure 6 reveals that what problems are solved by XRD. They are Assisted Synthesis, Biomedical Application, Biosynthesis of Silver Nanoparticles and so on.

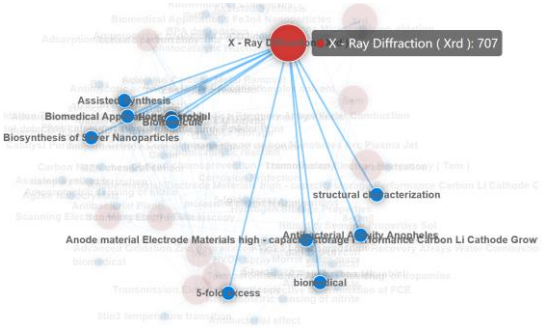

Figure 6: The problems solved by XRD method in Nano Structure

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 7 shows the hotspots in the field of Neural Networks. They are distinct from the scientific entities recognized in section 3.1, which have no semantic type but reflect the popularity degree of terms.

Figure 7: 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 Neural Networks, Nano Structure and Genetic Engineering 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.