Metadata Extraction for Scientific Papers

Binjie Meng[1,2], Lei Hou[3], Erhong Yang[1,2⋆], and Juanzi Li[3]

[1] Beijing Advanced Innovation Center for Language Resources,
Beijing Language and Culture University, Beijing, China
[2] School of Information Science,
Beijing Language and Culture University, Beijing, China
[3] Dept. of Computer Sci. and Tech, Tsinghua University, Beijing, China
{mllrose, yerhong}@126.com, {greener2009, lijuanzi2008}@gmail.com

Abstract. Metadata extraction for scientific literature is to automatically annotate each paper with metadata that represents its most valuable information, including problem, method and dataset. Most existing work normally extract keywords or key phrases as concepts for further analysis without their fine-grained types. In this paper, we present a supervised method with three-stages to address the problem. The first step extracts key phrases as metadata candidates, and the second step introduces various features, i.e., statistical features, linguistics features, position features and a novel fine-grained distribution feature which has high relevance with metadata categories, to type the candidates into three foregoing categories. In the evaluation, we conduct extensive experiments on a manually-labeled dataset from ACL Anthology and the results show our proposed method achieves a +3.2% improvement in accuracy compared with strong baseline methods.

Keywords: Metadata Extraction · Scientific Literature · Fine-grained Distribution · Classification

1 Introduction

As the number of scientific literature increases quickly, getting access to the core information of scientific papers easily and fast is becoming more and more important. With these core information, we can improve both the quality and efficiency of information retrieval, literature search engine and research trend prediction. In this paper, we aim at the mining of the core information of scientific, and refer to it as **metadata extraction**.

Metadata extraction is a complicated task and poses the following challenges:

- It is hard to determine whether an extracted item from a scientific paper is representative (i.e., belongs to the metadata) or not.
- To the best of our knowledge, there is not a public labeled dataset, even an effective and widely accepted annotation rules.

⋆ Corresponding author: Erhong Yang

– Although all scientific papers follow a common writing rule (e.g., organized as the motivation, background, innovation, contrast, solution and experimental results), the metadata may be flexible enough to appear in any section, making the metadata extraction very challenging.

Traditional metadata extraction[5] is to extract a set of controlled vocabularies with a fixed schema, which greatly depends on the hand-crafted extraction rules and lacks flexibility. Adit Krishnan et al.[7] extract scientific concepts via an automatic manner from paper titles only, which results in a large amount of loss of useful information.

In this paper, we divide metadata into 3 categories: Problem, Method and Dataset, which we believe can represent the main contents of one scientific together. We define the **metadata extraction** as the mining of key phrases belonging to these 3 categories. To extract these information, we construct a manual labeled dataset based on the ACL history data, and propose an supervised domain-specific framework.two

Our model can be divided into three phases. In Phase One, we feed the full paper context into Segphrase[9] to extract key phrases, named as metadata candidates. In Phase Two, we represent all metadata candidates using our proposed novel features, including semantic features and section-leveled tfidf features. In phrase Three, we feed the features into a classifier to predict which category they belong to (Problem, Method, Dataset and Not-Metadata).

To sum up, our contributions are as follows:

– We firstly put forward a pioneering task—metadata extraction—for scientific papers, which is very meaningful to many NLP applications.
– We construct a manual-labeled dataset for metadata extraction, hoping that it can push the development of related researches.
– Based on our dataset, we propose a three-phased supervised domain-specific framework to extract metadata.
– We conduct experiments on our dataset and demonstrate that our framework outperforms the baseline methods for metadata extraction of scientific papers.

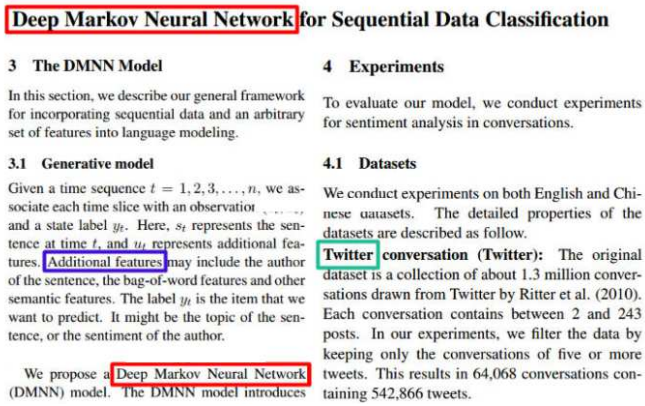
The rest of this paper is organized as follows. Section 2 presents problem formalization. In Section 3 we present our detailed approaches. The experimental results are introduced in Section 4. Section 5 reviews the related literature. Finally, we conclude our work in Section 6.

2 Problem Formalization

In this section, we formalize the problem of metadata extraction for scientific papers. Before that, we first introduce some related basic concepts.

Definition 1 (Scientific Paper). *Conceptually, a scientific paper is a report of intellectual work within several key integrant sections, including standardized argumentation structure, which varies slightly in different subjects. Formally,*

a paper p can be represented as a collection of sections $p = \{s_p, s_m, s_e, s_c\}$ with the subscript denoting problem, method, experiment, related work and conclusion respectively. Each section s is a word sequence $s = \langle w_1, w_2, \ldots, w_{|W|} \rangle$ with each word w chosen from the vocabulary V.



Deep Markov Neural Network for Sequential Data Classification

3 The DMNN Model

In this section, we describe our general framework for incorporating sequential data and an arbitrary set of features into language modeling.

3.1 Generative model

Given a time sequence $t = 1, 2, 3, \ldots, n$, we associate each time slice with an observation [illegible] and a state label y_t. Here, s_t represents the sentence at time t, and u_t represents additional features. Additional features may include the author of the sentence, the bag-of-word features and other semantic features. The label y_t is the item that we want to predict. It might be the topic of the sentence, or the sentiment of the author.

We propose a Deep Markov Neural Network (DMNN) model. The DMNN model introduces

4 Experiments

To evaluate our model, we conduct experiments for sentiment analysis in conversations.

4.1 Datasets

We conduct experiments on both English and Chinese datasets. The detailed properties of the datasets are described as follow.

Twitter conversation (Twitter): The original dataset is a collection of about 1.3 million conversations drawn from Twitter by Ritter et al. (2010). Each conversation contains between 2 and 243 posts. In our experiments, we filter the data by keeping only the conversations of five or more tweets. This results in 64,068 conversations containing 542,866 tweets.

Fig. 1. An example of part of scientific paper

For example, Figure 1 presents an example of scientific paper which contains method and experiment sections. Normally, a scientific paper aims at one or more research tasks, proposes specific methods, designs experiments for validation and achieves some conclusions, and we call these information as metadata.

Definition 2 (Metadata). *Metadata*[4] *is data that provides information about other data, and can be grouped for different purposes, including descriptive (e.g., title and author), structural (e.g., pages), and administrative (e.g., owner). Previously, metadata about scientific papers usually includes author, publisher, venue, title, abstract and so on. As mentioned above, we focus on more informative and fine-grained metadata in this paper. Formally, the metadata of paper p is described as a triple $md_p = (problem, method, dataset)$, and each element is composed of several continuous words, which can be also called key phrase.*

Also in Figure 1, "Deep Markov Neural Network" and "Twitter" denote the method and dataset of the example paper. Note that not all key phrases are metadata, e.g., "Additional features". Papers could have various kinds of metadata, and the above schema, i.e., $(problem, method, dataset)$, plays a very important role in understanding papers or inspiring follow-up research. Therefore, we propose the following problem:

[4] `https://en.wikipedia.org/wiki/Metadata`

Definition 3 (Metadata Extraction for Scientific Paper). *Given a collection of scientific papers $P = \{p_1, p_2, \ldots, p_n\}$, our goal is to extract the metadata of each paper, i.e., md_{p_i} for p_i. In particular, we formalize the task as a classification problem. For each paper p_i, we first separate it into disjoint sections $p_i = \{s_{p_i}, s_{m_i}, s_{e_i}, s_{c_i}\}$ and extract several key phrases $K_i = \{k_1, k_2, \ldots, k_m\}$ meanwhile. Then we characterize each key phrase k_i with well-designed syntax and structure features, and finally build a classifier to determine whether it is a kind of metadata (i.e., problem, method and dataset) or not.*

Note that our dataset is constructed from papers in PDF format, and section information is not totally identical to the original papers. Even in the original papers, the section organization is also various. Thus the section schema is manually defined as above.

3 The Proposed Approach

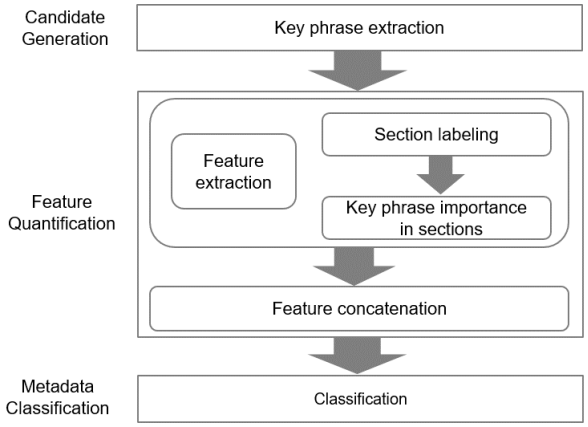


Fig. 2. The framework of the proposed approach

In this section, we describe the proposed method for scientific metadata extraction in details. Figure 2 shows the framework in a pipeline paradigm, involving three major steps:

- **Candidate Generation.** Perform a key phrase extractor on the paper collection P in which the abstract of each paper is excluded to get the key phrase set $K = \{k_1, k_2, \ldots, k_m\}$ for each paper $p \in P$, which serves as the candidate metadata mentions.
- **Feature Quantification.** For each key phrase, we characterize it with two types of features, basic features and fine-grained distribution features. The basic features focus on the statistics, position and syntax of the phrase, and

the fine-grained distribution features first group paper full text into several disjoint sections and evaluate the importance in different sections for each key phrase. Two types of features are concatenated to represent each key phrase for the following classification.
– **Metadata Classification.** Based on the above feature representation, build a classifier to classify extracted key phrases into one of the metadata categories, i.e., problem, method, dataset or not metadata(NOM).

In the above framework, key phrase extraction is to segment the paper text into a sequence of cohesive content units and select representative concept or process phrase as the candidate metadata mentions. In this paper, we emply a distant-supervised phrase segmentation algorithm named SegPhrase[9] with its released implementation[5], and other similar alternatives could also be applied. As for the final classifier, we try several typical classification models and decide to use the random forest (see the experiment part for details). Therefore, the core component of the framework is the design and quantification of different features, and we will demonstrate the details according to feature types, namely basic features and fine-grained distribution features.

3.1 Basic Features

From the perspective of the syntactic and grammar, the metadata phrase obviously and distinctively differs from other normal phrases. In this study, each key phrase is characterized by a set of features in terms of position, statistics and syntax are evaluated. Specifically, the following seven features are calculated for each key phrase $k \in K$:

– **(1) Frequency.** It is the number of times that key phrases appear within one paper, which is applied to evaluate the importance and possibility of a key phrase being metadata. It is based on a straightforward assumption that metadata is more frequently mentioned as compared with normal phrases.
– **(2) Length and (3) Max Word Length.** Intuitively, the length of phrase indicates the information it contains, and thus key phrase length and max word length are applied in this study. The number of characters of each key phrase is defined as key phrase length, and the length of longest word within key phrase is max word length.
– **(4) Leading Letter Capitalized.** Capitalizing the first letter of phrases is a typical and popular way in scientific paper writing to emphasize the importance of these words, highlight the authors points or ideas, and attract readers attention. If words in key phrases in accordance with this situation, such key phrases tends to be innovation and informative so that it is likely be related to be metadata. For instance, all the leading letters in each words of "Deep Markov Neural Network" are capitalized, which is labeled as "Method" in our dataset.

[5] `https://github.com/shangjingbo1226/SegPhrase`

- **(5) In Title and (6) In Abstract.** In general, the title and abstract are a condensation or summarization of a paper, which contains the most important and useful information. Based on above backgrounds, whether a key phrase appears in title or abstract reflects its significance for a paper, and thus is related to its probability of being metadata. For example, the metadata phrase "Deep Markov Neural Network" emerges in title of the paper.
- **(7) Lexical Cohesion.** Words in metadata phrases are usually consistent in lexicon, and thus we define lexical cohesion for quantification, which is computed as follows:

$$w_n \times ((1 + \lg f_k \frac{f_k}{\sum w_c}))$$

where w_n is the number of constituent word one key phrase contains, f_k is the frequencies of appearance in one paper for each key phrase, w_c is the frequencies of each constituent word of one key phrase in the whole paper content.

3.2 Fine-grained Distribution Features

In this subsection, we focus on the distribution of metadata phrases across the full paper content. We first use a manually labeled sample to demonstrate our assumption, based on which we introduce our proposed feature quantification.

Assumption 1 *Different types of metadata follow different distributions across the full paper content, and metadata and their more correlated sections tend to have consistent semantics, e.g., a key phrase labeled with "dataset" has higher probability of appearance in "experiment" and "introduction" as compared with "related work" and "problem formalization".*

To verify the above assumption, we randomly select 30 papers to statistics the distribution of key phrases in four predefined sections. Table 1 depicts the occurrences of several key phrases in different sections. These key phrases clearly show that the distribution of each key phrase among four sections is closely linked with its semantic meaning. Take "neural network" as an example, the highest occurrence located in section "Model" and not once in section "Experiment", which is in accordance with our expectation. As a contrast, an NLP task named "sarcasm detection" is referred as many as 11 times in section "Problem", highest among other sections. However, in Experiment section, experiment setting and experiment results will regularly be explained, and problem description commonly will not be mentioned. And the figure about "sarcasm detection" clearly supports this convention.

This assumption inspires us that the section-level distributions of candidate phrases are good indicators for metadata classification. To generate such fine-grained distribution features, we first segment a full paper several disjoint sections as defined in Section 2, and then quantify the importance of each candidate for all sections as the final feature vectors. In the following part, we will introduce the section segmentation and importance measurement.

Table 1. Examples on phrase distributions in different sections

Phrase \ Section	Conclusion	Experiment	Model	Problem
training set	0	20	1	0
neural network	1	0	11	7
hidden state	0	0	7	10
sarcasm detection	3	0	2	11
stochastic gradient descent	1	2	7	2

Section Segmentation. Scientific literature usually follows an acknowledged logical structure to organize contents (e.g., using sections, subsections). Each section has a different focus from others. To achieve our goal of metadata extraction, the problem, method and experiment sections are necessary. Besides, a paper normally includes introduction, related work and conclusion. All the six sections constitute the section schema. As mentioned after problem definition in Section 2, we need to group the paper content into the above sections due to the paper organization is usually various or even missing. To complete the segmentation, we apply a CRF-based parsing tool[13] to attach a section label for each sentence, and then group these sentences into sections. Note that the key phrases extracted by SegPhrase[9] are within one sentence not cross sentences, thus one key phrase mention corresponds to only one section label and key phrase mention with ambiguous section labels does not exist.

Importance Measurement. Various methods can be applied to measure the importance of a given key phrase in different sections. In this paper, we employ the most classical tf-idf measurement. Particularly, we treat all the sentences with the same section labels as a virtual documents, and a paper could be split into at most six documents, based on which the tf-idf is calculated. Note that the importance value is set be 0 if a section does not contain any sentence.

4 Experiment

In this section, we evaluate the proposed framework on a manually created dataset. We first briefly introduce the dataset and experiment settings, then present the detailed experimental results, and finally investigate the feature contributions and some other method details.

4.1 Dataset

To the best of our knowledge, there is no existing benchmark dataset which can be directly used for our metadata extraction evaluation. As such, we crawled the papers in PDF format from ACL Anthology[6], and employed GROBID tool[7] to

[6] `http://aclweb.org/anthology/`
[7] `http://grobid.readthedocs.io/en/latest/`

extract the textual content. Then we randomly selected 30 papers, from which 1,947 key phrases were extracted. Finally, we invited graduated students to manually grouped them into predefined metadata schema as defined in Section 2, i.e., "problem","method","dataset" and NOM. They were required to firstly read the corresponding paper to understand the "problem" and "method". After annotation, over 75% of the phrases (i.e., 1,478 key phrases) were labeled as NOM, leading to an unbalanced classification problem. This results make sense since a paper normally does not contain too many metadata. Note that key phrase labeled with NOM was excluded in experiments to keep balance among the remaining three groups. Detailed statistics are presented in Table 2.

Table 2. Dataset statistics

Metadata	**Problem**	**Method**	**Dataset**	**NOM**	**Total**
Numbers	253	165	46	1483	1947

4.2 Experiment Setting

Baseline Methods. We use **MESP** to denote the proposed method of Metadata Extraction for Scientific Papers, and compare it with the following baseline methods to verity its effectiveness:

- **Section Importance Ranking (SIR):** An unsupervised method that treats the most important key phrases in corresponding sections as results. In experiment, we use the section importance introduced in Section 3.2 to rank the key phrases in *problem*, *method* and *experiment* sections, and select the top 5 as results.
- **MESP-B:** The simplified version of our proposed method that only uses basic features in Section 3.1.
- **MESP-I:** The simplified version of our proposed method that only uses the section importance features and in Section 3.2.
- **MESP-P:** The fusion version of our proposed method that uses the section importance features in Section 3.1 and the fine-grained distribution features in Section3.2.

Comparison Metrics. To quickly have a basic knowledge about one scientific paper, only a few key phrases identified as correct metadata are better than many key phrases considered to be metadata, of which only several key phrases are accurate metadata. Therefore, "accuracy" is a most important and effective indicator than others, which is the comparison metric in this experiment. Let $\hat{y}_i$, y_i denote the predicted and corresponding true values of the i-th sample, then the accuracy over n samples is defined as

$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{I}(\hat{y}_i = y_i)$$

where $\mathbb{I}(x)$ is the indicator function.

Implementation Details. To find the most appropriate classifier, we tried six typical methods, i.e., Logistic Regression, Nearest Neighbors, Decision Tree, Nave Bayes, Adaboost, and Random Forest. In experiment, we used their python implementation in sklearn[12], and searched the best hyper-parameter space via gridsearhcv tool[12] based on standard train/test split and 5-fold cross validation.

4.3 Performance Results

As depicted in Table 3, Random Forest achieves the best performance among all classifiers previously mentioned. Random Forest is an ensemble of several weak learners, which can make full use of the data samples and the features of each sample while training. It implies that metadata extraction is a non-trivial problem which requires extracting different features in diverse ways instead. Besides, metadata phrases are located much closer with each other than NOM phrases in vector space, and that is why Nearest Neighbor classifier achieves good performance.

Table 3. Performance of various classifiers

Classifier	Accuracy
Nave Bayes	51.80%
Random Forest	67.63%
Nearest Neighbor	63.36%
Adaboost Classifier	62.28%
Decision Tree	54.61%
Linear Regression	57.54%

To validate the effectiveness of the proposed framework and designed features, we compare **MESP** with baseline methods using Random Forest and summarize the results in Table 4. From the results, we can see that **MESP** performs best among other baseline methods and achieves a considerable improvement of +3% over the strongest baseline in the accuracy because it is fully equipped with all the grammar, statistics, position and distribution features.The result of **MESP-B** demonstrates that the basic features are fairly good for the extraction task. **MESP** can further improves the performance, which verifies that section importance is a good feature that describes the distribution of key phrases. However, metadata extraction for scientific paper is such a complicated problem which requires much more other information from different perspectives, and this is why **SIR**, which distinguishes metadata merely by section importance value, performs poorest among all methods.Besides,the huge gap of results between **SIR** and other methods demonstrates that superviesed methods are more effective than un-supervised ones in this task.

Table 4. Performance comparison with baselines

Method	SIR	MESP-B	MESP-I	MESP
Accuracy	10.89%	64.44%	58.58%	67.63%

5 Related Work

Two lines of researches are related to the metadata extraction in the current paper, i.e., the definition of metadata and keyword extraction for scientific papers.

5.1 The Definition of Metadata

Metadata could be simply interpreted as *data about data*, and is used for providing information about other data. For metadata definition of scientific literature, the most widely accepted standard is *the Dublin Core*. It defines 15 elements[8] for resource description. From its scope, we can see that it is only used for building unified services for digital libraries and does not understand the deeper semantics within paper content. The metadata in this study is entirely different from the Dublin Core standard. Our metadata definition aims to understand the problem, method and experiment of scientific papers, and could be used for tasks that requires deep semantics, such as "hot ideas" detection in a scientific field [2, 4] and literature summarization [1]. For a similar goal, Simone Teufel et al. propose Argumentative Zoning (AZ) based on scientific papers, which tags the rhetorical structure of scientific literature on sentence level [14], and they further extend the AZ scheme to 15 fine-grained categories for annotation [15]. However, they do not provide any corresponding annotation tools, and manually labeling costs a lot of time and human resources. Therefore, automatic metadata extraction is necessary.

5.2 Keyword Extraction for Scientific Papers

Keyword extraction is to extract representative words that occur frequently in texts while other semantic words apart from stop words seldom occurring. Due to its significance to many downstream tasks, the problem attracts numerous research attention, e.g., Liu et al. [9] and Su Nam Kim et al. [7] study the extraction of phrases that capture the main topics discussed in a document from large corpus. However, there is normally no categorization for these keywords during the extraction.

In recent years, many researches turn their attention to the scientific literature. Li et al. extracts keywords in experiment-related graphs for chemistry metadata extraction [8]. Sonal Gupta et al. using dependency parsers to extract key phrases for three focus related categories from each sentence for the

[8] *The metadata of papers in the Dublin Core include: Title, Creator, Subject, Description, Contributor, Publisher, Date, Type, Format, Identifier, Source, Relation, Reference, Is referenced By, Language, Rights and Coverage*

analysis of dynamics of research focus of scientific papers [3]. Pan et al. employs a widely-used linguistic pattern introduced by Justeson and Katz [6], i.e., $((A|N)^+|((A|N)^*(NP)?)(A|N)^*))N$, to extract terminology concepts [11]. However, academic key phrases are more ambiguous and hard to type based on frequencies. Chen-Tse Tsai et al. uses noun-phrase chunking to extract concept mentions and local textual features and annotating concept mentions iteratively for the identifying scientific concepts [16], but the extracted concepts are too coarse so that they cannot satisfy the fine-grained demands in some analysis tasks. Adit Krishnan et al. introduces an unsupervised method for extraction of representative concepts from scientific literature based on titles scientific paper [7]. However, titles are not sufficient to provide fine-grained metadata for key phrase extraction and categorization[10], and thus we consider the full text of scientific paper for the metadata extraction.

Overall, the problem of metadata extraction for scientific papers is a non-trivial task due to it heavily depends on latent semantic information. Therefore, we address this task in a three-stage framework as described in Section 3. The metadata candidates are generated using a widely-accepted tool [9], and our focus is mainly on how to type the extracted metadata. Particularly, we design a novel fine-grained distribution feature based on preliminary data observation, whose effectiveness is verified in the experiments.

6 Conclusion

In this paper, we put forward a novel metadata extraction task on scientific papers and propose a three-stage framework for solution. In particular, we introduce an especially important fine-grained distribution feature for metadata typing. Besides, we perform extensive evaluation on a manually labeled dataset to validate the effectiveness of the proposed framework as well as the designed features. The results show that the performance of our method outperforms the strong baselines by an average accuracy of 3.2%. Our future work is to studying other representative and essential metadata for scientific papers and extending a discipline-independent extractor.

Acknowledgement This research project is supported by the Major Project of the National Language Committee of the 13rd Five-Year Research Plan in 2016 (ZDI135-3); supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University(17YCX148).

References

1. DAvanzo, E., Magnini, B.: A keyphrase-based approach to summarization: the lake system at duc-2005. In: Proceedings of DUC (2005)
2. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101**(suppl 1), 5228–5235 (2004)

3. Gupta, S., Manning, C.: Analyzing the dynamics of research by extracting key aspects of scientific papers. In: Proceedings of 5th international joint conference on natural language processing. pp. 1–9 (2011)
4. Hall, D., Jurafsky, D., Manning, C.D.: Studying the history of ideas using topic models. In: Proceedings of the conference on empirical methods in natural language processing. pp. 363–371. Association for Computational Linguistics (2008)
5. Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic document metadata extraction using support vector machines. In: Digital Libraries, 2003. Proceedings. 2003 Joint Conference on. pp. 37–48. IEEE (2003)
6. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering **1**(1), 9–27 (1995)
7. Krishnan, A., Sankar, A., Zhi, S., Han, J.: Unsupervised concept categorization and extraction from scientific document titles. CoRR **abs/1710.02271** (2017)
8. Li, N., Zhu, L., Mitra, P., Mueller, K., Poweleit, E., Giles, C.L.: orechem chemxseer: a semantic digital library for chemistry. In: Proceedings of the 10th annual joint conference on Digital libraries. pp. 245–254. ACM (2010)
9. Liu, J., Shang, J., Wang, C., Ren, X., Han, J.: Mining quality phrases from massive text corpora. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1729–1744. ACM (2015)
10. McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., Biran, O., Bothe, S., Collins, M., Fleischmann, K.R., et al.: Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology **67**(11), 2684–2696 (2016)
11. Pan, L., Wang, X., Li, C., Li, J., Tang, J.: Course concept extraction in moocs via embedding-based graph propagation. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). vol. 1, pp. 875–884 (2017)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
13. Prabhakaran, V., Hamilton, W.L., McFarland, D., Jurafsky, D.: Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1170–1180 (2016)
14. Teufel, S., Carletta, J., Moens, M.: An annotation scheme for discourse-level argumentation in research articles. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. pp. 110–117. Association for Computational Linguistics (1999)
15. Teufel, S., Siddharthan, A., Batchelor, C.: Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. pp. 1493–1502. Association for Computational Linguistics (2009)
16. Tsai, C.T., Kundu, G., Roth, D.: Concept-based analysis of scientific literature. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 1733–1738. ACM (2013)