Automatic identification and classification of integrated knowledge content in an interdisciplinary field: A case study on eHealth

Shiyun Wang[1[0000-0002-4409-7751]], Jin Mao[1,2,* [0000-0001-9572-6709]] and Hao Xie[1[0000-0002-1788-0468]]

[1] Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China
[2] School of Information Management, Wuhan University, Wuhan 430072, China
*Corresponding author. Email: danveno@163.com

Abstract. Investigation of knowledge integration characteristics, especially from the content perspective, is essential for understanding the formation and evolution of interdisciplinary fields. However, in previous studies, it involves considerable time and effort for researchers to recognize the integrated knowledge content in an interdisciplinary field and to analyze the content characteristics. Therefore, we have studied the automatic methods to identify the explicit integrated knowledge phrases from citation contexts in interdisciplinary field papers and recognize the functions of integrated knowledge phrases by utilizing word embedding techniques and deep learning models. To evaluate the performance of our methodology, we constructed an experimental dataset by taking the eHealth field as a case of interdisciplinary field. From the experimental results, we obtained Recall, Precision and F1 scores of 0.838, 0.989 and 0.907 for the explicit integrated knowledge identification process, and Recall, Precision and F1 scores of 0.856, 0.863 and 0.842 in the unknown phrases test dataset in knowledge functions classification.

Keywords: Knowledge Integration, Interdisciplinary Field, Semantic Function Recognition, Deep Learning.

1 Introduction

Interdisciplinary research is often considered as an important driver in modern science [1]. The essence of interdisciplinary research is successful recombination of existing disconnected knowledge units from various disciplines [2], which may lead to novel ideas and accelerate scientific breakthroughs. The increasing number of emerging interdisciplinary fields demonstrate that interdisciplinary research has become an important mode in science.

Scientists and policy makers have attempted to explore the characteristics of interdisciplinary research to promote the development of interdisciplinary research. Several bibliometric indicators, e.g., Rao-Stirling [3], have been proposed to measure the interdisciplinarity of research domains or publications. Most studies used citation analysis to investigate knowledge diffusion relations between an interdisciplinary field and its

source disciplines through references [4-6]. However, these studies only measure knowledge dissemination at the paper and journal level, rather than from the perspective of knowledge units. A few recent studies have explored knowledge integration and evolution of an interdisciplinary field from the content perspective to understand the formation and development of an interdisciplinary field [1,7-9]. Nonetheless, the identification of integrated knowledge content involved considerable human efforts in these studies. To foster the subsequent analysis and knowledge mining at a large scale, automatically identifying integrated knowledge content is in great demand.

In this study, we investigate the integrated knowledge content by an interdisciplinary field. We applied NLP techniques to automatically identify the integrated knowledge units from citation sentences and the texts of reference publications. And, we classified the functions of integrated knowledge through deep learning models. An experimental dataset of the eHealth field was constructed to validate the effectiveness of our methodology.

2 Methodology

2.1 Definitions and Problem Formulation

Scientific publications record various forms of knowledge integration, e.g., the cooperation of researchers and the citations to the references of different disciplines. In this article, we investigate the knowledge integrated in an interdisciplinary field, which can be reflected through citation relations. To this end, we propose integrated knowledge phrases and a classification scheme based on knowledge functions, which are defined as follows.

Integrated Knowledge Phrases. Citation contexts often express relevant information about cited articles [10]. To reflect the integrated knowledge units, we use the noun phrases extracted from citation sentences that also appear in the text of the corresponding cited publications. We contend that the shared phrases between the two counterparts explicitly signify the transferred knowledge.

Knowledge Classification Schema. Several classification frameworks have been proposed to annotate semantic functions of concepts or terms in scientific papers [11-12]. However, the classes in these frameworks, e.g., problems, solutions, and goals, are too general to analyze knowledge integration in a domain at a fine-grained level. In our previous study, we proposed a knowledge classification schema for the annotation of integrated knowledge content, which comprises seven categories, including *Research Subject*, *Theory*, *Research Methodology*, *Technology*, *Entity*, *Data*, and *Others* [9]. *Research Subject* is broader than other categories, which covers multiple kinds of domain-specific research subjects, e.g., diseases, drugs, and research themes. The identification of domain entities have become significant tasks in many recent studies, e.g., Biomedical NER task [13], and some tools have been developed, e.g., PubTator Central [14]. Therefore, we do not involve the *Research Subject* category, as well as the *Others* category. The final framework in this study is shown in Table 1.

Table 1. The classification framework of integrated knowledge phrases.

Category	Description	Exemplar phrases
Theory	Theory related phrases	e.g., *TAM*, *social cognitive theory*, *trans-theoretical model*
Research methodology	Methodology used in research	e.g., *systematic review*, *analysis*, *meta analysis*, *randomized control trial*
Technology	Technique, device and system that used in research	e.g., *mobile phone*, *web*, *smartphone*, *app*
Entity	Human related research object	e.g., *patient*, *woman*, *child*, *adolescent*
Data	Phrases related to dataset, data source and data material	e.g., *twitter*, *qualitative datum*, *clinical datum*

The research design of this study is summarized in Fig. 1. The task of integrated knowledge phrases identification includes knowledge phrases extraction from citation sentences and corresponding reference texts, and knowledge phrases matching between the two sources. For the task of classifying the functions of integrated knowledge phrases, several deep learning models are adopted.

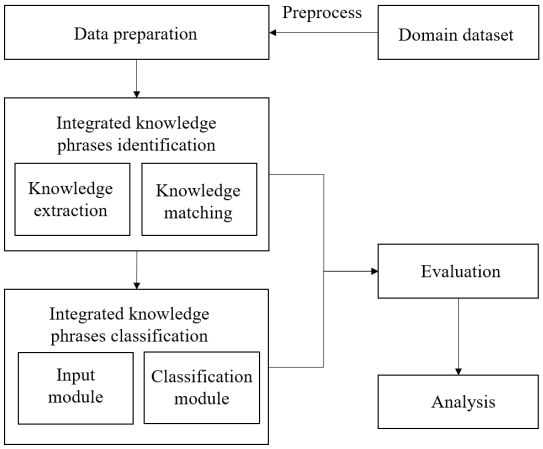

Fig. 1. Research framework of this study

2.2 Dataset Construction

We collected full texts of papers in an interdisciplinary field. For each paper, citation sentences and bibliography data (title, PMID, etc.) were extracted and linked via the in-text citation tags in the text, e.g., "*[1]*, *[2-8]*". Next, we complemented the metadata of the references, e.g., titles and abstracts, as the cited texts. Then, each in-text citation generates a pair of citation-reference. An in-text citation of several references was split

into multiple citation-reference pairs. The section titles of citation sentences were also fetched. The citation-reference pairs records are constituted as the initial dataset for the following processes.

2.3 Integrated Knowledge Phrases Identification

Knowledge Phrases Extraction. For each citation-reference pair, we used spaCy, an open-source Python natural language processing toolkit to extract noun phrases in the citation sentences and reference texts. We only selected phrases with 2 to 4 words rather than retained all the phrases with less than 7 words as in our previous study [9]. We also removed the phrases started or ended with numbers and the phrases with single characters. Moreover, scispaCy [15], a Python package for processing biomedical scientific text was applied to expand abbreviations in the text.

Knowledge Phrases Matching. The extracted knowledge phrases from the two sources were lemmatized and stemmed using the NLTK package before matching, so that different variations of the same word could be matched correctly. Next, we used a combination of three approaches to match the phrases from citation sentences with those from the corresponding references for each citation-reference pair.

Direct Matching. This approach only counts the identical knowledge phrases from the two sources as matched phrases.

Indirect Matching. In this process, the extracted knowledge phrases from the citation sentences will be exactly matched with the sentences in corresponding reference text using regular expressions, and vice versa. This approach could identify those phrases with the same meaning but with different collocations. For example, “focus group” and “focus group method” could be matched through this method.
The above two approaches were combined as the baseline method for the knowledge matching process. We further applied a phrase similarity calculation approach based on word embedding techniques to identify those phrases with similar meaning but are represented in different word collocations.

Word Embedding + Cosine Similarity. Word embedding technique was utilized to transform phrases into high dimensional vectors, and then cosine similarity of phrase vectors was calculated. We selected two word embedding models, GloVe and BERT (Bidirectional Encoder Representations from Transformers). In short, GloVe and BERT are both language representation models which can be used to vectorize the words. Word vectors involve rich semantic and contextual information of the words in the training corpus. However, compared to conventional word embedding models, such as GloVe, BERT introduces position encoding to describe sequence position information, and takes the jointly left and right contexts for each occurrence of a given word into account, which could capture more contextual information. In this paper, we used a 100 dimensions GloVe model [16] that was pre-trained on the dataset of Wikipedia

2014 and Gigaword 5, and a 12 layers, 768 hiddens BERT base model [17] pre-trained on Wikipedia and BookCorpus, a corpus with 11,038 unpublished books.

Then, we integrated the matched noun phrases identified by the above three approaches, and removed the deduplicated phrases. The retained phrases were denoted as the integrated knowledge phrases.

2.4 Integrated Knowledge Phrases Classification

We applied several deep learning models to classify integrated knowledge phrases, which consists of the following major modules:

Input Module. For the input module, we considered several contextual information of the phrases, which could be divided into semantic information and syntactic information.

Semantic Information. Both citation sentences and corresponding reference texts contain the semantic contextual information of phrases. Therefore, we included both of them as the input text features.

Syntactic Information. Different sections in the scientific text may have different functions [18]. For example, the Introduction section describes more information about the background of the study, while the Methods section depicts the methods applied. The section title of the citation sentences where the integrated knowledge phrases occur may cover some useful contextual information for the knowledge classification.
For each integrated knowledge phrase of each citation-reference pair, we combined the three features, i.e., the citation sentence, the corresponding reference text, and citation section title as the contextual information field of the integrated knowledge phrase, and spliced it with the phrase as the input sequence of the deep learning models.

Classification Module. We applied five deep learning models, including LSTM, TextCNN, BERT, BERT+SVM, and BERT+XGBoost for the phrase classification task.

LSTM. LSTM (Long-Short Term Memory) [19] is a type of Recurrent Neural Network (RNN), but designed to model chronological sequences and their long-range dependencies more precisely than conventional RNNs. Therefore, it could provide more long-distance contextual information. The contextual information and integrated knowledge phrases were embedded respectively in the embedding layers by using the GloVe model in this paper.

TextCNN. TextCNN [20] is a text classification technique using the Convolutional Neural Network(CNN). CNN is a kind of artificial neural network, in which the output of each layer is used as the input of the next layer of the neuron. Generally, it includes

four parts, including embedding layer, convolutional layer, pooling layer, and fully connected layer. The input layer of the model is the same as the LSTM model.

BERT. BERT [17] is not only a pre-trained language representation model, but can also act as a classifier, if we fine-tuned the model with just one additional output layer. For each integrated knowledge phrase, the text sequence with contextual information and integrated knowledge phrase was fed to the BERT model, and was embedded into vectors in the embedding layers, including token embedding, segment embedding, and position embedding. Then, these embeddings were spliced in the fully connected layer, and fed to the output layer, which acted as a classifier for predicting the function label of integrated knowledge phrase.

BERT + SVM. SVM (Support Vector Machine) [21] is a kind of generalized linear classifier, which determines the best decision boundary between vectors that belong to a category and that do not belong to it. In this model, we used the BERT to extract the pre-trained embeddings of our text features and fed them to the SVM classifier.

BERT + XGBoost. XGBoost [22] represents "Extreme Gradient Boosting", which is an implementation of gradient boosted decision trees. It is designed to push the limit of computations resources for boosted tree algorithms and has achieved great performance and speed in applied machine learning recently. We also used BERT to vectorize the input text sequence in this model.

2.5 Evaluation

We chose precision, recall and F1 score to measure the performance of the integrated knowledge phrases identification and classification methods.
For the identification task, precision is calculated as the number of correctly identified phrases divided by the total number of phrases identified through the automatic methods. Recall value is calculated by the number of correctly identified phrases divided by the number of phrases should be identified through the annotation.
In the classification task, we used weighted precision and recall scores of all the categories. For each category, precision is calculated as the number of correctly labelled integrated knowledge phrases by the total number of integrated knowledge phrases of this category the classified models recognized, and recall is calculated as the number of correctly labelled integrated knowledge phrases of this category by the number of phrases that should be labelled as this category. For the overall precision and recall score of the test dataset, we assign weight to the precision and recall value of each category by the proportion of phrases in each category, and then plus all the weighted precision scores and weighted recall scores of all categories.
Finally, each F1 score is twice of the multiplication value of precision and recall score divided by the sum of the two values.

3 Experiments and Results

3.1 Experimental Datasets

We constructed an eHealth dataset to test the performance of our methodology framework. XML files of 3,221 eHealth papers published from 1999 to 2018 were downloaded from two high impact journals, *Journal of Medical Internet Research* and *JMIR mHealth and uHealth*. The metadata of the references were complemented from Web of Science (WoS) and PubMed. Overall, we obtained 199,461 citation-reference pairs. Two datasets were constructed for the two tasks in our methodology. For the identification procedure, we randomly selected 100 citation-reference pairs to manually annotate the matched phrases and obtained 105 matched phrases in total. For the classification task, we selected 45,166 matched phrases identified in our previous study [9], which were labelled as "Research Methodology", "Technology", "Entity", "Data", and "Theory". It was randomly divided into ten folds, eight of them for training, which contains 36,133 phrases; one for validation, containing 4,517 phrases; and the remaining one, including 4,516 phrases, for test.

3.2 Results of Integrated Knowledge Phrases Identification

We chose the combination of direct and indirect matching approach, which was used in the previous study [9], as the baseline method. In this paper, we applied a new approach, *Word Embedding + Cosine Similarity,* for improvement. Two word embedding models, GloVe and BERT, were compared. The threshold of cosine similarity was tuning among 0.7, 0.8, and 0.9. The evaluation results in Table 2 show that the baseline approach of our method already has a good performance, with precision of 1.0 and F1 score of 0.900. However, the recall value is relatively low. Regarding the new approaches, all of them obtained a higher recall value than the baseline method. This means that we could recognize more integrated knowledge phrases with the new approach, which is more effective for the knowledge integration analysis. We observe that with the increase of the threshold of cosine similarity, the precision of the method is rising, while the recall value is decreasing. To comprehensively measure the performance of our new method, we further calculated the F1 score. It demonstrates that the Baseline + BERT approach with the 0.9 cosine similarity threshold has the greatest performance, with F1 value of 0.907.

Table 2. Identification results on the standard dataset.

Methods	Recall	Precision	F1
Baseline	0.819	1.000	0.900
Baseline + GloVe (0.7)	0.857	0.823	0.840
Baseline + GloVe (0.8)	0.857	0.836	0.846
Baseline + GloVe (0.9)	0.848	0.948	0.895
Baseline + BERT (0.7)	0.867	0.715	0.784
Baseline + BERT (0.8)	0.848	0.919	0.882
Baseline + BERT (0.9)	0.838	0.989	0.907

3.3 Results of Phrases Classification

Five models were trained in the training dataset, and the hyperparameters of the trained models were tuned in the validation dataset. Then, the test dataset was used for the evaluation. We calculated the weighted indicators on the overall test dataset as well as on the unknown phrases dataset. The unknown phrases dataset contains 291 phrases in the test dataset but neither occur in the training dataset nor in the validation dataset. We considered these unknown phrases would measure the generalization ability of the models better than the whole test dataset since some phrases in the test dataset have occurred in the training and validation dataset and the models may have remembered the features of these phrases, although with different contextual information.
As shown in Table 3, the models with BERT embedding, i.e., BERT, BERT+SVM and BERT+XGBoost, appeared more effective than other deep learning models, i.e., LSTM and TextCNN. The BERT model itself already has a powerful semantic understanding and syntax analysis capabilities, which achieved the greatest performance on the overall test dataset, with a precision of 0.980, a recall of 0.980 and a F1 score of 0.980. As for the unknown phrases dataset, BERT+XGBoost is the best model. This result reflects the high efficiency, flexibility and portability of the XGBoost algorithm.

Table 3. Classification results on the test dataset.

Methods	Test Dataset	Recall (weighted)	Precision (weighted)	F1 (weighted)
LSTM	Overall phrases	0.851	0.874	0.849
	Unknown phrases	0.649	0.669	0.647
TextCNN	Overall phrases	0.956	0.957	0.956
	Unknown phrases	0.797	0.813	0.795
BERT	Overall phrases	0.980	0.980	0.980
	Unknown phrases	0.811	0.811	0.810
BERT+SVM	Overall phrases	0.971	0.972	0.971
	Unknown phrases	0.842	0.838	0.839
BERT+XGBoost	Overall phrases	0.973	0.973	0.972
	Unknown phrases	0.856	0.863	0.842

4 Conclusion

In this paper, we provided a new methodology to automatically identify the explicit integrated knowledge phrases from citation contexts in interdisciplinary field papers through word embedding techniques, and then utilize several deep learning models to classify the functions of the integrated knowledge phrases. The eHealth field was taken as a case of interdisciplinary field to evaluate the performance of our methodology. The results show that BERT has a great performance, not only as a pre-trained language representation model but also as a classifier. In the integrated knowledge phrases identification process, it obtained Recall, Precision and F1 scores of 0.838, 0.989 and 0.907

respectively when combined with the baseline string match method. Meanwhile, it achieved the weighted Recall, Precision and F1 scores of 0.980, 0.980 and 0.980 on the overall test dataset in the classification task. Moreover, when utilized the XGBoost algorithm along with the BERT model, the model achieved the greatest performance on the unknown phrases dataset, with weighted Recall, Precision and F1 scores of 0.856, 0.863 and 0.842.

In general, this paper is one of the primary works to apply the word embedding techniques and deep learning models in the integrated knowledge phrases identification and classification of an interdisciplinary field. This automatic methodology would contribute to the deep investigation of knowledge integration in an interdisciplinary field from the content perspective. In addition, it could be applied to the knowledge interaction exploration between any source and target publications.

However, there are also some limitations in this study. First, in the integrated knowledge phrases identification step, we only considered the knowledge explicitly integrated from the references to the citation sentences, but did not include other knowledge integration forms to comprehensively investigate the knowledge integration of an interdisciplinary field. Second, the cited reference texts in this article were represented by the metadata of the references rather than cited texts identified in the full text of references. The integrated knowledge not contained in the metadata of references may be lost. Finally, the dataset we used just covers two journals in the eHealth field, more annotation datasets from various disciplines are further needed to test the performance of our methodology.

References

1. Ba, Z., Cao, Y., Mao, J., & Li, G.: A hierarchical approach to analyzing knowledge integration between two fields—a case study on medical informatics and computer science. Scientometrics 119(3), 1455-1486 (2019).
2. Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Barabási, A. L.: Science of science. Science 359(6379), eaao0185 (2018).
3. Leydesdorff, L., & Rafols, I.: Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. Journal of Informetrics 5(1), 87–100 (2011).
4. Borgman, C.L., & Rice, R.E.: The convergence of information science and communication: A bibliometric analysis. Journal of the American Society for Information Science 43(6), 397–411 (1992).
5. Chang, Y. W., & Huang, M. H.: A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. Journal of the American Society for Information Science and Technology, 63(1), 22-33 (2012).
6. Leydesdorff, L., & Probst, C.: The delineation of an interdisciplinary specialty in terms of a journal set: The case of communication studies. Journal of the American Society for Information Science and Technology 60(8), 1709–1718 (2009).
7. Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L: Understanding the formation of interdisciplinary research from the perspective of keyword evolution: A case study on joint attention. Scientometrics 117(2), 973-995 (2018).

8. Engerer, V.: Exploring interdisciplinary relationships between linguistics and information retrieval from the 1960s to today. Journal of the Association for Information Science and Technology 68(3), 660-680 (2017).
9. Mao, J., Wang, S., & Shang, X.: Investigating interdisciplinary knowledge flow from the content perspective of citances. In: EEKE@JCDL 2020, pp. 40-44 (2020).
10. Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D.: Blind men and elephants: What do citation summaries tell us about a research article? Journal of the American Society for Information Science and Technology 59(1), 51–62 (2008).
11. Kondo T, Nanba H, Takezawa T, et al.: Technical trend analysis by analyzing research papers' titles. In: Proceedings of the Language and Technology Conference on Human Language Technology, Challenges for Computer Science and Linguistics, pp. 512-521. Springer, Heidelberg (2009).
12. Heffernan K, Teufel S.: Identifying problems and solutions in scientific text. Scientometrics 116(2), 1367-1382 (2018).
13. Zhao, S., Su, C., Lu, Z., & Wang, F.: Recent advances in biomedical literature mining. Briefings in Bioinformatics, bbaa057 (2020).
14. Wei, C. H., Allot, A., Leaman, R., & Lu, Z.: PubTator central: automated concept annotation for biomedical full text articles. Nucleic acids research 47(W1), W587-W593 (2019).
15. Neumann, M., King, D., Beltagy, I., & Ammar, W.: Scispacy: Fast and robust models for biomedical natural language processing. arXiv preprint arXiv: 1902.07669 (2019).
16. Pennington, J., Socher, R., & Manning, C. D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543(2014).
17. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
18. Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V.: The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology 67(1), 164-177(2016).
19. Hochreiter, S., & Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735-1780 (1997).
20. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv: 1408.5882 (2014).
21. Cortes, C., & Vapnik, V.: Support-vector networks. Machine learning 20(3), 273-297 (1995).
22. Chen, T., & Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, New York, USA (2016).