Generating Topic Pages for Scientific Concepts Using Scientific Publications

Hosein Azarbonyad, Zubair Afzal, and George Tsatsaronis

Elsevier, The Netherlands
{h.azarbonyad, zubair.afzal, g.tsatsaronis}@elsevier.com

Abstract. In this paper, we describe *Topic Pages*, an inventory of scientific concepts and information around them extracted from a large collection of scientific books and journals. The main aim of *Topic Pages* is to provide all the necessary information to the readers to understand scientific concepts they come across while reading scholarly content in any scientific domain. *Topic Pages* are a collection of automatically generated information pages using NLP and ML, each corresponding to a scientific concept. Each page contains three pieces of information: a definition, related concepts, and the most relevant snippets, all extracted from scientific peer-reviewed publications. In this paper, we discuss the details of different components to extract each of these elements. The collection of pages in production contains over 360,000 *Topic Pages* across 20 different scientific domains with an average of 23 million unique visits per month, constituting it a popular source for scientific information.

Keywords: Scientific document processing · Definition extraction · Multi-document summarization.

1 Introduction

Technical terminology is an important piece of scientific publications [6,7]. Scientists and researchers use technical terminology and concepts to convey information concisely. As a result, there is an overwhelming and growing number of scientific concepts in any scientific domain, adding to the difficulties scientists have to catch up with the ever-growing list of technical concepts and new content. Knowledge sources such as *Wikipedia* can provide useful information on technical and scientific concepts to a large extent, however, due to their *"wisdom-of-crowds"* creation method there are many omissions and errors, and they may not always be a trustworthy source to understand and refer to a scientific concept. Our *Topic Pages*[1] proposition creates a knowledge source in a *"wisdom-of-experts"* fashion, as the information on scientific concepts is extracted from iconic scientific books in the domain, or from high-impact peer-reviewed scientific publications on the topic.

Each *Topic Page* is centered around one scientific concept and contains a definition for the concept, a set of related concepts, and a set of relevant snippets

---

[1] https://www.elsevier.com/solutions/sciencedirect/topics

Fig. 1. An example *Topic Page* presenting the concept "regression analysis"', with a definition, related concepts, and a set of relevant snippets extracted from articles and books.

all extracted from scientific peer-reviewed articles and books. The definition comprises one sentence extracted from books and journals that provides a brief, yet concise, description of the concept. Snippets are text excerpts from books or journals, relevant to the concept, and provide contextual information about the concept. Related concepts are a set of most relevant concepts to the given concept that can help users to explore the relevant terminology around their concept of interest.

The collection contains over $360,000$ *Topic Pages* in 20 different scientific domains. These topic pages are hyperlinked from publications in ScienceDirect[2], which is one of the largest scientific publication search engines and databases containing over 18 million full-text articles, helping users to navigate to the corresponding *Topic Page* when they encounter an unfamiliar scientific concept in an article with just one click. There are over 5.8 million articles that provide hyperlinks that we have created from scientific articles to topic pages. *Topic Pages* attract over 23 million unique visits per month.

In the remainder of the paper, we briefly review related work in Section 2, we describe the technical pipeline for generating *Topic Pages* in Section 3, we evaluate empirically the most challenging module of the pipeline, which is the definition extraction, in Section 4 and we conclude in Section 5 by arraying some limitations of the current technical solution and provide pointers to future work.

2   Related Work

To the best of our knowledge, there is no similar solution to the one introduced in this paper for automatically generating topic pages for scientific concepts. Most of the related work falls under the definition extraction task, and this is where we put the focus in this section. Early work on definition extraction task
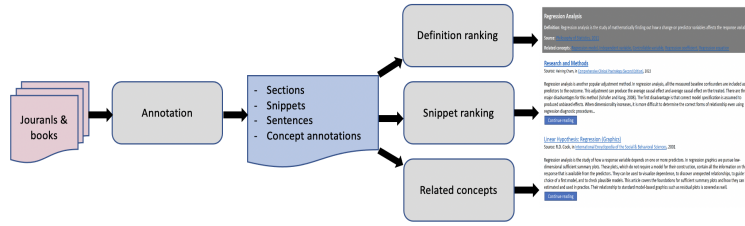
---

[2] https://www.sciencedirect.com/

Fig. 2. An overview of the topic pages generation pipeline including all essential components.

was focused on rule-based and pattern-matching approaches [3,8,20], often resulting in low recall given their limited coverage. Supervised models have also been proposed and shown to be more effective than the rule-based methods for this task [6,14,15,16,9]. These models use statistical information regarding concepts, as well as structural information of the sentences such as part of speech (POS) tags to distinguish definitional from non-definitional sentences. More recent work for definition extraction focused on using neural models for the task [11,2,5,18,7,13,19]. Notably, $LSTM$ [11] and a combination of CNN and $LSTM$ [2] have been used to learn the structure of definitional sentences. In our work, we also introduce and use a combined $LSTM+CNN$ model but, different from [2], we capture both semantic (learned from the sentence itself) and structural information within sentences (learned from POS tags). A joint model that encodes sentences and their structure has been used before in [7], but, unlike the task tackled in that work, our definition extraction component assumes that the term is known and tries to detect whether the given candidate sentence is a good definition for the term or not.

3 Topic Pages Pipeline

There are four main components for generating *Topic Pages*, as shown in Figure 2: an annotation module, a definition ranking module, a snippet ranking module, and a related concept extraction module.

3.1 Article Annotation

The annotation module receives content in *XML* format, finds concepts' mentions in articles and books, and then feeds the sentences and snippets mentioning a concept into the subsequent components. Each section in the article is considered a snippet. After we perform sentence splitting, we annotate concepts in sentences by using a simple dictionary look-up against the *Omniscience* taxonomy [12] which is a taxonomy of scientific concepts. If an abbreviation for the concept is proposed in the text, such as "Machine Learning (ML)", then the abbreviation (ML) is also added as an alias for the concept and is looked in the article. We use the Schwartz and Hearst method [17] to detect such abbreviations. If multiple concepts partially share some span (of an annotation), we annotate the span with the longest concept and ignore the short annotation.

3.2   Definition Ranking and Extraction

Definitions provide a concise description of the concept. For each concept, and per domain, we rank all the sentences where the concept was annotated and select the top-ranked one as the definition for the concept. We simplify the machine learning task to binary classification where, given a concept and a candidate sentence, the model predicts if it is a good definition for the concept or not. For a target concept, candidate sentences are ranked based on the score the classifier assigns to them and the top-ranked sentence is used as the definition. We use two different models for the definition classification task: an *LSTM+CNN* and a *SciBERT* model.

LSTM+CNN model Previous work [11,2] used *LSTM* [4] and *CNN* [10] models to classify sentences in the definition classification task. We use a combined approach that uses two *LSTMs* and two *CNNs*: one *LSTM* gets the actual sentence as the input and captures the sequential patterns of terms, and the other *LSTM* gets the *POS* tags of the words in the sentence as the input and captures the sequential patterns of syntax in the sentence. One *CNN* gets the actual sentence as the input captures the spatial distribution of terms, and the other *CNN* gets the *POS* tags of the words in the sentence as the input and captures the spatial distribution of grammatical elements. We concatenate the representations learned by each of these models and feed it to a feed-forward *MLP* layer which does the classification, using cross-entropy loss for training.

SciBERT model We use the *SciBERT* model [1] which is trained on scientific articles. As input, we feed the concept and the candidate sentence separated with a special token ([SEP]) to the *SciBERT* model and get the representation of the [CLS] token. This representation is then fed to a simple feed-forward layer which does the classification, using cross-entropy loss for optimization.

3.3   Snippet Ranking

For a given concept, all snippets annotated with the concept are collected and ranked by a snippet ranking method. The top 10 snippets are used for generating the *Topic Page* for the concept. We use a lexical matching model that scores snippets using a simple location-aware term frequency score as follows:

$$F(c,s) = \frac{tf(c,s)}{|s|} * (1 - \frac{l_1(c)}{|s|}) \tag{1}$$

where c and s are a concept and a snippet respectively, $tf(c,s)$ is the frequency of c in s, $|s|$ is the length of s, and $l_1(c)$ is the location of the first occurrence of c in s. Hence, the earlier the concept is mentioned in a snippet, the higher the score the snippet would receive.

3.4   Related Concept Extraction

To find the most relevant concepts to a given concept, we retrieve all co-occurring concepts in snippets. Concepts are then ranked based on the number of their

Model	Precision	Recall	F1
Jin et al. [6]	0.92	0.79	0.85
Li et al. [11]	0.90	0.92	0.91
Navigli and Velardi [14]	**0.99**	0.61	0.85
LSTM+CNN	0.94	0.91	0.92
SciBERT	0.94	**0.93**	**0.93**

Table 1. Performance of different definition classification models on the WCL dataset in terms of macro-averaged precision, recall, and F1.

co-occurrence with the target concept and the top 5 concepts are selected as the related concepts to the target concept.

4 Results

In this section, we describe the scientific content collection and the used taxonomy (*Omniscience*) that are the basis of the *Topic Pages*. We further discuss the results of the different definition ranking models on two datasets and provide some statistics of the generated *Topic Pages* and usage statistics over time. We leave a large-scale evaluation of the snippet ranking and the related concept extraction modules to future work.

4.1 Datasets and baselines

We use a collection of articles published in over $2,700$ journals as well as the content of $43,000$ books to generate the *Topic Pages*. This collection contains over 18 million articles and book chapters in *XML* format. All journals and books belong to different scientific domains. We use the *OmniScience* taxonomy to build the *Topic Pages*, which contain over 700K concepts for the 20 domains.

To evaluate the performance of the definition ranking module, we use the *WCL* dataset [14] which contains $4,619$ sentences labeled either as definitional ("good") or non-definitional ("bad") sentences regarding a concept. We follow the same setup as [11] for training and evaluating models on this dataset. We additionally use a proprietary dataset containing $43,368$ sentences extracted from articles and books distributed across 8 different domains and labeled by subject matter experts for the definition evaluation task as either "good" or "bad" definitions regarding a concept. We compare the performance of several models including the *LSTM+CNN*, *SciBERT*, Navigli and Velardi [14], Li et al. [11], and Jin et al. [6] on the *WCL* dataset. We further evaluate the performance of the best-performing models on the proprietary dataset. For the *LSTM+CNN* model, the batch size is set to 32, the number of hidden layers of the *LSTM* model is set to 128, and word embeddings are initiated with *GloVe* and fine-tuned during training. The *MLP* module has a hidden layer with 256 dimensions trained for 10 epochs. The *SciBERT* model is trained for 8 epochs with a batch size of 16. We perform 10-fold cross-validation and report the average performance.

4.2   Results of the definition extraction models

Table 1 shows the performance of different models on the *WCL* dataset. This dataset is extracted from Wikipedia and most of the Wikipedia-based definitions follow a similar structure, making them easy to classify. The *SciBERT* model achieves the best *F1* score on this dataset. Navigli and Velardi [14] have higher precision than all models but a very low recall compared to *SciBERT*. The higher performance of the *SciBERT* model compared to the *LSTM+CNN* model shows that *SciBERT* can learn both sequential and spatial distribution of words in definitional sentences as well as the structural information within such sentences.

We further evaluate the performance of the top-performing models (*SciBERT* and *LSTM+CNN*) on the proprietary dataset which is much larger than the *WCL* dataset; results are shown in Table 3. This dataset contains definitions from various sources. Unlike Wikipedia-based definitions, definitions extracted from different books and journals do not follow a similar structure which makes the classification task more difficult, hence the lower performance of the two models compared to the *WCL* dataset. The *SciBERT* model outperforms the *LSTM+CNN* model on this dataset as well across all domains. This again confirms the ability of the *SciBERT* model in modeling semantics and the structure of definitions. Moreover, *SciBERT* has consistently higher performance than the *LSTM+CNN* on all individual domains except *Social Sciences*. As *SciBERT* is pre-trained on publications in the biomedical and computer science domains the low performance of this model on domains such as *Social Sciences* may be attributed to this fact. On the other hand, as the results show, *SciBERT* performs better on domains such as *Chemistry* and *Material Sciences* as such domains are closer to its trained domains.

Concept	Definition	Error source
Association List	An association list is simply a list of name value pairs.	Too generic
Hierarchical DB	In a hierarchical DB relationships are defined by storage structure	Too generic
Habilitation	The acquisition of abilities not possessed previously.	Too specific
Sample Space	the set of all possible outcomes in a probability model	Partially good

Table 2. Example of errors (false positives) of the SciBERT-based models.

Other than the domain difference, the additional errors should be attributed to the inherent difficulty of the task. Based on our analysis, the biggest sources of errors are the false positives which are mainly caused by generic, specific, or partially good definitions. Table 2 shows examples of definitions wrongly labeled by the *SciBERT* model and the possible explanation for the errors. Generic definitions are good definitions but they cover a very broad aspect of the concept. Specific definitions are also good definitions but they contain unnecessary additional information. Partially good definitions cover only some essential aspects of the concept. All these cases are labeled as "bad definitions" by the subject-

domain	SciBERT			LSTM+CNN		
	Precision	Recall	F1	Precision	Recall	F1
Chemistry	0.78	0.80	0.79	0.69	0.68	0.68
Earth Sciences	0.80	0.84	0.82	0.66	0.64	0.65
Material Sciences	0.80	0.88	0.83	0.50	0.49	0.49
Computer Science	0.56	0.60	0.58	0.43	0.48	0.45
Social Sciences	0.39	0.43	0.41	0.38	0.46	0.42
All domains	**0.79**	**0.78**	**0.78**	0.70	0.69	0.69

Table 3. Performance of the *LSTM+CNN* and *SciBERT* models on five domains.

matter experts but detected as "good definitions" by the model. To handle such cases, the model should have an understanding of the generality or specificity of the concept which can be quite challenging to model.

The *Topic Pages* product contains over $363,000$ topic pages in 20 different scientific domains. Topic pages have over 23 million visits per month making them one of the popular knowledge bases among researchers and students. There are about $63,000$ concepts without a definition on *Topic Pages* mostly due to the bad performance of the current production model (*LSTM+CNN*) in some domains.

5   Conclusions and Discussion

In this paper, we introduced *Topic Pages*, a publicly available knowledge base for scientific concepts with their definitions, most relevant concepts, and snippets providing more context around them. We described all the major components combined to build this resource. The pipeline for generating *Topic Pages* can be used on top of any document collection as well as a taxonomy to build a similar resource in any domain. With over $363,000$ topic pages in 20 different scientific domains, and more than 23 million unique visitors per month, *Topic Pages* are one of the popular knowledge bases among researchers and students. We described all major components of the pipeline for extracting different pieces of information necessary to generate the pages. In this work, we mainly focused on building a high-performance definition extraction model. To this end, we used an *LSTM+CNN* and a *SciBERT* model. Empirical evaluation shows that both models can outperform existing models for the definition classification and extraction task. However, the *SciBERT* model still needs to be improved for domains such as *Social Sciences*. The biggest drawback of using *SciBERT* for such domains is that this model is pre-trained on mostly biomedical articles and, therefore, it cannot model all other domains as well. As a future work, we would like to exploit the concepts and their definitions extracted from Wikipedia as well as expand our dataset to further fine-tune the *SciBERT* model for such domains. As another future work, we are going to use the click-through data we have collected as a proxy to train supervised models for related concept extraction and snippet ranking components.

References

1. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: EMNLP-IJCNLP. pp. 3615–3620 (2019)
2. Espinosa-Anke, L., Schockaert, S.: Syntactically aware neural architectures for definition extraction. In: NAACL. pp. 378–385 (2018)
3. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: COLING. pp. 539–545 (1992)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
5. Jain, A., Gupta, N., Mujumdar, S., Mehta, S., Madhok, R.: Content driven enrichment of formal text using concept definitions and applications. In: Proceedings of the 29th on Hypertext and Social Media, pp. 96–100 (2018)
6. Jin, Y., Kan, M.Y., Ng, J.P., He, X.: Mining scientific terms and their definitions: A study of the acl anthology. In: EMNLP. pp. 780–790 (2013)
7. Kang, D., Head, A., Sidhu, R., Lo, K., Weld, D.S., Hearst, M.A.: Document-level definition detection in scholarly documents: Existing models, error analyses, and future directions. In: Proceedings of the First Workshop on Scholarly Document Processing. pp. 196–206 (2020)
8. Klavans, J.L., Muresan, S.: A method for automatically building and evaluating dictionary resources. In: LREC. pp. 231–234 (2002)
9. Kobyliński, Ł., Przepiórkowski, A.: Definition extraction with balanced random forests. In: International Conference on Natural Language Processing. pp. 237–247 (2008)
10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
11. Li, S., Xu, B., Chung, T.L.: Definition extraction with lstm recurrent neural networks. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 177–189 (2016)
12. Malaisé, V., Otten, A., Coupet, P.: Omniscience and extensions–lessons learned from designing a multi-domain, multi-use case knowledge representation system. In: European Knowledge Acquisition Workshop. pp. 228–242 (2018)
13. Murthy, S.K., Lo, K., King, D., Bhagavatula, C., Kuehl, B., Johnson, S., Borchardt, J., Weld, D.S., Hope, T., Downey, D.: Accord: A multi-document approach to generating diverse descriptions of scientific concepts. arXiv preprint arXiv:2205.06982 (2022)
14. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: ACL. pp. 1318–1327 (2010)
15. Reiplinger, M., Schäfer, U., Wolska, M.: Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In: ACL-2012 special workshop on rediscovering 50 years of discoveries. pp. 55–65 (2012)
16. Roig Mirapeix, M., Espinosa Anke, L., Camacho-Collados, J.: Definition extraction feature analysis: From canonical to naturally-occurring definitions. In: Proceedings of the Workshop on the Cognitive Aspects of the Lexicon. pp. 81–91 (2020)
17. Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: Biocomputing 2003, pp. 451–462 (2002)
18. Veyseh, A., Dernoncourt, F., Dou, D., Nguyen, T.: A joint model for definition extraction with syntactic connection and semantic consistency. In: AAAI. pp. 9098–9105 (2020)

19. Veyseh, A.P.B., Dernoncourt, F., Tran, Q.H., Nguyen, T.H.: What does this acronym mean? introducing a new dataset for acronym identification and disambiguation. In: COLING. pp. 3285–3301 (2020)
20. Westerhout, E.: Definition extraction using linguistic and structural features. In: Proceedings of the 1st Workshop on Definition Extraction. pp. 61–67 (2009)