Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature

Luca Foppiano, Thaer M Dieb, Akira Suzuki, Masashi Ishii

一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature

Luca Foppiano†‡ Thaer M. Dieb†‡ Akira Suzuki†‡ and Masashi Ishii†‡

†Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

E-mail: ‡{FOPPIANO.Luca, MOUSTAFADIEB.Thaer, SUZUKI.Akira3, ISHII.Masashi}@nims.go.jp

Abstract The automatic collection of materials information from research papers using Natural Language Processing (NLP) is highly required for rapid materials development using big data, namely materials informatics (MI). The difficulty of this automatic collection is mainly caused by the variety of expressions in the papers, a robust system with tolerance to such variety is required to be developed. In this paper, we report an ongoing interdisciplinary work to construct a system for automatic collection of superconductor-related information from scientific literature using text mining techniques. We focused on the identification of superconducting material names and their critical temperature (Tc) key property. We discuss the construction of a prototype for extraction and linking using machine learning (ML) techniques for the physical information collection. From the evaluation using 500 sample documents, we define a baseline and a direction for future improvements.

Keywords Material Informatics, Superconductors, Machine Learning, NLP, TDM

1. Introduction

Automatic information extraction from research papers using Natural Language Processing (NLP) is a highly required approach in many domains. In material research, the use of big data obtained experimentally, known as Material Informatics (MI), may give insight leading to new breakthroughs in materials discovery.

The increasing availability of scientific papers and the expertise costs to manually extract valuable data justify the needs of Text and Data Mining (TDM) automatic approaches. Despite the general understanding of the necessity of TDM, the wide variation of writing styles and formats within in the same research topics makes this task complex.

In this paper, we propose a framework for automatic data extraction and tried an application of this system to the superconductivity scientific field.

Superconductivity is a phenomenon of exactly zero electrical resistance and expulsion of magnetic flux fields occurring in certain materials called superconductors, under a characteristic critical temperature [1].

Historically, high-temperature superconductors have been suddenly discovered by intuition of scientists rather than systematic consideration because of the lack of theoretical understanding [2] [3]. In this situation, data-driven exploration [4][5][6][7] would be a feasible approach to discover new superconducting materials. Since it requires huge data sets for precise prediction, high-throughput experiments, first-principle calculations, existing material databases should be used as data sources.

Currently, several material databases are available for property search. However, when looking at the superconductor sub-domain, the main one is SuperCon [8] hosted and maintained by the National Institute for Materials Science (NIMS). The SuperCon contains about 32000 inorganic and about 558 organic superconductor material definitions. Although it is constantly updated with manual data collection, it cannot catch up with the massive fresh information from the increasing number of articles each year. It is our challenge to make SuperCon richer for data-driven science. In this paper, we describe the ongoing attempt to design a TDM system using NLP techniques. In particular, we focus on extracting superconducting materials names and their linking to the corresponding critical temperature (Tc) values.

Our system is built on an Open Source library for text mining for scholarly documents: Grobid [9]. We evaluate the performance of the system using precision, recall and F1-score. These quantitative values provide a baseline for measuring our progress in

solving the task. Similar attempts of mining scientific literature in materials domain had been conducted in [10] and [11].

This paper is organised as follows, Section 2 describes the details of a working prototype specialised in which development of annotated corpus for material name recognition is included. Section 3 presents the evaluation methods and results. Section 4 concludes this paper with summary and scopes.

2. System architecture

The scientific information in articles is often presented in tables and figures for easy understanding. However, in order to achieve higher order data structuring, it is necessary to link extracted entities. For example, the different expressions of superconductor's name should be appropriately linked to corresponding property values depending on various material and measurement parameters, where the material parameter is the dopant density, stoichiometry, etc. and measurement parameter is the applied magnetic field, pressure, and so on. Obviously, these relational entities are not fully included in the tables and figures. Therefore, we insist on extracting information existing in the main body text and in captions of tables and figures. In superconductor sub-domain, since a dataset of superconductor material and its corresponding Tc is crucial, we constructed a system for linking the material and Tc (material-Tc). Other properties, such as critical magnetic field (Hc) will be targeted in the upcoming work.

We built our implementation based on an open source library called Grobid [9]. Grobid is a sequence labelling and document segmentation library based on machine learning (ML), Conditional Random Field (CRF) [12]. It provides full support for extraction of data from PDF and built-in workflow for pre-annotating training data, machine learning models training and evaluation. The PDF support in Grobid was important because allowed us to focus on the processing of a single format instead of dealing with several XML flavours depending on external publishers. Among various available open source tools, the choice of Grobid is well justified by the fact that it is still actively developed and it is available as ready product employed in several large-scale research repositories, such as Mendeley [13]. In a recent benchmark study, Grobid performed best in citation extraction task [14]. Lastly, Grobid has been successfully extended to support several domain-specific problems, for example, astronomical entities recognition [15], dictionaries segmentation [16], software mention [17] and measurements extraction and normalisation [18]. Such open source measurement extractor, Grobid-quantities was fitting in our use case for Tc recognition.

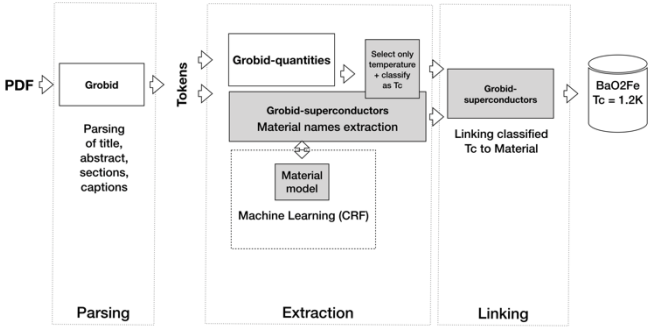


Figure 1: Schema of the system showing the interaction of different component and models when a document is processed.

Our developed system illustrated in Figure 1, beside the Parsing which is provided by Grobid, it consists mainly of two phases: (a) "Extraction phase" and (b) "Linking phase". We extract relevant entities, i.e., superconducting materials and Tc in (a), and link these extracted entities in (b). After parsing title, abstract, sections, and captions and following tokenisation, the (a) Extraction phase combines the entities resulting from a newly trained model for superconductor material recognition and a conventional module, Grobid-quantities for measurement extraction. The superconducting material recognition model was trained with five full documents manually annotated (42 entity in total). We also used domain-specific chemical recogniser, called ChemSpot [19] which extracts chemical entities from text and classify them into types: SYSTEMATIC, IDENTIFIER, FORMULA, TRIVIAL, ABBREVIATION, FAMILY, MULTIPLE and UNKNOWN.

In the following (b) Linking phase, three tasks are sequentially performed. Since Grobid-quantities supports extraction of a wide range of measurements (temperatures, lengths, pressures, etc.), we selected only temperatures (task 1, Selection). Then we classified each of them into "Tc" and the others (task 2, Classification). Finally, we linked "Tc" with the positionally closest material term (task 3, Linking).

The (2) classification was realised by word matching to an original dictionary which summarises commonly used as Tc-related words in superconducting literature (e.g. "Tc", "critical temperature"). We looked for the Tc-related words in the surroundings (within five words according to empirical probability) of a numerical temperature value and made a pair of Tc and value (Tc-value). At last, we (3) linked the closest material term to Tc-value, resulting in the required entity linking of material-Tc-value.

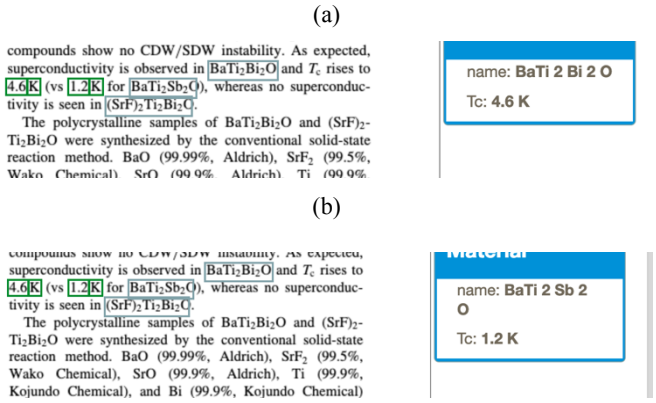


Figure 2: Example of a correct linking between material and Tc. The popup windows indicate the links of material (a) BaTi2Bi2O with Tc of 4.6K and (b) BaTi2Sb2O with 1.2K

In Figure 2 and Figure 3 we show two examples each of correct and incorrect linking respectively. Figure 2 indicates the automatic links of material (Figure 2a) BaTi2Bi2O with Tc of 4.6K and (Figure 2b) Bati2Sb2O with 1.2K. Figure 3, however, shows (Figure 3a) unlinked BaTi2Bi2O to the Tc-value and (Figure 3b) the incomplete linking of KOS2O6 to the Tc-value. In the former case, the misclassification of 1.2K in task 2 results in unlinked BaTi2Sb2O. In the latter case, misrecognition of a decimal point as a period provided an incomplete Tc value of 60K. The statistical investigation of the (b) Linking phase is discussed in the later part of Section 3.

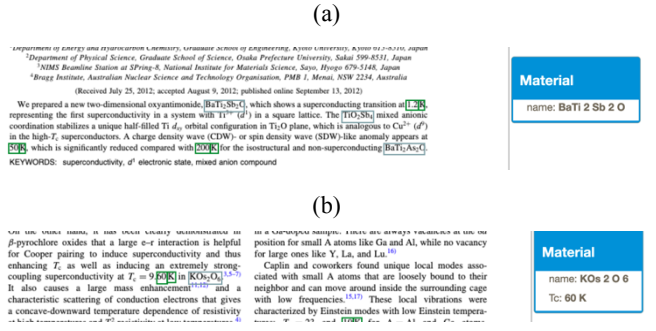
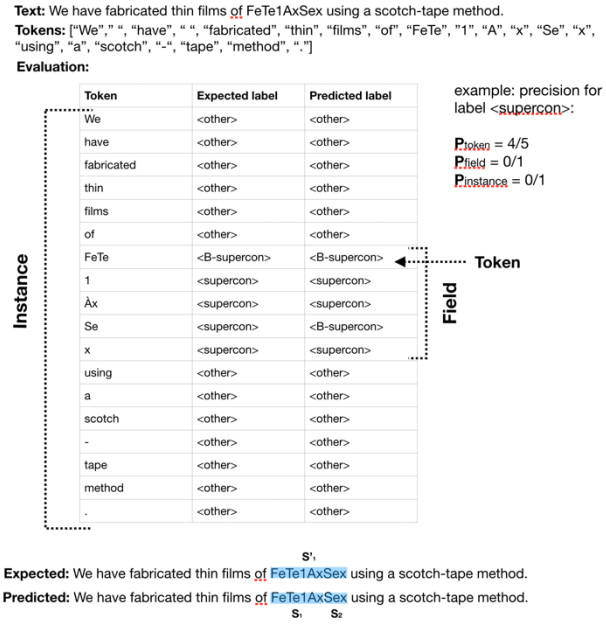


Figure 3: Example of an incorrect linking between material and Tc.

Although the materials (a) $BaTi_2Sb_2O$ and (b) $KO_{S2}O_6$ are correctly identified, Tc of $BaTi_2Sb_2O$ cannot be extracted and that of $KO_{S2}O_6$ is extracted incorrectly (correct Tc was 9.60 K while the extracted value was 60 K).

3. Experiments and results

The superconductor CRF model in the (a) Extraction phase was investigated using a corpus of five papers (four for training and one for testing) having a total of 42 entities classified with a superconducting material label <supercon>.



Token	Expected label	Predicted label
We	<other>	<other>
have	<other>	<other>
fabricated	<other>	<other>
thin	<other>	<other>
films	<other>	<other>
of	<other>	<other>
FeTe	<B-supercon>	<B-supercon>
1	<supercon>	<supercon>
Àx	<supercon>	<supercon>
Se	<supercon>	<B-supercon>
x	<supercon>	<supercon>
using	<other>	<other>
a	<other>	<other>
scotch	<other>	<other>
-	<other>	<other>
tape	<other>	<other>
method	<other>	<other>
.	<other>	<other>

Figure 4: Given a sentence (or paragraphs) the tokenisation transforms it in an array of character sequences. These are then labelled by the ML CRF model. This figure illustrates the three different levels of granularity on which measurements are calculated.

We estimated precision, recall and F1-score for the model using the evaluation framework built-in in Grobid. These measure indices are calculated at three different levels: tokens-level, field-level and instance-level. Figure 4 illustrates the concept through an example of prediction. In this table, expected (second column) and predicted (third column) labels are tagged on to each token. The token-level evaluates the predictability for each token P_{token}, in this example equals to 4/5. The field-level evaluates the continuity of a field with the same label. Although "$FeTe_1A_xSe_x$" should be recognised as a field, the predicted label indicated separation into two fields of "$FeTe_1A_x$" and "Se_x". Therefore, the predictability in field-level, P_{field} is 0/1. Finally, instance-level indicates predictability of the

continuity of all the fields within the same instance, where the instance is defined by a paragraph in this case. If "$FeTe_1A_xSe_x$" is only one field in the paragraph, the predictability in instance-level, $P_{instance}$ is 0/1. The predictability of P_{token}, P_{field}, and $P_{instance}$ are statistically extendable to more strict indices of precision, recall, and F1-score. The following listing shows precision, recall and F1-score for our trained model.

Table 1: Grobid evaluation framework results for the (a) material extraction model

Token-level results				
Label	Accuracy	Precision	Recall	F1
<supercon>	98.61	84.42	85.28	84.85
Field-level results				
Label	Accuracy	Precision	Recall	F1
<supercon>	72.94	66.67	56.25	61.02
Instance-level results				
Total expected instances		Correct instances	Instance-level recall	
22		15	68.18	

As shown in Table 1, recall at token-level, field-level, and instance-level were 85.28, 56.25, and 68.18%, respectively. The minimum recall at the field-level indicates insufficient training for recognition of superconducting compounds, rather than superconducting elements. The higher score at instance-level suggests that some specific superconducting compounds which are intensively appeared in a paragraph have never been trained at this stage. Moreover, the fact that the precision (66.67%) is better than the recall (56.25%) in the field-level indicates false positive (FP) is smaller than false negative (FN). Consequently, the true superconducting compounds are relatively difficult to recognise for our trained model (c.f., Figure 4). We have already found that missing annotation in the training papers, and so a quality improvement of training data is necessary to establish a practical extraction system.

Finally, we tested the proposed system on a larger corpus of papers. We processed 500 superconductor-related PDF papers from three publishers: American Institute of Physics (AIP), American Physical Society (APS) and Institute of Physics (IOP) and we manually evaluated the extracted critical temperatures and their link with the related material. As discussed in Figure 4 the material recognition in (a) the Extraction phase mistook in boundaries detection. The other examples of mistake are *missing notation*: predicted $LaFe_xO$ for expected $LaFe_xO_{1-x}$, and *irregular separation*: predicted single superconductor for expected two different materials separated by coordinating conjunctions like "and" or "comma".

Despite several imperfect extractions, we obtained unique material entities of 1644 from the 500 papers as summarised in Table 2. For Tc extraction, although the temperature could refer to unrelated experimental conditions or thermal treatment for sample preparation, the number of extracted Tc was 1173.

Table 2: Material and Tc extraction results from a corpus of 500 papers.

Material entities	Unique material entities	Temperature entities	Tc entities
5400	1644	7554	1173

Table 3 shows evaluations at *sentence-level* and *paragraph-level* for the (b) Linking phase. As shown in this table, 77 (sentence-level) and 109 (paragraph-level) correct links are obtained for the extracted Tc of 1173. From these values, in the case of sentence-level boundary, precision and recall were estimated to be 68.7% and 6.5%, respectively. The paragraph-level resulted in lower precision and higher recall, 57% and 10.7% respectively. Increasing the search span from sentence to paragraph, F1-score increased from 11.87% to 18.01%. The decrease in precision was compensated by the increase in recall.

The generally low recall is partly caused by wrong parsing prior to (a) Extraction phase. The conversion from PDF to text unavoidably produces irregular tokens originated from wrong UTF-8 characters, stream-ordering issues, and missing fonts. Considering that we used empirical rules for linking as described in Section 2, the irregular tokens increase FN, resulting in the low recall. The effects of irregular tokens can be reduced by using volumes of training data with a high-quality corpus.

Table 3: Result of linking materials and Tc.

Boundaries	Links	Correct links	Precision	Recall	F1-Score
Sentence level	112	77	68.7%	6.5%	11.87%
Paragraph level	191	109	57%	10.7%	18.01%

4. Conclusion

In this paper, we proposed an automatic extraction of superconductor related information from scientific publications. The proposed system consists of two phases: a machine learning sequence labelling process for entity extraction (Extraction phase) and a simple rule-based for linking (Linking phase).

Although we found feasible performance for the (a) Extraction phase using a sequence labelling approach, bulk corpus is necessary to reach practical performances. For the (b) Linking phase, the rule-based approach was limited by irregular tokens introduced in a conversion process from PDF to text, so that probability model and ML are required. In the next step, we plan to introduce deep neural networks like Bi-LSTM+CRF approach and embedding in the Extraction phase.

References

[1] W. contributors, "Superconductivity Wikipedia Page," https://en.wikipedia.org/wiki/Superconductivity.

[2] M. Klintenberg and O. Eriksson, "Possible high-temperature superconductors predicted from electronic structure and data-filtering algorithms," *Comput. Mater. Sci.*, vol.67, pp.282–286, 2013.

[3] T. Konno, H. Kurokawa, F. Nabeshima, R. Ogawa, M. Iwazume, I. Hosako, and A. Maeda, "Deep Learning of Superconductors I: Estimation of Critical Temperature of Superconductors Toward the Search for New Materials," *CoRR*, vol.abs/1812.01995, 2018.

[4] R. Matsumoto, Z. Hou, M. Nagao, S. Adachi, H. Hara, H. Tanaka, K. Nakamura, R. Murakami, S. Yamamoto, H. Takeya, T. Irifune, K. Terakura, and Y. Takano, "Data-driven exploration of new pressure-induced superconductivity in PbBi2Te4," *Sci. Technol. Adv. Mater.*, vol.19, no. 1, pp.909–916, 2018.

[5] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor," *Comput. Mater. Sci.*, vol.154, pp.346–354, 2018.

[6] R. M. Geilhufe, S. S. Borysov, D. Kalpakchi, and A. V. Balatsky, "Towards novel organic high-T_c superconductors: Data mining using density of states similarity search," *Phys Rev Mater.*, vol.2, no. 2, p.024802, Feb. 2018.

[7] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, "Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints," *Chem. Mater.*, vol.27, no. 3, pp.735–743, 2015.

[8] NIMS, "SuperCon," https://supercon.nims.go.jp/.

[9] G. Contributors, "GROBID," https://github.com/kermitt2/grobid.

[10] T. Dieb, M. Yoshioka, S. Hara, and M. C Newton, "Framework for automatic information extraction from research papers on nanocrystal devices," *Beilstein J. Nanotechnol.*, vol.6, pp.1872–1882, 2015.

[11] C. J. Court and J. M. Cole, "Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction," *Sci. Data*, vol.5, p.180111, 2018.

[12] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[13] P. Gooch and K. Jack, "How well does Mendeley's Metadata Extraction Work?," https://krisjack.wordpress.com/2015/03/12/how-well-does-mendeleys-metadata-extraction-work/.

[14] D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel, "Evaluation and Comparison of Open Source Bibliographic Reference Parsers: A Business Use Case," *CoRR*, vol.abs/1802.01168, 2018.

[15] G. Contributors, "grobid-astro: A machine learning software for extracting astronomical entities from scholarly documents.," https://github.com/kermitt2/grobid-astro.

[16] M. Khemakhem, L. Foppiano, and L. Romary, "Automatic extraction of TEI structures in digitized lexical resources using conditional random fields," electronic lexicography, eLex 2017, 2017.

[17] G. Contributors, "software-mentions: GROBID module to recognize in textual documents and PDF any mentions of software.," https://github.com/Impactstory/software-mentions.

[18] G. Contributors, "grobid-quantities: GROBID extension for identifying and normalising physical quantities.," https://github.com/kermitt2/grobid-quantities.

[19] T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: a hybrid system for chemical named entity recognition," *Bioinformatics*, vol.28, no. 12, pp.1633–1640, 2012.