Toward full automatic identification of superconducting materials and their properties in original papers: ambitious scope and current status

Luca Foppiano*, Sae Dieb*, Akira Suzuki*, Pedro Baptista de Castro+, Yan Meng+, Kensei Terashima+, Yoshihiko Takano+, Masashi Ishii*

***Material Database Group, MaDIS, NIMS**

+Nano Frontier Superconducting Materials Group, MANA, NIMS

E-mail: FOPPIANO.Luca@nims.go.jp

The importance of Text and Data Mining (TDM) processes is growing with the increasing popularity of new material discovery techniques, i.e., materials informatics (MI) and the expanding volume of research publications. Although fully automatic TDM processes of entities extraction with properties linking are still challenging, we can recognise its potential through related work on the collection of Curie and Néel temperatures[1] and recipes of inorganic material synthesis[2], from scientific articles.

The National Institute for Materials Science (NIMS) is working on developing a TDM framework toward fully automatic database construction. Ideally, the system produces an output dataset of material names and related properties, from scientific literature. In this presentation, we discuss the current status and the challenges of this ambitious work, applied to the superconductors domain.

We designed a framework[3] dividing out this task into two simpler steps: ”Extraction” and ”Linking”. In collaboration with domain experts, we produced the first superconducting corpus, SuperMat[4] for training and evaluation of our Machine Learning (ML) based ”Extraction” step. The ”Linking” phase is responsible to find relationships between pre-extracted entities. It is more challenging than ”Extraction” because there is not a general established methodology. In fact, traditional dependency parsers[5] are not effective for recognising such deep relationships. In order to overcome this limitation, we implemented a heuristics, relying on combinations and vicinity of different types of entities and complemented this method with a sequence labelling approach for simplified cases. Using our framework, we processed a dataset of 500 articles on superconductivity selected from Journals of American Institute of Physics (AIP), American Physical Society (APS) and Institute of Physics (IOP) and we manually verified the 600 extracted links. We confirmed that the system achieves an F1-score of 70% (Precision: 73%, Recall: 67%). The system is also applied to dataset extraction from articles in Journal of Superconductivity (IOP) published from 2015 to 2018, and 850 links (material - superconducting critical temperature *Tc*) were obtained from 1088 papers.

We are working on improving the system precision by extending SuperMat to have more training data, and by exploiting customised trained embeddings from unsupervised text analysis in the ”Linking” step.

1. Court, C., Cole, J. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. Sci Data 5, 180111 (2018). https://doi.org/10.1038/sdata.2018.111

2. Kononova, O., Huo, H., He, T. et al. Text-mined dataset of inorganic materials synthesis recipes. Sci Data 6, 203 (2019). https://doi.org/10.1038/s41597-019-0224-1

3. Foppiano et al. Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature. IEICE 2019

4. Foppiano et al. SuperMat: Corpus for Extraction of Superconductor Materials Data, JSAP Spring, 2020

5. Sandra Kübler, Ryan McDonald, and Joakim Nivre Dependency parsing, Synthesis Lectures on Human Language Technologies, 2009, Vol. 2, No. 1 , Pages 1-127 (https://doi.org/10.2200/S00169ED1V01Y200901HLT002)