SuperMat: Corpus for Extraction of Superconductor Materials Data

Luca Foppiano[*], Sae Dieb[*], Akira Suzuki[*], Kensei Terashima[+], Pedro Baptista de Castro[+], Suguru Iwasaki[+], Yoshihiko Takano[+], Masashi Ishii[*]

[*]Material Database Group, MaDIS (NIMS)

[+]Nano Frontier Superconducting Materials Group (NIMS)

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

The automatic collection of material information from research papers using Machine Learning (ML) and Natural Language Processing (NLP) is a milestone to establish a sustainable approach for creating or enriching domain-specific databases. In the field of superconductors materials, the manual data collection used to populate SuperCon[1] cannot cope with the massive fresh information from the increasing number of articles published every year. For this reason, an inter-disciplinary project is currently ongoing. It aims to develop a system to automatically extract superconductors materials and related properties from scientific literature[2]. Unfortunately, in this unexplored terrain, there is no record of previous attempts in the scientific literature, nor existing datasets in the public domain.

In this submission, we present our work and the methodology used for creating a corpus for extraction of superconductor material data: SuperMat, in collaboration with the Nano Frontier Superconducting Material Group.

The raw data is composed of scientific papers collected from three different sources. The (a) Open Access version of articles referenced in SuperCon records, (b) articles provided from domain-experts and (c) articles obtained by search, using keywords such as 'superconductor', 'critical temperature' and 'superconductivity', in the arXiv's "Condensed matter" category[3]. After creating the logical model, selecting which relevant information to collect and subsequently define the labels (or tag-set), we established the annotations guidelines and the correction / cross-validation process.

Papers	**Annotations**	**Labels**
60	4666	6
Paragraphs	**Sentences**	**Tokens**
1146	7243	365631

Table 1: Overview of the dataset.

material-tc	**tc-pressure**
115	42

Table 2: Overview of relationship information.

Currently, we have annotated and validated 60 papers whose main characteristics are summarised in Table 1. The dataset provides scientific text annotated with entities and relationship information (links). The entities are identified among 6 classes (or labels) as summarised in Figure 1 and linked using relationships `material-tc` and `tc-pressure` (Table 2).

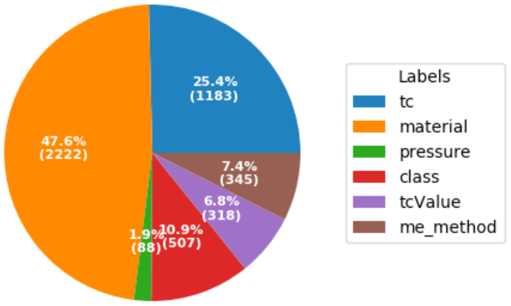

Figure 1: SuperMat labels distribution.

This corpus is designed for training sequence labelling statistical models and can be utilised for developing domain-specific systems for entity extraction, entity-relationship and clustering.

[1] `http://supercon.nims.go.jp`

[2] Luca Foppiano et al., "Proposal for Automatic Extraction Framework of Superconductors related Information from Scientific literature," *THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS*, 2019,

[3] `https://arxiv.org/archive/cond-mat`