# Machine readable extraction of chemically modified materials name

**Luca Foppiano*, Sae Dieb*, 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 text and data mining (TDM) processes are necessary to help research exploiting the vast available scientific knowledge. Although the extraction of material names from text is obviously a key functionality in material science, the ability to parse and segment the raw form is still challenging.

The main tools: pymatgen[1], and the Material Parser[2]can perform rule-based parsing of formulas, sophisticated analysis such as phase diagram, crystal structure analysis. However, material obtained from text can be presented in several forms due to writing style but also from extraction issues, for example text extracted from PDF documents can present missing or invalid encoded characters, or incorrect stream order. In these cases, the rule-based approach might not be effective. In fact, we evaluated the Material Parser[2] to recognise the formulas on materials names extracted from PDF documents and obtained a precision of only 18%. The low precision is mainly caused by that the names are complexly expressed with segments dependent on sample fabrication processes, e.g. *hole-doped La1 − x SrxOyFe1-yAs compound with (x = 0.1, 0.2 and 0.3 and y = 0.1, 0.4 and 0.5)*. In this study, we performed the segmentation of the complex material name and machine understandable materials recognition.

We present a machine-learning based material name parser for unstructured text on superconductors. The parser implements a Conditional Random Field (CRF) model that segments the raw material string in six components: name (*Metal diboride*, *hydrogen*, etc.), chemical formula (*La Fe $O_7$*, *$SiH_4$*, etc.), doping ratio (*Zn-doped*, *pure*, etc.), stochiometric variable names and values (*x = 1, 2; y = 3*), and shape (*thin film*, *powder*, etc.). We constructed the training data of 3000 material names, using all the material entities from the SuperMat[3] dataset. Currently the results as precision, recall and f1-score, using a holdout validation corpus of 4000 material names are as follow:

| Component | Label | Precision | Recall | F1 |
|---|---|---|---|---|
| Name | <name> | 64.29 | 46.15 | 53.73 |
| Doping ratio | <doping> | 35.71 | 35.71 | 35.71 |
| Chemical formula | <formula> | 83.82 | 87.02 | 85.39 |
| Stochiometric values | <value> | 90.91 | 83.33 | 86.96 |
| Stochiometric variable names | <variable> | 100 | 95.83 | 97.87 |
| Shape | <shape> | 96 | 92.31 | 94.12 |
| All (micro avg.) | | 82.26 | 78.16 | 80.16 |

We plan to cover additional element, such as substrate and fabrication process and to extend the data to other domains: thermoelectric, spintronic, magneto caloric.

1. Ong et al ”Python Materials Genomics (pymatgen) : A Robust, Open-Source Python Library for Materials Analysis”. Computational Materials Science, 2013, 68, 314-319. doi:10.1016/j.commatsci.2012.10.028

2. Kononova et. al ”Text-mined dataset of inorganic materials synthesis recipes”, Scientific Data 6 (1), 1-11 (2019) doi:10.1038/s41597-019-0224-1

3. Foppiano et al ”SuperMat: Construction of a linked annotated dataset from superconductors-related publications” (2021) doi:10.5281/zenodo.4422093