Leveraging Segmentation of Physical Units through a Newly Open Source Corpus

Luca Foppiano, Akira Suzuki, Thaer M. Dieb, Masashi Ishii[1] and Mikiko Tanifuji
MaDIS, National Institute for Materials Science (NIMS)
E-mail: FOPPIANO.Luca@nims.go.jp

The identification of physical measurements is a recurrent need in materials informatics (MI). For example, the extraction of superconductor materials and their properties[2] requires to identify and understand temperature, pressure, magnetisation. When designing automatic systems for information extraction from scientific literature, the identification of the raw measurement alone is not sufficient. Quantity transformations, such as normalisation, require the understanding of values and units, which are contained in unstructured text with ad-hoc conventions. String matching and lookups are failing with growing unit complexity and variability. Therefore a generic unit segmentation system is necessary.

This contribution is part of a larger project called Grobid-quantities[3], a machine learning (ML) based, Open Source system for extracting and normalising physical measurements from scientific and patent literature. In this submission, we present a general approach for units representation, and we introduce the public availability (Creative Commons licence) of a corpus of segmented physical units. Currently, there are no comparable results in scientific literature because no public datasets are available for this task. Our approach for the unit representation follows the International System of Measurement (SI), where each unit is represented as a product of triples: *prefix*, *base* and *power*. This straightforward approach offers the flexibility to support any combination of units from any system of measurements. Figure 1 illustrates an example where `kV²/cm` is tokenised and segmented as product of triples.

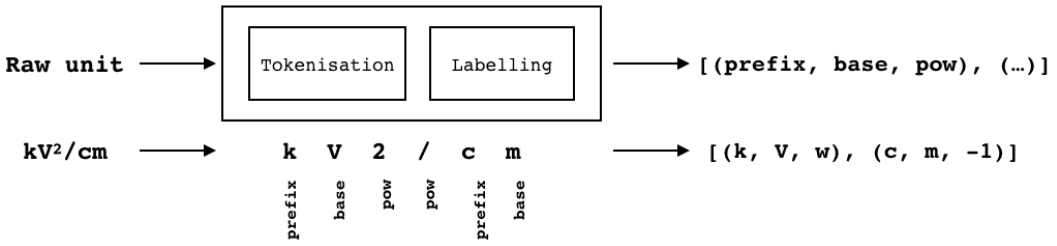

Figure 1: The process of parsing a raw unit into the product of triples. Notice that the label `pow` is used to identify both exponent and division marks (needed to correctly set the second triple's exponent, in this case negative).

We used the Grobid-quantities ML-based unit segmentation implementation to create a new corpus. We used data provided by previous work of some of the authors[4], where about 2000 units were extracted from 3490 papers of Journal of Applied Physics. The data was pre-annotated and manually corrected.

The resulting corpus contains approximately 700 simple and 1300 complex units, and it's available in XML format at the Grobid-quantities repository[3]. It is suitable for evaluating new or existing systems for unit segmentation. We plan to increase the coverage by adding new data from other domains.

---

1. Corresponding author: ISHII.Masashi@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. *grobid-quantities*, `https://github.com/kermitt2/grobid-quantities`, [Online; accessed 18-April-2019], 2016.
4. Suzuki Akira and Ishii Masashi, "Constructing a "Unit dictionary" from scientific articles," in *Third International Workshop on SCIentific DOCument Analysis (JSAI International Symposia on AI)* (Springer, 2018).