Lukrécia Mertová | 18.07.2024

Text- and Table-Based Language Models that Mine Materials Science Literature

ChemDataExtractor 2.0 and **TableDataExtractor:**
auto-populating ontology of crystallographic data

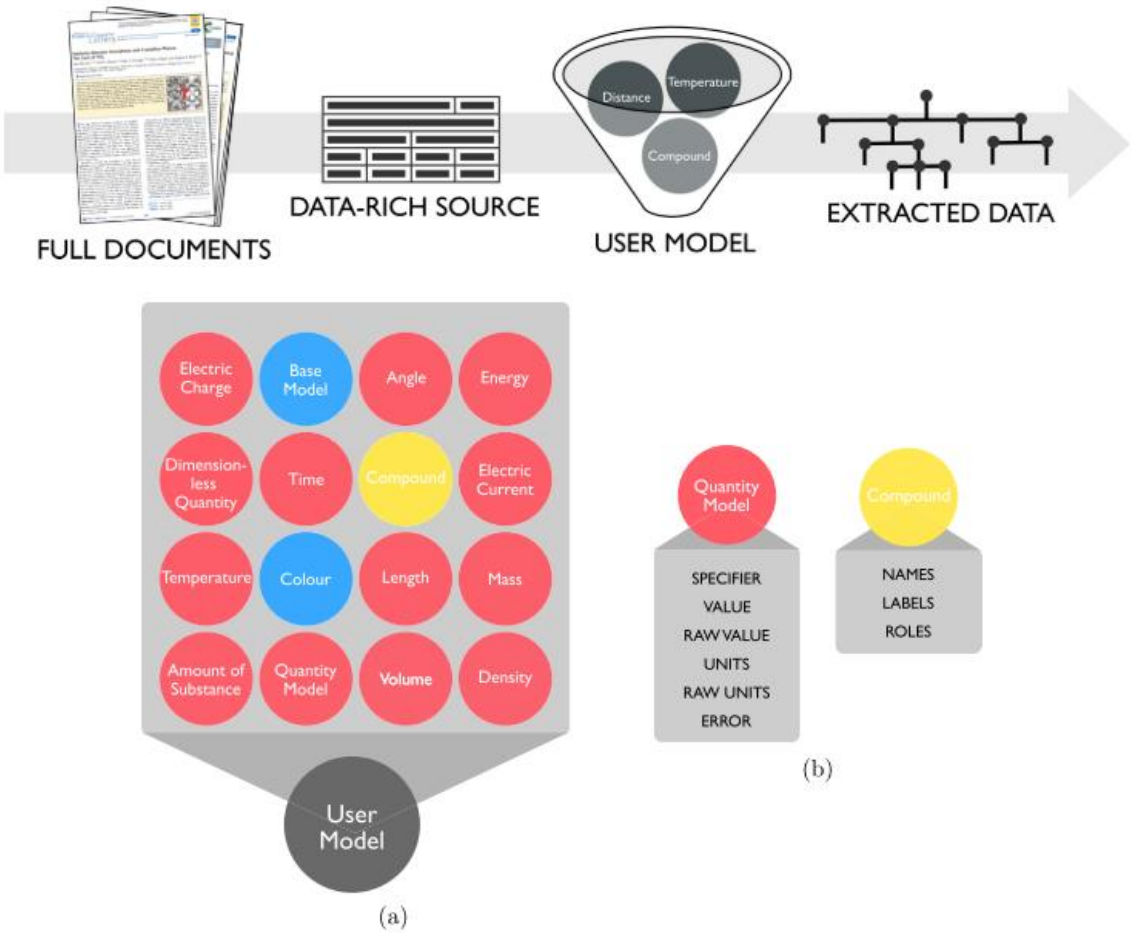

Figure 2. Model concept in ChemDataExtractor 2.0. (a) *User Model* is constructed by joining individual model templates. The individual models can be linked in a nested fashion. *Quantity Models* are shown in red, the *Compound Model* is in yellow, and *General Models* are in blue. New model templates can be created, either by combining existing ones, or from scratch, for example, by defining new quantities with appropriate base units for *Quantity Models*. (b) Most models contain relevant fields that define the model. The fields for the *Quantity Model* and the *Compound Model* are shown.

OpticalBERT and **OpticalTable-SQA:**
Text- and table-based language models for optical materials

The **BERT** model has a transformer-based architecture. Instead of the traditional left-to-right language-modelling procedure, BERT achieves bidirectional information propagation by predicting randomly masked tokens in sentences and predicting if two sentences follow each other (Next Sentence Prediction, NSP).

Task	Precision OpticalPureBERT OpticalBERT
Abstract Classification	93.5%
Question Answering (QA)	87.0%
Chemical-Named-Entity Recognition	80.90% - 82.06%
Question-Answering Tasks on Tables	45.34% - 62.32%, 82.16% - 90.00%

- For CNER, datasets are **CHEMDNER** and **Matscholar**
- Tapas-SQA: The publicly available table-parsing sequential question-answering model for general text.
- **OpticalTable-SQA**: enabled **question-answering** capabilities for **tabular data** by focusing on teaching the model to understand various symbols of optical properties that reside in the header of tables.
- 4,534 question-answering pairs – **What?** and **Which?**

Domain-specific BERT-based language models perform better in domain-specific tasks

J. Chem. Inf. Model. **2023**, *63*, 1961.