Automatic Identification and Normalisation of Physical Measurements in Scientific Literature

Luca Foppiano[1], Laurent Romary[2], Masashi Ishii[1] and Mikiko Tanifuji[1]

[1]Material Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Japan

[2]Inria, team ALMAnaCH, France

Content

- Background and motivation
- System overview
- Sytem architecture
- Evaluation
- Conclusions

Background

Text and Data Mining is a growing discipline

Collection of high quality data:

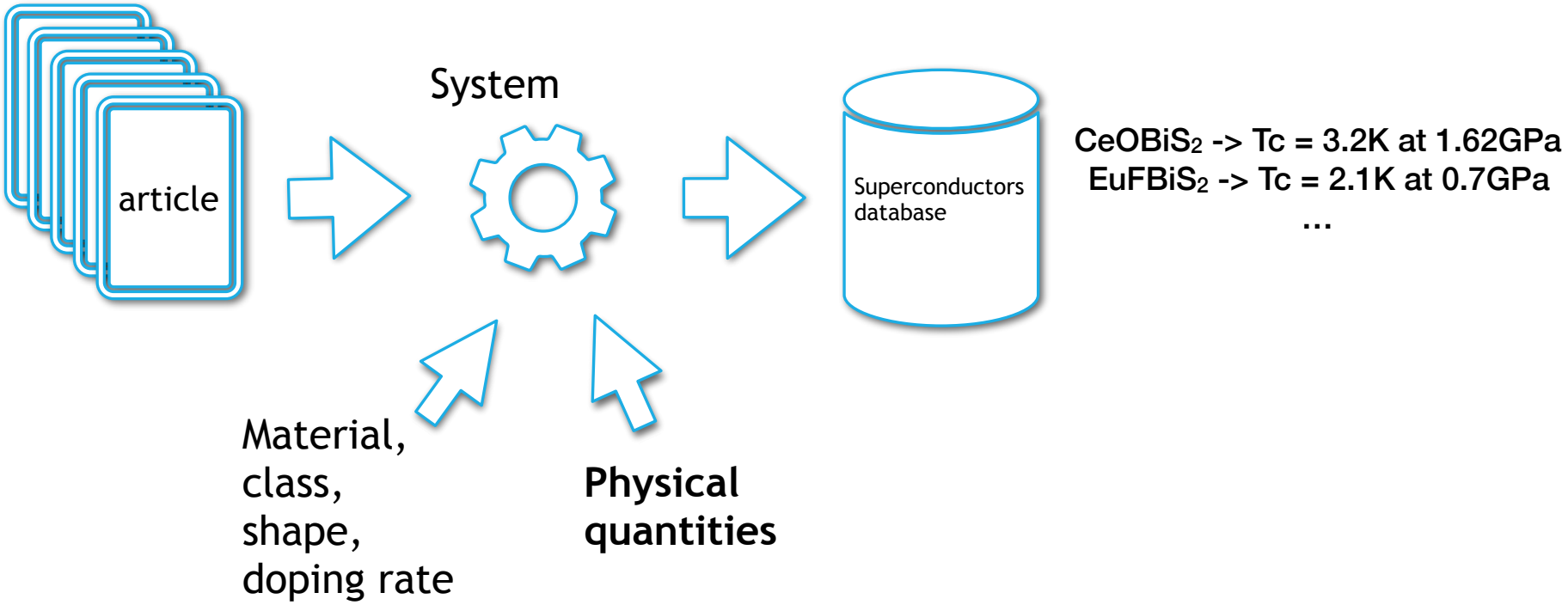
- availability of large quantities of data
- cheaper (and faster) than manual process

Many other applications:

- information retrieval
- tagging and categorization
- summarization

Example: Material Informatics

Automatic construction of superconductor database [1] using scientific articles describing experiments and results.



[1] L. Foppiano, T. M. Dieb, A. Suzuki, & M. Ishii (2019). Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature. In *Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature* (pp. 1-5). 信学技報, vol. 119, no. 66, SC2019-1 (no.66).

Grobid-quantities

— Grobid-quantities is an open-source system for automatic extraction and normalisation of physical quantities

— Based on Grobid (Generation of Bibliographic data), a library for extracting and structuring text from PDFs of scientific literature

— Developed in collaboration with Patrice Lopez (author of the Grobid library)

— Tools and data are available on github http://github.com/kermitt2/grobid-quantities

— Native support to PDF extraction and coordinates

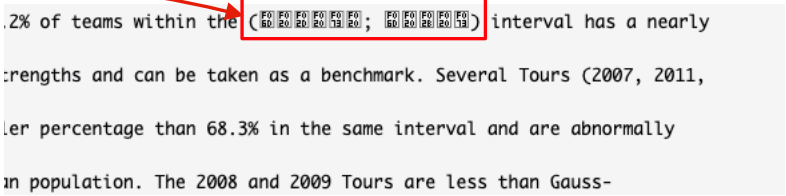
— Data processing via REST or local batch

— Measurement normalisation is implemented using Unit of Measurement Java library (https://unitsofmeasurement.github.io/) Standard JSR-385

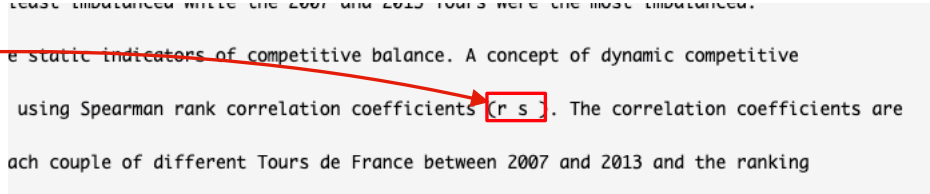
Data in real life

Some example how data looks like when is extracted from PDFs:

a 68.2% of teams within the $(\mu - \sigma;\ \mu + \sigma)$ interval has a nearly
strengths and can be taken as a benchmark. Several Tours (2007, 2011,
naller percentage than 68.3% in the same interval and are abnormally



competitive balance. A concept of dynamic competitive
correlation coefficients (r_s). The correlation coefficients are



ng protocol

wind (LW) and medium wind (MW) were defined as wind
ls ranging from 5 to 9 knots (2.57–4.63 $m \cdot s^{-1}$), and 10 to
nots (5.14–8.23 $m \cdot s^{-1}$), respectively. The races with too much
variation switching from LW to MW were excluded from the
sis. Wind speed was measured with an anemometer (Plastimo,

averaged weekly training programs included two running, two muscle strengthening, and four to six windsurfing sessions. The study was approved by the University ethics committee.Light wind (LW) and medium wind (MW) were defined as wind speeds ranging from 5 to 9 knots (2.57-4.63 mAEs)1), and 10 to 16 knots (5.14-8.23 mAEs)1), respectively.

Grobid-quantities and ML

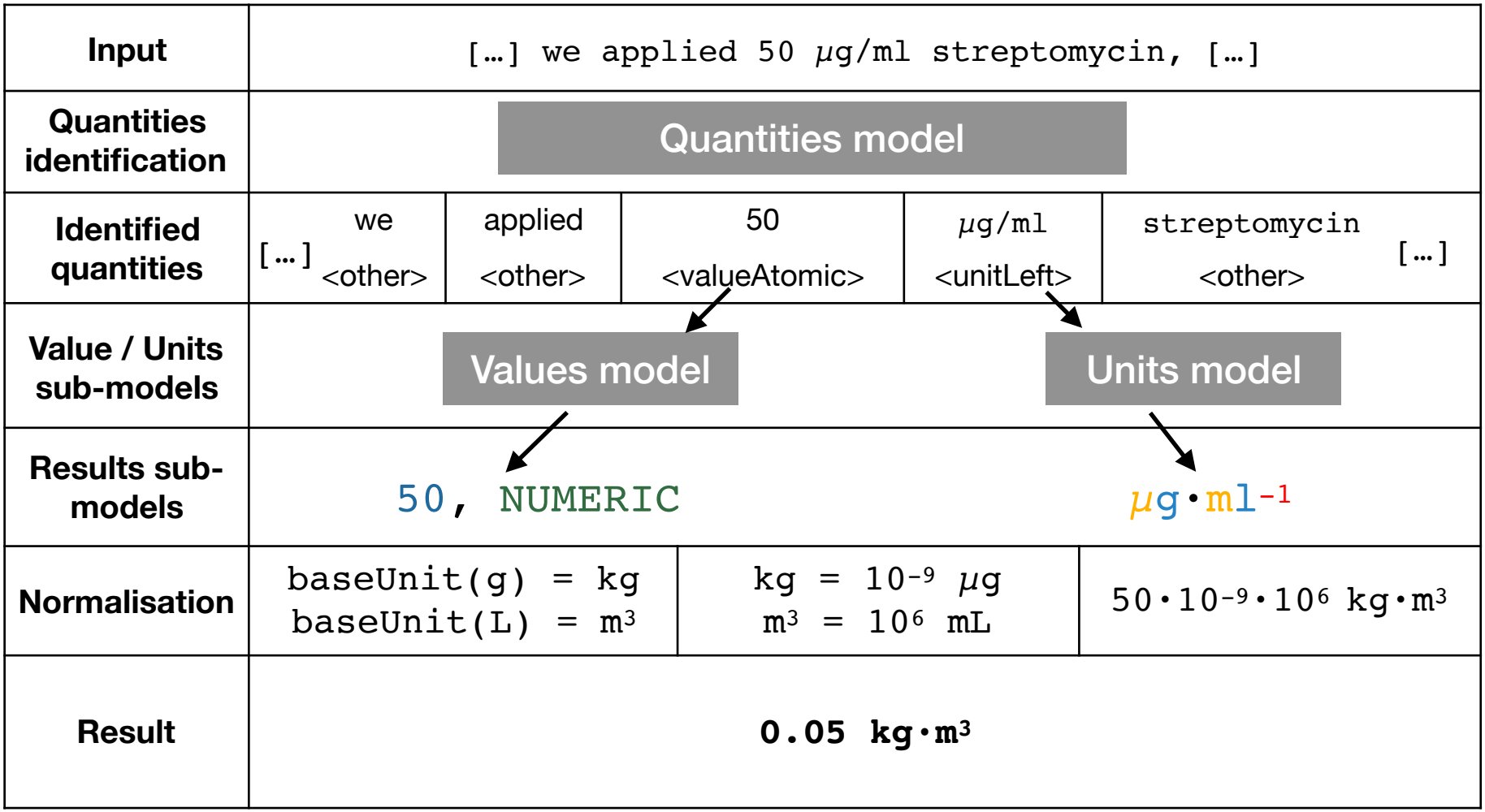
Machine Learning can assist to reduce the effect of noisy data

Grobid-quantities uses Conditional Random Field (CRF) algorithm

Machine Learning cascade architecture:

- Maximise the efficiency/Minimise the effort of each component
- Errors are propagated and (!!) amplified

Cascade architecture



Quantities model

Extract quantities as combination unit and values

Works at token/word level

Supports different type of quantities: atomic value, interval min/max, interval base+range, lists

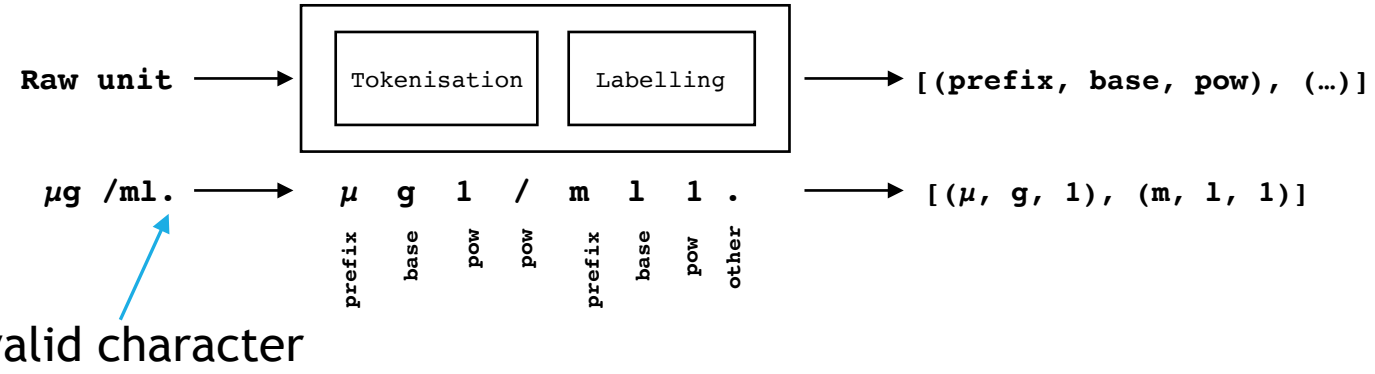
Label	Description	Example
<valueAtomic>	value of an atomic quantity	**2** m
<valueLeast>	least value in an interval	from **2** m
<valueMost>	max value in an interval	up to **7** m
<valueBase>	base value in a range	**20** ± 7 m
<valueRange>	range value in a range	20 ± **7** m
<valueList>	list of quantities	**2**, **3** and **10** m
<unitLeft>	left-attached unit	**pH** 2
<unitRight>	right-attached unit	2 **m**
<other>	everything else	-

Units model

Segmentation of Units works at character level

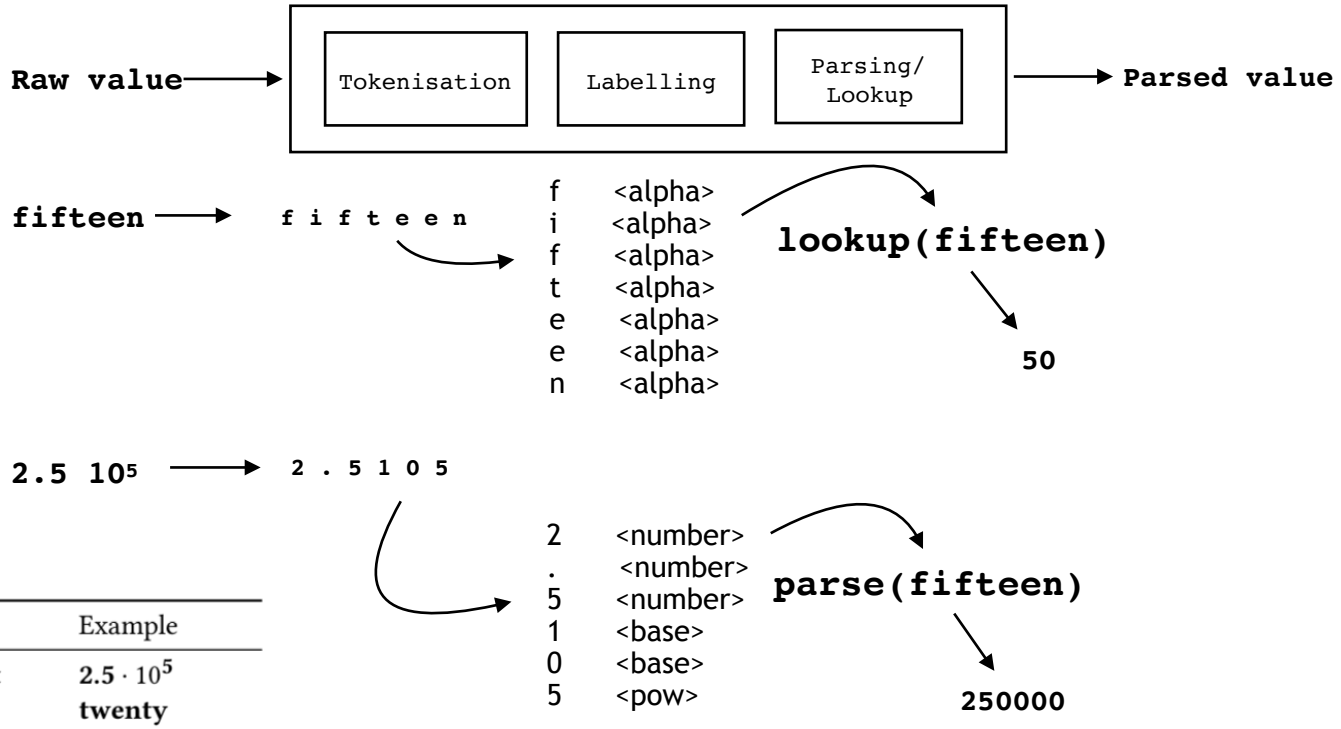
Model based on product of triples (from the SI): prefix, pow and base.

Label	Description	Example
<prefix>	prefix of the unit	$\mathbf{k}m^2$
<base>	unit base	$k\mathbf{m}^2$
<pow>	unit power	$km^{\mathbf{2}}$
<other>	everything else	-



Values model

Character level model



Label	Description	Example
<number>	numeric value / coefficient	$\mathbf{2.5} \cdot 10^5$
<alpha>	alphabetic value	**twenty**
<time>	time expression	in **1970-01-02**
<base>	base in scientific notation	$2.5 \cdot \mathbf{10}^5$
<pow>	exponent in scientific notation	$2.5 \cdot 10^{\mathbf{5}}$
<other>	everything else	-

Features

— General features: capital, digits, punctuation, etc..

— Unit Lexicon (standard notations, type, system, inflections, lemmas)

— **Typographical information** (superscript, subscript, fonts, etc.) **are ignored**


```
{
  "notations": [
    {
      "raw": "m^3",
      "product": [
        {
          "base": "m",
          "pow": "3"
        }
      ]
    },
    {
      "raw": "m³",
      "product": [
        {
          "base": "m",
          "pow": "3"
        }
      ]
    }
  ],
  "type": "VOLUME",
  "system": "SI_BASE",
  "supportsPrefixes": true,
  "names": [
    {
      "lemma": "cubic meter",
      "inflections": [
        "cubic meters"
      ]
    },
    {
      "lemma": "cubic metre",
      "inflections": [
        "cubic metres"
      ]
    }
  ]
}
```

Evaluation experiment

— Training and evaluation was done using 32 PDFs **Open Access** articles selected in domain of medicine, robotics, astronomy, and physiology (available on grobid-quantities github repository) and manually corrected.

— Evaluation metrics (precision, recall and f1-score) were calculated using 10-fold cross-validation

Evaluation evaluation results

Label	Precision	Recall	F1	Support
Quantities CRF model				
<unitLeft>	96.76	94.71	95.71	2805
<unitRight>	93.06	72.1	80.02	120
<valueAtomic>	85	84.77	84.84	3599
<valueBase>	78.82	76.52	77.53	94
<valueLeast>	85.05	77.39	80.94	862
<valueList>	72.09	54.87	61.33	494
<valueMost>	84.09	73.03	78.07	878
<valueRange>	84.56	81.5	82.68	93
all (macro avg.)	84.93	76.86	**80.14**	
Unit CRF model				
<base>	99.02	99.22	99.12	3075
<pow>	98.04	98.9	98.46	322
<prefix>	99.19	98.8	98.99	821
all (macro avg.)	98.75	98.97	**98.86**	
Values CRF model				
<alpha>	96.64	98.65	97.62	826
<base>	83.06	69.23	72.77	58
<number>	98.01	99.02	98.52	3858
<pow>	76.45	74.67	74.58	56
<time>	72.54	87.83	79.34	322
all (macro avg.)	85.34	85.88	**84.57**	-

Promising results with CRF considering the small training corpus

Unit evaluation is biased due to the nature of the data

Evaluation with an independent evaluation corpus [1] resulted in 81% F1-Score.

[1] Foppiano, L. & Suzuki, A. & Dieb, T. & Ishii, M. & Tanifuji, M. (2019). Leveraging Segmentation of Physical Units through a Newly Open Source Corpus.

Demo

Conclusions

We presented an open-source application for extracting and normalizing physical quantities with promising results

This application is engineered to support the processing of large quantities of data

Currently used in a project for extraction of superconductors material related properties

Future plans:

— increase the amount of training data (!)

— exploit typographical/layout information such as superscript/subscript/font

— add more contextualized information (e.g. article domain) to solve units ambiguities

Thank you