Leveraging Segmentation of Physical Units through a Newly Open Source Corpus

LUCA FOPPIANO, AKIRA SUZUKI, THAER M. DIEB, MASASHI ISHII AND MIKIKO TANIFUJI

MATERIAL DATA AND INTEGRATED SYSTEM (MADIS),

NATIONAL INSTITUTE FOR MATERIALS SCIENCE (NIMS)

Content

Introduction and motivation

Text and Data Mining in scientific literature requires inevitably to deal with **units of measurements and physical quantities**

— The units recognition sub-task is an important step (measurement normalisation)

— Extraction of physical quantities is not a new subject

  — different techniques have been already investigated

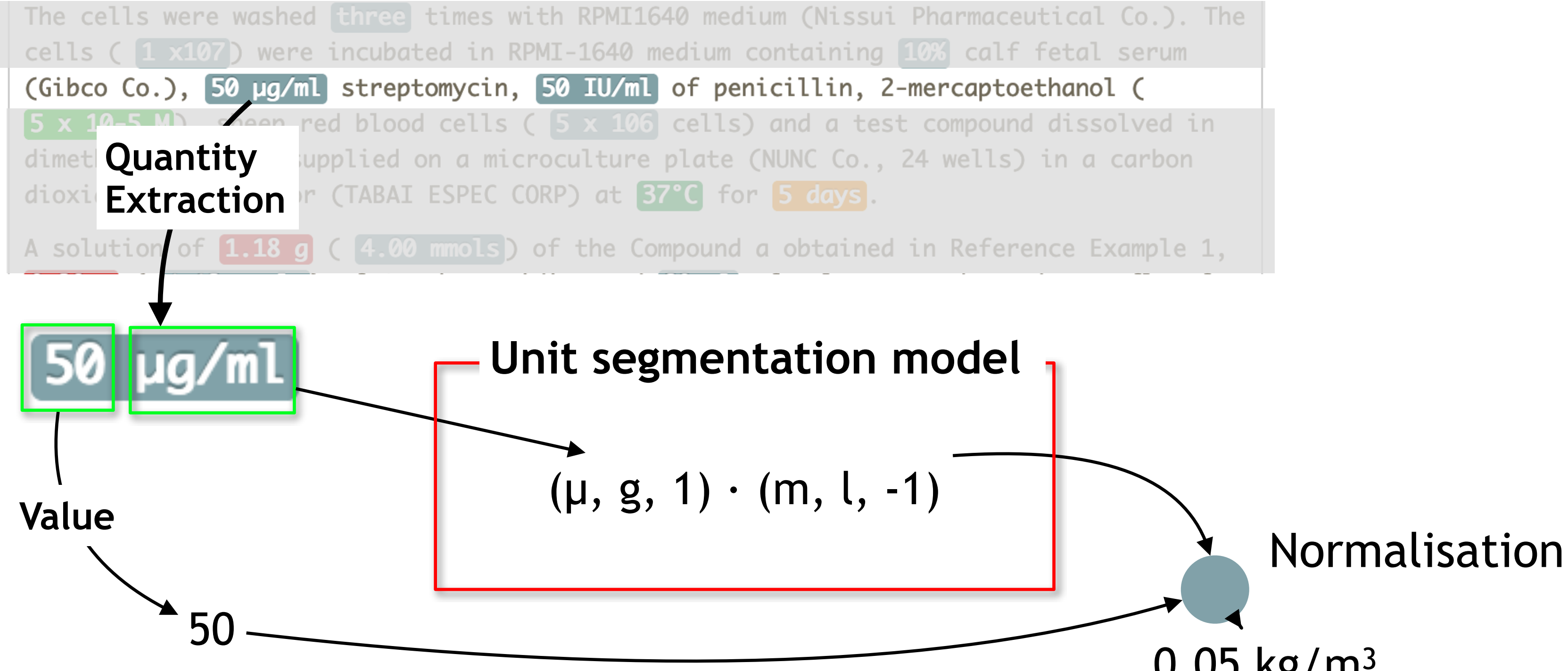
  — there is no benchmark to evaluate different approaches (**Reproducibility issue!**)

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Quantity extraction system

Use an open-source system called Grobid-quantities (developed in collaboration with P. Lopez)

Example data:



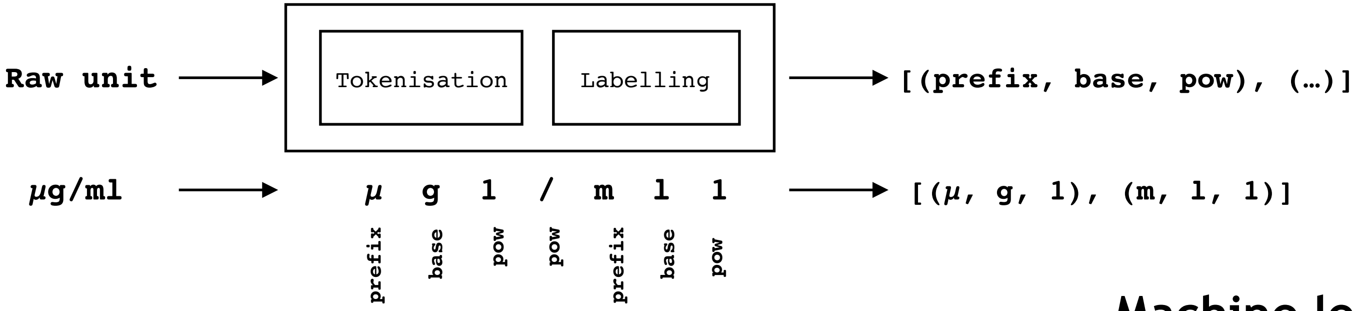
The cells were washed three times with RPMI1640 medium (Nissui Pharmaceutical Co.). The cells (1 x107) were incubated in RPMI-1640 medium containing 10% calf fetal serum (Gibco Co.), 50 μg/ml streptomycin, 50 IU/ml of penicillin, 2-mercaptoethanol (5 x 10-5 M), sheep red blood cells (5 x 106 cells) and a test compound dissolved in dimet supplied on a microculture plate (NUNC Co., 24 wells) in a carbon dioxi r (TABAI ESPEC CORP) at 37°C for 5 days.

A solution of 1.18 g (4.00 mmols) of the Compound a obtained in Reference Example 1,

Grobid-quantities is hosted on GitHub https://github.com/kermitt2/grobid-quantities

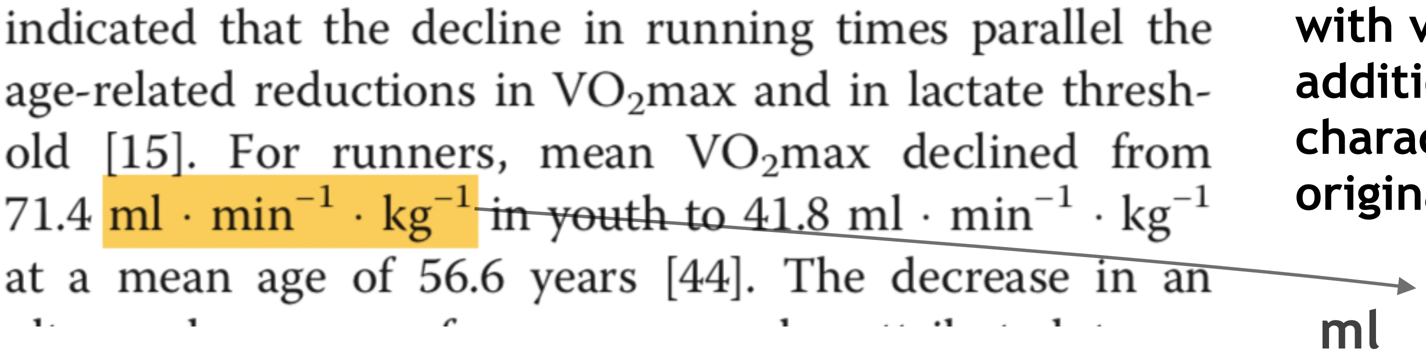
Unit segmentation model

Segments raw text to **product of triples (prefix, base, power)**, International System of Units



Raw unit → Tokenisation | Labelling → **[(prefix, base, pow), (…)]**

µg/ml → **µ g 1 / m 1 1** → **[(µ, g, 1), (m, 1, 1)]**



indicated that the decline in running times parallel the age-related reductions in VO_2max and in lactate threshold [15]. For runners, mean VO_2max declined from 71.4 ml · min^{-1} · kg^{-1} in youth to 41.8 ml · min^{-1} · kg^{-1} at a mean age of 56.6 years [44]. The decrease in an

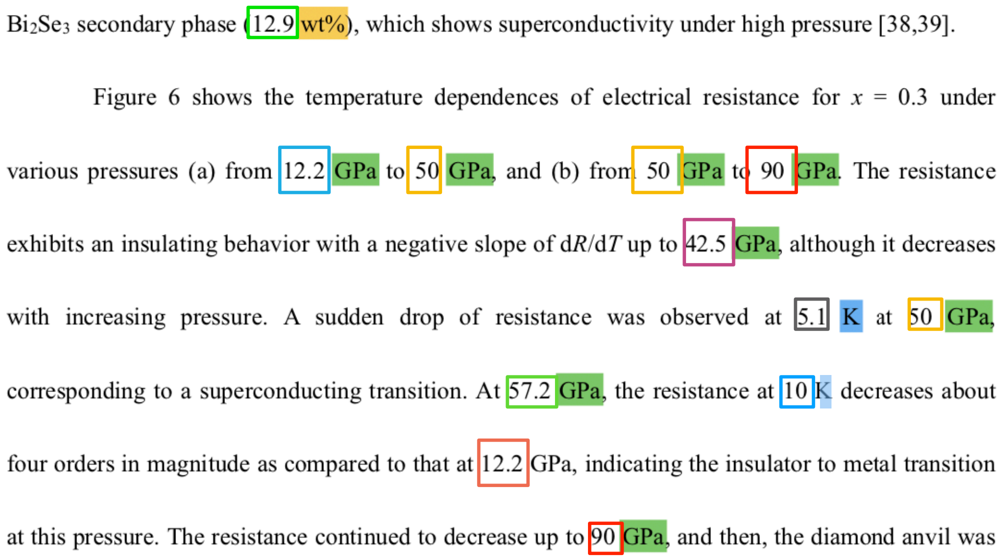
ml min – 1 kg –1

Machine learning is important for dealing with variation having additional or missing characters from the original text.

Problem

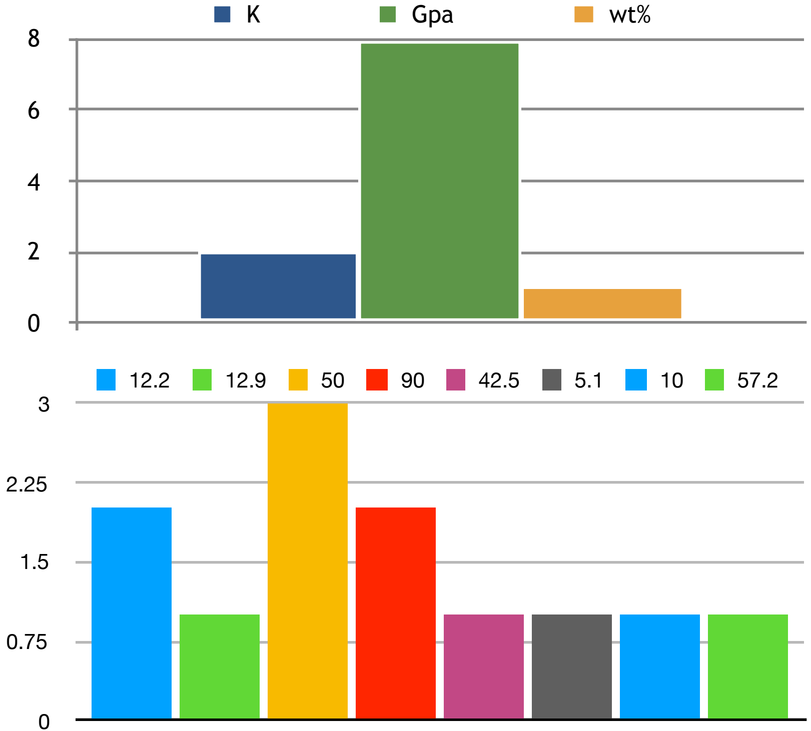
— No benchmark for evaluation

— The statistical distribution of units in specific subdomains creates biased evaluation results



Bi_2Se_3 secondary phase (12.9 wt%), which shows superconductivity under high pressure [38,39].

Figure 6 shows the temperature dependences of electrical resistance for $x = 0.3$ under various pressures (a) from 12.2 GPa to 50 GPa, and (b) from 50 GPa to 90 GPa. The resistance exhibits an insulating behavior with a negative slope of dR/dT up to 42.5 GPa, although it decreases with increasing pressure. A sudden drop of resistance was observed at 5.1 K at 50 GPa, corresponding to a superconducting transition. At 57.2 GPa, the resistance at 10 K decreases about four orders in magnitude as compared to that at 12.2 GPa, indicating the insulator to metal transition at this pressure. The resistance continued to decrease up to 90 GPa, and then, the diamond anvil was

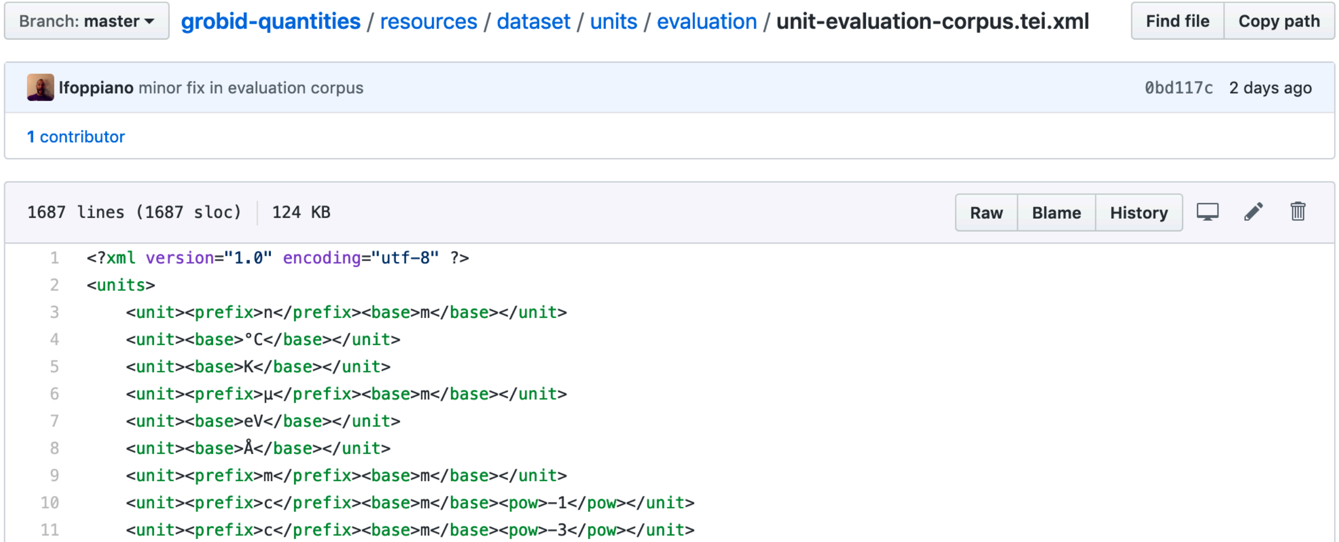


Unit segmentation corpus

We constructed a **UNIt Segmentation CORpus** [UNISCOR]

- “general dataset” covering broad area of Applied Physics

- open-source, available to be used (and improved) by anybody



grobid-quantities / resources / dataset / units / evaluation / **unit-evaluation-corpus.tei.xml**

lfoppiano minor fix in evaluation corpus 0bd117c 2 days ago

1 contributor

1687 lines (1687 sloc) 124 KB

```
<?xml version="1.0" encoding="utf-8" ?>
<units>
    <unit><prefix>n</prefix><base>m</base></unit>
    <unit><base>°C</base></unit>
    <unit><base>K</base></unit>
    <unit><prefix>μ</prefix><base>m</base></unit>
    <unit><base>eV</base></unit>
    <unit><base>Å</base></unit>
    <unit><prefix>m</prefix><base>m</base></unit>
    <unit><prefix>c</prefix><base>m</base><pow>-1</pow></unit>
    <unit><prefix>c</prefix><base>m</base><pow>-3</pow></unit>
```

UNISCOR construction

— Collected 3490 papers of Journal of Applied Physics

Suzuki Akira and Ishii Masashi, “Constructing a ”Unit dictionary” from scientific articles,” in Third International Work- shop on SCIentific DOCument Analysis (JSAI International Symposia on AI) (Springer, 2018).

— Automatic annotations using grobid-quantities

— Manually check the annotated data in collaboration with other researchers/engineers from NIMS

Corpus statistics

- extracted 1700 unique units:
  - 400 simple units (e.g. m, l, etc..)
  - 1300 complex units (e.g. m/s, etc..)

- Licence: **Open source** (CC-BY 4.0)

- Available at https://github.com/kermitt2/grobid-quantities/blob/master/resources/dataset/units/evaluation/unit-evaluation-corpus.tei.xml

Evaluation experiments

Experiment set-up:

— [GQ1] corpus created for training grobid-quantities (built with the application)

— [UNISCOR] evaluation corpus we are presenting (built independently)

Evaluation experiment 1

We compare results on the same system:

— Training + evaluation using [GQ1]

— Training using [GQ1] and evaluation using [UNISCOR]

Results from evaluation on [GQ1] using standard approach

precision	recall	F1-score
98.83	98.99	**98.91**

Results from evaluation using [UNISCOR]

precision	recall	F1-score
82.27	81.12	**81.64**

Evaluation experiment 2

Comparison of two systems.

- Training and evaluation with [GQ1]:
  - CRF: **98.86%**
  - BiLSTM + CRF: **98.38%**
- Training with [GQ1] and evaluation with [UNISCOR]:
  - CRF (lexicon + lexical features): F1 **81.64%**
  - BiLSTM + CRF (embeddings): F1 **74.09%**

[UNISCOR] provides a benchmark that can be used to compare different systems

Conclusions

— We presented our Unit segmentation approach which relies on Machine Learning

— We release a UNIt Segmentation CORpus [UNISCOR] as Open-source (CC-BY).

— UNISCOR can be used as benchmark to provide evaluation measurement for unit recognition.

— In future:

  — we extend the dataset to more units

  — we add more evaluation datasets (quantities and value segmentation)

Thank you