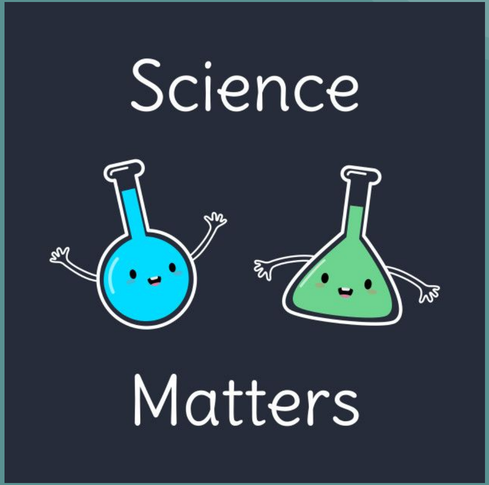
SemEval 2017 Task 10: Science IE

Extracting Keyphrases and Relations from Scientific Texts
Jennifer Storozum



The Task

Subtask (A): Identification of keyphrases

Given a scientific publication, the goal of this task is to identify all the keyphrases in the document.

Subtask (B): Classification of identified keyphrases

In this task, each keyphrase needs to be labelled by one of three types: (i) PROCESS, (ii) TASK, and (iii)

MATERIAL.

PROCESS
Keyphrases relating to some scientific model, algorithm or process should be labelled by PROCESS.

TASK
Keyphrases those denote the application, end goal, problem, task should be labelled by TASK.

MATERIAL
MATERIAL keyphrases identify the resources used in the paper.

The Task

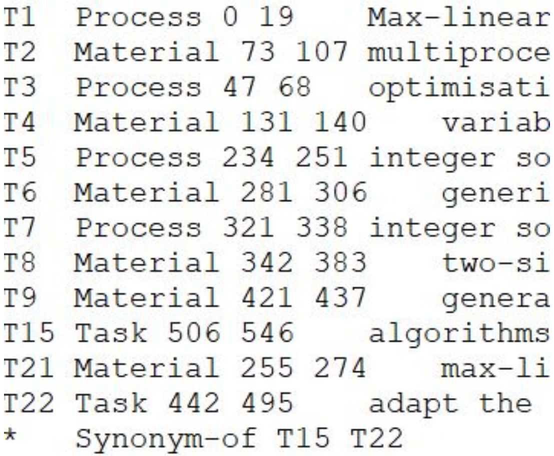
Subtask (C): Extraction of relationships between two identified keyphrases

Every pair of keyphrases need to be labelled by one of three types: (i) HYPONYM-OF, (ii) SYNONYM-OF, and (iii) NONE.

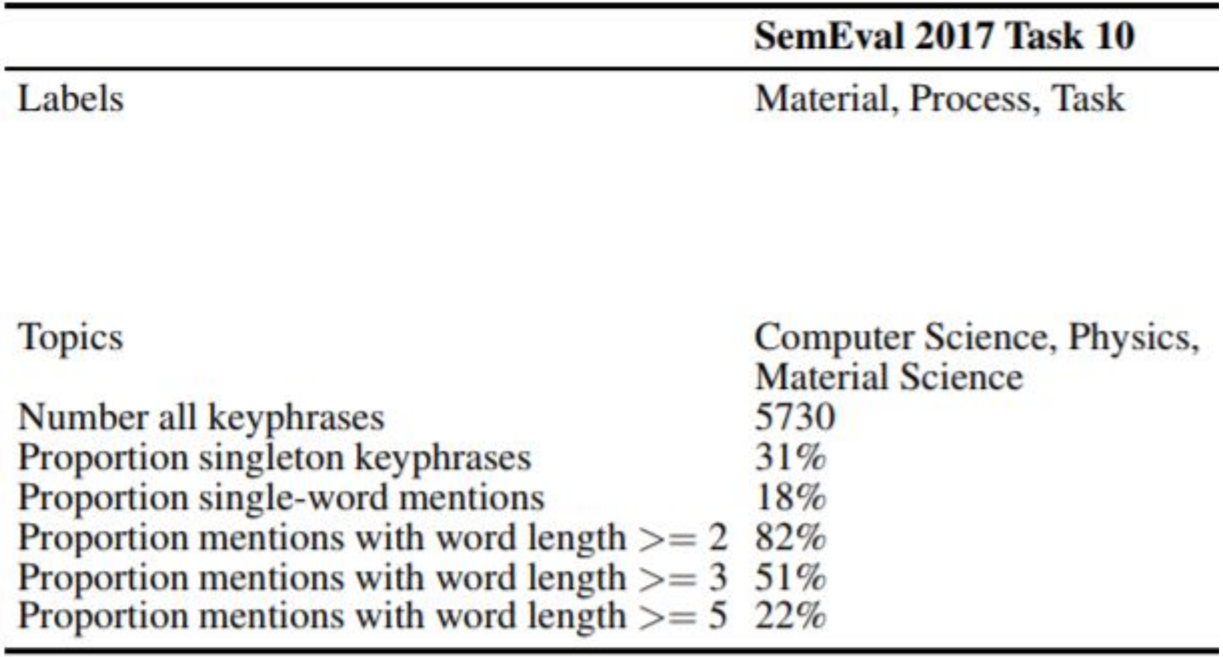
HYPONYM-OF
The relationship between two keyphrases *A* and *B* is HYPONYM-OF if semantic field of *A* is included within that of *B*. One example is *Red* HYPONYM-OF *Color*.

SYNONYM-OF
The relationship between two keyphrases *A* and *B* is SYNONYM-OF if they both denote the same semantic field, for example *Machine Learning* SYNONYM-OF *ML*.

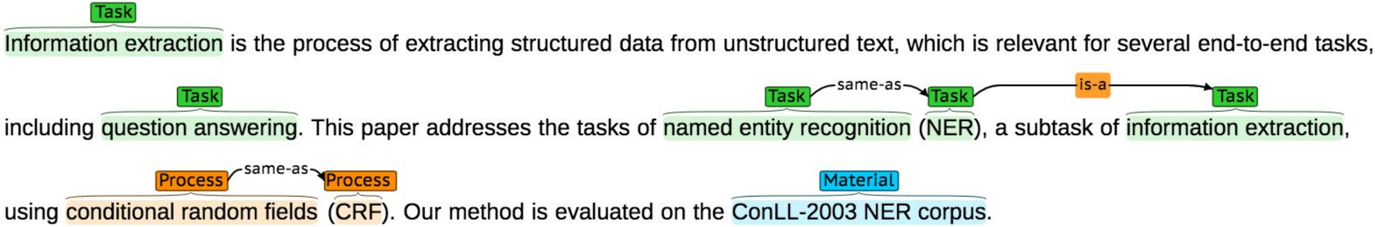
The Data



```
T1   Process 0 19     Max-linear
T2   Material 73 107 multiproce
T3   Process 47 68    optimisati
T4   Material 131 140     variab
T5   Process 234 251 integer so
T6   Material 281 306     generi
T7   Process 321 338 integer so
T8   Material 342 383     two-si
T9   Material 421 437     genera
T15  Task 506 546     algorithms
T21  Material 255 274     max-li
T22  Task 442 495     adapt the
*    Synonym-of T15 T22
```



	SemEval 2017 Task 10
Labels	Material, Process, Task
Topics	Computer Science, Physics, Material Science
Number all keyphrases	5730
Proportion singleton keyphrases	31%
Proportion single-word mentions	18%
Proportion mentions with word length >= 2	82%
Proportion mentions with word length >= 3	51%
Proportion mentions with word length >= 5	22%



Information extraction is the process of extracting structured data from unstructured text, which is relevant for several end-to-end tasks, including question answering. This paper addresses the tasks of named entity recognition (NER), a subtask of information extraction, using conditional random fields (CRF). Our method is evaluated on the ConLL-2003 NER corpus.

The Approach

BILOU tagging

CRF with Scikit learn CRF Suite

Feature engineering (w-1, w, w+1):

- Prefixes, suffixes (up to 4 letters)
- Upper/lower/titlecase, isDigit, isAlphaNum, contains AlphaNum
- POS tag
- Word length >= 2, 3, 4

Gazeteer: GO (Gene Ontology)

The Challenges

As usual, data heavily skewed by negative samples

The same span can (and often does) have more than one label
How to represent this? Examples:

```
T13  Material 835 848    simple metals
T14  Material 835 893    simple metals with sufficiently delocalized wave
functions

T2   Task 65 79     thermalization
T3   Process 65 79  thermalization

T12  Process 61 87  chemical vapour deposition
T14  Material 70 76 vapour
```

The Results - Dev Set

Material

	precision	recall	f1-score
B	0.476	0.365	0.413
I	0.339	0.308	0.323
L	0.585	0.448	0.507
U	0.776	0.232	0.357
avg	0.550	0.338	0.401

Process

	precision	recall	f1-score
B	0.454	0.334	0.385
I	0.292	0.278	0.285
L	0.515	0.376	0.435
U	0.432	0.182	0.256
avg	0.404	0.313	0.350

Task

	precision	recall	f1-score
B	0.284	0.180	0.220
I	0.340	0.184	0.238
L	0.338	0.211	0.260
U	0.000	0.000	0.000
avg	0.326	0.185	0.236

THANK YOU FOR COMING TO MY

SCIENCE

PARTY!