Extracting Key Phrases and Relations from Scientific Publications

Author: Thomas Clarke

Supervisor: Dr. Mark Lee

Main Goals

1

Provide a solution to ScienceIE (from SemEval 2017) – Natural Language Processing

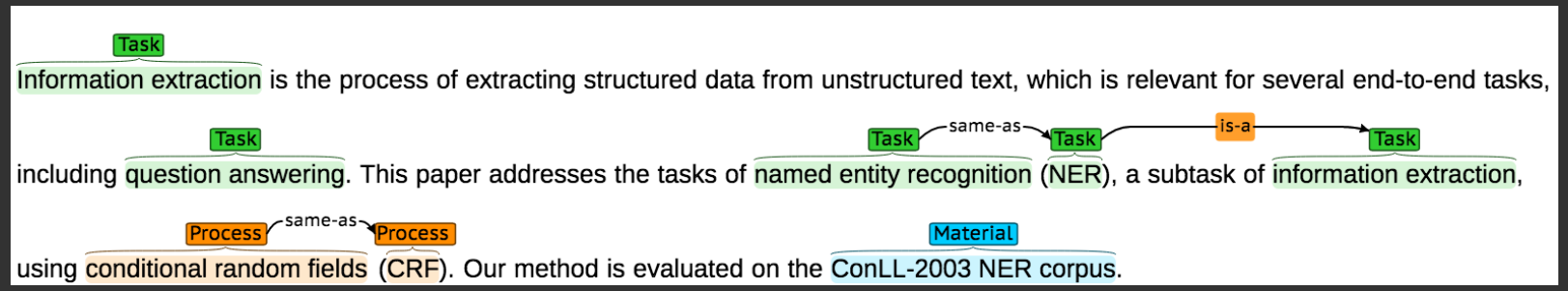
2

Produce a ‘product like’ proof of concept presenting a use of data produced

What is ScienceIE?

- Key Phrase (KP) identification (A)
- KP classification into (B)
  - Task
  - Process
  - Material
- Relation extraction (C)
  - Hyponym-of
  - Synonym-of

Example


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

Formatted Example

Original Text

In this paper a comparison between two popular feature extraction methods is presented. Scale-invariant feature transform (or SIFT) is the first method. The Speeded up robust features (or SURF) is presented as second. These two methods are tested on set of depth maps. Ten defined gestures of left hand are in these depth maps. The Microsoft Kinect camera is used for capturing the images [1]. The Support vector machine (or SVM) is used as classification method. The results are accuracy of SVM prediction on selected images.

BRAT Annotated Format

```
T1	Process 88 121	Scale-invariant feature transform
T2	Process 126 130	SIFT
T3	Process 157 183	Speeded up robust features
T4	Process 188 192	SURF
*	Synonym-of T1 T2
*	Synonym-of T3 T4
T5	Material 257 267	depth maps
T6	Process 398 420	Support vector machine
T7	Process 425 428	SVM
*	Synonym-of T6 T7
T8	Task 16 73	comparison between two popular feature extraction methods
T9	Process 332 355	Microsoft Kinect camera
T10	Process 492 495	SVM
T22	Task 47 65	feature extraction
T23	Process 441 455	classification
T27	Process 441 462	classification method
R2	Hyponym-of Arg1:T7 Arg2:T27
```

ScienceIE Data

- 50 development documents
- 350 training documents
- 100 test documents
- BRAT annotation format
- Short documents
  - Not full publications!

Background

- At ScienceIE
  - Support Vector Machine (SVM)
  - Neural Networks
    - Recurrent Neural Networks
    - Long Short Term Memory networks
  - Conditional Random Fields
  - Gazetters
- Other
  - Clustering
  - Rule based systems for relation extraction

Background – Word2Vec

- A vector space containing word vectors, with similar words being near each other
- E.g.: “knee is to leg” as “elbow is to [forearm, arm, ulna_bone]” (produced using Word2Vec)
- Several large pretrained models:
  - Google News (vocabulary size of 3 million different tokens)
  - Freebase (vocabulary size of 1.4 million)
  - Extension could be automatically build a model based on just scientific papers
    - All data from ScienceIE was too small to build a meaningful vector space
- I use the ‘Deep Learning for Java’ (DL4J) Word2Vec library

Project Architecture

- Java 8
- Maven 3
  - Supports importing libraries
  - Supports exporting this as a library (for part 2 of the project)

Preprocessing

- All data passed through Stanford's CoreNLP. This produces:
  - Sentences
  - Tokenisation
  - Parse trees of sentences
- I utilise Java serialisation to help cut down experiment run times

Key Phrase Extraction

- Support Vector Machine (libsvm implementation)
  - Radial Basis Function, $C = 50$, $\gamma = 0.5$
  - 12 support vectors
  - Training data tokens labelled as key phrases
  - Mainly based around position in text (and relation to other words)
  - Word2Vec distances and other token attributes (e.g. TF-IDF) also considered
  - Cross Validation to help with parameter optimisation
- Post processing
  - Remove badly formed key phrases
  - Remove low TF-IDF words

Key Phrase Extraction (Other)

- Also attempted hierarchal clustering
  - Bag-of-words approach
  - Word2Vec to define spacing
  - Didn’t work well…

Key Phrase Classification

- Word2Vec distances
  - Average and closest distance between bag-of-words tested
  - Option to ignore unimportant words
  - Default classification (catches words not in dictionary)
    - 221 couldn't be classified: 3 "task", 81 "process" and 137 "material"
    - Attempted to see if these were in WordNet but unfortunately not
- Also prove the position of words doesn't really give any indication of class (similar SVM to before)

Relation extraction

- Based on Word2Vec relative distances
- SVM
  - One for hyponyms (218 pairs), one for synonyms (207 sets)
    - Around 50,000 possible combinations
  - Relative distances between phrases
  - Angel and length of distance also tested between phrases
- Simple rules

Testing

- How to test key phrases?
- ScienceIE evaluation scripts use BRAT libraries
- My evaluation
  - Key phrase extraction comes in flavours:
    - Harsh (exact phrase matches)
    - Reasonable (matching gold and predicted phrases)
    - Generous (gold phrases are equal to or subphrases of predicted)
  - Classification compared gold to predicted (on gold key phrase extraction)
  - Relation correctness between gold key phrases

Results – ScienceIE Scripts (F1 Scores)

Section	ScienceIE		My Evaluation	
	Individual Best / Average	End-To-End Best / Average	Individual Best	End-To-End Best
KP Extraction	0.56 / 0.38	0.56 / 0.38	0.20	0.20
KP Classification	0.67 / 0.57	0.44 / 0.26	0.55	0.11
Relation extraction	0.64 / 0.43	0.28 / 0.07	0.1	0.02
Overall	N/A	0.43 / 0.25	N/A	0.11

Results – My Evaluation

- Key phrase extraction F1 scores:
  - Very strict: 0.2 (I agree with the ScienceIE score)
  - When matching on just the phrase: 0.36
  - When being generous: 0.74

The Product: The Website: ExtractorIE

- Java based web project, with the NLP discussed as a library through Maven
- MySQL database
- Uses Spring Boot 1.5
  - Produces convenient self-contained jar
    - Integrated Apache Tomcat
  - Useful libraries for MySQL
    - Hibernate
- Interesting Visualisations:
  - Donut chart (using d3.js)
  - Word Cloud (using jqcloud.js)
- Custom search prioritising search results with key phrases connected to query

See ExtractorIE

http://tomclarke.xyz

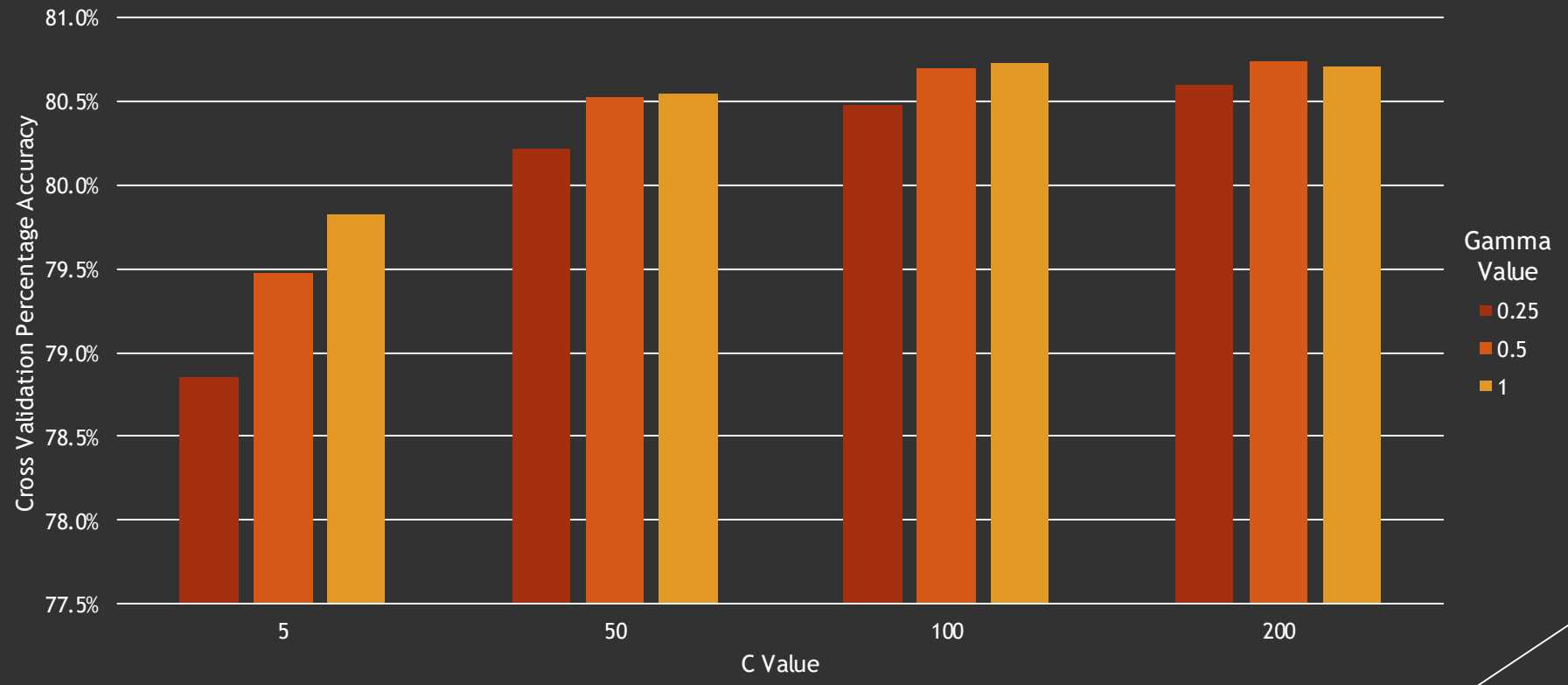
Thank you for listening!

Any questions?

Key Phrase Extraction SVM: Support Vectors

1. Word length / max word length
2. Part-Of-Speech type
3. TF-IDF
4. Whether the word is a stop word
5. Depth in document (position / total tokens)
6. Depth in sentence
7. Whether the token is in the first sentence
8. Whether the token is in the last sentence
9. Parse tree depth
10. Whether the previous token was a key phrase
11. Similarity to “task”
12. Similarity to “process”
13. Similarity to “material”

Key Phrase Extraction SVM: Cross Validation



ExtractorIE

Extracting Key Phrases and Relations from Scientific Publications

For more information on what the project is based upon, see the ScienceIE shared task.

Search the database | Add a new paper

8	14
Hyponyms	Synonyms

Search for papers

Enter terms to search for and focus on task, process or material related items:

Search terms or phrases

Submit

Focus on: Task Process Material

Search for papers

Enter terms to search for and focus on task, process or material related items:

support vector machine Submit

Focus on: Task Process Material

Search for "support vector machine" completed in 0.257 seconds, finding 7 papers.

ID	Paper Title	KPs / Rels
63	**S221267161400105X** Kinect camera is used for capturing the images [1]. The **support vector machine** (or SVM) is used as classif...	15 / 0
84	**S1877750315000460** and build HemeLB on any remote resource, to reuse **machine**-specific configurations, and to organize a...	22 / 6
26	**S0021999112003579** We order the discrete unknowns so that the **vector** of unknowns, xPS=[X,L], contains the nx unk...	14 / 0
70	**S0098300414000259** event that future versions of Hadoop are optimized to **support** paradigms other than MapReduce, Pig script...	19 / 0
81	**S107158191630074X** the simulation (e.g., the input buffer of a certain **machine** at the time of refilling it). A relatively...	22 / 0
97	**S002002551630384X** hulls, such as Vizster [22]. However, they do not **support** visualizing set overlaps. ...	12 / 0
18	**S1877750313001269** is itself very important. Grid and cloud computing **support** different interaction models. In grid comp...	20 / 0

Search for papers

Enter terms to search for and focus on task, process or material related items:

test Submit

Focus on: Task Process Material

Search for "test" completed in 0.613 seconds, finding 19 papers.

ID	Paper Title	KPs / Rels
6	**S0263822312000657** bamboo shows that they are all graded with their grea**test** strength on the outside, in areas where the g...	18 / 1
61	**S0301679X14003272** Q, is measured and recorded throughout the entire **test** by a piezoelectric load cell which is connect...	19 / 0
68	**S0021999113006955** The **test** cases confirm that the high-order discretisat...	34 / 0
79	**S0370269303017222** making the predicted Roper mass heavier than the ligh**test** negative parity baryon mass. Pairwise spin-de...	19 / 0
91	**S0011227514002136** thermometry (which was not available at the time of **test**ing but which will be used for the mKCC), as t...	13 / 0
85	**S0021999113004555** Three Runge–Kutta IMEX schemes were **test**ed by Ullrich and Jablonowski [23] for the HEV...	19 / 0
87	**S0257897213004131** Fig. 7 shows the relationship between the **test**ing time and friction coefficients of various ...	19 / 0
89	**S0011227515000648** #1/#2 and the others is the most influential on the **test** results. The redesign and upgrade to 110-nm p...	16 / 0
98	**S0301679X14000449** RH ceramics was smaller than Sc,critical under all **test**ed conditions during the initial stage of fric...	28 / 0
63	**S221267161400105X** SURF) is presented as second. These two methods are **test**ed on set of depth maps. Ten defined gestures ...	15 / 0

Show 9 more

Search for papers

Enter terms to search for and focus on task, process or material related items:

test Submit

Focus on: ☑ Task ☐ Process ☐ Material

Search for "test" completed in 0.597 seconds, finding 19 papers.

ID	Paper Title	KPs / Rels
6	**S0263822312000657** bamboo shows that they are all graded with their grea**test** strength on the outside, in areas where the g...	18 / 1
61	**S0301679X14003272** Q, is measured and recorded throughout the entire **test** by a piezoelectric load cell which is connect...	19 / 0
85	**S0021999113004555** Three Runge–Kutta IMEX schemes were **test**ed by Ullrich and Jablonowski [23] for the HEV...	19 / 0
98	**S0301679X14000449** RH ceramics was smaller than Sc,critical under all **test**ed conditions during the initial stage of fric...	28 / 0
63	**S221267161400105X** SURF) is presented as second. These two methods are **test**ed on set of depth maps. Ten defined gestures ...	15 / 0
68	**S0021999113006955** The **test** cases confirm that the high-order discretisat...	34 / 0
44	**S0370269304012638** SU(N?1)×U(1). There have appeared two independent F6 **test**s of this conjecture [19,20], with conflicting...	15 / 0
79	**S0370269303017222** making the predicted Roper mass heavier than the light**test** negative parity baryon mass. Pairwise spin-de...	19 / 0
91	**S0011227514002136** thermometry (which was not available at the time of **test**ing but which will be used for the mKCC), as t...	13 / 0
28	**S221450951400014X** According to Fig. 2 and the results of the Marshall **test**s, the optimum bitumen measures decrease signi...	28 / 0

Show 9 more

S0927025615006357

Download Paper | Download Extractions

> In this paper, crystal plasticity model, in combination with XFEM, has been applied to study cyclic deformation and fatigue crack growth in a nickel-based superalloy LSHR (Low Solvus High Refractory) at high temperature. The first objective of this research was to develop and evaluate a RVE-based finite element model with the incorporation of a realistic material microstructure. The second objective of this work was to determine the parameters of a crystal plasticity constitutive model to describe the cyclic deformation behaviour of the material by using a user-defined material subroutine (UMAT) interfaced with the finite element package ABAQUS. The model parameters were calibrated from extensive finite element analyses to fit the monotonic, stress relaxation and cyclic test data. The third objective was to predict crack growth by combining the XFEM technique and the calibrated crystal plasticity UMAT, for which accumulated plastic strain was used as the fracture criterion.

Key Phrases	
crystal plasticity model	Material
XFEM	Material
cyclic deformation	Material
superalloy LSHR	Material
high temperature	Material
objective of this research	Task
element model	Process
ncorporation	Material
material microstructure	Material
objective of this work	Task
parameters	Process
crystal plasticity constitutive model	Process
cyclic deformation behaviour of the material	Material
material subroutine (UMAT) interfaced	Material
finite element package ABAQUS	Process
model parameters	Process
calibrated	Process
element analyses	Material
monotonic	Material
cyclic test dat	Process
objective	Task
growth	Process
XFEM technique	Process


Key Phrases
tomclarke.xyz/kps
ExtractorIE
Home
Search for a paper
Key Phrases
Add a paper
All found key phrases, grouped by classification.
To generate the below graphs, key phrases are broken up into tokens, and the size of a token is relative to it's TF-IDF value.
dimensional
rgument
gale
rchaeological
self
bargmann
brane
squares
first
mages
slider
spring
duality
bsorbance
nucleus
camer
loop
nitial
vs
ssue
dashpot
nformal
nt
spectr
ads
uncert
pre
ccuracy
three
ndustry
principles
ll
su
least
stigmatism
processing
multi
View all Task key phrases
nterpretation
lisflood
na
http
www
kiber
assembly
nput
mplementations
ssessment
spe
doption
cm
although
mol
co
solid
cryo
su
rea
rving
de
htm
fp
step
com
tems
bility
spin
aa
al
energy
brillouin
nalysis
ccuracy
az
cumg
automat
improved
redshift
mplicit
nitial
ntensity
nfrastructure
kirkwood
ssociated
high
ccomplished
half
mproved
fferent
pseudo
re
pproach
distinct
View all Process key phrases
transconductance
benzyl
ggregate
pre
nformal
sigma
distance
catalysts
quasi
ends
algebr
early
nodizing
dvantages
full
xial
silic
ii
leading
short
dea
nput
dat
ndividual
poly
resin
non
self
di
rtefacts
grown
oriented
ntigens
wave
nitial
catalyzed
lloy
gluon
dinger
zn
nelasticity
mvs
gain
iii
pplication
el
az
assembles
complementary
posterior
coord
starchy
aa
medi
hydroxyethylpyrrolidinonium
nterface
community
View all Material key phrases