Spring 2024 CS7.401

Intro to Natural Language Processing

Project Report on

Extracting Key-phrases and Relations from Scientific Publications

By Team

Semantics Squad

Sahil Patel (2023201081)

Harsh Shah (2023202028)

Shubham Kathiriya (2023201050)

Table Of Contents

1. Abstract

This project revolves around the automated extraction of key phrases from scientific publications, with a focus on labelling and establishing relationships between these phrases. Within the realm of scientific literature, key phrases pertaining to PROCESS, TASK, and MATERIAL constitute fundamental objects. Consequently, the project aims to derive hyponym and synonym relations between keyphrases.

2. Objectives

To achieve all goals, we divided entire project on three tasks.

1. Key-phrase Extraction
2. Named Entity Representation (NER)
3. Hyponym-Synonym Discovery

3. Introduction

3.1 Dataset

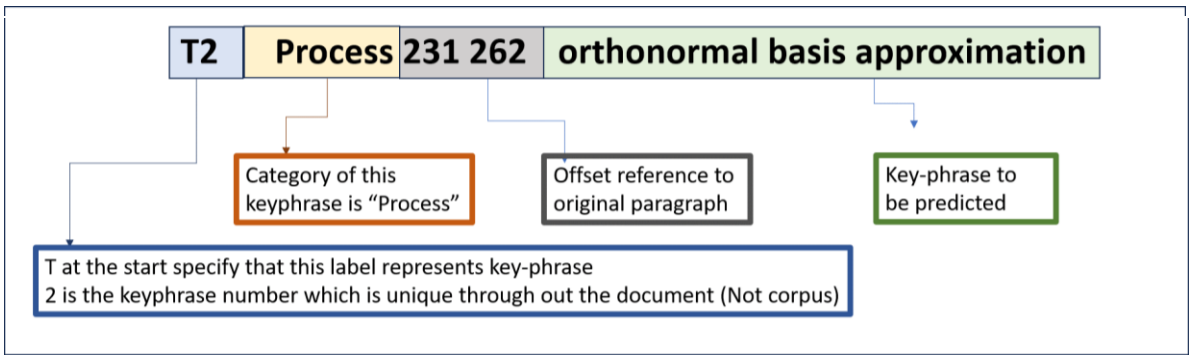
We are using very popular, structured and **Labelled Dataset** named **SeamEval2017**. This dataset is used by hundreds of research paper as considering it for standard benchmarking.

Given Dataset has three partitions – Train, Val, Split

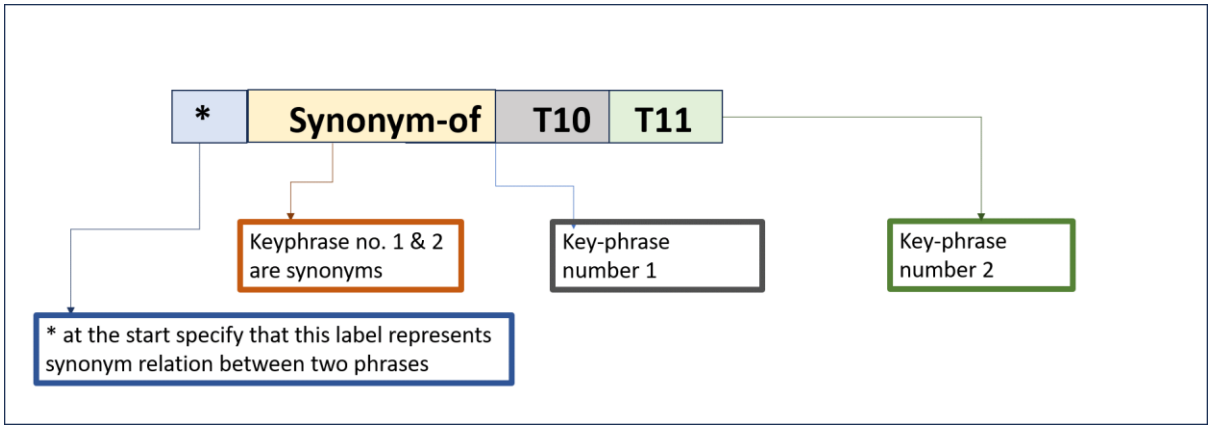
Split	Total Examples (paragraphs)
Train	350
Val	50
Test	100

Example of Dataset

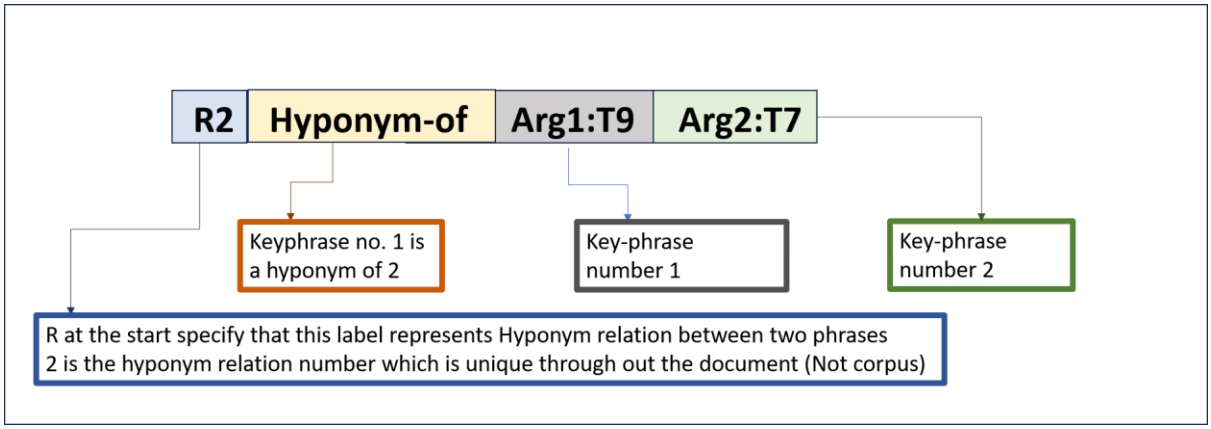
Given Dataset has two types of file. “.txt” format files have paragraphs from research materials. “.ann” format files contains labels of corresponding paragraphs. “.ann” has three types of labels.



Label Type 1: Represents Keyphrase and its Category



Label Type 2: Represents Synonym relation between two key-phrases



Label Type 3: Represents Synonym relation between two key-phrases

3.2 Task 1 – Key-phrase extraction

Task 1 aims to predict all possible key-phrases from given paragraph. We reviewed lots of **unsupervised** and **supervised techniques** as well as **prebuilt tools & libraries** for this task.

Prebuilt Tools:

- PKE (Python based keyphrase extraction toolkit)
- KeyBERT

Unsupervised Techniques:

- TextRank
- PageRank
- PositionRank
- SentenceRank
- HyperRank
- PromptRank (Highest scores among all Unsupervised Techniques)

Supervised Techniques:

- Using Deep bi-LSTM
- Positive Unlabeled Learning
- Using Pretrained LLMs (BERT, T5)

We can use prebuilt tools like PKE and keyBERT with few lines of code and can achieve score nearby SOTA for this task. But it is against the learning objectives of project, So we decided to not go with them.

We explored variety of Unsupervised Techniques for this task and gone through research paper and other literature of it. From which we found out that TextRank, PageRank and Sentence Ranks are more suitable for keyword extraction not keyphrase extraction. HyperRank and PromptRank can achieve highest scores among all Unsupervised Techniques, they have very complex implementations and take more time to code. We decided to go with PositionRank due to less complexity of algorithm and giving satisfactory score.

But the score was not up to the mark. So we went for Supervised Techniques and found out that training a bi-LSTM from our data only will not give enhanced score which we got from PositionRank. That's why we choose to use pretrained LLM like BERT, T5 for this task. We finetuned BERT on our corpus and achieved highest accuracies among all of our experiments.

3.3 Task 2 – Named Entity Representation

Given three categories for key-phrase **Process, Task & Material**, Task 2 aims to predict that, in which category the extracted keyphrase belongs to.

Technique: Trained bi-LSTM

We found out this task identical to Named Entity Representation (NER) Task. We trained bi-LSTM on the corpus.

3.4 Task 3 – Hyponym-Synonym discovery

Task 3 aims to predict weather predicted keyphrase is Hyponym or Synonym of some other key-phrase predicted for same paragraph or not.

We looked into many resources but not get any decent paper or technique for it. We observed that nothing we can do more than three class classification for each pair of the keyphrase predicted from single paragraph. The main problem here is there is very high inequality between datapoints. We have only few examples of hyponyms and synonyms.

Split	Total Synonyms	Total Hyponyms	No Relation
Train	263	532	71704
Val	41	222	15551
Test	135	179	23485

The task is too complex that we can not train model from scratch, so we decided to go with pretrained LLM BERT and finetune over limited data. We reduced count of “No Relation” points to be nearby equal to Synonyms and Hyponyms.

Split	Total Synonyms	Total Hyponyms	No Relation
Train	263	532	700
Val	41	222	100
Test	135	179	200

4. Experiments & Results

4.1 Task 1 – Key Phrase Extraction

Test Accuracies using PositionRank (top_k) = 15

Precision @k		Recall @k		F1-score @k	
Pr@1	0.026	Re@1	0.002	F1@1	0.004
Pr@2	0.034	Re@2	0.006	F1@2	0.010
Pr@3	0.034	Re@3	0.009	F1@3	0.013
Pr@4	0.034	Re@4	0.011	F1@4	0.016
Pr@5	0.036	Re@5	0.014	F1@5	0.019
Pr@6	0.035	Re@6	0.017	F1@6	0.021
Pr@7	0.035	Re@7	0.019	F1@7	0.023
Pr@8	0.034	Re@8	0.021	F1@8	0.025
Pr@9	0.033	Re@9	0.023	F1@9	0.025
Pr@10	0.032	Re@10	0.025	F1@10	0.026
Pr@11	0.032	Re@11	0.027	F1@11	0.028
Pr@12	0.032	Re@12	0.029	F1@12	0.029
Pr@13	0.033	Re@13	0.032	F1@13	0.031
Pr@14	0.032	Re@14	0.033	F1@14	0.031
Pr@15	0.032	Re@15	0.036	F1@15	0.032

Validation Accuracies using pretrained BERT

Best Validation Precision: 0.5348

Best Validation recall: 0.6076

Best Validation **f1**_score: **0.5614**

Testing Accuracies using pretrained BERT

Test Precision: 0.4912

Test recall: 0.5924

Test **f1**_score: **0.5371**

4.2 Task 2 – Named Entity Representation

We used bi-LSTM, trained it on dataset and got below scores

```
Accuracy is  0.6179337231968811
              precision    recall  f1-score   support

         0.0       0.67      0.48      0.56       954
         1.0       0.42      0.44      0.43       194
         2.0       0.62      0.81      0.71       904

    accuracy                           0.62      2052
   macro avg       0.57      0.58      0.56      2052
weighted avg       0.63      0.62      0.61      2052
```

Classification report for Task-2,
here 0.0 denotes Process, 1 .0 denotes Task & 2.0 denoted Material

The highest F1-score of the competition was 0.64 which they got with ensemble of three to four different models, comparatively our F1-score is 0.61 which seems fair to us.

Combine Inference Example for Task 1 & 2:

```
Abstract:  Fig. 9 displays the growth of two of the main corrosion products that
develop or form on the surface of Cu40Zn with time, hydrozincite (Fig. 9a) and Cu2O
(Fig. 9b). It should be remembered that both phases were present already from start
of the exposure. The data is presented in absorbance units and allows comparisons to
be made of the amounts of each species between the two Cu40Zn surfaces investigated,
DP and HZ7. The tendency is very clear that the formation rates of both hydrozincite
and cuprite are quite suppressed for Cu40Zn with preformed hydrozincite (HZ7)
compared to the diamond polished surface (DP). In summary, without being able to
consider the formation of simonkolleite, it can be concluded that an increased
surface coverage of hydrozincite reduces the initial spreading ability of the NaCl-
containing droplets and thereby lowers the overall formation rate of hydrozincite and
cuprite.
```

```
Key-Phrases:
Material   |     cuprite
Material   |     cup
Material   |     cu2o
Material   |     dp
Material   |     hydrozincite
Material   |     nacl
Material   |     containing droplets
Material   |     surface coverage
Material   |     cu40zn
Material   |     ##ance
Task       |     corrosion products
Material   |     cu40zn surfaces
Process    |     formation
Task       |     ##formed hydrozincite
Material   |     diamond polished surface
```

4.3 Task 3 – Hyponym-Synonym Discovery

We used BERT for this task. Below are the scores

	precision	recall	f1-score	support
0	0.61	0.66	0.64	200
1	0.54	0.71	0.61	135
2	0.66	0.45	0.53	179
accuracy			0.60	514
macro avg	0.60	0.61	0.59	514
weighted avg	0.61	0.60	0.59	514

Inference Example:

```
absorbance units, cu2o --> Hyponym
absorbance units, cu40zn --> Hyponym
```

5. References

KeyBERT: https://github.com/MaartenGr/KeyBERT

PKE: https://github.com/boudinfl/pke

PromptRank: https://aclanthology.org/2023.acl-long.545.pdf

HyperRank: https://aclanthology.org/2023.emnlp-main.997.pdf

PositionRank: https://aclanthology.org/P17-1102.pdf

Pretrained BERT: https://huggingface.co/docs/transformers/en/model_doc/bert