Extracting Keyphrases and Relations from Scientific Publications

Narendar Singh
Rohit Joshi
Aadya Ranjan

Introduction

A keyphrases is a simplex word or a sequence of words that characterizes the theme and content of a document.[You et al., 2013] Research papers are crucial in the field of Research, however the expanding digital libraries of research papers pose challenges for online searches to cater to individual needs effectively. However, a thoughtfully curated list of keyphrases can enhance search precision, enabling readers to swiftly determine the relevance of documents to their specific domains of interest. Extracting key phrases and relationships from scientific papers is a vital task in natural language processing. This process aids researchers in swiftly identifying pertinent articles and extracting meaningful insights.

This can be broken down into three main sub-problems:

- Key Phrase Extraction
- Keyword Classification
- Relationship Identification

Initially, the focus is on identifying and extracting key phrases from the text. These phrases are then categorized into three distinct types: PROCESS, TASK, or MATERIAL. Finally, the relationships between these categorized phrases are identified.

Corpus Details

We utilized the SemEval 2017 Task 10 dataset[Augenstein et al., 2017], which comprises scientific articles from various fields such as Chemistry, Computer Science, and Physics. The dataset is divided into training, development, and test sets, each containing .txt files with text excerpts and corresponding .ann files with keyword annotations.Annotations include keyword boundaries and types, along with some relationships labeled as Synonym-of and Hyponym-of.

Methodology

Pre-processing of the Data

1. For subtask 1 and subtask 2, our objectives are clear, in subtask 1, we determine whether a given token qualifies as a Keyword, while in subtask 2, we classify these keywords into one of three types: Task, Material, or Process. To better our approach, we've merged both tasks, considering their computational demands and the optimization of resources. Implementing two separate models would be inefficient. To preprocess the dataset for both subtasks, we've adopted the BIO scheme, utilizing chunk representation. In this scheme, each word in the text is tokenized and assigned a label from a set of seven possible labels. This strategy enables us to efficiently handle the classification tasks at hand.Please see the Table 1 for description

Description	**Label**
O	Not a Keyphrase/Keyword
B-Process	Beginning of the Keyphrase of type Process
I-Process	Inside of the Keyphrase of type Process
B-Task	Beginning of the Keyphrase of type Task
I-Task	Inside of the Keyphrase of type Task
B-Material	Beginning of the Keyphrase of type Material
I-Material	Inside of the Keyphrase of type Material

Table 1: Labels(12)

2. In relationship identification(subtask 3), we've extracted the relation between keyphrases from the annotation file. These relationship fall into two categories: Synonym-of and Hyponym-of. Synonym-of denotes a bidirectional relationship, while Hyponym-of is unidirectional. To address this, we've treated the absence of a reverse relationship in Hyponym-of as a distinct label, resulting in three labels overall,please see the table below.

Labels	Description
0	No Relation
1	Hyponym-of
2	Synonym-of

Table 2: Labels(For task 3)

Extracting and Classifying Keywords and Keyphrases

Keywords and keyphrases extraction and classification involved utilizing the SciBERT pretrained model, a specialized variant of BERT trained on scientific data. We fine-tuned this model on the corpus, treating the task similar to Named Entity Recognition (NER) and thus transforming it into a Token Classification task.

For implementation, we employed the SciBERT tokenizer from the AutoTokenizer package and loaded the pretrained SciBERT model using the AutoModelForTokenClassification package. Fine-tuning was conducted over 10 epochs,resulting in validation accuracy as depicted in the provided figure.

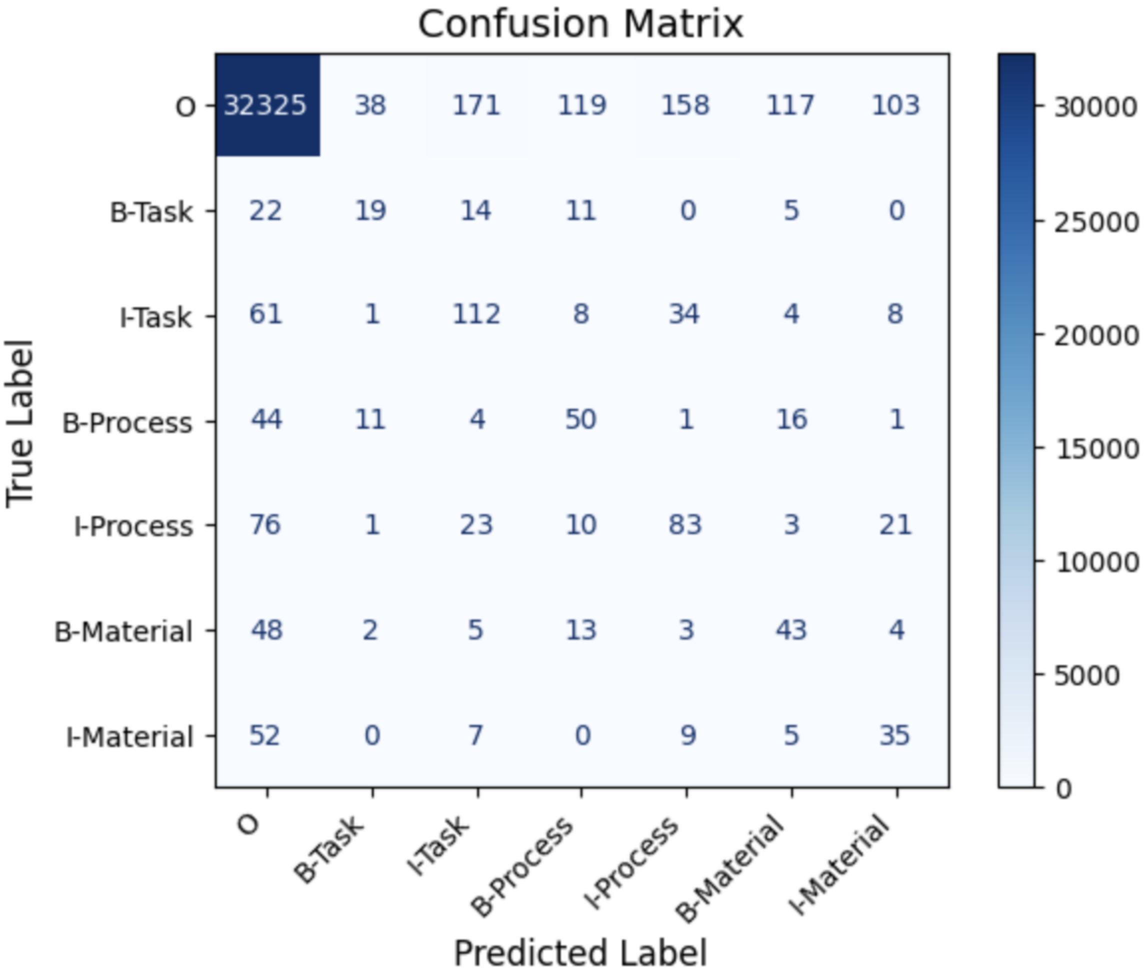
Identifying Relationships

We utilized sentence-transformers to generate embeddings for each phrase, resulting in embedding vectors of size 384 for each entity. These embeddings were concatenated to create a combined vector of size 768. The task then became a supervised classification problem, for which we employed SVM classifiers to assign one of three labels to each entity pair. Using MiniLM Sentence Transformer as our text encoder, we fed both entities into the model's encode method to obtain their encodings. These encodings were then merged based on their relationship and organized into a data frame for training and testing. Subsequently, we trained the SVM classifier on the training data and evaluated the predicted results using appropriate metrics.

Evaluation and Analysis

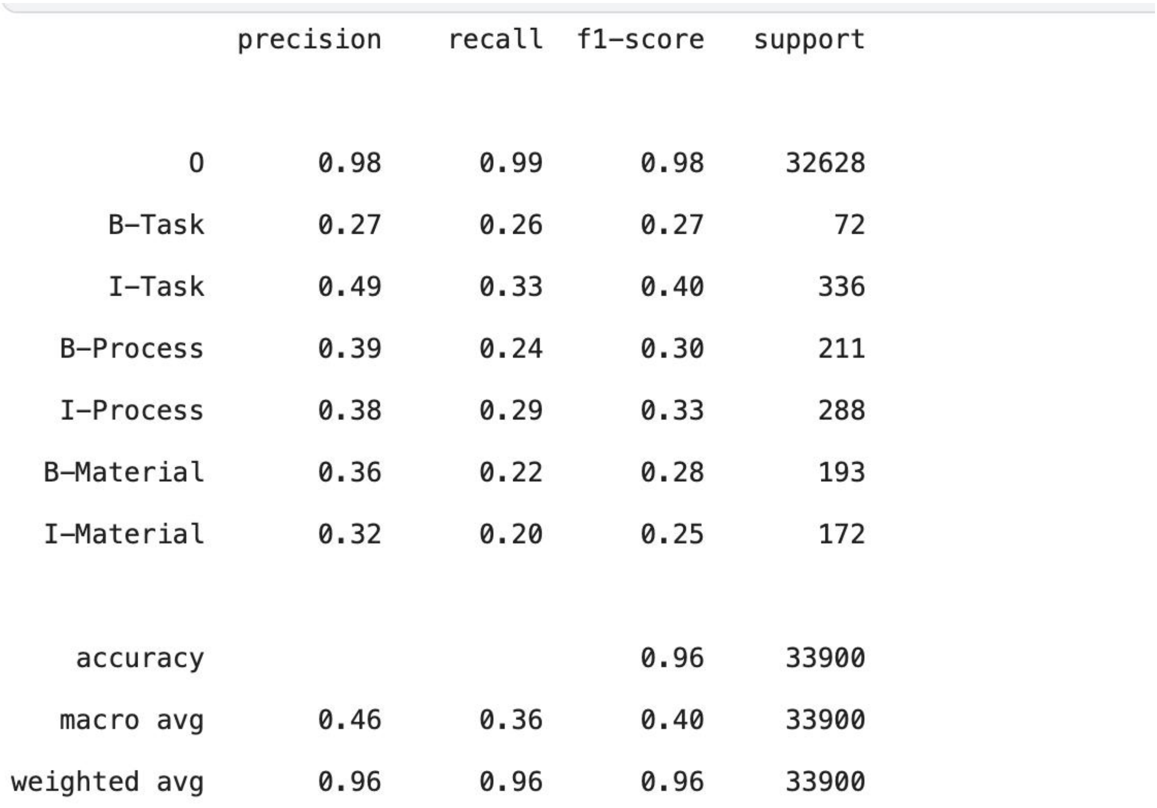
Confusion Matrix Task 1 and 2

```
Confusion Matrix :
```



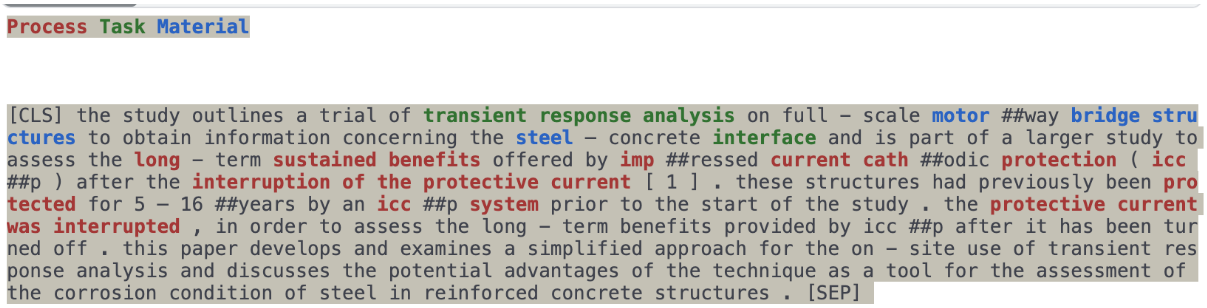
True Label \ Predicted Label	O	B-Task	I-Task	B-Process	I-Process	B-Material	I-Material
O	32325	38	171	119	158	117	103
B-Task	22	19	14	11	0	5	0
I-Task	61	1	112	8	34	4	8
B-Process	44	11	4	50	1	16	1
I-Process	76	1	23	10	83	3	21
B-Material	48	2	5	13	3	43	4
I-Material	52	0	7	0	9	5	35

Evaluation Task 1 and 2



```
              precision    recall  f1-score   support

           O       0.98      0.99      0.98     32628
      B-Task       0.27      0.26      0.27        72
      I-Task       0.49      0.33      0.40       336
   B-Process       0.39      0.24      0.30       211
   I-Process       0.38      0.29      0.33       288
  B-Material       0.36      0.22      0.28       193
  I-Material       0.32      0.20      0.25       172

    accuracy                           0.96     33900
   macro avg       0.46      0.36      0.40     33900
weighted avg       0.96      0.96      0.96     33900
```

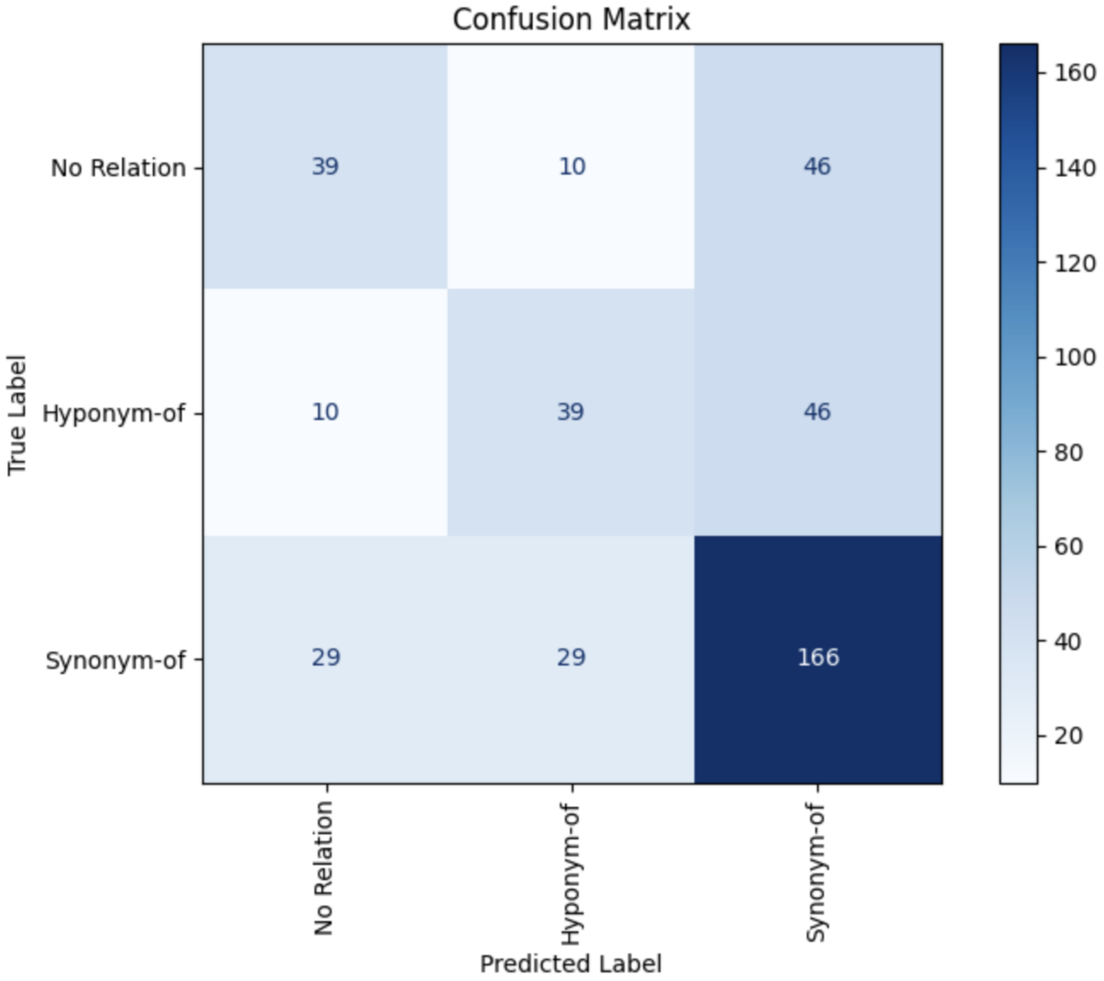
Result Task 1 and 2



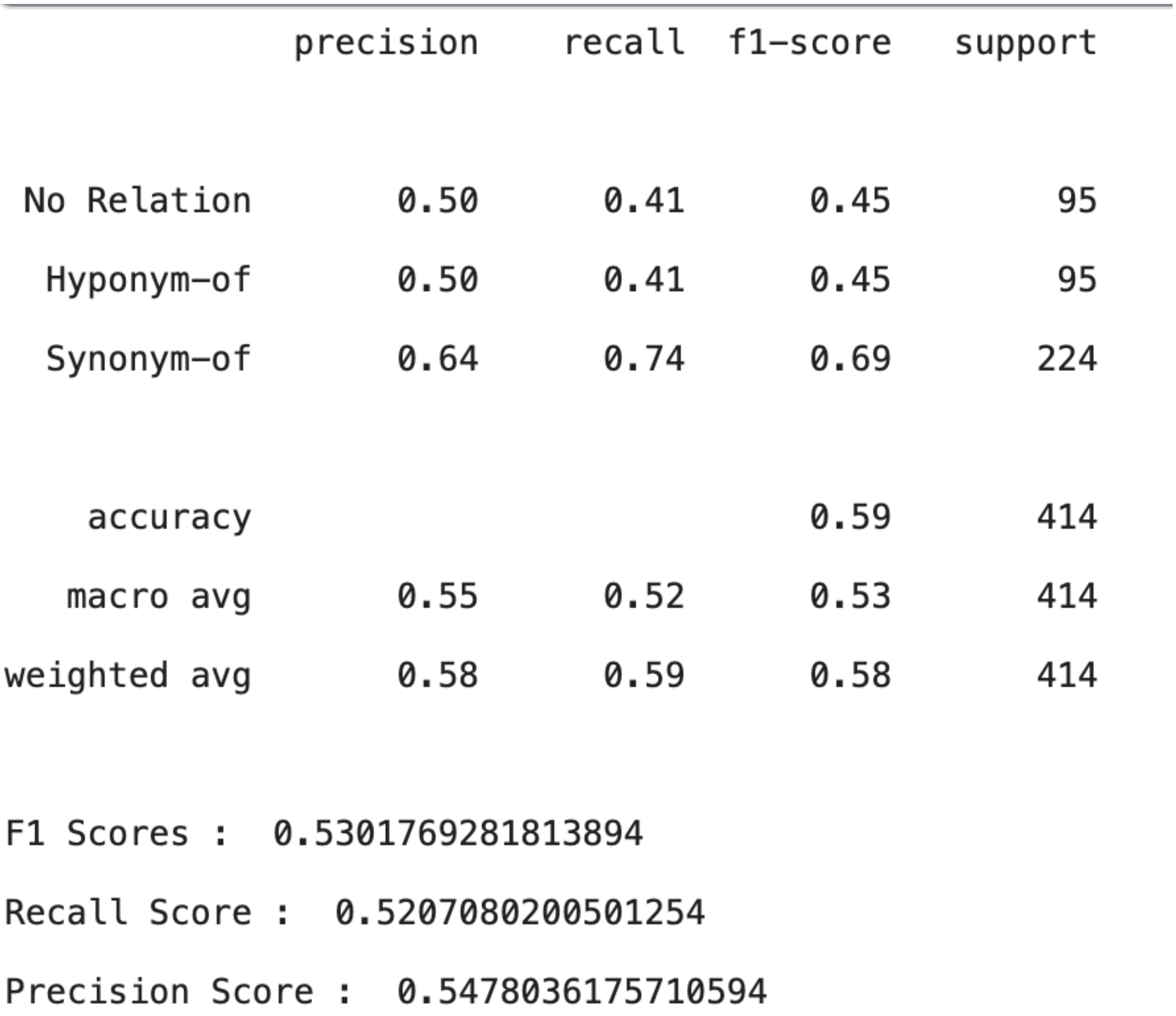
```
Process Task Material


[CLS] the study outlines a trial of transient response analysis on full – scale motor ##way bridge stru
ctures to obtain information concerning the steel – concrete interface and is part of a larger study to
assess the long – term sustained benefits offered by imp ##ressed current cath ##odic protection ( icc
##p ) after the interruption of the protective current [ 1 ] . these structures had previously been pro
tected for 5 – 16 ##years by an icc ##p system prior to the start of the study . the protective current
was interrupted , in order to assess the long – term benefits provided by icc ##p after it has been tur
ned off . this paper develops and examines a simplified approach for the on – site use of transient res
ponse analysis and discusses the potential advantages of the technique as a tool for the assessment of
the corrosion condition of steel in reinforced concrete structures . [SEP]
```

Confusion Matrix Task 3

```
Confusion Matrix:
```



Evaluation Task 3



```
              precision    recall  f1-score   support


 No Relation       0.50      0.41      0.45        95

  Hyponym-of       0.50      0.41      0.45        95

  Synonym-of       0.64      0.74      0.69       224


    accuracy                           0.59       414

   macro avg       0.55      0.52      0.53       414

weighted avg       0.58      0.59      0.58       414


F1 Scores :  0.5301769281813894

Recall Score :  0.5207080200501254

Precision Score :  0.5478036175710594
```

References

[Augenstein et al., 2017] Augenstein, I., Das, M., Riedel, S., Vikraman, L., and McCallum, A. (2017). Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

[You et al., 2013] You, W., Fontaine, D., and Barthès, J.-P. (2013). An automatic keyphrase extraction system for scientific documents. *Knowledge and information systems*, 34(3):691–724.