# EXTRACTING KEYPHRASES AND RELATIONS FROM SCIENTIFIC PUBLICATIONS

Narendar Singh
2023201018
Rohit Joahi
2023202016
Aadya Ranjan
2023814001

Introduction to NLP

# Outline


1. Introduction
2. Corpus Details
3. Data Pre-processing
4. Methodology
5. Evaluation

# Introduction

- Keyphrases are crucial for summarizing the theme and content of a document.
- Despite the importance of research papers, the proliferation of digital libraries poses challenges for effective online searches.
- Curating a list of keyphrases enhances search precision, aiding readers in quickly assessing document relevance.

# Introduction

- Extracting key phrases and relationships from scientific papers is vital in natural language processing, involving sub-tasks such as :

# Introduction

- Extracting key phrases and relationships from scientific papers is vital in natural language processing, involving sub-tasks such as :
  - Key Phrase Extraction

# Introduction

- Extracting key phrases and relationships from scientific papers is vital in natural language processing, involving sub-tasks such as :
  - Key Phrase Extraction
  - Keyword Classification

# Introduction

- Extracting key phrases and relationships from scientific papers is vital in natural language processing, involving sub-tasks such as :
  - Key Phrase Extraction
  - Keyword Classification
  - Relationship Identification

# Corpus Details

- Utilized SemEval 2017 Task 10 dataset, consisting of scientific articles from Chemistry, Computer Science, and Physics.
- Dataset divided into training, development, and test sets, each containing .txt files with text excerpts and .ann files with keyword annotations.

# Data Pre-processing

- Merged subtasks 1 and 2 to determine token as Keyword and classify it into Task, Material, or Process types. Utilized BIO scheme for dataset preprocessing, assigning each word a label from seven possible labels, see below.

| **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)

# Data Preprocessing

- Extracted relations between keyphrases from annotation file in subtask 3 with labels, see below.

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

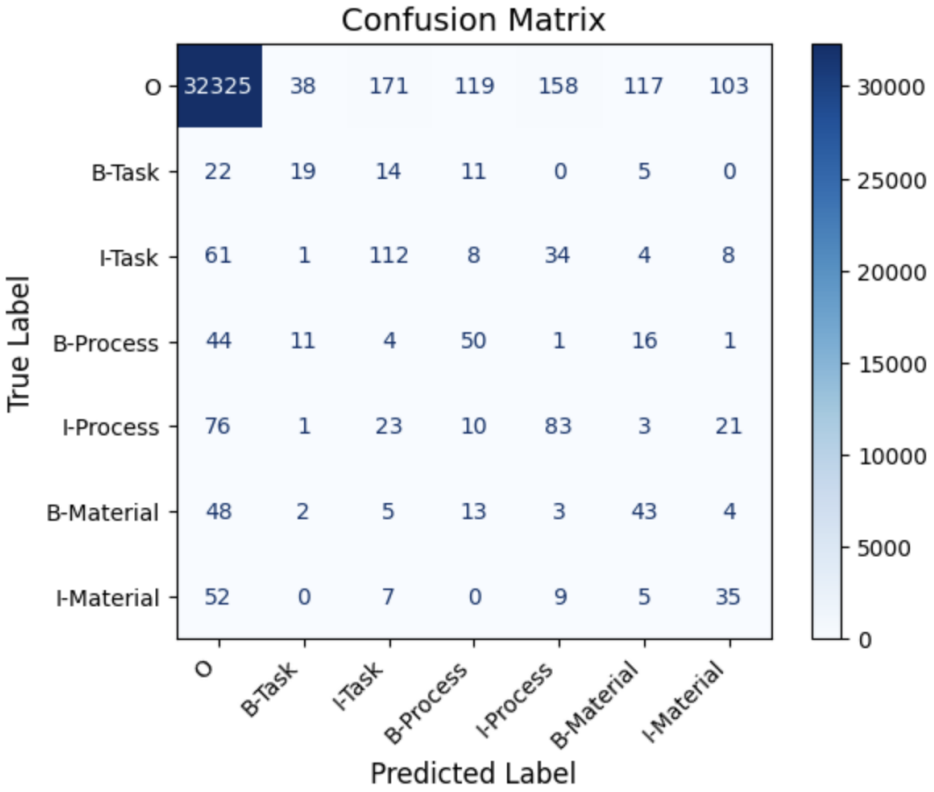
Table 2: Labels(For task 3)

# Methodology

- Subtask 1 (Keyword extraction) merged with Subtask 2 (Keyword Classification).
- Subtask 3 (Identifying relations) executed and evaluated separately from previous subtasks.
- Utilized SciBERT, a BERT variant pretrained on scientific data.
- SciBERT architecture mirrors BERT-base, trained specifically for scientific domain with scivocab vocabulary.

# Evaluation - Task 1 and 2 Confusion Matrix

Confusion Matrix :



# Evaluation

| | 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 |

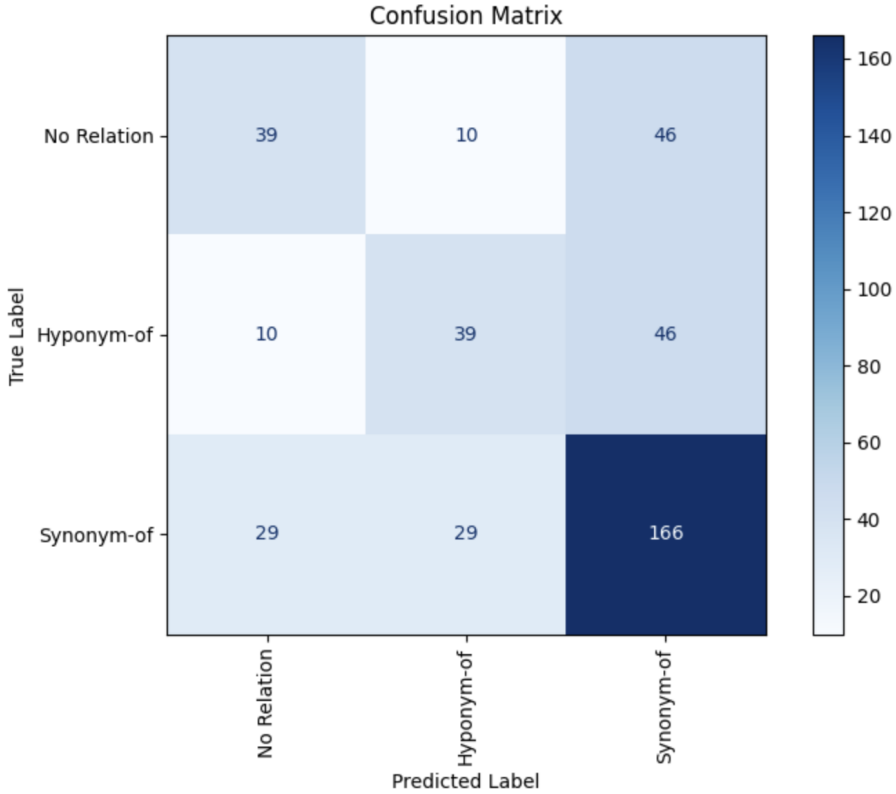
# Evaluation

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]

# Evaluation - Task 3

Confusion Matrix:



# Evaluation

| | 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