INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

EXTRACTING KEYPHRASES AND RELATIONS FROM SCIENTIFIC PUBLICATIONS

SARTHAK BANSAL (2021101134)
SARTHAK.BANSAL@STUDENTS.IIIT.AC.IN
ASHNA DUA (2021101072)
ASHNA.DUA@STUDENTS.IIIT.AC.IN
PRADHUMAN TIWARI (2020115016)
PRADHUMAN.T@RESEARCH.IIIT.AC.IN

MARCH 12, 2024

Contents

Contents 1

1 Problem Statement 2
- 1.1 Introduction . . . 2
- 1.2 Dataset . . . 2

2 Literature Review 4

3 Methodology 5
- 3.1 Keyword Extraction Module . . . 5
  - 3.1.1 POS Tags . . . 5
  - 3.1.2 Deep Learning . . . 5
- 3.2 Keyword Classification Module . . . 6
  - 3.2.1 Statistical Methods . . . 6
  - 3.2.2 Neural Methods . . . 6
- 3.3 Relationships Extraction Module . . . 7
- 3.4 Work Flow . . . 8

4 Timeline 9

1

Problem Statement

1.1 Introduction

The task comprises of extracting keyphrases and relations between them from scientific documents. The task is crucial for determining which papers describe which tasks and processes, use which materials and how those relate to one another . At its core , the task can be divided into 3 smaller subtasks:

- **Extracting Keyphrases**: This involves identifying and extracting descriptive phrases or terms from the document that encapsulate its main concepts and ideas. Keyphrases serve as important indicators of the content and focus of the document.
- **Classifying keyphrases**: This involves classifying the extracted keyphrases into TASK , MATERIAL AND PROCESS. For eg. named entity recognition is a TASK which is done using conditional random fields which is a PROCESS
- **Identifying semantic relations**: This includes identifying hyponyms (terms that are more specific versions of another term) and synonyms (terms that have similar meanings)

1.2 Dataset

The dataset for the task has the following:

- Paragraph of text drawn from a scientific paper (in plain text)
- Keyphrases of the text (start offset , end offset , id , surface form)
- Labels of the corresponding keyphrases

Some important findings about the dataset:

- Most of the phrases are **NOUN** phrases (around 93%)
- Long keyphrases (5 or more tokens) are present in the corpus in a decent amount
- 1/3rd of the keyphrases appear only once in train

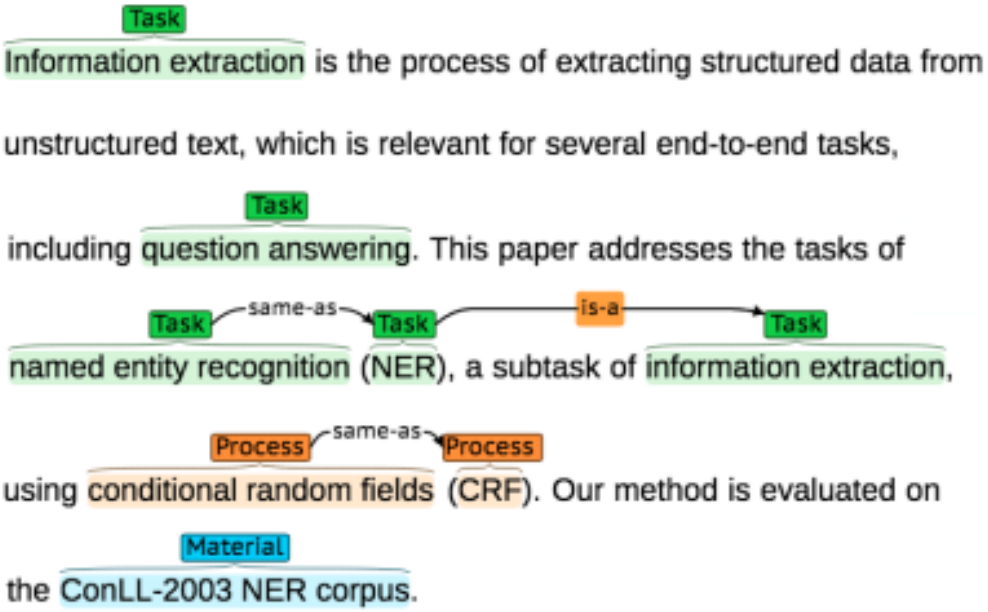

Figure 1.1: *Extracting keyphrases and classifying them*

ID	Type	Start	End
0	TASK	0	22
1	TASK	150	168
2	TASK	204	228
3	TASK	230	233
4	TASK	249	271
5	PROCESS	279	304
6	PROCESS	306	309
7	MATERIAL	343	364

ID1	ID2	Type
2	0	HYPONYM-OF
2	3	SYNONYM-OF
5	6	SYNONYM-OF

Figure 1.2: *Extraction Relations*

2

Literature Review

The paper, Anju et al., 2018 achieved state of the art results for the task of key phrase extraction from the scientific publications using the Conditional random field (CRF) approach based on 15 features including POS Tags , Named Entities , TF-IDf , Chunking etc. The sameval task of 2017, Prasad et al., 2017 which had these same objectives also used the above mentioned CRF approach for the task of Keyphrase Identification and Classification, while applied the sklearn.ensemble's random forest classifier, with syntactic similarity features for the task of keyphrase relation identification.

The paper Eger et al., 2017 uses three different models to solve the above tasks . One uses learnable embeddings and a character-level convolutional neural network (char-CNN) for classification . Other is a stacked learner which takes five base classifiers from scikit-learn and trains them repeatedly on 90% of the training data, extracting their predictions on the remaining 10% . The final model is an attention based Bidirectional Long Short-Term Memory network which uses convolutional layers to generate an attention vector which is supplied along with the pre-trained embeddings to the LSTM layer to generate softmax probabilities

The authors of the paper Garg et al., 2020, introduced a scholarly tool called SEAL, whose keyphrase extraction module consisted of two-stage neural architecture composed of Bidirectional Long Short-Term Memory cells augmented with Conditional Random Fields and the classification module consisting of a Random Forest classifier.

The paper, Saha, 2020 tested the hypothesis that the contextual information is sufficient for identifying a homonymous word. They did this by using BERT embeddings of the words to capture the context and then applying various clustering algorithms on them.

3

Methodology

Automatic scientific keyphrase extraction can be a challenging problem, facilitating several tasks like search, recommendation and ranking. The methodology propsed for Task A, B, and C, namely ***Keyword Extraction, Keyword Classification and Extracting Relationships*** is as follows:

3.1 Keyword Extraction Module

The Keyword Extraction Module is a crucial part of the methodology. It is responsible for extracting the most relevant words or phrases from the scientific documents. This module uses two methods to extract keywords:

3.1.1 POS Tags

Part of Speech (POS) tagging is a process of tagging a word in a text (corpus) corresponding to a particular part of speech, based on its definition and context. Using POS Tagging, we can generate a candidate set of keywords based on specific sequences of POS tags. Since most of the keywords are **Noun Phrases**, we can extract them using the POS Tags.

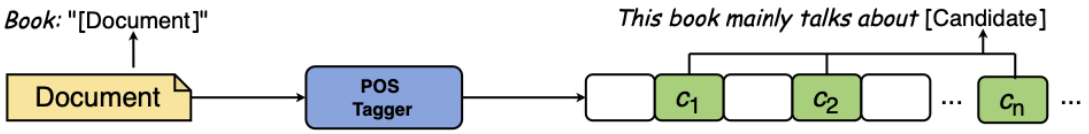

Figure 3.1: *Extracting Keywords using POS Tags*

3.1.2 Deep Learning

The Deep Learning approach uses pre-trained token level embeddings to train three layers of **Bidirectional Long Short-Term Memory (BiLSTM)** cells stacked on top of each other. The output is then sent to a linear layer, to downsize the large dimensional vector into a smaller one. Label of the token is predicted using **Conditional Random Field (CRF) Layer.**

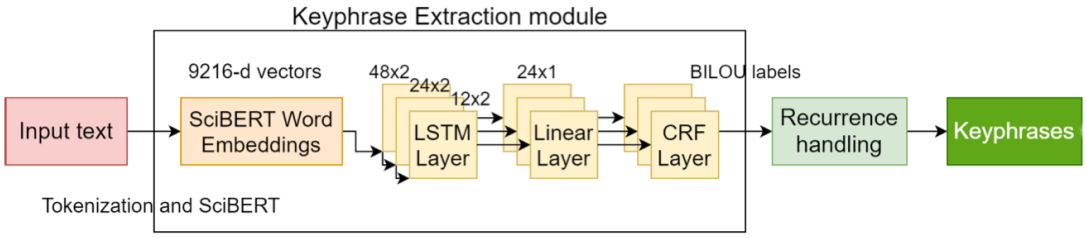

Figure 3.2: *Extracting Keywords using Deep Learning*

3.2 Keyword Classification Module

The Keyword Classification Module is responsible for classifying the extracted keywords into one of three categories: Task, Process, or Material. This classification is crucial for understanding the role of each keyword in the context of the scientific document. The module uses two methods for classification:

3.2.1 Statistical Methods

This method uses pre-trained embeddings, and for each token, the immediate neighboring tokens are considered as context. The concatenated vector of the embedding of the current token, previous token and next token is sent to Statistical Classifiers like Random Forest, etc. Candidate tokens without next or previous tokens will be appropriately padded with the embedding corresponding to the <UNK> tag. This is similar to pre-processing done for Assignment 2 Feed Forward Neural Network component.

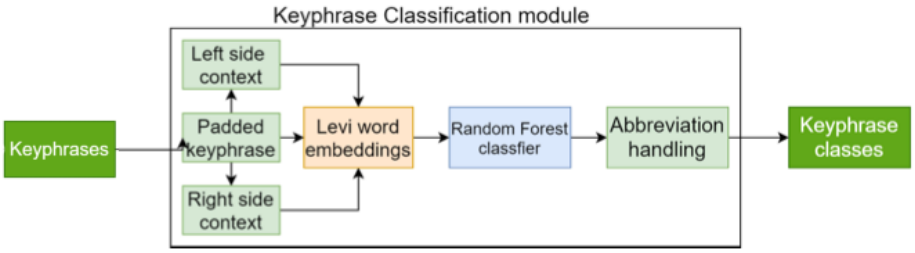

Figure 3.3: *Classification using Statistical Methods*

3.2.2 Neural Methods

This method also makes use of the pre-trained embeddings and uses the immediate neighbors of the current token as context. Candidate tokens without next or previous tokens will be appropriately padded with the embedding corresponding to the <UNK> tag. The concatenated vector of the embedding of the current token, previous token, and the next token will be sent to Neural Networks such as Recurrent Neural Network

(RNNs), Long Short-Term Memory(LSTMs) etc. for classification of the keyword into one of the three categories: Task, Process, Material.

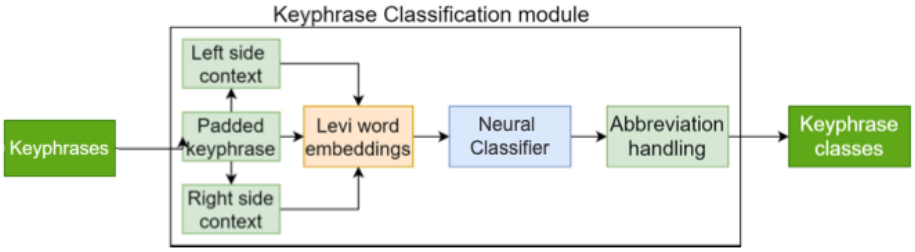

Figure 3.4: *Classification using Statistical Methods*

3.3 Relationships Extraction Module

The Relationships Extraction Module is responsible for extracting mention-level semantic relationships between keyphrases with the same keyphrase types. There are two types of relations that need to be extracted:

- **HYPONYM-OF**
- **SYNONYM-OF**

This module will use a clustering approach for extracting the above relations. The first step is to generate contextual embeddings for each keyword using BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language model, that can capture the context of a word in a sentence. Unlike Word2Vec, which combines senses into one vector, BERT embeddings are used to capture the context.

Once the embeddings are generated, various clustering algorithms can be applied to these embeddings, to group similar instances based on their characteristics, i.e. contextual embeddings of the keywords. The goal is to group together keywords that are synonyms (have similar meanings) and separate keywords that are homonyms (have different meanings but may have the same spelling or pronounciation).

By using this approach, the Extracting Relationships Module can effectively identify homonyms and synonyms among the classified keywords. These relationships can then be used for further analysis and can facilitate several tasks like search, recommendation, and ranking.

3.4 Work Flow

The diagram below provides a comprehensive overview of the entire methodology for Tasks A, B, and C. It illustrates the workflow and the interconnections between the different tasks and modules.

Each of these tasks is interconnected, forming a cohesive workflow that allows for the efficient processing and analysis of scientific documents. The output from each task serves as the input for the next, ensuring a smooth transition and continuity in the workflow.

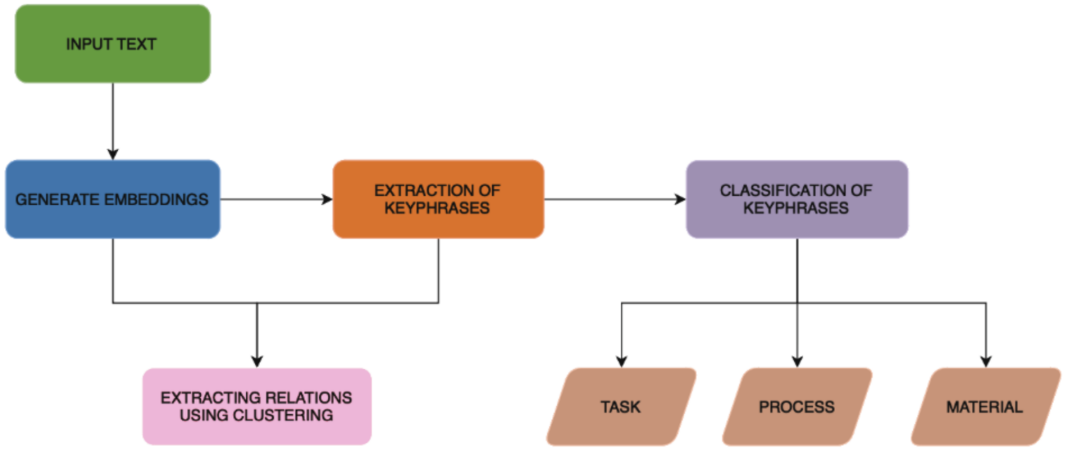

Figure 3.5: *Extracting Keywords, Classification and Extracting Relationships*

4

Timeline

Here is the proposed timeline of our project. If everything goes according to the plans , we will be able to experiment with multiple methods to achieve good performance on all the tasks mentioned in chapter 1

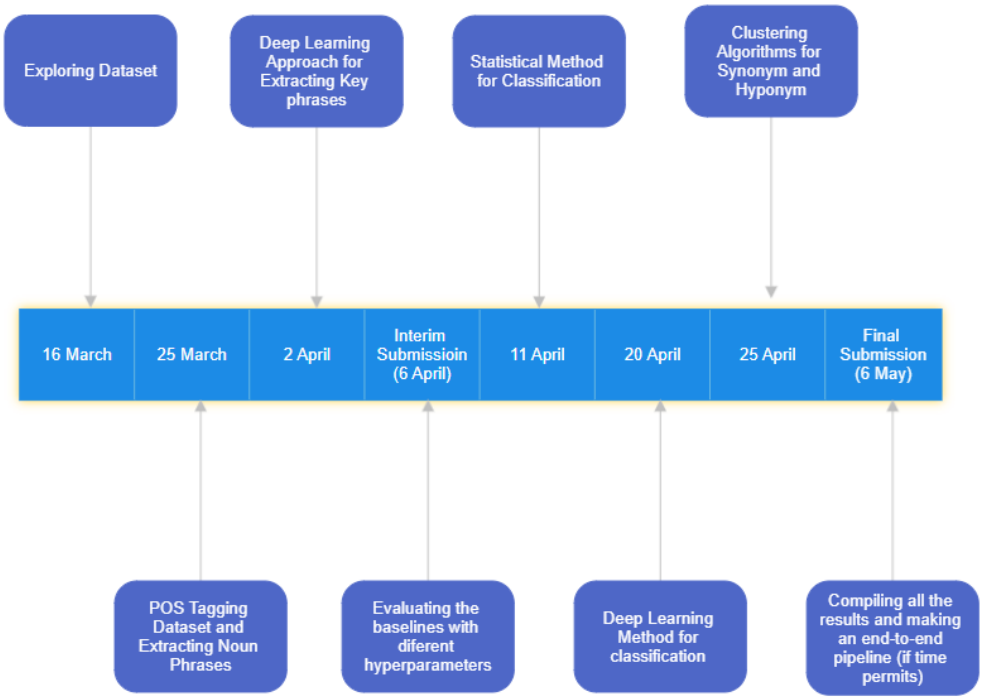

Figure 4.1: *Timeline for the project*

Bibliography

Anju, R. C., Sree Harsha Ramesh, and P. C. Rafeeque (2018). "Keyphrase and Relation Extraction from Scientific Publications". In: *Advances in Machine Learning and Data Science*. Ed. by Damodar Reddy Edla, Pawan Lingras, and Venkatanareshbabu K. Singapore: Springer Singapore, pp. 113–120. ISBN: 978-981-10-8569-7.

Eger, Steffen et al. (2017). *EELECTION at SemEval-2017 Task 10: Ensemble of nEural Learners for kEyphrase ClassificaTION*. arXiv: 1704.02215 [cs.CL].

Garg, Ayush, Sammed Shantinath Kagi, and Mayank Singh (2020). "SEAL: Scientific Keyphrase Extraction and Classification". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20. Virtual Event, China: Association for Computing Machinery, pp. 527–528. ISBN: 9781450375856. DOI: 10.1145/3383583.3398625. URL: https://doi.org/10.1145/3383583.3398625.

Prasad, Animesh and Min-Yen Kan (Aug. 2017). "WING-NUS at SemEval-2017 Task 10: Keyphrase Extraction and Classification as Joint Sequence Labeling". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Ed. by Steven Bethard et al. Vancouver, Canada: Association for Computational Linguistics, pp. 973–977. DOI: 10.18653/v1/S17-2170. URL: https://aclanthology.org/S17-2170.

Saha, Rohan (2020). "Homonym Identification using BERT - Using a Clustering Approach". en. In: DOI: 10.13140/RG.2.2.29120.07681. URL: http://rgdoi.net/10.13140/RG.2.2.29120.07681.