Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Review of Extraction and Classification of Key-Phrases in Scientific Publications using CRF and WEDP

Riya Tyagi[1]
[1]Galgotias University, School of computing Science and Engineering,
Greater Noida, Uttar Pradesh 203201, India

***Abstract*: Keyphrase extraction is a requisite task in natural language processing that aids the mapping of documents to a set of emblematic phrases. These phrases can be used for several applications such as publication ranking, query-based engines, such as Google Scholar etc. Key-phrases extracted abstracts the whole idea of the paper to a few selected candidates. This paper tries to review the task that was introduced in the SemEval 2017 task of extracting key-phrases and classifying them into three different categories which are most relevant when it comes to categorizing keywords from scientific articles and these categories are TASK, PROCESS and MATERIAL. These types can also be used as NER tags in NLP system dedicated to classifying key-phrase in scholarly articles. The methodologies discussed were tried to come to the discussed solution are also mentioned to provide evidence to the solution. We explored different methods that were being used for the designated tasks. The results mentioned here are the review of the extraction of keywords using CRF and classification using WEDP.**

Keywords: SVM, Nucleus noun, WEDP, CRF, word embeddings.

I. INTRODUCTION

The scientific research field is a very vast domain with new concepts being added to the pool daily. So, when it comes to doing research and reading up on on the existing developments for a particular topic. It can be very challenging to find the right ones which are most relevant to the topic under research. There are existing systems which have a vast repository of such publications but it still can be challenging to find the papers relevant to the topic. These existing systems are Google Scholar, Academia, IEEE Xplore etc. While these provide a fairly good system they lack to find the papers which will be most suitable to the search query based on keyphrase summarization. This paper tries to review the task by combining two methods that have been used before and then trying to see how these methods can be used to together to produce acceptable results. While discussing the extraction of keywords NER(Named Entity Relation) comes to mind and the current model only finds the NER tags related to noun phrases, a proper noun, for example, name, location, organizations etc. Tags to identify **task** mentioned, **processes** followed and **material** used are not yet included. Also, these tags are very domain-specific and are useful mostly when studying research paper to get insight into the paper.

The NER tagging task is more famous in the fields of biomedical and clinical terminologies identification. There are manifold methods using machine learning-based that are being employed in this direction, which comprise Hidden Markov Model (HMM), which uses a various biomedical feature in the data-set[1], Support Vector Machine (SVM)[2] and Conditional Random Field (CRF)[3] which is used for extracting key-phrases in this paper.

Word embeddings and their pre-trained models are coming into light to find the semantic similarity between the words using vectors. These pre-trained embeddings are also exploited to increase feature space[4]. For task A, extracting keywords, Conditional Random Field (CRF) was used and for task B, classification, word embeddings with linear SVM was used.

II. BACKGROUND

The most prevalent algorithm in the direction of important information extraction from a piece of text, NLP stands out but its developments in the field of scientific research and publication are rather contained but it is most widely used for clinical information extraction from medical records and entries. Other than NLP other unsupervised approaches to the problem of keyphrase extraction are Rake, clustering, TF-IDF filtering, which have been broadly used for such a process. It has been seen that it works very effectively for the scientific publications to extract the keywords by filtering out the most commonly used words like 'the' from the document and giving out words like 'metamorphosis' which is likely to be specific to publication. IDF has also been used with clustering to find insured results on a small dataset.

Apart from the unsupervised approaches, supervised approaches have also been explored and SVM which includes features like TF-IDF score, number of candidate tokens, the position of phrase in the text, paragraph wise or topic wise positioning is also considered but it is found to be more effective while extracting keywords from news articles. But creating a dataset to be used for learning can be quite tedious and generally have less number of data points but they are proved to be showing plausible results.
Another area which can be exploited for this task is the graph-based algorithms like Page-Rank in which key phrases and words are classified as nodes of the graph and the edges represent semantic similarity among the words, the words which are more semantically closer are more likely to be connected using the direct edge. Centrality measure of the keywords in the graph is also explored and compared with Text Rank used with additional features.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Word embeddings have also been used with different algorithms as they tend to determine the semantic intactness between different words using vectors. The most widely used methods for training of word embeddings are Glove, Word2Vec etc. Use of word embeddings with supervised algorithms like SVM has produced promising result using WEDP. They have also been used with weighted graph-based algorithms for extraction and ranking of keyphrases. Neural networks and deep learning concepts like LSTM, RNN are also proved to be effective for relationship extraction.

Wordnets have been used for finding out the semantic relationship between the keywords using the lexical representational knowledge like synset is used to find synonyms. The topic of extraction of synonyms using the antonyms is also explored. Hyponyms and synonyms extraction is looked upon by classifying words into concepts. Cosine similarity(formula-1) of words, also used by word2vec, is most commonly used similarity metric.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Formula 1: Cosine Similarity formula

III. MATERIAL

The SemEval [6] data-set is well structured and consists of 350 training, 50 development and 100 test papers. This division is reasonable and enables the fast results using a short set of development articles, and an extensive set of articles for testing and training to train the model to achieve performance as close to a real-world scenario. Separate testing and training data-set beforehand can be used to avoid the problem of over-fitting.

Some analysis conducted at Science IE are:

- Nonu phrases as keywords were abundant about 93%.
- 22% of key-phrases had 5 or more tokens,

This analysis proves the importance of tokens with NN, NP or NNP, NNS POS(Part-Of-Speech) tag is critical for extraction and classification.

IV. METHOD

A. *Preprocessing*

For preprocessing of the data nltk library is used to find Part-Of-Speech taggings, sentence tokenization and detection. POS tagging is crucial to find key-phrases because of the above mentioned analysis. The sentences were refined and words which do not add meaning to the keywords were removed like, a, the, etc.

B. *Extraction*

For the extraction of key-phrases RAKE[4] algorithm was adopted firstly but the results obtained were not apropos to the ScienceIE data. After this method SVM was also used with POS tag and TDF-ID etc. features as input but the method was not giving very good results on evaluation.

After trying the above methods and further research Conditional Random Fields (CRF) is used and to do so kleis-keyphrase-extraction is used which is specifically designed to classify the Science IE corpus.

Conditional random fields is a preferential model best befitted to prediction tasks where information derived from context or neighbour's state affect the current result of the prediction. CRFs find their applicability in noise reduction, object detection problems, named entity recognition, gene prediction, part of speech tagging, etc..

A Markov Network or Markov Random Field, as shown in Fig. 1, is a form of a graph-based model having a graph with nodes as random variables which is undirected. The structure of this graph decides the interdependence and relationships between the random variables.

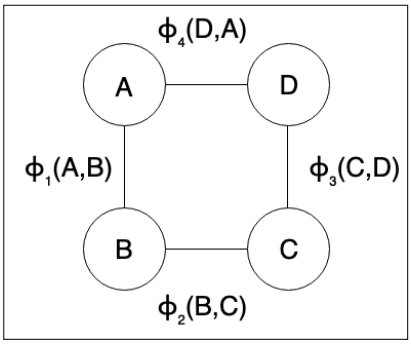

Figure 1: Markov Random Field with 4 random variables

C. *Extraction of Nucleus Noun*

Nucleus noun or head noun is the noun which is crucial to find the syntactic category of a noun phrase and with most of the keywords being noun phrases it becomes crucial to find the nucleus as it can be used for better results. Examples of Nucleus nouns are: "conversion of metals to alloys", here conversion is useful to determine the class of the key-phrase as a process. Another example can be seen in "classification of scientific materials", here classification can be defined as the task. Other examples of nucleus noun can be seen in Fig. 2.

Nucleus nouns are used to remove ambiguity in the context of the sentence or phrase while conserving the semantic intactness of the phrase or sentence. NLTK POS tagger and tokenization is used for this purpose using the stop word list, to maintain the overall structure while removing the unnecessary data.

Since Nucleus noun seems most appropriate the features for the relationship extraction are based on this noun only and all

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

other tokens were ignored. For phrases where no noun phrase was present the first word of the phrase was considered instead of the blank data-point.

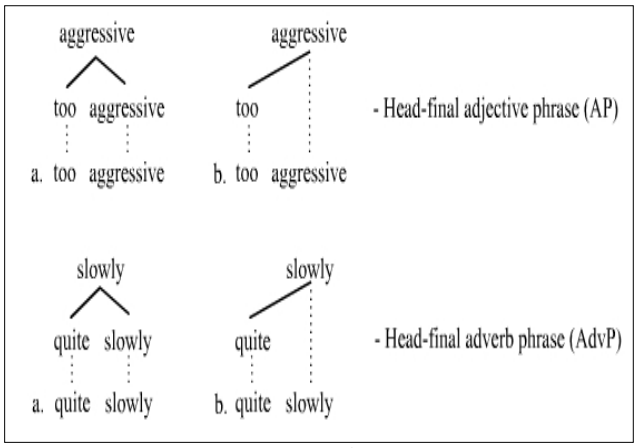

Figure 2: Examples of Nucleus Noun

D. Features

During the selection of features for the model it was found that the position of the keyphrase in the text was not of much importance and rather than the semantic and syntatic features combined with the word embeddings were very useful.

The model used is a 3 class classification and Linear plane and below mentioned features were supplied to it:

Boolean feature(syntactic):

- If first letter of nucleus is capital
- If it contains digits
- If it is alphabetic
- If it is in uppercase

E. Classification

WEDP[5] is used to find the relationship between the most commonly used word in the corpus to find similarity using Glove embedding vectors between nucleus and the words. Classes, Task, Material and Process were also added to the data-set. The most common words found are:

['system', 'surface', 'task', 'equation', 'method', 'film', 'material', 'problem', 'particle', 'alloy', 'effect', 'data', 'function', 'algorithm', 'model', 'reaction', 'structure', 'layer', 'process']

Glove embeddings with 50D were used to train the model.

Lemmatization and POS tags were considered but dropped to keep the model simple.

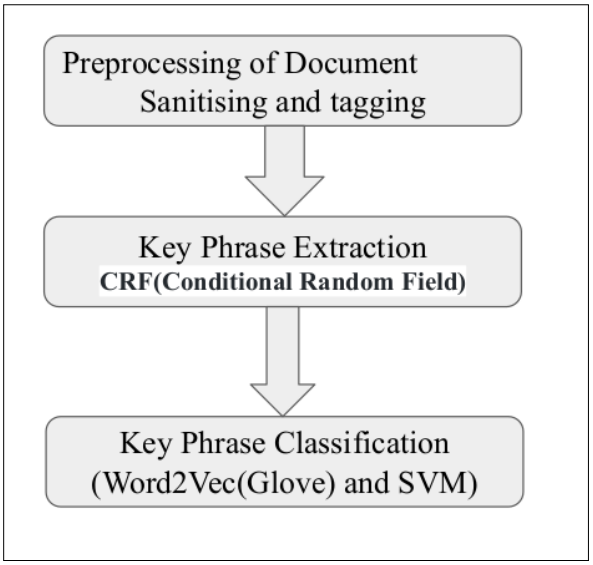

Figure 3: Architecture of Implementation

V. RESULTS

Various models were considered first but at last SVM with Linear model using scikit-learn was chosen, the architecture followed during the research is shown in Fig. 3, and the results are mentioned below:

Table 1: Architecture of Implementation

Category	P	R	F1	Support
Classification	0.30	0.42	0.38	2051
Material	0.71	0.73	0.72	567
Process	0.56	0.71	0.62	458
Task	0.28	0.28	0.29	137

VI. CONCLUSION

In this paper, different approaches were discussed and then supervised process with word embeddings was used. It was seen that classifying task has a lower score as the number of data entries for training was not adequate in comparison to other classes. The highest F1 score of 0.73 was achieved for Material class and for classification the highest F1 score of 0.38 was achieved for test articles and 0.48 for development articles.

For future, the task of relationship extraction will also be included and a system to also provide ranking to these papers on the basis of key phrases extracted will also be looked into. support for longer documents and other web resources will be looked into to make it closer to the real-world situation.

REFERENCES

[1] Zhang, J., Shen, D., Zhou, G., Su, J., & Tan, C. L. (2004). Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of biomedical informatics*, *37*(6), 411-422.

[2] Cheong, S., Oh, S. H., & Lee, S. Y. (2004). Support vector machines with binary tree architecture for multi-class classification. *Neural Information Processing-Letters and Reviews*, *2*(3), 47-51.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

[3] Tsai, R. T. H., Sung, C. L., Dai, H. J., Hung, H. C., Sung, T. Y., & Hsu, W. L. (2006, December). NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. In *BMC bioinformatics* (Vol. 7, No. S5, p. S11). BioMed Central.

[4] Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, *1*, 1-20.

[5] Liu, S., Shen, F., Chaudhary, V., & Liu, H. (2017, August). Mayonlp at semeval 2017 task 10: Word embedding distance pattern for keyphrase classification in scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 956-960).

[6] Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*. Available: http://nl1.vnunet.com/news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)