Keyphrases and Relations Extraction from Scientific Publications

Anju R C[1], Sree Harsha Ramesh[2], Rafeeque P C[3]

[1,3] Government Engineering College, Palakkad, Kerala, India

[2] Surukam Analytics Pvt. Ltd, Chennai, Tamilnadu, India
anju.malu5000@gmail.com

Abstract. This paper proposes a novel approach for extracting keyphrases and its relations from scientific published articles such as research papers using Conditional Random Fields(CRF). Keyphrase is a word or set of words that describe the close relationship of content and context in a particular document. Keyphrases may be the topics of the document which represent the key idea of that document. Automatic keyphrase extraction has very important role for the automatic systems like independent summarization, query or topic generation, question-answering system, Information retrieval, Document classification etc. The relationships of the keyphrases are also extracted. Two types of relations are considered-Synonym and Hyponyms. The result shows that our proposed system outperforms the existing systems.

Keywords: Keyphrase Extraction, Topic Extraction, Information Extraction (IE), Summarization, Question Answering (QA), Document Classification.

1 Introduction

Keyphrase is a word or set of words that describe the close relationship of content and context in the document. Keyphrases are sometimes simple nouns or noun phrases (NPs) that represent the key ideas of the document i.e., topic. Document keyphrases have enabled fast and accurate searching for a given document from a large text collection, and have exhibited their potential in improving many natural language processing (NLP) and information retrieval (IR) tasks, such as text summarization, text categorization, opinion mining and document indexing.

Scientific research is the systematic investigation of scientific theories and hypothesis such as gaining, maintaining and understanding the body of existing work in specific areas related to such fundamental objects. In such cases researchers and practitioners faces some typical questions.

- Which paper have addressed a specific task?
- Which paper have studied a process or variants?

- Which papers have utilized such materials?
- Which paper have addressed this task using variants of this process?

The existing works in this area does not address the above mentioned problems efficiently. Hence utilities are required to identify the keyphrases and relations. This paper propose the task of keyphrases extraction from scientific documents using CRF (Conditional Random Field). It also addresses the task of extracting type of keyphrases such as PROCESS, TASK, and MATERIAL and relation between keyphrases. Mainly two types of relations are considered.
Synonym_Of (Abbreviation):The relationship between two keyphrases A and B is said to be A =>Synonym_Of(B) if they both denote the some sematic field.
eg., Machine Learing => Synonym_Of(ML).
Hyponym_Of: The relationship between two keyphrases A and B is said to be A=> Hyponym_Of(B) if semantic field of A is included within that of B.
eg., Red =>Hyponym_Of(Colours).

Our proposed methods is compared with existing systems like Alchemi API[18] and Rake[22] and it shows better result than both of the systems.

2 Related Works

The methods for keyphrase extraction can be categorized into supervised and unsupervised approaches. Mihlceva and Tarau (2004)[15] proposed an unsupervised approach that considered single tokens are vertices of a graph and co-occurrence relation between tokens are edges. Candidate can be ranked using PageRank and adjacent keywords are merged into keyphrases in post processing step. KEA (Keyphrase Extraction Algorithm) is a supervised system that used all n-grams of a certain length, a naive bayes classifier and tf-idf and position features (Frank et. al. 1999)[6]. Turney (2000)[13] introduced Extractor, a supervised system that select a stem and stemmed n-grams as candidate and turns its parameter (mainly related to frequency, position and length) with a generic algorithm. Pinaki, Kishorjit, Sivaji (2002)[14] introduced a supervised method for keyword extraction as a part of SemEval 2010. They used CRF tool for finding keywords from the scientific publications.

KP_Miner system (2010)[2] is used for extracting the keyphrases in english and arabic document. When they are developing the system, the goal was to build a general purpose keyphrase extraction system that can be easily configured by users based on their understanding of the documents or the use of any sophisticated natural language processing or linguistics tools. KP_Miner system is an unsupervised system. Keyphrase extraction in the system is a three step process: candidate keyphrase selection, candidate keyphrase weight calculation and finally keyphrase refinement.

This system gives the result that precision (24.9%), recall (25.5%) and F measure (25.02%).

WINGNUS (2010)[12] developed a method for SemEval 2010. In this method they used test input format of PDF because logical structure recovery is much more robust. For this they used Google Scholar base Crawler to find the PDFs given plain text. For logical structure extraction SectLabel(Luong et al.,)[12] used. Precision, Recall and Fmeasure for the system were 24.9% , 25.5% and 25.2% respectively.

In 2010 S. Lahiri and Mihalcea developed a system for extracting keywords from emails[17]. They proposed a supervised keyword extraction system. The system contain mainly five steps. Email preprocessing, Candidate extraction, Pre-processing, Ranking/classification, Post processing. They got the accuracy of Precision-24.8%, Recall-25.4%, Fmeasure-25.1% .

In 2013 Kamal Sankar proposed a hybrid approach to extract keyphrase from medical document[7]. In this paper proposes amalgamation of two methods: first one assigns weights to candidates keyphrases based on effective combination of features such as position, term frequency, inverse document features and second one assign weights to candidate keyphrases using some knowledge about their similarities to the structure and characteristics of keyphrases available in the memory. This keyphrase extraction method consist of three primary components: Document preprocessing, candidate keyphrase identification and assigning scores to the candidates for ranking.

RAKE(Rapid Automatic Keyword Extraction)[22] is based on the keywords frequently contain multiple words but rarely contain standard punctuation or stop words. The input parameters for RAKE comprise a list of stop words, a set of phrase delimiters and set of word delimiters RAKE uses stop words and phrase delimiters to partition the document text into candidate keywords, which are sequences of content words as they occur in the text. Co-occurrences of words within these candidate keywords are meaningful and allow us to identify word co-occurrence without the application of an arbitrarily sized sliding window. Word associations are thus measured in a manner that automatically adapts to the style and content of the text, enabling adaptive and fine-grained measurement of word co-occurrences that will be used to score candidate keywords.

Alchemy-Api [18] is developed by IBM and they used a deep learning method to extract the keywords or keyphrases from the document.

Textacy[25] is a Python library for performing higher-level natural language processing (NLP) tasks, built on the high performance spaCy[23] library. This extract the abbreviations from the documents with use of basics tokenization, part-of-speech tagging, dependency parsing, etc.

3 Keyphrases and Relation Extraction

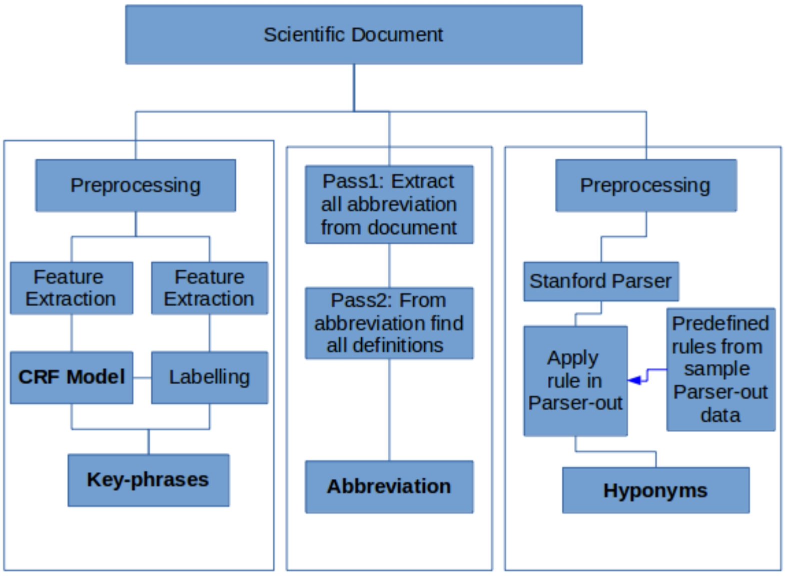

Figure 1 : Over all System Architecture.

3.1 System Design

The proposed architecture consist of three phrases:- Keyphrases Extraction, Abbreviation Extraction and Hyponym Extraction. Figure 1 shows the overall system diagram for the proposed system. Keyphrases extraction describe how CRF[19] used to extract the keyphrases and the type of the keyphrases from scientific documents. Two pass method is used to extract abbreviation and definition from the document. Finally rule based system is used for hyponym extraction.

The proposed work implements a system to extract all the keyphrases from the documents and also find out the type of keyphrases such as PROCESS, TASK and MATERIAL.
PROCESS : Keyphrases related to some scientific models, algorithm and process should be labeled as PROCESS.
TASK : Keyphrases related to application, end goal, problem and task should be labeled as TASK.
MATERIAL : Keyphrases identify the resource used in the paper

Then system finds all the abbreviation and its expansion and hyponyms from the document.

3.2 Keyphrases Extraction

The proposed system extract the keyphrases by using Conditional Random Field(CRF). From the 500 collected document 350 taken for the training purpose. After training a CRF Model is created for generate keyphrases. 15 features are considered for training.

1. POS tags : Find out the part of speech(POS) tags for all the words in the document. Nouns(NN,NNP,NNS etc.,) can be most probable of a keyphrases. Then that word feature is 1 and others 0.
2. Named Entities(NE): Find out all the named entities. From that a word will part of NE then its feature value is 1 otherwise 0.
3. Upper Case Strings : Upper case words are 1 others are 0.
4. Term Frequency Greater Than 2: A word occur more than 2 times then it will 1 otherwise 0.
5. TF Value Normalized: For each word find out the Term Frequency (TF) value. Normalized term frequency value is this feature.
6. IDF Value : For each word find the IDF (Inverse Document Frequency) value.
7. Chunking : Finds all chunks from the document. The word is part of that chunk then it will 1 otherwise 0.
8. Position of the word in the document.
9. Capitalized words: Capitalized word will 1 otherwise 0.
10. Reference in a Sentence: A sentence will contains the reference.
11. Section of Text : Word will be in abstract, Introduction or Body. If word in abstract feature value is 1. If word in introduction section then feature value is 2. Otherwise feature value is 3.
12. Article Type: The document will occur in which article type. Eg., Corrosion Science, Environmental science.
13. Contains Greek letter and chemical symbol: The word is Greek letter or chemical symbol then feature value is 1 otherwise 0.
14. Contains Hyphenated words: Feature value is 1 for hyphenated words.
15. Contains in Vocabulary list: Word contains in python vocabulary 0.0.5[21] then feature value is 1.

3.3 Abbreviation and Expansion Extraction

Abbreviation and expansion extraction is a two pass process. During the first pass the program will extract all the abbreviations from the document and it saves in a temporary list. In second pass taken each abbreviations from the list and finds its expansion.

3.4 Hyponym Extraction

A=>Hyponyms_Of(B) means the semantic field of A is included within that of B. eg., Red =>*Hyponym_Of (*Color). General rules are generated by the parsed sample documents.These general rules applied to the new document and words that satisfy the rules are the hyponyms.

4 Evaluation and Discussion

4.1 Corpus Collection

Data sets for this project is built from Science Direct open access publications. It consists of 500 journal articles evenly distributed among the domains Computer Science, Material Sciences and Physics.

- 350 labeled document from different scientific publications are collected for training purposes.
- 50 labeled document for testing case.
- 100 unlabeled document for testing case.

System	Recall	Precision	F-measure
Alchemi Api	0.3002	0.3006	0.3003
Proposed System	0.3305	0.3462	0.3381
Rake	0.3102	0.3130	0.3115

Table 1 Evaluation Result for Keyphrase Extraction (Top 15 Candidates).

4.2 Evaluation and Results of Keyphrases Extraction

The evaluation of Keyphrase Extraction is done by comparing the proposed system with existing system such as Alchemy Api [18]and Rake (Rapid Automatic Keyword Extraction)[22]. Table 1 show the recall, precision, f-measure result for three system. After analyzing the table it will be clear that my proposed system gives much good result than Alchemi Api and Rake system.

4.3 Evaluation and Results of Abbreviation and Expansion Extraction

Table 2 shows the evaluation result for abbreviation extraction. The proposed system is compared with Textacy [25]. The results shows that our system perform better than Textacy.

System	Recall	Precision	F-measure
Textacy	0.6002	0.6080	0.6040
Proposed System	0.6505	0.6888	0.6991

Table 2 Evaluation Result for Abbreviation Extraction.

5 Conclusion and Future Works

Keyphrases have very important role in most of the application. CRF based approach to keyphrase extraction has been attempted in the paper. Hyponyms and abbreviations are the two relations of the keyphrases. Rule based approach is used for the extraction of hyponyms and two pass process is used in abbreviation and definition extraction. Proper cleaning of the input documents and identification of more appropriate features could have improved the score.

One major limitation of Stanford Parser is time consumption so in future use better parser for parsing. Statistical and hybrid approaches are other alternatives for Rule based method for Hyponyms extraction. Such method may give better result.

References

1. C. Huang, Y. Tian, Z. Zhou, C.X. Ling and T. Huang Keyphrase extraction using semantic networks structure analysis in IEEE Int. Conf. on Data Mining, pp. 275-284 (2006).
2. EI-Beltagy, S.R., and Rafea A, KP-Miner. Participated in SemEval-2 Proceeding the 5 th International Workshop on Semantic Evaluation, ACL, pp. 190-193,Uppsala,Sweden, (2010)
3. Erwin Marsi, Pinar Ozturk, Extraction and Generalization of Variables from Scientific Publications, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon,Portugal, pp. 505-511(2015).
4. Georgeta Bordea and Paul Buitelaar, DERIUNLP: A Context Based Approach to Automatic Keyphrase Extraction. Proceedings of the 5th International Workshop on Semantic Evaluation ACL , pp:146-149 , Uppsala, Sweden(2010).
5. G. K. Palshikar, Keyword Extraction from a Single Document Using Centrality Measures in 2 nd Int. Conf. PReMI 2007 LNCS 4815, pp. 503-510 (2007).
6. Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin and Craig G. Nevill-Manning KEA: Practical Automatic Keyphrase Extraction (1999).
7. Kamal Sarkar, A Hybrid Approach to Extract Keyphrases from MedicalDocuments, International Journel for Computer Application (0975-8887) Vol.63,No:18 (2013).
8. Kathrin Eichler and Gunter Neumann, DFKIKeyWE: Ranking Keyphrases Extracted from scientific articles. proceedings of the 5 th International Workshop on Sematic Evaluation, ACL, pp: 150-153, Uppsala, Sweden (2010).

9. Mounia Haddoud, Aicha Mokhtari, Thierry Leiroq and Said Abdeddaim, Accurate Keyphrase Extraction from scientific papers by Mining Linguistics Information, CLBib (2015).
10. M. Litvak, M. Last, H. Aizenman, I. Gobits and A. Kande,l DegExt: A Language-Independent Graph-Based Keyphrase Extractor. Advances in Intelligent and Soft Computing, Vol 86, pp.121-130 (2011).
11. M. Litvak and M. Last, Graph-based keyword extraction for single-document summarization. Proceedings of the 2nd Workshop on Multi-source Multilingual Information Extraction and Summarization, pp.17- 24, Manchester, UK (2008).
12. Nguyen T D and MinYenKan, WINGNUS:Keyphrase Extraction Utilizing Document Logical Structure . Proceeding the 5th International Workshop on Semtic Evluation, ACL, pp. 166-169. Uppsala, Sweden (2010).
13. Peter D. Turney, Learning Algorithms for Keyphrase Extraction. Information Retrieval-INRT, pp.34-99 (2000).
14. Pinaki Bhaskar, Kishorjit Nongnieikapam and Sivaji Bandyopadhyay, Keyphrase Extraction in Scientific Articles: A Supervised Approach , Proceeding of COLING 2012, Mumbai, pp. 17-24 (2012).
15. Rada Mihalcea and Paul Tarau, TextRank: Bringing Order into Texts. Conference on Empirical Methods in Natural Language Processing (2004).
16. Ramesh Nallapati,James Allan and Sridhar Mahadevan, "Extraction of Keywords from News Stories" CIIR technical report, IR(345) (2013).
17. S. Lahiri, R. Mihalcea and P.H Lai,Keyword Extraction from Emails. Proceedings of 5 th International Workshop on Semantic Evaluation, ACL, pp.1-24, 2016 Cambridge University Press, UK (2010).
18. Alchemy-api, https://www.ibm.com/watson/alchemy-api.htm last accessed 2017/03/20
19. CRF++ http://taku910.github.io/crfpp last accessed 2017/01/28
20. NLTK , http://www.nltk.org , last accessed 2017/01/27
21. Python Vocabulary0.0.5, http://pypi.python.org/pypi/vacabulary/0.0.5 last accessed 2017/02/01
22. Rake: Rapid Automatic Keyword Extraction, https://hackage.haskell.org/package/rake last accessed 2017/03/23
23. Spacy, https://pypi.python.org/pypi/spacy last accessed 2017/02/25
24. Stanford CoreNLP, https://stanfordnlp.github.io/CoreNLP last accessed 2017/02/27
25. Textacy, https://pypi.python.org/pypi/textacy. last accessed 2017/03/13