# Acronym Disambiguation using Web Scraping

**K. Premkumar, V. Atchayaa, P. Idayavalli, R. Gayathri**

***Abstract***: ***Web Scraping is one of the current technologies that uses scraping tools to perform tasks similar to humans. It is adopted in many applications like e-commerce, dataset creating in machine learning, advertising etc. This work focuses on acronym disambiguation which is part of natural language processing. Acronym disambiguation is mainly used in chat bot, named entity recognition, natural language processing and so on. In this paper, an acronym disambiguation system is built by web scraping using Jsoup and cosine similarity score is used to identify the most suitable acronym. Our goal is to identify the acronym suitable for the abbreviation based on context of the paragraph where it lies. For this we use cosine similarity to calculate the score, the acronym which obtains maximum score is the concluded as suitable expansion***

***Keywords: Web scraping, JSoup, cosine similarity, acronym, abbreviation.***

## I. INTRODUCTION

An acronym is a developed abbreviation that consists of several word-initial elements to signify terms in a short format. The acronyms are identical to the abbreviation but the initial term number has to be greater than two. The abbreviation is a style that should be removed when the duration of a phrase is usually too long for ads, such as "ad."

. For context, "IPS" applies not only to Intrusion Prevention System in the area of computer science or artificial intelligence but also to Indian Police Service, which is police position in Indian government. This polysemous acronym which means an acronym that has several expansions, poses a major obstacle when a computer attempts to discern just what "IPS" implies. Over the years, several scholars have been researching what extensions of acronyms in the provided text details utilizing linguistic and syntactic methods to improve the performance of Data collection, Information Processing, and more. These studies focused on the issue of defining and extracting accurate expansions of acronyms in specific papers.

But this created a major issue in natural language processing and named entity recognition. For example, when a document-based retrieval chatbot have to be trained, large number of documents on various field will be given for the training of the chatbot to increase its score. But if the documents having acronyms in it, will be difficult for the chatbot to process and understand it.

So, in that cases there is a need for acronym disambiguation systems because it is difficult to manually find the expansion of all the acronym present in the documents and also it is time consuming. So, the acronym disambiguation systems with accuracy rates are needed. And there are many works related to this have been done in the past years.

## II. RELATED WORKS

Mathiew Roche et.al [1] proposed a paper which provides nine quality measures of the relevant definition, prediction based on mutual information [MI], Cubi MI and dice's coefficient. A combination of this statistical measure with cosine approach is proposed. Experiments have been done on biomedical domain where acronyms are numerous. The results on biomedical corpus showed that the proposed measures where the accurate devices predict relevant definition. There are two main measures used in that paper:

AcroDef measure: It computes dependencies between forming the expansions. It helps to choose the relevant definitions.

IaDef (Independency between acronyms and definitions): It computes the dependency between acronyms and definitions.

Yangli et.al [2] proposed a system which gives a solution for acronym disambiguation for enterprise. The system gives an end to end framework to tackle all these challenges faced in the enterprise. The framework produces a high-quality acronym disambiguation system as output. Their disambiguation model is trained via distant supervised learning without requiring any manually labeled training example. Their framework can be deployed to any enterprise to support high quality acronym disambiguation.

Dongjin choi et.al [3] proposed a method to recommend the most related expansion of acronym through analyzing co-occurrence words by using Wikipedia. Their goal is to find the most appropriate expansion for given acronyms. In this system, they first need to make a acronym dataset which contains acronyms, expansions, titles, and co-occurrence words, which are taken by analyzing the Wikipedia extended abstracts. To find the most appropriate expansion words for the given acronyms they applied WP(Wu-Palmer) metric similar to wordNet The drawbacks of the system was there were no recommendations for similar expansions, example National Aeronautics and Space Administration and National Aeronautics and Space Act both the expansions are similar in those cases which will not give the exact output as required.

**Revised Manuscript Received on March 17, 2020.**

***** Correspondence Author

**K. Premkumar***Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India.
Email: hodcse@smvec.ac.in

**V.Atchayaa**, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India.
Email: atchayaaravi33@gmail.com

**P. Idayavalli,** Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India.

**R. Gayathri,** Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India.

Akram Gaballah et.al [4] proposed a language modelling approach for acronym disambiguation using context information. A dictionary of all possible expansions of acronyms is generated automatically. The dictionary is used to search for all the possible expansions. Training data is automatically collected from downloaded documents of search engine results. The collected data is used to build a uni-gram language model, that models the context of each candidate expansion. Their approach has the option to reject the acronym if it is not confident on disambiguation. They have evaluated the performance of our language modelling approach and compared it with tf-idf discriminative approach. For building the model.

This is a major drawback in the previous system, in order to overcome the above drawbacks and further increase the efficiency of the acronym disambiguation system we proposed the new system model.

## III. PROPOSED SYSTEM

Our proposed system to rectify the existing systems disadvantages is as follows:

- The file containing acronym will be uploaded by the user, and text file be tokenized and the punctuations will be removed.
- After removing the punctuation, stop words will also be removed the acronym will be detected and it will be sent to acronym finder website, to detect the possible expansions.
- Then the relevant content for those expansions will be fetched from the internet. After which the similarity between context will be identified by using cosine similarity score.
- The score with maximum value will be considered as the appropriate expansion. And that expansion will be replaced in the user uploaded text.
- For parsing data from the search engine, we need an external tool, JSoup can be used for this purpose. It can scrap the data from the internet and easily manipulate and store them in excel, database or CSV format.
- JSoup[12] is the open source library which is used to work with HTML, it provides an efficient API to fetch the URLs and extracting , manipulating data by using HTML and CSS selectors.
- JSoup is created in such a way that it deals with all kinds of HTML found in the wild, starting from pristine and validating to invalid tag soup

Here we have used the cosine similarity [13] to find the similarity between the context, the cosine similarity is calculated by using the formula:

$$\text{Cosine similarity} = (A.B)/\|A\| \|B\|$$

where, cosine similarity is the scalar product of two vectors divided by the product of their Euclidean norms. To find the cosine similarity, initially context files should be converted in a word2vec format to create the word embedding. Word embedding [11] is most used representation for documents. By using this we can capture the context of words present in a document and also in addition to that we can also capture semantic and syntactic similar relation between two documents.

The below is our proposed architecture diagram which have five modules. The work flow of the system is mentioned clearly in the below architecture diagram.

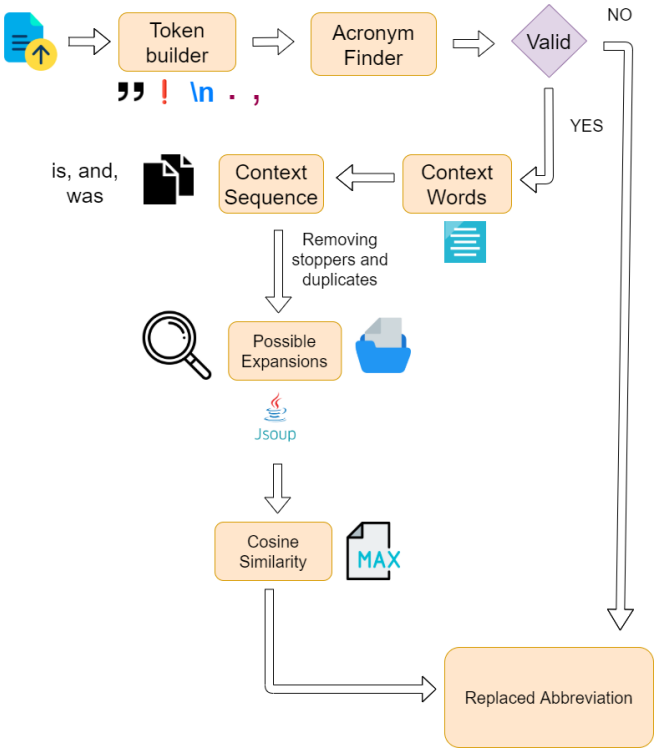


**Fig 1 : Proposed System Architecture**

There are five main modules in our system as follows:

- Token Builder
- Acronym Finder
- Context Sequencing
- Possible Expansions
- Cosine Similarity

### A. Token Builder

Token Builder is the process of converting upload text document into tokenized file by splitting the text, whenever new line starts. Fig .2 demonstrates the following steps:

1. The text file containing acronym is uploaded by the user to application, whose contents will be read and displayed in the text area.
2. For tokenizing the uploaded file, Regular Expression RegEx is used, which will tokenize the text on new line and store them in separate file
3. In the raw data noise mainly punctuation are removed. This process is called data cleaning. The PosixCharacter class present in java is used by to remove all the punctuations present in the data.

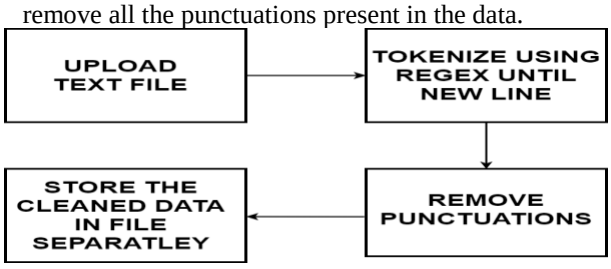


**Fig.2 Token Builder**

### B. Acronym Finder

In this module, first the acronym present in the text will be located and they are stored separately in the list. So, to find the valid acronym the text file is scanned, and the word in which all letters are in caps is considered as acronym initially, then we have to check whether the acronym is valid or not. For this, we use the website acrnoymfinder, and to scrap the data from the website we will use a JSoup. JSoup [14] is a java library which is used to extract the information from HTML page or any web page. And the scraped data can be manipulated and used. Now the words which are considered as acronym is checked to be valid or not. If they are valid, they will be stored separately in a list. The entire process is as follows, which is shown in fig.3.

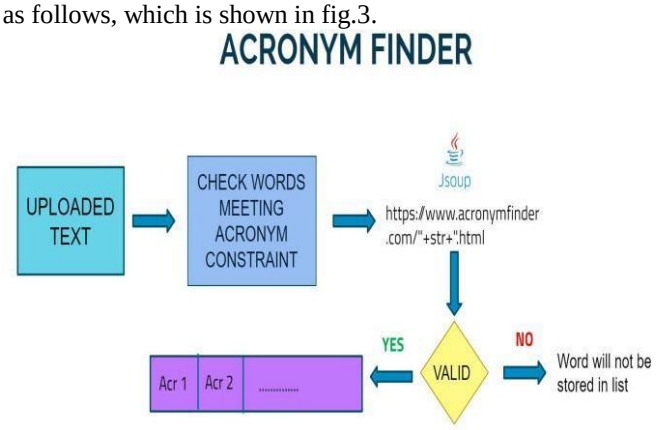


**Fig.3 Acronym Finder**

### C. Context Sequencing

The context sequencing is an important process, because only this context will be used to obtain the cosine similarity score. In this first we remove all the stop words from the context. And then the line in which the valid acronym is present will be located. After which only that line, before and after line in which the acronym is present will be retained. Because mostly the domain relevant words will be present among these lines. And it saves time to calculate the score of these lines instead of calculating for the entire paragraph uploaded.

In this module, we will also remove the duplicate acronyms which are stored in the string list. So only distinct acronyms will be retained.

### D. Possible Expansions

Now, the possible expansion for the acronym will be collected from the acronym finder website by using JSoup. For example, the word NASA itself have different possible expansions as listed in the table I.

**Table- I: Possible Expansions for NASA**

| | |
|---|---|
| NASA | National Auto Sport Association |
| NASA | National Aeronautics and Space Administration |
| NASA | National Aeronautics and Space Act |
| NASA | North America South America |
| NASA | Native American Student Association |
| NASA | Natural Athlete Strength Association |

And all these acronyms are fetched and will be stored in a separate text file. Now, every expansion will be searched on internet using web scraping method and their web context will be fetched by using the scraper program. All the expansions will be searched in internet and their respective context will be fetched by using JSoup. The fetched contents form the web will be stored in the separate files as shown in table II.

**Table- II: Possible Expansions for NASA**

| | |
|---|---|
| National Auto Sport Association | The **National Auto Sport Association** (NASA) is an American motor sports organization promoting road racing and high-performance driver education. Member ship in **National Auto Sport Association** will make you part of a large family of motorsports enthusiasts and will provide you many privileges. Digital subscription to Speed News, delivered to your inbox each month. Participate in more than 100 events across the country. |
| National Aeronautics and Space Administration | The National Aeronautics and Space Administration is an independent agency of the United States Federal Government responsible for the civilian space program, as well as aeronautics and aerospace research. NASA was established in 1958, succeeding the National Advisory Committee for Aeronautics. **NASA's** mission is to pioneer the future in space exploration, scientific discovery and aeronautics research. |

### E. Cosine Similarity

i. Cosine similarity is the important module which is used here, in this we convert the obtained web context file and the uploaded user context file as word2vec embedding.

ii. And now all the web context file of each expansion will be compared with the user uploaded context file, and the score is calculated by comparing similarity between both the uploaded and the web context file.
iii. The cosine similarity value ranges between 0 and 1. The web context file of an expansion having the maximum score is the suitable expansion.
iv. Finally, the suitable expansion will be replaced for the respective acronym in the user uploaded file which can be further used in any natural language processing or named entity recognition based upon the user's need.

## IV. RESULTS AND DISCUSSIONS

The result of our proposed model to find the correct expansion is as follows in the below table III, which shows the method in which we compare the score of two context, and the context with maximum score will be stores as max score.

**Table- III: Cosine Similarity Score comparing the context**

| MAX SCORE | CURRENT SCORE | CURRENT EXPANSION | MAX SCORE EXPANSION |
|---|---|---|---|
| 0.02914717 | 0.02914717 | National Auto Sport Association | National Auto Sport Association |
| 0.61649428 | 0.61649428 | National Aeronautics and Space Administration | National Aeronautics and Space Administration |
| 0.61649428 | 0.56773430 | National Aeronautics and Space Act | National Aeronautics and Space Administration |
| 0.63589799 | 0.63589799 | National Aeronautics and Space Agency | National Aeronautics and Space Agency |
| 0.63589799 | 0.04599328 | National Academyof Sciences of Armenia | National Aeronautics and Space Agency |

After checking the context score for all the expansions finally the context with maximum score will be chosen. And we occurred an accuracy of 93% by our proposed model and we are trying to improvise our model further in future by using machine learning algorithms.

## V. CONCLUSION

In the proposed system, we can find the apt expansion of the acronym based on the context, even if the expansions are too similar. This was a major drawback in the existing system. We overcame it by converting the text context into word2vector for which we find the cosine similarity. In future enhancements, we tend to decrease the time complexity and improve our proposed model using machine learning algorithms to improve its accuracy in finding the apt acronym.

## REFERENCES


1. Mathiew Roche, Violaine Prince (University of Montpellier 2, France) "A web-mining approach to disambiguate biomedical acronym expansion",2010.
2. Yangli,Bo zhao,Ariel euxman,Fango tao (Mountain view, San Faciso , Menlo Park ,CA,USA) 'Guess me if you can: acronym disambiguation for enterprises",2018.
3. Akram Gaballah Ahmed, Mohamed Farouk Abdul Hady, Emaf Nabil (faculty of computers and Information, Cairo Egypt) "A Language Modelling approach for acronym expansion disambiguation",2015.
4. Dongjin choi,Juhyum shin.Eunji lee,Pankoo kim (Department of computer engineering ,Chosum University,Gwangju,republic of Korea)" A Method for recommending the most appropriate expansion of acronyms using Wikipedia", 2018.
5. Paolo Atzeni,Fabio Polticelli,Daniele Toti(Dipartimento di Indformatica e Automazione, Universita Roma tre)"A Framework for semi-Automatic Identification, Disambiguation and Storage of Protein-related Abbeiviations in Scientific Literature",IEEE2011.
6. S.Pakhomov, "Semi-Supervised Maximum Entropy based Approach to Acronym and Abbreviation Normalization in Medical Texts, "Association for Computational Linguistics, pp. 160-267,2002.
7. M.Hwang, D.Choi, and P.Kim,"A Method for Knowledge Base Enrichment using Wikipedia Document Information," An International Interdisciplinary Journal,Vol.13,no.5,pp. 1599-1612,September 2010.
8. D.sanchez and D.Isern, "Automatic extraction of acronym definitions from the Web", Journal of Applied Intelligence, vol 34,no 2,pp.377-327,April,2011.
9. https://www.webharvy.com/articles/what-is-web-scraping.html
10. https://medium.com/@aziryasin/web-scraping-made-easier-with-jsoup-4c07734ec600
11. https://towardsdatascience.com/introduction-to-word-embedding -and-word2vec-652d0c2060fa
12. https://jsoup.org/


## AUTHORS PROFILE

**Mr. K. Premkumar** is head of the computer science and engineering department in Sri Manakula Vinayagar Engineering College, Puducherry affiliated to Pondicherry University, Puducherry, India. Mr. K. Premkumar pursed Bachelor degree from Adhipara Sakthi Engineering college and Master degree in Computer Science and Engineering from Sathyabama Deemed University, Chennai. He is pursuing his P.hd in the field of Vanet at Manonmaniam Sundaranar University, Tirunelveli. He has 16 years of teaching experience.

**V. Atchayaa** is currently pursuing her Under Graduate Degree in Sri Manakula Vinayagar Engineering College, Puducherry in the Field of Computer Science and Engineering. She is interested in the research works on web scraping, testing tools and Enterprise Resource Planning Systems She had worked in performance testing tool JMeter to check the efficiency of the websites. She had worked in SAP to automate the business work flow.

**P. Idayavalli** is currently pursuing Bachelor of Technology in Sri Manakula Vinayagar Engineering College, Puducherry. She is from Computer Science and Engineering Department. Her Current research interests id in the field of Web development, Data science, cognitive computing. She worked in the medical health alert project which helps to identify the patients affected by most prominent disease in that area.

**R. Gayathri** is currently pursuing Bachelor of Technology in the field of computer science & Engineering in Sri Manakula Vinayagar Engineering College, Puducherry. Her current research interest is in the area of cyber security. Her research interest in Machine Learning. She worked on the project "Spectrum sensing using cognitive radio for VSAT".