A web based approach: Acronym Definition Extraction

R. Menaha*, M.Barkavi , P.Guha Prashanthini , R.Narmadha

*Assistant Professor, Information Technology, Dr. MCET, Tamil Nadu, India

B.Tech - Information Technology, Dr. MCET, Tamil Nadu, India

Abstract

Acronyms are widely used in the tasks like web searches, tweets, text messages, and mails etc. Acronyms are typically ambiguous and often disambiguated by context words. Due to its dynamicity, different approaches has been experimented in the past decade to extract acronym definition. Different manual edited websites like acronym finder, acronyms. the free dictionary.com, nalingo.com is accessible to list the definition of an acronym. However an automated method is required to support this identification. This paper presents an automatic web based approach to extract the definitions of an acronym. The proposed system uses the web resources of Google, Wikipedia, Bing and Acronym Finder to identify the definitions of an acronym. From Google and Bing web pages, the snippets and titles are extracted and pattern extraction algorithm is applied to identify the definitions. Similarly the Wikipedia and acronym finder web pages are extracted to find the acronym definitions. The extracted definitions might be applied in information retrieval, question answering system and query expansion area as future work.

Keywords: *Acronym definition extraction, Abbreviation extraction*

----------***----------

1. INTRODUCTION

Acronyms are abbreviations formed from the initial components of words or phrases. Acronyms are textual forms used to stress the importance of entities and provide an alternative way to refer the same entity which is easier to understand .Some of the characteristics of acronyms are:

1. **Dynamicity:** New Acronyms are defined in every domain and every day. This is evident in social networks like twitter, LinkedIn, online chat.

2. **Ambiguity**: Each acronym has different meanings in different domain [e.g., SOAP has 35 definitions in acronym finder and 32 definitions in acronyms.thefreedictionary.com]. This can be disambiguated by using context words [e.g., SOAP XML refers Simple Object Access Protocol, SOAP Analysis refers Spectrometric Oil Analysis Program].

3. **Diverse of Generality:** Some acronym –definition pairs are commonly used, and some of them are very rare [e.g., NASA –National Aeronautics and space Administration [most common], NASA [Newspaper Advertising Sales Corporation].

Usage of acronyms is more in the applications like Biomedicine, Natural language Processing, Ontology population, question answering system and web search. Due to its dynamicity, it is difficult to maintain up-to-date lexical repository of all the acronyms and their meanings. Some of the manually edited websites like acronym finder, acronyms.thefreedictionary.com have millions of acronyms and meanings. Automatic discovery of acronym definitions also attempted in the past decade most of them are language dependent. This paper proposes an automated web based approach to identify the definitions of an acronym. It uses the web resources of google, Bing, Wikipedia, and acronym finder. The system extracts the snippets, titles and apply pattern extraction algorithm to discover the list of definitions. The rest of the paper is organized as follows. Section 2 defines previous works in the area of acronym definition discovery. Section 3 describes in detail the proposed methodology. Section 4 presents the results and evaluation, showing the obtained results and discussing them. The last section gives some conclusions and proposes some lines of future work.

2. RELATED WORKS

Andr'e Kempe [1], the report shown how the alignment-based approach to acronym-meaning extraction can be implemented by means of a 3-tape weighted finite-state machine (3-WFSM). The 3-WFSM will read a text chunk on tape 1 and an acronym on tape 2, and generate all possible alignments on tape 3, inserting dots to mark which letters are used in the acronym.

CvetanaKrstev, DuskoVitas, RankaStankovi'c[3] presented a comprehensive approach to acronyms for Natural-Language Processing (NLP) of Serbian texts. The procedure includes extraction of acronyms and their definitions that are usual Multi-Word Units (MWUs), shallow parsing of MWUs that enables MWU lemmatization and production of entries in morphological electronic dictionaries, both for MWU and acronyms, that are provided with grammatical, syntactic, semantic and domain information

Dana Dannels [4], the approach is based on that acronym-definition pairs follow a set of patterns and other regularities that can be usefully applied for the acronym identification task. Supervised machine learning was applied to monitor the performance of the rule-based method, using Memory Based Learning (MBL).

David Sanchez, David Isren [5], proposed methodology which has been divided into several part, first a simple algorithm generates the possible acronym by the combination of alpha numeric characters of a specific length. For all those generated candidates, the system tries to discover all possible definitions from the web. Finally the list of definitions found is filtered using a set of general rules.

Jun Xu.Yalou Huang [7], proposed a novel machine learning approach to extract acronyms from text. First, all likely acronyms are identified by heuristic rules. Second, expansion candidates against each likely acronym are generated from the surrounding text. He used the support vector machine (SVM) model to select the genuine expansions for acronyms.

Sunghwan sohn, Donald C Comeau, Won Kim and W John Wilbur [10], proposed an abbreviation identification algorithm that employs a number of rules to extract potential SF-LF pairs and variety of strategies to identify the most probable LFs.The reliability of the strategy can be estimated which they term pseudo – precision (P-precision).

3. METHODS

The list of acronyms are extracted from four different web resources. They are Google, Bing, Wikipedia and Acronym Finder. The method that are involved in the expansion identification differs for different web resources. The outline of proposed system implementation is given in fig [1.0]

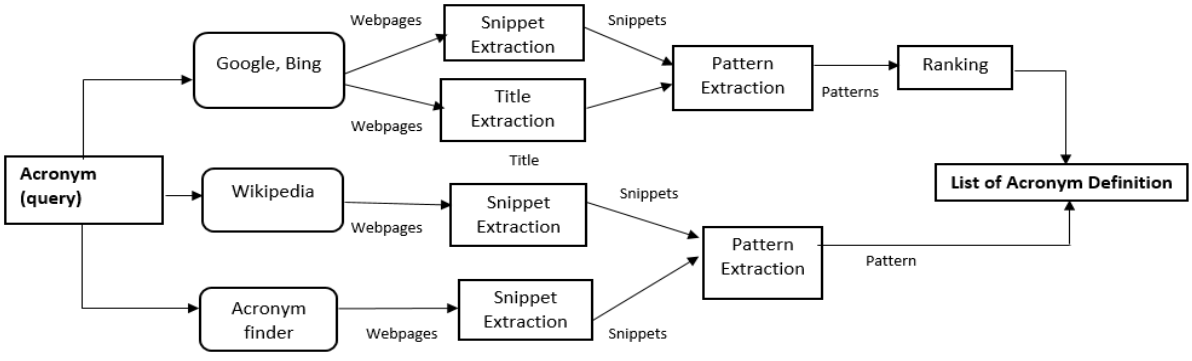

Fig 1.0 Outline of system implementatio

3.1 Google and Bing Web pages

Google and Bing pages contains snippets, titles, URL and other information sources. Usually the snippets and titles contains the acronym definitions of given input acronym query. Hence the snippets and titles are extracted from Google and Bing web pages they are stored in a local repository. Pattern extraction algorithm presented in fig 5.0 is applied to extract the patterns from snippets and titles.

3.1.1 Snippet Extraction

The snippet contains the small hint of information that the user request from the search engine. Snippets are useful for searches because most of the time, the user can read the snippet and decide whether particular search result is relevant without opening the URL. The processing of snippet is also efficient because it prevent from errors in downloading web pages, which might be time consuming depending on the size of the web pages. This extracted snippet is stored in the local database, once the new data comes existing data gets overridden.

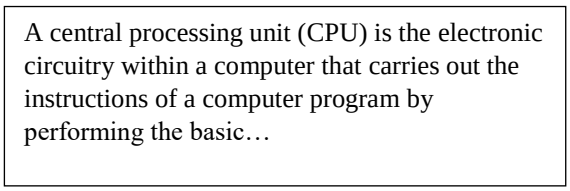

Fig 2.0 Snippet extracted from google for CPU

3.1.2 Title Extraction

Web pages contains the title which shows the detail that are available in the document may contains the definitions of an acronym.

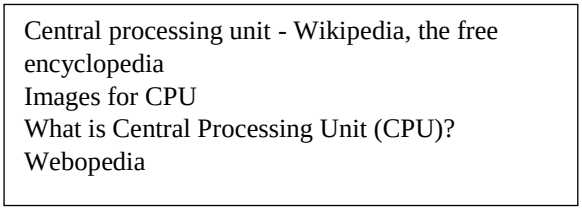

Fig 3.0 Title extracted from google

```
Input: Acronym
Output: List of Acronym defifntions
/* Google, Bing, Wikipedia and Acronym finder web pages are used for the acronym definition extraction */
/* Google and Bing */
do
For each acronym
        a. α1 ← Extract the google web pages
        b. S1 ← Extract the snippets from α
        c. T← Extract the titles from α
        d. P1 ← Extract the patterns from S and T. end
/* Wikipedia */
do
For each acronym
        a. α2 ← Extract the Wikipedia disambiguation web pages
        b. S2← Extract the required snippets from α
        c. P2 ← Extract the patterns from S2.
end
/* Acronym Finder */
do
For each acronym
        a. α3 ← Extract the Wikipedia disambiguation web pages
        b. S3← Extract the required snippets from α
        c. P3 ← Extract the patterns from S3.
End
/*** To extract patterns Pattern Extraction algorithm is applied ***/
```

Fig 4.0 Outline of System Implementation

3.2 Wikipedia

Wikipedia follows the procrastination principle regarding the security of its content. It started almost entirely open—anyone could create articles, and any Wikipedia article could be edited by any reader, even those who did not have a Wikipedia account. Modifications to all articles would be published immediately. Whenever an acronym is searched through Wikipedia, it retrieve the results in various contexts and meaning using fig.5.0. The pages can be stored in a local repository to extract snippets and titles. In Wikipedia the web pages gets extracted, and the Acronym definitions are extracted from the disambiguous webpages. The patterns for input acronym query is retrieved using the pattern extraction pseudocode given in fig.5.0

3.3 Acronym Finder

The web pages of Acronym Finder contains the expansion for input acronym. The Acronym Finder would contain five domains like Science and Medicine, Business, Military and Government, Organizations and Information Technology. First of all, the user would be asked to select the domain and then the acronym must be given. To extract web pages, an acronym is given as input to the search engine and the results of search engine are stored in local database. From the extracted web pages, the expansions are fetched using pattern extraction procedure using fig.5.0 and stored in lexical repository. In acronym finder the web pages are extracted, which contains the list of acronym expansion from different domains. Based on the user interest the corresponding meaning of an acronym in its domain are listed.

```
Input: Acronym
Output: List of Acronym defifntions
/* Google, Bing, Wikipedia and Acronymfinder web pages are used for the acronym definition extraction */
/* Google and Bing */
do
For each acronym
        a. α1 ← Extract the google web pages
        b. S1← Extract the snippets from α
        c. T← Extract the titles from α
        d. P1← Extract the patterns from S and T. end
/* Wikipedia */
do
For each acronym
        a. α2 ← Extract the Wikipedia disambiguation web pages
        b. S2← Extract the required snippets from α
        c. P2 ← Extract the patterns from S2.
end
/* Acronym Finder */
do
For each acronym
        a. α3 ← Extract the Wikipedia disambiguation web pages
        b. S3← Extract the required snippets from α
        c. P3 ← Extract the patterns from S3.
End.
/*** To extract patterns Pattern Extraction algorithm is applied ***/
```

3.4 Patterns Extraction

A Pattern is a series of action or event or behavior that together show how things normally happen or done. The following pattern extraction algorithm fig. 5.0 is devised to extract the acronym definitions from snippets and titles of Google, Bing, Wikipedia and Acronym finder.

```
Input: Snippets /Titles
Output : Lists of Acronyms
Variables used: L,S ,F, T flag;
    1. L←  Find the length of the Acronym
    2. S [0-L]←  Store each character of given acronym in an array.
    3. Read the snippets / titles.
    4. Replace all the numbers [0-9] and Special symbols [, " . ? etc.,] with empty space.
    5. Remove stop words from the snippets and titles. [e.g., a,an, the , between , etc.,]
    6. F← Read the snippet / title file and store it into buffer
          While (F! =eof)
          {   a. Read line from buffer F and apply tokenization.
              b. T← Read token
              c. If(T!==NULL)
              {       for (i=0;i<l;i++)
                        T[i]← Read the first character of token
                        if (T[i]==S[i])
                        {       flag++;
                                add the  token into wordlist
                                goto step b.
                        }
                        else
                           goto  step b.
                         if (flag==l)
                                add that wordlist into pattern file.
              }
        }
```

Fig 5.0 Pattern Extraction Pesudocode.

5. RESULTS AND DISCUSSION

Thus the expansions of acronyms have been extracted from four web resources like Acronym Finder, Bing and Google Pages and Wikipedia. For experimental purpose 100 acronym definitions are extracted from that 4 web resources. As an example 5 acronyms and extracted definitions from acronym finder, Bing, Google and Wikipedia is listed in table 1.0. For implementation, Java, JSON, Swing and Google search engine API were used. The extracted definitions will be applied in query expansion system as future work.

6. CONCLUSION

Acronyms are widely used in various applications. This paper proposed a method on extracting acronym definitions from Google, Bing, Wikipedia and acronym finder. Acronym definition extraction for about hundred acronyms is extracted successfully from the various web resources. As a future work extracted definitions would be applied in query expansion for effective information retrieval.

Acronym	Acronym Finder	Bing and Google	Wikipedia
CPU	Central Processing Unit Communist Party of Ukraine Chemical Production Unit Central Philippine University Commonwealth Press Union Central Policy Unit Computer Power Use Cost Per Unit Call Pick Up Critical Path Update	Central Processing Unit Computer Processing Unit Common Party User Cost Per Unit Columbia Pacific University	Central Processing Unit Central Philippine University Commonwealth Press Union Clark Public Utilities Columbia Pacific University China Pharmaceutical University Chemical Production Unit
SOAP	Subjective Objective Assessment Plan Simple Object Access Protocol Seal Of Approval Process Symbolic Optimal Assembly Program Spectrometric Oil Analysis Program Summary On A Page Small Operator Assistance Program Students Organized Against Prejudice Students Organized Against Poverty	Simple Object Access Protocol Summary On A Page Society Of Airway Pioneers Strategy On A Page Seal Of Approval Process	Simple Object Access Protocol Symbolic Optimal Assembly Program Spectrometric Oil Analysis Program Snakes On A Plane Students Organized Against Prejudice Students Organized Against Poverty
NSS	National Security Strategy Network Switching Subsystem National Service Scheme Names Service Switch National Service Switch Network Security Scanner Nadal Switching System Naval Surface Strike	National Security Service Network Switching Subsystem Novel Storage Services Not So Sure	National Shelter System National Security Service Network Switching Subsystem National Service Scheme National Scheme Services New Skies Satellites Nair Service society
NBA	National Basketball Association National Blood Association Next Best Alternative National Band Association Narmada Bachao Andolan National Book Award National Boxing Association National Business Association	National Book Award National Basketball Association No Boys Allowed Network Behavior Analysis Netbook Book Agreement N-Butyl Alcohol	National Basketball Association Narmada Bachao Andolan National Book Award National Boxing Association Nepal Basketball Association National Braille Association

Table 1.0 Extracted definitions for CPU, SOAP, NSS and NBA from Google, Bing, Wikipedia and Acronym Finder.

REFERENCE

[1] Andr´e Kempe, "Acronym-Meaning Extraction from Corpora Using Multitape Weighted Finite-State Machines", Springer Verlag, arxiv.org, Dec 2006, cs/0612033

[2].Bilyana Taneva1∗, Tao Cheng2, Kaushik Chakrabarti2, Yeye He "Mining Acronym Expansions and Their Meanings Using Query Click Log", WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil. ACM 978-1-4503-2035-1/13/05.

[3] Cvetana Krstev, Duško Vitas, Ranka Stankovi´ , "A Lexical Approach to Acronyms and their Definitions"

[4]. Dana Dannels, "Automatic Acronym Recognition", ACM DL, EACL '06, Pg:167-170

[5] David Sanchez, David Isren, "Automatic extraction of acronym definitions from theWeb" Apple Intell(2011) 34: 311-327 DOI 10.1007/s 10489-009-0197-4

[6] Jain A, Cuzerzan S, Azzam S, "Acronym –Expansion Recognition and Ranking on the web, IEEE International conference on Information Reuse and Integration 2007, pg:209-214. DOI: 10.1109/IRI.2007.4296622.

[7] Jun Xu.Yalou Huang, "using SVM to extract acronyms from text", Springer Verlag 2006, DOI 10.1007/s00500-006-0091-5.

[8] *Leah S. Larkey, Paul Ogilvie, M. Andrew Price, Brenden Tamilio, "Acrophile: An Automated Acronym Extractor and Server", ACM 2000, Pg:205-214, DOI:10.1145/336597.336664*

[9] Min song, Peishih Chang, "Automatic Extraction of Abbreviation for Emergency Management Websites",ISCRAM conference- Washington May 2008, Pg:93-100.

[10] Sunghwan sohn, Donald C Comeau, Won Kim and W John Wilbur, Abbreviation definition identification based on automatic precision estimates, BMC Bioinformatics 2008, 9:402 doi:10.1186/1471-2105-9-402. Sep 2008.

[11] www.acronymfinder.com