International Journal of Research in Advent Technology, Vol.7, No.4, April 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Deriving Meaning of Trending Words for Expanding Thesaurus

Bhoomi Joshi[1], Bhavesh Borisaniya[2], Darshankumar Modi[3]
Dept. of Information Technology
Shantilal Shah Engineering College, Bhavnagar
Gujarat Technological University, Ahmedabad, India
joshibhoomij@gmail.com[1], borisaniyabhavesh@gmail.com[2], darshan.modi91@gmail.com[3]

Abstract— Nowadays there is a huge increase in the types of new or trending words being generated and used across social media. Though these words are used frequently, neither these trending words i.e. buzzwords are present in any dictionary, nor are they declared as official word-forms. However, these are the terms that drive a language to a different level. Thus, to take a look at the linguistic changes occurring across the world, this study is carried out of the approaches that can be used to find such new words that are highly in use, but have not yet attained an official status. Finding such words can be helpful in tracking the language change. Therefore, an approach for finding such words and imparting semantic relationship to find out their possible meanings or opposites is proposed in this paper.

Index Terms — Big Data, Thesaurus, Thesaurus Generation, Trending Words, Buzzword

1. INTRODUCTION

Do languages change? Yes, they do. Overtime languages keep changing, adapting and evolving. The change is so gradual that it is very hard to predict changes occurring in a span of a few years. It takes decades to notice these evolution patterns.

So, the question arises that why the languages change? Change in languages takes place due to various reasons. First, it can be the demand or need of the speakers of the language. Second, it is the technological advancements, new experiences, new products and new innovations that are required to be called by different names or new names, for example, cell-phone, scanner which leads to innovation of new words in a language [1].

Third reason is that every person in their lifetime has been through a different set of words. No two persons' set of words can be exactly the same because they have been through different experiences [1] in their life, hence have dealt with a different set of words. Another reason for the language change can be the age group, education, regional influence of the speaker [2].

There has been a huge increase in the types of new words used across social media. Mostly the short forms i.e. abbreviations or the trending words used frequently are not present in the dictionary, even those words or abbreviations are not declared as official word-forms. The words that are so commonly used are directing language in a different direction. Hence, finding such trending words can be useful for linguistic study.

Efforts [3] have been made to find out such words and keep a track of the language change taking place around the world. The techniques used up till now were not up to the mark for automatic thesaurus generation. There are various techniques which can be used to optimize the resultant set of words, found out after the extraction process. By using new techniques it is possible to get quicker as well as optimal thesaurus. In this paper an approach is suggested which can be used for generating thesaurus automatically.

The rest of the paper is organized as follows: Section 2 discusses about thesaurus and thesaurus generation; Section 3 contains the related work carried out for this research; Section 4 consists of the proposed approach that can be carried out to perform thesaurus generation; section 5 concludes the paper.

2. THESAURUS AND THESAURUS GENERATION

A thesaurus is a work of reference that provides words' list and groups together in accordance to the nearness of context i.e. containing synonyms and sometimes antonyms [4]. Thesaurus varies from a dictionary. A dictionary lists the words in alphabetical order with their respective definitions. Whereas, the main purpose for this work of reference is that users can find out the words that can sufficiently express an idea.
The process of thesaurus generation has been carried out manually up till now. Though trending words were found out using frequency of their occurrences [3], their synonyms or antonyms were manually imparted. Thus, automated thesaurus generation can be carried out using following basic steps:

1) Recognition of concepts or terms [5]:
   This step recognizes and extracts the terms related to particular criteria or as desired to be retrieved.

2) Extraction of semantic relationships between those terms [5]:
   In this step the correspondence between the input and output is established and further extracted. The relationship justifies that why a particular word is generated as output to a specific input.

3) Derivation of the results:
   The relationship derived is the result of thesaurus generation process. The trending word and the word the related words generated are then added to the existing thesaurus. This is how thesaurus can be automatically generated.

3. RELATED WORK

This section contains the related work that has been carried out for attaining the goal of automatic thesaurus generation. Various approaches from different literatures have been referred to form a sequence of steps needed to be carried out for automatic thesaurus generation.

In an approach proposed by Jack Grieve et al. [3] words occurring minimum for 1000 times were extracted from 8.9 billion word-corpora for analysis. Multi-word units were not analyzed. Two forms of a word can have two different meanings, converting them to its base form (lemmatizing them) can harm the semantic relationship of the words. Lemmatizing words can remove some useful words, so lemmatization was not performed on words. [3]

Later the relative frequencies were calculated for those frequently occurring words and those relative frequencies were multiplied by 1 billion to get the per billion words (PBW) count of the word to obtain a normalized frequency range.

Relative frequency at the initiation of the rise and fall of the word had been calculated and the average relative frequency for a day was also calculated. Spearman correlation coefficient [6] is a technique to find monotonic patterns, where increase in value of a variable shows increase in another variable too. Spearman correlation coefficient was used to determine the shape formed due to rise and fall in use of a word.

Initially when a particular word was used, commonness of the form was represented by relative frequency. Spearman correlation coefficient represents the degree to which the form of the word usage has risen in frequency over the time course. With an average relative frequency of less than 1000 per billion words and with a Spearman correlation coefficient of larger than 0.80[3] (which is a strong correlation coefficient as assumed by Jack Grieve et al.) 131 word forms were extracted for the next stage of process. Then the spearman's coefficient of such words and relative frequencies were plotted on a graph to find out a set of new emerging word forms.

Proper nouns, advertisement, technical and medical terminologies were manually excluded from the resultant set. The remaining set of words was taken as a resultant set. Meanings to the resultant set were imparted manually and were represented.

In a method proposed by Dan Li et al. [7] phrases occurring more often than expected are referred as collocations. As shown in figure 1, the collocations are extracted using a particular window size sliding through the document. After selecting collocations which are syntactically correct, were forwarded to the next phase. In the next phase, those collocations were semantically checked and passed to the next phase. In the last phase of the analysis collocations were checked on the frequency of their occurrences, thus, deriving final set of collocations.

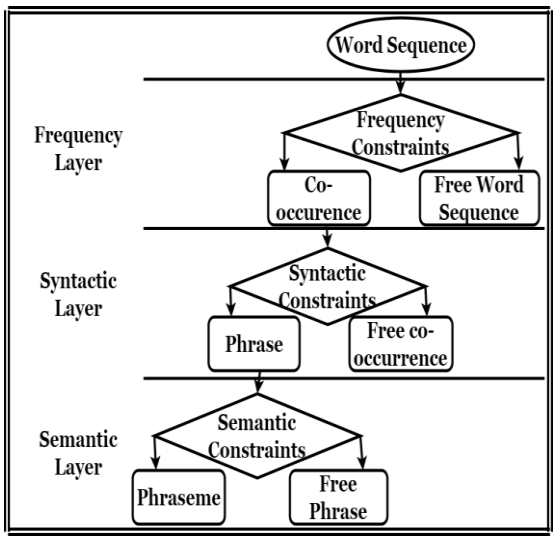

Figure 1: Three Layers of Collocation Extraction [7]

News headlines can be constructed from whole of the news article in [7]. For this purpose, 5 systems were used:

1) Preprocessing
2) Keyword/Key-phrase extraction
3) Parse tree generation
4) Compressed sentence generation
5) Headline construction

Underlying the above 5 systems worked the following techniques [7]:

1) Sentence segmentation
2) Tokenization
3) Stop words removal

4) Stemming
5) Keyword/ key-phrase extraction
6) Parsing of the sentence

Here, segmentation, tokenization, stopwords removal and stemming are a part of preprocessing system. Whereas, parse tree generation, compression of sentence and headline construction deal with parsing of the sentence.

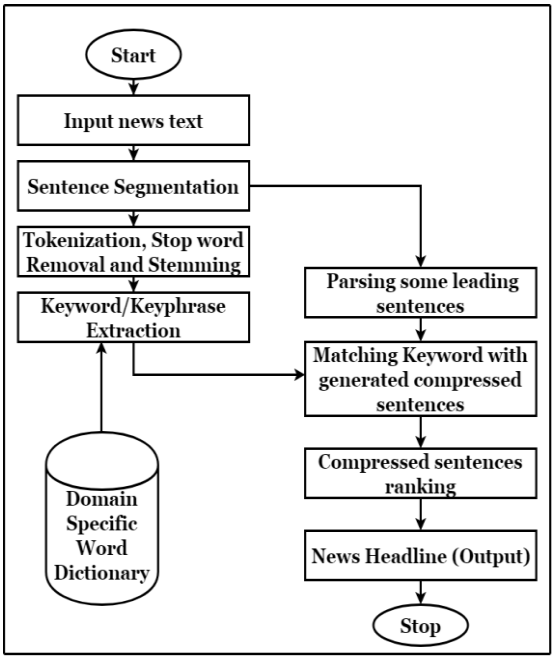

Figure 2: System Model [8]

It was assumed that the first sentence of the news article might contain most of the content of news in brief [8]. Hence, the first sentence was termed as the leading sentence of the article. As shown in Figure 2, the news article was taken as an input corpus. The article was segmented to sentences. Then leading sentence was taken on a different processing route, whereas, the other sentences were sent for further tokenization. After performing tokenization and other cleaning tasks like stop words removal and stemming, the result was directed to the key-phrase extraction algorithm which extracted terms and phrases relative to the news article in correspondence with the domain specific dictionary.

The words received as the output of key-phrase extraction algorithm and the words of the leading sentences are matched to generate a compressed sentence. Those compressed sentences are given rankings and are parsed to validate it syntactically and semantically into a correct sentence [8]. Thus, this approach generated news headlines from a descriptive news article.

From the methods studied from various literatures, we can combine a few methods to design an approach which helps in automatic thesaurus generation. Following proposed is an approach for the same.

4. PROPOSED APPROACH

Initially, thesaurus generation was carried out manually. Even extraction of the trending words was carried out just by finding out the relative frequency and spearman co-relation coefficient of the words [3]. Later, on the meanings of the words were imparted manually [3]. Figure 3, shows the steps of proposed approach which can be implemented in order to find out the trending words and give them appropriate synonym or antonym automatically.

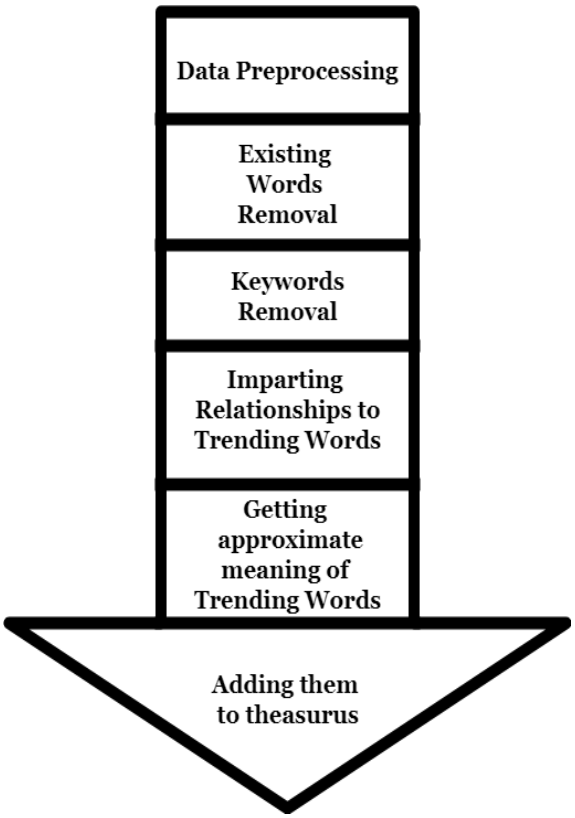

Figure 3: Proposed Workflow

1) Data Pre-Processing

Data Pre-processing: Preprocessing is an important step because it removes unwanted data which helps in increasing the speed of execution and hence optimizing the process.

There are three basic steps in Data Processing:

a. Data Extraction
b. Data Storing
c. Data Cleaning

Data Extraction operation extracts the data field from the unstructured data files.

Data Storing operation store extracted data field in a manner on which, further operations can be performed optimally.

Data Cleaning is done to remove the unwanted data which will be of no use in the future.

2) Stopwords Removal

In this phase, the stop words like *is, am, are, he, she, it, etc*. such articles, conjunctions and pronounces are to be removed. For this purpose, some libraries like *python NLTK* [9] could be used to remove stopwords hence, decreasing amount of text to be processed.

3) Existing Words Removal

After removing stopwords from the previous step, further a process could be started to remove the words already existing in thesaurus to further reduce the size of data.

4) Keywords Removal

From the left over data, in this phase we extract the keywords such as technical or medical or some special words with correspondence to domain dictionary to obtain the trending words.

5) Imparting Relationships To Trending Words

After obtaining trending words relationships can be established between words by algorithms that predict nearby word-form using classification or clustering. Levenstien distance [10] or Latent Semantic Analysis (LSA) [11] [12] can be used to establish relationships and it may be possible to get approximate meaning or synonym of the new words.

6) Generating Thesaurus

After deriving trending words and its meanings, those trending words with their respective related word can be added to thesaurus.

By implementing the above proposed method, addition to a current thesaurus can be made by adding the trending words to current thesaurus.

5. CONCLUSION

Automatic generation of thesaurus is a major challenge. It is a tedious task to impart accurate relationship between the trending words and the words which actually are synonym or antonym to those trending words. So, in this paper an approach is designed for the extraction of the possibly correct relationships between the new words and existing words. In future, we will make an attempt to implement the proposed workflow and show the results for automatic generation of thesaurus.

REFERENCES

[1] Bryson, Bill. (1991) Mother Tongue: The English Language. New York: Penguin Books, pp.1-20
[2] Aitcheson, lean. (1991) Language Change: Progress or Decay? Cambridge: Cambridge University Press, pp.20-30
[3] Jack Grieve, Andrea and Diansheng Guo (2018) Analyzing lexical emergence in Modern American English online University of South Carolina, pp. 1-44
[4] Roget Peter (1852) Thesaurus of English Language Words and Phrases.
[5] Ksenia Lagutina, Eldar Mamedov, Nadezhda Lagutina, Ilya Paramonov, Ivan Shchitov (2016) Analysis of Relation Extraction Methods for Automatic Generation of Specialized Thesauri: Prospect of Hybrid Methods P.G. Demidov Yaroslavl State University, Yaroslavl, Russia PROCEEDING OF THE 19TH CONFERENCE OF FRUCT ASSOCIATION, pp.138-144
[6] Spearman, Charles. (1906) 'Footrule' for measuring correlation.Brit. J. Psychol., pp.89–108
[7] Dan Li, Jingxiang Cao and Degen Huang Dalian (2015) A Hierarchical Collocation Extraction Tool Dalian University of Technology, IEEE Fifth International Conference on Big Data and Cloud Computing pp.52–55
[8] Urmila Shrawankar and Kranti Wankhede (2016) Contruction of News Headline from Detailed News Article G. H. Raisoni College of Engineering, Nagpur, Maharashtra , India, IEEE
[9] S Bird, E Klein, E Loper (2009) Natural Language Processing with Python
[10] G. Fazekas, V L Levenshtien (1995) On Upper Bounds For Code Distance And Covering Radius Of Designs In Polynomial Metric Spaces, Journal of Combinatorial Theory, Series A, Elsevier
[11] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman (1990) Indexing by Latent Semantic Analysis, John Wiley & Sons, Inc.
[12] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester and R. Harshman (1998) Using Latent Semantic Analysis to Improve Access to Textual Information, Washington D.C., USA