IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Semantics discovery of Short Text

Miss. P. G. Kamble [1*], Prof. S. B. Bhagate [2]

[1]Dept. of Computer Science and Engineering, D.K.T.E'S TEI (An Autonomous Institute), Ichalkaranji, India.
[2]Dept. of Information Technology, D.K.T.E'S TEI (An Autonomous Institute), Ichalkaranji, India.

[*]*Corresponding Author: pournimakamble05@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 23/Jul/2018, Published: 31/July/2018

***Abstract*—**Currently every person's use short text in real life for communication and chatting purpose. Short texts are also uses in social posts, news titles, events, search queries, tweets, conversations, keywords, Short text understanding is a confusing process in ideas deals with secret messages. The short text is produce that contain social posts, discussions, keywords and news titles which are restricted context and represent the significance of the text or insufficient information. As short text has more than one meaning, they are challenging to understand as they are noisy and ambiguous. The term can be any single or multi-word. Short texts do not contain satisfactory information. Some short texts have unique features. So these short texts are difficult to handle. It essential well understand the short text. Semantic analysis is necessary to understand the short text properly. Tasks such as part-of-speech tagging, concept labelling and segmentation are used for semantic analysis. Conduct short text uses in real life information. The prototype system is constructed and used to recognize the short text. These systems deliver the semantic knowledge from knowledge base and collection of written words that are automatically harvest. Creating construction of co-occurrence network and term extraction showing to better understand for short text.

***Keywords*—**Short Text, Semantics, Text segmentation, co-occurrence,Term Extraction .

I. INTRODUCTION

Data mining is to identify useful information from raw data and process to explore implicit, useful and understandable relationships in large amount of data. It extracts knowledge from large amount of data. Data mining is part of the knowledge discovering process. Knowledge discovering process consist of Data cleaning, Data integration, Data transformation, Data selection and Knowledge representation. Data cleaning deals with elimination of noise and irrelevant data from the collection of data. Data integration is combining data from several sources and provides unified data. Converting information from one format to another format is called Data transformation. Data selection is the process of gathering information from collection of data. Knowledge representation is the technique of designing computer representations of captured information.

The short text is a group of words or phrases with limited context generated by search queries, tweets, ad keywords, subtitles, document titles, and the like. Short texts are commonly used in social posts, news titles, events, search queries, tweets, conversations, keywords. Short text understanding is very confusing process. A better understanding of short text is to eliminate the hidden semantics in the text. In addition, much interest is in analyzing and conceptualizing short text.

Text mining is the analysis of data from natural language text. Text mining involves information and data retrieval for tagging, information extraction, and pattern recognition and predictive analytics. Text mining is the process of extracting non-trivial patterns from unstructured text documents. Selecting the correct keywords for search is the most important component. Even with many keywords, search results do not always deliver what is expected to the user. Improving the accuracy of search is important and one of the best ways to do this is to incorporate text mining. Semantics is the study of relationships between the words and construct the meaning of short text. It is interpretation of word, sign, and sentences. In conventional data mining techniques, the data is not available and semantic information of data is not accessible in detail. Semantics is a study of meaning of language and word. Semantics can be categorized as Formal semantics and lexical semantics. Formal semantics is the study of logical aspects of meaning. Lexical semantics is the study of meanings and relations between words. Text is human readable sequence of the characters. Text can be classified into different categories like organization, location, persons. Abbreviation of the text is also known as Short text. Sufficient information is not contained in short text to support text mining approaches. Short text may be noisy so it is difficult to handle. For examples app. (application), ATM. (Automatic tailor machine), i.e. (that is).Short text can be used in many applications like web search, message, query, tweets and news titles. Short text is difficult and ambiguous

to understand, because it has multiple meaning. There is need to better understand the meaning of short text and avoid ambiguity.

II. RELATED WORK

A Schutze and Y. singer proposed by Part-of speech tagging uses variable € memory Markov model (VMM) [1]. It is based on minimizing the statistical prediction error for a Markov model. It measured by instantaneous Kullback-Leibler and find a prediction suffix tree that has the same statistical properties as the sample, and it can be used to predict the next outcome for sequences generated by the same source. At each stage, it transforms the tree into a variable memory Markov process. It builds a prediction tree and measures the probability of equals the sample. VMM algorithm achieves average accuracy. It can be uses for pruning many of the tagging alternatives using its prediction probability; it does not complete tagging system. It is independence on assumption of tags and observes words.

M. Utiyama proposed by Text segmentation technique uses domain-independent model statistical approach [2]. It automatically partitions text into the related segment. It based on the technique that build an exponential model which, builds features of the text. It specifies the near boundary of word segment. It detects the occurrence of specific words. It only considers a surface of feature. It ignores the requirement of semantic coherence. It may lead to incorrect segmentation.

Nikita Mishra proposed by unsupervised query segmentation scheme uses query logs [3]. It as the effectively capture structural units of queries. It helps the understanding grammatically structure. It implemented a statistical model based on Hoeffding's difference to extract necessary word n-grams from doubts and subsequently use them for segmenting the queries. It technique can detect limited units that removed from queries conditions based on PMI baseline. Evaluation of segmented the queries across manually segmented queries.

Dong Deng proposed by Trie-based Method uses Approximate Entity Extraction with Edit-Distance Constraints [4]. It considers the smaller index size and its efficiency for large edit distance threshold. It is used to edit distance threshold. Each term evenly divides into a number of segments. A substring is similar to a term concerning the threshold. It must contain one segment of that term. Every substring of short text is considered. It checks whether text matches with the segment or not. It requires different edit distance threshold. Trie-based framework utilizes one specific edit distance threshold. The vocabulary contains a large amount of abbreviations and multiword instances. Longer terms may lead to misspell and mistakes in this system.

Peipei Li proposed Computing Term Similarity by Large Probabilistic isA Knowledge that uses Knowledgebase Approach [5]. It is used to knowledge base taxonomy to compute a similarity between two terms and find the shortest path from two terms in taxonomy graph. It is simple but low accuracy. Because taxonomy graph links represent uniform distance. It ignores the amount of information of terms.

W. Hua proposed Short text understanding through lexical-semantic analysis that uses the generalized framework to effectively and efficiently understand the short text [6]. It has used randomized approximation algorithm to achieve better accuracy. It has used the text segmentation that divides the text into a number of sub-text. It takes the text as input form bag of words. It is insufficient to express meaning semantically. Statistical and rule-based approaches depend on the assumption that a text is correctly structured, but not always for short texts. The work only considers lexical features and ignores semantics.

Zheng Yu proposed by Understanding Short Texts through Semantic Enrichment and hashing uses Semantic Enrichment and Hashing [7]. It has used semantic hashing approach. The meaning of a text is encode into a compact binary code. If two text has similar meaning, then there is need to check if they have similar code. Each a short text represents a dimensional semantic feature vector. It captures co-relationship from the short text and also captures abstract feature from the short text. Auto-encoder specific learning function is designed, to do semantic hashing on these semantic feature vectors for a short texts. The output of threshold is a binary code. It is regard as semantic hashing code for input a text. A compact binary code is created for every short text. It checks the similarity of short text and matches with binary code.

Wen Hua, Zhongyuan Wang proposed to Understand Short Texts by Harvesting and Analyzing Semantic Knowledge uses Analyzing Semantic Knowledge [8]. Construct the co-occurrence of related terms in the dataset the vocabulary. It scans the terms in vocabulary. It is calculated a frequency of appearing term. Approximate term extraction is done to locate the substring in a text that is contained in the vocabulary. Concept labeling eliminates the ambiguity of terms.

III. METHODOLOGY

The user understands semantic of word based on extraction of words. So that gets actual meaning of word and understands the meaning of short text.

Fig.1. shows architecture of ESSTM, which consists of three modules.

1. Construct Co-occurrence Network
2. Term Extraction
3. Concept labeling

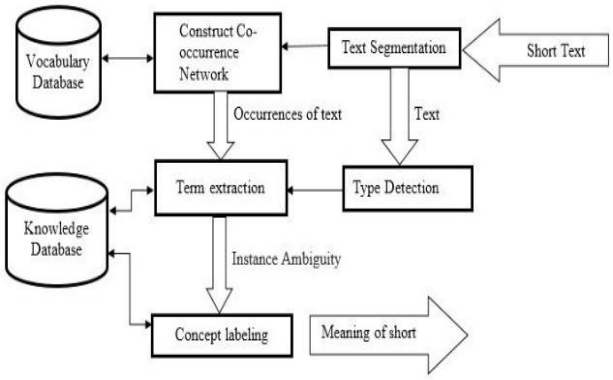

Fig 1.System Architecture

The conceptual view state that, it mainly consists of three modules first one is construction of co-occurrence network, second one is approximate term extraction and another is concept labeling. Before executing this modules system pre-process text data. For executing first module of construct co-occurrence network there is need to filter stop words and split the text into sub text for extracting term. System to need list of stop word for filtered text data. System have on semantic word dictionary having different words are stored.

A. Construct co-occurrence Network

I. Data Pre-processing

Consider each user's Short Text as a collection of words without considering the order and construct vocabulary then filter all Stop words from short text. After word filtering, get input text clear and without much interference. Then construct vocabulary V with all this unique words. This V is used to input for Construction of co-occurrence Network. Generate process for co-occurrence network.

In construct the vocabulary in ESSTM system short text are considered as a collection of words, without considering the order. Then filter all Stop words from text short text. After word filtering, get input text clear and without much interference. Then construct occurrences with all this unique words. This vocabulary is used for Input of Approximate Term extraction and Concept labeling module. The word is input of co-occurrence network. In the construction of co-occurrence network user want the meaning of the particular term to display the meaning of occurrence and its meaning. The one term belongs to different meaning of word. Apple is instance co-occurs with Concept Company and fruit. The co-occurrence network should be construct term instance and typed term.

Algorithm: Co-occurrence network

Input: Short Text
Output: Occurrence with weight

- **Generative process of Co-occurrence network**

Co-occurrence network assume that occurrence created with weight in the following way:

1. Determine number of words in document.
2. Divide the short text into typed term.
3. Calculate frequency of two typed term :
   $fs(\overline{x}, \overline{y}) := ns \cdot e\text{-}dist(x, y)$
4. Calculate aggregate frequencies among typed term :
   $f(\overline{x}, \overline{y}) := \sum_s f_s(\overline{x}, \overline{y})$
5. Calculate the weight of among typed term
   $w(\overline{x}, \overline{y}) := \frac{f(\overline{x},\overline{y})}{\sum_{\overline{z}} f(\overline{x},\overline{z})} \cdot \log\left(\frac{N}{Nnei(\overline{y})}\right)$

In construct co-occurrences of word from short text. In ESSTM first calculate the frequency of appearing typed term. In calculate the frequency divide the original text into number of typed term. Then calculate the aggregate frequency to use frequency among typed term and calculate the weight of that type term. Then find word and match to vocabulary. Calculate frequency using the following equation:

$$fs(\overline{x}, \overline{y}) := ns \cdot e^{-dist(\overline{x}\overline{y})} \quad (1)$$

Where ns denotes total number of times sentence s appear in the vocabulary.

–dist$(\overline{x}, \overline{y})$ denotes the total number of occurrence for typed term appear in the vocabulary .

After calculating the frequency system can using (eq.1). Calculate aggregate frequency the following equation.

$$f(\overline{x}, \overline{y}) := \sum_s f_s(\overline{x},\overline{y}) \quad (2)$$

Where fs $(\overline{x}, \overline{y})$ denotes frequency of typed term.

After calculating calculate aggregate frequency system can using (eq.2) by calculate the weightage typed term following equation.

$$w(\overline{x}, \overline{y}) := \frac{f(\overline{x},\overline{y})}{\sum_{\overline{z}} f(\overline{x},\overline{z})} \cdot \log\left(\frac{N}{Nnei(\overline{y})}\right) \quad (3)$$

Where $\frac{f(\overline{x},\overline{y})}{\sum_{\overline{z}} f(\overline{x},\overline{z})}$ denotes probability of aggregate frequency, N total number of typed term contain in the co-occurrence network, and Nnei$(\overline{y})$ is the number of occurrence neighbour of $\overline{y}$.

B. Term Extraction

The word is input of term extraction. In this term extraction can be done as per user requirement and actual meaning of the term is provided to the user. Related or similar term is provided to the user search. Find the all possible term for related text and determine the similarity between two strings and provide the exact meaning of short text. In the Tem Extraction extract the actual term in the obtaining co–occurrence Network. Construct co-occurrence network from calculating weigh pair of two typed term. These weight of typed term as input of Term extraction. In term extraction extract all pair of all typed term with its weigh. Then select probability of higher weigh of two typed term. After selecting pair of typed term remove all connected node. Then remove all connected edge to disconnected node. Get the actual typed term and also meaning of short text.

C. Concept labeling

The short text is input of concept labeling. Concept labeling is used to overcome the ambiguity of the term. Same name with different meaning is to be identified by specifying a label. So related term are used to avoid ambiguity. It is process of eliminating inappropriate short text behind ambiguous instance. Typed terms is obtained along with the weighted edges in between. Get the target instance term, related terms can be retrieved by comparing weight of edges connecting to the target instance.

Concept labeling is done using the formula

$$\bar{x}.W_i := \text{Vself(Ci)} \cdot \text{Vcontext(Ci)}$$

Here, $\bar{x}$ represent typed term, Vself(Ci) represent the term of occurrence and Vcontext(Ci) represent the weight of co-occurrence neighbour of term. It obtain from co-occurrence network and term extraction. Give an example of calculated weight of word in table 2

Algorithm: concept labeling

Input: Word.

Output: Occurrences of word with weightage.

1] Input as the word

2] Calculate the weightage of word with occurrences

$\bar{x}.W_i := V_{self}(C_i) \cdot V_{context}(C_i)$

3] Display the word with weight

Concept labeling is done using the formula

$$\bar{x}.W_i := V_{self}(C_i) \cdot V_{context}(C_i) \quad (4)$$

Here, $\bar{x}$ represent typed term, Vself(Ci) represent the term of occurrence and Vcontext(Ci) represent the weight of co-occurrence neighbour of term. It obtain from co-occurrence network and term extraction.

IV. Results and Discussion

A. *Timing required to process on text length of short text.*

Understanding short text is related to text mining applications. These applications process large amounts of short texts. The time requirement for short text increases linearly as the length of the text increases.

Following table shows the text length and timing required to process on different text length.

Table 1. Processing time for text length

Text length	Time (ms)	Time (ms)
1	6	9
2	8	11
3	10	15
4	11	16
5	14	18
6	14	19
7	15	19

Above table shows timing results to process the different length of texts. The timing depends on the actual length of the short text, and the size of the short text depends on the length of the text. This time varies for different short text.

Following figure 2 shows the line chart for different text length with corresponding time required for short text. The graph describes the impact of the text length on time. Observation shows that, the time decreases when the text length decreases. There is an improvement is less time required as compared to existing system.

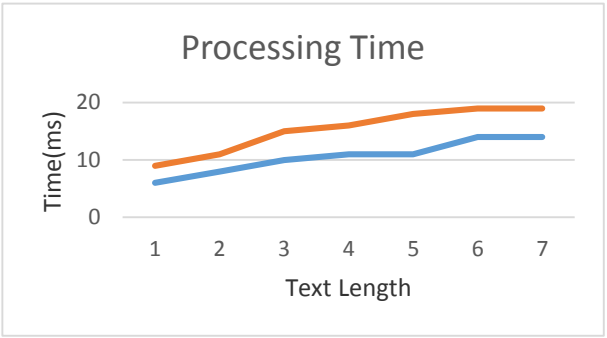

Fig.2. line graph of processing time for text length

Effectiveness of Short Text Understanding

The experiments are carried out to measure the effectiveness of short text understanding by concept labeling on different short text.

The system is evaluated for different occurrences of short text manually by concept labeling and the precision is

calculated. And conduct disambiguation in short text using concept labeling and the precision is calculated. The different short text and its corresponding precision is shown in table below

Table 2. Precision table for short Text understanding

Short Text	Time (ms)	Time (ms)
Nicknames	0.88	0.90
word	0.90	0.91
all	0.89	0.90

The comparison is done at term level and the precision is calculated for three term that are all, word, Nicknames manually by using concept labeling of occurrences of short text.

Following figure 3 shows the clustered column chart for different Short Text like nicknames, word and all. The graph describes the impact of the Short Text. Observation shows that, the instance ambiguity is decreases. There is precision improvement of an ambiguity in short text understanding.

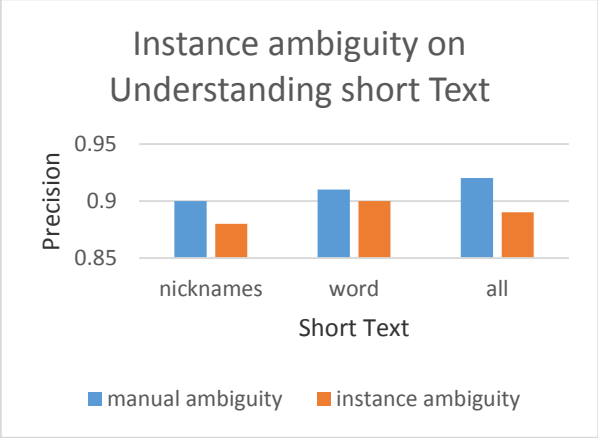

Fig.3. clustered column graph of Instance ambiguity

V. CONCLUSION AND FUTURE SCOPE

The conversion of short text in to long text is essential to understand short texts efficiently and effectively. Co-occurrence network, term extraction, and concept labeling methods are used for better understanding of short text with high accuracy, that too efficiently in less time. Concept labeling is a process of eliminating inappropriate short text from ambiguous instance. It avoids the ambiguity of the text which results into improved accuracy. Sometimes, the system may no detect all possible short text. In future, the system can be designed to attempt to analyse and combine the effects of spatial temporal features in understanding short text.

References

[1] M. Utiyama and H. Isahara, "A statistical model for domain-independent text Segmentation," in Proc. 39th Annu. Meeting Assoc. Comput. Linguistics, 2001, pp. 499–506.

[2] N. Mishra, R. Saha Roy, N. Ganguly, S. Laxman, and M. Choudhury, "Unsupervised query segmentation using only query logs," in Proc. 20th Int. Conf. Companion WorldWideWeb, 2011, pp. 91–92.

[3] D. Deng, G. Li, and J. Feng, "An efficient Trie-based method for approximate entity extraction with edit-distance constraints," in Proc. IEEE 28th Int. Conf. Data Eng., 2012, pp. 762–773.

[4] P. Li, H. Wang, K. Q. Zhu, Z. Wang, and X. Wu, "Computing term similarity by large probabilistic ISA Knowledge," in Proc. 22nd ACM Int. Conf. Inform. #38; Manage., 2013, pp. 1401–1410

[5] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Short text understanding through lexical-semantic analysis," in ICDE, pp. 495–506, 2015.

[6] Zheng Yu, Haixun Wang, Xuemin Lin, Senior Member, IEEE, and Min Wang, "Understanding Short Texts through Semantic Enrichment and Hashing". VOL. 28, NO. 2, FEBRUARY 2016

[7] Wen Hua, Zhongyuan Wang, Haixun Wang, Member, IEEE, Kai Zheng, Member, IEEE, and Xiao Fang Zhou, Senior Member, IEEE, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge" , VOL. 29, NO. 3, MARCH 2017

[8] *E. Brill, "A simple rule-b9ased part of speech tagger," in Proc. Workshop Speech Natural Language, 1992, pp. 112–116.*

[9] *R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer, and L. Ramshaw, "Coping with ambiguity and unknown words through probabilistic models," Comput. Linguistics, vol. 19, no. 2, pp. 361–382, 1993.*

[10] https://cracku.in/blog/nicknames-of-indian-states-and-cities-pdf/

[11] http://www.txtdrop.com/abbreviations.php

Authors Profile

Miss. Pournima Kamble pursed Bachelor of Engineering from SIT college of Engineering, Yadrav (Ichalkaranji) Shivaji University, Kolhapur, India in 2016. She is pursuing Master of Technology in Computer Science & Engineering from DKTE Society's Textile & Engineering Institute, (An Autonomous Institute), Ichalkaranji, 416115, India.

Mr. Suhas B. Bhagate pursed Bachelor of Engineering from Shivaji University, Kolhapur in 2003 and Master of Engineering from Walchand College of Engineering, Sangli in Shivaji University, and Kolhapur in year 2011. He is pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and Engineering, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji since 2004. He is IEEE Graduate Student Member. He has published more than 10 research papers in reputed international journals. His main research work focuses on Visual Cryptography Algorithms, Data Structures, Big Data Analytics and Data Mining.

