Impact Factor Value: 4.029 ISSN: 2349-7084
International Journal of Computer Engineering In Research Trends
Volume 4, Issue 11, November - 2017, pp. 519-523 www.ijcert.org

Various Mechanisms for understanding Short Text

Pournima G. Kamble*[1], S. B. Bhagate [2]

[1] *Dept. of Computer Science and Engineering,* DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute)*, 416115, India*

[2] *Dept. of Computer Science and Engineering, DKTE* Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute)*, 416115.*

pournimakamble05@gmail.com[1] , suhas.bhagate@gmail.com[2]

Abstract: -Now a day's all people use short text in real life for communication and chatting purpose. Short texts are also uses in news titles, social posts, tweets, conversations, keywords, search queries. Short text understanding is an ambiguous process in opinions, deals, events and private messages. The short text is produce that contain social posts, conversations, keywords and news titles which are limited context and represent the insufficient information or meaning of the text. As short text has more than one meaning, they are difficult to understand as they are ambiguous and noisy. The term can be either single or multi-word. Short texts do not contain sufficient data. Some short texts have unique characteristics. So these short texts are difficult to handle. It required better understand the short text. Semantic analysis is essential to understand the short text accurately. Tasks such as segmentation, part-of-speech tagging, and concept labeling are used for semantic analysis. Conduct short text uses in real life data. The prototype system is built and used to understand the short text. These systems provide the semantic knowledge from knowledge base and collection of written words that are automatically harvest. Creating construction of co-occurrence network showing to better understand for short text.

Keywords: Short Text, Part of speech tagger, Semantics, text segmentation, Term Extraction.

---

1. Introduction

All people use short text for chatting and communication in real life. A number of people use a huge amount of short text for chatting and conversations. Certain words in English have several parts of speech (POS).The machine needs to understand the short words. Web search and microblogging applications per day handle more amount of short text. The work focus on determining semantic from texts. Data mining is the process of sorting through large data sets to find patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends and it has the ability to select data on the multiple groups in data mining. Data mining is to identify helpful information from raw data. Data Mining is the process to examine unexpressed, useful and understandable relationships in big amount of data. It extracts knowledge from lots of data.

Text mining structures of the data from natural language text. Text mining includes information and data that are retrieving for tagging, information extraction, and pattern recognition and meaningful analysis. Text mining is the process of extracting patterns from unstructured text documents. The most important component is to select the correct keywords for search. Even with various words, search results never always convey what is expect of the user. Improving the accuracy of search is important, and one of the best ways to do this is to integrate text mining.

© 2017, IJCERT All Rights Reserved

Semantics is the study of relationships between the words and construct the meaning of the short text. The interpretation a word, sign, and sentences.

The text a human readable sequence of the characters. Identified clearly, text can be located and classified into different categories like organization, location, persons. Abbreviation of the text is 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. Abbreviations, misspellings, nicknames are used in short text. For examples: M.Tech ([Master of Technology], CM [Chief Minister], imp [important],) etc.

Appropriate information is not contained in short text to support text mining approaches. Short text may be noisy, so it is difficult to handle Nicknames, abbreviations, misspellings that are used in short text. Short text can be used in many applications like web search, message, query, tweets and news titles. The short text is ambiguous and hard to understand because it has more than one meaning. There is need to better understand, the meaning of short text and avoid ambiguity.

2. Literature Review

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 Knowledge-based 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

© 2017, IJCERT All Rights Reserved

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.

3. Proposed Work

3.1 Understanding Short Text Using Co-Occurrence Network

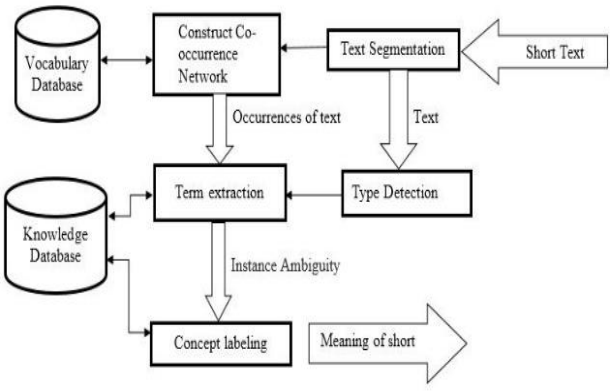

Fig.1 System architecture of exploring semantics of short text messages.

Fig.1 shows Text segmentation, construct co-occurrence network, Approximate term Extraction, type detection and concept labeling.
Define short text understanding to detect concepts mentioned in short text.

1. Ambiguous Segmentation:

"April in Pune paper" versus "vacation April in Pune" both term and its sub term contained into vocabulary, essentially to multiple possible segmentations for a short text. The valid segmentation maintains similarity of short text. According to the knowledge that Aril exam is related to the exam and April related to the month. Pune is related to the cities. It avoids the semantics thus the sometimes lead to incorrect segmentation. It ignores only stop word "in" given short text. In the text segmentation text is dividing into sub text. 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 the various meaning of word. The watch is instance co-occurs with concept buy and price, watch is verb co-occurs with concept movie. The co-occurrence network should be constructed term instance and typed term. The related or similar term is provide to the user search.

2. Noisy Short Text:

"Ichalkaranji city" versus "Ich" versus "Manchester" Identify the segmentation for short text in the vocabulary. Short text usually regulated format and error based nicknames, abbreviations and misspellings for example, "Ich" is an abbreviation of "Ichalkaranji" and also called as "Manchester" By approximate term extraction to extract possible term and its handles the misspelling of the short text. Using approximate term extraction identifies the meaning of short text. As per user requirement to find out meaning of the word and particular meaning of word provided by the user.

3. Ambiguous type:

"Blue songs" versus "blue bag" tag this term with lexical and semantic type. It describes the type of term to contribute short text understanding. For example "blue" in "blue songs" refers to singer and "blue shoes "refers as the color of the bag. Part of speech tagger resolve lexical type based on rules [9], [10], or lexical and probabilities learned from labeled [11]. The short text does not always observe syntax of the written language. Find the all possible term for related text. To determine the similarity between two strings, to provide the exact meaning of the short text.

4. Ambiguous instance:

"Watch two states" versus "Read two states"
An instance [e.g., "two states"] can associate to multiple concepts [e.g., movie, book]. It retrieves instance and

© 2017, IJCERT All Rights Reserved

concepts directly from knowledge. An instance can be referred to various concepts. Some methods try to discard instance ambiguity. Attributes, adjectives, verbs these terms can help to avoid ambiguity instance. To avoid instance ambiguity, all information know that attributes, adjectives, verbs. The short text is an input of concept labeling. In the type detection, obtain the collection typed of the term from the vocabulary. 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 is used to avoid ambiguity.

Text segmentation technique creates an exponential model. It automatically divides the text into a related segment. It ignores the requirement of semantic coherence and may lead to invalid segmentation it uses trie based model term equally divided into a number of segments. It checks either match with the segment or not. Longer term is more possible to mistakes and misspells. Knowledge-based taxonomy computes the similarity between two terms. It avoids the amount of information of the terms and also low accuracy. Randomized approximation algorithm to achieve better accuracy. The statistical approach depends upon the assumption of the short texts. It ignores semantics. It uses semantic hashing approach the meaning of text a compact into binary code. It captures co-relationship from the short text and also captures abstract feature from short text. It cheeks similarity of text with binary code. In every short text need compact binary code. By creating construction of co-occurrence network related words and its meaning can be find.

5. Conclusion

In this work, the main goal is to better understand short text used in real life. Using co-occurrence network, all possible terms can be evaluated, and it extracts the meaning of that term. Concept labeling is a process of eliminating inappropriate short text behind ambiguous instance and it avoids the ambiguity of the text. To give a label for the short text identifies correct meaning of the word. Using concept labeling better accuracy is achieved. It helps for better understanding of short text.

References

[1] Schutze and Y. Singer, "Part-of-speech tagging using a variable € memory Markov model," in Proc. 32nd Annu. Meeting. Assoc. Comput. Linguistics, 1994, pp. 181–187.

[2] 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.

[3] 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.

[4] 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.

[5] 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

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

[7] 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

[8] 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

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

[10] R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer, and L. Ramshaw, "Coping with ambiguity and unknown words

© 2017, IJCERT All Rights Reserved

through probabilistic models," Comput. Linguistics, vol. 19, no. 2, pp. 361–382, 1993.
[11] B. Merialdo, "Tagging English text with a probabilistic model," Comput. Linguistics, vol. 20, no. 2, pp. 155–171, 1994.
[12] D.P. Vardhini,. B.Ranjithkumar," An Efficient Way to Recommend Friends on Social Networks through Life-Style. "International Journal of Computer Engineering in Research Trends., vol.2, no.10, pp. 867-870, 2015.

Author 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.

© 2017, IJCERT All Rights Reserved