ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 9, September 2019

Discovering Symantics of Short Texts Using Knowledge Intensive Approach

Harish K. Barapatre[1] Miss.Vaishali H. Bhavsar[2]

Department of Computer Engineering, Yadavrao Tasgaonkar Institute of Engineering and Technology, Maharashtra, India[1 2]

ABSTRACT*:* The idea of this project is to implement a short text understanding, short texts is difficult to many applications. Short texts not follow the grammatical syntax of written language. Using the old natural language processing tools, identified by part-of speech of each word in short texts does give us precise results. Short texts do not contain sufficient information to identify its meaning. Short texts are more ambiguous and noisy, are generated in a conflict volume, which is more tedious to handle them. In this project, we develop a system for short text understanding which shows similar knowledge provided by well-known datasets and automatically detect from a huge standaford dictionary. Our approach is to less use of traditional methods for using such as text segmentation, part-of-speech tagging, and concept labelling. All these tasks focus on similar short text. We perform this method on real-time data. The results show that semantic knowledge for short text understanding.

KEYWORDS: Concept labelling, short text Understanding, text segmentation, text detection, event detection;

I. INTRODUCTION

Information explosion highlights the need for machines to better understand natural language texts. In this paper, we focus on short texts which refer to texts with limited context. Many applications, such as web search and micro blogging services etc., need to handle a large amount of short texts. Obviously, a better understanding of short texts will bring tremendous value. One of the most important tasks of text understanding is to discover hidden semantics from texts. Many efforts have been devoted to this field. For instance, named entity recognition locates named entities in a text and classifies them into predefined categories such as persons, organizations, locations, etc. Topic models attempt to recognize “latent topics”, which are represented as probabilistic distributions on words, from a text. Entity linking focuses on retrieving “explicit topics” expressed as probabilistic distributions on an entire knowledgebase. However, categories, “latent topics”, as well as “explicit topics “still have a semantic gap with human’s mental world. As stated in Psychologist Gregory Murphy’s highly acclaimed book, “concepts are the glue that holds our mental world together”. Therefore, we define short text understanding as to detect. A typical strategy for short text understanding which consists of three steps:

1.1) **Text segmentation** - divide a short text into a collection of terms contained in a vocabulary (e.g., “Book Magical hotel Goa” is segmented as book Magical Hotel Goa).
1.2) **Type detection** - determine the types of terms and recognize instances (e.g., “Magical” and “Goa” are recognized as instances, while “Book” is a verb and “hotel” concept).
1.3) **Concept labelling** - infer the concept of each instance (e.g., “Magical” a“Goa” refer to the concept theme park and state respectively). Overall, three concepts are detected from short text “Book Magical hotel Goa” using this strategy, namely theme park, hotel.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 9, September 2019

II. EXISTING SYSTEM

Understanding text to retrieve required content from a huge database is a critical task and more efforts have been devoted to this field. In current available system if a query is processed by user the entire query is considered to be keyword and processing of entire query will takes place. Therefore processing entire query leads to more time consumption and computation power because the machine learning does not understand which word is important or main key to search content. Short text identification is also difficult task in effective retrieval of data. Entity linking focuses on retrieving "explicit topics" expressed as probabilistic distributions on an entire knowledgebase. However, categories, "latent topics", as well as "explicit topics" still have a semantic gap with humans' mental world.The existing system employs the Bayesian Inference mechanism to conceptualize instances and short texts ,and eliminates instance ambiguity based on homogenous instances.. The system captures semantic relatedness between instances using probabilistic topic model and disambiguates instances based on related instances.

III. RELATED WORK

In this section, we discuss related work in three aspects: text segmentation, POS tagging, and semantic labeling.

TEXT SEGMENTATION:-

We can recognize all possible terms from a short text using the tried-based framework described. But the real question is how to obtain a coherent segmentation from the set of terms. We use two examples to illustrate our approach of text segmentation. Obviously, fapril in parislyricsg is a better segmentation of "april in paris lyrics" than faprilparislyricsg, since "lyrics" is more semantically related to songs than two months or cities. Similarly, fvacationaprilparisg is a better segmentation of "vacation april in paris", due to higher coherence among "vacation","april", and "paris" than that between "vacation" and "april in paris". We consider text segmentation as to divide a text into a sequence of terms. Existing approaches can be classified into two categories: statistical approaches and vocabulary based approaches. Statistical approaches, such as N-gram Model, calculate the frequencies of words co-occurring as neighbors in a training corpus. When the frequency exceeds a predefined threshold, the corresponding neighbouring words can be treated as a term. Vocabulary-based approaches extract terms in a streaming manner by checking for existence or frequency of a term in a predefined vocabulary. In particular, the Longest Cover method, which is widely-adopted for text segmentation due to its simplicity and real-time nature, searches for longest terms contained in a vocabulary while scanning the text.

POS TAGGING:-

Sometimes words can represent more than one part of speech at certain times and this makes the Part-of-speech tagging harder. Just having a list of words and their parts of speech is not sufficient. This condition is common because in natural languages a large Percentage of words are ambiguous. For example, in the case of the sentence "the sailor dogs the hatch", "dogs", which is usually thought as plural noun, can also be a verb. POS tagging determines lexical types (i.e., POS tags) of words in a text. Mainstream POS tagging algorithms fall into two categories: rule-based approaches and statistical approaches. Rule-based POS taggers attempt to assign POS tags to unknown or ambiguous words based on a large number of handcrafted or automatically learned linguistic rules. Statistical POS taggers avoid the cost of constructing tagging rules by building a statistical model automatically from a corpora and labeling untagged texts based on those learned statistical information.

SEMANTIC LABELLING:-

Semantic Labeling are use to disovers hidden semantics from a natural language text. Considering representation of semantics. Semantic labeling categorized as namely named entity recognition (NER), topic modelling and entity linking, NER detect and classifies named entities in a short text. Classifies them into predefined categories just like persons, organizations, locations, times, quantities and percentages etc. using linguistic grammar-based techniques as

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 9, September 2019

well as statistical models CRF and HMM. Latent topics are attempt to recognize by topic model which are represented as probabilistic distributions of words. Entity linking provides services to Knowledge bases as a retrieving "explicit topics" which is expressed as probabilistic distributions. High accuracy can be achieve by Semantic Labeling.

IV. PROPOSED SYSTEM

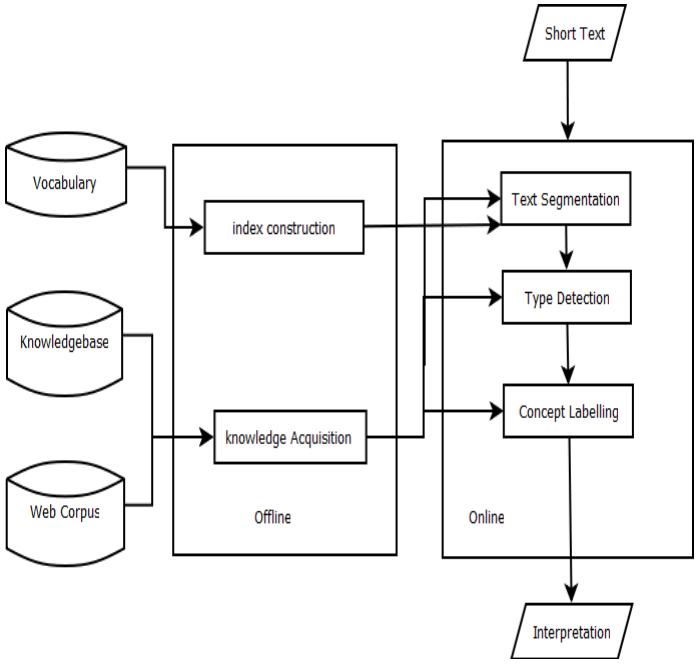

Fig 1. Proposed Architecture

Fig. 1 illustrates our framework for short text understanding. In the offline part, we construct index on the entire vocabulary and acquire knowledge from web corpus and existing knowledge bases. Then, we pre-calculate semantic coherence between terms which will be used for online short text understanding. In the online part, we perform text segmentation, type detection, and concept labeling, and generate semantically coherent interpretation for a given short text. Fig. 1 illustrates our framework for short text understanding. In the offline part, we construct index on the entire vocabulary and acquire knowledge from web corpus and existing knowledge bases. Then, we pre-calculate semantic coherence between terms which will be used for online short text understanding. In the online part, we perform text segmentation, type detection, and concept labeling, and generate a semantically coherent interpretation for a given short text.

Understanding short texts is crucial to many applications, but challenges abound. First, short texts do not always observe the syntax of a written language. As a result, traditional natural language processing tools, ranging from part-of- speech tagging to dependency parsing, cannot be easily applied. Second, short texts usually do not contain sufficient statistical signals to support many state-of-the-art approaches for text mining such as topic modelling. In this work, we build a prototype system for short text understanding which exploits semantic knowledge provided by a well-known knowledgebase and automatically harvested from a web corpus. Our knowledge-intensive approaches disrupt traditional methods for tasks such as text segmentation, part-of-speech tagging, and concept labelling, event labelling.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 9, September 2019

Proposed System Approach:-

Event Detection:-
There are two scenarios of event detection as follows-
a) Short-Term Event Detection:-
It extracts the most important events currently being posted in Twitter. In this scenario, we need to find out the synchronized words behavior. I.e. which of the words posted by the tweets present similar temporal patterns…

b) Long-Term Event Detection:-
It reviews the events that have occurred over a long time interval to synopsize what has mostly happened during that interval. To detect the most important events in the scenario, we need to find out similar words behaviour being invariant to time shifts and for this reason new similarity metrics are needed.

Advantages of Proposed System:-

1) The results show that semantic knowledge is indispensable for short text understanding and in this knowledge-intensive approach are both effective and efficient in discovering semantics of short texts.
2) Outperforms existing state-of-the-art approaches in the field of short text understanding…
3) User can search related words.

V. METHODOLOGY

OFFLINE PROCESSING:-

A prerequisite to short text understanding is the knowledge about semantic relatedness between terms. We describe how we construct the co- occurrence network and quantify semantic coherence. After that, we introduce the indexing strategy to allow for approximate term extraction on the vocabulary, as well as the approach to determine instance ambiguity.

ONLINE PROCESSING

There are basically three tasks in online processing of short texts, namely text segmentation, type detection, and concept labeling.

INDEXING OF VOCABULARY AND KNOWLEDGE ACQUISITION

Approximate term extraction aims to locate sub-strings in a text which are similar to terms
Contained in a predefined vocabulary. To quantify the similarity between two strings, many similarity functions have been proposed including token-based similarity functions (e.g., jaccard coefficient) and character-based similarity functions (e.g., edit distance)

TEXT SEGMENTATION

We can recognize all possible terms from a short text using the tried-based framework described. But the real question is how to obtain a coherent segmentation from the set of terms. We use two examples to illustrate our approach of text segmentation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 9, September 2019

TYPE DETECTION

Recall that we can obtain the collection of typed-terms for a term directly from the vocabulary. For example, term "watch" appears in instance-list, concept-list, as well as verb-list of our vocabulary, thus the possible typed-terms of "watch" are watch[c]; watch[e]; watch[v]. Analogously, the collections of possible typed-terms for "free" and "movie" are free[ad j]; free[v] and movie[c]; movie[e] respectively, as illustrated.

CONCEPT LABELLING

The most important task in concept labeling is instance disambiguation, which is the process of eliminating inappropriate semantics behind an ambiguous instance. We accomplish this task by re-ranking concept clusters of the target instance based on context information in a short text (i.e., remaining terms), so that the most appropriate concept clusters are ranked higher and the incorrect ones lower**.**

VI. APPLICATIONS

1) Use of social media increases rapidly in society also short text increased accordingly

2) The labeling of concept or text which is received on social site is crucial.

3) Short text must be easy to understand and real time nature, searches for longest terms contained in a vocabulary while scanning the text.

VII. RESULT ANALYSIS

It's a framework for short text understanding which can recognize best segmentations, conduct type detection, and eliminate instance ambiguity explicitly based on various types of context information. Therefore, we manually picked 11 ambiguousterms. Ambiguous in the sense for example "apple" it could be a fruit or a technically it could be a company. So general POS gives you the according to its understanding but to overcome this we proposed this model.
Foreach term derived from a short text, type detection determines the best typed-term from the set of possible typed-terms.
In the caseof "watch free movie", the best typed-terms for "watch", "free",and "movie" are watch as a verb,free as a adjective], and movie as a conceptual respectively.

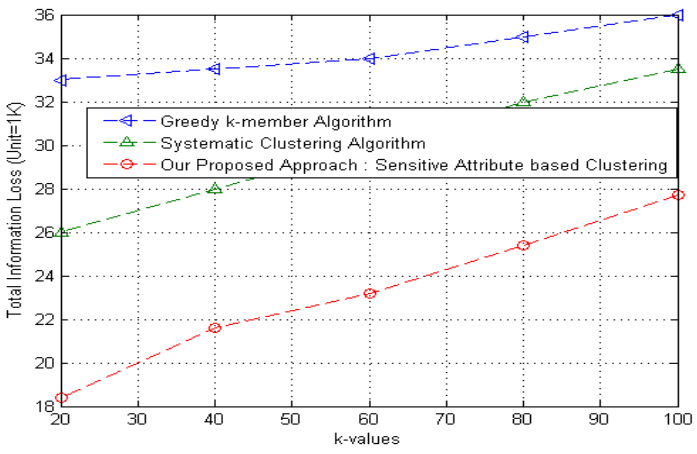

Fig 2.Result Analysis

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 9, September 2019

VIII. CONCLUSION

More specifically, we divide the task of short text understanding into three subtasks: text segmentation, type detection, and concept labeling.

We formulate text segmentation as a weighted Maximal Clique problem, and propose a randomized approximation algorithm to maintain accuracy and improve efficiency at the same time.

We introduce a Chain Model and a Pair wise Model which combine lexical and semantic features to conduct type detection.

We propose to better understanding and Hot event evolution of short text message in social media.

Feature:-

1) Short text understanding is to discover hidden semantics from texts.

2) Short text must be easy to understand and real-time nature, searches for longest terms contained in a vocabulary while scanning the text.

3) Semantic analysis is crucial to better understand short text.

REFERENCES

1)"Structural semantic relatedness: A knowledge-based method to named entity disambiguation," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics , ser. ACL '10, Stroudsburg, PA, USA, 2010, pp. 50–59.

[2] X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: A graph-based method," in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '11, New York, NY, USA, 2011, pp. 765–774

[3] W. Shen, J.Wang, P. Luo, and M.Wang, "Linden: Linking named entities with knowledge base via semantic knowledge," in Proceedings of the 21st International Conference on World Wide Web, ser. WWW '12, New York, NY, USA, 2012, pp. 449–458.

4) C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '12, New York, NY, USA, 2012, pp. 721–730.

5) D. Kim, H. Wang, and A. Oh, "Context-dependent conceptualization," in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI'13, 2013, pp. 2654–2661

6) G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 473–480.

7] D. M. de Oliveira, A. H. Laender, A. Veloso, and A. S. da Silva, "Fsner: A lightweight filter-stream approach to named entity recognition on twitter data," in Proceedings of the 22nd International Conference on

World Wide Web, ser. WWW '13 Companion, Republic and Canton ofGeneva, Switzerland, 2013, pp. 597–604.