International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.01, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Short text Understanding

Snehal Gurav[1], Urmila Lambe[2], Sonal Balanna[3], Puja Savarde[4], Aishwarya Suryavanshi[5], Prof. Mrs. Mulla Shagupta M[6]

[1,2,3,4,5,6]*CSE, Bharathi Vidyapeeth's College Of Engineering,Kolhapur, (India)*

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.[1]

Keywords: Concept labeling, semantic knowledge, Short text understanding , text segmentation, type detection

I. INTRODUCTION

Information technology is a need for machines to better understand language texts. In this project work, 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 [1].

II. REALATED WORK

We are going to represent literature overview of few papers that we have studied for choosing the topic as follows:

In this paper [1] models for many natural language tasks benefit from the flexibility to use overlapping, non-independent features. For example, the need for labeled data can be drastically reduced by taking advantage of domain knowledge in the form of word lists, part-of-speech tags, character grams, and capitalization patterns. While it is difficult to capture such inter-dependent features with a generative probabilistic model, conditionally-trained models, such as conditional maximum entropy models, handle them well. There has been significant work with such models for greedy sequence modelling in NLP. This paper describes Web Listing, a method that obtains seeds for the lexicons from the labelled data, and then uses the Web, HTML formatting regularities and a search engine service to significantly augment those lexicons.

In this paper [2] entity linking is a very important task for many applications such as web people search, question answering and knowledge base population. In this paper, they proposed LINDEN, a novel framework to link named entities in text with YAGO, knowledge base unifying Wikipedia and WordNet. By leveraging the rich semantic knowledge derived from the Wikipedia and the taxonomy of YAGO, LINDEN can obtain great results on the entity linking task. A large number of experiments were conducted over two public data sets, i.e., the CZ data set and the TAC-KBP2009 data set. Empirical results show that LINDEN significantly outperforms the state-of-the-art methods in terms of accuracy. Moreover, all features adopted by LINDEN are quite effective for the entity linking task.

In this paper [3] they proposed a statistical method that finds the maximum-probability segmentation of a given text. This method does not require training data because it estimates probabilities from the given text. Therefore, it can be applied to any text in any domain. An experiment showed that the method is more accurate than or at least as accurate as a state-of-the-art text segmentation system. Documents usually include various topics. Identifying and isolating topics by dividing documents, which is called text segmentation, is important for many natural language processing tasks, including information retrieval and summarization.

III. PROPOSED METHOD

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. Third, short texts are more ambiguous and noisy, and are generated in an enormous volume, which further increases the difficulty to handle them. We argue that semantic knowledge is required in order to better understand short texts. 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, in the sense that we focus on semantics in all these tasks. We conduct a comprehensive performance evaluation on real-life data. The results show that semantic knowledge is indispensable for short text understanding, and our knowledge-intensive approaches are both effective and efficient in discovering semantics of short texts.[1]

We used Windows 10 Operating System, Java used for project development which uses NetBeans IDE version and JDK 1.8.0 softwares. Also hardware requirements for our project implementation are system with 4GB RAM,I3 Processor,320 GB HardDisk.

3.1 BLOCK DIAGRAM

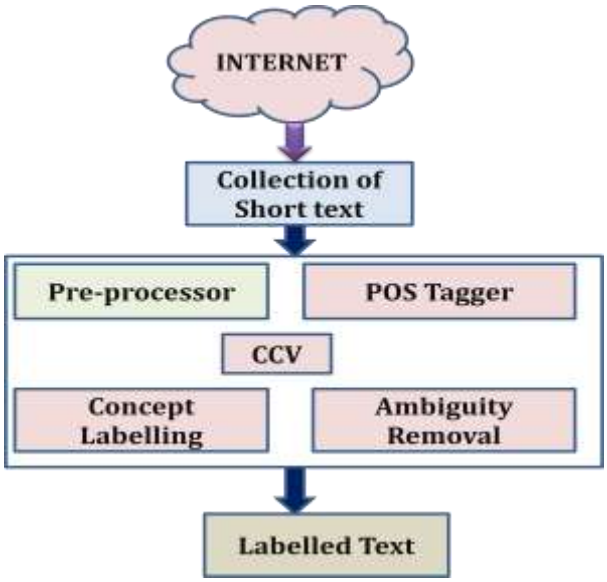
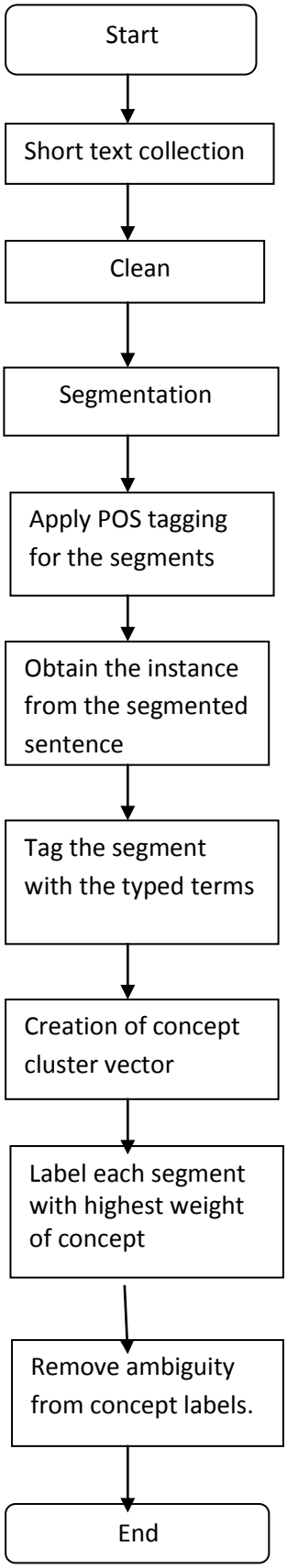


Fig. block diagram of Short text understanding

In these block diagram we collect the tweets through internet and put in the file. In pre processor the browse the tweet file, clean this particular file and remove the all stop words in that file. Using POS tagger find out the part of speech of each word in file. In CCV (concept cluster vector) obtaining the pos tags of each word the concept of each instance is obtained from the Microsoft Probase Engine and a concept vector is formed.In concept labelling is instance disambiguation, which is the process of eliminating inappropriate semantics behind an ambiguous instance..IN Ambiguity removal ambiguity between the labelled terms is reduced.Finally we get labelled text.

3.2 FLOW CHART



IV. IMPLEMENTATION

4.1 Modules & their Functionality

4.1.1 Module 1 –Browse tweet file

Collection of tweets in twitter site**.** Here we are browsing tweet file from folder and display that path into textbox.

4.1.2 Module 2 – Clean

This tweet file contain unnecessary data so remove that unnecessary data or stop words using clean button. After clicking clean button, it removes all stop words in tweet file

.

4.1.3 Module 3 – POS tag

After removing stop words get the appropriate sentence and apply that sentence POS tagging. Then that sentence each word categorized according to their part of speech.

4.1.4 Module 4 – Bigrams

To knowing each word POS then apply bigrams to get pair of short text. Each pair consist two words.

4.1.5 Module 5 - Co-occure

On obtaining the pos tags of each word the concept of each instance is obtained from the Microsoft Probase Engine and a concept vector is formed

Using the pos tags, and the similarity between the typed terms and the concept vector , the labelling of each keyword is done.

4.1.6 Module 6-Ambiguity Removal

In the testing and the implementation phase the ambiguity between the labelled terms is reduced.

V. CONCLUSION

In this work, we propose a generalized framework to understand short texts effectively and efficiently.

More specifically, we collect the tweets, Clean the tweets and remove the stop words from that tweets, after cleaning the tweets we apply POS tagging and tweets categorized according the their part of speech.[1]

Then short text is organized into pairs and each pairs consist two short text.

5.1 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.

4) micro blogging services and web search etc., are required to handle number of short text.

5.2 Limitation:-

1) Short texts refer to texts with limited context.

2) Short texts are more ambiguous and more noisy, and difficult to understand because it having more than one meaning, which increases the difficulty level to handle them.

REFERENCES

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

[2] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web enhanced lexicons," in Proc. 7th Conf. Natural Language Learn., 2003, pp. 188–191.

[3] W. Shen, J. Wang, P. Luo, and M. Wang, "Linden: Linking named entities with knowledge base via semantic knowledge," in Proc.21st Int. Conf. World Wide Web, 2012, pp. 449–458.

[4] 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.

[5] IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 3, MARCH 2017

[6] R. Mihalcea and A. Csomai, "Wikify! Linking documents to encyclopaedic knowledge," in Proc. 16th ACM Conf. Inf. Knowl. Manage. 2007, pp. 233–242.

[7] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti "Collective annotation of wikipedia entities in web text," in Proc. 15th ACM SIGKDD Int. Conf. Know. Discovery Data Mining, 2009, pp. 457–466.

[8] X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: A graph-based method," in Proc. 34th Int. ACM SIGIR Conf. Res.Develop. Inform. Retrieval, 2011, pp. 765–774.

[9]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.

[10] G. L. Murphy, The Big Book of Concepts. Cambridge, MA, USA: MIT press, 2004.