RECOGNIZING SHORT TEXTS BY GATHERING AND EXAMINING SEMANTIC KNOWLEDGE

[1]Likhitha H D, [2] Dr. K. Thippeswamy Ph.D.,
[1]Student, [2] Professor & Chairman
[1] Department of Computer Science and Engineering
[1]Visvesvaraya Technological University Department of PG Studies, Regional Office
Mysuru, Karnataka, India

Abstract: **Now a days short text understanding plays a major role in social media, chatting etc. Understanding short text is difficult to understand because it contains the hidden semantics. Short text are noisy and ambiguous and difficult to understand. As a result traditional ranging tools are used such as natural language processing, Data mining. These cannot be easily applied. Short text will be having more than one meaning. So to improve the efficiency we are using text segmentation, type detection and concept labeling. This improves the effectiveness and efficiency of the system. It provides the better understanding of the short text.**

***Index Terms*- Text segmentation, Short text understanding, Concept labeling, Part-of-speech tagging.**

I. INTRODUCTION

Short texts refer to texts with limited context. Many applications, like micro blogging services and web search etc., are required to handle number 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. Lots of efforts have been committed to this field. For instance, named entity recognition locates named entities in a text and classifies them into predefined topic models attempt to recognize "latent topics", which are represented as probabilistic distributions on words, from a text. Entity linking focuses on retrieving.

Short texts refer to texts with limited context. Many applications, like micro blogging services and web search etc., are required to handle number 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. Lots of efforts have been committed to this field. For instance, named entity recognition locates named entities in a text and classifies them into predefined topic models attempt to recognize "latent topics", which are represented as probabilistic distributions on words, from a text. Entity linking focuses on retrieving.

Short text must be easy to understand and easy to implement. Short text must be disambiguated. Therefore, we define short text understanding as to detect concepts mentioned in a short text

A. Text Segmentation

It divide a short text into a collection of word (i.e., terms and phrases) contained in a vocabulary (e.g., "Star Hotel" is segmented as {star Hotel});

B. Type Detection

It determines the types of terms and recognizes instances (e.g., both "Star" and "Hotel" are recognized as instances.

Although a number of methods have been identified for understanding short texts, there exist a number of challenges faced by these methods which include :

- ***Incorrect Segmentation***

A short text can have multiple possible Segmentations, While performing segmentation, it is highly essential that semantic coherence should be considered failed to which can result in incorrect segmentation. The most frequently used method;

Longest Cover method considers the longest possible segmentation as the best segmentation.

- ***Abbreviations And Noisy Short Texts***

Short texts often employ abbreviations and there are chances that more than one word can have same abbreviations. Short texts also employ nicknames of countries, famous personalities etc.

- ***Multiple type possibilities***

Same word can have multiple types. For example, 'sharp products' vs. 'sharp knives'. Sharp in 'sharp products' is an instance of the concept.

- ***Same instance- multiple concepts***

Consider the example; 'read jane eyre' vs. 'watch jane eyre' vs. 'age jane eyre'. Jane eyre is the instance of the concept, book in 'read jane eyre', instance of the concept, movie in'watch jane eyre' and instance of the concept, character in 'age jane eyre'. Since same instance can belong to multiple concepts, predicting the appropriate concept is a challenging task. Therefore, semantic coherence is highly essential in determining the most appropriate concept.

- ***Huge volume of data***

With the emergence of social media, the number of short texts generated is so high that their presence in the network cannot be ignored.

II. RELATED WORK

In this section, related works is being discussed that outlines some of the earlier used techniques.

1. ***TAGME: On-the-fly Annotation of Short Text Fragments***

The specialty of Tagme is that it may annotate texts which are short and poorly composed, such as snippets of search-engine results, tweets, news, etc.. This annotation is extremely informative, so any task that is currently addressed using the bag-of-words paradigm could benefit from using this annotation to draw upon (the millions of) Wikipedia pages and their inter-relations

.

2. ***Short Text Conceptualization Using a Probabilistic Knowledgebase.***

They introduce a method of conceptualizing short text using a probabilistic knowledgebase. We detect and map terms in short text to instances and attributes in the knowledgebase. Then we derive the most likely concepts using Bayesian inference. The conceptualization technique is applied to clustering Twitter messages. Results showed that our approach is highly effective compared to traditional bag-of-words based statistical methods.

3. Named Entity Recognition using an HMM-based Chunk Tagger

This paper proposes a Hidden Markov Model (HMM) and an HMM-based chunk tagger, from which a named entity (NE) recognition (NER) system is built to recognize and classify names, times and numerical quantities. Through the HMM, our system is able to apply and integrate four types of internal and external evidences: 1) simple deterministic internal feature of the words, such as capitalization and digitalization; 2) internal semantic feature of important triggers; 3) internal gazetteer feature; 4) external macro context feature. In this way, the NER problem can be resolved effectively.

4. Understanding short texts through semantic enrichment hashing

Clustering short texts by their meaning is a challenging task. The semantic hashing approach encodes the meaning of a text into a compact binary code. Thus, to tell if two texts have similar meanings, we only need to check if they have similar codes. The encoding is created by a deep neural network, which is trained on texts represented by word-count vectors. Unfortunately, for short texts such as search queries, such representations are insufficient to capture the underlying semantics. They propose a method to add more semantic signals to enrich short texts. Furthermore, we introduce a simplified deep learning network constructed by stacked auto-encoders to do semantic hashing. Experiments show that our method significantly improves the understanding of short texts, including text retrieval, classification and other general text-related tasks.

III. METHODOLOGY

In this section, we discuss in detail for short text understanding i.e., Text segmentation, text matching, and text recommendation.

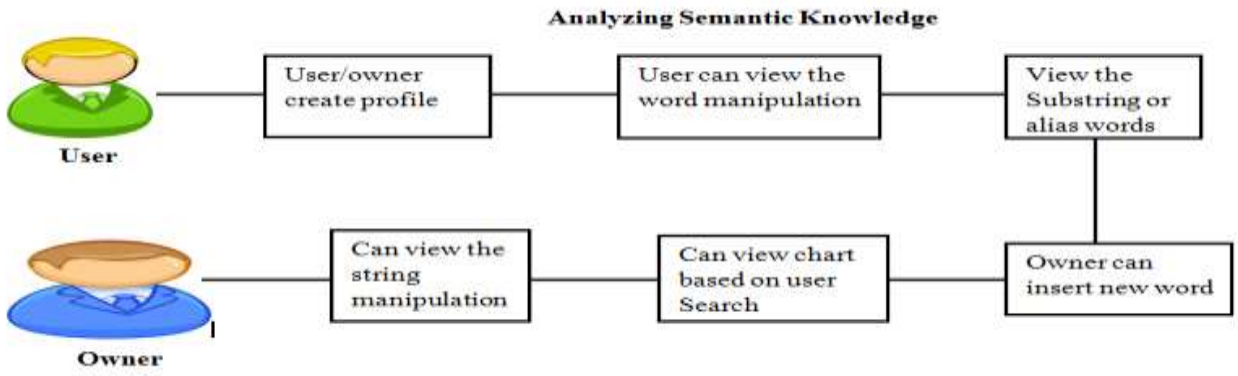

Figure: System architecture

3.1 Text Segmentation:

Text segmentation is the process of dividing the text or sentence into meaning full instances. In this project we are dividing the text based on the space between two charaters.

3.2 Text Matching

Text matching is the technique of finding the string or to match the pattern. Here, in this projectwe are using Rabin Karp algorithm for the text matching process.

Robin Karp algorithm uses hashing to find any one of pattern string in the text. A rolling hash allows an algorithm to calculate a hash value without having the rehash the entire string.

Algorithm

Compute hp (for pattern p)
Compute ht (for the first substring of t with m length)
For i = 1 to n –m
If hp =ht
Match t[i . . . .i + m] with p, if matched return1
Else
ht = (d ht − t[i + 1] .dm−1 + t[m + i + 1]) mod q
End;

3.3 Text Recommendation

For text recommendation we are using Microsoft association algorithm. Microsoft Association algorithm for use in creating data mining models that we can use for market basket analysis. Here the recommendation engine recommends items to customer based on items they have indicated an interest.

The Microsoft Association algorithm treasures a data to find items that appear together in a case. The algorithm then groups into Item sets any associated items that appear at a minimum, in the number of cases that are specified by MINIMUM_SUPPORT parameter.

3.3.1 Feature selection

The limits on the size of each item set or setting the Maximum and minimum support required to add an item set to the model includes:

- To filter out items and events that are too common and therefore uninteresting, decrease the value of MAXIMUM_SUPPORT to remove very frequent item set model.
- To filter out items and item sets that are rare, increase value of MINIMUM_SUPPORT.
- To filter out rules, increase the value of MINIMUM_PROBABILITY.

IV. EXPERIMENTAL OVERVIEW

This is the most generalized method for short text understanding. This experiment gives the most appropriate method for understanding the short text.

4.1 Effectiveness of Text segmentation

In order to overcome the noise in short text we are using text segmentation. Here we also adopt tier based method to allow for approximate term extraction with varying distances.

4.2 Effectiveness of text matching

Short text understanding is very ambiguous. So it is difficult for text matching because the text will be having some hidden semantics. So the robin karp algorithm for text matching. It improves the efficiency.

4.3 Effectiveness of understanding short text

Understanding short text is difficult to understand. Because it is more ambiguous and noisy. So it is an efficient way to understand the short text in this project

V. CONCLUSION

In this work, we propose a generalized framework to understand short texts effectively and efficiently. More specifically, we divide the task of short text understanding into three subtasks: text segmentation, type detection, and concept labeling. Here we are using SH1 algorithm to provide better security and we are using robin carp algorithm for text matching. The experimental results demonstrate that our proposed framework outperforms existing state-of-the-art approaches in the field of short text understanding. As a future work, we attempt to analyze and incorporate the impact of spatial-temporal features into our framework for short text understanding.

VI. ACKNOWLEDGMENT

I would like to thank Dr.Thippeswamy Ph.d., Professor & Chairman, Dept of studies in computer science &engineering, vtu regional office, mysuru and anonymous reviewers encouragement and constructive piece of advice that have prompted us for new round of rethinking of our research, additional experiments and clear presentation of technical content.

REFERENCES

[1] Wen Hua, Haixun Wang, Kai Zheng Zhong yuan Wang, and Xiao fang Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge," IEEE Transactions on Knowledge and Data Engineering, 2016.
[2] D. M. de Oliveira, A. H. Laender, A. Veloso, and A. S. daSilva, "Fsner: 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 of Geneva, Switzerland, pp. 597-604, 2013.
[3] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, ser. CONLL '03, Stroudsburg, PA, USA,
2003, pp. 188–191.
[4] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ser. ACL '02, Stroudsburg, PA, USA, 2002,pp. 473–480.
[5] R. Mihalcea and A. Csomai, "Wikify! linking documents to encyclopedic knowledge," in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ser. CIKM '07, NewYork, NY, USA, 2007, pp. 233–242.
[6] D. Milne and I. H. Witten, "Learning to link with wikipedia," in Proceedings of the 17th ACM conference on Information and knowledge management, ser. CIKM '08, New York, NY, USA, 2008, pp. 509–518.
[7] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of wikipedia entities in web text," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '09, New York, NY, USA, 2009, pp. 457–466.
[8] X. Han and J. Zhao, "Named entity disambiguation by leveraging wikipedia semantic knowledge," in Proceedings of the 18th ACM conference on Information and knowledge management, ser. CIKM '09, NewYork, NY, USA, 2009, pp. 215–224.
[9] "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.