International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 7 (2018) pp. 4855-4858

Interpretation of Short Text Using Semantic Knowledge

Arnav Gupta, Aditya Dave, Anita R

Department of Computer Science and Engineering SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

Abstract

Interpretation of short texts play a very vital role to many applications, but there are many challenges that come during their interpretation. First, short text is very frequently used in day-to-day life, which increases the ambiguity, and make the text-understanding complex making them difficult to handle. Second, the conventional and regular natural language tools cannot be implied on the text easily as short text do not always monitor the structure of the text. Third, discovery of hidden semantic structure in text body cannot be applicable on short text, as they do not hold sufficient statistical signals. Semantic search and information retrieval are the integral part of the search engine. We required a semantic knowledge for their understanding. In this work, we build a system, which uses the semantic analysis methodologies like part-of-speech tagging, text segmentation and concept labelling along with KNN algorithm. We applied KNN algorithm on a well-known knowledge base, which gives the semantic knowledge on the better interpretation on short text.

Keywords: KNN algorithm, text segmentation, concept labelling, semantic knowledge.

INTRODUCTION

Advancement in machines is requires for better understanding of natural language texts. In this paper, we focus mainly on short texts, which contain finite information. Searching web queries and conversation on social posts contain a large amount of short text, which makes them difficult to handle. The most important task is to understand the hidden meaning in the text. Many efforts are done in this field like named entity recognition (NER)[1][2] locates the named instances in the text and categories them as persons, locations etc. Topic modelling gives the 'latent topics' in the text and entity linking recognizes the 'explicit topics'. Because there is a sematic gap between the latent and explicit topics, we need a better understanding in short text. We use KNN algorithm to find the nearest neighbor of the given text and to provide the semantic knowledge from the well-known knowledge base harvested from collection of written words. For the better understanding of short text, we need a strategy divided into 3 steps.

Text segmentation- Short text is divided into the group of terms, which are present in the vocabulary. {eg – "comic wonderland hotel Rajvillas" is segmented as comic wonderland hotel Rajvillas.}

Type Detection- Types of terms are recognized and determined (eg- wonderland and Rajvillas are recognized as instances while comic is recognized as verb and hotel as concept.)

Concept Labeling- Concept of each term is infer. (eg- wonder and refers to amusement park and Rajvillas refers to a place.)

Although after applying, these techniques there are some challenges that needs to be handle. In the following we illustrate some challenges abound.

Challenge 1 (Matching Short Text)

- 'Udaipur city' vs 'Mewar' vs 'Venice of the east' vs 'The city of lakes'.

Text segmentation is performing on the given text through semantic coherence by extracting candidate terms. This can be perform by building a hash index on entire vocabulary. However, the short text is full of abbreviations, short-forms, acronyms, nicknames etc. For example, Udaipur city or mewar or Venice of the east or the city of lakes all correspond to the district Udaipur in India. This calls for vocabulary to have enough information about these abbreviations and acronyms as possible.

Challenge 2 (Incorporated Ambiguous Text)

- 'Toxicity in world lyrics' vs 'Toxicity of Delhi' vs 'read Toxicity in India'

The text segmentation of the given text should maintain semantic coherence. However, Traditional Longest cover methods seeks for longest term in vocabulary resulting in incoherent segmentation sometimes. 'Toxicity in world lyrics' is semantic incoherent segmentation by longest cover method.

After text segmentation, we tag the terms with lexical types (POS tag) and semantic type. Traditional POS taggers will mistakenly tag 'toxicity in world lyrics' as adjective rather than noun, which makes it a limitation.

Instance (toxicity) in the given text can belong to multiple concepts like (song, reports, books etc.). Some methods attempt to eliminate instance ambiguity based on related instances, but short text contain limited text, which makes these methods non-applicable on them. For example- 'toxicity in world' is a song, 'toxicity of Delhi' is a report, 'toxicity in India' is a book, which is easily recognize by humans but it is non-trivial for machines to disambiguate instances without knowledge.

Challenge 3 (Huge Volume)

Even in a single sentence, there are 3 to 4 short texts, which means that short texts generate in a much larger volume in any document or report. Google, which is the most popular web search engine of 2016, gets over 40,000 search queries every second, which means 3.5 billion searches per day and 1.2

trillion searches per year. Twitter more over detailed in 2017 that it pulled in more than 6,000 tweets every second, which corresponds to 350 million tweets per minute, 500 million tweets per day and around 200 billion tweets per year by 100 million people. However, short text can have many segmentation, tagging a term with multiple types and an instance can have many concepts. Therefore, the system should be able to handle them for better understanding and analysis of short text in real time. Thus, it is time-consuming to dispense with these ambiguities and accomplish the semantic understanding for short texts.

Contributions

In this work we, contend that semantic information is crucial for short text understanding, which in turn benefits numerous real-world applications that require handling an expansive sum of short texts. Agreeing to the over dialog, three sorts information are required to manage with the challenges in short text understanding a comprehensive lexicon, mappings[4][5] between instances and concepts, semantic coherence between terms. Based on the obtained information, we propose knowledge-intensive approaches to understand short texts both effectively and efficiently. Our contribution in this work is –

- We watch the prevalence of uncertainty in short texts and the restrictions of conventional approaches in handling them.
- We accomplish way better understanding of short text by collecting semantic information from web corpus [3] and existing information bases, and presenting knowledge-intensive approaches based on lexical-semantic analysis.
- We make the progress in the efficiency to facilitate the better understanding to short text.

RELATED WORK

In this section, we will discuss the related work, which divides into 3 categories - text segmentation, POS tagging and semantic labelling.

Text segmentation – Text segmentation is dividing the text into separate terms. Longest Cover Method, which is traditional method, searches for longest text in vocabulary, is the most widely adopted method for text segmentation. Semantic search cannot execute due to the simplistic nature of existing methods leading to incorrect text segmentation. We propose content semantic text segmentation to overcome this challenge.

POS Tagging – In POS tagging the terms are tag in the different parts of speech like nouns, verbs, adjective etc. As, we have discussed in challenge 2 (Toxicity in the world lyrics) tagging would be perform wrongly. In this work, we attempt to build a tagger, which considers both lexical features and semantic detection.

Semantic Labelling – Semantic Labelling [4][5]is a used to discover the hidden semantics in the natural language text. Existing work on the semantic labelling can be categorize as named entity recognition (NER) [1][2], topic modelling, and entity linking. NER locates named entity in text and classifies them as in different categories (person, places, organizations etc.). Topic modelling focuses on recognizing 'latent topics' and entity linking focuses on recognizing 'explicit topics' on the knowledge base. However, as short text does not observe the syntax of written language, named entity recognition (NER) tools cannot apply on them. In addition, short text does not have sufficient content for topic modelling and entity linking; these methods can give incorrect results. Therefore, we attempt to perform the type detection into our framework for the better analysis and understanding of short text.

PROBLEM STATEMENT

Pre-requisite-

Definition1 (Vocabulary)-

Vocabulary can be defined as a set of rules which are being followed by a language to form meaningful phrases or sentences. In other words, vocabulary is a collection of words and phrases, which constitute our day-to-day conversations. In the present day vocabulary, the sentence can be formed by a collection of words but the meaning of the sentence can be different as most of the time the collective meaning of the words in a sentence is not similar to what the sentence really means. This can be proved by a simple example as two negative words does not make the sentence more negative rather they mean a positive outcome.
e.g.- "not bad" =>it is a collection of two negative words but the overall meaning is positive. Similarly one positive and one negative word will make the product negative.

Definition2 (Term)-

A term is a single entity which is a part of a phrase or a sentence. A collection of terms is called a sentence and a collection of letters is called a term or a word. A term taken one at a time has a meaning of its own but when put together with more of its kind then they can have meaning of their own. The meaning may depend upon the collection of the terms.

Definition3 (Segmentation)-

Segmentation is the process of dividing the short text in different parts or segments. The semantic analysis of the short text depends mainly on the segmentation of the short text as dividing the sentence in different parts may present different meanings of the sentence.

Eg- "Lord of the rings movie" – this sentence can be divided in different formats and that will change the meaning of the statement and semantic knowledge will give obscure results and will cause a great deal of error.So to prevent this kind of situation we come with some kind of rules for our vocabulary which will decide whether our segmentation is valid or not .

Problem definition-

In this section we briefly introduce some of the major issues which have been resolved after our research work .

1. We observed that the existing system had many problems which occurred due to unsupervised

segmentation of the data or the short text. So for a given short text we devised a method to predict the most relevant segment of the data.

2. Next major problem was the correct type detection of the data or short text. Correct type detection [1][5] of a text very important otherwise the result will be meaningless. And the system will try to find the correct outcome in an incorrect category.
3. To deal with ambiguous data - It is a major problem as there is a lot of ambiguous data and it needs to be dealt with. The ambiguous data needs to be re ranked according to its context.

METHODOLOGY

In this we will discuss the details of our framework for better analysis and understanding of short text.

Constructing Co-occurrence model –

We create a co-occurrence relation [5] to connect all the semantic entities together. The semantic entities are related to each other by co-occurring relations which define their relationship among themselves and with other entities. It can be compared to an indirect graph with many interrelated nodes. The edge weight determines the strength of the relation between two nodes.

The following equation gives frequency of two typed terms-

$$fs(\bar{}x, \bar{}y) = ns \cdot e - dists(\bar{}x, \bar{}y)$$

Where ns = no of time the word appear in the corpus or the dataset.

Search

Extensive research studies have been done on structured queries as well as on text search over short keyword queries. In the view of difficulty of formulating the queries with precise structures over standard data, an IR-style querying, in particular, full text and keyword search is introduced. This approach has the merit of eliminating structures in the query.Maio.et.al presented an ontology based retrieval approach, which supports data organization and visualization and provides a friendly navigation model.

.

Top k Search-

This is also sub module of user module here admin added one URL. That URL search top k search. So here that URL search tops more by the user and first position. And algorithm also using top k search for the searching keyword on top position. In the top k search - KNN search algorithm is implemented. The KNN algorithm is used to find the k nearest neighbors of the given short text in the present knowledge base. Choosing the value of K is very critical in this case as if the value of is small then there will be a lot of noise in the output but if you take a much larger value of k then it would be computationally expensive.

Add Search Link-

Here in this module admin added link for the user so user searches on top position. That link show on top position that's why this depends on the algorithm. The search links are added in the dataset or the database. Whenever the user types a query the system searches the links present in the database and then find the most semantically relevant collection of links. This is achieved with the help of the search algorithm.

View Uploaded-

Here admin show all the uploaded files and view the files of users also this is only module modified by the admin not user this is restricted for the users. This module is only for the admin not for the user from here the admin can monitor the activities of the different users and check whether the system is working perfectly or not. The admin can receive feedback from any user if he finds any error in the system or any problem with the search results.

Most Searches-

Here the admin can check the live stats of the system such as the frequency of a single search or similar search results. The admin can get a graphical representation of the search results by the different user's and can improve the system data set on the basis of that information. The graphical representation can be on the basis of the number of searches per-person, similarity in search by the different users or the total no of times a query related to the same topic is searched.

PROPOSED SYSTEM

- We have observed that there are a lot of ambiguous results due to the limitations occurred while using traditional approaches. It is observed that the existing system had problem while classifying results with proper nouns or names. This can be seen whenever the system searches for proper nouns or names of person or a place it tends to get miss matched with its actual meaning rather than the user's perspective. In case of "Udaipur" it is often called "City of Lakes" which the system takes quiet literally and searches for a semantically incorrect and uncertain solution. This is resolved by increasing the dataset or the size of the corpus and using better algorithms.
- We have achieved better results from the given short text by implementing the process of harvesting data on a larger scale [3] and chronologically editing the data to fit the most recent events related to the text. This helps us to give better suited results and try new approaches to get the preferred outcome.
- We have achieved higher efficiency for predicting the relevant results for the given short text. We have improved the efficiency of our result by KNN algorithm as it is easy to implement and we have added more varied data in the database to get a more relevant result. The algorithm used is very efficient

and is less time consuming than the previous techniques used earlier.

- We are using basic nlp practices to scrap data from various web sources and linking them with the most significant search results [3][4][5]. This creates the corpus or the dataset more semantically reliable and reduces the time taken by the different search algorithms to get the correct result.

Finally, we have used KNN algorithm to find or to search the database or the corpus for any of the short text query and find the solution

KNN Algorithm

We proposed KNN (K-Nearest Neighbor) algorithm that can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. This KNN algorithm understands the user short text keyword and provides the result. KNN calculation is one of the least difficult classification calculation. Indeed, with such straightforwardness, it can allow exceedingly competitive comes about. KNN calculation can too be utilized for relapse issues. The as it were contrast from the examined strategy will be utilizing midpoints of closest neighbors or maybe than voting from closest neighbors.

CONCLUSION

In this work, we have successfully implemented KNN algorithm on the available corpus. The data has been successfully extracted from the various sources such as the Wikipedia pages of different subjects on which our system gives the most relevant output. We have taken data from Wikipedia pages – tokenize it – linked it to the most relevant entity in our co-occurrence network and all this has led to a very well categorized knowledge base. Thus when the user searches with a short text query our system uses the KNN search algorithm to find the most relevant result. The semantic analysis of the corpus makes the available data set more easily relatable to the searched short text. The short text understanding of our system has been increased by a bigger database and use of semantic knowledge. The present system has overcome all the previous problems, which were caused due to insufficient data in the corpus.

REFRENCES

[1] 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.

[2] G. Zhou and J. Su, “Named entity recognition using a 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.

[3] 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, New York, NY, USA, 2007, pp. 233–242.

[4] 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.

[5] 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.