International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Short Text Interpretation for User Classifications

Jeena Sara Viju[1], Sreehara B[2]
[1]*PG Scholar, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India*
[2]*PG Scholar, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India*

Abstract: ***The texts with limited context are referred as short text. Understanding and Interpreting short texts are important to many applications, but challenges appear. The problem when handling with short text is that firstly, 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. Secondly, short texts usually do not contain sufficient statistical signals to support many state-of-the-art approaches for text mining such as topic modelling. Thirdly, short texts are more ambiguous and noisy, and are therefore generated in a huge volume, which further increases the problem to handle them. Various methods for short text understanding and interpretation along with the limitations faced and, how the interpretation can be used in an effective way is discussed.***

Keywords: Short Text, Part of Speech Tagging, Natural Language Processing, Latent Dirichlet Allocation, N-Gram.

I. INTRODUCTION

The process of extracting information from large sets of data is termed as data mining. In the case of huge data sets, it is the computing process. Data mining is an essential process in which intelligent methods are used to extract data patterns.

Data mining being an interdisciplinary subfield of computer science is used in various situations. The main objective of the data mining process is to obtain information from large data sets and transform the data into an understandable format in order to use it efficiently in future. Apart from the basic analysis step, it also involves database and data management aspects, data pre-processing, model and the inference considerations, etc. In databases process, or KDD, data mining is the analysis step of knowledge discovery. Text mining, which is also referred as text data mining, is similar to text analytics. It is the process of obtaining high-quality information from text. High-quality information is obtained through the devising of patterns and trends through means such as statistical pattern learning. Text mining mainly consists of the process of framing the input text usually parsing, along with the addition of some derived linguistic features. It also consists of removal of others, and subsequent insertion into a database, deriving patterns within the structured data, and finally evaluation and interpretation of the output.

The term 'High quality' used in text mining usually refers to certain combination of relevance, novelty, and interestingness. Text categorization, text clustering, concept or entity extraction, production of granular taxonomies, etc are certain text mining tasks. The text with limited context is termed as short text. Each day billions of short texts are generated, which will take the form of search queries, ad keywords, tags, tweets, messenger conversations, social network posts, etc. Short texts have some certain properties and characteristics which make it different from normal documents. Understanding short text means deriving the concept hidden in the short text. Even though there are many challenges while handling short text, the short text understanding is important to many of the applications. The main three reasons which is considered as a problem for short text understanding are as follows. First, short text does not follow or obey the syntax of any written language.. Second, sufficient statistical signals in order to support many state-of-the-art approaches are not contained in short text. Third, since short texts are generated in an enormous volume, they are more ambiguous and noisy. This further causes problem in handling short text. To understand short text, Semantic knowledge is essential. Some of the challenges of understanding short text are mentioned below.

1) *Challenge 1 (Ambiguous Segmentation):* Consider the two short texts, “april in paris lyrics” versus “vacation april in paris”. A vocabulary contains both a term and its sub-terms which lead to multiple possible segmentations for a given short text. Semantic coherence should be maintained for a valid segmentation. For example, two possible segmentations can be derived from the short text “april in paris lyrics”, namely {april in paris lyrics} and {april paris lyrics}. The first segmentation is better one because the word “lyrics” is more semantically related to songs (“april in paris”) than months (“april”) or cities (“paris”).
2) *Challenge 2 (Noisy Short Text):* Consider the three short texts, “new york city” versus “nyc” versus “big apple”. It is essential to find out all the candidate terms for finding the semantic coherence in order to find the best segmentation for a particular text. This can be easily performed by building a hash index on the entire vocabulary. However, short texts are usually informal, full of abbreviations, etc. For example, in the above case “new york city” is usually abbreviated to “nyc” and known as “big apple”.

So it is important to find as much information possible about abbreviations and nicknames. Meanwhile, to handle misspellings occurred in short texts, approximate term extraction is required.

3) *Challenge 3 (Ambiguous Type):* Consider the two short texts "pink [singer] songs" versus "pink [adj] shoes". A word can belong to several types, and its best type in a short text depends on context semantics. For example, in the first short text, "pink" in "pink songs" refers to a famous singer and so it should be labelled as an instance, whereas in the second short text pink describes the color of the shoes and is therefore an adjective. Consider "pink songs" as an example. The probability of "pink" as an adjective and the probability of an adjective preceding a noun are relatively high, therefore traditional POS taggers will mistakenly label "pink" in "pink songs" as an adjective.
4) *Challenge 4 (Ambiguous Instance):* In the three cases, "read harry potter [book] versus "watch harry potter [movie] versus "age harry potter [character]". The instance (e.g., "harry potter") can belong to multiple concepts (e.g., book, movie, character, etc.). When the context varies, these similar instances might refer to different concepts.
5) *Challenge 5 (Enormous Volume):* Short texts are generated in a much larger volume when compared with normal document. Google being the most widely used search engine received over 3 billion search queries daily in 2014. Twitter reported in 2012 that it attracted more than 100 million users who posted 340 million tweets per day. Therefore, a feasible framework should be handled in real time for understanding short texts.

II. PROBLEM DEFINITION

Text segmentation, Type Detection and Concept Labelling which are the three steps for short text understanding sound quite simple, but challenges still abound. In order to face the main challenges which are being already discussed new approaches must be introduced to handle them. Given some short text, firstly text segmentation should be performed based on the semantic coherence. The best segmentation should be found out. Type Detection and Concept Label in order to understand and interpret the short text should be performed.

III.SHORT TEXT UNDERSTANDING APPROACHES

Various methods are used for understanding short text. Some of the methods are discussed below.

A. Transformation Based Error Driven Learning & Natural Language Processing

Linguistic information being encoded manually is challenged by automated corpus based learning. It is a method for providing a natural language processing system consisting of linguistic knowledge. Corpus based approaches have been successful and is used in many different areas of NLP, though these methods capture the linguistic information they are modelling indirectly in large opaque tables of statistics. All these make it difficult to analyze, understand and improve the ability of these approaches in order to model the underlying linguistic behaviour. In order to perform automated learning of linguistic knowledge, a simple rule-based approach is explained. The latest method for corpus based NLP called transformation based error driven learning is discussed. This algorithm was tested by applying to a number of language processing problems. Figure 1 describes the working of transformation based error driven learning. To start with firstly the unannotated text is passed through an initial state annotator. The ranges of complexity for the initial state annotator can vary from assigning random structure to assigning the output of a manually created sophisticated annotator. Various initial state annotators are used in parts of speech tagging. As indicated in the training corpus, it labels all the words with their most likely tag. Initial state annotations are explored for syntactic parsing. After passing the text through the initial-state annotator, it is then compared to the truth. For reference of truth, a manually annotated corpus is used. To better resemble the truth, an ordered list of transformations is learned that can be applied to the output of the initial-state annotator. There are two components to a transformation: a rewrite rule and a triggering environment.

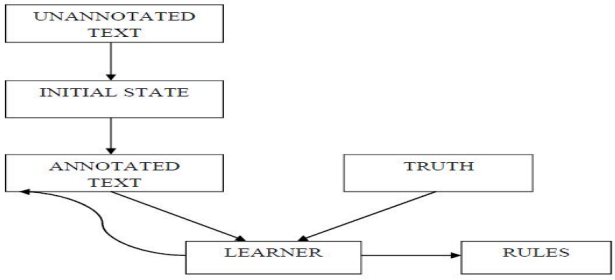

Fig-1: Transformation Based Error Driven Learning

A number of differences are present between transformation-based error driven learning and learning decision trees. One major difference between both is that during training a decision tree, the depth of the tree is increased every time, at the new depth, the average amount of training material available per node is halved (for a binary tree). While considering the case of transformation based learning, in order to find all transformations the entire training corpus is used. Transformations are being ordered, in a way that later transformations will be dependent on the result of applying earlier transformations. Therefore intermediate results help in classifying one object to be available in classifying other objects.

B. A Part-Of-Speech Tagger

A part-of-speech tagger which is based on the concept of hidden Markov model is discussed. It is a stochastic model and is used to design the randomly changing systems. In this particular case, it is assumed that the current state is responsible for the future states and not on the events that occurred previously. Most commonly, this prediction enables for reasoning and computation with the model that would otherwise be intractable. The Markov property is exhibited for this reason in the fields of predictive modelling and probabilistic forecasting. Hidden Markov Model (HMM) is a statistical Markov model and the system in which it is modelled is Markov process with neglected states. It is a generalization of a mixture model in which hidden variables that control the mixture component for each observation, are related through a Markov process rather than independent of each other. A type of Markov process consisting of either discrete state space or discrete index set is known as Markov chain. A Part-Of-Speech Tagger reads text and assign of parts of speech to each word, such as noun, verb, adjective, etc.

C. 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.

D. Computing Term Similarity By Large Probabilistic

While considering a set of documents or terms, the idea of distance between them means the likeliness of the meaning or semantic content which is termed as semantic similarity. These are mathematical tools which used to compute the strength of the semantic relationship between words, concepts or instances. Semantic relatedness is different from semantic similarity. Semantic relatedness means any relation between two terms, while semantic similarity only includes "is a" relations. In the case of text analyses, semantic relatedness can be estimated by a vector space model to correlate words and textual contexts from a suitable text corpus. The basic structure for finding semantic similarity between two terms is discussed. Given a pair of terms <t1, t2>, first determine the type of the terms, T (t1) and T (t2), and calculate the similarity between the two contexts.

$$SIM(t1, t2) = sim(T(t1), T(t2))$$

where sim (c1, c2) is a similarity function for contexts.

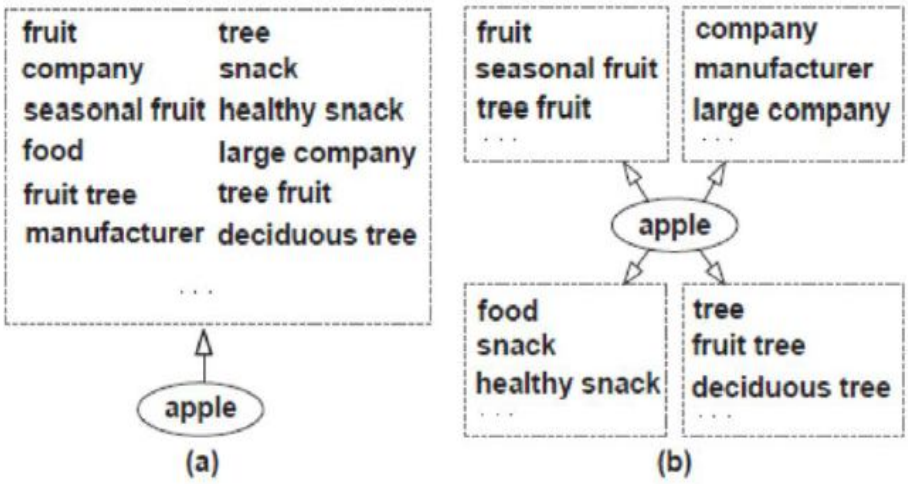

Fig-2: The concept context of “apple”

E. Graph Based N-Gram Language Identification

Language identification is an important task in natural language processing. In the case of N-gram, N can take many values, as per the values of N, the text will be divided. For example, if there are two texts and if the value of N is 3 i.e, trigram then the text will be divided into many segments in which each segment consists of three words each. After performing this to both the sentences, jaccard value will be found out. By taking the intersection and union count, it is possible.

IV. USER CLASSIFICATION

Different methods can be used for performing the three important steps in short text identification. The parts of speech tagging can be performed by using WordNet Tagger since Stanfford Tagger has some limitations and by experiment it is found that Stanfford tagger will tag some meaningless words as nouns, adverbs, etc...The Probase dataset can be used in order to conduct the experiment as it contains more than eighty five lakhs of IsA relations between different concept and instances which are obtained from the web corpus. Short text interpretation thus can lead a way to user classification. A community blog in which users can message each other in short text, then the short text will be interpretation and better understanding will be performed to group the users into clusters.

V. CONCLUSION

The various methods for interpreting short texts are studied. Short text interpretation can be better used to understand the meaning hidden in short text. A community blog can be developed in which users can message in short text and upon understanding the hidden meaning, the users can be further classified into clusters.

REFERENCES

[1] E. Brill, “Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging,” Comput. Linguistics, vol. 21, no. 4, pp. 543–565, 1995.
[2] E. Brill, “A simple rule-based part of speech tagger,” in Proc. Workshop Speech Natural Language, 1992, pp. 112–116.
[3] H. 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.
[4] 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.
[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; Knowl. Manage., 2013, pp. 1401–1410.
[6] 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.
[7] 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.
[8] G. Zhou and J. Su, “Named entity recognition using an hmmbased chunk tagger,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 473–480.
[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[10] 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.
[11] B. Merialdo, “Tagging english text with a probabilistic model,” Comput. Linguistics, vol. 20, no. 2, pp. 155–171, 1994.
[12] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, “Short text conceptualization using a probabilistic knowledgebase,” in Proc. 22nd Int. Joint Conf. Artif. Intell., 2011, pp. 2330– 2336.
[13] W. Wu, H. Li, H. Wang, and K. Q. Zhu, “Probase: A probabilistic taxonomy for text understanding,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2012, pp. 481–492.
[14] W. Wang, C. Xiao, X. Lin, and C. Zhang, “Efficient approximate entity extraction with edit distance constraints,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 759– 770.
[15] R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer, and L. Ramshaw, “Coping with ambiguity and unknown words through probabilistic models,” Comput. Linguistics, vol. 19, no. 2, pp. 361–382, 1993.