International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

A Survey on Short Text Conceptualization and Clustering

P. Dileep Kumar Reddy
Lecturer, Department of Computer Science & Engineering, JNTUACEA, Anantapuramu, Andhra Pradesh, India

ABSTRACT

The trend of social media and various online applications has rapidly increased over the past few years. These computer-mediated communications has resulted in the generation of large amount of short texts. A short text refers to the text with limited contextual information. Lots of interest lies in analyzing and conceptualizing short text for understanding user intents from search queries or mining social media messages. Consequently, the task of understanding short text is crucial to many online applications. But it is not ease to handle enormous volume of short texts, since they are relatively more ambiguous and noisy than normal data. The short texts do not follow the syntax of natural language. Thus, point out the necessity for an efficient text understanding technique. Short text understanding is an important but challenging task relevant for machine intelligence. The task can potentially benefit various online applications, such as search engines, automatic question-answering, online advertising and recommendation systems. In these kind of applications, the necessary basic step is to transform an input text into a machine-interpretable model namely to "understand" the short text. To achieve this goal, various approaches have been proposed to leverage external knowledge sources as a complement to the inadequate contextual information accompanying short texts. This survey reviews current progress in short text understanding with a focus on the vector based approaches, which aim to derive the vectorial encoding for a short text.

Keywords : Knowledge Mining; Short Text Understanding; Conceptualization; Semantic Computing.

I. INTRODUCTION

Short texts are different from long documents, they have unique characteristics which make very difficult to understand and handle. Everyday billions of short texts are generated in an enormous volume in the form of search queries, news titles, tags, chat bots, social media posts etc. Most of the generated short texts contain less than 3 words. These short texts, do not always examine the syntax of a written language. Hence, traditional NLP methods do not always apply to short texts. Many applications, including search engines, Question answering system, online advertising etc. rely on short texts. Short texts usually encounter data sparsity and ambiguity problems in representations for their lack of context. Understanding short texts processing, retrieval and classification become a very difficult task. An important challenge that would be faced while dealt with short texts is that they do not always follow the syntax of a written language. Also short texts generally do not have sufficient content to support statistical models. It may usually be informal and error-prone i.e., short texts are noisy and may have ambiguous types.

The fast development of the Internet, e-commerce and social networks brings about a large amount of user-generated short texts on the Internet, such as online question answer system, social media comments, tweets and micro-blogs. Such short texts as online reviews are usually subjective and semantic oriented. Huge explosion of information urge the need for machines that better understand the general language texts. The short text refers to those groups of words or phrases with limited context, that are generated via search queries, twitter messages, ad keywords, captions, document titles etc. So, a better understanding of a short text expose the hidden semantics from texts. Also lot of interests lies in analyzing and conceptualizing short text for understanding user intents from search

queries or mining social media messages for business insights. But understanding short text is a challenging task for machine intelligence meanwhile a very relevant concept on handling massive text data. Different from regular text data, the ambiguity of short text content brings challenge to traditional topic models because words are too few to learn and analyze from original corpus.

II. LITERATURE SURVEY

Xiang Wang et al. used Wikipedia concept to represent short document text. The mapping from document text to Wikipedia concepts is conducted using inverted index which is built from Wikipedia articles of concepts. The traditional classification method SVM is used to perform text categorization on the Wikipedia concept based document representation. The results obtained shows that the proposed method gives better performance than traditional SVM and MaxEnt method that is based on BOW model.

Pu Wang and Carlotta Domeniconi have made a strive to overcome the drawbacks of the BOW technique by using embedding background understanding derived from Wikipedia into a semantic kernel, which is then used to enhance the representation of documents. This method effectively achieves advanced classification accuracy with respect to the BOW technique, and to other recently developed methods. This methodology is able to keep multiword concepts unbroken; it captures the semantic closeness to synonyms, and performs word sense disambiguation for polysemy terms.

Abdullah Bawakid et al. present a system that performs automatic semantic based text categorization. The system reports on a simple analysis performed to evaluate the different implemented methods. The results obtained show that using WordNet based semantic approaches does yield to a better accuracy given that the right parameters (i.e. semantic similarity threshold) are selected.

Traditional statistics-based methods consistently fail to achieve satisfactory performance for short texts classification due to their sparsity of representations (Sriram et al., 2010). Based on external Wikipedia corpus, Phan et al. (2008) proposed a technique to find out hidden features using LDA and expand short texts. Chen et al. (2011) Proved that leveraging troubles at multiple granularity can version short texts greater exactly.

Traditional text retrieval methods, such as TFIDF, LSA, LDA and pLSA, have made significant achievements in most text-related applications. Recently, Salakhutdinov and Hinton propose a new information retrieval mechanism called semantic hashing. The model is stacked by RBMs and learns to map a document semantic to a compact binary code. Compared with traditional methods, such as TF-IDF and LSA, their semantic hashing model achieves comparable retrieval performance.

Development and application of a metric on semantic nets by R. Rada, H. Mili, E. Bichnell, and M. Blettner presents the method to find distance between two terms. Here the distance is nothing but average minimum path length between two subset of node. The distance can be used to find the conceptual distance between set of concepts. For finding the semantic similarity over semantic nets two tasks are performed. Firstly, the conceptual distance can be found. In second step distance judgment can be calculated. That determines whether semantic net s1 is better or worse than semantic net s2. People help to perform distance judgment over the semantic nets. The result shows that s1 is better than s2, if distance on s1 more like people than s2.

Allamanis and Sutton predicted an n-gram from a software program corpus with more than one billion tokens, but we regard the large scale as an organic smoothing approach. The method's effectiveness continuous to be problem to token distances in the corpus, in which clues at the back of the n-gram's tremendously short prefix (or "history") are elided from the model's context. Moreover, the large scale does not really solve the problem of considering tokens' semantic similarity. The approach for software language modeling is designed to consider an arbitrary number of levels of context, where context takes on a much deeper meaning than concatenated tokens in a prefix. In this work, the deep learning model encodes context in a continuous-valued state vector, encapsulating much richer semantics. Finally, Allamanis and Sutton conducted experiments where they collapsed the vocabulary by means of having the tokenizer update identifiers and literals with normal tokens, which was a novel way to measure the model's

performance on structural aspects of the code. However, we regard this approach as feature engineering. In this case, the token types in the corpus are engineered to solve the specific problem of modeling syntax. But the essence of deep learning, which underpins this work, is to design approaches that can automatically discover these feature spaces to—for instance—capture regularities at the syntactic, type, scope, and semantic levels.

Bengio et al. proposed a statistical model of natural language based on neural networks to learn distributed representations for words to allay the curse of dimensionality: One training sentence increases the probability of a combinatorial number of similar sentences. Sequences of words were modeled by agglutinating the word representations of consecutive words in the corpus into a single pattern to be presented to the network. Bengio also constructed model ensembles by combining a neural network language model with low-order n-grams and observed that mixing the neural network's posterior distribution with an interpolated trigram improved the performance. This work also measured the performance of the model after adding direct connections from nodes in the projection layer to output nodes, but the topology of this network does not constitute a deep architecture. This model represents history by presenting n-gram patterns to the network, whereas this work is based on a network which considers an arbitrary number of contextual levels to inform predictions.

Mikolov, who excised the projection layer in Bengio's architecture and added recurrent connections from the hidden layer back to the input layer to form a RNN. Representing context with recurrent connections rather than patterns of n-grams is what distinguishes Mikolov's recurrent architecture from Bengio's feed-forward architecture. Mikolov reported improvements using RNNs over feed-forward neural networks and implemented a toolkit for training, evaluating, and using RNN language models. The package implements several heuristics for controlling the computational complexity of training RNNs. Recently, Raychev et al. proposed a tool based in part on Mikolov's package, RNNs, and program analysis techniques for synthesizing API completions.

III. CONCLUSION

Many text mining applications like classification, conceptualization and clustering the task of understanding short text is considered as an underlying task or an online task. It is known that these applications need to handle millions of short texts at a time, signifies the importance of an efficient text conceptualization or text understanding task. A short text understanding can be more specifically divided into three steps, as text segmentation, type detection and concept labeling. Since the efficiency of short text understanding is extremely critical, each of these steps is required to be more precise.

IV. REFERENCES

[1]. M. Sahami and T. D. Heilman, "A web -based kernel function for measuring the similarity of short text snippets," in WWW, 2006, pp. 377-386.http://wwwconference.org/ www2006/programme/files/pdf/3069.pdf

[2]. W. tau Yih and C. Meek, "Improving similarity measures for short segments of text," in AAAI, 2007, pp. 1489-1494.https://pdfs.semanticscholar.org/33b5/ 8e4a7398ab3603c0918efab1e44a610835f6.pdf

[3]. D. Shen, R. Pan, J. -T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang, "Query enrichment for web-query classification," ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 320-352, 2006.https://pdfs.semanticscholar.org/5c62/ae64d72f4dfabdb5835a464f8aa3f49eb257.pdf

[4]. D. Kim, H. Wang, and A. H. Oh, "Context -dependent conceptualization," in IJCAI, 2013.https://arxiv.org/ pdf/1702.03342.pdf

[5]. B. Stein, "Principles of hash -based text retrieval," in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 527- 534.http://ceur-ws.org/ Vol-1536 /paper23.pdf

[6]. R. Salakhutdinov and G. E. Hinton, "Semantic hashing," Int. J. Approx. Reasoning, vol. 50, no. 7, pp. 969-978, 2009. https://esc.fnwi.uva.nl/thesis/centraal/files/f919407146.pdf

[7]. J. A. Anderson and J. Davis, An introduction to neural networks. MIT Press, 1995.

https://www.infor.uva. es/~teodoro/neuro-intro.pdf

[8]. Z. Harris, "Distributional structure," Word, vol. 10, no. 23, pp. 146-162.http://copec.eu/ congresses/ wccsete2016/ proc/works/11.pdf