International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 19
August 2018

Efficient Framework for Understanding Short Texts in Large-Scale Data Collection

[1] G.T.Vineela, [2] D.Gousiya Begum
[1]*M.Tech Student, Dept. of CSE, SKU Engineering College, A.P, India*
[2]*Assistant Professor, Dept. of CSE, SKU Engineering College, A.P, India*

Abstract: *Short texts are different from long documents, they have unique characteristics which make 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, chatbots, social media posts etc. Most of the generated short texts contain less than 5 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 retrieval, classification and processing become a very difficult task.*

In this paper, we propose a neural network based approach for understanding short text, where we perform texts as a vectors with Recurrent Neural Networks (RNN), and use a semantic network to determine our intention for clustering and understanding short texts. The task of short text understanding or conceptualization can be divided into three, as text segmentation, type detection, and concept labeling. In text segmentation, first the input text is pre-processed and removes all the stop words if any. Then it is divided into a sequence of terms. Type detection is incorporated into the framework for short text understanding and it help to conduct disambiguation based on various types of contextual information that present in the text. Finally, concept labeling is performed to discover the hidden semantics from a natural language text. The conceptualization can benefit from various online applications such as automatic question-answering, recommendation systems, online advertising, and search engines. All these applications requires an information extraction phase in which the prior step is to extract the concepts from the input text.

I. INTRODUCTION

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 natural 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.

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 usually 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. We focus on conceptualizing from texts or words. For example, given the word “India,” a person will form in his mind concepts such as country or region. Given two words, "India" and "Russia," the best ideas may move to Asian nation or biggest nation, and so on. Given yet another word, “Brazil,” the top concepts may change to BRICS or

emerging market, etc. Besides generalizing from instances to concepts, humans also form concepts from descriptions. For example, given words "body," "smell" and "color," the concept of wine comes into our mind. Certainly, concepts and instances may mix, for example, we conceptualize {"apple," "headquarter"} to company, but {"apple," "smell," "color"} to a natural product i.e. fruit.

II. RELATED WORK

M. Sahami and T. D. Heilman,

Determining the similarity of short text snippets, such as search queries, works poorly with traditional document similarity measures (e.g., cosine), since there are often few, if any, terms in common between two short text snippets. We address this problem by introducing a novel method for measuring the similarity between short text snippets (even those without any overlapping terms) by leveraging web search results to provide greater context for the short texts. In this paper, we define such a similarity kernel function, mathematically analyze some of its properties, and provide examples of its efficacy. We also show the use of this kernel function in a large-scale system for suggesting related queries to search engine users.

We have presented a new kernel function for measuring the semantic similarity between pairs of short text snippets. We have shown, both anecdotally and in a human-evaluated query suggestion system, that this kernel is an effective measure of similarity for short texts, and works well even when the short texts being considered have no common terms. Moreover, we have also provided a theoretical analysis of the kernel function that shows that it is well-suited for use with the web. There are several lines of future work that this kernel lays the foundation for. The first is improvement in the generation of query expansions with the goal of improving the match score for the kernel function. The second is the incorporation of this kernel into other kernel-based machine learning methods to determine its ability to provide improvement in tasks such as classification and clustering of text.

2) J. A. Anderson and J. Davis,

An Introduction to Neural Networks falls into a new ecological niche for texts. Based on notes that have been class-tested for more than a decade, it is aimed at cognitive science and neuroscience students who need to understand brain function in terms of computational modeling, and at engineers who want to go beyond formal algorithms to applications and computing strategies. It is the only current text to approach networks from a broad neuroscience and cognitive science perspective, with an emphasis on the biology and psychology behind the assumptions of the models, as well as on what the models might be used for. It describes the mathematical and computational tools needed and provides an account of the author's own ideas.

Students learn how to teach arithmetic to a neural network and get a short course on linear associative memory and adaptive maps. They are introduced to the author's brain-state-in-a-box (BSB) model and are provided with some of the neurobiological background necessary for a firm grasp of the general subject.

The field now known as neural networks has split in recent years into two major groups, mirrored in the texts that are currently available: the engineers who are primarily interested in practical applications of the new adaptive, parallel computing technology, and the cognitive scientists and neuroscientists who are interested in scientific applications. As the gap between these two groups widens, Anderson notes that the academics have tended to drift off into irrelevant, often excessively abstract research while the engineers have lost contact with the source of ideas in the field. Neuroscience, he points out, provides a rich and valuable source of ideas about data representation and setting up the data representation is the major part of neural network programming. Both cognitive science and neuroscience give insights into how this can be done effectively: cognitive science suggests what to compute and neuroscience suggests how to compute it.

3) B. Stein,

Hash-based similarity search reduces a continuous similarity relation to the binary concept "similar or not

similar": two feature vectors are considered as similar if they are mapped on the same hash key. From its runtime performance this principle is unequaled--while being unaffected by dimensionality concerns at the same time. Similarity hashing is applied with great success for near similarity search in large document collections, and it is considered as a key technology for near-duplicate detection and plagiarism analysis. This papers reveals the design principles behind hash-based search methods and presents them in a unified way. We introduce new stress statistics that are suited to analyze the performance of hash-based search methods, and we explain the rationale of their effectiveness. Based on these insights, we show how optimum hash functions for similarity search can be derived. We also present new results of a comparative study between different hash-based search methods.

4) W. Wu, H. Li, H. Wang, and K. Q. Zhu,

Knowledge is indispensable to understanding. The ongoing information explosion highlights the need to enable machines to better understand electronic text in human language. Much work has been devoted to creating universal ontology's or taxonomies for this purpose. However, none of the existing ontology's has the needed depth and breadth for “universal understanding”. In this paper, we present a universal, probabilistic taxonomy that is more comprehensive than any existing ones. It contains 2.7 million concepts harnessed automatically from a corpus of 1.68 billion web pages. Unlike traditional taxonomies that treat knowledge as black and white, it uses probabilities to model inconsistent, ambiguous and uncertain information it contains. We present details of how the taxonomy is constructed, its probabilistic modeling, and its potential applications in text understanding.

In this paper, we presented a framework which automatically inferences an open-domain, probabilistic taxonomy from the entire web. This taxonomy, to the best of our knowledge, is currently the largest and the most comprehensive in terms of the number of concepts included. Its probabilistic model allows the integration of both precise and ambiguous knowledge and even tolerates inconsistencies and errors which are common on the Web. More importantly, this model enables probabilistic inference between concepts and instances which will benefit a wide range of applications that require text understanding.

5) E. Gabrilovich and S. Markovitch,

Computing semantic relatedness of natural language texts requires access to vast amounts of common-sense and domain-specific world knowledge. We propose Explicit Semantic Analysis (ESA), a novel method that represents the meaning of texts in a high-dimensional space of concepts derived from Wikipedia. We use machine learning techniques to explicitly represent the meaning of any text as a weighted vector of Wikipedia-based concepts. Assessing the relatedness of texts in this space amounts to comparing the corresponding vectors using conventional metrics (e.g., cosine). Compared with the previous state of the art, using ESA results in substantial improvements in correlation of computed relatedness scores with human judgments: from r = 0.56 to 0.75 for individual words and from r = 0.60 to 0.72 for texts. Importantly, due to the use of natural concepts, the ESA model is easy to explain to human users.

We use Wikipedia and the ODP, the largest knowledge repositories of their kind, which contain hundreds of thousands of human-defined concepts and provide a cornucopia of information about each concept. Our approach is called Explicit Semantic Analysis, since it uses concepts explicitly defined and described by humans. Compared to LSA, which only uses statistical cooccurrence information, our methodology explicitly uses the knowledge collected and organized by humans. Compared to lexical resources such as WordNet, our methodology leverages knowledge bases that are orders of magnitude larger and more comprehensive. Empirical evaluation confirms that using ESA leads to substantial improvements in computing word and text relatedness. Compared with the previous state of the art, using ESA results in notable improvements in correlation of computed relatedness scores with human judgements: from r = 0.56 to 0.75 for individual words and from r = 0.60 to 0.72 for texts. Furthermore, due to the use of natural concepts, the ESA model is easy to explain to human users..

III. EXISTING SYSTEM

❖ Many approaches have been proposed to facilitate short text understanding by enriching the short text.
❖ More effectively, a short text can be enriched with explicit semantic information derived from external resources such as WordNet, Wikipedia, the Open Directory Project (ODP), etc.
❖ Salakhutdinov and Hinton proposed a semantic hashing model based on Restricted Boltzmann Machines (RBMs) for long documents, and the experiments showed that their model achieved comparable accuracy with the traditional methods, including Latent Semantic Analysis (LSA) and TF-IDF.

DISADVANTAGES OF EXISTING SYSTEM:

❖ Search-based approaches may work well for so-called head queries, but for tail or unpopular queries, it is very likely that some of the top search results are irrelevant, which means the enriched short text is likely to contain a lot of noise.
❖ On the other hand, methods based on external resources are constrained by the coverage of these resources. Take WordNet for example, WordNet does not contain information for proper nouns, which prevents it to understand entities such as “USA” or “IBM.”
❖ For ordinary words such as “cat”, WordNet contains detailed information about its various senses. However, much of the knowledge is of linguistic value, and is rarely evoked in daily usage. For example, the sense of “cat” as gossip or woman is rarely encountered.
❖ Unfortunately, WordNet does not weight senses based on their usage, and these rarely used senses often give rise to misinterpretation of short texts. In summary, without knowing the distribution of the senses, it is difficult to build an inferencing mechanism to choose appropriate senses for a word in a context.

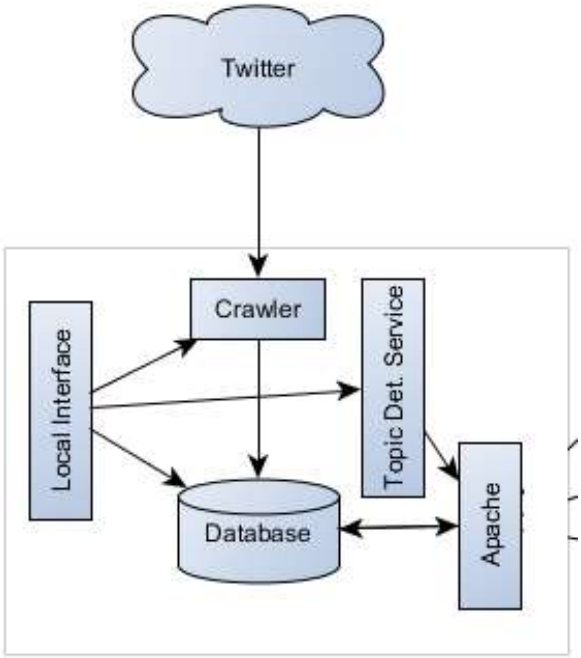
IV. PROPOSED SYSTEM

In this paper, we propose a novel approach for understanding short texts.
Our approach A semantic network based approach for enriching a short text;
We present a novel mechanism to semantically enrich short texts with both concepts and co-occurring terms, such external knowledges are inferred from a large scale probabilistic knowledge base using our proposed thorough methods.
For each auto encoder we design a specific and effective learning strategy to capture useful features from input data.
We provide a way to combine knowledge information and deep neural network for text analysis, so that it helps machines better understand short texts..

ADVANTAGES OF PROPOSED SYSTEM:

We carry out extensive experiments on tasks including information retrieval and classification for short texts.
We show significant improvements over existing approaches, which confirm that
concepts and co-occurring terms effectively enrich short texts, and enable better understanding of them;
Our auto-encoder based DNN model is able to capture the abstract features and complex correlations from the input text such that the learned compact binary codes can be used to represent the meaning of that text..

SYSTEM ARCHITECTURE



IMPLEMENTATION

Semantic Hashing:

In this project Semantic hashing is a new information retrieval method that hashes texts into compact binary codes using deep neural networks. It can be viewed as a strategy do transform texts from a high dimensional space into a low-dimension binary space, and meanwhile the semantic similarity between texts is preserved by the compact binary codes as much as possible. Therefore, retrieving semantically related texts is efficient: we simply return texts whose codes have small Hamming distances to that of query. Semantic hashing has two main advantages: First, with non-linear transformations in each layer of the deep neural network, the model has great expressive power in capturing the abstract and complex correlations between the words in a text, and hence the meaning of the text; Second, it is able to represent a text by a compact, binary code, which enables fast retrieval.

Probase:

Probase is a large-scale probabilistic semantic network that contains millions of concepts of worldly facts. These concepts are harvested using syntactic patterns (such as the Hearst patterns) from billions of webpages. For each concept, it also finds its instances and attributes. For example, company is a concept, and it is connected to instances such as apple and microsoft. Moreover, Probase scores the concepts and instances, as well as their relationships.

Backpropagation:

Backpropagation is a common method for training artificial neural networks. It is a robust approach to approximating real-valued, discrete-valued, and vector-valued target functions. The backward propagation of errors of backpropagation, is a common method of training artificial neural networks and used in conjunction with an optimization method such as gradient descent. The algorithm repeats a two phase cycle, propagation and weight update. When an input vector is presented to the network, it is propagated forward through the network, layer by layer, until it reaches the output layer. The output of the network is then compared to the desired output, using a loss function, and an error value is calculated for each of the neurons in the output layer. The error values are then propagated backwards, starting from the output, until each neuron has an associated error value which roughly represents its contribution to the original output.

Enriching Short Texts:

We propose a mechanism to semantically enrich short texts using Probase. Given a short text, we first identify the terms that Probase can recognize, then for each term we perform conceptualization to get its appropriate concepts, and further infer the co-occurring terms. We denote this two-stage enrichment mechanism as Concepts-and Co-occurring Terms .After enrichment, a short text is represented by a set of semantic features and is further denoted as a vector that can be fed to our DNN model to do semantic hashing. We focus on conceptualization and inferring co-occurring terms (do semantic enrichment) for noun phrases. Verbs and adjectives are also important as they can be useful for disambiguation and other tasks.

VI. CONCLUSION AND FUTURE WORK

In this paper, we advocate a unique approach for understanding brief texts. First, we introduce a mechanism to enrich quick texts with concepts and co-happening terms which might be extracted from a probabilistic semantic network, known as Probase. After that, each quick textual content is represented as a 3,000-dimensional semantic feature vector. We then design a more efficient deep learning model, which is stacked with the aid of three auto-encoders with specific and effective learning capabilities, to do semantic hashing on these semantic function vectors for quick texts. A two-stage semi-supervised schooling strategy is proposed to optimize the version such that it is able to seize the correlationships and summary capabilities from brief texts.

We carry out comprehensive experiments on brief text focused duties which includes statistics retrieval and class. The big upgrades on each tasks show that our enrichment mechanism ought to efficiently increase quick text representations and the proposed auto-encoder based deep getting to know model is able to encode complicated functions from input into the compact binary codes.

References

[1] M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 377–386.
[2] W. tau Yih and C. Meek, "Improving similarity measures for short segments of text," in Proc. 22nd Nat. Conf. Artif. Intell., 2007, pp. 1489–1494.
[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.
[4] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA, USA: MIT Press, 1998.
[5] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in Proc. 18th ACM Conf. Inf. Knowl. Manage., 2009, pp. 919–928.
[6] S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using wikipedia," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 787–788.
[7] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, pp. 1606–1611.
[8] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in Proc. 19th Int. Joint Conf. Artif. Intell., 2005, pp. 1048–1053.
[9] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in Proc. Int. Conf. Manage. Data, 2012, pp. 481–492.
[10] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledge base," in Proc. 22nd Int. Joint Conf. Artif. Intell., 2011, pp. 2330–2336.