ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

Towards Efficient Framework for Semantic Query Search Engine in Large-Scale Data Collection

G.Sneha[1], Kante Ramesh[2]

M.Tech Student, Dept. of CSE, GATES Engineering College, Affiliated to JNTUA, Andhra Pradesh, India[1]

Associate Professor, Dept. of CSE, GATES Engineering College, Affiliated to JNTUA, Andhra Pradesh, India[2]

ABSTRACT: 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 sematic 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.

KEYWORDS: Short text understanding, conceptualization, semantic labeling, text segmentation, Recurrent Neural Networks.

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

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

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 quick text snippets, which includes seek queries, works poorly with conventional document similarity measures (e.g., cosine), seeing that there are regularly few, if any, phrases in not unusual among short text snippets. We deal with this trouble by introducing a singular method for measuring the similarity between short text snippets (even those with none overlapping terms) with the aid of leveraging web search outcomes to offer greater context for the quick texts. In this paper, we define this type of similarity kernel function, mathematically examine some of its homes, and provide examples of its efficacy. We also display the use of this kernel function in a massive-scale system for suggesting related queries to search engine users.
We have provided a brand new kernel characteristic for measuring the semantic similarity between pairs of short textual content snippets. We have proven, both anecdotally and in a human-evaluated query concept device that this kernel is an effective degree of similarity for short texts, and works properly even if the quick texts being considered don't have any commonplace phrases. Moreover, we have additionally provided a theoretical evaluation of the kernel feature that suggests that it's miles well-appropriate for use with the web. There are numerous traces of destiny paintings that this kernel lays the inspiration for. The first is improvement inside the technology of query expansions with the intention of enhancing the match rating for the kernel function. The second is the incorporation of this kernel into other kernel-primarily based device gaining data of techniques to determine its ability to offer improvement in tasks including type and clustering of textual content.
2) J. A. Anderson and J. Davis,
An Introduction to Neural Networks falls into a new ecological niche for texts. Based on notes which have been elegance-examined for greater than a decade, it is aimed at cognitive science and neuroscience students who need to understand brain characteristic in phrases of computational modeling, and at engineers who need to head past formal algorithms to programs and computing techniques. It is the simplest cutting-edge textual content to technique networks from a extensive neuroscience and cognitive science perspective, with an emphasis at the biology and psychology in the back of the assumptions of the fashions, in addition to on what the models is probably used for. It describes the mathematical and computational gear wanted and offers an account of the author's very own ideas.
Students learn how to teach mathematics to a neural community and get a short direction on linear associative reminiscence and adaptive maps. They are introduced to the author's brain-state-in-a-box (BSB) model and are supplied with a number of the neurobiological history necessary for a firm hold close of the overall difficulty.
The discipline now called neural networks has cut up in latest years into two essential agencies, reflected within the texts which are presently available: the engineers who're basically inquisitive about realistic applications of the brand new adaptive, parallel computing technology, and the cognitive scientists and neuroscientists who are inquisitive about scientific programs. As the gap between those two agencies widens, Anderson notes that the academics have tended to waft off into irrelevant, regularly excessively abstract research while the engineers have misplaced contact with the supply of ideas in the area. Neuroscience, he points out, affords a wealthy and treasured supply of thoughts about records representation and putting in the facts representation is the important part of neural community programming. Both cognitive technology and neuroscience provide insights into how this could be carried out successfully: cognitive technology indicates what to compute and neuroscience indicates how to compute it.
3) B. Stein,
Hash- based similarity search reduces a non-stop similarity relation to the binary concept "similar or not similar": two characteristic vectors are taken into consideration as comparable if they're mapped on the same hash key. From its runtime overall performance this precept is unequaled--while being unaffected by dimensionality worries on the same time. Similarity hashing is applied with super success for close to similarity seek in big record collections, and it is taken into consideration as a key generation for close to-reproduction detection and plagiarism evaluation. This papers famous the design ideas in the back of hash-primarily based seek strategies and affords them in a unified way. We introduce new stress information that are suited to investigate the overall performance of hash-primarily based search

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

strategies, and we provide an explanation for the rationale of their effectiveness. Based on these insights, we show how surest hash functions for similarity search may be derived. We also gift new effects of a comparative examine among different hash-based totally seek methods.

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

Data is quintessential to know-how. The ongoing information explosion highlights the want to permit machines to higher recognize digital textual content in human language. Much work has been devoted to creating general ontology's or taxonomies for this reason. However, none of the existing ontology's has the needed intensity and breadth for "normal know-how". In this paper, we present a normal, probabilistic taxonomy that is greater complete than any current ones. It consists of 2.7 million principles harnessed routinely from a corpus of 1.68 billion net pages. Unlike conventional taxonomies that treat information as black and white, it uses probabilities to version inconsistent, ambiguous and unsure information it carries. We gift information of ways the taxonomy is constructed, its probabilistic modeling, and its ability packages in text know-how.

In this paper, we supplied a framework which mechanically inferences an open-domain, probabilistic taxonomy from the whole internet. This taxonomy, to the excellent of our understanding, is presently the most important and the maximum complete in phrases of the wide variety of concepts covered. Its probabilistic model permits the integration of each unique and ambiguous understanding or even tolerates inconsistencies and errors which can be commonplace on the Web. More importantly, this model permits probabilistic inference among ideas and times a good way to benefit a huge range of programs that require textual content know-how.

5) E. Gabrilovich and S. Markovitch,

Computing semantic relatedness of herbal language texts requires get right of entry to to great amounts of not unusual-sense and domain-unique world understanding. We endorse Explicit Semantic Analysis (ESA), a singular technique that represents the meaning of texts in a excessive-dimensional space of principles derived from Wikipedia. We use device studying strategies to explicitly represent the means of any text as a weighted vector of Wikipedia-based totally principles. Assessing the relatedness of texts on this space quantities to comparing the corresponding vectors the usage of conventional metrics (e.g., cosine). Compared with the preceding state of the art, the use of ESA effects in giant upgrades in correlation of computed relatedness scores with human judgments: from r = 0.56 to 0.Seventy five for individual words and from r = zero.60 to zero.72 for texts. Importantly, due to the use of natural concepts, the ESA model is simple to give an explanation for to human customers.

We use Wikipedia and the ODP, the most important understanding repositories of their kind, which incorporate hundreds of thousands of human-defined principles and provide a cornucopia of information about every idea. Our technique is called Explicit Semantic Analysis, because it uses ideas explicitly defined and described by means of human beings. Compared to LSA, which best makes use of statistical cooccurrence statistics, our technique explicitly uses the data collected and prepared by means of people. Compared to lexical sources including WordNet, our technique leverages expertise bases which are orders of importance large and more comprehensive. Empirical evaluation confirms that the usage of ESA ends in widespread improvements in computing word and text relatedness. Compared with the preceding state of the artwork, the use of ESA outcomes in great improvements in correlation of computed relatedness rankings with human judgements: from r = 0.56 to 0.75 for person words and from r = zero.60 to 0.Seventy two for texts. Furthermore, because of the usage of natural principles, the ESA model is simple to provide an explanation for to human customers.

III.EXISTING SYSTEM

Many applications have been proposed to facilitate short text understanding by enriching the short text.

Short texts introduce new challenges to many text-related tasks including information retrieval, classification, and clustering. Unlike long texts, two short texts that have similar meaning do not necessarily share many words. For example, the meanings of "upcoming apple products" and "new iPhone and iPad" are closely related, but they share on common words. The lack of sufficient statistical information leads to difficulties in effectively measuring similarity, and as a result, many existing text analytics algorithms do not apply to short texts directly.

A neural network model approach is introduced for understanding short texts. This model mainly consists two components: i) enriching short texts from semantic network; and ii) a deep neural network based method for revealing the semantics of a short text based on its enriched representation. A multiple inferencing mechanism called conceptualization is proposed to get the most appropriate sense for a term under different contexts. The concept space

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

that is employed is provided by Probase which contains millions of fine-grained, interconnected, probabilistic concepts. The concept information is more powerful in capturing the meaning of a short text because it explicitly expresses the semantic. However, conceptualization alone continues to be no longer enough for tasks such as comparing two short texts or classifying short texts. Consider the two short texts: “upcoming apple products” and “new iPhone and iPad”. After conceptualization, a set of concepts is obtained for each short text but there are still no common terms.
Semantic hashing is a new information retrieval method that converts texts into compact binary codes using deep neural networks (DNN). It could be viewed as a method to convert texts from a high dimensional vectors into a low-dimension binary vectors, and meantime the semantic relationship between texts is preserved by the compact binary codes as much as possible. Semantic hashing equip 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. A deep neural network (DNN) is constructed with 3-layer stacked auto-encoders to perform semantic hashing for short texts. Each auto-encoder has specific learning functions, and we implement a two-stage semi-supervised training strategy, including a hierarchical pre-training and an overall fine-tuning process, to train the model. This auto-encoder based deep neural network (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.

DISADVANTAGES OF EXISTING SYSTEM:

- Search-based strategies 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.
- It spends more time for classification and very less prediction rate in high dimensional space.
- It consumes more time to process the query.
- The model is inefficient in terms of classification accuracy.
- The disadvantage is that because of information overload problem, it cannot cluster the similar words.

IV. PROPOSED SYSTEM

In this paper, we propose a Deep Recurrent Neural Network (DRNN) model associated with stacked denoisingautoencoder (AE) to capture the semantics of the short text. Recurrent Neural Networks (RNNs) have shown great results in machine translation tasks. Unlike feed forward neural networks, RNNs are able to handle a variable-length sequence input by having a recurrent hidden state whose activation at each time is dependent on that of the previous time.For each autoencoder we design a specific and effective learning strategy to capture useful features from input data.The features generated from stacked denoisingautoecoders (AE) operation can be viewed as advanced features like n-grams. Since recurrent neural network (RNN) can process sequential input and learn the long-term dependencies, we take these features as the input of the recurrent neural network.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 includingsemantic similarity can be measured between multi-term 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 DRNN model is able to capture the better quality in measuring the similarity of short text segment. The model can work well under corrupted and unlabeled input data, also shows high-precision rate over large scale data. The embodied autoencoder with Recurrent Neural Network (RNN) promises they are best text categorizers and it searches queries easily.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

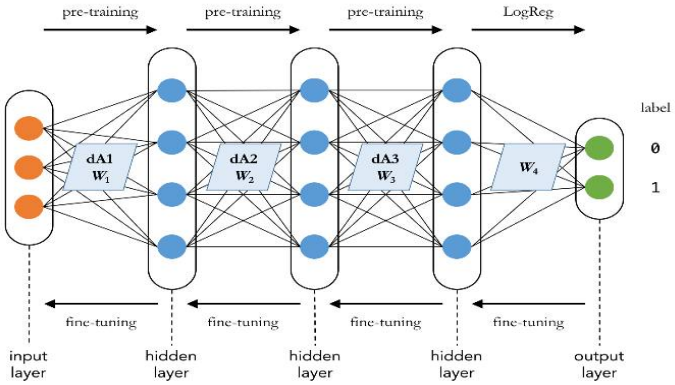
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

SYSTEM ARCHITECTURE



IMPLEMENTATION

Training Recurrent Networks:

Backpropagation Through Time (BPTT) is the common method is used to train the Recurrent Neural Networks (RNN). One of the common examples of a recurrent neural network is LSTM. BPTT algorithm updates the weights in the network based on the input time sequences. The algorithm promises high prediction rate and high-precision rate through time propagation. When compare to Backpropogation, BPTT is more stable and efficient in high dimensional space.

The ultimate goal of the Backpropagation design is to minimize the error rate of the neural network outputs.

The general algorithm is

1. First, present the input pattern and propagate it through the network to get the output.
2. Then compare the predicted output to the expected output and calculate the error.
3. Then calculate the derivatives of the error with respect to the network weights
4. Try to adjust the weights so that the error is minimum.

The Backpropagation Through Time is the modified version of Backpropagation algorithm which is applied to the sequence data like the time series. It is applied to the recurrent neural network. The recurrent neural network is shown one input each timestep and predicts the corresponding output. So, we will say that BPTT works by unrolling all the input timesteps. Each timestep has one input time step, one output time step and one replica of the neural network. Then the errors are calculated and expand for each timestep. The network is then rolled back to update the weights.

Probase:

Probase is a big-scale probabilistic semantic network that incorporates thousands and thousands of standards of worldly records. These principles are harvested the usage of syntactic patterns (together with the Hearst patterns) from billions of webpages. For every idea, it also reveals its times and attributes. For instance, organisation is an idea, and its miles connected to times including apple and microsoft. Moreover, Probase ratings the principles and times, in addition to their relationships.

Building Auto-encoders:

We have seen that the reconstruction criterion alone is unable to guarantee the extraction of usefulfeatures as it can lead to the obvious solution “simply copy the input” or similarly uninteresting ones that trivially maximizes mutual information. One strategy to avoid this phenomenon is to constrain the representation.

Here we propose and explore a very different strategy. Rather than constrain the representation,we change the reconstruction criterion for a both more challenging and more interesting objective: cleaning partially corrupted input, or in short denoising.Two underlying ideas are implicit in this approach:

• First it is expected that a higher level representation should be rather stable and robust under corruptions of the input.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

• Second, it is expected that performing the denoising task well requires extracting features that capture useful structure in the input distribution.

Enriching Short Texts:

We propose a mechanism to semantically enhance short texts using Probase. Given a short text, we first perceive the terms that Probase can apprehend, then for each term we perform conceptualization to get its appropriate concepts, and similarly infer the co-happening phrases. We denote this two-stage enrichment mechanism as Concepts-and Co-happening Terms .After enrichment, a short text is represented through a set of semantic functions and is further denoted as a vector that may be fed to our DRNN version to do semantic hashing. We perform onconceptualization and inferring co-happening phrases (do semantic enrichment) for noun phrases, Verbs and adjectives also are vital as they may be useful for disambiguation and other tasks.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a completely unique approach for understanding short texts. First, we introduce a mechanism to enrich short texts with concepts and co-going on phrases which is probably extracted from a probabilistic semantic community, called Probase. After that, every short text is represented as a 3,000- dimensional semantic feature vector. We then layout a greater than deep learning version that is stacked with the useful resource of three stacked denoising auto-encoders with unique and effective studying competencies, to do semantic hashing on those semantic function vectors for short texts. A two-level semi-supervised training approach is proposed to optimize the model such that it is able to capture the correlations and abstract features from short texts. When training is completed, the output is thresholded to be a 128-dimensional binary code that is seemed as a semantic hashing code for that enter textual content. We perform comprehensive experiments on short textual content centered obligations which includes statistics retrieval and sophistication. The big improvements on each task show that our enrichment mechanism ought to efficiently boom short textual content representations and the proposed auto-encoder based deep recurrent neural network getting to know model is capable of encode complex 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 data,” 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 data,” 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 data base,” in Proc. 22nd Int. Joint Conf. Artif. Intell., 2011, pp. 2330–2336.
[11] G.Pradeep Reddy "Combining Metric Cache and D-cache Mams for Maximizing Efficiency of Similarity Search" in IJERT, Vol 2-11 Nov 2013.