A Survey on Short Text Interpretation with Semantic Knowledge

Ms.Namrutha Mohan

Abstract— Nowadays the impact made by technology to the society is tremendous and this is a factor which increases the data accumulation day by day .Among the huge amount of data there is a very big consideration of short texts too. Short text classification is a very difficult task as they do not follow the semantics of written language and also they are sparse,ambiguous and multidimensional. So considerable means have to be evolved for the interpretation of short texts. Short text interpretation is an important phenomenon as it is the domain for applications such as Ads/search semantic match,Definition mining,Query recommendation ,Web table understanding ,Semantic search etc.As these applications rely on the understanding of short texts the semanticsw of the texts are to be identified and an interpretation of the same have to be made.This paper goes through various areas ,algorithms and also discussions made so far used for the short text interpretation.

***Index Terms*— Short texts, Knowledgebases, Semantic Knowledge, Clustering,Text segmentation**

I. Introduction

Short text volumes are exploding , more data accumulation is faced in the current era, as relation between technology and human is growing more intensive. So when our mind thinks beyond a machine it is necessary to make the machine understand the natural languages. This is a short text challenge as short text do not follow any semantics of the natural language and also as it is sparse and ambiguous the necessity to make interpretations or better understandings on short text are increasing.So we can come to a conclusion from the basics itself that semantic knowledge is an indispensible part for short text understanding.Even all the social networking sites uses the short texts for the commenting purposes and so.

So for the interpretation of short texts initially the text is to be segmented. Text segmentation is a precursor to text retrieval, automatic summarization,retrieval of the information(IR) and also it includes modelling of the language lexically and all the natural language processors(NLP). When considering the literal meaning text segmentation is not mere segmentation of the words but they the boundaries between words, phrases, or some other linguistic units that carries some meanings that can be some sentences or can be topics . Such segmented terms can be used understand human texts and can support computers to do artificial processes such as information retrievals or natural language processing [1] .This part goes to the clustering of the short texts.Different clustering means and algorithms are to be considered in this context. The relatedness and similarity between the texts is another category that comes in short text interpretation.The survey discusses on the relation of the semantic knowledge with short texts and also how to gather semantic knowledgebases, different knowledgebases,and also grouping the segmented text elements into clusters ,different clustering algorithms and also the systems already implemented and their effectiveness.

II. Semantic Knowledgebases

[1] WordNet [Stark et al. 1998]
WordNet is a lexical database for the English language.[2] The actual task include making up of the English words into some synonyms which is a word spelled set of abbreviations. This set is known as the synets ,they are able to provide all sorts of short definitions and also some examples of the usage of the work. So WordNet is thesaurus and also a dictionary. It is accessible for the outside word through internet that is the web browsers and the application or usage is in data mining, artificial intelligence ,text analysis etc.The license is under the BSD style and is available for free from the database of the WordNet site.

[2]KnowItAll[Bankoet al. 2007, Etzioniet al. 2011]
KnowItAll is also a knowledgebase from which all sorts of high quality data with the semantic relationship between the data is available. They are obtained from the natural language textys of the web scale.they include a very large quantity of extractions that is it sometimes exceeds about 5 billions over a web page.

[3]Probase [Wu et al. 2012]
Probase is a semantic network which makes the machine understand a bit of human communication.It is a huge collection of concepts,reations and instances. A probase may assist to collect the random data along with the relations and is easily available too.

[4]Google Knowledge Graph
Google knowledge graph is yet another knowledgebase that was announced by Google in the year of 2012[3].It was bought about by google inorder to support the search engines. From the internet live statistics we can understand the extend of usage of the search engines each day. So there is a need to

***Manuscript received March, 2017*.**

Namrutha Mohan, *M Tech Student,Network Computing,Toc H Institute of Science and Technology,APJ Abdul Kalam Technological University,Arakunnam,Kerala.*

bring about some enhancement methods.They are filled in with the semantic informations which are collected from extremely distinct sources.It also have details about each texts or topics and even the redirection links to various sites.These information can be used by the users to find their relevant resolutions for the queries and also can scan through different sites to gather the informations themselves.

III. Comparitive Surveyings And Clustering Techniques

Short Text Similarity with Word Embeddings[4]

The thought of finding the short text similarity only with semantics is the context that is exempted here that is no external sources for semantic matching is used here.The model uses word embeddings and also the vector representation of terms and thereby represents the terms in semantic space in which the semantic proximity of vectors can be interpreted as the semantic similarity.The approach moves from the word level to the text level.The meaning of longer texts is obtained by taking mean value of the short terms or the individual terms.Inorder to obtain the similarity between terms or short texts to be more specific, a supervised machine learning algorithm is used.A limitation of the approach is obtaining the meta features from the word vectors.

Word Relatedness Using Temporal Semantic Analysis[5]

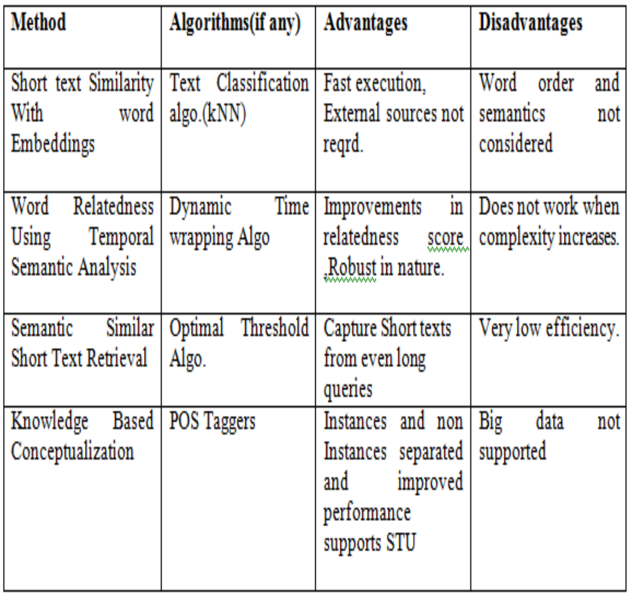
The former approach states that a very much considerable amount of relatedness information can be found in studying the patterns of word usage over the time that is the temporal aspects.The model captures the temporal temporal information.It focuses on two approaches one for rerpresenting the semantics of natural language and the other approach to compute the semantic relatedness between words.It associates a word with weighted vector of concepts and the3 vector of time series are manipulated considering the word.The algorithm is robust Dynamic Time Wrapping algorithm and assigns different weights to time periods.Time complexity is the factor which presides in the section for much consideration.

Semantic Similar Short Text Retrieval[6]

Semantic similarity is the consideration here and it makes out a framework for retrieving the top k semantic similar short texts.It accesses small size of candidates in the whole data collection.It consists of the preprocessing procedure which targets on the collection of the data and the second module targets of finding the similarity by the quick scan of the database and also the approach brings out the equation for the word orderings.Two representative similarity metrics are selectedknowledge based and also the corpus based.efficient Strategies are introduced to test few candidates in the querying process.

Knowledge Based Conceptualisation[7]

The main thought that bought about the query conceptualization is to map instances in a query to concepts defined in a certain ontology or knowledge base. Queries usually do not follow the regular syntax of the natural language or the written language and also no inference can be obtained statistically . However, the available context, i.e., the verbs related to the concepts alo the available attributes,shown instances etc do not provide any means to understand the concepts or instances. In the knowledge based approach mining of a variety of relations among terms from a large web corpus is done and map them to related concepts using a probabilistic knowledge base. Then, for a given query,the terms are conceptualised in the query using a random walk based iterative algorithm. A lexical knowledge base understands fine grained semantic signals and also determoine the types in the query.



Method	Algorithms(if any)	Advantages	Disadvantages
Short text Similarity With word Embeddings	Text Classification algo.(kNN)	Fast execution, External sources not reqrd.	Word order and semantics not considered
Word Relatedness Using Temporal Semantic Analysis	Dynamic Time wrapping Algo	Improvements in relatedness score .Robust in nature.	Does not work when complexity increases.
Semantic Similar Short Text Retrieval	Optimal Threshold Algo.	Capture Short texts from even long queries	Very low efficiency.
Knowledge Based Conceptualization	POS Taggers	Instances and non Instances separated and improved performance supports STU	Big data not supported

A discussion on two different subspace clustering techniques and the importance of using clustering in the particular context of understanding of short texts is given below.

PROCLUS algorithm

A Cluster here is formed from the subspace data points and a subset dimensions where the subspace points are closely related to the subspace dimensions. The PROCLUS algorithm uses a top-down approach which creates clusters that are partitions of the data sets, where each data point is assigned to only one cluster which is highly suitable for customer segmentation and trend analysis where a partition of points is required. The algorithm also finds the outliers and uses the sampling technique to select a sample dataset and sample medoid set.The algorithm uses the K-medoids method to find the centers of the cluster and his is used to find out the original cluster. The three phases that are used in this algorithm is initialization ,then the iteration and finally the cluster refinement.

CLIQUE algorithm

CLIQUE is the short term for CLustering In QUEst developed by R.Aggrawal[9] which uses a top-down approach based subspace clustering algorithm that starts by placing each object in its own cluster and then merges the small atomic clusters into bigger clusters and the limit is until all objects gets placed.In this algorithm there is no consideration on how the input data is processed but the output gives identical results there is no consideration of the canonical distribution of the data given as the input.

Clustering is an unsupervised learning method which groups up or clusters the unlabelled data.There is no particular criterions in which an independent final cluster is obtained in the case of clustering the user determines what is the criterion that is to be followed which will satisfy the needs of the user.It should pass the data reduction,should determine the natural data types and should also be able to find out data objects.

IV. Conclusion

Considering the huge accumulation of short texts nowadays and also the applications that are immensely relaying on the understanding of short texts include Ads/search, semantic match,Definition mining,Query recommendation ,Web table understanding ,Semantic search etc so there is a necessity to better make an interpretations of the short texts.Semantics is closely related to all the texts and they are the primary considerations to text understanding.Clustering algorithms can be used to make the performance better as they are categorized in unsupervised learning.For the actual understanding of a short text it has to be retrieved and then segmented to make out all the individual terms and further the type of the terms can be tackled out.Segmentation can be done with the assistance of the clustering algorithms then the concept of the text can be interpreted to make a better understanding for this to happen a semantic knowledge support is required.Concept labelling can also support to wipe out the ambiguity problems of the short text.

As short text understanding have hold on to the wings now,the target is to make overcome all the challenges bought up by their accumulation and also to make a proper understanding.

Acknowledgment

I gratefully thank the entire faculty of the Department of Information Technology, ToC H Institute of Science & Technology for their valuable support and encouragement in conducting this work .

REFERENCES

[1] Haoda Huang, Benyu Zhang,"Text Segmentation", Encyclopedia of Database Systems,Springer 2009.

[2] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244.

[3] Singhal, Amit (May 16, 2012). "Introducing the Knowledge Graph: Things, Not Strings". Official Blog (of Google). Retrieved May 18, 2012.

[4] Tom Kenter, Maarten de Rijke,"Short Text Similarity with Word Embeddings", University of Amsterdam.

[5] Kira Radinsky_, Eugene Agichteiny, Evgeniy Gabrilovichz, Shaul Markovitch"A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis",WWW,2011.

[6] Yanhui Gu, Zhenglu Yang,_ Junsheng Zhou, Weiguang Qu, JinmaoWei, Xingtian Shi," A Fast Approach for Semantic Similar Short Texts Retrieval", Proceedings of the AAAI Conference on Artificial Intelligence,2007.

[7] ZhongyuanWang, Kejun Zhao, HaixunWang , Xiaofeng Meng , Ji-Rong WenQuery "Understanding through Knowledge-Based Conceptualization",24th International Joint Conference on Artificial Intelligence,2015.

[8] C.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, J.S. Park, Fast algorithms for Projected clustering, in proceedings of the ACM SIGMOD international conference on management of data, ACM Press, 1999.

[9] R. Agrawal, J. Gehrke, D.Gunopulos, P. Raghavan, Automatic subspace clustering of high dimensional data for Data mining applications, in proceedings of the ACM SIGMOD conference of management of Data, Montreal, Canada, 1998.

Namrutha Mohan received B.Tech degree in Computer Science and Enginnering from KRGCEW,Kerala University,Kerala.Currently pursuing her Masters Degree in Network Computing from Toc H,APJ Abdul Kalam Technological University,Kerala,India.