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 Segmentation for Improved Query Processing

Dr. C.S. Kanimozhi Selvi[1], Dr. S.V. Kogilavani[2], Dr. S. Malliga[3] D. Jayaprakash[4]
[1, 2, 3, 4,] *Department of Computer Science and Engineering, Kongu Engineering College*

Abstract*: Understanding short texts is crucial to many applications, but challenges are many. Short texts do not always detect 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. Semantic knowledge is required in order to better understand short texts. In this work, a model for short text understanding which exploits semantic knowledge is used. The results show that semantic knowledge is indispensable for short text understanding, and our knowledge-intensive approaches are both effective and efficient in discovering semantics of short texts. A query data set containing short text is pre-processed by removing punctuation marks, numbers, stop words removal and stemming. Then POS tagging is performed and Co- Occurrence network is formed to obtain the semantic similarity. Term Similarities are calculated and finally clustering is done using K-Means, hierarchical, PAM, CLARA, DIANA and AGNES methods and the result reveals that hierarchical clustering gives optimal clusters.

Keywords: ***Short Text, Clustering, Co-occurrence, Similarity, Semantic***

I. INTRODUCTION

Text mining, which is sometimes referred to text analytics is to make unstructured data usable. Text mining, also referred to as text data mining, is the process of deriving high-quality information from text. High-quality information is typically derived through the generation of patterns and trends through means such as statistical pattern learning. Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. Mining unstructured data with natural language processing and machine learning techniques can be challenging, however, because natural language text is often inconsistent[5]. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double entendres and sarcasm[10]. Information explosion highlights the need for machines to better understanding natural language texts[11]. Many applications such as web search and microblogging sites etc., need to handle a large amount of short texts. One of the most important tasks of short text understanding is to discover hidden semantics from texts. For instance named entity recognition locates named entities in a text and classifies them into predefined categories such as persons[19].

II. LITERATURE REVIEW

The process of classifying words into their parts of speech and labeling them accordingly is known as part-of-speech tagging, Parts of speech are also known as word classes or lexical categories[6][7][8]. The collection of tags used for a particular task is known as a tag set. Some of the pos tags are Noun, adverb, adjective, conjunction etc[1][2][3]. POS tagging determines lexical types of words in a text. POS tags are mainly used for clustering the words in short text. There are two types of POS tagging algorithms: rule-based approaches and statistical approaches[4][9][12]. A Hidden Markov Model (HMM) and an HMM-based chunk tagger is proposed in [14], from which a named entity (NE) recognition (NER) system is built to recognize and classify names, times and numerical quantities[13]. Through the HMM, the system is able to apply and integrate four types of internal and external evidences: simple deterministic internal feature of the words, such as capitalization and digitalization, internal semantic feature of important triggers, internal gazetteer feature, external macro context feature. Latent Dirichlet allocation (LDA) descibe a generative probabilistic model for collections of discrete data such as text corpora[15]. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. The topic model introduced by [16] is a generative model for documents that extends Latent Dirichlet Allocation(LDA) to include authorship information. Each author is associated with a multinomial distribution over topics and each topic is associated with a multinomial distribution over words. A document with multiple authors is modeled as a distribution over topics that is a mixture of the distributions associated with the authors. Exact inference is intractable for these datasets and we use Gibbs sampling to estimate the topic and author distributions and compare the performance with two other generative models for documents, which are special cases of the author-topic model

In [17][18], the use of Wikipedia as a resource for automatic keyword extraction and word sense disambiguation is introduced. It shows how this online encyclopaedia can be used to achieve state-of-the-art results on both these tasks and shows how the two methods can be combined into a system able to automatically enrich a text with links to encyclopedic knowledge. Given an input document, the system identifies the important concepts in the text and automatically links these concepts to the corresponding Wikipedia pages. A graph-based collective EL method is discussed in [20], which can model and exploit the global interdependence between different EL decisions. Specifically, first propose a graph-based representation, called Referent Graph, which can model the global interdependence between different EL decisions. Then propose a collective inference algorithm, which can jointly infer the referent entities of all name mentions by exploiting the interdependence captured in Referent Graph. The key benefit of our method comes from: The global interdependence model of EL decisions, the purely collective nature of the inference algorithm, in which evidence for related EL decisions can be reinforced into high-probability decisions. A novel framework linden is proposed in [21] to link named entities in text with a knowledge base unifying Wikipedia and WordNet, by leveraging the rich semantic knowledge embedded in the Wikipedia and the taxonomy of the knowledge base. They extensively evaluate the performance of our proposed linden over two public data sets and empirical results show that linden significantly outperforms the state-of-the-art methods in terms of accuracy.

III. PROPOSED SYSTEM

Our proposed System consists of four modules. In architecture diagram, the short text is given as input and it gives clusters as output. First the system pre-processes the short text and also finds the parts of speech tagging for terms in queries. And the co-occurrence network is formed, here nodes are terms in the queries and the edge weights calculated using the formula. After that, the term similarities between the words are measured.. Based on the similarity values, the terms are clustered using various clustering methods. After clustering all the relevant terms are grouped under one cluster. The steps to build the model are listed below and shown in Figure 1.

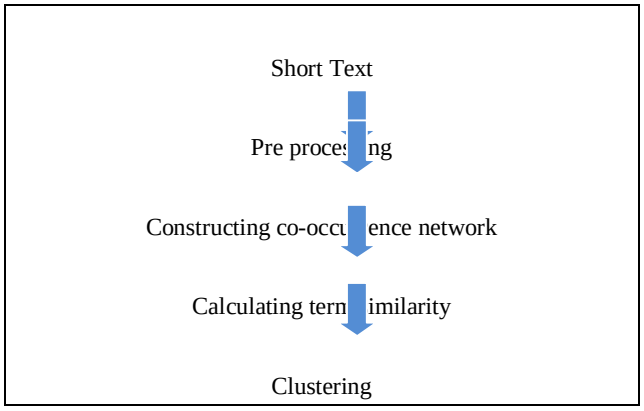

Fig. 1 Architectural Diagram

A. Pre Processing

Data set consists of 1276 queries. The proposed system uses query as dataset. A query containing a sequence of words describes the particular semantics. Short text contains complete sentences, websites. Abbreviations, misspellings etc.

The sample queries in dataset that are used in the proposed system are shown in Table I as follows

TABLE I SAMPLE QUERIES IN DATASET

1	007 omega watch
2	Population studies
3	Chainz birthday
4	Apple company
5	Birthday cake
6	Book lists
7	pink baby shower
8	population of new York state
9	state population
10	www.facebook.com

1) Data preprocessing is done which includes

a) Removing numbers
b) Removing punctuations
c) Stop words removal
d) Stemming
e) Pos tagging

When working with text mining applications, numbers and punctuations and stop words are removed from short text query. Stemming is the process of reducing inflected words to their root form generally a written word form. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation. Queries are stored in text file and it is used as input to the stemming after removing the number, stop words, punctuation. In text mining pos tagging is necessary to identify the types of words in text. Here, POS tagger is used to tag all the words of short text. Maxent POS tagger is used to tag each word in a review sentence. The noun word like 'watch' are tagged as "watch/NN/" and adjective word like omega are tagged as "omega/JJ/". Finally tagged sentences are stored in corpus.

B. Co-Occurrence Network

Co-occurrence networks can be created for any given list of terms in relation to any collection of texts. Co-occurring pairs of terms can be called neighbours and these often group into neighbourhoods based on their interconnections. Individual terms may have several neighbours. Neighbourhoods may connect to one another through at least one individual term or may remain unconnected. Co-occurrence can be regarded as undirected graph where nodes are terms in the text, edge weight can be formulated the strength of semantic relatedness between typed terms. Figure 2 shows the co-occurrence network generated from the list of terms.

Fig. 2 Co-occurrence Network

C. Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

1) k-means
2) Hierarchical
3) Clara
4) Agnes
5) Diana
6) Pam

IV. PERFORMANCE ANALYSIS

The performance analysis can be done with internal validation of clusters. When a clustering result is evaluated based on the data that was clustered itself, this is called internal evaluation. These methods usually assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. One drawback of using internal criteria in cluster evaluation is that high scores on an internal measure do not necessarily result in effective information retrieval applications additionally; this evaluation is biased towards algorithms that use the same cluster model. Internal validation consists of three measures Dunn index, silhouette, connectivity. After performing the internal validation the following results will be produced.

TABLE II
VALIDATION MEASURES

No of cluster		2	3	4	5	6
Hierarchical	Connectivity	2.9290	5.8579	8.7869	11.7159	14.6448
	Dunn	0.9925	0.8507	0.7659	0.6656	0.6751
	Silhouette	0.6114	0.5039	0.4254	.3866	0.3865
K means	Connectivity	3.9925	0.8507	0.7659	0.6656	0.6751
	Dunn	0.9925	0.8507	0.7659	0.6656	0.6751
	Silhouette	0.6114	0.5039	0.4254	0.3866	0.3763
Pam	Connectivity	2.9290	5.8579	9.7159	14.0028	18.2897
	Dunn	0.9925	0.8507	0.4541	0.3303	0.2832
	Silhouette	0.6114	0.5309	0.1998	0.0753	-0.0399
Diana	Connectivity	2.9290	5.8579	8.7869	11.7159	14.6448
	Dunn	0.9925	0.8507	0.7659	0.6656	0.6751
	Silhouette	0.6114	0.5039	0.4254	0.3866	0.3865
Agnes	Connectivity	2.9290	5.8579	8.7869	11.7159	14.6448
	Dunn	0.9925	0.8507	0.7659	0.6656	0.6751
	Silhouette	0.6114	0.5039	0.4254	0.3866	0.3865
Clara	Connectivity	2.9290	5.8579	8.7869	11.7159	14.6448
	Dunn	0.4553	0.4541	0.3941	0.3616	0.353
	Silhouette	0.3723	0.3700	0.2975	0.2643	0.2529

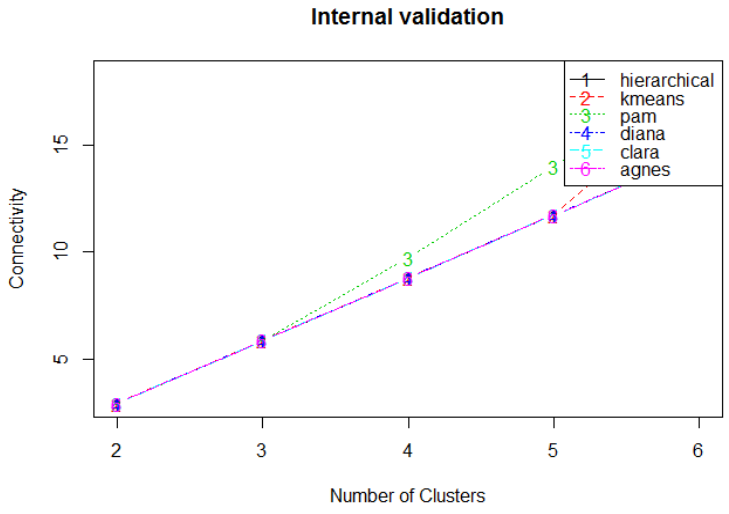

Fig. 3 Connectivity

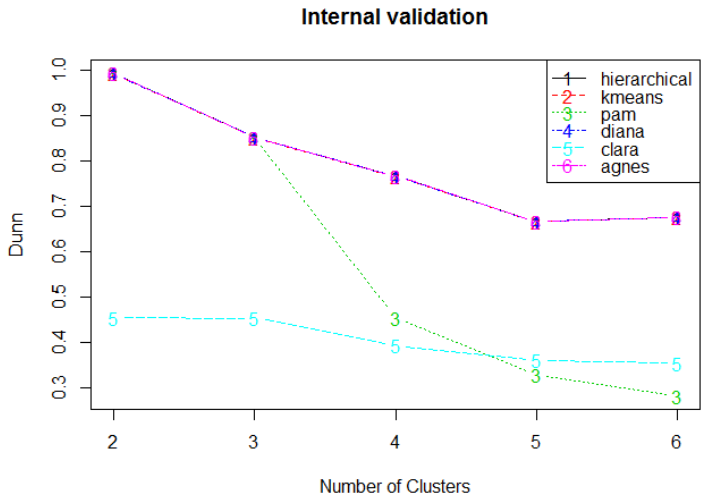

Fig. 4 Dunn Index

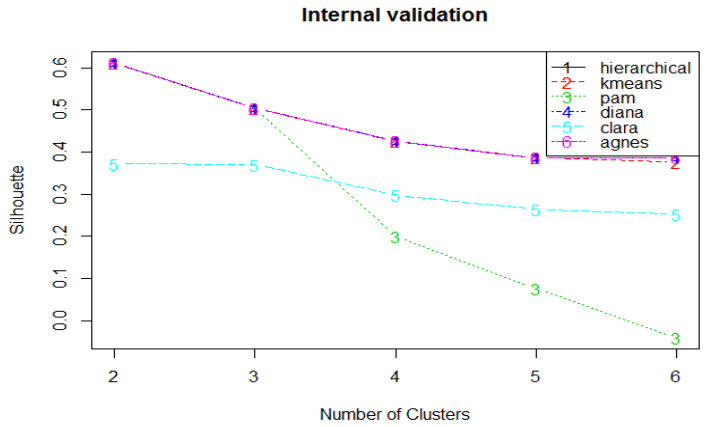

Fig. 5 Silhouette Index

TABLE III OPTIMAL SCORES

Hierarchical	Connectivity	2.9290
	Dunn	0.9925
	Silhouette	0.6114

Among all the clusters Hierarchical clustering with two clusters performs the best in each case. Optimal number of cluster is to be identified using the silhouette. It is high for hierarchical clustering with two clusters and also the k means clustering also has the same silhouette value but its connectivity value is higher than the hierarchical method. Hence the hierarchical clustering method is the best one.

V. CONCLUSION

The Proposed System clusters the words based on the similarities. In this system, text is preprocessed by removing numbers, punctuation, and stop words. Stop words are removed using the stop word dictionary and co-occurrence network is formed between the typed terms. Here the terms frequency is calculated to identify the centrality of the network. And term similarity is calculated based on the similarity between terms and is clustered using various clustering methods. And the clusters are validated by calculating the silhouette, Dunn, connectivity values to analyze the performance of the cluster. After analyzing the performance of the cluster, the result reveals that hierarchical clustering is best one with two clusters because it produces the optimal scores.

REFERENCES

[1] S. Klein and R. F. Simmons, "A computational approach to grammatical coding of english words," J. ACM, vol. 10, no. 3, pp. 334–347, 1963
[2] B. B. Greene and G. M. Rubin, Automatic grammatical tagging of English. Department of Linguistics, Brown University, 1971
[3] E. Brill, "A simple rule-based part of speech tagger," in Proceedings of the workshop on Speech and Natural Language, ser. HLT '91, Strouds burg,PA, USA, 1992, pp. 112–116.
[4] E. BRILL, "Some advances in transformation-based part of speech tagging," in National Conference on Artificial Intelligence, 1994, pp. 722–727
[5] E. Brill, "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging," Comput. Linguist. vol. 21, no. 4, pp. 543–565, 1995
[6] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," in Proceedings of the second conference on Applied natural language processing, ser. ANLC '88, Stroudsburg, PA, USA,1988, pp. 136–143
[7] S. J. DE Rose, "Grammatical category disambiguation by statistical optimization, "Comput. Linguist., vol. 14, no. 1, pp. 31–39, 1988
[8] C. G. de Marcken, "Parsing the lob corpus," in Proceedings of the 28thannual meeting on Association for Computational Linguistics, ser. ACL'90, Stroudsburg, PA, USA, 1990, pp. 243–251
[9] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of speech tagger," in Proceedings of the third conference on Applied natural language processing, ser. ANLC '92, Stroudsburg, PA, USA, 1992, pp.133–140
[10] R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer, and L. Ramshaw, "Coping with ambiguity and unknown words through probabilistic models, "Comput. Linguist. vol. 19, no. 2, pp. 361–382, 1993
[11] B. Merialdo, "Tagging english text with a probabilistic model," Comput. Linguist. vol. 20, no. 2, pp. 155–171, 1994.
[12] H. Schutze and Y. Singer, "Part-of-speech tagging using a variable memory markov model," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ser. ACL '94, Stroudsburg, PA,USA, 1994, pp. 181–187
[13] McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, ser. CONLL '03, Stroudsburg, PA, USA,2003, pp. 188–191
[14] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ser. ACL '02, Stroudsburg, PA, USA, 2002,pp. 473–480
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J.Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[16] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, ser. UAI '04, Arlington, Virginia, United States, 2004, pp. 487–494.
[17] R. Mihalcea and A. Csomai, "Wikify! Linking documents to encyclopaedic knowledge," in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ser. CIKM '07, New York, NY, USA, 2007, pp. 233–242.
[18] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of wikipedia entities in web text," in Proceedings of the 15thACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '09, New York, NY, USA, 2009, pp. 457–466.
[19] X. Han and J. Zhao, "Named entity disambiguation by leveraging wikipedia semantic knowledge," in Proceedings of the 18th ACM conference on Information and knowledge management, ser. CIKM '09, New York, NY, USA, 2009, pp. 215–224
[20] X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: A graph-based method," in Proceedings of the 34th International ACMSIGIR Conference on Research and Development in Information Retrieval,ser. SIGIR '11, New York, NY, USA, 2011, pp. 765–77
[21] W. Shen, J.Wang, P. Luo, and M.Wang, "Linden: Linking named entities with knowledge base via semantic knowledge," in Proceedings of the 21stInternational Conference on World Wide Web, ser. WWW '12, New York ,NYs, USA, 2012, pp. 449–458.