IMPROVED UNSUPERVISED FRAMEWORK FOR SOLVING SYNONYM, HOMONYM, HYPONYMY & POLYSEMY PROBLEMS FROM EXTRACTED KEYWORDS AND IDENTIFY TOPICS IN MEETING TRANSCRIPTS

Sheeba J.I1, Vivekanandan K2

1 Assistant Professor,sheeba@pec.edu
2Professor, k.vivekanandan@pec.edu
Department of Computer Science & Engineering,
Pondicherry Engineering College, Puducherry – 605 014, India

ABSTRACT

Keyword is the important item in a document that provides efficient access to the content of a document. In the Existing system, Synonym, Homonym, Hyponymy and Polysemy problems were solved from only trained extracted keywords in the meeting transcripts. Synonym problem means different words which have similar meaning they are grouped and single keyword is extracted. Hyponymy problem means one word denoting subclass that is considered and super class keyword is extracted. Homonym means a word which can have two or more different meanings.. A Polysemy means word with different, but related senses. Hidden topics from meeting transcripts can be found using LDA model. MaxEnt classifier is used for extracting keywords and topics which will be used for information retrieval Training the keyword from the dataset is separately needed for all the problems, it is not an automatic one .In this proposed frame work, a dataset has been designed to solve the above mentioned four problems automatically.

KEYWORDS:

Keyword, Meeting transcripts, LDA, MaxEnt, Synonym, Homonym, Polysemy, Hyponymy, Dataset

1 INTRODUCTION

Keyword is a word that occurs in text more often with some useful meaning. Keywords provide an efficient information and sharp access to documents concerning their main topics. It can be used for various natural language processes like text categorization and information retrieval. However, most documents will not provide keywords. In particular, spoken documents mostly may not have keywords. On comparing written text and other speech data with meeting speech, meeting speech is much different [1]. There is a sudden increase in Communication, E-marketing, online services and other entertainments, due to the availability of web data in many different forms, genres, and formats than before. This difference in data formats gives new challenges in mining and IR search.

The main challenges in this study are Synonym, Homonym, Hyponymy and Polysemy problems .Synonyms are natural linguistic phenomena which NLP and IR researchers commonly find difficult to cope with. Synonym, that is, two or more different words have similar meanings, causes difficulty in connecting two semantically related documents. For example, the similarity

DOI : 10.5121/ijcsea.2012.2508

between two (short) documents containing Tale and Story can be zero despite the fact that they can be very relevant. Hyponymy is a relation between two words in which the meaning of one of the words includes the meaning of the other word. For example Blue, and Green kinds of color and they are specific colors and color is a general term for them. Homonym means a word that can have two or more different meanings. For example, Bank of India (Institution), River of bank (River). A Polysemy means word with different, but related senses. For example bank has different related meanings: blood bank, financial institution.

Query expansion in IR [2] helps to solve the synonym problem so that retrieval precision and recall will be improved. It retrieves more relevant and better documents by representing user queries with additional terms using a concept-based thesaurus, word co occurrence statistics, query logs, and relevance feedback. Dimension reduction and synonym problem can be solved by mapping vector space model to compact space using mathematical tool by Latent Semantic Analysis(LSA)[3,4].Some studies use clustering as a means to cluster related words before classification and matching[5,6,7].The semantic correlation between words can be represented using taxonomy, ontology, and knowledge base for better classification or clustering.

The paper has deal with a general framework to overcome the above challenge by utilizing hidden topics discovered from data sets. The main idea behind the framework is that, to collect a bulk of data set, and then build a model on both a small set of data and a rich set of hidden topics discovered from the universal data set. A better similarity measure between the documents for more accurate classification, clustering, and matching/ ranking can be given by these hidden topics. Topics inferred from a global data collection help to emphasize and guide semantic topics hidden in the documents in order to handle synonym problem[8].

In this Existing framework, the above mentioned four problems were solved from the extracted trained keywords in the meeting transcripts[9]. In this proposed framework, a data set has been designed to solve these problems to an automatic one .The main advantage of this method is solve the problems from any type of un trained occurrence of keywords .This type of universal dataset is not available in the online.

2 RELATED WORKS

The number of related studies focused on solving Synonym problem. In this section, it will give a short introduction of several studies that found most related to the work. The first group of studies focused on the similarity between very short texts. Metzeler et al. [10] evaluated a large variety of similarity measures for short queries from Web search logs. Yih and Meek [11] considered this problem by improving Web-relevance similarity and the method. Sahami and Heilman [12] also calculated the relatedness between text snippets with the help of search engines and a similarity kernel function

Gabrilovich and Markovitch [13] computed semantic relatedness using Wikipedia concepts. Before topic analysis models, word clustering algorithms were introduced to get better text categorization in different ways. Baker and McCallum [5] tried to condense dimensionality by class distribution-based clustering. Bekkerman et al. [6] combined distributional clustering of words and SVMs. Dhillon and Modha [7] introduced spherical k-means for clustering sparse text data. The text categorization by boosting automatically from extracted concepts by Cai and Hoffman [14] is almost certainly the study most related to this framework. Their method attempts to evaluate topics from data using probabilistic LSA (pLSA) and uses both the original data and resulting topics to train two different weak classifiers for boosting and the difference is that they extracted topics only from the training and test data while they have discovered hidden topics from external large-scale data collections.

Another related work used topic-based features to improve the word sense disambiguation by Cai et al[15].Xuan-Hieu et al.[8] they proposed finding hidden topics using LDA model. J.I.Sheeba et al.[9] discovered for improving the accuracy of extracted keywords,to solve the four problems in the meeting transcripts and also find the topic from the transcripts. But in this framework, it is solved only from the trained extracted keywords, and also it is not an automatic one. To over come this problem here, it is proposed for a modified framework.

3 PROPOSED FRAMEWORK

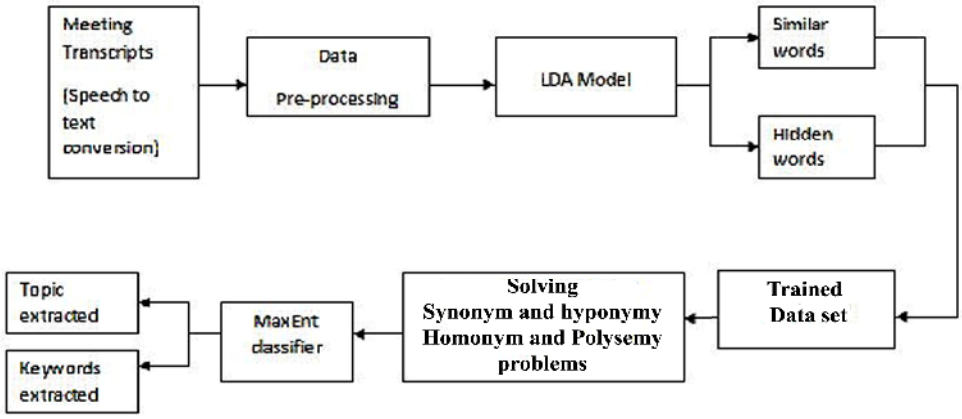

Fig .1. Improved Unsupervised Framework

This Fig 1 represents Unsupervised Hidden Topic Framework and it consists the following steps :

1. Meeting transcripts
2. Data Pre-processing
3. LDA Model
4. Synonym problem
5. Hyponymy problem
6. Polysemy problem
7. Homonym problem
8. MaxEnt classifier
9. Topic Extraction
10. Trained Dataset

3.1 Meeting transcripts

Meeting transcripts are the text file containing meeting speech in readable format. The audio dialogue from meeting is taken as an input to Nuance Dragon Naturally Speaking conversion software tool and speech is converted readable - written text file. Before conversion to text, the software is trained with some data.

3.2 Data Pre-processing

Here text file is taken as an input, in the file stem and stop words are removed to give only the meaningful words. Then using tf-idf frequency of each word is calculated.

3.2.1 Stem word- the form of a word after all affixes are removed; "thematic vowels are part of the stem". Stemming is the process to identify a word by its root which attempts to reduce a

word to its *stem* or root form. Generally, the key terms of a document are represented by stems rather than by the original words. The number of discrete terms are needed for representing a set of documents. The stemming process makes a word shorter by removing such things as prefix or suffix. Examples for stemming the words friendship can be changed into friend after the stemming process. The porter stemming algorithm is used to remove all affixes here.

3.2.2 Stop Word - Stop words are natural language words which have been filtered out after processing. Some search engines do not record extremely common words in order to save space or to speed up searches. These are known as "stop words."

3.2.3 TF-IDF - The **tf–idf** weight (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. Here, the weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

The importance increases proportionally to the number of times a word appears in the transcripts but is offset by the frequency of the word in the transcripts. The term count in the given transcript is simply the number of times a given term appears in that transcript.

3.3 LDA Model

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus.LDA documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document **w** in a corpus *D*:

α is the parameter of the Dirichlet prior on the per-document topic distributions.
β is the parameter of the Dirichlet prior on the per-topic word distribution.
θ_i is the topic distribution for document *i*,
φ_k is the word distribution for topic *k*,
z_{ij} is the topic for the *j*th word in document *i*, and
w_{ij} is the specific word.

LDA Algorithm

For each topic k $\in$ [1, K] do
Generate $\vec{\varphi}_k \sim Dir(\vec{\beta})$
End for
For each document m $\in$ [1, M] do
Generate $\vec{\vartheta}_k \sim Dir(\vec{\alpha})$
Generate $\vec{N}_m \sim poiss(\varepsilon)$
For each word n $\in$ [1,N_m] do
Generate $z_{m,n} \sim Multi(\vec{\vartheta}_m)$
Generate $w_{m,n} \sim Multi(\vec{\varphi}_z)$
End for
End for [8].

Gibb sampling Algorithm

Gibbs sampler is an algorithm to generate a sequence of samples from the joint probability distribution of two or more random variables. The function of such a order is to approximate the joint distribution to approximate the marginal distribution of one of the variables or some subset of the variables. Here the Gibbs sampling algorithm generates an instance from the distribution of each variable in turn and conditional on the current values of the other variables.

Algorithm (Gibbs Sampling)

Specify an initial value $\psi^{(0)} = \left(\psi_1^{(0)}, \ldots, \psi_p^{(0)}\right)$

Repeat for $j = 1,2,\ldots,M$

.

Generate $\psi_1^{(j+1)}$ from $\pi\left(\psi_1 | \psi_2^{(j)}, \psi_3^{(j)}, \ldots, \psi_p^{(j)}\right)$

Generate $\psi_2^{(j+1)}$ from $\pi\left(\psi_2 | \psi_1^{(j+1)}, \psi_3^{(j)}, \ldots, \psi_p^{(j)}\right)$

Generate $\psi_p^{(j+1)}$ from $\pi\left(\psi_p | \psi_1^{(j+1)}, \ldots, \psi_{p-1}^{(j+1)}\right)$

Return the values. $\left\{\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(M)}\right\}$

Using this Gibb sampling algorithm one can yield relatively simple algorithms for approximate inference in high dimensional models like LDA. LDA method uses Gibb Sampling Algorithm and this algorithm takes much iteration to find more relevant words and hidden words as output[8].

3.4 Synonym Problem

A Synonym means different words which have same meaning. From LDA, model similar words are grouped and their topic inference is made. Here, the topics of the similar words are extracted. Each similar group of words contains single topic and that topic is extracted as the output of this problem. By extracting topics, the synonym problem was solved.

3.5 Hyponymy Problem

A Hyponym is a lower class, specific term whose referent is included in the referent of higher class term. Hyponymy is not restricted to objects, abstract, concepts, or nouns . Here word's subclass is considered and its super class is extracted as the output. By extracting super class of each word, Hyponymy problem is solved.

3.6 Homonym Problem

Homonym means same word which has different meanings. In LDA model, keywords are grouped under hidden topics. These topics are labeled with generalized context. Homonym keywords are identified by comparing with hidden topic keywords. The corresponding topics name gives context of keywords and then calculated the frequency used for extraction. The outputs of this problem are keywords and different meaning words.

3.7 Polysemy Problem

Polysemy refers to a word that has two or more similar meanings. Different keywords are presented in the meeting transcripts. Related meaning keywords are identified by comparing with hidden keywords. These identified keywords are used for MaxEnt classifier.

3.8 MaxEnt Classifier

Maximum Entropy is a general technique for estimating probability distribution from data. The distribution is uniform as possible when nothing is known, will have Maximal Entropy. Constraints can be set using labeled training data .Constraints are represented as expected values of “features," any real-valued function of an example. It is a machine learning framework used in classification.

Max Ent takes single observation; it extracts features and groups to one set ,it is robust and has been applied successfully to a wide range of NLP tasks, such as part-of speech (POS) tagging, parsing, Named Entity Recognition (NER).It even performs better than SVM.

It is very fast in both training and inference. Max Ent classifier is trained and on the basis of probability estimation, high probability keywords are extracted from the meeting transcripts.

3.9 Topic Extraction

Topic Extraction means extracting overall topic of the transcript. First, labeled test data has been prepared that will contain topic name and keywords. This labeled data is used for topic extraction that is, labeled data compared with transcript keywords. The topic extraction can be done using LDA model. Most of the keywords in transcript are presented in a particular topic. That topic can be extracted as overall topic of the transcripts.

3.10 Trained Dataset

Trained Dataset is used to collect all the types of words, their actual, related meaning , super class meaning to the particular word from the dictionary and to train all the types of words in the dataset **.** In this proposed framework, it will compare any type of untrained extracted keywords to the dataset and then it solves the problems automatically .In Existing framework, the problem was solved only from trained extracted keywords. Using this proposed dataset, it can solve problems from any type of trained/untrained occurrence of keywords.

4 EXPERIMENTS AND RESULTS

Using this approach, keywords have been extracted from the meeting transcripts which describe about some topic. Nuance Dragon Naturally Speaking conversion software tool converts the audio dialogue to text format Data preprocessing is done and unwanted words are removed. LDA model provides more similar words under each topic as the result for finding hidden topic.

Synonym and Hyponym problems were solved by using Word net as training set. Homonym and Polysemy problems were solved by training dataset. In this framework, it can also solve the above mentioned four problems from untrained extracted occurring keywords occurs .Finally MaxEnt Classifier was trained using constraints and most probable keywords are extracted. Here Fig 2 represents after solving Synonym and Hyponymy problem results and Fig 3 shows after solving Polysemy & Homonym problem results.

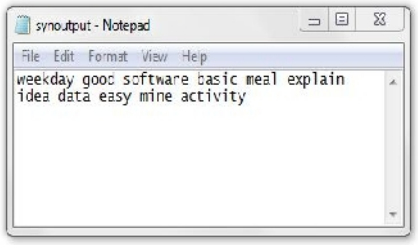

Fig.2.After solving Synonym & Hyponymy problem

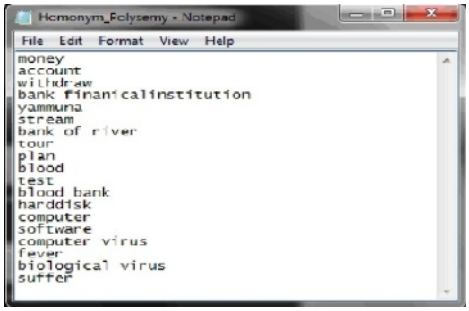

Fig 3.After solving Polysemy & Homonym problem

5 CONCLUSION

The unsupervised framework provides a solution to solve the Synonym, Hyponymy, Homonym and Polysemy problems from extracted keywords in the meeting transcripts. Discovering the hidden topics makes the framework more efficient and reduces the complexity. Topic and keywords can be extracted more accurately by solving those problems. Because, executing the iteration in LDA model takes less time compared to other hidden topic model.

This framework focused on solving Synonym and Hyponym problems by reducing the similar keywords and provides better results. Homonym and Polysemy problem give more accurate meaning to the keyword. In the Existing work, the above mentioned four problems solved only from trained extracted keywords. But in this framework the problems will be solved even untrained occurring keywords using proposed trained dataset. These four problems will be automatically solved by this dataset quickly compared to an Existing one .Thus, this framework extracts keywords in an effective and an efficient way and it will be solved any type of occurring keywords.

REFERENCES

[1] Feifan Liu, Deana Pennell, Fei Liu :Unsupervised Approaches for Automatic keyword extraction , Boulder, Colorado, ACM, June (2009).

[2] C.D. Manning, P. Raghavan, and H. Schutze,: Introduction to Information Retrieval, Cambridge Univ. Press, Springer (2008).

[3] S. Deerwester, G. Furnas, and T. Landauer :Indexing by Latent Semantic Analysis, J. Am. Soc. for Information Science, vol. 41, no. 6, pp. 391-407 (1990).

[4] T.A. Letsche and M.W. Berry : Large-Scale Information Retrieval with Latent Semantic Indexing, Information Science, ACM, vol. 100, nos. 1-4, pp. 105-137(1997).

[5] L. Baker and A. McCallum: Distributional Clustering of Words for Text Classification, Proc. ACM SIGIR(1998).

[6] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter : Distributional Word Clusters vs. Words for Text Categorization. Machine Learning Research, ACM, vol. 3, pp. 1183-1208 (2003).
[7] I. Dhillon and D. Modha : Concept Decompositions for Large Sparse Text Data Using Clustering, Machine Learning, ACM, vol. 42, nos. 1/2, pp. 143-175 (2001).
[8] Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen : A Hidden Topic-Based Framework toward Building Applications with Short Web Documents, IEEE Transactions on Knowledge and Data Engineering ,vol. 23, (2011).
[9] J.I.Sheeba, Dr.k.Vivekanandan: Unsupervised hidden topic framework for extracting keywords (Synonym ,Homonym, Hyponymy, Polysemy)and topics in meeting transcripts .The second international conference on Advances in Computing and Information Technology, ACITY ,Springer, July (2012).
[10] D. Metzler, S. Dumais, and C. Meek :Similarity Measures for Short Segments of Text, Proc. 29th European Conference IR Research (ECIR), ACM (2007).
[11] W.Yih and C. Meek : Improving Similarity Measures for Short Segments of Text, Proc. 22nd National Conference on Artificial Intelligence (AAAI) (2007).
[12] M. Sahami and T. Heilman : A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets, Proc. 15th International Conference on World Wide Web, ACM (2006).
[13] E. Gabrilovich and S. Markovitch: Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis, Proc. 20th Int'l Joint Conference. Artificial Intelligence (2007).
[14] L. Cai and T. Hofmann : Text Categorization by Boosting Automatically Extracted Concepts, Proc. ACM SIGIR (2003).
[15] J. Cai, W. Lee, and Y. The : Improving WSD Using Topic Features, Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL), pp. 1015–1023, Prague, June (2007).