Combination of Bayesian and Latent Semantic Analysis with Domain Specific Knowledge

Shen LU
Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33612 USA
E-mail: shenlu@mail.usf.edu
Richard S. SEGALL
Department of Computer & Information Technology, Arkansas State University, State University, AR 72467 USA
E-mail: rsegall@astate.edu

ABSTRACT

With the development of information technology, electronic publications become popular. However, it is a challenge to retrieve information from electronic publications because the large amount of words, the synonymy problem and the polysemi problem. In this paper, we introduced a new algorithm called Bayesian Latent Semantic Analysis (BLSA). We chose to model text not based on terms but associations between words. Also, the significance of interesting features were improved by expand the number of similar terms with glossaries. Latent Semantic Analysis (LSA) was chosen to discover significant features. Bayesian post probability was used to discover segmentation boundaries. Also, Dirchlet distribution was chosen to present the vector of topic distribution and calculate the maximum probability of the topics. Experimental results showed us that both P_k [8] and WindowsDiff [27] decreased 10% by using BLSA in comparison to the Lexical Cohesion with the original data.

KEYWORDS

Glossaries, Bayesian Latent Semantic Analysis (BLSA), Latent Semantic Analysis (LSA), Linkage Detection, Topic Segmentation

1. INTRODUCTION

The change of the way people publish, obtain and exchange information gives rise to new challenges that possibly there is chance that we can use modern technology to create or improve the way we obtain, manage, and integrate information. We have this idea is because of such a problem that, in the era of electronic publications, we can generate and deliver information easier and faster than before. On the other hand, in order to find the information we need, we have to collect and process more information, which is time-consuming. The solution is that, if we can automatically put information into different categories, we can efficiently reduce the total amount of information we need to go through and only keep the most related information.

For example, in research fields, conference proceedings have a lot of articles and most of researchers need to find the frontier research in several proceedings on a regular basis. There are lots of electronic magazines and blogs online and most of them have similar contents. Even in electronic books, there are many pages in one book. If we can find a way to efficiently put information into different categories, we can read less.

In this paper, we presented a way to efficiently group text into different topics by using latent semantic analysis and Bayesian theory, especially with the combination of domain glossaries which is downloaded from the WikiHyperGlossary project [1].

2. RELATED WORK

Below we discuss a summary of the work of researchers in which are discussed the applications of the concepts of Latent Semantic Analysis (LSA), Bayesian Latent Semantic Analysis (BLSA) and Latent Semantic Indexing (LSI). Document indexing and representation of term-document relations are very important issues for document clustering and retrieval. In Matreeva et al. (2005) [24], Generalized Latent Semantic Analysis (GLSA) is used a framework for computing semantically motivated term and document vectors. The experiments performed by Matreeva et al. (2005) [24] demonstrate that GLSA term vectors efficiently capture semantic relations between terms and outperform related approaches on the synonymy test.

In the Chien and Wu (2008) [4], an extension to Probabilistic Latent Semantic Analysis (PLSA) into a Bayesian framework is presented for the statistical modeling of documents. Chien and Wu (2008) [4] focuses on exploiting the incremental learning algorithm for solving the updating problem of new domain articles. The “adaptive Bayesian Latent Semantic Analysis (BLSA)” was developed by Chien and Wu (2008) [4] to improve document modeling by incrementally extracting up-to-date semantic information to match the changing domains at run time.

An incremental PLSA algorithm is constructed by Chien and Wu (2008) [4] to accomplish the parameter estimation as well as hyper-parameter updating. Compared to standard PLSA using maximum likelihood estimate (MLE), the proposed approach is capable of performing dynamic document indexing and modeling. Chien and Wu (2008) [4] also presents the maximum a posteriori PLSA for corrective training. The experiments performed by Chien and Wu (2008)[4] on information retrieval and document categorization demonstrate the superiority of using Bayesian PLSA methods.

Bayesian Latent Semantic Analysis (BLSA) was also discussed by DeFreitas and Barnard (2000, 2001)[6][7], Eisenstein and Barzilay (2008)[9] for Bayesian Unsupervised Topic Segmentation, Hoffman (1999, 2001, 2004) [17][18][19] for PLSA, and Yu et al.{2005)[32] for Dirichlet Enhanced Latent Semantic Analysis.

Probabilistic Latent Semantic Analysis (PLSA) has many applications most prominently in information retrieval, Natural Language Processing (NLP), machine learning from text, and in related areas. Hoffman in [18] presents that perplexity that result for different types of text and linguistic data collections and discussing an application of automated document indexing. The experiments conducted by Hofmann (2001) [18] indicate substantial and consistent improvements of the probabilistic method over standard Latent Semantic Analysis (LSA).

Hoffman (2003) in [19] described a new model-based algorithm designed for collaborative filtering via Gaussian Probabilistic Latent Semantic Analysis (PLSA). In [19], Hofmann (2003) used collaborative filtering at learning predictive models of user preferences, interests or behavior from a database of available user preferences. It is assumed that users can participate probabilistically in one of more groups. Hofmann (2003) concluded in [19] that with the experiments performed on each of the movie data sets that the proposed approach compared favorably with other collaborative filtering techniques.

Gong and Liu (2001) in [13] performed generic text summarization using relevance measures and latent semantic analysis. Gong and Liu (2001) proposed in [13] two generic text summarization methods that create text summaries by ranking and extracting sentences from the original documents, The first method uses standard information retrieval methods to rank sentence relevancies, while the second method uses the latent semantic analysis technique to identify semantically important sentences. Both methods in Gong and Liu (2001) [13] strive to select sentences that are highly ranked and different from one another as an attempt to create a summary with a wider coverage of the document's main content and less redundancy.

The following presents some highlights of related work on topic segmentation. Purver (2011)[28] of Queen Mary University of London discussed in a chapter the task of topic segmentation: automatically dividing single-long recordings or transcripts into shorter, topically coherent segments. Morris and Hirst (1991) [25] wrote a ground-breaking paper on topic segmentation that lead to a steadily growing research are in computational linguistics.

Eisenstein(2009)[10] discussed hierarchical text segmentation from multi-scale lexical cohesion in a Bayesian setting using collapsed variational Bayesian inference over the hidden variables. The resulting system by Eisenstein (2009)[10] is shown to be fast and accurate and compares well against heuristic alternatives.

Other investigators in the applications of topic segmentation to analysis of linguistics and the spoken discourse include Purver et al. (2006) [29], Buch-Kromann and Korzen (2010) [3], Hsueh et al. (2006)[20], Malioutov and Barzilay (2006) [23], Elsner and Charniak (2008) [11], Niekrasz and Moore (2009) [26], and Van der Vloet et al. (2011) [30].

3. CONCEPTS

3.1 DISTRIBUTION OF THE SEMANTIC STRUCTURE

3.1.1 Modeling Text Data

If we ignore the associations between words and documents, the text data can be simplified as a bag of words and each individual word is the data unit. The distribution of data belongs to univariate distributions. however, the performance of the term matching retrieval is not promising The reason is that the query may not use the words in the documents but the synonymies instead. The synonymy problem and the polysemy problem are main issues to cause the decrease of the accuracy in recall in word matching retrieval.

It is easy to think of a solution to fix this problem, such as automatic term expansion, the construction of thesaurus, and so on. However, given a set of individual words, since the text data set is incomplete and lack of reliable evidences, it is hard to take redundancy into consideration. Since there are some problems in term based retrieval, we look for higher order structure in which terms are replaced by more reliable indicates, such as the association between terms and documents. In a article, meaningful terms are not independent of the article. They are associated with the article.

3.1.2 Semantic Structure

The goal is to find a model and fit it with the association between terms and documents. In this way, we turn the information retrieval problem in to a statistical problem. The semantic structure of the term and document association is listed below. We choose multivariate polya distribution in Johnson et al., (1997) [21] to represent the semantic structure.

In multivariate polya distribution (MPD), given k groups, each group U contains the same number of features V, each feature belongs to different topics. (U, V) means there are U features of topics V. For group i, the distribution can be presented as the following: (U_{1i}, V_{1i}), (U_{2i}, V_{2i}), ... , (U_{mi}, V_{mi}). We use a large matrix of term document associations to construct the semantic space. The entire model can be presented as the following

groups	1st group	ith group
topics	V_{11}, V_{21}, ... , V_{mi}	V_{1i}, V_{2i}, ... , V_{mi}
row		
1		
2		
...		
n		
	S_{11}, S_{21}, ... , S_{m1}	S_{1i}, S_{2i}, ... , S_{mi}

In the model, each group has a set of features; in the column, each feature U_{ji} is corresponding to a topic V_{ji}, and S_{ji} is the frequency that topic V_{ji} is chosen from group i in the n drawings.

$$\sum_{j=1}^{m} Sji = n \quad \text{for all } i = 1,2, ..., k. \qquad (1)$$

$$\sum_{j=1}^{m} U_{ji} = N \quad \text{N is the number of features} \qquad (2)$$

$$\prod_{r=1}^{n} [N + (R - 1)g] \qquad (3)$$

$$\blacksquare \frac{\prod_{j_1=1}^{m_1} \prod_{r_{j1}=1}^{S_{j1}} \left[G_{j1}+\left(r_{j1}-1\right)g\right]}{\prod_{r=1}^{n} [N+(r-1)g]} \qquad (4)$$

When we randomly draw color balls from the urns, we work in this way. For example, if we want to draw n balls from the first urn, there are N balls in the urn at the first drawing, and then increasing by g at each drawing. The product for the denominators for the first urn is

We obtain the product of all ratios for the first urn as

similar expressions hold for all i = 1, 2, ..., k urns. We obtain the expression for all random variable X as

In which,

K: number of urns

N: number of balls in each urn before the first drawing

G_{ji}: number of balls with color A_{ji} in urn i before the first drawing

g: number of balls of the chosen color added to the respective urns.

m_i: number of balls of different colors in urn i

$$\frac{n!\cdot \prod_{i=1}^{k} \prod_{j_i=1}^{m_i} \prod_{r_{ji}=1}^{S_{ji}} \left[G_{ji} + \left(r_{ji} - 1\right)g\right]}{\prod_{j_1=1}^{m_1} \prod_{j_2=1}^{m_2} \left(\cdots \prod_{j_k=1}^{m_k} b_{j_1 \cdot j_2 \cdot j_3 \ldots}! \right) \left[\prod_{r=1}^{n} [N + (r - 1)g] \right] \blacksquare^{k}} \qquad (5)$$

n: number of drawings

S_{ji}: frequency with which a ball of color A_{ji} is drawn from the i^{th} urn in n drawings.

When k=1, m=1 and g=0, we can have the distribution for bag of words. In other words, bag of words is a special

$$\frac{\prod_{r_i=1}^{S_i} G_i}{\prod_{r=1}^{n} N} \qquad (6)$$

distribution of MPD.

3.2 LATENT SEMANTIC ANALYSIS

Latent Semantic Analysis (LSA) in Deerwester et al. (1990)[8] is a general theory of acquired similarity and knowledge representation. LSA can be used to discover knowledge from text with a general mathematical learning method without knowing prior linguistic or perceptual similarity knowledge. The motivation of LSA in terms of psychology is that people learn knowledge only from similarity of individual words taken as units, not with knowledge of their syntactical or grammatical function. LSA assumes that the dimensionality of the context in which all of the local words are represented is of great importance and the reduction of dimensions of the observed data from original text to a much small but still large number can improve human cognition.

LSA consists of two steps:

1. Represent the text as a matrix in which each row is a unique word and each column is a text message or other context. Each cell contains the frequency of the word in column of the corresponding passage. The frequency of the cell entry is weighted by a function that expresses both the importance of the word in the particular passage and how much information the word has in general.

2. LSA applies Singular Value Decomposition (SVD) to the matrix. Latent Semantic Analysis (LSA) is a Natural Language Processing (NLP) technique that is based on similarity of words but not grammatical or syntactical structure and extracts knowledge through the similarity of individual words. Document indexing and representation of term-document relations are very important issues for document clustering and retrieval.

We store documents into t*d matrix X.

$$X = T_0 \cdot S_0 \cdot (D')_0 \qquad (7)$$

In which, t is the vector of terms, d is the vector of documents. T_0 and D_0 have orthonormal columns and S_0 is diagonal. This is called singular value decomposition of X.

3.3 TEXT SEGMENTATION

Most of words in natural language have multiple meanings that can only be determined by considering the context in which they occur. Given a target word used in a number of different contexts, word sense discrimination is the process of grouping these instances of the largest word together by determining which contexts are the most similar to each other.

There are several ways to discover text segmentation. TextTiling which was discussed in Hearst (1997)[15], Hearst and Plaut (1993)[16]; Hearst (1994)[14] is one of the text segmentation method. We define a sliding window and then use it to go through the text. The lexical similarity is calculated for each pair of adjacent window by using the cosine score. Local minima are calculated by comparing the depth score for each point based on its relative depth below its nearest peaks on either side. The segment boundary starts at the point with the highest scores.

LSA is another method which is mention in the previous section. Both TextTiling and latent semantic analysis consider the connections of adjacent words. For TextTiling, we only consider the connections in a sliding window, but for LSA we consider the connections in the entire sentence.

The topics defined by our model is based on LSA. The reason we prefer to use LSA is that the words are not taken from predefined domain knowledge but rather emerge in a data-driven manner from the similarity of features in difference documents. Given a corpus of training documents, we do not have prior knowledge about the association of the content. But, we can find the similarity of different words with semantic analysis by using Latent Semantic Analysis. In this way, we can define feature sets. Each feature is a group of synonyms, which may appear in difference documents. In order to discover similar documents together, we assume that similar documents talk about similar topics so that they use synonyms but explain in different ways.

3.4 TOPIC DETECTION

Topic shift can be discovered by comparing the similarity of the new segment and the topic context of the segments. This comparison is performed in two steps: first, we calculate the similarity of the new segment and the topic context of the segments; then, the decision procedure shows if the two contexts are similar. As provided by Choi (2000)[5] and Kaufmann (1999)[22], we use cosine measure to evaluate the similarity between the new segment context and the topic segment context.

Algorithms for topic shift detection are mentioned by Ferret and Grau (2000)[12]. It works in this way: at each topic context, if the similarity between the topic context of the new segment and the topic context of the previous segments is rejected, a topic shift is generated and a new topic context is opened. Otherwise the active topic context is extended to the new segment. There are four states of segment comparison in the topic shift algorithm: NewTopicDetection state, InTopic state, EndTopicDetection state, and OutOfTopic state.

The process of segment starts with OutOfTopic state, after the end of the previous segment. If the focus window is stable enough between the two successive positions, it turns into the NewTopicDetection state. If the new segment context is consistent with the next position, the InTopic state can be reached. Otherwise, it assumes that it is a false alarm and return to OutOfTopic state. As soon as the current segment context begins to change significantly between two successive positions. Especially, when the algorithm stays in OutOfTopic state for too long, it creates a new topic which covers all of the concerned positions.

When the topic shift algorithm goes from the NewTopicDetection state to the InTopic state, it first go through all of the topic contexts of the previous segments and find out if the topic of the current context is similar to one of the topic contexts it has discovered. A specific threshold is used for decision making. If the similarity measure is greater than the threshold, the current topic context is linked to the existing topics; otherwise, a new topic is created.

4. COMBINATION OF GLOSSARIES WITH DOCUMENTS

Documents contain terms from the corresponding glossaries in the same domain areas. Glossaries include all of the domain knowledge from different areas, which are described by the significant terms and their corresponding definitions.

In text mining, one of the issues is how to extract significant terms from the text. All of the terms associated with domain knowledge are distributed everywhere in an article and are mixed with general words which have nothing to do with the domain knowledge. Latent Semantic analysis can provide the meanings of the terms based on the context. However, one article cannot include all of the domain knowledge and the definition extracted from the context where the term appears in that article is not accurate. But, in glossaries, all of the terms are defined clearly.

In this paper, we manually put the definitions of the terms in glossaries to those words in an article and use those definitions to improve the accuracy of the background knowledge we can extract from the context. In this way, we can define meaningful words and use them to decide the theme of the corresponding sections.

In this paper, we also defined the following process:

1. Specify the representation of the text: turn a plain text document into N*D dimensional matrix V. The value of each element is the conditional probability of the term calculated by above where N is the number of initial elements to be grouped into coherent segments and D the dimension of the element representation. The significant features can be discovered by using SVD

To find the similarity of the adjacent segments, we can measure the similarity of the two vectors by using Cosine similarity.

$$\frac{S_i}{\sum_{m=1}^{n-1} S_m} \qquad (8)$$

2. Score the candidate segments: scoring function can be used to specify the coherence of a segment of text. The score function can be used to represent the rank transformed vector of its cosine distances to each other element. In which S_i is the similarity between sentence i and sentence (i+1)

We choose to use Bayesian post probability to computer the candidate score because Bayesian post probability can be used to calculate the global probability of each sentence. The peak point is supposed to be the segmentation boundary.

3. Topic selection for each sentence. According to MPD score of each element in the sentence, we can choose the topic for each sentence by using

$$\theta(j,i) = \frac{n_{(j,i)} + \theta_0}{\sum_{m=i}^{W} n_{(j,i)} + W\theta_0} \qquad (9)$$

In which, n(j, i) is the vector of the sentence j in which each element is the MPD point estimation of the word, W is count of the words in sentence j. $\theta_j \sim Dir(\theta_0)$

5. EXPERIMENTS AND RESULTS

We did experiments the benchmark dataset (Walker et al., (1990)[31]). The dataset consists of 121 medical diagnoses in free text format. The task is to divide each chapter into the sections according to the difference of the content. The dataset contains 227 chapters and 1137 sections. We choose 50 definitions from medical glossaries. We create five different datasets by adding 10 definitions, 20 definitions, 30 definition, 40 definition and 50 definitions to the dataset as background knowledge.

$$P_k = \frac{\sum_{i=1}^{N-k} \left[\delta_H(j, j+k) \oplus \delta_R(i, j+k)\right]}{(N-k)} \qquad (10)$$

$$\delta_s(j,j) = \begin{cases} \text{1 if segmentation S assigns i and j} \\ \text{To the same segment} \\ \text{0 otherwise.} \end{cases}$$

All experiments are evaluated in terms of commonly-used P_K (Beeferman et al.(words in segmentation. K is the window size. 1999)[2]) and WindowDiff (WD) (Pevzner and Hearst, (2002)[27]) scores. H is the hypothesis segmentation, R is the reference segmentation. N is the the number of

$$WD = \frac{\left[\sum_{i=1}^{N-k} \left[\left|b_H(i,j) - b_R(i,i+k)\right| > 0\right]\right]}{(N-k)} \qquad (11)$$

P_k can be used to indicate the missing boundaries, but fails to indicate the false alarms. The lower P_k, the less missing boundaries. We can use WindowsDiff to present the false alarms.

$$\left|b_H(i,j) - b_R(i,i+k)\right| > 0 \qquad (12)$$

In which, the lower WindowsDiff, the less false alarms.

The two measurements can indicate whether the sentences on the edges of the window are properly segmented with respect to each other. P_K indicates whether the two sentences are in the same segment or not. WindowDiff shows the number of intervening segments between the two sentences identical in the hypothesized and the reference segments. P_K and WindowDiff are penalties, so lower values indicate better segments. In Figure 2, we can see that both P_K and WD decrease by adding more and more definitions to the text. That means we can divide the text into better segments. So, by combining glossaries with the text, we can efficiently improve the linkage discovery from the text.

Table 1 below shows P_K and WindowDiff values on the data set with no definition, the dataset with 10 definitions, the dataset with 20 definitions, the dataset with 30 definitions, the dataset with 40 definitions, the dataset with 50 definitions. Figure 2 below show the P_K and WindowDiff values on six different datasets.

Below are tables and figures for ten experiments that were conducted for different data sets. For each experiment, we randomly chose documents from the dataset, and then changed the data by adding to the documents 10 definitions, 20 definitions, 30 definition, 40 definitions, and 50 definitions. We measured the Pk values and WindowsDiff values on different datasets. We repeated the experiment for 10 times. The performance of the ten experiments were listed in table 1 to table 10.

We can see, by using BLSA model, P_k is 10% lower than the one using Lexical cohesion model. WindowsDiff is also 10% lower.

Table 1. Experiment 1

Data Set #1	p_k	WD
0 definitions	0.6263	0.6263
20 definitions	0.3704	0.3704
30 definitions	0.38	0.38
40 definitions	0.3305	0.3305
50 definitions	0.3353	0.3353

Table 2. Experiment 2

Data Set #2	p_k	WD
0 definitions	0.3723	0.3723
10 definitions	0.3889	0.4125
20 definitions	0.3878	0.4132
30 definitions	0.3871	0.3871

40 definitions	0.2024	0.324
50 definitions	0.2022	0.3238

Table 3. Experiment 3

Data Set #3	p_k	WD
0 definitions	0.5173	0.5235
10 definitions	0.6051	0.6513
20 definitions	0.5021	0.552
30 definitions	0.3767	0.4257
40 definitions	0.5373	0.5429
50 definitions	0.5422	0.5434

Table 4. Experiment 4

Data Set #4	p_k	WD
0 definitions	0.2537	0.2933
10 definitions	0.296	0.2963
20 definitions	0.2483	0.2993
30 definitions	0.2478	0.299
40 definitions	0.1437	0.2004
50 definitions	0.1437	0.2001

Table 5. Experiment 5

Data Set #5	p_k	WD
0 definitions	0.392	0.392
10 definitions	0.3602	0.3602
20 definitions	0.383	0.383
30 definitions	0.3753	0.3753
40 definitions	0.3718	0.3718
50 definitions	0.3643	0.3643

Table 6. Experiment 6

Data Set #6	p_k	WD
0 definitions	0.5049	0.5049
10 definitions	0.345	0.345
20 definitions	0.3519	0.3519
30 definitions	0.3564	0.3564
40 definitions	0.3628	0.3628
50 definitions	0.2057	0.2057

Table 7. Experiment 7

Data Set #7	p_k	WD
0 definitions	0.3016	0.3016
10 definitions	0.1832	0.2443
20 definitions	0.1856	0.2428
30 definitions	0.1601	0.2187
40 definitions	0.2544	0.2544
50 definitions	0.1463	0.2123

Table 8. Experiment 8

Data Set #8	p_k	WD
0 definitions	0.2042	0.2042
10 definitions	0.4596	0.4596
20 definitions	0.5355	0.5355
30 definitions	0.5305	0.5303
40 definitions	0.5219	0.5219
50 definitions	0.5235	0.5235

Table 9. Experiment 9

Data Set # 9	p_k	WD
0 definitions	0.5727	0.5727
10 definitions	0.6508	0.6508
20 definitions	0.5107	0.5107
30 definitions	0.4689	0.4689
40 definitions	0.4207	0.4207
50 definitions	0.4373	0.4373

Table 10. Experiment 10

Data Set #10	p_k	WD
0 definitions	0.5361	0.5361
10 definitions	0.6454	0.6454
20 definitions	0.6182	0.6182
30 definitions	0.6203	0.6203
40 definitions	0.596	0.596
50 definitions	0.5816	0.5816

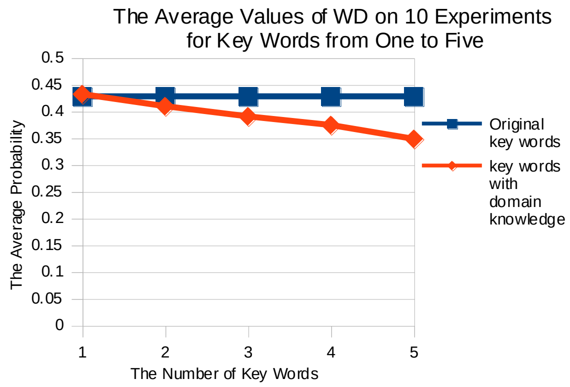

Figure 3(a). Average PK values on 10 different datasets.

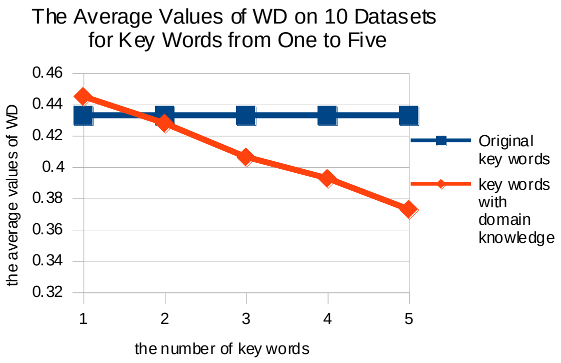

Figure 3(b). Average WindowDiff values on 10 different datasets.

6. CONCLUSIONS

This paper introduced a new algorithm called Bayesian Latent Semantic Analysis (BLAS) for topic detection. We also introduce data modeling for text segmentation and the distribution we choose to model text. We presented and compared the existing text segmentation and topic detection methods and the advantage of using BLSA for topic detection. We also used glossaries to increase the weights of the significant features in text data. We tested BLSA algorithm on 121 medical diagnoses documents with no structures. We increasingly add domain knowledge to the text and tested the performance of BLSA on 10 different datasets. Experimental results showed us that BLSA efficiently reduced P_k and WindowsDiff about 10%. By combining glossaries with Bayesian Framework, we can improve the linkage discovery from the text. In future, we will apply this idea to some specific domains by using domain specific glossaries. Combining this work with specific domains will bring this research of linkage discovery to a higher level for future exploration of data and glossaries.

7. REFERENCES

[1] Bauer, M.A., Berleant, D., Cornell, A.P., and R. E. Belford. WikiHyperGlossary (WHG): An information literacy technology for chemistry documents**.** ***Journal of Cheminformatics*** (2015), 7:22.

[2] Beeferman, D., Berger, A. and Lafferty, J.D., (1999), Statistical models for text segmentation, ***Machine Learning***, vol. 34, no. 1-3, pp. 177-210.

[3] Buch-Kromann, M. and Korzen, I. (2010), "The Unified Annotation of Syntax and Discourse in the Copenhagen Dependency Treebanks", ***Proceedings of the Fourth Linguistic Annotation Workshop***, Association of Computational Linguistics (ACL) 2010, Uppsala, Sweden, July 15-16, pp. 127-131

[4] Chien, J-T, and Wu M-S.(2008), Adaptive Bayesian Latent Semantic Analysis, **IEEE *Transactions on Audio, Speech, and Language Processing***, 16(1), Janauary pp. 198-207.

[5] Choi, F. (2000), Advances in domain independent linear text segmentation, **NAACL'00**, pp. 26-33.

[6] De Freitas, N. and Barnard, K. (2000). Bayesian Latent Semantic Analysis, University of California at Berkeley, Department of Computer Science,

[7] De Freitas, N. and Barnard, K. (2001). Bayesian Latent Semantic Analysis of multimedia databases, ***Technical Report TR 2001-15***, University of British Columbia, Department of Computer Science. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.9012

[8] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), 'Indexing by latent semantic analysis'**,** ***Journal of the American Society for Information Science***, vol. 41, n.6, pp. 391-407.

[9] Eisenstein, J. and Barzilay, R. (2008). Bayesian Unsupervised Topic Segmentation, ***EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*****,** Association for Computational Linguistics, Stroudsburg, PA, pp. 334-343.

[10] Eisenstein, J.(2009). "Hierarchical Text Segmentation form Multi-Scale Lexical Cohesion", Human Language Technologies: ***The 2009 Annual Conference of the North American Chapter of ACL***, Boulder, CO, June, pp. 353-361.

[11] Elsner, M. and Charniak, E.(2008). "You talking to me? A Corpus and Algorithm for Conversation Disentaglement", **Proceedings of ACL-08**, HLT, pp. 834-842.

[12] Ferret, O. and Grau, B. (2000). A topic segmentation of texts based on semantic domains, ***ECAI 2000***, pp.426-430.

[13] Gong, Y. and Liu, X.(2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, ***Proceedings of IGIR '01,*** September 9-12,New Orleans, LA, pp. 19-25.

[14] Hearst M. (1994) Multi-paragraph Segmentation of Expository Text. In ***Proceedings of the 32th Meeting of Association of Computational Linguistics (ACL),*** 9-16. Las Cruces, NM, June.

[15] Hearst M. (1997) TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. ***Proceedings of Association for Computational Linguistics (ACL).*** 23(1), 33-64.

[16] Hearst, M. & Plaunt, C. (1993) Subtopic structuring for full-length document access. In ***Proceedings of the 16th Annual International ACM/SIGIR Conference***, 59-68, Pittsburgh, PA

[17] Hoffmann, T. (1999). Probabilistic Latent Semantic Indexing, ***Proceedings of 22nd Annual International Conference SIGIR Conference on***

Research and Development in Information Retrieval.

[18] Hoffmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis***, Machine Learning***, 42, pp. 177-196

[19] Hofmann, T. (2003), Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis, ***Proceedings of SIGIR'03***, July 28-August 1, pp. 259-266.

[20] Hsueh, P-Y., Moore, J.D. and Renals, R. (2006). "Automatics Segmentation of Multiparty Dialogue", ***Proceedings of Spoken Language Technology Workshop***, IEEE, December 10-13, pp. 98-101.

[21] Johnson, N., Balakrishnan, N. & Kotz, S (1997) Discrete Multivariate Distributions. Wiley Series in Probability and Statistics. ***Applied Probability and Statistics***. pp 212-217. ISBN: 0471128449.

[22] Kaufmann, S. (1999). Cohesion and collocation: using context vectors in text segmentation, ***ACL'99***, pp. 591-595.

[23] Malioutov, I. and Barzilay, R. (2006). "Minimum Cut Model for Spoken Lecture Segmentation", ***Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL***, Sydney, Association of Computational Linguistics, pp. 25-32.

[24] Matveeva, I, Kevow G-A, Farahat, A, and Royer, C. (2005). Term Representation with Generalized Latent Semantic Analysis, ***Proceedings of RANLP*** 2005, http://faculty.washington.edu/levow/papers/SynGLSA_ranlp_final.pdf

[25] Morris, J. & Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1), 21-48.

[26] Niekrasz, J. and Moore, J. (2009). Participant Subjectivity and Involvement as a Basis for Discourse Segmentation, ***Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue***, pages 54–61.

[27] Pevzner, L. and Hearst, M.A. (2002). A critique and improvement of an evaluation metric for text segmentation, ***Computational Linguistics***, vol. 28, no. 1, pp. 19-36.

[28] Purver. M. (2011). Topic Segmentation. In Tur & de Mori (eds.), ***Spoken Language Understanding: Systems for Extracting Semantic Information from Speech***, Wiley.

[29] Purver, M., Kording, K.P., Griffiths, T.L., and Tenanbaum, J.B.(2006). "Unsupervised Topic Modelling for Multi-Party Spoken Discourse", ***Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL***, Sydney, July, pp. 17-24.

[30] Van der Vloet, N. Berzlanovich, I., Bouma, G., Egg, M. and Redeker, G. (2011). "Building a Discourse Dutch Text Corpus," ***Proceedings of the Workshop 'Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena'"***, Gottingen, Germany, February 22-25.

[31] Walker, K., Hall, D.W., and Hurst, J.W. (1990). ***Clinical Methods: The History, Physical and Laboratory Examinations***. Butterworths.

[32] Yu, K., Yu, S. and Tresp, V. (2005). Dirichlet Enhanced Latent Semantic Analysis, ***Proceedings of Conference in Artificial Intelligence and Statistics***, pp. 437-444.