Topic Segmentation and Evaluation Measures for E-learning based on Domain and Pedagogical Ontology

K.Sathiyamurthy
Research Scholar
DCSE, Anna University chennai
INDIA

T.V.Geetha
Professor
DCSE, Anna University chennai
INDIA

ABSTRACT

Block level segments obtained from e-learning material should form cohesive blocks of e-learning content not only from continuity perspective but also from concept coverage perspective. Therefore in this work two new evaluation measures specificity and proximity have been proposed to evaluate the segments from this concept coverage perspective. These two methods of segmentation have been compared using the standard evaluation measure pk and the two new proposed evaluation measures. Block level text segmentation for e-learning material has been implemented using texttiling method based on domain and pedagogical ontology and this has been compared with hierarchical Latent Dirichlet Allocation based on domain and pedagogical ontology.

General Terms

E-learning, Segmentation, Unsupervised learning;

Keywords

Latent Dirichlet Allocation (LDA), Ontology, Pedagogy;

1. INTRODUCTION

The growth of web based courses for education and training results in challenges to the e-learning systems especially to generate content suitable for effective learning. Learning objects are now the norm for representing cohesive units of learning material. Availability of cohesive segmented content can be used for constructing learning objects and for automatic annotation of learning objects. The constructed learning object from the segmented output can be used by search engines to prepare courseware suited to the learning task. Text segmentation which is the task of dividing text in to topically coherent segments [7] thus forms an important component for automatic content generation in e-learning.

The work described in this paper focuses on text segmentation of technical documents in the computer science domain for the purpose of e-learning. To make the segmentation process effective for e-learning two ontologies to check their domain and pedagogical flow needs to be incorporated. Domain ontology-ACM ontology is based on ACM computing classification a standard of computer science. In general domain ontology is used to indicate the domain specific relation between concepts which can be used in text segmentation for determining topic cohesiveness. However in e-learning in addition to topic cohesiveness the pedagogical role of the text segment is important. Therefore in this work segmentation is also based on pedagogical ontology which identifies the particular context or pedagogical role such as introduction, description, explanation, and example etc. specific to the e-learning context. The use of domain, context (pedagogical) and structural ontology for e-learning has been proposed by Stojanovic, L, Staab, S., and Studer [15]. In their work the ontologies have been used for tagging the content, context and structure of the e-learning materials. However in the work described in these paper concepts from the domain ontology and contextual clues from pedagogical ontology has been used for block segmentation. Segmentation has been carried out using a modified version of Latent Dirichlet Allocation. Latent Dirichlet Allocation is a generative probabilistic model which can be used for unsupervised learning of topic [4]. In LDA, documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. However in the work by K.Sathiyamurthy and T.V.Geetha [14] the basic LDA has been extended with hierarchical structure consisting of root topic mixture, subtopics and topics based on ACM ontology and contextual clue the pedagogical ontology. In this paper two new evaluation measures to determine the suitability for e-learning has been described. These measures essentially evaluate the concept cohesiveness and the pedagogical flow of the text segments.

This paper is organized as follows. The next section explains related work done for text segmentation and LDA based approaches. Section III explores the hierarchical LDA model with domain and pedagogical ontology for e-learning content segmentation. Section IV describes the evaluation measures followed by experimental results with discussion. The section VI deals with conclusion and future direction of this work.

2. RELATED WORK

Text segmentation is an important component of any language processing tasks. An important approach is the texttiling algorithm developed by Martin A.Hearst [8] describes a paragraph-level model of discourse structure based on the notion of subtopic shift. Supervised approaches to coherently segment text has been described by D. Beeferman, A. Berger, and J. Lafferty [2]. This approach incrementally builds an exponential model to extract features that are then associated with the presence of boundaries in the labeled training text. Unsupervised approach to text segmentation include the topic modeling approach. The LDA topic modeling has been used for text segmentation. This approach in addition to determining boundaries of semantically coherent segments also determines the topic associated with each segments[7]. Basic model of LDA is the three level hierarchical Bayesian model proposed by David M. Blei ,Andrew Y. Ng and Michael I. Jordan [4] which essentially model documents as a mixture of topics where each topic is considered as a multinomial distribution of over words in a vocabulary[4]. Multinomial distribution of topics of a document is identified from a dirichlet distribution. This model repeatedly samples a word to the topic from this multinomial. Efficient approximate inference techniques based on variational methods and EM algorithm for parameter estimation has been discussed in this paper. For text segmentation Hemant Misra, François Yvon and Joemon M. Jose used the LDA to determine the topic distribution of a segment. The fewer topics in a segment the more coherent it is. Another approach to text segmentation using LDA has been described by M. Mahdi Shafiei and Evangelos E. Milio[10] where in addition to using LDA for determining topic distribution over words they also use a hierarchical structure based on predefined number of topics and supertopics to determine correlation between word topics. A work on comparative study of mixture models for automatic topic segmentation of multiparty dialogues was attempted by Maria Georgescul,Alexander Clark and Susan Armstrong [10]. In this work a block level Topic segmentation on multi-party meeting recording transcript using LDA was done. Text Segmentation with LDA-Based Fisher Kernel was proposed by Qi Sun, Runxin Li, Dingsheng Luo and XihongWu,[11]. In this work Latent Dirichlet allocation (LDA) is employed to compute semantic distribution of words and semantic similarity is measured by the Fisher kernel method and segments are identified using dynamic programming. However all LDA based methods used for text segmentation considered only a bag of words to describe topics. The use of an ontology to bring about correlation between topics using hierarchical LDA was proposed by K.Sathiyamurthy and T.V.Geetha[14].

3. David M. Blei ,Thomas L. Griffiths, Michael I. Jordan and Joshua B. Tenenbaum [5]. In hLDA, each document is assigned to a path through the topic tree, and each word in a given document is assigned to a topic at one of the levels of that path. A work on hierarchical Panchinko allocation model using DAG-structured mixture model was attempted by (Li & McCallum, 2006) [15]. PAM is a family of generative models in which words are generated by a directed acyclic graph (DAG) consisting of distributions over words and distributions over other nodes. Another work on Pachinko allocation model representing nested hierarchy of topics with topical word distributions shared among several topics was attempted by David Mimno and Andrew McCallum [6]. Precision and recall are the most popular means of performance evaluation for most text processing and application tasks. In the paper on statistical models for text segmentation have by D. Beeferman, A. Berger, and J. Lafferty (1999) point out the short comings of precision and recall measures for text segmentation and have instead proposed an error metric (Pk) to evaluate the segmentation results. Pk is defined as the probability that two segments which are drawn randomly from a document are incorrectly identified as same segment. The lower the value of pk better is the segmentation. WindDiff is another evaluation measure which evaluates segments by moving a sliding window across the text and counts the number of times the referred segment boundaries are different from the determined segment boundaries within the window. Lower values of windDiff mean better segmentation.

In this paper the focus is on the evaluation of the text segmentation of e-learning material with a view to converting this text segments to learning objects in future. From this perspective the placement of the text segment as belonging to a topic in the domain ontology is important. In addition these text segments need to also belong to a specific pedagogical role. Therefore the use of domain ontology and pedagogical ontology for text segmentation was described by K.Sathiyamurthy and T.V.Geetha[14]. This paper describes two new evaluation measures to determine the effectiveness of segmentation for e-learning.

3. HIERARCHICAL LDA MODEL WITH DOMAIN AND PEDAGOGICAL ONTOLOGY

The hierarchical LDA model for topic segmentation of e-learning material using domain and pedagogical ontology proposed in paper[14] is depicted in figure 1. The boxes in the figure 1 illustrates plate notation [4]. Plate notation is a standard way of illustrating probabilistic models with repeated sampling steps. The abbreviation for the symbols used in figure1 are

α_{DC} - Dirchlet distribution for the corpus
θ - Root topic mixture
π - Super- topic mixture of the document
φ - Sub-topic mixture of the block
α_{STM} - Dirchlet distribution for the sub-topic mixture of the block
α_{con} – Dirchlet distribution for the pedagogical cue word vocabulary
α_{Voc} – Dirichlet distribution for the domain word vocabulary

The root topic mixture (θ) for the e-learning material corpus is restricted to computer science domain. The topics for root topic mixture is taken from level 1 of ACM classification ontology consisting of General literature, Hardware, Computer system organization, Software, Data, Theory of computation, Mathematics of computing, Information systems, Computing methodologies, Computer applications and Computing milieux. Super topic mixture (π) for the given document is extracted from level 2 of the ACM classification ontology for the corresponding topic of the root topic mixture. Since block level segmentation is performed, for the given block in the document sub-topic mixtures are determined for the corresponding super-topic. From the sub-topic mixture the topic which has high probability is assigned to the block. If consecutive blocks has same topic all blocks are combined together to form same segment. If adjacent blocks differ with different topic or with different cue-word from the pedagogical vocabulary then it is made as a separate segment. The segmentation process for the

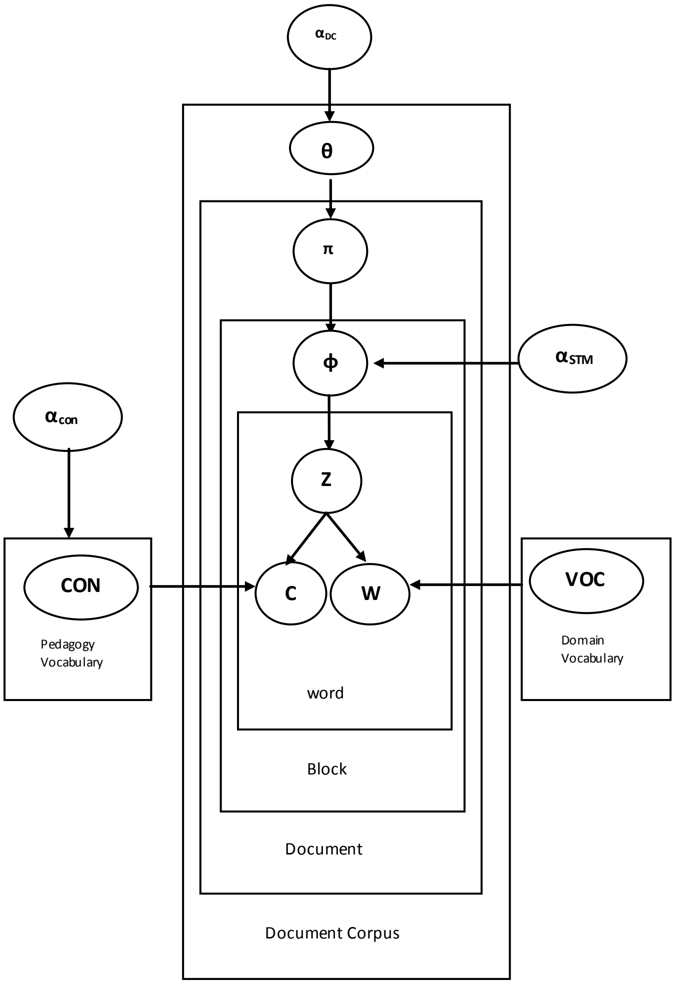

Fig1: Hierarchical LDA based on Domain and Pedagogical Ontology

proposed work is given below:

1. Choose B ~ Poisson(μ) : number of blocks in the document
2. Choose θ ~ Dir($\alpha_{DC)}$
3. For each document d, sample a distribution π over super-topics from the root-topic mixtures.
4. For each block B, sample a distribution over sub-topics from the super-topic mixtures of the document.
5. For each domain-word(W) and cue_word(C)
   (a) Choose a topic Z_n ~ Multinomial(φ)
   (b) Choose a domain word W_n from $P(W_n / Z_n$, voc), a multinomial probability conditioned on Z_n and domain vocabulary.

   (c) Choose a pedagogical cue_word C_n from $P(C_n / Z_n$, con), a multinomial probability conditioned on C_n and pedagogical vocabulary.
6. Segments of the blocks are formed based on the maximum log likelihood value of domain topics and pedagogical cue-word.

The hierarchical LDA model with domain and pedagogical ontology essentially tries to segment the text such that the segment is topic cohesive with respect to the domain ontology and concerned with only one pedagogical role. The measures proposed in this work attempts to validate these perspectives.

4. EVALUATION MEASURES FOR TEXT SEGMENTATION

Evaluation measures for text segmentation have already been described in section II. However this paper attempts to define evaluation parameters for text segmentation based on the performance of the segmentation with respect to the two ontologies used for the purpose. Metrics already exist for many ontology based applications including ontology based information retrieval. Metrics for evaluation of ontology based information retrieval was proposed by Diana Maynard, Wim

Peters and Yaoyong Li the precision and recall were augmented considering the semantic distance of the concepts in ontology along with the binary notion of correctness. In this section we describe the evaluation metric defined by D. Beeferman, A. Berger, and J. Lafferty [2]. In addition two new measures specificity and proximity are defined.

4.1 Pk Probability

The pk error metric was introduced by D. Beeferman, A. Berger, and J. Lafferty [2] formalizes the belief that one segmentation algorithm is better than another if it is able to identify that two sentences drawn randomly are correctly identified as belonging to the same document or not belonging to the same document. The value of pk error metric is a real number between zero and one. The segmentation algorithm that correctly assigns boundaries receives a score of zero. In the work with critique and improvement of pk evaluation metric for text segmentation was done by Lev Pevzner Marti A. Hearsty (2002). In their work a new metric for text segmentation with a simple modification to the pk metric called windDiff was proposed. In windDiff the false positives and false negatives encountered in pk metric are eliminated and it also distinguishes the near-miss error and penalizes it to a different amount which pk is unable to capture. However as pk and windDiff both deal with segment boundaries and not with the concept coverage in this work the basic pk metric is considered along with two new measures for evaluation. Pk probability determines that taken two blocks are correctly labeled as being related or being unrelated but it does not address suitability of the segment to the concept in domain ontology and lacks to determine the distance of the topic in the segment with concepts in the ontology. Hence to better of the evaluation measures suited for e-learning systems, two conceptual coverage measures specificity and proximity are proposed as given below

4.2 Specificity

Specificity is defined as the measure that the words from the segment contribute to the particularity of the concept in the domain ontology. Higher values of specificity indicates that the text segments has more terms that match with the terms of correct concept in the ontology and therefore the segment is more specific to the concept in the ontology.

$$\text{Proximity} = \frac{\text{Extracted terms from the block} \cap \text{Annotated conceptual labels of the concept from domain ontology}}{\text{Extracted Terms from the block}}$$

4.3 Proximity

Proximity is defined as the relative distance of the identified segment to the corresponding concept in the ontology by which the segment has been tagged. If the value of proximity is higher, it implies that the extracted segment is in close to the corresponding concept. Proximity determines the closeness of the segment to a concept in domain ontology.

$$D(X,Y) = \sum_{x,y\varepsilon\pi} d(x,y)$$

Where X= $(x_1,\ldots x_{k)}$ is the identified topic segments by the hierarchical LDA model

Y= $(y_1,..y_{k)}$ is the domain ontology based tagged segment

5. EXPERIMENTAL RESULTS

The dataset for this work is the collection of e-learning materials restricted to computer science domain. The articles consist of different subsets adhering to ACM classification ontology. After removal of list of stop words unique words are extracted from the documents. Figure2 below illustrate the topic assignments done for a random block drawn from e-learning corpus using the proposed hierarchical LDA model. The log likelihood values in figure1 depict the probability of different topics assigned for the block using the proposed model. After calculating the log likelihood of all topics the block is assigned to the topic having maximum likelihood value. Accordingly for the random block taken it is assigned to "Internet computing" topic. The baseline method taken to evaluate the proposed system performance is the texttiling[8] segmentation algorithm with domain and pedagogical ontology. A training model with document corpus of 700 e-learning documents related to computer science was created and to evaluate the system performance blocks are taken randomly from 100 documents. To evaluate our segmentation performance, Pk value is considered as the probability that two blocks drawn randomly from a document are incorrectly identified as belonging to the same topic.

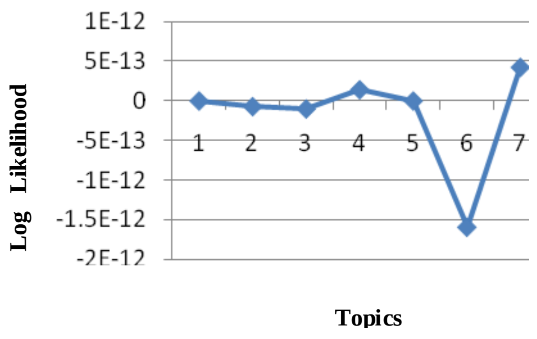

Fig2: Loglikelihood values of Topics

The average Pk probability for the hLDA model with domain and pedagogical ontology is lower than the baseline method indicated in table 1 and in figure3. Lower the Pk value indicates greater the segmentation performance. The evaluation of concept coverage is shown in table 2 and in figure 4. The value of hLDA with domain and pedagogical ontology has higher specificity and proximity compared with the baseline method. As our results shown in figure3 and figure 4 are significantly better than the baseline method it indicates that the system performs well but still the segmentation can be extended by adopting sentences instead of blocks.

Table 1. Pk Probability values of Baseline Method and with hLDA

Method	Pk Probability	
	3-5	6-9
BaseLine	0.35	0.5
hLDA	0.3	0.44

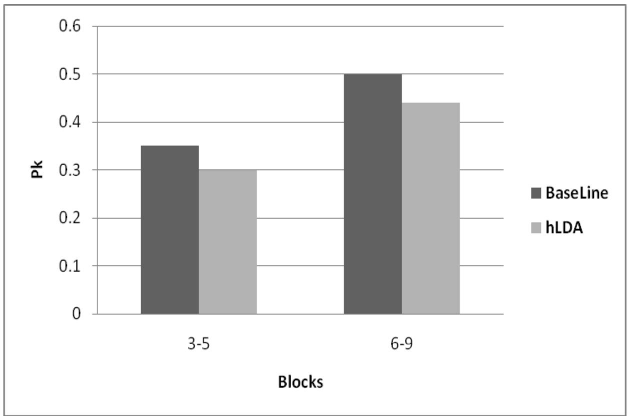

Fig3: Comparison of Pk probability values of Baseline method and hLDA

Table2. Specificity and Proximity values for Baseline and hLDA

Method	Concept coverage Measures	
	Specificity	Proximity
BaseLine	71.42	65.71
hLDA	80.61	78.58

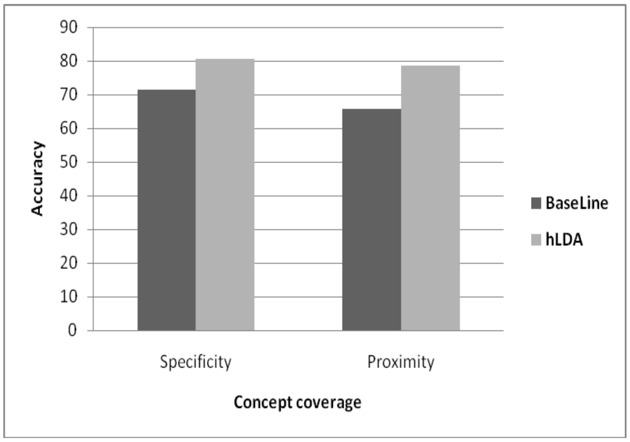

Fig4: Comparison of Specificity and Proximity values for Baseline method and hLDA

6. CONCLUSION

In this work, a hierarchical LDA model for segmenting e-learning materials using domain and context ontology with required ontology concept evaluation measures was proposed. The usage of this hierarchical LDA model makes the segmentation process flexible to accommodate the growth of large e-learning materials. The experimental results indicate that the proposed method performs better than the baseline approach. This work can be extended by adopting sentence level segmentation of e-learning contents compared to block level segmentation and Pachinko allocation model can be adopted to overcome the overlapping topics.

7. REFERENCES

[1] The ACM Computing Classification System–1998 Version valid in 2009, http:/www.acm.org/class/1998/

[2] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. Machine Learning, 34 (1–3):177–210, 1999.

[3] Chaitanya Chemudugunta, Padhraic Smyth, Mark Steyvers, Text Modeling using Unsupervised Topic Models and Concept Hierarchies , CoRR abs/0808.0973: (2008)

[4] David M. Blei ,Andrew Y. Ng ,Michael I. Jordan ,Latent Dirichlet Allocation, Journal of Machine Learning Research, 3:993–1022, 2003.

[5] David M. Blei ,Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, Hierarchical Topic Models and the Nested Chinese Restaurant Process, Advances in Neural Information Processing Systems , NIPS, December 8-13, 2003.

[6] David Mimno Andrew McCallum ,Mixtures of Hierarchical Topics with Pachinko Allocation, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvalis, Oregon, USA, June 20-24, 2007, 633-640

[7] Hemant Misra,François Yvon,Joemon M. Jose, Text Segmentation via Topic Modeling: An Analytical Study,

Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009

[8] Lev Pevzner Marti A. Hearsty, A Critique and Improvement of an Evaluation Metric for Text Segmentation, Computational Linguistics, 28 (1):19–36, 2002

[9] Marti Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics, and 23(1):33–64.

[10] M. Mahdi Shafiei and Evangelos E. Milios A Statistical Model for Topic Segmentation and Clustering, Canadian Conference on AI 2008: 283-295

[11] Maria Georgescul,Alexander Clark & Susan Armstrong, A Comparative Study of Mixture Models for Automatic Topic Segmentation of Multiparty Dialogues, The 3rd International Joint Conference on Natural Language Processing (IJCNLP), Hyderabad, India, January 7-12, 2008.

[12] Qi Sun, Runxin Li, Dingsheng Luo and XihongWu, Text Segmentation with LDA-Based Fisher Kernel, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA.

[13] Sathiyamurthy.K, T.V.Geetha, Association of domain concepts with Educational objectives for E-Learning, ACM compute 2010, Bangalore, India.

[14] Sathiyamurthy.K, T.V.Geetha, Topic Segmentation of E-Learning materials using Domain and Pedagogical Ontology, ICFIT December 2010, Changsa, China.

[15] Stojanovic, L. , Staab, S., and Studer, R., Elearning based on the Semantic Web, World Conference on the WWW and Internet (WebNet), Orlando, Florida, USA, 2001

[16] Takayuki Sekiya, Yoshitatsu Matsuda and Kazunori Yamaguchi, Analysis of Curriculum Structure Based on LDA, International Conference on Education and Information Technology 2009, pp.561-566.

[17] Wei Li weili,Andrew McCallum Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006. ICML 2006: 577-584

[18] Wang Wei, Payam Barnaghi and Andrzej Bargiela, Probabilistic Topic Models for Learning Terminological Ontologies, IEEE Transactions on Knowledge and Data Engineering , 22(7): 1028-1040 (2010)