Broadcast News Story Segmentation Using Manifold Learning on Latent Topic Distributions

Xiaoming Lu[1,2], **Lei Xie**[1*], **Cheung-Chi Leung**[2], **Bin Ma**[2], **Haizhou Li**[2]
[1]School of Computer Science, Northwestern Polytechnical University, China
[2]Institute for Infocomm Research, A*STAR, Singapore
luxiaomingnpu@gmail.com, lxie@nwpu.edu.cn, {ccleung,mabin,hli}@i2r.a-star.edu.sg

Abstract

We present an efficient approach for broadcast news story segmentation using a manifold learning algorithm on latent topic distributions. The latent topic distribution estimated by Latent Dirichlet Allocation (LDA) is used to represent each text block. We employ Laplacian Eigenmaps (LE) to project the latent topic distributions into low-dimensional semantic representations while preserving the intrinsic local geometric structure. We evaluate two approaches employing LDA and probabilistic latent semantic analysis (PLSA) distributions respectively. The effects of different amounts of training data and different numbers of latent topics on the two approaches are studied. Experimental results show that our proposed LDA-based approach can outperform the corresponding PLSA-based approach. The proposed approach provides the best performance with the highest F1-measure of 0.7860.

1 Introduction

Story segmentation refers to partitioning a multimedia stream into homogenous segments each embodying a main topic or coherent story (Allan, 2002). With the explosive growth of multimedia data, it becomes difficult to retrieve the most relevant components. For indexing broadcast news programs, it is desirable to divide each of them into a number of independent stories. Manual segmentation is accurate but labor-intensive and costly. Therefore, automatic story segmentation approaches are highly demanded.

Lexical-cohesion based approaches have been widely studied for automatic broadcast news story segmentation (Beeferman et al., 1997; Choi, 1999; Hearst, 1997; Rosenberg and Hirschberg, 2006; Lo et al., 2009; Malioutov and Barzilay, 2006; Yamron et al., 1999; Tur et al., 2001). In this kind of approaches, the audio portion of the data stream is passed to an automatic speech recognition (ASR) system. Lexical cues are extracted from the ASR transcripts. Lexical cohesion is the phenomenon that different stories tend to employ different sets of terms. Term repetition is one of the most common appearances.

These rigid lexical-cohesion based approaches simply take term repetition into consideration, while term association in lexical cohesion is ignored. Moreover, polysemy and synonymy are not considered. To deal with these problems, some topic model techniques which provide conceptual level matching have been introduced to text and story segmentation task (Hearst, 1997). Probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) is a typical instance and used widely. PLSA is the probabilistic variant of latent semantic analysis (LSA) (Choi et al., 2001), and offers a more solid statistical foundation. PLSA provides more significant improvement than LSA for story segmentation (Lu et al., 2011; Blei and Moreno, 2001).

Despite the success of PLSA, there are concerns that the number of parameters in PLSA grows linearly with the size of the corpus. This makes PLSA not desirable if there is a considerable amount of data available, and causes serious over-fitting problems (Blei, 2012). To deal with this issue, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been proposed. LDA has been proved to be effective in many segmentation tasks (Arora and Ravindran, 2008; Hall et al., 2008; Sun et al., 2008; Riedl and Biemann, 2012; Chien and Chueh, 2012).

Recent studies have shown that intrinsic dimensionality of natural text corpus is significantly lower than its ambient Euclidean space (Belkin and Niyogi, 2002; Xie et al., 2012). Therefore,

[*]corresponding author

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 190–195,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

Laplacian Eigenmaps (LE) was proposed to compute corresponding natural low-dimensional structure. LE is a geometrically motivated dimensionality reduction method. It projects data into a low-dimensional representation while preserving the intrinsic local geometric structure information (Belkin and Niyogi, 2002). The locality preserving property attempts to make the low-dimensional data representation more robust to the noise from ASR errors (Xie et al., 2012).

To further improve the segmentation performance, using latent topic distributions and LE instead of term frequencies to represent text blocks is studied in this paper. We study the effects of the size of training data and the number of latent topics on the LDA-based and the PLSA-based approaches. Another related work (Lu et al., 2013) is to use local geometric information to regularize the log-likelihood computation in PLSA.

2 Our Proposed Approach

In this paper, we propose to apply LE on the LDA topic distributions, each of which is estimated from a text block. The low-dimensional vectors obtained by LE projection are used to detect story boundaries through dynamic programming. Moreover, as in (Xie et al., 2012), we incorporate the temporal distances between block pairs as a penalty factor in the weight matrix.

2.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a generative probabilistic model of a corpus. It considers that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over terms.

In LDA, given a corpus $D = \{d_1, d_2, \ldots, d_M\}$ and a set of terms $W = (w_1, w_2, \ldots, w_V)$, the generative process can be summarized as follows:

1) For each document *d*, pick a multinomial distribution θ from a Dirichlet distribution parameter α, denoted as $\theta \sim Dir(\alpha)$.

2) For each term *w* in document *d*, select a topic z from the multinomial distribution θ, denoted as $z \sim Multinomial(\theta)$.

3) Select a term *w* from $P(w|z, \beta)$, which is a multinomial probability conditioned on the topic.

An LDA model is characterized by two sets of prior parameters α and β. $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ represents the Dirichlet prior distributions for each *K* latent topics. β is a $K \times V$ matrix, which defines the latent topic distributions over terms.

2.2 Construction of weight matrix in Laplacian Eigenmaps

Laplacian Eigenmaps (LE) is introduced to project high-dimensional data into a low-dimensional representation while preserving its locality property. Given the ASR transcripts of *N* text blocks, we apply LDA algorithm to compute the corresponding latent topic distributions $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ in $\mathbb{R}^K$, where *K* is the number of latent topics, namely the dimensionality of LDA distributions.

We use *G* to denote an *N*-node (*N* is number of LDA distributions) graph which represents the relationship between all the text block pairs. If distribution vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ come from the same story, we put an edge between nodes *i* and *j*. We define a weight matrix **S** of the graph *G* to denote the cohesive strength between the text block pairs. Each element of this weight matrix is defined as:

$$s_{ij} = cos(\mathbf{x}_i, \mathbf{x}_j)\mu^{|i-j|}, \quad (1)$$

where $\mu^{|i-j|}$ serves the penalty factor for the distance between *i* and *j*. μ is a constant lower than 1.0 that we tune from a set of development data. It makes the cohesive strength of two text blocks dramatically decrease when their distance is much larger than the normal length of a story.

2.3 Data projection in Laplacian Eigenmaps

Given the weight matrix **S**, we define **C** as the diagonal matrix with its element:

$$c_{ij} = \sum\nolimits_{i=1}^{K} s_{ij}. \quad (2)$$

Finally, we obtain the Laplacian matrix ***L***, which is defined as:

$$\boldsymbol{L} = \boldsymbol{C} - \boldsymbol{S}. \quad (3)$$

We use $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$ ($\mathbf{y}_i$ is a column vector) to indicate the low-dimensional representation of the latent topic distributions **X**. The projection from the latent topic distribution space to the target space can be defined as:

$$f : \mathbf{x}_i \Rightarrow \mathbf{y}_i. \quad (4)$$

A reasonable criterion for computing an optimal mapping is to minimize the objective as follows:

$$\sum_{i=1}^{K}\sum_{j=1}^{K} \parallel \mathbf{y}_i - \mathbf{y}_j \parallel^2 s_{ij}. \quad (5)$$

Under this constraint condition, we can preserve the local geometrical property in LDA distributions. The objective function can be transformed as:

$$\sum_{i=1}^{K}\sum_{j=1}^{K}(\mathbf{y}_i - \mathbf{y}_j)s_{ij} = tr(\mathbf{Y}^T\mathbf{L}\mathbf{Y}). \qquad (6)$$

Meanwhile, zero matrix and matrices with its rank less than K are meaningless solutions for our task. We impose $\mathbf{Y}^T\mathbf{L}\mathbf{Y} = \mathbf{I}$ to prevent this situation, where $\mathbf{I}$ is an identity matrix. By the Reyleigh-Ritz theorem (Lutkepohl, 1997), the solution can obtained by the Q smallest eigenvalues of the generalized eigenmaps problem:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{y} = \lambda\mathbf{X}\mathbf{C}\mathbf{X}^T\mathbf{y}. \qquad (7)$$

With this formula, we calculate the mapping matrix $\mathbf{Y}$, and its row vectors $\mathbf{y}'_1, \mathbf{y}'_2, \ldots, \mathbf{y}'_Q$ are in the order of their eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_Q$. $\mathbf{y}'_i$ is a Q-dimensional (Q<K) eigenvectors.

2.4 Story boundary detection

In story boundary detection, dynamic programming (DP) approach is adopted to obtain the global optimal solution. Given the low-dimensional semantic representation of the test data, an objective function can be defined as follows:

$$\Im = \sum_{t=1}^{N_s}\Big(\sum_{i,j\in Seg_t} \parallel \mathbf{y}_i - \mathbf{y}_j \parallel^2\Big), \qquad (8)$$

where $\mathbf{y}_i$ and $\mathbf{y}_j$ are the latent topic distributions of text blocks i and j respectively, and $\parallel \mathbf{y}_i - \mathbf{y}_j \parallel^2$ is the Euclidean distance between them. Seg_t indicates these text blocks assigned to a certain hypothesized story. N_s is the number of hypothesized stories.

The story boundaries which minimize the objective function $\Im$ in Eq.(8) form the optimal result. Compared with classical local optimal approach, DP can more effectively capture the smooth story shifts, and achieve better segmentation performance.

3 Experimental setup

Our experiments were evaluated on the ASR transcripts provided in TDT2 English Broadcast news corpus[1], which involved 1033 news programs. We separated this corpus into three non-overlapping sets: a training set of 500 programs for parameter estimation in topic modeling and LE, a development set of 133 programs for empirical tuning and a test set of 400 programs for performance evaluation.

In the training stage, ASR transcripts with manually labeled boundary tags were provided. Text streams were broken into block units according to the given boundary tags, with each text block being a complete story. In the segmentation stage, we divided test data into text blocks using the time labels of pauses in the transcripts. If the pause duration between two blocks last for more than 1.0 sec, it was considered as a boundary candidate. To avoid the segmentation being suffered from ASR errors and the out-of-vocabulary issue, phoneme bigram was used as the basic term unit (Xie et al., 2012). Since the ASR transcripts were at word level, we performed word-to-phoneme conversion to obtain the phoneme bigram basic units. The following approaches, in which DP was used in story boundary detection, were evaluated in the experiments:

- PLSA-DP: PLSA topic distributions were used to compute sentence cohesive strength.
- LDA-DP: LDA topic distributions were used to compute sentence cohesive strength.
- PLSA-LE-DP: PLSA topic distributions followed by LE projection were used to compute sentence cohesive strength.
- LDA-LE-DP: LDA topic distributions followed by LE projection were used to compute sentence cohesion strength.

For LDA, we used the implementation from David M. Blei's webpage[2]. For PLSA, we used the Lemur Toolkit[3].

F1-measure was used as the evaluation criterion.We followed the evaluation rule: a detected boundary candidate is considered correct if it lies within a 15 sec tolerant window on each side of a reference boundary. A number of parameters were set through empirical tuning on the developent set. The penalty factor was set to 0.8. When evaluating the effects of different size of the training set, the number of latent topics in topic modeling process was set to 64. After the number of latent topics was fixed, the dimensionality after LE projection was set to 32. When evaluating the effects of different number of latent topics in topic modeling computation, we fixed the size of the training set to 500 news programs and changed the number of latent topics from 16 to 256.

4 Experimental results and analysis

4.1 Effect of the size of training dataset

We used the training set from 100 programs to 500 programs (adding 100 programs in each step) to e-

[1] http://projects.ldc.upenn.edu/TDT2/

[2] http://www.cs.princeton.edu/ blei/lda-c/

[3] http://www.lemurproject.org/

valuate the effects of different size of training data in both PLSA-based and LDA-based approaches. Figure 1 shows the results on the development set and the test set.

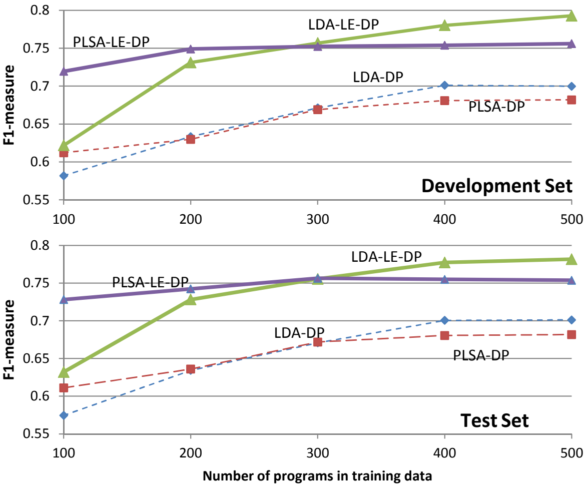

Figure 1: Segmentation performance with different amounts of training data

LDA-LE-DP approach achieved the best result (0.7927 and 0.7860) on both the development and the test sets, when there were 500 programs in the training set. This demonstrates that LDA model and LE projection used in combination is excellent for the story segmentation task. The LE projection applied on the latent topic representations made relatively 9.88% and 10.93% improvement over the LDA-based approach and the PLSA-based approach, respectively on the test set. We can reveal that employing LE on PLSA and LDA topic distributions achieves much better performance than the corresponding approaches without using LE.

We have compared the performances between PLSA and LDA. We found that when the training data size was small, PLSA performed better than LDA. Both PLSA-based and LDA-based approaches got better with the increase in the size of the training data set. All the four approaches had similar performances on the development set and the test set.

With the increase in the size of the training data, the LDA-based approaches were improved dramatically. They even outperformed the PLSA-based approaches when the training data contained more than 300 programs. This may be attributed to the fact that LDA needs more training data to estimate the parameters. When the training data is not enough, its parameters estimated in the training stage is not stable for the development and the test data. Moreover, compared with PLSA, the parameters in LDA do not grow linearly with the size of the corpus.

4.2 Effect of the number of latent topics

We evaluated the F1-measure of the four approaches with different number of latent topics prior to LE projection. Figure 2 shows the corresponding results.

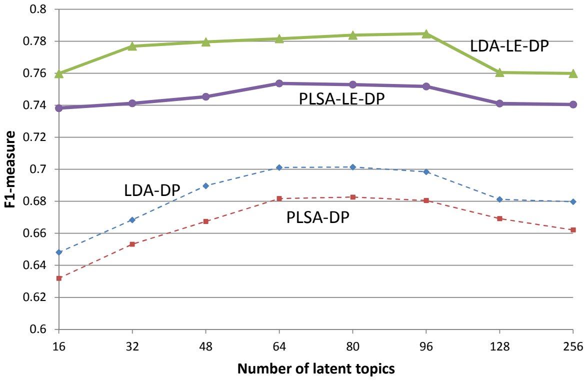

Figure 2: Segmentation performance with different numbers of latent topics

The best performances (0.7816-0.7847) were achieved at the number of latent topics between 64 and 96. When the number of latent topics was increased from 16 to 64, F1-measure increased. When the number of latent topics was larger than 96, F1-measure decreased gradually. We found that the best results were achieved when the number of topics was close to the real number of topics. There are 80 manually labeled main topics in the test set.

We observe that LE projection makes the topic model more stable with different numbers of latent topics. The best and the worst performances differed by relatively 9.12% in LDA-DP and 7.97% in PLSA-DP. However, the relative difference of 2.79% and 2.46% were observed in LDA-LE-DP and PLSA-LE-DP respectively.

5 Conclusions

Our proposed approach achieves the best F1-measure of 0.7860. In the task of story segmentation, we believe that LDA can avoid data overfitting problem when there is a sufficient amount of training data. This is also applicable to LDA-LE-LP. Moreover, we find that when we apply LE projection to latent topic distributions, the segmentation performances become less sensitive to the predefined number of latent topics.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61175018), the Natural Science Basic Research Plan of Shaanxi Province (2011JM8009) and the Fok Ying Tung Education Foundation (131059).

References

J. Allan. 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publisher, Norwell, MA.

Doug Beeferman, Adam Berger, and John Lafferty. 1997. *A Model of Lexical Attraction and repulsion*. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pp.373-380.

Freddy Y. Y. Choi. 2000. *Advances in Domain Independent Linear Text Segmentation*. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL)*, pp.26-33.

Thomas Hofmann. 1999. *Probabilistic Latent Semantic Indexing*. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp.20-57.

Mimi Lu, Cheung-Chi Leung, Lei Xie, Bin Ma, Haizhou Li. 2011. *Probabilistic Latent Semantic Analysis for Broadcast New Story Segmentation*. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp.1109-1112.

David M. Blei. 2012. *Probabilistic topic models*. Communication of the ACM, vol. 55, pp.77-84.

David M. Blei, Andrew Y. Ng, Michael I. Jordan. 2003. *Latent Dirichlet Allocation*. the Journal of Machine Learning Research, vol. 3, pp.993-1022.

Marti A. Hearst. 1997. *TextTiling: Segmenting Text into Multiparagraph subtopic passages*. Computational Liguistic, vol. 23, pp.33-64.

Gokhan Tur, Dilek Hakkani-Tur, Andreas Stolcke, Elizabeth Shriberg. 2001. *Integrating Prosodic and Lexicial Cues for Automatic Topic Segmentation*. Computational Liguistic, vol. 27, pp.31-57.

Andrew Rosenberg and Julia Hirschberg. 2006. *Story Segmentation of Broadcast News in English, Mandarin and Aribic*. In *Proceedings of the 7th North American Chapter of the Association for Computational Linguistics Conference (NAACL)*, pp.125-128.

David M. Blei and Pedro J. Moreno. 2001. *Topic Segmentation with An Aspect Hidden Markov Model*. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrival (SIGIR)*, pp.343-348.

Wai-Kit Lo, Wenying Xiong, Helen Meng. 2009. *Automatic Story Segmentation Using a Bayesian Decision Framwork for Statistical Models of Lexical Chain Feature*. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.357-364.

Igor Malioutov and Regina Barzilay. 2006. *Minimum Cut Model for Spoken Lecture Segmenation*. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.25-32.

Freddy Y. Y. Choi, Peter Wiemer-Hastings, Juhanna Moore. 2001. *Latent Semantic Analysis for Text Segmentation*. In *Proceedings of the 2001 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pp.109-117.

Rachit Arora and Balaraman Ravindran. 2008. *Latent Dirichlet Allocation Based Multi-document Summarization*. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data (AND)*, pp.91-97.

David Hall, Daniel Jurafsky, Christopher D. Manning. 2008. *Latent Studying the History Ideas Using Topic Models*. In *Proceedings of the 2008 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pp.363-371.

Qi Sun, Runxin Li, Dingsheng Luo, Xihong Wu. 2008. *Text Segmentation with LDA-based Fisher Kernel*. In *Proceedings of the 46th Annual Meeting of the Assocation for Computational Linguistics on Human Language Technologies (HLT-ACL)*, pp.269-272.

Mikhail Belkin and Partha Niyogi. 2002. *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*. Neural Computation, vol. 15, pp.1383-1396.

Lei Xie, Lilei Zheng, Zihan Liu and Yanning Zhang. 2012. *Laplacian Eigenmaps for Automatic Story Segmentation of Broadcast News*. IEEE Transaction on Audio, Speech and Language Processing, vol. 20, pp.264-277.

Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. 2008. *Modeling Hidden Topics on Document Manifold*. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pp.911-120.

Xiaoming Lu, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li. 2013. *Broadcast News Story Segmentation Using Latent Topics on Data Manifold*. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. 1999. *A Hidden Markov Model Approach to Text Segmenation and Event Tracking*. In *Proceedings of the 1999 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.333-336.

Martin Riedl and Chris Biemann. 2012. *Text Segmentation with Topic Models*. the Journal for Language Technology and Computational Linguistics, pp.47-69.

P. Fragkou , V. Petridis , Ath. Kehagias. 2002. *A Dynamic Programming algorithm for Linear Text Story Segmentation*. the Joural of Intelligent Information Systems, vol. 23, pp.179-197.

H. Lutkepohl. 1997. *Handbook of Matrices*. Wiley, Chichester, UK.

Jen-Tzung Chien and Chuang-Hua Chueh. 2012. *Topic-Based Hieraachical Segmentation*. IEEE Transaction on Audio, Speech and Language Processing, vol. 20, pp.55-66.