UNSUPERVISED BROADCAST NEWS STORY SEGMENTATION USING DISTANCE DEPENDENT CHINESE RESTAURANT PROCESSES

Chao Yang, Lei Xie and Xiangzeng Zhou

Shaanxi Provincial Key Laboratory of Speech and Image Information Processing
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
yang.placebo@gmail.com, lxie@nwpu.edu.cn, xzzhou@nwpu-aslp.org

ABSTRACT

Traditional unsupervised broadcast news story segmentation approaches have to set the segmentation number manually, while this number is often unknown in real-world applications. In this paper, we solve this problem by modeling the generative process of stories as distance dependent Chinese restaurant process (dd-CRP) mixtures. We cut a news program into fixed-size text blocks and consider these blocks in the same story are generated from a story-specific topic. Specifically, we add a dd-CRP prior which has an essential bias that the blocks' topic is more likely to be the same with the nearby blocks. Subsequently, story boundaries can be found by detecting the changes of topics. Experiments show that our approach outperforms both supervised and unsupervised approaches and the segmentation number can be automatically learned from data.

1. INTRODUCTION

Story segmentation refers to partitioning a stream of multimedia, e.g., a broadcast news program, into segments each belonging to a coherent story [1]. For tasks like broadcast news retrieval or topic tracking, it is an important prerequisite to segment each independent story from the entire news program. Manual segmentation is accurate but infeasible due to the explosive growth of the multimedia data. Therefore, automatic story segmentation approaches are highly in demand.

For this task, some supervised methods have been proposed using lexical and prosodic cues [2, 3, 4]. Specifically, topic-model-based approaches have drawn much interest recently. These approaches use a training corpus to learn topic models from texts and then map the term-frequency representation into the topic representation [5, 6, 7, 8]. The topic representation is used for story segmentation by some boundary detection methods. Although these approaches achieve superior segmentation results, they require a labeled training corpus which is difficult to obtain in real-world applications. Thus unsupervised methods are highly desired.

Unsupervised story segmentation systems rely upon the notion of lexical cohesion: each well-formed segment should be generated from a consistent and compact lexical distribution [9]. As one of the earliest approaches introducing this idea, TextTiling [10, 11, 12] measures adjacent sentence lexical similarities and identifies boundaries at local similarity minima. To reach a global optimum, MinCut [13] uses a dynamic programming method to detect story boundaries. These methods use handcrafted metrics such as cosine similarity or cross entropy for quantifying lexical cohesion, which may not generalize well across multiple datasets. Moreover, some parameters must be tuned individually for different datasets in order to achieve good segmentation results. To overcome this drawback, Eisenstein *et al.* [14] develop a Bayesian unsupervised method from the probability views, namely BayesSeg. In their models, words in each topic segment are drawn from a multinomial distribution associated with the segment. Maximizing the observation likelihood in such a model yields a segmentation.

However, all these methods have a restriction that the exact number of segments needs to be set manually. Note that, to count how many stories exist in a document, users need to browse the whole file from start to end, the effort of which almost equals to that of locating all story boundaries directly. On the other hand, if we choose a story number very different from the real one, the segmentation results may not be satisfactory.

In this paper, we solve this problem by developing an unsupervised non-parametrical Bayesian model [15] for story segmentation. The news program is cut into fixed-size text blocks. We consider that the blocks in the same story are generated from a story-specific topic just like that in BayesSeg [14]. Importantly, we add a distance dependent Chinese restaurant process (dd-CRP) prior which has an essential bias that a block's topic is more likely to be same with the nearby blocks. Subsequently, story boundaries can be found by detecting the changes of topics. The dd-CRP model [16] is a general form of Chinese restaurant process often treated as a description of Dirichlet process [17]. This prior relaxes the CRP's assumption of exchangeability, which the sequential data like text and audio don't have, and has been successfully applied in image segmentation [18] and 3D objects segmentation [19]. Meanwhile, our approach can be viewed as an extension of the BayesSeg approach without setting the segment number manually. Experiments show that the proposed approach outperforms the BayesSeg approach, the TextTiling approach and a recent supervised approach based on probabilistic latent semantic analysis (PLSA) [8].

2. MODEL

We propose to develop a method to segment a new broadcast without setting the segmentation number. A story is cut into small blocks where the terms in each block are exchangeable but the blocks are not. We assume that the blocks in the same story are generated from the same topic which is a multinomial distribution. We use distance dependent Chinese restaurant process as the prior of the partitions of topic. With the help of a sampling method, we can get the posterior of the partitions conditioned on the observed data. Then the changes of topic form the segmentation boundaries.

2.1. Chinese Restaurant Process

The Chinese restaurant process (CRP) [20] is a process that generates a distribution over partitions. It is described as follows. There is

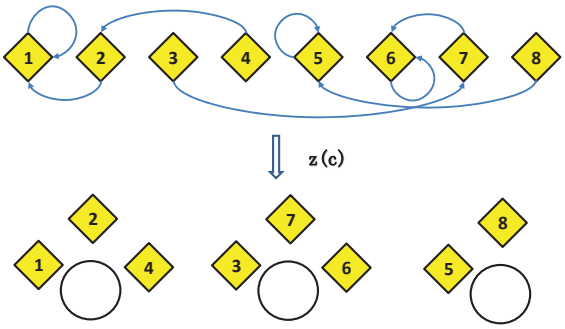

Fig. 1. An illustration of dd-CRP. The diamonds denote customers and the circles denote tables. The lines with arrow are the links between customers. In this example, the first customer c_1 links to itself. Customer c_2 links to c_1, c_4 links to c_2 and no other costumers link to them. Hence c_1, $c2$ and $c4$ sit at the same table.

a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers. The first customer comes in and sits at a table. The ith customer either sits at an already occupied table with probability proportional to the number of customers already sitting there or opens up a new table with probability proportional to a parameter α. After all customers have sat down, the tables define a partition.

For the exchangeability [17], we can draw each customer's table assignment z_i by assuming he/she is the last customer to sit down. Let n_k denote the number of customers sitting at table k and K denotes the number of occupied tables. This can be written as:

$$p(z_i = k|z_{1:(i-1)}, \alpha) \propto \begin{cases} n_k & \text{if } k \leq K \\ \alpha & \text{if } k = K+1. \end{cases} \quad (1)$$

2.2. Distance Dependent CRP

The distance dependent Chinese restaurant process (dd-CRP) [16] is a generalization of CRP that allows for a non-exchangeable distribution on partitions. Rather than representing a partition by customers assigned to tables, the dd-CRP models customers linking to other customers or themselves. Then the customers linked are considered sitting at the same table. Fig. 1 shows an example. The diamonds denote customers and the circles denote tables. The lines with arrow are the links between customers. In this example, the first customer c_1 links to itself. Customer c_2 links to c_1, c_4 links to c_2 and no other costumers link to them. Hence c_1, $c2$ and $c4$ sit at the same table. The dd-CRP can be formulated as

$$p(c_i = j|D, f, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{if } j = i \end{cases} \quad (2)$$

where D denotes the set of all distance measurements between customers, d_{ij} is the distance measurement between customers i and j and $f(d)$ is a non-increasing decay function.

We use the sequential distance, which is $d_{ij} = i - j$ for $j < i$ and $d_{ij} = \inf$ for $j > i$ and use a window decay function $f(d) = \mathbb{1}[d <= a]$, where a is the window size. This puts a prior on the segmentation that the current block's topic is either the same with previous n blocks' or a new one. Note that this prior does not restrict the size of segments, because any two reachable blocks will be in the same segment.

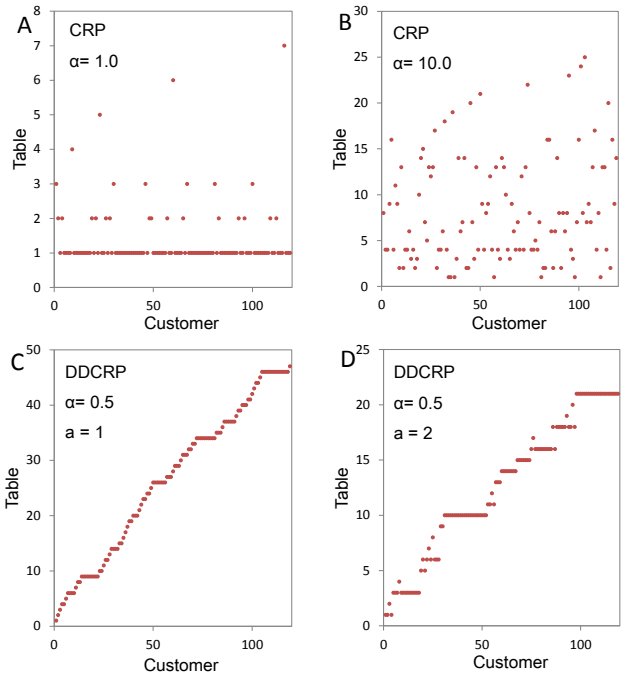

Fig. 2. Draws from CRP and dd-CRP. We can see that the draws of CRP (A and B) are dispersive, because CRP has no good property of segmentation. The draws of dd-CRP (C and D) form clear clusters according to the adjacent area in sequence data.

2.3. Story Segmentation

As mentioned at the beginning of this section, we view the observations of news program as draws of a generative process. Firstly, the topic of each block is drawn. Then the word in each blocks is drawn from its topic-specific distribution. If each block contains enough terms which provide adequate statistical information, the distribution of topic is not important. It is just introduced to make the posterior inference feasible. We could use a Dirichlet distribution or a CRP in order to find a suitable topic number. But in our task, the story with unknown size may be too short so that we can only cut the news into small blocks. In this situation, we will just get a result of clustering but not segmentation using the CRP prior. So we want to put a prior that the topic of current block is more likely to have the same cluster with the nearby blocks. The dd-CRP model is a good choice having this bias, as shown in Fig. 2.

We treat text blocks as customers and the topics as the tables in dd-CRP. According to the definition of dd-CRP, we can't draw the table directly. Instead, we draw the customer assignments for each block. Then the blocks linked together are viewed as being generated from the same topic. All topics are drawn i.i.d. from a base distribution G_0. Here the topic is a multinomial distribution and G_0 is a Dirichlet distribution. The full generative process for the news program is as follows:

1. For each block i, sample its customer assignment c_i $\sim$dd-CRP(α, f, D).
2. Map the customer assignment c_i to the table assignment (topic assignment) z_i. For each table k, sample parameters $\phi_k \sim G_0$.
3. For each block i, independently sample the observed feature x_i $\sim$Mult$(.|\phi_{z_i})$.

3. INFERENCE

We have built the probability generative process of the news program. Currently, the key problem that we need to solve is to compute the posterior distribution of the latent topic variables conditioned on the observed term frequency features:

$$p(c_{1:N}|x_{1:N},\theta,G_0)=\frac{p(c_{1:N},x_{1:N}|\theta,G_0)}{\sum_{c'_{1:N}}p(c'_{1:N},x_{1:N}|\theta,G_0)} \quad (3)$$

where θ is the short form for the parameters α, f, D in the dd-CRP prior ($\theta=[\alpha,f,D]$).

Unfortunately, this distribution is intractable to directly evaluate due to the combinatorial sum in the denominator. Instead of determined inference, we use Gibbs sampling which iteratively samples each latent variable conditioned on the others and the observations. In our model the latent variables are c_i and ϕ_k. Since we only care about the topic assignment and the base distribution G_0 is a conjugate prior, we marginalize ϕ_k analytically and get the collapsed Gibbs sampler [21]:

$$p(c_i|c_{-i},x_{1:N},\theta,G_0)\propto p(c_i|\theta)p(x_{1:N}|z(c_{1:N}),G_0). \quad (4)$$

The first term is the dd-CRP prior given in Eq. (2) and the second term is the likelihood which is factorized according to the table index. The likelihood term is:

$$p(x_{1:N}|z(c_{1:N}),G_0)=\prod_{k=1}^{|z(c)|}p(x_{z^k(c)}|G_0) \quad (5)$$

Here $|z(c)|$ is the number of tables and $z^k(c)$ is the set of customer indices that are assigned to table k. Correspondingly, $x_{z^k(c)}$ are the features of blocks whose table index is k. Because of this factorization, we do not need to compute the terms that are not influenced when we reassign c_i.

To sample c_i, we first remove the links of customer i and then we reassign c_i. Only two cases need to be considered – the reassignment joins two tables or causes no change in the partition, as shown in Fig. 3.

In the first case, let l and m denote the indices of joined tables. The two terms of z^l and z^m will be changed into one which is $z^l(c)\cup z^m(c)$. The other terms are not changed. So the likelihood term will be:

$$p(x_{z^l(c)\cup z^m(c)}|G_0)\prod_{k\neq m,l}p(x_{z^k(c)}|G_0). \quad (6)$$

In the second case, all terms remain unchanged. By removing the same factor $\prod_{k\neq m,l}p(x_{z^k(c)}|G_0)$,we get the details of the Gibbs sampler as follows:

$$p(c_i|c_{-i},x_{1:N},\theta,G_0)\propto\begin{cases}p(c_i|\theta)\Delta(x,z,G_0) & c_i \text{ joins } l \text{ and } m\\ p(c_i|\theta) & \text{otherwise}\end{cases} \quad (7)$$

where

$$\Delta(x,z,G_0)=\frac{p(x_{z^l(c)\cup z^m(c)}|G_0)}{p(x_{z^l(c)}|G_0)p(x_{z^m(c)}|G_0)} \quad (8)$$

The k-th factor term in the likelihood is given by

$$p(x_{z^k(c)}|G_0)=\int p(x_{z^k(c)}|\phi_k)p(\phi_k|G_0)d\phi_k. \quad (9)$$

which can be calculated analytically as

$$\frac{\Gamma(V\lambda_0)\prod_v\Gamma(n_k^v+\lambda_0)}{\Gamma(n_k+V\lambda_0)\Gamma^V(\lambda_0)}, \quad (10)$$

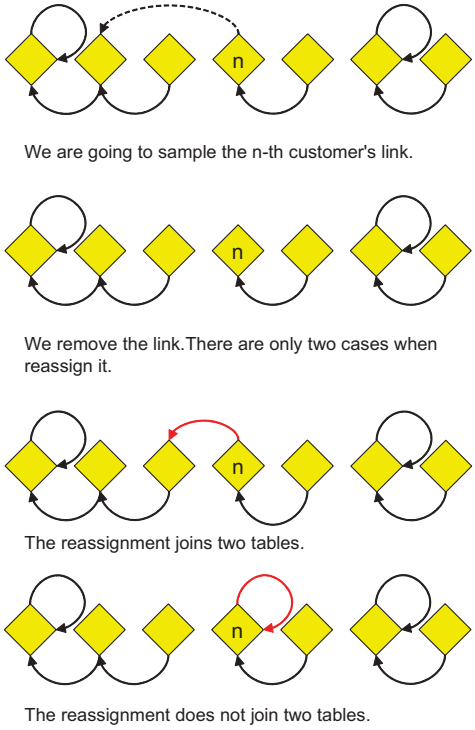

Fig. 3. An example of a single step of the Gibbs sampler.

where λ_0 is the parameter of the symmetrical Dirichlet distribution G_0, V is the size of the vocabulary, n_k^v is the number of word v assigned to table k and n_k is the number of all words assigned to table k.

4. EXPERIMENTS

We demonstrate the performance of our approach on the TDT2 VOA English broadcast news corpus[1] that includes speech recognition transcripts of 111 news programs with annotated story boundaries. We choose the dataset with the same test set used in [8] for evaluation. All texts are preprocessed by a Porter stemmer and stop words are removed. Then the texts are split into non-overlapping blocks with fixed-size. Performance is evaluated using the F1-measure according to the TDT2 standard. We compare our approach against three approaches: TextTiling [10], BayesSeg [14] and PLSA-DP-CE [8].

The PLSA-DP-CE approach first maps the term frequency feature into a topic representation feature using the pLSA model and then dynamic programming is used for story segmentation [8]. The BayesSeg approach models the topic and the word frequency feature uniformly by a probabilistic generative model. It also uses the dynamic programming method to find the final segmentation [14]. Therefore, both of the approaches can be viewed as special cases of the MinCut framework [13]. The difference is that PLSA-DP-CE uses a handcrafted cost function of the mapped topic feature while the BayesSeg uses a cost function of the term frequency induced from the view of probability.

[1] http://www.ldc.upenn.edu/Projects/TDT2

Approach	F1 Measure
TextTiling	0.5341
PLSA-DP-CE	0.6815
BayesSeg	0.7137
dd-CRP	0.7357

Table 1. Experimental results.

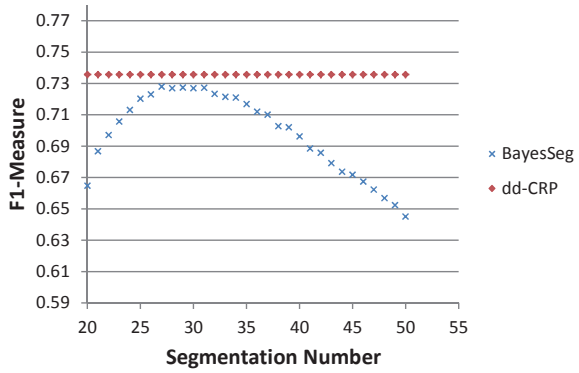

Fig. 4. F1 measure of a news program that contains 32 stories. BayesSeg is severely impacted by segment number setting while dd-CRP do not need this parameter.

Our approach is governed by the CRP concentration parameters α, the base Dirichlet distribution parameter λ_0 and the window size a. The parameter α contributes to the probability of opening up a new table, which means smaller values of α will encourage larger segmentation length. The parameter λ_0 is a smoothing parameter of term counts and larger λ_0 will make the two blocks less distinguishable. The window size a is a key parameter that can model the long distance dependence when it is set larger than 1. In our approach, we put non-informative priors on the parameters and use an EM-like process to optimize λ_0 and α. In the E step, we infer the table assignment and the segmentation bound by fixing the parameters, In the M step, λ_0 can be updated via the maximum likelihood method by fixing the segmentation. The parameter α can be updated via the auxiliary method by fixing the table assignments. The parameter a is tuned on a development set [2]. For the methods that need to know the story number beforehand, we assume the real story number is given according to the manual annotation. Parameter tuning in TextTiling and PLSA-DP-CE is the same with that in [8].

From Table 1, we can clearly see that our dd-CRP approach achieves the best result. The two Bayesian approaches (dd-CRP and BayesSeg) outperform the two approaches using handcrafted similarity metrics (TextTiling and PLSA-DP-CE). Although the BayesSeg outperforms the PLSA-DP-CE and its performance is comparable with our dd-CRP approach, it needs to set the number of storys manually. Fig 4 shows that the performance of BayesSeg will fall down when this number is set improperly. In contrast, our approach can learn the number automatically from data.

[2]This development set is small and a news program is enough. We tune the window parameter by the maximum likelihood criterion instead of the F1 measure.

5. CONCLUSIONS

This paper proposes an unsupervised approach for broadcast news story segmentation. We model the generative process of broadcast news using a fully probabilistic model. Specifically, we cut a broadcast news stream into text blocks and we consider the blocks in the same story are generated from a story-specific topic. To address the problem of unknown topic number and the non-exchangeability, we add a distance dependent Chinese restaurant process (dd-CRP) prior which has an essential effect that a block's topic is more likely to be the same with the adjacent blocks. Subsequently, story boundaries are discovered by detecting the change of topics. Experimental results show that our approach is superior to several recent approaches.

6. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (61175018) and the Fok Ying Tung Education Foundation (131059). We thank the anonymous reviewers for their valuable comments and suggetions.

7. REFERENCES

[1] James Allan, *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Academic Publishers, 2002.

[2] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, , and Gökhan Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1, pp. 127–154, 2000.

[3] Gökhan Tür, Dilek Hakkani-Tür, Andreas Stolcke, and Elizabeth Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Computational Linguistics*, vol. 27, no. 1, pp. 31–57, 2001.

[4] Andrew Rosenberg and Julia Hirschberg, "Story segmentation of broadcast news in English, Mandarin and Arabic," in *Proc. HLT-NAACL*, 2006, pp. 125–128.

[5] Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore, "Latent semantic analysis for text segmentation," in *Proc. EMNLP*, 2001, pp. 109–117.

[6] David Hall, Daniel Jurafsky, and Christopher D. Manning, "Studying the history of ideas using topic models," in *Proc. EMNLP*, 2008, pp. 363–371.

[7] Jen-Tzung Chien and Chuang-Hua Chueh, "Topic-based hierarchical segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 55–66, 2012.

[8] Mimi Lu, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Probabilistic latent semantic analysis for broadcast news story segmentation," in *Proc. Interspeech*, 2011, pp. 1301–1304.

[9] Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *Proc. ACL*, 2006, pp. 17–24.

[10] Marti A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[11] Banerjee, Satanjeev, and Rudnicky Alexander I, "A Texttiling based approach to topic boundary detection in meetings," in *Proc. Interspeech*, 2006.

[12] Xiaoxuan Wang, Lei Xie, Bin Ma, Eng Siong Chng, , and Haizhou Li, "Phoneme lattice based Texttiling towards multilingual story segmentation," in *Proc. Interspeech*, 2010, pp. 1305–1308.

[13] Igor Malioutov and Regina Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proc. ACL*, 2006, pp. 25–32.

[14] Jacob Eisenstein and Regina Barzilay, "Bayesian unsupervised topic segmentation," in *Proc. EMNLP*, 2008, pp. 334–343.

[15] Nguyen Viet-An, Boyd-Graber Jordan, and Resnik Philip, "SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations," in *Proc. ACL*, 2012, pp. 78–87.

[16] David M. Blei and Peter I. Frazier, "Distance dependent Chinese restaurant processes," *Journal of Machine Learning Research*, vol. 12, pp. 2461–2488, 2011.

[17] Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[18] Soumya Ghosh, Andrei B. Ungureanu, Erik B. Sudderth, and David M. Blei, "Spatial distance dependent Chinese restaurant processes for image segmentation," in *Proc. NIPS*, 2011, pp. 1476–1484.

[19] Soumya Ghosh, Erik B. Sudderth, Matthew Loper, and Michael Black, "Deformations to parts: Motion-based segmentation of 3D objects," in *Proc. NIPS*, 2012, pp. 2006–2014.

[20] Samuel J. Gershman and David M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.

[21] Radford M. Neal, "Markov chain sampling methods for Dirichlet processes mixture models," *Computational Linguistics*, vol. 9, no. 2, pp. 249–265, 2000.