Lexical Story Co-Segmentation of Chinese Broadcast News

Wei Feng[1], *Xuecheng Nie*[1], *Liang Wan*[2], *Lei Xie*[3], *Jianmin Jiang*[1]

[1]School of Computer Science and Technology, Tianjin University, Tianjin, China
[2]School of Computer Software, Tianjin University, Tianjin, China
[3]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{wfeng,xcnie,lwan}@tju.edu.cn, lxie@nwpu.edu.cn, jmjiang@tju.edu.cn

Abstract

We present an unsupervised technique, namely story co-segmentation, to automatically extract the common stories on the same topic within a pair of Chinese broadcast news transcripts. Unlike classical topic tracking that usually relies on previously trained topic models, our method is purely data-driven and is able to simultaneously determine the common stories of the input texts. Specifically, we propose an iterative four-step MRF solution to the problem of story co-segmentation using lexical cues only. We first construct a sentence-level graph formulation of the input news transcripts, and initialize foreground and background labeling by lexical clustering. We then update both foreground and background models based on the current labeling. We formalize story co-segmentation as a Gibbs energy minimization problem that balances the optimal objectives of foreground/background likelihood, intra-doc coherence, and inter-doc similarity. Finally, the labeling refinement is obtained by hybrid optimization with QPBO and BP. The effectiveness of our method has been validated on real-world CCTV corpus.

Index Terms: story co-segmentation, foreground and background story modeling, lexical clustering, MRF, QPBO, belief propagation (BP)

1. Introduction

Automatic extraction of common story segments on a same topic from multiple documents is very useful in practice, especially for semantic summarization [1] and user behavior analysis [2]. Under several proper conditions, this problem can be solved using the techniques of topic tracking and detection (TDT) [3]. For instance, if the common topic model is known and all story boundaries of input transcripts are available, topic tracking methods [4] can help us to detect all stories in the input streams focusing on the particular target topic. However, a general solution to this problem for unsegmented documents without any previously trained topic models is yet to be discovered.

In recent years, a new problem called image co-segmentation has rapidly attracted great attention in the area of computer vision and image analysis [5]. The success of image co-segmentation mainly attributes to its prominent capability of segmenting semantically related foregrounds in multiple images, without the need of either supervised training or user interaction [6].

In this paper, the concept of *co-segmentation* has been extended from image foreground/background labeling to lexical *co-story* extraction in a pair of Chinese broadcast news transcripts.[1] Our aim is to provide a general and unsupervised solution to automatically extract the common story segments from a pair of unsegmented Chinese broadcast news transcripts via lexical cues only. That is, our method is purely data-driven and relies only on the intra- and inter-doc dependencies and constraints to detect the semantically meaningful co-story.

To this end, we propose a four-step iterative approach based on Markov random field (MRF) to the problem of lexical story co-segmentation. For the input pair of Chinese news transcripts, we first construct a sentence-level graph formulation to encode both intra- and inter-doc dependencies. The first step also initializes foreground/background labeling via lexical clustering and common-cluster selection. Then, the current labeling is further used to update the foreground and background models. Next, we formalize story co-segmentation as a Gibbs energy minimization problem through regularizing the optimal objectives of foreground/background likelihood, intra-doc coherence, and inter-doc similarity. At last, the refined foreground/background labeling is obtained by hybrid optimization. Experiments on real-world CCTV corpus show that our method usually outperforms story-matching in extracting common co-stories from Mandarin broadcast news transcripts.

2. Lexical Story Co-segmentation

From the "bag-of-words" assumption, the semantics of a word stream $\mathcal{S}$ can be statistically represented by its word frequency distribution, or un-normalized histogram equivalently, over a common vocabulary $\mathcal{V}$. Therefore, in this paper, we measure the distance between any two

[1]For a pair of news transcripts $\mathcal{T}_1$ and $\mathcal{T}_2$, a co-story $\mathcal{S} = \{S_1, S_2\}$ refers to a pair of stories S_1 and S_2, each of which is extracted from one transcript and both of which discuss a same unknown topic.

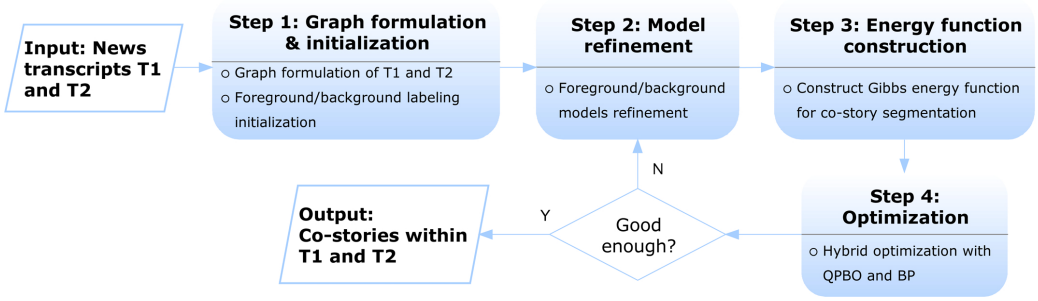

Figure 1: *The proposed four-step algorithm for lexical story co-segmentation.*

word sequences s_i and s_j by $\|\mathrm{H}(s_i) - \mathrm{H}(s_j)\|_2$, where $\mathrm{H}(s_i)$ and $\mathrm{H}(s_j)$ denote the word-over-vocabulary distributions of s_i and s_j, respectively, $\|\cdot\|_2$ is the Euclidean distance. The goal of lexical story co-segmentation is to extract common story segments, i.e., co-story, from the input pair of Chinese news transcripts, with maximum intra-doc foreground/background likelihood and minimum inter-doc distance. For this purpose, we propose a four-step iterative algorithm, as shown in Fig. 1.

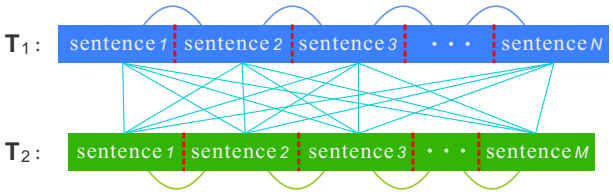

Figure 2: *An example of sentence-level graph formulation of a pair of news transcripts with cutoff value* $\tau = 1$.

2.1. Initialization

Our first step is to construct a sentence-level graph formulation to the input Chinese news transcripts $\mathcal{T}_1$ and $\mathcal{T}_2$. See Fig. 2 for an example. We split each input transcript into a sequence of sentences s_i with fixed length L. The graph is established with vertex set as all sentences in $\mathcal{T}_1$ and $\mathcal{T}_2$. The edge set of the graph is composed of *intra-doc edges* encouraging adjacent sentences in a doc belong to a same topic and *inter-doc edges* ensuring sentences in different docs with similar semantics be labeled as a same story. To control the complexity, two sentences are linked as an intra-doc edge iff their distance is lower than a cutoff threshold τ. Clearly, such graph formulation is an instance of MRF model [7].

With the graph formulation, story co-segmentation becomes a 0-1 labeling problem to all graph vertices, i.e., all sentences of $\mathcal{T}_1$ and $\mathcal{T}_2$. Specifically, those sentences belong to the common co-story are labeled as foreground 1, while others are labeled as background 0, i.e.,

$$\mathcal{T}_1 \cup \mathcal{T}_2 = \mathcal{F} \cup \mathcal{B}_1 \cup \mathcal{B}_2 \tag{1}$$

where $\mathcal{F} = \{s|\mathrm{Label}(s) = 1 \wedge s \in \mathcal{T}_j \wedge j \in \{1,2\}\}$ is the set of foreground sentences in $\mathcal{T}_1$ and $\mathcal{T}_2$, and $\mathcal{B}_j = \{s|\mathrm{Label}(s) = 0 \wedge s \in \mathcal{T}_j\}$ is the set of background sentences of $\mathcal{T}_j$ ($j \in \{1,2\}$).

To obtain a reasonable initialization to the common foreground $\mathcal{F}$, backgrounds $\mathcal{B}_1$ and $\mathcal{B}_2$, we agglomerate all sentences of $\mathcal{T}_1$ and $\mathcal{T}_2$ into K clusters, and select the most possible common-cluster by minimizing the following *discrepancy score*:

$$\mathrm{Ds}(\mathcal{C}_k) = \gamma\left|\frac{\min(\mathcal{C}_k^1, \mathcal{C}_k^2)}{\max(\mathcal{C}_k^1, \mathcal{C}_k^2)} - 1\right| + (1-\gamma)\left\|\mathrm{H}(\mathcal{C}_k^1) - \mathrm{H}(\mathcal{C}_k^2)\right\|_2 \tag{2}$$

where $\mathcal{C}_k = \{\mathcal{C}_k^1, \mathcal{C}_k^2\}$ refers to the kth nonempty clusters, $\mathcal{C}_k^j$ represents the subset of sentences in $\mathcal{C}_k$ belonging to transcript $\mathcal{T}_j$ ($j \in \{1,2\}$), $\mathrm{H}(\cdot)$ denotes the word-over-vocabulary distribution for a given set of words, γ is a linear modulation parameter controlling the relative importance of size and semantic discrepancies. According to (2), we can easily select the optimal common-cluster $\hat{\mathcal{C}} = \arg\min_k \mathrm{Ds}(\mathcal{C}_k)$ via enumerating the discrepancy scores of all K clusters. The foreground/background labeling can then be initialized accordingly.

2.2. Foreground/Background Story Modeling

According to the current foreground/background labeling $\mathcal{F}^{(t)}$, $\mathcal{B}_1^{(t)}$ and $\mathcal{B}_2^{(t)}$, we directly update their models as $\mathrm{H}(\mathcal{F}^{(t)})$, $\mathrm{H}(\mathcal{B}_1^{(t)})$ and $\mathrm{H}(\mathcal{B}_2^{t})$, respectively, with t depicting the current number of iterations.

2.3. Gibbs Energy for Story Co-segmentation

Generally, a good story co-segmentation should at least satisfy two conditions: (i) the extracted foreground segments should be good stories in their own transcripts (i.e., the foreground/background likelihood and the neighboring coherence prior should be properly balanced); (ii) the extracted foreground stories should be semantically similar enough. Accordingly, we present the following Gibbs energy function for lexical story co-segmentation:

$$\mathrm{E}(X) = \sum_{d=1}^{2} \mathrm{E}_{\mathrm{intra}}(X_d) + \beta\,\mathrm{E}_{\mathrm{inter}}(X) \tag{3}$$

where $X = X_1 \cup X_2 = \{x_i\}_{i=1}^{N+M}$ is the set of label variables of all sentences in news transcripts $\mathcal{T}_1$ and $\mathcal{T}_2$,

Table 1: *Lexical story co-segmentation results (F1-measure) on "cctv-66-s" dataset.*

Method	Score	Word		Unigram		Bigram		Trigram		Quadgram	
		Char.	Syll.	Char.	Syll.	Char.	Syll.	Char.	Syll.	Char.	Syll.
Story-matching	F1-measure	0.49	0.54	0.54	0.58	0.53	0.52	0.53	0.53	0.55	0.52
	Precision	0.53	0.58	0.51	0.61	0.61	**0.52**	0.50	0.41	0.62	**0.51**
	Recall	0.46	0.51	0.57	0.55	0.47	0.52	**0.58**	0.79	0.49	0.53
Our method	F1-measure	**0.60**	**0.63**	**0.74**	**0.66**	**0.63**	**0.56**	**0.60**	**0.60**	**0.77**	**0.53**
	Precision	**0.56**	**0.80**	**0.89**	**0.85**	**0.63**	0.48	**0.87**	**0.46**	**0.77**	0.50
	Recall	**0.66**	**0.52**	**0.64**	**0.64**	**0.63**	**0.67**	0.48	**0.87**	**0.77**	**0.58**

$x_i \in \{0,1\}$ is the label of ith sentence, X_1 (with N variables) and X_2 (with M variables) are the labeling of sentences in transcripts $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, coefficient β is used to balance the role of intra- and inter-doc energies.

Intra-doc energy $\mathrm{E}_{\mathrm{intra}}(\cdot)$ measures the goodness of foreground/background labeling within a same transcript:

$$\begin{aligned}\mathrm{E}_{\mathrm{intra}}(X_d) = \sum_{i=1}^{|X_d|} \big(& x_{i,d} \left\|\mathrm{H}(\mathcal{F}^{(t)}) - \mathrm{H}(s_{i,d})\right\|_2 \\ & + (1-x_{i,d}) \left\|\mathrm{H}(\mathcal{B}_d^{(t)}) - \mathrm{H}(s_{i,d})\right\|_2 \big) \\ & + \alpha \sum_{i\sim j} |x_{i,d} - x_{j,d}|\end{aligned} \tag{4}$$

where X_d is the labeling of the dth document ($d \in \{1,2\}$), $s_{i,d}$ represents the ith sentence in document d, $x_{i,d}$ is its label variable, $i \sim j$ indicates that sentences i and j are adjacent. Intra-doc energy $\mathrm{E}_{\mathrm{intra}}(\cdot)$ is composed of two parts. The first part represents the foreground/background labeling cost, while the second part reflects the adjacent coherence prior, with α as the parameter modulating their relative influences.

Inter-doc energy $\mathrm{E}_{\mathrm{inter}}(\cdot)$ manages the similarity between the foreground stories of two input transcripts:

$$\mathrm{E}_{\mathrm{inter}}(X) = \sum_{k=1}^{K} \Big(\textstyle\sum_{p\in\mathcal{C}_k^1} x_p - \sum_{q\in\mathcal{C}_k^2} x_q \Big)^2 \tag{5}$$

Note that inter-doc energy $\mathrm{E}_{\mathrm{inter}}(\cdot)$ penalizes the difference between the un-normalized histograms of the potential foreground stories in two transcripts. Fig. 3 shows an example of the effectiveness of such measurement. We can clearly see that stories on same topic have very close lexical distributions.

2.4. Hybrid Optimization and Refinement

Since (5) contains lots of submodular and supermodular items at the same time, it is generally NP-hard to minimize the Gibbs energy function defined in (3) [7, 8]. For the purpose of both accuracy and efficiency, in this paper, we use a hybrid energy minimization method to solve (3). We first minimize the original energy function (3) by QPBO [8]. Due to the persistency and partial optimality properties of QPBO, we reserve all 0-1 labels produced by QPBO, and then use BP [9] to approximately optimize the unlabeled variables by minimizing the simplified energy function of (3), thus obtaining a suboptimal full labeling to all variables in X.[2] Based on current X, we update the foreground/background labeling $\mathcal{F}^{(t+1)}$, $\mathcal{B}_1^{(t+1)}$ and $\mathcal{B}_2^{(t+1)}$ accordingly.

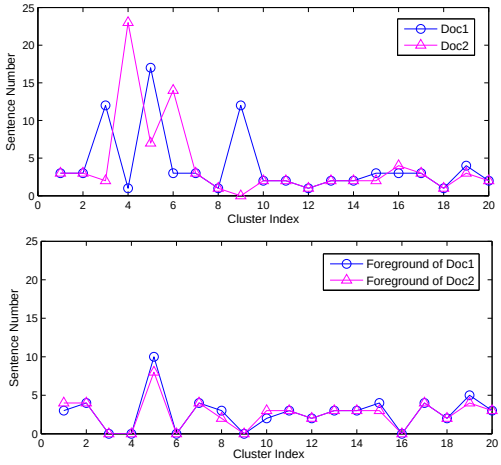

Figure 3: *Row 1 and 2 show the lexical distributions of two randomly selected docs, and of their corresponding foreground co-stories, respectively. The distance of the two whole docs is* 31.6*, while the distance between the two foreground stories is only* 3.0*.*

3. Experimental Results

Our experiments were carried out based on the real-world CCTV corpus, which covers 71 news episodes with 27 hours of Mandarin broadcast news and contains three different ASR rates 59%, 66% and 75%. We compared our method with *story-matching* in lexical story co-segmentation of Chinese broadcast news.[3] Since the n-gram subword representation is robust to speech recognition and OOV errors in Mandarin broadcast news [10], we tested our method and story-matching at word level

[2] The simplified Gibbs energy function $\mathrm{E}_{\mathrm{simp}}(X_{0.5}) = \mathrm{E}(X = X_{0.5} \cup X_{0/1})$, where $X_{0.5}$ and $X_{0/1}$ indicate the unlabeled variables and variables certainly labeled by QPBO, respectively.

[3] Story-matching extracts the co-story from two broadcast news transcripts through establishing the story-level matching with minimum discrepancy score (2).

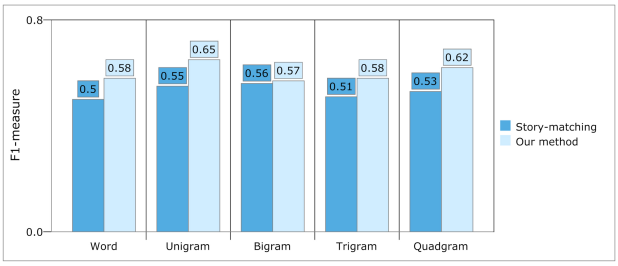

Figure 4: *Average performance of story-matching and our method on four datasets in CCTV corpus.*

and four character/syllable subword levels. All results, i.e., the F1-measure, reported in this paper were obtained on a testing dataset using empirically tuned parameters based on corresponding tuning dataset. In our experiments, we used ground-truth story boundaries in story-matching to remove the possible bias caused by particular story segmentation methods.

Table 1 summarizes the co-segmentation results on the "cctv-66-s" dataset of the proposed method and story-matching. We observe that our method is better than story-matching at all levels. For instance, using character unigram, our method achieved 37.5% relative improvement over story-matching. Fig. 4 shows the average F1-measure of story-matching and our method on four datasets "cctv-59-f", "cctv-59-s", "cctv-66-f" and "cctv-66-s", which also validates the superior performance of our method. For clarity, in the following, only the best result using character or syllable is reported at each level. We can clearly see that the accuracy of the proposed method is usually superior to that of story-matching. This is mainly because our method achieved better overall balance of foreground/background likelihood, intra-doc coherence and inter-doc similarity than story-matching did.

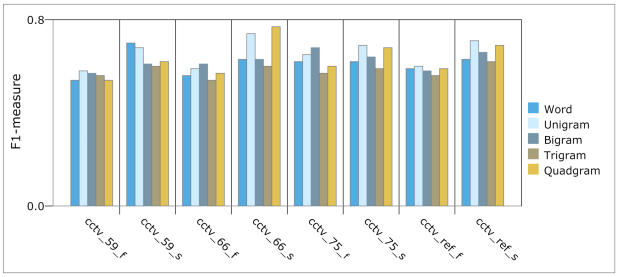

Figure 5: *Co-segmentation results of our method on all datasets in CCTV corpus.*

We have tested out method on all datasets in CCTV corpus, the results are shown in Fig. 5. We observe that the results at subword levels are usually better than that of word-level. This is mainly due to the fact that the most frequently used words in Chinese are bi-character and bi-gram subwords, which are robust to speech recognition errors and OOV words [10].

4. Conclusions

This paper has extended the concept of *co-segmentation* to lexical *co-story* extraction in a pair of Chinese broadcast news transcripts. Different from classical topic tracking and detection, story co-segmentation requires unsupervised approach able to accurately determine the co-stories in multiple unsegmented documents without the guidance of pre-trained topic models. To this end, we have proposed a general solution within the framework of MRF. Our main contributions are: (1) a feasible criterion for common lexical cluster selection and foreground/background initialization considering both size and semantic discrepancies; and (2) a general and extendible Gibbs energy function and corresponding hybrid optimization algorithm for lexical story co-segmentation. Experiments on real-world CCTV corpus show that our method usually outperforms story-matching in extracting common co-stories from Mandarin broadcast news transcripts at different ASR error rates.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (61100121, 61100122 and 61175018), the Research Fund for the Doctoral Program of Higher Education (20110032120036 and 20110032120041) and the Program for New Century Excellent Talents in University (NCET-11-0365).

6. References

[1] L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 42–60, 2005.

[2] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior," in *IJCAI*, 2009.

[3] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report," in *The DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 194–218.

[4] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi, "Topic tracking using subject templates and clustering positive training instances," in *COLING*, 2002.

[5] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs," in *CVPR*, 2006.

[6] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *ECCV*, 2010.

[7] W. Feng, J. Jia, and Z.-Q. Liu, "Self-validated labeling of Markov random fields for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1871–1887, 2010.

[8] V. Kolmogorov and C. Rother, "Minimizing nonsubmodular functions with graph cuts - a review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 7, pp. 1274–1279, 2007.

[9] J.-S. Yedidia, W.-T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 51, no. 7, pp. 2282–2312, 2005.

[10] J. Zhang, L. Xie, W. Feng, and Y. Zhang, "A subword normalized cut approach to automatic story segmentation of Chinese broadcast," in *AIRS*, 2009.