Text Segmentation with LDA-Based Fisher Kernel

Qi Sun, Runxin Li, Dingsheng Luo and Xihong Wu
Speech and Hearing Research Center, and
Key Laboratory of Machine Perception (Ministry of Education)
Peking University
100871, Beijing, China
{sunq,lirx,dsluo,wxh}@cis.pku.edu.cn

Abstract

In this paper we propose a domain-independent text segmentation method, which consists of three components. Latent Dirichlet allocation (LDA) is employed to compute words semantic distribution, and we measure semantic similarity by the Fisher kernel. Finally global best segmentation is achieved by dynamic programming. Experiments on Chinese data sets with the technique show it can be effective. Introducing latent semantic information, our algorithm is robust on irregular-sized segments.

1 Introduction

The aim of text segmentation is to partition a document into a set of segments, each of which is coherent about a specific topic. This task is inspired by problems in information retrieval, summarization, and language modeling, in which the ability to provide access to smaller, coherent segments in a document is desired.

A lot of research has been done on text segmentation. Some of them utilize linguistic criteria (Beeferman et al., 1999; Mochizuki et al., 1998), while others use statistical similarity measures to uncover lexical cohesion. Lexical cohesion methods believe a coherent topic segment contains parts with similar vocabularies. For example, the TextTiling algorithm, introduced by (Hearst, 1994), assumes that the local minima of the word similarity curve are the points of low lexical cohesion and thus the natural boundary candidates. (Reynar, 1998) has proposed a method called dotplotting depending on the distribution of word repetitions to find tight regions of topic similarity graphically. One of the problems with those works is that they treat terms uncorrelated, assigning them orthogonal directions in the feature space. But in reality words are correlated, and sometimes even synonymous, so that texts with very few common terms can potentially be on closely related topics. So (Choi et al., 2001; Brants et al., 2002) utilize semantic similarity to identify cohesion. Unsupervised models of texts that capture semantic information would be useful, particularly if they could be achieved with a "semantic kernel" (Cristianini et al., 2001) , which computes the similarity between texts by also considering relations between different terms. A Fisher kernel is a function that measures the similarity between two data items not in isolation, but rather in the context provided by a probability distribution. In this paper, we use the Fisher kernel to describe semantic information similarity. In addition, (Fragkou et al., 2004; Ji and Zha, 2004) has treated this task as an optimization problem with global cost function and used dynamic programming for segments selection.

The remainder of the paper is organized as follows. In section 2, after a brief overview of our method, some key aspects of the algorithm are described. In section 3, some experiments are presented. Finally conclusion and future research directions are drawn in section 4.

2 Methodology

This paper considers the sentence to be the smallest unit, and a block b is the segment candidate which consists of one or more sentences. We employ LDA

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 269–272,
Columbus, Ohio, USA, June 2008.

model (Blei et al., 2003) in order to find out latent semantic topics in blocks, and LDA-based Fisher kernel is used to measure the similarity of adjacent blocks. Each block is then given a final score based on its length and semantic similarity with its previous block. Finally the segmentation points are decided by dynamic programming.

2.1 LDA Model

We adopt LDA framework, which regards the corpus as mixture of latent topics and uses document as the unit of topic mixtures. In our method, the blocks defined in previous paragraph are regarded as "documents" in LDA model.

The LDA model defines two corpus-level parameters α and β. In its generative process, the marginal distribution of a document $p(d|\alpha, \beta)$ is given by the following formula:

$$\int p(\theta|\alpha)(\prod_{n=1}^{N}\sum_{k} p(z_k|\theta_d)p(w_n|z_k,\beta))d\theta$$

where d is a word sequence $(w_1, w_2, ...w_N)$ of length N. α parameterizes a Dirichlet distribution and derives the document-related random variable θ_d, then we choose a topic $z_k,\ k \in \{1...K\}$ from the multinomial distribution of θ_d. Word probabilities are parameterized by a $k \times V$ matrix β with V being the size of vocabulary and $\beta_{vk} = P(w = v|z_k)$. We use variational EM (Blei et al., 2003) to estimate the parameters.

2.2 LDA-Based Fisher Kernel

In general, a kernel function $k(x, y)$ is a way of measuring the resemblance between two data items x and y. The Fisher kernel's key idea is to derive a kernel function from a generative probability model. In this paper we follow (Hofmann, 2000) to consider the average log-probability of a block, utilizing the LDA model. The likelihood of b is given by:

$$l(b) = \sum_{i=1}^{N} \widehat{P}(w_i|b) \log \sum_{k=1}^{K} \beta_{w_i k}\theta_b^{(k)}$$

where the empirical distribution of words in the block $\widehat{P}(w_i|b)$ can be obtained from the number of word-block co-occurrence $n(b, w_i)$, normalized by the length of the block.

The Fisher kernel is defined as

$$K(b_1, b_2) = \bigtriangledown_\theta^T l(b_1) I^{-1} \bigtriangledown_\theta l(b_2)$$

which engenders a measure of similarity between any two blocks b_1 and b_2. The derivation of the kernel is quite straightforward and following (Hofmann, 2000) we finally have the result:

$$K(b_1, b_2) = K_1(b_1, b_2) + K_2(b_1, b_2), \quad with$$
$$K_1(b_1, b_2) = \sum_k \theta_{b_1}^{(k)}\theta_{b_2}^{(k)}/\theta_{corpus}^{(k)}$$
$$K_2(b_1, b_2) =$$
$$\sum_i \widehat{P}(w_i|b_1)\widehat{P}(w_i|b_2) \sum_k \frac{P(z_k|b_1,w_i)P(z_k|b_2,w_i)}{P(w_i|z_k)}$$

where $K_1(b_1, b_2)$ is a measure of how much b_1 and b_2 share the same latent topic, taking synonymy into account. And $K_2(b_1, b_2)$ is the traditional inner product of common term frequencies, but weighted by the degree to which these terms belong to the same latent topic, taking polysemy into account.

2.3 Cost Function and Dynamic Programming

The local minima of LDA-based Fisher kernel similarities indicate low semantic cohesion and segmentation candidates, which is not enough to get reasonably-sized segments. The lengths of segmentation candidates have to be considered, thus we build a cost function including two parts of information. Segmentation points can be given in terms of a vector $\vec{t} = (t_0, ..., t_m, ..., t_M)$, where t_m is the sentence label with m indicating the mth block. We define a cost function as follows:

$$\begin{aligned} J(\vec{t};\lambda) &= \sum_{m=1}^{M} \lambda F(l_{t_m+1,t_{m+1}}) \\ &+ K(b_{t_{m-1}+1,t_m}, b_{t_m+1,t_{m+1}}) \end{aligned}$$

where $F(l_{t_m+1,t_{m+1}})$ is equal to $\frac{(l_{t_m+1,t_{m+1}}-\mu)^2}{2\sigma^2}$ and $l_{t_m+1,t_{m+1}}$ is equal to $t_{m+1} - t_m$ indicating the number of sentences in block m. The LDA-based kernel function measures similarity of block $m-1$ and block m, where block $m-1$ spans sentence $t_{m-1}+1$ to t_m and block m spans sentence $t_m + 1$ to t_{m+1}

The cost function is the sum of the costs of assumed unknown M segments, each of which is made up of the length probability of block m and the similarity score of block m with its previous block $m-1$. The optimal segmentation $\vec{t}$ gives a global minimum of $J(\vec{t};\lambda)$.

3 Experiments

3.1 Preparation

In our experiments, we evaluate the performance of our algorithms on Chinese corpus. With news documents from Chinese websites, collected from 10 different categories, we design an artificial test corpus in the similar way of (Choi, 2000), in which we take each n-sentence document as a coherent topic segment, randomly choose ten such segments and concatenate them as a sample. Three data sets, Set 3-5, Set 13-15 and Set 5-20, are prepared in our experiments, each of which contains 100 samples. The data sets' names are represented by a range number n of sentences in a segment.

Due to generality, we take three indices to evaluate our algorithm: precision, recall and error rate metric (Beeferman et al., 1999) . And all experimental results are averaged scores generated from the individual results of different samples. In order to determine appropriate parameters, some hold-out data are used.

We compare the performance of our methods with the algorithm in (Fragkou et al., 2004) on our test set. In particular, the similarity representation is a main difference between those two methods. While we pay attention to latent topic information behind words of adjacent blocks, (Fragkou et al., 2004) calculates word density as the similarity score function.

3.2 Results

In order to demonstrate the improvement of LDA-based Fisher kernel technique in text similarity evaluation, we omit the length probability part in the cost function and compare the LDA-based Fisher kernel and the word-frequency cosine similarity by the error rate P_k of segmenting texts. Figure 1 shows the error rates for different sets of data. On average, the error rates are reduced by as much as about 30% over word-frequency cosine similarity with our methods, which shows Fisher kernel similarity measure,with latent topic information added by LDA, outperforms traditional word similarity measure. The performance comparisons drawn from Set 3-5 and Set 13-15 indicates that our similarity algorithm can uncover more descriptive statistics than traditional one especially for segments with less sentences due to its prediction on latent topics.

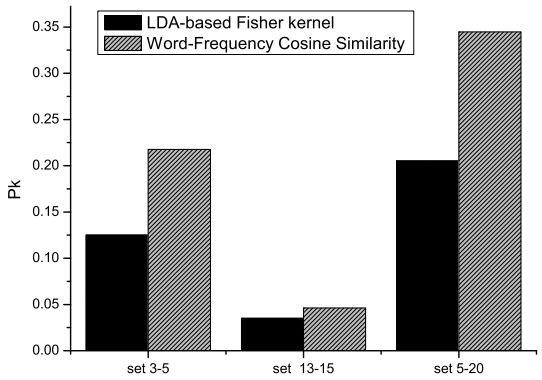

Figure 1: *Error Rate P_k on different data sets with different similarity metrics.*

In the cost function, there are three parameters μ , σ and λ. We determine appropriate μ and σ with hold-out data. For the value of λ, we take it between 0 and 1 because the length part is less important than the similarity part according to our preliminary experiments. We design the experiment to study λ's impact on segmentation by varying it over a certain range. Experimental results in Figure 2 show that the reduce of error rate achieved by our algorithm is in a range from 14.71% to 53.93%. Set 13-15 achieves best segmentation performance, which indicates the importance of text structure: it is easier to segment the topic with regular length and more sentences. The performance on Set 5-20 obtains the best improvement with our methods, which illustrates that LDA-based Fisher kernel can express text similarity more exactly than word density similarity on irregular-sized segments.

Table 1: *Evaluation against different algorithms on Set 5-20.*

Algo.	P_k	Recall	Precision
TextTiling	0.226	66.00%	60.72 %
P. Fragkou Algo.	0.344	69.00%	37.92 %
Our Algo.	0.205	59.00%	62.27 %

While most experiments of other authors were taken on short regular-sized segments which was firstly presented by (Choi, 2000), we use comparatively long range of segments, Set 5-20, to evaluate different algorithms. Table 1 shows that, in terms of

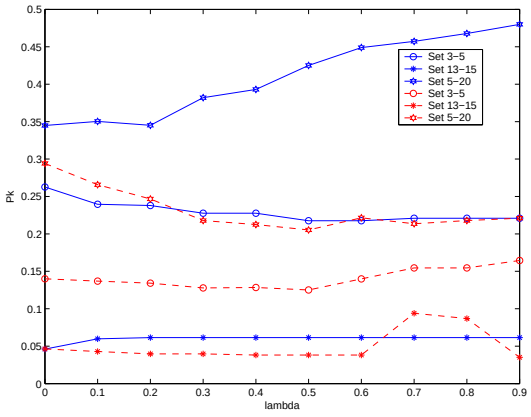

Figure 2: *Error Rate P_k when the λ changes. There are two groups of lines, the solid lines representing algorithm of (Fragkou et al., 2004) while the dash ones indicate performance of our algorithm, and each line in a group shows error rates in different data sets.*

P_k, our algorithm employing dynamic programming as P. Fragkou Algo. achieves the best performance among those three. As for long irregular-sized text segmentation, although local even-sized blocks similarity provides more exact information than the similarity between global irregular-sized texts, with the consideration of latent topic information, the latter will perform better in the task of text segmentation. Though the performance of the proposed method is not superior to TextTiling method, it avoids thresholds selection, which makes it robust in applications.

4 Conclusions and Future Work

We present a new method for topic-based text segmentation that yields better results than previously methods. The method introduces a LDA-based Fisher kernel to exploit text semantic similarities and employs dynamic programming to obtain global optimization. Our algorithm is robust and insensitive to the variation of segment length. In the future, we plan to investigate more other similarity measures based on semantic information and to deal with more complicated segmentation tasks. Also, we want to exam the factor importance of similarity and length in this text segmentation task.

Acknowledgments

The authors would like to thank Jiazhong Nie for his help and constructive suggestions. The work was supported in part by the National Natural Science Foundation of China (60435010; 60535030; 60605016), the National High Technology Research and Development Program of China (2006AA01Z196; 2006AA010103), the National Key Basic Research Program of China (2004CB318005), and the New-Century Training Program Foundation for the Talents by the Ministry of Education of China.

References

Doug Beeferman, Adam Berger and John D. Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210.

David M. Blei and Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning Research* 3: 993–1022.

Thorsten Brants, Francine Chen and Ioannis Tsochantaridis. 2002. Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis. *CIKM '02*211–218

Freddy Choi, Peter Wiemer-Hastings and Johanna Moore. 2001. Latent Semantic Analysis for Text Segmentation. *Proceedings of 6th EMNLP*, 109–117.

Freddy Y. Y. Choi. 2000. Advances in Domain Independent Linear Text Segmentation. *Proceedings of NAACL-00*.

Nello Cristianini, John Shawe-Taylor and Huma Lodhi. 2001. Latent Semantic Kernels. *Proceedings of ICML-01, 18th International Conference on Machine Learning* 66–73.

Pavlina Fragkou, Petridis Vassilios and Kehagias Athanasios. 2004. A Dynamic Programming Algorithm for Linear Text Segmentation. *J. Intell. Inf. Syst.*, 23(2): 179–197.

Marti Hearst. 1994. Multi-Paragraph Segmentation of Expository Text. *Proceedings of the 32nd. Annual Meeting of the ACL*, 9–16.

Thomas Hofmann. 2000. Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization. *Advances in Neural Information Processing Systems* 12: 914–920.

Xiang Ji and Hongyuan Zha. 2003. Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. *Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Informaion Retrieval,* 322–329.

Hajime Mochizuki, Takeo Honda and Manabu Okumura. 1998. Text Segmentation with Multiple Surface Linguistic Cues. *Proceedings of the COLING-ACL'98*, 881-885.

Jeffrey C. Reynar. 1998. Topic Segmentation: Algorithms and Applications. PhD thesis. University of Pennsylvania.