Applied Mechanics and Materials *Online: 2014-05-23*
ISSN: 1662-7482, Vols. 556-562, pp 4018-4022
doi:10.4028/www.scientific.net/AMM.556-562.4018

Text Segmentation Based on PLSA-TextTiling Model

YuChao Zheng[1, a]

[1]School of East China Jiaotong University, Nanchang 330013, China

[a]zyc0791@163.com

Keywords: text segmentation; probabilistic latent semantic analysis (PLSA); similarity metric; boundary discovering

Abstract. Text segmentation is very important for many fields including information retrieval, summarization, language modeling, anaphora resolution and so on. Text segmentation based on PLSA-TextTiling associates different latent topic swith observable pairs of word and sentence. In the experiments, the whole sentences are taken as elementary blocks. PLSA model is used to calculated similarity metric basing on the idea of TestTiling and several approaches to discovering boundaries are tried. The results show the Pμ value is 0.87, which is better than that of other algorithms of text segmentation.

Introduction

Text segmentation is very important for many fields including information retrieval, summarization, language modeling, anaphora resolution and so on. Text segmentation means combining adjacent paragraphs of one sub-topic as a semantic paragraph. This semantic paragraph is called a segment unit or a segment. In this way, a text could be linear segmented into several semantic paragraphs, forming a semantic paragraph sequence. The essence of which is to find boundaries of each topic within the text according to their similarity with the topic so that the internal segments are of maximum similarity and the inter-segments are of the minimum. Therefore, a text segmentation algorithm is to solve the fundamental problem of similarity metric and boundary discovering[1]. Since 1990s, overseas researchers began researching on auto-segmentation of long English texts and put forward some segmentation methods, which could be divided into three categories in general, namely segmentation method on the basis of lexical cohesion[2], on linguistic features[3] and on statistics[4]. Method adopted in this paper falls into the third category. It presented a segmentation method of PLSA-TextTiling to solve two essential issues of text segmentation algorithm: similarity metric and boundary discovering. The method is to determine the minimum interval point of chapter compactness in order to achieve the goal of the division of discourse structure, borrowing the basic ideas of TextTiling and applying PLSA measuring the similarity of adjacent text unit.

PLSA Model

Model Description. Many applications of Natural Language Processing (NLP) needs to explore meaning hidden inside words. Simple literally match hardly works, so the crux lies in the grasp of the synonyms and polysemy. Latent Semantic Analysis (LSA) provides some methods of solving the problems, that is, using Singular Value Decomposition (SVD) to project the high dimensional lexicon-text co-occurrence matrix to a low dimensional latent semantic space, so that uncorrelated lexicons reflet deeper similarity[5]. As a ramification of LSA, Probabilistic Latent Semantic Analysis (PLSA) owns a solider mathematics foundation and a more applicable data model generation, and it has been proved to provided better vocabulary match for information retrieval[6].

The basic idea of PLSA is the same with LSA, but PLSA provides an appropriate probability model. This model initially hypothesis a hidden variable $z \in Z = \{z_1, z_2, \cdots, z_k\}$, which can be seen as topic. Certainly, the topic variable z cannot be observed directly but the variable pair <d, w> can, that is, word w appears in text d, $d \in D = \{d_1, d_2, \cdots, d_N\}$, $w \in W = \{w_1, w_2, \cdots, w_M\}$, then Probabilistic Generation Model corelates a topic variable z with each <d, w>. This model is called Aspect Model. It can be shown as joint probability of lexicon and text:

$$P(d_i, w_j) = \sum_{K=1}^{K} P(z_k)P(w_j \mid z_k)P(d_i \mid z_k) \quad (1)$$

$P(z_k)$ is the latent semantic probability distribution, $P(w_j|z_k)$ is the latent semantic probability distribution of lexicon, and $P(d_i|z_k)$ is the latent semantic probability distribution of text.

EM Fitting Algorithm. LSA uses SVD to estimate "lexicon-text" co-occurrence matrix, while PLSA uses Expectation Maximization (EM) algorithm to fit latent semantic structure. Obviously, latent topic variety z is the bottleneck of calculating each conditional probability component, so actually observed statistics of lexical appearance in text is needed to fit its probability distribution. EM algorithm conducts Expectation Step (E step) and Maximization Step (M step) alternatively to do the iterative computations and progressive optimization of the whole model.

Before using EM algorithm on symmetry model, four matrixes should be constructed and initialized: a) constructing text and lexicon co-occurrent matrix $N=[n(d, w)]_{M\times N}$ which means lexicon w's frequency of occurrence in text d; b) constructing PLSA matrixes $U=[P(w, z)]_{M\times K}$, $V=[P(d, z)]_{N\times K}$ and $\sum=[diag(P(z))]_{k\times k}$, each component of matrix U, V and $\sum$ is generated randomly, so summation of each column of matrix U and V is 1, and summation of diagonal element of matrix $\sum$ is 1.

E step: calculating the posterior probability of each variable < d, w> generating underlying topic variable z.

$$P(z_k \mid d_i, w_j) = \frac{P(z_k)P(d_i \mid z_k)P(w_j \mid z_k)}{\sum_{t=1}^{k} P(z_t)P(d_i \mid z_t)P(w_j \mid z_t)} \quad (2)$$

M step: reestimating the model utilizing the following formula.

$$\begin{cases} P(w_j \mid z_k) = \dfrac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P(z_k \mid d_i, w_m)} \\ P(d_i \mid z_k) = \dfrac{\sum_{m=1}^{M} n(d_i, w_m)P(z_k \mid d_i, w_m)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P(z_k \mid d_i, w_m)} \\ P(z_k) = \dfrac{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P(z_k \mid d_i, w_m)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)} \end{cases} \quad (3)$$

According to the principle of maximum likelihood estimate, by maximizing the following log-likelihood function to set the iteration termination condition of EM algorithm.

$$L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log P(d_i, w_j) \quad (4)$$

When the calculated value increment of the log-likelihood function of the above formula converge gradually until less than a given threshold to stop the iteration, then an optimal solution is got. This means construction of the whole PLSA model has been completed. Similar to SVD of LSA, after finishing statistical estimating the symmetrical PLSA model, by using the constructed matrixes U, V and Σ, a "lexicon-text" joint probability matrix H_{d_w} can be acquired, which reflects the relevancy between lexicon and text.

$$H_{d\_w} = U\Sigma V' = [P(d, w)]_{M\times N} \quad (5)$$

Meanwhile, a similarity matrix H_{w_w} of "lexicon-lexicon" and a similarity matrix H_{d_d} of "text-text" can be acquired easily:

$$H_{w\_w} = H_{d\_w}H'_{d\_w} = [P(w_i, w_j)]_{N\times M} \quad (6)$$

$$H_{d\_d} = H'_{d\_w}H_{d\_w} = [P(d_i, d_j)]_{N\times N} \quad (7)$$

Text Segmentation Method Based on PLSA-TextTiling

The task of text segmentation is to automatically recognize the boundaries of segments with independent meanings (semantic paragraph) in the text. A semantic paragraph is generally greater than or equal to natural paragraph, and it does not have clear boundaries as natural paragraphs but formed in the process of expressing the topic. In fact, semantic paragraph is consists of several continuous text units (natural paragraphs or sentences). However, the chosen words and the frequency of the words used in the natural paragraphs or sentences of the same semantic paragraph are always extremely similar, because together they support the main idea of the semantic paragraph. So the text segmentation unit should be a "sentence" from the perspective of text topic segmentation.

Similarity Metric. TextTiling is a classical text segmentation algorithm, the basic idea of which is to find the transition point of one topic to another in a text[7]. Despite the fact that the topics or lexicons of different paragraphs fluctuate frequently and randomly in a multi-paragraph text, the most basic change is still in consistent with the segmental boundary.

If the point between sentences called a Gap, then Cohesion Scorer is to measure the continuity of topic in the gap. The former Cohesion Scorer of TextTiling algorithms are: Vector Space Scoring, Vector Space Scoring, and Block Comparison etc.

Assuming text D contains n sentences and k semantic paragraphs, and H stands for text semantic paragraph, s stands for sentence, the composition relationship is as follows:

$$D = \{H_1H_2 \cdots H_k\} = \{s_{i_1} \cdots s_{i_2-1}\}\{s_{i_2} \cdots s_{i_3-1}\} \cdots \{s_{i_k} \cdots s_{i_{k+1}-1}\} \quad (8)$$

$i_1 = 1 \le i_2 \le \cdots i_k \le i_{k+1} - 1 = n$. It is simplified as s_i , s_{i+1} hereafter.

If the similarity $sim(s_i, s_{i+1})$ of adjacent sentences in the text is taken as the measurement of text cohesion, the similarity of adjacent sentences can also reflect the continuity of topic. Obviously, text structure segment point should be gap with low similarity, which means a poor continuity and can be used as the candidate segmentation point. Therefore, attention needs to be focused on those gaps with lower similarity.

If a text is regarded as a complete text vector space, a whole sentence from the text can be taken as basic processing unit and formalized as a text vector. Taking variable <s, w> as lexical item w appeared in sentence s, $s \in S = \{s_1, s_2, \cdots s_n\}$, $w \in W = \{w_1, w_2, \cdots w_m\}$. Constructing "lexicon-sentence" co-occurrence matrix $N=[n(s,w)]_{m\times n}$, n(s, w) means the appearance frequency of lexical item w in sentence s. Using PLSA model to construct similarity probability matrix $H_{s\_s} = H'_{s\_w}H_{s\_w} = [P(s_i, s_j)]_{n\times n}$ between sentences. Extracting $P(s_i, s_{i+1})$, $1\le i\le n-1$ in the similarity probability matrix H_{s_s}, component value as similarity $sim(s_i, s_{i+1})$ of adjacent sentences. Higher similarity means higher cohesion, and oppositely, lower similarity means lower cohesion which implies the end of a topic.

Boundary discovering. The aim to boundary discovering is to determine candidate text segmentation point according to the cohesion between gaps. Boundary discovering method are listed as follows:

Threshold Value Method. Setting a constant θ, if similarity between sentences $sim(s_i, s_j)<\theta$, then s_i, s_j belong to different segments. This method is easy to achieve. Lower error rate could be acquired if a proper θ is picked.

Dynamic constant method. Although threshold value method is easy, some people think an optimum value can hardly be acquired because the θ is set factitiously. So θ can be changed dynamically according to the similarity of adjacent sentences. Assuming that the text to be segmented has n complete sentences, so similarity of adjacent sentences can be expressed as:

$$SimTable = \{Sim_1, Sim_2, \ldots, Sim_i, \ldots, Sim_{n-1}\} \quad (9)$$

$Sim_i = sim(s_i, s_{i+1})$, $1 \le i \le n-1$, given

$$\text{avgSim} = (\text{Sim}_1 + \text{Sim}_2 + \ldots + \text{Sim}_i + \ldots + \text{Sim}_{n-1})/(n-1) \quad (10)$$

$$\text{avgmSim} = ((\text{Sim}_2 - \text{Sim}_1) + \ldots + (\text{Sim}_{n-1} - \text{Sim}_{n-2}))/(n-2) \quad (11)$$

If $\text{avgmSim} \le \text{sim}(s_i, s_{i+1}) \le \text{avgSim}$, then s_i , s_{i+1} are considered to be belonging to different segments.

Partial Minimum Method. Choosing minimum $\text{sim}_{\min}(s_i, s_j)$ from adjacent sentences SimTable, looking for nearest Sim_l and Sim_r respectively from the left to the right, begins with each partial minimum, then using formula (12) to calculate relative depth:

$$d_{rel}(s_i, s_j) = (\text{sim}_l + \text{sim}_r)/(2 \times \text{sim}_{\min}(s_i, s_j)) - 1 \quad (12)$$

Assuming a threshold value α, if relative depth $d_{rel}(s_i, s_{i+1}) > \alpha$, then s_i , s_{i+1} belong to different segments. After the test, Partial Minimum Method is the best algorithm.

Experiment

The evaluation standard of text segmentation algorithm is relatively subjective because there hasn't been any consistent agreements on the position of segment boundary and text segmentation granularity, which add great difficulty to the segmentation result judgment. To overcome this difficulty, some studies connect texts with different contents and set the segment boundary artificially, others set it according to people's evaluation and used majority's opinions as standard. This paper adopts the former method to assure that the performance of all types of text segmentation models are evaluated correctly.

Evaluation Corpus. There has not been any generally open Chinese text segmentation corpus up till now. Therefore, this paper constructed a Chinese corpus consists of 106 articles as evaluation database. These articles are derived from electronic edition of People's Daily covering an extensive range of genre and content like science and technology exposition, biographies, and commentary. Considering the length of Chinese texts, 5 to 8 semantic paragraphs are chosen from each article in the corpus which consists of 25.8 natural paragraphs on average. Each text's standard segmentation model are given in an artificial way. In this experiment, the tested corpus fall into two parts: tested corpus 1 (5 semantic paragraphs) and tested corpus 2 (6 to 8 semantic paragraphs).

Evaluation Method. This paper first used traditional precision and recall F1 value to evaluate the performance of text segmentation method. Precision refers to the proportion that the correct segmentation points takes in all segmentation points in the segmentation results; recall refers to the proportion that the segmentation points correctly judged by the algorithm takes segmentation points in the standard answer. The value of F1 is computed according to the following formula:

$$\text{F1} - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

However, for text segmentation evaluation, traditional precision and recall cannot evaluate the performance of segmentation comprehensively and impartially. The reason mainly lies in that the major concern of precision and recall is absolute matching results, but in fact, the wrong segmentation points near the right segmentation point performs better than distant one, while precision and recall could not tell the difference by treating these two equally.

To overcome the above difficulties, this paper adopted Beeferman et al.'s Pμ assessment method to assess the performance of the text segmentation system. The following is the Pμ assessment method:

$$P_\mu(\text{ref}, \text{hyp}) = \sum_{1 \le i \le j \le n} D_\mu(i, j)(\delta_{\text{ref}}(i, j) \overline{\oplus} \delta_{\text{hyp}}(i, j)) \quad (14)$$

Wherein ref refers to segmentation model in standard answer; hyp refers to segmentation model given by the system; n refers to the number of sentences in the text; $\delta_{ref}(i, j)$ refers to indicator function, when sentence i and j belongs to the same segment in ref, the value of $\delta_{ref}(i, j)$ is 1, otherwise is 0; in a similar way, when sentence i and j belongs to the same segment in hyp, the value of $\delta_{hyp}(i, j)$ is 1, otherwise is 0. The above formula conducts a XOR on the value of $\delta_{ref}(i, j)$ and $\delta_{hyp}(i, j)$. Dμ refers to

the distance probability distribution of the randomly picked sentence in the text, which value depends on parameter μ, and Dμ refers to exponential distribution of μ:

$$D_{\mu}(i, j) = r_{\mu} e^{-\mu|i-j|} \quad (15)$$

In which r_{μ} is a normalizing factor. Parameter $1/\mu$ is set as 11, which means the average number of sentences in the concentrated segment in the tested text.

Evaluation Result. In this experiment, TextTiling algorithm by Hearst and LSA by Choi are used as Baseline system. Table 1 shows that TextTiling algorithm provides better Recall and LSA provides better Precision, but as for F1 value and Pμ value, PLSA-TextTiling works better, which is consistent with the above analysis.

Table 1 Three models comparing

Algorithm	Evaluation etric	Testing corpus 1 (5 segments)	Testing corpus 2(6~8segment)	Average
PLSA-TextTiling	Precision	0.485	0.452	0.470
	Recall	0.455	0.544	0.495
	F1	0.470	0.494	0.482
	Pμ	0.840	0.906	0.87
TextTiling (Hearst)	Precision	0.425	0.350	0.391
	Recall	0.452	0.568	0.504
	F1	0.438	0.433	0.441
	Pμ	0.781	0.842	0.808
LSA(Choi)	Precision	0.485	0.457	0.473
	Recall	0.440	0.514	0.474
	F1	0.462	0.483	0.473
	Pμ	0.799	0.870	0.832

Summary

Text segmentation system is consisted of three parts: basic block setting, similarity metric and boundary discovering. This paper takes the complete sentence as basic block, use PLSA model to calculated similarity metric basing on the idea of TestTiling. Several approaches to discovering boundaries are tried. Finally, it puts forward a PLSA-TextTiling text segmentation model. The result of the experiment shows that this model operates better than TextTiling and LSA algorithm in general.

References

[1] Choi F Y Y: Proc. the North American Chapter of the Association for Computational Linguistics (Seattle, USA, 2000). Vol. 1, p.26.
[2] A Kehagias, A Nicolaou, P Fragkou: Mathematical and Computer Modelling, Vol. 39 (2004) No.2, p.209.
[3] Gina-Anne Levow: Proc. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (Boston, USA, 2004). Vol. 1, p.137.
[4] D Blei, P Moreno: Proc. the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, USA,2001). Vol. 1, p.343.
[5] Choi F Y Y, Wiemer-Hastings P, Moore J: Proc. the 6th Conference on Empirical Methods in Natural Language Processing (Pittsburgh, USA, June 3-4, 2001). Vol.1, p.109.
[6] Thomas Hofmann: Proc. the 15th Annual Conf on Uncertainty in Artificial Intelligence (San Francisco, USA, 1999). Vol.1, p.289.
[7] MA Hearst: Computational Linguistics, Vol. 23 (1997) No.1, p.33.

Mechatronics Engineering, Computing and Information Technology
10.4028/www.scientific.net/AMM.556-562

Text Segmentation Based on PLSA-TextTiling Model
10.4028/www.scientific.net/AMM.556-562.4018