

In this paper, we use the PPM framework to identify text segments. To this end, we obtain the posterior joint probability of an observed text and its segmentation as a product of two terms:

- (a) the probability of the segmentation, described by appropriate prior cohesions, and
- (b) the conditional (given the segmentation) probability of the *sentence similarity matrix*, described by an appropriate homogeneity function.

Note that we use PPMs to assign probabilities to *two-dimensional* structures (the sentence similarity matrices) rather than to one-dimensional sequences. The negative logarithm of the joint probability is the *segmentation cost*, which is minimized by a fast dynamic programming algorithm (compare to the use of computationally demanding Markov chain Monte Carlo algorithms in [11,12]). The homogeneity function we use depends on some parameters which are estimated from training data; as far as we know, this approach has not been previously used in conjunction with PPMs.

In Section 2, we describe the PPMs we use to tackle the text segmentation problem. In Section 3, we present a dynamic programming algorithm to solve the problem. In Section 4, we present some experiments to evaluate our algorithm. Finally, in Section 5, we discuss our results, review some work related to our own, and present future research directions.

2. PROBLEM FORMULATION

2.1. Representation

Consider a text with T sentences. A *segmentation* of the text is a partition of $\{1, 2, \dots, T\}$ into K contiguous *segments*: $\{1, 2, \dots, t_1\}$, $\{t_1 + 1, t_1 + 2, \dots, t_2\}$, \dots , $\{t_{K-1} + 1, t_{K-1} + 2, \dots, T\}$; and t_0, t_1, \dots, t_K are the *segment boundaries*¹ which satisfy

$$0 = t_0 < t_1 < \dots < t_{K-1} < t_K = T.$$

A concise representation of the segmentation is given by the vector $\mathbf{t} = (t_0, t_1, \dots, t_K)$; note that vector length K (i.e., the number of segments) is variable, but it satisfies $K \leq T$. We will denote the set of all possible segmentations of $\{1, 2, \dots, T\}$ by Φ_T .

Assume that the text has a vocabulary of L distinct words (common *uninformative* words such as “and”, “the”, etc., are not included). The text can be represented by a $T \times L$ matrix \mathbf{c} , where (for $t = 1, 2, \dots, T$ and $l = 1, 2, \dots, L$)

$$c_{t,l} = \begin{cases} 1, & \text{iff the } l^{\text{th}} \text{ word appears in the } t^{\text{th}} \text{ sentence,} \\ 0, & \text{else.} \end{cases}$$

The *sentence similarity matrix* of the text is a $T \times T$ matrix \mathbf{d} , where

$$d_{t,t} = 0, \quad \text{for } 1 \leq t \leq T, \quad \text{and} \quad d_{s,t} = \begin{cases} 1, & \text{if } \sum_{l=1}^L c_{s,l}c_{t,l} > 0, \\ 0, & \text{if } \sum_{l=1}^L c_{s,l}c_{t,l} = 0, \end{cases} \quad \text{for } 1 \leq s \neq t \leq T.$$

In other words, $d_{s,t} = 1$ when the s^{th} and t^{th} sentence have at least one word in common (but note that the diagonal elements $d_{t,t}$ are set equal to zero). We will denote the set of all possible sentence similarity matrices (of dimension $T \times T$) by Ψ_T . The resulting matrix \mathbf{d} has zeros and ones arranged in a characteristic pattern which corresponds to the structure of the text. In

¹We assume that segment boundaries always appear at the ends of sentences.

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات