Modeling the Statistical Behavior of Lexical Chains to Capture Word Cohesiveness for Automatic Story Segmentation

Shing-kai Chan, Lei Xie, Helen Mei-ling Meng

Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{chansk, lxie, hmmeng}@se.cuhk.edu.hk

Abstract

We present a mathematically rigorous framework for modeling the statistical behavior of lexical chains for automatic story segmentation of broadcast news audio. Lexical chains were first proposed in [1] to connect related terms within a story, as an embodiment of lexical cohesion. The vocabulary within a story tends to be cohesive, while a change in the vocabulary distribution tends to signify a topic shift that occurs across a story boundary. Previous work focused on the concept and nature of lexical chains but performed story segmentation based on arbitrary thresholding. This work proposes the use of the log-normal distribution to capture the statistical behavior of lexical chains, together with data-driven parameter selection for lexical chain formation. Experimentation based on the TDT-2 Mandarin Corpus shows that the proposed statistical model leads to better story segmentation, where the F1-measure increased from 0.468 to 0.641.

Index Terms: story segmentation, spoken document retrieval, Chinese

1. Introduction

Story segmentation is the task of segmenting a text into distinctive units known as stories, each of which is coherent within itself. It is a prerequisite for a wide range of speech and language information retrieval tasks, namely topic tracking, clustering, indexing and retrieval. In particular, broadcast news, which is delivered in continuous video/audio streams, needs segmentation before information can be retrieved. Segmentation done manually requires human segmenters to watch/listen through the whole video/audio stream, which takes so huge an amount of time that makes it an intractable task. To perform story segmentation, there are three categories of cues available: lexical cues from transcriptions, prosodic cues from the audio stream and video cues such as anchor face and color histograms.

Among the three types of cues, lexical cues are the most generic since they can work on text and multimedia sources. The main approaches include word cohesiveness [2], use of cue phrases [3] and Hidden Markov Models [4]. We focus on an approach based on lexical chaining, that embodies word cohesiveness[1]. A lexical chain links up related words in a textual document. Intuitively, the vocabulary used in the same story is more cohesive, while a shift in vocabulary can be observed across a story boundary. Consequently most lexical chains should be embedded within a story and few chains straddle a story boundary. Taking advantage of this feature one can perform story segmentation. [1] proposed using lexical chains computed by a thesaurus to determine textual structure. Lexical chaining was further studied by Stokes [5], who incorporates a comprehensive set of semantic relationships for story segmentation. However, Stoke reveals that it is "counterintuitive and disappointing" that employing more semantic relationships has a negative effect on the performance of segmentation, owing to noise incurred by additional semantic relationships. In this work, we decide to build lexical chains based on word repetitions only. Repetitions, according to Stokes, exhibit the best performance for story segmentation.

Stokes' work discovers boundaries by chaining up terms and finding points where the count of chain starts and ends (known as boundary strength) achieves local maxima. A number of parameters including the maximum chaining distance between related terms and the boundary strength threshold under which a hypothesized boundary is discarded are fixed at some values determined by manual observations. The values are not determined in a rigorous manner and do not guarantee optimality across different corpora. Our work extends Stokes' work by defining a statistically robust parameter determination procedure.

2. Corpus

We experiment with the TDT-2 Mandarin Corpus,[1] which contains about 46 hours of Voice of America (VOA) Mandarin Chinese broadcast news from February 1998 to June 1998. The audio files are accompanied with textual transcripts from the Dragon Automatic Speech Recognizer and metadata manually marked story boundaries. In the corpus there are two main types of news programs – (a) the long programs are about one hour in duration, (b) the short programs are five to ten minutes long. Our analysis shows that the mean story length of long programs is different from that of short programs. Hence we estimate distribution parameters and perform boundary discovery separately.

We divide the corpus into three portions: a half as training set for determining probability distribution parameters, a quarter as development set for obtaining *a priori* information about number of boundaries in each type of program and the remaining quarter as testing set for final evaluation. We allocate large share as training data set to ensure adequacy for parameter estimation relating to the lexical chains. To ensure uniformity of data, we maintain the same fraction of long and short stories across the three sets. In evaluating the segmentation per-

[1] Refer to http://projects.ldc.upenn.edu/TDT2/ for details.

formance, we consider a detected story boundary as correct if it lies within a fifteen-second window on each side of a hand-marked true boundary. This is the standard tolerance defined in the TDT-2 tasks.

3. Parameter Estimation for Maximum Chaining Distance

Lexical chains aim to connect related terms *within* a story. However, some terms in a news story may re-appear in another story. Hence it is necessary to impose a *maximum chaining distance* ϕ beyond which no lexical chains are formed. If ϕ is too long, we may have many lexical chains spanning across two or more stories. On the other hand, if ϕ is too short, repeating terms remain disconnected. St.Onge and Hirst [6] and Stokes [5] imposed a fixed ϕ by manual observation. This tends to be a bit arbitrary for experimentation. Hence we devise a data-driven method to estimate an appropriate value for ϕ. First we define a "link" to be bonding of two terms adjacent in a chain. A correct link is one whose endpoints lie within the same story. An incorrect link is one whose endpoints lie in different stories. We vary ϕ from zero seconds to the maximum program length of the training set. For each parameter value, we compute the following based on the training set:

$$\text{recall} = \frac{\text{correct links captured with } \phi}{\text{correct links in perfect link formation}} \quad (1)$$

$$\text{precision} = \frac{\text{correct links captured with } \phi}{\text{all links captured with } \phi}. \quad (2)$$

We also compute the F1-measure, i.e., the harmonic mean of recall and precision, for each ϕ value. The value achieving highest F1-measure is "the optimal ϕ" for lexical chain formation in the development set and test set. The ϕ is determined separately for short programs (31.6sec) and long programs(130.9sec).

4. Statistical Behavior of Lexical Chains

A lexical chain links up repeating (related) terms where a chain starts at the first appearance of an informative term and ends at the last appearance of the term. Morris and Hirst [1] pointed out that a high concentration of starting and/or ending points of lexical chains is a good indication as a strong boundary. A story boundary may be detected by locating temporal landmarks before which many lexical chains end and after which many chains start. We conceive that as a news story begins and progresses, we should observe a dwindling number of lexical chain starts. As the story nears its end, we should observe a rising number of lexical chain ends. This motivates us to search for a probability model that can properly capture the statistical behavior of lexical chains. We believe that such a model is essential for improving the performance of automatic story segmentation using lexical chains. We propose the following mathematic formulation:

B denotes the set of all story boundaries. At any time instance t in a news program, we define *feature points* $\mathcal{C}(t)$ as the union of set of all starts following t ($\mathcal{F}(t) = \{t_1, t_2, \ldots, t_{n_F}\}$), and all chain ends preceding t ($\mathcal{P}(t) = \{t_{-1}, t_{-2}, \ldots, t_{-n_P}\}$):

$$\mathcal{C}(t) = \mathcal{F}(t) \cup \mathcal{P}(t). \quad (3)$$

At a story boundary $t_0 \in B$, we expect start chains $\mathcal{F}(t_0)$ to spawn off from this point, and preceding t_0 end chains $\mathcal{P}(t_0)$ to taper off towards t_0. Also, the concentration of feature points $\mathcal{C}(t_0)$ in the vicinity of t_0 should be higher than the concentration of feature points farther apart. This inspires us to model the arrival of start chains following t_0 by a probability distribution. Likewise, we can model the arrival of end chains preceding t_0 by a probability distribution. We assume that each feature point is generated by an i.i.d. from a boundary point t_0, governed by a model ω following a distribution $D(\cdot)$ and parameters $\boldsymbol{\theta}$ (i.e., $P(t_i|\chi_B(t);\omega) \sim D(\boldsymbol{\theta})$):

$$P(\mathcal{C}(t_0)|\chi_B(t_0);\omega) = \prod_{i=1}^{|\mathcal{P}(t_0)|} P(t_{-i}|\chi_B(t_0);\omega) \prod_{i=1}^{|\mathcal{F}(t_0)|} P(t_i|\chi_B(t_0);\omega), \quad (4)$$

where $\chi_B(t_0)$ is an indicator function showing the existence of boundary at t_0, i.e., $\chi_B(t_0) = 1$when $t_0 \in B$, and 0 otherwise. Under this assumption, we should observe that the feature points $\mathcal{C}(t)$ follow a probability distribution consistently at all boundaries $t \in B$. This assumption is supported by our observation in section 4.1.

The existence of boundary can be estimated by maximum likelihood estimation (MLE) under this generative assumption. Alternatively, and more reasonably, maximum a posterior (MAP) estimation is a more direct and straightforward measure of the probability of boundary occurrence:

$$P(\chi_B(t_0)|\mathcal{C}(t_0);\omega) \propto P(\mathcal{C}(t_0)|\chi_B(t_0);\omega) P(\chi_B(t_0);\omega) \quad (5)$$

In this paper, we assume that boundaries are equally likely to occur anywhere in a program. The uniformity of prior $P(\chi_B(t_0);\omega)$ implies the equivalence of MAP estimation and MLE. The boundary discovery problem can therefore be reduced to finding points t_0 whose likelihood of boundary existence $P(\mathcal{C}(t_0)|\chi_B(t_0);\omega)$ given our observation of distribution of feature points $\mathcal{C}(t_0)$ are at local maxima.

4.1. Selecting a Probability Distribution

In order to estimate the likelihood of boundary existence, we must first find out how lexical chain starts/ends are "generated" from a boundary $P(\mathcal{C}(t_0)|\chi_B(t_0);\omega)$. To address this problem, we have to find a suitable probability distribution $D(\cdot)$ together with the model parameters $\boldsymbol{\theta}$. We identified a list of candidate distributions which are possibly in accord with the feature point distribution. The list includes the two most common distributions:

- Single-sided Normal distribution
- Exponential distribution

and the following common continuous time distributions, whose shapes are similar to that observed from the chain distribution data (Fig.1):

- Generalized Pareto distribution
- Gamma distribution
- Weibull distribution
- Log-normal distribution

Table 1 shows the log-likelihood that the training data is generated from each of the candidate distributions. We see that in most cases the log-normal distribution best fits the data. The Weibull distribution obtains the next-to-best result in terms of log-likelihood. Hence in the subsequent experiments we applied both the log-normal and Weibull distribution for comparison.

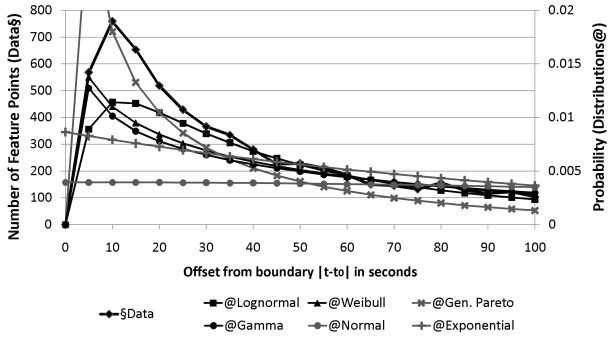

Figure 1: This graphs shows an example of the distribution of feature points (in this case, the feature points shown are the chain starts following real boundaries in long programs) and the hypothesized distributions (Normal, Exponential, Generalized Pareto, Gamma, Weibull and Log-normal). The hypothesized distributions are parameterized at their maximum likelihood estimation of the data. It is clear that normal and exponential distributions do not appear to fit the data well, while the other four distributions have shapes more similar to the distribution of real data. In this example, the best fit is by the log-normal distribution while the second best fit is by the Weibull distribution (data shown in Table 1).

Table 1: *MLE estimation of the data with common distributions. Here* $\mathbf{P} = \bigcup_{t \in B} \mathcal{P}(t)$ *and* $\mathbf{F} = \bigcup_{t \in B} \mathcal{F}(t)$.

Distribution	Log-likelihood (MLE)				
	Short Programs		Long Programs		Sum
	P	F	P	F	
Normal	-16340	-15866	-61026	-61198	-154430
Exponential	-13861	-12929	-52188	-51757	-130735
Pareto	-13821	-12910	-51568	-50534	-128833
Gamma	-13636	-12928	-51783	-50478	-128825
Weibull	-13690	-12923	-51637	**-50303**	-128553
Log-normal	**-13612**	**-12903**	**-51536**	-50467	**-128518**

5. Story Segmentation Methodology

5.1. Candidate Term Extraction

Lexical chain formation in Chinese spoken document transcripts faces special challenges. First, Chinese does not have explicit word delimiters. Word tokenization alone is a research topic in and of itself. Second, news stories often contain a large number of out-of-vocabulary (OOV) words, most of which are proper nouns, e.g. person names and place/organization names. Failing to identify these words may lead to errors in chaining. To solve these two problems, we follow the algorithms in [7]. We first perform word segmentation by matching words in the CALLHOME lexicon with a greedy algorithm. Since OOV words often appear in recognition transcriptions as a series of single Chinese characters, we implemented an algorithm to combine the singletons together, followed by a filtering mechanism, to form candidate terms. We further applied POS tagging to constrain the candidate terms to nouns, as they are more indicative of the topic. The extracted candidate terms are used to form lexical chains by following the procedures described in sections 3 and 4.

5.2. Boundary Hypothesis and Selection

To quantify the likelihood of boundary $P(\mathcal{C}(t_0)|\chi_B(t_0);\omega)$ at a point t_0, we define *boundary score* $S(t_0)$ by measuring how much the occurrence of chains around t_0 matches the log-normal distribution. The boundary score is defined by:

$$S(t_0) = \sum_{t \in \mathcal{P}(t_0)} s_{D(\theta_1)}(t, t_0) + \sum_{t \in \mathcal{F}(t_0)} s_{D(\theta_2)}(t, t_0), \quad (6)$$

depending on the choice of distribution $D(\cdot)$ and the parameters $\boldsymbol{\theta}$. For example, if the distribution chosen is log-normal $LN(\mu, \sigma)$, we have

$$s_{LN(\mu,\sigma)}(t, t_0) = \frac{e^{-(\ln|t-t_0|-\mu)^2/(2\sigma^2)}}{|t-t_0|\sigma\sqrt{2\pi}}. \quad (7)$$

Likewise, for Weibull distribution $W(k, \lambda)$, we define

$$s_{W(k,\lambda)}(t, t_0) = \frac{k}{\lambda}\left(\frac{|t-t_0|}{\lambda}\right)^{k-1} e^{-(|t-t_0|/\lambda)^k}. \quad (8)$$

An example showing how the boundary score at a hypothesized point t_0 is calculated is illustrated in Fig.2. We compute the boundary score at each utterance boundary (short pause) and plot the boundary score over time (Fig.3). This score needs to be normalized before we can find boundaries based on the score. The reason is that towards both ends of the news program, the observation we can make is less than what we can observe in the middle of the program. As a result the boundary scores near both ends of the stories are lower and we need to boost the value for fair scoring. In addition, in some parts of the news, the output of speech recognizer makes so many errors that the number of words available for training is reduced. In this case, fewer feature points are available for calculating the boundary score. To tackle these problems, we minus the moving average of the boundary score from $S(t)$ to obtain *normalized score* $S_N(t)$. Boundaries are hypothesized when we observe local maxima in $S_N(t)$. We assume *a priori* knowledge of the total number of story boundaries (n) in a program and pick the n-best candidates among the hypothesized boundaries. Fig.3 illustrates the boundary scoring scheme. To find n for each type of program, we perform grid search to find the value that achieves the highest F1-measure among the 5-minute, 10-minute and 1-hour programs.

6. Experimental Results

We compare four approaches to automatic story segmentation (see Table 2) (1) All utterance boundaries with a pause exceeding a threshold tuned from the training and development set are hypothesized as a story boundary; (2) A re-implementation of Stokes' scheme[5]; (3) Our proposed approach using the log-normal distribution; (4) Our proposed approach using the Weibull distribution. The results in terms of F1-measure are shown in Table 2.

Results indicate that the log-normal distribution achieves the best segmentation performance overall. The Weibull distribution achieves second-best, followed by the simple approach of pause-based story segmentation. The use of pauses is particularly effective for short programs, because the duration of each story is relatively short, it is unlikely that the anchor inserts a long within-story break. Story boundaries can possibly be detected by human behavior of inserting longer break across topic

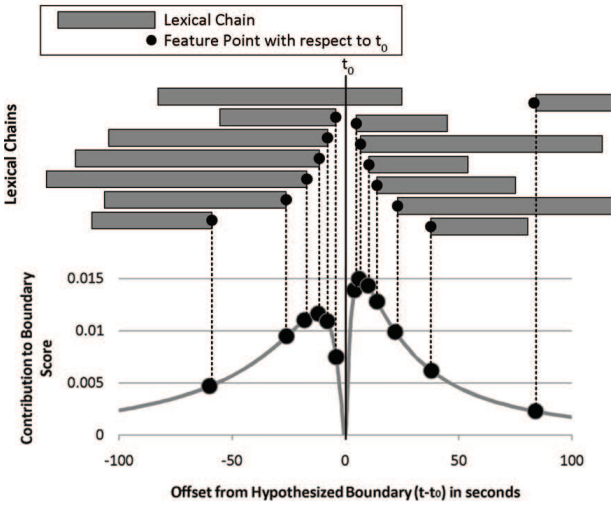

Figure 2: At each utterance boundary t_0, we apply equation (6) to calculate the boundary score $S(t_0)$. The figure shows an example of how this score is calculated. Given the utterance boundary t_0, we can locate the feature points $\mathcal{C}(t_0)$. Each $t \in \mathcal{C}(t_0)$ contributes an amount $s_{D(\cdot)}(t, t_0)$to the boundary score by equation (7) or (8). According to equation (6), the boundary score at t_0 is the sum of all these contributions.

Table 2: *Results of experiment comparing different story segmentation schemes*

Scheme	Long Prog.	Short Prog.	Overall
(1) Pause only	0.372	**0.777**	0.590
(2) Stokes	0.445	0.505	0.468
(3) Proposed (Log-normal)	**0.535**	0.746	**0.641**
(4) Proposed (Weibull)	0.460	0.733	0.590

shift with high accuracy. However the performance declines sharply for long programs, as it is impossible for the anchor to speak uninterruptedly from the beginning to the end, thus longer within-story pauses can be observed. This reduces the reliability of using pauses to mark topic shift. The log-likelihood of the distribution of feature points to the log-normal distribution is higher than that to Weibull distribution in the training set, and the segmentation result is in accord with this observation. We conclude that using the proposed scheme with the log-normal distribution provides the best story segmentation result.

7. Conclusions and Future Work

The paper has presented a scheme to perform Chinese broadcast news segmentation. By assuming the production of lexical chain endpoints in a generative manner near story boundaries, we have modeled the occurrence of lexical chain endpoints by the log-normal distribution. We take advantage of this observation to determine story boundaries by maximum likelihood estimation (MLE). Furthermore, we have identified the weakness of previous works in determining parameters and we have clearly laid down the procedure to (1) estimate maximum chaining length, (2) find the probability distribution of chain feature points and estimate the parameters, and (3) calculate and normalize boundary scores to discover story boundaries. The experimental results show our approach outperforms previous approach using lexical chaining. Our approach also performs

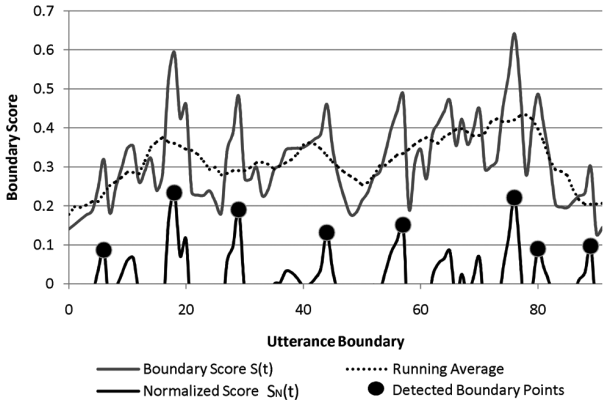

Figure 3: This figure summarizes how to locate boundaries with the scoring scheme. Initially the boundary score $S(t)$ is calculated, whose value is subtracted by the running average to obtain the normalized score $S_N(t)$. We have the *a priori* knowledge that the number of story boundaries is eight, and therefore the utterance boundaries corresponding to the eight highest local maxima are selected as the story boundaries.

more stably across news programs of different durations. In the future, we aim at salvaging the missing boundaries we are not able to capture in this work by incorporating other sets of feature, in particular prosodic features such as long pauses, speaker changes and pitch resets, and possibly experimenting the segmentation scheme with datasets of other languages.

8. Acknowledgments

This research is partially supported by the CUHK Shun Hing Institute of Advanced Engineering and is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

9. References

[1] Morris, J. and Hirst, G., "Lexical cohesion computed by thesaural relations as an indicator of the structure of text", Computational Linguistics, 17(1), 21–48, 1991.

[2] Hearst, M.A., "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages", Computational Linguistics, 23(1), 33–64, 1997.

[3] Reynar, J.C., "Statistical models for topic segmentation", Proc. 37th annual meeting of the ACL on Computational Linguistics, 357–364, 1999.

[4] Yamron, J.P., Carp, I., Gillick, L., Lowe, S., van Mulbregt, P. and D.S. Inc. Newton, M.A., "A hidden Markov model approach to text segmentation and event tracking", Proc. ICASSP 1998, 333–336, 1998.

[5] Stokes, N., Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain, PhD thesis, University College Dublin, 2004.

[6] Hirst, G. and St-Onge, D., "Lexical chains as representations of context for the detection and correction of malapropisms", WordNet: An Electronic Lexical Database, 305–332, 1998.

[7] Li, D., Lo, W.K. and Meng, H., "Initial Experiments on Automatic Story Segmentation in Chinese Spoken Documents Using Lexical Cohesion of Extracted Named Entities", Proc. ISCSLP 2006, 693–703, 2006.