Automatic Story Segmentation using a Bayesian Decision Framework for Statistical Models of Lexical Chain Features

Wai-Kit Lo
The Chinese University of Hong Kong, Hong Kong, China
wklo@se.cuhk.edu.hk

Wenying Xiong
The Chinese University of Hong Kong, Hong Kong, China
wyxiong@se.cuhk.edu.hk

Helen Meng
The Chinese University of Hong Kong, Hong Kong, China
hmmeng@se.cuhk.edu.hk

Abstract

This paper presents a Bayesian decision framework that performs automatic story segmentation based on statistical modeling of one or more lexical chain features. Automatic story segmentation aims to locate the instances in time where a story ends and another begins. A lexical chain is formed by linking coherent lexical items chronologically. A story boundary is often associated with a significant number of lexical chains ending before it, starting after it, as well as a low count of chains continuing through it. We devise a Bayesian framework to capture such behavior, using the lexical chain features of start, continuation and end. In the scoring criteria, lexical chain starts/ends are modeled statistically with the Weibull and uniform distributions at story boundaries and non-boundaries respectively. The normal distribution is used for lexical chain continuations. Full combination of all lexical chain features gave the best performance (F1=0.6356). We found that modeling chain continuations contributes significantly towards segmentation performance.

1 Introduction

Automatic story segmentation is an important precursor in processing audio or video streams in large information repositories. Very often, these continuous streams of data do not come with boundaries that segment them into semantically coherent units, or stories. The story unit is needed for a wide range of spoken language information retrieval tasks, such as topic tracking, clustering, indexing and retrieval. To perform automatic 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. Previous approaches include TextTiling (Hearst 1997) that monitors changes in sentence similarity, use of cue phrases (Reynar 1999) and Hidden Markov Models (Yamron 1998). In addition, the approach based on lexical chaining captures the content coherence by linking coherent lexical items (Morris and Hirst 1991, Hirst and St-Onge 1998). Stokes (2004) discovers boundaries by chaining up terms and locating instances of time where the count of chain starts and ends (boundary strength) achieves local maxima. Chan *et al.* (2007) enhanced this approach through statistical modeling of lexical chain starts and ends. We further extend this approach in two aspects: 1) a Bayesian decision framework is used; 2) chain continuations straddling across boundaries are taken into consideration and statistically modeled.

2 Experimental Setup

Experiments are conducted using data from the TDT-2 Voice of America Mandarin broadcast. In particular, we only use the data from the long programs (40 programs, 1458 stories in total), each of which is about one hour in duration. The average number of words per story is 297. The news programs are further divided chronologically into training (for parameter estimation of the statistical models), development (for tuning decision thresholds) and test (for performance evaluation) sets, as shown in Figure 1. Automatic speech recognition (ASR) outputs that are provided in the TDT-2 corpus are used for lexical chain formation.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 265–268,
Suntec, Singapore, 4 August 2009.

The story segmentation task in this work is to decide whether a hypothesized utterance boundary (provided in the TDT-2 data based on the speech recognition result) is a story boundary. Segmentation performance is evaluated using the F1-measure.

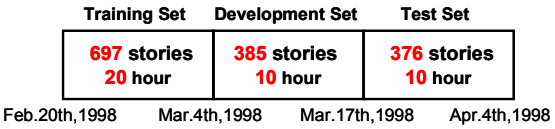

Figure 1: Organization of the long programs in TDT-2 VOA Mandarin for our experiments.

3 Approach

Our approach considers utterance boundaries that are labeled in the TDT-2 corpus and classifies them either as a story boundary or non-boundary.

We form lexical chains from the TDT-2 ASR outputs by linking repeated words. Since words may also repeat across different stories, we limit the maximum distance between consecutive words within the lexical chain. This limit is optimized according to the approach in (Chan *et al.* 2007) based on the training data. The optimal value is found to be 130.9sec for long programs.

We make use of three lexical chain features: chain starts, continuations and ends. At the beginning of a story, new words are introduced more frequently and hence we observe many lexical chain starts. There is also tendency of many lexical chains ending before a story ends. As a result, there is a higher density of chain starts and ends in the proximity of a story boundary. Furthermore, there tends to be fewer chains straddling across a story boundary. Based on these characteristics of lexical chains, we devise a statistical framework for story segmentation by modeling the distribution of these lexical chain features near the story boundaries.

3.1 Story Segmentation based on a Single Lexical Chain Feature

Given an utterance boundary with the lexical chain feature, X, we compare the conditional probabilities of observing a boundary, B, or non-boundary, $\overline{B}$, as

$$P(B \mid X) \gtrless P(\overline{B} \mid X). \quad (1)$$

where X is a single chain feature, which may be the chain start (S), chain continuation (C), *or* chain end (E).

By applying the Bayes' theorem, this can be rewritten as a likelihood ratio test,

$$\frac{P(X \mid B)}{P(X \mid \overline{B})} \gtrless \theta_x \quad (2)$$

for which the decision threshold is $\theta_x = P(\overline{B})/P(B)$, dependent on the a priori probability of observing boundary or a non-boundary.

3.2 Story Segmentation based on Combined Chain Features

When multiple features are used in combination, we formulate the problem as

$$P(B \mid S,E,C) \gtrless P(\overline{B} \mid S,E,C). \quad (3)$$

By assuming that the chain features are conditionally independent of one another (i.e., $P(S,C,E|B) = P(S|B)\,P(C|B)\,P(E|B)$), the formulation can be rewritten as a likelihood ratio test

$$\frac{P(S \mid B)P(E \mid B)P(C \mid B)}{P(S \mid \overline{B})P(E \mid \overline{B})P(C \mid \overline{B})} \gtrless \theta_{SEC}. \quad (4)$$

4 Modeling of Lexical Chain Features

4.1 Chain starts and ends

We follow (Chan *et al.* 2007) to model the lexical chain starts and ends at a story boundary with a statistical distribution. We apply a window around the candidate boundaries (same window size for both chain starts and ends) in our work. Chain features falling outside the window are excluded from the model. Figure 2 shows the distribution when a window size of 20 seconds is used. This is the optimal window size when chain start and end features are combined.

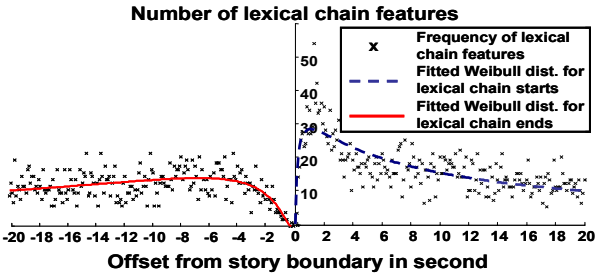

Figure 2: Distribution of chain starts and ends at known story boundaries. The Weibull distribution is used to model these distributions.

We also assume that the probability of seeing a lexical chain start / end at a particular instance is independent of the starts / ends of other chains. As a result, the probability of seeing a sequence of chain starts at a story boundary is given by the product of a sequence of Weibull distributions

$$P(S \mid B) = \prod_{i=1}^{N_s} \frac{k}{\lambda}\left(\frac{t_i}{\lambda}\right)^{k-1} e^{-\left(\frac{t_i}{\lambda}\right)^k}, \quad (5)$$

where S is the sequence of time with chain starts ($S=[t_1, t_2, \ldots t_i, \ldots t_{Ns}]$), k_s is the shape, λ_s is the scale for the fitted Weibull distribution for chain starts, N_s is the number of chain starts. The same formulation is similarly applied to chain ends.

Figure 3 shows the frequency of raw feature points for lexical chain starts and ends near utterance boundaries that are non-story boundaries. Since there is no obvious distribution pattern for these lexical chain features near a non-story boundary, we model these characteristics with a uniform distribution.

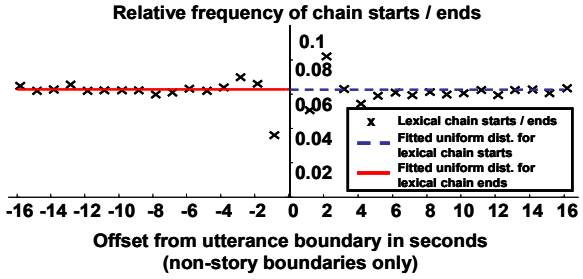

Figure 3: Distribution of chain starts and ends at utterance boundaries that are non-story boundaries.

4.2 Chain continuations

Figure 4 shows the distributions of chain continuations near story boundary and non-story boundary. As one may expect, there are fewer lexical chains that straddle across a story boundary (the curve of $P(C \mid B)$) when compared to a non-story boundary (the curve of $P(C \mid \overline{B})$). Based on the observations, we model the probability of occurrence of lexical chains straddling across a given story boundary or non-story boundary by a normal distribution.

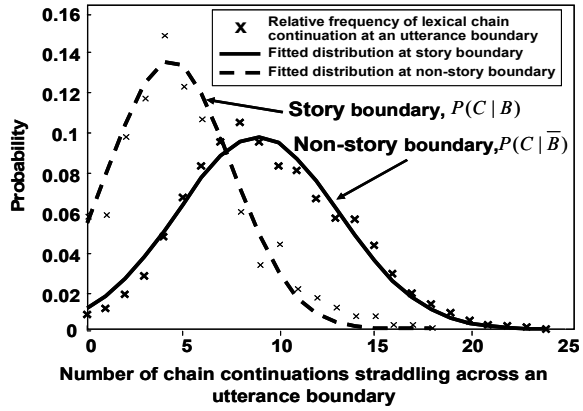

Figure 4: Distributions of chain continuations at story boundaries and non-story boundaries.

5 Story Segmentation based on Combination of Lexical Chain Features

We trained the parameters of the Weibull distribution for lexical chain starts and ends at story boundaries, the uniform distribution for lexical chain start / end at non-story boundary, and the normal distribution for lexical chain continuations. Instead of directly using a threshold as shown in Equation (2), we optimize on the parameter n, which is the optimal number of top scoring utterance boundaries that are classified as story boundaries in the development set.

5.1 Using Bayesian decision framework

We compare the performance of the Bayesian decision framework to the use of likelihood only $P(X|B)$ as shown in Figure 5. The results demonstrate consistent improvement in F1-measure when using the Bayesian decision framework.

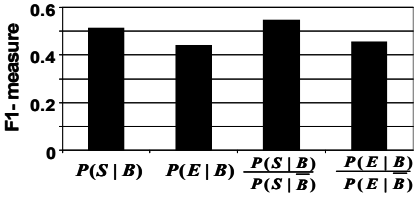

Figure 5: Story segmentation performance in F1-measure when using single lexical chain features.

5.2 Modeling multiple features jointly

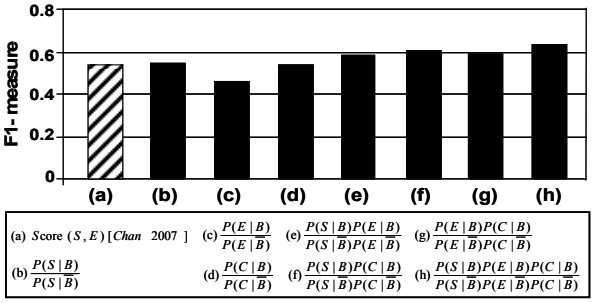

Figure 6: Results of F1-measure comparing the segmentation results using different statistical models of lexical chain features.

We further compare the performance of various scoring methods including single and combined lexical chain features. The baseline result is obtained using a scoring function based on the likelihoods of seeing a chain start or end at a story boundary (Chan *et al.* 2007) which is denoted as *Score(S, E)*. Performance from other methods based on the same dataset can be referenced from Chan *et al.* 2007 and will not be repeated here. The best story segmentation performance is achieved by combining all lexical chain features which achieves an F1-measure of 0.6356. All improvements have been verified to be statistically significant (α=0.05). By comparing the results of (e) to (h), (c) to (g), and (b) to (f), we can see that lexical chain continuation feature contributes significantly and consistently towards story segmentation performance.

5.3 Analysis

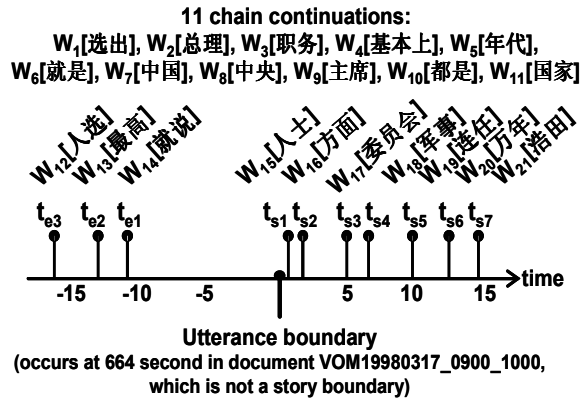

Figure 7: Lexical chain starts, ends and continuations in the proximity of a non-story boundary. W_i[xxxx] denotes the *i*-th Chinese word "xxxx".

Figure 7 shows an utterance boundary that is a non-story boundary. There is a high concentration of chain starts and ends near the boundary which leads to a misclassification if we only combine chain starts and ends for segmentation. However, there are also a large number of chain continuations across the utterance boundary, which implies that a story boundary is less likely. The full combination gives the correct decision.

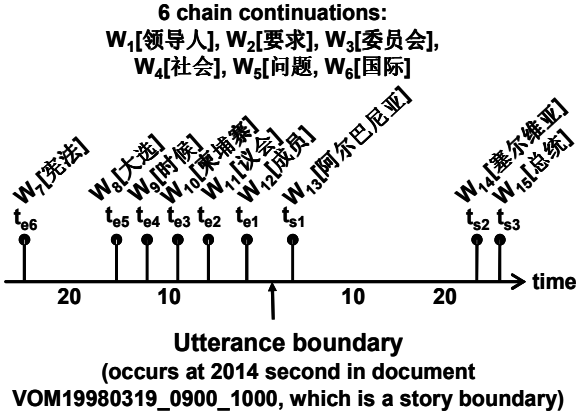

Figure 8: Lexical chain starts, ends and continuations in the proximity of a story boundary.

Figure 8 shows another example where an utterance boundary is misclassified as a non-story boundary when only the combination of lexical chain starts and ends are used. Incorporation of the chain continuation feature helps rectify the classification.

From these two examples, we can see that the incorporation of chain continuation in our story segmentation framework can complement the features of chain starts and ends. In both examples above, the number of chain continuations plays a crucial role in correct identification of a story boundary.

6 Conclusions

We have presented a Bayesian decision framework that performs automatic story segmentation based on statistical modeling of one or more lexical chain features, including lexical chain starts, continuations and ends. Experimentation shows that the Bayesian decision framework is superior to the use of likelihoods for segmentation. We also experimented with a variety of scoring criteria, involving likelihood ratio tests of a single feature (i.e. lexical chain starts, continuations or ends), their pair-wise combinations, as well as the full combination of all three features. Lexical chain starts/ends are modeled statistically with the Weibull and normal distributions for story boundaries and non-boundaries. The normal distribution is used for lexical chain continuations. Full combination of all lexical chain features gave the best performance (F1=0.6356). Modeling chain continuations contribute significantly towards segmentation performance.

Acknowledgments

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. We would also like to thank Professor Mari Ostendorf for suggesting the use of continuing chains and Mr. Kelvin Chan for providing information about his previous work.

References

Chan, S. K. *et al*. 2007. "Modeling the Statistical Behaviour of Lexical Chains to Capture Word Cohesiveness for Automatic Story Segmentation", *Proc. of INTERSPEECH-2007*.

Hearst, M. A. 1997. "TextTiling: Segmenting Text into Multiparagraph Subtopic Passages", *Computational Linguistics*, 23(1), pp. 33–64.

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

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

Reynar, J.C. 1999, "Statistical models for topic segmentation", *Proc. 37th annual meeting of the ACL*, pp. 357–364.

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

Yamron, J.P. *et al.* 1998, "A hidden Markov model approach to text segmentation and event tracking", *Proc. ICASSP 1998*, pp. 333–336.