# Initial Experiments on Automatic Story Segmentation in Chinese Spoken Documents Using Lexical Cohesion of Extracted Named Entities

Devon Li, Wai-Kit Lo, and Helen Meng

Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong
{ycli, wklo, hmmeng}@se.cuhk.edu.hk

**Abstract.** Story segmentation plays a critical role in spoken document processing. Spoken documents often come in a continuous audio stream without explicit boundaries related to stories or topics. It is important to be able to automatically segment these audio streams into coherent units. This work is an initial attempt to make use of informative lexical terms (or key terms) in recognition transcripts of Chinese spoken documents for story segmentation. This is because changes in the distribution of informative terms are generally associated with story changes and topic shifts. Our methods of information lexical term extraction include the extraction of POS-tagged nouns, as well as a named entity identifier that extracts Chinese person names, transliterated person names, location and organization names. We also adopted a lexical chaining approach that links up sentences that are lexically "coherent" with each other. This leads to the definition of a lexical chain score that is used for story boundary hypothesis. We conducted experiments on the recognition transcripts of the TDT2 Voice of America Mandarin speech corpus. We compared among several methods of story segmentation, including the use of pauses for story segmentation, the use of lexical chains of all lexical entries in the recognition transcripts, the use of lexical chains of nouns tagged by a part-of-speech tagger, as well as the use of lexical chains of extracted named entities. Lexical chains of informative terms, namely POS-tagged nouns and named entities were found to give comparable performance (F-measures of 0.71 and 0.73 respectively), which is superior to the use of all lexical entries (F-measure of 0.69).

**Keywords:** Story boundary detection, lexical cohesion, informative terms extraction, named entities.

## 1 Introduction

Story segmentation plays a critical role in spoken document processing. Spoken documents often come in a continuous audio stream (e.g. in news broadcasts) without explicit boundaries related to stories or units. It is important to be able to automatically segment these audio streams into coherent units. The segmentation process is non-trivial since the physical audio contents of a story boundary may be

Q. Huo et al. (Eds.): ISCSLP 2006, LNAI 4274, pp. 693 – 703, 2006.

very diverse – it may be a silent pause, a short duration of music, a commercial break, etc. A simple approach for detecting story boundaries may be based on cue word matching, but the cue words may be specific to the television/radio program and its period. Changes in the cue words will present a need to alter the heuristics in the system. Previous approaches have used a combination of prosodic, lexical, semantic and structural cues for story segmentation. They include audio energy levels and their changes [1], timing and melody of speech [2], novel nouns appearing in a short look-ahead window [3], word repetitions, synonyms and other associations [4]. In particular, Stokes et al. [4] proposed the use of lexical chaining that does not depend on specific cue word entries. Hence the approach is robust towards changes in the program and time. Previous work was done mostly in English text or speech recognition transcripts of English audio. Limited results were presented for Chinese. This paper reports on our initial attempt in the development of an automatic story segmentation system based on recognition transcripts of Chinese news audio. The approach includes extraction of informative lexical terms, including nouns and named entities; followed by the insertion of "lexical chains" that connects repeated informative terms. These chains are then used in scoring sentences for story boundary hypothesis. Figure 1 depicts an overview of the task of audio extraction from an audio/video news program, the process of recognition transcription, the process of story boundary detection and the use of detected boundaries for story segmentation.

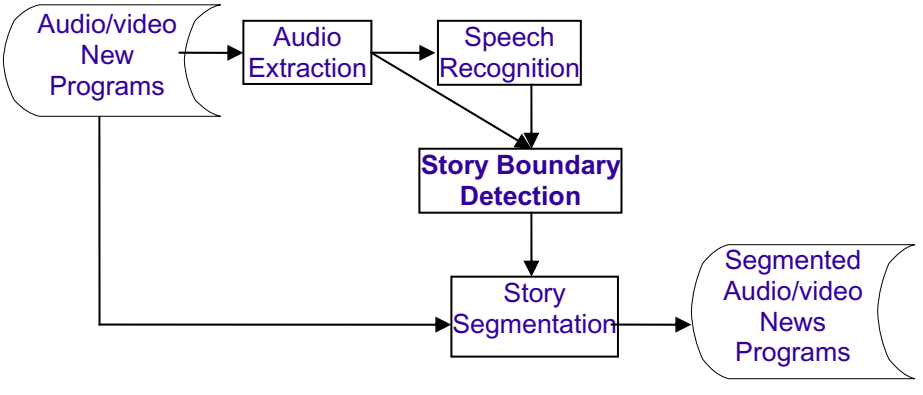


**Fig. 1.** Overview of the story segmentation task

## 2 Experimental Corpus and Evaluation

The experimental data is based on TDT2 Voice of American (VOA) Mandarin Speech corpus from February to June 1998 [5]. The VOA corpus contains radio news broadcasts in Mandarin and the corresponding recognition transcripts in GB-encoded Chinese characters. The recognition transcripts include pauses. Story boundaries are manually annotated in TDT2. Among 177 programs in TDT2, 54 programs (Feb to April) are used for training and parameter tunings and the rest of 123 programs (May

to June) are used for evaluation. From the training set, there are 1,549 stories in the corpus where 1,106 are annotated as news stories. The remaining are classified as *miscellaneous*. Miscellaneous stories contain filler content during story transition (e.g. silence, music, both silence and music, etc.), a news summary, an advertisement or introductory and conclusive comments from the newscaster. For the evaluation set, there are 1,456 stories where 1,159 are annotated as news stories. According to the TDT2 evaluation plan version 3.7 [6], a hypothesized story boundary is considered "correct" if it is placed with 15 seconds of the manually defined reference boundary.

## 3 Overview of the Approach

Our approach for automatic story segmentation includes three phases: (i) informative lexical term extraction; (ii) lexical chaining and (iii) story boundary hypothesis. Details on each phase are presented in the following subsections.

### 3.1 First Phase – Extraction of Informative Lexical Terms (Nouns and Named Entities)

Our work on informative lexical terms extraction can draw heavily from previous work in the MUC (Message Understanding Conference) and MET (Multilingual Entity Task Evaluations) that focused on named entities (NE) [7]. Informative lexical terms refer to terms that carry useful content related to its story. Previous approaches have emphasized the use of *nouns* (see section 1). Other examples of informative terms are *named entities* include person names, location names, organization names, time and numeric expressions. For our experiments, we use an existing part-of-speech (POS) tagger [8] to perform part-of-speech tagging in Chinese. We extract the tagged nouns informative terms. We also develop a named entity extraction approach to extract Chinese person names, transliterated foreign person names, location names and organization names. It is well-known that the Chinese language presents a special research challenge for automatic lexical analysis due to the absence of an explicit word delimiter. A Chinese word may contain one or more characters and the same character set is used for both Chinese names and transliterated foreign names. Automatic lexical analysis of speech recognition transcripts faces the additional challenges of recognition errors and word segmentation errors. The latter arises because the speech recognizer's output is based on its (constrained) vocabulary, which is different from the open vocabulary in news audio. In view of this, we propose a lexicon-based approach and a purely data-driven approach for word tokenization followed by a series of NE filters to extract informative terms.

#### 3.1.1 Word Tokenization

Figure 2 illustrates our word tokenization algorithm. Lexicon-based word tokenization involves a greedy algorithm that maximizes the length of the matching string as it references the CALLHOME lexicon. However, out-of-vocabulary (OOV) words absent from the lexicon will be tokenized as a series of singleton characters. Proper names are often OOV. For example:

贸易代表巴尔舍夫斯基
*(translation: trade representative Barshefsky)*

The character string is tokenized as:
［贸易］［代表］［巴］［尔］［舍］［夫］［斯］［基］

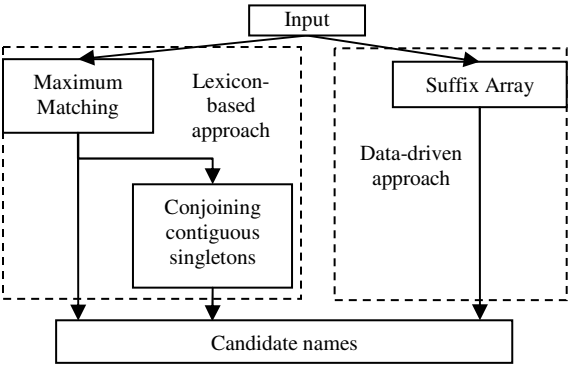


**Fig. 2.** Word tokenization using lexicon-based maximum matching as well as suffix array for generating candidate informative terms

Therefore, in order to salvage the OOV words that may potentially be a name, we conjoin the contiguous singletons to form a candidate name for names filtering in the subsequent module, i.e.:

［贸易］［代表］〈巴尔舍夫斯基〉
*(translation: [trade] [representative] <Barshefsky>)*

In other words, the lexicon-based approach leverages available lexical knowledge to extract OOV words that may be candidate names. However, since a given Chinese character sequence may have multiple possible word segmentations, the greedy algorithm may be biased by the lexicon and may miss out on other possible tokenization options. Hence, we supplement with a purely data-driven approach for word tokenization.

We used the suffix array structure [9] to extract the longest recurring string patterns in a radio program. The algorithm generates all substrings at all lengths within all sentences / utterances.[1] These substrings are then sorted. Only substrings that occur more than once and with lengths greater than one character are preserved, stop characters on either ends are removed and the resulting strings are treated as candidate names for subsequent names extraction. Analysis shows that the suffix array uncovers substrings such as "克林顿在" (*translation: "Clinton in")* and "８万马克" (*translation*: *80,000 Deutsch Mark*), which may contain useful transliterated names.

[1] For the k$^{th}$ sentence with $N_k$ characters ($C_1$, $C_2$,...,$C_{Nk}$), possible output is {$C_{ki}$...$C_{kj}$; $\forall i$=*1 to* $N_k$, $\forall j$=*i to* $N_k$}.

### 3.1.2 Names Extraction

Four types of names will be extracted from the list of tokenized words: Chinese person names, transliterated person names, location names and organization names. For Chinese person name extraction, we apply two simple heuristics in this step: (a) the most common 100 surnames with reference to the surname list from [10], augmented with other surnames we found from the Web (i.e. 219 Chinese surnames in all); and (b) the popular Chinese names structure that consists of the surname (in one or two characters), followed by the given name (in one or two characters). Valid name structures include: SG, SGG, SSG and SSGG (where 'S' denotes a surname character, e.g. 陈; and 'G' denotes a given name character, e.g.红). Hence, in the Chinese name filtering procedure, a name candidate must be of two to four characters in length and must follow the pre-defined name structures in order to be qualified as a Chinese name.

Extraction continues with a transliterated name character bigram model that is trained on the MEI transliterated name list of 42,299 items [11]. We used Good-Turing discounting and backoff smoothing. By thresholding the normalized log probability at above -3, a 99% recall can be obtained from the training data. Log probability score are calculated for each word tokens, those scores above the threshold (i.e. -3) are extracted as transliterated person name.

A list of commonly used location and organization suffix characters are used to further extract word token which contains special suffix characters, e.g. "厅 署 中学 公司 城 池 村 道" (*translations: department, office, school, company, city, lake, village, road*). We also used a list of well-known location and organization names as basis for the extraction of known organization and location names.

## 3.2 Second Phase – Lexical Chaining

We extended the lexical chaining approach in [4] to Chinese with a focus on the repetition of informative lexical terms as an indicator of lexical cohesion. Lexical cohesion is represented as lexical chains that connect repeated occurrences of informative lexical terms. As mentioned previously, informative lexical terms include nouns or named entities. Other terms are deemed non-informative. If we observe a point in the story transcriptions where many existing lexical chains end and new lexical chains begin, then we consider it to be an indicator of a possible topic shift that is related to the occurrence of a story boundary.

More specifically, we group all contiguous words in the recognition transcripts of TDT2 into a "sentence". Hence the "sentences" are separated by pauses. Each sentence is labeled with an index number. Every informative lexical term in a sentence is tracked with regards to its occurrences in other sentences. This tracking process is conducted sequentially across the sentences that lie within a fixed window length, i.e. the window length for chain formation. Choice of values for this window length should reference the story length. In our training set, 94% of the stories range between 20 and 400 seconds in duration. At every 20 second we take a value and a total of twenty values are obtained. From these twenty values, the one that optimizes the training performance is selected. A lexical chain is inserted if the current sentence contains an informative lexical term that has occurred in the previous sentence. We define a "start chain" (i.e. starting lexical chain) to occur where a lexical term is

chained to the following sentences but not the preceding sentences. Conversely, we define an "end chain (i.e. ending lexical chain) to occur where a lexical term is chained to the preceding sentences but not the following sentences. This is illustrated in Figure 3.

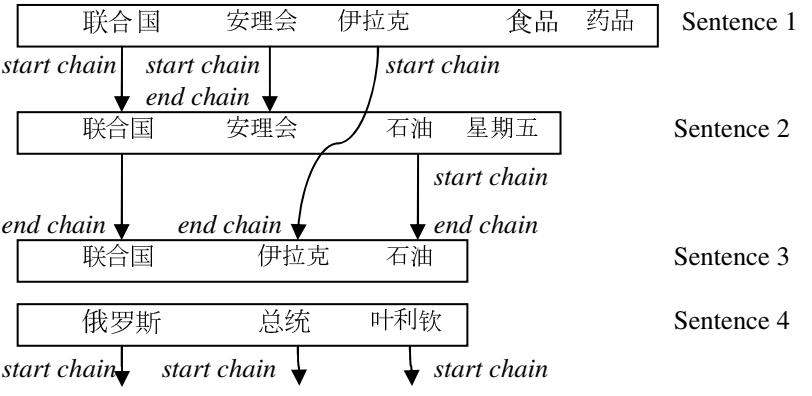


**Fig. 3.** Illustration of lexical chaining, Sentence 1 contains three start chains (联合国, 安理会, 伊拉克); Sentence 2 contains one end chain (安理会) as well as one start chain (石油); Sentence 3 contains three end chains (联合国, 伊拉克, 石油) and Sentence 4 contains three start chains (俄罗斯, 总统, 叶利钦)

### 3.3 Third Phase – Story Boundary Hypothesis

After all lexical chains are established, we assign a "chain score" to every sentence, defined as:

$$chainScore_{sent\_i} = num\_start\_chain_{sent\_i} + num\_end\_chain_{sent\_i-1} \quad \text{(Equation 1)}$$

Based on Equation (1), the chain scores for sentences 1 to 4 in Figure 3 should be 3, 1, 1 and 6 respectively. A sentence with a high chain score generally has a high number of start chains and its preceding sentence contains a high number of end chains. This should be an indication of a likely occurrence of a story boundary.

In this work, we take an *"over-hypothesize and filter"* approach to story boundary hypothesis. We will first hypothesize the occurrence of a story boundary if the chain score exceeds a tuned threshold as obtained from parameter tuning. We observed from the training data that story boundary existed at sentence with chain scores between one to nine. A tuned threshold can be obtained within this range. However, given that the evaluation criterion can tolerate offsets of 15 seconds for a story boundary (see Section 2), we also follow up with a *filtering mechanism* that selects the highest-scoring proposed boundary within a fixed window length. For example, in Figure 4 we see two sentences that lie 5.68 seconds apart but with two hypothesized boundaries. The filtering mechanism will remove the boundary at time 272.1 (with lower chain score=4) and keep the boundary at time 266.42 (with higher chain score =5). This parameter can be obtained by tuning in the development set

Before filtering hypothesized boundaries

<sen id="25" time="266.42" score="5" boundary="**yes**">
"苏联""格鲁吉亚""共和国""当局"</sen>
<sen id="26" time="272.1" score="4" boundary="**yes**">
"联合国""军队""观察员"</sen>

After filtering hypothesized boundaries

<sen id="25" time="266.42" score="5" boundary="**yes**">
"苏联""格鲁吉亚""共和国""当局"</sen>
<sen id="26" time="272.1" score="4" boundary="**no**">
"联合国""军队""观察员"</sen>

**Fig. 4.** Illustration of the filtering mechanism for hypothesized story boundaries. We take an "over-hypothesize and filter" approach to story boundary detection. Sentences with a chain score exceeding a trained threshold will be hypothesized with a story boundary. Given that the evaluation criterion can tolerate boundary offsets up to 15 seconds, our filtering mechanism uses a fixed window length within which only the highest-scoring boundary is preserved.

where given that there is a 15 seconds offsets during evaluation, a value smaller than 30 seconds will be a reasonable candidate for this parameter.

## 4 Experimental Results

We have a series of comparative experiments on automatic story segmentation. The various experimental setups are:

1. **Baseline performance using pauses:** The first baseline segments stories based on the occurrence of pauses. A story boundary is hypothesized whenever a pause occurs in the recognition transcripts. This is a very aggressive baseline segmenter, since pauses may also result from breath breaks, turn-taking in a dialog, etc. which do not correspond to a story boundary.
2. **Baseline performance using all lexical terms:** The second baseline includes all lexical terms found in the recognition transcripts for lexical chaining. Hence the vocabulary used for lexical chaining is identical to that of the speech recognizer.
3. **Performance of lexical chaining with POS-tagged nouns:** An existing POS tagger [8] which is trained on another text corpus is used for tagging nouns (including locations). These are categorized as "informative lexical terms" and only such terms are used for lexical chaining and subsequent story boundary hypothesis.
4. **Performance of lexical chaining of extracted named entities:** In this setup, informative lexical terms are defined as extracted named entities, including Chinese personal names, transliterated personal names, location names and organization names. The method of extraction is described in Section 3.1. The extracted named entities are used in the lexical chaining experiments.

Table 1 shows the tuned values for each parameter from the training corpus. The window length for chain formation is consistent across all units and 80 is actually closed to the average story length (100 seconds) in the training corpus. The window length for boundary removal also consistent across all units at 25 seconds while the chain score differ with each other where the chain score value is in proportional to the number of terms available during chain formation. From the training corpus, the best performance is obtained by using named entities for chaining where it also achieved the best precision among other chaining units.

**Table 1.** shows tuned parameters for each lexical chaining unit in the training data set as well as their corresponding performance on story boundary detection

| | **All lexical terms** | **POS-tagged nouns** | **Named entities** |
|---|---|---|---|
| Window length for chain formation (seconds) | 80 | 80 | 80 |
| Chain score threshold | 4 | 3 | 2 |
| Window length for boundary removal (seconds) | 25 | 25 | 25 |
| Number of terms | 191,371 | 115,110 | 43,417 |
| Precision(P) | 0.55 | 0.58 | 0.63 |
| Recall (R) | 0.64 | 0.69 | 0.66 |
| F-measure (F) | 0.59 | 0.63 | 0.64 |

We applied the trained thresholds to the evaluation corpus and results are shown in Table 2 and Figure 5. Total number of terms for each chain unit in the evaluation corpus is 67,380, 43,051 and 18,872 respectively.

**Table 2.** Performance on story boundary detection based on (i) use of pauses in recognition transcripts; (ii) lexical chaining of all vocabulary items in recognition transcripts; (iii) lexical chaining of POS-tagged nouns; (iv) lexical chaining of extracted named entities

| | Pauses only | **All lexical terms** | **POS-tagged nouns** | **Named entities** |
|---|---|---|---|---|
| Precision(P) | 0.04 | 0.86 | 0.87 | 0.88 |
| Recall (R) | 1 | 0.57 | 0.63 | 0.59 |
| F-meas. (F) | 0.08 | 0.69 | 0.73 | 0.71 |

It can be observed from Figure 5 that story segmentation based on pauses produces very high recall (R=1) but very low precision (P=0.04), leading to an F-measure of 0.08. This is because there are over 14,700 pause segments in the corpus but only 1,159 correspond to story boundaries. The filtering mechanism removes a fraction of false alarms in story boundary hypotheses occurring within a 25-second window.

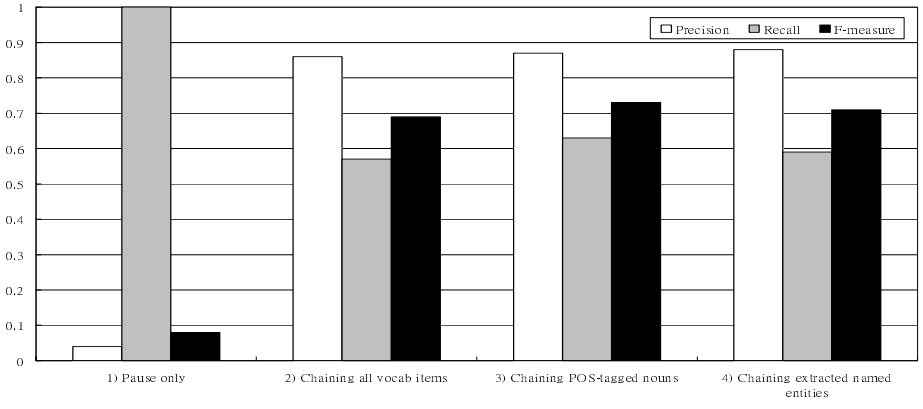


**Fig. 5.** Performance on story boundary detection based on (i) use of pauses in recognition transcripts; (ii) lexical chaining of all vocabulary items in recognition transcripts; (iii) lexical chaining of POS-tagged nouns; (iv) lexical chaining of extracted named entities

As we migrated to the use of lexical chaining of all vocabulary items in the recognition transcripts, performance values are P=0.86, R=0.57 and the overall F-measure improved to 0.69. The lexical chains offer lexical constraints for story segmentation. Lost recall is generally due to having too few lexical chains, causing the chain score to fall below the threshold, thereby missing the hypothesis of a story boundary. For example, one of the boundary sentences was "在 危地马拉" *(translation: in Guatemala),* where there is only one lexical chain, leading to a missed story boundary.

The use of POS-tagged nouns attains the performance of P=0.87, R=0.63 and further improved the F-measure to 0.73. We believe that these selected informative terms offers more focus in ascertaining lexical coherence. For example, the sentence "一个 发电厂 一个 核反应堆" (*translation: one power plant, one nuclear reactor*) contains three terms ""一个""发电厂""核反应堆"". Two of these are tagged as nouns, i.e. ""发电厂", "核反应堆"". The term "一个" is rather general and is generally not significant in the determination of lexical cohesion. It may even be possible for such terms to give rise to insignificant lexical chains that generate inaccurate story boundaries. This is an illustration of the possible benefits of using POS-tagged nouns for lexical chaining.

The use of extracted named entities gave the performance of P=0.88, R=0.59 and the F-measure of 0.71, which suggests that these are comparable with POS-tagged nouns for ascertaining lexical cohesion for story segmentation, with slightly better precision and slightly lower F-measure. In our analysis, we found that named entities are often more descriptive of lexical cohesion than general nouns (hence achieving better precision). For example, the sentence with the named entities,"美国 中东 特使 罗斯 星期一 以色列" (*translation: US, Middle East, special envoy, Roth, Monday, Israel*) contains five nouns "美国 中东 罗斯 星期一 以色列". The term "星期一" (Monday) was lexically chained with a preceding sentence, which suggests lexical cohesion of this sentence with the preceding sentences. The remaining four

terms were chained with following sentences which suggests lexical cohesion with following sentences and thereby outweighing the effect of the term "星期一". Since named entities generally contain more distinctive information for describing lexical cohesion, they provide a better precision value for story segmentation. On the other hand, when compared with POS-tagged nouns, named entities achieve a lower recall for in both the training and evaluation corpora. This may be related to the use of 18.872 unique named entities in the data set, as compared with 43,051 POS-tagged nouns.

## 7 Conclusions and Future Work

This paper presents our initial experiments in automatic story segmentation of recognition transcripts of Chinese spoken documents. This is an important problem since spoken documents often come in a continuous audio stream without explicit boundaries that indicate the transition from one story (or topic) to another. Our approach consists of three phases:

- Automatic term extraction that includes lexicon-based maximum matching for word tokenization, followed by POS tagging and nouns extraction. We also develop a named entity extraction approach, involving lexicon-based maximum matching to uncover out-of-vocabulary words as singleton characters, together with purely data-driven suffix array approach that identify recurring strings. These extracted terms are then passed through a series of filters for Chinese names, transliterated names, location and organization names.
- A lexical chaining algorithm that connects repeated informative lexical terms as an indication of lexical cohesion among sentences. Story boundaries tend to occur where many existing lexical chains end and new lexical chains begin.
- A story boundary hypothesis component that adopts an "over-hypothesize and filter" paradigm – the lexical chain score (based on the total number of ending and starting lexical chains) of each sentence is compared with a trained threshold, above which a story boundary will be proposed. This is followed by a filtering mechanism that checks whether multiple boundaries are hypothesized within a small time window (25 seconds), upon which only the highest-scoring boundary hypothesis is preserved.

We conducted story segmentation experiments based on TDT2 Voice of America Mandarin news data. We observe increasing F-measures in story segmentation performance as we migrate from using only pauses for story segmentation; using all vocabulary items in the recognition transcripts with lexical chaining; using informative terms with lexical chaining. These results suggest that named entities serve well as informative lexical terms that can effectively describe lexical cohesion for automatic story segmentation. Future work will incorporate the use of both POS-tagged nouns and named entities, synonyms and other word associates in HowNet [12] for lexical chaining; as well as the incorporation of other prosodic features, e.g. fundamental frequencies for story segmentation.

## 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.

## References

1. Greiff, W., Hurwitz, L., Merlino, A. MITRE TDT-3 Segmentation System, TDT Evaluation System Summary, 1999
2. Shriberg, E., Stolcke, A., Hakkani-Tur, D. & Tur, G., Prosody-Based Automatic Segmentation of Speech into Sentences and Topics, *Speech Communication* 32(1-2), 127-154, September 2000
3. Franz, M., McCarley, J.S., Ward T., Zhu, W.J., Segmentation and Detection at IBM: Hybrid Statistical Models and Two-tiered Clustering. TDT Evaluation System Summary, 1999
4. Stokes, N., Carthy, J., Smeaton, A.. SeLeCT: A Lexical Cohesion based News Story Segmentation System. In the Journal of AI Communications, Vol. 17, No. 1, pp. 3-12, March 2004.
5. TDT2 Main page, http://projects.ldc.upenn.edu/TDT2/
6. TDT2 Evaluation Plan 1998, v 3.7. http://www.nist.gov/speech/tests/tdt/tdt98/doc/tdt2.eval.plan.98.v3.7.ps
7. Palmer, D. and Ostendorf, M. Improved word confidence estimation using long range features, In EUROSPEECH-2001, 2117-2120.
8. Meng, H. and Ip, C. W., An Analytical Study of Transformational Tagging on Chinese Text, Proceedings of the 1999 ROCLING conference, August 1999.
9. Manber, U. and Myers, E.W. Suffix arrays: a new method for on-line string searches. SIAM Journal of Computing, 22(5), pp. 953-948., 1993
10. Yuan, 新百家姓出炉：李王张继续位列姓氏前三甲, http://news.sina.com.cn/c/2006-01-10/09097941017s.shtml, January 2006.
11. Meng, et al., Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval, http://www.clsp.jhu.edu/ws2000/groups/mei/, 2000
12. HowNet, http://www.keenage.com