Unsupervised Identification of Story Boundaries in Malay Spoken Broadcast News

Zainab A. Khalaf Aleqili[1,2]
[1]School of Computer Sciences, Universiti Sains Malaysia (USM)
11800 Pinang, Malaysia
[2]Department of Computer Science, College of Science, University of Basrah
BASRAH, IRAQ
Email: zainab_ali2004@yahoo.com

Tien Ping Tan
School of Computer Sciences, Universiti Sains Malaysia (USM)
11800 Pinang, Malaysia
Email: tienping@cs.usm.my

***Abstract* — This paper describes a spoken document retrieval system for processing Malay spoken broadcast news that uses an approach to enhance retrieval performance. An automatic speech recognition (ASR) system was adapted to reduce the impact of ASR transcription errors on retrieval performance. The performance of unsupervised learning was evaluated using Malay broadcast news as the data source. A latent semantic analysis was used to reduce the impact of synonymous words and to identify the story boundaries within the news segments. Among other things, the current system proved to be a powerful instrument to identify news story boundaries automatically.**

***Index Terms* — spoken document retrieval; broadcast news transcription; clustering; latent semantic analysis; Singular Value Decomposition (SVD)**

I. INTRODUCTION

A spoken document retrieval (SDR) system uses both automatic speech recognition (ASR) and information retrieval (IR) technologies to analyze and process multimedia documents. Researchers usually use ASR systems to translate spoken documents (speech) into text transcripts. In SDR research, well-known text-based search algorithms are applied to time-aligned ASR transcripts so that the spoken content can be automatically indexed and retrieved from various multimedia documents, including radio/television broadcasts, digital library archives, call-center recordings, meetings, academic lectures, and internet user audio/video. For this reason, SDR systems are designed to integrate both ASR and IR technologies. An effective SDR system usually provides access to both spoken textual content and to rich information that reflect the intended meaning and the speaker's emotional state. Efficient SDR systems are needed because of the constantly increasing volume of multimedia content and the demand to access information in these multimedia collections. Although audio searching has become very popular, a number of obstacles make the process of retrieving spoken documents challenging. These problems include the lack of overt punctuation and formatting and the difficulty in detecting story boundaries or segments. Moreover, identifying word errors generated by ASR is one of the major challenges facing SDR. To address these problems, we designed a system to reduce the impact of ASR transcription errors on segmentation performance and tested it using Malay spoken broadcast news [1-4].

II. DATA SOURCE

A transcript produced manually from spoken broadcast news [5] was used in this study to identify the story boundaries. This process was applied to Malay broadcast news documents already collected at University Sains Malaysia as the output of the Malay ASR system. Thus, the main data source was Malay broadcast news stories that were recorded from different Malay television broadcasts [5]. The database included ~25 hours of transcribed speech. The ASR system was trained using a ~15 hour portion of the database, and the SDR test sets included ~10 hours of Malay broadcast news. None of the test sets overlapped with the ASR training set. Table 1 shows the data source details.

TABLE 1:

DATA SOURCE DETAILS

Number of news shows	18
Number of news stories in all news shows	379
Number of sentences in news database	4698
Number of words in the news	81116
Word error rate before adaption	34.5
Word error rate after adaption	33.9
Story length	Around 1 to 167 Sentence
The rate of the audio signal extract	10 s

doi:10.4304/jetwi.5.1.28-34

III. PROPOSED SYSTEM

In this study, the spoken document system was designed and used to identify news story boundaries and the number of stories within dataset. The broadcast news transcription was split into sentences, and then the cluster model was used to cluster the sentences into stories using the latent semantic analysis (LSA) algorithm and translation program. In short, a given spoken document was converted to text using the ASR system. The ASR result was improved using the Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR) algorithms, which are part of the ASR system. Next, the story boundaries were identified using LSA, and these were considered to be the baseline results. To improve the clustering results, we propose that the text be translated into several languages and then back to Malay in order to reduce the impact of synonymous words on clustering.

The proposed system process proceeds in five stages. Figure 1 shows the flow diagram of the process, and the five stages are described in more detail below:

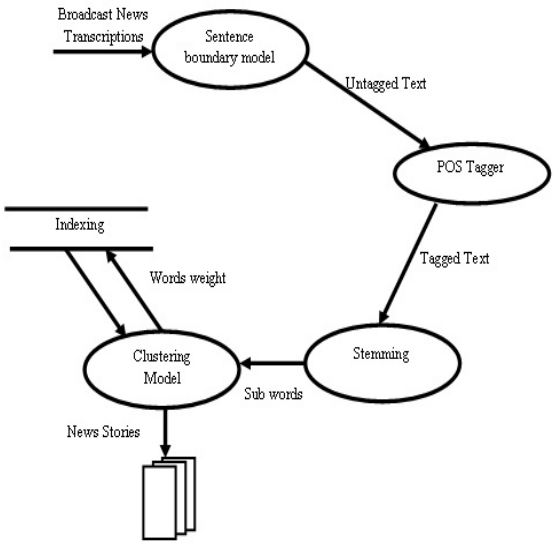

Figure 1: Data flow diagram

Stage 1: Using ASR, Malay spoken broadcast news stories are converted to text files.

Stage 2: Words are marked in a text to correspond to a particular part of speech (POS) via tagging. This process is based on the definition and context relationship of a given word with related and adjacent words in a sentence [6]. Figure 2 shows an example of POS output.

```
ert/N health/N . faks/N menang/V tiga/NUM_CARD gangsa/N art/N
structure/N the/DET art/N . gas/N pengangkutan/N tan/N sri/N
chan/N kong/N choy/N tidak/NEG-PART bersalah/A diperketat/N
sessions/N tercetus/V tiga/NUM_CARD . unit/N yang/CS haru/N .
partners/N akhbar/N the/DET star/N dan/CC negeri/N terjamin/A .
substance/N penjaring/N . gas/N akan/AU_V terus/V meresap/V
dengan/PREP cepat/A . sons/N campur/V charles/N didakwa/V_EN
menyimpulkan/V abdullah/V ahmad/N badawi/N tidak/NEG-PART
berdaya/AU_INF tanya/V . untuk/AU_INF memberikan/V
persetujuan/N supaya/AU_INF kesan/N estetika/N kepada/PREP
kualiti/N ginseng/N yang/CS berhasrat/N templer/N park/N
untuk/AU_INF memajukan/V projek/N pembangunan/N dan/CC
shipman/N hak/N di/PREP pulau/N tiga/NUM_CARD resort/N .
empat/NUM_CARD itu/DET kontrak/N telah/AU_V menyembunyikan/V
fakta/N bahawa/CS pembiayaan/N kos/N pembangunan/N itu/DET .
akan/AU_V dibiayai/V_EN melalui/PREP terbitan/N bon/N oleh/PREP
kuala/N dimensi/N sdn/N bhd/N . dan/CC syarikat/N asing/A
mengenai/PREP tahap/N bhd/N . dengan/PREP sokongan/N pelajar/N
malaysia/N . hadir/V menteri/N pengangkutan/N malaysia/N .
dengan/PREP saiz/N tempatan/A itu/DET benar/A abdullah/V
ahmad/N badawi/N bersetuju/V meluluskan/V projek/N
berkenaan/DET . perbuatan/N itu/DET dilakukan/V_EN pada/PREP
tempat/N dan/CC doktor/N lima/NUM_CARD februari/N dua/NUM_CARD
ribu/NUM empat/NUM_CARD . sebelas/N oktober/N pondok/N
lapan/NUM_CARD oktober/N diterima/V_EN . dan/CC dua/NUM_CARD
```

Figure 2: POS output

Stage 3: A number of words, usually called stopping words, are frequently repeated in speech and have no real meaning themselves but may be used as auxiliary verbs or prepositions (e.g., is , are, at, in, the, a, an) [6]. Such words are deleted during this indexing stage to avoid problems created by the presence of such generic terms and to reduce the size of the index. Examples include such prepositions as “ke, to,” “dari, from,” and “pada, at,” which provide no real information significant to the document’s topic but appear frequently in almost every document collection. When such words are removed, the index size decreases and the quality of the search results improves. The idea is that only the words that contribute specific information to each document are retained.

Stage 4: “Stemming” describes a process used to improve the effectiveness of IR. During this process, variant forms of the same word with different endings are reduced to a common stem. Stems are useful in IR conducted using techniques that unify vocabularies. The use of stems reduces term variants and storage space and, at the same time, increases the matching probability of the documents [7].

Stage 5: Clustering the spoken broadcast news transcription using the LSA algorithm and translation approach is performed in this stage. It is described in greater detail in the next section.

IV. CLUSTERING MODEL

Clustering of sentences can be used to find repeated information, and this is done by grouping similar sentences together. Previous research has identified a number of different methods that can be used to identify similar sentences. Some methods use shallow techniques to detect similarities in sentences (e.g., word or n-gram overlap). Other methods use a deeper syntactic or semantic analysis. The resulting clusters represent subtopics of the document set, whereby one cluster ideally describes one subtopic. The important point here is that clustering using LSA is largely independent of the language, unlike other approaches that rely heavily on a deeper linguistics analysis. The advantage of this technique is that the similarity estimation is not based on

shallow methods such as word matching. LSA takes underlying latent semantic structures in the form of word usage patterns into account. Accordingly, the problem of synonymy is avoided [8, 9].

Documents are usually read and split into sentences during the preparation process, and these sentence sets then are indexed and term-by-sentence matrices (TSMs) are built. One of the payoffs of using LSA is that it reduces dimensionality, which in turn results in a quicker search. When the matrix is ready, it is subjected to Singular Value Decomposition (SVD) [10, 11]. SVD is a method from the field of linear algebra that is used to diagonalize any t_d matrix A. The diagonalization process corresponds to a transition to a new coordinate system through which the latent semantic structure of a document set is brought forth. The SVD formula is as follows:

$$A = AT * S * AD$$

Any rectangular matrix A is squared by multiplying it by AT. The eigenvectors of $AT = A\ A^T$ are the left singular vectors of A, and the eigenvectors of $AD = A^T\ A$ are the right singular vectors of A. Singular values are the square roots of the common eigenvalues of AT and AD and are written in descending order in S. The eigenvectors in T and D are ordered correspondingly.

The TSM A is created, in which each row stands for a term and each column stands for a sentence. The cells of such a matrix contain weighted term frequencies. SVD is applied to A to break down the original TSM into r base vectors, which are linearly independent [10, 11]. The result of SVD is three matrices (AT, S, and AD) (Figure 3), which basically are used to calculate a ranking matrix by using cosine distance, which is a very popular similarity measure, to compute the distance between two sentences. Later, similar sentences are placed together to create a sentence cluster. Next, a new matrix is created from this cluster and the rest of the sentences. After applying SVD, all sentences are compared pair-wise. This process is repeated until the distance of similarity between the document sentences is larger than threshold.

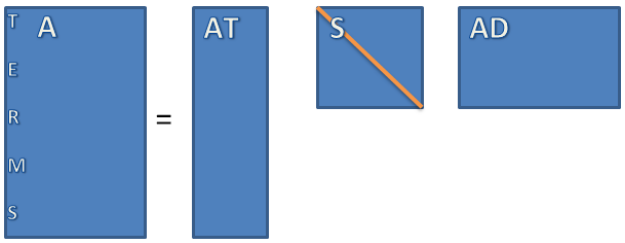

Figure 3: SVD description

V Evaluation Metrics For Segmentation

Because of the need for a strong evaluation strategy to estimate the influence of different parameters on clustering quality, two metrics were used: the sentence error rate and the the F-measure.

A. Sentence Error Rate

In order to compute semantic distance between automatic segmentation and manual segmentation, the proposed framework implements a statistical measure called the sentence error rate (SER). The SER is similar to the word error rate (WER) approach, but here it is applied to sentences instead of words. Thus, the proposed metric for evaluating the performance of the algorithms for determining spoken news story boundaries is the SER comparison strategy, which is outlined as follows:

$$\text{SER} = 100\% \times \frac{\text{Dels} + \text{Ins}}{\text{No. of sentences in the correct story}}$$

where Dels is the number of sentence deletions (when an accurate sentence is omitted within the recognized story) and Ins refers to the number of sentence insertions (when an additional sentence is added into the recognized story). Using an automatic programming algorithm, the sentence alignment process between the story in the reference transcription (manual transcription) and the hypothesis transcription (automatic story identification) is performed. However, some types of errors (Ins, Dels) might be more critical than others and thus have a greater impact on the final metric of the spoken clustering task.

B. F-Measure

In information retrieval, relevance scores of the documents during the ranking step are sorted from high to low and provide a list of the sorted document to the user as a result to the user query. To evaluate the IR system, two statistics terms are used to describe the results of a query. First, precision explains the fraction of the returned results (clusters) that is relevant to the data needed (classes).

$$precision = \frac{|Clusters \cap Classes|}{|Clusters|}$$

The second term is recall, which denotes the fraction of the relevant documents in the corpus that was actually returned by the system to the user.

$$recall = \frac{|Clusters \cap Classes|}{|Classes|}$$

The F-measure considers both the precision and the recall of the IR test to compute the score.

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

VI. Experimental Results

In this study, we designed a SDR system and tested its ability to detect boundaries in Malay spoken broadcast news stories and to determine the number of stories automatically using unsupervised clustering. Once the recognition decoder output was generated, the MAP and

MLLR algorithms were used to improve the ASR transcription. Story boundaries were identified using the clustering algorithm, which also was used to prepare the output for other applications, such as classification, summarization, and title classification.

After constructing a document-term matrix, filtering stop words, and stemming, SVD was completed and the cosine similarity was computed for each possible sentence couple. The proposed system was tested on a sample of 4698 sentences and 397 stories from 18 Malay spoken broadcast news show segments containing different kinds of news, including local news, political news, sport news and so on. The range of story length was 1 to around 160 sentences. Figure 4 shows a diagram of the analysis system, and Table 2 shows the results when the LSA method was used to process the 18 news shows.

From table 2, the first column: refers to the number of news shows, the second column: refers to the number of sentences inside each show, the third column: refers to the number of total stories in each show. The two former values (column two and three) are obtained from the reference document. The forth column acts the accuracy to determine these stories boundary in each show. The next column represents number of stories that get it from the proposed system automatically. The sixth column acts the F-measure for each show, and the next column acts the number of the correct stories that the proposed system was recognized it correctly. The last column acts the sentence error rate values for each show.

As result, in our approach we identified 152 out of 379 stories correctly with F-Measure 0.46 and accuracy 0.93.

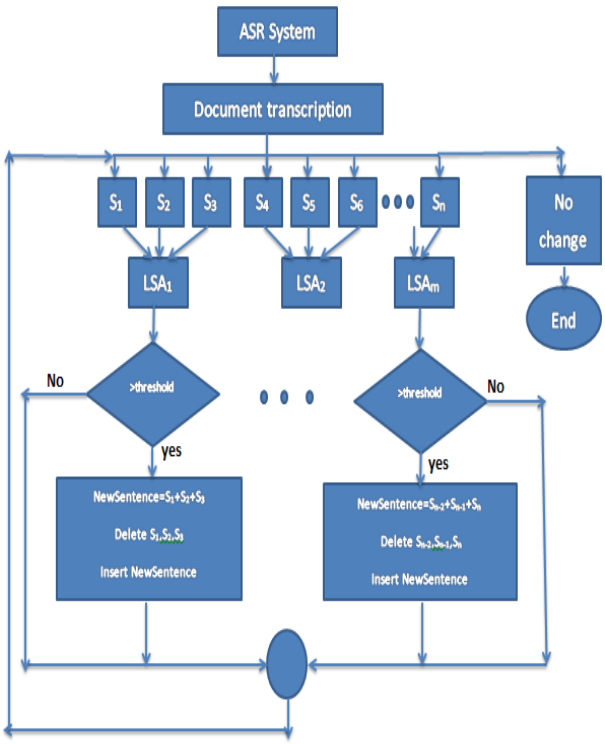

Figure 4: The proposed system diagram

TABLE 2:
THE RESULTS WHEN THE LSA METHOD WAS USED TO PROCESS THE 18 NEWS SHOWS.

No. of Shows	No. of Sentences	No. of stories in Reference	Accuracy	No. of stories in Hypothesis	F-Measure	No. of correct stories	SER
1	161	17	0.93905	32	0.49391	10	56.08465608
2	210	17	0.94005	24	0.49503	6	56.73469388
3	303	25	0.95532	34	0.50291	10	52.85285285
4	168	23	0.95961	15	0.61382	12	51.88284519
5	258	25	0.96097	18	0.62051	14	47.87535411
6	216	17	0.93231	21	0.46637	6	58.60655738
7	358	27	0.95008	36	0.44375	8	58.58823529
8	329	11	0.81828	17	0.3154	3	62.56684492
9	299	22	0.9406	34	0.42232	12	55.95238095
10	300	32	0.95287	38	0.43152	16	54.96183206
11	286	26	0.95848	48	0.45013	13	58.55072464
12	280	22	0.94602	20	0.5412	6	57.35660848
13	277	28	0.94325	29	0.44036	8	65.14745308
14	313	20	0.93345	35	0.39685	7	61.11111111
15	159	15	0.94359	11	0.58108	7	46.80851064
16	326	19	0.85767	25	0.32067	6	60.10781671
17	284	23	0.9334	23	0.48701	4	64.95535714
18	171	10	0.85828	18	0.33781	4	62.76595745
Total	4698	379		478		152	
Average			0.929071		0.464481		57.38387733

Figures 5, 6, and 7 show the results when the LSA method was used to process the 18 news shows following the principles of SER, accuracy, and F-measure to generate the Malay broadcast news story boundaries.

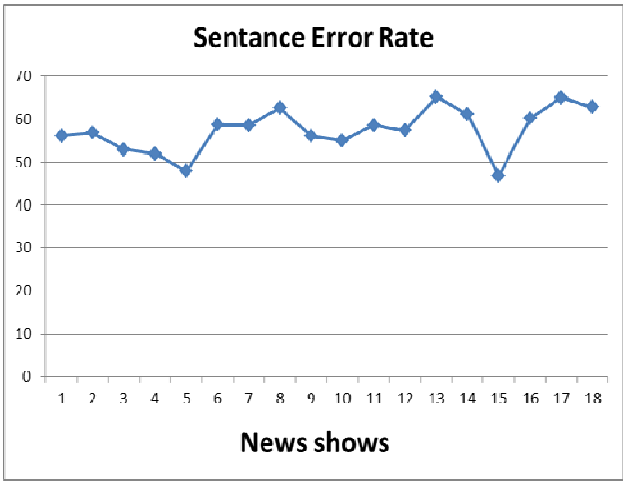

Figure 5: The sentence error rate (SER) for the 18 spoken broadcast news shows that were analyzed

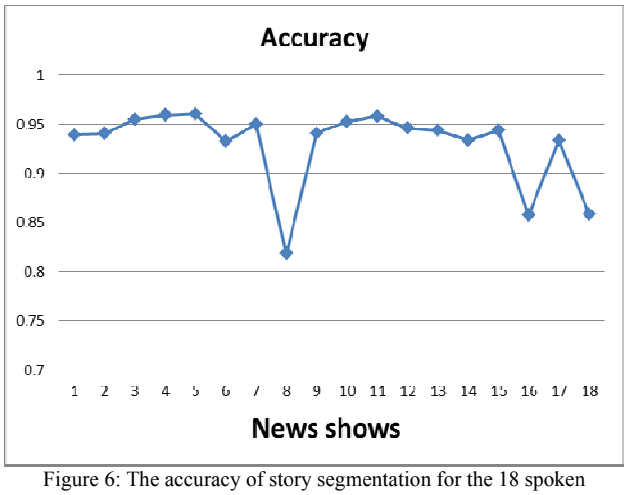

Figure 6: The accuracy of story segmentation for the 18 spoken broadcast news shows that were analyzed

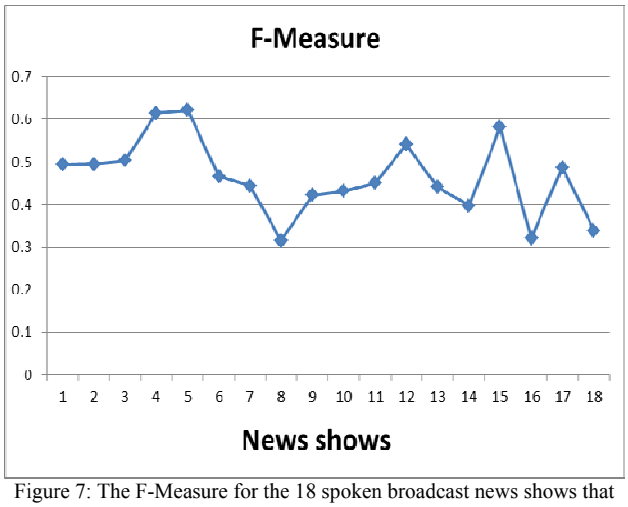

Figure 7: The F-Measure for the 18 spoken broadcast news shows that were analyzed

The performance of LSA in story clustering in this experiment was automatically compared with manual story boundary identification. The experimental results showed that the use of LSA to segment the Malay broadcast news into stories yielded better accuracy than that of our previous system [12].

VII. Related Works

Processing spoken documents is challenging because of the word errors generated by the ASR process [1, 13]. Determining the boundaries of broadcast news stories is another obstacle to processing spoken documents. The lack of overt punctuation and formatting contributes to this problem. For the purpose of conducting any operation like search , document retrieval, classification , summarization and topic detection, IR must first determine the beginning and the end of the segments or paragraphs [4, 8, 13-16]. Overall, the process of determining the number of segments within the spoken document automatically is the most difficult problem in data clustering [9].

Broadcast news contains a tremendous amount of information in the form of audio, video, and text. This information, like other non-organized groups of data, is not very useful unless relevant data can be easily retrieved from it. Current technology allows for extraction, indexing, and searching for information and data within broadcasts. The problem in spoken document is that the individual news stories that include in broadcast news are not indexed and searched separately. This causes inaccurate search results, so the user must scan the entire broadcast to find the relevant information. For example, an user searching for news associated with "war in Iraq" would prefer to receive short clips concerning the war rather than every broadcast that contains a story concerning the war [13, 17, 18].

The absence of punctuation and capitalization in spoken documents makes automatic story segmentation in multimedia documents very challenging. Previous attempts at multimedia system segmentation have concentrated on three types of cues: visual cues, such as the presence of an anchor person's face [19] and motion changes [20]; audio cues, such as significant pauses and resets of the pitch [21, 22]; and lexical cues, such as word similarity measures of the speech recognition transcripts or closed captions of video [23, 24]. Cues from completely different modalities (audio, video, and text) are often consolidated to achieve a higher segmentation performance [19, 25].

For story segmentation, Hearst et al. (1997) proposed the TextTiling approach, which is based on the straightforward observation that different topics usually employ different sets of words and that shifts in vocabulary usage are indicative of topic changes [23]. As a result, pairwise sentence similarities are measured across the text and a local similarity minimum implies a story boundary. Stokes et al. (2004) embodied word cohesion in a lexical chaining approach in which related words in a text were linked into chains and a high concentration of chain starting and ending points is an indication of a story boundary [24]. These two approaches have been recently used to segment multimedia documents such as broadcast news [22, 25] and meetings [26]. Rosenberg et al. (2006) presented results from a broadcast news story segmentation system developed for the SRI NIGHTINGALE system that was applied to English, Arabic, and Mandarin news shows to provide input for subsequent question-answering processes [13]. Xie (2008) used word and subword multiple scales for story segmentation and showed the robustness of subwords for reducing the impact of errors and improving story broadcast news segmentation [22]. Wu et al. (2009) used decision tree and maximum entropy to identify the positional story broadcast news boundaries locally and then used a genetic algorithm to identify the final story boundaries [27]. The BASRAH [6] and MAHER [12] systems were designed to detect story boundaries in multilingual (English and Malay) and Malay broadcast news stories using confidence measures of the ASR to minimize the WER in ASR transcription and a Euclidean distance algorithm for clustering, respectively. Finally, the current system uses LSA, which depends on deeper syntactic or semantic analysis rather than shallow techniques to identify the news story boundaries and the number of stories inside Malay

broadcast news segments. This is first attempt to process Malay broadcast news and also to determine the number of stories inside the spoken document automatically. The proposed system was tested on 18 Malay news transcriptions with accuracy 0.93 and F-measure 0.46.

VIII. Conclusions

SDR is the science that studies how to improve search and retrieval performance of data within ASR output. The identification of broadcast news story boundaries is one of the complicated tasks involved in SDR. Identifying broadcast news story boundaries plays an important role in many natural language processing applications, such as topic identification and story classification. There are not many generic frameworks for spoken document segmentation, and most of those that do exist are domain specific. To the best of our knowledge, the current study is the first one proposed for the Malay language. It utilizes the computer to help identify the stories boundary and the number of news stories automatically without first training the program. The proposed system was tested on 18 Malay news transcriptions with accuracy 0.93 and F-measure 0.46.

Acknowledgment

ZAK would like to thank Faris Mahdi Alwan (University of Sciences Malaysia, USM) for important suggestions and for helpful discussions. She owes her deepest gratitude to USM for its support of her PhD research. She also would like to extend thanks to Basrah University for their helpful support.

References

[1] Chelba, C., T.J. Hazen, and M. Salaclar, *Retrieval and Browsing of Spoken Content.* IEEE Singal Processing Magazine, 2008. **25**(3): p. pp. 39-49.

[2] Lo, W.-K., H.M. Meng, and P.C. Ching, *Multi-Scale Spoken Document Retrieval for Cantonese Broadcast News.* International Journal Of Speech Technology, 2004. **v. 7**(2-3): p. pp. 203-219.

[3] Parlak, S. and M. Saraclar, *Performance Analysis and Improvement of Turkish Broadcast News Retrieval.* IEEE Transactions on Audio,Speech and Language Processing, 2011. **v. 20**(3): p. pp. 731 - 741.

[4] LU, M.-m., et al., *Multi-Modal Feature Integration for Story Boundary Detection in Broadcast News.* IEEE, 2010: p. 420-425.

[5] Tien-Ping, T., et al., *Mass: A Malay Language LVCSR Corpus Resource.* Cocosda'09. Urumqi, China, 2009.

[6] Khalaf, Z.A. *The BASRAH System: A Method for Spoken Broadcast News Story Clustering*. in *NDT,Springer-Verlag Berlin Heidelberg*. 2012: Springer-Verlag Berlin Heidelberg

[7] Yeong, Y.L. and T.P. Tan, *Language Identification Of Code Switching Malay-English Words Using Syllable Structure Information*, in *Spoken Languages Technologies for Under-Resourced Languages (SLTU'10)*. 2010: Penang.Malaysia. p. pp. 142-145.

[8] Vinciarelli, A. and S. Favre. *Broadcast news story segmentation using social network analysis and hidden Markov models*. in *Proceedings of the 15th international conference on Multimedia*. 2007: ACM.

[9] Jain, A.K., *Data Clustering: 50 Years Beyond K-Means1.* Pattern Recognition Letters, 2010. **31**: p. pp. 651-666.

[10] Geiß, J., *Latent semantic sentence clustering for multi-document summarization*. 2011, University of Cambridge.

[11] Landauer, T.K., P.W. Foltz, and D. Laham, *An Introduction to Latent Semantic Analysis.* Discourse Processes, 1998. **25**: p. 259-284.

[12] Khalaf, Z.A. *MAHER: A Clustering System for Analyzing Spoken Broadcast News*. in *2012 4th International Conference on Electronics Computer Technology (ICECT 2012) on IEEE Catalog , Print ISBN: 978-1-4673-1850-1*. 2012. India: IEEE.

[13] M. Ostendorf, et al., *Speech Segmentation and its Impact on Spoken Document Processing.* 2007.

[14] Li, D., W.-K. Lo, and H. Meng, *Initial Experiments on Automatic Story Segmentation in Chinese Spoken Documents Using Lexical Cohesion of Extracted Named Entities.* ISCSLP, 2006: p. 693-703.

[15] Terol, R., et al., *An Application of NLP Rules to Spoken Document Segmentation Task*, in *Natural Language Processing and Information Systems*, A. Montoyo, R. Muńoz, and E. Métais, Editors. 2005, Springer Berlin Heidelberg. p. 376-379

[16] Wechsler, M., E. Munteanu, and P. Schauble, *New Approach to Spoken Document Retrieval* Information Retrieval, 2000. **v. 3**: p. pp. 173-188.

[17] Xie, L., Y.L. Yang, and Z.Q. Liu, *On the effectiveness of subwords for lexical cohesion based story segmentation of Chinese broadcast news.* Information Sciences, 2011. **181**(13): p. 2873-2891.

[18] Mengle, A.V. and D. Palmer, *Trainable News Broadcast Boundary Identification Using Feature Density.* 2004.

[19] Hsu, W., et al. *Discovery and fusion of salient multimodal features toward news story segmentation*. in *Proceedings of SPIE*. 2004.

[20] Winston, H., H.M. Hsu, and S.F. Chang. *A statistical framework for fusing mid-level perceptual features in news story segmentation*. in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. 2003: IEEE.

[21] Xie, L., C. Liu, and H. Meng. *Combined use of speaker- and tone-normalized pitch reset with pause duration for automatic story segmentation in Mandarin broadcast news*. in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. 2007: Association for Computational Linguistics.

[22] Xie, L., *Discovering salient prosodic cues and their interactions for automatic story segmentation in Mandarin broadcast news.* Multimedia Systems, 2008. **14**(4): p. 237-253.

[23] Hearst, M.A., *TextTiling: Segmenting text into multi-paragraph subtopic passages.* Computational linguistics, 1997. **23**(1): p. 33-64.

[24] Stokes, N., J. Carthy, and A.F. Smeaton, *SeLeCT: a lexical cohesion based news story segmentation system.* AI COMMUNICATIONS, 2004. **17**(1): p. 3-12.

[25] Rosenberg, A. and J. Hirschberg. *Story segmentation of brodcast news in English, Mandarin and Arabic*. in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. 2006. Stroudsburg, PA, USA: Association for Computational Linguistics.

[26] Banerjee, S. and A.I. Rudnicky, *A TextTiling based approach to topic boundary detection in meetings.* 2006.

[27] Wu, C.H. and C.H. Hsieh, *Story segmentation and topic classification of broadcast news via a topic-based segmental model and a genetic algorithm.* Audio, Speech, and Language Processing, IEEE Transactions on, 2009. **17**(8): p. 1612-1623.