MAHIR System: Unsupervised Segmentation for Malay Spoken Broadcast News Stories

Zainab Ali Khalaf and Tan Tien Ping

***Abstract*—Current studies on spoken document retrieval (SDR) systems concentrate on building strong systems using an approach that reduces the impact of automatic speech recognition (ASR) on retrieval performance. Herein we tend to propose the SDR system, the main goal of that is to reduce the effect of ASR transcription errors on retrieval performance. An automatic speech recognition system is employed to convert the Malay spoken broadcast news to text. The performance of unsupervised learning is evaluated on the Malay broadcast news using apriori algorithm.**

***Index Terms*—Spoken document retrieval, unsupervised learning, apriori algorithm, broadcast news segmentation.**

I. Introduction

A spoken document retrieval (SDR) system uses automatic speech recognition and information retrieval technologies to analyze and process multimedia documents [1]-[4]. Automatic speech recognition (ASR) systems are used to convert spoken documents (speech) into text transcription. In SDR research, well-known text-based search algorithms are applied to time-aligned ASR transcripts. This application helps automatically index and retrieve spoken content from various multimedia documents, such as radio/television broadcasts, digital library archives, call-center recordings, meetings, tutorial lectures, and web Existing call-center recordings, meetings, academic lectures, and internet user audio/video. Thus, SDR systems are designed to integrate both ASR and information retrieval (IR) technologies. Effective SDR systems provide access not only to spoken textual content but also to rich information that reflects the intended meaning and the speaker's emotional state. Efficient systems are needed because of the continually increasing amount of multimedia content and the demand to access information in these multimedia collections. Accordingly, audio searching has become more popular [1], [5], [6]. However, a number of obstacles, including the lack of overt punctuation and formatting and the difficulty in detecting story boundaries or segments, make the process of retrieving spoken documents challenging [6]-[8]. In particular, identifying word errors generated by ASR is one of the major challenges facing SDR. To ascertain that the handling and management of large news video text is done in efficacy, the spoken broadcast news need to be segmented into units of stories. Separating the news into units enable us to detect the one part where the story ends and the part when one begins in a stream of medium, like text, video or speech [5], [9]. Entirely being conscious of the fact that the story boundary detection has undergone various phases of research, this is deemed impractical because of its less-than-satisfactory performance. As the problem is thought to be difficult and too general, no specific single feature is considered adequate to handle the story boundary detection process for an abundance of broadcast news. Some amounts of efforts that have been exerted recently have demonstrated that the integration of varying features is able to enhance the performance of detection [6]. To address these problems, we conducted a case study using Malay broadcast news. Specifically, our goal was to design a system to identify news story boundaries.

Manuscript received May 20, 2014; revised July 27, 2014. This work was supported by the CS Department of Universiti Sains Malaysia (USM).

Zainab Ali Khalaf was with Basra University, Basra, IRAQ. She is now with the Department of Computer sciences, Universiti Sains Malaysia (USM), Penang, Malaysia (e-mail: zainab_ali2004@yahoo.com).

Tan Tien Ping is with the Department of Computer sciences, Universiti Sains Malaysia (USM), Penang, Malaysia (e-mail: tienping@cs.usm.my).

II. Data Source

A transcript produced manually from spoken broadcast news [10] 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.

III. Proposed System

In this study, the spoken document system was designed and used to enhance identification of news story boundaries. The document then was split into sentences. Finally, the cluster model was used to cluster the sentences into stories using the apriori algorithm.

The MAHIR system process proceeds in five stages:

Stage 1: This stage describes the ASR output. Once the recognition decoder output was generated, Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR) commands, were used to improve the ASR transcription. Then, the text document is transformed into tokens separated by spaces.

Stage 2: The algorithm described in Section III-A is used to assign a particular part of speech (POS) to each word in the document.

Stage 3: The algorithm described in Section III-B is used to remove stopping words.

Stage 4: The stemming algorithm described in Section C is applied.

Stage 5: Indexing (Section III-D) and identification of broadcast news story boundaries are conducted using a clustering algorithm (Section IV).

A. POS Tagger

POS tagging is the process of marking words in a text to correspond to a particular part of speech, and it is based on the definition and context relationship of a given word with related and adjacent words in a sentence. The POS tagger model contains a set of tags, including verb (V), noun (N), adverb (A), number (NUM_CARD), negation (NEG-PART), conjunction (CC), preposition (PREP), determiner (DET), and auxiliary verb (AU_V). Each word from the input sentence is matched with one of the tags present in the tag set. The output of the POS Tagger is then stored in an XML document. Fig. 1 shows an example of POS output [11]-[13].

B. Removal of Stopping Words

Stopping words are words that are repeated frequently in speech and have no real meaning themselves but maybe used as auxiliary verbs or prepositions (e.g., is , are, at, in, the, a, an). To avoid the noise problem created by the presence of such generic terms and to reduce the size of the index, these words are regularly deleted during the indexing stage. The Malay prepostions "ke, to," "dari, from," and "pada, at," for example, provide no real information significant to the document's topic, yet they appear frequently in almost every document collection. Removing such words helps to decrease the index size and improve the quality of the search results by retaining only the words that contribute specific information to each document [13], [14].

```
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
```

Fig. 1. POS output.

C. Stemming

Stemming describes a process used to improve the effectiveness of information retrieval. In this process, variant forms of the same word with different endings are reduced to a common stem. Stems are useful in information retrieval conducted using techniques that unify vocabularies; use of stems reduces term variants and storage space and increases the matching probability of the documents [15], [16].

The Malay language is an agglutinative language; its base words are used to form new words with new meanings by the addition of a prefix, suffix, or circumfix [17], [18]. Many different suffixes can be placed at the end of words, but only five prefixes usually are appended to the start of words (ber-, per-, ter-, me-, and pe-). These five native prefixes may produce deletions, insertions, or assimilation contact with the base word [13], [17], [18].

Fig. 2 and Table I show Malay word structure, some example of Malay prefix, suffix and circumfix respectively .

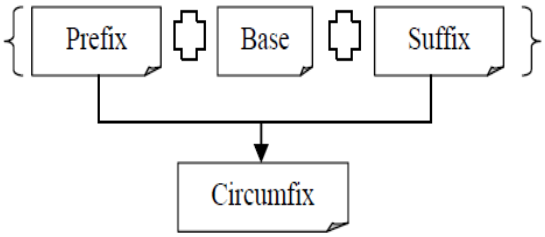

Fig. 2. A Malay word structure [13], [18].

TABLE I: SHOWS SOME EXAMPLE OF MALAY PREFIX, SUFFIX AND CIRCUMFIX [18]

Term	Example
Prefix	'ber','di','juru','ke','pem', 'meng', 'peng','per', 'ter', 'mem', 'men', 'pen', 'me', 'pe', 'be', 'se', 'te'
Suffix	'nya','kan','an','i','kah','lah','tah'
Circumfix	'ber…an', 'per…an', 'ter…kan', 'mem…kan', 'pem…an', 'pen…an', 'pe…an', 'ke…an', 'se…an', 'te…kan', 'di…kan', 'ber…kan', 'me…i','men…i', 'meng…i','menge…kan', 'penge…an','peng…an'

D. Indexing

The result of the previous stage is a list of words that represent documents in the collection. To facilitate an efficient search through these words, an index for every word in the document is created during the information retrieval stage. A word's index represents its frequency of occurrence [13], [19].

IV. CLUSTERING MODEL

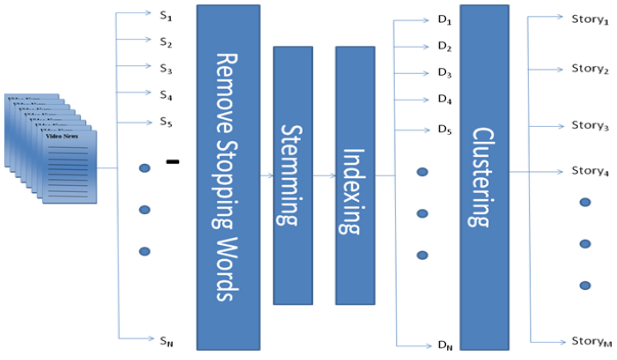

Fig. 3. Clustering model using apriori algorithm.

Varied types of fundamental modes of understanding and

learning exist, one of which is the organization of data into meaningful groupings. Specifically, cluster analysis is a formalized mode of reading into the methods and algorithms for grouping or clustering objects following their measured or perceived intrinsic characteristics or the common properties they possess [20], [21]. Fig. 3 is a data flow diagram showing how the proposed MAHIR system proceeds, ultimately using apriori algorithm in the clustering step.

V. The Apriori Algorithm

Apriori algorithms are those that are accepted for association rule mining. The apriori algorithm generates candidate set throughout every pass. It decreases the dataset by removing the infrequent itemsets that do not satisfy the minimum support value from the candidate sets. However, apriori algorithms have two limitations. First, the complex candidate generation process requires a great deal time, space, and memory. A second disadvantage is that the apriori algorithm requires multiple scans of the database to generate the candidates. Many algorithms based on existing apriori algorithms have been designed with modifications or enhancements [8], [22]. In this study, we modified and enhanced an existing apriori algorithm to meet our specific aims.

The apriori algorithm manipulates prior knowledge with regards to a significant quality of recurrent item sets. Thus, the a priori algorithm of an item set rules out that all non-empty subgroups of a recurrent item set should be recurrent. Put simply, if an item set is considered to be infrequent (implying that if it does not fulfill the minimum threshold level of support), any groups within this item set will also be infrequent because they cannot occur any more recurrently than the original counterpart [20]. Each object is denoted as a vector, which highlights which items are included in multiple objects. The vectors are scrutinized to search for which items are recurrently joined together by various objects (i.e., linked or associated together). These co-occurrences are given in the form of the rules of association [8], [20]:

$$LHS \Rightarrow RHS[Support, Confidence]$$

where LHS is the left hand side and RHS is the right hand side, with a particular value of the confidence and support.

Confidence and support serve as a yardstick for determining the quality of a particular rule in terms of its utility (strength) and certainty [8], [20]. Support provides an idea of how many of the examples (transactions) of a dataset are used to produce the rule, including elements of both LHS and RHS. Confidence further denotes the number of examples (transactions) that contain elements from the LHS and the RHS. Normally, measured values are given in percentages. An association rule is deemed to be useful if the minimum values of confidence and support are fulfilled, and these are specified by the user in question (domain expert).

More specifically, support is defined as the probability (frequency) of the complete rule concerning D, where D is a group of transactions referred to as the transactional database. It is the percentage of transactions that contain A and B within the total number of transactions (i.e., the probability of both A and B that occur in shared D) [20]:

$$Support(A \Rightarrow B) = P(A \cup B) = \frac{\|T \in D \mid A \cup B \subseteq T\|}{\|D\|}$$

Confidence refers to the strength of implication in the rule. . It is defined as the percentage of transactions that contain A, B within the number of transactions that contain A (i.e., conditional probability of B given A) [20]:

$$Confidence(A \Rightarrow B) = P(B \mid A) = \frac{\|T \in D \mid A \cup B \subseteq T\|}{\|T \in D \mid A \subseteq T\|}$$

VI. Experimental Results

In this study, we designed an SDR system and tested its ability to detect boundaries in Malay spoken broadcast news stories. Once the recognition decoder output was generated, the MAP and MLLR commands 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.

The current system was tested on a sample of 4698 sentences and ~400 stories from 18 Malay spoken broadcast news shows containing different kinds of news, including local news, political news, and sport news. The range of story length was 1 to 167 sentences. The process of the spoken broadcast news with word errors rate ~ 34%.

In order to compute the story boundaries between automatic segmentation (Clusters) and manual segmentation (Classes), the proposed system implements F measures.

$$F_1 = \frac{2 \times P \times R}{P + R}$$

where, Precision (P) is defined as a number of sentences common in both Clusters and Classes divide with number of sentences in Clusters.

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

Recall (R) is defined as a number of sentences common in both Clusters and Classes divide with number of sentences in Classes.

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

After splitting document sentences, filtering stop words and stemming, apriori algorithm was computed for P sentence groups with minimum support = 45%. The following algorithm attempts to find item sets which are common between the groups P of the given sentences.

1) Find the first candidate item sets (C_1).
2) Determine the set of all points in C_1 that satisfy minimum support to generate the first frequent item sets (L_1).
3) Create C_2 from L_1 joins with itself.
4) Repeat the process to obtain L_4, L_5, …, L_h until no more candidates are obtained.
5) Combine the candidates (the sentences) that satisfy the minimum support together.
6) Repeat step 1 to 6 until completes the entire document.

The performance of apriori algorithm in sentence clustering in this experiment was automatically compared with the reference file (Manual segmentation). We have achieved average F1 measure as 0.524 for automatic segmentation of the 18 spoken broadcast news shows compared with 0.7 for manual reference segmentation of the same 18 spoken broadcast news shows. Fig. 4 shows the results when apriori method was used to process the 18 news shows.

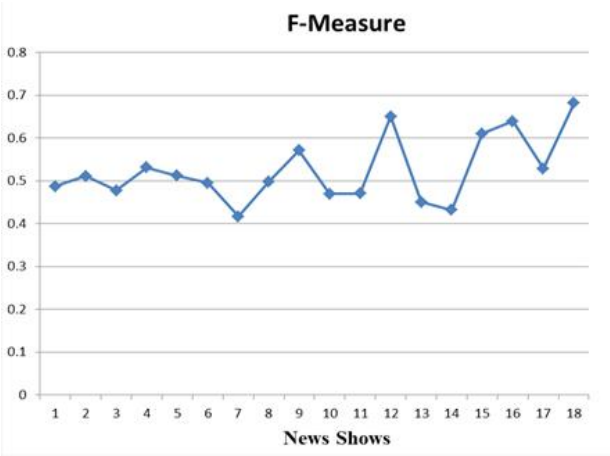

Fig. 4. Shows F-Measure for 18 spoken broadcast news shows were analyzed.

VII. Related Works

Spoken documents (speech) are converted into text transcription by using a type ASR systems The process of retrieving information from spoken documents has been described as a challenging task facing multimedia retrieval systems by Schauble (1997) [23]. The other important point deals with the development of efficient information retrieval techniques. Such technique usually faces a number of challenges when providing a mechanism for assessing and extracting information from spoken audio formats, including TV, radio, and video materials. More importantly, the interest in such processes is said to have increased because of the initial SDR track within the Text REtrieval Conference (TREC) [24].

It is quite evident that there are other attempts, in addition to the present study, that have been conducted to detect stories boundaries (i.e., to cluster stories) within spoken documents. Among such attempts, the study carried out by Hazen et al. [3] where they used two techniques. In the first one, the researchers tried to decrease the recognition errors by applying an explicit model for the identification of the presence of out-of-vocabulary (OOV) words. In the second technique a confidence scoring model is applied in order to identify potentially misrecognized words from a group of confidence features already extracted from the recognition process. In another study, the language leveraging algorithms within a spoken information search system has been used by Akbacak system [25] to initially process multilingual (English and Spanish) broadcast news for which training resources are limited. In order to process Turkish language documents recognition and indexing units are used as sub-words units by Parlak *et al.* to reduce both the OOV rate and the index alternative recognition hypothesis to handle ASR errors [26]. Some researcher such as Lo *et al*. [4] concentrated on the application of a multi-scale paradigm for Chinese SDR to simply improve retrieval performance. BASRAH [18] system has been designed to detect story boundaries in multilingual (English and Malay) using Confidence Measures (CMs) of the ASR. This process is basically done to reduce the rate of word error in ASR transcription and a Euclidean distance algorithm for clustering news stories. Finally, the current system, MAHIR, was designed to detect story boundaries in Malay broadcast news stories. This system uses MAP and MLLR of the ASR to minimize the word error rate in ASR transcription and apriori algorithm for clustering.

VIII. Conclusions

The importance of SDR systems has increased with the passage of time because of the urgent daily need to access spoken archives on the web and in other archive sources.

Our designed system is a useful tool for retrieving information from broadcast news. In future studies, we will apply the current system to another application, such as classification broadcast news stories, to test its usefulness in other aspects of document retrieval.

Acknowledgement

ZAK owes her deepest gratitude to USM for financial support in her PhD study.

References

[1] E. Arisoy *et al.*, "Turkisk broadcast news transcription and retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 874-883, 2009.
[2] C. Chelba *et al.*, "Retrieval and browsing of spoken content," *IEEE Singal Processing Magazine*, vol. 25, pp. 39-49, 2008.
[3] H. Jiang *et al.*, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech and Language*, vol. 16, pp. 49–67, 2002.
[4] W.-K. Lo *et al.*, "Multi-scale spoken document retrieval for Cantonese broadcast news," *International Journal of Speech Technology*, vol. 7, pp. 203-219, 2004.
[5] G. Senay *et al.*, "A segment-level confidence measure for spoken document retrieval," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, vol. 11, pp. 5548 - 5551.
[6] M. Ostendorf *et al.*, "Speech segmentation and its impact on spoken document processing," *Signal Processing Magazine*, vol. 25, issue 3, 2007.
[7] C. Amitkumar *et al.*, "Multimedia data mining: State of the art and challenges," *Multimedia Tools and Applications*, vol. 51, issue 1, pp. 35-76, 2011.
[8] M.-M. Lu *et al.*, "Multi-modal feature integration for story boundary detection in broadcast news," presented at International Symposium on Chinese Spoken Language Processing (ISCSLP2010), Tainan, Taiwan, 2010.

[9] H. Diao *et al*., "The application of improved K-nearest neighbor classification in topic tracking," *Proceeding IEEE*, 2010.
[10] T. Tien-Ping *et al*., "Mass: A Malay language LVCSR corpus resource," in *Proc. 2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, 2009, pp. 25-30.
[11] B. Megyesi, "Brill's PoS tagger with extended lexical templates for Hungarian," in *Proc. the Workshop (W01) on Machine Learning in Human Language Technology: ACAI'99*, 1999, pp. 22-28.
[12] B. Megyesi, "Brill's rule-based part of speech tagger for Hungarian," Department of Linguistics, Stockholm University, 1998.
[13] Z. A. Khalaf, "MAHER: A clustering system for analyzing spoken broadcast news," presented at 4th International Conference on Electronics Computer Technology (ICECT 2012), India, 2012.
[14] S. E. Johnsont *et al.*, "The Cambridge University spoken document retrieval system," in *Proc. International Conference on Acoustics, Speech And Signal Processing (ICASSP)*, March 1999, vol. 1, pp. 49-52.
[15] M. Adriani *et al.*, "Stemming Indonesian: A confix-stripping approach," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 6, no. 4, pp. 1-33, 2007.
[16] A. Hartl. (2010). Other tips & tricks: Word stemming in Java with word net and JWNL. [Online]. Available: http://tipsandtricks.runicsoft.com/Other/JavaStemmer.html.
[17] B. R. Malacon, "Computational analysis of affixed words in Malay language," *Internal Publication*, vol. USM, 2004.
[18] Y. L. Yeong and T. P. Tan, "Language identification of code switching Malay-English words using syllable structure information," presented at the Spoken Languages Technologies for Under-Resourced Languages (SLTU'10), Penang, Malaysia, 2010.
[19] Z. A. Khalaf, "The BASRAH system: A method for spoken broadcast news story clustering," in *Proc. NDT*, April 2012, pp. 126–134.
[20] K. J. Cios *et al*., *Data Mining a Knowledge Discovery Approach, Springer*, 2007, ch. 9-10, pp. 257-306.
[21] A. K. Jain, "Data clustering: 50 years beyond K-Means1," *Pattern Recognition Letters*, vol. 31, pp. 651-666, 2010.
[22] R. Agrawal *et al*., "Mining association rules between sets of items in large databases," in *Proc. Conf. ACM SIGMOD*, 1993, pp. 207-216.
[23] P. Schäuble, *Multimedia Information Retrieval: Content-based Information Retrieval from Large Text and Audio Databases*, Boston: Kluwer Academic Publishers, 1997.
[24] J. S. Garofolo *et al.*, "The TREC spoken document retrieval track: A success story," presented at the RIAO, 2000.
[25] M. Akbacak, "Rebust spoken document retrieval in multilingual and nosiy acoustic envernments," Ph.D. dissertation, Dept. of Electrical, Computer, and Energy Engineering, Colorado Univ., USA, 2009.
[26] S. Parlak and M. Saraclar, "Performance analysis and improvement of Turkish broadcast news retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 731 - 741, 2011.

Zainab A. Khalaf obtained her M.Sc. in computer science in 2001 from Basra University, Basra, Iraq. During 2001-2010, she was an assistant professor at Basra University. Her research interests are natural language processing, information retrieval, data mining and speech recognition. She has published papers at top conferences and journals.

She is currently a candidate for a PhD fellowship at the School of Computer of Sciences at the Universiti Sains Malaysia (USM). She awards USM fellowship in 2011. Her current research interests include spoken document retrieval and speech recognition.

Tien-Ping Tan obtained his PhD in computer science from Université Joseph Fourier, France in 2008. His research interest is in the field of multilingual automatic speech recognition, speech search and Malay speech processing.

He is currently a senior lecturer and researcher in Universiti Sains Malaysia, Penang, Malaysia.