BROADCAST NEWS SEGMENTATION USING AUTOMATIC SPEECH RECOGNITION SYSTEM COMBINATION WITH RESCORING AND NOUN UNIFICATION

by

ZAINAB ALI KHALAF

Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy

JULY 2015

DEDICATION

To my beloved Mother, for her prayers to me.

To the soul of my father, God bless his soul,

who encouraged me to be the best I can be, to have high expectations and to fight hard for what I believe. He always provided me with best opportunities in life. I feel he is always with me supporting and guiding.

To my great young brother

(Maher)

who passed away from this life early in June 18, 2008.

He was a role model for me

and

I will keep him in my heart forever.

ACKNOWLEDGEMENTS

بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيم

In the name of Allah the most gracious the most merciful first and foremost, Praise Allah, Lord of the worlds and prayers and peace be upon the master of messengers the Prophet (PBUH) Mohammed, our leader in this life until the closing.

I would like to thank my thesis advisors Dr. Tan Tien Ping and Dr. Wong Li Pei for their valuable suggestions, guidance and patience throughout the lengthy process of preparing this thesis. I would also like to thank the thesis committee members, my research group colleagues for the discussions on many issues. I would like also to thank the sponsors: TWAS-USM, and my family, who provided financial support during the course of this research. In addition, especial thanks my former advisor Dr. Cheah Yu-N, who has continued to support me throughout the years. Lastly, I owe a huge debt of gratitude to my mother, my wonderful sisters, brothers and my friends for all of their support.

TABLE OF CONTENTS

CHAPTER 4: PROPOSED BROADCAST NEWS SEGMENTATION APPROACHES

CHAPTER 5: EXPERIMENTAL RESULTS

CHAPTER 6: CONCLUSION AND FUTURE STUDIES

LIST OF TABLE

LIST OF FIGURES

LIST OF ABBREVIATIONS

AM	Acoustic Model
ASR	Automatic Speech Recognition
BAYCOM	Bayesian Combination
CMU	Carnegie Mellon University
CN	Confusion Network
CNC	Confusion Network Combination
CS	Confidence Score
DP	Dynamic Programming
DT	Decision Trees
DTW	Dynamic Time Warping
EM	Expectation-Maximization
EPA	Expected Phone Accuracy
FFT	Fast Fourier Transform
GS	Gold Standard
HMM	Hidden Markov Model
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LE	Laplacian Eigenmaps
LM	Language Model
LPC	Linear Predictive Coding
LPCC	Linear Predictive Coding Cepstrum
LSA	Latent Semantic Algorithm
LVCSR	Large Vocabulary Continues Speech Recognition

MAP	Maximum A Posteriori
MFC	Mel-Frequency Cepstral
MFCC	Mel-Frequency Cepstral Coefficients
MLE	Maximum Likelihood Estimate
MLLR	Maximum Likelihood Linear Regression
MLP	Multi-Layer Perceptron
NIST	National Institute of Standards and Technology
OOV	Out-Of-Vocabulary
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
PM	Pronunciation Model
POS	Part-Of-Speech tagger
ROVER	Recognized Output Voting Error Reduction
SCTK	Scoring Toolkit
SDR	Spoken Document Retrieval
SPSS	Statistical Package for Social Sciences
SVD	Singular Value Decomposition
VSM	Vector Space Model
WCN	Word Confusion Network
WER	Word Error Rate
WTN	Word Transition Network

SEGMENTASI BERITA SIARAN MENGGUNAKAN PENDEKATAN GABUNGAN SISTEM PENGECAMAN PERTUTURAN AUTOMATIK DENGAN PENILAIAN SEMULA DAN PENYATU KATA NAMA

ABSTRAK

Siaran berita memaklumkan perkembangan terbaru, peristiwa dan isu-isu terkini yang berlaku di dunia kepada penonton. Pada masa kini, berita yang disiarkan boleh diakses dengan mudah atas talian. Peningkatan yang pesat dalam jumlah siaran berita daripada media massa tradisional seperti radio, televisyen, dan televisyen kabel boleh dicapai menerusi Internet. Selain itu, dengan adanya telefon bimbit dengan kamera yang baik, perkara ini telah membolehkan pengguna untuk merakam video yang menarik dan dikongsi dengan semua orang. Kini, terdapat keperluan terhadap sistem yang boleh mengakses dan mencari kandungan siaran berita dengan berkesan dan cepat. Untuk membolehkan pencarian kandungan pertuturan dalam siaran berita, kandungan pertuturan tersebut perlu terlebih dahulu ditukar kepada teks. Pemprosesan automatik sumber siaran berita memerlukan suatu sistem pengecaman pertuturan automatic (PPA) untuk menyahkod pertuturan ke dalam transkripsi teks bertulis.

Transkrip PPA biasanya ialah dokumen tidak berstruktur yang terdiri daripada perkataan, tanpa pemformatan (iaitu tanda baca, dan huruf besar/kecil). Selain itu, sistem ASR menghasilkan kesilapan yang banyak disebabkan oleh beberapa faktor yang mengurangkan prestasi PPA. Oleh sebab itu, masalah-masalah ini boleh mengurangkan prestasi pemprosesan peringkat tinggi seperti pencarian, perumusan dan penterjemahan. Sistem segmentasi dokumen pertuturan (SDP) ialah sistem yang menyahkodkan pertuturan kepada transkrip dan seterusnya membahagikannya kepada unit logikal sebelum membolehkan pemprosesan seterusnya seperti pencarian, penyimpulan dan terjemahan dilakukan.

Pentranskripan berita secara manual adalah terlalu mahal dan mengambil masa yang lama. Oleh itu, tanpa sistem SDP, pengaksesan dan pemprosesan arkib audio akan terhad kepada dokumen teks yang telah disalin secara manual dan dibahagikan oleh manusia.

Pendekatan penggabungan sistem PPA dicadangkan untuk mentranskripkan siaran berita Melayu secara automatik. Pendekatan ini menggabungkan hipotesis yang dihasilkan oleh sistem pengecaman pertuturan automatik selari untuk menghasilkan hipotesis yang lebih tepat. Setiap sistem PPA menggunakan model bahasa yang berbeza, salah satu merupakan model domain generik dan model yang lain ialah model domain khusus. Idea utama adalah untuk mengambil kesempatan terhadap pengetahuan ASR yang berbeza untuk meningkatkan kejituan keputusan penyahkodan ASR. Pendekatan yang dicadangkan dibandingkan dengan pendekatan gabungan konvensional, pengurangan kesilapan pengundian output pengecaman (ROVER). Pendekatan yang dicadangkan mengurangkan kesilapan penyahkodan daripada 34.5% kepada 30.6% dan 30.1%, dan pendekatan ini adalah lebih baik daripada pendekatan ROVER konvensional.

Tambahan pula, untuk mengenal pasti sempadan topik dalam transkrip PPA ialah satu cabaran kerana kesilapan yang dijana daripada sistem PPA serta ketiadaan tanda bacaan dan pemformatan. Oleh itu, pendekatan segmentasi topic tradisional (contohnya algoritma TextTiling) tidak boleh menghasilkan sempadan topik yang baik dengan dokumen-dokumen yang dihasilkan daripada sistem PPA. Bagi menangani kesilapan-kesilapan dalam penyahkodan transkrip PPA yang boleh menyebabkan masalah yang ketara dalam padanan perkataan dan hubungan saling berkait dalam segmentasi topik, dua pendekatan dicadangkan: penyatuan kata nama dan pengubahsuaian pendekatan TextTiling. Penyatuan kata nama adalah berdasarkan maklumat fonologi untuk mengenal pasti kata nama yang dan sebutan yang serupa, menggabungkan kata nama dan kemudian digunakan untuk segmentasi topik. Pengubahsuaian TextTiling adalah berdasarkan kepada algoritma apriori. Keputusan yang dikutip daripada segmentasi topik menunjukkan bahawa penyatuan kata nama dan algoritma TextTiling yang diubah suai memberikan prestasi yang lebih baik berbanding dengan algoritma TextTiling asal. Pendekatan TextTiling yang diubahsuai dengan penyatuan kata nama mencapai F-ukuran 0.71; sementara pendekatan tanpa penyatuan kata nama mencapai F-ukuran 0.62.

BROADCAST NEWS SEGMENTATION USING AUTOMATIC SPEECH RECOGNITION SYSTEM COMBINATION WITH RESCORING AND NOUN UNIFICATION

ABSTRACT

Broadcast news keeps viewers informed about the latest developments, events and issues occurring in the world. Nowadays, broadcast news can be easily accessed online. There is a rapid growth in the amount of news broadcasted from the traditional mass media such as radio, television, and cable television that are made available on the Internet. Besides that, with the availability of mobile phones with a good camera, it has allowed users to record interesting videos and shared them with everyone. Now more than before, there is a need for systems capable of accessing and searching the contents of the broadcast news effectively and quickly. To allow the searching for the spoken contents in broadcast news, the spoken contents have to be first converted to text. Automatic processing of broadcast news sources requires automatic speech recognition (ASR) system in order to decode speech into a written text transcription.

Typical ASR transcription is an unstructured document that includes only words, without further formatting (i.e. punctuations, and capitalization). Moreover, ASR system produces substantial errors due to several factors that are degrading the ASR performance. Thus, these problems reduce the performance of a high-level processing such as searching, summarizing, and translation. Spoken document segmentation (SDS) is a system that decodes broadcast news to transcription and then segments the transcription to the logical unit before allows subsequent high-level processing to be carried out.

Manual news transcription and topic segmentation are too expensive and take a long time. Hence, without an SDS system, access to audio archives and searches within them would be restricted to the limited number of textual documents that have been manually transcribed and segmented by humans. Multiple hypotheses are useful because the single best recognition output still has numerous errors, even for state-of-the-art systems.

Two ASR system combination approaches are proposed for automatic transcribing Malay broadcast news. These approaches combine the hypotheses produced by parallel automatic speech recognition (ASR) systems. Each ASR system uses different language models, one which is generic domain model and another is domain specific model. The main idea is to take advantage of different ASR knowledge to improve ASR decoding result. The proposed approaches are compared with a conventional combination approach, the recognizer output voting error reduction (ROVER). The proposed approaches reduce the decoding error from 34.5% to 30.6% and 30.1, and these approaches are better than the conventional ROVER approach.

Moreover, identifying the topic boundaries in ASR transcription is a challenge because of the errors generating from ASR system as well as the absence of overt punctuation and formatting. Thus, the traditional topic segmentation approaches (e.g. TextTiling algorithm) cannot work properly with these documents that result from ASR system. To address the decoding errors in ASR transcripts that can cause significant difficulties in word matching and interlinked relationships in topic segmentation, two approaches are proposed: noun unification and modified TextTiling approach. Noun unification is based on phonological information to identify similarly pronounced nouns, unified the nouns and then in turn is used for topic segmentation. A modified TextTiling text segmentation algorithm is based on an apriori algorithm. The results collected from topic segmentation provide the evidence that noun unification and the modified TextTiling algorithm give better performance compared to the original TextTiling algorithm. The modified TextTiling with noun unification achieved an F-measure of 0.71; without the noun unification, it achieved an F-measure of 0.62.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Nowadays, multimedia data can be easily accessed online due to the rapid growth in the amount of data from conventional mass media (e.g. radio, television and cable television) available on the Internet. In addition, mobile phones with high-quality cameras and internet facilities allow its users to record and share interesting videos. Therefore, there is an increased need for systems capable of accessing and searching the contents of multimedia data effectively and quickly. The search for specific multimedia content (speech, text, video, and image data) is useful to find interesting information of implicit knowledge from multimedia documents. However, multimedia analysis can be argued is still a challenging problem. To allow the processing of speech content contained in online multimedia documents, it have to be first converted to text. Spoken document segmentation (SDS) system is a system that transcribes spoken files to texts and subsequently segments the transcription into logical units before allowing processing applications to be performed such as searching, summarizing (Christensen et al., 2005), and translation (Cho et al., 2012; Petiz et al., 2012). For example, spoken document retrieval (SDR) achieves the functions of transcription, segmentation and processing by combining spoken document segmentation system and information retrieval system (Chelba et al., 2008; Bhatt et al., 2011).

The earliest work on SDR can be traced back to the Informedia project at Carnegie Mellon University (CMU). The Informedia project focuses on the automated retrieval of broadcast television news using various techniques that include speech recognition, natural language processing, and image processing in order to improve search and discovery performance in the video medium. The Informedia project allows the user to match the queries to automatically transcribed speeches. However, the project applies only the most possible decoding of an acoustic signal, which is selected from a wide range of hypotheses evaluated during the ASR process. The researchers involved in the Informedia project concluded that n-best ASR hypotheses seem most promising to improve information retrieval performance. Nonetheless, words incorrectly identified may reduce matches with query terms, and consequently decrease retrieval performance (Hauptmann et al., 1997; Kanade et al., 1998). Since then, several related projects have adopted this framework, particularly the SDR system of TREC (Text Retrieval Conference) evaluations (Hauptmann et al., 1997; Garofolo et al., 1998 ; Chen et al., 2012).

TREC-SDR uses the broadcast news as a platform to test its performance. There are 550 hours of speech from the “Topic Detection and Tracking” phase 2 collections (TDT-2) of broadcast news. These collections of broadcast news were compiled between 1998-1999 from different sources, such as ABC, CNN and the Voice of America, and then manually segmented into approximately 21,500 stories. The word error rate (WER) was 14.5% for the closed caption-quality transcripts for video broadcasts and 7.5% for radio broadcasts (Garofolo et al., 1999; Voorhees et al., 2000). For IR evaluation in TREC, two different types of the query containing 50 topics are applied: single-sentence descriptions and keyword descriptions. Given the large data set, only a subset of the files was manually transcribed. In addition, closed captions were used as a reference (gold standard) for text-based retrieval. The major observation from the SDR system of the TREC evaluations was that no marked degradation in retrieval performance was observed when using ASR transcripts rather than approximate manual transcripts (Garofolo et al., 2000; Voorhees et al., 2000). Fig. 1.1 shows the typical SDR architecture.

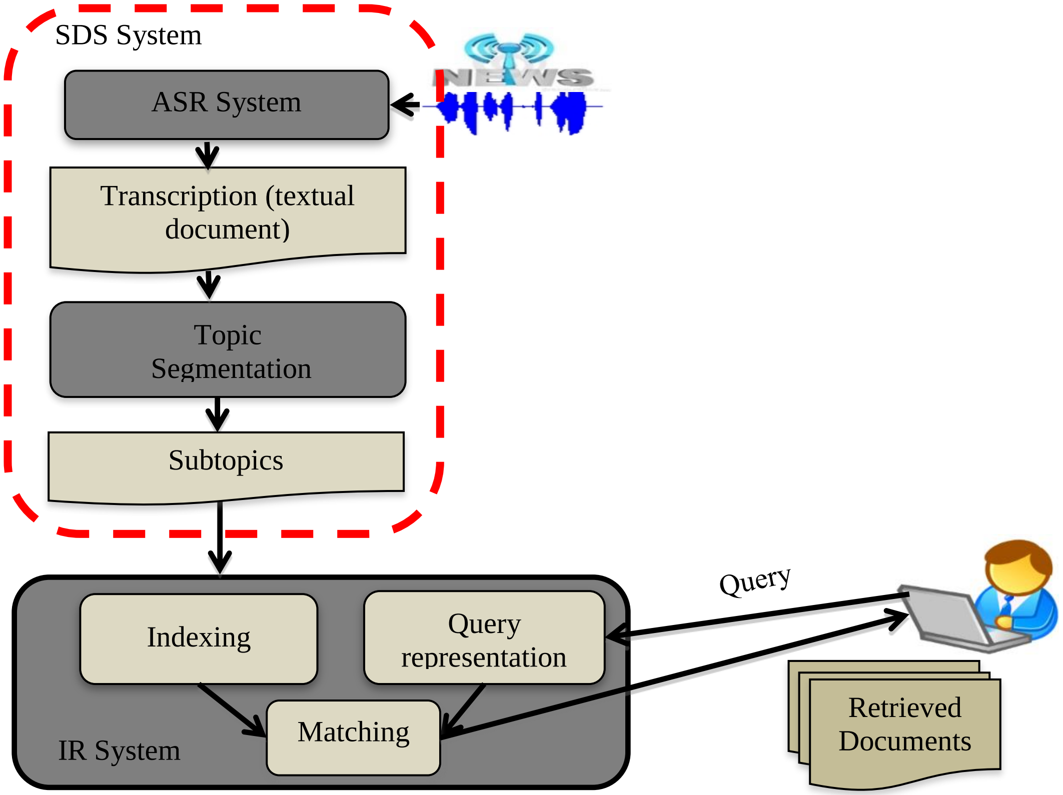


Figure 1.1: The typical SDR architecture

1.2 Problem Statement

An SDS system consists of ASR system, and topic segmentation. ASR system produces text transcripts of speech files, which are divided into coherent units using topic segmentation algorithms. The units are processed afterward with standard information retrieval algorithms adapted to this task. The intrinsic problems of the SDS system are discussed in the following subsections.

1.2.1 ASR Problems

One of the foremost challenges facing a SDS system is the WER produced by ASR, which can degrade the effectiveness of the SDS system. Fig. 1.2 shows a transcription decoded by the ASR system (HYP) compared to a manual reference (REF).

HYP	timbalan perdana menteri berucap pada majlis pertubuhan rundingan undang undang asia afrika salcon di new delhi (1) majlis ini memberi tumpuan kepada tadbir urus kelompok dalam *** pada itu persatuan kemungkinan isu dan cabaran (2) *** *** tanzimuddin berkata di malaysia kerajaan memperkenalkan gagasan tua yang tidak bercahaya (3) ia penting sebagai landasan kawasan bentong dan *** kebangsaan (4) beliau berkata malaysia percaya bahawa bagi mengekalkan dengan pas tidak berani urus dasar awam itu mesti digerakkan ke arah memenuhi keperluan rakyat (5) pada majlis ini timbalan perdana menteri melancarkan buku *** *** pihak tentera berhubung perundangan antarabangsa (6) penulisnya setiausaha agung yang perlu yang berpangkalan di new delhi *** rahman mohd yang juga ahli akademik (7) selepas mengakhiri lawatan ke *** india tinting berikut timbalan perdana menteri *** (8)

	jas pindah tv tiga kini menjadi (9) syarikat berkaitan kerajaan glc disaran membantu syarikat bumiputera yang mengalami masalah kewangan ekoran ketidaktentuan ekonomi global (10)
REF	timbalan perdana menteri berucap pada majlis pertubuhan perundingan undang-undang asia afrika aalco di new delhi (1) majlis ini memberi tumpuan kepada tadbir urus global dalam abad kedua puluh satu kemunculan isu dan cabaran (2) tan sri muhyiddin berkata di malaysia kerajaan memperkenalkan gagasan rakyat didahulukan pencapaian diutamakan (3) ia penting sebagai landasan dasar dan program kebangsaan (4) beliau berkata malaysia percaya bahawa bagi mengekalkan demokrasi dan tadbir urus dasar awam mesti digerakkan ke arah memenuhi keperluan rakyat (5) pada majlis ini timbalan perdana menteri melancarkan buku mengenai perspektif asia afrika berhubung perundangan antarabangsa (6) penulisnya setiausaha agung aalco yang berpangkalan di new delhi profesor rahmat mohammad yang juga ahli akademik (7) selepas mengakhiri lawatan ke new delhi destinasi berikut timbalan perdana menteri chennai (8) josephine daz tv tiga new delhi (9) syarikat berkaitan kerajaan glc disaran membantu syarikat bumiputera yang mengalami masalah kewangan ekoran ketidaktentuan ekonomi global (10)

Figure 1.2: An example of ASR transcription

1. Data quantity: This is a major factor that influences ASR performance. Large amount of data are needed to train robust ASR models: hundreds of hours of transcribed speech data for training acoustic model and Gigabytes of text in order to train the language

model. For example, as shown in Fig. 1.2, the word sequence "*profesor rahmat mohammad*" might be a popular name, thus the occurrence of the string in text may be low. Therefore, this word sequence may be misrecognized as the more frequently occurring common word sequence "*rahman mohamad.*"

2. Environmental conditions: The presence of noise in the recording environment is an important factor that reduces the performance of ASR systems. Hence, the noisy condition of speakers typically increases WER.

3. Lexical content disparity: Lexical content variation is related on the accent of the speaker, gender, emanation status, speech type (read speech or spontaneous). For example, each speaker pronounces sound differently when talking (Akbacak, 2009). For example, as illustrated in Fig. 1.2, the word "*global*" is recognized correctly, but in sentence 2 misrecognized with word "*kelompok*" for the same speaker due to pronunciation similarity. Moreover, some words sound seems very similar, which renders it very harder to distinguish them (e.g. "word" and "world").

4. Out-of-vocabulary problem: Out-of-vocabulary (OOV) words are unknown words that are unrecognized by an ASR system. These words bring about recognition errors that are usually deleted or substituted by other words included in recognition vocabulary, referred to as in-vocabulary (IV) words, or can cause the insertion of non-uttered words (Hetherington, 1995; Bazzi, 2002; Saraclar, 2004). As observed in Fig. 1.2, the word "*aalco*" is not recognized correctly in two sentences (1 and 7), because it is an OOV word and misrecognized with in-vocabulary words "*salcon*" and "*perlu*" respectively.

1.2.2 Topic Segmentation Problems

Topic segmentation is the process of segmenting and dividing a text transcription into units that are related to one topic (Hearst, 1997). Determining the topic boundaries in broadcast news transcription is a difficult process because of the impact of WER resulting from ASR system as well as the absence of explicit punctuation and formatting (Ostendorf et al., 2008). In addition, directly implementing conventional text document similarity measures, such as the widely applied cosine score, to error-document often leads to poor results (Ostendorf et al., 2008; Diao et al., 2010; Claveaua et al., 2015).

1.3 Research Questions

This research is designed to deal with the following questions:

1. How to minimize the word error rate in the ASR transcription?
2. How to detect and reduce the OOV words in ASR transcription?
3. What is a better topic segmentation algorithm that can better deal with the decoding errors?

1.4 Research Objectives

Based on the challenges highlighted above, the main objectives of this study are as follows:

1. To propose an approach capable of minimizing the WER in ASR transcription with rescoring approach at time-level.
2. To propose an approach that reduces the WER and OOV words in ASR transcription using rescoring approach at phrase-level and phone-level.
3. To propose a topic segmentation approach that can deal with ASR errors better.

1.5 Scope of the Study

The general aim of the research is to improve Malay broadcast news transcription and reduce the impact of the decoding errors on topic segmentation, in order to support useful applications such as information retrieval. This thesis is concerned with the transcription of Malay broadcast news comprising speeches of newscasters, reporters, and interviewers in noisy environments. The collection of broadcast news were acquired from local news channels in Malaysia (e.g. Television 3 (TV3) and Natseven Television (NTV7)), including different types of news. This data were collected in March 2011.

However, the code-switching (multilingual) is out of the scope in the current study.

1.6 Thesis Layout

This thesis consists of six chapters. **Chapter One** provides a general introduction to the thesis topic, including motivation for the study, significance of the study, problem statement, questions and objectives of the study, and scope of the study. **Chapter Two** presents theoretical background of the research area. **Chapter Three and Four** describe the research methodology and proposed framework, respectively. **Chapter Five** presents the results of experiments and analyses performed in the study. **Chapter Six** concludes with a summary of the results and a discussion of future work.

Chapter 2: Theoretical Background

This chapter presents a theoretical background and an overview of the ASR system combination. In addition, a background of topic segmentation will be discussed.

Chapter 3: Research Methodology

This chapter presents broadcast news segmentation approaches adapted for this research, which comprises experimental setup, datasets and evaluation metric.

Chapter 4: Proposed Broadcast News Segmentation Approaches

This chapter deals with the framework of the methodology proposed to answer the research questions. The proposed methodology consists of two sections: improving in ASR results and topic segmentation will be described.

Chapter 5: Experimental Results

The experimental results of ASR combination and topic segmentation are discussed in this chapter.

Chapter 6: Conclusion and Future Studies

The final chapter summarizes the conclusions of this thesis and provides suggestions for future studies.

CHAPTER 2

THEORETICAL BACKGROUND

2.1 Introduction

Spoken document segmentation (SDS) system is a system that transcribes spoken files to text and the text is then divided into textual units. The segmented text is subsequently be transformed to content that can be used in subsequent processing such as information retrieval, text summarization, and translation. The following figure shows a typical SDS system.

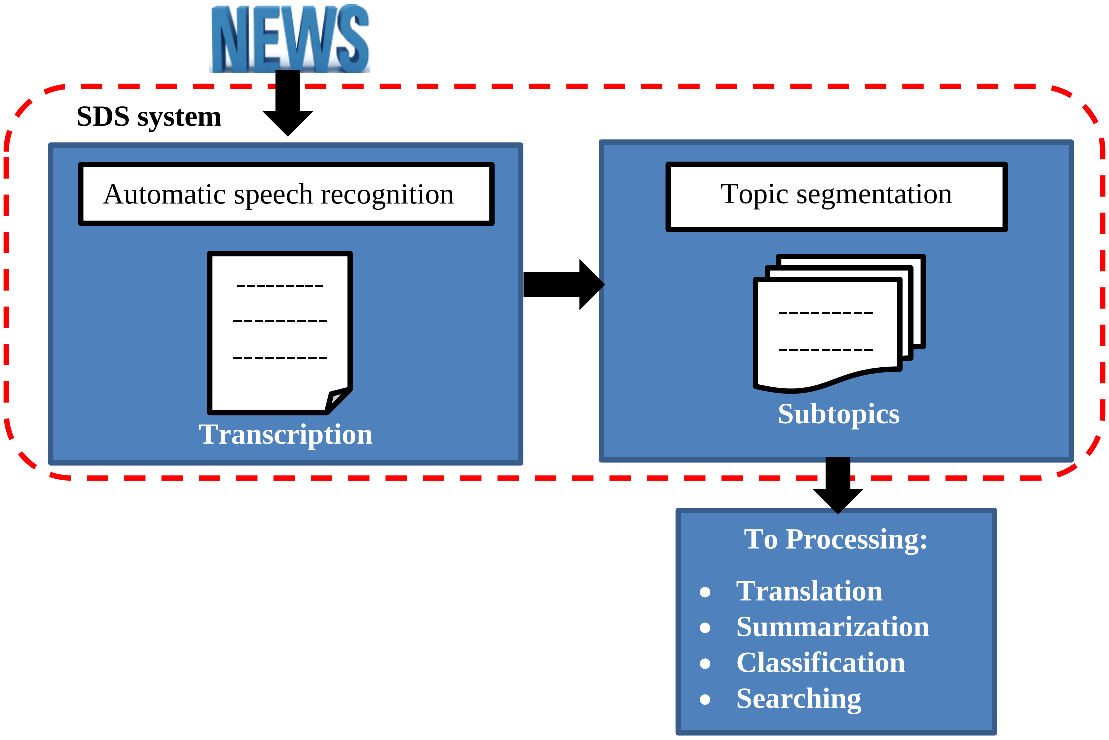
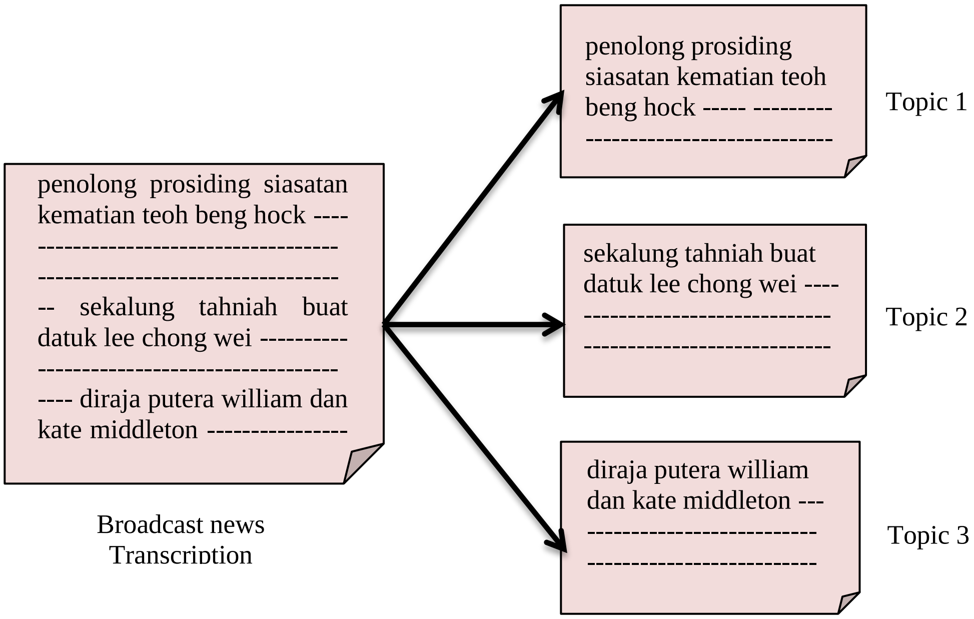


Figure 2.1: A typical SDS system

SDS system consists of an automatic speech recognition system (ASR) and a topic segmentation. The main goal of ASR is to decode speech to text transcription. ASR system is used because manual news transcription is too expensive and takes a long time, in addition, the access to multimedia archive and searches within them would be restricted to the limited number of textual documents that have been manually transcribed by humans or indexed with keywords (Grangier et al., 2005; Wu et al., 2009; Lu et al., 2010).

The resulting transcription generated by ASR system is then divided into subtopics that are related by using topic segmentation. Topic segmentation approach needed because the ASR transcription is an unstructured document (Ostendorf et al., 2008; Chelba et al., 2011). Fig. 2.2 shows an example of topic segmentation. The resulting subtopics can then be converted to indexing content that is searchable via information retrieval (IR) system. Thus, IR aims to locate information inside the indexed subtopics that is relevant to a user's query.


penolong prosiding siasatan kematian teoh beng hock -- -- sekalung tahniah buat datuk lee chong wei -- ---- diraja putera william dan kate middleton ----------------
Broadcast news Transcription
penolong prosiding siasatan kematian teoh beng hock ----- ---------
----
Topic 1
sekalung tahniah buat datuk lee chong wei ----

Topic 2
diraja putera william dan kate middleton ---

Topic 3

Figure 2.2: An example of topic segmentation approach

The growing interest in broadcast news in the community of the ASR and SDR systems is related to the following issues:

1. Broadcast news is one of the resources that have rich information.
2. Broadcast news data is considered as an ideal test platform because it contains vast different topics.
3. Broadcast news is considered a challenging domain for the ASR community because:

   - It may have different styles within (e.g., reading, two-way conversation, interviews, spontaneous speaking).
   - Broadcast news contains a different number of speakers and genders which makes it a harder task during processing.
   - There are no acoustic pauses between spoken words in continuous speech file in lieu of blanks in texts.
   - The broadcast news is recorded in a different environment that can include noise and music background. It can also contain advertisements.

There are two major problems in SDS system. The first problem is that the ASR system may have a high word error rate (WER) (e.g., more than 30%), which will significantly impact topic segmentation (Lo et al., 2004; Senay et al., 2011). The most direct method to enhance the accuracy of an SDS system is by improving ASR performance and consequently reducing recognition errors in ASR hypothesis. Researchers have addressed the problem in ASR in many ways. One way is to combine multiple ASR systems to improve the decoding by exploiting different ASR outputs to take advantage of their potential complementarities to improve the transcript (Breslin et al., 2007; Hoffmeister et al., 2008). Furthermore, the ASR can be run in different machines with different knowledge sources. Considering the above reasons, the two primary aim of this thesis is to improve ASR using two independent system combinations.

The second problem is in processing ASR outputs, which are unstructured data. Textual documents typically have punctuation and capitalization to format and segment words into sentences and their corresponding sub-sentential units. Sentences are then further organized into high-level groups in the forms of paragraphs by means of formatting. Contrastingly, when a spoken language is processed through an ASR, the output comprises an raw stream of words, and consequently, the ideal solution would be to automatically segment the spoken document and convey the results to users (Ostendorf et al., 2008; Diao et al., 2010).

Some efforts have been made by researchers to develop methods to segment a spoken document (especially in the broadcast news domain) into separate topics (stories) to enhance the performance of the SDS task (Ostendorf et al., 2008; Lu et al., 2010). Due to the importance of spoken document segmentation and the difficulty of manual segmentation, which is time-consuming and expensive, the focus of the current thesis is on improving topic segmentation. Therefore, another aim of this thesis is to improve the segmentation of the transcription into meaningful coherent topics/stories. By separating the ASR text into stories, relevant sentences that describe an event can be grouped together. This task is also considered as an essential step for subsequent high-level processing in the natural language-processing field such as machine translation, topic detection, and classification. Thus, this thesis is focused on SDS system (see the red dotted rectangle in Fig. 2.1).

This chapter is structured around two main parts: The first part will present, in section 2.2, an overview of the ASR architecture and the previous studies of the ASR system combination. In section 2.3, literature review of topic segmentation is provided. Finally, this chapter concludes with a discussion to show the limitations of existing algorithms.

2.2 Automatic Speech Recognition System

An ASR system is a system that decodes speech signals into a sequence of words. Statistical ASR systems produce the most likely words given the observed acoustic features (Ahmed, 2014). Fig. 2.3 shows a statistical ASR architecture.

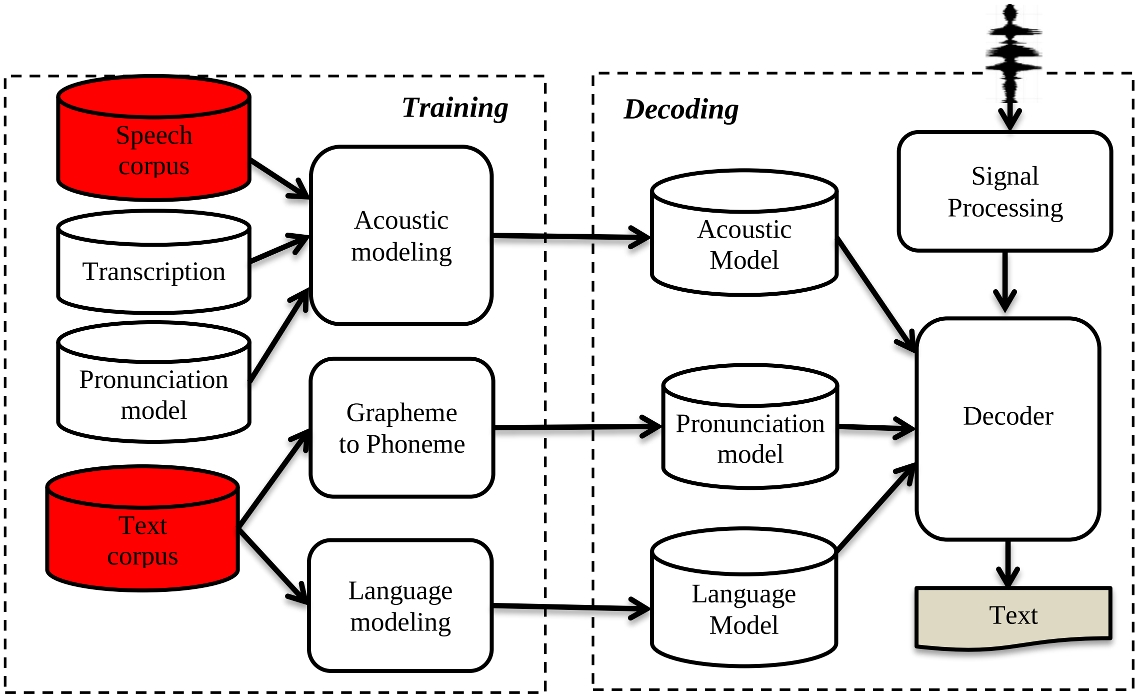


Figure 2.3: An ASR Module (Tan, 2008)

As shown in Fig. 2.3, an ASR consists of two main modules: training (learning) and decoding (recognition). Three models, an acoustic, a language and a pronunciation model, are created during training and they used to decode speech to text. The acoustic model represents elementary speech units such as phones, words, and syllables. The pronunciation model defines larger linguistic units such as syllables or words using the acoustic units. In turn, the statistical language structure and syntax are represented in the language model (Tan, 2008).

The ASR decoding module consists of two components: a signal processing front-end and a decoder. The main role of the signal processing front-end is to digitize the analog signal, extract feature observations from the signal, and then convert the signal into distinct features for recognition (Rosdi, 2008; Tan, 2008). A decoder is a search process that aims to capture the sequence of the most probable words given a series of feature observations (Mengusoglu, 2004). Fig. 2.4 shows the framework of a decoding module.

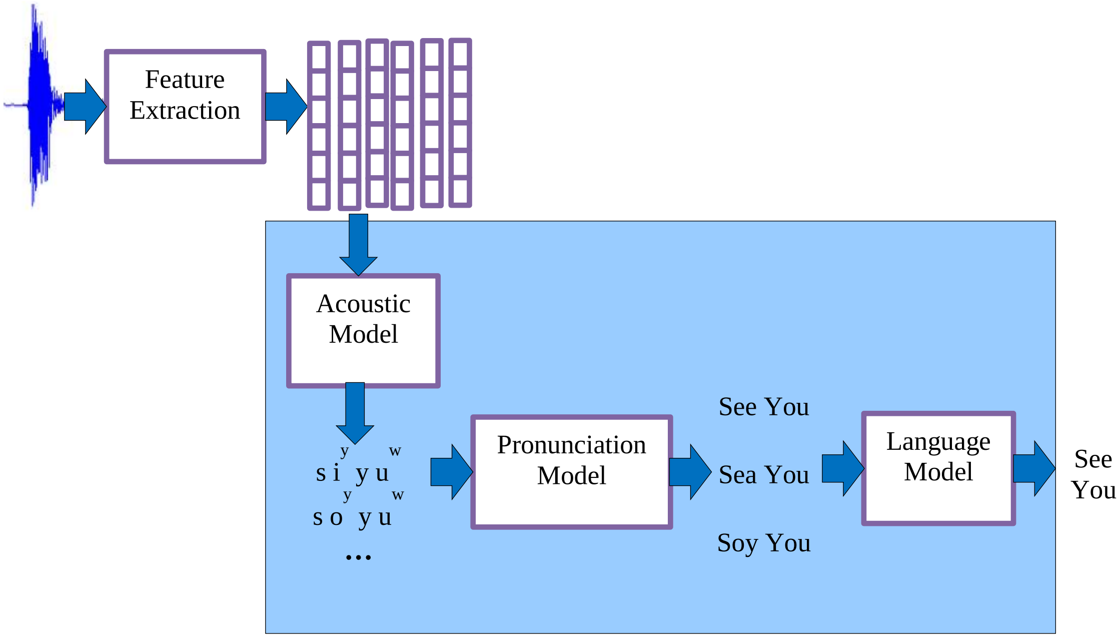


Figure 2.4: The decoding module

Feature extraction is the task of a signal processing front-end of an ASR system that translates the speech waveform into a compact acoustic representation. Acoustic model defines the phones of a language. Conversely, the pronunciation model represents the word that make up from phones. The language model provides the a priori likelihood of a hypothesized word string based on the syntax of the language to be recognized (Tan, 2008).

Some of the essential components of the ASR system will be explained in more details in the following subsections.

2.2.1 Signal Processing Front-End

Feature extraction is the task of the signal processing front-end of an ASR system that converts the speech waveform to some type of parametric representation. The aim of the signal processing front-end is to derive distinguishing features that are discernibly significant (Rosdi, 2008). Firstly, the signal processing front-end digitizes an analog signal into a compact acoustic representation. This process comprises a number of stages: filtering, pre-emphasis, quantization, and sampling. A sampling frequency of 16 kHz has been shown to be adequate for comprehensibly characterizing human speech (Rosdi, 2008; Tan, 2008). However, parameterization is generally applied to speech at the frame level using a sliding window average with a short inter-frame step size of 10 ms. Then, the digitized signal is transformed into feature vectors that are more pertinent for speech processing. The potential types of features (feature vectors) consist of short-time spectral envelopes, zero crossing rates, energy, level-crossing rates, and so on. Frequency-domain features such as short-time spectral envelopes are more descriptive for analyzing speech compared with time-domain features (Tan, 2008; Akbacak, 2009).

The spectral analysis methods include the linear predictive coding (LPC) (Atal et al., 1967), perceptual linear prediction (PLP) (Hermansky, 1990), and mel-frequency cepstrum (MFC) spectral analysis models. Mel-frequency cepstral coefficients (MFCCs) are one of the most extensively employed features in ASR (Davis et al., 1980). MFCCs are related to the real cepstrum of a windowed short-time signal extracted from a series of the fast Fourier transform (FFT) of that signal. Non-linear predictors are proposed to approximate the behavior of the auditory response of the human ear instead of utilizing traditional linearly spaced frequency scale bands (Qin, 2013).

2.2.2 Decoder

In ASR, the decoder is the module that reveals the word sequences embedded in the speech signal or specifically the feature vectors. Finding the most probable word sequence can be accomplished by optimizing the posterior probabilities of the specified feature vectors. Statistical ASR aims to decode a given acoustic observation X to the corresponding word sequence $W' = w_{1,}\ w_{2,}\ . . .,w_m$ that has the maximum expected posterior probability $P(W|X)$,

$$W' = arg\ \max_w \quad P(W|X) \qquad (2.1)$$

Rather than calculating the posterior probability directly, it can be replaced with another form with the use of Bayes theorem:

$$W' = arg\ \max_w \quad P(W)P(X|W)\ /P(X) \qquad (2.2)$$

Since the augmentation in equation (2.2) is performed with the observation X fixed, it is identical to the following maximization:

$$W' = arg\ max_w \quad P(W)P(X|W) \qquad (2.3)$$

Generally, directly calculating the probability $P(X|W)$ of each word in the small vocabulary systems of tracking models of the acoustic signal may be possible, but the calculation becomes more complicated with increasing vocabulary. For this reason, the probability $P(X|W)$ of equation (2.3) is rewritten as:

$$P(X|W) = \sum_U P(X,U|W) \qquad (2.4)$$

where U is sequences of the phonetic units. The probability of the acoustics given a sequence of words can be marginalized in equation (2.4) over all possible sequences of phonetic units. Assuming that the acoustics X are independent of the word sequence W given the phonetic sequence U:

$$P(X|W) = \sum_U P(X|U)P(U|W) \qquad (2.5)$$

Typically, by decomposing the sequences of words *W* into sequences of phonetic units *U*, the search for the most likely word hypothesis *W* given the acoustics observation *X* can be represented as follows:

$$W' = arg\ \max_w \ \ P(X|U)P(U|W)P(W) \qquad (2.6)$$

Equation (2.6) indicates the three models that are used in ASR: P(X|U) refers to the acoustic model, P(U|W) refers to the pronunciation model, and P(W) refers to the language model (Gales et al., 2007; Tan, 2008). The approach that uses these models (acoustic, pronunciation and language) to produce the most likely sequence as the ASR output is called the decoder.

The decoder is the engine of an ASR system that decodes feature vectors into text (Shi, 2008). The decoding process of a statistical ASR system refers to a search process that aims to capture the sequence of words whose corresponding acoustic, pronunciation and language models best match the input speech signal (Mengusoglu, 2004). Taken together, the acoustic representation of the speech signal is compared and matched during the search against the acoustic model which encodes the acoustic realization of the signal speech in the form of likelihoods. Then, these probabilities are integrated with the prior likelihoods for word sequences (the language model), leading to a vast "complex network" of possible word sequences. The term "search space" is used to designate this complex network (Chelba et al., 2011).

Fig. 2.5 illustrates the process of decoding an utterance, using a decoder with a vocabulary of just three words {A, B, C}. The linked small circles denote hidden Markov models (HMMs) that define the phones. The acoustic score *P(X|U)* of the observation is calculated given the HMMs. The dotted circles signify the word pronunciation models *P(U|W)*. The arrows connecting the dotted circles specify the language probability *P*(W), while the arrow linking connected circles symbolizes the shift from one phoneme model to another (Tan, 2008).

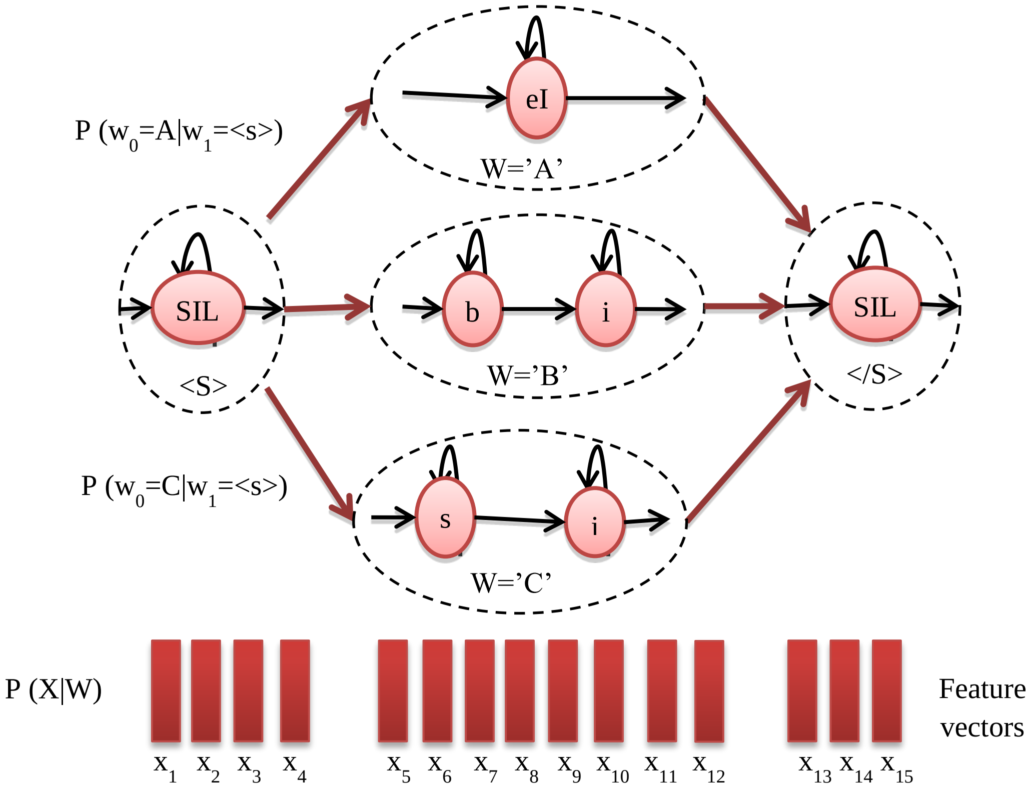


Figure 2.5: Speech recognition as a search problem.

The most plausible word gives the highest score of *P(W)P(U|W)P(X|U)*

2.2.3 Acoustic Model

The decoder determines the *P(X|U)* probability from the acoustic model of equation (2.6). An acoustic model defines the basic speech units, such as phones, words, or syllables. Acoustic models were created from the training set of observed features. This training set is a large set of transcribed speech data. Baum-Welch algorithm is used to train the acoustic models.

There are several potential ways of modeling the acoustic units, such as HMMs, artificial neural networks, template models and so on. An HMM is one of the most extensively utilized approaches in statistical ASR due to its versatility (Mengusoglu, 2004).

The theory of HMMs was developed in the late 1960s. IBM has employed it in ASR since the 1970s. A Markov chain is a stochastic procedure with a short memory in which the current state is based solely on the preceding state. In a Markov chain, the observations are actually the state sequence. An HMM is an extension of the Markov chain where the observation is a function of the state; thus, the state sequence is hidden in an HMM (Ming, 2007; Tan, 2008; Zhu et al., 2012).

The $P(U|W)$ and $P(X|U)$ terms from equation (2.6) can be provided within an HMM framework.

2.2.4 Pronunciation Model

The pronunciation or lexical model, denoted with $P(U|W)$ in equation (2.6), is also called the lexicon. Typically, a pronunciation model represents words/phrase and their permissible pronunciations, which are often defined as a sequence of phonemes (Chang, 2012). The pronunciation model defines the permissible vocabulary. Consequently, an ASR framework can only recognize a limited number of words that are contained in the pronunciation model. Table 2.1 shows some example words and their phoneme sequences(Qin, 2013).

Table 2.1: Examples of lexical mappings from words to phonemes

Word	Pronunciation	Meaning
berselawat	/ b ə r s ə l a w a t /	bless
ketinggalan	/ k ə t i ŋ g a l a n /	miss
spontannya	/ s p o n t a n ŋ a /	spontaneous

2.2.5 Language Model

The language model specifies the probable distribution of words. Statistical language model for an ASR system is the n-gram language model (Bahl et al., 1983). An n-gram model is used to approximating the probability of a sentence *P(W)*, where *W* is a sequence of words $\{W=w_1,w_2,w_3,..\ w_n\}$. By applying a chain rule (Stolcke, 2002), on both *W* and *P(W)*, the probability of a sentence can be calculated as follows:

$$\begin{aligned} P(W) &= P(w_1, w_2, w_3 ... w_n) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)...P(w_n \mid w_1,...w_n) \\ &= \prod_{k=1}^{n} P(w_k \mid w_1, w_2,...w_n) \end{aligned} \tag{2.7}$$

The probability of the n-gram model for the word sequence "*selamat hari raya Aidilfitri*" can be estimated as follows**:**

For a Unigram model:

*P(selamat hari raya Aidilfitri) =P(selamat)*P(hari)*P(raya)*P(Aidilfitri)*