論文 / 著書情報
Article / Book Information

Title	Proper Noun Adaptation for Improving a Spoken Query-based Indonesian Information Retrieval System
Author	Dessi Puji Lestari, Sadaoki Furui
Journal/Book name	International Conference on Rural Information and Communication Technology (r-ICT) 2009, , , pp. 366-371
発行日 / Issue date	2009, 6

Proceedings of International Conference on Rural Information and Communication Technology (r-ICT) 2009, Bandung, Indonesia, pp. 366-371, June 2009

Proper Noun Adaptation for Improving a Spoken Query-based Indonesian Information Retrieval System

Dessi Puji Lestari[1], Sadaoki Furui[2]
Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Insitute of Technology, 2-12-1 Ookayama Meguro-ku Tokyo 152-8552 Japan
Tel. +81-3-5734-3480
[1]dessi@furui.cs.titech.ac.jp, [2]furui@cs.titech.ac.jp

Abstract— **Proper noun recognition is one of the challenging problems in the automatic speech recognition. In many languages, proper nouns pronunciation often does not follow general grapheme-to-phone conversion rules. Our experimental work reported in this paper shows that proper noun recognition in the Indonesian LVCSR (Large Vocabulary Continuous Speech Recognition) system is also more difficult than other regular words such as verbs and adjectives. However, in most of the information retrieval (IR) systems, proper nouns are usually keywords of the queries. Thus, high proper noun recognition error significantly decreases the performance of the IR. In order to increase the proper noun recognition accuracy, we propose a proper noun adaptation method based on the MLLR (Maximum Likelihood Linear Regression) approach. This technique reduces the recognition error rate by 2.38%, and gains 1.96% improvement of the IR MAP (mean average precision) score and 2.43% improvement of the IR MRR (mean reciprocal rank) score comparing to our baseline system.**

Index Terms—**Indonesian Spoken Query, Information Retrieval, LVCSR, MLLR Adaptation.**

I. INTRODUCTION

The widespread access to the Internet has made information resources easily accessible from everywhere. However, there are cases where it is impossible or not convenient to use keyboard as the input device, such as for car drivers, and blind or partially-sighted users. In addition, a very large part of the world population does not have access to computers or Internet, e.g. in the rural area, in which the only available or the most convenient communication mean can be a telephone or a mobile phone. In all of these cases, if we want to let users take advantage of the large amount of information stored in digital repositories, it is necessary to enable voice access to the system as the query input rather than by using the text input. However, sometimes a high error rate of the speech recognition system severely decreases the information retrieval (IR) system's effectiveness. When the spoken terms are incorrectly recognized, a number of relevant documents containing the correct terms cannot be retrieved, while a number of irrelevant documents containing the wrong terms are retrieved. The larger the number of misrecognition terms becomes in the transcribed query, the lower the ranking of the retrieved documents becomes. In most of the information retrieval systems, the proper nouns are usually keywords of the queries. However, the proper noun recognition in the Indonesian LVCSR (Large Vocabulary Continuous Speech Recognition) system is more difficult comparing to other regular words. Thus, the high proper noun recognition error significantly drops the IR performance.

Several different automated methods, such as Boltzmann machines [1,2] and Decision Trees [3] have been proposed to reduce the proper noun error rate, however they have not yet achieved an acceptable error rate. In this paper we propose a proper noun adaptation method based on the MLLR (Maximum Likelihood Linear Regression) approach to reduce the proper noun error of Indonesian spoken queries.

II. INDONESIAN LANGUAGE

The Indonesian national language called Bahasa Indonesia is a variant of Malay language and categorized as the Austronesian or the Malayo-Polynesian language. Indonesian language is written using the Latin alphabet consisting of 26 characters from A to Z. The space symbol is used to separate words and some punctuation symbols e.g. ".", ",", "!", and "?", are used to separate sentences as in English. The correspondence between sounds and their written forms is generally regular. However there are some exceptions in proper nouns especially for old written style proper nouns or proper nouns that came from regional languages. The Indonesian standard phoneme set as described by Darjowidjojo [4] can be seen in Appendix 1. Indonesian language has borrowed many words from many languages, including Sanskrit, Arabic, Persian, Portuguese, Dutch, Chinese and many other languages, including other Austronesian languages. The basic word order in Indonesian sentences is Subject-Verb-Object. The adjective, demonstrative pronoun and possessive pronoun are written to follow the modified noun.

III. Large Vocabulary Continuous Speech Recognition (LVCSR)

Automatic speech recognition is a technology that allows computers equipped with a device for sound input such as a microphone to transform human speech into a sequence of words. The system consists of three main components: an acoustic model, a language model and a decoder. The acoustic model represents how a given word or a phoneme ("phone") is pronounced. The language model predicts likelihood of a given word sequence appearing in the language. When the size of vocabulary that can be recognized is large (more than thousands of words), and they are spoken continuously, the task is referred to as large vocabulary continuous speech recognition or LVCSR.

The decoder estimates a word sequence $\hat{W}$ that generated a given acoustic observation sequence O. The speech recognition process can be formulated as a process of maximizing a posteriori probability of W given O, $P(W|O)$, as follows:

$$P(\hat{W} \mid O) = \max_{W} P(W \mid O) \tag{1}$$

By Bayes rule, the a posteriori probability can be converted into:

$$P(W \mid O) = \frac{P(O \mid W) P(W)}{P(O)} \tag{2}$$

Since $P(O)$ is independent of W, the recognition process becomes to:

$$\hat{W} = \arg\max_{W} P(O \mid W) P(W) \tag{3}$$

where $P(O|W)$ is a probability of producing O given W. It represents the probability of an acoustic observation sequence conditioned on the given word sequence. This probability is usually represented by Hidden Markov Models (HMMs). $P(W)$ is a language model which represents a probability of a word sequence observed in the language. Word n-gram models are generally used to estimate this probability.

A. Hidden Markov Models (HMMs)

In a HMM-based phone model, each phone is represented by an HMM. Each HMM phone model usually has a left-to-right topology and has five states. These states are an entry state, three emitting states, and an exit state. To produce a word, phone HMMs are combined together by joining the exit state of an HMM with the entry state of another HMM.

A basic principle of a Markov model is that it works as a finite state machine which changes the state once every time unit t. At each time t, a state j is entered and a speech vector O_t is generated based on a probability density $b_j(O_t)$. The transition from state i to state j is also probabilistic with a probability a_{ij}. In the generation process, only the observation sequence is known, and the underlying state sequence is hidden. This is why it is called a Hidden Markov Model. By summing the product of transition probabilities a_{ij} and output probabilities $b_j(O_t)$, a joint probability of a speech vector O_t and a state sequence X given a model M can be computed, and by considering only the most likely state sequence, the likelihood can be approximated as follows:

$$\hat{P}(O \mid M) = \max_{x} \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(O_t) a_{x(t)x(t-1)} \right\} \tag{4}$$

where $x(0)$ is the model entry state and $x(T+1)$ is the model exit state. Equation (4) assumes that the transition probabilities a_{ij} and the output probabilities $b_j(O_t)$ are trained for each model M_i. These parameters can be calculated by conducting a robust and efficient re-estimation procedure using a set of training examples corresponding to a particular model. Thus, to build an HMM-based model, a training set of acoustic utterances with its transcription is needed. Usually, the larger the available training data is, the higher the recognition accuracy.

To determine the parameters of each model M_i, it is necessary to make an initial model of what they might be until more accurate parameters in the sense of maximum-likelihood can be found. This can be achieved by applying Baum-Welch algorithm. This algorithm is based on an expectation maximization (EM) algorithm. It can compute maximum likelihood estimates and posterior model estimates for the parameters (transition and emission probabilities) of an HMM, even when only emissions are given as training data. Current LVCSR systems use mixtures of Gaussian distribution to represent each state in the HMM. Parameters of each Gaussian distribution are estimated from training data.

B. N-gram Models

In LVCSR systems, n-gram language models are used to provide the recognizer with an a priori likelihood $P(W)$ of a given word sequence W. It is usually derived from a large training text that shares the same language characteristics as expected input. N-gram language models rely on the likelihood of sequences of words, such as word pairs (in the case of bi-grams) or word triplets (in the case of tri-grams). If we assume that W is a sequence of words and q is the number of words in W, i.e., $W = (w_1, w_2, ..., w_q)$. The language model $P(W)$ can be generated as follows:

$$P(W) = P(w_1, w_2, ..., w_q) = \prod_{i=1}^{q} P(w_i \mid w_{i-n+1}, ..., w_{i-1}) \tag{5}$$

where n is the order of the Markov process. In particular, models with $n = 2$ and $n = 3$ are widely used and are called bigrams and trigrams, respectively. In order to estimate the probability of $P(w_i \mid w_{i-2}, w_{i-1})$ in the trigram case, simple counts of each word-triplet in the training corpus are used as follows:

$$P(w_i \mid w_{i-2}, w_{i-1}) = \frac{N(w_{i-2}, w_{i-1}, w_i)}{N(w_{i-2}, w_{i-1})} \quad (6)$$

where $N(a, b)$ denotes the number of times one observes a and b continuously in the training data.

N-gram models have been popular because they yield simple and highly reasonable models. However, a problem arises when a word-pair or a triplet is not observed in the training text. It will give a 0 probability for the unseen word pair or triplet and, hence will make the probability of the entire sentence W to be 0. Due to data sparseness, this happens very often in practical situations, and it causes speech recognition errors. Another problem of n-gram is that an unreliable probability is assigned to infrequent words such as infrequent proper nouns. To avoid this problem, several smoothing techniques are usually applied. The fundamental idea of smoothing techniques is to subtract some small probability mass from the relative frequency estimates by Equation (6) for the probabilities of infrequent n-grams, and to redistribute this probability to unseen n-grams. There are several proposed smoothing techniques. These methods differ according to how much is subtracted out, called discounting, and how it is redistributed, called back-off. Some of the smoothing techniques that worked well in ASR systems include Linear interpolation proposed by Jelinek [5], Good-Turing discounting [6], Witten-Bell discounting [7], and Katz back-off [8].

C. MLLR Adaptation

The maximum likelihood linear regression (MLLR) method computes a set of linear transformations that reduce the mismatch between an initial model set and adaptation data. This technique estimates transformations for mean and covariance of Gaussian mixture HMMs so that they maximize the likelihood of the adaptation data. For speech recognition, this method was originally proposed by Leggetter et al. [9] and has been widely used.

IV. Spoken Query Information Retrieval

Spoken query information retrieval refers to the IR which uses spoken queries to retrieve textual or spoken documents. The spoken query processing has a number of challenges compared with the text query processing. The most important ones are:

- Misrecognition of spoken query terms produced by the ASR. This may cause the terms to disappear from the query representation and also to be replaced by different terms. Thus, a large set of potentially relevant documents may not be retrieved.
- The additional time required to pre-process the query. A spoken query needs to be recognized and a transcript needs to be produced at the time the query is submitted. It has been observed that user satisfaction with an IR system also depends upon the time the user needs to wait for the system to display the results [10]. This includes the time to process the query.
- Spoken style queries tend to be longer than text queries and they frequently contain less important words, such as function words. Moreover they may also increase the searching confusion. Thus, they require some additional processing to remove words in the query that do not give important information for search.

Some studies show that the use of classical IR techniques for spoken query is quite robust to considerably high levels of WER (up to about 47%) [11]. However, there still exists the room for improvement. The tf-idf (term frequency – inverse document frequency) weighting method [12] is often used in IR. It is a statistical technique to evaluate how important a term is in a document. The importance increases proportionally to the number of times a word appears in the document but is offset by how common the word is in the document collection. There are many variations of the tf-idf formula depending on the way the tf and idf weights are computed [13]. In this experiment we used the most standard tf-idf formula as shown below:

$$tf(i, j) = \frac{n_{i,j}}{length_j} \quad (7)$$

$$idf(t_i) = \log \frac{N}{n_i} \quad (8)$$

where $n_{i,j}$ is the number of occurrences of a term t_i in a document d_j, $length_j$ is the number of words in the document d_j, N is the total number of documents in the collection, and n_i is the number of documents in which the term t_i occurs in the document collection.

The retrieval status value (RSV) is evaluated by applying the dot product to the document and query representations obtained using the tf-idf weighting schema. The score for each document is calculated by summing the tf-tdf weights of all query terms found in the document as shown below:

$$RSV(d_j, q) = \sum_{t_i \in q} idf(t_i) \cdot tf(i, j) \quad (9)$$

Some IR systems enable to perform Relevance Feedback (RF). The RF enables the system to reformulate the query after gaining some feedbacks. There are two kinds of the RF regarding the way it gains the feedback. The first one is the standard RF which enables a user to interactively express his information requirement by modifying his original query by explicitly confirm the relevance of some documents retrieved by the system. The second one is the Pseudo RF. In this method the system assumes that its top-ranked documents are relevant, and uses these documents in the RF algorithm. If RF performs well the final search should contain more relevant documents than the initial search.

The main IR evaluation measures are Recall and Precision. Recall is defined as the portion of all the relevant documents in the collection that has been retrieved. Precision is the portion of retrieved documents that is relevant to the query. To give more accurate measure, Average Precision is commonly used, defined as the average of 11 Precision values of the preset different levels of Recall. MAP is a mean of Average Precision scores for a group of queries. Another common IR evaluation measures is the mean reciprocal rank (MRR). The reciprocal rank is defined as the inverse of the rank of the first correct answer. The MRR is the average of the reciprocal ranks of results for a sample of queries Q as shown below:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i} \quad (10)$$

where $|Q|$ is the number of queries and $rank_i$ is the ranking of the first relevant document in the search list.

V. EXPERIMENTS

A. Baseline System

For the ASR, we employ the Bahasa Indonesia LVCSR system that we built previously [14]. For the acoustic model, context-dependent HMMs were trained using 32 Gaussian mixtures per state from a 14.5 hours Indonesian phonetically balanced-speech corpus recorded in our laboratory. For the language model, the bigrams and trigrams were trained using the ILPS corpus [15]. The articles in the corpus were taken from the two popular Indonesian newspaper [1] and magazine [2] sites. Both bigrams and trigrams were smoothed using the Good-Turing back off technique.

[1] http://www.kompas.com

[2] http://www.tempointeraktif.com

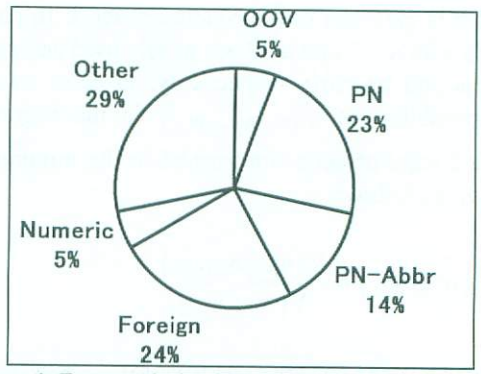

Figure 1. Error analysis of the transcribed queries for the baseline system (PN: proper nouns)

For building the dictionary, words that occur in the ILPS corpus for more than 3 times were selected. There were 2.65K words in the dictionary.

The spoken query was first transcribed using the Bahasa Indonesia LVCSR. We used Julius version 3.4[3] as the speech decoder. After removing the stop words in Bahasa Indonesia [16], the transcribed query with a speech recognition confidence score for each term in the query was fed into the IR system. We used the Lemur toolkit [4] provided by Carnegie Mellon University and the University of Massachusetts, Amherst, to build the Indonesian IR system.

B. Proper Noun Adaptation

From the baseline evaluation, we found that a majority of the misrecognized words was caused by proper nouns (23% error was regular proper nouns, and 14% error was abbreviated proper nouns) as shown in Figure 1. Assuming that the difficulties of recognizing proper nouns came from the acoustic variation, we tried to solve the problem by enhancing the acoustic model. To model the acoustic variations in uttering proper nouns by Indonesian speakers, an adaptation technique was used to create proper-noun specific acoustic models. We conducted supervised adaptation based on the MLLR technique using 8 regression classes. The adaptation data was proper noun utterances (14,840 words) extracted from the Indonesian speech corpus that was used to train the baseline acoustic model.

C. Evaluation

Since there is no standard evaluation corpus for spoken query IR in Bahasa Indonesia, we recorded spoken queries from 20 native Indonesian speakers (11 males, 9 females), each uttering 35 queries with different topics. The queries were derived from the Bahasa Indonesia IR collection developed by the ILPS [15]. There are 35 query topics available for the magazine corpus and the newspaper corpus in the ILPS corpus. In the experiment

[3] http://julius.sourceforge.jp/index.php

[4] http://www.lemurproject.org

in this paper, we only used the corpus taken from the magazine. For each of the 35 topics of the query, we developed three kinds of spoken queries in terms of the length: short query (2-4 words), medium-length query (4-8 words), and long query (8-16 words). There are 2100 Indonesian spoken queries in total. The document collection was taken from the portion of the Indonesian text corpus provided by ILPS that was not used in training the language model of Bahasa Indonesia LVCSR. The trigram language model had a test-set perplexity of 61.04 and an OOV (out of vocabulary) rate of 1.75% for this test set.

The average ASR accuracy of the baseline system was 80.66% and by applying the proper noun adaptation described above, the average accuracy increased 2.38% and became to 83.04%. The accuracy of each speaker can be seen in Figure 2.

Both transcribed queries from the baseline ASR and that using the proper-noun adapted acoustic models were input to the IR system. The text queries were also given into the IR to compare the results with that obtained using ASR. The average MRR score and MAP score for spoken queries using the ASR baseline system and the proper-noun adapted system, and using the text queries can be seen in Table I.

In addition, we applied a query expansion technique using the Pseudo-Relevance Feedback by choosing the top 5 documents in the search list and adding 5 most frequent words in those documents to expand each query. By doing this, we gained 2.7% improvement of the IR MAP score and 1.6% improvement of the IR MRR score.

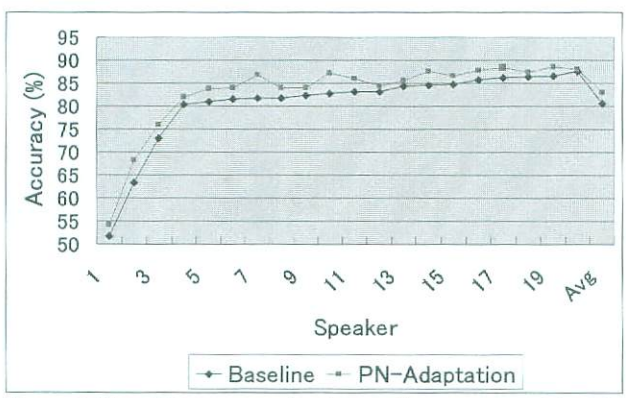

Figure 2. The ASR accuracy for the baseline system and the PN-adapted system for 20 speakers (sorted)

Table I
MRR and MAP score for spoken queries using baseline ASR, spoken queries using PN-adapted ASR, and the text queries

	MRR	MAP
ASR Baseline	0.6983	0.5127
PN-Adapted ASR	0.7226	0.5323
Text Query	0.8242	0.608

VI. CONCLUSION

In order to increase the proper noun recognition rate in the Indonesian LVCSR to increase the IR performance, we proposed a proper noun adaptation method based on the MLLR approach. This technique could reduce 2.38% of the recognition error rate of the spoken query, and gained 1.96% improvement of the IR MAP (mean average precision) score and 2.43% improvement of the IR MRR (mean reciprocal rank) score comparing to our baseline system.

As can be seen in Figure 1, there are several other sources of errors in the transcribed queries, such as foreign words (e.g. English words), numeric words, OOVs, and others. Since all foreign words that appear in the queries were misrecognized in our experiments, we are planning to improve the foreign word recognition performance as our future work.

APPENDIX

Phonetic Category	**Phoneme**	**Example**	
		Word	**Phoneme Sequence**
Vowels			
	/a/	saya	/s a y a/
	/e/	enak	/e n a k/
	/E/	kEmana	/k E m a n a/
	/i/	ingin	/i n g i n/
	/o/	orang	/o r a ng/
	/u/	untuk	/u n t u k/
Diphtongs			
	/ai/	sungai	/s u ng ai/
	/au/	danau	/d a n au/
	/oi/	amboi	/a m b oi/
Semi-vowels			
	/w/	wanita	/w a n i t a/
	/y/	saya	/s a y a/
Consonants			
Plosives	/b/	berapa	/b e r a p a/
	/p/	petani	/p e t a n i/
	/d/	dia	/d i a/
	/t/	teman	/t e m a n/
	/g/	giat	/g i a t/
	/k/	kamu	/k a m u/
	/kh/	khairul	/kh a i r u l/
Africates	/j/	juga	/j u g a/
	/c/	cinta	/c i n t a/
Fricatives	/v/	video	/v i d e o/
	/f/	maaf	/m a a f/
	/z/	jenazah	/j e n a z a h/
	/s/	saya	/s a y a/
	/sy/	syahdu	/sy a h d u/
	/h/	hujan	/h u j a n/
Liquids	/r/	ramai	/r a m a i/
	/l/	lambat	/l a m b a t/
Nasals	/m/	mana	/m a n a/
	/n/	mana	/m a n a/
	/ny/	nyanyian	/ny a ny i an/
	/ng/	lambang	/l a m b a ng/

Appendix 1. Indonesian phoneme set

VII. ACKNOWLEDGMENT

The authors would like to thank F. Z. Tala for providing a Bahasa Indonesia text corpus for information retrieval.

REFERENCES

[1] D. Neeraj, A. Le, J. Ngan, J. Hamaker and J.Picone, 1997. "An advanced system to generate multiple pronunciations of proper nouns," in *Proc. IEEE ICASSP*, pp. 1467-70, Munich, Germany, April I997.

[2] D, Neeraj, M. Weber, and J.Picone, "Automated generation of N-best pronunciations of proper nouns," in *Proc. IEEE ICASSP*, pp. 283-6, Atlanta, Georgia, May 1996.

[3] J. Ngan, A. Ganapathiraju, and J. Picone, "*Improved surname pronunciations using decision trees*," in Proc. ICSLP, pp. 3285-8, Sydney, Australia, November 1998.

[4] S. Darjowidjojo, "Indonesian syntax," *Ph.D dissertation*, Georgetown, University, Washington, 1966.

[5] F. Jelinek, Statistical Methods for Speech Recognition. MIT Press, Cambridge, Massachusetts, 1998

[6] I.J. Good, "The population frequencies of species and the estimation of population parameters," in *Biometrika, no. 40*, pp. 16-264, 1953.

[7] I.H.Witten, and T.C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," in *IEEE Transaction Information Theory, vol. 37, no. 4*, pp. 1085-1094, 1991.

[8] S.M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," in *IEEE Trans. ASSP, vol. 35, no. 3*, pp. 400-401, 1987.

[9] C.J, Leggetter, and P.C, Woodland, " Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," in *Computer Speech and Language*, vol. 9, no. 2, pp. 171-185, April 1995.

[10] C. Cleverdon, J. Mills, and M. Keen, "ASLIB Cranfield Research Project: factors determining the performance of indexing systems," *ASLIB*, 1966.

[11] F. Crestani, "Spoken query processing for interactive information retrieval," in Data Knowl. Engineering. 41(1), pp. 105-124, 2002.

[12] G. Salton, and C. Buckley, "Term weighting approaches in automatic text retrieval," in *Technical report, Ithaea*, NY, USA, 1987.

[13] D. Harman, Ranking algorithms, in: W. Frakes, R. Baeza-Yates (Eds.), Information Retrieval: data structures and algorithms, Prentice Hall, Chapter 14, Englewood Cliffs, New Jersey, USA, 1992.

[14] D. P. Lestari, K. Iwano, and S. Furui, "A large vocabulary continuous speech recognition system for Indonesian language," in Proc. *15th Indonesian Scientific Conference in Japan (ISA-Japan)*, pp.17-22, Hiroshima, Japan, 2006.

[15] F .Z Tala, J. Kamps, K. Muller, and M. de Rijke, 'The Impact of Stemming on Information Retrieval in Bahasa Indonesia," *14th Meeting of Computational Linguistics in the Netherlands (CLIN-2003)*, Netherland, 2003.

[16] F .Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis*, Appendix D, pp. 39-46, University of Amsterdam, 2003.