

# Research on spoken query-based Indonesian information retrieval

○Dessi Puji Lestari, Sadaoki Furui (Tokyo Institute of Technology)

## 1 Introduction

Spoken query processing plays an important role in the interactive information retrieval system which enables users to input the query by speech rather than by typing. One of the important issues in the spoken query processing system is the term misrecognition caused by the ASR that reduces the effectiveness of the IR. Our experiments show that proper nouns and foreign words are the type of words in Indonesian which are most frequently misrecognized. In this paper we propose a proper noun adaptation of acoustic models to improve their recognition accuracy. To increase the foreign word (English word) recognition accuracy, rule-based phone mapping from English to Indonesian is proposed.

In the speech recognition system, a confidence score of each recognized word can be used as a measure of how certain the system is about the recognized word [1]. To raise the ranking of the correctly retrieved documents, we use a speech recognition confidence score based on posterior probabilities of the words to weight the term in the query, processed by the inference network (IN-based) retrieval model. The IN-based information retrieval has an advantage in that it allows structured query operators and explicit term weighting.

## 2 Indonesian Language

The Indonesian national language called Bahasa Indonesia is a variant of Malay language and categorized as an Austronesian or a 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 basic word order in Indonesian sentences is Subject-Verb-Object. Adjectives, demonstrative

pronouns and possessive pronouns are written to follow the modified nouns.

The Indonesian phoneme set contains 6 vowels, 4 diphthongs and 22 consonants. 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. Indonesian language has borrowed many words from many languages, such as Sanskrit, Arabic, Portuguese, Dutch, and many other languages, including other Austronesian languages. The Indonesian government defined some rules about how to transform foreign words into Indonesian words. However, there are many cases where Indonesian people tend to use original foreign words especially English words in both formal and informal occasions, even when the suitable translated Indonesian words exist for those foreign words. This is mainly because the suitable Indonesian words are still not familiar for Indonesian people.

## 3 Term Weighting

The tf-idf (term frequency – inverse document frequency) weighting method [2] 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. Although there are many variations of the tf-idf formulae, in this experiment we use the most standard tf-idf formula as shown below:

$$tf(i, j) = \frac{n_{i,j}}{length_j} \quad (1)$$

$$idf(t_i) = \log \frac{N}{n_i} \quad (2)$$

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-idf weights of all query terms found in the document.

#### 4 Inference Network Model

The Inference Network (IN) model is basically a directed acyclic graph (DAG) of Bayesian Network. The network is used to model documents and document contents (the document sub-network), and also to model queries (the query sub-network) [3].

The document sub-network consists of three layers of nodes: the document nodes; the text representation nodes; and the representation nodes. A causal link represented as a down arrow between nodes indicates that the parent nodes are represented by the children node. Each link contains a conditional probability, or weight. The evaluation of a node is performed using the value of the parent nodes and the conditional probabilities. This specification is basically a indexing weight such as the tf-idf term weighting as previously described.

The query sub-network consists of three layers of nodes: the query concept nodes; the query nodes; and the user information need node. Each query node contains a specification in the form of the link matrices to describe the dependence of the query on its parent query concepts.

In the retrieval process, to form the complete IN, the query sub-network is attached to the document sub-network when the concepts in both networks are the same. The complete IN is evaluated for each document node to form the probability of the relevance to the query. For all non-root nodes in the IN, the probability of each node is calculated using its parent values. It usually uses a link matrix to provide diagnostic

information to the set. This link matrix can be used to implement a variety of weighting schemes, including the tf-idf weighting schema.

#### 5 English to Indonesian Phone Mapping

One way to recognize foreign words is mapping the phoneme symbol of the foreign words that exist in the lexicon to the phoneme symbol of the target language. There are several ways to map the phoneme symbols across languages: a knowledge-based or a data-driven approach. Since the data driven approach requires many training data that is difficult to be collected for resourced-lacked languages such as Bahasa Indonesia, the most intuitive and straightforward approach is by using the linguistic knowledge-based phonetic mappings [4]. This approach has been used by a previous researcher working on rapid development of Indonesian speech recognition using a cross-language approach [5]. In our experiment, we use some English to Indonesian phoneme mapping rules found in [5] and add/delete some rules as shown in Table 1.

Table 1. English to Indonesian Phonemes Mapping

Eng	Ind	Eng	Ind	Eng	Ind
<b>aa</b>	a	hh	h	ow,ao	o
<b>ay</b>	ay	Ih,iy,ix	i	oy	oy
<b>aw</b>	aw	jh	j	p	p
<b>b</b>	b	k	k	r	r
<b>ch</b>	c	k+h	kh	s	s
<b>d,dx,dh</b>	d	l	l	sh	sy
<b>eh,ae,ah,ax</b>	e	m	m	t,th	t
<b>ey</b>	ey	n	n	uh,uw	u
<b>f,v</b>	f	ng	ng	w	w
<b>g</b>	g	n+y	ny	y	y
				z,zh	z

To recognize English words in the lexicon, we use the CMU lexicon as the reference and then re-filter the resulted English word list by using the most standard Indonesian dictionary managed by Indonesian government, the Kamus Besar Bahasa Indonesia (KBBI), as the reference. Using that dictionary, some words which are recognized as English words but also exist in the KBBI are

deleted from the list to avoid ambiguity in the recognition. Some English words which have the same pronunciation as Indonesian words are also deleted from the list to avoid redundancy.

## 6 Experiments

### 6.1 Baseline System

For the ASR, we employ the Bahasa Indonesia LVCSR system that we built previously [6]. A spoken query is first transcribed using the Bahasa Indonesia LVCSR. After removing the stop words in Bahasa Indonesia [7], the transcribed query is fed into the IR system. The correct query text is also given to the IR to compare the results with that obtained using ASR. We use the standard tf-idf weighting schema as described in Section 3. The MRR (mean reciprocal rank) scores for the baseline system are shown in Table 3.

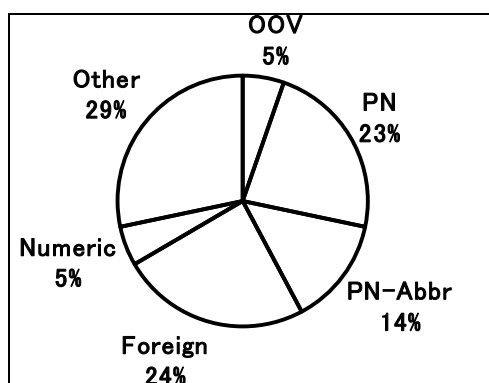


Figure 1. Error analysis of the transcribed queries for the baseline system (PN: proper nouns)

### 6.2 Proper Noun Adaptation

From the baseline evaluation, we found that a majority of misrecognized words was caused by proper nouns as shown in Figure 1. To model the acoustic variations in uttering proper nouns by Indonesian speakers, an adaptation technique was employed to create proper-noun specific acoustic models. We conducted supervised adaptation based on the MLLR technique using 8 regression classes. The adaptation data consisted of 14,840 proper nouns extracted from the Indonesian speech corpus that was used to train the baseline acoustic model.

### 6.3 English Words Correction

To correct the pronunciation of English words in the Indonesian lexicon, the pronunciation of English words was mapped using the rule described in Table 1. There are 4050 words in the lexicon recognized as English words by using the CMU lexicon as the reference. After filtering the resulted English word list using the KBBI, and removing the words with the same pronunciation as Indonesian, the number of English words became 1939 words.

### 6.4 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 [8]. 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 IR 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%. The averaged accuracy for each query-length for the baseline system, the PN-adapted system, and the English words corrected lexicon is shown in Table 2.

Table 2. The ASR accuracy for the baseline system (Base), the PN-adapted system (PNA), and the English words corrected lexicon (ECL) for each type of query.

Query Type	Base	PNA	ECL
Short	79.36	82.05	84.51
Medium	81.80	84.23	84.56
Long	79.07	83.69	83.84
Average	80.08	83.32	84.30

Table 3. MRR scores for spoken queries using baseline ASR (Base), that using PN-adapted ASR (PNA), that for the English words corrected lexicon (ECL), and the text queries. Results are compared for the standard tf-idf vector space model (VSM) and the IN-based tf-idf (IN)

	VSM	IN
Base	69.83	80.10
PNA	72.26	82.20
ECL	72.70	82.65
Text Query	82.42	–

### 6.5 IN-based IR

The transcribed queries from the baseline ASR, that using the proper-noun adapted acoustic models, and that using the English words corrected lexicon with their speech recognition confidence scores were input to the IR system. The average MRR scores for all transcribed queries both using the standard tf-idf and the IN-based IR are shown in Table 3.

## 7 Conclusions and Ongoing Work

This paper presents our research on a spoken query-based Indonesian information retrieval. In order to increase the proper noun recognition rate in the Indonesian LVCSR, we proposed a proper noun adaptation method based on the MLLR approach. This technique could reduce 3.24% of the recognition error rate of the spoken query. To increase the English word recognition rate, the rule-based English to Indonesian phoneme mapping was applied to the English words in the lexicon. This technique could reduce 0.98% of the recognition error rate of the spoken query.

We also compared the standard tf-idf using the vector space model and the inference network-based IR with Indonesian spoken queries. The term weighting strategy using the speech recognition confidence score showed its great potential to improve document ranking in the inference network-based IR. It works well especially for the spoken query with low accuracy of recognition. However, further work using confidence score weighting needs to be

conducted to improve the performance of the spoken query with high correctness of recognition.

### Acknowledgements

The authors would like to thank ILPS, University of Amsterdam for giving us the Kompas and Majalah Tempo collections.

### References

- [1] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in Proc. EUROSPEECH, vol. 2, pp. 827-830, Greece, September 1997.
- [2] G. Salton, and C. Buckley, "Term weighting approaches in automatic text retrieval," in Technical report, Ithaea, NY, USA, 1987.
- [3] H. R. Turtle and W. B. Croft, "Inference networks for document retrieval," in ACM Transactions on Information Systems, vol. 9, no 3, pp. 187-222, July 1991.
- [4] C. Nieuwondt and E.C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Speech Communication*, vol. 38, pp. 101-113, 2002.
- [5] S. Sakti, K. Markov, S. Nakamura, "Rapid Development of Initial Indonesian Phoneme-based Speech Recognition Using The Cross-Language Approach," in Proc. Oriental COCOSDA, pp. 38-43, Jakarta, Indonesia, 2005.
- [6] 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.
- [7] 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.
- [8] F.Z Tala, J. Kamps, K. Muller, and M. de Rijke, "The Impact of Stemming on Information Retrieval in Bahasa Indonesia," In CLIN, Netherland, 2003.