論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Automatic abbreviation detection using syllable composition rules for Indonesian spoken query-based information retrieval
著者(和文)	DESSI PUJI LESTARI, 古井 貞熙
Authors(English)	Dessi Puji Lestari, Sadaoki Furui
出典(和文)	日本音響学会2011年春季講演論文集, , No. 3-5-13, pp. 105-108
Citation(English)	, , No. 3-5-13, pp. 105-108
発行日 / Pub. date	2011, 3

Automatic abbreviation detection using syllable composition rules for Indonesian spoken query-based information retrieval*

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

1 Introduction

This paper presents our research on a spoken query-based Indonesian information retrieval (IR) system. The system has two main components, an automatic speech recognizer (ASR) which transcribes spoken queries into text queries and an IR which searches for relevant documents stored in digital repositories according to the user's information need. In the spoken query-based IR, speech recognition errors often decrease the IR performance, in particular when the misrecognized terms are keywords in the query. In our previous work, we found that ASR made significant errors in recognizing abbreviations. Abbreviations are shortened form of words or phrases, e.g. *W3C* (W three C) for World Wide Web Consortium or *e-mail* for electronic mail. One of the main problems in recognizing abbreviations is grapheme-to-phoneme conversion errors for abbreviations in the lexicon. In order to give correct pronunciation, it is essential to correctly detect abbreviations.

This paper proposes using syllable-based word composition rules for abbreviation detection. We assume that if syllables in a word do not follow standard Indonesian word composition rules, the word is difficult to pronounce, and hence it is pronounced by the letters. Our experiments show that this technique improves the recognition accuracy of abbreviations in the spoken query, and it improves the IR performance. This paper also proposes using word occurrence weight in the n-best transcribed query to explicitly weight the term importance modeled by the inference network (IN-based). Our result shows that these techniques improve mean reciprocal rank (MRR) score of the retrieved documents.

2 Syllable composition rules in Indonesian language

Indonesian government defined standard rules to compose words by syllable units. To segment each word by syllables, rules describes in [1] is used in our experiments as follows:

1. If there are two consecutive vowels (except diphthong) in the middle of the word, syllable separation is conducted between the two vowels, e.g. in a word *ta-at* which means obedient in English.
2. If there are two consonants (except digraph) with one vowel in between, syllable separation is conducted before the second consonant, e.g. in word *ka-mu* which means you in English.
3. If there are two consonants or more in the middle of the word, syllable separation is conducted after the first consonant, e.g. in word *struk-tur* which means structure in English.
4. If there are two vowels with one consonant in between, syllable separation is conducted after the first vowel, e.g. in word *a-nak* which means a child in English.
5. Affix and particle are separated from their basic word.

Table 1 describes the standard Indonesian syllables.

3 Proposed method

In our previous work, we found that an abbreviation list defined by Indonesian government and organizations is not adequate for detecting abbreviations. As the number of abbreviations in the Indonesian language is growing rapidly through a variety of social, business, science, and technology influence, and there is no rule for the formation of new abbreviations, their correct detection and transcription by ASR is increasingly getting difficult.

* インドネシア語の音声クエリによる情報検索のための音節構成規則による略語の自動検出法、レスタリ・デッシ　プジ、古井 貞熙（東工大）

Table 1: Indonesian standard syllables (V: vowel; C: consonant).

Syllable	Example
V	a-nak, ba-u
VC	an-da, da-un
CV	se-bab, man-di
CVC	lan-tai, ma-kan
CCV	pra-ha-ra, sas-tra
CCVC	frik-si, kon-trak
VCC	eks, ons
CVCC	pers, kon-teks
CCVCC	kom-pleks
CCCV	in-stru-men
CCCVC	struk-tur, in-struk-si

In this paper, we propose an automatic abbreviation detection method using syllable-based word structure rules to improve the recognition accuracy of abbreviations. We handle only abbreviations pronounced by the letters, e.g. BBC for British Broadcasting Corporation, and USA for United States of America. Abbreviations having the same pronunciation as words such as the opening letter initialization, e.g. *Radar* for radio detection and ranging, are excluded from our focus. The procedure to add a correct pronunciation to each abbreviation in the lexicon is as follows:

1. Detect abbreviations in the lexicon using an Indonesian abbreviation list provided by one of the Indonesian newspapers.
2. After excluding abbreviations detected in the previous step from the lexicon, other abbreviations are detected as follows:
   - Segment each word into syllables as described in Section 2.
   - Check syllables of each word by Indonesian standard rules (Table 1). If one or more syllables in the word do not follow the rules, the word is categorized as an abbreviation.
3. Letter pronunciations are then added to each detected abbreviation.

4 ASR Experiments

4.1 Baseline

Hidden Markov Model (HMM)-based acoustic models and n-gram language models are trained to develop the LVCSR system for the Indonesian language [2]. The first through 12th order Mel-Frequency Cepstral Coefficients (MFCC) are extracted every 10 ms by using a 25-ms-wide window. Delta features of MFCC coefficients and energy are also incorporated. 32-Gaussian mixture is used for each state to train context-dependent HMMs. The total number of states is 1,746, and the number of context-dependant models is 6,088. Both bigrams and trigrams are smoothed using the Good-Turing back-off technique. The 3-gram language model has test-set perplexity of 61.0 and an OOV rate of 1.75% for the spoken queries described in the next subsection.

4.2 Indonesian spoken queries

We have developed a test set of spoken queries for our experiments. The queries were derived from the Bahasa Indonesia IR collection developed by the ILPS [3] and from the Bahasa Indonesia IR collection developed by the School of Computer Science and Information Technology, RMIT University, Australia [4]. There are 35 query topics available for the magazine corpus (Tempo-Tala corpus), 35 query topics available for the newspaper corpus 1 (Kompas-Tala corpus), and 20 query topics available for the newspaper corpus 2 (Kompas-Jelita corpus). In total, there are 90 query topics. For each topic of the query, we have developed three kinds of spoken queries in terms of length: short queries (2-4 words), medium-length queries (4-8 words), and long queries (8-16 words). We have recorded these queries spoken by 20 native Indonesian speakers (11 males, 9 females), each uttering 90 queries on different topics. These speakers are different from those used for training the acoustic model. There are 5400 Indonesian spoken queries in total.

4.3 Experimental results

By employing the methods described in Section 3, recognition accuracy was increased by 1.1% on average compared with the baseline ASR as shown in Table 2. Figure 1 shows ASR accuracy by using three kinds of testing corpora: the Tempo-Tala; Kompas-Tala; and Kompas-Jelita corpora.

Table 2: ASR accuracy (%) by the baseline and abbreviation corrected lexicon (ACL) systems

Baseline	ACL
77.8	78.9

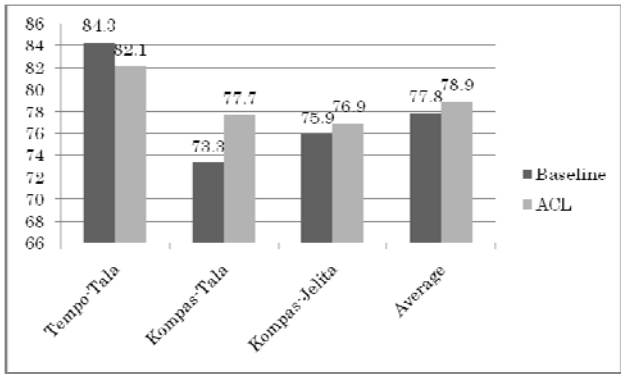

Figure 1: ASR accuracies (%) of the baseline and the abbreviation corrected lexicon (ACL) systems

4.4 Discussions

Figure 1 shows that average accuracy of the ACL system using Tempo-Tala corpus as testing data is significantly decreased (2.2%) from that of the baseline system. One of the reasons is that the Tempo-Tala testing queries contain a lot of "di" word (1,673 words) which has two meanings with different pronunciations depending on the context:

- If it corresponds to "di", which is a particle of place such as "at" in English, it should be pronounced as "d-i".
- If it corresponds to "DI", which is Darul Islam (Islamic Organization in Indonesia), it should be pronounced as "d-e-i".

All speakers in the Tempo-Tala corpus pronounced it following the first rule, however in the lexicon the available pronunciation resulted from the ACL method follows the second rule. Thus, all "di" words were incorrectly recognized. The ACL method gives negative effects to the word which has two pronunciations, as a regular word and as an abbreviation (letter pronunciation).

In the test data consisting of 43,797 words, 2,667 words are abbreviations (6.1% of the testing data). By using a pre-defined abbreviation list, 1,985 abbreviations (74.4% from the total number of abbreviations in the testing data) were correctly recognized. By adding the syllable rules, 2,101 (78.8%) abbreviations were correctly recognized. Incorrectly recognized abbreviations include foreign abbreviations, OOVs, abbreviations that could not be detected, and other errors caused by unclear pronunciation, etc. Abbreviations that still could not be detected include:

- Abbreviations pronounced as a combination of letters and words, e.g. W3C (W three C) for World Wide Web Consortium.
- Abbreviations pronounced as a word or letters, depending on speakers or context, e.g. FAQ: ([f-ae-k] or F A Q) frequently asked questions.
- Abbreviations pronounced by the letters but following Indonesian syllable patterns and not defined in the abbreviation list, e.g. DPA [de-pe-a] which means high advisement board in English.

5 IR Experiments

5.1 Experimental condition

We use Inference Network-based (IN-based) IR [5]. A transcribed query is fed to the IR system after removing stop words in the Indonesian language. Each correct query text is also given to the IR in order to compare the result with that obtained using ASR. We use the mean reciprocal rank (MRR) as the evaluation measure. The ad-hoc retrieval task was conducted using IR collections stored with the TREC format. The collections contain documents, queries, and exhaustive relevance judgments.

5.2 Occurrence weight

In order to enrich the transcribed queries, we also use n-best results instead of only using the 1-best as the input to the IR system. In addition, the number of occurrence for each term in the n-best list is used to explicitly weight each term in the transcribed query as a simple measure of certainty for each recognized word. The bigger the occurrence score is, the more certain the recognized word is. The following is an example of how we use the term occurrence in the 5-best list. Suppose we have 5-best outputs as follows:

- 1-best result : w_1 w_2 w_3
- 2-best result : w_1 w_2 w_4
- 3-best result : w_1 w_2 w_4
- 4-best result : w_1 w_2 w_3
- 5-best result : w_1 w_2 w_3

The query to be fed to the IR system is (5 w_1, 5 w_2, 3 w_3, 2 w_4). The number indicates the occurrence of each term in the 5-best list.

5.3 Evaluation

Table 2 shows MRR scores for the spoken queries using the abbreviation-corrected lexicon (ACL) ASR and the baseline ASR, as well as the text queries. Although the recognition accuracy for the Kompas-Tala corpus was decreased by using the ACL ASR comparing with the baseline system as shown in Fig. 1, Table 2 shows that their MRR score is increased. This is because the "di" word which was misrecognized by the ACL ASR is a word included in the stop list, and therefore does not affect the IR performance.

The occurrence weight for the n-best results works well for all the testing data as shown in Figure 2. The MRR score is improved comparing to the 1-best result. On average 0.9% improvement was achieved by including the occurrence weight in the retrieval process.

Table 2: MRR score (%) for spoken queries (1-best result) using baseline ASR (Base) and abbreviation-corrected lexicon (ACL) ASR, as well as text queries.

ASR system	**Tempo -Tala**	**Kompas -Tala**	**Kompas -Jelita**	**Ave-rage**
Base	82.7	75.5	61.4	73.2
ACL	83.4	76.8	62.7	74.3
Text query	86.7	85.5	69.2	80.5

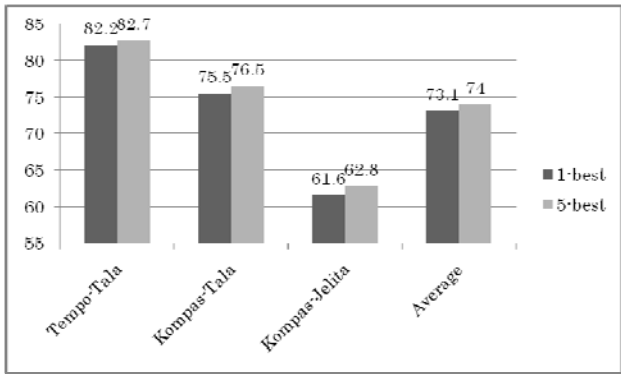

Figure 2: Comparison of MRR score (%) using 1-best results and that using 5-best results weighted by word occurrence.

6 Conclusions and future work

We have proposed an automatic abbreviation detection method using syllable-based word composition rules to improve the recognition accuracy of abbreviations. In order to incorporate a simple measure of the certainty of the ASR results in the retrieval process, a confidence score based on the word occurrence is used in the inference network (IN)-based IR. Our experimental results show that this technique improves the IR performance.

In the Indonesian language, there are specific rules to make a consonant cluster (a group of consonants). These rules need to be incorporated in the future to improve the accuracy in detecting abbreviations.

Acknowledgements

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

References

[1] Department of Cultural and Education of Republic of Indonesia, Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan, 2nd edition, Keputusan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 0543a/U/1987, 9 September 1987.

[2] D.P Lestari, S. Furui, "Adaptation to Pronunciation Variations in Indonesian Spoken Query-based Information Retrieval", IEICE Trans. Information and System, Vol.E93-D, No.9, pp.2388-2396, September 2010.

[3] 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.

[4] J. Asian, H.E. Williams, and S.M.M. Tahaghoghi, "A Testbed for Indonesian text retrieval", Proc. 9th Australasian Document Computing Symposium (ADCS 2004), pp. 55-58, Melbourne, Australia, 13 December 2004.

[5] H.R. Turtle and W.B. Croft, "Inference networks for document retrieval", ACM Transactions on Information Systems, vol. 9, no. 3, pp. 187-222, July 1991.