論文 / 著書情報
Article / Book Information

Title	Language models and dialogue strategy for a voice QA system
Author	Daewoo Kim, Sadaoki Furui, Hideki Isozaki
Journal/Book name	18th international congress on acoustics (ICA2004), Vol. , No. 5, pp. 3705-3708
発行日 / Issue date	2004, 4

Language Models and Dialogue Strategy for a Voice QA System

Daewoo Kim, Sadaoki Furui, Hideki Isozaki

Department of Computer Science, Tokyo Institute of Technology,
and NTT Communication Science Laboratories
daewoo@furui.cs.titech.ac.jp, furui@furui.cs.titech.ac.jp, isozaki@cslab.kecl.ntt.co.jp

Abstract

This paper describes a method of generating effective language models for voice query recognition and a new dialogue strategy for a voice activated QA system. By using multiple domain language models, better recognition accuracy is obtained for query utterances. In the proposed interactive dialogue strategy using multimodal user interface, users are requested to indicate correct keywords and incorrect keywords are automatically replaced by most probable keywords in the n-best list based on word co-occurrence probabilities. A preliminary QA system using voice input and graphical user interface has been implemented using the SAIQA open-domain QA system.

1. Introduction

Although various QA systems (Eg. [1][2]) have recently been investigated, most of them use keyboards to input queries to the systems. If voice can be used instead, the systems are expected to become more useful and easier for many users. In the voice QA systems that respond to the queries given by voice to retrieve articles from a wide range of domains [3][4], how to reduce the effect of speech recognition errors in the query utterances on the answers produced by the systems is one of the most important research issues from usability point of view.

This paper proposes a new method in which multiple language models for recognizing voice queries are trained using clustered news paper articles corresponding to multiple domains, such as economy, entertainment, international affairs, general public and sports. Each domain-dependent language model is interpolated with a class language model constructed using transcribed query utterances. The interpolated multiple language models are used in parallel in recognizing each query utterance and a hypothesis having the largest likelihood score is chosen. Recognition experiments using many query utterances are conducted to evaluate the effectiveness in improving the recognition accuracy. This paper also reports a new dialogue strategy that effectively dissolves the problem of speech recognition errors by a small number of human-computer interactions. Structure of a prototype of a multimodal voice QA system that is now being built is also presented in the paper.

Section 2 describes how to make effective language models for voice activated QA systems. Section 3 explains our dialogue strategy for voice activated QA and its effectiveness. Section 4 presents our multimodal voice QA system which is now being built and the paper is concluded in Section 5.

2. Language models

Almost all the voice activated QA systems that have recently been investigated use a domain-independent language model for recognizing query utterances. Although the domain-independent language model can handle every kind of queries, its recognition rate is not always high enough for getting correct answers and its processing speed is relatively low. It is well known that speech recognition performance can be significantly improved by using a domain/topic-dependent language model that matches the input utterance.

2.1. Domain-dependent language models

In this paper, five domains are defined as subsets of the whole open domain: economy(eco), entertainment(ent), international affairs(int), general public(soc) and sports(spo). Mainichi newspaper corpus from 1994 to 2001 is used for generating a domain-independent language model and each domain-dependent language model is generated by a subset of the corpus. Clustering of the database is performed using domain indexes given in the corpus. CMU-Cambridge Toolkit is used for generating the language models.

Table 1: *Effectiveness of the domain-dependent language models*

LM	Correctness(%)
Domain-independent	67.49
Eco domain	67.82
Ent domain	68.29
Int domain	68.12
Soc domain	67.91
Spo domain	67.75
MAX	69.35

Table 1 shows speech recognition results for query ut-

terances by seven speakers, each uttering 100 utterances, using the domain-independent language model and a specific language model corresponding to the domain of each utterance. MAX indicates the method in which the language model achieving the maximum likelihood is automatically selected for each utterance. Julius speech recognition engine and PTM (phonetically-tied mixture) HMMs having 3000 states and 64 Gaussian mixtures in each state were used. The insertion penalty and the language model weight were optimized by a development set of utterances.

The word correctness is used in the evaluation instead of word accuracy, since insertion error causes relatively small problems due to the dialogue strategy for the QA described in Section 3.

The experimental results show that every domain-dependent language model achieves better recognition performance than the domain-independent model. Even if the domain that each query belongs to is unknown, the best performance can be obtained by choosing the domain-dependent language model that maximizes the likelihood (MAX). The correctness was improved by approximately 2% in absolute value by this method in comparison with the domain-independent method.

2.2. Interpolation of language models

Since the domain-dependent language models are generated using the newspaper corpus, the models do not fit to the spontaneous query utterances. The models are therefore improved by interpolating with the language model made by a corpus of 2000 query utterances. Since the coverage of keywords, especially named entities in the collected queries, is very low, a class language model is built before interpolating with the domain-dependent models. The model interpolation is performed using the SRI toolkit.

Figure 1 shows variation of the word correctness of the query utterances according to the interpolation ratio of the domain-dependent language model and the query language model.

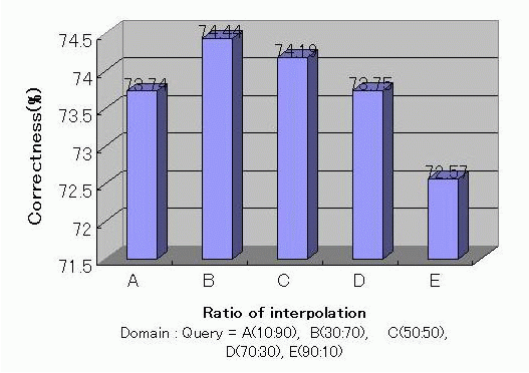

Figure 1: *Word correctness of the query utterances with different interpolation ratios.*

The largest improvement of the correctness was obtained when the interpolation ratio was set at 30(domain-dependent):70(query). At this interpolation condition, 5% improvement was obtained compared with the domain-independent model.

3. Dialogue Strategy

Speech recognition errors in query utterances cause errors in QA. Therefore, it is important to design an efficient strategy of making a dialogue with the computer system to avoid the effect of recognition errors on the QA. This paper proposes a new strategy using GUI (graphical user interface) in which speech recognition errors are effectively corrected in the dialogue process. After recognizing each query, keywords are automatically extracted and displayed on a screen. The user is then requested to indicate correct keywords, and the incorrect keywords are automatically replaced by other candidate keywords by using a proposed method described below. A set of correct keywords are transmitted to SAIQA QA system made by NTT Lab[1]. The SAIQA system was originally built for accepting keyed-in queries.

3.1. Keyword extraction

Since Japanese words have inflections and there is no spacing between words in the written form of Japanese sentences, morphemes are usually used as units for Japanese speech recognition. In our system, CHASEN morphological analysis tool was used for segmenting sentences into morphemes in building language models. Therefore, speech recognition results are given as a sequence of morphemes.

Table 2 shows an example of a sequence of morphemes. The sentence is ”日本の首相は誰ですか。” which means ”Who is the prime minister of Japan?”.

Table 2: *Morphological analysis and keyword selection*

Word	Word class	Selection
日本 ”Japan”	Noun	*Selected*
の ”of”	Postpositional particle	Discarded
首相 ”minister”	Noun	*Selected*
は ”is”	Postpositional particle	Discarded
誰 ”who”	Noun	*Selected*
です ”is”	Auxiliary verb	Discarded
か ”?”	Postpositional particle	Discarded

The SAIQA QA system uses only three kinds of word classes: noun, adjective and verb. In the example shown in the table, selected keywords are ”日本”, ”首相” and ”誰”.

By using a GUI, the user is requested to indicate whether the keywords automatically selected from the recognition results by the system are correct or not. If there exist incorrect words indicated by the user, they are auto-

matically replaced by other keywords in the N-best list obtained as a result of query utterance recognition.

3.2. Co-occurrence probability

In order to achieve good performance in replacing incorrect keywords with other candidate keywords in the N-best list, it is necessary to use some linguistic knowledge that could be extracted from information sources in the QA system. This paper proposes to use domain-dependent probabilities of keyword co-occurrences in each paragraph for this purpose.

The co-occurrence probabilities are calculated for each of the five domains using the same clustered databases used in generating the domain-dependent language models. Only the words occurring more than one thousand times in each domain are used for calculating the co-occurrence probabilities to save the cost of both probability calculation and execution time in using the probabilities.

The number of words observed more than one thousand times in each domain is 8071(eco), 7717(ent), 7939(int), 9245(soc) and 8463(spo). The co-occurrence probabilities are calculated by equation (1), in which N indicates the number of appearances of word pair (i, j) in the same paragraph and NA indicates the total number of paragraphs.

$$C_d(i,j) = \frac{N_d(i,j)}{NA_d} \tag{1}$$

After calculating the co-occurrence probabilities for all the pairs of keywords by equation (1), the probabilities are normalized by equation (2).

$$C_d(i,j)' = \frac{C_d(i,j)}{CMAX_d} \tag{2}$$

where $CMAX_d$ is the maximum value of $C_d(i,j)$ over every combination of i and j.

3.3. Replacement

In order to obtain a candidate keyword set to replace incorrect keywords, N-best sentence hypotheses are generated for each query utterance, and nouns, adjectives and verbs are selected from all the hypotheses.

The following score is calculated for each keyword c extracted from the hypotheses using the co-occurrence probabilities, and a keyword having the maximum value is selected to replace the incorrect keyword.

$$Score(c) = \sum_{i=1}^{n} \log C_d(k_i, c) \tag{3}$$

where k_i denotes each keyword in the query that is indicated as a correct keyword by the user.

In the evaluation experiment, the number of N-best hypotheses was controlled by the beam-width used in the recognition decoder under the constraint that the maximum number of hypotheses was 100. Table 3 shows the improvement of the correctness(%) of keyword recognition by the keyword replacement procedure as a function of the beam-width. In the table, "baseline" indicates the results before the keyword replacing process, "upper limit" indicates ideal results that could be obtained when all the correct keywords in the hypotheses are correctly selected, and "replaced" indicates the results after replacing incorrect keywords using the proposed method. These results indicate that 14 to 19 % of correct keywords can be selected from the hypotheses by the proposed method.

Table 3: *Improvement of the correctness (%) of keyword recognition by the keyword replacement procedure at each beam-width condition*

Beam-width	Default	1000	10000
Baseline(B)	67.12	68.79	70.61
Upper limit(U)	70.97	72.73	74.50
After replace(R)	67.84	69.34	71.26
$\frac{R-B}{U-B} * 100$	18.70	13.96	16.71

4. Structure of the QA system

Structure of our voice QA system that is now being built is shown in Fig. 2. The system consists of three components: GUI, a set of multiple speech recognition engines, and the SAIQA QA system.

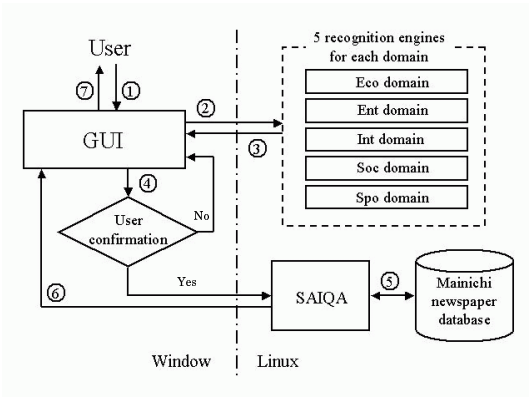

Figure 2: *Structure of our QA system.*

The GUI is made by Visual C++ with TVML (TV Making Language)[1] which is a tool that can show and move several 3D characters and objects on a screen. By using this tool, the system can make the user feel as if he/she is talking to humans. This makes the system more comfortable than text-based systems for the user. Actions of the characters help the conversation between users and the system.

[1] http://www.strl.nhk.or.jp/TVML

Figure 3 shows the GUI having five 3D characters corresponding to the five domains described above. Each character speaks and makes actions when its corresponding domain is selected by the system to be the most relevant to the query according to the likelihood value.

When the system has recognized the input query and extracted keywords from the recognition result, the keywords are shown as buttons in a window used in the user interface. The user is requested to select incorrect keywords by clicking the buttons, and if at least one button is pushed, the incorrect keywords are automatically replaced using co-occurrence probabilities. Results of the replacement are also shown in the same window as soon as it is done so that the user does not feel interruption of the conversation.

The whole process of the human-computer interaction using this structure is as follows.

1. Voice query input
   A user inputs a voice query to the system.

2. Voice query recognition
   The input voice query is transmitted to the multiple recognition engines by TCP/IP protocols. Five recognition engines having domain-dependent language models recognize the input voice query in parallel and one result having the maximum likelihood is chosen.

3. Keywords extraction
   Keywords to be used in the QA system are extracted from the recognition result.

4. Showing keywords and requesting the user to select incorrect keywords
   The selected incorrect keywords are automatically replaced by keywords selected from N-best hypotheses based on co-occurrence probabilities and displayed to the user.

5. Receiving and showing a list of answers from the QA system SAIQA
   Ranked multiple answers given by the QA system are presented on the screen, since the top-ranked answer is not always the best for the query.

6. Asking correctness of the answers to the user
   If the user is not satisfied with the answer, the process returns to step 4 and the user can replace or add keywords until satisfactory answers are obtained.

7. Synthesizing the answer by a commercial TTS (speech synthesizer) and showing evidences and reasons why the answer is obtained from the database

Figure 3: *GUI of the prototype QA system.*

5. Conclusions

This paper has proposed a method for generating effective language models and a useful and convenient dialogue strategy for voice activated QA systems. By using multiple domain-dependent language models interpolated with a query language model, better recognition results for query utterances have been obtained. The proposed system structure has an advantage in that new domains can be easily added by simply adding new domain-dependent language models without requiring any modification to the existing models.

Incorrect keywords in the speech recognition results indicated by the user using GUI are automatically replaced by keyword candidates extracted from N-best hypotheses based on domain-dependent probabilities of co-occurrences with correct keywords.

Our future work includes evaluating our methods in obtaining correct answers to voice queries. It is expected that speech recognition method using multiple domain-dependent language models and the GUI-based user-interface including the efficient error correction strategy will play important roles in making efficient, practical and useful voice QA systems. Our future work also includes a new recognition engine that directly recognizes keywords instead of recognizing all words in the query utterances using various linguistic information extracted from knowledge resources.

6. References

[1] S. Harabagiu et. al, “Experiments with Open-Domain Texutual Question Answering”, COLING 2000, pp.292 - 298, August 2000.

[2] Y. Sasaki et. al, “SAIQA : A Japanese QA System Based on a Large-Scale Corpus”, IPSJ SIG notes FI-2001-9-11, 2001.

[3] S. Harabagiu et. al, “Open-Domain Voice-Activated Question Answering”, COLING 2002, Vol I, pp.321 - 327, 2002.

[4] Chiori Hori, Takaaki Hori and Hideki Isozaki, “Deriving Disambiguous Queries in a Spoken Interactive ODQA System”, ICASSP 2003 I-624, 624 - 627, 2003.