ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

AN INFORMATION GAIN AND GRAMMAR COMPLEXITY BASED APPROACH TO ATTRIBUTE SELECTION IN SPEECH ENABLED INFORMATION RETRIEVAL DIALOGS

Haiping Li, Haixin Chai,

IBM China Research Lab, Beijing 10085, China

ABSTRACT

Effective dialog driven method is required for today's speech enabled information retrieval systems, such as name dialer. Similar to dynamic sales dialog for electronic commerce scenarios, the information gain measure based approaches are widely used for attribute selection and dialog length reduction. However for speech enabled information retrieval systems, another important factor influencing attribute selection is speech recognition accuracy. Too low accuracy will result in a failed dialog. Recognition accuracy varies with many issues including acoustic model performance, grammar's complexity. Acoustic model is fixed for a whole dialog, while grammar is different for each interaction round, thereby grammar complexity will influence the attribute selected for next question. In this paper, an approach combining both information gain measurement and grammar complexity is present for dynamic dialog driven. Off-line evaluations show that this approach can give a trade-off of faster discriminating the candidates for retrieval target and higher recognition accuracy.

1. Introduction

Speech enabled information retrieval system is widely used as automation service, to provide effective customer interaction and reduce the manual work load. Name dialer is such a typical system. It takes human operator's responsibility by automatically recognizing and dispatching most incoming calls to the target destination. Normally such a system relies on pre-recorded prompts (or a speech synthesizer) and speech recognizer, use a dialog other than single round interaction to communicate with the caller, to identify end-user's target. In addition, since current speech recognition accuracy is not perfect enough, some fuzzy technologies need to be applied to shorten the dialog interaction. For example, when a caller speaks a name to the system, a typical way of the output of recognizer is not a single name, but a Target Name Set (TNS), which includes the top N candidates. Then by asking questions on the most distinctive attribute of names in this set, the TNS will be narrowed down and the target can be eventually pinpointed.

Normally, the system developer is responsible for designing the call flow and interaction questions for dialogs in the information retrieval system. The selected attributes for questions and corresponding asking order are pre-defined, which may potentially affect the system efficiency or accuracy, either because more interaction rounds with the caller may be required for reducing the size of the TNS, or the grammars used for the corresponding question may grow too complicated. There should be an optimal attribute among all the candidates, which best satisfies the discriminating task, meanwhile keeps the complexity of its corresponding grammar relatively low. By selecting this optimal attribute, the target could be found with possible fewer questions, and also possible improvement in recognition accuracy.

For dynamic sales dialogs in electronic commerce scenarios, it is very important to ask as few questions as possible adapted to the customers' knowledge about the product space. This is similar to the scenario in speech enabled information retrieval systems. It has to be taken into account that online customers are very quickly annoyed and/or bored. Recently, a couple of case based reasoning approaches to automated sales dialogs have been suggested [1, 2]. The ideas that can be found have in common that their aim is the reduction of the number of questions (dialog length) a customer is asked by the sales system. Most of the approaches are based on an information gain measure that is used to select the next attribute to ask which is maximally discriminating the product database. Our approach applies similar measures in speech enabled information retrieval dialogs.

In speech enabled information retrieval systems, another key factor influencing the dialog length is recognition accuracy. If a recognized result is incorrect, the call flow will be direct away from an optional path. If the accuracy is so poor that recognized results are frequently wrong, an automatic dialog will be failed, and it will tend to a manual operator.

Speech recognition accuracy varies according to different technologies like acoustic modeling, searching algorithm etc. However, for a given finite technology implementation, the accuracy may still vary according to complexity of given grammar. Various grammar types, sizes and members may result in quite different accuracy. Thus a new concept, grammar complexity, is introduced in our approach to represent the relationship between such grammar properties and recognition accuracy, and a measure, Grammar Complexity Cost (GCC), is used in cost function for attribute selection.

In this paper, a name dialer system is used as an example to describe the proposed approach. This approach can be easily implemented in other wide range speech enabled systems, such as Directory Dialer, Directory Assistance, to automatically select attributes for question-generation through a dialog manager, or a simply designed call flow control component.

The paper is structured as following. Section 2 described the name dialer system used as the deployment environment. In section 3, the information gain and the GCC defined in our approach are introduces in details. In Section 4, the results of off-line experiments are reported. The future work is presented in section 5.

2. System Overview

Our name dialer system is illustrated in Figure 1. When a user calls in, the system prompts to ask for the staff name. After the user speaks the name, his / her utterance is fed into the speech recognizer, and the recognition results are further processed by fuzzy methods to improve the hit rate of target name. The recognized results with top n scores are output as a TNS. The Call Flow Controller receives this TNS and sends it to the Attribute Selector.

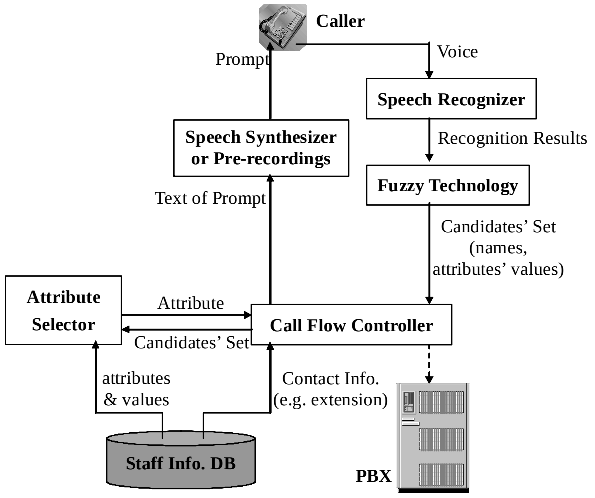

Figure 1: System Architecture

If the number of names in TNS is below a given threshold, defined as Terminal Set Size (TSS), the Attribute Selector takes this set as a terminal state. In this case, no attribute is needed for generating further question(s) to reach the target name. And the selector will search the Staff Information DB for the attributes to distinguish all the members in TNS. After such information is prompt, the caller can select the target by pressing the selection key in the telephone panel or speaking it. For example, there are two names in the TNS, Harry Cheung and Henry Cheung, and their attributes in the DB are listed in table 1.

Table 1: Example of Attribute Information in DB

Names	Gender	Group	Title	Ext.
Harry Cheung	M	Speech	Engineer	318
Henry Cheung	M	Image	Engineer	319

The Group information will be selected for generating name options like,

1, Harry Cheung, Speech Group
2, Henry Cheung, Image Group

If the caller presses 1, the Call Flow Controller searches the DB to get the extension 318, and dispatches the call by sending a connection request to the PBX.

According to user experiences, the TSS should not be too large since all the name options with distinguishable attribute values will be read out. Otherwise, most callers won't have enough patience to finish listening all the prompts. Our experience range is from three to five.

If the TNS size is larger than TSS, the approach to select attribute will be applied. An example for this case is given in the following. The TSS is set to be three, there are four names in TNS, and their attributes' values are listed as table 2.

Table 2: Example of TNS

Names	Gender	Group	Title
Harry Cheung	M	Speech	Engineer
Henry Cheung	M	Image	Engineer
Henry Cheng	M	Market	Administrator
Henry Cheng	M	Support	Administrator

If the Attribute Selector selects the attribute Group, a question like which group is the employee in will be prompt to caller. And then the caller's oral response is fed into recognition engine again. This time, the grammar does not consist of the staff names, but the attribute values. In this example, the grammar in the second iteration round consists of four entries, Speech, Image, Market and Support. With the same fuzzy processes on recognition results, the candidates' set becomes a set of values of the corresponding attribute. The Attribute Selector reduces the TNS by only keeping the names having attribute values in this value set. Then it compares the size of reduced TNS with the TSS again. Such a process is kept iterating, until the terminal state is reached, or the attribute selector fails to select a suitable attribute. In the second case, the call flow controller will transfer the call to the operator for manual service. In this example, suppose the caller's answer is Speech, and the candidates' set after fuzzy processes consists of Speech and Support, the reduced TNS will only include Harry Cheung and Henry Cheng. At this time its size is two, so it will be a terminal state now.

3. Attribute selection approach

If an attribute is selected, the question for getting a value for it will be generated in the next interaction round. And the corresponding grammar used in speech recognizer consists of all the values of this attribute. In our approach, two criteria are used for selecting attribute. One is reducing the number of continuous interaction round (dialog depth), the other is the grammar corresponding to this attribute should not be too complicated otherwise the recognition accuracy will be low. Our approach applies decision trees to classify the names in TNS. It uses information gain of the first layer nodes in decision tree to measure the first criteria, uses the GCC to measure the second one, and combines these two measures using cost function.

3.1 Problem description

Decision trees are known as an efficient classification technique, which commonly selects the attribute reaching the maximal gain or gain ratio on information to partition the sample space [3, 4]. In this name dialer scenario, the attribute selection problem is described as:

Given a TNS T, and suppose there are |T| names in it. Use a non-categorical attribute X to partition the set T, and it is partitioned into sub-sets T1, T2, …, Tn on the basis of the value of X. Select the X having the maximal score defined as:

$$\text{Score}(X) = \text{Gain}(X, T) - SF * GCC(g(X)) \qquad (1)$$

Where Gain(X, T) is the gain in information due to attribute X, and GCC(g(X)) is the GCC of grammar g(X), which consists of values of attribute X for all the staffs in the TNS. The SF is introduced for adjusting the weight of Gain(X, T) and GCC.

Gain(X, T) represents the difference between the information needed to identify an element of T and the information needed to identify an element of T after the value of attribute X has been obtained. According to the common definition in the Decision Tree Classifier, the quantity Gain(X,T) is,

$$Gain(X,T) = Info(T) - Info(X,T)$$

Where Info(T) is the information needed to identify the class of an element of T, if the set T is partitioned into disjoint exhaustive classes C1, C2, .., Ck on the basis of the value of the categorical attribute. It equals the information conveyed by the probability distribution P of the partition (C1, C2, .., Ck):

$$P = (p_1, p_2, \cdots, p_k) = (\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \cdots, \frac{|C_k|}{|T|})$$

Thus,

$$\begin{aligned} Info(T) &= I(P) \\ &= -\sum_{i=1}^{k} p_i * \log(p_i) \\ &= -\sum_{i=1}^{k} \frac{|C_i|}{|T|} * \log(\frac{|C_i|}{|T|}) \end{aligned}$$

In our attribute selection problem, there is not a categorical attribute. The partitions in which the number of elements in each class (|Ci|) is less than TSS are acceptable. Thus, we can select the partition with the least entropy as the target partition,

$$(|C1|, |C2|, .., |Ck|) = (TSS, TSS, \ldots, L) \qquad (2)$$

Where L is the number of left names,

$$L = |T| - TSS*(k-1)$$

Thus,

$$Info(T) = -\sum_{i=1}^{k-1} \frac{TSS}{|T|} * \log(\frac{TSS}{|T|}) - \frac{L}{|T|} * \log(\frac{L}{|T|}) \qquad (3)$$

If we partition T on the basis of the value of a non-categorical attribute X into sets T1, T2, …, Tn, then the information needed to identify the class of an element of T becomes Info(X,T), which is the weighted average of the information needed to identify the class of an element of Ti,

$$Info(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} * Info(T_i) \qquad (4)$$

The calculation of Info(Ti) is similar to that of Info(T). Thus the Gain(X, T) is calculated. For the other item in formula (1), its definition and calculation are given in the following section.

3.2 GCC definition

The GCC(g(X)) reflects the difficulty of recognizing entries in the grammar g(X), which eventually influence the recognition accuracy. Many factors relate to GCC, such as the structure of grammar, the acoustic distance of entries in grammar, the number of path in grammar. A more complicated grammar results in a lower accuracy, so a higher GCC is assigned. In our name dialer system, the GCC is defined as,

$$GCC = TypeCost * ScaleCost \qquad (5)$$

The TypeCost is used to measure the GCC brought by grammar type. Classified according to recognition accuracy, grammar types include common words, mix-language, Alpha-Numeric, digit, etc. If the type of a grammar is common words, all the terminal words in it are in the same language. An example for such type is the grammar for the attribute Gender in a Chinese name dialer,

<gender> = 男 |
女 .

If there are words in another language, the grammar type is mix-language. Such grammars exist in bilingual systems. A grammar consisting of letters and numbers is called Alpha-Numeric grammar, and if only digits are contained, the grammar type is digit. The accuracy of recognizing words in grammars of the above three types is much lower than that of words in common words grammar.

Our name dialer system is a bilingual case, Chinese Mandarin and English, and the TypeCost is defined in table 3.

Table 3: Example of TypeCost Definition

Type	Cost
Mix-language	2.0
Alpha-Numeric	1.0
Digit	1.0
Common Words	1.0

A way to apply this definition is, firstly travel each word in a grammar for judging the grammar type. Then add up all the corresponding type costs. For example, a grammar is,

<group> = 语音组 |
Speech 组 |
S1 组 .

There are Chinese words 语音组 and 组, English word Speech, alphabet S, and number 1, thus its type is Common Words + Mix-language + Alpha-Numeric, and the TypeCost is calculated as 1.0+2.0+1.0 = 4.0.

The ScaleCost reflect the GCC brought by the size of grammar. The more paths in the grammar result in the lower accuracy. Thus the ScaleCost is in direct proportion with the size of grammar. In our name dialer, the grammar is a list of the values of an attribute thereby we use the following definition,

ScaleCost = lg (number of values)

In previous grammar example, the attribute Group has three values, so the ScaleCost is lg(3), and the GCC(Gender) is 1.91.

3.3 An example

In the following, an example is given to describe how to calculate score. Suppose in a Chinese-English name dialer system, there are six members in TNS, and the TSS to be three. The used attributes and their values for those members are listed in table 4.

Table 4: Attribute Values for the names in TNS

Names	Gender	Group	Title
李海萍	女	TTS 组	research member
李海平	男	TTS 组	research member
黎海平	男	TTS 组	advisory researcher
林海平	男	SR 组	manager
黎海萍	女	SR 组	research member
李孩平	男	SR 组	research member

For the attribute Gender, the partition (2) becomes (3, 3), so according to formula (3) and (4),

$$Info(T) = -2 * \frac{1}{2} * \log(\frac{3}{6}) = 1$$

$$Info(gender,T) = 2 * \frac{1}{6} * I(1) + \frac{4}{6} * I(4)$$

$$I(1) = 0, I(4) = -\frac{3}{4}\log(\frac{3}{4}) - \frac{1}{4}\log(\frac{1}{4}) = 0.81$$

Thus Gain(gender, T) = 0.46. According to formula (5), GCC(gender)=1.0*lg(2)=0.30

In formula (1), if the SF is set to be 1.00, the Score(gender) will be 0.16. The scores for the other two attributes are listed in table 5. The attribute Gender has the maximal score 0.16, and it will be selected.

Table 5: Gain, GCC and Scores

Attributes	Gender	Group	Title
Gain	0.46	1.00	0.46
GCC	0.30	0.90	1.43
Score	0.16	0.10	-1.00

4. Off-line Experiments

4.1. Trade-off of fewer questions and lower GCC

Off-line evaluations were performed in a bilingual (Chinese Mandarin and English) name dialer system with various number of attribute, various number of value, different type of grammar, and various number of staff in TNS. An example is shown in the following:

The TSS is 3.

Totally 10 attributes and their corresponding number of values are (12, 2, 3, 5, 5, 5, 5, 5, 5, 5).

The grammar types for the first three attributes are mix-language, Chinese common words, and mix-language respectively. Others are all digits.

The TNS contains 6 staffs.

Table 7: Attribute Score

Attribute Index	Info. Gain	GCC	Score (SF=0.5)	Score (SF=1.0)
1	1.00	1.81	0.10	-0.81
2	0.46	0.30	0.31	0.16
3	1.00	1.43	0.29	-0.43
4,7,9,10	1.00	0.95	0.53	0.05
5,6	1.00	1.20	0.40	-0.20
8	0.19	0.60	-0.11	-0.41

From the attribute evaluation results shown in table 7, it can be seen that the SF can be adjust to balance the information gain and the GCC. If the accuracy is considered to be a critical factor in a name dialer system, a larger SF can be set. In this example, when SF is 0.5, the attribute 4, 7, 9 or 10 with the largest score 0.53 is selected, as shown in column 4. And these attributes correspond to digit grammars. Works in speech recognition showed that the recognition accuracy of digit grammar is relatively low compared to common word grammar. In order to avoid digit grammar, the SF should be enlarged, say to be 1.0. Then we got another group of scores for these 10 attributes as shown in column 5. In this case, the attribute 2 which corresponds to a Chinese common words grammar is selected. And it will result in higher recognition accuracy.

So before running on-line name dialer system, the SF can be tuned to get optimized results.

4.2. Time Cost

The setting in tests for computation time was the same as that in the previous example, only the size of TNS was changed. For each TNS size, the recorded time was the average of 10 tests, in which the staffs in TNS were different. The testing machine was configured with the CPU of 1.8GHz, and the memory of 512M.

As the results shown in figure 2, the approach needed little computation time. Even the TNS is enlarged to 512, the computation time is only 41 ms, thereby such an approach can be applied in real time.

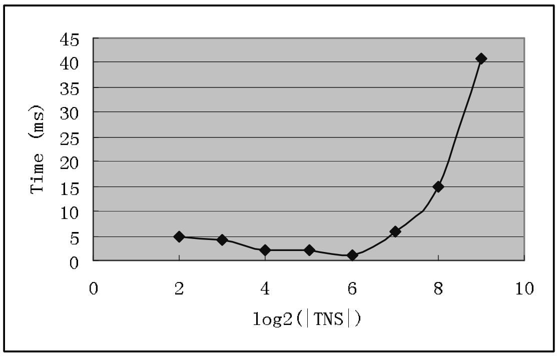

Figure 2: Computation Time

5. Future Work

This approach will be applied in our name dialer system for on-line testing, thus the rate of correct identification, number of queries per identification can be got to compare with the system without this approach. Improving work underway includes exploring new approaches for combining the information gain and the GCC, instead of merely using a cost function as described in this paper. Future research work will also put much focus on the GCC measurements.

6. Reference

[1] M. Doyle, P. Cunningham, "A Dynamic Approach to Reducing Dialog in On-Line Decision Guides", Proc. of the 5th European Workshop on Case-Based Reasoning, EWCBR 2000, Trento, Italy. LNAI 1898, Springer.

[2] A. Stahl, R. Bergmann, "Applying Recursive CBR for the Customization of Structured Products in an Electronic Shop", Proc. of the 5th European Workshop on Case-Based Reasoning, EWCBR 2000, Trento, Italy. LNAI 1898, Springer.

[3] R. Lopez de Mantaras, "A Distance--Based Attribute Selection Measure for Decision Tree Induction", Machine Learning, 6(1), 81--92, (1991).

[4] U. M. Fayyad and K. B. Irani, "The attribute selection problem in decision tree generation", in Proc. of Tenth National Conference on Artificial Intelligence, pages 104--110, Cambridge, 1992. MA: AAAI Press/MIT Press.