CASE BASED REASONING SOLUTION TO THE PROBLEM OF SUSTAINED LEARNING IN KEYWORD SPOTTING

Tieran Zheng, Jiqing Han, Guibin Zheng, Shiwen Deng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ABSTRACT

In some practical keyword spotting applications, users or service providers are willing to provide spotting-result feedback to help improve system performance. To do so, they require a keyword spotting technique with a sustained learning ability. This paper presents a new Chinese keyword spotting method based on a case based reasoning framework. Two level keyword case representations are adopted based on a set of symbols that are discriminative both in acoustic feature vector space and in semantic space. Then case bases are indexed with a tree structure and searched for test speech based on an elastic matching strategy. Finally, the feedback is used to adjust the statistics attached to the cases or to append new cases. Two experiments were conducted to compare our approach with a syllable lattice based method and to test the sustained learning ability.

Index Terms— Keyword spotting, sustained learning, case based reasoning, acoustic symbol clustering

1. INTRODUCTION

Keyword spotting plays a very important role in recognizing spontaneous conversation speech and finding speech of interest in spoken documents. A number of Hidden Markov Model (HMM)-based keyword spotting methods have been proposed in the past several decades, and they can be roughly divided into three general categories: filler model based [1,2], Large Vocabulary Continuous Speech Recognition (LVCSR) based [3] and word or subwordlattice based [4,5]. However, in some practical applications, users remain troubled with the presence of large amounts of false alarms or false rejects, which are mainly caused by the mismatch between acoustic model and input speech. Users remain troubled by many false alarms or rejections mainly caused by acoustic model/speech mismatches.

In HMM-based approaches, the acoustic model, which is the main basis for spotting hypothesized keywords, represents the knowledge of pronunciation variations covered by training data in a statistical manner. However, human pronunciations and their acoustic representations are easily influenced by many factors such as speaking styles, accents, environments and channels. Wide varieties of real applications introduce infinite pronunciation variations and make it impossible to train such a model to cover all of the necessary knowledge in advance. In a specific application, the existence of pronunciation variations not covered by the acoustic model results in the problem of model mismatch, which degrades the model's performance.

To improve the performance of their keyword spotting systems, some users or service providers willingly provide consistent, accurate spotting-result feedback. The feedback, which refers to unknown pronunciation variations knowledge, can be used to gradually extend system knowledge bases and therefore achieve better performance. In such case, a kind of keyword spotting technique with a sustained learning ability is needed.

Case Based Reasoning (CBR) [6] solution is a better choice than model based solutions for constructing a sustained learning algorithm. CBR takes the specific knowledge of previously experienced as cases, and new problem can be solved by finding a similar past case, which is a powerful and frequently applied method of human problem solving. Moreover, being different from statistical model methods, it is an approach to incremental, sustained learning in nature, as a new experience can be directly retained and made immediately available to help solve future problems.

This paper presents a Chinese keyword spotting approach based on a CBR framework. First, a set of discrete symbols is obtained to represent the acoustic feature vector space using an agglomerative hierarchical clustering algorithm. A composite clustering criterion is adopted in the algorithm after accounting for three factors, including intra-cluster similarity, discriminative pronunciation representations and the size of the symbol set. Two level keyword case representations are then designed based on the symbols. After each case is subjected to an endpoint relaxation process, two level case bases are constructed and indexed with tree structures. A case searching algorithm based on an elastic matching strategy is then proposed to find matched cases for the test speech, and the keyword occurrence probability is estimated simultaneously. In our approach, we implement user feedback to adjust the statistics attached to the cases or to append new cases. Finally, we present the results from two experiments conducted to evaluate our approach.

2. CASE REPRESENTATION

In speech recognition, speech is usually represented in terms of acoustic feature vectors such as MFCC, by which each pronunciation can be viewed as a sequence of feature vectors. A direct idea is to take the sequences corresponding to keywords as keyword cases in our CBR based approach. However, indexing and searching the case base is computationally expensive and impractical for its huge and increasing size. Mapping the vectors

into a set of discrete symbols and then taking a sequence of the symbols as a case is a more reasonable idea. The symbols, which we term "acoustic symbols," refer to distinct groups that exhaust all of the vectors in the feature vector space. Determining a suitable set of acoustic symbols is obviously a clustering issue.

2.1. Acoustic symbol clustering

Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_M\}$ be the set of all feature vectors in training data. A clustering of X into N clusters that are labeled with the acoustic symbols $S = \{s_1, s_2, \dots, s_N\}$ needs to account for the following requirements.

1) *The requirement of high intra-cluster similarity.* The vectors contained in each cluster should be similar to each other. The sum of squared error (SSE) is a widely used evaluation for this requirement, and can be expressed as

$$E = \sum_{n=1}^{N} \sum_{\boldsymbol{x} \in s_n} \|\boldsymbol{x} - \boldsymbol{m_n}\|^2, \quad (1)$$

where $\boldsymbol{m_n}$ is the mean of the vectors in cluster s_n. The smaller the E value, the higher the intra-cluster similarity that can be achieved for S. If only this requirement must be satisfied, the K-mean clustering algorithm is a very good choice for minimizing the SSE. However, other requirements also need to be considered in our clustering task.

2) *The requirement of highly discriminative representations of pronunciations.* When using a sequence of symbols to represent speech, different pronunciations should be highly discriminative. In other words, they should be less confusing. In our Chinese speech recognition approach, the confusion can be evaluated on the set of all syllables due to its fixed and manageable size. The set of all syllables is denoted by $Y = \{y_1, \dots, y_U\}$, and let $L = \{l_1, \dots, l_Z\}$ denotes the set of all possible sequences belonging to a syllable. In L, each $l_z, 1 \le z \le Z$ is a distinct sequence of acoustic symbols, and it may appear as different syllables in speech. For this, a conditional probabilities matrix $P(y_u|l_z), 1 \le z \le Z, 1 \le u \le U$ can be estimated using the maximum likelihood estimation method. The syllable confusion that arises from adopting S can be measured with the conditional entropy $H(Y|S)$:

$$H(Y|S) = H(Y|L) = -\sum_{z=1}^{Z} P(l_z)\, H(Y|l_z)$$
$$= -\sum_{z=1}^{Z} P(l_z) \sum_{u=1}^{U} P(y_u|l_z) log P(y_u|l_z), \quad (2)$$

where $P(l_z)$ is the probability of l_z. This value is easy to estimate in training data. A reasonable symbol set should introduce a very low $H(Y|S)$ value.

Eq. (1) evaluates the distinguishability of the clusters in the feature vector space, and Eq. (2) does the same in the semantic space of spoken language. When both of these requirements are satisfied, a composite clustering criterion J_1 is given:

$$J_1 = E^{\lambda_1} \cdot H(Y|S), \quad (3)$$

where λ_1 is a weighted factor to balance the two requirements.

We must also decide the N value. Among the clusterings that can introduce a sufficiently small value of J_1, the clustering with a smaller N value should be more suitable to be adopted in our approach, since less computation and store resources would be taken based on its corresponding symbol set. Therefore, the clustering criterion can be modified as

$$J_2 = N^{\lambda_2} \cdot E^{\lambda_1} \cdot H(Y|S). \quad (4)$$

The clustering procedure can be described as follows. First, a sufficient amount of initial clusters are obtained using the K-mean clustering algorithm. An agglomerative hierarchical clustering algorithm is then adopted to create a hierarchy of clusterings from the initial clusters step by step. At each step, each pair of clusters is merged into one cluster to get an alternative of the successive clustering. The E value and $H(Y|S)$ value of those alternatives are calculated in training data, and then the alternative with the minimum J_1 value is taken as the successive clustering. Finally, within the hierarchy, the clustering that takes the minimum J_2 value is determined as our symbol set.

A Gaussian probability density function (pdf) is estimated for each symbol (cluster). The probabilities that feature vectors are contained in the corresponding symbols are calculated based on the pdfs, and then each vector is assigned to the symbol with the maximum probability.

2.2. Two level keyword case representations

- Word-level case

In the training data, all frames are mapped to acoustic symbols and all the symbol sequences that refer to predefined keywords are extracted. Among them, each kind of the sequences is taken as a word-level case. For each case may be shared by different keywords, a group of statistics can be collected for it. For case $c_i, 1 \le i \le I$, its statistics B_i can be expressed as

$$B_i = \left\{N^i_{kw_1}, N^i_{kw_2}, \dots, N^i_{kw_Q}, N^i_O\right\}, \quad (5)$$

where $N^i_{kw_q}, 1 \le q \le Q$ is the number of instances of keyword kw_q that belong to c_i, N^i_O is the number of other pronunciations whose representations match c_i but are not among the keywords. N^i_O can be obtained by searching the entire training data with c_i. For the need of revising the cases online, the absolute numbers rather than some ratio values are adopted here.

- Syllable-level case

There are just a few keyword instances in the training data. It is clear that the word-level cases extracted from the training data are inadequate to support even a simple keyword spotting task. Thus, syllable-level cases are also adopted in our approach. The pronunciations of the syllables that are part of the predefined keywords are labeled with accurate boundaries in the training data and their corresponding symbol sequences are extracted as syllable-level cases. A group of statistics is also attached to each case, and includes the number of instances of all syllables that belong to the case and the number of other pronunciations that match the case but do not exactly match any syllables in the training data.

3. CASE BASE INDEXING

3.1. Endpoint relaxation

The starting and ending symbols in the cases are easy to be influenced by their surrounding utterances. To avoid the influence in case matching and reduce the size of our case base, we adopt an endpoint relaxation process. A universal starting symbol is defined as a collection of the cases' starting symbols. Matching of the universal starting symbol is different from the normal symbols. If any element in the collection can be matched with a target symbol, the universal starting symbol is considered matched with it. A universal ending symbol is similarly defined as a collection of the cases' ending symbols. In each case, the starting and ending symbols are replaced with the universal starting and ending symbols. After that endpoint relaxation process, some cases that become exactly same now can be merged together to reduce the number of the keyword cases. Moreover, the new cases can be more flexible in case matching.

3.2. Index tree

Two level case bases are constructed: a word-level case base and a syllable-level case base. All of the word-level cases converge to form a word-level case base. To support efficient searching, the cases are merged into a tree structure. We assign the k-th symbols of the cases to be a node in the k-th layer of the index tree. The root node is obviously the universal starting symbol, all of the leaf nodes are the universal ending symbol and each symbol takes the previous symbol in its case as its parent node in the index tree. Merging starts at the root and is implemented downward layer by layer. In each layer of the index tree, except for the leaf nodes, the nodes that share the same symbols and parent nodes are merged into a single node. Every branch from the root node to a leaf node in the tree corresponds to a case, and each leaf node is thus linked with the statistics of the corresponding case. The syllable-level case base is also indexed according to the aforementioned procedure.

4. KEYWORD DETECTION

4.1. Elastic matching

In our approach, pronunciation variations are covered by involving a number of cases that refer to the same pronunciation. However, when matching a specific case, a flexible strategy must be implemented to overcome the effects of different speaking speeds. Firstly, two operations are defined for symbol sequences. Deleting the second symbol from a pair of identical symbols is defined as "shortening", and replacing a symbol with a pair of the symbols is defined as "stretching". If a sequence is matched with another sequence after a set of stretching and shortening operations, we refer to the matching as elastic matching, meaning that although their durations may differ, the categories and orders of the included symbols are identical. For example, sequence "aab" and "abb" can be viewed as matched in elastic matching, since they are identical after a shortening and a stretching operations.

If utterance $s_b, \dots, s_e$ and word-level case c_i are matched, the probabilities that the utterance is keyword kw_q can be estimated as follows:

$$P(kw_q | s_b, \dots, s_e) = P(kw_q | c_i) P(c_i | s_b, \dots, s_e), \quad (6)$$

where $P(kw_q | c_i)$ can be estimated by the statistics of c_i

$$P(kw_q | c_i) = \frac{N^i_{kw_q}}{\sum_{k=1}^{Q} N^i_{kw_k} + N^i_O}, \quad (7)$$

and

$$P(c_i | s_b, \dots, s_e) = \alpha^{N_e}, \quad (8)$$

where N_e is the number of elastic operations on matching of $s_b, \dots, s_e$ and c_i. α is a constant, set to 0.96 in our approach. Eq. (8) shows that the more elastic operations, the smaller the probability that the utterance is c_i.

4.2. Keyword detection based on the word-level case base

For the test utterance $X = \{s_1, s_2, \dots, s_T\}$ and the word-level case base C_w, the detection procedure is described as follows.

Step 1: scores $L(b, e|C_w)$ and their corresponding keyword hypothesis $H(b, e|C_w)$ are calculated for all of possible starting time b and ending time e in X.

$$L(b, e|C_w) = \max_{1 \le q \le Q} \max_{c \in C_w} P(kw_q | c) P(c | s_b, \dots, s_e), \quad (9)$$

and

$$H(b, e|C_w) = \operatorname*{argmax}_{1 \le q \le Q} \max_{c \in C_w} P(kw_q | c) P(c | s_b, \dots, s_e). \quad (10)$$

Step 2: the scores $L(b, e|C_w)$, $1 \le b < e \le T$ are compared with a threshold. The hypotheses whose scores are lower than the threshold are deleted.

Step 3: if the overlap between two hypotheses is more than one-third of the length of either of them, the one with lower score is deleted. Finally, the remaining hypotheses are reported as spotting results.

4.3. Keyword detection based on the syllable-level case base

A similar procedure is implemented to detect syllable hypotheses. Subsequently, keywords are detected by searching the syllable hypotheses. A rule about the orders and spans of the syllable hypotheses is designed here to decide whether a keyword exists.

4.4. Fast case searching algorithm

Because only the matched cases need to be considered in Eq.(9), a breadth-first search algorithm is adopted to search the index tree and find all of matched cases for a specific starting time b. Some special considerations of the algorithm are described here.

(1) Only when the searching node s_P in the index tree is matched with the current symbol s_t of the utterance, its children nodes need to be searched to match the next symbol in the utterance. There are three kinds of "matched": they are identical, they are different but s_t is identical with the parent node of s_P, and they are different but s_P is identical with the parent node of itself.

(2) If a leaf node is matched, then a matched case is found.

(3) The number of elastic operations is counted and reported for each matched case. The counts are also used to prune the searching paths.

5. FEEDBACK PROCESSING

Three kinds of feedback are processed as follows.

- *There is a correct spotting*. Adjust the statistics of the word- or syllable-level cases that comprise the spotting result by increasing the numbers of the correct keyword or syllables by one, so that their probability in the cases can be promoted.
- *There is a false alarm*. Adjust the statistics of the cases comprising the spotting result by increasing the number of other pronunciations by one, so that the probability of the correct keyword or syllables in the cases can be reduced.
- *There is a false rejection*. Search the matched cases for rejected utterance in the case base. If there is a matched case, increase the number of the keyword by one. If that case cannot be found in the case base, append a new case and merge it into the index tree.

6. EXPERMENTS

Two corpora were used in our experiments. Corpus A, which was taken from "Microsoft Mandarin Speech Toolbox" [7], contained read speeches merely from male speakers. The training set included about 20,000 sentences read by 100 male speakers, and the testing set included 500 sentences read by 25 male speakers. Corpus B, which was recorded from an Internet broadcast of a China Central Television talk show, contained actual spontaneous utterance about 40 hours in duration given by male and female speakers, some of whom had strong dialect accents. We selected 20 two-syllable words as the keywords.

The first experiment was conducted to evaluate our approach on Corpus A. For comparison, a syllable lattice based keyword spotting method was also evaluated on the same Corpus. In the approach, a Chinese syllable recognizer, also taken from "Microsoft Mandarin Speech Toolbox" [7], was used to output the syllable lattice, and the keywords were then spotted by searching the hypotheses in the lattices and calculating their posterior probabilities. In our approach, 39-dimensional MFCCs were taken as the feature vectors. The training set of corpus A was used to generate the case bases. We then searched the keywords in the testing set of corpus A. Both approaches were adjusted to their best performances. The figure of merit (FOM) values of the two approaches were 63% and 71% respectively. Our CBR approach was not as effective as the syllable lattice based approach due to its lower detection rate. However, a detection rate of 55% was achieved under the condition of a very low false alarm (1.3 FA/H/W). The character of our approach was sometimes very useful for monitoring type applications.

In the second experiment, we trained on Corpus A and tested two approaches on corpus B. Due to the severe mismatch arising from the speakers' genders, speaking styles, dialect accents and environments, both approaches performed very poorly. We tested the sustained learning ability by gradually increasing the proportion of user feedback. For comparison, the feedback on correct spotting and false rejection were also used to modify the HMM model according to a traditional MAP adaptation method. Thus, the performances of both approaches changed with the proportion of user feedback. The results are plotted in Fig. 1.

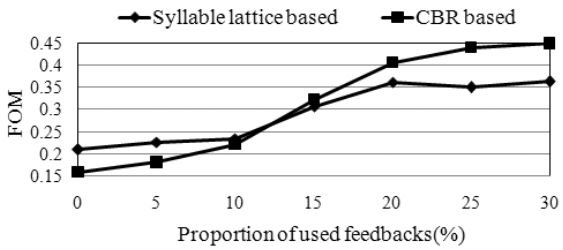

Fig. 1 Comparison of a syllable lattice based approach and our CBR approach on sustained learning ability

Our CBR approach worked better in this situation than the syllable-lattice-based approach, again mainly due to the lower false alarm rate. Moreover, in this experiment, the feedback was batch processed for the MAP method, which would be difficult to implement in a real application.

7. CONCLUSION

In this paper, we propose a CBR based Chinese keyword spotting method as an approach to sustained learning. This method can use the feedback offered by users or service providers to accumulatively learn new knowledge of pronunciation variations. Although the CBR based approach is slightly less effective than the HMM based approach when test and training condition is matched, a very low false alarm can be obtained along with an acceptable detection rate. When test and training condition is mismatched, our approach can consistently improves performance by constant feedback processing.

8. RELATION TO PRIOR WORK

At present, Chinese keyword spotting researches mainly focus on HMM based approaches. While much progress has been made, one obstacle remains in the way of worldwide practical application. It is impossible to train HMMs to satisfy all applications, and to train one specific model for each application. Developing a technique that can evolve through the user-driven sustained learning mechanism may be the solution. Based on this consideration, we propose a new Chinese keyword spotting framework with a sustained learning ability.

9. ACKNOWLEDGEMENT

This work is supported by the National Nature Science Foundation of China (91120303).

10. REFERENCES

[1] R.C. Rose and D.B. Paul, "A Hidden Markov Model based Keyword Recognition System," in Proceedings of ICASSP, Vol. 1, pp. 129-132, 1990.

[2] S. Zhang, Z. Shuang, Q. Shi and Y. Qin, “Improved Mandarin Keyword Spotting using Confusion Garbage Model,” in Proceedings of ICPR, pp. 3700-3703, 2010.

[3] J. Garofolo, G. Auzanne and E. Voorhees, “The TREC Spoken Document Retrieval Track: A Success Story,” in Proceedings of TREC-9, 2000.

[4] T. Mertens and D. Schneider, “Efficient Subword Lattice Retrieval for German Spoken Term Detection,” in Proceedings of ICASSP, pp. 4885–4888, 2009

[5] A. Norouzian and R. Rose, "Facilitating Open Vocabulary Spoken Term Detection Using A Multiple Pass Hybrid Search Algorithm," in Proceedings of ICASSP, pp. 5169 - 5172, 2012

[6] A. Aamodt and E. Plaza, “Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches,” AI Communications, Vol. 7:1, pp. 39-59, 1994.

[7] E. Chang, Y. Shi, J.L. Zhou and C. Huang, “Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research,” in Proceedings of the 7th European Conference on Speech Communication and Technology, pp. 2779-2782, 2001.