Topic Mining based on Word Posterior Probability in Spoken Document

Lei Zhang, Guo-xing Chen, Xue-zhi Xiang, Jing-xin Chang
Information and Communication Engineering College, Harbin Engineering University, Harbin, China
Email: { zhanglei , xiangxuezhi}@hrbeu.edu.cn

***Abstract*—For speech recognition system, there are three kinds of result representations as one-best, *N*-best and Lattice. Since lattice has multi-path which can reduce the effect of recognition error rate, it is widely applied nowadays. In fact, there are amount of redundancies in lattice, which leads to the increasing of complexity of latter algorithm based on it. Additionally, for the decoding algorithm, it is acted as maximum a posterior probability (MAP) which can only guarantee the posterior probability of the whole sentence is of maximum. For MAP does not mean the highest syllable recognition rate, here, confusion network is introduced in topic mining system. In the clustering during confusion network, the minimum word error rule is adopted, which is proper to topic mining system since the least meaningful unit is word in Chinese and word information is most important in topic mining.**

In this paper, a simplified confusion network generation algorithm is proposed to handle some problems caused by insertion error during recognition. Then based on the confusion network, a word list extraction approach is proposed, in which, the dictionary is adopted to judge whether the consecutive arc in confusion sets is a word. At this stage, the error word information produced by error recognition rate can be corrected to some extent. After the competition part in word list extraction on confusion network, a final word list with posterior probability can be obtained. Furthermore, this kind of posterior probability can be combined in topic mining system. SVD and NMF are adopted here to decompose the term-document matrix on the word list of confusion network. From the experiments, it can be drawn that the proposed approach based on confusion network can achieve better performance than that of one-best and *N*-best. Additionally, the modified weight which combined posterior probability into term-document matrix can further improve the system performance.

***Index Terms*—topic mining, spoken document, posterior probability, confusion network, modified weight**

I. Introduction

Vector space model [1] and latent semantic indexing (LSI) or latent semantic analysis (LSA) technology [2] are widely applied into text documents classification and text information mining. In LSA or LSI, it is operated under the assumption that there is some underlying latent semantic structure in the data. It can map the words and documents onto a continuous parameter space, named term-document matrix. During the analysis procedure, it can also reduce the dimension of matrix and mine the semantic meaning in the structure. Since the successful applying in text data mining, these kinds of approaches are also adopted in speech signal processing, such as spoken document indexing and spoken document retrieval [3-9]. For text documents, the term-document matrix can be easy constructed by many approaches as TF-IDF, MPP and others [10,11]. When LSA or LSI is combined with spoken document processing, the main problem is how to represent the term-document matrix based on speech recognition result. Some approaches are based on 1-best recognition result and lattice which has multi-path candidates[12].

Since the accurate rate of speech recognition system is still far away from real application, the research about spoken document is also hot. In order to reduce the effect of low recognition rate, the retrieval and classification of spoken document are mainly based on lattice [13,14]. Although lattice can avoid the effect of low recognition rate to some extent by multi-path, the storage space of lattice is large. And this kind of structure contains a lot of redundancies, which make the algorithm based on it complex and ineffective. Furthermore, the rule of decoding algorithm on traditional lattice is based on maximum a posterior (MAP) probability, which can only guarantee the error rate of the final sentence is of minimum.

In Chinese, the meaningful semantic unit is word, especially for topic mining system. Word is composed by at least two syllables in Chinese. Some study [15] shows that, although there are some relations between sentence error rate and syllable error rate, the maximum of posterior probability of whole sentence does not mean the minimum of syllable error rate. Although for maximum posterior probability decoding in lattice, it can get the lowest sentence error rate, to obtain words information, it still need Chinese segmentation. Under most condition, the lowest sentence error rate can not guarantee the best segmentation results. Like ‘ren2min2fa3yuan4’, which is a four-syllable word, if the second syllable ‘min2’ is misrecognized as ‘min4’ or other syllable, this four-syllable word can not be correctly segmented in Chinese segmentation. Then the most useful information can be missed only by one syllable’s error recognition.

Confusion network is proposed by Mangu [16]. It can carry out a practical approximate syllable error minimization on lattice. It will find alignment of all syllables in the lattice, and form the new sentence by concatenating the syllable with maximal posterior

doi:10.4304/jsw.6.11.2292-2299

probability from different non-overlapping classes of syllable in the lattice. There are also many researches on confusion network, as some simplified approaches [17] and applying it in speech recognition [18,19] and speech transformation [20,21]. Based on the analysis above, confusion network is also proper to spoken document classification and topic mining.

Based on syllable lattice, in this paper, a simplified algorithm to generate syllable confusion network is proposed. Furthermore, the dictionary instead of language model is applied to extract word information from syllable confusion network. A final word list with posterior probability can be obtained in the new proposed word extraction approach. This kind of posterior probability of word after normalization can express the information similar with the word frequency, which is called modified tf-idf weight in term-document matrix, so it can be combined into the topic mining system based on latent semantic indexing.

The follow sections are organized as: section II gives the simplified generation algorithm of confusion network and in section III, a new approach to extract the word information based on confusion network is presented. Furthermore, the latent semantic analysis combined with word posterior probability is proposed in section IV. At last, the experiments and results are listed in section V, and section VI gives the conclusion.

II. SIMLPIFIED GENERATION ALGORITHM OF CONFUSION NETWORK

Since there exist more than 80,000 commonly used words and more than 10,000 commonly used characters in Mandarin Chinese, it is hard to construct the recognition model based on words or characters. Furthermore, all characters are monosyllabic, and for Chinese, there are many homophones, then the total number of phonologically allowed syllables with tone is only 1345 [22]. So in our system, the recognition model is built on syllable. Combined with language model smoothed by modified Katz, the system can output the syllable lattice instead of syllable sequence known as 1-best result.

The lattice based on syllable is a directed acyclic graph as $G=(V,E)$. V is the node set in G, and each node corresponds to a time moment uniquely. E is the arc set. Here, the label of lattice is on arc, and then each arc has the attribute as syllable label, start node number, end node number, acoustic score and language score. In the lattice, the amount of arc is large, and in the similar start moment, there are a lot of similar arc. All of these increase the complexity and redundancy of lattice. In confusion network, the optimization can be conducted during clustering the arcs. In this optimization, two aspects must be considered, the one is the rule of clustering of arcs, and the other one is whether the partial order between arcs is still unchanged after clustering.

In [16], the clustering procedure can be divided into three stages. In the initial stage, the arcs starting and ending the same time and with the identity syllable label can be combined into an equivalence class. Thus the arcs in lattice can be divided into many classes which need to be combined further. In the second step, the equivalence classes in step 1 can be further combined for those with not only the same label, but also with no partial order. The order of combination starts from equivalence class with the least distance, where the distance depends on the overlap of the arcs and the posterior probability. In the final stage, the combination is mainly for the classes with different label. The whole combination procedure is similar with the second step, except the distance is computed based on the phonetic similarity. After all the steps of combination, the equivalence class can be named as confusion set. And the final whole network is turned into a few of concatenating confusion sets.

It can be seen that during the equivalence class generation, the judge of partial order is important. In [16], as long as there are arcs with partial order which belong to different class, then the classes preserve the partial order. This kind of judge is conducted for each arc in the equivalence, which increase the computation burden. Additionally, this kind of handling can not deal with the error caused by insertion. In fig. 1, if there is kind of insertion error in lattice like the dashed ellipse shown, then arc labeled with ‘ji2’ and arc with ‘ji4’ have the partial order. This can result that whatever how long the arc with ‘ji4’ has the overlap with other arc with ‘ji2’ and how similar ‘ji4’ is with ‘ji2’ based on phonetic similarity, ‘ji4’ can not combined with the same class with ‘ji2’.

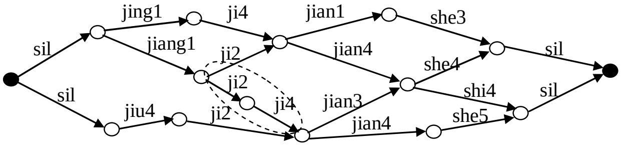

Fig. 1 Sketch of insertion error in lattice

In [16], pruning is conducted on lattice to handle this kind of question, in which, the insertion error is expected to be dropped. In fact, the width of pruning is hard to control under different conditions. To avoid this problem, another method of judging the partial order is proposed here.

It needs to define the start time and end time of the equivalence class. Here, the start and end time of the equivalence class after clustering is defined as the minimum start time and the maximum ending time of all arcs in it. Then the partial order can be judged by the overlap between the equivalence classes. If there is overlap between the equivalence classes, then there is no partial order between these two equivalence classes.

For example, like in figure 1, in the second step, ‘ji4’ and ‘ji2’ can form two different equivalence classes. Since these two equivalence classes have the overlap, which represents that there is no partial order between these two sets, so these two can be combined further based on the edit distance in the final stage.

Confusion network generated by above approach is with faster computation speed. At the same time, it can also effectively reduce the number of nodes and arcs in lattice. Table I gives the statistical result about 8703 spoken document of conversation in six categories. It can be seen that confusion network can spare the storage space, and can avoid the waste of computing time of further processing for smaller arc number.

TABLE I.
COMPARE OF COMPLEXITY BETWEEN LATTICE AND CONFUSION NETWORK

	Average number of nodes	Average number of arcs	Average storage space
Lattice	481	1687	99.07KB
Confusion network	18	189	7.24KB

Fig. 2 gives the recognition system performance of lattice and confusion network. In the decoding algorithm based on lattice, an improved Katz approach is applied. And in the confusion network, it only selects the maximum probability arc in each equivalence class as the final results, which is called consensus recognition rate based on confusion network.

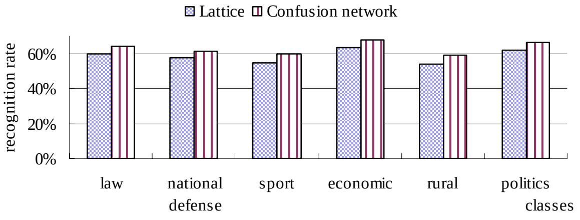

Fig. 2 Tonal syllable recognition rate based on lattice and confusion network for six categories

From fig. 2, it can be seen that the decoding performance of confusion network is better than that of lattice for each category. The main reason is that for confusion network, the equivalence class is constructed as minimum syllable error rate, which is identical with that in fig. 2. Contrarily, in the decoding of lattice, the basic rule is based on the maximum posterior probability of sentences.

III. EXTRACTION OF WORD INFORMATION ON CONFUSION NETWORK

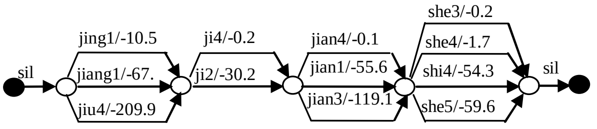

Fig. 3 Sketch of on confusion network

Syllable confusion network is shown as fig.3. There are four confusion sets in this figure. In every confusion set, there are some arcs candidates with posterior probability, start time and end time. As shown in fig. 3, consensus recognition rate is to select the arc with largest posterior probability. Since here, the posterior probability is in log domain, the result of decoding on confusion network as in fig. 3 is 'jing1ji4jian4she3'. The correct answer is 'jing1ji4jian4she4', which means the construction of economic in English. Whether decoding on lattice or confusion network, it can not extract this kind of information exactly. But from fig. 3, the correct syllable 'she4' is in the last confusion set, and it is the second candidate. It means that if adopting proper approach, it is hoped to dig this correct information.

Here, in order to correct this kind of error, the consensus recognition result is not used in topic mining directly. Instead, dictionary is applied to get the accurate word information based on confusion network. Since the word information is important to topic mining or spoken document classification, for the confusion network, it is aimed to dig more information of word instead of to get the best recognition result.

The whole procedure of this part is as fig. 4. If confusion network is input, then after pre-processing, word information extraction and competition parts, the final word list will be output.

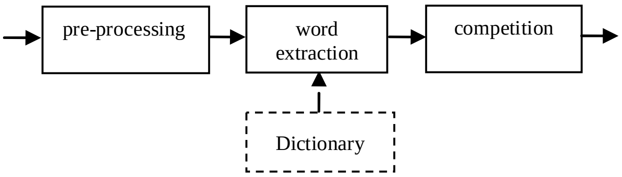

Fig. 4 Word information extraction based on confusion network

A. pre-processing

During the generation of lattice and confusion network, there is a factor to influence the final result, which is the pruning threshold. Fig. 2 gives the results of proper pruning threshold after adjusting. Because we hope to correct the error recognition results in this stage, the recall rate must be guaranteed in speech recognition. That is, we hope to keep more information during speech recognition. Thus in following procedure, the pruning threshold can be decreased so that more arcs are kept. In the analysis experiments, it is found that with the information of confusion network increased, there are also many insert errors. In the confusion network, there are some confusion sets with only one arc. At the same time, the duration of this kind of confusion set is very short. Additionally, for the unique arc in the confusion set, the posterior probability is very small.

From the analysis of 1000 confusion sets selected by random, it can be drawn that if the log posterior probability is less than -3.0 and the duration is less than 0.1 second, meanwhile, there is only one arc in the confusion set, this confusion set is produced by insert error which can be dropped. So in pre-processing, such confusion sets are filtered.

B. word extraction based on dictionary

After the pre-processing, the number of confusion sets can be reduced to some extent. Then the word information can be extracted based on dictionary. Here, the dictionary is generated by text corpus of the similar categories according to some approaches.

For mining all the possibilities of word information from the confusion network, each arc in the confusion set and all the possible combination of the following arcs must be considered.

Take four-syllable-word for example, supposing the average number of arcs in confusion set is 10, then traversing four confusion sets needs 10^4 times to consult the dictionary to determine whether it is a word. It is a huge computation. In order to avoid this kind of computation, another approach as follow is proposed.

Step 1: Traversing all arcs in the first two confusion sets, and look up the dictionary to determine whether it is a two-syllable word. All possible two-syllable words are extracted to compose a word set *A*.

Step 2: Look up the dictionary, finding all words started by the two-syllable word in set *A*. Then word set *B* is generated which including two-syllable words in word set *A*, three-syllable words and four-syllable words which are started by the two-syllable words in *A*.

Step 3: For three-syllable words in *B*, determine if the third syllable happens in the next confusion set connected to the first two. Similarly, for four-syllable words in *B*, except the third syllable happen in the third confusion set, it also needs to confirm that if the forth syllable is in the forth confusion set. Additionally, in Chinese, if it can find three-syllable words or four-syllable words in confusion network, it will happen with high probability. In order to ensure the quality of three-syllable words or four-syllable words, here, another limit is added that at least one syllable in three-syllable words or four-syllable words is the arc with the largest posterior probability in corresponding confusion set.

Step 4: Shift the confusion set, repeat step 1 until all confusion set is decided.

Based on the four steps, all latent words in confusion network are found directly with no need of Chinese segmentation. Although in the final word list, the extracted words may be more than the words by traditional approach, it can correct the error of recognition and dig all the information in confusion network to some extent.

Fig. 5 gives the part result of one confusion network. In fig. 5, the first column is the posterior probability of words which is sorted by descending order, and the second one represents the words. The number in last column is the mark of confusion set. For an example, the first line is 'hen2nan2ren2', and the following numbers mean that 'hen2' is from the second confusion set, and 'nan2' is from the third one and so on.

-0.000001	he2nan2ren2	2+3+4
-0.000001	san1shi2si4	5+6+7
-0.000001	you2guan4che1	11+12+13
-0.023803	shen2me5	1+2
-0.107507	guan4che4	12+13
-0.272582	san1shi2	5+6
-0.276518	zhe4yang4	9+10
-0.282238	shi2si4	6+7
-0.442879	si4qian1	7+8
-0.481976	zhe4zhong3	9+10
-0.483192	nan2ren2	3+4

Fig. 5 Example of word list based on confusion network

For the three-syllable and four-syllable words, since there are additional limit during extraction, they will appear in real condition with high probability. So the posterior probability of these words can be fixed as a larger value. In this system, it is -0.000001. As for two-syllable word *i* in spoken document *j*, the posterior probability is computed as

$$P_{ij} = \log P(w_1 w_2) = \lambda[\log P_{ac}(w_1) + \log P_{ac}(w_2)] + (1-\lambda)[\log P_{lm}(w_1) + \log P_{lm}(w_2)] \quad (1)$$

Where *ac* means acoustic probability, and *lm* is the language probability in confusion network. λ is the adjusting parameter, which can determine the effect of acoustic and language model. In (1), all the posterior probabilities are normalized.

In fig. 5, since every confusion set has the accurate start and end time, the mark of confusion set in last column can describe the time information. There are many extra words in the word list, and some of them may happen at the same time. Like 'zhe4yang4' and 'zhe4zhong3' in fig. 5, they are both from the ninth and tenth confusion sets. That means they compete the same moment in the word list. So after extraction of word information, there is also a competition part next.

C. Competition

In this section, it is aimed to filter the word with less competition ability.

Step 1: Select the word with the highest posterior probability as the basic word. And add this word in the basic word list.

Step 2: Get other words according to the posterior probability by descending order. If the word has no overlap with the words in basic word list, then add the word in basic word list. Otherwise, determine whether the syllables of overlap are the same or not. If they are the same, add this word in basic word list, otherwise, delete this word.

Step 3: Repeat step 2 until all the words in the word list have been processed.

After all the processing of section *A*, *B* and *C*, the final word list can be generated. In this word list, at any moment, there is only one syllable happen. That is, for each confusion set, only one syllable is kept. The kept syllable may not be the arc with highest posterior probability. It is only partly dependent on the posterior probability, and is also influenced by the dictionary. Apart from the word list, there is also corresponding posterior probability for each word.

IV. Latent Semantic Space based on Word Posterior Probability

A. Modified Weight in Vector Space

Supposing each term in document is independent, then the document can be expressed as a vector in vector space. It is the main idea of vector space model (VSM). If using $w(i, j)$ as the weight of term *i* in document *j*, in most text retrieval or mining system, it is represented as

$$w(i, j) = TF \times IDF = \frac{t_{ij}}{n_j} \times \log(\frac{N}{N_i} + L) \quad (2)$$

Where t_{ij} is the number of term *i* in document *j*, and n_j is the whole number of words in document *j*. N is the whole number of documents, and N_i is the number of document that term *i* occurring in it. In (2), *L* is a fix value to avoid the log function overflow, and here it is fixed as 0.01.

For traditional tf-idf (term frequency-inverse document frequency), it works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus, which is similar with (2).

This kind of feature can represent the distinguishing ability of term *i*. If term frequency of *i* in one topic is high, instead, it is low for other topics, then term *i* has the high relevance with the topic.

Furthermore, in (2), since if the term occurs in document one time, it will add 1 to t_{ij}. It is easy to compute t_{ij} for text documents. But for spoken document, in fact, every term happens in some document at a certain posterior probability. For example, in fig. 5, ‘guan4che4’ and ‘nan2ren2’ both happen in the same document. If t_{ij} is computed as (2), then although the posterior probabilities of these two words are different, they play the same role in t_{ij}. This is unreasonable. Since the posterior probability can reflect the acoustic difference to some extent, this kind of difference should be acted in the weight in term-document matrix.

In order to reflect the difference of probability, (2) now is turned into

$$\hat{w}(i,j) = \frac{P_{ij}}{\sum_i P_{ij}} \times \log(\frac{N}{N_i} + L) \tag{3}$$

The weight in (3) is called modified tf-idf. Since term *i* may happen more than one time, where P_{ij} is the accumulation of posterior probability of term *i* in document *j*. Then $\sum_i P_{ij}$ is the whole posterior probability of all words in document *j*. The first term in (3) can express the ratio of word *i* in the whole corresponding document.

B. Latent Semantic Analysis

For the whole word number of all documents are *N*, the term-document matrix is as $M \times N$. And each document can be expressed as a vector in the matrix. Since the number of words in each document is limited, and the whole word number *N* maybe huge, the term-document matrix in VSM is a high-dimensional and sparse one. It is sensitive to noise. Additionally, term-document matrix with high dimension is presumed too large for the computing resources.

In order to handle these problems, a linear function is adopted that can map the vector of document in high dimension space into a space with low dimension. During the mapping, it can also be supposed that the similar words may be mapped into the same dimension. That means, the synonym and polysemy problem in Chinese can be handled to some extent.

Two kinds of mapping function are adopted in our system. The one is singular value decomposition (SVD) [23], and the other one is Non-negative Matrix Factorization (NMF)[24].

Let $\boldsymbol{X}_{M \times N}$ as the term-document matrix, then for SVD, it is decomposed as

$$\boldsymbol{X}_{MN} = \boldsymbol{U}_{MM} \boldsymbol{S}_{MN} \boldsymbol{V}_{NN}{}^T \tag{4}$$

Where $\boldsymbol{U}$ is the left singular matrix with row vector u_i. $\boldsymbol{V}$ is the right singular matrix with row vector $\boldsymbol{v}_j$. As is well known, $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$.

S is the diagonal matrix with non-negative real numbers on the diagonal. A common convention is to order the diagonal entries in descending order. Then in **S**, only those numbers with high values are kept, and the parts with small number can be dropped off. Then (4) is approximated as

$$\hat{\boldsymbol{X}}_{MN} \approx \boldsymbol{U}_{MR} \boldsymbol{S}_{RR} \boldsymbol{V}_{RN}{}^T \tag{5}$$

Where *R* is the order of decomposition, which is less than the rank of ***S***.

After the mapping, it can be interpreted as that $\hat{\boldsymbol{X}}$ captures the major structural associations in ***X*** and ignores higher-order effects.

For different topics, let $\boldsymbol{U}^t$, $\boldsymbol{V}^t$ and $\boldsymbol{S}^t$ are the left and right singular matrix corresponding with topic *t*. These three matrixes can be obtained by training procedure. Furthermore, a topic center vector $\boldsymbol{c}^t$ can be obtained easily. Supposing these matrixes are known for different topics, then the test document $\boldsymbol{q}$, which is needed to mine the topic information, can be mapped as follow.

$$\hat{\boldsymbol{q}} = \boldsymbol{q}^T \boldsymbol{U}^t \boldsymbol{S}^{t^{-1}} \tag{6}$$

And the distance between test document after mapping $\hat{\boldsymbol{q}}$ and the topic center vector $\boldsymbol{c}^t$ can represent how possible the test document coming from this topic *t*. Among all topics, the one with minimum distance is the final result.

Another kind of decomposition approach is NMF. NMF is a new matrix decomposition algorithm based on semantic in the recent years, with which there is a limit that in the decomposed matrix, all elements are all non-negative. This kind of limit makes it easy to explain the meaning of decomposition.

$$\boldsymbol{X}_{MN} \approx \boldsymbol{B}_{MR} \boldsymbol{H}_{RN} \tag{7}$$

Similar with SVD, *R* is selected to be smaller that *M* or *N*. Where matrix ***B*** is called left matrix and ***H*** is the right matrix.

Because the original and decomposed matrices only contain non-negative elements, each column vectors of the original matrix is approximated by a linear combination of the columns of the left matrix ***B***, weighted by the components of the columns in right matrix ***H***. Therefore ***B*** can be regarded as a basic approximation of data in ***X***. That is, ***B*** is the latent space that can reveal the essence of original space ***X***.

B and ***H*** can be trained by different rules. Here, the Kullback-Leibler divergence is adopted as the approximation rule. That is

$$D(\boldsymbol{X} \| \boldsymbol{Y}) = \sum_i \sum_j [x_{ij} \log(\frac{x_{ij}}{y_{ij}}) - x_{ij} + y_{ij}] \tag{8}$$

Where $\boldsymbol{Y} = \boldsymbol{BH}$, and y_{ij} is the corresponding element in matrix ***Y***. Additionally, $\sum_{ij} x_{ij} = \sum_{ij} y_{ij} = 1$, that means for matrix ***X*** and ***Y***, it is normalized matrix.

Once ***B*** and ***H*** are obtained during training procedure, the new test document $\boldsymbol{q}$ can be projected into the basic space represented by ***B***. that is

$$\hat{\boldsymbol{q}} = \boldsymbol{B}^T \boldsymbol{q} \tag{9}$$

V. Experiments and Resultes

A. Corpuses in experiments

There are three corpuses here. The first one is to train the basic recognition system called basic speech corpus. It is from 863 corpus which is reading style pronunciation and consists of 90,821 utterances from 156 speakers.

Another one is for training language model, which is the text corpus. It is from Renmin newspaper, which has about 300M storage. And the language model is trained by modified Katz smoothing approach.

The third corpus is for topic mining, which is recording of radio programs, especially for conversation and talking with little restriction. It includes 8703 utterances here, and some of them which can not define the topic are used as adaptation data. MLLR and MAP adaptation approaches are applied here to get better performance of speech recognition system for radio data. For the other data, they are manually labeled and pre-defined for different topic. Here, six topics as national defense, sport, countryside, law, economy and politics. It includes 5924 utterances from different kinds of broadcast programs.

B. Evaluation approach

Precision rate (*P*) and recall rate (*R*) have been used regularly to measure the performance of information retrieval and information extraction system. Precision deals with substitution and insertion errors while recall deals with substitution and deletion errors. These two rates can be combined together into *F*-measure. It is a weighted combination of *P* and *R* as follow. Here, the weight is equal to 0.5 for both *P* and *R*.

$$F = \frac{2PR}{P + R} \tag{10}$$

C. N-best speech recognition result processing

For *N*-best results, there are top *N* results with maximum a posterior probability. Normally, the top 1 is the result of one-best. Here, letting W_1 and W_2 range over the *N*-best hypothesis output by speech recognizer, then the real result W_c called center hypothesis is found as follow.

$$W_c = \arg\min_{i=1,N} \sum_{k=1}^{N} P(W_1^{(k)} \mid A) WE(W_2^{(i)}, W_1^{(k)}) \tag{11}$$

Here, $WE(\cdot,\cdot)$ means the edit distance of two different strings.

D. Topic mining experiments

The first experiment is conducted to compare the system performance with traditional speech recognition results as one-best and *N*-best with the result of proposed approach, and here the weight in term-document matrix is as (2). In fig.6, CN represents the proposed approach, in which the word list is extracted based on confusion network.

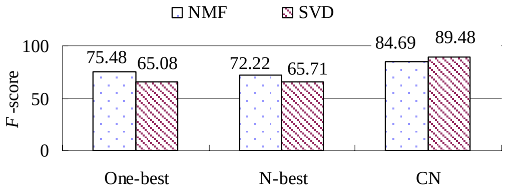

Fig. 6 performance comparison of three methods

From fig. 6, it can be seen that for *F*-score, the performance of proposed approach are better than one-best and N-best results no matter for NMF and SVD. For NMF, the improvements are 9.21 and 12.47 for those of one-best and N-best, and for SVD, they are 24.4 and 23.77 respectively. Form another aspect, for one-best and N-best, the topic mining results of NMF are better than those of SVD. This is reasonable that NMF can construct the basic space which is more meaningful that that of SVD. But during the experiment, it can be seen that for proposed approach, the performance of SVD is a little better than that of NMF. The reason may partly be that for the word list obtained on confusion network will be larger that the real results of words in one-best and N-best. That is, the original space is some different from that of one-best and N-best. The effect of this kind of difference on NMF and SVD need to be deeply analysis in future work.

Another experiment is to verify the influence of modified weight in term-document matrix. Combined the posterior probability into (2), then the weight here is selected as (3). Fig. 7 gives the comparison of these two kinds of weights.

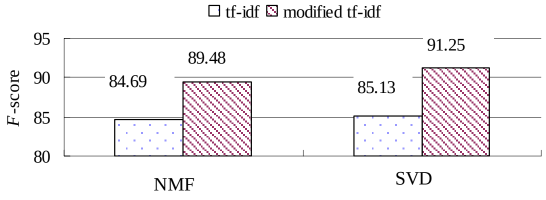

Fig. 7 Comparison of tf-idf and modified tf-idf

From fig. 7, it can be drawn that for (3), since it combines the posterior probability into weight, it can achieve better performance than the weight of (2) both for NMF and SVD. Additionally, the best topic mining results can be 91.25 for *F*-score by modified tf-idf and SVD.

VI. Conclusion

Since the decoding rule of syllable confusion network is based on minimum syllable error, which is better than maximum a posterior probability for extracting word information, it is applied in topic mining system. A new approach to extract the word list on confusion network is proposed here. At the extraction procedure, some kinds of word error caused by speech error recognition can be corrected to some extent. Furthermore, the final word list with posterior probability can be combined in term-document matrix. A new weight in this matrix is proposed to make use of the posterior probability, which include some knowledge about acoustic information. A set of experiments are designed to verify the performance of the proposed approach, and results show that the word extraction approach based on confusion network can effectively improve the system performance. Additionally, combined the posterior probability into weight of term-document matrix, the performance can be enhanced further.

In the decomposition of term-document matrix, SVD and NMF present different results for one-best, N-best and the proposed approach. It may be partly caused by the enlarging word list in proposed approach. The more details about this kind of effect need to be discussed in the future.

Acknowledgment

It is a project supported by National Natural Science Foundation of China #60702053, also supported by young teacher supporting plan by Heilongjiang province in China #1155G17 and young teacher supporting plan by Harbin Engineering University.

References

[1] Salton G Wong A, Yang C, “A Vector space model for automatic indexing”. *Communications of ACM*, vol. 18(11), pp. 613-620, 1975.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis”, *Journal of the American Society of Information Science*, 41(6): pp. 391-407, 1990

[3] Lin-shan Lee, Chen B. “Spoken document understanding and organization”, *IEEE Signal Processing Magazine*, vol. 22, issue 5, pp. 42-60, 2005.

[4] Ya-chao Hsieh, Yu-tsun Huang, Chien-chih Wang, Lin-shan Lee, “Improved spoken document retrieval with dynamic key term lexicon and probabilistic latent semantic analysis (PLSA)”, in Proceeding ICASSP, vol. I, pp. 961-964, 2006.

[5] Berlin Chen, Yi-ting Chen, “Extractive spoken document summarization for information retrieval”, *Pattern Recognition Letter*, vol. 29, Issue 4, pp. 426-437, March, 2008.

[6] Pere R. Comas and Jordi Turmo, “Spoken document retrieval based on approximated sequence alignment”, *Lecture Notes in Computer Science*, vol.5246, pp. 285-292,2008.

[7] Chung-Hsien Wu, Chia-Hsin Hsieh, Chien-Lin Huang, “Speech sentence compression based on speech segment extraction and concatenation”, *IEEE Transactions on Multimedia*, vol. 9 Issue 2, pp. 434-438, 2007.

[8] Xinhui Hu, Ryosuke Isotani, Satoshi Nakamura, “Spoken document retrieval using topic models”, *ACM International Conference Proceeding series: Proceeding of the 3rd International Universal Communication Symposium*, vol. 338 pp. 400-403

[9] Welly Naptali, Masatoshi Tsuchiya, Seiichi Nakagawa, “Topic-dependent language model with voting on noun history”, *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 9, Issue 2, article 7.

[10] Y-M Yang, and J. O. Pedersen. “A comparative study on feature selection in text categorization.” in *Proc.ICML-14*, 1997. pp.12-420

[11] Timothy, J. Hazen, Fred Richardson, and Anna Margolis. “Topic identification from audio recordings using word and phone recognition lattices.” In *ASRU* 2007, pp.659-664

[12] C-H Meng, H-Y Lee and L-S Lee, “Improved lattice-based spoken document retrieval by directly learning from the evaluation measures,” in *Proc. ICASSP*, Taipei, Taiwan 19-24 April 2009 pp.4893-4896.

[13] Zheng Tie-ran, Han Ji-qing, “Syllable lattice based Chinese speech retrieval techniques and removing redundancy method from indice”, *Acta Acoustic*, vol. 33, Issue 6, pp. 526-533, 2008.

[14] Zheng Tie-ran, Han Ji-qing, “Study on Chinese speech retrieval based on posterior probability”, *Chinese High Technology Letters*, vol. 19, Issue 2, pp. 119-124, 2009.

[15] Stolcke, A., Konig, Y. & Weintraub, M. (1997). Explicit word error minimization in N-best list rescoring. *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, vol. 1, pp. 163-166

[16] Lidia Mangu, Eric Brill, Andras Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 2000, 14: pp. 373-400

[17] Jian Xue, Yunxin Zhao, “Improved confusion network algorithm and shortest path search from word lattice”, In *ICASSP*, 2005, vol. 1, pp. 853-856.

[18] Gokhan Tur, Dilek Hakkani-Tur, Giuseppe Riccardi, “Extending boosting for call classification using word confusion networks”, In *ICASSP*, vol. 1, 437-440, 2004.

[19] Dilek Hakkani-Tur, Frederic Bechet, Giuseppe Riccardi, Gokhan Tur, “Beyond ASR 1-best: using word confusion networks in spoken language understanding”, Computer Speech and Language, vol. 20, pp. 495-514, 2006.

[20] Nicola Bertoldi, Marcello Federico, “A new decoder for spoken language translation based on confusion network”, In *ASRU* 2005, pp. 86-91.

[21] Nicola Bertoldi, Richard Zens, Marcello Federico, Wade Shen, “Efficient speech translation through confusion network decoding”, IEEE Transactions on Audio, Speech and Language Processing, vol. 16, No. 8, pp. 1696-1705, 2008.

[22] B. Chen, H.M. Wang, and L.S. Lee, “A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents,” *ACM Trans. Asian Lang. Inform. Process*., vol. 3, no. 2, pp.128-145, 2004

[23] Deerwester S,Dumais S T,Furnas G W, *et al*, “Indexing by latent semanitc analysis”. *Journal of the American Societyof infortnation Science*, vol. 41, No. 6, pp. 391-407, 1990.

[24] LEE D D , SEUNG H S, “Learning the parts of objects by non-negative matrix factorization”, *Nature*, vol. 401, No. 6755, pp. 788-791, 1999.

Lei Zhang Harbin, China, 1973. Received Master degree and Doctor in computer applying field in 2000 and 2004 in Harbin Institute of Technology, Harbin, China. Her main research fields are about spoken document classification and retrieval, robust speech recognition, speaker recognition and other fields about speech signal processing.

Currently, she is professor in Information and Communication College in Harbin Engineering University. She has published more than 20 papers in journal and international meetings, and also published a book about speech signal processing by QingHua University Press.

Dr. Zhang is the member of IEEE.

Guo-xing Chen Harbin, China, 1987. Receive Bachelor degree in signal processing in Harbin Engineering University in 2008. Currently, he is Master candidate in Harbin Engineering University.

Xue-zhi Xiang Harbin, China, 1979. Received Master degree and Doctor in computer applying field in 2004 and 2008 in Harbin Engineering University, Harbin, China. His main research field is about signal and information processing,

Currently, he is the associated professor in Information and Communication College in Harbin Engineering University.

Dr. Xiang is the member of IEEE..

Jingxin Chang Harbin, China, 1987. Receive Bachelor degree in signal processing in Harbin Engineering University in 2008. Currently, he is Master candidate in Harbin Engineering University.