A Novel Method for Spoken Text Feature Extraction in Semantic Video Retrieval*

Juan Cao[1,2], Jintao Li[1], Yongdong Zhang[1], and Sheng Tang[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences
Beijing 100080, China
[2] Graduate University of the Chinese Academy of Sciences
Beijing 100039, China
{caojuan, jtli, zhyd, ts}@ict.ac.cn

Abstract. We propose a novel method for extracting text feature from the automatic speech recognition (ASR) results in semantic video retrieval. We combine HowNet-rule-based knowledge with statistic information to build special concept lexicons, which can rapidly narrow the vocabulary and improve the retrieval precision. Furthermore, we use the term precision (TP) weighting method to analyze ASR texts. This weighting method is sensitive to the sparse but important terms in the relevant documents. Experiments show that the proposed method is effective for semantic video retrieval.

Keywords: ASR texts, relevant documents, HowNet, TREVID.

1 Introduction

Video is a rich source of information. It includes visual, acoustical and textual content. Among them, the textual information of video mainly includes the texts extracted by optical character recognition (OCR) and the transcripts obtained from automatic speech recognition (ASR). The visual features have been applied widely in the video retrieval [1,2,3], but the text features extracted from ASR didn't be deeply studied . One reason is that the ASR technique has a moderate word error rate. Another reason is that the OCR texts exist in the target shots, while the relevant ASR texts can be spoken in the vicinity near the target shots. However, these reasons can't deny the text feature's superiority to the visual feature, for a semantic gap remains between the low-level visual features that can be automatically extracted and the semantic descriptions that users desire, and the text features free from this problem. The semantic gap has became a significant stumbling block in the semantic video retrieval. Every year's results of the TREC and TRECVID display that the mean average precision(MAP) of text retrieval is far above that of the video retrieval. So taking full advantages of the

* This work was supported by Beijing Science and Technology Planning Program of China (D0106008040291), the Key Project of Beijing Natural Science Foundation (4051004), and the Key Project of International Science and Technology Cooperation (2005DFA11060).

Y. Zhuang et al. (Eds.): PCM 2006, LNCS 4261, pp. 270 – 278, 2006.

video's text resources to support the visual retrieval is imperative and feasible. This is the paper's aim.

1.1 Related Work

Information retrieval from speech recognition transcripts has received a lot of attention in recent years in the spoken document retrieval (SDR) track at TREC7, TREC 8 and TREC 9 [4]. But SDR is analyzing the texts independently to the videos, and retrieving the texts on the story level. In this paper we focus on mining the valuable information from the ASR texts as one feature to assist the semantic video retrieval, not retrieve the spoken texts.

Many TRECVID participants have done a lot of work in the video's text information analysis[1.2.3]. In the high-level feature extraction task, most of them used the basic and simple statistic methods to build the concept lexicons. These methods didn't consider the imbalance of concept's relevant documents and irrelevant documents. The calculated weights can only mean the importance degree of the terms to the documents, not the importance degree of the terms to their special concepts.

1.2 The Characters of ASR Texts

In the 2005 TRECVID's English development data, the concepts' proportions of the relevant documents to irrelevant ones are 1/1000 to 1/2, and the average proportion is 1/16. This is a problem of data imbalance. In this circumstance, our term weighting method should be more sensitive to the occurrence in relevant documents than in the irrelevant documents. It is a pity that many weighting methods can't make a difference between the relevant documents and irrelevant documents.

ASR texts are transformed from speech. Firstly the spoken language has freer rules than written language. As a previous study noted, only about 50% of the words that appear as written text in the video are also spoken in the audio track [5]. Besides, the outputs of ASR systems contain a moderate word error rate (WER). Therefore, it is important to select the few informative terms without sacrificing the query accuracy of the concept.

These characteristics raise the following two questions:

- How to resolve the problem of sparse data?
  In the experiments, the term precision model performs well.
- How to build a smaller dimension with more informative concept lexicon?

In the later experiment results, we'll found that combining the statistic and rule-based methods can get a good answer.

In the following sections, we'll describe three methods and their improved versions for processing spoken documents in video retrieval in section 2. In section 3 we'll present the experiments and results. In section 4 we'll discuss this paper's major findings. Finally we'll get the conclusion in section 5.

2 ASR Text Feature Extracting Methods

2.1 The Representation of Extraction of ASR Text Feature

ASR text feature extraction in semantic video retrieval includes three parts:

- Training data: A ASR documents collection labeled with semantic concepts.
- Testing data: A candidate documents collection waiting for labeling with concepts.
- Retrieval list: The candidate documents ranked according their similarity to a semantic concept. Documents that are likely to be relevant to the concept should be ranked at the top and documents that are unlikely to be relevant should be ranked at the bottom of the retrieval list.

We represent a concept's relevant documents collection with a terms collection (Concept Lexicon) $D=\{t_1, t_2, \ldots, t_n\}$, each term has it's weight. D is not the simple compounding of the document terms, whose terms are selected by many terms selecting methods, and the terms' weights reflect their importance to the concept.

We represent a shot document in the test data as a query Q: $Q=\{t_1, t_2, \ldots, t_k\}$.

Then the ASR text feature extraction in semantic video retrieval is transformed to a text retrieval problem: Computing the similarity between D and Q: similarity (Q, D).

Following we'll introduce three methods realized in our experiments.

2.2 TF*IDF Model (Model-TF*IDF)

TF*IDF is the most common and simple method in text retrieval, used by many TRECVID participants [1]. TF is the number of terms in a shot document. DF is the number of documents including the term. IDF is the inverse of DF. In our experiments, we realized the TF*IDF method for comparison, and called it as Model-TF*IDF.

To learn the relationship between the spoken terms and concepts, firstly this method will construct a concept lexicon. We select the lexicon terms based on the DF in the concepts' relevant documents collection. The basic assumption is that rare terms in relevant documents collection are either non-informative for the concept, or not influential in global performance.

To indicate the term's importance to the concept, we weight the terms with TF*IDF. A higher occurrence frequency in a given shot document indicates that the term is more important in that document; at the same time, a lower document frequency indicates that the term is more discriminative.

Finally we represent the ASR text feature as the sparse vectors, whose i^{th} item reflects the weight of the i^{th} term in the concept lexicon.

These ASR feature vectors will be trained by Support Vector Machine(SVM). We rank the shot documents according to the probabilities produced by the SVM model.

2.3 BM25 Model (Model-BM25)

Recent TREC tests showed that BM25 was the best known probabilistic weighting schemes in text retrieval[11]. This is a simple effective approximation to the 2-Poisson Model for probabilistic weighted retrieval [8,12]. We realized the Cornell/BM25

algorithm in our experiments to verify whether it performs as good in ASR text feature extraction as in common text retrieval.

The exact formulas for the Cornell version of BM25 method are:

$$similariy(Q,D)=\sum_{k=1}^{l} w_{qk}\bullet w_{dk}$$

$$w_{qk}=tf(t_k,q) \tag{1}$$

$$w_{dk}=\frac{tf(t_k,d)}{2\cdot(0.25+0.75L)+tf(t_k,d)}\bullet\log\frac{(N-df(t_k)+0.5)}{(df(t_k)+0.5)}$$

Where N is the total number of documents. L is the normalized document length (i.e. the length of this document divided by the average length of documents in the collection). $df(t_k)$ is the document frequency of term t_k. $tf(t_k,d)$ is the occurrence frequency of term t_k in documents collection. $f(t_k,q)$ is the occurrence frequency of term t_k in query.

The comparison of experiments results indicates that the performance of BM25 in ASR text feature extraction is not the best.

2.4 Term Precision Model (Model-TP)

In nature both TF*IDF and BM25 weight the terms with DF. The DF weighting method is the simplest method and to some extent has good performance. But a term's high DF value can't tell us the information that it frequently occurs in the relevant documents or in the irrelevant ones. Especially in the training data labeled with semantic concepts, the imbalance between the relevant documents and irrelevant documents is salient. Then we can't treat the DF in relevant documents and the ones in irrelevant documents equally. We apply the TP weighting method [6] to amend the data imbalance problem. TP method respectively computes the DF in relevant documents and irrelevant documents. Moreover, it is more sensitive to the relevant documents. In the common text retrieval, there are not positive and negative samples annotations, so the TP weighting method is rarely applied to the text retrieval. In our experiments, we realized the TP methods, called as Model-TP.

The application of TP weighting method in the Model-TP is as following:

$$similariy(Q,D)=\sum_{k=1}^{l} w_{qk}\bullet w_{dk}+k_2 n_q\frac{1-L}{1+L}$$

$$w_{qk}=\frac{tf(t_k,q)}{k_3+tf(t_k,q)} \tag{2}$$

$$w_{dk}=\frac{tf(t_k,d)}{k_1 L+tf(t_k,d)}\bullet\log\frac{(A+0.5)(SumN-B+0.5)}{(B+0.5)(SumP-A+0.5)}$$

Where k_1, k_2, k_3 are constants. A is the number of relevant documents containing the term t_k. B is the number of non-relevant documents containing the term t_k.

$SumP$ is the total number of relevant documents. $SumN$ is the total number of non-relevant documents. n_q is the length of the query.

According to the speech text's properties, we set the parameters as:k_1=1, k_2=0, k_3=1,L=1. The Model-TP performs better than the former two methods in our experiments.

2.5 The Models of Integrating Statistical Methods and Rule-Based Methods

Weighting the terms of concept lexicons with the TP method can effectively amend the data's imbalance problem. But the TP method does not perform as well as in selecting terms to build a concept lexicon, for the statistic methods can't effectively and quickly reduce the vocabularies. We try a HowNet-rule-based method to resolve this problem. HowNet is an on-line common-sense knowledge base that provides the inter-conceptual relations and attribute relations of concepts for lexical senses of the Chinese language and their English equivalents [9]. It is the most famous semantic lexical database currently.

We compute the concept relevancy based on the HowNet knowledge dictionary, and build the concept's rule-lexicon based on the values of the relevancy. Simultaneity we build the concept's statistic lexicon separately use the above three statistic weighting methods. The rule-lexicon's terms are correlative with the concept in meaning, and the statistic lexicon's terms are frequently co-occurrence with the concept. We extract the intersections of the rule-lexicons and statistic lexicons as the final concept lexicons, and use them to the semantic concept retrieval experiments. According different weighting methods, we call the models as Model-DF-HN, Model-BM25-HN, and Model-TP-HN correspondingly.

Experiments show that the introduction of the HowNet-rule-based method improved the former three statistic methods.

3 Empirical Validation

We choose the TRECVID 2005's English development data as our experimental data, each shot has been labeled with 40 concepts by IBM. We randomly divided the data to two parts according to the proportion of 1 to 2. The training data include 19935 shot documents, while the testing data include 9540 shot documents.

The version of HowNet we used in our experiments includes 158849 words.

3.1 Evaluation Method

To determine the accuracy of the methods' retrieval results, we use *average precision*, following the standard in TRECVID evaluations. Let $L^k = \{l_1, l_2, ..., l_k\}$ be a ranked version of the retrieve results set S. At any given rank k let R_k be the number of relevant shots in the top k of L, where R is the total number of relevant shots in a set of size S. Then average precision is defined as:

$$\text{average precision} = \frac{1}{R}\sum_{k=1}^{A}\frac{R_k}{k}f\left(l_k\right) \tag{3}$$

Where function $f\left(l_k\right) = 1$ if $l_k \in$ R and 0 otherwise. The average precision favors highly ranked relevant shots.

3.2 Pre-processing

Firstly, we matched the ASR texts with the shots based on the shot temporal boundaries. At the same time, we expanded the shot boundaries to include up to 3 immediate neighbors on either side to a shot document. This shot expansion results in overlapping speech segments and attempts to compensate for speech and visual misalignment.

Then we removed the words in a standard stop word list which includes 524 terms. Besides, we computed all the terms' document frequency occurred in the whole train data, and add the terms with highest DF to the user-defined stop word list. This preprocessing is a good supplement to the standard stop list.

We also used the Porter's well-established stemming algorithm [10] to unify terms, which allows several forms of a word to be considered as a unique term.

3.3 Primary Results

Without losing generality, we choose the following five concepts to experiment: Outdoor(1:3), Building(1:15), Military(1:30), Sports(1:100), Snow(1:250). The numbers in the brackets are the proportions of relevant documents to irrelevant ones.

Fig. 1. displays the performance curves for the retrieval average precisions of "sports" after using Model-TF*IDF, Model-BM25, and Model-TP respectively. From the figure, we found that Model-TP's performance is best.

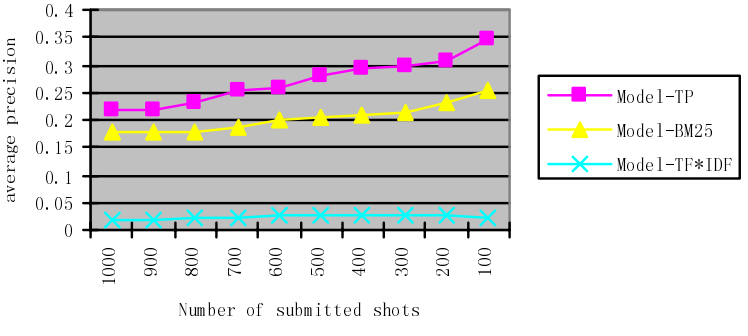

Fig. 1. Average precision of Model-TP,Model-BM25,Model-TF*IDF vs. sports

Considering that Model-TF*IDF is the traditional method, and the Model-TP is the best method, we chose the two models to validate the HowNet's rule-based lexicons. Fig.2 and Fig.3 separately displays the compare of the Model-TF*IDF and the

Model-TF*IDF-HN, the Model-TP and the Model-TP-HN. The improvement after using the rule-based lexicons in the Model-TF*IDF and the Model-TP is obvious.

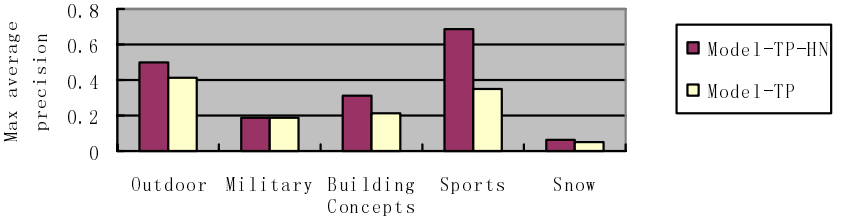

Fig. 2. Comparison of the Model-TP and Model-TP-HN

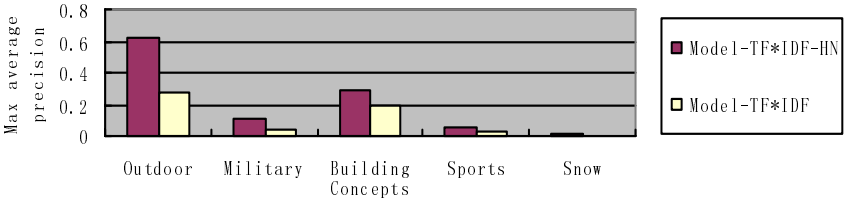

Fig. 3. Comparison of the Model-TF*IDF and Model-TF*IDF-HN

We used the value of "Max Average Precision" in Fig. 2 and Fig. 3. In general, the average precision increases as the number of submitted shots. After reaching the peak, then decreases as the number of submitted shots. The above Max Average Precisions are the peak values with the number of submitted shots between 10 and 1000. The number of submitted shots with peak values is small, which reflects that the most likely relevant documents are ranked at the top of the rank lists, vice versa.

Table1 and Table2 display the number of submitted shots with peak values for the Model-TF*IDF and the Model-TF*IDF-HN, The Model-TP and the Model-TP-HN. From the results, we can find that the convergence of using rules lexicon is better than the common methods.

Table1 and Table2 display the number of submitted shots with peak values for the Model-TF*IDF and the Model-TF*IDF-HN, The Model-TP and the Model-TP-HN. From the results, we can find that the convergence of using rules lexicon is better than the common methods.

Table 1. the number of submitted shots with max average precision for Model-TF*IDF and Model-TF*IDF-HN

Concept / Model	Outdoor	Military	Building	Sports	Snow
Model-TF*IDF	300	1000	50	100	50
Model-TF*IDF-HN	100	1000	10	10	30

Table 2. The number of submitted shots with max average precision for Model-TP and Model-TP -HN

Concept / Model	Outdoor	Military	Building	Sports	Snow
Model-TP	1000	1000	10	500	300
Model-TP-HN	10	300	10	500	300

4 Discussion

The weighting algorithms of TF*IDF and BM25 don't distinguish between the relevant documents and irrelevant ones, and suppose that the more rare terms are more valuable. The weight curve with the DF is as Fig.4. The TP method's weight curve with the DF is as Fig.5. The weight firstly increases as the DF, it indicates that the term is frequently occur in the data. After reaching a peak, the weight curve decreases as the DF, it indicates that the term is too frequently occurs in the data to became non-discriminative.

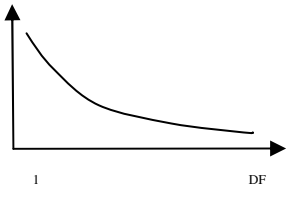

Fig. 4. The variation curve of term weight in Model-TF*IDF and Model-BM25

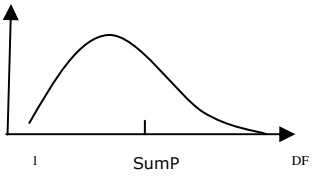

Fig. 5. The variation curve of term weight in Model-TP

In our experiments, the size of the intersections of the rule-based lexicon and the statistic lexicon is about 100, which indicates that the rule-based method can rapidly narrow the vocabulary. But there are some concept lexicons with the very small size. E.g. the size of lexicon for "Snow" is only 14, which badly influence the concept retrieval results(The max average precision of Snow in Model-TP-HN is 0.0588). We analyze the cause is that the proportion of relevant documents to irrelevant documents is too small (1:250), and few terms correlative with "Snow" are spoken. So we get a conclusion that the ASR text features is not effective to all the semantic concepts. We can weight all kinds of features in feature fusion step.

5 Conclusion

In this paper, we propose a method of integrating statistic method and rule-based method to build concept lexicon. The introduction of HowNet rapidly and effectively narrows the vocabulary. Simultaneously we apply the TP weighting method to analyze the ASR texts. Experiments show that this method is more suited to the extraction of

the ASR text features. Future work we will study how to use the ASR text features better to support the video semantic concept retrieval.

References

1. Arnon Amir, Janne Argillander, Murray Campbell, Alexander Haubold, Shahram Ebadollahi, Feng Kang, Milind R. Naphade, Apostol Natsev, John R. Smith, Jelena Teˇsi´c, and Timo Volkmer, "Ibm research trecvid-2005 video retrieval system," in *NIST TRECVID-2005 Workshop*, Gaithersburg, Maryland, November 2005.
2. Cees G.M. Snoek, Jan C. van Gemert, Jan-Mark Geusebroek, Bouke Huurnink, Dennis C. Koelma, Giang P. Nguyen, Ork de Rooij, Frank J. Seinstra, Arnold W.M. Smeulders, Cor J.Veenman, and Marcel Worring, The MediaMill TRECVID 2005 Semantic Video Search *Engine*. In Proceedings of the 3rd TRECVID Workshop , Gaithersburg, USA, November 2005.
3. Tat-Seng Chua, Shi-Yong Neo, Ke-Ya Li, Gang Wang, Rui Shi, Ming Zhao and Huaxin Xu "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS" In TRECVID 2004, NIST, Gaithersburg, Maryland, USA, 15-16 Nov 2004.
4. The TREC Video Retrieval Track Home Page,http://www-nlpir.nist.gov/projects/trecvid/
5. Hauptmann, A., Ng, T.D., and Jin, R. "Video Retrieval using Speech and Image Information," Proceedings of 2003 Electronic Imaging Conference, Storage and Retrieval for Multimedia Databases, Santa Clara, CA, January 20-24, 2003
6. K.LAM , G.SALTON, Term Weighting in Information Retrieval Using the Term Precision Model , January 1982 ,ACM Vol29, No 1, pp 152-170
7. S E Robertson and S Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In W B Croft and C J van Rijsbergen, editors, SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 345–354. Springer-Verlag, 1994.
8. D. Hiemstra. A probabilistic justification for using tf idf term weighting in information retrieval. International Journal on Digital Libraries, 3(2), 2000.
9. Z. Dong and Q. Dong, HowNet, http://www.keenage.com/
10. M.F. Porter An Algorithm for Suffix Stripping Program, 14 pp. 130-137, 1980
11. TREC 2001 National Institute of Standards and Technology, Text Retrieval Conference web page, http://www.trec.nist.gov/, 2001.