ENHANCING SPARSE VOICE ANNOTATION FOR SEMANTIC RETRIEVAL OF PERSONAL PHOTOS BY CONTINUOUS SPACE WORD REPRESENTATIONS

Yuan-ming Liou[1], *Hung-tsung Lu*[2], *Yi-sheng Fu*[2], *Winston Hsu*[2] *and Lin-shan Lee*[12]

[1]Graduate Institute of Communication Engineering, National Taiwan University
[2]Graduate Institute of Computer Science and Information Engineering, National Taiwan University
r02942070@ntu.edu.tw, r03922011@ntu.edu.tw, mayaplus@speech.ee.ntu.edu.tw

ABSTRACT

It is very attractive for the user to retrieve photos from a huge collection using high-level personal queries (e.q. uncle Bill's house), but technically very challenging. The previous work proposed a set of approaches to achieve the goal assuming only 30% of the photos are annotated by sparse spoken descriptions when the photos are taken. This includes fusing the sparse spontaneously spoken features with visual features of the photos by non-negative matrix factorization (NMF), and enhancing the results with two-layer mutually reinforced random walk. However, because the speech annotation is very sparse, the retrieval is very often dominated by the very complete visual features. In this paper, we propose to use continuous space word representations to extend the sparse speech information and expand the photo representation to enhance the retrieval model. Very encouraging improvements were observed in the preliminary experiments.

Index Terms— image retrieval, speech annotation, non-negative matrix factorization, word representation, fused features

1. INTRODUCTION

With the popularity of digital cameras and smart phones, many people saved huge collections of personal photos, but found it challenging to browse across the collection to find a desired photo. Users usually prefer to use personal words as queries to look for photos (e.g. who, where, when, what (objects/events), such as "uncle Bill's house" or "wedding ceremony"). Content-based image retrieval [1, 2] is not useful here, because it requires an example photo as the query. The huge number of annotated photos over the Internet can be useful in identifying photos of publicly known objects (such as "White House") [3, 4], but not necessarily for the personal photo descriptions considered here. Manual annotation of each individual photo is certainly useful, but not attractive at all. This led to the idea of annotating photos with speech [5, 6], and this task seems to be simply the spoken document retrieval [7, 8, 9].

A major issue in spoken document retrieval is that the query and its relevant documents may use different set of words. Latent topic or factor analysis can handle this issue to some extent, with probabilistic latent semantic analysis (PLSA) and non-negative matrix factorization (NMF) as two typical examples [10, 11]. But PLSA and NMF may not be able to solve the problem here, because the query and the labels for related photos may be in several different categories (e.g. some photos by where and some by who, while the query by event) or use different sets of words, and the latent relationships among different terms, specially in different categories, very possibly cannot be trained with very sparse personal annotations. This led to the concept of using image features jointly with speech annotations [12]. Related photos may be linked by image features if annotated very differently, or even not annotated at all.

In a recent work, we proposed to fuse local image features (e.g. visual words by clustering low level image features [13, 14]) and global image concepts(e.g. "people" or "outdoor" by Columbia 374 detector [15]) with the sparse, free-form, and spontaneously spoken annotations [16] and model the relationships among photos and their labels with NMF, and enhance the retrieval process considering different types of features with two-layer mutually reinforced random walk (MRRW) [17, 18]. Only a few words of annotation regarding the photos were needed for only a small percentage of the photos. The sparse speech annotations served as the user interface, while photos without annotation were automatically related by fused feature semantics from NMF and two-layer MRRW.

However, in the above work, the training process of NMF is dominated by the very heavy image features while the very sparse speech annotations carry relatively less weight, even though the words in speech annotations actually bring directly semantic and personal information. We therefore propose in this paper to use continuous space word representations to enhance the NMF model. Continuous space word representations obtained by neural networks have been shown to be able to properly characterize the semantic and syntactic behavior of words [19, 20, 21]. Various approaches such as recurrent neural networks (RNN) [22], continuous bag-of-word model (CBOW) and continuous skip-gram model [23] have been carefully considered and analyzed. We use these word representation approaches to find semantically/syntactically related words to extend the very sparse speech annotation and enhance the whole retrieval model. Substantial improvements were obtained in the preliminary experiments.

2. THE PROPOSED APPROACH

2.1. Overview of the proposed approach

As shown in Figure 1, in the preparation phase on the left, for each photo at the lower left corner, we first extract the visual words (Block (B)) and global visual concept features (Block (C)) from the photo, and speech features (Block (D)) from the annotation, if available. We also train the word representation model from a large text corpus (Block (E)(F)), based on which extended visual concepts and speech features (Block (G)) are obtained. Each photo with these visual words, global visual concept features, speech features and extended visual concepts and speech features is then taken as a document (Blocks (H)). A matrix is constructed for all photos in the archive which is further factorized into an NMF model (Blocks (I)(J)). In the retrieval phase on the right, the user query includes only very few words in text form (or transcribed if spoken). The NMF model gives the first-pass retrieved results (Block (K)), over which two-layer MRRW is performed (Block (L)) to give the final results. Blocks (E)(F)(G) are new in this paper.

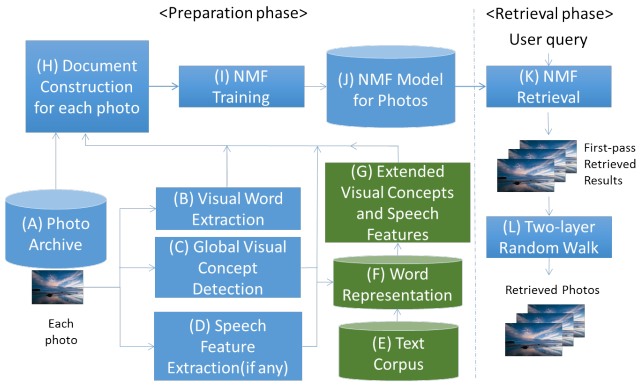

Fig. 1: The proposed approach

2.2. Visual Words/Columbia 374 as Local/Global Visual Features

We use the Scale-Invariant Feature Transform (SIFT) [24] to extract the feature vectors from images, and then produce the codebook of visual words by k-means clustering. The centers of the learned clusters are taken as codewords called visual words, each representing some similar patches (or local features) on images. In this way, an image is represented as the term frequencies of the visual words. On the other hand, Columbia374 detector developed by Columbia University [15] is able to categorize each photo among 374 possible high-level global visual concepts (e.g. "people", "outdoor", "streets"). We take each of these global visual concepts as a term, and the score for the term as the term frequency.

2.3. Speech Features

The speech annotation is the key information here because it provides the core personal semantic concepts such as "uncle Bill's house" or "wedding ceremony". But the speech annotation can be very spontaneous under varying acoustic conditions including out-of-vocabulary (OOV) words. The one-best recognition accuracy can be low, so each utterance is represented as a lattice. But we never know whether a term is present in an utterance represented by a lattice or not. We thus evaluate the expected term frequency $F[t|x]$ for each possible term t in an utterance x as:

$$F[t|x] = \sum_{all\ u} N(u,t)P(u|x) \quad (1)$$

where $N(u,t)$ is the occurrence count of the term t in a path u in the lattice for the utterance x, and $P(u|x)$ is the posterior probability of u based on acoustic and language models. Here we take a word or a subword n-gram (a segment of n consecutive subword units) as the speech term t and evaluate the expected term frequencies. The subword n-gram is to take care of the OOV words to some extent.

2.4. Non-negative matrix Factorization (NMF) and Semantic Retrieval

With the visual and speech features discussed in Sections 2.2 and 2.3 mentioned above, each photo is a document consisting of discrete image terms (visual words and Columbia374 visual concepts) and speech terms with term frequencies. The whole photo archive is then represented as a target matrix $\mathbf{A}$ in Figure 2, in which each row is a vector representing a photo (a document d_i, i=1,...,N), each column is a term (t_j,j=1,...M, speech terms on the left (can be empty) and image terms on the right), and each element in the matrix is the corresponding term frequency for the photo.

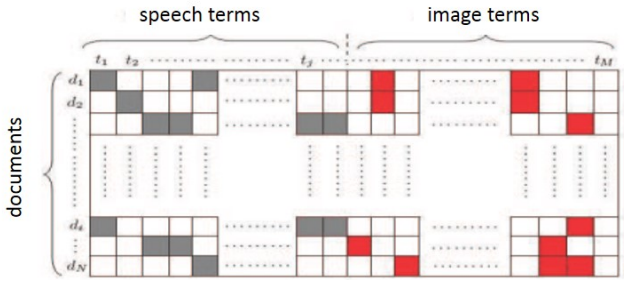

Fig. 2: Matrix representation for the photo archive. The image and speech features are all represented by discrete terms in term frequencies.

The non-negative matrix $\mathbf{A}$ in Figure 2 is then factorized into two non-negative metrics $\mathbf{W}$ and $\mathbf{H}$,

$$\mathbf{A_{NxM}} \approx \mathbf{W_{NxD}H_{DxM}}, \quad or\ \mathbf{A_{ij}} \approx \mathbf{W_iH_j} \quad (2)$$

where A_{ij} is the (i,j) element of $\mathbf{A}$, $\mathbf{W_i}$ the i-th row of $\mathbf{W}$ and $\mathbf{H_j}$ the j-th column of $\mathbf{H}$, and $D \ll N, M$ is the number of latent topics or factors, and $\mathbf{WH}$ is the compressed approximation of $\mathbf{A}$. For example, each column of $\mathbf{A}$ is approximated by a linear combination of columns of $\mathbf{W}$ weighted by the elements of $\mathbf{H}$, etc. In this way the speech and image features are fused.

During retrieval, the query Q can be in either text or speech form, represented as a sequence of L observed speech terms (words or subword n-gram), $Q = \{q_1, q_2, ..., q_L\}$. The documents d (or photos) are then sorted by the relevance

score $S(Q,d)$:

$$S(Q,d) = \sum_{i=1}^{L} W_d H_i \quad (3)$$

where W_d is the row vector of $\mathbf{W}$ corresponding to document d, and H_i the column vector of $\mathbf{H}$ corresponding to $q_i \in Q$. In this way, it becomes possible to retrieve the photos without speech annotation, or with sparse speech annotation in words in different categories from the query (e.g. where and who), because the matching is not based on the appearance of the terms, but on the latent relationships among fused features.

2.5. Two-Layer MRRW

The relevance scores (3) from NMF can be further enhanced by the two-layer MRRW. Each node in the lower layer represents a photos in the first-pass retrieved list from NMF, while that in the upper layer represents a photos in the first-pass retrieved list having speech annotation, or represents one of the D topics obtained by NMF in (2). Let $S_U^{(0)}$, $S_L^{(0)}$ represent the vectors for the relevance scores $S(Q,d_i)$ from NMF in (3) for nodes in upper and lower layers, and $S_U^{(t)}$, $S_L^{(t)}$ represent the enhanced version of them at the t-th iteration. The score propagation can be expressed as random walk in (4) below and shown in Fig 3,

$$\begin{cases} S_U^{(t)} = (1-\alpha)S_U^{(0)} + \alpha \cdot E_{UU}^T E_{UL} S_L^{(t-1)} & (4-1) \\ S_L^{(t)} = (1-\alpha)S_L^{(0)} + \alpha \cdot E_{LL}^T E_{LU} S_U^{(t-1)} & (4-2) \end{cases} \quad (4)$$

where E_{UU}, E_{UL} are respectively the upper-to-upper, upper-to-lower row-normalized cosine similarity matrices , similarly for E_{LL}, E_{LU}. For example in (4-2) the scores of upper layer $S_U^{(t-1)}$ are weighted first by the lower-to-upper similarity E_{LU} then by the lower-to-lower similarity E_{LL}, and then contribute to the scores of the lower layer $S_L^{(t)}$.

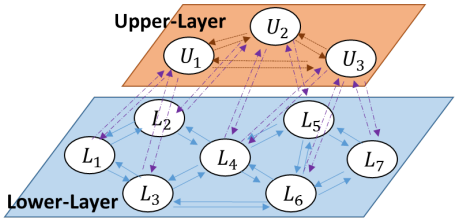

Fig. 3: A simplified example of the two-layer MRRW.

2.6. Continuous Space word Representation for Enhancing the Sparse Matrix

The matrix $\mathbf{A}$ for factorization in (2) is very sparse specially in the part of speech terms. In our experiment, only 30% of photos are annotated but all the photos have visual features, and only 0.5% of elements in the speech term part of matrix $\mathbf{A}$ are non-zero, but 10% for the image term part. So the matrix $\mathbf{A}$ is dominated by the visual features, though the speech terms are primarily words carrying directly semantic information. So data sparsity is an important problem.

Many different models were developed for representing words as vectors in continuous space [19, 20, 21, 22, 23], among which those learned with neural networks have been very successful. In the feedforward neural network language model (NNLM) [25], a linear projection layer and a non-linear hidden layer were used to learn the word vector representations. Recurrent neural network language model as in Fig.4(a) used the hidden layer at the previous time, h(t-1), with a recurrent structure to take into account the previous context, while the word representation vectors [22] can be obtained from the transformation for the hidden layer. Recently, new log-linear models were proposed and shown useful. Continuous bag-of-words model (CBOW) as in Fig.4(b) learned to predict the present word w(t) based on the preceding and following words such as w(t-2), w(t-1), w(t+1), w(t+2) via a projection layer without non-linear elements. The word representation can be obtained from the transformation for the output layer. Continuous Skip-gram model as in Fig.4(c) is very similar to CBOW, but with the layers reversed. The word representation can be obtained from the transformation for the projection layer [23]. It was shown that CBOW and continuous Skip-gram models are better than NNLM and RNNLM for both syntactic and semantic tasks.

With the word representation, for each photo, the Columbia 374 visual concept represented by a word (e.g. "people", "outdoor", etc.) is used to find the top L similar words with word representation cosine similarity above a pre-defined threshold. The term frequencies for these similar words (as speech terms) are then added by the corresponding cosine similarity properly weighted by the Columbia 374 visual concept score. This is repeated for each of the Columbia 374 visual concepts with scores above a pre-defined threshold for the photo. This is also done for top K word arcs in the lattices of the speech annotation for the photo. This is a kind of "document expansion" to extend the visual concepts and speech features for each photo, based on which the NMF model is enhanced.

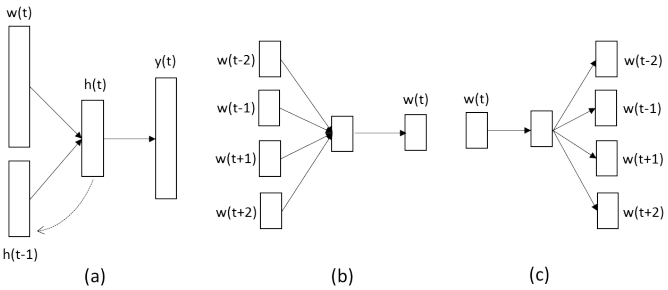

Fig. 4: Neural networks for modeling word representations: (a) RNNLM (b) CBOW (c) Skip-gram

3. EXPERIMENTS

3.1. Experiment Setup

The photo archive was taken from a Flickr user who has more than ten thousand photos on the web with diversified topics.

We randomly selected 7777 from them to be used here. Several students generated the annotation text (primarily in Chinese) spontaneously, most indicating one or two categories of information (e.g. where or who) about the photos explicitly or implicitly, many including OOV words. The audio for these annotations were recorded by 57 students without constraints on the microphone or the acoustic conditions.

The speaker independent (SI) acoustic models were adapted by 30 utterances for each speaker to generate the speaker adapted (SA) models. A language model interpolated from two models respectively trained by news corpora and Plurk corpora was used. The recognition accuracy for the very free speech annotations was only 40.3% for words. Syllable bigrams (segments of two consecutive syllables as the subword n-gram) and words were used for speech terms when evaluating the expected term frequency as in (1). Only 30% of the photos (2100) randomly selected out of the 7777 were allowed to have speech annotations, while the other 70% were assumed to have image features only. In the matrix $\mathbf{A}$, each row includes approximately 32 thousand speech terms, 10 thousand visual words and 374 Columbia concepts as image terms. Another five students generated 32 queries (4 where, 4 who, 4 event and 20 object, different from the previous work [16]) and labeled their ground truths for evaluation. Each query is a Chinese word composed of 2 or 3 syllables. For word representation, we used a corpus of 46 millions words collected from facebook, plurk, news, including photo annotations, to train the RNNLM, CBOW and Skip-gram models for 740k words, each with a 100-dimensional word vector. For NMF we empirically set $D = 90$ topics. For each query we retrieved the top 200 photos with NMF and used them in the two-layer random walk (200 nodes). All performance was in mean average precision for top 50 (MAP@50) [26]. In each experiment, the results below are the average of 10 tests, in each of which the NMF is randomly initialized.

3.2. Experimental Results

The results are listed in Table 1, in which Section (A) is for the baseline NMF, Section (B) is when the matrix $\mathbf{A}$ was enhanced by word representations obtained with RNNLM, Skip-gram and CBOW, and Section (C) is when random walk was used in addition to the best of Section (B). From Section (B) we see the extra words found by word representations really helped, and were much more helpful when based on the Columbia concepts (rows (c)(e)(g)) than on word arcs of lattices (rows (b)(d)(f)), obviously because the Columbia concepts gave clearer global concepts for the photos but the word arcs of the lattices were very noisy. Also RNNLM and Skip-gram were close, but CBOW was much better. So in row (h) for CBOW, we tried to combine the extra words based on Columbia concepts and word arcs of lattices ((f) plus (g)), with a result slightly better than row (f), but worse than row (g), obviously because the noisy lattices disturbed the Columbia concepts.

In the first part of Section (C), only the lower layer was used in single-layer random walk, in which the photo scores are propagated and smoothed based on the similarity matrices evaluated from expected term frequencies (row (i)), those based on word entries extended with word representations (row (j)), and from the visual word features (row (k)). We see re-ranking by random walk offered very good improvement, and the word representations proposed here really helped, even with a single layer. In the second part of Section (C) for two-layer, in row (l) the upper layer includes only those photos from NMF having speech annotations, with the row-normalized similarity matrices based on the expected term frequencies for E_{UU}, Columbia374 features for E_{UL} and E_{LU}, and visual word features for E_{LL}. Row (m) is the same as row (l) except the upper layer also included those word items extended by word representation proposed here. Row (n) is very similar to row (l), except the nodes on the upper layer are the D topics obtained in NMF (columns of W or rows of H), with the row-normalized cosine similarity between the corresponding rows of $\mathbf{H}$ as E_{UU}, between the corresponding columns of $\mathbf{W}$ as the matrix E_{LU}, and $S_U^{(0)} = [1, 1, ..., 1]^T/D$, or assuming all topics have equal scores initially, and those based on visual word features for E_{LL}. We see the two-layer MRRW was significantly better than the single-layer random walk, and the speech features extended by word representaion based on Columbia visual concepts (row (m)) can achieve better performance (rows (m) vs (l)). The best MAP@50 (CBOW word representation based on Columbia concepts and two-layer MRRW, row(m)) achieved 1.95 times higher performance than the NMF baseline of 12.88% (row(a)).

Table 1: Experimental results: (A) NMF baseline, (B) plus word representation, and (C) best of (B) (row (g)) plus random walk

	Methods	Types	MAP@50
(A)	Baseline	(a) NMF	12.88%
(B)	RNNLM	(b) Lattices	13.05%
		(c) Columbia concepts	14.64%
	Skip-gram	(d) Lattices	13.74%
		(e) Columbia concepts	14.66%
	CBOW	(f) Lattices	13.07%
		(g) Columbia concepts	15.72%
		(h) Both ((f) plus (g))	13.32%
(C)	Single-layer	(i) Lattices	19.17%
		(j) Extended Lattices	19.53%
		(k) Visual words	21.85%
	Two-layer	(l) Speech in upper	23.87%
		(m) Extended speech in upper	25.12%
		(n) Topic in upper	24.85%

4. CONCLUSION

This paper considers the enhancement of the very sparse voice annotation for semantic retrieval of personal photos. We propose to use word representations to find semantically/syntactically related words for the word arcs in the lattices and the Columbia visual concepts to extend the very sparse speech annotation and enhance the model. Very good improvements were observed in the preliminary experiments.

5. REFERENCES

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom, "Query by image and video content: the QBIC system," IEEE Computer, Sep. 1995.

[2] Smith, John R., and Shih-Fu Chang. "VisualSEEk: a fully automated content-based image query system." Proceedings of the fourth ACM international conference on Multimedia. ACM, 1997.

[3] Naphade, Milind, et al. "Large-scale concept ontology for multimedia." MultiMedia, IEEE 13.3 (2006): 86-91.

[4] Yi-Hsuan Yang, Po-Tun Wu, Ching-Wei Lee, Kuan-Hung Lin, Winston H. Hsu, "ContextSeer: Context Search and Recommendation at Query Time for Shared Consumer Photos," ACM Multimedia 2008(full paper), Vancouver, Canada.

[5] J. Chen, T. Tan, P. Mulhem, and M. Kankanhalli, "An improved method for image retrieval using speech annotation," Proceedings of the 9th International Conference on Multi-Media Modeling 2003.

[6] Timothy J. Hazen, Brennan Sherry and Mark Adler, "Speech-based annotation and retrieval of digital photographs," Interspeech 2007.

[7] C. Chelba, J. Silva and A. Acero," Soft indexing of speech content for speech in spoken documents," Computer Speech and Language, vol. 21, no. 3, pp.458-478, July 2007.

[8] Yi-chen Pan, Hung-lin Chang and Lin-shan Lee, "Analytical comparison between position specic posterior lattices and confusion networks based on words and subword units for spoken document indexing," Automatic Speech Recognition & Understanding, pp.677-682 , Dec 2007.

[9] Ya-chao Hsieh, Yu-tsun Huang, Chien-chih Wang and Lin-shan Lee, "Improved spoken document retrieval with dynamic key term lexicon and probabilistic latent semantic analysis(PLSA)," ICASSP 2006, vol. 1, May 2006.

[10] T. Hofmann, "Probabilistic latent semantic indexing," Proc. ACM SIGIR Conf. R&D in Informational Retrieval, 1999.

[11] Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788-791.

[12] Fu, Yi-sheng, Chia-yu Wan, and Lin-shan Lee. "Latent semantic retrieval of personal photos with sparse user annotation by fused image/speech/text features." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009.

[13] Tirilly, Pierre, Vincent Claveau, and Patrick Gros. "Language modeling for bag-of-visual words image categorization." Proceedings of the 2008 international conference on Content-based image and video retrieval. ACM, 2008.

[14] Yang, Jun, et al. "Evaluating bag-of-visual-words representations in scene classification." Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007.

[15] Yanagawa, Akira, et al. "Columbia universitys baseline detectors for 374 lscom semantic visual concepts." Columbia University ADVENT technical report (2007): 222-2006.

[16] Liou, Yuan-ming, Yi-sheng Fu, Hung-yi Lee, and Lin-shan Lee. "Semantic Retrieval of Personal Photos using Matrix Factorization and Two-layer Random Walk Fusing Sparse Speech Annotation with Visual Features," Interspeech 2014.

[17] Cai, Xiaoyan, and Wenjie Li. "Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization." Audio, Speech, and Language Processing, IEEE Transactions on 20.5 (2012): 1597-1607.

[18] Chen, Yun-Nung, and Florian Metze. "Two-layer mutually reinforced random walk for improved multi-party meeting summarization." Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, 2012.

[19] Hinton, Geoffrey E. "Distributed representations." (1984).

[20] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." Cognitive modeling (1988).

[21] Elman, Jeffrey L. "Finding structure in time." Cognitive science 14.2 (1990): 179-211.

[22] Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." HLT-NAACL. 2013.

[23] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013)

[24] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

[25] Bengio, Yoshua, et al. "Neural probabilistic language models." Innovations in Machine Learning. Springer Berlin Heidelberg, 2006. 137-186.

[26] Garofolo, John S., Cedric GP Auzanne, and Ellen M. Voorhees. "The TREC Spoken Document Retrieval Track: A Success Story." NIST SPECIAL PUBLICATION SP 246 (2000): 107-130.