FAST LATENT SEMANTIC INDEXING OF SPOKEN DOCUMENTS BY USING SELF-ORGANIZING MAPS

Mikko Kurimo

IDIAP CP-592
Rue du Simplon 4, CH-1920 Martigny, Switzerland
mikko.kurimo@idiap.ch

ABSTRACT

This paper describes a new latent semantic indexing (LSI) method for spoken audio documents. The framework is indexing broadcast news from radio and TV as a combination of large vocabulary continuous speech recognition (LVCSR), natural language processing (NLP) and information retrieval (IR). For indexing, the documents are presented as vectors of word counts, whose dimensionality is rapidly reduced by random mapping (RM). The obtained vectors are projected into the latent semantic subspace determined by SVD, where the vectors are then smoothed by a self-organizing map (SOM). The smoothing by the closest document clusters is important here, because the documents are often short and have a high word error rate (WER). As the clusters in the semantic subspace reflect the news topics, the SOMs provide an easy way to visualize the index and query results and to explore the database. Test results are reported for TREC's spoken document retrieval databases. (www.idiap.ch/ kurimo/thisl.html)

1. INTRODUCTION

As large audio resources have become easy to access, the problem of indexing them for automatic information retrieval has become very important. Examples of such spoken audio collections are recordings of radio and TV broadcasts, dictation tapes and telephone conversations. For many potential applications it is not feasible to provide enough manpower to manually index the collections of thousands of audio hours. Recent developments in computational power and LVCSR, however, are making it possible to scan the audio and build the index automatically.

This paper was prepared in the framework of the THISL (Thematic Indexing of Spoken Language) project [1]. The application of the project is to index broadcast speech from radio and TV news. Good LVCSR systems [11] [2] and NLP and IR [11] [9] systems have been developed in different languages. However, even the best speech recognizers make errors, and in broadcast news the WER can be quite high in some parts. The motivation for this paper is to present methods to successfully index even high WER parts by applying latent semantic analysis (LSA).

LSI [4] is often associated with the SVD of the term-document matrix and selecting index terms by using similarity measures in the semantic space [3]. Several improvements have been suggested, for example, word entropy weighting [5], document and word clustering [3], probabilistic LSA [6] and RM+SOM [9] to overcome the practical and theoretical deficiencies of the LSI. The method presented in this paper (sections 2,3) differs in several ways from the traditional LSI, and some of the results (section 6) show that LSI can be good even without SVD.

SOM is a neural network that performs unsupervised, non-parametric regression to represent nonlinear, high-dimensional data in a low-dimensional display [7]. A good example of using SOMs for data mining in text document collections is WEBSOM [8]. In the our paper the primary use of SOM is to smooth the documents and words in the semantic space so as to improve the indexing of short and noisy documents (section 3). However, the visualization of the index and the query results in a SOM display (section 4) has turned out to be a useful feature as well [9] [6].

2. FAST LSI

LSI attempts to reduce the word noise by projecting the vectors in the original space into a much lower dimensional semantic space. Reducing the dimensionality is done by selecting the subspace spanned by the most important semantic dimensions. The dimensions projected away are considered as irrelevant, i.e. noise, for the indexing task.

Using the RM prior to the SVD analysis provides a

This work was supported by ESPRIT Long Term Research Project THISL.

good way to speed up the LSI, because the final representation after SVD will still be very close to that given by the direct SVD [10]. The accuracy and speed are easily controlled by adjusting the number of random dimensions. The approximation of the SVD is important, because the SVD gets computationally very difficult as the size of the document collection increases. The complexity of the direct sparse SVD is $\mathcal{O}(mnc)$, where n and m are the dimensions of the term-document matrix with c non-zero elements per row. For the RM and the non-sparse SVD after it, the complexity is just $\mathcal{O}(mc\log n + m\log n^2)$ [10]. In practice, the RM dimension and the SVD rank are between $100-300$ whereas the term-document matrix dimensions are between $10000-100000$ or even higher. Other easy ways to approximate the LSI are sampling only a subset for documents or terms (or both). The effects of these on the LSI accuracy are more difficult to analyze, however.

Before SVD, stop lists and Porter stemming are used as in [11], and the entropy based importance weighting for the words as in [5]. After SVD, the semantic word vectors are composed from the projections of the original random dimensions to the obtained semantic subspace, and the document vectors are weighted sums of the word vectors.

3. SMOOTHING THE SEMANTIC VECTORS

Smoothing is important, because the spoken documents are often short, which, together with the high decoding WER, generates a lot of noise in the word count spectra. The smoothing provides, as well, a way to augment the rather small amount of index terms present in a decoded document by terms from other documents that are near in the semantic space.

Smoothing by finding the K nearest neighbors for each document is straight-forward, but this is too slow for large document collections, if no major optimization is made to reduce its complexity ($\mathcal{O}(m^2k)$ for k dimensional vectors). Clustering can be regarded as a way approximating this KNN smoothing, as the cluster centroids are averages of the neighboring documents. To get the mapping more continuous, the smoothed vector can also be computed as the weighted average of the K nearest clusters. This also supports the fact that as the clusters may well represent the typical topics in the collection, a single document can often be relevant for several topics. Clustering is often considerably faster than the KNN search in the input data. For a SOM of s units the complexity of the smoothing is only $\mathcal{O}(mks)$ and of the training $\mathcal{O}(ks^2)$, or even less, after some efficient approximations [8].

In this paper SOM was used to first extract the latent document topics from the collection by clustering the document vectors in the semantic space and then smoothing the vectors. Thus, the smoothed projection of the term t to document d with the nearest clusters $C_1, \ldots, C_K$ is

$$g(t,d) = \sum_{i=1}^{K} p(t, C_i) p(C_i, d) / \sum_{i=1}^{K} p(C_i, d) \;, \qquad (1)$$

after the projections $p()$ in the semantic space are normalized between $[0,1]$. The clustering makes the computation of the term-document projections faster as well. The complexity improves from $\mathcal{O}(mnk)$ to $\mathcal{O}((m+n)sk)$ as we only need to project the terms and documents to the clusters and not to each other.

The index terms for each document are stored with probabilistic weights describing the strength of the association. The probabilistic index weight w_{td} for term t and document d is determined as a convex combination of the Okapi term weighting function $CW(t,d)$ [11] and of the smoothed vector projection $g(t,d)$

$$w_{td} = (1-\lambda) CW(t,d) + \lambda g(t,d) \;, \qquad (2)$$

where the LSI weight $\lambda \in [0,1]$ can be optimized experimentally. This combination can be interpreted as balancing the importance between the smoothing and the decoding. Thus, with $\lambda = 0$ this would equal the basic ranking [11] with no semantic weighting.

As well as semantic document vectors, the semantic word vectors can be smoothed by an SOM. This is motivated by representing more reliably the rare words, which are generally more affected by the word noise. This could be interpreted as a probabilistic grouping of index terms"synonyms", i.e., clustering words that have similar existence patterns in the collection.

4. EXPLORING THE AUDIO COLLECTION

For smoothing the index, there are other clustering methods besides SOM that could probably achieve results very close in precision, but without such inherent visualization of the index. This is especially important, if the collection of the audio documents is very big and the user is not very familiar with its content or how to describe the documents to be retrieved. A topological map of document topics can help to get an overview of the collection and to explore the neighborhood of the obtained query results or other interesting areas [8].

The LSI proposed in this paper can be used as well for displaying the latent semantic axis, the latent semantic topics (labeled document clusters) and the term clusters in a 2D map. The U-matrix of the map [7]

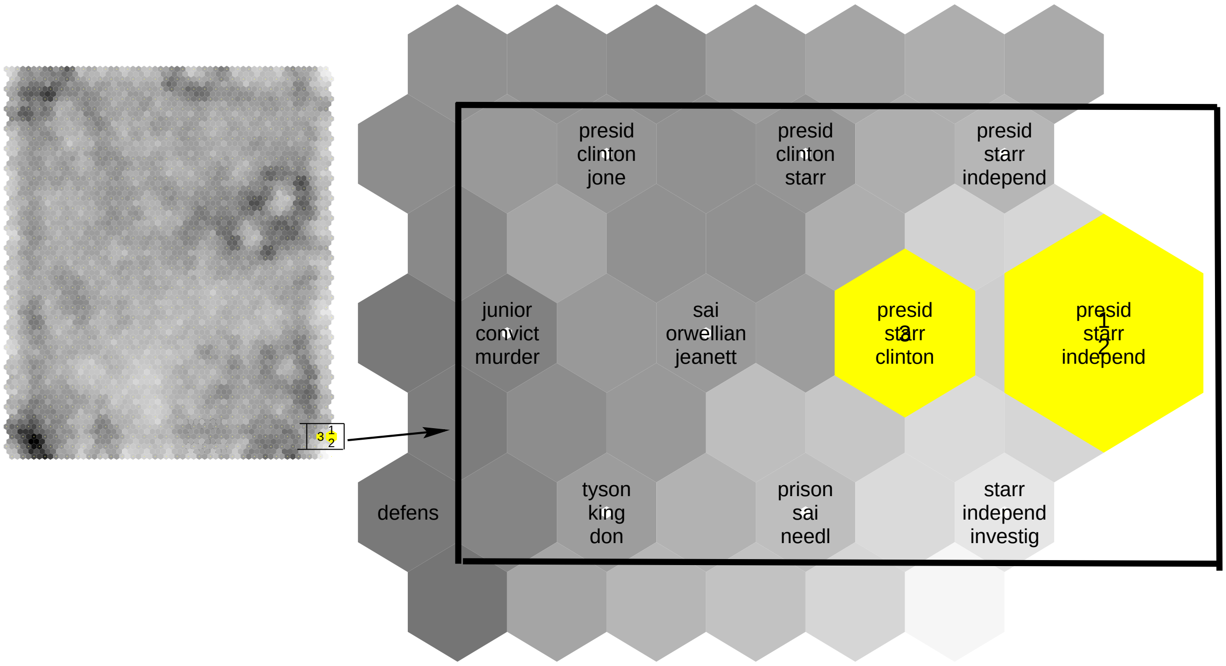

Figure 1: Displaying the latent document topics in a 2D map of hexagons. The original query was "Lewensky" (sic.). The closest topics for the 3 best documents are shown with magnified hexagons.

shows the relations and hierarchies among the topics (see Figure 1). The lighter the color between neighboring clusters, the closer they are to each other. Thus, the large areas of light colors show large topic structures. The document collection can be further explored by plotting useful information on the map, for example, labeling the topics and showing in which document clusters the query results would be mapped. Since the LSI already describes each document by a set of weighted index terms, the clustering based topics were labeled by index terms ranked over all the documents in the cluster.

5. EXPERIMENTS

Two broadcast news databases with standardized evaluations were used in testing the proposed indexing system. The databases are the evaluation sets for TREC-7 and TREC-8 spoken document retrieval tracks. The TREC-7 has approximately 100 hours of news segmented into 3000 stories and the TREC-8 550 hours in 22000 stories. The relevance judgments by human experts are provided for the results of 23 and 50 test queries, respectively.

The speech recognition was done using the THISL speech recognizer, which is a specialized version of the Abbot HMM/ANN hybrid [11] (S1). Results are also given for the reference ASR decodings provided by TREC (B1) and for the reference transcripts with no ASR errors (PF). The baseline method for indexing the decoded documents was the thislIR-0.2 [11], which applies the same stemming, stop list and Okapi term weighting function as the LSI system, but indexes the documents using just the stems found in the decoding.

The LSI+SOM method of Table 1 is described in Section 2. Table 2 gives the results for testing the robustness for values of the main parameters: the combination weight λ (2), the significance threshold S for term selection, the size of the SOM for document (SOM_d) and word vectors (SOM_w) the number of units for smoothing (K_d, K_w) (1), and the RM dimension (while the SVD rank was kept at 200). Additional tests were: to apply the Okapi term weighting CW_2 for the full combination (2) instead of just for the term frequency CW_1, to tune the Okapi parameters (K and b) [11] for each database instead of using the defaults ($K = 2$ and $b = 0.7$), to substitute the entropy W_e by inverted document frequency W_f [9] in word weights, to use the KNN smoothing instead of SOM, and to use RM and SOM without SVD.

6. DISCUSSIONS

The test results are given in terms of the average precision over all standard recall levels [11], which was the most used criterion for the TREC-7 evaluation. In

	WER %	thislIR	LSI+SOM
TREC-7/S1	35.9	0.374	0.381
TREC-7/PF	-	0.434	0.429
TREC-8/S1	32.0	0.400	0.423
TREC-8/B1	27.5	0.404	0.424
TREC-8/PF	-	0.438	0.454

Table 1: Results for the indexing systems in different broadcast news sets and decodings (see section 6).

Index variations:	TREC-7/S1	TREC-8/B1
$\lambda = 0.3$ (0.1)	0.379	0.423
$S = 99\%$ (99.9%)	0.381	0.424
CW_2 (CW_1)	0.374	0.409
Okapi tuned (def)	0.389	0.428
For document vectors:		
$K_d = 3$ or 20 (10)	0.381	0.424
$SOM_d = 1200$ (600)	0.382	0.424
KNN (SOM)	0.372	0.410
For word vectors:		
$SOM_w = 1200$ (-)	0.383	0.424
$K_w = 3$ or 20 (10)	0.382	0.424
$SOM_w = 2000$ (-)	0.381	0.424
W_i (W_e)	0.381	0.421
without SVD	0.381	0.423
$RM = 300$ (200)	0.380	0.423

Table 2: Results for the variations of LSI+SOM index (see section 6). The default values (used for LSI+SOM in Table 1) are given in parenthesis.

many applications the lowest recall levels (i.e. the highest ranked documents) are more important. It seems that for these levels the effect of the WER is smaller and the ranking bigger, but the differences between the methods do not change much, anyhow. The best systems in the TREC evaluations often exploit external text databases by query expansions or other methods, but this has not yet been tried with the current method. It is expected that, e.g., training the SOMs with a large error-free material, would be very helpful in smoothing. Naturally, the queries expanded using traditional indexes as well, could be helpful when using the current LSI index. Table 2 suggests that the average performance is very robust for most of the parameters. It is interesting to note that using the much slower KNN smoothing instead of SOM clustering actually degrades the results, but leaving the SVD out, for example, which makes the indexing even faster, does not. In this sense the SOM is here a very essential part of the LSI.

7. CONCLUSIONS

A novel method for latent semantic indexing is described and tested for spoken audio. The motivation for developing this method was to gain robustness for recognition errors and word noise in short documents as well as to improve the speed and visualization of the LSI. This method includes random mapping (RM) for rapid and controlled dimensionality reduction, entropy based word weighting, probabilistic indexing weights by combined Okapi term weighting and semantic matching, and the use of self-organizing maps (SOMs) to smooth the document and word vectors. In addition to computing the index, the clustering of the documents into latent topic models by SOM provides an easy way to visualize results. Test results are given for two standard evaluation databases for broadcast news. Although there is a lot of data for the speech recognition, the tasks are not very big as IR evaluations. However, the results seem to be good compared to the simple reference index and not very far from the results obtained without any speech recognition errors.

8. REFERENCES

[1] D. Abberley et al. The THISL broadcast news retrieval system. In *ESCA workshop on Accessing Information in Spoken Audio*, 1999.

[2] J. Andersen. Baseline system for hybrid speech recognition on French. COM 98-7, IDIAP, 1998.

[3] J.R. Bellegarda. A statistical language modeling approach integrating local and global constraints. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[4] S. Deerwester et al. Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.*, 41:391–407, 1990.

[5] J.R. Bellegarda et al. A novel word clustering algorithm based on latent semantic analysis. In Proc. ICASSP, pages 172–175, 1996.

[6] T. Hofmann. Probabilistic topic maps: Navigating through large text collections. In Proc. Symposium on Intelligent Data Analysis (IDA), 1999.

[7] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1997. 2nd extended ed.

[8] T. Kohonen et al. Self organization of a massive text document collection. In *Kohonen Maps*. Elsevier, 1999.

[9] M. Kurimo and C. Mokbel. Latent semantic indexing by self-organizing map. In *ESCA workshop on Accessing Information in Spoken Audio*, 1999.

[10] C. Papadimitriou et al. Latent semantic indexing: A probabilistic analysis. In *17th ACM Symposium on the Principles of Database Systems*, 1998.

[11] S. Renals et al. THISL spoken document retrieval. In *Proc. TREC-7*, 1998.