A SEGMENT-LEVEL CONFIDENCE MEASURE FOR SPOKEN DOCUMENT RETRIEVAL

Gregory Senay, Georges Linares, Benjamin Lecouteux

A SEGMENT-LEVEL CONFIDENCE MEASURE FOR SPOKEN DOCUMENT RETRIEVAL

Grégory Senay, Georges Linarès, Benjamin Lecouteux

University of Avignon - Laboratoire Informatique d'Avignon (LIA-CERI)
{gregory.senay,georges.linares,benjamin.lecouteux}@univ-avignon.fr

ABSTRACT

This paper presents a semantic confidence measure that aims to predict the relevance of automatic transcripts for a task of Spoken Document Retrieval (SDR). The proposed predicting method relies on the combination of Automatic Speech Recognition (ASR) confidence measure and a Semantic Compacity Index (SCI), that estimates the relevance of the words considering the semantic context in which they occurred. Experiments are conducted on the French Broadcast news corpus ESTER, by simulating a classical SDR usage scenario : users submit text-queries to a search engine that is expected to return the most relevant documents regarding the query. Results demonstrate the interest of using semantic level information to predict the transcription *indexability*.

Index Terms— Speech recognition, confidence measures, spoken document retrieval

1. INTRODUCTION

Typical approaches of speech retrieval associate Automatic Speech Recognition (ASR) and information retrieval techniques. One of the major issues of this processing scheme is the impact of recognition errors on the SDR system performance : ASR systems suffer from a lack of robustness on unexpected conditions where Word Error Rates (WER) may be upper than 30%, impacting significantly the accuracy of speech search [1, 2, 3]. On controlled conditions, TREC-SDR track concluded that errors do not corrupt SDR engines results [4].

Considering that the perfect ASR system is not yet a short-term perspective, many recent studies on SDR focus on error-tolerant indexing methods, based on word-lattices or N-Best list representations of ASR outputs [5, 6], indexing strategies [7, 8, 9] and the handling of out-of-vocabulary words.

For industrial applications, a realistic way could be to identify the speech segments where ASR system fails, not only in terms of WER but also by considering the final objective of speech retrieval. The erroneous segments could then be checked and corrected by an human operator. In this semi-automatic scenario, the availability of a self-diagnostic tool -that helps the operator to identify corrupted segments- is critical for the global cost of the indexing process. This paper describes such a method that is expected to predict how much an erroneous transcription impacts the global SDR system performance.

First section describes the task and the metric we use for evaluating retrieval errors due to the ASR. Section 2 introduces a segment-level Confidence Measure (CM) that aims to predict this metric. This method relies on the combination of ASR confidence measure and a local semantic compacity index. In Section 3, we present the experimental protocol. Section 4 reports results of experiments. Finally, the paper ends with conclusion and some perspectives.

2. TASK AND EVALUATION METRIC

Evaluation of the WER impact on SDR have been studied and discussed in many papers ([10]). TREC evaluation protocols compare the results provided by the SDR system and a reference ranking provided by human experts. Another way consists in comparing the ranking obtained by automatic transcription of spoken documents to the ones obtained by correct transcriptions. These evaluations are conducted by using a large set of queries, that are submitted to a search engine operating on the whole test set. The performance of the SDR system are obtained by Mean Average Precision (MAP) or R-Precision scores.

Here, our goal is to predict, at a the segment level, the impact of errors on the global SDR process. The next section presents how this indexability measure is estimated, the method we propose to predict indexability being described in section 5.

2.1. Indexability Estimate

Segments are extracted according to the speech pauses, with a maximum length of 30s. Each of them is considered as a document by the SDR system. Considering that one (ore more) error in the segment potentially impacts all search results (for all queries), each segment indexability estimate requires a full run of SDR evaluation.

Therefore, the indexability $Idx(s)$ of a segment s is computed by a 3 step process:

This research was supported by the ANR agency (Agence Nationale de la Recherche), AVISON project (ANR-007-014)

- (1) the targeted speech segment s is automatically transcribed by the ASR system,
- (2) for each test-query, search is performed on the whole speech database by using correct transcriptions for all segments, except for s which is automatically transcribed,
- (3) the resulting ranks are compared to the ones obtained by searching the full reference transcription set. Finally, indexability $Idx(s)$ of the segment s is obtained by computing the F-measure on the top-20 ranked segments, relatively to the top-20 ranking reference (i.e. the ranking on the correct transcripts).

This algorithm estimates the individual impact of the targeted segment transcription on the global SDR process, knowing the targeted ranking. The next section presents a method to predict this *indexability* metric.

3. PREDICTING INDEXABILITY

The proposed method aims to predict the impact of recognition errors on the indexing process. This is achieved by combining word-level confidence measures and a semantic compacity index on the one-best hypothesis from ASR. Combination is achieved by using a classical multi-layer perceptron. These main components are described in the next 3 sections

3.1. Confidence measure from ASR

The ASR confidence scores are computed in 2 stages. The first one extracts low level features related to acoustic and search graph topology, and high level features related to linguistics. In the second step, a first *error* detection hypothesis is produced by a classifier based on the boosting algorithm. Each word from the hypothesis is represented by a feature vector composed of 23 features, that are grouped into 3 classes:

- **Acoustic features** consist of the acoustic log-likelihood of the word, the averaged log-likelihood per frame, the difference between the word log-likelihood and the unconstrained acoustic decoding of the corresponding speech segment.
- **Linguistic features** are based on probabilities estimated by the 3-gram language model used in the ASR system. We use the 3-gram probability, the word perplexity and the unigram probability. We also add an index that represents the current back-off level of the targeted word.
- **Graph features** are based on the analysis of the word confusion networks: the number of alternative paths in the word section and values related to the distribution of posteriors probabilities.

We use a boosting classification algorithm in order to combine word features, as detailed in [11]. The algorithm consists in an exhaustive search for a linear combination of classifiers by overweighting misclassified examples. The classifier is trained on a specific training corpus, that was not included in the ASR system training. Each word from this corpus is tagged as *correct* or *erroneous*, according to the ASR system reference. This measure obtains a Normalised Cross Entropy of 0.373 on dev and 0.282 on test.

The confidence score of a document is computed by filtering the low meaningful words with a stop-list, scores of remaining words being averaged to obtain the segment-level confidence measure.

3.2. Semantic Compacity Index

The use of semantic-level information for indexability prediction is motivated by the fact that a query usually targets documents according to their semantic contents (topic or finer granularity concepts). Some papers proposed to use such high level-features for the estimate of confidence measures [12, 13]. Most of the authors concluded that such an approach does not significantly improve the CM accuracy for ASR. Nevertheless, meaningful words are critical for SDR and WER criterion does not evaluate the interpretability of transcriptions.

Our proposal is to estimate a semantic compacity index $SCI(s)$ for each segment s and to use it as an input feature of the predictor. This segment score is obtained by averaging the local semantic correlations $sc(w_i, w_j)$ of its word pairs (w_i, w_j) estimated on a large corpus.

We focus on short-term correlations between meaningful words. Therefore, cue words are removed according to a stop-list. Moreover, the remaining terms are lemmatised in both corpus and transcription segments. Then, word-pair semantic scores are computed by using lemma co-occurences frequencies weighted by a TF-IDF index:

$$sc(w_i,w_j) = TF(l_i,c).IDF(l_i).\delta^c(w_j)+TF(l_i,c).IDF(l_j).\delta^c(w_i) \quad (1)$$

where l_i is the lemmatised form of w_i word, $TF(l_i, c)$ the frequency of the lemma l_i in the context c, $IDF(l_k)$ the inverse frequency of the lemma l_k on the whole corpus, and $\delta^c(w_i) = 1$ if $w_i \in c$, 0 else.

Semantic compacity $sci(c)$ are estimated in a sliding window of 5 lemma, each corresponding to a context c:

$$sci(c) = \sum_{c_k} \sum_{(w_i,w_j)\in c_k} \sqrt{sc(w_i,w_j) * \frac{IDF(w_i)IDF(w_j)}{\sum_{k=1}^{n} IDF(w_k)}} \quad (2)$$

In our experiments, statistics are computed on a the French part of Wikipedia corpus which offers the advantage of covering large topics and subjects.

3.3. Scores combination

ASR confidence measures and semantic index are combined to predict the indexability score. In order to determine this mapping function from CM and SCI measures, a classical multilayer perceptron is trained [14]. This neural network is 3-layer network trained by using the back-propagation algorithm. Input, medium and output layers contain respectively two, ten and one cells. One of the eight hours of the ESTER test set is dedicated to train (220 of the 1694 segments).

4. EXPERIMENTAL FRAMEWORK

4.1. Speech database

Experiments are conducted on the ESTER database. This database is composed of about 100 hours French radio broadcasts that were manually annotated. The test set is composed by 8 hours from 4 different radio stations. This corpus is split in two parts: the first part is one hour to learn the neural predictor of indexability. The second part is used to test the system on 7 hours (1474 documents)

4.2. ASR system

Experiments are conducted by using the LIA ASR system (Speeral). It is an asynchronous decoder operating on a phoneme lattice; acoustic models are HMM-based, context-dependent with cross word triphones. The language models are classical trigrams estimated on about 200M words from the French newspaper *Le Monde* and from the ESTER broadcast news corpus (about 1M words). The lexicon contains 67K words. Since the full system runs 3 passes including unsupervised speaker adaptation, we use here only the system 2xRT, without any speaker adaptation, that performs 35.1% word error rate on the 8 hours test set of ESTER campaign.

4.3. Search engine and query set

As our goal is to evaluate the data quality rather than the search strategy, we used the standard TF-IDF-based search engine *Lucene* in our experiments. Queries are directly extracted from the headlines of the newspaper *Le Monde* published during the same period as the test corpus. Indeed, this query set is composed of 160k unique queries, every one of which can produce at least one result.

5. EXPERIMENTS

The first experiment consists in evaluating the prediction error rates (PER) on the 7 hours test corpus. In order to estimate the individual contribution of each feature, we trained neuronal predictors base respectively on confidence measure (CM), semantic compacity (SCI) and the combination of CM and SCI, noted as $CM + SCI$ in the table 1.

PER the distortions between $PIdx$ and Idx are evaluated as follow:

$$Q1 = \frac{1}{\tau}\sum_{j=1}^{\tau}\frac{|PIdx(j) - Idx(j)|}{Idx(j)} \quad (3)$$

$$RMS = \sqrt{\frac{1}{\tau}\sum_{j=1}^{\tau}(PIdx(j) - Idx(j))^2} \quad (4)$$

$Q1$ and RMS respectively presents the general distortion and root mean squared (RMS) error. In all results, the $CM + SCI$ score significantly outperforms both confidence measures and semantic consistencies. We can even see that the absolute difference of semantic prediction is much better than confidence measure (27% relative).

	CM	SCI	$CM + SCI$
$Q1$	1,65	1,74	1,59
RMS	0,25	0,32	0,22

Table 1. Mean Prediction error by using neural prediction based respectively on confidence measure only (CM), on semantic compacity index SCI and the combination of the 2 index ones ($CM + SCI$). Prediction errors rates (PER) are computed by using arithmetic error ($Q1$) and root mean square metrics (RMS).

In a second experiment, we investigate the interest of the proposed indexability predictor in the particular usage scenario where a metric is supposed to indicate, to a human operator, the segments that should be manually checked. It is a document classification task where each document is tagged as *indexable* or *unindexable* by the system. We estimate the classification performance by comparing the 2-classes tagging performed by using the real indexability score and predicted score, according to a common fixed threshold. A document is tagged as well-classified, if indexability and predicted indexability are both under or above the same threshold T. Threshold fluctuates between 10% and 90%. Results (cf. fig. 1) at the limit-conditions correspond to expectations: in the [10%,40%] interval, the worst documents are detected; beyond 70%, the system detects only the best ones. Depending of the chosen tradeoff between indexing-quality and cost, the threshold can be adjusted.

The confidence measure CM yields good classification for an indexability threshold under 50% (never below 80% of classification), but classification performances dramatically decrease under 60%. CM predictor outperforms the SCI-based system, but the combined approach $CM + SCI$ clearly provides the best results. It enables to classify documents correctly with T in $[50, 90]$ %. The best improvement is observed for a T value at 70%: confidence measure and semantic prediction obtain only 62% at 61% at this point, but the combination reaches 74%, corresponding to an improvement of about

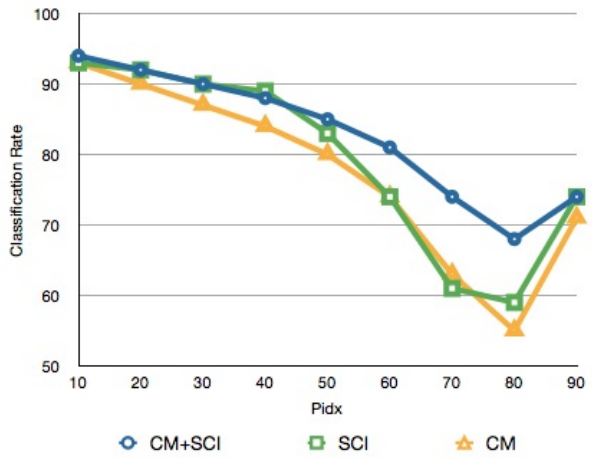

Fig. 1. Indexable/unindexable document classification according to the indexability threshold, by using the predicted indexability based on confidence measure (CM), semantic compacity index (SCI) and the combination of CM and SCI (CM+SCI).

13% relative. In most of the cases, this combined system correctly tags more than 70% of the speech segments.

6. CONCLUSION AND PERSPECTIVES

In this study, we investigated the interest of semantic level information to prior estimate of the transcription quality in the specific scope of spoken document retrieval. We introduced an method for indexability prediction that combines confidence measure from ASR and a local semantic compacity index. Results demonstrate that semantic information is a useful feature for estimating data-quality for SDR; even if its own prediction performance is worse than the one based on confidence measure, complementarity yields a significant improvement: the prediction error rates are improved of about 13% relative with the combined approach. We plan now to investigate various strategies for semantic modeling; here, we focused on local semantic compacity. Context widening and the use other modeling paradigm (such Latent Dirichlet Allocation) could improve the extraction and identification of the latent concepts in the speech stream.

7. REFERENCES

[1] D. W. Oard, D. Soergel, D. Doermann, X. Huang, G. C. Murray, J. Wang, B. Ramabhadran, M. Franz, S. Gustman, J. Mayfield, L. Kharevych, and S. Strassel, "Building an information retrieval test collection for spontaneous conversational speech," in *SIGIR '04*, New York, USA, 2004, pp. 41–48, ACM.

[2] S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, and S. Gary, "Scanmail: a voicemail interface that makes speech browsable, readable and searchable," in *Proceedings of CHI2002*. pp. 275–282, ACM Press.

[3] J.H.L. Hansen, R. Huang, B. Zhou, M. Seadle, J.R. Deller, A.R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 712–730, sep. 2005.

[4] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees, "The TREC spoken document retrieval track: A success story," in *in TREC 8*, 2000, pp. 16–19.

[5] Murat Saraclar, "Lattice-based search for spoken utterance retrieval," in *In Proceedings of HLT-NAACL 2004*, 2004, pp. 129–136.

[6] H. l. Chang, Y. c. Pan, and L. s. Lee, "Latent semantic retrieval of spoken documents over position specific posterior lattices," in *SLT Workshop, 2008. SLT 2008. IEEE*, dec. 2008, pp. 285–288.

[7] C. Chelba, J. Silva, and A. Acero, "Soft indexing of speech content for search in spoken documents," *Computer Speech and Language*, vol. 21, pp. 458–478, 2007.

[8] M. Kurimo and V. Turunen, "Retrieving speech correctly despite the recognition errors," in *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, July 2005.

[9] M. A. Siegler, *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance*, Ph.D. thesis, 1999.

[10] C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 39–49, may. 2008.

[11] P. Moreno, B. Logan, and B. Raj, "A boosting approach for confidence scoring," in *Interspeech, Aalborg, Denmark*, 2001, pp. 2109–2112.

[12] S. Cox and S. Dasmahapatra, "High-level approaches to confidence estimation in speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 7, pp. 460–471, oct. 2002.

[13] D. Hakkani-Tür, G. Tur, G. Ricardi, and H. Kook Kim, "Error prediction in spoken dialog: from signal-to-noise ratio to semantic confidence scores," 2005, vol. I, pp. 1041–1044.

[14] F. Rosenblatt, "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms," in *Spartan Books*, 1962.