Topic Identification from Audio Recordings using Rich Recognition Results and Neural Network based Classifiers

Roberto Gemello, Franco Mana, Pier Domenico Batzu

Loquendo, Torino, Italy

roberto.gemello@loquendo.com, franco.mana@loquendo.com, pierdomenico.batzu@loquendo.com

Abstract

This paper investigates the use of a Neural Network classifier for topic identification from conversational telephone speech, which exploits rich recognition results coming from an automatic speech recognizer. The baseline features used to feed the neural classifier are produced using the words extracted from the 1-best sequence. Rich recognition results include the word union of the first *n*-best sequences, the consensus hypothesis and the full or pruned Word Confusion Network generated from the *n*-best sequences. Different probabilistic information attached to the words, including confidence and word posterior probabilities, is investigated together with classical and probabilistic feature weighting schemes. A large experimentation on conversational telephone speech of Fisher corpus is reported, showing significant improvements when compared to the state of the art.

Index Terms: topic identification, speech recognition, neural networks classifiers, word confidence, word confusion network, consensus hypothesis, word posterior probabilities.

1. Introduction

Topic identification (topic ID) from audio recordings is the problem of categorizing an audio document into a set of predefined topics. The applications of this technology are related to the increasing need for automatic processing and indexing of audio/video media, in particular to the classification of the audio component of radio/tv news and documentaries, youtube videos, etc. The association of topic labels with a spoken audio conversation is also important for security and investigative aims.

Categorization of written texts has been widely studied in the text processing area, by using word-based feature extraction and statistical classifiers [6]. The most common extension of text categorization to topic ID of audio documents consists in applying word-based automatic speech recognition (ASR) to the audio recordings, using a large vocabulary and language modeling, and then processing the recognized words using text-based categorization techniques. This approach works effectively for tasks in which a reasonably accurate speech recognition performance can be achieved. Speech recognition errors can degrade topic identification performance, and this degradation becomes more severe as the accuracy of speech recognition decreases.

Topic ID from audio recordings has been recently investigated in the literature by several works. Hazen et alii compares the use of word and phone recognition lattices [2], and experiments Support Vector Machine (SVM) classifiers with enhanced Minimum Classification Error (MCE) training in [3] and [4], using Fisher telephonic conversational corpus. The same corpus is employed by Wintrode and Kulp [5] that also use SVM and incorporate confidence information to obtain a more robust identification.

Building on these works, we investigate the use of a Neural Network (NN) based classifier for vector based topic ID. Neural Networks have been widely used for classification [7] as offer many theoretical and practical advantages, and are particularly robust to noisy data. Our hypothesis is that NN can be very suited to classify the transcripts of spoken audio automatically obtained from a speech recognition system, and in particular, can exploit the informative contents of rich recognition results. In this work we explore this hypothesis by experimenting a NN classifier on Fisher topic ID. Beside the baseline 1-best sequence, we test richer ASR outputs, namely a *word union* of *n*-best sequences, the consensus hypothesis of the *word confusion net* (WCN) [11] generated from the *n*-best sequences, the full WCN word set and a pruned one. Different probabilistic information coming from ASR, namely confidence and word posterior probabilities are used. The experimental activity shows that results obtained with NN classifier on Fisher topic ID are comparable and in same cases better than those reported in the recent literature using state-of-art classifiers (SVM trained with MCE). Furthermore, the use of confidence and word posterior probabilities, together with rich ASR results, still improves performances.

The paper is organized as follows: Section 2 introduces the NN based topic ID approach. Section 3 describes the topic ID corpus employed and details the ASR and feature creation process. Section 4 describes experimental activity and discusses the results.

2. Neural Network based Topic ID

We follow the vector based topic ID approach coupled with the use of a statistical classifier. This approach is widely used in automated text categorization [6]. Common classifiers employed are Naïve Bayes, SVM and Maximum Entropy. In this work we propose instead a Neural Network based classifier.

2.1. Vector based Topic Identification

In vector based topic ID the document, represented as a sequence of words, is transformed into a vector of numerical features. This action, called *feature vector generation*, follows an approach common in information retrieval and text classification [6] that includes the steps of feature selection and feature weighting. The representation employed is the *bag-of-words*, a vector representing the presence or absence of the words into the text without considering their position. The features, after a selection based on word and document frequencies, can be weighted to improve the effectiveness of text classifier. In fact different terms have different importance in a text, and weighting helps the classifier to take into account this information. A common weighting is *term frequency – inverse document frequency* (TF-IDF) [6] and its probabilistic extension to incorporate word confidence or posterior probability [5].

Copyright © 2011 ISCA

2.2. Neural Network classifier

Neural Networks have been widely used for classification. As outlined in [7], the advantage of neural networks lies in many theoretical and practical aspects. First, neural networks are self-adaptive methods that adjust themselves to the data without any explicit specification for the underlying model. Second, they can approximate any function with arbitrary accuracy [8]. Third, neural networks are nonlinear models, which makes them flexible in modeling non-linearly separable classification problems, without the need of any input space transformation. Finally, differently from other classification methods, neural networks estimates the posterior probabilities of classes [9], which provides the basis for performing a grounded statistical analysis and simplifies the definition of rejection thresholds. A neural network for a classification problem can be viewed as a mapping function $F : R^d \rightarrow R^m$ where a *d*-dimensional input *x* is submitted to the network and an *m*-vectored output containing posterior probabilities $P(C_j|x)$ for each class *j* is obtained to make the classification decision. The network can be single or multi-layered, with sigmoidal units in the hidden layers and sigmoidal or softmax units in the output layer, and is trained to minimize an overall error measure such as mean squared error or cross-entropy.

2.2.1. Multi-class single-label NN classifier

The most common way to build a multi-class classifier by using a NN is to use a network with *m* outputs, one for each class. During training the target is set to 1 for the unit corresponding to the right class and to 0 for the others. During recall, the class corresponding to the winner output unit is chosen as the right classification. The posterior probability of the winning class can be used as confidence measure and to perform some form of rejection. Rejection is performed by sieving the winner class output or the distance between the winner and the second class outputs. Usually, softmax activation function is used in the output layer units, and the cross-entropy error function is minimized in training.

2.2.2. Multi-class multi-label NN classifier

The extension of this model to the multi-label case can be obtained by dropping the winner-takes-all classification policy and substituting it with a sieve based policy.
During the training phase a 1 target is assigned to output units associated to the classes corresponding to the labels of the training pattern and a 0 target is assigned to the other output units. During recall, all and only the classes corresponding to output units that exceed a given sieve (usually 0.5) are chosen as classifications of the input pattern. In this way a test pattern can be associated to a variable number of labels. The sieve can also be made class dependent, and estimated on a development set to minimize a classification measure (precision, recall, F_1). The multi-label case requires the use of sigmoidal output units, instead of the softmax output units used in the single-label case.

3. Experimental Task Description

3.1. Corpus

For the experiments we employed the English Phase 1 portion of the Fisher Corpus [1]. This corpus consists of 5851 recorded telephone conversations between two people that were instructed to discuss a specific topic for 10 minutes. There are 40 different topics, some of them are relatively distinct (e.g. "Movies", "Hobbies", "Education", etc.) while others are quite similar (e.g. "Issues in Middle East", "Arms Inspections in Iraq", "Foreign Relations").
We employed this corpus for closed set topic identification, i.e. identify the topic from the closed set of 40 topics.
The corpus was subdivided into three subsets, according to [4]:

1. Recognizer training set (3104 calls; 553 hours)
2. Topic ID training set (1375 calls 244 hours)
3. Topic ID test set (1372 calls; 226 hrs)

We used exactly the same lists employed in [4], thanks to a private communication of the authors, in order to produce comparable results.
We treated each conversation as an independent audio document. According to [4], in our experiments we provide results on a *whole call* basis, i.e. the two call side audio files for each call are processed independently by ASR, and than the ASR transcripts are merged into a unique document before classification.
As proposed in [2-5], individual call sides are further subdivided into individual audio segments, which are typically a few seconds in length, for processing by ASR. This subdivision can be done by using the manual annotated time limits or automatically. In this work we have used a functionality of Loquendo ASR that automatically detects the voice segments and process them sequentially until the end of the file is reached. This introduces some additional inaccuracy caused by automatic limit detection of voice segments, but can be performed in a fully automatic way without the need of any manual segmentation.

3.2. Speech Recognition details

In our topic ID experiments, the first stage of processing is the application of ASR to each voice segment of each audio document. Loquendo ASR (LASR) was used to automatically detect segments and recognize them.
In the first experiment the generated output included the best sequence that was used as transcribed text for the following text classification phases.
In further experiments richer information was extracted from LASR, as described in section 3.4.2. and 3.4.3:

- Union of *n*-best sequences with there associated confidence value, computed as described in [10];
- Words extracted from a WCN generated from the *n*-best sequences with their associated *word posteriors* probabilities. Both the *consensus hypothesis* and the full and pruned set of words of the WCN are experimented.

This richer information is expected to provide significant improvements in topic identification, as reported in [2] and [3].
The acoustic models of LASR are hybrid ANN/HMM, modelling stationary-transitional units (phonemes and transitions between them). The training material employed includes Fisher recognizer training set.
Recognition took place without any form of speaker normalization or adaptation. Decoding was performed with a FST algorithm employing a trigram language modelling with a 22.8K word vocabulary trained using the transcripts of the recognizer training set. Words with a frequency less than 3 were discarded.

3.3. Speech Recognition results

In this experiment the recognizer applies very basic modelling techniques with a single step and no adaptation. That makes the recognition faster than real time (on a current workstation) but word error rates is high (typically around 50%).

As a comparison, in the cited work by Hazen [4] the MIT SUMMIT ASR was employed, with an acoustic model trained on Fisher Recognizer training set and a language model trained using the transcripts of the same recognizer training set. Also in that case a single recognition step and no adaptation were used, with a reported error rate over 40%.

3.4. Features vector generation

Each audio file is recognized by LASR using a large vocabulary with trigram language modeling.
The words recognized from all the voice segments, with their attached probabilistic information when present, are concatenated into a unique text.
Then this text is reduced:

1) by eliminating the words from a stop-list, containing the most common functional words, like "a", "is", "don't", "have", etc.
2) by eliminating terms that are present in less that *n* documents (*n* is set to 10).

The retained words form the so called *bag-of-words*, i.e. a vector where words are positionally represented. The features of the *bag-of-words* are then weighted to improve the effectiveness of text classifier. Two weighting methods are used: the first one is classical TF-IDF [6], borrowed from the information retrieval field, and the second one is its extension proposed in [5] to deal with confidence weighted words or word posteriors derived from a lattice.

3.4.1. 1-best sequence with TFIDF

When using words from the 1-best sequence, the words are weighted with standard TF-IDF.

$$tfidf\left(t_k, d_j\right) = tf\left(t_k, d_j\right)\log\left(\frac{|Tr|}{df\left(t_k\right)}\right)$$

where *tf* is the frequency of term t_k in document d_j and *df* is the number of documents in the training set *Tr* where the term is present.

3.4.2. n-best union with confidence and ETF-IEDF

Here we consider the case where a confidence is associated with each word in the transcript, created from the 1-best sequence or the union of the first *n*-best sequences. The *n*-best union is created starting from the 1-best and moving through the following bests, adding only words that are not already present, keeping also into account the position.
In the case of 1-best we would like to discount words in the transcript using their probability of actually having occurred, and in the case of *n*-best union we would like to take into account in the same way also additional hypotheses that do not occur in the 1-best transcript.
For this purpose, following [5], we approximate *tf* and *df* calculations probabilistically as the expected term frequency (*etf*) and estimated document frequency (*edf*).

$$etf\left(t_k, d_j\right) = \sum_{i=1}^{N} P\left(t_i = t_k \mid d_j\right)$$

$$edf\left(t_k\right) = \sum_{i=1}^{|Tr|} \min\left(1, etf\left(t_k, d_i\right)\right)$$

$$etfiedf\left(t_k, d_j\right) = etf\left(t_k, d_j\right)\log\left(\frac{|Tr|}{edf\left(t_k\right)}\right)$$

where N is the number of terms in the given document, T_r is the training set, and $|T_r|$ its cardinality.

We take the ETF of a term to be the confidence-weighted sum of the *n* hypothesized occurrences of term t_k in the document. For 1-best transcripts, the expected term frequency is the term frequency weighted by the word-level confidence for each hypothesized word.

3.4.3. Word Confusion Network WPP

In this case we consider the words extracted from a WCN generated from the *n*-best sequences (*n* = 100) with their associated word posteriors. A WCN [11] is a time-ordered sequence of clusters where each cluster contains competing words and their *word posterior probabilities* (WPP). The probabilities in a cluster sum to one. A WCN is constructed using the time-overlap of words in the recognizer's word lattice or *n*-best output. The *consensus hypothesis* is the best path of the word confusion network.
From the WCN we extract four kinds of features:

a) words from the consensus hypothesis, weighted with TF-IDF;
b) words from the consensus hypothesis with their associated WPP, weighted with ETF-IEDF;
c) words from the full WCN with their associated WPP, weighted with ETF-IEDF;
d) words from the pruned WCN. A simple absolute pruning has been used: words with a WPP lower than a predefined threshold θ were deleted. The threshold θ was estimated on a development set (extracted from the train set) in a preliminary experiment, and then used in the final experiment. The estimated value was θ = 0.2.

3.5. NN Classifier employed

As Fisher topic ID is a single label problem, i.e. each call is always labeled with one category, the multi-class single-label NN classifier is employed. In particular a Multi-layer Perceptron (MLP) with an input layer of size equal to the number of words, an hidden layer of 200 hidden, and an output layer of size equal to the number of classes (40 units) is employed in the following experiments. An example of NN architecture used is 5105-200-40 (case of pruned WCN).

4. Experimental Results

Two experiments are reported hereinafter. The first one, outlined in table 1, is aimed at testing the topic ID performances of the MLP classifier, obtained with different kinds of inputs coming from the ASR:

- **1-best** unweighted and weighted with TF-IDF;
- ***n*-best union** (n=1,2,3,10) with confidence weighted with ETF-IEDF;
- **consensus hypothesis** words, weighted with TF-IDF;
- **consensus hypothesis** words with their **WPP**, weighted with ETF-IEDF;
- **Full WCN words** with their WPP weighted with ETF-IEDF. For comparison also WCN words without WPP are tested.
- **Pruned WCN words** with their WPP (pruning takes place for WPP < θ, θ = 0.2)

The results show that:

a) The baseline result obtained with 1-best and TF-IDF is CER = 9.6.
b) Feature weighting is necessary: 1-best without weighting obtains very bad results (CER = 18.6);
c) Using *n*-best union with confidence weighted with probabilistic TF-IDF (ETF-IEDF) slightly improves performances. In this case 2-best union obtains the best

result (CER = 9.2), while using more *n*-bests lead to a confusion that worsens performances.

d) Using WCN consensus hypothesis significantly improves 1-best results (CER = 9.3 vs. 9.6).
e) Using consensus hypothesis with WPP further improves performances (CER = 9.0).
f) Full WCN words with WPP are better than 1-best but inferior to consensus hypothesis (CER = 9.1).
g) Pruned WCN words with WPP obtain the best result (CER = 8.6).

Table 1. *Closed-set topic identification Classification Error Rate (CER) on Fisher test set with MLP classifier, using different kinds of ASR output, word probability information and feature weightings.*

Experimental Conditions			Topic ID test CER (%)
ASR output	*Prob info*	*Feature weigthing*	
1-best	none	none	**18.6**
1-best	none	TF-IDF	**9.6**
1-best	conf	ETF-IEDF	**9.3**
2-best union	conf	ETF-IEDF	**9.2**
3-best union	conf	ETF-IEDF	**9.7**
10-best union	conf	ETF-IEDF	**10.5**
WCN consensus	none	TF-IDF	**9.3**
WCN consensus	WPP	ETF-IEDF	**9.0**
Full WCN	none	TF-IDF	**17.7**
Full WCN	WPP	ETF-IEDF	**9.1**
Pruned WCN	WPP	ETF-IEDF	**8.6**

In the second experiment, reported in table 2, we compare our most significative results with those recently published in [4]. Only TF-IDF results of [4] are considered.

Table 2. *A comparison with state-of-art results on Fisher topic ID from ASR output.*

Classifier	*Features Extracted*	*Feature weighting*	*CER %*
MLP	1-best	TF-IDF	**9.6**
MLP	2-best + conf	ETF-IEDF	**9.2**
MLP	WCN consensus + WPP	ETF-IEDF	**9.0**
MLP	Pruned WCN + WPP	ETF-IEDF	**8.6**
SVM-MCE	Word posterior scores from lattice	ETF-IEDF	**9.8**
SVM-MCE	"	ETF-IEDF + MCE opt.	**9.1**

The common points in the two works are the following:

- The experiments have been performed using the same corpus and exactly the same training and test lists;
- The weighting scheme is the same (ETF-IEDF) [5], but in [4] a MCE optimization of features weights is also performed.

The main differences are the following:

- Feature selection: we filter text with a stopword list and perform a *document frequency* selection with $df > n$, with n=10, while in [2-4] a more sophisticated method is applied (see [2] for details).
- The ASR: Loquendo ASR vs. MIT SUMMIT;
- The ASR outputs and probability information associated to words: we use 1-best, *n*-best union and words from WCN (consensus hypothesis, full WCN, pruned WCN) with their associated WPP, while the expected count of each word within a call is computed by summing posteriors scores from lattices in [4].
- The classifiers: MLP vs. SVM with MCE optimization of parameters.

The results obtained with MLP are quite comparable and in same cases better than those reported in the recent literature with state-of-art classifiers (SVM trained with MCE). In particular the best result is obtained with pruned WCN, using WPP and ETF-IEDF weighting.

5. Conclusions

In this paper we have faced the task of topic identification of audio recordings. A multilayered NN classifier has been used, and it has proven to be comparable and in some cases better than SVM classifiers on the same task. Several kinds of ASR outputs have been used to feed the classifier, ranging form the best recognized sequence without any score information, to richer ASR results, like the union of *n*-best hypotheses, with their attached confidence, and the words extracted from a WCN (consensus hypothesis, full set, pruned set) with their posterior probabilities. The obtained results compares well with the state of the art.

6. Acknowledgements

We are indebted to T.J. Hazen from MIT Lincoln Labs for his sharing of the Fisher topic corpus definitions.

7. References

[1] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generation of speech-to-text," in *Proc. of Int. Conf. on Language Resources and Evaluation*, Lisbon, Portugal, May 2004.

[2] Timothy J. Hazen, Fred Richardson and Anna Margolis "Topic Identification from audio recordings using word and phone recognition lattices". *Proc. of ASRU*, Kyoto, December 2007.

[3] Timothy J. Hazen and Fred Richardson, "A hybrid SVM/MCE training approach for vector space topic identification of spoken audio recordings", *Proc. of Interspeech*, Brisbane, Australia, September 2008.

[4] Timothy J. Hazen, "Multi-Class SVM optimization using MCE Training with application to Topic Identification", *Proc. of ICASSP*, Dallas, Texas, 2010

[5] J. Wintrode and S. Kulp, "Confidence-based techniques for rapid and robust topic identification of conversational telephone speech", in *Proc. Interspeech*, Brighton, England, 2009.

[6] Fabrizio Sebastiani: Machine learning in automated text categorization. *ACM Comput. Surv.* 34(1): 1-47 (2002)

[7] Guoqiang P. Zhang, "Neural networks for classification: a survey" In *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, IEEE Transactions on, Vol. 30, No. 4. (Nov 2000), pp. 451-462.

[8] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251–257, 1991.

[9] M. D. Richard and R. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, vol. 3, pp. 461–483, 1991.

[10] D. Colibro, L. Fissore, C. Vair, E. Dalmasso, P. Laface, "A confidence measure invariant to language and grammar", *Proc. of Interspeech 2005*, pp.1001-1004, 2005.

[11] Lidia Mangu, Eric Brill, Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks" *Computer Speech and Language*, 14:373–400 (2000).