Deliverable D5.2

Final report on the implementation and evaluation of genre classification

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287678.

SEVENTH FRAMEWORK PROGRAMME

Participant no.	**Participant organisation name**	**Part. short name**	**Country**
1 (Coordinator)	University of Edinburgh	UEDIN	UK
2	Aalto University	AALTO	Finland
3	University of Helsinki	UH	Finland
4	Universidad Politécnica de Madrid	UPM	Spain
5	Technical University of Cluj-Napoca	UTCN	Romania

Project reference number	FP7-287678
Proposal acronym	SIMPLE4ALL
Status and Version	Draft1
	Deliverable title
Final report on the implementation and evaluation of genre classification	
Nature of the Deliverable	Report (R)
Dissemination Level	Public (PU)
This document is available from	`http://simple4all.org/publicati`
WP contributing to the deliverable	WP5
WP / Task responsible	WP5 / Task T5.2
Editor	Juan M Montero (UPM)
Editor address	juanmanuel.montero@upm.es
Author(s), in alphabetical order	Roberto Barra, Julian D. Echeverry, Javier Giurgiu, Juan M. Montero, Ruben San-Segund
EC Project Officer	Maqua Leonhard

Abstract

This document describes work on ...

Contents

1 Introduction

The main objectives of the work in WP5 of the SIMPLE4ALL project are:

- to select a set of genres and speaking styles which are acoustically and textually different
- to automatically extract text features to be used to identify the genre of a given text.
- to train classifiers to predict the genre of a text
- to train and test speaking style TTS models which can be associated to the predicted genre [2]

1.1 State of the Art in Automatic Genre Identification

With the rapid growth of the information available online, Automatic Genre (or topic) Identification (AGI) has become a key technique in text data classification [?]. This technique addresses the problem of identifying which genre or topic best matches a certain text or document, given a limited predefined set of genres or topics. It is currently been used in many domains of applications when document classification is needed: document indexing, automatic metadata generation, message filtering... Other alternative names for this area of research are: Text Categorization [?]Manning-2009] Text Classification [?] or Topic Spotting [?]. The task of AGI falls at the intersection of information retrieval and machine learning systems. In the last years a growing number of statistical learning methods have been applied in AGI from these research fields [?]. A standard text identification framework comprises several steps: preprocessing, feature extraction, feature selection and classification. The preprocessing module usually contains several stages: tokenization, stopword removal, stemming and term categorization. Regarding the feature extraction module, the most usual approach is the vector space model [?]. This model is based on the use of a bag of words [?]. The feature selection module generally uses filtering methods (such as weighting schemes for the term and document frequency [?]), or techniques for computing the mutual information of terms [?], information gain [?] and chi-square statistical metrics [?]. The classification module can use well-established techniques from the fields of information retrieval and machine learning. Other approaches include Latent Semantic Analysis [1], Rocchio's method [?], Decision Trees [?], naive Bayesian classifiers and Support Vector Machines [?]. AGI techniqies have been successful in several application domains: topic detection [?], metadata generation [?], text filtering [?], sentiment analysis [?], entity resolution [?]...

1.2 Automatic Genre Identification systems

Generally speaking, AGI can be described as an Information Retrieval process. AGI algebraic models assume that any text document can be represented as a set of terms called index terms. This set of terms can be automatically extracted from each document. In order to extract the feature vector from the natural language document, several steps of data pre-processing must be carried out.

1.2.1 Text Preprocessing

The vocabulary for building an IR model (the set of index terms which represent the documents) can be automatically obtained by preprocessing the text of the documents in order to extract the terms that could be relevant for discriminating between documents of different genres. The preprocessing stage allows converting the documents into a conciser way. The impact of this stage on the success of the genre identification process [?]. The most typical preprocessing modules are:

- Structural processing: it removes any structural non-relevant element in the document such as titles, sections, paragraphs or markup labels (such as XML). If the text is the manual or automatic transcription of an oral document or speech database, this structural processing could be unnecessary. For some genres, some structural elements could be relevant.

- Lexical analysis: it is performed with the objective of converting non-standard elements into a standard alphabetical representation: digits, dates, acronyms, abbreviations...
- Tokenization: it is the process of breaking a text into a stream of smaller tokens such as words, sentences or other task-meaningful text elements. In our research, tokens are words. The simplest way of tokenizing is to split the text by space characters and punctuation marks.
- Stopword removal: it removes the words that provided less information and are too frequent in the text documents. These function words are unlikely to contribute to the distinctiveness of the genres. Articles, prepositions, pronouns and conjunctions are types of words which are typically included in the stopword list. An appropriate stopword list filters out noise from the vocabulary, reducing the size of the indexing structure and contributing to speed up the decision processes.
- Stemming: it morphologically analyse each word, locating the stem and removing prefixes and suffixes. Stemming also minimize the size of the indexing structure by reducing the number of terms to index. For this step, we have used the Freeling Toolkit [**?**]. Due to a few errors in the stemming process, we have modified some of the original Spanish stemming rules in the toolkit.
- Term Categorization: it would build a thesaurus of the terms in the vocabulary. A thesaurus is mostly composed of synonyms and semantically-related words, and reveals hierarchical relationships between terms. This step increases the size of the indexing structure by adding additional terms.

In this work, we have followed all the preprocessing steps, except for the Term Categorization.

1.2.2 Data representation for genre or topic identification

The data representation is based on the vector space model (VSM) proposed by [**?**]. In this VSM, terms and documents are assumed to be independent and documents can be represented as vectors in a space formed by the index terms. This model is often used in Information Retrieval modelling because of its conceptual simplicity and robustness. The relationship between an index-term t_i and a document d_j can be quantified as the number of times the term appears in that document. Processing every term and every document in the training set, we can compute the Term-Document Matrix (TDM):

$$\begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{pmatrix}$$

where V = $\{t_1, t_2, t_3, ..., t_m\}$ is the set of the terms selected after preprocessing , m is the number of terms, n is the number of documents, t_i are the terms; and D = $\{d_1, d_2, d_3, ...d_n\}$ is the whole training set of documents. Each element c_{ij} represents the number of times the term t_i is in the document d_j. The document to be classified can also be represented, using the same terms, as a transposed vector q =[$cq_1 cq_2 ... cq_m$], where cq_i counts the number of times the term t_i was included in the document.

1.2.3 Weighting schemes

To improve the performance of the vector space method, weights can be applied to the index-terms in the Term-Document Matrix. The goal of a weighting scheme is to associate each occurrence of an index-term with a weight that represents its relevance with respect to the genre of the document it appears in [**?**]. A weighting scheme is composed of two different types of term weighting: local weights and global weights. Local weights depend on the estimated frequency of each term in a document. Global weights depend on how many times a term is included in the entire training set. By applying weighting schemes to the TDM, a new matrix WTDM is obtained. In this matrix each element w_{ij} is computed by multiplying two components: the local weight l_{ij} of the term t_i in the document d_j and the global weight g_i of the term t_i over all training documents. The global weight applied to the test set is the same applied to the terms in the training documents.

A standard weighting scheme in IR applications is the Term Frequency-Inverse Document Frequency (TF-IDF) scheme. In this method, the local weight TF accounts for the relative frequency of a term in a document and the global weight IDF accounts for the number of occurrences of a word in the entire corpus (generally on a log scale). This TF-IDF method is the one we have selected as the baseline weighting scheme for comparing the results obtained for the genre identification task in this deliverable. TF components for a document d_j are computed as c_{ij} over the sum of c_{kj} for every value of k. The global IDF g_i is computed as the logarithm of n over df_i, where df_i is the document frequency of the term t_i and it is equal to the number of documents in the collection containing the term t_i. The performance of term weighting schemes has not greatly increased throughout the last years. However, it is not definitive what form of term weighting scheme performs better than others and it has been suggested that these schemes are strongly dependent on the nature of the data [**?**, **?**]. In our work we have conducted experiments using different weighting schemes that are commonly used in the IR field. The local schemes we have used are TF, binary TF, log TF and augmented and normalized TF. The global schemes we have experimented with are IDF, probabilistic IDF, global frequency IDF and entropy [**?**]. Among these, entropy is the most sophisticated scheme. It is based on an Information Theory approach and it analyses the distribution of terms over documents [**?**]. The entropy weighting scheme is defined as follows:

$$\text{entropy} = 1 - \sum_{j=1}^{n} \frac{p_{i,j} \cdot \log(p_{i,j})}{\log(n)}, \quad \text{where} \quad p_{i,j} = \frac{c_{i,j}}{gf_i}$$

Where gf_i is the global frequency of the term t_i measured over all the documents in the collection. Nevertheless, this scheme may lead to a log zero calculation if an index-term is not present in a document. To avoid this problem some authors had suggested to include a smoothing parameter a, resulting in $p_{ij} = (a + c_{ij})/gf_i$. Indeed, it solves the log zero calculation, but the evaluation that we have performed of this scheme has shown that it does not significantly improve the TF-IDF baseline weighting scheme. We propose a pseudo entropy calculation based on the entropy formula in which the parameter pij is calculated as the weighted sum of c_{ij} and the inverse of gf_i. The main idea in this scheme is to assign less weight to the terms that are equally distributed over the documents in the collection and assign more weight to terms that are concentrated in a few documents. p_{ij}=betac_{ij}+gamma/gf_i. The proposed scheme not only solves the log zero problem, but also improves the topic identification accuracy. The parameters beta and gamma must be experimentally adjusted. For the evaluation proposed in the final part of this work, the best results were obtained by adjusting beta = 2 and gamma = 2.47.

1.2.4 Generalized Vector Model

In this model, both documents and queries are represented as vectors in a m-dimensional space, being m the number of index terms considered for the vocabulary of the topic identification task. All terms are assumed to be independent. The document d_j will be represented by the vector d_j (the j-th column of the WTDM matrix) and the query will be represented by the vector wq. Thus, we have the transposed vectors $d_j = [w_{1j} w_{2j} w_{3j} ... w_{mj}]$ and $wq = [wq_1 wq_2 wq_3 wq_m]$. In this model, the similarity between a document vector d_j and a query vector w_q can be computed using an euclidean distance function, usually the cosine distance, calculated as follows:

$$sim\left(\vec{d_j}, \vec{wq}\right) = \cos\theta = \frac{\sum_{i=1}^{m} (w_{i,j} \times wq_i)}{\sum_{i=1}^{m} w_{i,j}^2 \times \sum_{i=1}^{m} wq_i^2}$$

According to this distance, each document is ranked on how close it is to the query.

1.2.5 Latent Semantic Analysis

Although the Generalized Vector Model have been well developed and applied in many practical cases it has some drawbacks that are worthwhile to mention: the first of them is the synonymy, that is the possibility of a concept to be expressed by different words. The inability to aggregate the same concept expressed in different word forms handicap the effectiveness of genre identification. Another drawback is the polysemy, that is the property of some words to have several meanings. Counting all the senses of a words as one sense and generalizing words under the wrong sense impairs the precision of a topic identification system. Latent Semantic Analysis (LSA) proposed by [1], tries to overcome these problems. This method exploits the concept of Vector Space Model and Singular Value Decomposition (SVD) by applying a linear space transformation of the Term-Document Matrix to derive conceptual indices instead of individual words for retrieval. LSA assumes that there is some underlying structure in word usage that is partially obscured by variability in word choice. To reveal this structure, this technique projects documents and queries into a space with latent semantic dimensions. In this space a query and a document can have high cosine similarity even if the do not share any terms. This is possible as long as their terms are conceptually similar or as long as they have been used to express similar concepts. The latent semantic space has fewer dimensions than the original space (which has as many dimensions as index terms) and for this reason it can also be studied as a dimensionality reduction technique.

The SVD is computed by decomposing the Term-Document Matrix $WTDM_{mxn}$ into the product of three matrices T_{mxm}, S_{mxn} and transposed DT_{nxn}, with m as the number of index terms and n the number of documents in the collection. T and D are the matrices of left and right singular vectors, which have orthonormal columns, fulfilling the condition TTtransposed = DDtransposed = Identity matrix. These matrices contain information about the index terms and documents, respectively, in the latent semantic space. The diagonal matrix S contains the singular values of WTDM in descending order such as diag(S) = $[lambda_1, lambda_2, ...lambda_r]$ where $lambda_1 > lambda_2 > ... > lambda_r$ ¿ 0 and r ¡= min(m, n) is computed as the rank of the matrix WTDM. One of the advantages of the SVD is that it allows to obtain an approximate fit using smaller matrices. By selecting the first k largest singular values and their related rows in matrices T and D it is possible to obtain an approximate representation of terms and documents using fewer dimensions. $WTDM_{mxn}$ is approximately equal to $T_{mxk}\ \ S_{kxk}\ D_{kxn}$ transposed. In this method, the weighted query vector wq must be represented as a q_0 vector in k-dimensional space, as follows: wq transposed T S inverted. The product of wq by T gives the appropriate term weights and the right multiplication by S inverted differentially weights the separate dimensions. Then, the query vector q_0 can be compared to all existing document vectors d_j in the matrix D, by computing the similarity between them.

1.2.6 Generation of the stopword list

Stopword removal has proven to be one of the most important stages in the text preprocessing [**?**]. Not only the size of vocabulary directly depends on the stopword removal, but also the computational effort involved in processing the data. There are several stopword lists available online for different languages and for general applications in IR systems. These lists contain the most common words in general domain documents. Words such as articles, prepositions, pronouns, etc., are commonly included in these lists. However, for specific domains, generic stopwords lists do not contemplate terms, that in fact, are very frequent in the specific documents. For that reason, we performed the evaluation using two lists: a generic stopword list with 421 stopwords (List-1) and a domain specific stopword list (List-2), that we created by adding, to the generic list, those terms with an IDF value below a threshold. We performed different experiments on the Test set in order to find the optimal threshold. The lowest identification error was obtained by setting the threshold to 3.4. The List-2 has 525 stopwords.

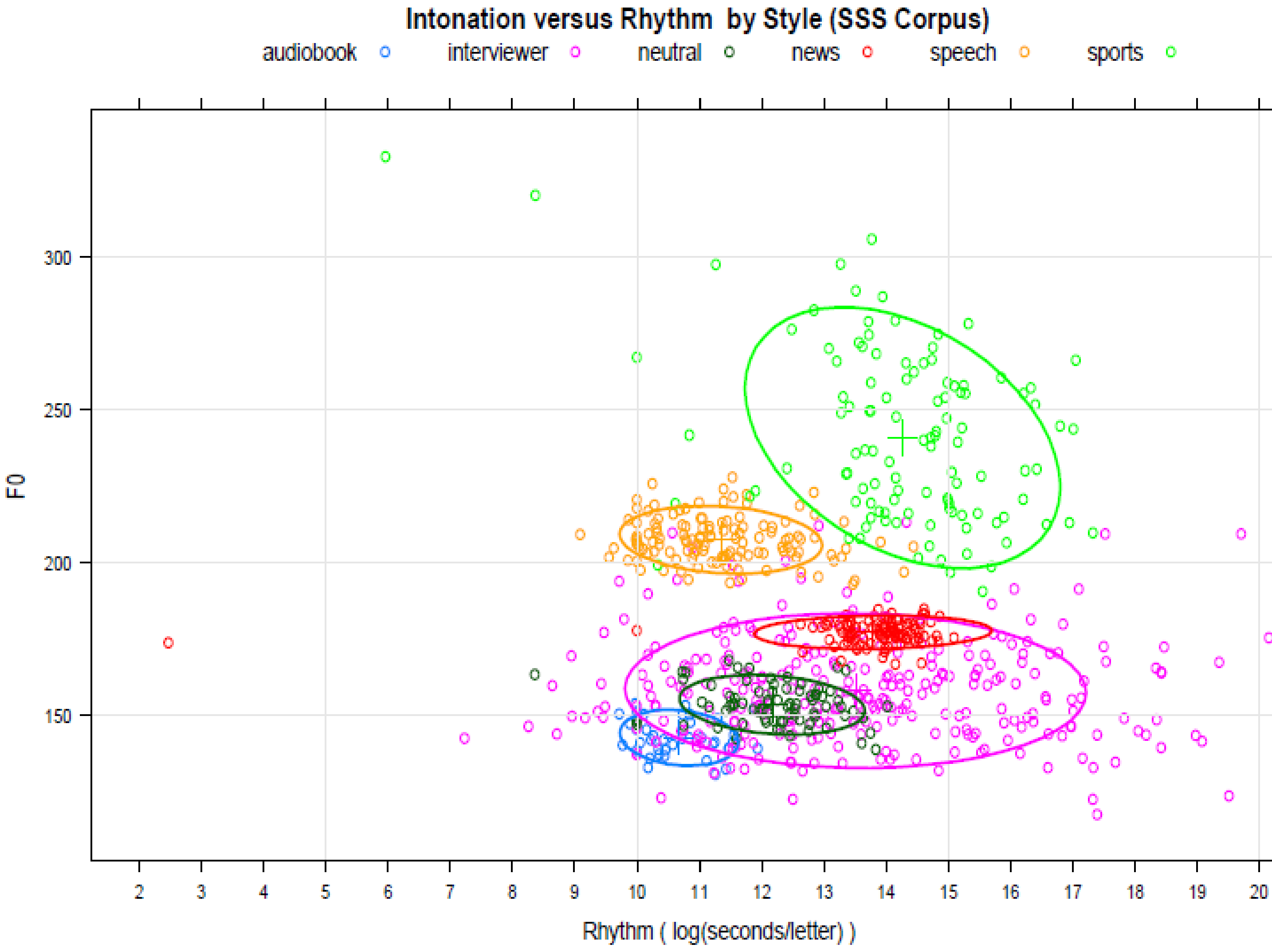
2 Datasets

2.1 Genre Identification Datasets

Our initial dataset was taken from the Spanish C-ORAL-ROM database [3]. Although in the deliverable D5.1, we established a set of genres which could be used for the final training and testing of speaking style synthesis, experiments carried out in the second year have introduced slight modifications in this set of genres, and a new speaking style database has been recorded.

In this new database SSS (Spanish Speaking Styles), we have recorded one male speaker in several speaking styles (including a reference read style). The main styles are: news broadcasting, interviews, political speech, live sport broadcasting and audio-book. We have also recorded several secondary styles for future use child, witch, ogre and old man. For each main style, we have recorded about one hour. The prompted styles (news and political speeches) were recorded in paragraphs; improvised styles (interviews ans live sports) were recorded continuously, and segmented offline. Three types of pauses were hand labelled: intra-sentence pauses, inter-sentence pauses and filled pauses. This labelling was imposed by difficulties when modelling interviews, due to its extremely high number of filled pauses and repetitions.

A prosodic analysis shows the speaking styles are separable, except for the conversational interviews, which is quite broad.



2.2 Topic (or subgenre) Identification Dataset

We have used the EPPS Spanish Database (European Parliament Plenary Sessions) of the TC-STAR Project to study the performance of the proposed system. Due to the fact that the evaluation we are proposing is focused on the supervised topic identification, it is necessary to extract from the database the partition in which there are labels for the topics that are discussed in the speeches. Among the training, development and evaluation datasets, the training dataset of the database is the only one that includes distinct labels for the topics. For this reason we use this dataset for both training and evaluation purposes. We selected a 70% partition of this dataset for training both acoustic and language models and a 30% partition for evaluation.

We believe that identifying the topic on short sentences can be ambiguous because few words do not provide semantic information about the topic that is being addressed. For that reason we decided to perform the evaluation over segments of audio of the same speaker in turns of intervention with a length no less than a minute. We have applied two different criteria for audio segmentation. By these criteria we have generated two sets of audio segments for the evaluation of the system: i) Set 1 is created with audio segments with a minimum length of approximately one minute. Segments that are significantly larger than a minute are not segmented and therefore, the whole turn of intervention of the same speaker remains complete. By this criterion, we obtained 252 audio segments for the evaluation; all of them belong to different topics. ii) Set 2 is created based on the same segments of the Set 1, except that in this case, audio segments significantly larger than one minute are segmented into smaller segments. By following this criterion we have obtained 754 audio segments for the evaluation.

The language of the dataset is Spanish; it contains both male and female speakers; the overall domain is parliamentary speech. The training set contains 67 topics or subgenres, with 23529 sentences grouped in 1908 speaker turns. The lexicon size is 17.4 k words. Test Set 1 contains 3738 sentences grouped in 252 audio segments and Test Set 2 contains 3738 sentences grouped in 754 audio segments. For improving the coverage of the background language model and topic-based language models we use the EUROPARL [**?**] 2 text database in addition to the existing sources of data for language modelling. We have taken advantage of this database in two distinct levels: i) We added it to the text of the training set for generating the background language model. ii) And we also used it for creating the topic-specific LMs. Using the IR models previously trained for identifying the topics in the collection, we applied the LSA approach to automatically classify each of the sentences in the EUROPARL database into one of the available topics. Then, we used these topic labelled sentences for improving the robustness of the topic-specific language models by merging the classified sentences of the EUROPARL database with the topic labelled sentences in the training set.

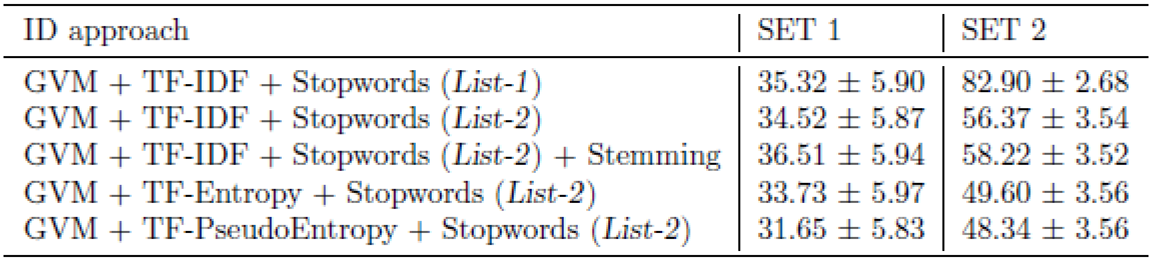
3 Experimental Evaluation

3.1 Genre Identification experiments in the C-ORAL-ROM database

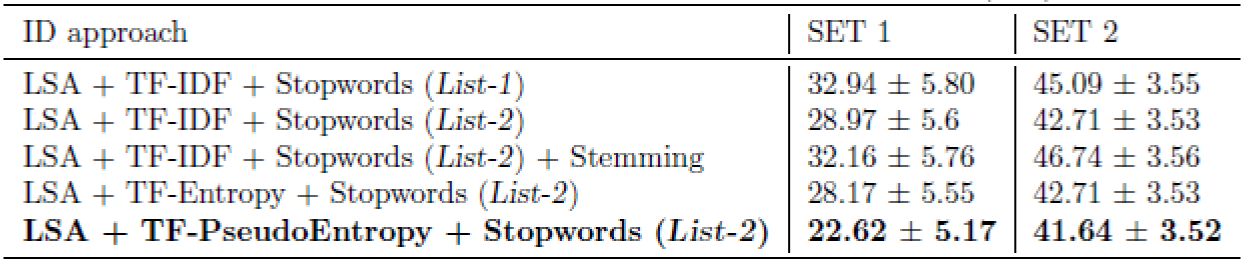
3.2 Experimental Evaluation on TC-STAR topic identification

Our evaluation on TC-STAR focuses on the evaluation of the identification approach by carrying out an evaluation for the TI task consists of identifying the topic that is discussed in each of the transcriptions.

For the topic identification task, the initial performance of the system was obtained by using the Generalized Vector Model, a classic TF-IDF weighting scheme and a general domain stopword list (SW List-1). We will use this configuration as the baseline to discuss the improvements in the different approaches that we have applied. Different tests were performed on both test data sets. We compared different lists of stopwords, which generation was described in section 3.6. We also compared different weighting schemes and the influence of pre-processing stages like stemming in the topic identification error. Table 2 shows the results obtained in topic identification using the generalized vector model. Table 3 show the results when using Latent Semantic Analysis. In general, LSA outperforms the Generalized Vector Model. In both topic models, the combination of TF and pseudo term entropy (TF-PseudoEntropy) reduces the topic identification error when compared to TF-entropy and to TF-IDF weighting schemes. For both models, Stemming does not significantly contribute in error reduction. The criterion that we followed for creating the List-2 of stopwords contribute in most of the cases in reducing topic identification error. The best combination of parameters is obtained for the LSA model, using the List-2 of stopwords and weighting the terms with TF-PseudoEntropy scheme. This configuration presents a relative improvement of 35.96% when compared to the baseline approach.



ID approach	SET 1	SET 2
GVM + TF-IDF + Stopwords (*List-1*)	35.32 ± 5.90	82.90 ± 2.68
GVM + TF-IDF + Stopwords (*List-2*)	34.52 ± 5.87	56.37 ± 3.54
GVM + TF-IDF + Stopwords (*List-2*) + Stemming	36.51 ± 5.94	58.22 ± 3.52
GVM + TF-Entropy + Stopwords (*List-2*)	33.73 ± 5.97	49.60 ± 3.56
GVM + TF-PseudoEntropy + Stopwords (*List-2*)	31.65 ± 5.83	48.34 ± 3.56



ID approach	SET 1	SET 2
LSA + TF-IDF + Stopwords (*List-1*)	32.94 ± 5.80	45.09 ± 3.55
LSA + TF-IDF + Stopwords (*List-2*)	28.97 ± 5.6	42.71 ± 3.53
LSA + TF-IDF + Stopwords (*List-2*) + Stemming	32.16 ± 5.76	46.74 ± 3.56
LSA + TF-Entropy + Stopwords (*List-2*)	28.17 ± 5.55	42.71 ± 3.53
LSA + TF-PseudoEntropy + Stopwords (*List-2*)	**22.62 ± 5.17**	**41.64 ± 3.52**

4 Sentiment Polarity prediction

In addition to identifying the genre a a text, we have also worked on identifying the sentiment polarity of an utterance. This is a text classification problem with many similarities to the genre classification problem. The pre-processing steps are the same, although adjectives, adverbs and verbs are usually more important than nouns in polarity prediction.

We have evaluated two approaches. The first method is based on a 3-dimensional model which takes into account text expressiveness in terms of valence, arousal and dominance. The second one determines the words semantic orientation according to Chi-square and Relevance factor statistic metrics. Several machine learning algorithms, Nave Bayes, SVM and C4.5 have been tested.

The evaluation is performed on four English emotional datasets:

- Semeval 2007 (containing news headlines and five emotional tags: anger, disgust, fear, sadness and joy)
- ISEAR (International Survey on Emotional Antecedents and Reactions: about human experiences and reactions, with 6 emotional tags: anger, disgust, fear, joy,sadness, shame, guilt)
- childrens fairy-tales (the emotional tags are: angry-disgust, fearful, happy, sad and surprised)
- movie reviews (just polarity: positive or negative).

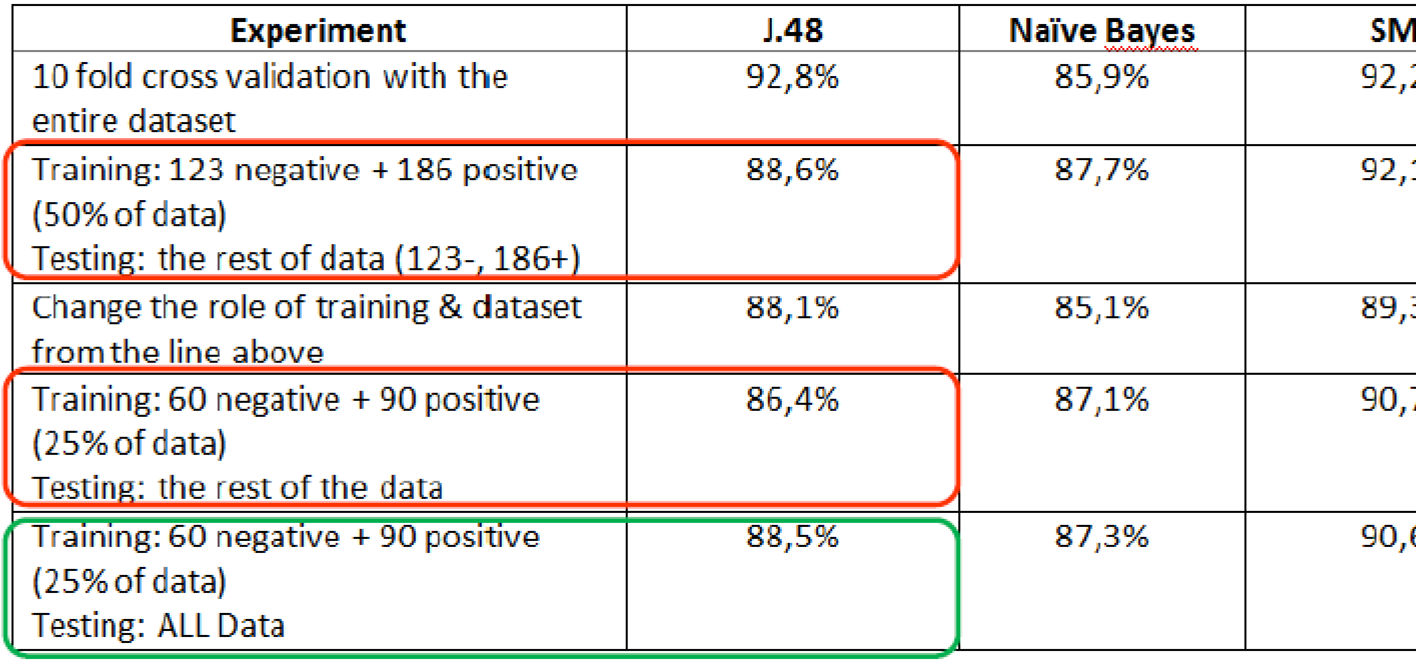
Some examples taken from the datasets are:

- Bombers kill shoppers (Semeval, negative)
- Kate is marrying Doherty (Semeval, positive)
- When I did not speak the truth (ISEAR, negative)
- During the period of falling in love (ISEAR, positive)
- It feels sad (fairy tales, negative)

- When Jemima alighted he quite jumped (fairy tales, positive)
- What a script, what a story, what a mess (movie review, negative)
- You will find this movie extremely funny (movie review, positive)

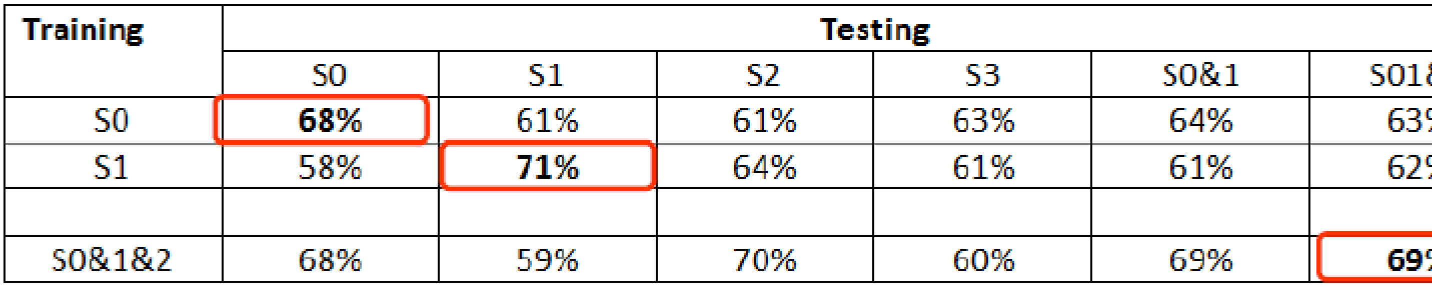
The results show a high correlation of the prediction performance with the database content, as well as to the average number of words within the classified text instances. Both models performed well, with an average 0.75 F-measure on all datasets.

Similar unpublished experiments were carried put on two Spanish corpora. The first one is a collection of reviews about the movie Avatar taken from an Internet site (http://filmaffinity.com). There are 246 positive reviews (65688 words) and 372 negative ones (103722 words). In a 10-fold cross-validation experiment the accuracy was very high as 92



Experiment	**J.48**	**Naïve Bayes**	**SM**
10 fold cross validation with the entire dataset	92,8%	85,9%	92,
Training: 123 negative + 186 positive (50% of data) Testing: the rest of data (123-, 186+)	88,6%	87,7%	92,
Change the role of training & dataset from the line above	88,1%	85,1%	89,
Training: 60 negative + 90 positive (25% of data) Testing: the rest of the data	86,4%	87,1%	90,
Training: 60 negative + 90 positive (25% of data) Testing: ALL Data	88,5%	87,3%	90,

The second dataset was provided by one of our collaborators, the company Daedalus. It contains 38786 Spanish tweets, that were divided into 4 subsets. Tweets are considerably more difficult than movie reviews because of their short length. When training the system on the first 2 subsets and testing on the other 2, the accuracy was quite good (about 65%).



Training	**Testing**					
	S0	S1	S2	S3	S0&1	S01
S0	**68%**	61%	61%	63%	64%	63
S1	58%	**71%**	64%	61%	61%	62
S0&1&2	68%	59%	70%	60%	69%	**69**

[4]

5 Conclusions on entropy and LSA

We have addressed the task of topic identification and by means of that we have improved the language model adaptation in an automatic speech recognition system, by proposing a framework to create topic-based language models and to dynamically adapt the language model used in the final stage of the proposed architecture). Several criteria have been used for creating the stopword list and therefore for selecting the index terms used in the topic identification task. Although stemming does contribute in reducing the size of the indexing structure, it does not contribute in reducing the topic identification error. This may be caused because of the lost of semantic information when reducing words to their stems and thus the relationships between terms and documents may be distorted for this approximation. The complexity of the task is determined in part, for the number of different topics included in the database. The reader should take into account, that this evaluation was performed on a single domain: political. The topic differs, but the domain remains the same throughout the whole evaluation. The results in the ASR task have shown that a small but statistically significant improvement in word recognition accuracy can be obtained by dynamically adapting a topic-dependent language model with interpolation weights computed after the first pass of a multi-pass recognition strategy. Adapting the LM by taking only into consideration the most close topic, improves the baseline performance, but does not take advantage of all the sources of information available.

References

[1] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[2] J. Yamagishi O. Watts J. M. Montero J. Lorenzo-Trueba, R. Barra-Chicote. Towards speaking style transplantation in speech synthesis. Barcelona, (Spain), 2013.

[3] A. Moreno-Sandoval, G. De la Madrid, M. Alcántara, A. Gonzalez, JM Guirao, and R. De la Torre. The spanish corpus. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, Amsterdam: John Benjamins Publishing Company*, pages 135–161, 2005.

[4] Ioana Muresan, Adriana Stan, Mircea Giurgiu, and Rodica Potolea. Evaluation of sentiment polarity prediction using a dimensional and a categorical approach. In *Speech Technology and Human-Computer Dialogue*, 2013.

Appendix: Submitted Conference Paper

Topic identification conference paper

Dynamic topic-based adaptation of language models: a comparison between different approaches.

J.D. Echeverry-Correa, B. Martínez-González, R. San-Segundo,
R. Córdoba, J. Ferreiros-López

Speech Technology Group, Universidad Politécnica de Madrid, Spain
{jdec,beatrizmartinez,lapiz,cordoba,jfl}@die.upm.es

Abstract

This paper presents a dynamic language model adaptation based on the topic that has been identified on a speech segment. We use Latent Semantic Analysis and the given topic labels in the training dataset to obtain and use the topic models. We propose a dynamic language model adaptation in order to improve the recognition performance in a two stages ASR system, in which the final one makes use of the topic identification with two variants: the first one uses just the most probable topic and the other one is dependent on the relative distances of the topics that have been identified. We perform the adaptation of the LM as a linear interpolation between a background model and topic-based LM. The interpolation weight between the models is dynamically adapted according to different parameters: the similarity between the language models and the result obtained by the topic identification process. The proposed method is evaluated on the spanish partition of the EPPS database, a multitopic speech database in a closed political domain. In our evaluation we have achieved a relative reduction in WER of 11.59% over the baseline system which uses a single background language model.

Index Terms: language model adaptation, topic identification, automatic speech recognition, information retrieval

1. Introduction

The performance of a speech recognition system depends significantly on the similarity between the language model (LM) used by the system and the context of the speech that is being recognized. This similarity is even more important in scenarios where the statistical properties of the language fluctuates throughout the time, for instance, in application domains involving spontaneous speech from multiple speakers on different topics. One representative example of this kind of domain is the automatic transcription of *political speeches*. Within this domain, the usage of content words (i.e. those that describe grammatical relationships between other words and hence are not considered function words) depends on several factors, such as the topic the speaker is addressing, the style of the speech, the vocabulary used by the speaker and the scenario in which the speech is taking place. Regarding these factors, in this paper we are focusing on studying the identification of the topic and its application in the adaptation of language models. The importance of the topic identification process in adapting LMs relies on the fact that the probability of usage of content words changes depending on the topic of the speech. The performance of the speech recognition system will depend, among other elements, on its capacity to update or dynamically adapt the LMs.

In this paper we propose a dynamic LM adaptation based on an information retrieval (IR) approach. We used IR techniques for identifying the topics that are related to the content of the speech segment under evaluation. This information enables the system to perform an adaptation of the language model. We explore different approaches for the dynamic language model adaptation. These approaches are based on the interpolation between a background model and topic-based language models.

The remainder of the paper is organized as follows. Section 2 presents a general overview of the proposed framework for the evaluation. Section 3 reviews related work on topic identification systems and language model adaptation. Section 4 provides a description of the topic identification task. Section 5 describes the language model adaptation procedures, with experimental results reported in Section 6. Section 7 discusses some conclusions that can be drawn from this work.

2. General overview

In this paper two major tasks can be distinguished: **topic identification** and **dynamic language model adaptation**. Both tasks pursue one common goal, that is improving the performance of an automatic speech recognition system for multitopic speech. We integrate these tasks in a two stages ASR framework as presented in Figure 1. In the first stage, an initial speech recognition of an audio segment is performed using a background language model built from the entire training set. Then, the IR module automatically identify the topic based on the results of the initial recognition pass. This module uses topic models that have been previously trained for each of the topics available in the database. Using the information provided by the topic identification system and topic-specific language models, a dynamic adaptation of the background language model is performed. In this paper we present different approaches for the dynamic adaptation of language models. In the final stage of the framework, the adapted LM is used to re-decode the utterance.

3. Related work

The task of topic identification (TI) falls at the intersection of information retrieval and machine learning systems. In the last years a growing number of statistical learning methods have been applied in TI from these research fields [1]. Common approaches includes Latent Semantic Analysis [2], Rocchio's method [3], Decision Trees [4] and Support Vector Machines [5]. TI has been succesfully applied in many contexts and disciplines, ranging from topic detection [6], automated metadata generation [7], document and messages filtering [8] and the recently developed area, sentiment analysis [9], among

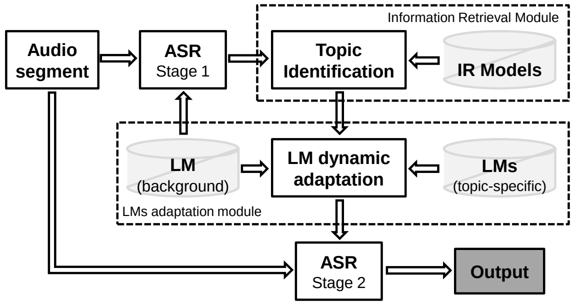


Figure 1: *Two stages ASR framework*

many other fields of application. Nevertheless it is interesting to review the influence of TI in the field of language model adaptation. Within this field, TI has been used to study the changes that the language experiences when moving towards different domains [10]. In that sense, TI is able to contribute to LM adaptation by adding new sources of information to previously existent models with the objective of enriching them. This leads to a diversity of approaches in the field of LM adaptation that can be distinguished regarding the origin of the new sources of information. Some LM adaptation approaches are based on the specific context of the task that they are addressing. In these approaches, the new data is used to generate a context-dependent LM which is then merged with a static LM. These new sources of information can proceed, for instance, from text categorization systems as in [11], from speaker identification systems [12], from linguistic analysis systems [13] or from the application context itself [14]. Other approaches are based on analysis and extraction of semantic information. Latent Semantic Analysis (LSA) is an example of this type of approach. In [15], the use of LSA is proposed to extract the semantic relationships between the terms that appear in a document and the document itself. More robust techniques in the field of information retrieval, as Latent Dirichlet Allocation (LDA) [16], have also been used for adapting LMs [17].

When using data available online it is possible to find information related to a large variety of topics. In this regard, clustering algorithms have been proposed to group together those elements that share some properties. Topic-based language modeling is an example of this clustering criterion [18, 19].

4. Topic identification

In a broad sense, topic identification is the task of automatically identify which of a set of predefined topics are present in a document. To perform topic identification some steps must be followed. These steps are: preprocessing, document representation, term weighting and topic modeling and identification.

4.1. Preprocessing

The preprocessing stage allows to convert both, documents and queries, to a more precise and concise format. This stage has a substantial impact on the success of the topic identification process [20]. Typical preprocessing steps include: structural processing, lexical analysis, tokenization, stopwords removal, stemming and term categorization. We provide a small description of the steps in which we made special considerations:

- *Stopword removal.* There are several stopword lists available online for different languages and for general applications in IR systems. However, for specific domains, generic stopwords lists do not contemplate terms, that in fact, are very frequent in the specific documents. For that reason, we performed the evaluation using two lists: a generic stopword list with 421 stopwords (*List-1*) and a domain specific stopword list (*List-2*), that we created by adding, to the generic list, those terms with an IDF value below a threshold. We performed different experiments on the Test set in order to find the optimal threshold. The lowest topic identification error was obtained by setting the threshold to 3.4. The *List-2* has 525 stopwords.
- *Stemming.* This step refers to the transformation of a word to its stem or root form. For this step, we have used the Freeling Toolkit [21]. Due to few errors in the original stemming process, we have modified some of the stemming rules for the spanish language of the toolkit.

4.2. Document representation

The document representation is based on the widely known bag-of-words model. In this model the relationships between the index-terms and each of the documents in the collection are represented by a Term-Document Matrix, that describes the frequency of ocurrence of the index-terms in the documents.

4.3. Term weighting

To improve the capacity of discrimination of the index-terms, weights can be applied to the elements of the Term-Document Matrix by associating the ocurrence of an index-term with a weight that represents its relevance with respect to the topic of the document. These weights are applied to both documents and queries. We have selected the combination of TF (*Term Frequency*) and IDF (*Inverse Document Frequency*) as the baseline weighting scheme for comparing the results obtained for the topic identification task in this paper. Among the most common weighting schemes, *term entropy* (*te*) is based on an information theory approach and it exploits the distribution of terms over documents [22]. For the index-term t_i in the document d_j, it is defined as follows:

$$te_{i,j} = 1 - \sum_{j=1}^{n} \frac{p_{i,j} \cdot \log(p_{i,j})}{\log(n)}, \text{ where } p_{i,j} = \frac{c_{i,j}}{gf_i} \tag{1}$$

Where $c_{i,j}$ represents the term frequency of the index-term t_i in the document d_j. gf_i is the global frequency of the index-term t_i measured over all the documents in the collection. This scheme may lead to a log zero calculation if an index-term is not present in a document. It has been suggested to include a smoothing parameter a, resulting in $p_{i,j} = (a + c_{i,j})/gf_i$. Indeed, it solves the log zero calculation, but the evaluation that we have performed on the combination of TF and this scheme has shown that it does not significantly improve the TF-IDF baseline weighting scheme. We propose a *pseudo term entropy* calculation based on the *term entropy* formula. Our idea is to assign less weight to the terms that are equally distributed over the documents in the collection and assign more weight to terms that are concentrated in a few documents. In this *pseudo term entropy* the parameter $p_{i,j}$ is calculated as the weighted sum of $c_{i,j}$ and the inverse of gf_i.

$$p_{i,j} = \beta \cdot c_{i,j} + \frac{\gamma}{gf_i} \tag{2}$$

The proposed scheme not only solves the log zero problem, but also improves the topic identification accuracy as shown in section 6. We performed different experiments on the Test set in order to adjust the parameters β and γ. For the evaluation proposed in this paper, the best results were obtained by adjusting $\beta = 2$ and $\gamma = 2.47$.

4.4. Topic models

In this paper we compare two topic models: the Generalized Vector Model (GVM) and Latent Semantic Analysis (LSA) [2]. Both models represent documents and queries as vectors in a multi-dimensional space, in which the number of dimensions is determined by the number of index-terms in the GVM or the number of latent dimensions in the LSA approach.

In both models, the similarity $sim(\vec{d}, \vec{q})$ between a document $\vec{d}$ and a query $\vec{q}$ can be computed using the cosine distance. According to this distance, each document is ranked on how close it is to the query. In our approach, we have gathered all documents in the collection belonging to the same topic into one document. We have done the same for all the topics. By doing this, each document represents a distinct topic. So, when computing the similarity between the query and a document, we are actually computing the similarity between the query and a topic.

5. Topic-based language model adaptation

Topic-based LM adaptation becomes a strategie to lower the word error rate of the transcription given by the ASR by providing language models with a higher expectation of words and word-sequences that are typically found in the topic or topics of the story that is being analyzed. LM interpolation is a simple and widely used method for combining and adapting language models [23, 24].

5.1. Language model interpolation

Given a background model $P_B(w|h)$ and a topic-based model $P_T(w|h)$ it is possible to obtain a final model $P_I(w|h)$, to be used in the second decoding pass, as

$$P_I(w|h) = (1 - \lambda)P_B(w|h) + \lambda P_T(w|h) \tag{3}$$

where λ is the interpolation weight between both models, which has to fulfill the condition $0 \leq \lambda \leq 1$. The topic-based LM is generated by combining several topic-specific LMs $P_t(w|h)$ in general. In our case, the background model, as well as the topic-specific models are static models. They are trained once and remain unchanged during the evaluation. The topic-based LM could be either static or dynamic. It depends on the adaptation scheme followed, as we will see later in this paper. This model, as well as the final model $P_I(w|h)$, are generated during the evaluation of each audio segment.

5.2. Interpolation schemes

Two question arises at this point. How to generate the topic-based model $P_T(w|h)$? and, how to determine the interpolation weight λ with the background model? For solving these questions, we propose different approaches:

- **Hard approach.** In this approach, the topic-based LM $P_T(w|h)$ is built by considering only one of the topic-specific language models ($P_t(w|h)$). This model is selected as the one related to the topic ranked in the first position by the TI system. For estimating the interpolation weight λ we define a distance measure δ between this LM and the background LM. In this approach, our hypothesis is that the greater the distance between both models, the greater the contribution of the topic specific model to the final one. This distance is computed by considering the average diference in the unigram probabilities of both models.

$$\delta_T = \frac{1}{N} \sum_{\forall w_i \in P_T} |P_T(w_i) - P_B(w_i)| \tag{4}$$

Where N is the number of unigrams in the topic-based LM $P_T(w|h)$. To ensure the interpolation weight fulfills the condition $0 \leq \lambda \leq 1$, we include the summation of the distances of all the topic-specific LMs to the background model as a normalization constant. Then, the interpolation weight is computed as the relative distance between δ_T and this normalization constant.

$$\lambda = \frac{\delta_T}{\sum_{j=1}^{n} \delta_j} \tag{5}$$

Where n is the number of topics and δ_j the distance of the *j-th* topic-specific LM to the background LM.

- **Soft-1 approach.** In this case, instead of using only one specific-topic LM for generating the topic-based LM, this model is built on a dynamic basis by the interpolation of a different number of topic-specific LMs. The **Soft-1 approach** tries to gather the dynamic of the specific-topic models $P_t(w|h)$ depending on the similarity of the audio segment to each of the topics. By doing this, more relevance is given to the topics ranked in the first positions by the TI system. The topic-based LM is then computed as follows:

$$P_T(w|h) = \alpha_1 P_{t_1}(w|h) + \alpha_2 P_{t_2}(w|h) + \cdots + \alpha_k P_{t_k}(w|h) \tag{6}$$

where k is the number of models considered for obtaining the topic-based LM. The interpolation weight α_i is calculated as the normalized value of the similarity measure of the TI system.

$$\alpha_i = \frac{sim(\vec{d_i}, \vec{q})}{\sum_{j=1}^{k} sim(\vec{d_j}, \vec{q})} \tag{7}$$

The interpolation weight λ between the background LM and the topic-based LM was set experimentally in this case.

- **Soft-2 approach.** This approach is similar to the previous one, but instead of setting λ experimentally, we have computed it by weighting the relevances of the topic-specific LMs according to the cosine distance. That is:

$$\lambda = \sum_{i=1}^{k} \frac{sim(\vec{d_i}, \vec{q})}{\sum_{j=1}^{k} sim(\vec{d_j}, \vec{q})} \cdot \frac{\delta_i}{\sum_{j=1}^{k} \delta_j} \tag{8}$$

In Soft-1 and Soft-2 approaches, we have considered two additional possibilities: a) to create the topic-based LM using all the topic-specific LMs, that is by setting k as the total number of topics, and b) to create the topic-based LM by selecting the 10 topics with higher positions in the TI ranking.

6. Experimental evaluation

Our evaluation focuses in two aspects: the evaluation of the topic identification approach and the evaluation of the dynamic language model adaptation by means of evaluating the performance of the speech recognition system. Before discussing the results obtained, we describe the dataset used for the evaluation.

6.1. Dataset

We have used the spanish partition of the EPPS Database (*European Parliament Plenary Sessions*) of the TC-STAR Project to study the performance of the proposed system. Due to the fact that the training dataset of the database is the only one that includes distinct labels for the topics, we use it for both training and evaluation purposes. We selected a 70% partition of this dataset for training both acoustic and language models and a 30% partition for evaluation. We believe that identifying the topic on short sentences can be ambiguous because few words do not provide semantic information about the topic that is being adressed. For that reason we decided to perform the evaluation over segments of audio with a length no less than a minute. We extracted these segments from turns of intervention of just one speaker. By this criterion, we obtained 252 audio segments for the evaluation. Some details of the corpus: The language of the corpus is Spanish. There are both male and female speakers (approx. 75% - 25% distributed). The domain of the corpus is political speeches. Training set is composed of 23539 sentences grouped in 1908 speaker turns of intervention. The lexicon size is 17.4k words and the Test set is composed of 3738 sentences grouped in 252 speaker interventions. Each of the speaker interventions (both training and test) belongs to one of 67 different topics. We also use the EUROPARL [25] text database for training both background and topic-specific LMs.

6.2. Topic Identification evaluation

For the topic identification task, the initial performance of the system was obtained by using the Generalized Vector Model, a classic TF-IDF weighting scheme and a general domain stopword list (SW *List-1*). We will use this configuration as the baseline to discuss the improvements in the different approaches that we have applied. We compared the two different lists of stopwords. We also compared different weighting schemes and the influence of preprocessing stages like stemming in the topic identification error. Table 1 shows the results obtained in topic identification using both GVM and LSA approaches. In general, LSA outperforms the Generalized Vector Model. In both topic models, the combination of TF and *pseudo term entropy* (TF-PseudoEntropy) reduces the topic identification error when compared to TF-*entropy* and to TF-IDF weighting schemes. For both models, Stemming does not significantly contribute in error reduction. The criterion that we followed for creating the *List-2* of stopwords contribute in most of the cases in reducing topic identification error. The best combination of parameters is obtained for the LSA model, using the *List-2* of stopwords and weighting the terms with TF-PseudoEntropy scheme. This configuration presents a relative improvement of 35.96% when compared to the baseline approach.

Table 1: *Topic Identification error (T.ID. error) using GVM and LSA topic models approaches*

Topic identification approach	T.ID. error (%)
GVM + TF-IDF + SW (*List-1*)	35.32 ± 5.90
GVM + TF-IDF + SW (*List-2*)	34.52 ± 5.87
GVM + TF-IDF + SW (*List-2*) + Stemming	36.51 ± 5.94
GVM + TF-Entropy + SW (*List-2*)	33.73 ± 5.97
GVM + TF-PseudoEntropy + SW (*List-2*)	31.65 ± 5.83
LSA + TF-IDF + SW (*List-1*)	32.94 ± 5.80
LSA + TF-IDF + SW (*List-2*)	28.97 ± 5.60
LSA + TF-IDF + SW (*List-2*) + Stemming	32.16 ± 5.76
LSA + TF-Entropy + SW (*List-2*)	28.17 ± 5.55
LSA + TF-PseudoEntropy + SW (*List-2*)	**22.62 ± 5.17**

6.3. Dynamic LM evaluation

For the evaluation of the dynamic LM adaptation we have used the best configuration of parameters obtained in the previous section. The initial performance of our baseline system (i.e. without the dynamic LM adaptation) achieved a WER of 21.75. In Table 2 the results of the speech recognition performance when using the proposed approaches for the dynamic LM adaptation are compared. Although there is no significant difference between the **Soft-1** and the **Soft-2** approaches when comparing both variants (all topics and top-10), there is, in fact, a significant difference between the results obtained by the **Soft-1 - top 10** and the **Hard** approach, and even better results can be found when compared to the baseline approach. In general, with this soft integration we manage to reduce 11.59% of the initial WER.

Table 2: *Comparison between the word error rate obtained for different LM adaptation approaches*

LM Adaptation approach	WER	Relative Improv.
Baseline (no adaptation)	21.75 ± 0.26	
Hard	19.88 ± 0.25	8.59
Soft 1 - all	19.60 ± 0.24	9.89
Soft 1 - top 10	**19.23 ± 0.24**	**11.59**
Soft 2 - all	19.63 ± 0.24	9.74
Soft 2 - top 10	19.47 ± 0.24	10.48

7. Conclusions

In this paper we have presented a framework for dynamic language model adaptation based on topic identification. The results in the ASR task have shown that a small but statistically significant improvement in word error rate can be obtained by the adaptation strategie that has been proposed. Adapting the LM by taking only into consideration the most close topic, improves the baseline performance, but does not take advantage of all the sources of information available. The proposed criterion for selecting stopwords and the proposed weighting scheme have contributed in reducing the topic identification error.

8. Acknowledgements

The work leading to these results has received funding from TIMPANO (TIN2011-28169-C05-03), and MA2VICMR (Comunidad Autonóma de Madrid, S2009/TIC-1542) projects. It has also been supported by the European Union under grant agreement number 287678.

J.D. Echeverry-Correa also acknowledges the support from Colciencias and from the Universidad Tecnológica de Pereira, Colombia.

9. References

[1] Sebastiani, F. “Machine learning in automated text categorization”. ACM Computing Surveys (CSUR), 34(1):1–47, 2002.

[2] Deerwester, S. et al. “Indexing by latent semantic analysis”. Journal of the American Society for Information Science, 41 (6): 391–407, 1990.

[3] Rocchio, J. “Relevance Feedback in Information Retrieval”, in G. Salton [Ed], The SMART retrieval system: experiments in automatic document processing. Prentice-Hall, Inc., 1971.

[4] Lewis, D., and Ringuette, M. “A comparison of two learning algorithms for text categorization”. In Proc. of 1994 Symposium on Document Analysis and Information Retrieval, pages 81–93. 1994.

[5] Joachims, T. “Text categorization with Support Vector Machines: Learning with many relevant features”. Machine Learning: ECML-98. Springer Berlin Heidelberg, pages 137–142. 1998.

[6] Qiu, Y. “A keyword based strategy for spam topic discovery from the internet”. In Proc. of 2010 Fourth International Conference on Genetic and Evolutionary Computing (ICGEC)., pages 260–263. 2010.

[7] Cheng, N. and Chandramouli, R. and Subbalakshmi, K.P. “Author gender identification from text”. Digital Investigation, 8 (1):78–88, 2011.

[8] Günal, S. et al. “On feature extraction for spam e-mail detection”. Multimedia Content Representation, Classification and Security. Springer Berlin Heidelberg, 2006

[9] Maks, Isa and Vossen, Piek. “A lexicon model for deep sentiment analysis and opinion mining applications”. Decision Support Systems, 53: 680–688, 2012.

[10] Bellegarda, J. “Statistical language model adaptation: review and perspectives”. Speech communication, 42 (1): 93–108, 2004.

[11] Seymore, K. and Rosenfeld, R. “Using story topics for language model adaptation”. In Proc. of EUROSPEECH, 1997.

[12] Nanjo, H. and Kawahara, T. “Unsupervised language model adaptation for lecture speech recognition”. In ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.

[13] Liu, Y. and Liu, F. “Unsupervised language model adaptation via topic modeling based on named entity hypotheses”. In Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2008., pages 4921–4924, 2008.

[14] Lucas-Cuesta, J. et al. “On the dynamic adaptation of language models based on dialogue information”. Expert Syst. Appl., 40 (4): 1069–1085, 2013.

[15] Bellegarda, J.“Exploiting latent semantic information in statistical language modeling”. Proceedings of the IEEE, 88 (8): 1279–1296, 2000.

[16] Blei, D. and Ng, A. and Jordan, M. “Latent dirichlet allocation”. Journal of Machine Learning Research, 3: 993–1022, 2003.

[17] Chien, J.T. and Chueh, C.H. “Dirichlet class language models for speech recognition”. Audio, Speech, and Language Processing, IEEE Transactions on, 19 (3): 482–495, 2011.

[18] Florian, R. and Yarowsky, D. “ Dynamic nonlocal language modeling via hierarchical topic-based adaptation”. In Proc. of the ACL, pages 167–174, 1999.

[19] Iyer, R. and Ostendorf, M. “Modeling long distance dependence in language: Topic mixtures versus dynamic cache models”. Speech and Audio Processing, IEEE Transactions on, 7 (1): 30–39, 1999.

[20] Uysal, A. and Günal, S. “The impact of preprocessing on text classification”. Information Processing and Management, 50: 104–112, 2014.

[21] Padró, L. and Stanilovsky, E. “Freeling 3.0: Towards Wider Multilinguality”. Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, May 2012. ELRA.

[22] Dumais, S. “ Improving the retrieval of information from external sources”. Behavior Research Methods, Instruments, & Computers, 23 (2): 229–236, 1991.

[23] Federico, M. and Bertoldi, N. “Broadcast news LM adaptation over time”. Computer Speech & Language, 18 (4): 417–435, 2004.

[24] Chiu, H. and Chen, B. “Word topical mixture models for dynamic language model adaptation”. In Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2007, volume 4, pages 169–172, 2007.

[25] Koehn, P. “Europarl: A Parallel Corpus for Statistical Machine Translation”. In Proc. of the 10th Conference on Machine Translation (MT Summit’05), 2005.