Expert Systems with Applications 42 (2015) 101–112

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition

J.D. Echeverry-Correa *, J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba, J.M. Montero

Speech Technology Group, E.T.S.I. de Telecomunicación, Av. Complutense, 30. Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Article history:
Available online 10 August 2014

Keywords:
Language model adaptation
Topic identification
Automatic speech recognition

ABSTRACT

In this paper we present an efficient speech recognition approach for multitopic speech by combining information retrieval techniques and topic-based language modeling. Information retrieval based techniques, such as topic identification by means of Latent Semantic Analysis, are used to identify the topic in a recognized transcription of an audio segment. According to the confidence on the topics that have been identified, we propose a dynamic language model adaptation in order to improve the recognition performance in 'a two stages' automatic speech recognition system. The scheme used for the adaptation of the language model is a linear interpolation between a background general LM and a topic dependent LM. We have studied different approaches to generate the topic dependent LM and also for determining the interpolation weight of this model with the background model. In one of these approaches we use the given topic labels in the training dataset to obtain the topic models. In the other approach we separate the documents in the training dataset into topic clusters by using the *k-means* algorithm. For strengthening the adaptation models we also use topic identification techniques to group non topic-labeled documents from the EUROPARL text database in order to increase the amount of data for training specific topic based language models. For the evaluation of the proposed system we are using the Spanish partition of the European Parliament Plenary Sessions (EPPS) Database; we selected a subset of the database with 67 labeled topics for the evaluation. For the task of topic identification our experiments show a relative reduction in topic identification error of 44.94% when compared to the baseline method, the Generalized Vector Model with a classic TF–IDF weighting scheme. For the task of dynamic adaptation of LMs applied to ASR we have achieved a relative reduction in WER of 13.52% over a single background language model.

1. Introduction

The performance of a speech recognition system depends significantly on the similarity between the language model (LM) used and the task that is being addressed. This similarity is even more important in scenarios where the statistical properties of the language fluctuates throughout the time, for example in application domains involving spontaneous and multitopic speech. Over the last years there has been an increasing effort in improving the speech recognition systems for such domains. In spontaneous and multitopic speech the grammar models are changing constantly and therefore the performance of the speech recognition system will depend, among many other parts of the system, on its capacity to update or adapt the LMs. In this paper we propose a dynamic LM adaptation based on an information retrieval (IR) approach. We used IR techniques as a tool for identifying the topic of the speech, thus enabling the system to perform an adaptation of the language model according to the topic that is being discussed.

1.1. Related work in topic identification systems

With the rapid growth of the information available online, topic identification (TI) has become one of the key techniques in the field of text data classification. This technique addresses the problem of identifying which of a set of predefined topics or themes are present in a document. It is currently been applied in many contexts and disciplines, ranging from document indexing to automated metadata generation, document and messages filtering and, in an general sense, in applications that need document separation and organization. Depending on the research discipline in which this task is being carried out, topic identification is also known as *text categorization* (Manning, Raghavan, & Schütze, 2008), *text classification*

* Corresponding author.
E-mail addresses: jdec@die.upm.es (J.D. Echeverry-Correa), jfl@die.upm.es (J. Ferreiros-López), acoucheiro@die.upm.es (A. Coucheiro-Limeres), cordoba@die.upm.es (R. Córdoba), juancho@die.upm.es (J.M. Montero).

http://dx.doi.org/10.1016/j.eswa.2014.07.035
0957-4174/

(Baeza-Yates & Ribeiro-Neto, 2011) or *topic spotting* (Wiener, Pedersen, & Weigend, 1995).

A conventional topic identification framework consists of preprocessing, feature extraction, feature selection and classification stages. The preprocessing stage is usually composed of several tasks such as tokenization, stopword removal, stemming and term categorization. In the feature extraction stage it is common the use of the Vector Space Model (Salton, Yang, & Yu, 1975). This model makes use of the bag of words approach (Baeza-Yates & Ribeiro-Neto, 2011). The feature selection stage generally utilizes filter methods such as weighting schemes for the term and document frequency (Dumais, 1991), techniques for obtaining the mutual information of terms (Liu, Sun, Liu, & Zhang, 2009), information gain (Lee & Lee, 2006) and chi-square statistical metric (Chen & Chen, 2011). The classification stage uses well known techniques from the fields 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 (Sebastiani, 2002). Common approaches includes Latent Semantic Analysis (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990), Rocchio's method (Rocchio, 1971), Decision Trees (Lewis & Ringuette, 1994), naive Bayesian classifiers (Kim, Rim, Yook, & Lim, 2002) and Support Vector Machines (Joachims, 1998).

TI has been successfully applied in many contexts and disciplines, ranging from topic detection (Qiu, Xu, Li, & Li, 2010), automated metadata generation (Cheng, Chandramouli, & Subbalakshmi, 2011), document and messages filtering (Günal, Ergin, Gülmezoglu, & Gerek, 2006). It has also been applied in recently developed areas, such as sentiment analysis (Maks & Vossen, 2012), genre classification (Petrenz & Webber, 2011) and entity resolution (Pereira et al., 2009), among 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 (Bellegarda, 2004). 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.

1.2. Related work in language model adaptation strategies applied to an ASR system

Over the past 30 years, the amount and diversity of information available online has exponentially grown and this tendency appears to remain unaltered in the near future. As a result, the quality of language models has increased in certain domains where such data became available. Nevertheless, this behavior seems to be reaching an upper limit and it is possible that this continuous increase of information does not lead to any significant improvement in language models (Rosenfeld, 2000). For this reason it is important to find new sources of information that increase the capacity of the data to describe and model the type of language that is being used in an automatic speech recognition application.

For an optimal adaptation of language models in specific domains it is required that the system has a previous knowledge of data belonging to the same domain or, at least, to a related one. Precisely, the aim in statistical language model adaptation is to add new sources of information to the previously existent models with the objective of enriching them. The goal in LM adaptation is to reflect the changes that the language experiences when moving towards different domains or, as in some applications, when dealing with multiple speakers (Bellegarda, 2004).

LM adaptation techniques can be classified according to different criteria. Rosenfeld (2000) proposes a classification based in the domain of the data. Bellegarda (2001), on the other hand, suggests that the classification must be done according to the system requirements. However, there is not a distinct separation between these criteria. Nowadays LM adaptation techniques are jointly based not only on the origin and domain of the data but also on the system requirements and the objective of the adaptation scheme. Some LM adaptation approaches are based on the specific context of the task that they are addressing. In these approaches, new sources of information are used to generate a context-dependent LM which is then merged with a static LM. These new sources of information may come, for instance, from text categorization systems as in Seymore and Rosenfeld (1997), from speaker identification systems (Nanjo & Kawahara, 2003), from linguistic analysis systems (Liu & Liu, 2008) or from the application context itself (Lucas-Cuesta, Ferreiros, Fernández-Martínez, Echeverry, & Lutfi, 2013).

Other approaches are based on analysis and extraction of metadata, which means extraction of information that it is not explicitly described in the text. The topic of a document or semantic information related to it are examples of metadata. Latent Semantic Analysis (LSA) is an example of the type of techniques that exploits this kind of information. The work presented in this paper uses a LSA-based approach in order to obtain information about the topics that are being discussed in an audio segment. This information is then used to dynamically adapt the language model used by an ASR system with the objective of improving the recognition accuracy. Similar works have been proposed in the same domain. In Bellegarda (2000), 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) (Blei, Ng, & Jordan, 2003), have also been used for adapting LMs in an automatic speech recognition task (Chien & Chueh, 2011). A keyword extraction strategy to determine the LM to be used in a multi-stage speech recognition system is proposed in (Chen, Gauvain, Lamel, Adda, & Adda, 2001). In contrast to LSA, which do not explicitly consider the exact word order in the history context, in Liu, Gales, and Woodland (2013b) a history weighting function is used to model the change in word history during LM adaptation.

There are also techniques based on information originated from different subsystems or domains (cross adaptation). In Liu, Gales, and Woodland (2013a) a linear combination of two different subsystems (syllable and words) is performed to obtain an adapted LM. Another example is cross-lingual adaptation which uses information in a language to adapt LMs in another language (Kim & Khudanpur, 2004; Tam & Schultz, 2009).

All these techniques have one thing in common and that is the importance of the selection of reliable sources of information for refining the existent models. One of the most common sources of data for adapting language models is the internet. When using data available online it is possible to find information related to a large variety of topics. Nevertheless, this broad coverage leads to a loss of specificity when estimating LMs (Lucas-Cuesta et al., 2013). To avoid this drawback, 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 (Chen, Seymore, & Rosenfeld, 1998; Iyer & Ostendorf, 1999).

Techniques, in the line of Latent Semantic Analysis (Deerwester et al., 1990) such as Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation have been proposed to group documents into topic clusters. Topic based language models can be found in a broad spectrum of applications, such as in information retrieval systems as part of the ranking function (Zhai, 2008), in spoken dialogue systems for adapting the speech recognizer to the dialogue context (López-Cózar & Callejas, 2006; Lucas-Cuesta, 2013), in dynamic language model adaptation for Large Vocabulary Continuous Speech Recognition

Systems (LVCSR) (Gollan, Bisani, Kanthak, Schlüter, & Ney, 2005; Saon & Chien, 2012) and in Statistical Machine Translation for creating context dependent LMs form monolingual corpora (Lu, Wei, Fu, & Xu, 2012), among other applications.

1.3. General overview

The work presented in this paper comprises two main tasks: *topic identification* and *dynamic language model adaptation*. The main objective is improving the performance of an *ASR*. These tasks are combined in a *multipass recognition architecture* as represented in Fig. 1. The method we propose, performs a second decoding based on the topic detected in the initial recognition pass. The initial recognition is performed with a background language model built from the entire training set. Then, the topic is identified based on the results of the initial recognition pass. Using the information provided by the TI system and previously trained topic-specific language models, a dynamic adaptation of the background language model is performed. In the final stage, the adapted LM is then used to decode again the utterance.

1.4. This work

In this work we propose to dynamically adapt the LM used in an automatic speech recognition system. The underlying adaptation approach is a linear interpolation between a background LM and several topic-specific LMs. All of these models are estimated offline for each of the topics available (therefore they are static since no adaptation is performed), but the selection of the most suitable models to estimate a topic-based LM takes place at each decoding turn. Therefore, our topic-based LM is also dynamic.

The interpolation weights between the different components of the topic-based LM are also obtained dynamically on a turn basis, using the information provided by the topic identification system. One of our claims is that the system itself can provide the LM generation module enough information to obtain accurate interpolation weights that depend on the current recognition result. This way, the interpolation weights will depend on certain values obtained by the system, such as distance measures, similarities to the topics, among others.

We propose different approaches to obtain the LMs to be interpolated: they could be related only to the topic that has been identified with the highest similarity, or they could be related to a group of topics. In this regard we will also assess automatic clustering strategies, which are applied to the documents in the training dataset. The objective is to create automatic topic clusters in order to verify whether these clusters can improve the topic-based LM when compared with the manually assigned topic labels of the training dataset.

Finally, in our evaluation we will measure the improvement achieved not only for the topic identification system, but also for the automatic speech recognition system. That is, we will determine to what extent an improvement in the LMs leads to an improvement in the automatic speech recognition module of the system.

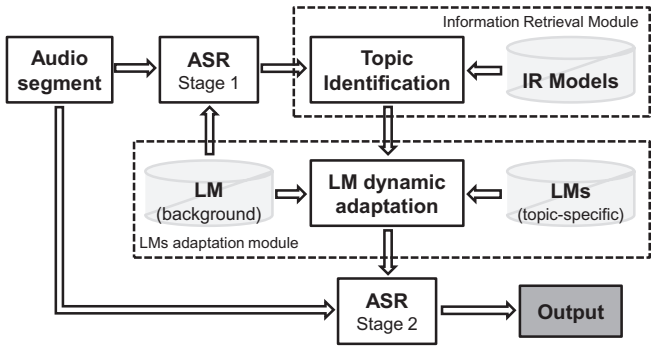

Fig. 1. Multipass recognition architecture.

This paper is structured as follows. Section 2 provides a description of the topic identification task and describes the preprocessing stages and the document representation. An entropy based weighting scheme and strategies for generating the stopword list are proposed in this section. Section 3 describes the language model adaptation procedures, with experimental results reported in Section 4. Finally, in Section 5 the results are discussed and the conclusions of this work are presented. In Section 6 the future research trends are described.

2. Topic identification

In this section we define the topic identification task and we formalize the steps involved in this process using classic IR models. This section is intended to focus on the Information Retrieval module of the architecture presented in Fig. 1. In a broad sense, topic identification is the task of automatically identifying which of a set of predefined topics are present in a document. Fig. 2 shows the steps that must be followed in order to perform TI as an information retrieval process. IR classic models assume that a document in a collection can be represented by a set of terms called index terms. This set can be automatically extracted from each document. In order to convert the natural language document to the vector space, there are some steps employed for data preprocessing. Their functions are described in the next subsections.

2.1. Preprocessing of documents and queries

The word inventory used in the construction of the IR model, that is, the set of index terms used to represent the documents and queries, can be obtained automatically by means of preprocessing the content of the documents in order to extract relevant terms that can be useful for discriminating between documents on different topics. The preprocessing stage allows to convert both, documents and queries, to a more concise and convenient format. This stage has a substantial impact on the success of the topic identification process (Uysal & Günal, 2014). Typical preprocessing steps include: structural processing, lexical analysis, tokenization, stopwords removal, stemming and term categorization. We provide a small description of each of these steps:

- *Structural processing* removes any structural element in the document such as labels for titles, sections, paragraphs or other kind of labels (which are common for XML and tagged markup languages).
- *Lexical analysis* is performed with the objective of converting digits into an alphabetical representation, treating hyphens, acronyms, punctuation marks and letter case.
- *Tokenization* is the process of breaking a stream of text into tokens that can be words, sentences, phrases, symbols or other meaningful elements according to the task that is under study. In a general way, tokenization occurs at the word level. The simplest way of tokenizing is to separate by white-space characters. Nevertheless there are some limitations, for example in word collocations like *"Castilla-La Mancha"* which must be considered as a single token. Overcoming these limitations depends on the availability of dictionaries or catalogs of predefined tokens. As far as we know, there is no an available token dictionary for the database we have used in the evaluation of this paper (described in Section 4.1). In the work presented, tokens like the one mentioned before were separated by each of its components and treated like separate words.

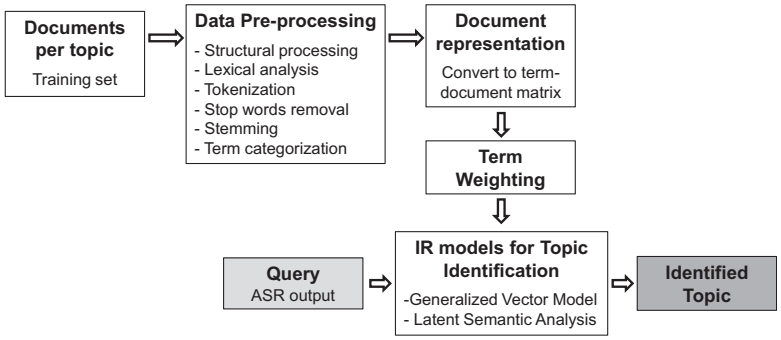

Fig. 2. Topic identification based on an IR approach.

- *Stopword removal* is performed to remove the words that have little lexical meaning and are too frequent among the documents in the collection. These words (also known as function words) are unlikely to contribute to the distinctiveness of the topics. Articles, prepositions, pronouns and conjunctions are examples of words that are typically included in the stopword list. An appropriate stopword list eliminates noise form the vocabulary, reduces the size of the indexing structure and contributes to speed up the clustering and decision processes. For these reasons, we have experimented with different strategies for creating the stopword list as we will describe in Section 2.7.
- *Stemming* refers to the transformation of a word to its stem or root form. It is done with the objective of removing prefixes, suffixes, plurals and morphological derivations of the words. Like stopword removal, it compresses the size of the indexing structure by reducing the number of distinct terms to index. For this step, we have used the Freeling Toolkit (Padró & Stanilovsky, 2012). Due to a few errors in the original stemming process, we have modified some of the stemming rules for the Spanish language of the toolkit.
- *Term categorization* consists of building a thesaurus for the terms in the word inventory. A thesaurus is mostly composed of synonyms and semantic related words, and reveals semantic relationships between terms. The motivation for building and using a thesaurus for indexing and searching is based on the idea of using a controlled and extended vocabulary. Despite the fact that this step increases the size of the indexing structure by adding additional terms, it presents important advantages such as reduction of noise, identification of index terms with a clear semantic meaning and retrieval based on concepts rather than on words.

In the work presented in this paper, we have followed all of the previously described preprocessing steps, except for the term categorization step.

2.2. Data representation for topic identification

The data representation is based on the widely known Vector Space Model (VSM) proposed by Salton et al. (1975). In this model, terms are assumed to be independent and both documents and queries can be represented as vectors in a space formed by the index-terms. This model is often used in information retrieval modeling because of its power contrasted to its conceptual simplicity. The representation of a document d_j considering each of the index-terms t_i can be quantified by the number of times the term appears in the document. For the whole document collection, this representation forms the Term-Document Matrix (TDM), which is:

$$TDM = \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & \dots & d_n \end{array} \\ \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & & c_{1,n} \\ c_{2,1} & c_{2,2} & c_{2,3} & & c_{2,n} \\ c_{3,1} & c_{3,2} & c_{3,3} & & c_{3,n} \\ & & & \vdots & \\ c_{m,1} & c_{m,2} & c_{m,3} & & c_{m,n} \end{bmatrix} \end{array} \begin{array}{c} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m \end{array}$$

where $V = \{t_1, t_2, t_3, \dots, t_m\}$ is the set of the index-terms that have been selected after the preprocessing stage, m is the number of index-terms that are considered, t_i are the index-terms for $1 \leqslant i \leqslant m$; and $D = \{d_1, d_2, d_3, \dots, d_n\}$ being the whole document collection containing n documents. Each element $c_{i,j}$ represents the number of times the term t_i appears in the document d_j. The query can also be represented, using the same indices, as a vector

$$\vec{q} = \begin{bmatrix} c_{1,q} & c_{2,q} & \dots & c_{m,q} \end{bmatrix}^T$$

where $c_{i,q}$ represents the number of times each term t_i appears in the query q.

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 topic identification capacity of the document it appears in Spärck-Jones (1972). A weighting scheme is usually composed of two different types of term weighting: local weights and global weights. Local weights are functions of the frequency of appearance of each term within a specific document. Global weights are functions of how many times a term appears in the entire collection.

By applying weighting schemes to the TDM, a new matrix Weighted TDM (WTDM) is obtained. In this matrix each element $w_{i,j}$ is composed by two components, usually computed as a product:

$$w_{i,j} = l_{i,j} \cdot g_i \tag{1}$$

where $l_{i,j}$ is the local weight of the term t_i in the document d_j and g_i is the global weight of the term t_i over all documents in the collection. These weights are computed not only for the documents in the collection but also to the query. The global weight applied to the query is the same applied to the terms in the documents. The weighted query is then a vector, in which each element is equal to:

$$w_{i,q} = l_{i,q} \cdot g_i \tag{2}$$

where $l_{i,q}$ is the local weight applied using the element $c_{i,q}$ in the query vector $\vec{q}$.

A widely used weighting scheme in IR applications, such as topic identification, is the TF–IDF (*Term Frequency–Inverse Document Frequency*) scheme. In this method the TF component (local weight) measures the relative frequency of a term in a document and the IDF component (global weight) measures the number of documents that use a term considering its specialty (higher weight is given to more specialized words that appear only in a few documents). This method is the one we have selected as the baseline weighting scheme for comparing the results obtained for the topic identification task in this paper. Local and global components of the TF–IDF scheme are calculated as follows:

$$l_{i,j} = \frac{c_{i,j}}{\sum_{k=1}^{m} c_{k,j}} \text{ for each document } d_j \tag{3}$$

$$g_i = \log_{10}\left(\frac{n}{df_i}\right) \tag{4}$$

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.

It has been suggested that the performance of term weighting schemes has not increased throughout the last years, and also that due to the fact that these schemes are strongly dependent on the nature of the data it is not definitive what form of term weighting scheme performs better than others (Chisholm & Kolda, 1999; Cummins, 2008). For that reason, 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* (see Dumais (1991) for more details on these schemes). Among these, *entropy* is the most sophisticated scheme. It is based on an information theory approach and it exploits the distribution of terms over documents (Dumais, 1991). The *entropy* weighting scheme is defined as follows:

$$\text{entropy} = 1 - \sum_{j=1}^{n} \frac{p_{i,j} \cdot \log_{10}(p_{i,j})}{\log_{10}(n)}, \quad \text{where} \quad p_{i,j} = \frac{l_{i,j}}{gf_i} \tag{5}$$

Where gf_i is the global frequency of the term t_i measured over all the documents in the collection.

2.4. Generalized Vector Model – GVM

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 topic identification task. All terms are assumed to be independent. The document d_j will be represented by the vector $\vec{d}_j$ (the jth column of the WTDM matrix) and the query will be represented by the vector $\vec{wq}$. Thus, we have:

$$\vec{d}_j = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ w_{3,j} \\ \vdots \\ w_{m,j} \end{bmatrix} \quad \text{y} \quad \vec{wq} = \begin{bmatrix} w_{1,q} \\ w_{2,q} \\ w_{3,q} \\ \vdots \\ w_{m,q} \end{bmatrix}$$

In this model, the similarity between a document $\vec{d}_j$ and a query $\vec{wq}$ can be computed using the cosine distance, calculated as follows:

$$sim\left(\vec{d}_j, \vec{wq}\right) = \cos\theta = \frac{\sum_{i=1}^{m}(w_{i,j} \times w_{i,q})}{\left(\sum_{i=1}^{m} w_{i,j}^2 \times \sum_{i=1}^{m} w_{i,q}^2\right)^{1/2}} \tag{6}$$

According to this distance, each document is ranked on how well aligned it is to the query. In one of the approaches considered in this work, the one that uses the given topic labels, 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.

2.5. Latent Semantic Analysis – LSA

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 topic identification. Another drawback is the *polysemy*, that is the property of some words to have several meanings. Counting all the senses of a word as one sense impairs the precision of a topic identification system.

Latent Semantic Analysis proposed by Deerwester et al. (1990), 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 and this consideration should improve over our first independence assumption. 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 similarity even if they 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 $WTDM_{m\times n}$ into the product of three matrices $T_{m\times m}, S_{m\times n}$ and $D^T_{n\times n}$, as follows:

$$WTDM = T \cdot S \cdot D^T \tag{7}$$

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 $TT^T = DD^T = I$. These matrices define the term and document vector spaces, respectively, in the latent semantic space. The diagonal matrix S contains the singular values of $WTDM$ in descending order such as $diag(S) = [\lambda_1, \lambda_2, \ldots, \lambda_r]$ where $\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0$ and $r \leqslant min(m, n)$ 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.

$$WTDM_{m\times n} \approx \widehat{T}_{m\times k} \cdot \widehat{S}_{k\times k} \cdot \widehat{D}^T_{k\times n} \tag{8}$$

From Eq. (7) it is possible to express the matrix D, which defines the document vector space in the latent semantic space, as:

$$D = WTDM^T \cdot T \cdot S^{-1} \tag{9}$$

Each of the rows of $WTDM^T$ represents a document of the collection. Therefore, by multiplying the $i-th$ row of $WTDM^T$ by $T \cdot S^{-1}$ we obtain the vector representation, in the latent semantic space, of the $i-th$ document. In order to compare the similarity between the query and the documents in the collection, the query must be represented as a vector in the same latent semantic space in which the documents are represented. Thus, the query can be treated as a pseudo-document and the vector representing the query can be created by multiplying the transpose of the weighted query vector $\vec{wq}$ by $T \cdot S^{-1}$. If fewer dimensions are being used, then the multiplica-

tion must be done by $\widehat{T} \cdot \widehat{S}^{-1}$, and the new query vector in the latent semantic space $\vec{q}_{new}$ can be computed as follows:

$$\vec{q}_{new} = \vec{wq}^T \cdot \widehat{T} \cdot \widehat{S}^{-1} \tag{10}$$

Then, the query, represented by $\vec{q}_{new}$, can be compared to all existing document vectors $\hat{d}_j$ in the matrix $\widehat{D}$, by computing the cosine distance between them as in (6). According to this distance, each document is ranked on how well aligned it is to the query.

2.6. Training set clustering

The corpus used for evaluation is provided with topic labels that were manually assigned. One of the approaches proposed in this paper is intended to evaluate the topic identification system based on those labels and to generate topic-based language models by grouping the documents which belong to each of those topics. Nevertheless, another goal of this work is to evaluate whether or not the use of these labels to produce LMs is optimal in terms of recognition accuracy. For this reason, in this section we present an alternative approach, in which the objective is to group the data in the training dataset into automatic topic clusters based on the semantic similarity between the documents.

Although several methods for clustering have been investigated for decades, accurately clustering documents without background information, nor predefined document categories or a given list of topics is still a challenging task (Xu, Liu, & Gong, 2003). To override this drawback, and in order to include some semantic information as our starting point, we propose to perform the clustering based on the projection of the documents in the latent semantic space. By clustering the documents in this space we expect to obtain a more cohesive association of the documents that are related by similar concepts. By doing this, the association of a document to a topic cluster will not depend on the manually assigned labels. This will increase the conceptual similarity between documents in the same cluster and allows us to expect an improvement of the LM within that cluster.

Among all the available techniques for data clustering, we decided to use the *k-means* algorithm due to its simplicity and its proven efficiency in text classification tasks (Sebastiani, 2002). To determine the optimal number of clusters we use the Silhouette Coefficient (SC) proposed by Rousseeuw (1987) that we are explaining below. This value is helpful in denoting the cohesiveness of the data in one cluster and the separation of data in one cluster from those in the other clusters. This coefficient has been used in text classification not only to analyze the quality of the clustering but also as a feature selection technique (Dey, Solorio, Gómez, & Escalante, 2011). In clustering tasks, the SC is calculated for each of the documents in the clusters in order to evaluate the clustering solution. Let $|c_k|$ denote the number of documents from the $k-th$ cluster and $dist(d_i, d_j) = 1 - \cos(d_i, d_j)$ indicate the distance between documents d_i and d_j. The Silhouette Coefficient $sc(d_i)$ for document d_i is computed as follows:

$$sc(d_i \in c_k) = \frac{b(d_i) - w(d_i)}{\max(b(d_i), w(d_i))} \tag{11}$$

where $w(d_i)$, the *within distance*, computes the average distance of the document d_i with all the documents in its own cluster, by using the following formula:

$$w(d_i \in c_k) = \frac{1}{|c_k| - 1} \sum_{\substack{\forall d_j \in c_k \\ d_j \neq d_i}} dist(d_i, d_j) \tag{12}$$

And $b(d_i)$, the *between distance*, is used to calculate the average distance of d_i with the documents of the other clusters. The minimum of all these average values is considered as $b(d_i)$, as shown in the following formula

$$b(d_i \in c_k) = \min_{j \neq k} \left[\frac{1}{|c_j|} \sum_{\forall d_m \in c_j} dist(d_i, d_m) \right] \tag{13}$$

The SC can have values from -1 to $+1$. Thus, if a document has SC value near $+1$, it implies that the *within distance* $w(d_i)$ is much smaller than the smallest *between distance* $b(d_i)$. In that case, it is possible to say that there appears to be little doubt that d_i has been assigned to a very appropriate cluster. It is also feasible to calculate the *overall average* SC $\bar{s}(k)$ for all the documents grouped in the k clusters. In general, different values for k will yield a different *overall average* $\bar{s}(k)$. Then, one way to select an appropriate value of k is to select that value of k for which $\bar{s}(k)$ is as large as possible. We have performed different clustering experiments on the training dataset by varying the number of clusters. At this point, it is important to notice that there are 67 hand-labeled topics in the original database, as we will describe later in this paper in Section 4.1. The largest value $\bar{s}(k)$ was found for $k = 20$. Therefore, the results for the clustering approach we will show in Section 4.3 were obtained by clustering the dataset in 20 clusters.

Once the clustering process has been done, each cluster will represent an "automatic topic" and the similarity between a query and a topic will be computed as the cosine distance between the query and the centroid of the cluster.

2.7. Generation of the stopword list

Stopword removal has proven to be one of the most important stages in the text preprocessing (Uysal & Günal, 2014). Not only the size of the index-terms inventory 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 computed the IDF by using a Term-Document Matrix composed of the 1802 documents in the training dataset and the 16,528 terms in the word inventory.

We performed different experiments on the Development set in order to find the optimal threshold. The lowest topic identification error on this dataset was obtained by setting the threshold to 0.4356, which means removing the terms that appear, at least, in 661 documents. The *List-2* has 445 stopwords.

3. Language model adaptation

In this section we present the motivation of the work regarding dynamic language model adaptation, as well as the approaches based on topic dependent language models after automatic topic identification.

3.1. Motivation

When analyzing spontaneous and multitopic spoken language the usage of content words is limited by 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, among other factors. There are words related to specific topics that would appear more frequently in a discourse related to those topics than in other audio segments. Therefore, the probability of usage of some words is increased depending on the topic of the speech.

As long as the sources of information for generating language models remain the same, the model will remain static. That is, regardless of the addressed topic, domain or style, the probability of events will not change. However, a static model is not the best option for modeling language in multitopic speech. As said before, in a natural conversation between humans, the topic, subject, genre, style, etc. changes often, and therefore the word usage changes accordingly. For this reason, the language model should be adapted dynamically (Kim, 2004).

In an ASR, dynamic LM adaptation becomes a strategy 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. This technique has shown to be effective in tasks that comprise a large amount of documents on different topics and also for processing data from multidomain applications such as broadcast news transcription (Federico & Bertoldi, 2004; Chiu & Chen, 2007).

3.2. Language model interpolation

LM interpolation is a simple and widely used method for combining and adapting language models. It consists of taking a weighted sum of the probabilities given by the component models. Let $P(w|h)$ be the probability of observing the word w given the previous sequence of words in its history h. 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 follows:

$$P_I(w|h) = (1-\lambda)P_B(w|h) + \lambda P_T(w|h) \tag{14}$$

where λ is the interpolation weight between both models, which has to fulfill the condition $0 \leqslant \lambda \leqslant 1$.

The scheme followed in this work for the generation of the LMs in the different stages of the process is presented in Fig. 3. In our approach, model interpolation occurs at two different levels: for generating the topic-based LM and, in a final instance, for creating the final dynamic LM used in the second decoding pass.

In the first level, model interpolation consists of generating a topic-based LM by selecting just one or combining several topic-specific LMs – $P_t(w|h)$. For the generation of the topic-specific language models, we have followed two approaches: (a) to group the whole training data according to the topic labels of the original documents and (b) to group the whole training dataset according to the clusters found in the clustering stage. For both strategies, if additional sources of information are available, an automatic labeling of the new data, based on the proposed clustering strategy, can be done in order to include them in each specific topic training data. In the second level, the topic-based LM is then merged with the background LM, generating the dynamic LM that the speech recognizer will use in the second decoding pass.

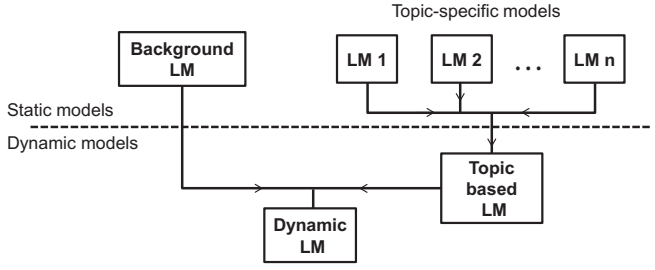

Fig. 3. Scheme of interpolation of language models.

The background model is a general model. It is trained with more, but not specific, data. On the other hand, the topic-based model is trained with more specific data related to the topic or topics we want to adapt the model to, thus enhancing the statistics of those words that better match the discussed topic.

In our case, the background model and the topic-specific models are static models. They are trained once and they 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 online during the processing of each audio segment.

LM adaptation strategies tested in this work differ in two ways: how to build or derive topic-based LMs and how to combine these models with the static background LM. In the next section we will address these issues and we will detail the interpolation schemes proposed for the dynamic LM adaptation.

3.3. Interpolation schemes

Two questions arise at this point. How to generate the topic-based model $P_T(w|h)$? and, how to determine the interpolation weight λ? 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_T 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 simply computed by considering the average difference in the unigram probabilities of both models.

$$\delta_T = \frac{1}{N}\sum_{\forall w_i \in P_T} |P_T(w_i) - P_B(w_i)| \tag{15}$$

where N is the number of unigrams in the topic-based LM $P_T(w|h)$. To ensure the interpolation weight fulfills the condition $0 \leqslant \lambda \leqslant 1$, 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.

$$\lambda = \frac{\delta_T}{\sum_{j=1}^{n}\delta_j} \tag{16}$$

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 dynamics of the right combination 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:

$$P_T(w|h) = \alpha_1 P_{t_1}(w|h) + \alpha_2 P_{t_2}(w|h) + \cdots + \alpha_k P_{t_k}(w|h) \tag{17}$$

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.

$$\alpha_i = \frac{sim(\vec{d}_i, \vec{q})}{\sum_{j=1}^{k} sim(\vec{d}_j, \vec{q})} \tag{18}$$

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 means of the sum, for all the topics, of the similarity measure provided by the TI system weighted by the relevance of the topic-specific LM. That is:

$$\lambda = \sum_{i=1}^{k} \alpha_i \cdot \frac{\delta_i}{\sum_{j=1}^{k} \delta_j} \tag{19}$$

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.

4. Experimental evaluation and implementation issues

Now that we have detailed our strategy for topic identification and the methodology for the dynamic language model adaptation, we present an experimental evaluation of them. Our evaluation focuses in two aspects: the evaluation of the topic identification approach itself and the evaluation of the dynamic language model adaptation by means of evaluating the performance of the two stages speech recognition system. First of all, the evaluation we have carried out for the TI task consists of identifying the topic that is discussed in each of the transcriptions provided by the first decoding pass (ASR Step 1). This evaluation is done for the original distribution of the database, that is the sentences with the topic labels that were manually assigned. The results presented are obtained by measuring the topic identification error. Additionally, we have evaluated the improvement in the ASR system by measuring the recognition performance in terms of word error rate. We have used the topic information conveyed by the speech to adapt the LM and recognize the same audio segment again in a second decoding pass. This way we can establish the first decoding pass as the baseline for the ASR task (without LM adaptation). All the confidence intervals presented in this work are stated at the 95% confidence level.

Regarding the implementation issues, the HTK Toolkit (Young et al., 2006) was used for training acoustic models and for the ASR decoding stages within the system architecture. The training dataset is composed of approximately 40 h of speech recordings. A detailed description of the composition of the database can be seen in Table 1. The SRILM Toolkit (Stolcke, 2002) was employed for creating and interpolating the language models that the system uses in both ASR stages. We use trigram models for both stages. The background LM is composed of nearly 2.8 M trigrams.

Before discussing the results obtained, we describe the dataset used for the evaluation.

4.1. Dataset

We have used the EPPS Spanish Database (*European Parliament Plenary Sessions*) of the TC-STAR Project (Mostefa, Hamon, Moreau, & Choukri, 2007) to study the performance of the proposed system. Due to the fact that the evaluation we are proposing for the topic identification system is focused on the automatic 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 originally defined 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 training, development and evaluation purposes defining new partitions.

Table 1
Details of the database used for the evaluation.

Language	Spanish
Gender of the speakers	Male and female
Domain	Political
Number of topics	67
Training set	21127 sentences grouped in 1802 speaker turns
Development set	2402 sentences grouped in 106 speaker turns
Lexicon size:	16528 words
Test *Set 1*	3738 sentences grouped in 252 audio segments
Test *Set 2*	Same 3738 sentences as in *Set 1* grouped in 754 audio segments

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 not less than a minute. We have applied two different criteria for audio segmentation. By these criteria we have generated two configurations for the set 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. Each of these segments belongs to one of the available 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 details of the database we have obtained after the new partitions are shown in Table 1.

For improving the coverage of the background language model and topic-based language models we have used the EUROPARL text database in addition to the existing sources of data for language modeling (Koehn, 2005). This database is composed of a parallel corpus in different languages for statistical machine translation. We have extracted approximately two million sentences in Spanish. We have taken advantage of this database in two distinct levels: (*i*) We added the extracted sentences to the text of the training set for generating the background language model. (*ii*) And we also used them for creating the topic-specific LMs. Using the TI 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 (or into one of the automatically derived clusters). Then, we used these automatic topic labeled sentences for improving the robustness of the topic-specific language models by merging the classified sentences of the EUROPARL database with the topic labeled sentences in the training set.

4.2. 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

Table 2
Topic identification error (T.I.E.) using GVM and LSA topic models approaches.

Topic identification approach	T.I.E. for *Set 1*	T.I.E. for *Set 2*
GVM + TF–IDF + SW (*List-1*) – Baseline	35.71 ± 5.91	84.08 ± 2.61
GVM + TF–IDF + SW (*List-2*)	34.13 ± 5.85	62.86 ± 3.44
GVM + TF–IDF + SW (*List-2*) + Stemming	36.11 ± 5.93	63.26 ± 3.44
GVM + TF-Entropy + SW (*List-2*)	34.52 ± 5.87	55.97 ± 3.54
LSA + TF–IDF + SW (*List-1*)	32.54 ± 5.78	47.22 ± 3.56
LSA + TF–IDF + SW (*List-2*)	30.56 ± 5.68	46.95 ± 3.56
LSA + TF–IDF + SW (*List-2*) + Stemming	34.13 ± 5.85	48.14 ± 3.56
LSA + TF-Entropy + SW (*List-2*)	30.16 ± 5.66	**46.29** ± **3.55**

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 configurations of the test dataset (*Set 1* and *Set 2*). We compared the performance with 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 2 shows the results obtained in topic identification using both GVM and LSA approaches.

We can separate the analysis of the results regarding each of the configurations of the test dataset used in the evaluation. It is important to notice that the *Set 1* contains less samples than *Set 2*, and therefore larger confidence intervals are obtained when analyzing the results for *Set 1*.

Despite the fact that there is a slight reduction in topic identification error for *Set 1* when comparing the baseline system to the LSA approach, this reduction is not statistically significant. Therefore the analysis of the results regarding the performance of the system for this particular configuration of the test dataset is not conclusive. On the other hand, for *Set 2* significant results are obtained when comparing not only both TI approaches, but also when comparing the Generalized Vector Model itself with the different stopword lists. By including the use of the extended list of stopwords (*List-2*) to the baseline GVM approach, a relative error reduction of 25.23% can be achieved. Although the TF-Entropy weighting scheme shows the minimum error for the GVM approach in the *Set 2*, this result is not significant when compared with the TF–IDF weighting scheme. Thus, this weighting scheme does not provide a significant improvement of the topic identification accuracy when used with the GVM.

Compared with the Generalized Vector Model, all the variants of the LSA approach improves the topic identification error for *Set 2*. Nevertheless, among them, no significant reduction can be obtained in the different configurations of the LSA approach. In this approach, neither the use of the *List-2* of stopwords, nor the stemming nor the TF-Entropy weighting scheme show a significant reduction when compared with the *List-1* of stopwords and the TF–IDF weighting scheme.

In general, when comparing the topic identification error obtained for both *Set 1* and *Set 2*, the minimum error was obtained when evaluating *Set 1*, because for larger audio segments, larger transcriptions are obtained and therefore, more semantic information is provided to the system.

Despite the fact that *stemming* reduces the number of index terms it does not provide a significant variation in the topic identification error for both sets. We believe that significant semantic differences can exist between a stem and its derivatives. Thus, by stemming we could be removing semantic information that might be useful for the topic identification objective.

When compared to the baseline, the best combination of parameters is obtained for the LSA model, using the *List-2* of stopwords and weighting the terms with TF-Entropy scheme. This configuration presents a relative improvement of 15.54% when compared to the baseline approach for the *Set 1* and a relative improvement of 44.94% for the *Set 2*.

4.3. 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. Now we describe the two strategies we have followed in the evaluation of the dynamic LM adaptation.

4.3.1. First strategy – using original topic labels for generating topic-specific LMs

First, in Table 3 we present the results of the speech recognition performance when using topic-specific language models trained according to the original topic labels of the documents in the training dataset. These results are shown for both configurations of the test dataset and for each of the proposed approaches for the dynamic LM adaptation.

In general terms, for both configurations of the test dataset, *Set 1* and *Set 2*, a statistically significant reduction of the word error rate, when compared to the baseline system, can be obtained by using the dynamic language model adaptation approaches proposed in this paper. However, among them, there are some differences that are worth mentioning. In *Set 1*, although there is not a 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 a relative improvement of 11.49% can be achieved when comparing this soft integration to the baseline approach.

The *Hard* approach takes only into consideration the most probable topic according to the TI system. Thus, in this approach the topic based LM is created by using only one of the topic-specific LMs. This significantly reduces the capacity of the Dynamic LM when compared to the *Soft-1 - top 10* approach.

In general, if we compare the results obtained when considering all the topics to the results obtained when considering only the top-10 topics, we can conclude that the system does not need to be too strict in selecting the closest topics. Actually, there is not a significant variation in the word error rate among both variants.

In *Set 2*, all the LM adaptation approaches present a similar result in WER and there are not significant differences between them.

In general, *Set 2* exhibits a lower word error rate when compared to *Set 1*. In *Set 2*, audio segments are equal or shorter than in *Set 1*. Thus, by processing shorter audio segments, a more refined LM adaptation can be done for each of them. Nevertheless, we believe that there must be a lower limit for the length of the segment, because it must contain, at least, enough information in order to perform the topic identification task. However, analyzing the effect of the length of the audio segment in the performance of the system is not one of the objectives of this work and it is just mentioned as an empirical conclusion. It should be covered in detail in future works.

In *Set 2*, the best LM adaptation approach is the *Soft-2 – top 10* approach. This means that the interpolation weight between the background general LM and the topic-based LM, computed automatically by the system, yields to an improvement in the recognition performance of the system. With this soft integration we manage to reduce 12.73% of the initial WER when compared to the baseline method.

It is important to notice that for this strategy, the topic-specific LMs have been generated using the original topic labels of the documents in the training dataset.

Table 3
Word Error Rate (WER) and Relative Improvement (Rel. Imp.) for the different LM adaptation approaches when training the topic-based LMs with the original topic labels of the documents.

Adaptation approach	SET 1		SET 2	
	WER	Rel. Imp. (%)	WER	Rel. Imp. (%)
Baseline (no adaptation)	21.75 ± 0.26		21.75 ± 0.26	
Hard	19.91 ± 0.25	8.45	19.27 ± 0.25	11.40
Soft 1 – all	19.58 ± 0.25	9.97	19.17 ± 0.25	11.86
Soft 1 – top 10	**19.25 ± 0.25**	**11.49**	19.08 ± 0.25	12.27
Soft 2 – all	19.62 ± 0.25	9.79	19.17 ± 0.25	11.86
Soft 2 – top 10	19.48 ± 0.25	10.43	**18.98 ± 0.25**	**12.73**

Table 4
Word Error Rate (WER) and Relative Improvement (Rel. Imp.) for the different LM adaptation approaches when performing the document clustering for the generation of the topic-based LMs.

Adaptation approach	SET 1		SET 2	
	WER	Rel. Imp. (%)	WER	Rel. Imp. (%)
Baseline (no adaptation)	21.75 ± 0.26		21.75 ± 0.26	
Hard	19.87 ± 0.25	8.64	19.23 ± 0.25	11.58
Soft 1 – all	19.60 ± 0.25	9.88	19.12 ± 0.25	12.09
Soft 1 – top 10	19.29 ± 0.25	11.31	18.96 ± 0.25	12.82
Soft 2 – all	19.26 ± 0.25	11.44	18.82 ± 0.25	13.47
Soft 2 – top 10	**19.21 ± 0.25**	**11.67**	**18.81 ± 0.25**	**13.52**

4.3.2. Second strategy – document clustering for generating topic-cluster LMs

We also performed the LM dynamic adaptation based on the automatic "topic clusters" created in the clustering process. By evaluating the clustering strategy, we obtain the results shown in Table 4.

By clustering the documents, the conceptual similarity is increased between documents in the same cluster and therefore an improvement of the LM within that cluster is achieved. The results obtained with this clustering strategy are promising (since in general the recognition performance tends to improve), but none of the results is statistically significant when compared to the previous strategy. However, there is a significant reduction of the word error rate when compared to the baseline system. This is true for the different dynamic LM adaptation approaches and for both configurations of the test dataset. The best result obtained with this clustering strategy outperforms the best result obtained with the previous one.

In *Set 1*, as in the previous strategy, there is not a significant difference between the Soft approaches, but there is a significant improvement when comparing their top 10 variants to the Hard approach. This shows that taking into consideration only one of the topic specific LMs for adapting the Dynamic LM is not optimal in terms of improving the speech recognition performance.

For *Set 1* the best result is obtained for the *Soft-2 – top 10* approach and a relative reduction of 11.67% of the initial WER is achieved.

As in the previous strategy, in *Set 2*, all the LM adaptation approaches present a similar result in WER and there are not significant differences between them. As well as in the previous strategy, *Set 2* exhibits a lower word error rate when compared to *Set 1*.

In *Set 2*, the best LM adaptation approach is the *Soft-2 – top 10* approach. When using this dynamic integration we manage to reduce 13.52% of the initial WER.

5. Discussion and conclusions

5.1. Discussion

Our contribution was focused on studying the capacity of the proposed system to dynamically adapt the LM according to the changes experienced by the grammar when dealing with spontaneous and multitopic domains. In this regard, a set of experiments were conducted to evaluate the performance of the dynamic LM adaptation techniques presented in this paper. Experimental results lead to the following findings:

- According to the results, the interpolation schemes presented as part of our dynamic language model interpolation strategies (Section 3.3) are useful to improve the performance of an ASR system within a multipass architecture.
- As it is already known a speech recognizer in conjunction with a topic identification system are able to capture additional information relevant to the topic that is being discussed in a decoding turn. Our contribution is intended to make use of this information in order to perform a dynamic adaptation of the language models used by an ASR system.
- Results presented in Tables 3 and 4 show that the performance of the ASR is enhanced when adapting LMs to shorter audio segments and significant reduction of word error rate can be achieved when compared to larger audio segments. This may seem counterintuitive because one could expect to obtain more useful information from larger audio segments. Nevertheless, the language model adapts better to short segments even though the topic identification error is increased for those segments.

5.2. Conclusions

In this paper we proposed a framework to create topic-based language models and to dynamically adapt the language model used in the final stage of a multipass recognition architecture. In this sense we addressed the tasks of automatic topic identification and document clustering for enhancing the language model adaptation in order to improve the performance of the automatic speech recognition system.

Regarding the document preprocessing, the proposed criterion for selecting stopwords has contributed in reducing the topic identification error.

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 loss 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, by 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.

Regarding the interpolation schemes, we have shown that 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. In this sense to compute the interpolation weight based on the similarity of the audio segment to several topics, as it is done in the soft approach, increases the sources of information and therefore contributes in the dynamic adaptation of the language models.

The strategy for clustering the documents into new automatic "topic clusters" allows us to improve the cohesiveness of the documents that are related to similar concepts. The topic based LMs created by following this strategy improve the speech recognition performance of the proposed architecture.

The results in the ASR task have shown that a small but statistically significant improvement in word recognition accuracy can be obtained in this hard task where topics do not change as much as in a conversational task. This is achieved by dynamically adapting a topic-dependent language model with interpolation weights computed after the first pass of a multipass recognition strategy.

We are not specifically analyzing the influence of the length of the audio segment in the adaptation of the LM, but according to the results shown for both configurations of the dataset, it can be suggested that there is a relation between the capacity of the LM adaptation system and the length of the analyzed audio segment.

6. Future work

Here are some of the research guidelines that we are currently working on regarding topic identification and dynamic LM adaptation for speech recognition.

As future work, we plan to study a new approach for topic clustering of the documents, by means of techniques such as PLSA (Probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation). These techniques would be studied in order to evaluate if different clustering strategies for non labeled documents are useful to obtain a more suitable adaptation of language models.

The database we have used for evaluating this work contains documents belonging to the same domain: political speeches. In the future we plan to use the methodology described in this paper to address the same problem in multidomain databases.

Regarding the preprocessing stages and the vocabulary selection we believe that better results can be achieved by exploring deeper relationships between the terms. We could use not only the list of index-terms extracted from the documents but also their morphological information by using a thesaurus for constructing more robust lists of index-terms. It is also worth asking whether a more detailed study and a selective application of stemming rules could improve the contribution of this preprocessing stage in the overall performance of the system.

We are aware that for including the system in a real application, the system proposed is not able to select which of the recognition stages performs better. Although the adaptation scheme used for the second stage outperforms the first stage in a general sense, sometimes the first stage has a lower word error rate when compared to the second stage. For that reason, we plan to include confidence measures for speech recognition in order to evaluate the reliability of recognition results in both stages and select the best output.

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.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.eswa.2014.07.035.

References

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval* (2nd. edition.). Pearson Education Ltd..

Bellegarda, J. R. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE, 88*(8), 1279–1296.

Bellegarda, J. R. (2001). An overview of statistical language model adaptation. Invited Lecture. In *Adaptation-2001* (pp. 165–174).

Bellegarda, J. R. (2004). Statistical language model adaptation: Review and perspectives. *Speech Communication, 42*, 93–108.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Chen, L., Gauvain, J., Lamel, L., Adda, G., & Adda, M. (2001). Using information retrieval methods for language model adaptation. In *Proceedings of the 7th european conference on speech communication and technology (EUROSPEECH'01)* (pp. 255–258).

Chen, S. F., Seymore, K., & Rosenfeld, R. (1998). Topic adaptation for language modeling using unnormalized exponential models. In *Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing (ICASSP'98)* (Vol. 2, pp. 681–684).

Chen, Y., & Chen, M. C. (2011). Using chi-square statistics to measure similarities for text categorization. *Expert Systems with Applications, 38*(4), 3085–3090.

Cheng, N., Chandramouli, R., & Subbalakshmi, K. (2011). Author gender identification from text. *Digital Investigation, 8*(1), 78–88.

Chien, J., & Chueh, C. (2011). Dirichlet class language models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(3), 482–495.

Chisholm, E., & Kolda, T. G. (1999). New term weighting formulas for the vector space method in information retrieval. Technical report, Oak Ridge National Laboratory, USA.

Chiu, H., & Chen, B. (2007). Word topical mixture models for dynamic language model adaptation. In *Proceedings of the 2007 IEEE international conference on acoustics, speech and signal processing (ICASSP'07)* (Vol. 4, pp. 169–172).

Cummins, R. (2008). The evolution and analysis of term-weighting schemes in information retrieval (Ph.D. thesis). National University of Ireland.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Dey, D., Solorio, T., Gómez, M., & Escalante, H. (2011). Instance selection in text classification using the silhouette coefficient measure. In *Proceedings of the 10th Mexican international conference on artificial intelligence (MICAI'11)* (pp. 357–369).

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers, 23*(2), 229–236.

Federico, M., & Bertoldi, N. (2004). Broadcast news LM adaptation over time. *Computer Speech & Language, 18*(4), 417–435.

Gollan, C., Bisani, M., Kanthak, S., Schlüter, R., & Ney, H. (2005). Cross domain automatic transcription on the TC-STAR EPPS corpus. In *Proceedings of the 2005 IEEE international conference on acoustics, speech and signal processing (ICASSP'05)* (pp. 825–828).

Günal, S., Ergin, S., Gülmezoglu, M. B., & Gerek, O. N. (2006). On feature extraction for spam e-mail detection. In *Proceedings of the international workshop on multimedia content representation, classification and security (MRCS'06)* (pp. 635–642).

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'99)* (pp. 50–57).

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

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European conference on machine learning (ECML'98)* (pp. 137–142).

Kim, S., Rim, H., Yook, D., & Lim, H. (2002). Effective methods for improving Naive Bayes text classifiers. In *Proceedings of the 7th Pacific rim international conference on artificial intelligence (PRICAI'02)*.

Kim, W. (2004). Language model adaptation for automatic speech recognition and statistical machine translation (Ph.D. thesis). The Johns Hopkins University.

Kim, W., & Khudanpur, S. (2004). Cross-lingual latent semantic analysis for language modeling. In *Proceedings of the 2004 IEEE international conference on acoustics, speech and signal processing (ICASSP'04)*, (Vol. 1, pp. 257–260).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th conference on machine translation (MT Summit'05)*.

Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing and Management, 42*(1), 155–165.

Lewis, D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of the 1994 symposium on document analysis and information retrieval* (pp. 81–93).

Liu, H., Sun, J., Liu, L., & Zhang, H. (2009). Feature selection with dynamic mutual information. *Pattern Recognition, 42*(7), 1330–1339.

Liu, X., Gales, M., & Woodland, P. (2013a). Language model cross adaptation for {LVCSR} system combination. *Computer Speech & Language, 27*(4), 928–942.

Liu, X., Gales, M., & Woodland, P. (2013b). Use of contexts in language model interpolation and adaptation. *Computer Speech & Language, 27*(1), 301–321 [Special issue on paralinguistics in naturalistic speech and language].

Liu, Y., & Liu, F. (2008). Unsupervised language model adaptation via topic modeling based on named entity hypotheses. In *Proceedings of the 2008 IEEE international conference on acoustics, speech and signal processing (ICASSP'08)* (pp. 4921–4924).

López-Cózar, R., & Callejas, Z. (2006). Combining language models in the input interface of a spoken dialogue system. *Computer Speech & Language, 20*(4), 420–440.

Lu, S., Wei, W., Fu, X., & Xu, B. (2012). Translation model based cross-lingual language model adaptation: From word models to phrase models. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL'12)* (pp. 512–522).

Lucas-Cuesta, J. M. (2013). Contributions to the contextualization of human–machine spoken interaction systems (Ph.D. thesis). Department of Electronic Engineering, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid.

Lucas-Cuesta, J. M., Ferreiros, J., Fernández-Martínez, F., Echeverry, J. D., & Lutfi, S. L. (2013). On the dynamic adaptation of language models based on dialogue information. *Expert Systems with Applications, 40*(4), 1069–1085.

Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems, 53*(4), 680–688.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mostefa, D., Hamon, O., Moreau, N., & Choukri, K. (2007). Evaluation report for the technology and corpora for speech to speech translation (TC-STAR Project). deliverable n. 30.

Nanjo, H., & Kawahara, T. (2003). Unsupervised language model adaptation for lecture speech recognition. In *Proceedings of the 2003 ISCA & IEEE workshop on spontaneous speech processing and recognition (SSPR'03)*.

Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 2012 language resources and evaluation conference (LREC'12)*.

Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., Laender, A. H., Gonçalves, M. A., & Ferreira, A. A. (2009). Using web information for author name disambiguation. In *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries, JCDL '09* (pp. 49–58).

Petrenz, P., & Webber, B. (2011). Stable classification of text genres. *Computational Linguistics, 37*(2), 385–393.

Qiu, Y., Xu, Y., Li, D., & Li, H. (2010). A keyword based strategy for spam topic discovery from the internet. In *Proceedings of the fourth international conference on genetic and evolutionary computing (ICGEC'10)*.

Rocchio, J. (1971). *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, Inc. [chapter Relevance Feedback in Information Retrieval].

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE, 88*(8), 1270–1278.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Salton, G., Yang, C., & Yu, C. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science, 26*(1), 33–44.

Saon, G., & Chien, J. (2012). Large-vocabulary continuous speech recognition systems: A look at some recent advances. *IEEE Signal Processing Magazine, 29*(6), 18–33.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34*(1), 1–47.

Seymore, K., & Rosenfeld, R. (1997). Using story topics for language model adaptation. In *Proceedings of the 5th European conference on speech communication and technology (EUROSPEECH'97)*.

Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Stolcke, A. (2002). SRILM-An extensible language modeling toolkit. In *3rd international conference on speech and language technology (INTERSPEECH'02)*.

Tam, Y., & Schultz, T. (2009). Incorporating monolingual corpora into bilingual latent semantic analysis for crosslingual LM adaptation. In *Proceedings of the 2009 IEEE international conference on acoustics, speech and signal processing (ICASSP'09)* (pp. 4821–4824).

Uysal, A. K., & Günal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management, 50*(1), 104–112.

Wiener, E., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. In *Proceedings of the 4th annual symposium on document analysis and information retrieval (SDAIR'95)*.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 267–273).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2006). The HTK book. Cambridge University Engineering Department, 12.

Zhai, C. (2008). Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval, 2*(3), 137–213.