Universidad Tecnológica de Pereira

Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación

Contributions to speech analytics based on speech recognition and topic identification

Julian David Echeverry Correa

Editorial UTP

Colección Tesis Laureadas

Julián David Echeverry Correa, (Pereira, Risaralda, Colombia, 1981). Doctor en Ingeniería de Sistemas Electrónicos y Máster en Sistemas Electrónicos de la Universidad Politécnica de Madrid. Magíster en Ingeniería Eléctrica de la Universidad Tecnológica de Pereira. Ingeniero Electrónico de la Universidad Nacional de Colombia. Profesor Asociado adscrito a la Facultad de Ingenierías de la Universidad Tecnológica de Pereira.

Ha publicado artículos en revistas especializadas nacionales e internacionales.

Pertenece al Grupo de Investigación en Automática.

jde@utp.edu.co

Contributions to speech analytics
Based on speech recognition and topic identification

Julian David Echeverry Correa

Universidad Tecnológica de Pereira

Colección Tesis Laureadas
Facultad de Ingenierías
2016

Echeverry Correa, Julián David
Contributions to speech analytics based on speech recognition and topic identification / Julián David Echeverry Correa. -- Pereira : Editorial Universidad Tecnológica de Pereira, 2016.

143 páginas : ilustrado. – (Colección Tesis Laureadas)

ISBN: 978-958-722-267-8

1. Reconocimiento automático de la voz 2. Teoría de las señales (telecomunicaciones) 3. Sistemas digitales 4. Procesamiento de señales 5. Filtros electrónicos 6. Sistemas de procesamiento de la voz.

CDD. 621.3822

Esta Tesis se desarrolló en el Departamento de Ingeniería Electrónica de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid, España. Fue financiada por el Grupo de Tecnología del Habla de la UPM, por Colciencias y por la Universidad Tecnológica de Pereira.

Primera edición

Universidad Tecnológica de Pereira
Vicerrectoría de Investigaciones, Innovación y Extensión
Editorial Universidad Tecnológica de Pereira
Pereira, Colombia

Coordinador editorial:
Luis Miguel Vargas Valencia
luismvargas@utp.edu.co
Teléfono 3137381
Edificio 9, Biblioteca Central "Jorge Roa Martínez"
Cra. 27 No. 10-02 Los Álamos
Pereira, Colombia
www.utp.edu.co

Montaje y producción:
Centro Recursos Informáticos y Educativos
Universidad Tecnológica de Pereira
Pereira

A mis padres por creer y seguir creyendo.

A mi esposa Beatriz por haberle dado un nuevo sentido a mi vida,
a ella especialmente le dedico esta Tesis.
Por ser mi compañera en la vida, por su paciencia,
comprensión y sobre todo por su amor.

“Parecían dos niños, me dijo. Y esa reflexión la asustó,
pues siempre había pensado que sólo los niños son capaces de todo”.

Gabriel García Márquez - Crónica de una muerte anunciada - 1981

“Science is unreasonably effective, it’s generated knowledge beyond all expectation. It’s also delivered perspective. Yes, we are an insignificant speck in an infinite universe, but we’re also rare. And because we’re rare, we’re valuable. So, what are we to do to secure our future? Well, we must learn to value the acquisition of knowledge for its own sake, and not just because it grows our economy or allows us to build better bombs. We must also learn to value the human race and take responsibility for our own survival. Why? Because there’s nobody else out there to value us or to look after us. And finally, most important of all, we must educate the next generation in the great discoveries of science and we must teach them to use the light of reason to banish the darkness of superstition, cos if we do that, then at least there’s a chance that this universe will remain a human one”

Brian Cox - BBC Series Human Universe - 2014

Agradecimientos

Quiero agradecer a todas y a cada una de las personas que me apoyaron en el transcurso de esta Tesis. Primero que todo, quiero expresar mi más profunda gratitud a mi Director de Tesis, Profesor Javier Ferreiros. Quiero agradecerle el haberme hecho crecer como investigador. Sin sus palabras y su orientación, esta Tesis no habría sido posible.

Quiero agradecer también a los demás profesores del Grupo de Tecnología del Habla: Manolo, Rubén, Ricardo y muy especialmente a Juancho. Su invitación a visitar el grupo, allá en 2008, fue el comienzo de todo este camino. Gracias a todos por su constante ayuda, sus críticas constructivas y por haber compartido conmigo en tantas ocasiones sus valiosos comentarios acerca de esta Tesis.

Quiero dar también las gracias a Ascensión, Doroteo y a Rubén, que revisaron esta Tesis en la fase de prelectura. Sus comentarios y consejos han permitido que esta Tesis tome una mejor forma y que el resultado final sea notablemente mejor.

Gracias a Fernando González por haberme permitido conocer no sólo al profesor sino también a la gran persona que es él. Por sus palabras, por su ayuda y por haber estado ahí en momentos difíciles, mil gracias Fernando.

Gracias a mis compañeros de despacho y amigos Syaheerah y Juanma. Syaheerah, mi hermana mayor en la distancia, la voz líder de nuestras sesiones de canto. Juanma, la amabilidad hecha persona, siempre dispuesto a echar un cable cuando lo necesitabas. Qué solo se sentía el despacho sin ellos. Gracias a Robert por las sesiones de filosofía que compartíamos. Igual me quedaba con las energías renovadas o igual me quedaba hecho polvo por sus comentarios. Aún así, las charlas con él eran catarsis necesarias. A Luisfer gracias también por sus palabras y por evocar constantemente tantos recuerdos de Colombia. Gracias por haber sido ese puente que, incluso en el trabajo, no me dejaba olvidar de dónde veníamos.

A Fer, gran amigo, excelente profesor y compañero e incomparable persona. No olvidaré tus clases hasta las tantas en pleno mundial de fútbol. De verdad, no lo olvidaré. Fer, no pierdas la esperanza que estoy seguro que algún día veremos levantar la copa a tu Almería y a mi querido Deportivo Pereira.

Gracias también a mis compis de labo, a Christian, Jaime y especialmente a Vero. Hemos compartido muy buenos momentos imposibles de olvidar. A Silvia, gracias por toda la ayuda que me ha prestado. Y a Mariano, mil gracias por todo el papeleo que me ha ahorrado y por la gestión de tanto trámite relacionado con la Tesis.

A Lore y a Rafa, amigos que conocí en la Escuela y cuya amistad permanece intacta a pesar del paso del tiempo. A Ramona, por ser mi mejor amiga, por saberlo todo sobre mí. Por los viajes, por las fiestas, por los buenos recuerdos. Gracias Rami.

A toda la Minipandi "gracias" por echar por tierra mis esfuerzos de hacer dieta. Con vosotros es imposible no comer hasta la saciedad. Gracias por vuestro cariño y por haber hecho que mis años en Madrid estuvieran llenos de buenísimos recuerdos e historias. A todos vosotros, Nat, Vero, Albert, Mori, Bel, Jose, Moni, Chiquitín, Sandra y David, mil gracias por ser las excelentes personas que sois. Os quiero ver en Colombia muy pronto.

A Sagrario, Julian, Isa y Diego, mil gracias porque desde el primer momento me habéis hecho sentir parte de vuestra familia. Gracias por vuestra constante ayuda y por todo vuestro cariño. Os extrañaré muchísimo (y os esperaremos en Colombia una vez al año, como mínimo).

Para mis padres no tengo suficientes palabras para agradecerles todo lo que han hecho por mí. Nada de esto hubiera sido posible sin su constante esfuerzo, dedicación, paciencia, comprensión y amor. Gracias por haberme apoyado en todas y cada una de las decisiones que he tomado en mi vida. Gracias por haber confiado y por haber creído en mí. A mis padres se lo debo todo.

A mi hermanito Mau, gracias por siempre hacerme ver el lado bueno bueno y alegre de la vida. Mil gracias por ser la maravillosa persona que eres. Gracias anticipadas también por los sobrinos que me darás.

Gracias a mi profesor de cálculo Francisco Javier Acosta "Pacho", quien sembró la semilla de mi interés por la Ingeniería. A pesar de que la violencia y la intolerancia se lo llevaron de este mundo, su recuerdo sigue vivo en muchos de nosotros, sus estudiantes. Estoy seguro que su ejemplo de dedicación como profesor seguirá vivo por muchos años más.

Gracias totales a mi primer y mejor amigo, Federico. Han sido 28 años de amistad llenos de buenos recuerdos e historias que resisten al paso del tiempo. Gracias por todas las llamadas de cumpleaños y por haberme ayudado a encontrar el mejor anillo de compromiso que pude imaginar para mi esposa.

Quiero también agradecer a la Universidad Tecnológica de Pereira, y especialmente al profesor Alberto Ocampo por su constante apoyo.

Gracias a Alejandro por su buen trabajo, su constante esfuerzo y sus interminables ganas de hacer las cosas siempre bien y mejor. Hay un futuro brillante esperando por ti, te lo aseguro.

Y finalmente, quiero agradecer a mi Tesis en sí. Si no fuera por esta aventura no habría conocido a la mujer más maravillosa, tierna y encantadora que hay. A mi esposa Beatriz, gracias por todo tu amor, por toda tu paciencia y por todo el apoyo que me has dado en estos años y que me han permitido llevar a buen puerto este trabajo. A ti Bea, especialmente a ti, gracias por todo.

Symbols and abbreviations

d_j, $\vec{d}_j$	j-th document, vector representation of the j-th document
q_i, $\vec{q}$	vector representation of the query
t_i	i-th index-term
$c_{i,j}$	Raw frequency of the i-th term in the j-th document
V, $\|V\|$	Term inventory, size of the term inventory
m	Number of index-terms
n	Number of documents in the collection
$l_{i,j}$	Local weight applied to the i-th term in the j-th document
g_i	Global weight applied to the i-th term
$tf_{i,j}$	Term frequency of the i-th term in the j-th document
df_i	Document frequency of the i-th term
gf_i	Global frequency of the i-th term
$\overrightarrow{wd}$, $\overrightarrow{wq}$	Weighted document vector, weighted query vector
$P(X)$	Probability of X
$P(X\|Y)$	Conditional probability of X given the occurrence of Y
A, $\|A\|$	Set of predefined classes (categories), size of this set (i.e. number of classes)
$P(\mathtt{w}\|\mathtt{h})$	Probability of word $\mathtt{w}$ given the history $\mathtt{h}$
$\vec{C}_i$	i-th centroid vector
ASR	Automatic Speech Recognition
BOW	*bag-of-words*
EPPS	European Parliament Plenary Sessions
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LM	Language Model
LSA	Latent Semantic Analysis
LVCSR	Large-Vocabulary Continuous Speech Recognition
ML	Machine Learning
NLP	Natural Language Processing
PP	Perplexity
SC	Silhouette Coefficient
TDM	Term-Document Matrix
VSM	Vector Space Model
WER	Word Error Rate

Abstract

The last decade has witnessed major advances in speech recognition technology. Today's commercial systems are able to recognize continuous speech from numerous speakers, with acceptable levels of error and without the need for an explicit adaptation procedure. Despite this progress, speech recognition is far from being a solved problem. Most of these systems are adjusted to a particular domain and their efficacy depends significantly, among many other aspects, on the similarity between the language model used and the task that is being addressed. This dependence 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 enhancing the speech recognition systems for such domains. This has been done, among other approaches, by means of techniques of automatic adaptation. These techniques are applied to the existing systems, specially since exporting the system to a new task or domain may be both time-consuming and expensive.

Adaptation techniques require additional sources of information, and the spoken language could provide some of them. It must be considered that speech not only conveys a message, it also provides information on the context in which the spoken communication takes place (e.g. on the subject on which it is being talked about). Therefore, when we communicate through speech, it could be feasible to identify the elements of the language that characterize the context, and at the same time, to track the changes that occur in those elements over time. This information can be extracted and exploited through techniques of information retrieval and machine learning. This allows us, within the development of more robust speech recognition systems, to enhance the adaptation of language models to the conditions of the context, thus strengthening the recognition system for domains under changing conditions (such as potential variations in vocabulary, style and topic).

In this sense, the main contribution of this Thesis is the proposal and evaluation of a framework of **topic-motivated contextualization** based on the **dynamic and non-supervised adaptation of language models** for the enhancement of an automatic speech recognition system. This adaptation is based on an combined approach (from the perspective of both information retrieval and machine learning fields) whereby we identify the topics that are being discussed in an audio recording. The topic identification, therefore, enables the system to perform an adaptation of the language model according to the contextual conditions. The proposed framework can be divided in two

major systems: a *topic identification system* and a *dynamic language model adaptation system*.

This Thesis can be outlined from the perspective of the particular contributions made in each of the fields that composes the proposed framework:

- Regarding the *topic identification system*, we have focused on the enhancement of the document preprocessing techniques in addition to contributing in the definition of more robust criteria for the selection of *index-terms*.
  - Within both information retrieval and machine learning based approaches, the efficiency of topic identification systems, depends, to a large extent, on the mechanisms of preprocessing applied to the documents. Among the many operations that encloses the preprocessing procedures, an adequate selection of *index-terms* is critical to establish conceptual and semantic relationships between terms and documents. This process might also be weakened by a poor choice of *stopwords* or lack of precision in defining stemming rules. In this regard we compare and evaluate different criteria for preprocessing the documents, as well as for improving the selection of the *index-terms*. This allows us to not only reduce the size of the indexing structure but also to strengthen the topic identification process.
  - One of the most crucial aspects, in relation to the performance of topic identification systems, is to assign different weights to different terms depending on their contribution to the content of the document. In this sense we evaluate and propose alternative approaches to traditional weighting schemes (such as *tf-idf*) that allow us to improve the specificity of terms, and to better identify the topics that are related to documents.
- Regarding the *dynamic language model adaptation*, we divide the contextualization process into different steps.
  - We propose supervised and unsupervised approaches for the generation of topic-based language models. The first of them is intended to generate topic-based language models by grouping the documents, in the training set, according to the original topic labels of the corpus. Nevertheless, a goal of this Thesis is to evaluate whether or not the use of these labels to generate language models is optimal in terms of recognition accuracy. For this reason, we propose a second approach, an unsupervised one, in which the objective is to group the data in the training set into automatic topic clusters based on the semantic similarity between the documents. By means of clustering approaches we expect to obtain a more cohesive association of the documents that are related by similar concepts, thus improving the coverage of the topic-based language models and enhancing the performance of the recognition system.
  - We develop various strategies in order to create a context-dependent language model. Our aim is that this model reflects the semantic context of the current utterance, i.e. the most relevant topics that are being discussed.

This model is generated by means of a linear interpolation between the topic-based language models related to the most relevant topics. The estimation of the interpolation weights is based mainly on the outcome of the topic identification process.

- Finally, we propose a methodology for the dynamic adaptation of a background language model. The adaptation process takes into account the context-dependent model as well as the information provided by the topic identification process. The scheme used for the adaptation is a linear interpolation between the background model and the context-dependent one. We also study different approaches to determine the interpolation weights used in this adaptation scheme.

Once we defined the basis of our topic-motivated contextualization framework, we propose its application into an automatic speech recognition system. We focus on two aspects: the contextualization of the language models used by the system, and the incorporation of semantic-related information into a topic-based adaptation process. To achieve this, we propose an experimental framework based in 'a two stages' recognition architecture. In the first stage of the architecture, Information Retrieval and Machine Learning techniques are used to identify the topics in a transcription of an audio segment. This transcription is generated by the recognition system using a background language model. According to the confidence on the topics that have been identified, the dynamic language model adaptation is carried out. In the second stage of the recognition architecture, an adapted language model is used to re-decode the utterance.

To test the benefits of the proposed framework, we carry out the evaluation of each of the major systems aforementioned. The evaluation is conducted on speeches of political domain using the EPPS (*European Parliamentary Plenary Sessions*) database from the European TC-STAR project. We analyse several performance metrics that allow us to compare the improvements of the proposed systems against the baseline ones.

Resumen

La última década ha sido testigo de importantes avances en el campo de la tecnología de reconocimiento de voz. Los sistemas comerciales existentes actualmente poseen la capacidad de reconocer habla continua de múltiples locutores, consiguiendo valores aceptables de error, y sin la necesidad de realizar procedimientos explícitos de adaptación. A pesar del buen momento que vive esta tecnología, el reconocimiento de voz dista de ser un problema resuelto. La mayoría de estos sistemas de reconocimiento se ajustan a dominios particulares y su eficacia depende de manera significativa, entre otros muchos aspectos, de la similitud que exista entre el modelo de lenguaje utilizado y la tarea específica para la cual se está empleando. Esta dependencia cobra aún más importancia en aquellos escenarios en los cuales las propiedades estadísticas del lenguaje varían a lo largo del tiempo, como por ejemplo, en dominios de aplicación que involucren habla espontánea y múltiples temáticas. En los últimos años se ha evidenciado un constante esfuerzo por mejorar los sistemas de reconocimiento para tales dominios. Esto se ha hecho, entre otros muchos enfoques, a través de técnicas automáticas de adaptación. Estas técnicas son aplicadas a sistemas ya existentes, dado que exportar el sistema a una nueva tarea o dominio puede requerir tiempo a la vez que resultar costoso.

Las técnicas de adaptación requieren fuentes adicionales de información, y en este sentido, el lenguaje hablado puede aportar algunas de ellas. El habla no sólo transmite un mensaje, también transmite información acerca del contexto en el cual se desarrolla la comunicación hablada (e.g. acerca del tema sobre el cual se está hablando). Por tanto, cuando nos comunicamos a través del habla, es posible identificar los elementos del lenguaje que caracterizan el contexto, y al mismo tiempo, rastrear los cambios que ocurren en estos elementos a lo largo del tiempo. Esta información podría ser capturada y aprovechada por medio de técnicas de recuperación de información (*information retrieval*) y de aprendizaje de máquina (*machine learning*). Esto podría permitirnos, dentro del desarrollo de mejores sistemas automáticos de reconocimiento de voz, mejorar la adaptación de modelos del lenguaje a las condiciones del contexto, y por tanto, robustecer al sistema de reconocimiento en dominios con condiciones variables (tales como variaciones potenciales en el vocabulario, el estilo y la temática).

En este sentido, la principal contribución de esta Tesis es la propuesta y evaluación de un marco de **contextualización motivado por el análisis temático** y basado en la **adaptación dinámica y no supervisada de modelos de lenguaje** para el robustecimiento de un sistema automático de reconocimiento de voz. Esta adaptación toma

como base distintos enfoque de los sistemas mencionados (de recuperación de información y aprendizaje de máquina) mediante los cuales buscamos identificar las temáticas sobre las cuales se está hablando en una grabación de audio. Dicha identificación, por lo tanto, permite realizar una adaptación del modelo de lenguaje de acuerdo a las condiciones del contexto. El marco de contextualización propuesto se puede dividir en dos sistemas principales: un *sistema de identificación de temática* y un *sistema de adaptación dinámica de modelos de lenguaje*.

Esta Tesis puede describirse en detalle desde la perspectiva de las contribuciones particulares realizadas en cada uno de los campos que componen el marco propuesto:

- En lo referente al *sistema de identificación de temática*, nos hemos enfocado en aportar mejoras a las técnicas de pre-procesamiento de documentos, asimismo en contribuir a la definición de criterios más robustos para la selección de *index-terms*.
  - La eficiencia de los sistemas basados tanto en técnicas de recuperación de información como en técnicas de aprendizaje de máquina, y específicamente de aquellos sistemas que particularizan en la tarea de identificación de temática, depende, en gran medida, de los mecanismos de preprocesamiento que se aplican a los documentos. Entre las múltiples operaciones que hacen parte de un esquema de preprocesamiento, la selección adecuada de los términos de indexado (*index-terms*) es crucial para establecer relaciones semánticas y conceptuales entre los términos y los documentos. Este proceso también puede verse afectado, o bien por una mala elección de *stopwords*, o bien por la falta de precisión en la definición de reglas de lematización. En este sentido, en este trabajo comparamos y evaluamos diferentes criterios para el preprocesamiento de los documentos, así como también distintas estrategias para la selección de los *index-terms*. Esto nos permite no sólo reducir el tamaño de la estructura de indexación, sino también mejorar el proceso de identificación de temática.
  - Uno de los aspectos más importantes en cuanto al rendimiento de los sistemas de identificación de temática es la asignación de diferentes pesos a los términos de acuerdo a su contribución al contenido del documento. En este trabajo evaluamos y proponemos enfoques alternativos a los esquemas tradicionales de ponderado de términos (tales como *tf-idf*) que nos permitan mejorar la especificidad de los términos, así como también discriminar mejor las temáticas de los documentos.
- Respecto a la *adaptación dinámica de modelos de lenguaje*, hemos dividimos el proceso de contextualización en varios pasos.
  - Para la generación de modelos de lenguaje basados en temática, proponemos dos tipos de enfoques: un enfoque supervisado y un enfoque no supervisado. En el primero de ellos nos basamos en las etiquetas de temática que originalmente acompañan a los documentos del corpus que empleamos. A partir de estas, agrupamos los documentos que forman parte de la misma

temática y generamos modelos de lenguaje a partir de dichos grupos. Sin embargo, uno de los objetivos que se persigue en esta Tesis es evaluar si el uso de estas etiquetas para la generación de modelos es óptimo en términos del rendimiento del reconocedor. Por esta razón, nosotros proponemos un segundo enfoque, un enfoque no supervisado, en el cual el objetivo es agrupar, automáticamente, los documentos en *clusters* temáticos, basándonos en la similaridad semántica existente entre los documentos. Por medio de enfoques de agrupamiento conseguimos mejorar la cohesión conceptual y semántica en cada uno de los *clusters*, lo que a su vez nos permitió refinar los modelos de lenguaje basados en temática y mejorar el rendimiento del sistema de reconocimiento.

- Desarrollamos diversas estrategias para generar un modelo de lenguaje dependiente del contexto. Nuestro objetivo es que este modelo refleje el contexto semántico del habla, i.e. las temáticas más relevantes que se están discutiendo. Este modelo es generado por medio de la interpolación lineal entre aquellos modelos de lenguaje basados en temática que estén relacionados con las temáticas más relevantes. La estimación de los pesos de interpolación está basada principalmente en el resultado del proceso de identificación de temática.
- Finalmente, proponemos una metodología para la adaptación dinámica de un modelo de lenguaje general. El proceso de adaptación tiene en cuenta no sólo al modelo dependiente del contexto sino también a la información entregada por el proceso de identificación de temática. El esquema usado para la adaptación es una interpolación lineal entre el modelo general y el modelo dependiente de contexto. Estudiamos también diferentes enfoques para determinar los pesos de interpolación entre ambos modelos.

Una vez definida la base teórica de nuestro marco de contextualización, proponemos su aplicación dentro de un sistema automático de reconocimiento de voz. Para esto, nos enfocamos en dos aspectos: la contextualización de los modelos de lenguaje empleados por el sistema y la incorporación de información semántica en el proceso de adaptación basado en temática. En esta Tesis proponemos un marco experimental basado en una arquitectura de reconocimiento en 'dos etapas'. En la primera etapa, empleamos sistemas basados en técnicas de recuperación de información y aprendizaje de máquina para identificar las temáticas sobre las cuales se habla en una transcripción de un segmento de audio. Esta transcripción es generada por el sistema de reconocimiento empleando un modelo de lenguaje general. De acuerdo con la relevancia de las temáticas que han sido identificadas, se lleva a cabo la adaptación dinámica del modelo de lenguaje. En la segunda etapa de la arquitectura de reconocimiento, usamos este modelo adaptado para realizar de nuevo el reconocimiento del segmento de audio.

Para determinar los beneficios del marco de trabajo propuesto, llevamos a cabo la evaluación de cada uno de los sistemas principales previamente mencionados. Esta evaluación es realizada sobre discursos en el dominio de la política usando la base de datos EPPS (*European Parliamentary Plenary Sessions* - Sesiones Plenarias del Parlamento Europeo) del proyecto europeo TC-STAR. Analizamos distintas métricas acerca

del rendimiento de los sistemas y evaluamos las mejoras propuestas con respecto a los sistemas de referencia.

Contents

List of Figures

List of Tables

Introduction

Speech and natural language are the most natural ways of communication between humans. From a few decades ago they also have emerged as a means of communication between humans and machines. This has lead to the design of modern large vocabulary continuous speech recognition systems to the point where their application covers nowadays a broad set of domains including speaker dependent systems, automatic broadcast news transcription, and lectures and meetings transcriptions in speaker-independent environments, just to name a few.

The degree of performance of such systems depends crucially on the knowledge they have about human language. A way to acquire this knowledge is leveraging existing sources of information. However, today this represents a major challenge: the volume of information that is available to us is continuously growing (e.g. Web contents) and it tends to diversify into several fields each day more and more. Not only the contents of the Web have enlarged, also has grown the number of repositories and databases that cover larger content and diversity. For this reason, there is an increasing interest not only in processing the available information, but also in selecting the appropriate information sources. In this regard it is not feasible to commit exclusively to humans to process it all. Therefore, automatic systems developed in different fields of knowledge for the analysis of information are compelled to evolve and specialize.

This Thesis is about the topic based adaptation of language models for automatic speech recognition. Therefore we will focus our attention on the fields of knowledge that relate to this Thesis, these are: the *identification of topics* from the perspective of information retrieval and machine learning systems; and the *dynamic adaptation of language models* for the enhancement of automatic speech recognition systems.

In response to different challenges of providing information access, the field of *Information Retrieval* (IR) evolved to give major approaches to searching various forms of content. Even though the IR systems did not begin with the Web, it must be acknowledged that this has been a major driver of innovation, releasing web documents at the scale of tens of millions. This explosion of available information would be unresolved if the information could not be found, indexed and analysed in a way that each user can expeditiously find information that they may find both relevant and diverse for their needs. Within this field, web search systems have witnessed the exponential growth of their indexing schemes and have been forced to adapt on the basis of daily volumes of consultations. Due to the continuously increasing size of data and the urgent need for the queries to be solved in a shorter period of time, the systems have evolved into

specialized data structures which aim is to provide fast access to the data and allow speeding up query processing.

Despite the fact that the analysis of queries and the search of relevant information (for instance in web search engines) is the field of application more extended and known for IR systems, there are other application domains in which these systems are equally important and in which it is possible to take advantage of the huge potential of the models that are developed in this field. One of these domains is *topic identification*.

The task of topic identification addresses the problem of identifying which of a set of predefined *topics* are present in a document. This task emerged as an important field of IR systems at the end of the last century and since then has made its way into the field of *Machine Learning* (ML). Nowadays, we can think of topic identification as the meeting point between IR and ML disciplines, and as such it shares a number of common characteristics. There is still a considerable debate on where to draw the boundaries between these disciplines. However, on the sidelines of this debate, both disciplines have much to contribute in the development of more robust systems for topic identification.

Within the field of Machine Learning, topic identification has become one of the key solutions for text data classification. It is currently been applied in many contexts and disciplines, ranging from document indexing to automated metadata generation, document and messages filtering and, in a general sense, in applications that comprise document separation and organization.

Nowadays, topic identification is one of the most challenging research topics due to the necessity to organize and categorize the increasing number of electronic documents worldwide. There is a further obstacle and it is the fact that these documents may take several forms (e.g. web pages, email, newspaper stories or scientific articles), therefore the systems should be versatile and adaptable, so they can cope with all types of input documents.

In recent years contributions to the field of topic identification have improved substantially, allowing for the processing of huge amounts of textual information with an acceptable level of efficacy. Some examples of this is the application of ML techniques to various domains such as topic detection and tracking, spam filtering, plagiarism detection, web page classification and sentiment analysis, among others.

In the area of *Speech Technology*, there have been substantial improvements in the capability and performance of speech processing systems over the last few decades. Definitely, the evolution of speech recognition systems over the past years has been noteworthy. Along with the progress that emerge from new technologies, *Speech Technology* is changing the way information is accessed. This evolution may be largely attributed to advances in statistical language modeling techniques and the refinement of automatic speech recognition systems for large vocabularies. In the modern systems we can find possible application domains as diverse as can be the automatic transcription of news, the transcription of meetings, conferences or phone calls, the voice search on mobile devices and the provision of call center services, just to name a few of them. A feature that is common to the application domains we have just mentioned, is that

all of them are framed within what is known as *speech analytics*, that is the process by which certain information can be gathered from speech and audio recordings.

Despite the fact that it has become common the use of the term *speech analytics* to define a set of very specific applications that gather information from the dialogues and service processes for commercial purposes, the truth is that the concept of *speech analytics* encompasses a broader spectrum of speech and audio analysis, ranging from the study of *what* has been said to *how* has been said and *who* has said that.

Speech analytics has emerged as a branch of speech processing that aims for collecting data that can be extracted from speech. These data, also known as metadata, which are not part of the message itself to be transmitted, provide information about the topic that is being discussed, the gender of the speaker, the emotion expressed by the speaker and may even allow the biometric identification of the speaker via the fingerprint of the voice. In this Thesis our interest is not to explore all the fields in the area of *speech analytics*, but to explore those that allow us to extract contextual information of the speech with the aim of identifying the topic on the audio recording. Our final goal is to use this information within a framework for the dynamic adaptation of language models in an automatic speech recognition system.

Although speech recognition systems are gaining increasing prominence not only in commercial applications but also in a considerable number of electronic devices of daily use, the development of a system that is reliable, accurate and efficient in multiple domains is still critical if we want to bring this technology to a larger number of applications. This development depends crucially, among other things, on the availability of large corpora of transcribed speech and annotated text specific to the language and the application domain. Most of the advances in this field have taken place in languages such as English, German and Japanese, and in domains such as travel information and broadcast news transcription, for which such linguistic resources have been largely developed. Construction of accurate systems for languages with deficient resources has recently started receiving attention. And this is the case of the Spanish language, for which there is a limited amount of training resources.

Another aspect to take into account in the development of reliable speech recognition systems is, that in a real environment, speech includes temporal variations commonly caused by changes of speakers, speaking styles, environmental noises, and changes of topics. Thus, these systems are required to track temporal changes in both acoustic and language environments. Regarding the changes that occur at the grammatical level, the grammar models are changing constantly in domains that involve spontaneous and multitopic speech, 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.

The optimal adaptation of language models for specific domains requires data that either belong to the specific domain or a similar one; and the Web can be a good place to go in the search of these data. One advantage of using the material available on the Web for training language models is that it can cover countless topics in the documents collected. However, this variety is at the same time a problem since the dispersion of the data can be so high that the language models could be poorly estimated. There is a

need for clustering techniques which allow to selecting, separating and grouping those documents that share similar properties, thus narrowing the data sets belonging to each specific domain. And it is precisely in this sense that the Information Retrieval and Machine Learning systems can contribute; these systems can be used to extract semantic relationships between terms and documents, as well as the relationships between documents in different collections, making it possible to establish levels of relevance of a document (or group of them) with a certain topic. The semantic analysis, therefore, allows grouping documents of similar topics and estimating models depending on a specific domain. Research in the field of adaptation of language models have taken advantage of topic identification techniques in this regard.

We could continue listing the changes of paradigms that modern technologies have brought in different areas of knowledge, but the interest in this work is focused on contributing to those already mentioned. From the point of view of the IR and ML techniques the goal is to make some contributions in the field of topic identification. From the perspective of automatic speech recognition systems, we want to provide a framework of topic-motivated contextualization based on the dynamic and non-supervised adaptation of language models.

1 Scientific and Technological context

The topic-motivated contextualization we aim to apply to automatic speech recognition systems primarily involves two major systems: a system for topic identification and a system for the adaptation of language models. In this chapter we review the scientific and technological context of each of these systems.

First, in Section 1.1 we present an overview of the current trends in the development of topic identification systems. We then review the fundamentals of the topic identification task (1.1.1) and we present an overview of the main techniques used for the representation (1.1.2) and pre-processing (1.1.3) of documents. We make a review of the most known techniques developed for term weighting (1.1.4). We present the fundamentals of Information Retrieval systems and their application into the topic identification task (1.1.5), and we also discuss some of the approaches in this regard from the field of Machine Learning (1.1.6).

In Section 1.2 we introduce the adaptation of language models and we present the motivation for this task (1.2.1). We also present an overview of the current trends, and the main challenges and limitations in these systems (1.2.2).

1.1 On Topic Identification

As more and more information services have become available, there has been an increasing interest in processing the information they provide. However, due to the fact that the amount of data is so overwhelming it is not feasible to commit exclusively to humans to process it all. Techniques for automatic text processing are an obvious solution to the information overload problem. Automatic text processing techniques can help people explore through large volume of texts, classify them into different categories, route them to relevant destinations and even make summaries of them. To achieve this, a central step is to identify the main topics of the texts.

In this sense, techniques for topic identification have emerged as a fundamental part and of great importance within the existing information systems. Topic Identification is a research area that arises in the field of Information Retrieval systems and as such, shares many of the fundamentals of these systems. Nowadays, it is an active research area not only in Information Retrieval, but also in Machine Learning and Natural Lan-

guage Processing, and it is currently motivated by many real world applications. We will mention some of the most common applications of topic identification from a general perspective and then we will focus on those applications that bring together speech technologies along with topic identification systems.

Automatic indexing may be considered as the application that stimulated the research in topic identification few decades ago [Salton and Yang, 1973]. This task consists of automatically generate, by means of a controlled dictionary, an index of the terms that better describe the contents of a document. Also, in the same domain, a closely related application is the automated metadata generation [Kim and Ross, 2006]. The metadata are normally used to describe the documents under a variety of aspects (e.g. publication date, field of knowledge, document type, etc.), among which one of the most relevant aspects to identify in a document is the topic.

Another application that is widely used is document organization. This can be considered one of the main applications of Topic Identification. Its objective is to classify a set of text documents into a set of predefined categories. A typical example of this application is the classification of news stories according to predefined tags (Political, Sports, Culture, etc.). Indeed, some of the text benchmark datasets that are most employed in the evaluation of topic identification systems are collections of news articles (Reuters-21578, Reuters Corpus Volume I and Volume II, Thomson Reuters Collection, AP Titles and UseNet data, among others) [1].

Another very common application of topic identification systems is text filtering; this is the activity of classifying a stream of incoming documents according to some elements in which the user is interested. Typical cases of filtering systems are e-mail filters [Upasana and Chakravarty, 2010] or filters of unsuitable content, such as spam or adult-only content [Chandrinos et al., 2000, Guzella and Caminhas, 2009].

From the NLP field, one of the most common applications is the Word Sense Disambiguation (WSD). This is the process of finding, given the occurrence in a text of an ambiguous word (i.e. polysemous or homonymous), the sense of this particular word occurrence [Escudero et al., 2000].

While it is true that the nature of the topic identification systems is to process text, the original source of information is not limited to being a source of written information; this can be of multiple types of content, such as speech, music, images, videos, etc. Systems that are capable of processing information of a different nature than text, usually employed intermediate systems to transform information from its original state to text to be further processed by them. For instance, applications based on topic identification on speech are based on the combination of an automatic speech recognition and a topic identification system. The ASR system is used to obtain the transcript of the audio before being processed by the topic identification system.

In this Thesis our interest is focused on the application of a topic identification system in conversational speech, not as a stand alone application, but as a complementary module within a contextualization framework for an automatic speech recognition system. In this sense many contributions have been made integrating both fields of topic

[1] In http://trec.nist.gov/data/reuters/reuters.html there is a detail description of the Reuters corpora.

identification and automatic speech recognition. A distinction can be made among these contributions. On the one hand, there are systems that use speech recognition systems to obtain a transcript of the speech and extract from it, parameters that are relevant to the application for which they are used. For instance, systems for topic identification in telephone conversations [Cieri et al., 2004, Hazen et al., 2008, Wintrode, 2013, Wintrode and Kulp, 2009], systems of *Spoken Term Detection* in which a search is made of a spoken term in an audio corpus [Abad et al., 2013, Echeverry-Correa et al., 2014, Senay et al., 2013], topic identification systems on dialogues segments [Myers et al., 2000], and summarization and indexing of speech corpus [Lamel and Gauvain, 2008, Mandal et al., 2013].

On the other hand, there are systems that use topic identification as a tool to adapt the models of the speech recognition systems to the conditions of the context [Myers et al., 2000]. Within this field, Topic Identification has been used to study the changes that the language experiences when moving towards different domains [Bellegarda, 2004]. One of the areas in which a large number of contributions have been made is the field of spoken dialogue systems. These systems allow the interaction between a human and a machine through a natural means of communication such as the voice. The main purpose of a dialogue system can be very diverse, however in accordance with McTear [2002] it is worth highlighting the dialogue systems applied to the retrieval of information, services and transactions. Examples of these systems are services for booking train tickets by phone [San-Segundo et al., 2005], information and recommendation systems for movies [Chu-Carroll, 2000], and control systems for household appliances [Fernández et al., 2005, Lucas-Cuesta et al., 2013].

The joint development of speech processing systems and information retrieval systems has generated a work area that combines both fields, known as *Speech Retrieval* or *Spoken Document Retrieval - SDR* [Glavitsch, 1995]. The objective of SDR is to recover spoken material in digital audio files that are relevant to a user's information need. In the most common scenario the user query consists of a typed sequence of words or a spoken query. The documents to be retrieved are previously indexed audio recordings, which were automatically transcribed by a speech recognition system [Hauptmann, 2006]. Similar approaches have also been proposed to recover not only spoken documents but also multimedia objects [Brown et al., 1994, Lamel and Gauvain, 2008].

1.1.1 Fundamentals of Topic identification

The task of *Topic Identification*, basically, addresses the problem of automatically classifying a new unseen document between different classes (in this case, a class corresponds to a topic from a predefined set of topics within the collection). Topic Identification is mainly a supervised classification task, where a training set composed of documents with previously assigned classes is provided, and a testing set is used to evaluate the system. This task is executed in two main steps: A learning step in which models of topics are learned from the labelled training dataset of documents in the collection, and an evaluation step in which these models are applied to the evaluation

dataset and one or more topics are assigned by the system.

A conventional topic identification framework consists of several stages, including: document representation, preprocessing, term weighting and the learning/evaluation stage. In turn, the preprocessing stage is usually composed of several procedures such as text normalization, stopword removal, stemming, index-term selection and thesaurus expansion. A typical scheme for topic identification is depicted in Figure 1.1.

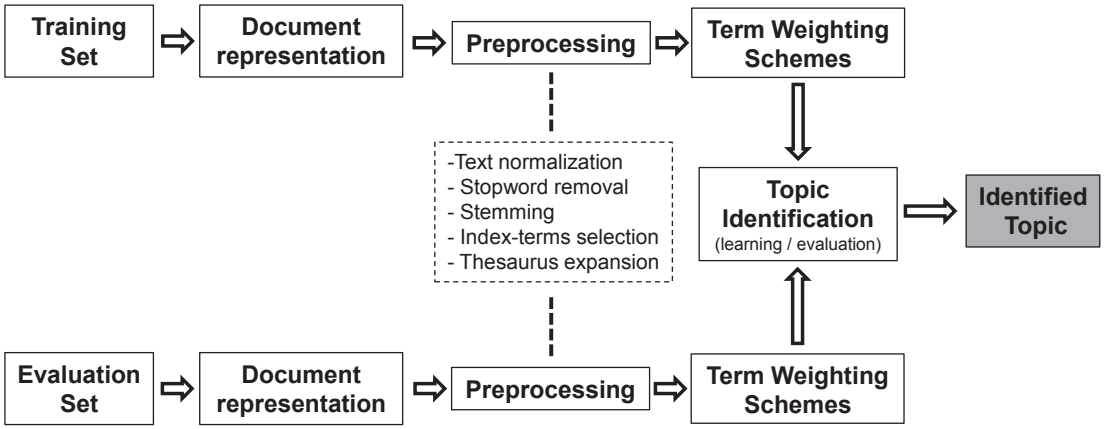


Figure 1.1: Topic identification general scheme

In the learning step the system "learns" the topic models automatically by examining the documents labelled by an expert under the set of topics (classes) in the collection. This makes topic identification a subjective problem, since the labels in the training documents that an expert can attribute to a document may vary with the purpose of the classification and personal experience. For instance, in a document organization application, a document on the minister of economy of Brazil can be classified under the class of *Politics*, or *Economy*, or under the class of *Latin-america* or under a combination of the three. Therefore the purpose of a topic identification system is to capture this subjectivity by examining the documents classified by the expert under a specific class.

Topic Identification may be formalized as the task of approximating an unknown target function $f : D \times A \rightarrow \{-1, 1\}$ that corresponds to how documents would be classified by an expert. The function f is the classifier, $A = \{a_1, a_2, \ldots, a_{|A|}\}$ is the predefined set of $|A|$ topics and $D = \{d_1, d_2, \ldots, d_n\}$ is the collection of documents. When $f(d_j, a_z) = 1$, the document d_j is a positive example of topic a_z, while when $f(d_j, a_z) = -1$ it is a negative example of topic a_z. In this sense, Topic Identification can be seen as a binary classification problem. Depending on the application, Topic identification may be either a single-label task, meaning that every document belongs to exactly one topic, or a multi-label task, in which each document can be classified in multiple topics. Most multi-label tasks are usually tackled using multiple binary classifiers.

In Sections 1.1.2, 1.1.3 and 1.1.4 we will review some of the most relevant research concerning different stages of the topic identification process (document representation, preprocessing and term-weighting schemes), and in Sections 1.1.5 and 1.1.6 we

introduce some of the Information Retrieval and Machine Learning techniques, respectively, used in the topic identification task.

1.1.2 Document representation

Each of the documents in the collection can be represented by a set of keywords called *index-terms*. In general, an index-term is a term that represents a key concept in a document. A carefully selected set of index-terms could either summarize a document or condense its main concepts [Goyal et al., 2013].

To select the set of index-terms we must first consider the words (or groups of words) that contribute in carrying the semantic content of a document. While all the words in a sentence are used with a particular purpose, it can be argued that most of the semantic is carried by nouns and verbs, although the latter to a lesser extent. The relationships between them create the basis for defining semantic concepts. Thus, an intuitive strategy for selecting index-terms is to use all the nouns and verbs in the text. This can be done by means of the elimination of adjectives, adverbs, articles, pronouns and connectives; these words are less useful when they are used as index-terms, not only because its meaning is not related to the topic of the document, but also because its function is basically connective (as in the case of conjunctions, articles, prepositions, etc.) and complementary (as adverbs, adjectives, pronouns) [Baeza-Yates and Ribeiro-Neto, 2011]. However, one of the drawbacks of this strategy is that it requires a syntactic analysis of the text. Also, depending on the conditions of the collection, a systematic elimination of all words aside from nouns and verbs, may not be the best strategy to adopt.

The *bag-of-words* (BOW) model has always been considered the starting point for the selection of index-terms. In this model the syntactic structure of the sentences and the order of words within the context is ignored. However, there have been some attempts to move forward this model and major advances have been achieved from the Natural Language Processing (NLP) field.

The best unit for matching a query and documents is often not an individual word. In Spanish, as well as in most languages, a group of words (also known as phrase) like "*Castilla y León*" or "*Parlamento Europeo*" lose much of its meaning if it is broken up into words. Thus, the main motivation for considering phrases is that a sequence of adjacent terms may be more discriminative than the individual terms in some cases.

A lot of NLP research has been devoted to detecting such phrases in text documents [Lewis, 1992b, Mladenic and Grobelnik, 1998]; approaches like n-gram indexing and Part-of-speech (POS) tagging have been employed to generate useful phrases [Manning et al., 2008]. As another alternative the use of the co-occurrences of terms (regardless of the order and position) has also been studied [Figueiredo et al., 2011]. Moreover, some experiments have been conducted in concatenating both approaches, the phrase-based index-terms with the BOW-based representation [Boulis and Ostendorf, 2005].

Despite the fact that these techniques have proven to be optimal in NLP applications

such as word sense disambiguation [Kilgarriff and Rosenzweig, 2000] and in many systems of document retrieval [Strzalkowski et al., 1999], surprisingly for the task of topic identification they have not improved significantly the BOW model [Moschitti and Basili, 2004, Silva and Ribeiro, 2010].

According to Figueiredo et al. [2011], although some works have reported gains when using n-grams as index-terms, these gains are only marginal or subject to specific circumstances. In this sense, Boulis and Ostendorf [2005], for instance, argues that considering both phrases and words as index-terms might produce an undesirable redundancy between the index-terms of the BOW model and more complex representations (e.g. phrases or co-occurrences). This redundancy adds more complexity to the systems and hinders the identification process. Zhang et al. [2011], on the other hand, consider that the effectiveness of these approaches (phrases, n-grams or co-occurrences) is strongly dependent on the types of topics of the collection. He claims that choosing multi-words as index-terms is effective for documents, in which fixed expressions (terminologies, collocations) are usually used, such as academic or scientific papers, but may not be effective for domains with extensive topics, in which fixed expressions are not used.

The discussion on whether it is appropriate or not to use a multi-words approach for the topic identification task is mainly based on evaluations performed on databases in English language. Most of the research that have led to raise these questions has been performed on common evaluation datasets (TREC collection, OHSUMED, Reuters, among others). To the best of our knowledge, there are no corpus in Spanish that have been deeply investigated. Some works, like Amini et al. [2009] had reported results on the Spanish partition of the Reuters Corpus (RCV2); and Bel et al. [2003], for instance, studied the multi-lingual text categorization task using the ILO corpus, but we believe that no significant conclusions can be drawn regarding the application of these approaches on databases in Spanish language.

The set of index-terms that are used to represent a document is commonly known as vocabulary. Nevertheless, with the aim of differentiating it from the vocabulary used in the generation of Language Models, we will use the concept of *term inventory* when referring to the set of index-terms in the topic identification system.

Different stages are involved in the preprocessing of the documents and the selection of the term inventory, as we shall see below in the next section.

1.1.3 Document preprocessing

The term inventory used by the topic identification system can be obtained in two ways: either a specialist proposes a set of index-terms that describe the documents; or the set of index-terms is automatically extracted from documents. In regard to the systems that concern us, the latter option is the most appropriate.

If the system is robust enough to process the text directly without any preprocessing, then the term inventory can be obtained directly from the original texts that compose the documents. Nevertheless, in most systems it becomes necessary to preprocess the

collection of documents to obtain it. The preprocessing stage of a topic identification system involves applying a set of well-known techniques not only to the documents in the training dataset, but also to the documents that are used for the evaluation of the system. These techniques attempt to reduce the size of the term inventory, controlling the computational cost involved, whilst maintaining or improving the performance of the system.

In the literature, multiple schemes can be found for this stage; the number of steps and the function of each one of them may vary slightly, but the aim is the same: to convert all documents to a more refined and concise format. However, preprocessing must be managed with care, since potentially useful information may be removed. Common preprocessing procedures include: *Text Normalization*, *Stopword Removal*, *Stemming*, *Index-terms Selection* and *Thesaurus expansion*, as depicted in Figure 1.2.

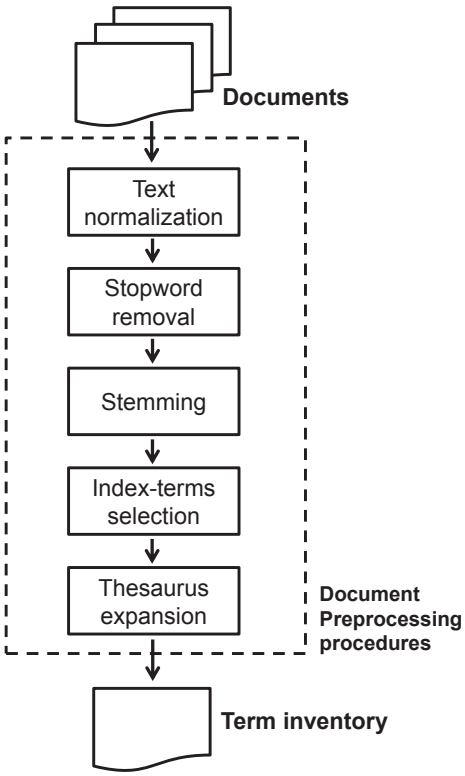


Figure 1.2: Common document preprocessing procedures

Some of the preprocessing procedures may be optional. The configuration of the preprocessing scheme will depend basically on the special needs of each application and the particular conditions of the collection of documents. These procedures are briefly described below:

- *Text normalization:* This preprocessing stage comprises three sub-stages: *structural processing*, *tokenization* and *normalization*. The objective in the sub-stage of structural processing is to analyse any structural element in the document such as labels for titles, sections, paragraphs, speakers, topics or other kind of extra-linguistic elements (which are common for XML and markup languages

in general). This stage decides whether to keep these elements or not depending on the information that can be extracted from them. An excerpt of an XML file, is shown below. This file is part of the document collection we use for the evaluation of the contextualization framework. In this excerpt are shown some structural elements within a XML file.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="sergio-701" audio_filename="20040503_163243_170254_ES_INT">
<Topics>
<Topic id="to001" desc="EPPS 03. May 2004 - Formal opening of the first sitting
of the enlarged European Parliament"/>
</Topics>
<Speakers>
<Speaker id="spk2" name="interpreter#2&lt;-speaker#2" check="no" type="male"/>
<Speaker id="spk5" name="interpreter#1&lt;-speaker#2" check="no" type="female"/>
</Speakers>
<Episode>
<Section type="nontrans" startTime="0" endTime="108.94">
<Turn startTime="0" endTime="108.94">
<Sync time="0"/>
</Turn>
</Section>
<Section type="report" startTime="108.94" endTime="730.079" topic="to001">
<Turn speaker="spk6" startTime="108.94" endTime="201.311">
<Sync time="108.94"/>
en el exterior delante del edificio Louis Weiss para la ceremonia de apertura
<Sync time="114.408"/>
<Event desc="b" type="noise" extent="instantaneous"/>
 de esta nueva Unión Europea
<Event desc="pause" type="noise" extent="instantaneous"/>
 ampliada con los nuevos ehh miembros .
<Sync time="120.954"/>
...
```

Regarding the tokenization process, this sub-stage consists of breaking a stream of text into tokens that can be words, sentences, phrases, symbols or other meaningful elements. 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, as we previously stated in Section 1.1.2. Overcoming these limitations depends, among other aspects, on the availability of dictionaries or catalogues of predefined tokens.

Finally, the normalization sub-stage aims to treat digits, hyphens, acronyms, punctuation marks and the case of the letters (lower case and upper case). Numbers are usually not good index-terms because, without a surrounding context, they are inherently vague. Hyphens represent an additional obstacle; breaking up hyphenated words might become a problem if words include hyphen as an integral part. This occurs more often in languages like English, in which these word constructions are common, nevertheless, in Spanish we can also find hyphenated words, such as names (e.g. *Echeverry-Correa*), relations between concepts (e.g. *calidad-precio*), grouped adjectives (e.g. *lingüístico-literario*), among others.

Regarding acronyms, they can be identified using external glossaries of terms. And finally, punctuation marks and the case of letters are usually not important for the selection of index-terms, however, particular scenarios might require the distinction to be made.

- *Stopword removal:* This procedure is performed to remove the non-informative words, i.e. words that have little lexical meaning and are too frequent among the documents in the collection. These 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 (the list of words to be removed). An appropriate list of stopwords eliminates noise from the term inventory, reduces the size of the indexing structure and contributes to speed up the clustering and decision processes.

  The use of standard lists of stopwords has been a general trend in IR systems over the years, however, some systems such as web search engines, do not use stopword lists, since some special searches may be disproportionately affected (e.g. a search like "To be or not to be" consists entirely of words that are commonly on these lists) [Manning et al., 2008]. However, in application domains other than search engines, the use of stopword lists is widely established. It should be noted that for an application in a specific domain, words that are to be included in the stopword list must be language and task dependent.

  A stopword list that is designed for a specific informative domain may not perform well in a different one. For instance, in a *political* domain, the word "law" could be a non-informative word since it may be too frequent among the documents. However, the same word could provide more lexical meaning in a different domain like *culture* or *sports* in which it could be less frequent.

  For English there is a standard stopword list, which is commonly used in many applications, called the SMART list [2]. Nonetheless, for Spanish language there is not a standard list, though there can be found different lists that usually come from NLP applications, suchs as the stopword list that come with the Snowball stemmer [3] or the stopword list from the Google code project [4].

- *Stemming:* This stage comprises not only the *stemming* process but also the *lemmatization* process. Both stemming and lemmatization aim to reduce inflectional forms and derivationally related forms of a word to a common base form.

  Stemming usually refers to the reduction of inflected words to their stem or root form. It is done with the objective of removing prefixes, suffixes, plurals and some morphological derivations of the words. *Lemmatization* process, is a more sophisticated procedure. It may involve more complex tasks such as understanding context and determining the part of speech of a word in a sentence (it is required, for instance, a previous knowledge of the grammar of the language).

[2] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
[3] http://snowball.tartarus.org/algorithms/spanish/stop.txt
[4] https://code.google.com/p/stop-words/downloads/list

For instance, if confronted with the word "*saw*" stemming might return just the word "*s*", while lemmatization would attempt to return either "*see*" or "*saw*", depending on whether the use of the word was a verb or a noun.

Lemmatization depends on the use of a vocabulary and morphological analysis of words. Its goals are both to remove inflectional endings and to return the base form of a word, which is known as the lemma.

These procedures are thought to be useful for improving the performance of the systems because they reduce variants of the same word; and consequently have the effect of compressing the size of the indexing structure by reducing the number of distinct terms to index.

However, there is an extensive debate in the literature regarding the benefits of stemming and lemmatization. For instance, in Manning et al. [2008] it is suggested that both stemming and lemmatization tend not to improve English information retrieval performance. In Méndez et al. [2006], the use of stemming on spam e-mail filtering is analysed. They reported that the application of stemming reduces the performance of a SVM-based classifier. Hollink et al. [2004], provide detailed results on the application of stemming in document retrieval for European languages. They conclude that there are only significant improvements in Finnish and Spanish, but for most languages, including English the results are poor. Uysal and Günal [2014] studied several preprocessing procedures on text classification in different domains and languages. They claim that stemming does not always lead to a significant improvement and its use must depend on both domain and language. Stemming algorithms are indeed specific to the language being studied.

We have found little research, in the literature, that addresses the problem of both stemming and lemmatization for topic identification in Spanish. A study that is worth mentioning is the one carried out by Fernández-Anta et al. [2013]. In their work, they studied the application of both procedures (stemming and lemmatization) for the tasks of sentiment analysis and topic detection over Spanish tweets. They concluded that lemmatization outperforms stemming, nevertheless they did not provide comparative results without using these techniques.

- *Index-terms selection:* In the first place, the selection of the index-terms depends on the term inventory remaining after performing the previous preprocessing stages. Different criteria have been proposed for the selection of the index-terms. In Section 1.1.2 we presented a preliminary discussion regarding this subject. Besides the bag-of-words model (BOW), NLP has offered different alternatives for the index-terms selection (phrases, ngrams, POS tags, etc.), but none seems to be as effective, for the task of topic identification, as the BOW model.

  Approaches from the fields of Machine Learning and Computational Linguistics have also emerged in this regard. Techniques of feature selection and dimensionality reduction employ some statistical measures over the training corpus and rank index-terms in order of their amount of information with respect to the topic labels of the identification task (typical measures are *Information Gain*,

Mutual Information and Chi-square among others). The objective in these techniques is to reduce the size of the term-inventory by selecting a subset of all index-terms to represent the documents. In Section 3.1.4 we present a more detailed description of these techniques.

- *Thesaurus expansion:* A thesaurus is, basically, a collection of synonyms and semantic related words, that can be used with the objective of revealing semantic relationships between terms. In this sense, this preprocessing procedure makes use of a thesaurus mainly with two purposes. First, it may perform a categorization of terms, which consists of grouping the terms, in the term inventory, into semantic categories. This presents important advantages such as reduction of noise, and retrieval based on concepts rather than on words. Second, it may expand the term inventory by adding similar and related terms to the existing index-terms. In some Information Retrieval tasks, this represents a major advantage, since it allows to expand the terms in the query to match additional documents. The expansion of the term inventory involves finding synonyms and sometimes even various morphological forms of words.

  The motivation for using a thesaurus for indexing and searching is based on the idea of using a controlled and extended vocabulary. The main difficulty in its application is that for some domains, a well known body of knowledge, which can be associated with the documents in the collection, might not exist.

The preprocessing stages reduce the complexity of documents and allow the transition between the original representation and the acquisition of the term inventory that is to be used in the subsequent stages of processing.

Once the term inventory is defined, the next step in the topic identification process is the definition and generation of the mathematical model for document representation. Among the IR systems, the most common models for document representation are the so-called *classic models*. These models are divided into three different categories: boolean models, vector (or algebraic) models and probabilistic models [Baeza-Yates and Ribeiro-Neto, 2011]. Later in this Chapter (1.1.5.1) we will briefly describe each of these models. For now, we just want to make a short introduction to the vector model, also known as Vector Space Model, since the techniques that we will review in the next section are directly related to it.

The Vector Space Model is the most common model for document representation. It is used not only in IR applications, also in machine learning approaches for topic identification, and in general text processing applications. This model offers a natural way to represent documents as vectors in a space formed by the index terms.

The development of this model and its application to the task of topic identification are part of the central body of this work. For this reason, later in this Thesis, particularly in Chapter 3, we will go into more detail on it. Basically, this model is based in the BOW (*bag-of-words*) representation, in which each document can be represented by the number of times the index-terms appear in the document. A more robust version of the Vector Space Model can be obtained by giving weights to the terms according to their significance within both each document and the document collection. In the

next section we will review relevant research regarding the most common weighting techniques.

1.1.4 Term weighting schemes applied to documents

Not all index-terms are equally useful for describing the document contents. There are distinct reasons for this:

i) There are semantic differences between terms; there are terms with a vaguer meaning which are not directly related to any of the topics of the documents, and there are also terms that evidently identify a concept relevant to a topic. Then, a distinction must be made between these terms in order to differentiate their contribution in describing the topic of a document.

ii) The index-terms are not uniformly distributed throughout the collection. As well as there are terms that appear in all documents, there are terms that only appear in a few of them. It may seem obvious, but actually this is one of the most important properties of the terms, since term distribution gives a notion of how informative an index-term is.

iii) The number of occurrences of the terms may be biased by the length of the documents. In a long document, the number of occurrences of a term may be larger than in a short document, thus the average contribution, without weighting, of its terms is increased. Long documents also have numerous different terms, increasing not only the number of matches between a query and a long document, but also the chances of retrieval of long documents in preference over shorter documents. From this point of view the raw frequency of terms would not be a reliable indicator of the ability of a term to represent a topic or a document.

In order to overcome these obstacles a weighting scheme can be applied to the index-terms. The goal of a weighting scheme is to associate each index-term with a weight that represents its relevance with respect, not only to the document it appears in, but also to the documents in the collection in which it does not appear. The success or failure of the Vector Space Model depends on the application of an appropriate term weighting scheme to the documents in the collection. There has been much research on term weighting techniques but little consensus on which method is the best.

Luhn [1957] described one of the earliest reported applications of term weighting. His work dwells on the importance of medium frequency terms and may be though as pioneer of tf-idf (*term frequency - inverse document frequency*) and related weighting schemes.

Dennis [1965] and then Salton and McGill [1983] proposed the noise as a measure of the term occurrence within a collection. In their work, the noise refers to how much a term can be considered useful for retrieval versus being simply a noisy term, and examines the concentration of terms within documents rather than just the number of occurrences.

Spärck-Jones [1973] proposed the inverse document frequency, with the aim of weighting a term according to the number of documents it appears in. She also explored different types of term weighting schemes involving term frequency within a collection, along with normalization measures for document length.

Salton and Yang [1973] proposed the tf-idf, doubtlessly, the most popular term weighting scheme. It uses weights that combine term frequency with inverse document frequency.

Another method for term weighting is based on user's judgements of relevant items. This method, called Relevance Weighting, proposed by Robertson and Jones [1976], is intended to include the user feedback in the process by weighting the terms according to the number of relevant and non-relevant documents in which they are contained.

Other approaches from the fields of machine learning and computational linguistics use some more sophisticated statistical measures. One of these measures, chi-squared statistics [Schütze et al., 1995], is intended to measure the lack of statistical independence between two variables (in this case, the terms and the documents in the collection).

In the same direction Fano and Wintringham [1961] proposed the mutual information as a measure of the relative entropy between the distributions of two variables (as in the previous scheme, these variables are the terms and documents). Based on the mutual information criteria, Church and Hanks [1990] proposed a measure which also encodes the linear precedence of terms, i.e. the order in which they appear in the text. Another metric, complementary to mutual information, is information gain [Lewis, 1992a], which not only considers the occurrence of terms in a document but also the absence of terms in documents. This metric balances the effects of term occurrences with the effect of term absences.

We could continue enumerating the contributions that have been done in this field, but clearly, that is beyond our objective. Besides, there are a great number of comparative studies in each field, from both theoretical and empirical points of view [Aizawa, 2003, Chisholm and Kolda, 1999, Harman, 1986, Liu et al., 2009, McGill, 1979, Salton and Buckley, 1988, Wintrode and Kulp, 2009].

The extensive history of such variety of measures is in itself an evidence of the difficulty of determining the preference of specific measures. This also suggests that the selection of the optimal scheme may be related to specific aspects of the collection, such as the size of the collection, the distribution of topics along the collection or the application domain [Cummins, 2008].

1.1.5 Information Retrieval Systems

Topic Identification emerged, few decades ago, as an important field of Information Retrieval systems. Since then it has evolved and specialized and has made its way into other different fields of knowledge. However, it is important to know and understand the fundamental aspects of Information Retrieval systems, since the various advances made in this field can provide us knowledge about how information should be repre-

sented and processed.

Information Retrieval (IR) is a scientific discipline that focuses on providing means to find relevant information according to user's information needs. Typically, general users tend to limit the spectrum of IR systems to the development of search engines, but it is obvious that the scope of these systems has gone far beyond this application, and nowadays it reaches all kinds of application domains.

The main objective of an IR system is retrieving the documents that can satisfy an information need from a large collection of documents. However, modern IR systems, involve more functions than just retrieving information. Nowadays, these systems can be implemented in multiple domains such as data modelling and representation, sentiment analysis, user interfaces and email/spam filtering, and evidently, the domain that is in the scope of this Thesis, topic identification [Baeza-Yates and Ribeiro-Neto, 2011, Manning et al., 2008]. Despite the fact that traditional IR research deals with text, retrieval of speech, images and video are becoming increasingly common.

In terms of research, IR can be seen from two different, though complementary, points of view: one focused on the development of computational algorithms and another focused on human-machine interfaces (interaction with the end user). From the first point of view, IR basically consists in the construction of indexes, query processing and in the development of algorithms of classification and ranking. From the point of view of the human-machine interaction, IR focuses on the study of the behaviour of users, understanding their needs and identifying ways in which subjectivity can affect the operation of a retrieval system. In this Thesis, we will focus on the computational point of view of IR systems.

1.1.5.1 Models for Information Retrieval

In Information Retrieval, a model establish a relation between a query formulated by an user and each of the documents in the collection. This relation is usually expressed as the relevance of the documents with respect to the query. The procedure for establishing such relation is often a mathematical procedure which encodes, by means of different approaches, the way in which words capture topic information of documents. In Topic Identification we can think of this relation as the evaluation of previously trained models on new documents.

There are different types of models in IR. Depending on the nature of the source of information, these models can be arranged into three major categories: those based on *text*, those based on *hyper-links* and those based on *multimedia* objects [Baeza-Yates and Ribeiro-Neto, 2011]. Among these, the so-called classic models for text processing are the most commonly used, these are: Boolean model, Vector Space Model and probabilistic model.

Boolean model is considered as the basic IR model. This model does not consider the number of occurrences of the index-terms in a document; it simply considers that a term is present or absent within the document. Queries in this model are formulated in terms of boolean expressions. If a document in the collection satisfies the condition

indicated by the boolean expression of the query, then that document will be retrieved as a relevant one. Thus, the Boolean model considers that each document is either relevant or non-relevant in accordance to the formulated query. Since there is no ranking for the documents retrieved as relevant, they are all considered equally relevant, which is slightly problematic, at least, if the number of retrieved documents is large. This causes the user may find it difficult to make a decision based on the results provided by the system. One of the drawbacks in this model is that it is not simple to express an information need as a Boolean expression. It requires prior knowledge of the ways, which are accepted by the system, to formulate a query. In fact, according to Baeza-Yates and Ribeiro-Neto [2011] most users find it difficult to express a query in terms of Boolean expressions. Most of the problems of the Boolean model can be solved using best match retrieval models, such as Vector Space Model or Probabilistic Model.

In the **Vector Space Model** both documents and queries can be represented by the number of times the index-terms appear in them. This means that, unlike the Boolean model, this model allows to take into account the frequency of occurrence of terms. Besides, this models offers a natural way to encode documents and queries into vectors. The matching of documents and queries is made using distance or similarity calculation between vectors.

An advantage of this model over the previous one (the Boolean model), is that it quantifies the relevance in a continuous range of values and not in a binary form. This means that it is possible to retrieve similar documents while assessing their degree of relevance. This, in turn allows to build a ranking of documents, by arranging them from the highest to the lowest in regard to the relevance measure. Furthermore, in this model, the user is free to enter the query on his own terms (a feature known as *free text query*) without being limited by the use of logical operators and predefined expressions [Manning et al., 2008]. The Vector Space Model is the most common model for document representation. It composes the theoretical basis of more advanced techniques such as LSA, which is commonly used in the adaptation of language models.

Although there are several models based on a probabilistic approach, the name of **Probabilistic Model** is used to refer to the model based on the Probability Ranking Principle (PRP) proposed by Robertson [1977]. This principle states that the retrieved documents that are presented to the user should be ranked by their estimated probability of relevance with respect to the information need. Since true probabilities are not available to an IR system, the Probabilistic Model estimates the probability of relevance of documents for a specific query. This estimation is a key part of the model, and this is where most probabilistic models differ from one another. The initial idea of a probabilistic model for information retrieval was proposed by Maron and Kuhns [1960]. Since then, many probabilistic models have been proposed, each based on a different probability estimation technique. Among these models, the Binary Independence Model (BIM) [Robertson and Jones, 1976] is the model that has traditionally been used with the Probabilistic Ranking Principle. This model does not consider the number of occurrences of the index-terms in a document; it simply considers that a term is present or absent within the document. In comparison, Vector Space Model may consider various important aspects: *i*) the frequency of the terms along documents

in the collection, *ii*) the relative frequency of terms in documents, and *iii*) normalization of the document length. The probabilistic model, does not integrate these components. With the aim of filling this gap in probabilistic models, Robertson et al. [2004] has introduced some alternative models which include *term frequency* factors and *length normalization*. These models are known as *Best Match* models and are common models BM1, BM11, BM15 and BM25.

1.1.6 Machine learning for document categorization

In the last few decades *Topic identification* has made its way into the field of *Machine Learning* (ML). Today, we can think of topic identification as the meeting point between IR and ML disciplines, and as such it shares a number of common characteristics. For instance, the document preprocessing techniques described in Section 1.1.3 are equally applicable in the machine learning approach for topic identification. Also, the techniques of ML, in their vast majority, are based on the *bag-of-words* model for the representation of documents. Therefore we cannot think of IR and ML as independent approaches, but as complementary within the task of topic identification.

The two broad types of classification methods for the topic identification task in machine learning are often characterized as being generative or discriminative. Generative classifiers, also known as probabilistic classifiers, are intended to train a model that learn the probability of a document belonging to a specific category or topic. Often, in these approaches the Bayes' theorem is applied to determine this probability. This type of classifiers include naive Bayes [Lewis, 1998], the Aspect model [Hofmann, 1999] and Latent Dirichlet Allocation [Blei et al., 2003].

Discriminative classifiers, in contrast, do not have a probabilistic framework. Discriminative methods include Support Vector Machines (SVM) [Joachims, 1998], Rocchio's method [Hull, 1994], k-nearest neighbor (KNN) [Yang and Liu, 1999], decision trees [Lewis and Ringuette, 1994] and centroid based classifiers [Han and Karypis, 2000].

Among these techniques, in this work we make use of the centroid based classifier not only for evaluating both the Topic Identification System and the contributions we propose, but also within the proposed architecture for classifying the transcripts provided by the first stage of the ASR. Since this is an important part of the work that we propose in this Thesis, this classification technique will be presented in more detail in Chapter 3.

1.2 On Language Model Adaptation

Throughout this chapter, we have presented the fundamentals of the Topic Identification task and some of the most important techniques in the fields of Information Retrieval and Machine Learning for topic identification. Recall that the purpose, within this Thesis, of the aforementioned techniques is to provide information regarding the context of the speech within the contextualization framework. This information allows

us to perform a dynamic adaptation of the language models that are used by a speech recognition system. In this Section we will present a review on the current trends in language model adaptation.

1.2.1 Motivation for language model adaptation

Language modeling aims to create models that are able to capture the regularities of a natural language. The objective of this task is to improve the performance in various natural language applications, such as speech recognition [Rosenfeld, 2000], handwriting recognition [Bunke et al., 1995], optical character recognition [Hahn et al., 1999], machine translation [Zhang, 2009] and information retrieval [Ponte and Croft, 1998].

Among these applications, language modeling for automatic speech recognition (ASR) systems has got a special interest in recent decades. Speech is the most natural way of interaction between humans and it is becoming an alternative mean of communication for the interaction of humans with computers. This has motivated the fast growth and evolution of ASR-based applications. The degree of performance of such systems depends crucially on the knowledge they have about human language and the way this language is modelled.

It may come as no surprise that performance of speech recognition suffers when evaluating language models on a domain which differ from the training corpus in topic, style, or genre. In an ideal scenario, we would like to have a language model trained with texts from the same domain as the one of the speech we are analysing. However, this is not always achievable and there might appear some obstacles. On the one hand, language experiences changes, sometimes even within the same domain. A change of topic, of speaker, of style, could make the language model close to useless. On the other hand, the amount of data available for some specific domains is usually only a small fraction of the corpora used for training general language models. For this reason, the quality of language models has only increased in certain domains where a significant amount of training data has become available. Nevertheless, more data does not necessarily lead to any significant improvement in the quality of language models [Rosenfeld, 2000], therefore 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.

LM adaptation is an approach to cope with those difficulties. It allows to model the changes that the language experiences when moving towards different domains. Precisely, one of the aims in language model adaptation is to find, analyse and use new sources of information with the objective of enriching the previously existent models.

LM adaptation techniques offer a major solution, for instance, in application domains involving spontaneous and multitopic speech. In such domains, grammar models are varying constantly; there are words that appear more frequently in a discourse related to some topics than in other audio segments. Therefore, the probability of usage of some words is increased depending on the topic of the speech. The performance of the

speech recognition system, for such domains, will depend among many other parts of the system, on its capacity to update or adapt the LMs.

LM adaptation becomes a strategy to lower the word error rate of the transcription given by an 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 [Chiu and Chen, 2007, Federico and Bertoldi, 2004].

Over the last years there has been an increasing effort in improving speech recognition systems by means of LM adaptation techniques. These techniques can be classified according to different criteria. Rosenfeld [2000] proposed a classification based on the domain of the data. Bellegarda [2001], on the other hand, suggested that the classification must be done accordingly to 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 and Kawahara, 2003], from linguistic analysis systems [Liu and Liu, 2008] or from the application context itself [Lucas-Cuesta et al., 2013].

Other approaches are based on analysis and extraction of metadata, i.e. 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. Topic based language modeling is a representative example of language model adaptation based on the context of the speech. This technique has become very popular mostly because the adaptation unit, i.e. the topic, is specific enough to capture distinctive aspects of language.

Although this kind of adaptation includes numerous and very different techniques, all of them are based on the assumption that the distribution of words depends on the topics of the text. Therefore a question that arises in this regard is how to identify the topic of a document. Fortunately, this question can be easily solved: throughout this chapter we have presented a broad spectrum of the most common techniques for identifying the topic of a document. In fact, classic IR models, such as the Vector Space Model, and more specialized techniques such as Latent Semantic Analysis (LSA), are among the first IR techniques to be applied within this category of language model adaptation [Bellegarda et al., 1996].

Clarkson [1999] proposed a mixture-based language modeling approach. In his work, a clustering technique is proposed to group documents into topic clusters. Each topic is then modeled by a single language model, which are linearly interpolated to produce a mixture based LM. 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 et al., 2003], have also been used for adapting LMs in an automatic speech recognition task [Chien and Chueh, 2011]. A keyword extraction strategy to determine the LM to be used in a multi-stage speech recognition system is proposed in [Chen et al., 2001a]. In contrast to LSA, which do not explicitly consider the exact word order in the history context, in [Liu et al., 2013a] 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 et al., 2013b] 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 and Khudanpur, 2004, Tam and 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 et al., 1998, Iyer and 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 and Callejas, 2006, Lucas-Cuesta, 2013], in dynamic language model adaptation for Large Vocabulary Continuous Speech Recognition Systems (LVCSR) [Gollan et al., 2005, Saon and Chien, 2012] and in Statistical Machine Translation for creating context dependent LMs from monolingual corpora [Lu et al., 2012], among other applications.

In the next section we review some of the most common techniques for LM adaptation.

1.2.2 Language model adaptation techniques

1.2.2.1 Cache-based models

These models exploit the fact that words which have occurred recently are more likely to occur in the near future. This makes the adaptation dependent on the recent history of the speech (i.e. the previous word sequences that have been recognized). The idea behind this technique is to increase the likelihood of a word in case it has been ob-

served in previous recognition steps [Kuhn and De Mori, 1990]. Different variants of cache-based models have been proposed. Jelinek et al. [1991] proposed that the cache must not be limited to containing single words, but containing recent bigrams and trigrams. Rosenfeld [1994] claims that function words provide little information and therefore suggests a selective unigram cache, where only content words are stored in the cache. Clarkson [1999] proposed a decaying history approach by considering that the more recent a word is, the higher its chance to re-occur.

While significant decrease in perplexity has been obtained [Clarkson, 1999, Kuhn and De Mori, 1992], according to Oparin [2008] most of the research on the application of this technique in large vocabulary continuous speech recognition has not shown a significant improvement of WER.

1.2.2.2 Trigger models

These models serve to capture the long-span relations between sequences of words. Although, theoretically, these sequences may contain any number of words, the most widespread models take account of pairs of words. These models can be consider as an extension of the cache-based models in the sense that they make use of the recent history of the recognition process. The underlying idea in this adaptation technique is very simple. For a pair of words, the likelihood of the second word (i.e. the *triggered* word) is increased if the first word occurred during recognition (i.e. the *trigger* word) [Tillmann and Ney, 1996]. It has been observed that much of the potential of trigger models lies in words that trigger themselves, called *self-triggers*. These words are virtually equivalent to the cache-based approach [Lau et al., 1993]. Trigger models have shown to reduce perplexity when interpolated with a background model [Rosenfeld, 1994], however according to Troncoso and Kawahara [2005] very little is gained from their use as compared to the basic cache-based approach.

1.2.2.3 Mixture-based models

The use of words and word sequences may vary greatly in terms of style and topic. This information is lost in standard language models due to the fact that these models calculate global statistics over a heterogeneous dataset. Mixture models try to recover this information by identifying subsets in the data and building models for each of these subsets.

A variety of methods has been used to explore mixture-based LMs. In general, modeling starts with partitioning the data, for instance, by using a manually tagged dataset or some form of automatic clustering [Clarkson, 1999]. Then for each partition a n-gram model is trained. In the automatic clustering approach, the number of partitions has to be chosen and it involves a tradeoff. If too many clusters are used, individual models may be under-trained on sparse datasets, and hence each of the cluster LMs will be poorly estimated. Conversely, few clusters will result in a model which may be unable to distinguish between topics or linguistic styles. Note that soft-clustering may be used, meaning that a document may belong to more than one cluster.

Usually not a separate cluster model is used in the recognition process. The most common approach is to use a general model interpolated with smaller cluster models. Commonly, a heterogeneous corpus (of a considerable size) is used to train the general model. As stated before, depending on the number of partitions, cluster models may have too little training data to be reliably estimated. Therefore, interpolation with the general model is done in order to maintain an optimum data coverage.

When the model is to be used, each component model must be assigned an interpolation weight. In this regard, there are different ways in which the interpolation weight can be selected: it can be set empirically by minimizing the perplexity in a development stage with data not seen during training [Clarkson, 1999, Tur and Stolcke, 2007]; it can also be estimated under some optimization algorithm, such as Expectation Maximization [Daumé et al., 2010] or Maximum A Posteriori (MAP) adaptation [Wang and Stolcke, 2007]; or it can be set dynamically depending on the current situation of the interaction (the topic of the speech, a specific speaker, etc.) [Haidar and O'Shaughnessy, 2012, Seymore and Rosenfeld, 1997].

In this Thesis we propose some contributions on the field of language model adaptation, related to the proposal of different interpolation schemes for the topic-motivated contextualization of a speech recognition system. These contributions are presented in Chapter 5.

2 | Objectives

Throughout this chapter we will present our research hypothesis, the main objective, the sub-objectives we want to achieve and the main contributions that we pursue in this Thesis.

In spoken language we are not only communicating a message, we are also providing information about the contextual circumstances in which the message occurs. This contextual information can reveal us many things, such as the gender of the speaker, his age, his identity or even the emotion expressed by him, among many other things; but above all, it can give us information on the subject on which the speaker is talking about. If we could incorporate this contextual information into the speech recognition technology, then we could adapt the recognition to the context of the speech that is being decoded, enhancing thus the performance of the recognizer. Therefore, the **hypothesis** we are addressing in this work is based on the possibility of identifying the semantic elements of the spoken language that give us information about the context, and particularly, about the topic. These elements could be extracted by means of Information Retrieval and Machine Learning techniques, and would allow us, within a contextualization framework, to adapt the language models used by a speech recognizer to the contextual conditions of the speech. This contextual adaptation could improve the recognition performance when compared to the results achieved by unadapted systems.

In accordance to our research hypothesis, the **primary objective** of this Thesis is to **propose and evaluate a framework of topic-motivated contextualization based on the dynamic and non-supervised adaptation of language models for the enhancement of an automatic speech recognition system**.

To achieve this objective we have divided the proposed framework in two principal technologies. These technologies, at least from a theoretical point of view, can be developed separately; but then, when merged into the contextualization framework, are closely linked to each other. On the one hand we propose the use of *topic identification* technology to detect the context we would like to adapt to. This technology is based on a combination of different techniques (from the IR, and ML fields). On the other hand we introduce a methodology for the *dynamic language model adaptation* to the detected topic context in order to enhance the performance of an automatic speech recognition system.

The particular contributions we pursue in each of these technologies, that composed the proposed framework, are:

i) Evaluate the impact of different criteria for preprocessing documents and for the selection of index-terms on the performance of a topic identification system.

ii) Compare and evaluate alternative approaches to traditional weighting schemes with the aim of improving the specificity of terms and to better differentiate the topic associated to documents.

iii) Introduce and evaluate different approaches for the generation of topic-based language models. To do this, we focus on improving the cohesiveness of the documents that are related by similar concepts thus improving the coverage of the language models.

iv) Develop various strategies in order to create a context-dependent model. These strategies are based on the combination of the topic-based language models and the outcome of the topic identification process. The context-dependent model is expected to reflect the semantic context of the utterance.

v) Integrate all the components of the contextualization framework into a dynamic adaptation process of the language model used by the system. For this integration we propose and evaluate different strategies based on linear interpolation between models.

With the aim of evaluating the previous contributions, we propose a system architecture based on two stages of recognition as depicted in Figure 2.1. This architecture integrates different modules: two Automatic Speech Recognition (ASR) modules, a Topic Identification module and a Language Model Adaptation module. This modular design allows us to study and adapt each module separately.

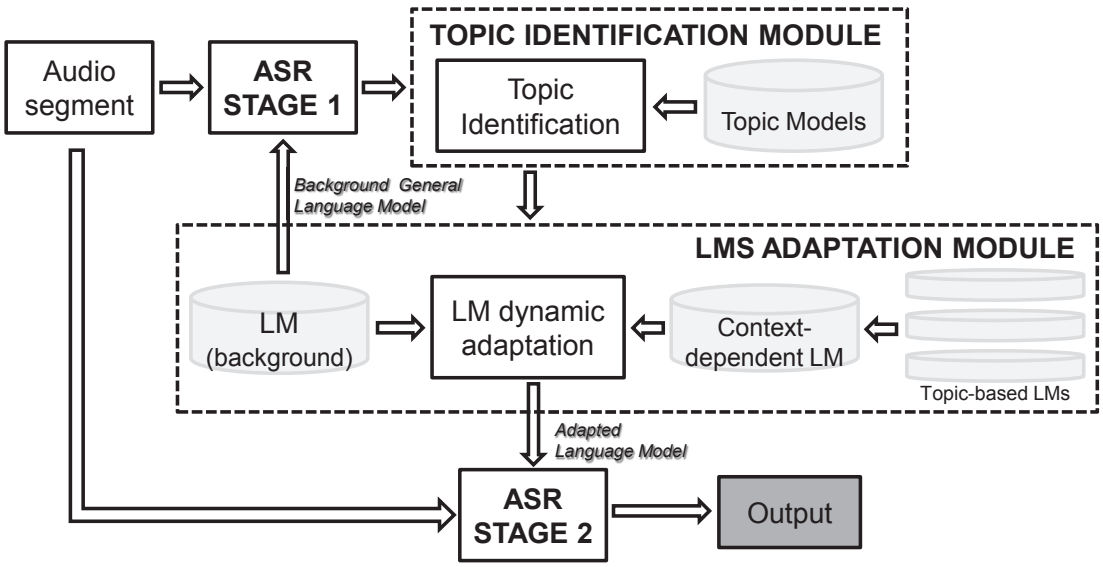


Figure 2.1: Experimental framework based in a 'two-stages' recognition architecture

The interaction between these modules is described as follows: the first stage of the ASR performs an initial decoding of the audio segment using a background general

language model. The output of this stage, i.e. the transcription of the audio segment, is processed by the Topic Identification module. Within this module, the Topic Identification system is responsible for identifying the topics that are relevant to this transcription. To do this, this system makes use of Topic Models that have been previously trained. Once the topic identification has been done, the LMs adaptation module makes use of that information and performs several procedures. First, it generates the context-dependent model combining topic-based models appropriately. Then, it performs the interpolation of the background language model with the contextual information provided by the context-dependent model and the Topic Identification module. In the final stage, the adapted LM is used to re-decode the utterance.

In the next Sections we present the sub-objectives and the contributions for each of the major technologies involved in this Thesis.

2.1 Proposal for improving the capabilities of the topic identification technology

Within our topic-motivated contextualization framework, we use the topic identification technology to extract and gather information related to the subjects of the speech. Regarding this technology, we focus our objectives specifically on the enhancement of document preprocessing techniques, in addition to contributing in the definition of more robust criteria for the selection of index-terms.

- The efficiency of IR systems, and particularly of those who carry out the task of topic identification, depends considerably on the mechanisms of preprocessing that are applied to the documents in the corpora used by those systems. These mechanisms allow to convert documents to a more concise and convenient format and have a substantial impact on the success of the topic identification process. Although several preprocessing procedures can be found in the literature, we can group all of them into five main operations: Text Normalization, Stopwords Removal, Stemming, Selection of Index-terms and Thesaurus Expansion.

  We are aware of the importance of all these operations within an IR system, and we will use most of them in the experimental evaluation of this Thesis, nonetheless in this work we will focus mostly on the adequate selection of the words that will be used as indexing elements (index-terms). A proper selection of index-terms in a document collection is essential to establish conceptual and semantic relationships not only between terms but also between terms and documents.

  In this regard we compare and evaluate different criteria for index-term selection. We want the selection to be dependent on the application domain and the particular conditions of the corpora; in this sense the selection must rely on information provided by metrics designed for this purpose. We also want the criteria selection to be optimal in terms of the topic identification system performance. We study the impact of the proposed criteria in reducing both the size of the indexing structure and the computational cost; we also evaluate its impact on the

performance of the topic identification process.

- The effectiveness of an IR system depends crucially on the identification and selection of significant terms in a corpus. This significance may be determined by how useful these terms are in order to identify the topic of a document. In many IR systems, specially in those based on the *Vector Space Model*, this significance can be quantified by assigning weights to terms based on the statistics of occurrence of the terms within a collection of documents.

  There have been proposed several weighting schemes, and it is by no means definitive what form of scheme consistently performs better than others. The utilization of the so-called *tf-idf* (term frequency - inverse document frequency) weighting scheme has been rather straightforward and intuitive and it has become the default choice within most of modern IR systems. However, its performance may be conditioned to the particular properties of the database, or to the specific task in which it is being applied.

  In this sense our aim is to compare and evaluate alternative approaches to traditional term weighting schemes that allow us not only to constrain the selection of the most significant terms but also to improve the properties of the term as a descriptor of a document topic. In this regard, different approaches from both IR and machine learning fields are analysed. By applying these approaches we expect to enrich the specificity of terms and enhance the topic identification results by improving the way the system prioritizes documents according to relevant terms.

2.2 Contributions on the dynamic adaptation of Language Models

In our work, the topic-motivated contextualization takes place in the adaptation process of the language models. This topic-motivated adaptation of language models is a strategy to lower the word error rate of the recognition, by providing language models with a higher expectation of words and word sequences that are typically found in the topics of the speech that is being analysed. To apply this strategy, our bottom-line idea is to make use of the available contextual information to dynamically update the LMs. This information may come from different sources and can be used in different stages of the adaptation process: The contextual information may emerge from the topic identification as well as from the topic-based language models; and it can be used during the generation of the context-dependent model as well as in the interpolation with the background language model. Our contributions in this regard are focused on the selection of the information that is used to train the topic-based language models as well as on the proposal of different strategies to estimate the interpolation weights in the different stages of the adaptation process.

The methodology we propose in this Thesis for the dynamic adaptation of language models is based on the scheme depicted in Figure 2.2.

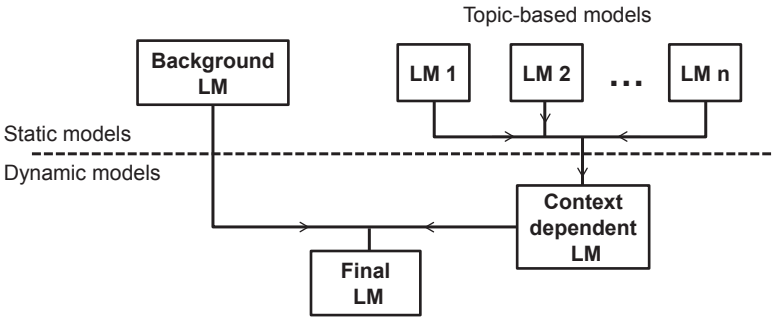


Figure 2.2: Scheme of adaptation of language models

Within the contextualization framework, the LM adaptation strategies that we propose differ in three ways: how to build or derive topic-based language models, how to combine them into a context-dependent model, and finally, how to create a final language model by means of the interpolation of the static background LM with the context dependent model. These contributions can be summarized as follows:

- We propose different approaches for the generation of topic-based language models. In the first place we propose a supervised approach intended to generate topic-based language models by grouping the documents, in the training dataset, according to the original topic labels of the corpus used for the evaluation of the system. It is worth mentioning that these topic labels were manually assigned according to the main topic of the debate session. In political speeches, as well as in many domains, different segments of the speech may not be directly related to the main topic of the discussion. In this sense, the semantic content of the utterance may convey information regarding different topics.

  One objective of this Thesis is to evaluate whether or not the use of these labels to generate language models is adequate in terms of recognition accuracy. For this reason, we propose a second approach, an unsupervised one, 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. 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 coverage of the topic-based language model within that cluster, consequently enhancing the performance of the recognition system.

- The next step in the contextualization framework is to generate a context-dependent model, which will be used as the underlying model in the adaptation process. In the proposed framework, the background model as well as the topic-based models are *static* models. That means that they are trained once and they remain unchanged during the evaluation. The context-dependent model, however, could be *static* or *dynamic*. This depends on the adaptation scheme followed. This model, as well as the adapted model used in the final ASR stage, are generated online during the processing of each audio segment. Our aim is that the context-dependent model reflects the semantic context of the current

utterance, i.e. the most relevant topics that are being discussed. To generate this model, we develop various strategies by means of linear interpolation between topic-based language models. The interpolation is performed between those models related to the most relevant topics, and the estimation of the interpolation weights is based mainly on the outcome of the topic identification process.

- As a final step in our contextualization process, we propose a methodology for the dynamic interpolation of a background language model. The scheme used for the adaptation is a linear interpolation between the background model and the context-dependent one. We study different approaches to determine the dynamic interpolation weights. In this regard, we claim that it is possible to gather enough information from the modules of the system to obtain these weights. The information may be provided not only by the context-dependent model but also by the topic identification process. The contextualized model will be used in the final stage of the recognition architecture.

2.3 Proposal for the evaluation and integration of the system modules

At this point, we have already presented the objectives that we pursue and the main contributions for each of the major systems that compose the contextualization architecture. Now, we propose a methodology for the evaluation of each of the modules and the integration of these modules into the 'two-stages' speech recognition architecture. In this regard, our work focuses on the evaluation of the *topic identification system*, and the evaluation of the *dynamic language model adaptation*.

- In the first place, the evaluation we propose for the Topic Identification module, consists basically of measuring the effectiveness of the system when identifying the topic that is being discussed in each audio segment. The set of topics is predefined and the topic reference labels we use to evaluate the effectiveness of the system are the original topic labels of the corpus. We evaluate the performance of our system when compared to a baseline system. In the evaluation we focus, especially, on those aspects of the system on which we are proposing improvements, these are: the preprocessing procedures, the selection of index-terms and the term weighting schemes. Evaluation of the system is performed on the transcriptions provided by the first decoding pass (ASR Stage 1, see Figure 2.1).

- The overall performance of the topic identification system may be evaluated not only by measuring the effectiveness of the identification process itself, but also by considering its effect on the speech recognition performance.

- Besides, system effectiveness is not the only variable we want to measure in the topic identification system. We also evaluate the impact of the proposed strategy

for creating domain-related stopword lists and term inventories in the size of the indexing structure.

- We also present a general and exploratory analysis on how different lengths of the audio segments impact both the topic identification effectiveness, and the overall performance of the adaptation strategies. We apply two different criteria for grouping individual and consecutive utterances of the same speaker into turns of intervention with different lengths. By these criteria we generate two configurations for the set of audio segments used in the evaluation of the system. Our aim in this sense, is to analyse, in a exploratory way, if the length of the audio segments has an impact on the performance of the systems.

- Regarding the language model adaptation module, our evaluation proposal should assess the quality of the language models. Statistical measures as the perplexity, allow us to perform a first evaluation of the robustness of the adapted models and their ability to represent the specific language of the context.

- However, perplexity is not as directly correlated with recognition accuracy. For that reason, the overall performance of the speech recognition system gives us a more realistic assessment of the usefulness of the adaptation. We propose to evaluate the final integration of the major modules in the ASR system by measuring the recognition performance in terms of word error rate. We use the topic information conveyed by the speech to adapt the LMs and recognize the same audio segment again in a second decoding pass. This way we can establish the first decoding pass (without LM adaptation) as our baseline for the ASR task.

3 Thesis work on Topic Identification

This chapter presents our **main contributions** and details the experimental conditions under which the work in the area of Topic Identification was carried out. For a clear presentation of the techniques that we use, our contributions and the experimental results, we have divided this chapter into three main sections (*Foreground*, *Contributions* and *Experiments on Topic Identification*.

First, in Section 3.1 we present the foreground material that was employed in conducting the experiments regarding this task. We introduce the models we used for document representation, i.e. the *Vector Space Model* (3.1.1) and the *Latent Semantic Analysis* approach for document modeling (3.1.2). We present the theoretical background of the *Centroid-based classifier* (3.1.3), which is the machine learning classification technique that we use for the supervised topic identification task. We also introduce various *term selection techniques* (3.1.4) with which we explore different alternatives to find an adequate set of index-terms for document representation.

In Section 3.2 we present our contributions regarding the topic identification task. Mainly these are focused on the enhancement of document preprocessing techniques and in the definition of more robust criteria for the selection of index-terms. We evaluate and propose alternative approaches to traditional weighting schemes that allow us not only to improve the document representation but also to enhance the identification of topics related to a document.

With the purpose of assess our contributions, we conducted different experiments that aim to compare and evaluate distinct criteria for document preprocessing, index-terms selection and term weighting. In Section 3.3 we present the results of these experiments, as well as the experimental framework which includes a description of the database used for the topic identification task.

3.1 Foreground on Topic Identification

Topic Identification (TI), basically, consists of learning models for a given set of topics (or classes) and applying these models to new unseen documents for topic assignment. Topics have been previously assigned by manually labeling each of the documents in the corpora; for this reason TI is mainly considered a supervised classi-

fication task.

A conventional framework involves several stages, of which the first stage is the definition of a document representation model. Typically, text documents are unstructured data, which facilitates their transformation into a representation that is suitable for computing. The first step into the document representation stage is to define the type of index terms to be used. As described in Section 1.1.2 (Document representation), there are different approaches for selecting the type of index terms: the *bag-of-words* model, or more complex representation models such as *phrases*. Once the type of index-terms is defined and terms are extracted from the training set of the document collection, each document may be represented using a model for this purpose.

The basic model for document representation we use in this Thesis is the *Vector Space Model* which we describe in the next section.

3.1.1 Vector Space Model

In this model, each document d_j can be represented by the number of times the index-terms appear in the document. For the whole document collection, this representation forms the Term-Document Matrix (TDM), which is:

$$TDM = \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & \dots & d_n \end{array} \\ \left[\begin{array}{ccccc} c_{1,1} & c_{1,2} & c_{1,3} & & c_{1,n} \\ c_{2,1} & c_{2,2} & c_{2,3} & & c_{2,n} \\ c_{3,1} & c_{3,2} & c_{3,3} & & c_{3,n} \\ & & & \vdots & \\ c_{m,1} & c_{m,2} & c_{m,3} & & c_{m,n} \end{array}\right] \end{array} \begin{array}{c} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m \end{array} \tag{3.1}$$

where $V = \{t_1, t_2, t_3, ..., t_m\}$ is the term inventory (i.e. the set of 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 $1 \leq i \leq m$; and $D = \{d_1, d_2, d_3, ..., d_n\}$ 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 ($c_{i,j}$ is commonly known as the raw frequency).

The Term-Document Matrix shown in Eq. (3.1), leads to a natural view of the document collection as a collection of vectors in a m-dimensional space. This data representation is known as *Vector Space Model* (VSM) and was 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. Note that the concept of query is a concept that has been adopted in the field of Information Retrieval. Originally, a query is an information need stated by an user. However, in the topic identification task, there is not actual user of the system, and therefore there is not actual query. Thus, the query can be understood as a document in the evaluation dataset that is being tested against the documents in the collection (the training dataset). Despite this, we will maintain this notation in order to keep the original formulation of the models.

Thus, each document can be represented as a vector, in the form:

$$\vec{d}_j = [c_{1,j}\ c_{2,j}\ c_{3,j} \ldots\ c_{m,j}]^T$$

And also a query can be represented, using the same index-terms, as a vector in the form:

$$\vec{q} = [c_{1,q}\ c_{2,q}\ c_{3,q} \ldots\ c_{m,q}]^T$$

where $c_{i,q}$ represents the number of times each term t_i appears in the query q. This model has the advantage of being effective, efficient and easy to implement and it is often used in Information Retrieval modeling because of its potential contrasted to its conceptual simplicity. By means of this representation we could directly use the distance between vectors to compute document similarity. However, a more robust version of the Term-Document Matrix can be obtained by weighting the terms in the matrix according to the significance of each term within both each document and the document collection.

3.1.1.1 Term weighting schemes

In the Vector Space Model, not all index-terms are equally useful for describing the document contents. There are three main reasons for this. First, we have to consider that there are semantic differences between terms. Not all the terms are appropriate to identify a concept relevant to a topic. Second, the distribution of the index-terms throughout the documents in the collection is not uniform. This means that while there are terms that appear on all documents, there are other terms that only appear in a few of them. And finally, the length of the documents bias the number of occurrences of the terms. In a long document a term is more likely to appear than in a short one. This indicates that the number of occurrences of a term is not a reliable indicator of the ability of a term to represent a topic by itself.

In order to overcome these obstacles and to improve the performance of the Vector Space Model, weights can be applied to the index-terms in the Term-Document Matrix. The goal of a weighting scheme is to associate each index-term with a weight that represents its relevance with respect, not only to the document it appears in, but also to the documents in the collection in which it does not appear. In Section 1.1.4 we presented relevant research regarding weighting techniques. We can infer, given the existence of such variety of weighting approaches, that determining the importance of the terms is not a trivial matter. At least, there is one thing on which most authors agree, and that is that the importance of terms depends mainly on the semantic relationships of each term with the documents in the collection; these relationships can be measured from the number of occurrences of a term.

For instance, a term that appears in all documents, has less probability to provide information to decide on what document a user might be interested in. Instead, a term present in only a small number of documents can narrow the search, simplifying the selection of documents that might be relevant to the user query. To understand how

to assess the importance of a term, it is necessary to outline two key concepts: the *specificity* and *exhaustivity* of terms and documents respectively [Spärck-Jones, 1972].

Within the field of Information Retrieval and text document processing, *specificity* is considered a semantic property of each term and it may be described as how well the term describes the topic of a document. For instance, the term "art" may be used in documents about "music", "theatre" and "painting". It is expected then that the more general term "art" appears in more documents than the separate terms "music", "theatre" and "painting". Therefore, the term "art" is less specific since it has a larger distribution in the collection than the more specific terms.

Exhaustivity, on the contrary, within the field of IR, is considered a property of the documents. It is related to the number of index-terms assigned to a given document and it may be described as the coverage the document provides for the main topics. The more index-terms are assigned to a document, the more exhaustive its description becomes; which in turn increases the probability that the document is retrieved in the event of a query. However, it must be noticed that increasing exhaustivity does not necessarily lead to an improvement of the system. Actually, for some cases this may suggest a drawback for the system, since if many index-terms of a document are considered, this document could be retrieved even for queries for which it was not relevant.

Therefore, there is a trade-off between specificity and exhaustivity. If more terms are used in the document description, (i.e. increasing the exhaustivity of documents), then the specificity of terms becomes lower. In this sense, the term weighting schemes aims to transform the properties of the Term-Document Matrix while adjusting the trade-off between specificity and exhaustivity.

A term weighting scheme is usually composed of two different types of term weighting: local weights and global weights.

Local weights are intended to modify, mainly, the exhaustivity of documents by transforming each of the $c_{i,j}$ elements in TDM matrix, independently for each document; they do not depend on inter-document frequencies.

Global weights, however, depend on how many times a term appears in the entire collection. These schemes are intended to modify, mainly, the specificity of the terms, indicating the overall importance of a term. They transform each element $c_{i,j}$ in the TDM matrix, independently for each term, based on its occurrence in all documents. Generally, these schemes are based on the idea that the less a term occurs in the collection the more discriminating it is.

Theoretically, the use of global schemes might replace the need to remove stopwords in the preprocessing stage, since terms that appear in most documents (such as stopwords) would have a small global weight, and could be pruned out of the final inventory of index-terms, which would mean that they are not significant in the later stages of the process. In practice, however, is not only easier to remove the stopwords in the preprocessing stage, but in doing so also the initial size of the term inventory is reduced, thus optimizing the calculations in the following stages. Below, we present some of the most used schemes for term weighting.

3.1.1.1.1 Local weighting schemes.

Each of these schemes is defined as a function of the raw frequency $c_{i,j}$ of each term t_i, inside each document d_j in the collection and independently of the frequencies of the term in other documents. So far, we should understand these local weights as substitutes of the original counts in a new matrix representing the terms in the different documents, on a similar way as the original TDM. We will use the notation $l_{i,j}$ to represent the local weight of the term t_i in the document d_j.

- **Binary:** This scheme assigns the same weight, specifically the value "1", to every term that appears in a document, regardless of how many times it appears. Otherwise, the weight is zero.

$$l_{i,j} = \begin{cases} 1 & \text{if } c_{i,j} > 0 \\ 0 & \text{if } c_{i,j} = 0 \end{cases}$$

- **Raw frequency:** In this scheme the weight is the same as the raw frequency of the term, that is the number of times a term appears in a document. Therefore this scheme gives more weight to words that appear more frequently. A drawback of this scheme is that it considers that a word that appears, for instance, ten times in a document is ten times more important than a word that appears only once; and usually, that is not true.

$$l_{i,j} = c_{i,j}$$

- **Log frequency:** A common modification to the previous scheme is to use instead the logarithm of the raw frequency. By doing this, the effects of large differences in frequencies are diminished, as can be seen in Figure 3.1. This scheme assigns a weight given by:

$$l_{i,j} = \begin{cases} 1 + \log(c_{i,j}) & \text{if } c_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

  The logarithm can in principle be computed in any base, since the selection of a specific base just constitutes a constant factor towards the overall result. In the area of Information Theory it is common to use logarithm base 2, due to the intrinsic relation with the amount of information in "bits" that can be measured by this base. We believe that in our work it is convenient to adapt the same criterion. We use logarithm base 2 in all our calculations, unless otherwise explicitly mentioned.

- **Term frequency (tf):** This scheme calculates the relative frequency of each term in the document. It is defined by the expression:

$$l_{i,j} = tf_{i,j} = \frac{c_{i,j}}{\theta_j}$$

  Where θ_j is a normalization factor computed for the document d_j. This normalization factor is used both to fairly retrieve documents of all lengths and to

remove the advantage that the long documents have in retrieval over the short ones. Below, we present two of the most common normalization factors that are used within this weighting scheme.

$$\theta_j = \begin{cases} \sum_{k=1}^{m} c_{k,j} & \text{Document length normalization} \\ \sqrt{\sum_{k=1}^{m} c_{k,j}^2} & \text{Euclidean normalization} \end{cases}$$

- **Augmented and Normalized Term Frequency:** This scheme gives a weight K to any word that appears in the document; additionally it gives some additional weight to words that appear frequently. This scheme is defined by the expression:

$$l_{i,j} = \begin{cases} K + (1-K)\left(\frac{c_{i,j}}{\max_i(c_{i,j})}\right) & \text{iff } t_i \in d_j \\ 0 & \text{otherwise} \end{cases}$$

Based on experiments on different datasets, Singhal et al. [1996] suggested that K must be set to a low value (e.g. 0.3) for large documents, and to higher values (e.g. 0.5) for shorter documents, in order to balance the impact of the length of documents.

Figure 3.1 shows a comparison of the different local weighting schemes for frequencies ranging from 0 to 20. We assume that the maximum frequency (used in both term-frequency and augmented and normalized term frequency schemes) is 100. Note that the raw frequency count grows very quickly while the other local schemes grow more slowly. Every local weight assigns a value of 0 to $l_{i,j}$ if term t_i does not appear in document d_j.

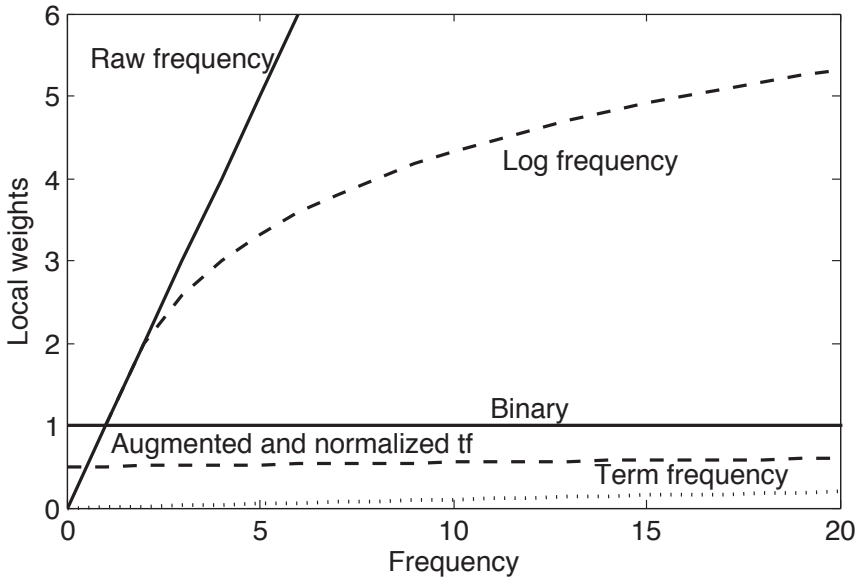


Figure 3.1: Comparison between local weighting schemes

3.1.1.1.2 Global weighting schemes.

These schemes are often defined as a function of other parameters, such as: document frequency df_i, computed as the number of documents in which the term t_i appears; global frequency gf_i, computed as the number of occurrences of the term t_i in the entire collection; and the term frequency $tf_{i,j}$, previously defined as a local scheme. Now, we should consider these global weights as multiplying the matrix obtained after applying one of the local weighting schemes presented before. We will consider g_i as the global weight for the term t_i for all documents.

- **Unitary weight:** This basically means that no global weighting scheme is applied. It is useful to emphasize the term frequencies in a document.

$$g_i = 1$$

- **Inverse Document Frequency (idf):** This weight will be zero if the given term appears in every document in the collection; and will increase as the number of documents in which the term appears decreases. It is defined as the logarithm of the ratio of the number of documents in the collection to the number of documents containing the given term.

$$g_i = \log\left(\frac{n}{df_i}\right)$$

  We are assuming that all the terms in the term inventory appear at least once in the collection. In cases for which this condition is not fulfilled, a smoothing parameter λ must be included in the denominator (i.e. $df_i + \lambda$) to avoid division by zero. In our work, all the terms in the term inventory appear at least once in the document collection, and therefore the idf is always computed as in the previous equation.

- **Probabilistic Inverse Document Frequency:** This is one variant of the Inverse Document Frequency. This weighting scheme arises from the classic probabilistic model where it is used as part of the ranking function for the retrieval of documents. It assigns weights ranging from $-\infty$ for a term that appears in every document to $\log(n-1)$ for a term that appears in only one document. It differs from the previous scheme in that probabilistic idf actually gives positive and negative weights.

$$g_i = \log\left(\frac{n - df_i}{df_i}\right)$$

  Despite it is used in the probabilistic model, it has not been proved its efficiency as weighting scheme for the Vector Space Model [Baeza-Yates and Ribeiro-Neto, 2011].

- **Global Frequency Inverse Document Frequency - gfidf:** Computes the ratio of the total number of times a term t_i appears in the collection (global frequency

- gf_i) to the number of documents it appears in (document frequency - df_i). If a term appears once in every document or once in only one document, it is given a weight of one, the smallest possible weight.

$$g_i = \frac{gf_i}{df_i} \quad \text{where} \quad gf_i = \sum_{k=1}^{n} c_{i,k}$$

- **Term Entropy:** What differentiates this scheme from the previous ones, is that this scheme takes into account not only the number of times a term appears in the collection, but also the number of times it appears in each document. This scheme is based on the normalized entropy of a term t_i, defined as:

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

Note that $\log(n)$ normalizes the entropy in order to limit its value to the range $[0, 1]$. The reason for this normalization will be explained in the next paragraph. For now, let us say that this is only a normalization factor and does not affect the reasoning below. It is interesting to analyze the extreme values of the normalized entropy; it allows a better understanding of this weighting scheme. ϵ_i takes a value of 0 if and only if the term t_i appears in only one document (that is $c_{i,j} = gf_i$, for the document d_j in which the term appears), and takes a value of 1 if and only if the term appears the same number of times in all documents (that is $c_{i,j} = gf_i/n$). Any other combination of frequencies will yield a weight between zero and one.

A value of ϵ_i close to 1 indicates a term distributed across many documents throughout the collection, while a value of ϵ_i close to 0 means that the term is present only in a few specific documents. In this sense, the aim of a weighting scheme is to give more weight to the latter case, i.e. those terms present in only few specific documents. So, it makes sense to consider the value $1 - \epsilon_i$ to weight the term t_i, rather than the value of the normalized entropy. The reason for which the entropy had to be normalized was to ensure that the value of $1 - \epsilon_i$ is within a positive range. The term entropy weighting scheme is then defined as follows

$$\begin{aligned} g_i &= 1 - \epsilon_i \\ g_i &= 1 + \frac{1}{\log(n)} \sum_{j=1}^{n} p_{i,j} \log(p_{i,j}) \end{aligned}$$

As previously mentioned, these weighting schemes must be applied not only to the documents in the collection but also to the query. However, note that given that the query is a single document (i.e. represented as a single vector), it is not sensible to calculate the global weight for each of its terms on just this query; since this weight should be calculated over a number of documents, as discussed above. For this reason the same global weight calculated for the terms in the collection is often applied to the terms in the query.

By applying weighting schemes to the TDM, a new matrix is obtained. We call this matrix Weighted Term-Document Matrix (from this point, and in order to simplify the mathematical notation, we will refer to this matrix as matrix W). In matrix W each element $w_{i,j}$ is composed by two components (local and global schemes), computed as a product:

$$w_{i,j} = l_{i,j} \times g_i \tag{3.2}$$

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 for the query. The weighted query is then a vector, in which each element is equal to:

$$w_{i,q} = l_{i,q} \times g_i \tag{3.3}$$

Analogously, $l_{i,q}$ and g_i are the local and global schemes respectively, applied to the query vector.

After applying the weighting schemes, the document d_j, as well as the query q, are represented by a *weighted* document vector $\overrightarrow{wd_j}$ (i.e. the $j-th$ column of the W matrix) and a *weighted* query vector $\overrightarrow{wq}$, respectively, as follows:

$$\overrightarrow{wd_j} = [w_{1,j}\ w_{2,j}\ w_{3,j} \ldots\ w_{m,j}]^T \tag{3.4}$$

$$\overrightarrow{wq} = [w_{1,q}\ w_{2,q}\ w_{3,q} \ldots\ w_{m,q}]^T \tag{3.5}$$

Among all weighting schemes presented, one of the most used in general information retrieval tasks is the one formed by the relative term frequency (as a local weight) and the inverse document frequency (as a global weight). This scheme, proposed by Salton and Yang [1973], is known as *tf-idf*. Its properties are well described in Salton et al. [1975]. In this work, it is shown that for large collections, *tf* and *idf* weights balance each other (preserving the trade-off between specificity and exhaustivity). Terms with high values of *tf* tend to be associated with low values of *idf* and terms with low values of *tf* are normally associated with high values of *idf*. As a result, the maximum *tf-idf* values are obtained with intermediate values of both *tf* and *idf*. Therefore, the terms that perform better in an IR task in a large collection of documents are not those with the maximum *idf* values, but those with an intermediate value.

3.1.1.2 Similarity measure for the Vector Space Model

In this model, the similarity between two documents (whether they are two documents in the collection or a document and a query) can be computed using the cosine distance. This distance measures the cosine of the angle between two vectors. It ranges from 1.0 for vectors pointing in the same direction (since $\cos(0°) = 1.0$) over 0.0 for orthogonal vectors to -1.0 for vectors pointing in opposite directions.

We will define it for measuring the similarity between a document $\overrightarrow{wd_j}$ and a query $\overrightarrow{wq}$ as follows:

$$sim\left(\overrightarrow{wd_j}, \overrightarrow{wq}\right) = \cos\theta = \frac{\overrightarrow{wd_j} \bullet \overrightarrow{wq}}{\|\overrightarrow{wd_j}\|\ \|\overrightarrow{wq}\|} \tag{3.6}$$

where the numerator represent the *dot product* (also known as the *inner product*) of the vector $\overrightarrow{wd_j}$ and $\overrightarrow{wq}$ and the denominator is the product of their *Euclidean lenghts*.

Figure 3.2 shows an example of a two dimensions representation of two documents and a query. In this example the term inventory is composed by two index-terms (*mythology* and *jupiter*), that is $m = 2$, and the vectors represent the documents and a query. In the figure, the cosine of the angle θ measures the similarity between document d_1 and the query q.

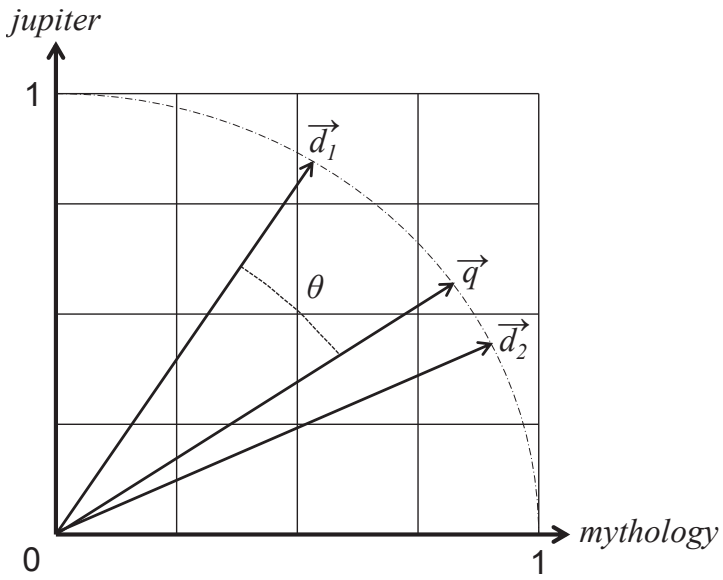


Figure 3.2: Example of representation of documents and query in the vector space. The cosine of the angle θ measures the similarity between the document d_1 and the query q

An advantage of this model over the Boolean model, is that it quantifies the similarity in a continuous range of values between -1 and 1 and not in a binary form. This means that it is possible to retrieve similar documents while assessing their degree of similarity. This, in turn allows to build a ranking of documents, by arranging them from the highest to the lowest in regard to the similarity measure. However, for the topic identification task, a hard decision is usually adopted. This means that the category of the document retrieved in the first position of the ranking could be the class assigned to the query (or in this case, to the evaluation document).

3.1.2 Latent Semantic Analysis - LSA

In the previous sections we described the Vector Space Model of documents and queries. This vector representation has a number of advantages including the treatment of queries and documents as vectors, the possibility to weight terms differently and the use of a simple metric as the cosine distance to measure the similarity between documents. In this section we will describe an improvement to this model known as Latent Semantic Analysis. Although the Vector Space Model has been well developed and applied in many practical cases, it has some drawbacks that are worthwhile to mention:

- The Vector Space Model assumes that there is a mutual independence between the terms of a document, that is to say, that there are not any semantic or conceptual relationships between them. In practice, that is not true. The text of a document is an ordered sequence of terms that relate to each other, while building ideas and concepts, therefore establishing semantic connections between terms.

- This model is unable to deal with two common problems that appear with the use of natural languages: *synonymy* and *polysemy*.

  - *Synonymy* is the property of a concept to be expressed by different words. For instance, the words "picture" and "photograph" are synonyms, but the Vector Space Model fails to capture this relationship and simply assigns each word to a separate dimension in the vector space. Thus, the similarity between a query containing the word "picture" and a document containing the word "photograph" certainly, will not take advantage of the similarity between these words.

  - *Polysemy*, describe words that have multiple meanings, which is a common property of language. For instance, the word "light" can occur in the context of the radiation that comes from the sun as well as in the context of the food that we should eat to lose weight. The problem in this case is that the Vector Space Model does not consider the co-occurrences of terms. Therefore the word "light" will be considered out of any context, as a separate dimension, regardless of the terms with which co-occurs.

- Finally, the Vector Space Model usually involves a high dimensional representation due to the number of index-terms in an entire collection.

Latent Semantic Analysis (LSA) proposed by [Deerwester et al., 1990], tries to overcome these problems.

Basically, this technique assumes that there is some underlying structure in the co-occurrence of terms; this consideration improves in the first place the independence assumption of the Vector Space Model.

To reveal this structure, this technique projects documents into a space with "latent" semantic dimensions. In this space, documents containing co-occurring terms are likely to be found in the same vicinity. In particular, this alleviates some of the problems of the Vector Space Model since documents that are somehow related will still be "close" even if they share no terms in common.

The latent semantic space has fewer dimensions than the aforementioned TDM (which has as many dimensions as index-terms) and for this reason, LSA can also be considered as a dimensionality reduction technique.

We first present a graphical representation of a simple example of LSA and then introduce a formal description of the method. Consider the TDM shown in Table 3.1.

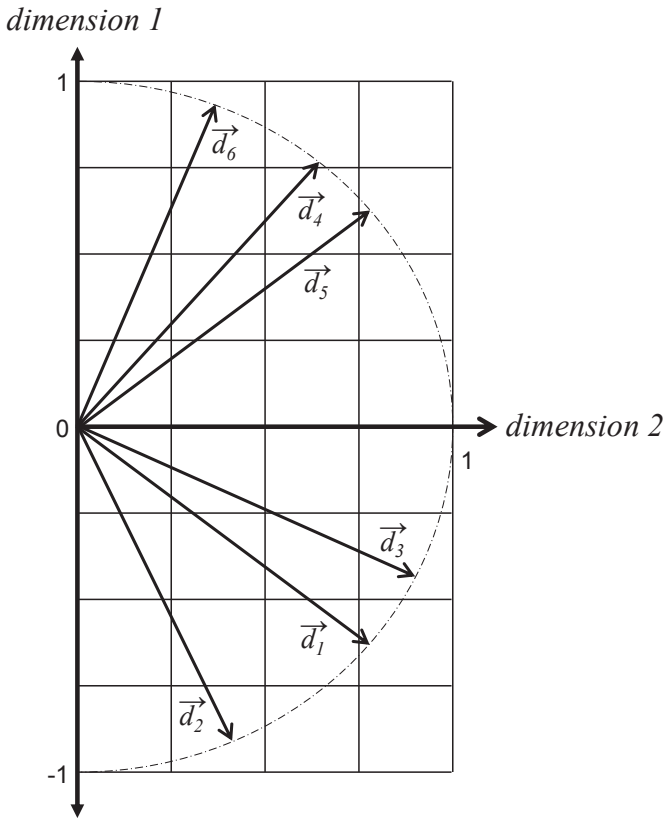


Figure 3.3: Latent Semantic Analysis technique applied to the Term-Document Matrix in Table 3.1

	d_1	d_2	d_3	d_4	d_5	d_6
light	1	0	1	0	1	0
jupiter	1	1	0	0	0	0
mythology	0	1	0	0	0	0
picture	1	0	0	1	0	0
photograph	0	0	0	1	1	1

Table 3.1: Example of a TDM

In this TDM there are five index-terms (i.e. it is a five-dimensional space), and six documents. In this example, we are neither considering weighting nor normalization schemes. The application of LSA produces the two dimensional space depicted in Figure 3.3 (for visualization purposes the length of the vectors is normalized and only two dimensions of the latent semantic space are considered).

It is worth to mention that in this extremely reduced space it is possible to observe some relations between the documents, for instance, there is some similarity between documents d_2 and d_3 despite they do not share any terms. It is also noticeable the similarity between documents d_4, d_5 and d_6, which have one term in common.

We can appreciate in a very general way, the appearance of groups of documents. Documents d_1, d_2 and d_3 could be gathered into one group due to the proximity be-

tween them; and documents d_4, d_5 and d_6 into another different group. Thus we could think that documents in which terms co-occur may be related by similar concepts. In fact, each of these groups could be considered as a topic cluster, in which documents that deal with the same topic are grouped together. Remarkably, LSA is capable to do this job also considering synonymy between index-terms. We are aware that this is a premature conclusion, and the analysis of this simple example does not allow us to assert that result, but this is something we want to note at this point. In Section 3.2 we will discuss in more detail the automatic topic clustering of documents.

The Latent Semantic Analysis starts with a Term-Document Matrix. If weighting schemes have been applied to the TDM, then the analysis must be performed on the weighted version of the TDM, that is the matrix W we obtain after applying the weighting and normalization schemes. LSA makes use of the Singular Value Decomposition (SVD) method by decomposing the Weighted Term-Document Matrix ($W_{m\times n}$) into the product of three matrices ($T_{m\times m}$, $S_{m\times n}$ and $D^T_{n\times n}$), as shown in Figure 3.4.

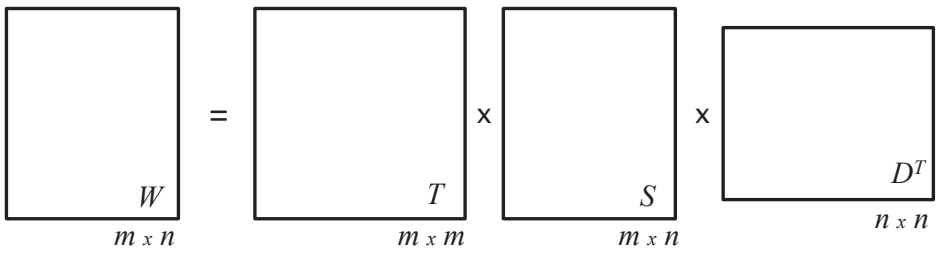


Figure 3.4: Singular Value Decomposition of the Weighted Term Document Matrix

Accordingly, we have:

$$W_{m\times n} = T \cdot S \cdot D^T \tag{3.7}$$

where m is the number of index-terms, n the number of documents in the collection and D^T means the transpose of the matrix D. The elements of this decomposition are described below:

- T is a square matrix whose columns are the eigenvectors of WW^T. These eigenvectors must be orthonormalized so they fulfill the condition $T^T \cdot T = I$. The matrix T define the term vector space in the latent semantic space.

- D is a square matrix whose columns are the eigenvectors of W^TW. These eigenvectors must be orthonormalized so they fulfill the condition $D^T \cdot D = I$. The matrix D define the document vector space in the latent semantic space.

- S is a rectangular matrix whose main diagonal contains the singular values of W. These values must be arranged in descending order such as $diag(S) = [\lambda_1, \lambda_2, \ldots, \lambda_r]$ where $\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0$. Where $r \leq min(m, n)$ is computed as the $rank$ of the matrix W. The singular values λ_i indicate the "weight" or the "contribution" along the new i-th dimension of the latent semantic space.

At the beginning of the analysis, there was only one matrix (matrix W), and now the decomposition has lead to the appearance of three new matrices (T, S and D). It has been apparently increased the size of the representation. Nonetheless, it should be noticed that matrices T and D are linked to matrix W and therefore do not represent new parameters. The key point in the LSA method is that, after the singular value decomposition, a number of the linearly independent components is very small, and may be ignored (the corresponding λ_i of the main diagonal of S are considered negligible), leading to an approximate model that could contain many fewer dimensions.

By selecting the first k largest singular values with some criterion, and their related columns and rows in matrices T and D^T respectively, it is possible to obtain a truncated representation of the latent space using fewer dimensions, as shown in Figure 3.5.

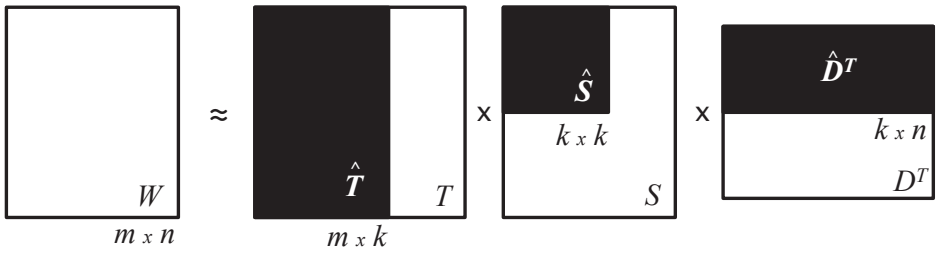


Figure 3.5: Approximate representation of the Weighted TDM by the LSA technique

This truncated representation is an approximation of the original matrix:

$$W_{m\times n} \approx \hat{T}_{m\times k} \cdot \hat{S}_{k\times k} \cdot \hat{D}^T_{k\times n} \tag{3.8}$$

The value of k fixes the number of dimensions in the truncated latent semantic space. The selection of the value for k may be conditioned by the size of the document collection and the particular conditions of the task for which LSA is being used. It should be noted that in many research developed in this area, the value for k has been suggested under the assumption that the collections are composed of thousands (even tenths of thousands) of documents. However, there are not well defined criteria to select this value. Thus, we can observe in the literature a certain variance in the value for this parameter depending on the particular collections and original index-term inventories of their experiments. Some authors suggest that the best value may be in the range of 50-100 [Deerwester et al., 1990]. Other authors had pointed out that it may be approximately 300 [Landauer and Dumais, 1997], while others claim that frequently chosen values are 100 and 150 [Manning and Schütze, 1999].

One of the major drawbacks that have been reported over the years regarding the application of this technique, is the large computational cost of implementing the SVD for large collections. Nowadays, thanks to the computational capability of modern systems, these limitations are minimized. However, the application of LSA will obviously still be dependent on the application domain and the specific characteristics of the document collections.

3.1.2.1 Similarity measure for the Latent Semantic Analysis

In the Vector Space Model, the documents, as well as the query, are represented in a m-dimensional space (m is determined by the number of index-terms of the TDM). Each document has as many dimensions as the query, and therefore the similarity between them can be computed as a mathematical operation (the cosine distance for example) between two vectors in the same dimensional space.

In the latent semantic space, nevertheless, the initial m-dimensional space has been mapped into a lower dimensional space (being k the number of dimensions). In consequence, to compute the similarity between a query and a document, the query must be first represented as a vector in this k-dimensional space.

We can consider the query as a separate document. Therefore, to represent the query in the k-dimensional space, we must first derive an equation that allows us to express a document, as a function of the index-terms in the W matrix, in the latent space.

Once this equation is derived, then we can extend the same set of operations to represent the query in the latent space as a function of its index-terms. From Eq. (3.7), we have:

$$W = T \cdot S \cdot D^T$$

Solving for D, we have:

$$\begin{aligned} W^T &= (T \cdot S \cdot D^T)^T \\ W^T \cdot T &= (T \cdot S \cdot D^T)^T \cdot T \\ W^T \cdot T &= D \cdot S \cdot T^T \cdot T \end{aligned}$$

we know that $T^T \cdot T = I$, then

$$W^T \cdot T = D \cdot S$$

and hence

$$D = W^T \cdot T \cdot S^{-1} \tag{3.9}$$

Since we are considering a truncated representation, then we have:

$$\hat{D} \approx W^T \cdot \hat{T} \cdot \hat{S}^{-1} \tag{3.10}$$

Each of the rows of W^T represents a document in the $m-$dimensional initial space. Therefore, by multiplying the i-th row of W^T by $\hat{T} \cdot \hat{S}^{-1}$ we obtain the vector representation, in the latent semantic space, of the i-th document.

As we said previously, to compare the similarity between a query and a document, the query must be mapped in the same latent semantic space in which the documents are represented. Thus, we can consider the query as a separate document and by using the same transformation we previously used to obtain the i-th document in the latent

space (that is $\hat{T} \cdot \hat{S}^{-1}$) we obtain the vector representation of the query in the latent semantic space. That is:

$$\vec{q}_{lsa} = \overrightarrow{wq}^T \cdot \hat{T} \cdot \hat{S}^{-1} \tag{3.11}$$

where $\overrightarrow{wq}$ is the weighted version of the query vector obtained in Eq. (3.5).

Then, the query, finally represented by $\vec{q}_{lsa}$ in the truncated latent semantic space, can be compared to all existing document vectors in the matrix $\hat{D}$, by measuring the similarity between them using the cosine similarity described in the Eq. (3.6), as follows

$$sim\left(\vec{\hat{d}}_j, \vec{q}_{lsa}\right) = \frac{\vec{\hat{d}}_j \bullet \vec{q}_{lsa}}{\|\vec{\hat{d}}_j\| \, \|\overrightarrow{q_{lsa}}\|} \tag{3.12}$$

Similarly as in the Vector Space Model, for an IR application, a ranking of the documents can be done by arranging them from the highest to the lowest in regard to the similarity measure. Recall that for the topic identification task, a hard decision is usually adopted. This can be done by means of a classifier. There are several approaches for implementing a classifier for the topic identification task. Some of these approaches were introduced in Section 1.1.6. In this Thesis we make use of the Centroid based classifier, which we examine in the next section.

3.1.3 Centroid based classifier

Centroid based classifier is one of the most popular algorithms in text classification. In this method, documents are represented using the vector model approach. Note that the documents can be represented in the original m dimensional space of the Vector Space Model or in the space of the Latent Semantic Analysis model.

The idea behind this classifier is extremely simple. Let us assume that $A = \{a_1, a_2, \ldots, a_{|A|}\}$ is the predefined set of $|A|$ topics (which in this case we can also refer to as classes) that are available in the document collection. The data that belong to a class a_z (i.e. all the documents in the collection labelled as a_z) are represented by a unique vector that is at the center of the class (commonly known as the centroid vector). Thus, based on the assumption that the centroid vector of a set of data is the best representative of these data, each topic is represented by a single centroid vector. In centroid based classification, the training consists of calculating the centroid vector of each class.

If there are $|A|$ topics in the training set, this leads to $|A|$ centroid vectors $\{\vec{C}_1, \vec{C}_2, \ldots, \vec{C}_{|A|}\}$ where each $\vec{C}_i$ is the centroid vector for the i-th topic.

There are different ways to compute the centroid vector for topic a_z. One of these approaches, known as Cumuli Geometric Centroid (CGC) considers the sums of the weights of the various terms present in the n_z documents d_j that belong to class a_z [Guan et al., 2009]. In this approach the centroid vector for topic a_z is computed as

follows:

$$\vec{C}_z = \sum_{d_j \in a_z} \vec{d}_j \tag{3.13}$$

A different approach considers the average of the weights rather than their sum. This averaging approach is known as Arithmetic Averaging Centroid (AAC), where the elements are simply the mean values of the corresponding term weights [Han and Karypis, 2000]. In this approach we define the centroid vector for topic a_z to be:

$$\vec{C}_z = \frac{1}{|n_z|} \sum_{d_j \in a_z} \vec{d}_j \tag{3.14}$$

For evaluating a new document (let us call this new document q which is represented by the vector $\vec{q}$), we simply use the similarity measure defined for the Vector Space Model, i.e. the *cosine* similarity. Therefore, we compute the cosine between the vector of the document we want to evaluate, $\vec{q}$, and all the centroid vectors. Based on these similarities, we assign q to the class corresponding to the most similar centroid, that is the class of q given by

$$\arg\max_{z=1,\ldots,|A|}(\cos(\vec{q}, \vec{C}_z)) \tag{3.15}$$

The Centroid based classifier is a simple and efficient method and its properties make it preferable when compared with mathematically more complex classifiers. [Han and Karypis, 2000] compared the centroid based classifier with k-nearest neighbor (k-NN), and Naive Bayes classifiers while evaluating a topic identification system on different document collections, and showed that it yields comparable results with better time complexities.

Despite centroid-based classification is an efficient approach in general, it has two main problems:

- The tendency to be affected from small variations in the data; it was shown that, for some domains, filtering the outliers in the data improves the classification performance of the classifier by about 10% when compared to the classical centroid-based approach [Shin et al., 2006].
- It is sensitive to term-weighting, thus an initial tuning of the weighting parameters must be performed in order to obtain the best performance for this classifier.

3.1.4 Term selection strategies

Index-terms selection is an important step in topic identification. It encompasses several methods that aim to choosing, from the available term inventory, a smaller set of terms that more efficiently represents the documents. Index-terms selection has two main objectives: First, it makes the identification process more efficient by decreasing

the size of the effective index-terms. This is of particular importance for classifiers that are expensive to train. Second, index-terms selection often increases topic identification accuracy by eliminating noisy index-terms, thus increasing the robustness of the model by minimizing the number of parameters. However, in removing terms the risk is to remove potentially useful information of the documents. It is clear that, in order to obtain optimal effectiveness, the selection process must be performed with care.

A common approach to select the index-terms is the so-called *filtering approach*. This approach is based on keeping the terms that receive a higher score according to a specific function, or metric, that measures the relative importance of a term within the collection. In the next sections we will present some of these functions.

There are different ways of implementing a filtering approach for index-terms selection. One way is to select as index-terms those terms whose metric exceed a predefined threshold. These terms would compose the term-inventory that is used for document representation. Terms below the threshold would be discarded and therefore not included in the term-inventory. However, setting the threshold to a specific value is not a trivial decision to make and care must be taken since discriminant index-terms could be discarded.

Another way to implement a filtering approach is by ranking the terms according to the value of the specific function. The ranking is usually performed by sorting the terms in ascending order (that is from the smallest to the largest value). Starting from an initial term inventory, a new set of terms may be generated by removing the term in the first position of the ranking, that is the term with the lowest value. Once the new term inventory is generated, a classifier based on it is built and then tested on the development dataset. This procedure is repeated a number of times, removing the next term in the ranking in each iteration. The term inventory that results in the best effectiveness of the system is finally chosen. In this work we followed this latter approach for evaluating the index-terms selection strategy. We are aware that this is computationally more expensive than the former approach (by setting a threshold) but it allows us to perform a complete evaluation of all metrics by considering several index-terms inventories, and besides it allows us to evaluate the impact of the index-term reduction in the performance of the topic identification system.

Before proceeding, it is useful to define a term-class incidence table (see Table 3.2). These values are needed in next sections to define the different strategies for index-terms selection. In this table n_i is the number of documents from the training dataset that contain the term t_i, and n_z the number of documents from the training dataset assigned to topic a_z. Note that the index-term selection is performed on the training dataset, so the topic assignments correspond to the topic hand labeling of the original corpus. The number of documents that contain term t_i and are assigned to class a_z is given by $n_{i,z}$. The remaining quantities are calculated analogously.

Different methods have been proposed, either from the information theory and machine learning fields, to select the set of index-terms. Next, we briefly introduce the methods that we apply in the index-terms selection strategy. We are not deeply covering the details of these metrics (see Baeza-Yates and Ribeiro-Neto [2011] and Manning et al. [2008] for a detailed description of these techniques).

Case	Docs in a_z	Docs not in a_z	Total
Docs that contain t_i	$n_{i,z}$	$n_i - n_{i,z}$	n_i
Docs that do not contain t_i	$n_z - n_{i,z}$	$N_t - n_i - (n_z - n_{i,z})$	$N_t - n_i$
All docs	n_z	$N_t - n_z$	N_t

Table 3.2: Term-class incidence table

3.1.4.1 Inverse document frequency

A simple but efficient metric is to consider the inverse document frequency of terms. It can be computed, for each term t_i, by using the expression for the inverse document frequency as it was presented for the idf weighting scheme. Though it must be noticed that for ranking purposes, that is for ranking the value of the idf of the terms, the logarithm function is not essential. For simplicity, we will keep the same expression as for the weighting scheme (in this case adapted to the values in Table 3.2). This metric is then given by:

$$idf_i = \log\left(\frac{N_t}{n_i}\right)$$

where N_t is the number of documents in the collection and n_i is the number of documents containing the term t_i. The idf of a frequent term is likely to be low whereas the idf of a rare term is high. If we rank the idf values from the smallest to the largest values, we will find in the first positions of the ranking, frequent terms that appear in a high number of documents; in the filtering approach we will follow for the index-terms selection, those terms will be the first to be removed from the term inventory.

3.1.4.2 Mutual information

This technique measures the relative entropy between two variables; in our case these variables refer to a term and a topic. If those variables are independent, then their mutual information is zero. This utility measure is defined, for a term t_i and a class a_z as the expected value of

$$I(t_i, a_z) = \log\left(\frac{P(t_i, a_z)}{P(t_i)P(a_z)}\right) = \log\left(\frac{\frac{n_{i,z}}{N_t}}{\frac{n_i}{N_t} \times \frac{n_z}{N_t}}\right) = \log\left(\frac{n_{i,z} \times N_t}{n_i \times n_z}\right) \quad (3.16)$$

computed across all classes. That is

$$MI(t_i, A) = \sum_{z=1}^{|A|} P(a_z) I(t_i, a_z) \quad (3.17)$$

$$= \sum_{z=1}^{|A|} \frac{n_z}{N_t} \log\left(\frac{n_{i,z} \times N_t}{n_i \times n_z}\right) \quad (3.18)$$

3.1.4.3 Information gain

This metric is complementary to mutual information. It considers not only the presence of terms in the documents but also their absence. It is defined for a term t_i over all classes A as

$$IG(t_i, A) = H(A) - H(A|t_i) - H(A|\neg t_i) \tag{3.19}$$

Where $H(A)$ is the entropy of the set of topics A, computed as follows

$$H(A) = -\sum_{z=1}^{|A|} P(a_z) \log P(a_z) \tag{3.20}$$

and $H(A|t_i)$ and $H(A|\neg t_i)$ are the conditional entropies of A in the presence and the absence of term t_i, respectively, computed as follows

$$H(A|t_i) = -\sum_{z=1}^{|A|} P(t_i, a_z) \log \left(\frac{P(t_i, a_z)}{P(t_i)} \right) \tag{3.21}$$

$$H(A|\neg t_i) = -\sum_{z=1}^{|A|} P(\neg t_i, a_z) \log \left(\frac{P(\neg t_i, a_z)}{P(\neg t_i)} \right) \tag{3.22}$$

Eq. (3.19) can be rewritten as a function of the elements of the contingency table as:

$$IG(t_i, A) = -\sum_{z=1}^{|A|} \left(\frac{n_z}{N_t} \log \left(\frac{n_z}{N_t} \right) - \frac{n_{i,z}}{N_t} \log \left(\frac{n_{i,z}}{n_i} \right) - \frac{n_z - n_{i,z}}{N_t} \log \left(\frac{n_z - n_{i,z}}{N_t - n_i} \right) \right) \tag{3.23}$$

3.1.4.4 Chi-square

Chi-square is another popular index-terms selection method. In statistics, the Chi-square test is applied to test the independence of two events, where two events A and B are defined to be independent if $P(A, B) = P(A)P(B)$ or, equivalently, if $P(A|B) = P(A)$ and $P(B|A) = P(B)$. In our case, the two events are occurrence of the term t_i and occurrence of the topic a_z. It is defined as follows

$$\chi^2(t_i, a_z) = \frac{N_t(N_t n_{i,z} - n_z n_i)^2}{n_z n_i (N_t - n_z)(N_t - n_i)} \tag{3.24}$$

In order to apply this metric for index-term selection, we can apply two different criteria: to compute either the average or the maximum term value of Chi-square as follows:

$$\chi^2_{avg}(t_i) = \sum_{p=1}^{|A|} \frac{n_p}{N_t} \chi^2(t_i, a_p) \tag{3.25}$$

$$\chi^2_{max}(t_i) = \max_{p=1}^{|A|} \chi^2(t_i, a_p) \tag{3.26}$$

3.2 Contributions on Topic Identification

Within the contextualization framework that we propose for the enhancement of a speech recognition system, the identification of topics in a transcript of an audio segment is a fundamental step in the topic-motivated adaptation of language models. In this regard, a reliable topic identification system could provide adequate information concerning the topic that is being addressed in the speech while enabling an appropriate adaptation of language models.

In the topic identification task, we have focused mainly on the enhancement of preprocessing procedures, in addition to contributing in the definition of more robust criteria for the selection of *index-terms*.

Our main contributions on Topic Identification are:

- The evaluation of different strategies for the selection of index-terms. In this sense we want the selection of index-terms to be dependent on the specific domain of the task.
- To evaluate and compare different weighting schemes. There have been proposed different weighting schemes in the literature, but there is little consensus on which method performs better. Our evaluation aims to shed some light on the matter.
- To compare and evaluate different models for document representation that allow us not only to reduce the document space representation but also to enhance the topic identification system.
- The proposal of an *ad-hoc* global weighting scheme that may lead to a reduction of the topic identification error.

3.2.1 On the proposal of an *ad-hoc* weighting scheme

Among the most common weighting schemes, *term entropy* is based on an information theory approach and it exploits the distribution of terms over documents [Dumais, 1991]. This weighting scheme was previously introduced in Section 3.1.1.1.2. For the index-term t_i in the document d_j, it is defined as follows:

$$te_i = 1 + \frac{1}{\log(n)} \sum_{j=1}^{n} p_{i,j} \cdot \log(p_{i,j}), \text{ where } p_{i,j} = \frac{c_{i,j}}{gf_i} \tag{3.27}$$

Where $c_{i,j}$ represents the raw frequency of the index-term t_i in the document d_j. gf_i is the raw global frequency of the index-term t_i, which is equal to the number of times the term t_i appear in the documents of the collection. From the implementation point of view, this scheme may lead to a log zero calculation if an index-term is not present in a document. That is

$$p_{i,j} = 0, \quad \text{if } c_{i,j} = 0$$

Different solutions have been suggested to solve this problem. One possible solution is to approximate $p_{i,j} \cdot \log p_{i,j} \approx 0$. Another possible solution is to include a smoothing parameter a, resulting in $p_{i,j} = (a + c_{i,j})/gf_i$. Indeed, both approaches solve the log zero calculation. However, in different experiments that we have performed evaluating both solutions, combining this scheme with *term frequency* as local scheme, the results have shown that they do not significantly improve the *tf-idf* baseline weighting scheme.

In this sense, we propose an *ad-hoc* weighting scheme, that we called *pseudo-entropy*. In this scheme, which is based on the *term entropy* weighting scheme, the parameter $p_{i,j}$ is calculated as the weighted sum of $c_{i,j}$ and the inverse of gf_i, as follows:

$$p_{i,j} = \beta \cdot c_{i,j} + \frac{\gamma}{gf_i} \tag{3.28}$$

The aim of the β parameter is to emphasize the count of the term t_i in document d_j. The γ parameter, and in general the expression γ/gf_i, aims to avoid the value of zero for $p_{i,j}$ in those cases in which the term $t_i$is not present in the document d_j. This expression, at the same time, provides a value which is a function of the inverse of the global frequency gf_i. For terms that are very frequent in the collection (and thus have a high global frequency gf_i), this expression will have a low contribution in computing $p_{i,j}$. In contrast, for rare terms (usually those with a low gf_i) the contribution of this expression to $p_{i,j}$ will be higher.

By computing $p_{i,j}$ this way, we are not only avoiding the log zero calculation but we are also accounting a small value γ/gf_i for those terms that do not appear in the document. The value of the term entropy highly depends on the values of β and γ parameters. For this reason, these parameters must be carefully adjusted on the development set prior to conducting the evaluation of the topic identification system.

The evaluation of this scheme has shown a significant reduction of the topic identification error for one of our evaluation datasets, although no significant differences were achieved when compared to existing weighting schemes. The absolute minimum error obtained in our whole experimental framework was achieved by using this *ad-hoc* weighting scheme. These results are described later in Section 3.3.11.

The benefits of the application of the *pseudo-entropy* scheme on the performance of the topic identification system might be explained by analyzing the Figure 3.6. In this figure, we aim to compare the values of different global weights (*idf*, *term entropy* and the proposed *pseudo-entropy*) for the terms in the collection. The values of the distinct weights have been ordered and are presented in ascending order.

We can see, for instance, that the *term entropy* weight takes values near zero for a large number of terms in the collection. Recall from Section 3.1.1.1.2, that these terms are those that are distributed across many documents in the collection. The *idf* weight, in turn, takes a value that remains the same for those terms that appear in only one document (this is the value that appears in the figure as the maximum *idf* value).

In contrast to these schemes, the *pseudo-entropy* aims to take a broader range of different values for those terms distributed along the collection as well as for those terms that appear in only a few number of documents. Recall that the *pseudo-entropy* takes into account small values even in the case for which the term is not present in

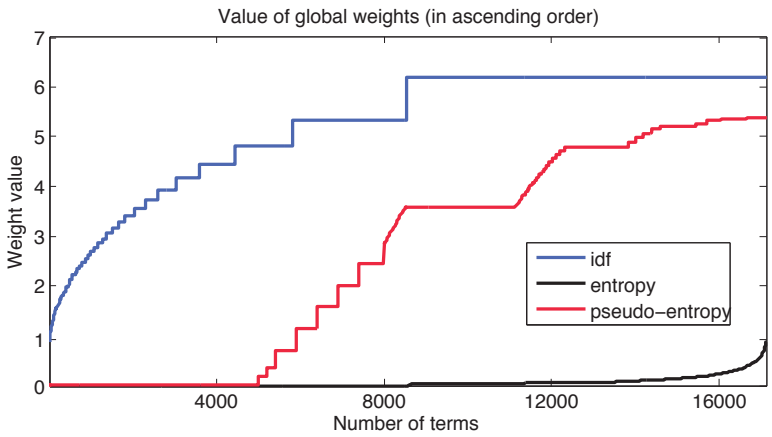


Figure 3.6: Comparison of the value of global weights for the terms in the collection (in ascending order)

the documents. We can appreciate in the figure that this scheme modifies the dynamic range of the weight values for the terms in the collection when compared to *idf* and *term entropy*. We believe that this property of the *pseudo-entropy* weighting scheme enhances the topic identification process by modifying the global weight of terms and thus modifying the way in which documents are represented.

3.3 Experiments on Topic identification

Our principal objective in this Thesis is to propose and evaluate a framework of topic-motivated contextualization based, ultimately, on the dynamic and non-supervised adaptation of language models for the enhancement of an automatic speech recognition system. To achieve this objective we have divided the framework in two principal systems: a *topic identification* system and a *dynamic language model adaptation* system. In this section we present our main contributions regarding the topic identification system.

Within our topic-motivated contextualization framework we propose different approaches for the adaptation of language models. These approaches depend on how the identification of topics is performed. We will detail these approaches later in this Chapter, but for now, let us note that one of these approaches requires an automatic and supervised topic identification system to decide which topic-based model to use in the adaptation process. In this Section we present the experiments conducted in the development of this system and the results that we obtained.

The experiments carried out during this research focus on speech recognition in a domain where multiple topics are covered. In this sense, the Spanish partition of the EPPS (*European Parliament Plenary Sessions*) database was appropriate for a number of reasons. First and foremost, it contains a wide range of topics. This suggests that there are potential gains to be obtained from a topic-inspired language model adaptation methodology. Second, this database contains many different styles of language,

ranging from natural and spontaneous to prepared, read and speech with long pauses (like the common pauses that are heard when an interpreter is translating someone else's speech). It also contains a wide variety of acoustic conditions ranging from high quality audio speech to speech with background noise. This may indicate that any findings made during this research could be hopefully applicable to different domains. Finally, it contains a considerable amount of hours of transcribed recordings, which allows the contextualization framework to be evaluated on a quite large data set.

The experiments on topic identification and language model adaptation were conducted on the EPPS database which is described in the following section.

3.3.1 The EPPS database

We have used the Spanish partition of the EPPS database (*European Parliament Plenary Sessions*) of the TC-STAR (*Technology and Corpora for Speech to Speech Translation*) European Project [Mostefa et al., 2007] to evaluate the systems proposed in this Thesis.

Compared with other research and other databases, in which the contextualization is performed on multiple and varied domains (e.g. sports, economy, culture, science, politics, etc.) in this research we focus on a single domain, the political domain. Within this domain, the EPPS database offers a broad set of topics. Table 3.3 shows some examples of the topics found in this database.

- *Formal opening of the first sitting of the enlarged European Parliament.*
- *Enlarged Europe and its neighbours.*
- *Situation in Ukraine.*
- *Resumption of the session.*
- *Work programme of the Netherlands presidency.*
- *United Nations Framework Convention on Climate Change.*

Table 3.3: Examples of topic labels in the EPPS database.

Due to the fact that the original *training* dataset of this database is the only one that includes distinct labels for the topics, we partitioned it into new training, development and evaluation datasets for our experimentation. This part of the EPPS database comprises approximately 61 hours of audio recordings of the European Parliament plenary sessions (and their corresponding transcriptions) recorded from 2004 to 2007. The language of the corpus is Spanish. There are both male and female speakers (approx. 75% - 25% respectively distributed).

We have selected a typical 70-10-20 distribution for the training, development and final evaluation datasets. The development dataset will be used for tuning some model parameters and to compare the performance between different systems. This allows us to decide which system, or which combination of parameters, performs better in order to choose the model for the final evaluation of the system on the evaluation dataset.

We believe that identifying the topic on short sentences can be ambiguous because few words may not provide enough semantic information about the topic that is being addressed. For instance, a short sentence like "*Gracias señor presidente, pasamos ahora a otro tema*" ("Thank you president, going on to a different subject") can be related to any of the topics of the collection. For this reason we decided to split the database into segments of audio corresponding to a complete *turn of intervention* of a speaker.

Besides, we fixed a minimum length for selecting the turns of intervention that compose both development and evaluation datasets. By these criteria we obtain a *training* set composed of 21127 sentences, grouped in 1802 turns of intervention. *Development* set is composed of 2402 sentences grouped in 106 turns of intervention, and the *evaluation* set is composed of 3738 sentences.

We have applied two different criteria for breaking the sentences in this evaluation set into turns of intervention. By these criteria we have generated two configurations for the *evaluation* set:

i) *Evaluation Set 1* is created with turns of intervention with a minimum length of approximately one minute. Turns 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 turns of interventions for this evaluation set. Each of these turns of intervention belongs to one of the available topics as specified on the original hand labeling of the database.

ii) *Evaluation Set 2* is created based on the same turns of intervention of the *Evaluation Set 1*, except that in this case, turns that are significantly longer than one minute are segmented into smaller segments. By following this criterion we have obtained 754 audio segments for the same evaluation data.

    We create this evaluation dataset for two main reasons: in the first place, we want to do an exploratory analysis of the length of the turn of intervention on both topic identification and speech recognition performance; and besides we want to increase the number of the evaluation elements in order to reduce the confidence intervals of our results. By having a larger number of audio segments (in this case segments of turns of intervention) we increase the possibility of finding significant improvements by reducing the confidence intervals.

It is important to notice that both evaluation datasets are composed of the same sentences, which means that they are also composed of the same audio segments. The difference between them is the way in which we have grouped the individual audio segments together to form interventions with different lengths.

The lexicon size is 16528 words. Each of the turns of intervention belongs to one of 67 different topics. The summarized details of the database are shown in Table 3.4.

Language:	Spanish
Speakers gender:	Male (approx. 75%) and female (approx. 25%)
Domain:	Political
Number of topics:	67
Training set:	21127 sentences grouped in 1802 turns of intervention
Development set:	2402 sentences grouped in 106 turns of intervention
Lexicon size:	16528 words
Evaluation *Set 1*:	3738 sentences grouped in 252 turns of intervention
Evaluation *Set 2*:	Same 3738 sentences as in *Set 1* grouped in 754 segments of turns of intervention

Table 3.4: Details of the database used for the evaluation

Documents in the collection are not uniformly distributed along all the topics. Figure 3.7 shows the distribution of documents in the training dataset ordered by number of documents. From Figure 3.7, we can appreciate that there is a noticeable difference

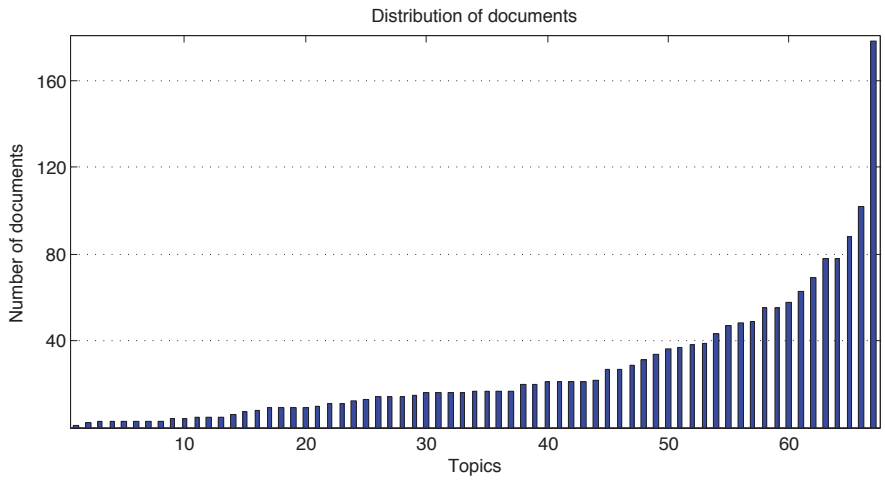


Figure 3.7: Distribution of documents along the topics in the collection

between the number of documents assigned to the most frequent and the less frequent classes in the collection.

However, it must be considered, that for the largest classes, in general, the average length of the document belonging to that class is clearly smaller than the average length of documents for the less common classes (in Figure 3.8 the average length of the documents for each topic is shown). Actually, in Figure 3.8 we can see that the maximum average length values are obtained for some of the less frequent classes. This suggests some kind of balance between the number of documents assigned to each class and the average length of these documents.

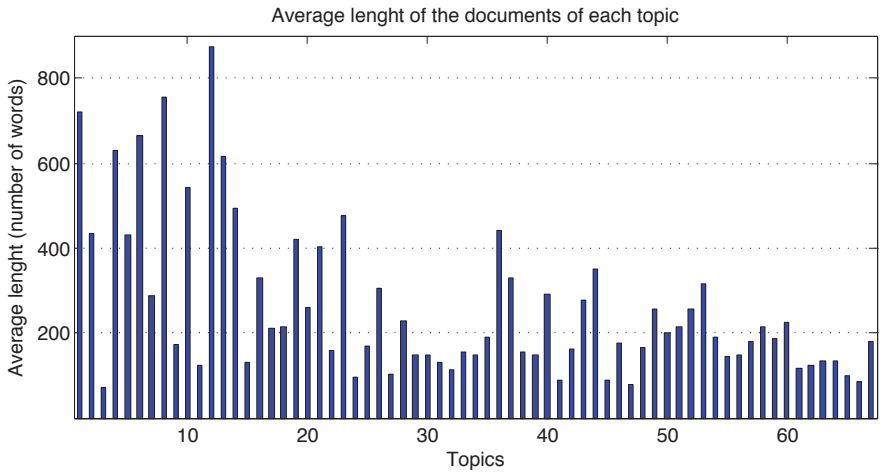


Figure 3.8: Average length of the documents assigned to each topic

Besides the average length, we should also highlight the total length of the documents that belong to each topic, according to the original topic labels. This length is shown in Figure 3.9. The topics are displayed in the same order as in the previous figures.

The total length of documents is clearly not the same for all topics. There are considerable differences between the most frequent topics and the less frequent topics. These differences may hinder the topic identification process.

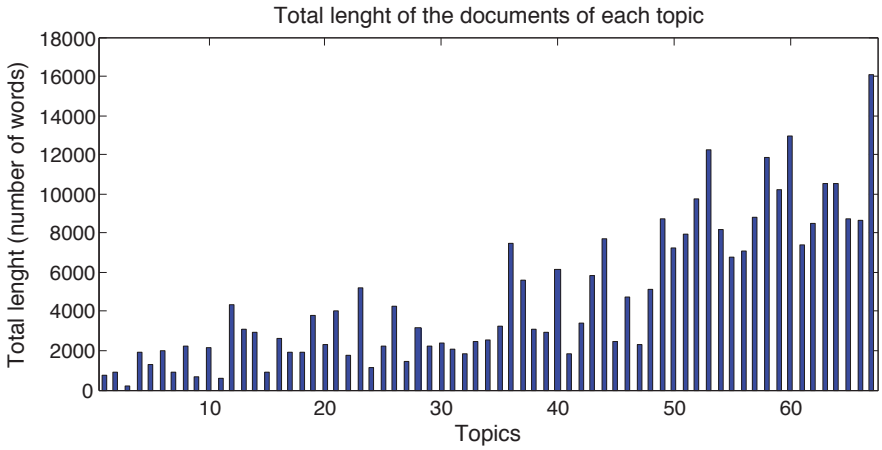


Figure 3.9: Total length of the documents assigned to each topic

3.3.2 Evaluation metrics

In this section we review the evaluation metrics that are used to measure the capacity of a classifier to take the right decisions. These metrics can be used to compare the performance of a classifier (in our case, the topic identification system) under different configurations. This would allow us not only to evaluate whether the feature selection approaches and the term weighting schemes enhance the topic identification system, but also which of the different approaches we propose can be considered the best for the topic identification problem that we are addressing in this Thesis.

To evaluate the performance of a classifier, the classification problem can be broken into several one-against-all binary classification problems. In this scenario, a contingency table for each class a_z, representing the condensed possible outcomes for the classification, should be defined as shown in Table 3.5.

Class a_z		Reference labels	
		YES	**NO**
Classifier decision	**YES**	TP_z	FP_z
	NO	FN_z	TN_z

Table 3.5: Contingency table for class a_z

In this table TP_z (*true positives*) is the number of test documents correctly assigned to class a_z. FP_z (*false positives*) is the number of test documents incorrectly assigned to class a_z. FN_z (*false negatives*) is the number of test documents that actually belong to class a_z but were misclassified, and finally TN_z (*true negatives*) is the number of test documents correctly assigned to a class other than a_z.

Next, we present different types of measures based on this contingency table.

3.3.2.1 Accuracy and Error.

These measures are common in the machine learning literature and have been used in several evaluations of text categorization systems. *Accuracy* is the fraction of the test documents that have been correctly classified. In turn, *Error* is the fraction of the test documents assigned to incorrect classes by the classifier. Note that the sum of both measures must be one. They can be defined for measuring the effectiveness of the classifier for each individual class a_z as

$$Acc_z = \frac{TP_z + TN_z}{n_t} \tag{3.29}$$

$$Err_z = \frac{FP_z + FN_z}{n_t} = 1 - Acc_z \tag{3.30}$$

Where $n_t = TP_z + TN_z + FP_z + FN_z$, is the number of test documents. Or, they can also be defined globally, that is for measuring the global effectiveness of a classifier, as

$$Acc_{global} = \frac{\text{Num. of test documents correctly classified}}{\text{Num. of test documents } (n_t)} \tag{3.31}$$

$$Err_{global} = 1 - Acc_{global} \tag{3.32}$$

3.3.2.2 Precision and Recall.

These measures are the most frequent and basic tools for measuring the effectiveness of a classifier. *Precision* (Pre_z) is the fraction of test documents assigned to the class

a_z that really belong to class a_z. *Recall* (Rec_z) is the fraction of test documents that belong to class a_z that were correctly assigned to class a_z. Both measures are defined, for a given class a_z, as follows

$$Pre_z = \frac{TP_z}{TP_z + FP_z} \tag{3.33}$$

$$Rec_z = \frac{TP_z}{TP_z + FN_z} \tag{3.34}$$

These measures are to some extent complementary, since *precision* puts emphasis on false positives and *recall* draws attention to false negatives. Thus, these measures may be misleading when examined alone and therefore it is convenient to combine them into a single one.

3.3.2.3 F-measure.

A single measure that trades off *precision* versus *recall* is the F-measure, which is calculated as the weighted harmonic mean of precision and recall. This measure allows to give different weights to both *precision* and *recall* and it is useful in such cases the system is intended to give more importance to *false positives* or *false negatives*. The F-measure is defined for a class a_z as

$$F_z = \frac{(\alpha^2 + 1)Pre_z Rec_z}{\alpha^2 Pre_z + Rec_z} \tag{3.35}$$

where α defines the relative importance of *precision* (related with false positives) and *recall* (related with false negatives). The most common value for α equally weights both measures, i.e. $\alpha = 1$. This particular measure is called $F1$-measure and is computed, for class a_z, as follows

$$F1_z = \frac{2Pre_z Rec_z}{Pre_z + Rec_z} \tag{3.36}$$

Recall that these measures (*precision*, *recall* and F-measure) are intended to evaluate the effectiveness of the classifier for a particular class a_z. In order to evaluate the performance of the classifier across all classes, these measures may be averaged in two distinct ways: by *macro-averaging* all the relative measures or by *micro-averaging* them, as we describe below.

3.3.2.4 Macro and Micro-averaging.

Macro-averaging computes a simple average over classes. In this sense, *macro-average* performance scores are computed by first computing the scores for the per-class contingency tables and then averaging these per-class scores to compute the global means. In turn, *micro-averaged* performance scores are computed by first creating a global contingency table whose cell values are the sums of the corresponding

cell in the per-class contingency tables, and then use this global contingency table to compute the *micro-average* performance scores.

There is a distinction between these averages. *Micro-average* performance scores give equal weight to every document, and is therefore considered a per-document average. Analogously, *macro-average* performance scores give equal weight to every class, regardless of the number of documents belonging to each class, and is therefore a per-class average.

3.3.2.5 Final considerations on evaluation metrics.

None of the described measures is perfect or even appropriate for every problem. What type of measure is more preferable, depends entirely on the application. For example, *recall*, if used alone might show deceptive results (imagine a system that classifies all test documents as belonging to a given class; it will show perfect *recall* for that class, since false negatives will be zero, and therefore *recall* will reach its maximum).

Accuracy as well as *error* (since they are complementary), works well when the number of documents in all classes is balanced, but in extreme conditions they might be deceptive too. If the number of documents in a class is very large compared to the number of documents in the other class, a very simple classifier that simply rejects the small category would have a good accuracy.

This is not to suggest that a trivial rejector classifier is good, but that accuracy or error may not be a appropriate measure of the effectiveness of a classifier when the classes are extremely skewed.

3.3.3 Experimental framework

The evaluation we have carried out for the Topic Identification task consists of identifying the topic that is discussed in each of the transcriptions provided by the first decoding pass (i.e the output of the ASR stage 1 - see Figure 2.1). Note that these transcriptions contain recognition errors, which means that the topic identification task must be performed on text to which index-terms, that appear in the original transcription of the audio segment, may be missing. Precisely, the objective of the contextualization framework is to reduce these recognition errors in the second decoding stage.

We focus our objectives on different aspects with the aim of improving the effectiveness of the topic identification system:

- The enhancement of document preprocessing techniques. We are aware that the efficiency of a topic identification system, depends considerably on the mechanisms of preprocessing that are applied to the documents in the corpora used by the system. These mechanisms allow to convert documents to a more concise and convenient format and have a substantial impact on the success of the topic identification process.

- The comparison between different weighting schemes. Our aim in this regard is to compare and evaluate alternative approaches to traditional term weighting schemes that allow us not only to constrain the selection of the most significant terms but also to improve the properties of the term as a descriptor of a document topic.
- The definition of more robust criteria for the selection of index-terms. A proper selection of index-terms in a document collection is essential to establish conceptual and semantic relationships not only between terms and documents but also between terms. It also allows to reduce the size of the term-inventory.

All results are obtained by measuring the topic identification error (an evaluation metric described in Section 3.3.2). We selected this metric because it allows a more quickly appreciation of the different outcomes between experiments.

To conduct the experiments on Topic Identification we developed and used our own code written in MATLAB.

3.3.4 Vector Space Model for topic identification - baseline method

For the topic identification task, the initial performance of the system, i.e. the baseline system, was obtained by using the generalized Vector Space Model for document representation, a generic stopword list composed of 278 words[1], and a classic *tf-idf* weighting scheme.

The procedure for the evaluation, in general terms, is similar to the one described in the centroid based classifier approach. Basically, we obtain a representative vector (centroid) for each topic and then classify each document in the evaluation set based on its similarity with respect to these representative vectors. This process, and in general the steps that we followed in the baseline evaluation procedure, are described as follows:

1. An initial inventory of terms (i.e. the set of index-terms) is obtained by considering all words in the training set of the document collection and removing those terms that are present in the generic stopword list. Hereinafter in this document, this term inventory will be referred as *Term inventory 1*. The size of the *Term inventory 1* is 16250 terms. Note that only the terms that appear in the documents have non-zero entries in the corresponding document vector representation. Then we could expect, by now, a high sparseness in the Term-Document matrix, which may hinder the identification process. This level of sparseness depends on both the length of documents and the narrowness of the term inventory.
2. Each document (in both training and evaluation datasets) is represented using the Vector Space Model. The representation space dimension is determined by the number of index-terms obtained in the previous step.

[1] We use the stopword list available in http://snowball.tartarus.org/algorithms/spanish/stop.txt

3. A representative vector $\vec{C}_z$ for each topic is obtained by the accumulation of the data of all original document vectors in the training dataset belonging to the same topic. With these representative vectors, we built a Term-Document matrix (TDM) by arranging each vector $\vec{C}_z$ as a column in the TDM. Thus, each column of this matrix represents a topic. This TDM has as many columns as topics in the collection and as many rows as index-terms.

4. A *tf-idf* weighting scheme is applied to the elements of this matrix. By doing this, a new matrix, called W (a weighted version of the TDM) is obtained. As in the previous step, each column of the matrix W represents a topic.

5. To classify a document vector $\vec{q}$ (one of the documents in the evaluation set), the next steps are followed:

   (a) A *tf-idf* weighting scheme is applied to the vector $\vec{q}$. By doing this we obtain a weighted document vector $\vec{wq}$.

   (b) The vector $\vec{wq}$ is then classified by calculating the similarity between $\vec{wq}$ and each topic vector in the matrix W, and then selecting the most similar topic as the resultant topic for that document. To do this, the cosine distance is used for similarity measurement.

   (c) These steps must be followed for all the documents in the evaluation set. The topic identification error is calculated as the percentage of miss classifications among all the documents in the evaluation set.

We performed the evaluation of the baseline system on both evaluation datasets. The topic identification error for the **baseline system** is $35.71 \pm 5.91\%$ for the **Evaluation Set 1** and $84.08 \pm 2.61\%$ for **Evaluation Set 2**. All the confidence intervals presented in this Thesis are set at the 95% confidence level.

First of all, it is important to notice that the *Evaluation Set 1* contains less samples than *Evaluation Set 2*, and therefore larger confidence intervals are obtained when analyzing the results for *Set 1*.

Audio segments in *Evaluation Set 1* are larger than in *Set 2*. For larger audio segments, larger transcriptions are obtained and therefore, more index-terms have non-zero entries. This contributes in reducing the sparseness of the document representation.

In the Vector Space Model, the dimensions of the representation space is determined by the number of index-terms. So, in this case we have a representation space of 16250 dimensions. We expect to reduce this number of dimensions by the use of Latent Semantic Analysis as we will see in the next section.

In Chapter 5 we will see how these results affect the language model adaptation and therefore the speech recognition performance. It is worth to anticipate that the result will be just the opposite. This means that we will have a better performance for the speech recognition system in the *Evaluation Set 2* rather than in *Set 1* despite the significant differences in the topic identification error.

3.3.5 Latent Semantic Analysis for topic identification

Latent Semantic Analysis is an alternative technique for document representation that tries to overcome some of the problems of the generalized Vector Space Model (VSM). This technique tries to capture the latent structure in the co-occurrence of terms improving in the first place the independence assumption of the VSM.

Unlike VSM, in which each index-term is considered as a dimension in the representation space, LSA approximates the original space with fewer dimensions, reducing not only the size of the representation space but also the sparseness that may exist in the original TDM.

The objective of the experiment proposed in this section is to evaluate the performance on the topic identification task comparing the generalized Vector Space Model and the Latent Semantic Analysis model for document representation. The results we obtained in the previous section are the baseline results.

In order to apply the LSA model we followed a similar scheme as the one described in the previous section. The steps that we followed are described below.

1. We use the same term inventory as in the baseline approach (*Term inventory 1*). This means that we consider all terms in the training dataset except for those terms that appear in the stopword list.

2. Each document (in both training and evaluation datasets) is represented using the Vector Space Model. The representation space is determined by the index-terms obtained in the previous step.

3. Following the same directions as in the baseline procedure we obtain a representative vector $\vec{C}_z$ for each topic in the document collection. With these representative vectors, we built a Term-Document matrix (TDM), whose dimension are $m \times n$ where m is the number of index-terms, and n the number of distinct topics. Representing the documents that belong to a topic in the collection by means of a topic centroid vector is a first step in dimensionality reduction.

4. A *tf-idf* weighting scheme is applied to the elements of this matrix. By doing this, a new matrix, called W (a weighted version of the TDM) is obtained.

5. Then, LSA is applied to the matrix W. This transforms this matrix into three matrices T, S and D^T. The dimension of matrices T and D is determined by the number of rows and columns in W respectively. In this model we are not truncating the latent semantic space (different experiments on the development set did not show a significant improvement by truncating the number of dimensions), so the resultant dimensions of matrix D are $n \times n$. Recall that the matrix D define the document vector space in the latent semantic space.

6. To classify a document vector $\vec{q}$ (one of the documents in the evaluation set), the next steps are followed:

(a) A *tf-idf* weighting scheme is applied to $\vec{q}$. By doing this, a weighted version ($\overrightarrow{wq}$) of the document vector $\vec{q}$ is obtained.

(b) Vector $\overrightarrow{wq}$ is projected into the latent semantic space by applying the transformation described in Eq. 3.11, i.e. $\overrightarrow{wq}_{lsa} = \overrightarrow{wq}^T \cdot T \cdot S^{-1}$, in which vector $\overrightarrow{wq}_{lsa}$ represents the document q in the latent semantic space.

(c) The document vector $\overrightarrow{wq}_{lsa}$ is classified by calculating the similarity between $\overrightarrow{wq}_{lsa}$ and each topic vector in the matrix D, and then selecting the most similar topic as the resultant topic for that document. To do this, the cosine distance is used for similarity measurement.

(d) These steps must be followed for all the documents in the evaluation set. The topic identification error is calculated as the percentage of miss classifications among these documents.

This classification scheme will be used in different parts along the experimental framework of this thesis. We will refer to it hereinafter as the *LSA classification procedure*. In later stages we will execute these same steps, considering the variations of the schemes we are evaluating.

The LSA model has several advantages when compared to the Vector Space Model. On the one hand, it offers a more compact representation of the document vectors in both training and evaluation datasets reducing sparseness. Note that in the Vector Space Model each document is represented in a m dimensional space, where m is the number of index-terms in the term inventory (in this case the term inventory 1 has 16250 terms, so $m = 16250$). In contrast, in the LSA model the number of dimensions in which each document is represented is constrained to the dimensions of matrix D. Since in the procedure we describe above, we are not truncating the latent semantic space, the resultant dimensions of matrix D are $n \times n$, which means that each document is represented in a n dimensional space. The value of n, which in this case corresponds to the number of topics in the document collection, is 67.

Figure 3.10 presents the results in terms of the topic identification error comparing both models for document representation (VSM and LSA). Recall that in these experiments we are analyzing the transcriptions provided by the first stage of the ASR, thus we are identifying the topic on transcriptions that may contain recognition errors, which increases the complexity of the task. In general, we can see in the figure that the topic identification error is much lower for the *Evaluation Set 1* when compared to the error obtained for the *Evaluation Set 2*. Recall that the audio segments in the *Evaluation Set 1* are larger than in *Evaluation Set 2* (which means that also the text transcriptions are larger); therefore they contain a higher number of index-terms and semantic elements that enhances the topic identification process. In *Evaluation Set 2*, as we said, the audio segments are shorter, therefore their transcriptions contain fewer index-terms. This results in a much more sparse term document matrix, which in turn hinders the topic identification process for this evaluation dataset.

Analyzing each evaluation dataset separately, we can appreciate that there is a slight reduction in topic identification error for *Evaluation Set 1* when comparing the baseline system to the LSA approach; however it must be noticed that this reduction is not

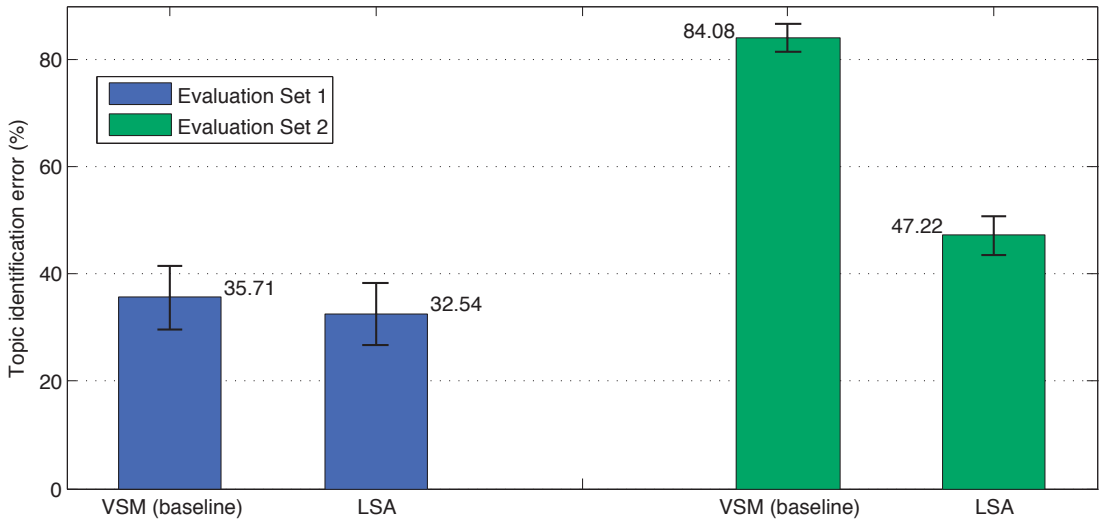


Figure 3.10: Topic Identification error for different document representation models

statistically significant. In contrast, for *Evaluation Set 2*, significant differences can be obtained due to the application of the LSA approach.

The use of the LSA model for document representation not only reduces the dimensional space but also reduces the effect of the high sparseness of the term document matrices, specially in *Evaluation Set 2*, for which the sparseness is higher. Results have shown that a reduction in the topic identification error can be obtained with this model.

It is worth mentioning that these results are obtained by identifying the topic on the transcriptions produced by the first decoding stage. Recall that these transcriptions contain recognition errors, and that the topic identification is performed on text to which some index-terms, that appear in the original transcription of the audio segment, may be missing (due to deletions, insertions or substitutions generated by the recognizer).

With the aim of comparing these results with the topic identification performance on the original transcriptions, i.e. the literal transcriptions of the audio segments (without recognition errors), we conducted an additional experiment. The objective of this experiment is mainly illustrative, since we are not expecting to use this topic identification result in the adaptation of language models. As we said, our aim is to compare the performance of the system analyzing text with and without recognition errors. In this experiment we use the *Term Inventory 1*, *tf-idf* as weighting scheme and the *LSA classification procedure*. The topic identification error for the transcriptions of the ASR is the same that we presented in Figure 3.10. The comparative results are shown in Table 3.6.

From Table 3.6, we can see that there are not significant differences between both experiments. It is worth mentioning at this point that the system has a Word Error Rate in the first ASR stage of 21.75% (this result will be described in more detail later in Section 5.3.2). Despite of this error, the topic identification performance is practically the same for both types of transcriptions. Usually, a substantial proportion of speech

	Topic identification error	
Transcription	Eval. set 1	Eval. set 2
ASR output	32.54 ± 5.78	47.22 ± 3.56
Reference transcription	30.95 ± 5.71	45.49 ± 3.55

Table 3.6: Comparison of the topic identification performance considering the ASR output and the reference transcription

recognition errors come from function words rather than content words, because of their inclination to be shorter, not well articulated and acoustically confusable [Bellegarda, 2000]. In this sense, the recognition errors do not affect considerably the topic identification models we are evaluating, (specifically in this case the LSA model), since these models do not take into account the function words into the topic identification process. Actually, most of the function words are removed in the stopword removal preprocessing stage. We believe that because of this there are not statistically significant differences in the performance of the system for both types of transcriptions.

3.3.6 Additional experiments comparing VSM and LSA

Besides the experiments we have presented, we have also conducted different experimental approaches with the aim of comparing the VSM and LSA models. Below, we describe these experiments and present the results that we have obtained.

- In the first place, for both models, we have built the TDM without computing a centroid vector for each topic. This means that we did not follow a centroid based approach. Instead, we built the TDM with each of the 1802 documents and from that point we proceeded according to the steps described in the baseline and the *LSA classification* procedures. By doing this, our aim is to evaluate whether the complete TDM offers a better representation for the document collection. We have found that considering the single documents to build the TDM, rather than the topic centroid vectors, increases the sparseness of the matrix hindering the topic identification process. Experiments with both VSM and LSA models showed that the topic identification error is increased with this approach for both evaluation datasets.

- We also performed experiments by truncating the latent semantic space in a different number of dimensions. In this regard we have followed two approaches:

  i) In the first of them, we consider the TDM formed with the topic centroid vectors. After applying LSA, we start by considering an initial number of dimensions of 67, which corresponds to the number of topics. Then, we performed different experiments truncating the latent semantic space, that is, considering fewer dimensions to represent the documents. The result of these experiments can be seen in Figure 3.11; in which we present the

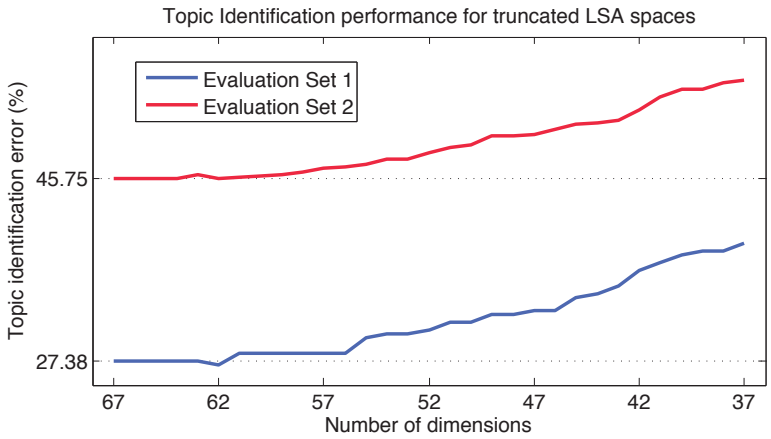


Figure 3.11: Topic Identification error considering a different number of dimensions in the LSA space

topic identification error for both evaluation datasets. The leftmost value in the figure corresponds to the topic identification error obtained without truncating the LSA space, that is, with 67 dimensions. It is shown that the error tends to increase as the number of dimensions are reduced. Nevertheless, we can appreciate that it can be achieved a reduction in the number of dimensions of the LSA space without a significant loss of performance. For instance, the topic identification error for *Evaluation Set 1*, considering 38 dimensions is 38.49 ± 6.01, which is not significantly higher than the error obtained with 67 dimensions (i.e. 27.38 ± 5.51).

ii) In the second approach, we consider the TDM built with each of the 1802 documents in the training dataset. After applying LSA to this TDM, we truncated the latent semantic space in a different number of dimensions, ranging from the initial dimensional space, i.e. 1802 dimensions, to 67 dimensions. Not only none of these experiments led to a reduction of the topic identification error but all of them provided an error significantly higher than the lowest error in the previous approach. We believe that the high sparseness in the TDM as a result of considering the single documents, does not allow an adequate representation of the documents or topics in the collection.

3.3.7 Considerations on the EPPS database

In the EPPS database (we described some details of this database in Section 3.3.1), an obstacle for the topic identification process lies in the high resemblance between some of the topics of the collection.

If we take a look to some of the topic labels that can be found in the database, we can see that there is a high closeness between many of the topics. For instance, the following topics: "*Election of the president*", "*Election of the vice-presidents*" and "*Election of the Quaestors*" are highly related to the same underlying subject.

Actually, documents belonging to these topics share a large number of index-terms and the similarity between them is higher than the similarity when they are compared to other topics. For instance, the similarity between the centroid vectors of the topics "*Election of the president*" and "*Election of the vice-presidents*" is 0.970.

In contrast, if we compare the similarity of the topic "*Election of the president*" with a different topic, such as "*Situation in Ukraine*" the similarity between them is 0.777.

Nevertheless, in the EPPS database there are also topics that can be clearly differentiated one from another. Examples of these topics are "*Situation in Colombia*" and "*Statement of the president*", whose similarity, measured between their topic centroid vectors is 0.5713.

Figure 3.12 shows the centroid vectors in the LSA space of the topics in the collection. In this figure we are considering only the two dimensions with the largest

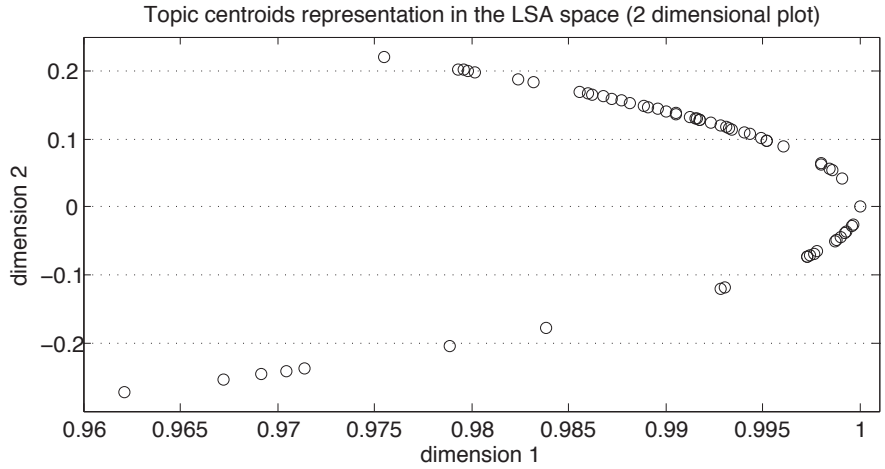


Figure 3.12: Representation of the topic centroid vectors in a 2-dimensional plot of the LSA space

contribution in the LSA space (i.e. the two dimensions with the largest singular values λ_i). We are aware that, in this figure, we are displaying an extremely reduced representation of the topic vectors, but it may help us to shed some light on the complexity of this task. We can see in this figure that some of the centroid vectors are highly overlapped and that the close distance between some of them could clearly hinder the discrimination between different topics.

To illustrate this better, it would be interesting to analyze some of the confusion matrices that are result of the topic identification process. Table 3.7 shows the confusion matrices for a representative group of topics. For this example, we have selected topics with distinct characteristics, ranging from those with a high accuracy rate in the topic identification process to those that are not identified by the classifier in this experiment.

The topic identification results that we show in this table have been obtained for the *Evaluation Set 1* applying the LSA model; actually these results are a small part of the global results presented in Figure 3.10 for all topics. For now, let us call these topics A, B, C and D, although later in this section we will describe them more in detail.

Topic A		Reference	
		YES	NO
Classified	YES	6	1
	NO	0	245

Topic B		Reference	
		YES	NO
Classified	YES	16	11
	NO	4	221

Topic C		Reference	
		YES	NO
Classified	YES	2	4
	NO	5	241

Topic D		Reference	
		YES	NO
Classified	YES	0	0
	NO	3	249

Table 3.7: Confusion matrices for some of the topics in the collection

We can see, for instance, that transcriptions belonging to Topic A are, in general, well classified, while elements from Topic D are not identified by the classifier. There are also elements, like those belonging to Topics B and C that have intermediate rates of accuracy. In Table 3.8 we present the Precision, Recall and F_1 measures for these selected topics. To understand the performance of the classifier for these topics it is

Topic	Precision	Recall	F_1 measure
A	0.857	1.0	0.923
B	0.592	0.8	0.681
C	0.333	0.285	0.307
D	0	0	0

Table 3.8: Precision, recall and F_1 values for the selected topics

convenient to analyze the labels and the content of the document belonging to these topics. Below, we present the topic labels for each of these topics.

- Topic A - *Dutch boat belonging to "Women on Waves" association.*
- Topic B - *Statement by the president designate of the Commission.*
- Topic C - *Towards an European constitution.*
- Topic D - *Programme of the Luxembourg presidency.*

The content of the Topic A is related to a pro-abortion association, called "Women on waves". In the collection, this topic is the only one related to this subject. Documents of this topic include unique index-terms that are not present in any other topics, such as "*aborto*" (abortion), "*reproductivo*" (reproductive) and many others. For this reason, audio segments in the evaluation datasets containing any of those terms could be classified in this topic; and for the same reason, documents that do not contain any of those unique index-terms are unlikely to be classified in this topic.

In contrast, Topics B and C are related to general subjects in the collection and do not cover any specific matter that can be undoubtedly identified as such.

Finally, and despite Topic D covers a specific matter within the political domain

of the collection, some of the elements of this topic are misclassified in the Topic "*Programme of the Netherlands presidency*", which is a topic with a high similarity.

The examples that we have just mentioned are intended to illustrate the complexity of the topic identification task for our document collection. In other application domains in which the document collections could encompass a broader variety of topics, such as sports, culture, political, financial, etc., there could be deeper semantic differences between the topics, which in turn would facilitate the topic identification task. However, in our case, as we have seen, in a fully political domain, the high similarity between documents hampers the identification process.

3.3.8 Experiments on index-terms selection

Stopword removal is an initial step in defining an adequate list of index-terms. This preprocessing procedure allows to remove the non-informative words, i.e. words that have little lexical meaning, that are too frequent among the documents in the collection and are unlikely to contribute to the distinctiveness of the topics. A stopword list typically includes words such as articles, prepositions, pronouns and conjunctions.

There are some benefits derived from the elimination of stopwords. There is a reduction of the size of the term inventory, which in turn contributes to speed up the topic identification process; and there is also a reduction of the sparseness, due to the fact that stopword removal may be seen as a filtering technique that removes noise from the document representation space. This noise can mislead the learning process by defining non-existent correlations between documents.

Nonetheless, stopword removal is only an initial step in the refinement of the definition of the term inventory. It must be accounted that generic stopwords lists are designed for general domains, therefore such lists may not cover all the non-informative words of a specific domain. Stopwords lists particularly designed for a specific informative domain may not perform well in a different one. For these reasons a more robust procedure is needed in order to detect and remove words that may have little lexical meaning in specific domains.

In this regard we propose a set of experiments in order to find an optimal term inventory for the specific domain we are analyzing. In these experiments we apply the term selection strategies described in Section 3.1.4. The procedure we followed in this experimental setup is commonly known as a *filtering* method [Silva and Ribeiro, 2010] and is described below:

1. Starting from an initial term inventory we compute a metric (e.g. the *idf* value) for each of the terms in the initial term inventory.

2. We sort all the terms in ascending order with respect to the value of the metric, meaning that the first term will be the term with the smallest value. For instance, if we are computing the *idf* metric, the term in the first position of the ordered list will be the term with the greatest document frequency, which is the term that appears in more documents, more than any other term.

3. We generate a new term inventory by removing the term in the first position of the ordered list.

4. We perform the *LSA classification procedure* described in Section 3.3.5 on the development dataset. The term inventory obtained in step 3 is used.

5. We repeat from step 3) removing, every iteration, an additional term (i.e. first two terms in the ordered list, then the first three, and so on). There are different stop criteria for this iterative procedure. For instance, it can be repeated until a predefined number of terms have been removed or until the system performance falls below a predefined value. We decided to repeat it until there were no terms left to represent a document in the TDM. We selected this criterion because one of our objectives in this experiment is to analyze the performance of the system when there is a drastic reduction of index-terms (these results will be discussed in the next Section).

6. We compare the topic identification results for all the iterations and select the optimal term inventory.

In our experiments we apply these steps for each of the index-terms selection strategies among those appearing in Section 3.1.4.

First, we define a **reference experiment**. The results that we obtain by applying the index-terms selection strategies will be compared against this reference experiment. The **reference experiment** is conducted using the *Term Inventory 1* on the development set. Recall that the *Term Inventory 1* is the set of index-terms in the training documents of the collection to which stopwords in a generic list have been previously removed. We also want to emphasize on the fact that the reference experiment is conducted on the development dataset. We took this decision mainly because this index-terms selection procedure is a parameter tuning procedure that must be adjusted outside the evaluation dataset.

Our aim in these experiments is to determine whether by applying different criteria for selecting the index-terms, a more adequate set of terms can be obtained. We also want to determine if this methodology improves the performance that is achieved by using a simple generic stopword list.

The best results for each technique, i.e. those with the term inventory that lead to the minimum error, are shown in Figure 3.13. These results were selected as the best result of each technique, for all the iterations.

In Figure 3.13, in the leftmost column, we can appreciate the results of the reference experiment. In the rightmost column we can appreciate the results obtained in an experiment combining all techniques. For this combined experiment, in each iteration, instead of removing only one term, we removed 5 terms, one for each metric. Our aim regarding the combined experiment is to combine all the index-terms selection strategies into a single experiment.

Among these strategies, only the *idf* selection strategy reduces the error of the reference experiment, although it has to be noticed that this reduction is not statistically

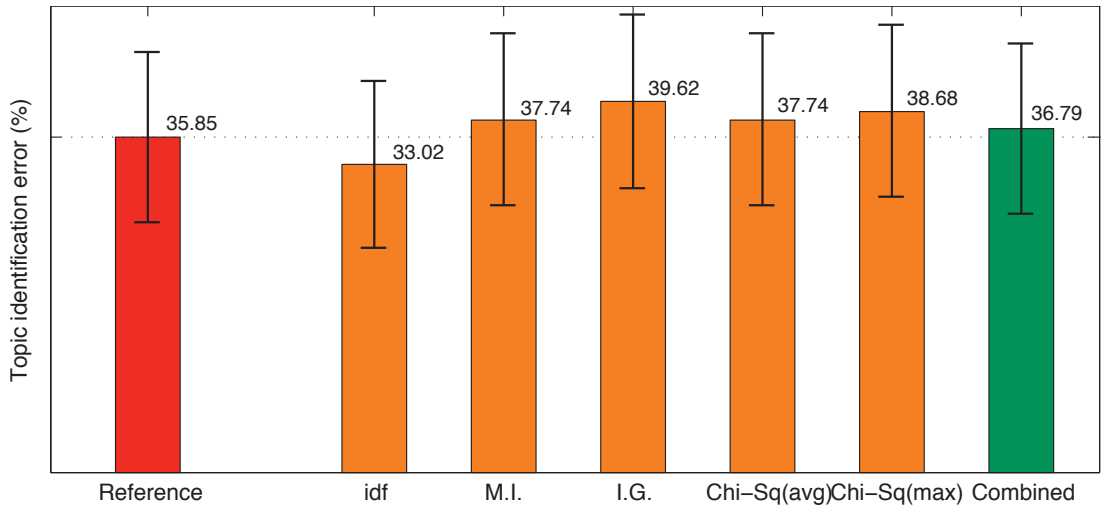


Figure 3.13: Minimum topic identification error obtained with different index-terms selection strategies. The compared metrics are: *idf - inverse document frequency*, M.I. - *Mutual Information*, I.G. - *Information Gain*, Chi-Sq - *Chi-Square* and a combination of all the techniques. These results are obtained on the development dataset.

significant. This result is achieved when discarding the first 24 terms of the ordered list obtained from the *idf* technique. These terms are shown in Table 3.9 and correspond to those terms that appear in most documents. For instance, the term "*señor*" appears in 1338 out of 1802 documents, being the term that appears in the largest number of documents, followed by terms "*gracias*" and "*presidente*", that appear in 1037 and 717 documents respectively.

Position	Term	N.D.App.	Position	Term	N.D.App.
1	SEÑOR	1338	13	EUROPEO	351
2	GRACIAS	1037	14	AHORA	341
3	PRESIDENTE	717	15	USTED	291
4	EUROPEA	530	16	CREO	289
5	MUCHAS	513	17	GRUPO	288
6	UNIÓN	498	18	PAÍSES	286
7	COMISIÓN	477	19	CONSEJO	285
8	SEÑORA	442	20	PARTE	274
9	PALABRA	431	21	MIEMBROS	266
10	PARLAMENTO	416	22	POLÍTICA	260
11	EUROPA	385	23	AÑOS	258
12	MINUTOS	359	24	HECHO	255

Table 3.9: List of Index-terms to be removed from the term inventory according to the *idf* index-terms selection technique. Table present the position of the term in the sorted listed and the number of documents it appears in (N.D.App.)

Table 3.10 shows the first terms to be removed according to each term selection technique. We can see from this table that these techniques present rather different term selection results.

For instance, by applying the Mutual Information metric the first terms to be removed are terms whose distribution is very similar in the class as it is in the collection as a whole. Information Gain metric has a similar criterion. Actually, by applying both metrics there can be found some similar terms. In general, both techniques select terms that are not the most common among documents, but neither are the more rare within the collection.

Mutual info.	N.D.App.	Info. gain	N.D.App.
INFORME	97	MINUTOS	359
CINCO	171	BARROSO	98
AYUDA	75	CONSTITUCIÓN	154
CUATRO	157	GRACIAS	1037
MIL	210	INFORME	97
SITUACIÓN	205	COMISIÓN	477
ADEMÁS	163	UCRANIA	49
SIETE	78	PALABRA	431
SEIS	114	VOTACIÓN	101
GRUPO	288	PRESIDENCIA	160
DIPUTADOS	148	CONSEJO	285
APOYO	104	PRODI	71
DEBATE	198	COMISARIO	145
CONSEJO	285	GRUPO	288
NOMBRE	217	FINANCIERAS	55
ELECCIONES	109	MIL	210
Chi-Sq (avg)	**N.D.App.**	**Chi-Sq (max)**	**N.D.App.**
HAREMOS	17	EJECUTIVO	6
FIRME	19	ROUCEK	4
HICIMOS	10	CAMBIE	9
EXPRESADO	22	DEJE	9
PELIGROS	11	HAREMOS	17
TEMOR	12	AHÍ	55
HECHA	6	ALREDEDOR	8
CONOCIDO	11	PRINCIPIO	89
VIVA	9	COMPETITIVOS	7
CONCRETA	17	PENA	25
CREEN	16	CEDER	7
MICRÓFONO	13	IGLESIA	5
ACABAN	5	MANO	38
ROUCEK	4	FUTURO	11
COMPROBAR	8	GUSTA	11
ALEMANIA	21	CONFIRMAN	3

Table 3.10: First terms to be discarded according to each term selection strategy and the number of documents they appear in (N.D.App.)

In contrast, Chi-square technique selects terms that are significant to a specific class but not too common in the whole collection, for example, rare terms. A rare term is a term that occurs few times in a large collection. That occurrence would be sta-

tistically significant for the class to which the term belongs, but at the same time a single occurrence would not be very informative from an information-theoretic perspective. Because the criterion of this technique is significance, Chi-Square therefore select more rare terms than Mutual Information. For instance, the term <ROUCEK> appearing in both variants of the Chi-Square criteria only appears in 4 documents.

Despite results of the index-terms selection strategies have not shown significant improvements compared to the **reference experiment**, they should not be neglected. Note that there has been a reduction of the error, although not significant considering the confidence intervals of the development dataset. In this dataset the confidence intervals are larger than for the evaluation sets, due to the reduced number of turns of intervention in this set.

We decided to consider the best result of these experiments and to evaluate them in the evaluation datasets. For doing this, we removed the terms listed in Table 3.9 from the Term inventory 1, and we generate a second term inventory (from now on we will refer to it as **Term inventory 2**).

We conducted experiments on the evaluation datasets by using the **Term inventory 2** and compared them with the previous results (obtained by using Term inventory 1). The results are shown below in Table 3.11. Results for Term inventory 1 are the same results we presented in Figure 3.10. We include them in this table for comparative purposes.

Model	Term Inventory	Eval. set 1	Eval. set 2
VSM	1	35.71 ± 5.91	84.08 ± 2.61
VSM	2	34.13 ± 5.85	62.86 ± 3.44
LSA	1	32.54 ± 5.78	47.22 ± 3.56
LSA	2	30.56 ± 5.68	46.95 ± 3.56

Table 3.11: Topic identification error for different term inventories

As shown in Table 3.11 the **Term inventory 2**, in general, reduces the topic identification error in the evaluation datasets. This reduction is statistically significant for the *Evaluation Set 2* when Vector Space Model is used for document representation. There is a slight reduction in error when using the LSA model, although this reduction is not statistically significant.

3.3.9 Impact of term inventory reduction on topic identification

In the previous Section (Section 3.3.8) we performed a procedure aiming to find an optimal term inventory for the specific domain we are analyzing. By means of this procedure we evaluated the topic identification system for different term inventories. These term inventories were obtained by applying various index-terms selection strategies. Figure 3.13 showed the best results among all possible iterations for each technique on the development dataset. From these experiments, some questions arise: how does the system perform for all possible term inventories on the evaluation datasets?,

and what is the impact of the term inventory reduction in the topic identification performance?

To solve these questions, in this Section we evaluate the topic identification system performance on the evaluation datasets, for each of the term inventories obtained in the iterative procedure described in the previous section. Recall that in this procedure, for each iteration, an index-term (according to the ordered list) is discarded from the term inventory. This means that the term inventory in each iteration is a reduced version of the inventory obtained in the previous iteration.

Our goal in these experiments is to find the maximum number of index-terms that can be discarded without the system to have a significant increase of the error. First, we have to define what a significant increase of the error is. We are considering that a result is significantly higher than a reference result, when its lower bound is greater that the upper bound of the reference result. So, taking as reference values the results presented in Table 3.11, for the *Term inventory 1* and applying the LSA classification procedure we can compute then the minimum error to be consider significantly higher for both evaluation datasets, as follows.

For *Evaluation set 1*, the upper bound of the reference result is given by $32.54\%+5.78$ which is 38.32%. To consider that a result is significantly higher, its lower bound has to be greater than 38.32%. The minimum error value that fulfils this condition is $44.84\% \pm 6.14$, for which the lower bound is 38.69%. So, an error equal or greater than 44.84% will be considered significantly higher than the reference value for the *evaluation set 1*.

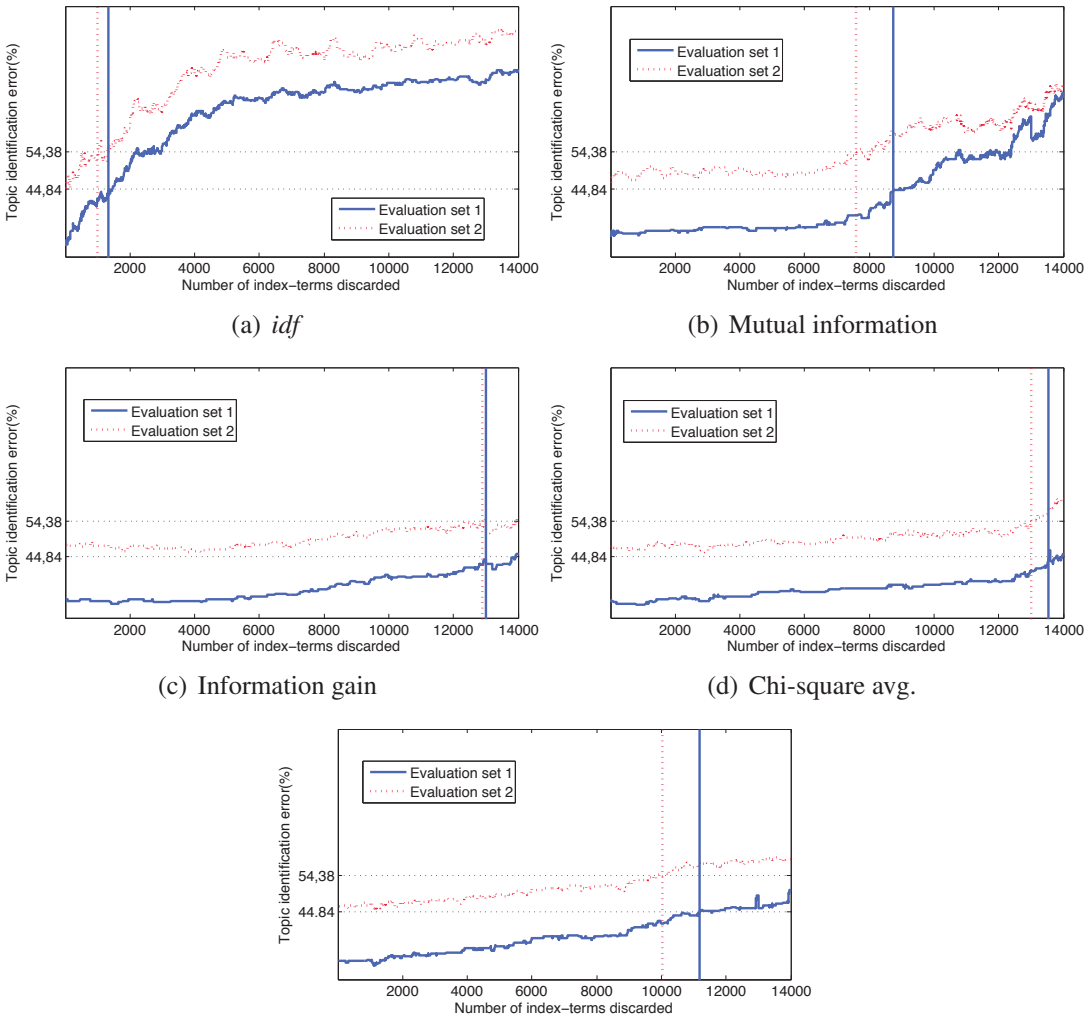
For *Evaluation set 2*, the procedure is the same. The upper bound of the reference result is given by $47.22\% + 3.56$ which is 50.78%. An error is considered significantly higher, if its lower bound is greater than 50.78%. In this evaluation dataset, the minimum error value that fulfils this condition is $54.38\% \pm 3.55$, for which the lower bound is 50.82%. An error equal or greater than 54.38% will be considered significantly higher than the reference value for the *evaluation set 2*.

Once we have defined these values, we have to look for the minimum term-inventory that lead to these results for each selection technique.

In Figure 3.14 we present the results of these experiments. In each sub-figure we present the topic identification error versus the number of index-terms discarded. The vertical lines in each figure (one for each evaluation dataset) mark the minimum term inventory for which the error is significantly higher.

From the results in Figure 3.14 we can draw some interesting conclusions. First of all, all techniques allow a considerable reduction of the term inventory with a limited loss in effectiveness. By means of the *idf* technique, the topic identification error becomes significantly higher when 1338 index-terms are removed from the initial term inventory for the *Evaluation set 1* and 988 index-terms for the *Evaluation set 2*. According to these values, this technique allows a reduction of only 8.2% and 6.1% of the initial size of the term inventory for evaluation sets 1 and 2 respectively.

In contrast, by applying the *Chi-square average* technique, we can remove up to 13607 index-terms for the *Evaluation set 1* and 13312 for the *Evaluation set 2* without



(a) *idf*

(b) Mutual information

(c) Information gain

(d) Chi-square avg.

(e) Chi-square max.

Figure 3.14: Topic identification system performance by applying distinct term reduction techniques

a significant loss of performance. This implies a reduction of 83.73% and 81.92% of the initial term inventory for evaluation sets 1 and 2 respectively.

In Table 3.12 we present the summary of the term reduction results for all techniques.

	Eval. set 1		*Eval. set 2*	
Technique	Num. terms disc.	(%)	Num. terms disc.	(%)
idf	1334	8.20	988	6.08
Mutual info.	8726	53.69	7585	46.67
Info. gain	13100	80.61	13008	80.04
Chi-sq (avg)	13607	83.73	13312	81.92
Chi-sq (max)	11180	68.80	10231	62.96

Table 3.12: Summary of the results for the term reduction for all index-terms selection techniques in both evaluation datasets. Table includes: *a)* the number of index-terms that can be discarded in each technique (Num. terms disc.) without a significant loss of performance, and *b)* The percentage that this reduction represents in the initial term inventory.

In general, the term inventories for which a loss of performance is achieved are larger for the *Evaluation set 2*. Techniques for index-terms selection based on information-theoric metrics allow a further reduction in the size of the inventory compared with techniques such as *idf*.

3.3.10 Comparison on different weighting schemes

One of our goals in this work is to compare and evaluate different weighting schemes and their performance on the topic identification system. We have performed experiments by combining several local and global weighting schemes. Although we performed a large number of experiments, in Table 3.13 we present only the best results. These results were obtained applying the *LSA classification procedure* and using the *Term Inventory 2* that was previously obtained in Section 3.3.8.

Among the local schemes, *term-frequency* outperforms both *log-frequency* and *augmented and normalized term frequency*. The error obtained for *term-frequency* is significantly lower for *Evaluation set 1*. For *Evaluation set 2*, it is achieved a smaller error, although there are not significant differences when comparing the three local schemes.

These results show that the document length normalization contributes to the enhancement of the system performance. Both schemes that normalize the length of the document (*term frequency* and *augmented and normalized term frequency*) reduce the error when compared to *log-frequency*, although there are only significant improvements for *term-frequency*.

Regarding the global schemes, there are no significant differences between them, although in general, *entropy* is the scheme that provides the lowest error for both evaluation datasets.

Considering the combination of local and global weighting schemes, the best com-

Weighting scheme		Topic identification error	
Local scheme	Global scheme	Eval. set 1	Eval. set 2
log-freq	idf	43.65 ± 6.12	53.58 ± 3.55
log-freq	gfidf	44.84 ± 6.14	52.65 ± 3.56
log-freq	entropy	42.85 ± 6.11	51.98 ± 3.56
aug.norm.tf	idf	43.25 ± 6.11	52.52 ± 3.56
aug.norm.tf	gfidf	44.04 ± 6.12	53.84 ± 3.55
aug.norm.tf	entropy	39.68 ± 6.04	51.85 ± 3.56
tf	idf	30.56 ± 5.68	46.95 ± 3.56
tf	gfidf	34.13 ± 5.85	49.20 ± 3.56
tf	entropy	30.16 ± 5.66	46.29 ± 3.55

Table 3.13: Comparison on different weighting schemes. The local schemes are: *log-frequency* (log-freq), *augmented and normalized term-frequency* (aug.norm.tf) and *term-frequency* (tf). The global schemes are: *inverse document frequency* (idf), *global frequency inverse document frequency* (gfidf) and *entropy*

bination is the one formed by the *term-frequency* and the *entropy* as local and global schemes respectively. This combination shows the lowest topic identification error for both evaluation datasets, although this result is not significantly different to the results obtained by other combinations of *term-frequency* and other global weighting schemes.

3.3.11 Performance of the proposed *ad-hoc* weighting schemes

In order to evaluate the proposed global weighting scheme - *pseudo-entropy* (described in Section 3.2.1), we first performed different experiments on the *development set*. The objective of these experiments is to tune the parameters β and γ with which the best performance of the system for this dataset is obtained. The best results for the development set were obtained by adjusting $\beta = 1.5$ and $\gamma = 2.1$.

Once these parameters were adjusted, we conducted experiments on the evaluation datasets by combining the proposed global weighting scheme with different local weighting schemes.In these experiments we applied the *LSA classification procedure* and we used the index-terms in the *Term Inventory 2* to represent the documents. The results of these experiments are shown in Table 3.14.

Local scheme	Global scheme	Evaluation set 1	Evaluation set 2
log-freq	pseudo-entropy	37.30 ± 5.97	51.06 ± 3.56
aug.norm.tf	pseudo-entropy	35.71 ± 5.91	50.79 ± 3.56
tf	pseudo-entropy	27.38 ± 5.50	45.75 ± 3.55

Table 3.14: Topic identification error applying the *ad-hoc* pseudo-entropy scheme

Among these combinations of weighting schemes, the combination of *term-frequency* as local scheme and *pseudo-entropy* as global scheme provides the lowest topic identification error, although there are not significant differences between them.

When compared to the combination of *term-frequency* and *entropy* (results presented in Table 3.13), there is a reduction of the topic identification error for both evaluation datasets, although this reduction is not statistically significant.

There is indeed a significant reduction of the error for the *Evaluation set 2* when compared to the baseline system (results presented in Figure 3.10). Our proposed weighting scheme allows a relative reduction of topic identification error of 23.33% for *Evaluation set 1* (although not significant) and 45.59% for *Evaluation set 2* when compared to the baseline approach.

3.3.12 Impact of stemming in the topic identification

We have performed experiments by stemming the term-inventory. Stemming is done with the objective of removing prefixes, suffixes, plurals and morphological derivations of the words. It compresses the size of the indexing structure by reducing the number of distinct terms to index. For this experiment, we have used the Freeling Toolkit [Padró and Stanilovsky, 2012]. By stemming, the initial term inventory was reduced to 8680 index-terms. 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. Table 3.15 presents a comparison between the best results obtained so far for the identification task and the results obtained by stemming the term inventory and applying the *LSA classification procedure*.

	Evaluation set 1	Evaluation set 2
Without stemming	27.38 ± 5.50	45.75 ± 3.55
Stemming	33.73 ± 5.83	47.61 ± 3.56

Table 3.15: Comparison between stemming vs. no-stemming

By stemming, the topic identification error was increased in both evaluation datasets, although this increase is not statistically significant. We conducted different experiments by applying the stemming procedure and different weighting schemes and also applying stemming along with the generalized Vector Space Model. None of these experiments lead to a significant reduction of the error.

3.3.13 Summary of results on Topic Identification

Table 3.16 shows a summary of the different experiments conducted in the topic identification task. In this table we want to compare the results obtained by using different models for document representation, different weighting schemes (including *pseudo-entropy*, which is the scheme we propose) and different term inventories (either obtained by means of the index-terms selection techniques or by means of the stemming process).

We present the results of the topic identification process using different models for document representation as well as different term inventories.

Topic identification approach	T.I.E. for Eval. Set 1	T.I.E. for Eval. Set 2
VSM + tf-idf + Term_inv 1 - *Baseline*	35.71 ± 5.91	84.08 ± 2.61
VSM + tf-idf + Term_inv 2	34.13 ± 5.85	62.86 ± 3.44
VSM + tf-entropy + Term_inv 2	34.52 ± 5.87	55.97 ± 3.54
VSM + tf-pseudo entropy + Term_inv 2	33.33 ± 5.82	54.24 ± 3.55
VSM + tf-pseudo entropy + Stemming	36.11 ± 5.93	57.56 ± 3.52
LSA + tf-idf + Term_inv 1	32.54 ± 5.78	47.22 ± 3.56
LSA + tf-idf + Term_inv 2	30.56 ± 5.68	46.95 ± 3.56
LSA + tf-entropy + Term_inv 2	30.16 ± 5.66	46.29 ± 3.55
LSA + tf-pseudo entropy + Term_inv 2	**27.38 ± 5.50**	**45.75 ± 3.55**
LSA + tf-pseudo entropy + Stemming	33.73 ± 5.83	47.61 ± 3.56

Table 3.16: Summarized results

Throughout Section 3.3 we have described the experiments that have been conducted in the development of a topic identification system. It has been shown that the different strategies that we followed concerning the preprocessing of documents and the index-terms selection lead to a considerable reduction in the classification error, and that this reduction has been significant for one of the evaluation datasets (*Evaluation set 2*).

Recall that we are identifying the topic on the transcriptions provided by the first ASR stage of the architecture; this means that these transcriptions contain recognition error, which in turn increases the difficulty of this task.

We are not deeply analysing the impact of the audio length segmentation on the topic identification effectiveness. However, we can separate the analysis of the results regarding each of the configurations of the evaluation dataset.

It is important to notice that the *Evaluation Set 1* contains less samples than *Evaluation Set 2*, and therefore larger confidence intervals are obtained when analyzing the results for *Evaluation Set 1*.

Despite the fact that there is a slight reduction in topic identification error for *Evaluation 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 *Evaluation Set 2* significant results are obtained when comparing not only both document representation models (VSM and LSA), but also when comparing the VSM itself with the different term inventories. By including the use of the *term inventory 2* to the baseline VSM approach, a relative error reduction of 25.23% can be achieved.

The *tf-pseudo entropy* weighting scheme shows the minimum error for the VSM approach in the *Evaluation Set 2*, and this result improves the performance when compared with the *tf-idf* weighting scheme. Thus, the *ad-hoc* proposed weighting scheme does provide a significant improvement of the topic identification accuracy when used with the VSM.

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.

Compared with the VSM, all the variants of the LSA approach improves the topic identification error for *Evaluation Set 2*. Nevertheless, among them, no significant reduction can be obtained in the different configurations of the LSA approach. We may need more data to reach significant conclusions on the differences among LSA experiments. In this approach, neither the use of the *Term inventory 2*, nor the stemming nor the *tf - pseudo-entropy* weighting scheme show a significant reduction when compared with the use of the *Term inventory 1* and the *tf-idf* weighting scheme.

In general, when comparing the topic identification error obtained for both evaluation datasets, the minimum error was obtained when evaluating *Evaluation Set 1*. Since for larger audio segments, larger transcriptions are obtained, this result suggests that for *Evaluation set 1*, more semantic information is provided to the system.

When compared to the baseline, the best combination of parameters is obtained for the LSA model, using the *Term inventory 2* and weighting the terms with *tf - pseudo-entropy* scheme. This configuration presents a relative improvement, although not significant, of 23.32% when compared to the baseline approach for the *Evaluation Set 1* and a relative and significant improvement of 45.58% for the *Evaluation Set 2*.

4 Thesis work on Automatic Document Clustering

Within our topic-motivated contextualization framework, our aim is to explore different approaches for generating *topic-based* language models. One of these approaches focuses in generating these LMs based on an automatic document clustering of the training documents in the collection.

In this chapter we present the methodology we follow in this regard. In Section 4.1 we present the document clustering techniques that we use to perform this clustering approach. Particularly, we present the k-means (4.1.1) and the Latent Dirichlet Allocation (LDA) (4.1.2) techniques for document clustering. We also describe the Silhouette Coefficient (4.1.3), which is the criterion we chose to find an adequate number of clusters.

In Section 4.2 we present our contributions regarding the application of automatic document clustering techniques to the language model adaptation process. Our contributions focus on the reduction of the model complexity and on the comparison between supervised and unsupervised techniques for the generation of topic-based language models.

Finally, in Section 4.3 we describe the experiments on document clustering along with the results obtained.

4.1 Foreground on Document Clustering

As we mentioned in Chapter 2, one of the aims of this work is to evaluate two different strategies in the generation of topic-based language models (recall that within the contextualization framework, these models will be merged into a context-dependent model which in turn will be interpolated with the background LM - see Figure 2.1).

In the first strategy we make use of the original topic labels of the documents of the collection to generate a specific topic-based language model for each of these topics. Thus, in this strategy we take advantage of the manually assigned labels of the collection and we generate as many topic-based LMs as topics in the training database.

In the second strategy we group the data in the training dataset into automatic topic clusters based on the semantic similarity between the documents. Therefore in this

strategy we ignore the topic labels, i.e. we assume the training documents are unlabelled, and from this point we perform an automatic clustering of documents. In this section we introduce the document clustering techniques that we have used in the development of this strategy.

Basically, document clustering consists of the assignment of similar documents to a priori unknown groups or clusters. In general, similar documents may be related not only by their topic, but also by other characteristics such as language, genre, authorship, etc., however, in this work, our goal is to find clusters among the documents in the collection that are related by their topic.

Despite the criterion we use to cluster the documents is based on their semantic similarity, the criterion we will use to evaluate the clustering technique is based on the speech recognition performance of the system.

From Section 3.1.1 we know that a document can be represented as a vector, using the Vector Space Model, in a m-dimensional space formed by the index-terms. Initially we could think that the clustering could be done in this vector space, however, this representation does not typically work well for clustering documents. The reason is that text data has a couple of properties that we must consider:

- The dimensionality of the documents is usually very large and the underlying data is sparse. In other words, the size of the term inventory may be of tenths of thousands, but a given document may contain only a few hundreds of index-terms. This problem requires even more attention when documents are short sentences since the sparsity becomes higher.

- Index-terms are typically related between them by means of semantic relationships. We can consider these semantic relationships as the concepts that underlie behind documents. Usually, the number of concepts in a document is much smaller than the number of the terms that appear in it. Therefore algorithms must take into account these term relationships in the clustering process.

For these reasons, it is often preferred a reduced space representation such as the LSA space. Recall that in this space a lot of dimensions in the data, which can be noisy for similarity based applications such as clustering, may be removed. This removal also helps in magnifying the semantic relationships in the underlying data.

The clustering methods can be categorized in two classes: *hierarchical* and *iterative*. In this Thesis we employed an iterative algorithm, the k-means clustering due to its simplicity and its proven efficiency in text classification tasks [Sebastiani, 2002]. Nonetheless, we briefly describe the hierarchical approach, then we examine the k-means clustering approach.

Hierarchical clustering (also known as agglomerative clustering) proceeds either bottom-up, by starting with the individual documents and grouping the most similar ones, or top-down, by starting with all the collection and dividing it into groups so as to maximize a given objective function. Since the underlying idea in both approaches is practically the same, we briefly describe the bottom-up approach.

In this clustering technique, the process of agglomerating documents into successively higher levels of clusters creates a cluster hierarchy (or *dendogram*) for which the leaf nodes correspond to individual documents, and the internal nodes correspond to the merged groups of clusters.

In this sense, when two groups are merged, a new node is created in this tree corresponding to this larger merged group. The two children of this node correspond to the two groups of documents which have been merged to it. A fundamental property of the hierarchical clustering is that whenever two documents are in the same cluster at some level, they remain together at all higher levels.

This technique has been widely used in the adaptation of language models [Chen et al., 2001b, Florian and Yarowsky, 1999, Iyer and Ostendorf, 1999]; however, and spite of its good performance, its main drawback is the high computational effort required when compared to the iterative approach.

4.1.1 *k*-means clustering

This is one of the most used techniques in document clustering [Jain et al., 1999], and along with the hierarchical clustering technique is one of the most common approaches applied to language model adaptation[Bellegarda, 2000, Clarkson, 1999, Wu, 2002]. The objective in this technique is to partition the documents in the collection into k clusters in which each document belongs to the cluster with the closest centroid. The number k of clusters, must be provided as an input, and it is in fact a parameter that must be carefully selected.

The algorithm for computing k-means clustering is iterative and is composed of two main steps: an *assignment* step, in which each document is assigned to the cluster with the closest centroid, and an *update* step, in which the centroids are adjusted to consider the documents newly assigned to the clusters. This process is repeated until no centroids changes. In Jain et al. [1999] there is a complete description of the algorithm along with many variants that have been proposed for it.

This technique assumes that each document is represented as a vector in a multi dimensional space. In this regard, there are different criteria for selecting such space of representation. For instance in Wu [2002], Kim and Khudanpur [2004] and Clarkson [1999] the generalized Vector Space Model is used, that means that the clustering is performed over a m-dimensional space, being m the size of the term inventory.

A major problem of this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function if the initial partition is not properly chosen. In different runs, the final result of the algorithm may not be the same, since it depends on the initial selection of the centroids.

The choice of k, the number of clusters, is a critical step in the algorithm. Different methods have been proposed to find the optimal number of clusters. In this work we use the Silhouette Coefficient, which will be described in Section 4.1.3.

4.1.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic topic modeling algorithm that has become commonly used in various Natural Language Processing and Information Retrieval applications in recent years. It was originally proposed by Blei et al. [2003] with the aim of providing a probabilistic framework that allows to infer the latent structure behind a collection of documents.

The main purpose of LDA is to find topics in a document collection and assign distributions of these topics over each document (as well as distributions of words over topics). This technique is based on two principal assumptions: that each document contains a mixture of different topics, and that each topic contains a mixture of different words (although in our case instead of words we should refer to them as index-terms since we are not considering all the words in the documents, only a subset of them, that is the term inventory of index-terms).

LDA is an unsupervised learning process, and as such it can be understood as an automatic clustering technique, in which each of the topics to be discovered can be seen as a topic cluster that groups a number of documents. LDA is based on the *bag-of-words* model, in which the word order in a document is ignored and only the number of times a word appears in a document is considered. So, as in the Vector Space Model, the starting point of the LDA algorithm is to consider the Term-Document Matrix, which is composed of the raw frequency of terms in the documents of the collection.

Next, we introduce the underlying idea behind the LDA algorithm. In this description T_k is each of the topics to discover and $|K|$ the number of topics to discover. LDA learns how topics and documents are represented in the following form:

1. First, the number of topics to discover $|K|$ must be given as an input.

2. Once the number of topics is selected, LDA will go through each document d_j in the training dataset, and it will randomly assign each index-term t_i in the document to one of the $|K|$ topics, as shown in Figure 4.1. This step gives both an initial topic representation of all the documents, and an initial index-terms distribution of all the topics.

3. It must be noticed that this first assignment of index-terms to topics, was done in a random form, so it is evident that this representation is not accurate and should be improved. To improve this assignment, LDA computes the following probabilities:

   - For each document d_j, the document distribution over the topics $p(T_k|d_j)$ is computed as the proportion of index-terms in the document that are assigned to topic T_k.
   - For each index-term t_i, the topic distribution over the terms $p(t_i|T_k)$ is computed as the proportion of the index-term t_i in all documents that is currently assigned to topic T_k.

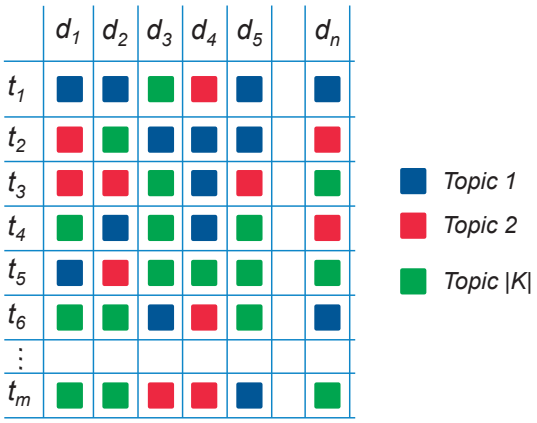


Figure 4.1: Initial random assignment of index-terms to topics

4. For each document it is evaluated whether an index-term must be reassigned to a new topic or not. In this step, the index-term t_i in document d_j is reassigned from topic A to topic B if and only if the following condition is fulfilled:

$$p(t_i|T_A, d_j) < p(t_i|T_B, d_j) \tag{4.1}$$
$$p(t_i|T_A) \times p(T_A|d_j) < p(t_i|T_B) \times p(T_B|d_j) \tag{4.2}$$

5. Steps 3 and 4 must be repeated until it eventually reaches a steady state in which the assignments do not change significantly between iterations.

The assignments obtained by this algorithm can be used for instance to estimate the topic mixtures of each document (by counting the ratio of index-terms assigned to each topic within that document) or the index-terms associated to each topic (by counting the ratio of index-terms assigned to each topic).

To classify a new document d_{NEW}, the first step is to assign topics to each index-term t_i in the new document according to the distribution $p(t_i|T_k)$, as shown in Figure 4.2. Then, we evaluated whether an index-term of d_{NEW} must be reassigned to a new topic, exactly as in Eq. (4.1). These assignments can then be used to estimate the topic mixture of the new document.

Above, we have presented the basic principle about how LDA works. In this description of the LDA algorithm we are not fully covering the details on how to compute the posterior probability that is needed in order to reassign index-terms to topics. The main difficulty in this algorithm is that the posterior cannot be computed directly [Blei et al., 2003], thus it has to be approximated. A wide variety of approximate inference algorithms can be considered. These algorithms are either based on sampling approaches (typically based on Markov Chain Monte Carlo (MCMC) methods [Hoffman et al., 2010]) or optimization approaches [Asuncion et al., 2009].

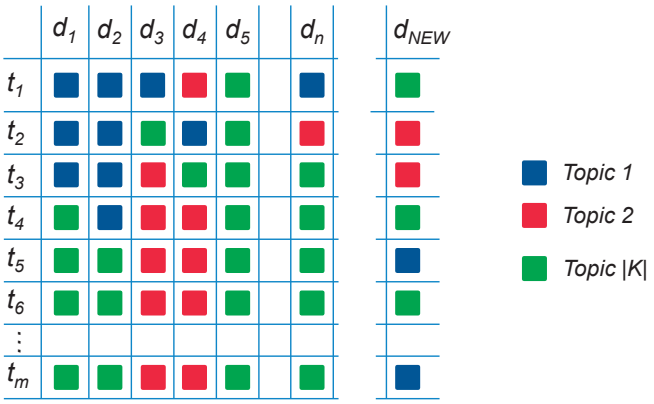


Figure 4.2: Assignment of topics for a new document d_{NEW}

4.1.3 Finding the optimal number of clusters

To determine the optimal number of clusters we use the Silhouette Coefficient (SC) proposed by Rousseeuw [1987]. 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 et al., 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 $dist(\vec{d}_i, \vec{d}_j) = 1 - \cos(\vec{d}_i, \vec{d}_j)$ indicate the distance between document vectors $\vec{d}_i$ and $\vec{d}_j$. The Silhouette Coefficient $sc(\vec{d}_i)$ for document d_i is computed as follows:

$$sc(\vec{d}_i \in c_k) = \frac{b(\vec{d}_i) - w(\vec{d}_i)}{\max(b(\vec{d}_i), w(\vec{d}_i))} \tag{4.3}$$

where $w(\vec{d}_i)$, the *within distance*, computes the average distance of the document vector $\vec{d}_i$ with all the document vectors in its own cluster, by using the following formula:

$$w(\vec{d}_i \in c_k) = \frac{1}{|c_k| - 1} \sum_{\substack{\forall d_j \in c_k \\ d_j \neq d_i}} dist(\vec{d}_i, \vec{d}_j) \tag{4.4}$$

And $b(\vec{d}_i)$, the *between distance*, is used to calculate the average distance of $\vec{d}_i$ with the document vectors of the other clusters. The minimum of all these average values is considered as $b(\vec{d}_i)$, as shown in the following formula

$$b(\vec{d}_i \in c_k) = \min_{j \neq k} \left[\frac{1}{|c_j|} \sum_{\forall d_m \in c_j} dist(\vec{d}_i, \vec{d}_m) \right] \tag{4.5}$$

The SC can have values from -1 to $+1$. Thus, if a document has SC value near $+1$, it implies that the *within distance* $w(\vec{d}_i)$ is much smaller than the smallest *between*

distance $b(\vec{d}_i)$. In that case, it is possible to say that there appears to be little doubt that document 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.

4.2 Contributions on Document Clustering

Our contributions in this Thesis regarding the application of automatic document clustering techniques are mainly focused on the comparison between different approaches for the generation of topic based language models and on the reduction of the number of parameters of the system's model. In this sense we can summarize our contributions as follows:

- We compare the performance of the speech recognition system under different strategies for the generation of topic-based language models. An initial strategy considers the original topic labels of the documents. In this strategy each of the topic-based LMs is generated from the documents that belong to each of these labels. Our aim is to evaluate whether the use of these labels in the generation of topic-based LMs is optimal in terms of recognition performance. In this sense we propose a generation of topic-based LMs by means of automatically clustering the documents of the training dataset. By doing this, the system generates topic-based language models that do not depend on the original topic labels.

  This strategy allows us to obtain a more uniform distribution of documents within the topic clusters. Besides, it also yields to an increase of the conceptual similarity between documents in the same cluster, which also would allow us to expect an improvement of the coverage of the topic-based language model within that cluster.

- The generation of topic-based LMs by means of the automatic document clustering strategies allows us to select a smaller number of topic-clusters compared to the supervised approach. In the supervised approach, there are as many topic-based LMs as there are topics in the collection; this means that there are 67 topic-based LMs (recall that in the document collection there are 67 original topic labels). In the automatic document clustering strategies that we propose, the number of topic-based LMs can be reduced.

  This implies a simplification of the system's model by reducing the total number of parameters involved in the system. As we will see in the next section, the number of topic-based language models can be significantly reduced by this strategy.

4.3 Experiments on Document Clustering

Within the topic-motivated contextualization framework we propose in this work, one of our objectives is to evaluate different approaches for the generation of topic-based language models. Until now, we have already seen the fundamentals of one of these approaches, in which we made use of the original topic labels of documents to perform an automatic and supervised topic identification.

Now, we explore a second approach, an unsupervised one, 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. By automatically clustering the documents, the association of a document to a topic cluster will not depend on the manually assigned labels. By doing this we expect to increase the conceptual similarity between documents in the same cluster. Also, we can expect to achieve an improvement of the coverage of the topic-based language models for each cluster.

In this regard, we have performed different clustering experiments on the training dataset.

4.3.1 Experimental framework

A first step in the application of the automatic document clustering techniques is to select a representation space for the documents that are to be clustered. For each of the techniques we use in this work for the clustering of documents, i.e. *k*-means and LDA, a different document representation space is used.

- For the application of the *k*-means algorithm, the documents can be initially represented either in the generalized vector space or in the latent semantic space. Due to the fact that the latent semantic space has fewer dimensions, it allows a more compact representation of the original document space. Besides, in contrast to the Vector Space Model, the LSA space also takes into account the semantic relationships between terms and it facilitates the clustering process. For these reasons, we select the latent semantic space to represent the documents.

  To do this, we first generate a Term Document Matrix composed of all the documents in the training dataset. We decide to use the **Term Inventory 2** for a couple of reasons. First, it allows an initial document representation using fewer dimensions. Second, the results of topic identification have shown that this term inventory may contribute in a more accurate document representation.

  Once the TDM is built, we apply a weighting scheme to each document vector. The scheme that we use is the one formed by the combination of *term-frequency* as local weight and *pseudo-entropy* as global weight. Then, we apply LSA to this matrix in order to obtain the document vector representation in the latent space. And finally, this vector representation is used to perform the automatic clustering of documents.

- For the application of the LDA algorithm, the documents are represented in the

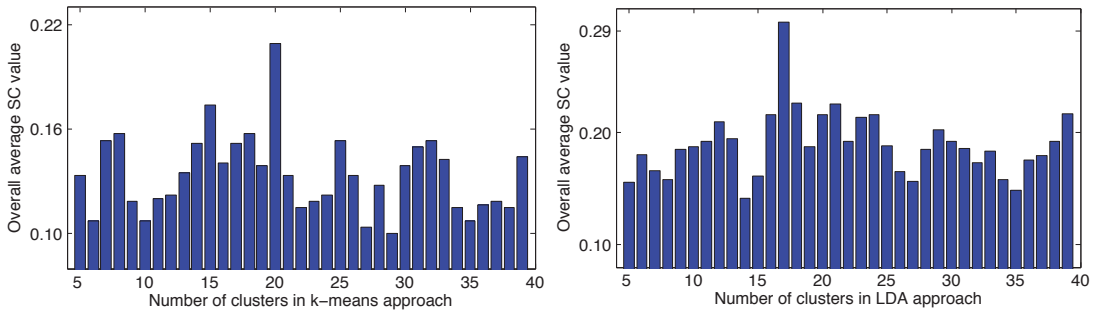
generalized vector space formed by the terms in the **Term Inventory 2**. We decided to use this term inventory for the same reasons mentioned in the previous item. The LDA algorithm is based on the raw frequency of terms in documents [Blei et al., 2003], therefore in this approach weighting schemes are not necessary. For the implementation of the LDA algorithm, we have used the Mallet toolkit [McCallum, 2002].

4.3.2 Experiments on finding the optimal number of clusters

These experiments are mainly focused in finding the optimal number of clusters (according to the criterion we selected).

It must be noticed that the results we will obtain in these experiments do not guarantee us that the resultant clusters will be optimal in terms of speech recognition. The outcome of these experiments only maximizes the *overall average Silhouette coefficient*, the real success of these experiments will not be known until the speech recognition experiments are done using this technique in generating topic-based language models. We are aware that if too many clusters are used, individual topic-based language models may be under-trained on sparse datasets, and hence each of the cluster language model will be poorly estimated. Conversely, a reduced number of clusters will result in a language model which may be unable to distinguish between topics.

We have conducted these experiments for both clustering approaches (k-means and LDA). Basically, we have clustered the documents in the training dataset under a different number of clusters and we have computed the *overall average Silhouette coefficient*. Figure 4.3 shows the overall average value $\bar{s}(k)$ for different numbers of clusters in both approaches.



(a) Overall average SC value for k-means (b) Overall average SC value for LDA

Figure 4.3: Overall average SC values for both clustering approaches

We have conducted experiments varying the number of clusters from 5 clusters to 65. In the figure we present the number of clusters for which we obtained the maximum $\bar{s}(k)$ value and the surrounding results. The largest value $\bar{s}(k)$ for the k-means approach was found for $k = 20$ clusters and for the LDA approach for $k = 17$ clusters.

Figure 4.4 shows the original distribution of documents in the collection and the distribution of documents after the clustering is performed.

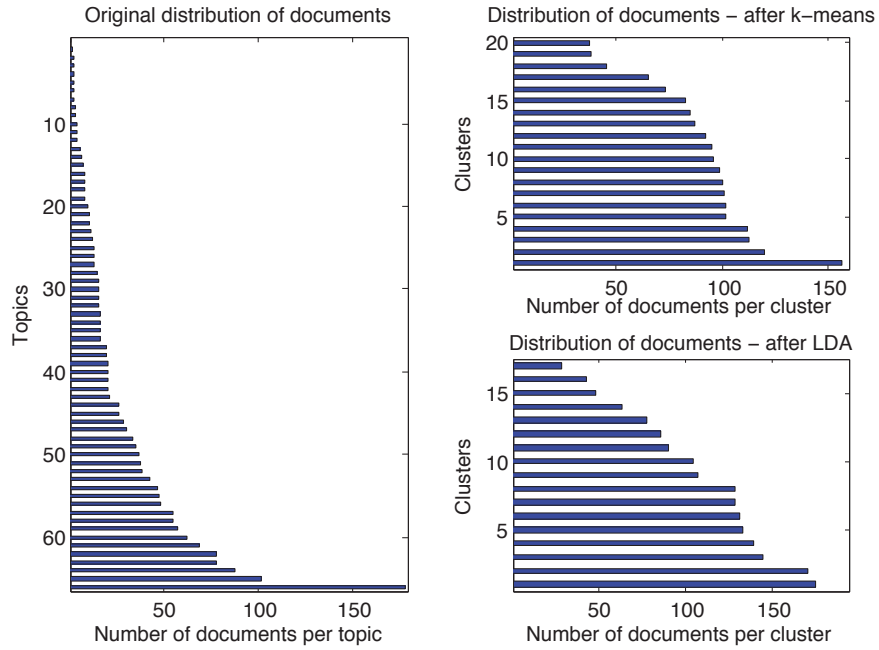


Figure 4.4: Distribution of documents before and after the application of clustering (comparative between k-means and LDA)

The distribution of documents in the clusters shows a more uniform distribution than the original, although there are still differences between the largest and the smallest clusters for both approaches. This difference can be better appreciated in Figure 4.5, in which we present the total length of the documents assigned to each topic compared to the total length of the documents assigned to each cluster for both clustering techniques.

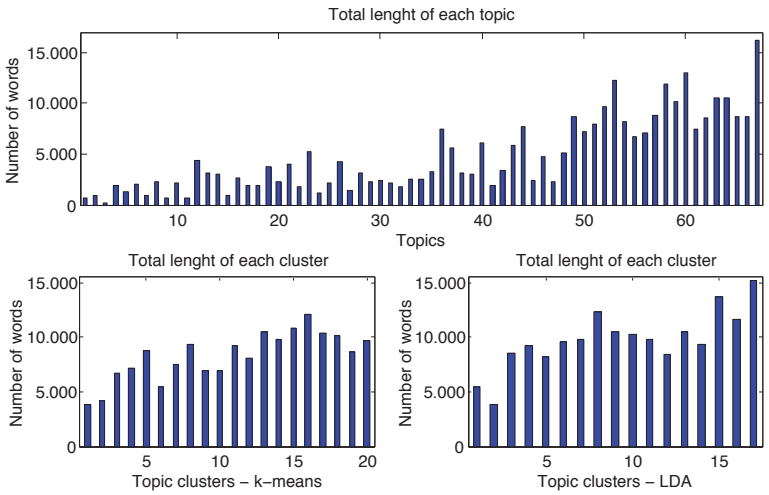


Figure 4.5: Total length of the documents assigned to each topic according to the original distribution of topics (figure on top); and to each topic clusters according to the automatic document clustering techniques (figures on bottom).

4.3.3 Discussion

As previously noted, the results of the application of the clustering techniques on the performance of the speech recognition system will not be known until they are integrated into the unsupervised strategy for the generation of topic-based LMs. In the next chapter, in Section 5.3.4, we will present the experiments and results on this regard.

However, for now, it is worth noting that one of the most remarkable consequences of the application of the automatic document clustering technique is the reduction of system parameters, and particularly, the reduction in the number of topic-based language models. In the supervised approach there are as many topic-based LMs as there are topics in the collection. In this document clustering strategy, the number of topic-based LMs can be reduced to 20 and 17 models, depending on the use of the k-means approach or LDA, respectively.

Obtaining a more uniform distribution of documents within the topic clusters is another advantage of the application of the document clustering strategy. It allows us to increase the conceptual similarity between documents in the same cluster, which also would allow us to expect an improvement of the coverage of the topic-based language model within that cluster.

5 Thesis work on Language Model Adaptation

This chapter presents our **main contributions** and details the experimental conditions under which the work in the area of Language Model Adaptation was carried out. This is the final step of our topic-motivated contextualization framework.

Section 5.1 presents foreground material on language modeling, and the performance metrics that are typically used when evaluating language models (5.1.2).

In Section 5.2 we present our contributions regarding the language model adaptation task. In this regard, we present the two different approaches (supervised and unsupervised) that we propose in this Thesis for the generation of *topic-based* language models (5.2.1). We propose a methodology for the dynamic adaptation of language models, by means of a linear interpolation between a background general LM and a *context-dependent* LM. Within this methodology we propose various strategies to create the *context-dependent* LM as well as different approaches for the adaptation of the model used in the final ASR stage of our architecture (5.2.2).

Finally, in order to evaluate our contributions in this task, in Section 5.3 we present the experimental framework as well as the results obtained not only for the dynamic adaptation LM methodology but also on the application of the proposed contributions in speech recognition.

5.1 Foreground on Language Modeling

Language modeling aims to improve the performance in various natural language applications by assigning a probability to a given sequence of words. One of its main properties is that it reduces the search space; since many of the problems related with natural language search for a solution among multiple candidates, language modeling allows to assign probabilities to all possible candidate paths and therefore, the search will be restricted only to those with a high probability of occurrence. Looked at in another way, language modeling aims to provide information on the context; in this way the probability of one of the multiple candidates will be conditioned by the context in which it occurs.

The most commonly used method in speech recognition for estimating word proba-

bilities is N-gram language modeling. To illustrate this method, let us consider a string of words $w_1, w_2, \ldots, w_{n-1}, w_n$ (which can be also represented as $w_1 \ldots w_n$ or w_1^n). In N-gram language modeling we are interested in knowing the probability of a word given the previous words in the string. To solve this problem, we can first consider the probability of the whole string, given by $P(w_1, w_2, \ldots, w_{n-1}, w_n)$. This probability can be computed by the chain rule of probability, which decomposes the probability of the string into

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2|w_1)P(w_3|w_1w_2)P(w_4|w_1w_2w_3)\ldots P(w_n|w_1^{n-1}) \\ &= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \end{aligned} \tag{5.1}$$

Since it is not feasible to compute the probability of a word given an arbitrarily long sequence of preceding words, Eq. (5.1) can be approximated by computing the probability of the word given only the previous $N-1$ words [Jurafsky and Martin, 2006]. This approximation is called a N-gram. To illustrate a typical case, we take into account the probability of a word given the previous one, (that is $N = 2$), which is called a bigram; the probability of the string then is approximated as:

$$\begin{aligned} P(w_1^n) &\approx P(w_1)P(w_2|w_1)P(w_3|w_2)P(w_4|w_3)\ldots P(w_n|w_{n-1}) \\ &\approx \prod_{k=1}^{n} P(w_k|w_{k-1}) \end{aligned} \tag{5.2}$$

A trigram model ($N = 3$) has the same underlying concept, except that it has to be computed on the two previous words. The same applies for high order N-grams.

N-gram models are usually estimated from large text corpora containing thousands (even millions and more) of words covering a broad range of topics. They can be trained by counting the number of occurrences of a word given its preceding words. From these counts the probability is calculated using the maximum likelihood estimate. For instance, for a particular bigram, we have:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} \tag{5.3}$$

Where $C(x)$ is the number of occurrences of the sequence of words x. Since the sum of all the bigrams that start with a given word w_{n-1} is the same as the number of occurrences of word w_{n-1}, then the Eq. (5.3) can be simplified as

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})} \tag{5.4}$$

In general, for any value of N, the N-gram model is expressed by

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})} \tag{5.5}$$

In order to simplify the notation, we will define all the previous words w_{n-N+1}^{n-1} simply as the history h_n given the word w_n. Thus, the N-gram $P(w_n|w_{n-N+1}^{n-1})$ can be simply rewritten as $P(w_n|h_n)$ or in a more general case as $P(w|h)$.

The key difficulty with using N-gram language models is the data sparsity. Even training N-gram models (particularly models for which $N \geq 2$) with large amounts of data, it is virtually impossible to collect a training corpora that would cover all the instances, therefore it is impossible in practice to avoid the problem of unseen events.

5.1.1 Smoothing

If a N-gram never occurs in the training text, then the method of maximum likelihood estimation will assign any string which contains the trigram, a probability of zero. Thus some method must be used to assign non-zero probabilities to events that have not been seen in the training text. This method is known as *smoothing*.

This method basically assigns a low value to probabilities of unseen N-grams. This cannot be done directly since the probabilities of N-grams sharing the same $N-1$ history should sum up to 1. The basic idea is to take little amount of probabilistic mass from seen N-grams and distribute it over unseen (but possible) ones.

One simple way to perform smoothing might be just to take the counts of words, before converting them into probabilities, and add one to all the counts. This simple method is called **add-one** smoothing. However, there is a number of smoothing techniques which correspond to different methods to take out probabilistic weight. Smoothing methods differ according to how much is subtracted out (discounting) and how it is redistributed (back-off). Both discounting and distribution of probabilistic weight form the notion of *smoothing*.

Among most well known techniques for N-gram smoothing is Good-Turing [Good, 1953], Witten-Bell [Witten and Bell, 1991] and Kneser-Ney [Kneser and Ney, 1995]. In Chen and Goodman [1999] it is presented a complete summary with detailed comparisons between different techniques.

5.1.2 Performance metrics

5.1.2.1 WER

Since the main goal for language models is to improve the speech recognition, the most straightforward way to measure the performance of a language model is to test them in a speech recognition system. The most intuitive direct measure of recognition performance is the word error rate (WER), which is a measure of how accurately an ASR system recognizes speech utterances. Simply, the WER is computed as the ratio of word errors in the ASR output to the total number of words in the correct reference transcription.

In detail, the WER is calculated for a reference transcription (with N_r as the total number of words) as the percentage of the number of substituted (S), deleted (D), and

inserted (I) words in the output text generated by the speech recognizer.

$$WER = \frac{S + D + I}{N_r} \tag{5.6}$$

Evaluating the speech recognizer separately with different language models and calculating the WER on the evaluation transcription is a good measure when comparing performances between language models. However, its main drawback is that it is very time consuming. It must be considered that one recognition run on a big evaluation dataset can take several hours to complete. In those cases for which computational resources are limited, a criterion to prioritize experiments should be followed. This should be understood as a personal consideration of the author of this Thesis taking into account the current available technology and, obviosuly, the particular conditions of the experimental framework.

5.1.2.2 Perplexity

A common way to evaluate the effectiveness of a N-gram language model is to measure how well the language model predicts the word sequence in a given evaluation text. One way to do it is by means of the *cross entropy* between the language model and the evaluation text. This information theoretic metric, is defined, in this case, as the average number of bits needed to encode each word in an evaluation text (T), given the language model $P(w|\mathrm{h})$.

$$H_P(T) = -\frac{1}{N}\sum_{i=1}^{N} \log P(w_i|\mathrm{h}_i) \tag{5.7}$$

where N is the size of the test set. From the cross entropy we can derive the more commonly used measure of perplexity, which is defined as:

$$PP_P(T) = 2^{H_P(T)} = \frac{1}{\sqrt[N]{\prod_{i=1}^{N} P(w_i|\mathrm{h}_i)}} \tag{5.8}$$

Perplexity can be intuitively thought as the approximate number of equally probable words the language model has to choose from when predicting the next word in the evaluation text [Kim, 2004].

The use of perplexity is mainly motivated for practical reasons; it is easier to manage absolute values in the usual range of perplexity, that is between 100-200, than numbers in terms of bits (as values of 6.64 and 7.64 bits corresponding to perplexity values of 100 and 200 respectively). It must be noticed that the absolute perplexity value is usually not so important when evaluating a language model. This absolute value depends on the model but also on the evaluation text. The relative perplexity reduction is a more important measure when compared to a baseline model.

However, perplexity is not a perfect metric, it has not been totally accepted that a reduction in perplexity correlates with a reduction in the WER. There have even been numerous reports which show higher error rates even though the language model presented a lower perplexity [Rosenfeld, 2000].

5.2 Contributions on Language Model Adaptation

5.2.1 Language Model Interpolation

When analyzing spontaneous and multitopic spoken language the election of content words is driven 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. There are also syntactical structures, such as phrases or named entities, that are specific to certain topics. For these reasons, a system that works in a multitopic domain should be able to use these characteristics of language and take advantage of them. In this sense, within statistical language modeling, a way to take advantage of these characteristics is by increasing the probabilities of some words, or some sequences of words, depending on the topic of the speech.

If we do not include new sources of information to our systems or otherwise exploit existing information in a different way, the models will remain static. That is, regardless of the addressed topic, domain or style, the probability of events and sequences of events, will not change.

A static model is not the best option for modeling language in multitopic speech. In a natural conversation between humans, the topic, subject, genre, style, etc. changes often, and therefore the language 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.

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. Given a *background* model $P_B(w|h)$ and a *context-dependent* model $P_{CD}(w|h)$, it is possible, by means of their interpolation, to obtain a *final* model $P_F(w|h)$, to be used in the second decoding pass of our recognition architecture, as follows:

$$P_F(w|h) = \lambda P_B(w|h) + (1-\lambda)P_{CD}(w|h) \tag{5.9}$$

where λ is the interpolation weight between both models, which weights the contribution of P_B, compared to the contribution of P_{CD}, to the *final* model P_F. The interpolation weight has to fulfill the condition $0 \leq \lambda \leq 1$. Later in this section we describe how we have created each of these models and what sources of information were taken into account for their generation.

There are different ways in which the interpolation weight can be selected: it can be set empirically by minimizing the perplexity in a development stage with data not seen during training [Clarkson, 1999, Tur and Stolcke, 2007]; it can also be estimated under some optimization algorithm, such as Expectation Maximization [Daumé et al., 2010]

or Maximum A Posteriori (MAP) adaptation [Wang and Stolcke, 2007]; or it can be set dynamically depending on a specific situation of the recognition process (related to the topic, the speaker, etc.) [Seymore and Rosenfeld, 1997]. In this work, we explore different options, including the last of the options mentioned above, to determine the interpolation weight of the models.

The underlying idea in our work is to exploit the information provided by the Topic Identification system in the generation of the *context-dependent* and the *final* language models. The scheme followed in this work for the generation of the LMs in the different stages of the process is presented in Figure 5.1. In our approach, model interpolation occurs at two different levels: for generating the context-dependent LM and, in a final instance, for creating the final dynamic LM used in the second decoding pass.

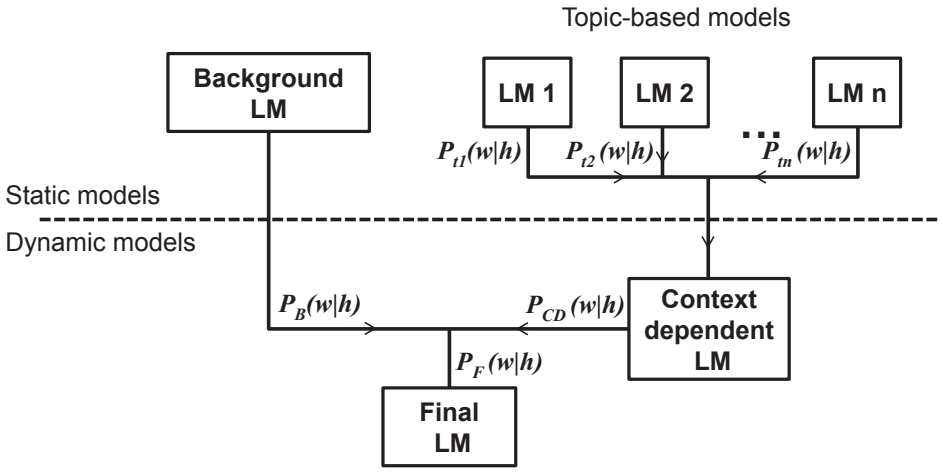


Figure 5.1: Scheme of interpolation of language models

In the first level, model interpolation consists of generating a *context-dependent* LM by selecting just one or combining several *topic-based* LMs - $P_t(w|h)$ through some balancing weights.

In the second level, the *context-dependent* LM is then interpolated with the *background* LM, generating the dynamic *final* LM that the speech recognizer will use in the second decoding pass. The *background* model is a general model. It is trained with more, but not specific, data. On the other hand, the *context-dependent* 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, and sequences of words, that better match the discussed topic.

For the generation of the topic-based language models, we propose two different approaches, to which we refer to as *supervised approach* and *unsupervised approach*. Each of these approaches has its own variants, which we describe later in this section.

The first of these approaches, the *supervised* one, is intended to generate topic-based language models by grouping the documents, in the training set, according to the original topic hand labelling of the document collection, as it is depicted in Figure 5.2. By doing this we create a topic-based LM P_{tz} from the documents that belong to topic z, where z is one of the available topics in the original database; this means that by

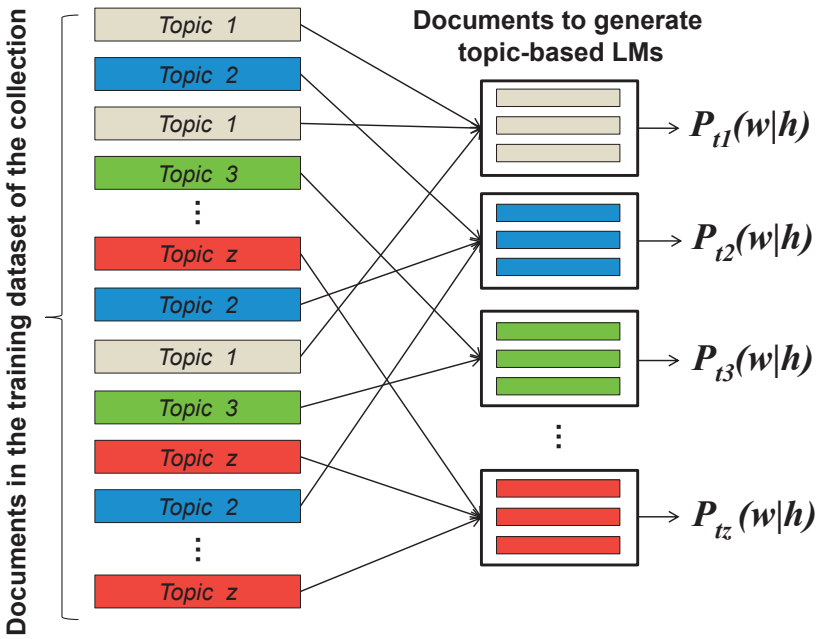


Figure 5.2: First approach for the generation of *topic-based* models - supervised approach

this approach we generate as many topic-based LMs as there are topics in the original collection. This approach, despite it may seem intuitive, generated some questions:

i) Considering the distribution of the documents in the collection (see Figure 3.7), will some topic-based LMs be estimated with much more data than others?

ii) Can we be totally confident of the topic hand labeling of the documents in the collection? This question is motivated for two reasons. First, the labeling process is a subjective process and as such may be subject to errors. Secondly, there are interventions in the middle of a political debate that, despite being labeled within a specific topic (the very topic of the debate), may be related to other different topics. For this reason, the topic label of the entire document may not describe the topic to which all the interventions of the document belong.

iii) Could a smaller number of topics represent in a more concise and compact way the topic content of all documents in the collection?

In order to try to find a solution for these questions, we propose a second approach, an *unsupervised* one, in which the objective is to group the data in the training set into automatic topic clusters as depicted in Figure 5.3. By means of this unsupervised clustering approach we expect to obtain not only a more uniform distribution of the documents, but also a cohesive association of documents that are related by similar concepts. This approach would also allows us to obtain a reduced number of topic-based LMs, reducing the number of parameters involved in the whole system.

For both approaches, when additional sources of information are included in our experiments, an automatic labelling of the new data can be done in order to include

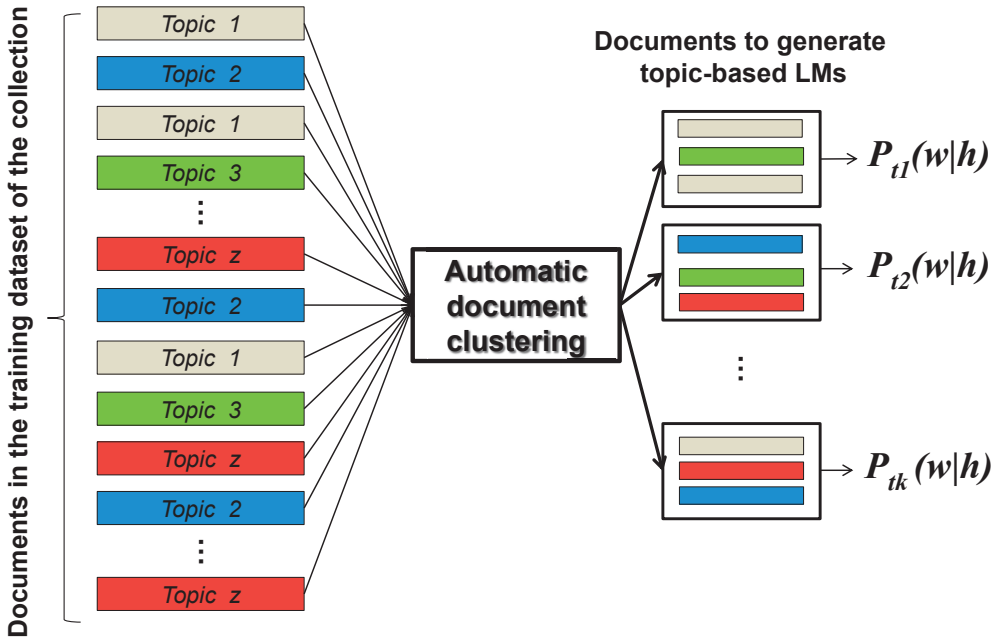


Figure 5.3: Second approach for the generation of *topic-based* models - unsupervised approach

them in each specific topic training data. In this regard we have explored different alternatives to include data. In Section 5.3.1 we will present the different alternatives we evaluated and the selected option.

In our case, the *background* model and the *topic-based* models are static models. They are trained once, and they remain unchanged during the evaluation. We use the same vocabulary to train the background model as well as the topic-based language models (general details of the database we used to extract the lexicon can be seen in Table 3.4). The *context-dependent* LM could be either static or dynamic. It depends on the adaptation scheme followed, as we will see later. This model, as well as the final model $P_F(w|h)$, are generated online during the processing of each audio segment.

LM adaptation strategies proposed in this work differ in two ways: how to build or derive *context-dependent* 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.

5.2.2 Interpolation Schemes

Two questions arise at this point. How to generate the *context-dependent* model? and, how to determine the interpolation weight λ between the background model and the context-dependent model? Well, for solving these questions, we can think of different possible solutions. Let us consider first what are our options when it comes to generating the *context-dependent* model.

- In our contextualization framework, we want this model to be dependent on the topic that is being addressed, therefore a natural approach is to use the topic-

based models in its generation.

- We could consider one or more of the topic-based LMs in the generation of the context-dependent model. The choice of the model, or models, can be conditioned by the outcome of the topic identification system. For instance, we could select a model depending on how similar the audio segment is to the topics in the collection.

- When using several topic-based LMs other questions arise: how many models should be considered? and how to combine them into the context-dependent model? We explore different alternatives in order to give solution to these new questions:

  - We believe that all topic-based models can be considered in the generation of the context-dependent model. Actually, one of our approaches is to consider all of them. However, we explore a different alternative and it is to consider a subset of them. There are many possible subsets, so we had to make a hard decision in this regard and it was to consider only the top-10 LMs related to the most similar topics.
  - A straightforward solution to combine the topic-based LMs is by means of linear interpolation. To compute the interpolation weight between them there are different alternatives. For instance, it can be set experimentally by evaluating the performance of the speech recognition system on the development dataset. Or we can use the similarity measure provided by the topic identification system. By doing this, we could give more weight to the topics ranked in the first positions by the topic identification system.

Regarding the interpolation weight between the background model $P_B(w|h)$ and the context-dependent model $P_{CD}(w|h)$ we also explored different alternatives:

- We can look for the value that minimizes the word error rate of the speech recognition system on the development set.

- We can estimate it by taking into account different metrics like the similarity measure of the topic identification system or some distance between the topic-based LMs and the background model.

In the next sections we present the different interpolation schemes we propose to obtain the context-dependent model $P_{CD}(w|h)$, as well as the dynamic interpolation weight between this model and the background model. These interpolation schemes take into account the previous considerations and are intended to include the information provided by the topic identification system into the contextualization of the language models.

5.2.2.1 Hard interpolation scheme

In this scheme, we built the context-dependent LM $P_{CD}(w|h)$ by considering only one of the topic-based language models, that is:

$$P_{CD}(w|h) = P_t(w|h)$$

Where $P_t(w|h)$ is the topic-based model related to the topic ranked in the first position by the topic identification system. For obtaining the final model $P_F(w|h)$ we need first to estimate the interpolation weight λ (i.e. the interpolation weight between the context-dependent model and the background model). In this regard we explore different options.

We can set λ experimentally as the value that minimizes the word error rate of the speech recognition system on the development set. Despite this approach implies to perform a large number of experiments, it allows us to obtain an optimal value of the interpolation weight (for the development set).

We can also estimate λ by considering a distance measure between the context-dependent LM and the background LM. In this sense, we propose a distance measure δ_T between these models. In the proposal of this distance, our hypothesis is that the greater the distance between both models, the greater the contribution of the context-dependent model should be to the final one. We can compute this distance by considering the average difference in the unigram probabilities of both models as follows:

$$\delta_T = \frac{1}{N} \sum_{\forall w_i \in P_{CD}} |P_{CD}(w_i) - P_B(w_i)| \tag{5.10}$$

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

$$\lambda = \frac{\delta_T}{\sum_{j=1}^{n} \delta_j} \tag{5.11}$$

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

5.2.2.2 Soft-1 interpolation scheme

In this case, instead of using only one topic-based LM for generating the context-dependent LM, this model is built on a dynamic basis by the interpolation of a different number of topic-based LMs. The **Soft-1 interpolation scheme** tries to gather the dynamics of the right combination of the topic-based models $P_t(w|h)$ depending on

the similarity of the audio segment to each of the topics. This similarity is provided by the topic identification system. By doing this, we can expect to give more relevance to the topics ranked in the first positions by the topic identification system. In this approach we compute the context-dependent LM as follows:

$$P_{CD}(w|h) = \alpha_1 P_{t_1}(w|h) + \alpha_2 P_{t_2}(w|h) + \cdots + \alpha_k P_{t_k}(w|h) \tag{5.12}$$

where k is the number of models considered for obtaining the topic-based LM. To compute the interpolation weight α_i we consider the similarity of the audio segment to the topic i. As the sum of all α_i must be one, we consider the normalized value of the similarity measure of the TI system.

$$\alpha_i = \frac{sim(\vec{d}_i, \vec{q})}{\sum_{j=1}^{k} sim(\vec{d}_j, \vec{q})} \tag{5.13}$$

Where $sim(\vec{d}_i, \vec{q})$ is the similarity measure (i.e. the cosine similarity) computed by the topic identification system between a topic (represented by a vector $\vec{d}$) and a test document (represented by the vector $\vec{q}$). The interpolation weight λ between the background LM and the topic-based LM can be obtained by different ways. We can set it experimentally as the value that minimizes the word error rate of the speech recognition system for the development set (as we proposed in the previous scheme), or we can estimate it by taking into account different metrics like the similarity measure of the topic identification system or the distance δ_T between the topic-based LMs and the background model. In this scheme we choose the former option, that is by setting λ experimentally. The latter option will be explored in the next section.

5.2.2.3 Soft-2 interpolation scheme

This scheme is similar to the previous one, but instead of setting λ experimentally, we proposed to estimate it by combining different metrics. Our objective in this scheme is to consider, not only the similarity measure of the audio segment to each of the topics, but also the distance δ_T between the topic-based LMs and the background LM.

In this sense, our proposal is to compute it by means of the sum, for all the topics, of the similarity measure provided by the topic identification system weighted by the normalized distance δ_T. That is:

$$\lambda = \sum_{i=1}^{k} \alpha_i \cdot \frac{\delta_i}{\sum_{j=1}^{k} \delta_j} \tag{5.14}$$

This scheme allows us to estimate dynamically and automatically both context-dependent LM and the interpolation weight between this model and the background LM for each audio segment we are analyzing.

In Soft-1 and Soft-2 schemes, we have considered two additional possibilities:

a) To create the context-dependent LM using all the topic-based LMs, that is by setting k as the total number of topics.

b) To create the context-dependent LM by selecting a subset of the topic-based LMs. Since there are many possible subsets, we had to make a hard decision in this regard and it was to consider only the 10 topics with higher positions in the topic identification ranking.

5.3 Experiments on Language Model Adaptation

5.3.1 Additional databases - The EUROPARL corpus

For improving the coverage of the background language model and topic-based language models we looked for additional text data that might be related to similar topics in the same domain. Despite there are limited resources in Spanish in this regard, we found two possible sources of new data.

On the one hand, we found the *Spanish Parliament database* also known as PARL corpus, which belongs to the same project as the EPPS database. PARL transcriptions consist of 38 hours of speech of members of the Spanish Parliament speaking in the Spanish Parliament and Spanish Congress during plenary sessions and commissions. By adding the documents in this database to the EPPS database and by training a background LM with these data, the performance of the system did not improve neither in the development nor in the evaluation datasets when compared to a background LM trained only with data from EPPS database. Despite this database is related to the same domain, i.e. the political domain, the topics differ to a large extent, which may impoverish the estimation of the language models.

We also found the EUROPARL corpus [Koehn, 2005], which consists of sentences extracted from the debates of the European Parliament in the period between the years 2006 and 2011. This corpus is composed of a parallel corpus in different languages for statistical machine translation. We have extracted approximately two million sentences in Spanish. Preliminary experiments on the development dataset showed that this corpus, when added to the EPPS database in the estimation of the background LM, significantly improve the performance of the system. For this reason we decided to take advantage of this corpus in two distinct levels: In a first level we added the extracted sentences to the text of the training set for generating the background language model. And we also used them for enhancing the robustness of the topic-based LMs. Depending on the approach that is being evaluated (supervised or unsupervised) we make use of this database in the following ways.

- **Supervised approach:** To take advantage of the EUROPARL corpus in this approach, we classify each sentence of the corpus into one of the available topics. To do this, we make use of the Topic Identification system, particularly the system with the best combination of parameters, i.e. the one that uses the **Term**

Inventory 2, *tf - pseudo-entropy* as weighting scheme and LSA as model for document representation.

The first step in this process is to treat each sentence of the EUROPARL corpus as a document and represent each of them in a vector space using the **Term Inventory 2**. Recall that this inventory was created by applying the index-terms selection techniques as described in Section 3.3.8. The next step is to apply a weighting scheme to each document vector (in this case, as previously mentioned, the scheme that we use is the one formed by the combination of *term-frequency* as local weight and *pseudo-entropy* as global weight).

And finally, to classify each of the sentences of this corpus, we followed the *LSA classification procedure* as described in Section 3.3.5. In this procedure, each sentence is projected into the latent semantic space. Then, it is computed the similarity between each sentence and the vectors that represent each of the original topics. Each sentence will be classified according to the topic vector with the highest similarity. By doing this to all the sentences in the EUROPARL corpus we can automatically classify them into one of the hand labelled topics. Once they are classified into a specific topic, we can add these sentences to the training text of each specific topic, and use them to estimate topic-based language models.

- **Unsupervised approach:** In a similar way as in the supervised approach, our objective is to classify each sentence of the EUROPARL corpus into one of the automatic generated topic-clusters, and use them to train topic-based language models. In this sense the first step is to represent each sentence in the representation space that is more adequate depending on the automatic document clustering technique that we use (i.e. *k*-means or LDA). For instance, in the case of the *k*-means technique, we have to represent each sentence in the latent semantic space. Then, it is computed the similarity between each sentence and each of the cluster centroid vectors that represent the topic clusters. Each sentence will be classified according to the cluster centroid vector with the highest similarity.

  In the case of the LDA technique, each sentence of the EUROPARL corpus is represented as a document vector using the Vector Space Model. To classify each sentence into an automatic topic cluster, it is followed the procedure for new documents described in Section 4.1.2.

  Once the sentences in the EUROPARL corpus have been classified into one of the automatic topic clusters (either by means of *k*-means or LDA), we can merge these sentences with the training text of each specific topic cluster, and use them to train the topic-based language models that are used in the unsupervised approach.

5.3.2 Introduction to Speech Recognition experiments

We have evaluated the topic contextualization strategy by measuring the improvement in the ASR system. To do this we measure the recognition performance in terms of

word error rate in both ASR stages. We established the first decoding pass (ASR Stage 1) as the baseline performance of the system (in this stage only the background LM was used). The relative improvements of the adaptation strategy on the recognition performance were also calculated.

The details of the speech recognizer are described below:

- Acoustic models: The feature vectors we used for the acoustic model training consisted of the first 13 PLP coefficients, as well as their first and second order time derivatives. The phone models were composed of three hidden states each. We used cross-word triphone models in order to account for contextual information and we consider up to 16 Gaussians per state during training. We use the same acoustic models for both ASR stages.

- Language models: The background language model was trained with the original transcription of the sentences of the training dataset and the sentences of the EUROPARL corpus. The background LM is composed of nearly 2.8M trigrams. We use trigram models for both ASR stages. The vocabulary that we used to train the LMs was extracted from the transcriptions of the training dataset (see Table 3.4 for more details on this dataset). We use the same vocabulary to train the background model as well as the topic-based LMs.

The OOV rate (out-of-vocabulary words rate) is 1.2%. The performance of the baseline system (without dynamic LM adaptation) achieved a WER of 21.75%. This result is the same for both evaluation datasets, since the individual audio segments that are recognized are the same for both sets. Recall that the difference between the evaluation datasets is the way in which we have grouped the individual audio segments together to form interventions.

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 SRILM Toolkit [Stolcke, 2002] was employed for creating and interpolating the language models that the system uses in both ASR stages.

Below we present the experiments of the different approaches we have followed in the evaluation of the dynamic LM adaptation. For a clear presentation of the results, we have divided them into the two main approaches we have followed in the generation of topic-based LMs.

In Section 5.3.3 we present the results we obtained by applying the *supervised* approach to the generation of the topic-based LMs. Recall that in this approach topic-based LMs are generated by grouping the training documents according to the original hand label topics.

In Section 5.3.4 we present the results we obtained by applying the *unsupervised* approach, that is by automatically clustering the documents to generate topic-based LMs.

For the evaluation of the dynamic LM adaptation we have used the best configuration of parameters obtained in the topic identification experiments.

5.3.3 Results on the *supervised* approach for the generation of topic-based LMs

In this approach topic-based LMs have been trained according to the original topic labels of the documents in the training dataset.

In Table 5.1 we present the results of the speech recognition performance within this approach. These results are shown for both configurations of the evaluation dataset and for each of the proposed interpolation schemes for the dynamic LM adaptation.

	Eval. Set 1		Eval. Set 2	
Adaptation approach	WER	Rel.Imp. (%)	WER	Rel.Imp. (%)
Baseline (no adapt.)	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	$\mathbf{19.25 \pm 0.25}$	**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	$\mathbf{18.98 \pm 0.25}$	**12.73**

Table 5.1: 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

Figure 5.4 presents the best results of each interpolation scheme for both evaluation datasets. In this figure we can appreciate that in general terms, for both configurations of the test dataset, *Evaluation Set 1* and *Evaluation Set 2*, a statistically significant reduction of the word error rate, when compared to the baseline system, can be obtained by applying the dynamic language model adaptation schemes proposed in this Thesis.

However, among these results, and particularly among both evaluation datasets, there are some differences that are worth mentioning.

On the Evaluation Set 1. Although there is not a significant difference between the **Soft-1** and the **Soft-2** schemes 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** (for which we obtained the lower word error rate) and the **Hard** scheme. A relative improvement of 11.49% can be achieved when comparing this soft integration to the baseline.

The **Hard** scheme takes only into consideration the most probable topic according to the topic identification system. Thus, in this scheme the topic-based LM is created by using only one of the topic-specific LMs. This significantly reduces the performance of the dynamic LM when compared to the **Soft** schemes.

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.

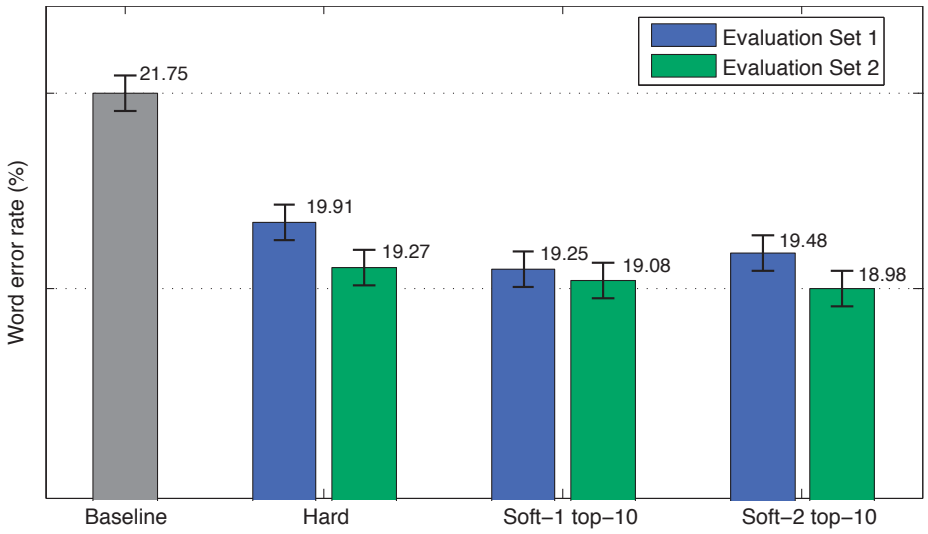


Figure 5.4: Best results for the *supervised* approach

Actually, there is not a significant variation in the word error rate among both variants.

On the Evaluation Set 2. In this dataset all the LM adaptation approaches present a similar result in WER and there are not significant differences between them. In general, *Evaluation Set 2* exhibits a lower word error rate when compared to *Evaluation Set 1*. In *Evaluation Set 2*, audio segments are equal or shorter than in *Evaluation Set 1*. Thus, by processing shorter audio segments, a more refined LM adaptation can be done for each of them.

In future work we expect to study more deeply the relation between the length of the segment and the performance of the system. Based on our results, 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.

In *Evaluation Set 2*, the LM adaptation scheme with the absolute minimum error is the **Soft-2 - top 10** approach. With this soft integration we manage to reduce significantly 12.73% of the initial WER when compared to the baseline method.

5.3.4 Results on the *unsupervised* approach for the generation of topic-based LMs

In this section we present the results we obtained by performing the LM dynamic adaptation based on the automatic "topic clusters" created in the clustering process described in Chapter 4.

Recall that in this approach, topic-based LMs are generated by automatically clustering the documents in the collection. By means of this unsupervised clustering approach

we expect to obtain not only a more uniform distribution of the documents, but also a more cohesive association of documents that are related by similar concepts.

Within this *unsupervised* approach we evaluated two different clustering strategies, k-means and Latent Dirichlet Allocation (LDA). The first step in both strategies is to determine an optimal number of clusters in which group the documents. In Section 4.3 we presented the experiments we performed in this regard for both clustering strategies. The number of clusters we obtained (i.e. 20 for the k-means strategy and 17 for the LDA strategy) is optimal according to the criterion we use to get it (the Silhouette Coefficient), but we do not know if it will be optimal in terms of improving speech recognition. For this reason we have conducted some experiments to evaluate the performance of the system by using a different number of clusters to generate the topic-based LMs.

Below, we **first** present the results of the experiments conducted to **evaluate the SC criterion**. **Next**, we will present the experiments and the results of the unsupervised approach for the dynamic language model adaptation. For a clear presentation we will introduce the experiments on the application of k-means and then the results on the application of LDA.

5.3.4.1 Exploratory evaluation of the SC criterion

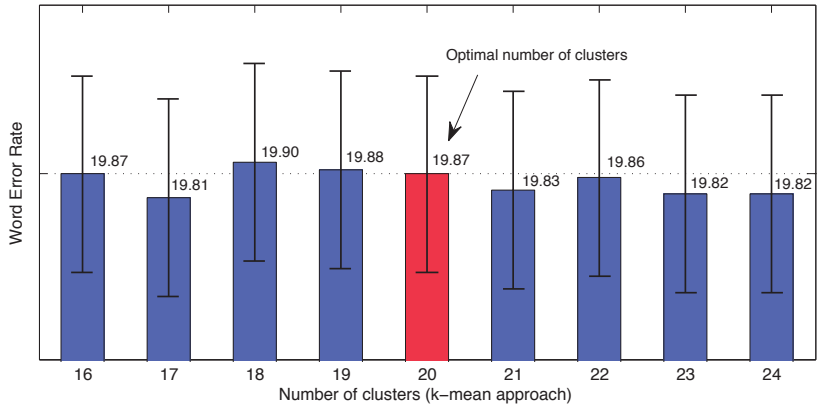
It is important to note that the criterion we use to find the optimal number of clusters (the overall average Silhouette coefficient) is based on minimizing the *within distance* of documents in the same cluster and at the same time maximizing the distance between clusters, but it does not undoubtedly suggest an improvement of the language models generated for each cluster, nor an enhancement of the speech recognition performance.

In this sense we have conducted some preliminary, and also exploratory, experiments to analyze the performance of the automatic speech recognition system by clustering the documents in a number of groups other than the suggested by the Silhouette coefficient criterion. We have performed these experiments varying the number of clusters around the maximum overall average Silhouette coefficient. We have decided to conduct them by applying only the Hard interpolation scheme, since this scheme allows us to see the general performance of the system. Figure 5.5 shows the results of the experiments by applying the *Hard* interpolation scheme on the *Evaluation Set 1*.

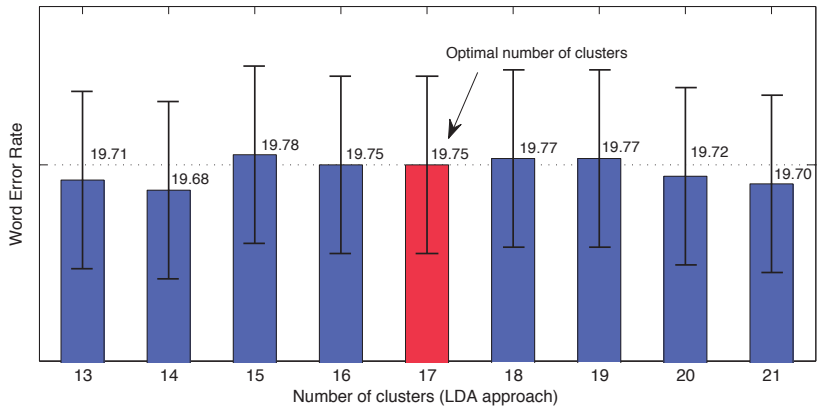
These results suggest that, in terms of the recognition performance, the number of clusters provided by the Silhouette Coefficient offers an appropriate solution since it improves the recognition performance when compared with the baseline system. However, they also suggest that this number of clusters is not optimal since similar results can be obtained with less clusters without a significant loss of performance.

5.3.4.2 *k*-means strategy for generating topic-based LMs

In this section we present the results obtained by applying the k-means strategy in the generation of topic-based LMs. The number of clusters, and therefore, the number



(a) WER obtained for a different number of clusters in the k-means approach



(b) WER obtained for a different number of clusters in the LDA approach

Figure 5.5: Speech recognition experiments conducted by varying the number of clusters around the optimal point

of topic-based LMs was chosen following the SC criterion, which means that for this strategy we partition the documents in the training dataset into 20 clusters. Table 5.2 presents the results obtained by applying k-means to automatic document clustering.

Figure 5.6 shows the best results of Table 5.2. We can appreciate a significant reduction of the word error rate for all the dynamic adaptation schemes when compared to the baseline system without adaptation.

The result for *Evaluation Set 1* shows that 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 suggests that in the Hard approach, in which only one of the topic-based LMs is considered for adapting the dynamic LM is not optimal in terms of improving the speech recognition performance.

For *Evaluation 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.

For the *Evaluation Set 2*, all the LM adaptation schemes present a similar result in WER and there are not significant differences between them. This evaluation dataset

	SET 1		SET 2	
Adaptation approach	WER	Rel.Imp. (%)	WER	Rel.Imp. (%)
Baseline (no adapt.)	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	$\mathbf{19.21 \pm 0.25}$	**11.67**	$\mathbf{18.81 \pm 0.25}$	**13.52**

Table 5.2: Word Error Rate (WER) and Relative Improvement (Rel.Imp.) for the different LM adaptation approaches when performing the k-means document clustering for the generation of the topic-based LMs

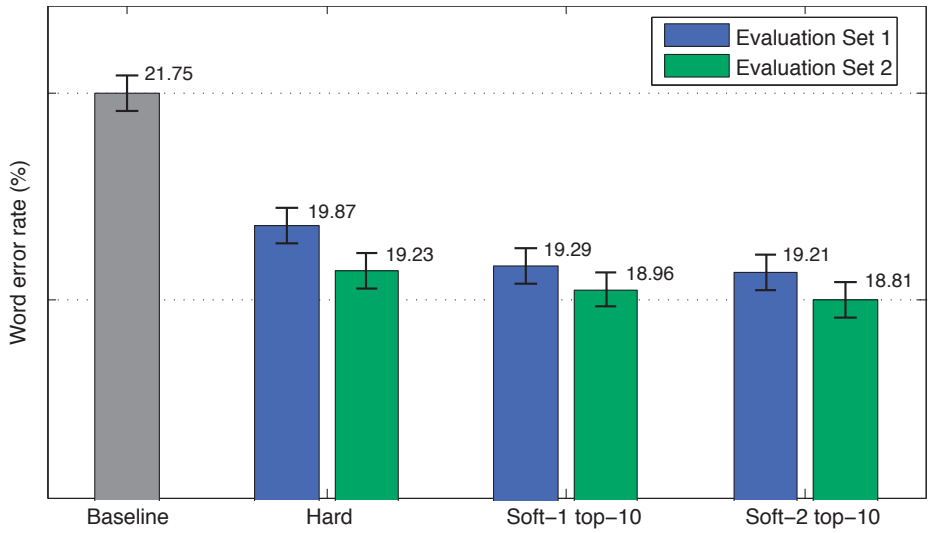


Figure 5.6: Best results for the unsupervised approach using k-means as clustering strategy

exhibits a lower word error rate when compared to *Evaluation Set 1*.

In *Evaluation 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.

It must be noticed that this strategy reduces the number of topic-based language models when compared to the supervised approach. Recall that in the supervised approach there are as many topic-based LMs as there are topics in the collection. In this unsupervised clustering approach, the number of topic-based language models is reduced to 20, which suggest a clear reduction on the number of parameters and an overall simplification of the system's model.

5.3.4.3 LDA strategy for generating topic-based LMs

In Table 5.3 we present the results obtained by means of applying LDA to automatic document clustering.

	SET 1		SET 2	
Adaptation approach	WER	Rel.Imp. (%)	WER	Rel.Imp. (%)
Baseline (no adapt.)	21.75 ± 0.26			
Hard	19.75 ± 0.25	9.1	19.11 ± 0.25	12.13
Soft 1 - all	19.63 ± 0.25	9.74	19.09 ± 0.25	12.22
Soft 1 - top 10	19.31 ± 0.25	11.21	19.03 ± 0.25	12.50
Soft 2 - all	$\mathbf{19.21 \pm 0.25}$	**11.67**	$\mathbf{18.87 \pm 0.25}$	**13.24**
Soft 2 - top 10	19.23 ± 0.25	11.58	18.92 ± 0.25	13.01

Table 5.3: Word Error Rate (WER) and Relative Improvement (Rel.Imp.) for the different LM adaptation approaches when performing the LDA document clustering for the generation of the topic-based LMs

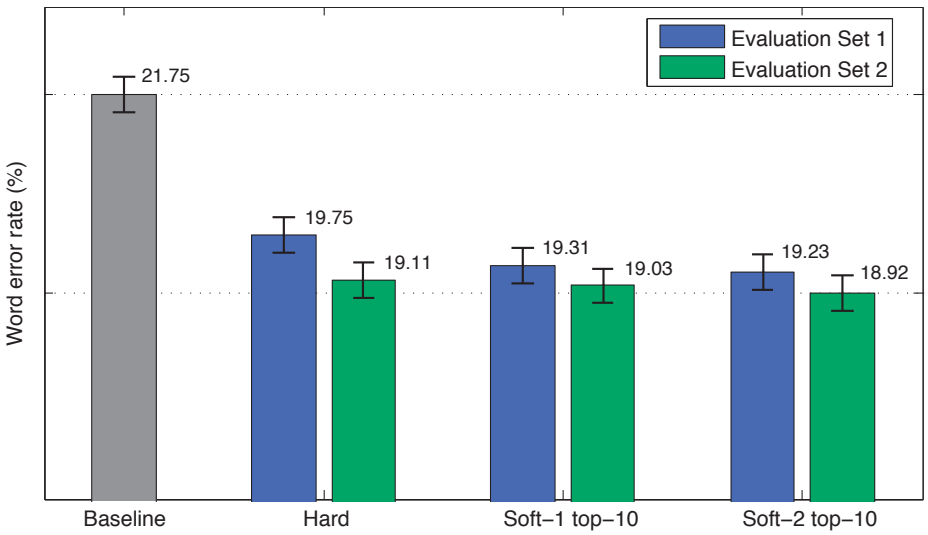


Figure 5.7: Best results for the unsupervised approach using LDA as clustering strategy.

Regarding the LDA approach, the results are similar to those obtained with the k-means approach. We do not appreciate significant differences in the WER between the LDA and the k-means approaches. Neither a significant reduction of the WER is obtained when compared to the supervised approach. Nevertheless, it must to be noticed that this technique provides an alternative clustering solution that reduces the number of topic-based language models, allowing a reduction of the number of parameters of the system.

Note that a significant reduction of the word error rate is obtained when compared to the baseline system. Figure 5.7 shows the best results of Table 5.3. None of the

different approaches shown in Table 5.3 improve significantly the best result obtained in the k-means approach.

As a general conclusion we can say that by clustering the documents, the conceptual similarity may be increased between documents in the same cluster. Therefore an improvement of the coverage of the LM within that cluster may also be expected.

In terms of WER, none of the results obtained with the clustering strategies is statistically significant when compared to the *supervised* approach in which we use the original topic labels of the collection. However, we believe that these results are promising for two reasons: i) in general the recognition performance tends to improve for each of the interpolation schemes we evaluated, particularly the scheme Soft-2; and ii) the clustering strategies that we propose and evaluate allows a significant simplification of the system's model by reducing the number of topic-based language models.

This has led us to believe that the application of the proposed interpolation schemes and our methodology on other application domains and other databases can yield not only similar results but also significantly better.

Finally, it also must be noticed that there is a significant reduction of the word error rate when compared to the baseline system. This is true for both clustering approaches, for the different dynamic LM adaptation schemes and for both configurations of the evaluation dataset. With the clustering strategy we obtained the absolute minimum error among all the systems we evaluated in this Thesis. This was achieved for the k-means strategy by applying the **Soft 2 - top-10** interpolation scheme. A relative reduction of 13.52% of the baseline word error rate was achieved.

5.3.5 Example of the system performance

In this section we present an example of the application of our methodology for language model adaptation. In this example we compare the original transcription of an audio segment of the *Evaluation set 1* with the output of both ASR stages.

At first, below is shown the original transcription of an intervention turn. Recall that an intervention turn is composed by different sentences of the same speaker. We present the transcription for each of the sentences. In this example we are neither considering the punctuation marks nor the case of the letters.

Original transcription of an intervention turn
Topic: Transatlantic relations
Length: 2m10s
1) gracias presidente
2) presidente me consta que las relaciones transatlánticas también incluyen aquellas relaciones con canadá a parte de con los estados unidos
3) y escuchando este debate no lo parecería pero como presidente de la delegación del parlamento europeo con canadá reconozco la gran importancia que tiene el gobierno canadiense y que da a esta relación con los veinticinco estados miembros en la unión europea
4) es importante que nosotros mantengamos este nivel de diálogo a nivel de cumbres y de otro tipo de niveles para enfrentarnos a los retos de la unión europea

5) la nueva agenda se lanzó el dieciocho de marzo del año pasado y establece un grupo de coordinación para asegurar la aplicación efectiva y rápida en las decisiones tomadas a nivel político y que incluyen todos los elementos de la relación entre canadá y la unión europea
6) este diálogo intensificado nos permitirá tener un enfoque más duradero para todos los aquellos candidatos de la unión europea especialmente en diferentes ámbitos como pueden ser temas de seguridad de integración de pesca de culturas seguridad en el transporte etcétera
7) todos trabajamos juntos para mejorar la frecuencia y el calidad del contrato entre esta agencia canadiense y las agencias europeas responsables de la ayuda al desarrollo para aunar nuestros enfoques
8) yo creo que mantener la paz y la seguridad se cumple mejor en un sistema multilateral
9) en la unión europea el gobierno de estadounidense y el gobierno canadiense trabajan juntos para luchar contra el terrorismo internacional para luchar contra la pobreza mundial y para promover la democracia
10) sabemos cuáles son nuestros retos conjuntos así que vayamos a su encuentro gracias

Below we present the output of both ASR stages. We have used some marks in the ASR output so the differences between the output produced by the system without adaptation and the system with the dynamic LM adaptation can be appreciated. The marks we used are:

word for substitutions
DEL for deletions
(word) for insertions

Recall that the first ASR stage makes use of the background language model and therefore no language model adaptation is performed. Below we present the output of the first ASR stage.

Output of the first ASR stage - No LM adaptation is performed
1) gracias presidente
2) presidente me consta ***queda*** DEL DEL transatlánticas ***tomen incluye nadie*** relaciones con canadá a parte de ***cual*** los estados unidos
3) y escuchando este debate no lo parecería pero como ***presidenta*** DEL DEL delegación del parlamento europeo con canadá reconozco la gran importancia que tiene el gobierno canadienses y que da a esta relación con los ***últimos dos*** miembros ***de*** la unión europea
4) es importante que nosotros mantengamos este nivel de diálogo a nivel de cumbres ***ideas*** DEL otro tipo de niveles para (ser) ***órganos*** a los retos de la unión europea
5) (en) la nueva agenda ***ser lanzo*** el dieciocho de marzo ***el*** año pasado y ***establecer*** un grupo de coordinación para asegurar la aplicación efectiva DEL rápida ***y*** las decisiones tomadas a nivel político y que incluyen ***todo*** los elementos de la relación entre canadá y la unión europea
6) este diálogo ***identificado*** nos permitirá tener un enfoque más (sea) (el) duradero para todos los ***días*** candidatos ***dinero*** DEL DEL DEL ***perfectamente*** en diferentes ámbitos como pueden ser (el) ***tema*** de seguridad DEL integración de pesca de ***cultura asegurar*** DEL DEL transporte etcétera
7) todos trabajamos juntos para mejorar (y) la frecuencia y DEL ***élite el*** contrato entre esta agencia (en) ***oriente*** y las agencias europeas ***responsable*** de la ayuda al desarrollo para aunar nuestros enfoques
8) yo creo que mantener la paz y la seguridad se ***cumplen*** mejor en un sistema multilateral

9) en la unión europea (el) el ***gobernanza unir ese si un fenómeno canarias*** (he) ***trabajado*** juntos para ***lucha*** contra el terrorismo internacional para luchar contra la pobreza mundial y para promover la democracia
10) sabemos cuáles son nuestros retos conjuntos así que vayamos DEL ***son cuando*** gracias

The **WER** of the system for this audio segment is 23.42%

In the final stage of the architecture, the ASR makes use of the adapted model. Below in this example we present the output of the final ASR stage applying the unsupervised approach (using the *k*-means algorithm for the generation of topic-based language models).

Output of the final ASR stage - LM adaptation performed using Soft 2 - top 10 scheme and automatic topic clustering for the generation of topic-based LMs
1) gracias presidente
2) presidente me consta ***queda*** DEL DEL transatlánticas ***tomen*** incluyen ***nadie*** relaciones con canadá a parte de con los estados unidos
3) y escuchando este debate no lo parecería pero como presidente de DEL delegación del parlamento europeo con canadá reconozco la gran importancia que tiene el gobierno canadienses y que da a esta relación con los ***últimos dos*** miembros ***de*** la unión europea
4) es importante que nosotros mantengamos este nivel de diálogo a nivel de cumbres DEL ***ideas*** otro tipo de niveles para (ser) ***órganos*** a los retos de la unión europea
5) (de) la nueva agenda se lanzó el dieciocho de marzo ***el*** año pasado y ***establecer*** un grupo de coordinación para asegurar la aplicación efectiva DEL rápida ***y*** las decisiones tomadas a nivel político y que incluyen todos los elementos de la relación entre canadá y la unión europea
6) este diálogo ***identificado*** nos permitirá tener un enfoque más (sea) (el) duradero para todos los ***días*** candidatos de la ***dinero*** DEL ***perfectamente*** en diferentes ámbitos como pueden ser temas de seguridad ***y*** integración de pesca de culturas **asegurar** en DEL transporte etcétera
7) todos trabajamos juntos para mejorar la frecuencia y DEL ***élite*** del contrato entre esta agencia (en) ***oriente*** y las agencias europeas responsables de la ayuda al desarrollo para aunar nuestros enfoques
8) yo creo que mantener la paz y la seguridad se cumple mejor en un sistema multilateral
9) en la unión europea el ***gobernanza*** de ***ese si un fenómeno canarias trabajado*** juntos para luchar contra el terrorismo internacional para luchar contra la pobreza mundial y para promover la democracia
10) sabemos cuáles son nuestros retos conjuntos así que vayamos a su ***cuando*** gracias

The **WER** of the system after processing this audio segment is 14.86%

As we can see in this example, the adapted language model contributes in reducing the Word Error Rate in the final ASR stage. The number of deletions in the output transcription of this stage has a relative reduction of 50% when compared with the first stage. In this same sense, the number of insertions ans substitutions have a relative reduction of 44% and 33.3% respectively.

In the final ASR stage, the system performs a dynamic adaptation of the language model and it allows an improvement of the performance of the recognition process.

6 | Conclusions

In this Thesis we have presented a framework for **topic-motivated contextualization** of automatic speech recognition. The contextualization on which we focused is based on the analysis and identification of semantic elements in speech, particularly, the topic that is being discussed in an utterance; and it was accomplished by making use of the topic-related information in the dynamic adaptation of the language models used by a speech recognition system.

Within this framework, we had to tackle with two areas of work: the area of *identification of topic*, which is in charge of document processing and analysis and extraction of semantic information from text documents (which can be either documents in the training set or speech transcriptions generated by automatic speech recognition in the evaluation set); and the area of *language model adaptation*, which is responsible for exploiting such semantic information into the generation of context-dependent models and the dynamic adaptation of language models for the speech recognizer.

According to these areas of work, and in order to be able to evaluate the performance of the proposed framework, we included into the architecture two principal systems (a *topic identification* and a *language model adaptation* systems), and for each of them we proposed a methodology that combines existing techniques along with our own contributions.

On the one hand, we developed a *topic identification* system which is based on the combination of different techniques from the fields of Information Retrieval and Machine Learning. On the other hand, we also developed a system for the *dynamic adaptation of language models* which is based on different interpolation schemes which were proposed and evaluated in this Thesis.

To integrate these technologies into the contextualization framework we developed an architecture based on two stages of recognition. This architecture allowed us to develop and evaluate each of the systems separately, and then assess their performance within the whole system.

Throughout this research we have conducted different experiments to evaluate the proposed systems; the results obtained in these experiments have led us to draw the conclusions that follow.

6.1 On Topic Identification

Our contributions for the topic identification task were focused on the enhancement of document preprocessing techniques and in the definition of more robust criteria for the selection of index-terms.

In this sense we proposed an *ad-hoc* weighting scheme, the *pseudo-entropy*, that tries to overcome the problems found when applying the *term entropy* weight. The proposed scheme tends to improve the performance of the system. The best results for the topic identification task were obtained by applying *term frequency* as a local weighting schemes along with *pseudo-entropy* as global; although it must be noticed that the reduction of the topic identification error we obtained by this combination of weights is not significant when compared to the *tf-idf* classic weighting scheme.

The selection of an adequate list of index-terms must not rely only on the definition of a stopword list. The selection must be dependent on the specific domain and must take account of the actual distribution of terms in the collection. In this regard, we evaluated different criteria for the index-terms selection. The results have shown that a reduced term inventory can be obtained by these criteria. For the *Evaluation Set 2* we obtained a significant reduction of the topic identification error when using the Vector Space Model along with the reduced term inventory that was obtained by the index-terms selection strategies.

When compared to the results obtained by the baseline approach, the best combination of parameters for the topic identification systems is obtained for the LSA model, using the term inventory that we have obtained by the idf index-terms selection technique and weighting the terms with the proposed *pseudo-entropy* scheme. This configuration presents a relative improvement of 23.32% when compared to the baseline approach for the *Evaluation Set 1* (although this improvement is not statistically significant) and a relative significant improvement of 45.58% for the *Evaluation Set 2*.

LSA offers a number of advantages when compared to the generalized Vector Space Model for document representation. It allows to reduce the dimensions of the representation space. In our experiments we reduced the space from 16250 dimensions to 67, which means a reduction of 99.58% of the original space. LSA also outperforms Vector Space Model in the topic identification task for the *Evaluation Set 2*.

Despite the *stemming* procedure reduces the size of the term inventory it does not provide a reduction in the topic identification error for both evaluation sets. We believe that this may be caused because of the loss of semantic information when reducing words to their stems. Thus the relationships between terms and documents may be distorted for this approximation. By stemming we could be removing semantic information that might be useful for the topic identification objective.

To the best of our knowledge this work has been the first to tackle the topic identification problem in the Spanish partition of the EPPS database. We have found a number of factors that can influence the difficulty of the topic identification task. Among these factors, some are related to particular conditions of the database.

Multiple topics in a single domain. Although the database contains multiple topics, they are all related to a single domain, *politics*. This is a clear difference in comparison with tasks in which databases from multiple domains are analyzed (e.g., sports, culture, science, politics, etc.). The high similarity of the topics within the EPPS database hinders the topic identification process.

Length of the audio segment. The length of the audio segment to analyze has a direct influence on the performance of the system. Larger audio segments, which as transcripts can be seen as larger text documents, involve, at least potentially, more useful information to process. However, our results suggests that there is a trade-off between the length of the audio segment and the capability of the system to improve recognition performance which is a different but central objective for us.

Recognition errors. The topic identification system processes the transcripts that are delivered by the first stage of the recognizer, which means that it processes transcripts containing recognition errors. This clearly hinders the identification process because words that might be important to identify a topic, or a concept within a topic, could not be recognized and therefore not appear in the transcript.

Some of our contributions in this Thesis are precisely intended to tackle some of these particular conditions.

We have proposed an unsupervised strategy for automatically clustering the documents in the collection with the aim of increasing the conceptual similarity of documents within the same cluster and generate topic-based LMs from these topic clusters. As a consequence of this clustering strategy, the number of topics, and therefore, the number of topic-based LMs is reduced.

Regarding the length of the audio segment, we have made a general and exploratory analysis on how different lengths of the audio segment can influence both the topic identification process and the performance of the speech recognition system. And finally, we have developed a robust topic identification system that, despite the recognition errors which are an inherent part in today's speech recognition systems, is capable of identifying, reliably, the topics in audio segments.

6.2 On Document Clustering

Our contributions on automatic document clustering were focused not only on on the simplification of the system's models by reducing the total number of parameters involved in the system but also on the comparison of different strategies for the generation of topic-based language models and their effect on the performance of the speech recognition system.

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,

thus improving the coverage of the topic-based language models generated by this strategy.

We evaluated two different techniques for document clustering, *k*-means and LDA. There are not significant differences in the performance of the ASR system when comparing both clustering techniques, therefore no conclusions can be drawn regarding which of these clustering techniques performs better when compared to each other in our case.

The application of automatic document clustering into our methodology for language model adaptation allows us to reduce the number of parameters that are involved in the system. Particularly, this strategy allows a reduction in the number of topic-based language models when compared to the supervised approach in which there are as many topic-based LMs as there are topics in the collection. By means of the unsupervised document clustering techniques the number of topic clusters, and therefore of topic-based LMs, is reduced to 17 and 20, depending on the use of the *k*-means or LDA approaches, respectively.

Among all the speech recognition experiments, the absolute best results, in terms of relative reduction of WER, were obtained by applying the automatic document clustering strategy for the generation of topic-based LMs. Among the techniques we evaluated within this strategy, the best results were obtained by clustering the documents with *k*-means, although as we previously mentioned, there are not significant differences when compared to LDA. For the best configuration we obtained a relative reduction of WER of 13.52% for the *Evaluation Set 2* (Table 5.2).

The *overall average* Silhouette Coefficient (SC) criterion allows us to find a number of clusters, for which the distribution of documents minimizes their distance within the same cluster and at the same time maximizes the distance between documents in different clusters (as it was described in Section 4.1.3). We have conducted different experiments by setting manually the number of clusters around the number suggested by the SC criterion and similar results have been obtained without a significant difference among them. These results suggest that this automatic selection criterion of the number of clusters is effective in the sense that reduces the error (although not significantly), but from a speech recognition performance perspective, we can not conclude that it is optimal, since similar results can be obtained with a different number of clusters.

6.3 On Language Model Adaptation

Contributions presented in this thesis were focused on studying the capacity of the proposed system to dynamically adapt the language model used by a speech recognizer according to the changes experienced by the grammar when dealing with spontaneous and multitopic domains.

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.

In this regard, a set of experiments were conducted to evaluate the performance of the dynamic LM adaptation techniques presented in this Thesis. According to the results, the interpolation schemes presented as part of our dynamic language model interpolation strategies (Section 5.2.2) tend to improve the performance of an ASR system within a multipass architecture.

The results in the ASR task have shown that a 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.

Results have also shown that the performance of the ASR is enhanced when adapting LMs to shorter audio segments (audio segments in *Evaluation Set 2*) and a significant reduction of word error rate can be achieved when compared to larger audio segments (*Evaluation Set 1*). 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.

Regarding the interpolation schemes, we have shown that adapting the LM by taking only into consideration the closest 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 to the dynamic adaptation of the language models.

In this sense we believe that our methodology, and the results obtained in its evaluation, are promising considering that the performance of the speech recognition system tends to improve for each of the interpolation schemes we evaluated, particularly the scheme Soft-2. We believe that the application of this methodology in a different domain could provide similar or even significantly better results.

In the final ASR stage we perform a second decoding of the audio segment using the final adapted language model. We are aware that in order to implement our methodology in a real application, an alternative and more adequate approach could be to perform a lattice rescoring stage instead of fully re-decoding the audio segment. This approach will be taken into account for future improvements of the proposed architecture.

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. In this sense, our results seem to indicate that there is a trade-off between the length of the audio segment and the error of both topic identification and speech recognition systems. On the one hand, long segments of audio (in our case a long segment means that the segment is significantly greater than one minute) lead to smaller topic identification error when

compared for short segments, however, the best recognition results were obtained for short segments.

7 | Future work

The work in this Thesis can be extended in several ways. Here are some of the research guidelines that we can follow regarding the main areas of our work.

The database we have used for evaluating this work contains documents belonging to the same domain: political speeches. The methodology we propose can be evaluated in different domains. We believe that the results of our experimental framework as well as the conclusions we achieved regarding these results, show that the application of this methodology may be promising in domains for which there are much more remarkable and discernible semantic differences.

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. We are aware that this may increase the size of the term inventory, but it also could improve the capability of terms to describe the topic of documents. 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.

From an application point of view, it would be useful to determine if the proposed global weighting scheme (*pseudo-entropy*) is useful in other domains.

There may be multiple benefits derived from the implementation of this scheme in different areas. Not only document clustering methods and document classification techniques are those most likely to benefit from new term weighting schemes, but also methods of feature extraction and measures of information content can also take advantage of it.

Although we have selected a close domain, such as the political domain, to evaluate the topic-motivated contextualization framework for speech recognition systems that we propose in this Thesis, the methodology presented here can also be extended to different application domains. Indeed, we believe that domains involving a wider semantic variety of topics would be appropriate to apply the proposed methodology. Such domains would involve, for instance, broadcast news automatic transcription systems, speech retrieval in audio corpus and dialogue managers in open-context applications, among many others.

A comparison could be made between the original transcriptions of the audio seg-

ments and the output of the first ASR stage. By means of comparing both transcriptions (e.g. on the development dataset) we could estimate a model that relates the desired topic and the recognition errors produced by the first recognition stage. This model could represent, for example, the missing index-terms in the output transcription for a certain topic, and could lead us, somehow, to compensate the recognition errors in the evaluation dataset in order to enhance the topic identification process of the system.

One of the aspects that can be analyzed to improve the effectiveness of the proposed system in a real application, is that the system should be 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, the first stage may provide a lower word error rate than the second stage, for some of the utterances. For that reason, it could be studied to include confidence measures for speech recognition in order to evaluate the reliability of recognition results in both stages and select the best output.

In the Soft-1 approach for LM interpolation as well in the Soft-2 approach, we evaluated the generation of the context dependent model by considering the top-10 most similar topics. There was not a specific reason for considering exactly this number of topics. This was the result of an *ad-hoc* criterion that we specify for the evaluation. In future work, we would like to explore automatic decision criteria for selecting the number of topic-based LMs to be considered.

We performed a second decoding in the final ASR stage using the adapted language model. To implement our methodology in a real application, an adequate alternative is to perform a lattice rescoring stage instead of fully re-decoding the audio segment. This approach will be taken into account for future improvements of the proposed architecture.

The length of the audio segments has a non trivial but direct influence in the performance of the system. As we already mentioned there is a trade-off between the length of the audio segment and the error of both topic identification and speech recognition systems. A further analysis of this trade-off and a more detailed study on the segmentation of audio segments are required to draw deeper conclusions in this regard.

8 | Publications

Next there is a list of the publications that are directly related with the Thesis' objectives and are a result of the work developed in this Thesis.

Journal Article

J.D. Echeverry-Correa, J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba, J.M. Montero, "Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition". Expert Systems with Applications, Vol. 42, pp. 101-112, January 2015, doi: 10.1016/j.eswa.2014.07.035

International Conferences

J.D. Echeverry-Correa, A. Coucheiro-Limeres and J. Ferreiros-López, "GTH-UPM System for Search on Speech Evaluation", Albayzin Evaluation Special Session, Proceedings of the Iberspeech 2014, pp. 299-305, november 2014, Las Palmas de Gran Canaria, Spain.

J.D. Echeverry-Correa, B. Martínez-González, R. San-Segundo, R. Córdoba and J. Ferreiros-López, "Dynamic Topic-Based Adaptation of Language Models: A Comparison Between Different Approaches". Proceedings of the Iberspeech 2014, pp. 139-148, november 2014, Las Palmas de Gran Canaria, Spain.

Chapter in book

J.D. Echeverry-Correa as author of the chapter: "Audio-Speech segmentation and Topic Detection for a Speech-based Information Retrieval System" in book "Applications of Speech Technologies: Talks and Contributions presented at the summer course: Applications of Speech Technologies". M.C. Benítez-Ortúzar, J.L. Pérez-Córdoba (eds.). Ed. Universidad de Granada, ISBN 978-84-338-5596-1, 2013, pp. 279-291.

Other publications

Next, there is a list of other publications which also involve, in some extent, the work we have developed in this thesis

J. Tejedor, D. Toledano, P. López, L. Docio, C. García, A. Cardenal, **J.D. Echeverry-Correa**, A. Coucheiro, J. Olcoz and A. Miguel "Spoken Term Detection ALBAYZIN 2014 evaluation: overview, systems, results and discussion", EURASIP Journal of Audio, Speech and Music processing. 2015. *To be published.*

J.M. Lucas-Cuesta, J. Ferreiros, F. Fernández-Martínez, **J.D. Echeverry-Correa** and S. Lebai Lutfi, "On the Dynamic Adaptation of Language Models based on Dialogue Information", Expert Systems With Applications, Vol. 40, Issue 4, march 2013, pp. 1069-1085.

B. Martínez-González, J.M. Pardo, **J.D. Echeverry-Correa**, J.M. Montero, "New experiments on speaker diarization for unsupervised speaking style voice building for speech synthesis", Procesamiento del Lenguaje Natural, Vol. 52, pp. 77-84, ISSN: 1989-7553, march 2014.

L.F. D'Haro, R. Córdoba, C. Salamea and **J.D. Echeverry-Correa**. "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition". Proceedings of the IEEE ICASSP 2014, Florence, Italy.

J. Lorenzo-Trueba, **J.D. Echeverry-Correa**, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, A. Gallardo-Antolín, J. Yamagishi, S. King and J. M. Montero, "Development of a Genre-Dependent TTS System with Cross-Speaker Speaking-Style Transplantation". ISCA/IEEE Proceedings of the 2nd International Workshop on Speech, Language and Audio in Multimedia (SLAM 2014), Penang, Malaysia, 2014.

B. Martínez-González, J.M. Pardo, **J.D. Echeverry-Correa**, J.A. Vallejo-Pinto, R. Barra-Chicote, "Selection of TDOA Parameters for MDM Speaker Diarization", InterSpeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13, 2012, pp. 2158-2161.

V. López-Ludeña, R. San-Segundo, S. Lutfi, J.M. Lucas-Cuesta, **J.D. Echeverry-Correa**, B. Martínez-González, "Source Language Categorization for improving a Speech into Sign Language Translation System" in SLPAT 2011 Workshop on Speech and Language Processing for Assistive Technologies, Edinburgh, UK July 30, 2011, pp. 84-93.

V. López, R. San-Segundo, R. Martín, J.M. Lucas, **J.D. Echeverry-Correa** "Spanish generation from Spanish Sign Language using a phrase-based translation system", FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop, Vigo, Spain, 10-12 November 2010.

V. López, R. San-Segundo, R. Martín, **J.D. Echeverry-Correa**, S. Lutfi "Sistema de traducción de lenguaje SMS a castellano", XX Jornadas Telecom I+D, Valladolid, 27-29 September 2010.

F. Fernández-Martínez, J. Ferreiros, J.M. Lucas-Cuesta, **J.D. Echeverry-Correa**, R. San-Segundo, R. de Córdoba, "Flexible, Robust and Dynamic Dialogue Modeling with a Speech Dialogue Interface for Controlling a Hi-Fi Audio System", Proceedings of the IEEE Workshop on Database and Expert Systems Applications (DEXA 2010), isbn 978-3-642-03572-2, issn 1529-4188, Bilbao Spain, 1-3 September 2010, pp. 250-254.

References

A. Abad, L. J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, and G. Bordel. On the calibration and fusion of heterogeneous spoken term detection systems. In *14th International Conference on Speech and Language Technology (INTERSPEECH'13)*, pages 20–24, 2013.

A. Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39:45–65, 2003.

M. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in neural information processing systems*, pages 28–36, 2009.

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2009.

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

N. Bel, C. H. Koster, and M. Villegas. Cross-lingual text categorization. In *Research and Advanced Technology for Digital Libraries*, pages 126–139. Springer, 2003.

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

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

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

J. R. Bellegarda, J. W. Butzberger, Y.-L. Chow, N. B. Coccaro, and D. Naik. A novel word clustering algorithm based on latent semantic analysis. In *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*, volume 1, pages 172–175. IEEE, 1996.

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

C. Boulis and M. Ostendorf. Text classification by augmenting the bag-of-words representation with redundancy compensated bigrams. In *Proc. of the International Workshop in Feature Selection in Data Mining*, pages 9–16. Citeseer, 2005.

M. Brown, J. Foote, G. Jones, K. S. Jones, and S. Young. Video mail retrieval by voice: An overview of the cambridge/olivetti retrieval system, 1994.

H. Bunke, M. Roth, and E. Schukat-Talamazzini. Off-line cursive handwriting recognition using hidden markov models. *Pattern Recognition*, 28(9):1399 – 1413, 1995.

K. Chandrinos, I. Androutsopoulos, G. Paliouras, and C. Spyropoulos. Automatic web rating: Filtering obscene content on the web. In J. Borbinha and T. Baker, editors, *Research and Advanced Technology for Digital Libraries*, volume 1923 of *Lecture Notes in Computer Science*, pages 403–406. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-41023-2.

L. Chen, J. Gauvain, L. Lamel, G. Adda, and M. Adda. Using information retrieval methods for language model adaptation. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, pages 255–258, 2001a.

L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda. Language model adaptation for broadcast news transcription. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, 2001b.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

S. F. Chen, K. Seymore, and R. Rosenfeld. 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)*, volume 2, pages 681–684, 1998.

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

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

H. Chiu and B. Chen. 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)*, volume 4, pages 169–172, 2007.

J. Chu-Carroll. Mimic: An adaptive mixed initiative spoken dialogue system for information queries. In *Proceedings of the sixth conference on Applied natural language processing*, pages 97–104. Association for Computational Linguistics, 2000.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

C. Cieri, D. Miller, and K. Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *Proc. of Intl. Conf. on Language Resources and Evaluation*, 2004.

P. R. Clarkson. *Adaptation of statistical language models for automatic speech recognition*. PhD thesis, University of Cambridge, 1999.

R. Cummins. *The evolution and analysis of term-weighting schemes in information retrieval*. PhD thesis, National University of Ireland, 2008.

H. Daumé, III, A. Kumar, and A. Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 53–59. Association for Computational Linguistics, 2010.

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

S. F. Dennis. The construction of a thesaurus automatically from a sample of text. In *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation, Washington, DC*, pages 61–148, 1965.

D. Dey, T. Solorio, M. Gómez, and H. Escalante. Instance selection in text classification using the silhouette coefficient measure. In *Proceedings of the 10th Mexican International Conference on Artificial Intelligence (MICAI'11)*, pages 357–369, 2011.

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

J. Echeverry-Correa, A. Coucheiro-Limeres, and J. F. López. GTH-UPM System for Search on Speech Evaluation. In *Proceedings of the Iberspeech 2014*, pages 299 – 305, November 2014.

G. Escudero, L. Màrquez, and G. Rigau. Boosting Applied to Word Sense Disambiguation. In *Proceedings of the 11th European Conference on Machine Learning*, ECML '00, pages 129–141. Springer-Verlag, 2000.

R. M. Fano and W. Wintringham. Transmission of information. *Physics Today*, 14:56, 1961.

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

F. Fernández, J. Ferreiros, V. Sama, J. M. Montero, R. San-Segundo, and J. Macıas-Guarasa. Speech interface for controlling an hi-fi audio system based on a bayesian belief networks approach for dialog modeling. In *6th International Conference on Speech and Language Technology (INTERSPEECH'05)*, pages 3421–3424, 2005.

A. Fernández-Anta, L. Núnez-Chiroque, P. Morere, and A. Santos. Sentiment Analysis and Topic Detection of Spanish Tweets: a comparative study of of NLP techniques. *Procesamiento del Lenguaje Natural*, 50:45–52, 2013.

F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and W. Meira Jr. Word co-occurrence features for text classification. *Information Systems*, 36(5):843–858, 2011.

R. Florian and D. Yarowsky. Dynamic nonlocal language modeling via hierarchical topic-based adaptation. In *In Proceedings of the ACL*, pages 167–174, 1999.

U. Glavitsch. A First Approach to Speech Retrieval. Technical report, Department of Computer Science - Swiss Federal Institute of Technology, 1995.

C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney. 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)*, pages 825–828, 2005.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

P. Goyal, L. Behera, and T. M. McGinnity. A context-based word indexing model for document summarization. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1693–1705, 2013.

H. Guan, J. Zhou, and M. Guo. A class-feature-centroid classifier for text categorization. In *Proceedings of the 18th international conference on World wide web*, pages 201–210. ACM, 2009.

T. S. Guzella and W. M. Caminhas. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206 – 10222, 2009.

S.-H. Hahn, J. H. Lee, and J.-H. Kim. A study on utilizing ocr technology in building text database. In *Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on*, pages 582–586. IEEE, 1999.

M. Haidar and D. O'Shaughnessy. Topic n-gram count language model adaptation for speech recognition. In *Proceedings of the Spoken Language Technology Workshop (SLT)*, pages 165–169, Miami, FL, 2012.

E.-H. S. Han and G. Karypis. Centroid-based document classification: Analysis and experimental results. In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 424–431, 2000.

D. W. Harman. An experimental study of factors important in document ranking. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 186–193. ACM, 1986.

A. Hauptmann. Automatic spoken document retrieval. Technical report, Carnegie Mellon University, 2006.

T. Hazen, F. Richardson, and A. Margolis. Topic identification from audio recordings using word and phone recognition lattices. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2008.

M. Hoffman, F. R. Bach, and D. M. Blei. Online Learning for Latent Dirichlet Allocation. In *Proceedings of Advances in Neural Information Processing Systems, NIPS'10*, pages 856–864, 2010.

T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

V. Hollink, J. Kamps, C. Monz, and M. De Rijke. Monolingual document retrieval for european languages. *Information retrieval*, 7(1-2):33–52, 2004.

D. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–291, 1994.

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

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss. A dynamic language model for speech recognition. In *Proceedings of Speech and Natural Language DARPA Workshop*, pages 293–295, 1991.

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

D. Jurafsky and J. Martin. *Speech and Language Processing*. Prentice Hall, 2006.

A. Kilgarriff and J. Rosenzweig. English senseval: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC*, Athens, Greece, 2000.

W. Kim. *Language model adaptation for automatic speech recognition and statistical machine translation*. PhD thesis, The Johns Hopkins University, 2004.

W. Kim and S. Khudanpur. Cross-lingual latent semantic analysis for language modeling. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, volume 1, pages 257–260, 2004.

Y. Kim and S. Ross. Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, C. Thanos, M. Verdejo, and R. Carrasco, editors, *Research and Advanced Technology for Digital Libraries*, volume 4172 of *Lecture Notes in Computer Science*, pages 63–74. Springer Berlin Heidelberg, 2006.

R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, volume 1, pages 181–184. IEEE, 1995.

P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Conference on Machine Translation (MT Summit'05)*, 2005.

R. Kuhn and R. De Mori. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6): 570–583, 1990.

R. Kuhn and R. De Mori. Corrections to a cache-based language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14 (6):691–692, 1992.

L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, pages 4–15. Springer, 2008.

T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: A maximum entropy approach. In *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'93)*, volume 2, pages 45–48. IEEE, 1993.

D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *Proceedings of the 1994 Symposium on Document Analysis and Information Retrieval*, pages 81–93, 1994.

D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50. ACM, 1992a.

D. D. Lewis. *Representation and learning in information retrieval*. PhD thesis, University of Massachusetts, 1992b.

D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15, Germany, 1998.

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

X. Liu, M. Gales, and P. Woodland. Use of contexts in language model interpolation and adaptation. *Computer Speech & Language*, 27(1):301 – 321, 2013a. Special issue on Paralinguistics in Naturalistic Speech and Language.

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

Y. Liu and F. Liu. 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)*, pages 4921–4924, 2008.

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

S. Lu, W. Wei, X. Fu, and B. Xu. 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)*, pages 512–522, 2012.

J. M. Lucas-Cuesta. *Contributions to the Contextualization of Human-Machine Spoken Interaction Systems*. PhD thesis, Department of Electronic Engineering, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid, 2013.

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

H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

A. Mandal, J. van Hout, Y.-C. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri, L. Ferrer, M. Graciarena, A. Kathol, and H. Franco. Strategies for high accuracy keyword detection in noisy channels. In *14th International Conference on Speech and Language Technology (INTERSPEECH'13)*, pages 15–19, 2013.

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244, 1960.

A. K. McCallum. Mallet: A machine learning for language toolkit, 2002. URL http://mallet.cs.umass.edu.

M. McGill. An evaluation of factors affecting document ranking by information retrieval systems. Technical report, Report from the School of Information Studies, Syracuse University, New York., 1979.

M. F. McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90–169, 2002.

J. R. Méndez, E. L. Iglesias, F. Fdez-Riverola, F. Díaz, and J. M. Corchado. Tokenising, stemming and stopword removal on anti-spam filtering domain. In *Current Topics in Artificial Intelligence*, pages 449–458. Springer, 2006.

D. Mladenic and M. Grobelnik. Word sequences as features in text-learning. In *In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98*, pages 145–148, 1998.

A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In *Advances in Information Retrieval*, pages 181–196. Springer, 2004.

D. Mostefa, O. Hamon, N. Moreau, and K. Choukri. Evaluation Report for the Technology and Corpora for Speech to Speech Translation (TC-STAR Project). deliverable n. 30, 2007.

K. Myers, M. J. Kearns, S. P. Singh, and M. A. Walker. A boosting approach to topic spotting on subdialogues. In *ICML*, pages 655–662, 2000.

H. Nanjo and T. Kawahara. Unsupervised language model adaptation for lecture speech recognition. In *Proceedings of the 2003 ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR'03)*, 2003.

I. Oparin. *Language Models for Automatic Speech Recognition of Inflectional Languages*. PhD thesis, University of West Bohemia, 2008.

L. Padró and E. Stanilovsky. Freeling 3.0: Towards Wider Multilinguality. In *Proceedings of the 2012 Language Resources and Evaluation Conference (LREC'12)*, 2012.

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM, 1998.

S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM, 2004.

S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.

R. Rosenfeld. *Adaptive statistical language modeling: A maximum entropy approach*. PhD thesis, Carnegie Mellon University, 1994.

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

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

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1983.

G. Salton and C.-S. Yang. On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372, 1973.

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

R. San-Segundo, J. M. Montero, J. Macías-Guarasa, J. Ferreiros, and J. M. Pardo. Knowledge-combining methodology for dialogue design in spoken language systems. *International Journal of Speech Technology*, 8(1):45–66, 2005.

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

H. Schütze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 229–237. ACM, 1995.

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

G. Senay, B. Bigot, R. Dufour, G. Linarès, and C. Fredouille. Person name spotting by combining acoustic matching and lda topic models. In *14th International Conference on Speech and Language Technology (INTERSPEECH'13)*, pages 1584–1588, 2013.

K. Seymore and R. Rosenfeld. Using story topics for language model adaptation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, 1997.

K. Shin, A. Abraham, and S. Han. Enhanced centroid-based classification technique by filtering outliers. In *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 159–163. Springer, 2006.

C. Silva and B. Ribeiro. *Inductive Inference for Large Scale Text Classification: Kernel Approaches and Techniques*, volume 255 of *Studies in Computational Intelligence*. Springer, 2010.

A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29. ACM, 1996.

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

K. Spärck-Jones. Index term weighting. *Information storage and retrieval*, 9(11): 619–633, 1973.

A. Stolcke. SRILM-An extensible Language Modeling Toolkit. In *3rd International Conference on Speech and Language Technology (INTERSPEECH'02)*, 2002. URL http://www.speech.sri.com/projects/srilm/.

T. Strzalkowski, J. P. Carballo, J. Karlgren, A. Hulth, P. Tapanainen, and T. Lahtinen. Natural language information retrieval: Trec-8 report. In *TREC*, 1999.

Y. Tam and T. Schultz. 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)*, pages 4821–4824, 2009.

C. Tillmann and H. Ney. Selection criteria for word trigger pairs in language modeling. In *Grammatical Interference: Learning Syntax from Sentences*, pages 95–106. Springer, 1996.

C. Troncoso and T. Kawahara. Trigger-based language model adaptation for automatic meeting transcription. In *6th International Conference on Speech and Language Technology (INTERSPEECH'05)*, pages 1297–1300, 2005.

G. Tur and A. Stolcke. Unsupervised language model adaptation for meeting recognition. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 4, pages IV–173. IEEE, 2007.

Upasana and S. Chakravarty. A survey on text classification techniques for e-mail filtering. In *Proceedings of the 2nd International Conference on machine Learning and Computing*, 2010.

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

W. Wang and A. Stolcke. Integrating MAP, marginals, and unsupervised language model adaptation. In *8th International Conference on Speech and Language Technology (INTERSPEECH'07)*, pages 618–621. Citeseer, 2007.

J. Wintrode. Leveraging locality for topic identification of conversational speech. In *14th International Conference on Speech and Language Technology (INTERSPEECH'13)*, pages 1579–1583, 2013.

J. Wintrode and S. Kulp. Techniques for rapid and robust topic identification of conversational telephone speech. In *10th International Conference on Speech and Language Technology (INTERSPEECH'09)*, pages 1471–1474, 2009.

I. H. Witten and T. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.

J. Wu. *Maximum entropy language modeling with non-local dependencies*. PhD thesis, Johns Hopkins University, 2002.

Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49. ACM, 1999.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al. *The HTK book*. Cambridge University Engineering Department, 2006. URL http://htk.eng.cam.ac.uk/.

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

W. Zhang, T. Yoshida, and X. Tang. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765, 2011.

Y. Zhang. *Structured language models for statistical machine translation*. PhD thesis, Carnegie Mellon University, 2009.

Este libro terminó de diagramarse en diciembre del 2016, en la Oficina de Recursos Informaticos y Educativos CRIE de la Universidad Tecnológica de Pereira, bajo el cuidado del autor.
Pereira, Risaralda, Colombia.

La Editorial de la Universidad Tecnológica de Pereira tiene como política la divulgación del saber científico, técnico y humanístico para fomentar la cultura escrita a través de libros y revistas científicas especializadas.

Las colecciones de este proyecto son: Trabajos de investigación, Ensayos, Textos Académicos y Tesis Laureadas.

Este libro pertenece a la Colección Tesis Laureadas.

Los últimos tiempos han sido testigos de importantes avances en el campo de la tecnología de reconocimiento de voz. Sin embargo, y a pesar del buen momento que vive esta tecnología, hay que reconocer que esta tarea dista de ser un problema resuelto. La mayoría de sistemas de reconocimiento automático de voz se ajustan a dominios particulares, y su eficacia depende de manera significativa, entre otros muchos aspectos, de la similitud existente entre el modelo de lenguaje y la tarea específica para la cual se está empleando. Esta dependencia cobra aún más importancia en aquellos escenarios en los cuales las propiedades estadísticas del lenguaje varían a lo largo del tiempo, como por ejemplo, en dominios de aplicación que involucren habla espontánea y múltiples temáticas.

En los últimos años se ha evidenciado un constante esfuerzo por mejorar los sistemas de reconocimiento para tales dominios. Esto se ha hecho, entre otros muchos enfoques, a través de técnicas automáticas de adaptación. Estas técnicas requieren fuentes adicionales de información y en este sentido, el lenguaje hablado puede aportar algunas de ellas. El habla no sólo transmite un mensaje, también transmite información acerca del contexto en el cual se desarrolla la comunicación hablada.

La principal contribución de este trabajo consiste en la propuesta y evaluación de un marco de contextualización motivado por el análisis temático y basado en la adaptación dinámica y no supervisada de modelos de lenguaje para el robustecimiento de un sistema automático de reconocimiento de voz. Esta adaptación toma como base distintos enfoques basados en sistemas de recuperación de información y de aprendizaje de máquina.

ISBN: 978-958-722-267-8

Universidad Tecnológica de Pereira