Un Método para Recuperación de Información en Documentos Orales basado en Codificación Fonética

por

Manuel Alejandro Reyes Barragán

Tesis sometida como requisito parcial para obtener el grado de

MAESTRO EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica
Agosto 2008
Tonantzintla, Puebla

Supervisada por:

Dr. Luis Villaseñor Pineda
Investigador Titular del INAOE

Dr. Manuel Montes y Gómez
Investigador Titular del INAOE

Resumen

La cantidad de información disponible gracias a los avances de la tecnología se ha incrementado enormemente en los últimos años. Los medios y formatos en que esta información está conservada, puede ser de muy variada naturaleza. Actualmente existen enormes colecciones de textos, de imágenes, de audio, de video, etc. Sin embargo, estas colecciones no son útiles si no llegamos a organizarlas para identificar aquellos elementos pertinentes a una determinada necesidad de información. Es justamente este problema el que aborda la Recuperación de Información. Por supuesto, los métodos de recuperación varían dependiendo de la naturaleza de la colección. En este trabajo de tesis nos enfocamos a la recuperación de información en documentos orales. Por documentos orales nos referimos a grabaciones de habla tales como discursos políticos, conferencias, noticieros radiofónicos, etc.

El enfoque de esta tesis para abordar el problema parte de las transcripciones automáticas de esas grabaciones. Desafortunadamente, las transcripciones generadas por un reconocedor automático del habla no son perfectas, de tal forma que es común encontrar errores de inserción, eliminación o substitución de palabras en las transcripciones automáticas. Esta situación tiene por resultado documentos con características propias, diferentes a las que se encuentran en texto escrito manualmente. De ahí, que los métodos tradicionales de recuperación de información no funcionen adecuadamente y sea necesario definir métodos enfocados al tratamiento de este tipo de documentos.

En esta tesis se propone un método original basado en la codificación fonética de las transcripciones automáticas. La idea del método consiste en enriquecer la representación de los documentos para tratar de abordar los errores inducidos por el reconocedor. Gracias a la codificación fonética es posible representar con un mismo código palabras cuya pronunciación es similar. El método utilizado fue evaluado utilizando el mismo conjunto de datos proporcionado por el foro de evaluación del CLEF CL-SR, lo cuál nos permite comparar objetivamente el desempeño de nuestro sistema. En comparación con los resultados de otros equipos en este foro, nuestro sistema se ubicó en el segundo lugar, demostrando que el método es adecuado para esta tarea.

Abstract

The amount of information available thanks to the advances on the technology has been increased in the last years. The storage media and formats can be of varied nature. The amount of available multimedia repositories is increasing; there are collections of text, images, audio and video. However these collections are not useful if we can't organize them to identify the pertinent elements to a necessity of information. It's exactly this problem the one that deal the Information Retrieval. Of course, the information retrieval methods vary depending of the collection. In this thesis we focus on the task of Spoken Document Retrieval. By the term Spoken Documents we mean to speech recordings as political speeches, conferences, news.

The approach of this thesis is to deal with the problem generated by the automatic transcriptions of speech recordings. Unfortunately, the transcriptions generated by the automatic speech recognition are not perfect, so is common to find several transcription errors (such as word substitutions, insertions and deletions). The result of this situation are documents with own characteristics, different to those find in text written manually. Therefrom, that the traditional methods of information retrieval don't work appropriately, and become necessary to define methods focused to the treatment of this type of documents.

In this thesis we propose an original method based on the phonetic codification of the automatic speech transcriptions. The idea of the method consists on enriching the representation of the documents to deal with the

errors generated by the automatic speech recognition. Thanks to the phonetic codification it's possible represent with the same code words whose pronunciation is similar. The method was evaluated using the same set of data proportionate for the evaluation forum CLEF CL-SR. These allow us to compare the performance of our system objectively. In comparison with the rest of the teams that participate in the forum, our system was located in the second place, demonstrating that the method is appropriate for this task.

Agradecimientos

Se agradece al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca para estudios de maestría no. 212715.

Un muy especial agradecimiento a todos los integrantes del Laboratorio de Tecnologías del Lenguaje del INAOE. En especial al Dr. Luis Villaseñor Pineda y el Dr. Manuel Montes y Gómez, quienes dirigieron mi tesis y gracias a su conocimiento, experiencia, paciencia y buen humor lograron llevar a buen término el trabajo. Gracias por hacer esto posible.

Dedicatoria

Para mi mama y abuelitos.

Para Liliana.
Por su cariño, amor y comprensión.

Para Alberto.
Por apoyarme desde que llegue y hasta la fecha.

Para mis amigos Antonio, Barbie, Esaú, Milton, Noemí, Pato.
Por hacerme reir, distraerme y ayudarme cuando lo necesité.

Índice General

Lista de Figuras

Lista de Tablas

Capítulo I

Introducción

La cantidad de información disponible gracias a los avances de la tecnología se ha incrementado enormemente en los últimos años. Cada vez un mayor número de personas participa en la creación de nuevos documentos digitales y gracias a Internet estos documentos son puestos a nuestra disposición. Los medios y formatos en que esta información está conservada, puede ser muy variada naturaleza. Por ejemplo, actualmente existen enormes colecciones de textos, de imágenes, de audio, de video, etc. Sin embargo, dados los tamaños de esas colecciones buscar una pieza de información se vuelve una difícil tarea. De tal manera, que estas colecciones no son útiles si no llegamos a organizarlas para identificar aquellos elementos pertinentes a una determinada necesidad de información. El problema de organizar y acceder a la información contenida en una colección es abordado por la Recuperación de Información (RI). Esta área de investigación propone métodos para organizar y buscar información en grandes colecciones no estructuradas de documentos digitales.

El objetivo final de la RI es resolver una necesidad de información. De tal forma que dada una petición de un usuario se entreguen un conjunto de documentos que presumiblemente satisfagan la necesidad expresada en esa petición. En el transcurso de los años, diferentes métodos de organizar y acceder a los documentos han buscado solucionar el problema de la recuperación de información en forma rápida, eficiente y accesible a cualquier persona. La primera discusión sobre cómo acceder en forma automática a una gran cantidad de información apareció en un artículo llamado “As We May Think”, publicado por Vanner Bush en 1945. Esta idea se trató de poner en práctica

cinco años después y desde entonces han surgido diversos enfoques en el área de RI.

Como se mencionó anteriormente, la tecnología actual ha permitido generar documentos electrónicos de forma sencilla y a un costo accesible. Esto ha impactado particularmente en la creación y almacenamiento de documentos multimedia: documentos conteniendo audio, imágenes y/o video. En consecuencia las técnicas tradicionales de la recuperación de información se han adaptado y/o extendido para tratar con este tipo de documentos. En esta tesis nos enfocamos a la recuperación de información utilizando la sección oral de estos documentos multimedia, donde una o varias personas se expresan oralmente. Para efectos prácticos, en esta tesis sólo consideraremos grabaciones de habla y no toda la información multimedia presente en este tipo de documentos. Ejemplos del tipo de documentos al que se aboca este trabajo de tesis son: emisiones de noticias, discursos políticos, debates, conferencias, reuniones de negocios, entrevistas, llamadas telefónicas solicitando información o soporte técnico, etc.

Para realizar la recuperación de información en documentos orales es necesario realizar una transcripción a texto y efectuar la búsqueda sobre dicha transcripción. Existen al menos dos opciones para efectuar esto, uno de ellos es la transcripción manual, es decir, una persona escribe el contenido de la grabación. El resultado es una transcripción de muy alta calidad, sin embargo es un proceso costoso que requiere una gran cantidad de tiempo y recursos. Otra forma de realizar una transcripción es por medios automáticos a través de un programa de reconocimiento automático del habla (RAH).

El presente trabajo parte de las transcripciones automáticas de los documentos orales. Es decir, se da por hecho que un proceso de reconocimiento de habla ha sido aplicado a las grabaciones resultando en una representación textual del

contenido de dichas grabaciones. Será a partir de estas transcripciones automáticas que se iniciará el proceso de recuperación de información.

Sin embargo, los programas de RAH distan de ser perfectos y la calidad de la transcripción impacta la recuperación. Los primeros reconocedores automáticos de habla contaban con un error a nivel palabra (WER) muy alto, este error puede variar según la calidad del audio y el tipo de contenido. Cuando se reporta un error a nivel palabra del 50% significa que la mitad de las palabras contenidas en la transcripción han sido mal reconocidas, con el consecuente mal rendimiento de la recuperación de información. Actualmente los reconocedores tienen errores entre el 45% y el 20% o menor, dependiendo de la grabación y el ruido de fondo.

1.1 Motivación

Actualmente la cantidad de datos almacenados aumenta día a día, incluyendo los llamados documentos multimedia. Entre los documentos multimedia encontramos: emisiones de noticias por radio y televisión, discursos políticos, debates, conferencias, reuniones de negocios, entrevistas en estudios a celebridades, testigos, personas afectadas en accidentes, lecciones en un salón de clases, etc.

Toda esta información, almacenada gracias al avance de la tecnología, es de gran interés y surge la necesidad de acceder a ella de forma sencilla y eficiente. Una manera de acceder la información de este tipo de documentos es enfocándonos en su parte oral, es decir, donde una persona se expresa oralmente. Realizar una consulta sobre un documento multimedia no es sencillo, ya que una grabación de este tipo puede durar horas y la información que nos resulta útil podría consistir únicamente de un minuto, y encontrarse en cualquier parte del documento. Por lo que lo deseable es obtener sólo aquel

segmento, en el cual se encuentra la respuesta a nuestra consulta. Ahora bien, si hablamos de colecciones de varios miles de documentos la tarea sólo es posible por medios automáticos. Por ello resulta importante poder encontrar una forma sencilla y práctica de recuperar información en colecciones de documentos orales.

1.2 Descripción del Problema

Cómo se mencionó anteriormente, el problema de la tesis se enfoca en la recuperación de información en transcripciones automáticas, las cuales son resultado de aplicar a documentos orales un proceso de reconocimiento automático de habla.

A pesar de no contar con un RAH perfecto se ha demostrado que es posible realizar la RI sobre transcripciones automáticas. De hecho con un WER entre 20% y 45% es posible aplicar las técnicas de recuperación de información desarrolladas para texto. En [1] se hace una revisión sobre RI en transcripciones de habla y muestra que el impacto del error introducido por un reconocedor automático es mínimo en la tarea de recuperación. Sin embargo, esta conclusión sólo es válida para cuando se trata de tareas de recuperación de información donde las consultas son extensas y los documentos recuperados son grandes. Sin embargo, cuando se efectúa una consulta cuya longitud es pequeña o se desean recuperar documentos pequeños o segmentados (donde no existe redundancia), los errores del RAH impactan fuertemente en la tarea de recuperación.

Los errores que introducen los RAH pueden agruparse en tres:

- *Inserción*: cuando una palabra extra fue agregada en la frase reconocida;
- *Borrado*: cuando una palabra fue omitida en la frase reconocida;
- S*ustitución*: cuando una palabra fue sustituida por otra palabra.

Estos errores impactan seriamente el proceso de recuperación y de forma más significativa cuando la colección consta de pequeños pasajes. Por ejemplo, si en una grabación tenemos la frase "*Unix Sun Workstation*" y el RAH transcribe erróneamente en: "*unique set some workstation*", imposibilitará la correcta recuperación del pasaje si se desea buscar usando el término Unix o Sun. Los errores de substitución son mucho más notorios cuando se trata de entidades nombradas (v. gr. nombres propios). Sobretodo por el hecho de que es imposible contar con un diccionario completo de estos términos. De ahí que el RAH aproxime la realización fonética de la entidad a una palabra fonéticamente cercana dentro de su diccionario.

En la actualidad los métodos propuestos para la RI en documentos orales (Spoken Document Retrieval o más conocidos con la abreviación SDR por sus siglas en inglés), se han enfocado principalmente en probar los métodos existentes en RI para texto escrito. Existen pocas propuestas para solventar los problemas generados por el RAH. Una de estas propuestas consiste en el uso de la codificación fonética.

La codificación fonética busca representar con un mismo código aquellas palabras cuya pronunciación es similar partiendo de su forma escrita. Este método fue propuesto para resolver el problema de diferentes variaciones en la ortografía de un mismo nombre (v. gr. Lewinsky vs Lewinskey). En la actualidad es comúnmente usado para la identificación de nombres personales al acceder a bases de datos, tales como sistemas de censos, archivos médicos, registros de empleados, etc.

El presente trabajo propone explorar la codificación fonética, al aplicarla a una transcripción automática, para aminorar el impacto de los errores del reconocedor. El objetivo es apoyar la recuperación con información a nivel fonético. La codificación fonética nos permitirá llevar la transcipción automática a una representación donde es posible descubrir palabras fonéticamente

similares. La Tabla 1.1 muestra un ejemplo de una codificación fonética y muestra su potencial al comparar una transcripción manual y una automática de una pequeña grabación. Como puede observarse es fácil reconocer la similitud fonética entre las palabras substituidas por el reconocedor. En este caso las palabras *Unix* y *unique* tienen el mismo código así como las palabras *Sun* y *some*.

Transcripción		**Codificación Fonética**
Manual	UNIX Sun Workstation	U520 S500 W623
Automática	*unique set some workstation*	U520 S300 S500 W623

Tabla 1.1. Ejemplo de similitud fonética

1.3 Objetivos

Objetivo General:

- Proponer un método para recuperación de información en documentos orales enriqueciendo su representación a través de codificación fonética.

Objetivos Específicos:

- Comprobar el alcance de las técnicas actuales de recuperación de información en el contexto de documentos orales
- Determinar la pertinencia de la codificación fonética en el ámbito de la RI en documentos orales
- Comprobar la complementariedad de la codificación fonética de los documentos orales con su representación textual
- Evaluar diferentes métodos de codificación fonética de las transcripciones en el proceso de RI

1.4 Organización de la Tesis

En el capítulo 2 se aclaran los conceptos básicos concernientes a las diferentes áreas de la tarea, como el reconocimiento automático del habla, y los problemas que tienen los reconocedores automáticos de habla. También se describe el proceso de recuperación de información, los problemas actuales de la recuperación de información, los modelos existentes de RI y la forma como se evalúan dichos sistemas. Además se presentan los foros de evaluación de RI como son el CLEF y el TREC. En la parte final del capítulo se describen los algoritmos de codificación fonética y se detallan los dos utilizados en esta tesis.

En el capítulo 3, se presenta el estado del arte, las etapas por las que ha pasado la recuperación de información en documentos orales y se muestran los resultados obtenidos en el foro de evaluación CLEF. También se hace un breve análisis de los métodos hasta ahora utilizados en esta tarea. Finalmente se mencionan algunos trabajos que han utilizado la codificación fonética, en específico para la consulta a nombres de personas.

En el capítulo 4 se describe el método propuesto, los experimentos realizados y los avances obtenidos. Finalmente en el capítulo 5 se dan las conclusiones finales de la tesis y se presenta el trabajo futuro que se desprende de este trabajo.

Capítulo II

Conceptos Básicos

En este capítulo se describen los conceptos básicos necesarios para familiarizar al lector con los procesos y terminología utilizados a lo largo de la tesis.

2.1 Reconocimiento automático del habla

Poder controlar con órdenes de voz un dispositivo ha sido uno de los sueños de la humanidad, incluso antes de que existieran las computadoras. Para las personas comunicarse usando el habla es algo fácil e intuitivo ya que por naturaleza el humano tiene la capacidad de comunicarse por medio del habla con las demás personas.

Con el surgimiento de las primeras computadoras se exploró la posibilidad de que estos dispositivos pudieran reconocer y entender el habla, no sólo por facilidad de los usuarios, sino también por cuestiones de seguridad como interceptar automáticamente llamadas de terroristas. Sin embargo, la capacidad de procesamiento de las primeras computadoras fue bastante limitado como para obtener una aplicación real.

Fue hasta 1970 que investigadores de ATT, BBN, CMU, IBM, Lincoln labs, MIT y SRI hicieron contribuciones importantes en la investigación sobre el reconocimiento del lenguaje hablado. En 1971 DARPA (Agencia de Proyectos de Investigación Avanzada para Defensa) invirtió 15 millones de dólares en un

ambicioso proyecto de 5 años, el objetivo era un sistema que aceptara habla continua de varios hablantes con mínima adaptación a los hablantes, un vocabulario de 1000 palabras, sintaxis artificial y de dominio restringido. Harpy y Hearsay-II ambos de la universidad Carnegie-Mellon lograron los objetivos originales e incluso los rebasaron [2].

El problema que representa el reconocimiento del habla es hacer cooperar información procedente de diversas fuentes del conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática) en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido [3].

Los seres humanos nos podemos comunicar bajo muy diversas circunstancias. Como ejemplos: podemos estar hablando con otras personas incluso cuando el ambiente está viciado por ruidos ambientales, cuando estamos escuchando música, cuando alguien está martillando o podando el césped, e incluso cuando otras personas hablan a nuestro alrededor.

Una de las razones por las que los humanos podemos entendernos tan bien, es que usamos información del contexto donde nos encontramos. Sin embargo, un sistema de RAH solo cuenta con modelos acústicos, un diccionario y un modelo de lenguaje, del cual se obtiene una secuencia de palabras probables y limitadas a aquellas que se encuentran dentro del diccionario y del modelo de lenguaje. Esta es una de las razones por las que un RAH genera diversos errores. A continuación se muestra de forma general la arquitectura básica de un sistema de reconocimiento automático de habla y se describe cada una de las partes que lo componen.

2.1.1 Arquitectura básica

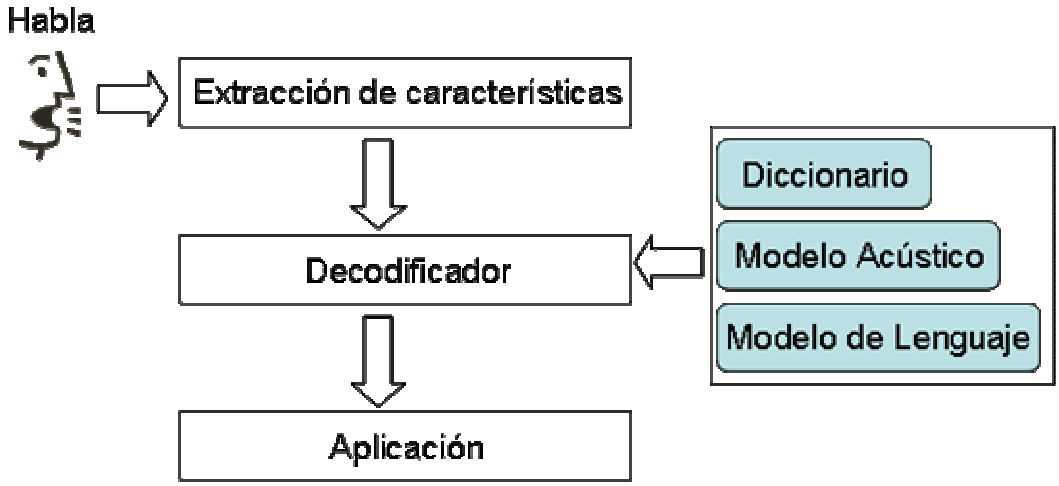


Figura 2.1 Arquitectura básica de un sistema de RAH

En mayor detalle un sistema típico de reconocimiento de habla consiste de los siguientes componentes básicos [4]:

- **Extracción de características acústicas**: extrae características, para el reconocimiento del habla, que se encuentran en el audio. Típicamente incluye un análisis cepstral de corto tiempo, este genera un vector de características de baja frecuencia (10-16) coeficientes cepstrales[1] cada 10ms.

- **Modelo acústico**: Son representaciones estadísticas de los sonidos que forman cada palabra. Se obtienen a partir de una grabación de habla y su correspondiente transcripción fonética La unidad del modelo puede basarse en: unidades con significado semántico (como palabras o unidades fonéticamente significativas). Los modelos ocultos de Markov es la opción prevaleciente para este propósito, aunque las redes neuronales también son utilizadas en varios sistemas.

[1] Coeficientes cepstrales: "Modelo matemático de extracción de patrones". Son la representación de la magnitud logarítmica de la transformada de Fourier del espectro.

En otras palabras el modelo acústico, proporciona al RAH información sobre las propiedades y características de los sonidos del habla.

- **Modelo del lenguaje**: provee restricciones gramaticales y lingüísticas a la secuencia del texto. A menudo está basado en modelos de lenguaje estadísticos como n-gramas. Un modelo de lenguaje de n-gramas es de la forma $P(w_n|w_1,\ldots,w_{n-1})$, lo cuál significa la probabilidad de observar la palabra w_n dadas las palabras anteriores $w_1,\ldots,w_{n-1}$.

  En otras palabras el modelo de lenguaje, contiene la información de cómo se deben combinar las palabras para formar frases. Puede ser pequeño, calculado a partir de unas cuantas frases o extenso a partir de grandes cantidades de texto.

- **Máquina de decodificación**: busca la mejor secuencia de palabras, dadas las características y el modelo del lenguaje. Para el reconocedor de habla basado en modelos ocultos de Markov, esto es alcanzado a través de la decodificación Viterbi[2].

- **Diccionario de pronunciaciones**: indica el conjunto de sonidos con los que se forma cada palabra del vocabulario, y en su caso, las variantes en la pronunciación de una misma palabra.

[2] Decodificación de Viterbi: Nos permite encontrar las secuencia de estados más probable en un Modelo oculto de Markov

2.1.2 Problemas en el reconocimiento del habla

Como mencionamos anteriormente existen diversos problemas en el reconocedor automático de habla. A continuación mencionaremos algunos de estos problemas, los cuáles generan errores provocando que las técnicas tradicionales para recuperación de información no tengan el desempeño que tendrían sobre texto escrito manualmente.

Palabras fuera del vocabulario

Este problema es ocasionado principalmente por un diccionario de pronunciaciones incompleto. Esto se debe a la constante evolución del lenguaje. La cual es mucho más notoria sobre aquellas palabras representando nombres propios. La introducción de nuevos nombres propios es tan rápida que es prácticamente imposible contar con un diccionario completo, el cual enumere todos los posibles nombres de empresas, personas, lugares, etc. Al no existir la entrada correcta en el diccionario de pronunciaciones, el reconocedor intentará aproximar el habla a las palabras conocidas. Así la transcripción de una palabra fuera del diccionario será substituida por una palabra fonéticamente similar o, peor aún, por un grupo de palabras cuya pronunciación conjunta sea fonéticamente similar.

Delimitación de palabras

Otro problema es la correcta composición de las palabras. Un reconocedor tiende a “adivinar”, dada una cadena de fonemas, cual es el corte más apropiado para determinar las palabras correspondientes. Esto provoca que no siempre se obtenga la o las palabras correctas. Este error se produce principalmente por el modelo de lenguaje. Un modelo de lenguaje nos dice la probabilidad de que ocurra una determinada secuencia de palabras. Para el modelo de lenguaje usualmente se generan trigramas (secuencias de 3 palabras consecutivas). El modelo de lenguaje intenta resolver el problema de delimitación de las palabras a partir de las probabilidades de los n-gramas. De

esta manera, a partir de probabilidades es posible identificar una palabra a partir de las palabras anteriores.

El cálculo de los modelos de lenguaje se realiza sobre grandes corpus de documentos. Normalmente las palabras muy comunes, en especial las llamadas palabras vacías, tienden a afectar este paso, ya que se repiten con mayor frecuencia acompañando a diversas palabras y se les otorga una gran probabilidad de ocurrencia. Además tienden a ser partículas pequeñas que fácilmente podemos encontrar como parte de una grande. Por ejemplo, la palabra “enlatados”, contiene las partículas *en*, *la*, *dos*; y el modelo de lenguaje tiende a darle mayor probabilidad al bigrama “en la”. De esta manera, palabras poco probables son erróneamente transcritas.

A continuación se presentan algunos ejemplos de los errores generados por un RAH. Que tradicionalmente se han clasificado en función del error a nivel palabra.

Inserción: una palabra extra fue agregada en la frase reconocida

Correcta	EL PRESIDENTE VICENTE FOX *** REGRESO A MÉXICO
Transcrita	EL PRESIDENTE VICENTE FOX QUE REGRESO A MEXICO

Borrado: una palabra fue omitida en la frase reconocida

Correcta	NO DEJA DE SER UNA CUESTIÓN MUY IMPORTANTE
Transcrita	NO **** DE SER UNA CUESTION MUY IMPORTANTE

Sustitución: una palabra fue sustituida por otra palabra

Correcta	EL DERECHO DE LA UNIÓN EUROPEA
Transcrita	EL DERECHO DE LA UN EUROPEA

Finalmente, también existen otro tipo de errores generados por el tipo de grabaciones a tratar. Por ejemplo, en el caso de transcripciones provenientes de entrevistas se tiene una grabación en habla espontánea. En el habla espontánea existen muchos fenómenos lingüísticos que complican la correcta transcripción por parte del reconocedor. La tabla 2.1 muestra algunos ejemplos. Este tipo de fenómenos, a pesar de pasar inadvertidos para una persona, son fuente de error para un reconocedor.

Fenómeno	**Ejemplo**
Repeticiones	Presiona **el botón**… **el botón** derecho
Auto correcciones	Cierra **el escaparate**... **la ventana**
Elipsis	[mueve] La puerta izquierda aquí
Comentarios	Dibuja una línea ... **eso es**... a la derecha
Agramatical	Reserva boletos 2 personas función de 10
Expresión idiomática	**Genial**!!! No tiene nada que hacer
“Pausas llenas”	Es **mmm** el que está **mmm** a la izquierda

Tabla 2.1 Ejemplos de fenómenos lingüísticos del habla espontánea

2.2 Recuperación de información

En esta sección se describe el concepto de recuperación de información (RI), se muestran algunos de los enfoques usados y se describen las medidas utilizadas para su evaluación.

El término recuperación de información es usado muy frecuentemente y, sin embargo, su definición puede ser muy amplia, llevando a causar cierta confusión en su definición.

Meadow describe la recuperación de la información como “una disciplina que involucra la localización de una determinada información dentro de un almacén de información o base de datos” [5].

Grossman y Frieder indican que “la recuperación de información es encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits” [6].

Greengrass define la recuperación de información como “la disciplina que trata con recuperar datos no estructurados, especialmente documentos de texto, en respuesta a una consulta o tema, la cual también puede estar no estructurada, como una oración u otro documento” [7].

La definición que nosotros retomamos para este trabajo es la propuesta por Manning, la cual dice que la RI “consiste en encontrar material (usualmente documentos) de naturaleza no estructurada (usualmente texto) que satisface una necesidad de información dentro de una gran colección (usualmente servidores de computadora locales o en Internet)” [8].

El término “no estructurado” se refiere a datos que no tienen una estructura fácil de leer para una computadora. Así, “no estructurado” no se refiere a la estructura inherente del lenguaje humano presente en cualquier texto. El texto, desde el punto de vista computacional, es diferente a los datos estructurados que se encuentran almacenados en una base de datos con campos bien definidos.

Por otro lado, cabe destacar que el uso de la palabra “información” podría considerarse un abuso, ya que lo que realmente entregamos al usuario son “documentos”. No obstante, el término “recuperación de información” ha sido ampliamente aceptado y es comúnmente usado para describir este tipo de trabajos.

2.2.1 Objetivo de la recuperación de información

Como se mencionó, la recuperación de información se enfoca en recuperar documentos basándose en el contenido de éstos. Una consulta puede hacer referencia tanto a datos no estructurados, como es el cuerpo del documento, y a datos estructurados como el nombre del autor o fecha del documento.

La recuperación de información pretende encontrar documentos dentro de una colección y presentarlos como resultado a una petición. Estos documentos están relacionados temáticamente a la petición y presumiblemente satisfacen la necesidad de información expresada por la petición. De ahí que existan dos tipos de documentos recuperados: los documentos que satisfacen la necesidad de información llamados *documentos relevantes*; y aquellos no relacionados con el tema o que no satisfacen esa necesidad de información llamados *documentos no relevantes*.

2.2.2 Arquitectura general

Chowdhury identifica el siguiente conjunto de funciones principales en un Sistema de Recuperación de Información [9]:

1. Identificar las fuentes de información relevantes a las áreas de interés de las solicitudes de los usuarios.
2. Analizar los contenidos de los documentos.
3. Representar los contenidos de las fuentes analizadas de una manera adecuada para compararlas con las preguntas de los usuarios.
4. Analizar las preguntas de los usuarios y representarlas de una forma que sea adecuada para compararlas con las representaciones de los documentos de la base de datos.
5. Realizar la correspondencia entre la representación de la búsqueda y los documentos almacenados en la base de datos.

6. Recuperar la información relevante
7. Realizar los ajustes necesarios en el sistema basados en la retroalimentación con los usuarios

Uno de los problemas principales de la RI es la representación de los documentos y de las consultas, la cual determinará como medir las similitudes entre ellas para determinar los documentos relevantes.

Modelo	**Descripción**
Clásicos	Incluye los tres más comúnmente citados: booleano, espacio vectorial y probabilístico
Alternativos	Están basados en la Lógica Fuzzy
Lógicos	Desarrollados en la década de los noventa, basados en la Lógica Formal.
Basados en la interactividad	Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la relevancia de los documentos recuperados [10]
Basados en Otras Técnicas	Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 2.2 Clasificación de los Modelos de Recuperación de Información según Dominich.

Fue en 1957 que Luhn propone utilizar las palabras como unidades para indexar los documentos y medir la ocurrencia de las palabras para saber que tan relevante es el documento para una consulta. Desde ese entonces se desarrollaron varios modelos para representar los documentos y la consulta.

La tabla 2.2 muestra una clasificación de los modelos de recuperación propuestos hasta ahora. Como puede observarse las propuestas son muchas y muy variadas, esto nos da una idea de la complejidad de esta tarea. Pretender evaluar los diferentes métodos y determinar el más pertinente de entre ellos va

más allá del propósito de esta tesis. El método propuesto es independiente del modelo de recuperación usado. No obstante, hemos escogido una máquina de recuperación cuyos resultados reportados sean relevantes y que nos brinde las herramientas suficientes para la realización de los experimentos necesarios. En nuestro caso se escogió la máquina de recuperación INDRI. La siguiente sección describe este modelo de recuperación.

2.2.3 Máquina de recuperación INDRI

Para la realización de los experimentos se utilizó la máquina de recuperación INDRI, la cuál ha demostrado obtener buenos resultados en el área de RI [40,20]. Esta máquina de RI combina un modelado estadístico del lenguaje y una red de inferencia, con esto nos permite que consultas similares a las del lenguaje INQUERY[3] [39] puedan ser efectuadas, pero que sean evaluadas usando los estimados del modelo de lenguaje dentro de la red. La figura 2.3 muestra la representación de esta red de inferencia.

Al final, los documentos son ordenados de acuerdo a la siguiente probabilidad $P(I|D,\alpha\beta)$: que es la creencia de que la necesidad de información I, es encontrada, dado el documento D. Usando como parámetros de suavizado α y β [40,41]. INDRI además incluye un mecanismo de retroalimentación de seudo-relevancia, el cuál está basado en una adaptación del modelo de relevancia de Lavrenko [44]. La forma como está implementado en INDRI no deja mucha libertad al realizar algunas tareas, por ejemplo: no nos permite saber los términos que agregan para expandir la consulta. Sin embargo nos apoyó lo suficiente para realizar los experimentos pertinentes a nuestra tarea.

[3] Este lenguaje de consulta permite realizar búsquedas con lenguaje natural o estructurado. Provee de operadores para incluir en la búsqueda sinónimos, frases completas, buscar por campos o asignar un peso mayor a un término.

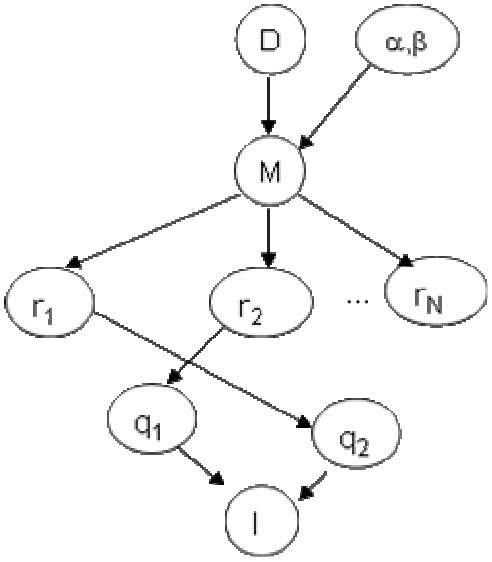


Figura 2.3. Modelo general de recuperación del INDRI

La red consiste en los siguientes tipos de nodos:

- El nodo del documento (D)
- Los nodos de suavizado (alfa, beta)
    - El suavizado es un método usado para superar el obstáculo de una probabilidad cero y el problema de la escasez de datos.
    - Tipos de Suavizado: Jelinek-Mercer, Dirichlet, Two-Stage
- Los nodos modelo (M)
    - Representan las características del modelo de lenguaje (representación del documento)
- Los nodos de representación de conceptos (r)
    - Ejemplo de eventos asociados a estos nodos son:
        - Que aparezca el término dog
        - Que aparezca la frase exacta ‘casa blanca’
- Los nodos de creencia (q)
    - Son variables aleatorias binarias, que son usadas para combinar creencias (probabilidades) dentro de la red.
    - Operadores de creencia: *#combine, #weight, #not, #max, #or*

- El nodo de necesidad de información (I)
    - Combina toda la evidencia dentro de la red en un solo valor de creencia.

Al realizar una consulta se pueden usar diferentes tipos de operadores, los más útiles e importantes son el operador *#combine* y *#weight*. A continuación se muestra un ejemplo de cada uno de ellos. En el caso del operador combine, la creencia es la multiplicación de cada una de las creencias de cada palabra en la consulta.

Operador #combine

Es un operador de "creencia", acerca de términos y/o frases. Este operador pertenece a los operadores no pesados, es decir que todos los términos aportan lo mismo. Realiza una multiplicación de las creencias, desde el primer término hasta el último de la consulta.

$$b_{\#combine} = \prod_{i=1}^{n} b_i^{\left(\frac{1}{n}\right)}$$

Ejemplo de consulta:

#combine(child survivors sweden)

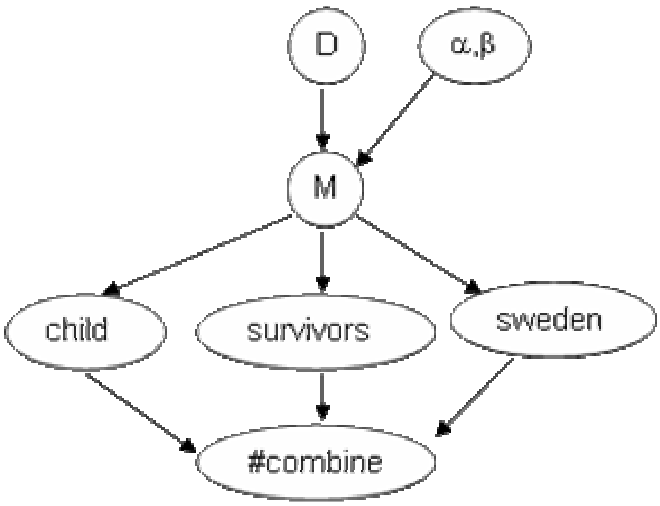


Figura 2.4. Consulta usando *#combine*

Operador #weight

Este operador de creencia, pertenece a los que admiten asignar ciertos pesos a las expresiones según queramos ó a términos. También es una multiplicación de las creencias, desde el primer término hasta el último de la consulta. Sin embargo a diferencia de combine, donde la creencia siempre se eleva a una misma potencia, aquí la creencia se eleva a una potencia que depende, del peso que tenga el término.

$$b_{\#weight} = \prod_{i=1}^{n} b_i^{\left(\frac{wi}{W}\right)} \qquad W = \sum_{i=1}^{n} w_i$$

Ejemplo de consulta:
#weight(1.0 child 1.0 survivors 2.0 sweden)

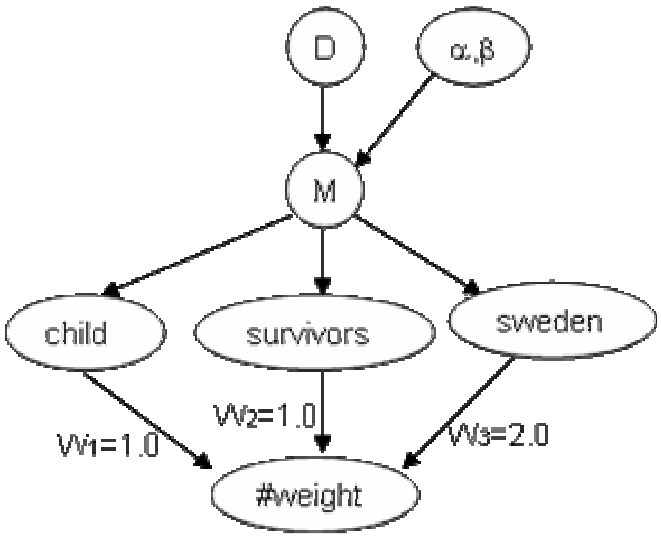


Figura 2.5 Consulta usando *#weight*

En ambos casos, se da por echo que existen n nodos padre, cada uno con creencia b_i y para el operador *#weight* un peso w_i. En caso de que todos los

nodos padre tengan un peso con valor 1, el comportamiento de *#weight* es igual al del operador *#combine*. Para nuestro caso *#weight* nos será útil, ya que permitirá asignarle un mayor peso a lo que consideremos más confiable.

2.2.4 Medidas de evaluación

Al final del día, la medida significativa es la felicidad del usuario. Tiempo de respuesta y precisión son factores importantes para conseguir esa felicidad. Claro que la felicidad del usuario no es algo que se pueda medir, por eso la parte medular de la evaluación en RI es el concepto de relevancia.

La relevancia es un concepto subjetivo, en el sentido que es la satisfacción de la necesidad humana, la última meta, solo que en ocasiones las personas tienen diferente punto de vista sobre la relevancia de un mismo documento. Lo que una persona considera relevante, para otra puede no serlo. Por lo que se puede decir que un documento es relevante para una consulta, usuario y colección en particular. Para otro usuario, otro tipo de colección o en otra consulta, el documento puede no tener la misma relevancia.

Para medir la efectividad de un sistema de recuperación de información con fines específicos y de forma estándar se necesita:

1. Una colección de documentos de prueba.
2. Una colección de casos de prueba de necesidad de información, expresable como consultas.
3. Un conjunto de juicios de relevancia, normalmente una evaluación binaria o de relevancia y no relevancia para cada par consulta-documento.

En la figura 2.6 se muestra como ejemplo una colección de documentos, y los conjuntos existente al efectuar una consulta. Para saber el éxito de la

recuperación de información de un sistema, existen varias medidas de evaluación, entre ellas la precisión y el recuerdo [12].

Precisión es la proporción de documentos relevantes recuperados de todos los documentos recuperados.

$$precisión = \frac{|\{documentos\_relevantes\} \cap \{ducumentos\_recuperados\}|}{|\{documentos\_recuperados\}|}$$

Recuerdo es la proporción de documentos relevantes recuperados de todos los relevantes.

$$recuerdo = \frac{|\{documentos\_relevantes\} \cap \{documentos\_recuperados\}|}{|\{documentos\_relevantes\}|}$$

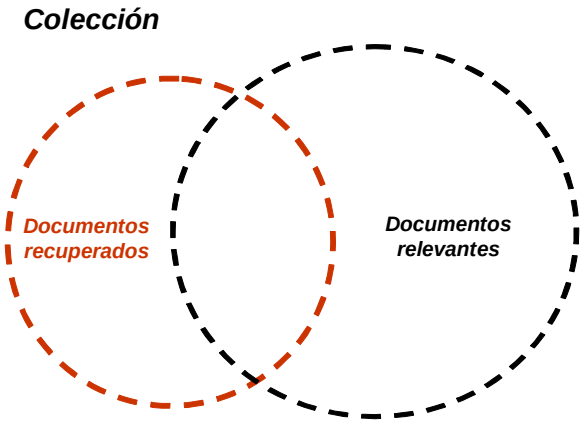


Figura 2.6 Conjuntos de documentos para una consulta

Precisión y recuerdo pueden tener una relación muy cercana, como puede observarse en la figura 2.6. Por ejemplo si se usa una técnica para obtener la raíz de una palabra como “teething troubles” que junta significa problemas iniciales y obtenemos “teeth troubl”, aumentaremos la cantidad de documentos recuperados, sin embargo varios de los nuevos documentos recuperados, no

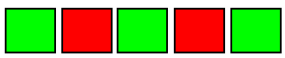
van a mantener la idea de la frase original, ya que por separado tienen otro significado. Es decir, obtendremos un mayor recuerdo pero tendremos una pérdida en la precisión o viceversa.

Precisión Promedio (average precision AvP): es otra medida para tratar de evaluar el comportamiento de un sistema de RI. Una característica importante de este comportamiento es donde quedan colocados los documentos relevantes de entre la lista de documentos recuperados. Recuerde que un sistema de RI regresa una lista ordenada por un valor de relevancia asignado por el sistema. El mejor sistema será el que coloque los documentos relevantes entre las primeras posiciones de la lista. La precisión promedio (AvP) pretende medir el orden de los documentos relevantes en la lista de documentos recuperados. La siguiente fórmula mide la precisión promedio:

$$AvP = \frac{\sum_{r=1}^{N} P(r)}{Número\_de\_documentos\_relevantes}$$

Donde r es el rango y N el número de documentos recuperados y P(r) es la precisión a un corte dado del rango.

Por ejemplo, supóngase una consulta donde se recuperan cinco documentos, de los cuales tres son relevantes. Los documentos relevantes están intercalados tal y como se muestra el siguiente ejemplo donde los documentos marcados en verde son los relevantes. La precisión promedio en este caso resulta ser 0.76, como se ve en el siguiente ejemplo:



$$\frac{1}{3}\cdot\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}\right)\approx 0.76$$

Como vemos esta medida integra el número de documentos relevantes junto con el orden en la lista de documentos recuperados. Sin embargo, esta medida

no es fácil de interpretar. Para observar algunas de las limitantes de esta medida supongamos un ejemplo donde se han recuperado diez documentos, y en total existen cuatro documentos relevantes.

Ahora bien, mostremos dos casos. En el caso 1 se recuperan 2 documentos relevantes en las posiciones 2 y 3. En el caso 2 se recuperan los 4 documentos relevantes, pero todos ellos en las últimas posiciones. Sin embargo como se puede observar el AvP es casi igual para ambos casos.

Caso 1:

$$\frac{1}{4}\cdot\left(\frac{1}{2}+\frac{2}{3}\right)\approx 0.29$$

Caso 2:

$$\frac{1}{4}\cdot\left(\frac{1}{7}+\frac{2}{8}+\frac{3}{9}+\frac{4}{10}\right)\approx 0.28$$

Como podemos observar en el caso 1 y caso 2 si sólo utilizamos la presición promedio, no podemos concluir si el resultado se obtiene porque se recuperaron muchos documentos o porque se recuperaron pocos pero en una buena posición. A pesar de la dificultad en la interpretación de los resultados, esta medida es de gran ayuda en RI y en conjunto con la precisión y el recuerdo, nos puede dar elementos para comparar diferentes sistemas de recuperación.

MAP (*mean average precision*) es otra medida utilizada en recuperación de información, esta medida nos da una idea global del sistema a través de un conjunto de consultas. Para ello se calcula el promedio de las precisiones promedio para ese conjunto de consultas. Esta medida es ampliamente usada para evaluar los sistemas de recuperación, es muy utilizada en los foros de evaluación de RI y es la misma que usaremos para mostrar los resultados obtenidos con nuestro método.

2.3 Foros de evaluación en recuperación de información

El departamento de defensa de Estados Unidos y el Instituto Nacional de Estándares y Tecnología (NIST[4]) en 1992, apoyaron para que se realizara la Conferencia en Recuperación de Texto (TREC[5]) con el propósito de impulsar la investigación dentro de la comunidad de recuperación de información y con el propósito de mantener la infraestructura necesaria para la evaluación a gran escala de metodologías de recuperación de texto.

Una vez creado el TREC, se promovieron diversos foros de evaluación con el fin de promover el intercambio de ideas y de lograr avances más rápidamente. En estos foros se observó que las técnicas para RI empleadas hasta esas fechas no eran adecuadas cuando se empleaban en colecciones grandes, por lo que tuvieron que ser modificadas, surgiendo nuevas aproximaciones.

Durante varios años el TREC [42] ha tenido diferentes talleres como son búsqueda en Web, en blogs, en video, enfocados en precisión, en encontrar información nueva, búsqueda en documentos multilingüe, en documentos legales, en buscar información sobre el genoma y recuperación de documentos orales. La complejidad de las tareas ha aumentado con el tiempo, las primeras tareas se enfocaron en la recuperación de información sobre documentos de texto, y poco a poco conforme las soluciones avanzaban las tareas se incrementaron en dificultad con el fin de avanzar en puntos específicos. Precisamente, una de las tareas recientes es recuperar información en transcripciones automáticas de habla.

Al otro lado del mundo, en el año 2000 surge en Europa otro foro de evaluación llamado CLEF[6] (Cross Language Evaluation Forum) cuya mira es promover la

[4] http://www.nist.gov
[5] http://trec.nist.gov/
[6] http://www.clef-campaign.org/

investigación y desarrollo en recuperación de información a nivel multilingüe, especialmente en lenguas europeas. Para conseguir esto, aporta a la comunidad lo siguiente:

1) Una infraestructura para la experimentación y evaluación de sistemas de RI en lenguas europeas.
2) Crear un conjunto de pruebas de datos reusables, los cuales puedan ser empleados para desarrollar sistemas con el propósito de medir su eficiencia.

En este foro han surgido una gran cantidad de “talleres” en el campo de recuperación de información multilingüe, bilingüe, monolingüe y evaluación de sistemas de dominio específico. Han surgido talleres como búsqueda de respuestas, recuperación de documentos orales, recuperación de imágenes, búsqueda en Web, recuperación geográfica, desambiguación de palabras, validación de respuestas.

Es gracias a estos foros que hoy en día contamos con los elementos apropiados para la evaluación de las diferentes soluciones propuestas en el área de recuperación de información. En nuestro caso, es gracias al esquema de evaluación del foro CLEF que nuestro trabajo puede ser evaluado y comparado objetivamente a otras propuestas de solución en la problemática de recuperación de información en documentos orales.

2.4 Algoritmos de Codificación fonética

La codificación fonética permite representar con un mismo código aquellos nombres que son similares en cuanto a su pronunciación. Es utilizada para identificar variantes en la pronunciación de nombres de personas, necesidad que surgió debido a que los nombres personales cambian su escritura con el

tiempo y/o la región, pero siguen pronunciándose igual. Este problema se presenta sobretodo en aplicaciones como: bases de datos de pacientes, empresas, sistemas de censo entre otros.

La tarea no es sencilla, debido a que hay que encontrar una correlación entre como suena un nombre y como se escribe. Por lo tanto, estos algoritmos son dependientes del idioma y es necesario cambiar las reglas para usarlos en idiomas diferentes. Muchos de los algoritmos fonéticos han sido desarrollados para el idioma inglés, son algoritmos complejos que necesitan manejar muchas excepciones, debido a cambios históricos en la pronunciación del idioma y palabras tomadas de otros idiomas.

La tabla 2.3 enlista los algoritmos fonéticos más comunes así como la fecha en que fueron propuestos y en [34] se realiza una comparación de varios algoritmos fonéticos. Los dos primeros métodos de la tabla 2.3, son los más conocidos y documentados. Por lo que son los que se utilizaron en este trabajo de tesis.

Método	**Año**
Soundex	1918, 1930
Daitch-Mokotoff Soundex	1985
NYSIIS	1970
Phonix	1988, 1990
Metaphone	1990
Double Metaphone	2000

Tabla 2.3 Algoritmos fonéticos

Expresado con teoría de conjuntos, los algoritmos fonéticos dividen el conjunto de todos los posibles apellidos en diversos subconjuntos, agrupando en cada subconjunto apellidos fonéticamente cercanos.

Soundex por ejemplo crea subconjuntos disjuntos: es decir, un mismo apellido no puede pertenecer simultáneamente a dos o más de estos subconjuntos (con lo que se crea una partición del conjunto de todos los apellidos). Otros algoritmos, como Daitch-Mokotoff, crean subconjuntos no disjuntos, lo cual significa que un mismo apellido puede estar simultáneamente en varios de estos conjuntos [13]. A continuación se detallan estos dos métodos.

2.4.1 Soundex

Desarrollado en 1918 por Robert Russell y Margaret Odell [43]. Inicialmente el método fue utilizado para manipular el censo y datos de inmigración. Se volvió popular con el volumen tres de “The Art of Computing”. Actualmente también es parte de los algoritmos de búsqueda, se emplea en programas de manejo de bases de datos y programas para comprobar ortografía, entre otros [14].

El método usado por Soundex está basado en la clasificación fonética de los sonidos del habla humana, los cuales se dividen en 6 clases: bilabial, labiodental, dental, alveolar, velar y glotal. Esta categorización depende de donde se colocan los labios y la lengua para generar un sonido [15].

Este algoritmo transforma los apellidos ingleses en un código de cuatro caracteres. El primer carácter es una letra mayúscula y los tres restantes son dígitos. A continuación se presentan los pasos del algoritmo.

Algoritmo:

1. La primera letra de la cadena se deja en mayúsculas y con la cadena restantese aplican las reglas de abajo.
2. Se remueven todas las ocurrencias de las siguientes letras (a, e, h, i, o, u, w, y), de la cadena restante
3. Se asigna números a las letras restantes como se muestra en la tabla 2.4
4. Si dos o más letras con el mismo número fueran adyacentes, entonces se omiten las repeticiones, menos la primera
5. Regresa los primeros cuatro caracteres y se rellena con ceros a la derecha si son menos de cuatro.

Código	Caracteres
0	a e h i o u w y
1	b f p v
2	c g j k q s x z
3	d t
4	L
5	m n
6	R

Tabla 2.4 Tabla de asignación de números a las letras

Usando Soundex en los nombres Robert y Rupert obtenemos el mismo código R163. Si lo aplicamos al nombre Rubin obtenemos el código R150. Sin embargo, para nombres como Catherine y Katherine obtenemos códigos diferentes C365 para Catherine y K365 para Katherine.

Para los experimentos se utilizó una versión mejorada de Soundex, llamada Enhanced SoundEx [15] que permite elegir el tamaño de la codificación final, de cuatro a diez códigos, además de aumentar unas reglas para tratar de forma apropiada ciertas combinaciones de letras. En la siguiente sección se explica otra versión de Soundex que genera una codificación diferente a la de Soundex.

2.4.2 Daitch-Mokotoff Soundex

Este algoritmo fue desarrollado en 1985 por el genealogista Gary Mokotoff y fue publicado en el primer número de Avoyaynu (el diario de la genealogía judía) en un artículo titulado “The Jewish Soundex: a revised format”. Posteriormente Randy Daitch expandió las reglas del algoritmo creado por Mokotoff. Tanto Daitch como Mokotoff pertenecían a la Sociedad Genealógica Judía [14]. Al final, este algoritmo es una mejora del algoritmo Soundex creado por Russell y Odell.

Este algoritmo surgió por la necesidad de indexar los nombres de unas 28,000 personas que vivían en Palestina entre 1921 y 1948. Al realizar el censo, se observó que una gran cantidad de apellidos de gente judía erán apellidos alemanes y eslavos, los cuales tenían variaciones ortográficas y era necesario homogeneizarlos. Al utilizar el sistema Soundex que tenía el gobierno de Estados Unidos basado en el sistema de Russell y Odell, se observó que nombres diferentes con igual pronunciación no eran convertidos en un mismo código [16]. Fue cuando surgió el método inventado por G. Mokotoff.

Características:

- El código resultante tiene una longitud de seis dígitos.
- El código puede ser almacenado como valor numérico.
- La letra inicial es codificada como cualquier otra letra en el nombre. Si es una vocal, tiene el valor de cero.
- Se agregaron varias reglas en el algoritmo para codificar secuencias de caracteres como dígitos individuales (el apéndice A muestra la tabla de codificación).
- Las letras tienen diferentes valores si están al principio de la palabra, en el centro y variando si anteceden a una vocal o no.
- Puede regresar varios códigos para un mismo nombre.

Las nuevas reglas del método D-M Soundex son independientes de consideraciones geográficas o étnicas. Y se ha convertido en un estándar como en la Organización Genealógica Judía y la Sociedad de Ayuda a inmigrantes hebreos. La tabla 2.5. muestra algunos ejemplos de la conversión de apellidos con ambos métodos.

Apellido	**Soundex**	**D-M Soundex**
Peters	P362	739400, 734000
Peterson	P362	739460, 734600
Moskowitz	M232	645740
Moskovitz	M213	645740
Auerbach	A612	097500, 097400
Uhrbach	U612	097500, 097400
Jackson	J250	154600, 454600, 145460, 445460

Tabla 2.5 Ejemplo de Apellidos codificados a Soundex y D-M

Capítulo III

Estado del Arte

En este capítulo se muestran los avances en el área de recuperación de información en transcripciones de habla.

3.1 RI en transcripciones de habla

Recuperar información en transcripciones de habla es una tarea reciente, comparada con otras. En los siguientes trabajos [35,36,37,38] se proponen diferentes métodos para realizar la RI en documentos orales, algunas de esas ideas han vuelto a retomarse en algún momento en los foros de evaluación y en otros trabajos. Nos centraremos en los avances más recientes, surgidos desde el TREC de 1997, foro en el cuál se realizó esta tarea por primera vez. La tarea recibió el nombre de recuperación en documentos orales (Spoken Document Retrieval o SDR), en esa ocasión el corpus sobre el cual se realizaron los experimentos fue de noticias. En la última versión de la tarea SDR realizada en 1999, se observó que los resultados eran comparables con la recuperación en texto escrito. Los participantes del TREC-9 [17,18,19] en sus conclusiones mencionan que con los resultados obtenidos se demuestra que el rendimiento de los sistemas de RI no se veía afectado, por lo cuál fue considerada una tarea resuelta. Y en 1999 fué la última vez en que este foro se llevó a cabo. Sin embargo, la robustez que se obtuvo en el foro es debido principalmente a la repetición de palabras importantes en la transcripción automática. Allan menciona que se ha demostrado que aún con errores a nivel palabra del 40%, la efectividad de un sistema de RI cae menos del 10%, en [1] se realiza un

análisis más profundo de por qué la recuperación de información en documentos orales se considera un problema resuelto.

Por supuesto esta aseveración es cierta para un cierto tipo de escenario. En el caso de la tarea SDR, se trataba de grabaciones de noticias, éstas tienen la característica de que la pronunciación es bien articulada con una especial preocupación por parte del reportero en hablar correctamente, además estas grabaciones están claramente delimitadas por tópicos.

En otros escenarios, estas condiciones no se cumplen. Por ejemplo cuando se desea realizar búsqueda de respuestas, o cuando la consulta o el segmento a recuperar es pequeño –del orden de una o dos frases–, los errores del reconocedor impactan fuertemente en la efectividad de los sistemas. Además, depende del tipo de grabaciones de audio consideradas. En el caso de grabaciones en habla espontánea como entrevistas o reuniones, las personas no usan la misma entonación, el lenguaje es coloquial, y están presentes infinidad de fenómenos del habla que causan que el rendimiento del RAH sea bajo (la tabla 2.1 muestra algunos ejemplos de los fenómenos presentes en el habla espontánea).

Recuperar información de documentos orales provenientes de entrevistas o reuniones ha resultado ser una tarea muy compleja, esto debido principalmente al habla espontánea. Entre los factores que intervienen en este caso tenemos:

- La gente usa expresiones coloquiales
- Emplea mal las palabras
- Las ideas se presentan desordenadamente tal y como llegan a la mente del entrevistado
- No existen pausas claras a diferencia de texto leído
- La transcripción es un texto sin límites definidos entre temas o entre los diversos hablantes.

Es precisamente el reto de la recuperación de información en habla espontánea que ha dado origen a otras tareas en los foros de evaluación. En un principio, el CLEF retomó la tarea del TREC, con el objetivo de comprobar si las conclusiones obtenidas en un ambiente monolingüe (i.e. en inglés) podían ser extendidas a un ambiente multilingüe, se creó una tarea llamada CL-SDR (Spoken Document Retrieval). Esta tarea comprobó nuevamente la efectividad de los sistemas multilingües comprobando las conclusiones alcanzadas en el TREC [21]. Posteriormente para analizar el contexto del habla espontánea se propuso una nueva tarea llamada CL-SR (Speech Retrieval). Es justamente bajo el contexto de esta segunda tarea que nuestro trabajo se evaluará. A continuación se detallan las dos tareas y se presentan los resultados alcanzados y métodos utilizados en estas tareas.

3.2 CLEF SDR (Spoken Document Retrieval)

A continuación mostraremos una breve descripción de esta tarea, la cual se realizó durante dos años (2003 y 2004). Es importante remarcar que esta tarea, a diferencia de esta tesis, no aborda el problema del habla espontánea. No obstante, se presentan las ideas de los métodos utilizados así como los resultados alcanzados en esta tarea como antecedentes de la tarea CL-SR.

Los recursos utilizados en el CLEF SDR 2003 fueron tomados del NIST, los mismos que se utilizaron en la tarea SDR del TREC 8 y 9. Estos son:

- Una colección de transcripciones de 557 horas de noticias en inglés americano difundidas por ABC, CNN, PRI (Radio internacional Pública), VOA (Voz de América) realizadas entre febrero y junio de 1998. Las transcripciones se proporcionaron con límites conocidos de las historias y sin límites conocidos de las historias. Con un total de 21,754 historias en la colección.

- Dos conjuntos de 50 tópicos en inglés (uno del TREC 7 y otro del TREC 8) además de sus respectivas traducciones a francés, alemán, italiano, español y holandés.
- Juicios de relevancia para cada uno de estos tópicos.
- Software de evaluación.

Dadas las conclusiones del TREC, el reto principal en esta tarea se enfocó en comprobar el impacto de la traducción automática en la recuperación. Además se probaron diferentes soluciones para tratar con la segmentación temática de las transcripciones. Hay que recordar que una transcripción automática es una secuencia de palabras sin signos de puntuación, ni marcas que indiquen un cambio de tema. Es necesario segmentar para que el sistema de recuperación nos entregue aquellos documentos más relevantes. Desafortunadamente, no existen criterios claros para determinar el tamaño ideal de un documento.

La tabla 3.1 muestra los resultados alcanzados durante el 2003. En esta ocasión se les brindó a los participantes tanto las transcripciones segmentadas como no segmentadas. Ningún participante presentó alguna representación o técnica específica para el tratamiento de los errores introducidos por el RAH. Todos ellos se abocaron a aplicar las técnicas usadas comúnmente a texto. Por ejemplo, la Universidad de Alicante participó con un sistema de recuperación de información llamado IR-n, el cuál fue diseñado para realizar la búsqueda en documentos escritos manualmente, por lo que tuvo que ser adaptado para esta tarea. El objetivo de ellos fue probar la robustez de su sistema IR-n y cómo resultaría su funcionamiento en una colección donde no se conocen los límites de las frases. Como puede observarse los resultados fueron adecuados llegando a un MAP de 0.3569. En [24] se puede consultar con mayor detalle las modificaciones y el funcionamiento.

Otro ejemplo es la Universidad de Exeter, la cual utilizó un sistema clásico basado en el pesado de términos Okapi BM25[7] con retroalimentación de pseudo relevancia, además aplicó procedimientos estándares como remover palabras vacías y el truncado de términos. Simplemente optimizaron los parámetros de pesado de Okapi para la colección de prueba SDR. En [22,23] se pueden consultar mayores detalles.

Equipo	**Resultados**
Alicante	.3569
JHU	.3184
Exeter	.3824
ITC-irst	.3944

Tabla 3.1 MAP de los resultados de consultas en inglés 2003

La recuperación de Información sobre documentos de texto (escritos manualmente) va de 0.4 a 0.5. Como puede verse en la tabla 3.1, los resultados son comparables a los alcanzados en recuperación de información en documentos de texto. Estos resultados comprobaron la validez de las técnicas de RI tradicionales sobre documentos orales. Por supuesto hay que recordar que se trata de transmisiones de noticieros de radio, es decir, no se trata de habla espontánea y tanto las consultas como los documentos son relativamente grandes.

Durante 2004 se realizó nuevamente el foro CLEF SDR. En ese año se realizó la evaluación con un ligero cambio, los sistemas no podían estar basados en límites temáticos conocidos. Es decir, cada sistema tenía que segmentar las transcripciones automáticas y conservar un índice temporal para determinar la posición del segmento dentro de la grabación. En 2004 sólo participaron la Universidad de Chicago y el ITC-IRST

[7]Okapi BM25: es una función que sirve para ordenar los documentos, basada en modelos probabilísticos. Se utiliza el concepto de bolsa de palabras, mediante la cual se representan los documentos que deseamos ordenar.

Ambos trabajos aplicaron diferentes técnicas de RI para texto, ninguno de ellos realizó algo específico para intentar recuperarse de los errores de reconocimiento. Por ejemplo, la Universidad de Chicago [25] se enfocó en mejorar la recuperación y para ello usó expansión con retroalimentación de pseudo-relevancia. Para resolver el problema de la segmentación, las grabaciones se dividieron en segmentos de 30 segundos, con un traslape de 10 segundos y éstos segmentos fueron indexados para realizar la recuperación.

Equipo	Resultado
ITC-irst	.3059
Chicago	.2963

Tabla 3.2 MAP de las ejecuciones del 2004

En la tabla 3.2 se pueden observar los resultados obtenidos ese año. Aunque los resultados disminuyeron con relación al año anterior, los resultados fueron considerados suficientes por lo que la tarea se consideró resuelta. A pesar de que aún no se sabe como segmentar apropiadamente las transcripciones, los métodos de recuperación aplicados en texto escrito eran suficientes para este tipo de documentos orales.

3.3 CLEF CL-SR (Speech Retrieval)

A partir del 2005 se inició la evaluación de una tarea con un reto mucho mayor, la recuperación de información en habla espontánea. Para esta tarea se consideraron grabaciones de entrevistas. Esta tarea se distingue de la otra por la carencia de un tema claro en la conversación –ya que durante una entrevista usualmente se tratan diversos temas–, el error introducido por el RAH es mucho mayor.

Tres factores influyeron para lanzar la propuesta de esta tarea:

- La mejora de los reconocedores automáticos de habla, los cuales llegan a obtener un error a nivel palabra (WER) para el reconocimiento del habla de 20% en conversaciones telefónicas y de 30% en grabaciones de reuniones (en ambos casos se trata de habla espontánea).
- Se contaba con el acceso a un corpus de grandes dimensiones y donde era necesario resolver el problema de recuperación de información. La Fundación de Historia Visual (VHF) recolectó una gran cantidad de entrevistas la cual se digitalizó y anotó. Se tienen 116,000 horas de entrevistas a sobrevivientes del holocausto, testigos y rescatistas, de las cuales un subconjunto de 10,000 horas fue extensivamente anotado.
- Los resultados alcanzados en RI en documentos orales presuponían que era posible alcanzar resultados apropiados en este escenario.

3.3.1 Descripción del corpus

La colección de la Fundación de Historia Visual (VHF) es enorme, por eso inicialmente se utilizó un subconjunto de 10,000 horas. Aún así resultó ser muy grande, y en otra etapa se realizó otra selección, sobre todo debido a que el sistema de reconocimiento automático del habla es un proceso iterativo, donde los resultados del sistema inicial son usados para guiar el desarrollo de un sistema más refinado. Al momento de realizarse la evaluación en el CLEF, solamente una porción de 272 entrevistas había sido procesada por dos sistemas de RAH. En promedio, una entrevista de la Fundación de Historia Visual (VHF) se extiende por más de dos horas Al final se obtuvieron 8,104 segmentos de entrevistas, con 589 horas de habla. Esto da en promedio segmentos de 4 minutos, equivalentes a unas 503 palabras por segmento.

Aunque una colección de este tipo resulta pequeña comparada con los experimentos realizados por la recuperación de información moderna usada en recursos escritos, es comparable a las usadas en las tareas de SDR. Cada segmento fue acompañado de metadatos. En especial se adicionaron campos con palabras claves obtenidas tanto por procesos manuales como automáticos. Para una descripción detallada del corpus puede consultarse [26].

3.3.2 Resultados del 2005 y descripción de los tres mejores sistemas

En 2005 para esta tarea participaron siete grupos de investigación:

- Universidad de Alicante
- Universidad de Dublín
- Universidad de Maryland
- Universidad Nacional de Educación a Distancia (UNED)
- Universidad de Pittsburgh
- Universidad de Ottawa
- Universidad de Waterloo

Aunque cada equipo podía realizar la recuperación aprovechando todos los campos de los documentos, se solicitó a todos los equipos una corrida base para contar con resultados fácilmente comparables. En esta corrida base sólo se permite usar los campos generados automáticamente, es decir, las transcripciones automáticas y las palabras claves generadas automáticamente. Esto le da un ámbito de realismo al problema. La tabla 3.3 muestra los resultados alcanzados. Como puede observarse el MAP es muy bajo evidenciando la complejidad de la tarea.

Universidad	MAP
Ottawa	0.1653
Maryland	0.1288
Waterloo	0.1121
UNED	0.0934
Alicante	0.0768
Pittsburgh	0.0757
Dublín	0.0654

Tabla 3.3 Resultados del CLEF CL-SR 2005.

A continuación se describen los sistemas presentados por los tres equipos con mejores resultados en ese año.

Universidad de Ottawa

Usó un sistema construido con componentes comerciales, el sistema de RI usado se llama SMART (Salton's Magic Automatic Retriever of Text, por sus siglas en inglés). En él se emplearon varias configuraciones de pesado para el indexado de los documentos y la representación de las consultas. También se removieron las palabras vacías y se aplicó un truncado sobre las palabras restantes.

Los esquemas de pesado son combinaciones de frecuencia de términos por documento, frecuencia de términos en la colección y componentes de normalización dada las diferencias de longitud entre los documentos. SMART está basado en el modelo de espacio vectorial. A pesar de probar con varios esquemas de pesado, sólo se reportaron los mejores resultados.

En este caso si se realizó un intento por abordar el problema del reconocimiento del habla. Para ello se generó la transcripción fonética de la colección. Se calcularon n-gramas de tamaño 4 sobre las secuencias de fonemas y esta información acompaño a los documentos al momento de indexar la colección.

Los resultados agregando la información fonética al texto resultaron mejores que al utilizar únicamente texto.

Universidad de Maryland

Usó el sistema de RI llamado InQuery y aplicó técnicas de recuperación de información automáticas (como retroalimentación de pseudo-relevancia). También las palabras vacías de cada consulta y de los documento fueron removidas automáticamente por InQuery. El truncado de las consultas y documentos también se realizó con InQuery.

Maryland utilizó una colección alterna para realizar la expansión del documento. Esta colección alterna cotaba con 168,584 segmentos de entrevistas restantes de la colección. También se experimentó con diferentes configuraciones para la expansión. Por ejemplo, se tomaron 10, 20, 50 y 100 documentos y de ellos se tomaron los 10, 20, 30, 40 y 50 términos más adecuados con la restricción de que el término apareciera en al menos tres de los mejores documentos.

Además usaron un tesauro de sinónimos para expandir el documento. Aunque al realizar un análisis de estos sinónimos se observó que son términos relacionados más que sinónimos. Como ejemplo la palabra *tíos* tiene los siguientes términos relacionados: primos, abuelos, cuñada, cuñado, nuera, suegro, suegra.

Finalmente, realizaron la expansión de la consulta[8] usando retroalimentación de relevancia ciega. Para ello, primero se realizó la expansión de los documentos y estos documentos fueron los utilizados para realizar la expansión de la consulta. Se utilizaron las mejores 5, 10, 15 y 20 palabras tomando los primeros 10, 20, 30 documentos. Y se encontró que el mejor resultado se alcanzó

[8] Expansión de la consulta: es el proceso de reformular la pregunta original, con el fin de mejorar el desempeño en la recuperación de información.

usando las 5 mejores palabras de los mejores 20 documentos y que la palabra aparezca en al menos 2 de esos documentos.

Universidad de Waterloo

Usó un sistema de RI llamado Wumpus. Muchas de sus corridas usaron retroalimentación de pseudo-relevancia. Para llevar a cabo la retroalimentación el corpus original se incrementó con un corpus secundario. Este corpus secundario de 2.5 GB fue recuperado de la Web por un motor de búsqueda (crawler[9]) enfocado en tópicos relacionados, con una semilla inicial de 17 sitios dedicados al holocausto judío.

Al igual que la Universidad de Ottawa se utilizaron n-gramas de fonemas de tamaño 4. Se experimentaron varias formulaciones de la consulta y técnicas de expansión, incluyendo el uso de los n-gramas fonéticos y la expansión de la consulta usando retroalimentación sobre el corpus extendido.

Al observar los resultados de estos tres métodos es interesante constatar que el primero no usó fuentes externas de conocimiento. Por otro lado, no fue posible concluir sobre la utilidad de la representación de los documentos usando n-gramas fonéticos. Los resultados demostraron que aún falta mucho por hacer en esta complicada tarea.

3.3.3 Resultados del 2006 y descripción de los tres mejores sistemas

Para la evaluación del 2006 se agregaron 30 nuevos temas de búsqueda y se contó con un mejor sistema de reconocimiento automático del habla, el cuál proporcionó una trasncripción con un error a nivel palabra más bajo. En este año, la colección se distribuyó incluyendo la transcripción de un nuevo

[9] Crawler: Programa que visita sitios Web y recoge información de acuerdo a algunos criterios generales

reconocedor, llamado ASR2006. En esta ocasión también se incluyó una colección en checo, pero en esa no profundizaremos, ya que nuestro objetivo por el momento sólo se enfoca en la recuperación monolingüe en inglés.

En esta ocasión fueron seis los equipos participantes:

- Universidad de Alicante
- Universidad de Dublín
- Universidad de Maryland
- Universidad Nacional de Educación a Distancia (UNED)
- Universidad de Ottawa
- Universidad de Twente

Cabe destacar que los tres equipos que obtuvieron los mejores lugares en este año (Dublín, Ottawa y Maryland), ya habían participado en la competencia del año 2005. A continuación se resumen los métodos empleados por estos tres equipos.

Universidad de Dublín

Usó dos sistemas basados en el modelo de recuperación. La tarea se enfocó en explorar la combinación de múltiples campos asociados con los documentos de habla.

Dos métodos estándar de combinar múltiples campos de documentos son:

- Combinar todos los campos representándolos en un solo documento.
- Indexar los campos por separado, ejecutar corridas por cada campo y mezclar los resultados sumándolos en un proceso final de fusión de datos.

Realizaron la eliminación de palabras vacías (usaron una lista de 260 palabras) de los documentos y la consulta. Se utilizó truncado y los términos son indexados usando un pequeño conjunto estándar de sinónimos. Usaron dos

variantes de retroalimentación de pseudo-relevancia seleccionando diferentes criterios para la selección de los términos de expansión.

Universidad de Ottawa

En este segundo año de participación, utilizó dos sistemas de RI: el sistema que había empleado anteriormente: SMART, además de agregar otro, Terrier. Los dos sistemas fueron usados con diferentes esquemas de pesado para la indexación de los segmentos y en las consultas. Además en las consultas se usaron diversas técnicas de expansión. Usaron diferentes trascripciones de los RAH para indexar los segmentos y varias combinaciones de las transcripciones automáticas.

El sistema SMART incluye un mecanismo de expansión de la consulta, que sigue la idea de extraer palabras relacionadas a cada palabra en el tema.

La idea de usar el sistema Terrier era comprobar el uso de un sistema de recuperación basado en un modelo diferente al SMART. El sistema Terrier está basado en un modelo probabilístico. Además, el mecanismo de expansión de la consulta en Terrier es diferente, en este caso se usa el modelo de Kullback-Leibler (KL).

Universidad de Maryland

A diferencia del año anterior, no utilizó información externa. En este año se limitaron a usar la combinación de las trascripciones obtenidas por los RAH del 2004 y 2006, además de los campos de palabras clave generados automáticamente. Las palabras vacías fueron removidas tanto de la pregunta, como de los documentos, de forma automática por InQuery, que es la máquina de RI que utilizaron. También se obtuvo la raíz de las palabras tanto de las consultas, como de los documentos de forma automática por InQuery.

Universidad	MAP
Dublín	.0733
Ottawa	.0565
Maryland	.0543
Twente	.0381
UNED	.0376
Alicante	.0375

Tabla 3.4 Resultados en el CLEF CL-SR 2006

Como puede observarse en la tabla 3.4, los resultados bajaron respecto a los del 2005. Esto se debió principalmente a que se utilizaron otros tópicos de prueba los cuales resultaron ser más complicados. En el apéndice D, se pueden observar algunos ejemplos de estos tópicos.

3.3.4 Resultados del 2007 y descripción de los tres mejores sistemas

La tarea del 2007 fue idéntica a la del 2006 usando el mismo conjunto de consultas. Sin embargo, se realizó una revisión completa de los juicios de relevancia y se eliminaron ciertas inconsistencias. Debido a esta revisión, existen variaciones entre los juicios de relevancia del 2006 y 2007. En el documento descriptivo de la tarea para 2007 [33], se menciona que los resultados del 2007 no son estrictamente comparables con los del 2006.

En este año participaron cinco equipos, entre los que se encontraba la universidad de Dublín y Ottawa, quienes ya habían competido en años pasados. Las universidades de Brown, Ámsterdam y Chicago fue su primera participación en está tarea.

- Universidad de Brown
- Universidad de Dublín
- Universidad de Amsterdam

- Universidad de Chicago
- Universidad de Ottawa

En esta ocasión las Universidades de Dublín y de Ottawa, ambas participantes del año pasado, obtuvieron nuevamente el primer y segundo lugar en las evaluaciones de este año.

Universidad de Ottawa

Esté es el tercer año de su participación, de manera similar a su participación anterior usó dos sistemas de RI: SMART y Terrier. Ambos sistemas fueron usados otorgando diferentes esquemas de pesado a los campos en la indexación de las transcripciones. Para el tratamiento de las consultas se usaron diversas técnicas de expansión, por tesauro y retroalimentación de pseudo-relevancia. Se usaron dos métodos de fusión para combinar las salidas de los sistemas. Estos métodos consideraron el comportamiento del sistema en el año anterior. Es decir, se tomó el MAP alcanzado el año anterior para determinar sobre el peso de los sistemas para lograr una fusión apropiada. Desafortunadamente, al tratarse del mismo conjunto de evaluación del año anterior, es de esperarse que los buenos resultados se deban a un sobreajuste sobre esos datos de evaluación. Para la fusión se usó un total de 15 diferentes listas [31].

Universidad de Dublín

También fue el tercer año que compitió esta universidad, en esta ocasión como primer tratamiento para los documentos se agregaron sinónimos, se aplicó un truncado y se eliminaron las palabras vacías. Después se indexaron separadamente los campos y a través del método BM25F se ordenan los documentos recuperados, integrando el aporte de los diferentes campos. Cabe mencionar

que realizaron pruebas con dos variantes de retroalimentación de pseudo-relevancia [30].

Universidad de Brown

Para indexar los documentos aplicó truncado y para su recuperación usó retroalimentación de pseudos-relevancia. Usaron una máquina de recuperación probabilista usando modelos basados en unigramas y bigramas. Además mezclaron la colección con dos grandes corpus (40,000 frases del Wall Street Journal y 450,000 frases del North American News Corpus), con el objetivo de evitar los problemas de falta de datos al trabajar con colecciones pequeñas [32].

Universidad	**MAP**
Ottawa	.0855
Dublín	.0787
Brown	.0785
Chicago	.0571
Ámsterdam	.0444

Tabla 3.5 Resultados obtenidos en el CLEF CL-SR 2007

Cabe hacer notar que ninguno de estos métodos propuestos en el 2007 trata de resolver el error introducido por el reconocedor, ni tampoco realiza algún tratamiento de la información a nivel fonético.

También es relevante hacer mención que el uso de información fonética para solucionar esta tarea fue abandonada en los últimos años y ninguna otra propuesta fue explorada para tratar de evitar el error introducido por el RAH. A continuación se presentan algunos de los pocos trabajos que han usado la codificación fonética en diversas tareas de RI. La tabla donde aparecen juntos los resultados obtenidos del 2005 al 2007 se puede observar en el apéndice C, aunque no son comparativos los de un año con los de otro, da una idea de cómo se comportó la tarea y los equipos participantes.

3.4 RI usando algoritmos fonéticos

Los algoritmos fonéticos se han utilizado principalmente para realizar la recuperación de apellidos de personas. Esto debido a una mala escritura del nombre o a variaciones fonéticas y culturales que llevan a la "deformación" de los nombres.

En [27] los autores mencionan que asignar un sólo código fonético a cada nombre, presume la existencia de un algoritmo que se ajusta a todas las situaciones, como eso no se cumple, por eso proponen mejorar el recuerdo sin perder mucha precisión utilizando una fusión de varios algoritmos fonéticos. El sistema resultante obtiene buenos resultados, sobre un corpus de apellidos llamado COMPLETE y el cual contiene 14,972 apellidos distintos.

En [29] Zobel y Dart realizan una comparación de varias técnicas fonéticas, aunque también sólo se orienta a recuperar nombres. Ellos cuentan con una lista de 30,000 apellidos distintos ya depurados, obtenidos de la Web. Y generaron 100 consultas para las cuales generaron juicios de relevancia. Los mejores resultados se obtuvieron utilizando un algoritmo de distancia fonética, llamado Editex. Este algoritmo combina las propiedades de la distancia de edición con la estrategia de agrupamiento de letras usada por Soundex.

El trabajo más cercano a esta tesis es el propuesto a Allan y Raghavan [28], ellos mencionan el problema que se tiene en la ortografía de los nombres y que es muy remarcado en el caso de los RAH. Ellos proponen usar Soundex en documentos obtenidos por un reconocedor automático de habla, sin embargo al igual que los trabajos anteriores, sólo es utilizado para indexar nombres. Y mencionan que hasta ese año 2003 todos los trabajos se concentraban en recuperar nombres de bases de datos preexistentes. Este era el primer intento en reconocer nombres de forma automática usando códigos fonéticos, para después indexarlos en una base de datos.

Capítulo IV

Método propuesto

La idea central del método de recuperación propuesto consiste en enriquecer la descripción de los documentos al incorporar información fonética. La intención de usar esta nueva información es resolver, en cierto grado, los errores introducidos por el RAH. Con esta nueva representación se espera incrementar el número de documentos relevantes recuperados.

Ahora bien, como consecuencia de la codificación fonética, la representación resultante es una representación más compacta, es decir el número de códigos resultantes es menor que el vocabulario de la colección. En consecuencia, se obtienen códigos con altas frecuencias. El tener códigos frecuentes impacta en el éxito de la recuperación, de la misma manera que las palabras vacías impactan en la recuperación en texto. De ahí que es necesario aplicar un proceso para eliminar aquellos códigos frecuentes. El método propuesto también incluye un proceso de filtrado de la representación con el objeto de eliminar los códigos frecuentes.

Como es de imaginar, dado el cambio en la representación de los documentos también es necesaria la modificación de la consulta y extenderla con su representación fonética. Con el fin de integrar esta nueva información en la recuperación de documentos, se realizaron un conjunto de experimentos para determinar un método adecuado. Por último, tal como lo demuestran los métodos propuestos en el CLEF, un proceso de expansión de la consulta es utilizado para mejorar el rendimiento del sistema.

La siguiente sección describe a detalle la construcción de la nueva representación a partir de la codificación fonética. Posteriormente en la sección de experimentos se muestran los resultados alcanzados con los diferentes conjuntos de datos así como la mejora al utilizar el proceso de expansión de la pregunta.

4.1 Extendiendo la representación de los documentos orales.

Para transformar la representación de los documentos se realizan los siguientes pasos:

1) Se filtra cada una de las transcripciones automáticas eliminando las palabras vacías.
2) Cada transcripción es codificada fonéticamente usando el algoritmo Soundex mejorado (con códigos tamaño 6).
3) Una vez conocidas las frecuencias de los códigos en toda la colección, se realiza un filtrado eliminando aquellos códigos frecuentes.
4) La codificación filtrada se incluye como parte de la descripción del documento.

La eliminación de palabras vacías es importante pues este tipo de partículas son las más abundantes en el lenguaje y por ende las más fácilmente introducidas por el reconocedor (véase la sección 2.1). El conjunto de palabras vacías que se eliminaron de la trascripción automática y de las palabras clave, se tomó de la lista que aparece en los recursos de la Universidad de Glasgow, del departamento de ciencias computacionales[10].

[10] http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

El paso de eliminar las palabras vacías en toda la colección de 8,104 documentos se realiza en unos segundos. El paso de convertir todos los documentos de la colección a la codificación Soundex se lleva unos 30 minutos. Las 63 consultas de entrenamiento, se realizan en menos de 2 minutos.

Como se indicó anteriormente se utilizó el algoritmo Soundex Mejorado para la codificación fonética, el cuál permite generar códigos de tamaño 4 a 10. Este algoritmo nos permitió extender la representación base que ofrece Soundex, de un código de tamaño 4 a uno de 6 posiciones, esto permite incrementar la granularidad de la representación.

Hay que recordar que la longitud del código está relacionada con la longitud de la palabra codificada. Así, para codificar adecuadamente palabras largas necesitamos códigos de más posiciones. En nuestro caso particular se observó que el tamaño promedio de las palabras en la colección es de 7.2 caracteres. Debido a que la codificación Soundex elimina vocales y otros códigos, se consideró que una codificación de tamaño 6 es suficiente para representar las palabras.

Para resolver el problema de la eliminación de los códigos frecuentes se realizaron diferentes experimentos. El método con mejores resultados consistió en la aplicación de una idea sencilla: conservar la misma tasa de eliminación que la utilizada a nivel texto. Para ello se calculó el volumen total de la colección, visto como el número total de palabras en la colección. Después de calculó aquella fracción de este volumen correspondiente a las palabras eliminadas en el texto (i.e. las palabras vacías).

La cantidad de palabras vacías utilizada fue de 319, y estas palabras corresponden al 75% del volumen total de la colección de texto. Para conservar esta misma relación en la representación fonética, fue necesario tomar los 263 códigos más frecuentes. La tabla 4.1 muestra algunos de estos códigos, el

apéndice E muestra la lista más amplia de algunos de los códigos más frecuentes que fueron eliminados.

Codificación fonética	Texto	Codificación fonética	Texto
U00000	U, uh, uhhuh	C50000	came, cane, can, coin, coma, come
N00000	no, nowa, nowe, nieuw, nah, neo, new, now, knee, knew, know	W53000	won't, wound, weaned, wanda, wendy, windy, wined, want, went, wind, window
D30000	dad, did, dot, daddy, death, data, dead, deed, died, diet, dude, duty, ditto	T50000	teeny, teeny, taiwan, tinny, tummy, ten, tim, tin, tom, tommy
P14000	pupil, people, popiel, pablo, papal	T52000	teams, tango, tanks, teens, thomas, teeing, times, tongue, tony's
L20000	lucia, lucky, league, louise, laughs, locks, lodge, leg, log, looks, loose, louis, lousy	S30000	shade, shady, sheet, pseudo, shoot, shoddy, saudi, sweaty

Tabla 4.1 Ejemplos del colapso de texto a Soundex

La última línea de las tablas 4.3 y 4.4, muestra la representación enriquecida de los documentos, en ella se mezclan tanto la representación textual (i.e. la transcripción oral) y los códigos fonéticos correspondientes. Está mezcla resulta importante debido a la complementariedad que se obtiene al usar la codificación fonética y al usar solo la representación textual.

Respecto a la consulta también es necesario transformarla fonéticamente, la tabla 4.2 muestra un ejemplo de ello. Como puede notarse en el ejemplo, el uso de la codificación fonética permite incrementar las posibilidades de alineación

entre la transcripción oral y la consulta. Por ejemplo, entre el documento ejemplo de la tabla 4.4 y la consulta de la tabla 4.5 las palabras "*roll*" y "*Raoul*" fueron representadas con el mismo código, incrementando las similitudes entre ambos. En este caso particular, "roll" fue una substitución equivocada producida por el reconocedor.

A continuación se muestran una serie de ejemplos donde es pertinente utilizar la codificación fonética, y donde la representación enriquecida ayuda a la recuperación. La tabla 4.3 muestra un ejemplo de una transcripción, con su respectiva codificación fonética, donde la palabra *homework* fue transcrita en lugar de la palabra *hamburg*, que tiene más sentido en el contexto y tiene el mismo código H56200. Eso se supone, porque es un segmento de un pasaje que es relevante a la consulta que aparece en la tabla 4.2

Transcripción automática	neuengamme concentration camps, external women camps of neuengamme concentration camps in ***hamburg***
Codificación fonética	N52500 C52536 C52000 E23654 W55000 C52000 O00000 N52500 C52536 C52000 I50000 ***H56200***

Tabla 4.2 Ejemplo de la codificación de la consulta

Transcripción automática	… and he went to homework …
Codificación fonética	… A53000 H00000 W53000 T00000 H56200 …
Representación enriquecida	{ and he went to ***homework*** A53000 H00000 W53000 T00000 ***H56200*** }

Tabla 4.3 Ejemplo de una codificación enriquecida

En el ejemplo anterior, la representación enriquecida se muestra con todo y palabras vacías, sin embargo como se mencionaba las palabras vacías fueron

eliminadas del documento. En las tablas 4.4 y 4.5, se muestra un ejemplo de otra consulta, la transcripción en texto y su representación fonética, así como la representación del documento enriquecido, una vez que se retiraron las palabras vacías.

Transcripción automática	…just your early discussions was roll wallenberg's uh any recollection of of uh where he came from and so…
Codificación fonética	... J23000 Y60000 E64000 D22520 W20000 R40000 W45162 U00000 A50000 R24235 O10000 O10000 U00000 W60000 H00000 C50000 F65000 ...
Representación enriquecida	{just, early, discussions, roll, wallenberg's, uh, recollection, uh, came, E64000, D22520, R40000, W45162, R24235}

Tabla 4.4 Ejemplo de una codificación enriquecida del documento

Consulta original	wallenberg and eichmann eyewitness accounts that describe the personalities and actions of raoul wallenberg and adolf eichmann
Codificación fonética	W45162 A53000 E25500 E35200 A25320 T30000 D26100 T00000 P62543 A53000 A23520 O10000 R40000 W45162 A53000 A34100 E25500
Representación enriquecida	wallenberg eichmann eyewitness accounts personalities actions raoul wallenberg adolf eichmann W45162 E25500 E35200 A25320 P62543 A23520 R40000 W45162 A34100 E25500

Tabla 4.5 Ejemplo de una codificación enriquecida de la consulta

Para este ejemplo, es interesante notar que el uso de la codificación fonética permite mejorar el emparejamiento entre la transcripción y la consulta, ya que las palabras "*roll*" y "*Raoul*" son representadas con el mismo código fonético (R40000). En este caso suponemos que *roll* fue obtenido de una transcripción incorrecta, ya que Raoul Wallenberg fue uno de los que salvo muchas vidas de judíos en la segunda guerra mundial y fue una de las primeras victimas de ella.

Otro ejemplo de un segmento relevante a la consulta anterior de Raoul Wallenberg es el siguiente, donde las palabras eating yes, pudieron haber sido una separación generada por el RAH. En este caso la palabra eyewitness, tiene

la misma codificación fonética que eating (E3520), permitiendo realizar la recuperación del documento. La tabla 4.6 muestra un ejemplo de la representación del documento.

Transcripción automática	… have you been eating yes you know wherever the russians were allies …
Codificación fonética	... H10000 Y00000 B50000 E35200 Y20000 Y00000 N00000 W61600 T00000 R25200 W60000 A42000 ...
Representación enriquecida	{ eating yes know wherever russians allies E35200 W61600 R25200 A42000}

Tabla 4.6 Ejemplo de una codificación enriquecida del documento

Como se verá en la sección de experimentos, la consulta se realiza combinando los términos originales así como su codificación fonética a través de un esquema de pesado. Dado que las palabras aportan información más precisa que las codificaciones, se otorga un peso mayor a los términos originales que a su respectiva codificación fonética.

4.2 Resultados experimentales

A continuación se explica el corpus sobre el cuál se realizaron los experimentos y cuales de ellos se llevaron a cabo, para lograr nuestros objetivos.

4.2.1 Corpus CL-SR 2007

La colección sobre la que se realizaron los experimentos cuenta con 8,104 documentos y fue obtenida del foro de evaluación CLEF, para mayor detalle de los campos que son proporcionados en la colección ver apéndice B. Cada uno de los documentos contenidos en la colección, cuenta con los siguientes campos:

- 4 transcripciones generadas automáticamente
- 2 conjuntos de palabras claves generadas automáticamente
- 1 conjunto de palabras clave generadas manualmente
- 1 resumen generado manualmente

En los experimentos no se utilizó ninguno de los campos generados manualmente, con la intención de enfrentar un escenario lo más real posible. Las transcripciones son generadas por diferentes reconocedores automáticos de habla. Se cuenta con cuatro RAH, uno del 2003 con un error a nivel palabra (WER) del 45%, otro del 2004 con un WER del 38% y dos del 2006 con un WER del 25%

El conjunto de palabras clave[11] es generado automáticamente por dos clasificadores de vecinos más cercanos y cada campo contiene los mejores 20 términos obtenidos por los clasificadores. Nos referiremos al conjunto de estos términos como AK1 y AK2.

Para realizar los experimentos el corpus cuenta con:

Preguntas de entrenamiento:	63
Total de documentos relevantes, para las 63 preguntas:	5,229
Total de entrevistas con su respectiva transcripción automática:	8,104
Total de entrevistas con palabras clave:	8,104

Cabe mencionar que las transcripciones utilizadas fueron las proporcionadas por el reconocedor ASR2006B, el cuál solo tiene la transcripción de 7,378 documentos (con un WER del 25%), los otros 726 documentos, fueron los generados por el reconocedor ASR2004 (con un WER del 38%). De aquí en adelante nos referiremos a ellas como ASR06, para mayor detalle de la colección consultar [46]. La razón por la que se utilizan diferentes tazas de error

[11] Son palabras relacionadas al documento, pero obtenidas de una fuente externa.

es debido a que al momento de la competición, aún faltaba aplicar el nuevo reconocedor automático de habla a algunas grabaciones, por lo que se usaron las que se habían obtenido con anterioridad.

4.2.2 Experimentos y Resultados

En esta sección se presenta una serie de experimentos que permiten evaluar el uso de la codificación fonética.

La primera serie de experimentos se realizó para comprobar la *utilidad* de la codificación fonética de las transcripciones. Para ello se realizó la recuperación usando únicamente la información textual (transcripciones del campo ASR06 y las palabras clave generadas automáticamente AK1 y AK2). Por el otro lado, se codificaron las transcripciones orales y se compararon las tazas de recuperación para estos dos casos, en la tabla 4.7 se muestran los resultados alcanzados, los cuales aún son más bajos que los obtenidos en la tarea CL-SR que alcanza el 0.0855.

Como puede observarse la codificación fonética llega a recuperar una cantidad menor de documentos a los recuperados al usar el texto, sin embargo más abajo se demuestra que hay complementariedad entre ambos conjuntos de documentos recuperados. En los resultados se muestra el MAP, que es la medida utilizada en el foro de evaluación y que se describe en la sección 2.2.4, la precisión que se obtiene a las 10 primeras posiciones (P@10) y la cantidad de documentos relevantes recuperados (Rel. Rec.).

	Recuperación tradicional (texto)	**Recuperación usando únicamente la codificación fonética**
MAP	0.0864	0.0651
P@10	24.4%	18.6%
Rel. Rec.	2045	1923

Tabla 4.7 comparación de RI en Texto y usando Soundex

Como bien lo demuestran los trabajos del CLEF [30,31,32], realizar una expansión de la consulta favorece la recuperación de documentos. Así que se realizaron otra serie de experimentos aplicando una expansión de la consulta. En nuestro caso se utilizó retroalimentación ciega. El método incorpora a la consulta original los 140 términos más frecuentes de los primeros 10 documentos recuperados. La tabla 4.8 muestra los resultados alcanzados al usar retroalimentación.

	Recuperación tradicional (texto)	**Recuperación usando únicamente la codificación fonética**
MAP	0.1089	0.081
Precision@10	25.4%	19.4%
Rel. Rec.	2389	2079

Tabla 4.8 Utilizando retroalimentación ciega

Para obtener los valores de retroalimentación, se realizaron varios experimentos, en la tabla 4.9 se pueden observar los resultados. Los resultados mostrados son aplicando retroalimentación sobre los documentos de texto (transcripción automática y palabras clave), se utilizaron los 10 primeros documentos y se varió la cantidad de términos frecuentes a agregar. Cabe notar que las consultas originales tiene en promedio 18 palabras, por lo que se inició a probar como se comportaban los resultados al agregar 10 términos y se fue aumentando la cantidad de términos a agregar, hasta quedarnos con 140 términos. Se puede observar que esto fue a prueba y error. Se puede observar que con 180 términos no se notaba una gran mejora, además la cantidad de documentos relevantes recuperados empezó a caer.

	10 términos	**20 términos**	**40 términos**	**80 términos**	**140 términos**	**180 términos**
MAP	0.0864	0.0973	0.1019	0.1049	0.1089	.1090
P@10	22.8%	23.97%	23.6%	24.4%	25.4%	25.5%
Rel. Rec.	2166	2252	2287	2357	2389	2384

Tabla 4.9 Selección de términos frecuentes

En la tabla 4.8 se observa que la codificación fonética tiene un MAP menor al obtenido sobre texto, sin embargo, la información recuperada por ambas representaciones no es exactamente igual, esto demuestra que si es de utilidad el uso de la codificación fonética. En la Tabla 4.10 y 4.11 se muestra el nivel de complementariedad y redundancia de los documentos recuperados al usar las distintas representaciones. En las tablas se muestra la cantidad de documentos complementarios tomando en cuenta la cantidad de documentos relevantes recuperados.

	Solo en Texto	**Solo en Soundex**	**En ambos**
Núm. de Doc. Rec.	371	249	1674

Tabla 4.10 Relevantes Recuperados sin retroalimentación

	Solo en Texto	**Solo en Soundex**	**En ambos**
Núm. de Doc. Rec.	557	252	1832

Tabla 4.11 Relevantes Recuperados con retroalimentación

Como se observa en las tablas 4.10 y 4.11, existe cierta complementariedad entre ambos métodos. En otras palabras ambos métodos recuperan 2,240 documentos relevantes, de los cuales comparten 1674, indicando que un 27% aún son complementarios. Se realizó otra serie de experimentos con la intención de demostrar la utilidad de combinar la información textual y fonética.

La descripción de un documento queda integrada por su descripción textual (campos ASR06, AK1 y AK2) y por su descripción fonética (codificación Soundex de los campos ASR06, AK1 y AK2). Esto es, el nuevo documento contiene dos campos, en uno se encuentra solo la representación textual y en el otro campo sólo la representación fonética.

Como se mencionó anteriormente, la consulta está compuesta de una parte textual y otra la codificación fonética de los términos textuales. Sin embargo,

dado que los términos textuales no tienen colisiones y permiten una mayor distinción entre palabras, se les asignó un peso mayor que a los códigos correspondientes.

Se realizaron varios experimentos, en la tabla 4.12 se muestran los resultados obtenidos al otorgar diferentes pesos a las palabras de la consulta y manteniendo el peso de la codificación fonética en 1.0. Se puede observar que antes de usar la retroalimentación, el mejor caso es al otorgar un peso de 2 a las palabras, sin embargo con la retroalimentación el mejor resultado (MAP) fue cuando se le otorgó un peso de 4 a las palabras, pero si observamos la cantidad de documentos recuperados, se observa que fue mayor al otorgar un peso de 1.0 a las palabras y a los códigos.

Un ejemplo de una consulta es el siguiente:

#weight(2.0 palabra1 2.0 palabra2 1.0 código1 1.0 código2).

En este ejemplo se le asigna un peso de 2.0 a las palabras y un peso de 1.0 a la codificación fonética.

				Retroalimentación		
Peso al texto	**1**	**2**	**4**	**1**	**2**	**4**
MAP	0.0854	0.0877	0.0874	.1026	.1021	.1027
P@10	0.2397	0.2317	0.2381	25.4%	24.4%	24.9%
Rel. Rec.	2055	2080	2082	2346	2307	2311

Tabla 4.12 Uso de diferentes pesos al texto

Hasta el momento los resultados obtenidos al realizar la RI utilizando el texto y la representación fonética mejoran respecto a utilizar solo el texto, pero esto es antes de usar la retroalimentación. Al utilizar la retroalimentación los resultados son menores. Como se mencionó la retroalimentación se la estamos dejando al Indri, En el caso anterior estamos codificando a Soundex tanto la transcripción

automática, como las palabras clave. Sin embargo como se observa en la tabla 4.1 la codificación nos ayuda a corregir ciertos errores del reconocedor, pero también nos puede generar una gran confusión y hacer que los códigos sean excesivamente frecuentes. Además las palabras clave están bien escritas y no es necesario codificarlas ya que solo agregaran más ruido. Por lo que en el siguiente experimento se omitieron las codificaciones de las palabras clave y solo se agregó la codificación en Soundex de la transcripción automática. En la tabla 4.13 se muestran los resultados obtenidos al usar retroalimentación y sin usarla.

				Retroalimentación		
Peso al texto	**1**	**2**	**4**	**1**	**2**	**4**
MAP	0.0837	0.0876	0.0886	0.1085	0.1098	0.1108
P@10	23.6%	23.97%	21.6%	26.1%	26.2%	26.8%
Rel. Rec.	1930	2005	2036	2228	2308	2317

Tabla 4.13 resultados alcanzados con la representación combinada 2

Si se comparan los resultados con los obtenidos en las tablas 4.12 y 4.13, se observa una ligera mejora cuando se utiliza la retroalimentación. Como se comentaba, debe tomarse en cuenta que la codificación Soundex compacta el vocabulario y aunque las palabras claves ya no son codificadas a Soundex, la codificación de la transcripción automática aún está generando muchas colisiones, por lo que existe una gran cantidad de códigos fonéticos frecuentes que no ayudan en el proceso de recuperación (ni tampoco al método de retroalimentación).

La siguiente serie de experimentos demuestra la utilidad de eliminar dichos códigos frecuentes. Para identificar los códigos a eliminar se utilizó una analogía con el texto y las palabras vacías. La idea consistió en conservar la misma tasa de eliminación utilizada en el texto debido a las palabras vacías. Para ello se observó el número total de ocurrencias de las palabras vacías. El

volumen total de las 319 palabras vacías correspondió el 75% del volumen total de la colección. Para conservar esta relación en la codificación fonética es necesario quitar los 263 códigos más frecuentes.

				Retroalimentación		
Peso al texto	**1**	**2**	**4**	**1**	**2**	**4**
MAP	0.0842	0.0895	0.0902	0.1105	0.1163	0.1147
P@10	24.3%	24.4%	25.5%	26.98%	28.1%	27.3%
Rel. Rec.	1928	1998	2052	2259	2355	2371

Tabla 4.14 resultados alcanzados con la representación combinada 3

En la tabla 4.14 se observa que el método mejora a la recuperación sobre texto para los casos donde se le asigna un peso mayor a las palabras. El MAP más alto se obtuvo al asignar un peso de 2.0 a las palabras, sin embargo con un peso de 4.0 provoca que se recuperen más documentos. Sin embargo por el momento, nos interesa más mejorar el MAP, por lo que seguiremos utilizando un peso de 2.0 para el texto. Como la cantidad de códigos a eliminar coincidía casualmente a eliminar aquellos códigos con frecuencias mayores a 1000 veces y ahora para observar la pertinencia de la regla anterior, se realizaron experimentos con otros umbrales de frecuencia: mayor a 500 repeticiones (417 palabras) mayor a 2000 repeticiones (141 palabras) y mayor a 4000 repeticiones (71 palabras). La tabla 4.15 muestra los resultados obtenidos con los diversos umbrales para la eliminación de códigos frecuentes. Los pesos de la consulta fueron iguales en todos los casos: el doble a las palabras del asignado a los códigos. De igual forma, en todos los casos se aplicó retroalimentación de seudo relevancia[12].

[12] Método con el que se expande la consulta al agregar nuevos términos, que aparecían de forma frecuente en los documentos obtenidos al realizar una primera recuperación.

Frecuencia	**> 500**	**> 2000**	**> 4000**
MAP	.0922	.1115	.1114
P@10	23%	26.8%	26.3%
Rel. Rec.	2221	2342	2314

Tabla 4.15 Diferentes casos de eliminación de códigos frecuentes.

Los resultados demuestran la utilidad de eliminar códigos frecuentes. La Tabla 4.16 muestra la mejora obtenida con respecto a la recuperación al usar únicamente texto.

	Recuperación tradicional (texto)	**Combinación**	**Cambio relativo**
MAP	0.1089	0.1163	+6.8%
P@10	25.4%	28.1%	+10.6%
Rel. Rec.	2389	2355	-1.4%

Tabla 4.16 Comparación de resultados

Como puede observarse, el método nos permite un mejor ordenamiento de los documentos relevantes, alcanzando un mejor MAP y una mejor P@10 (precisión a las primeras 10 posiciones).

Por último, se realizó una serie de experimentos para comprobar el alcance del método propuesto. En este caso, se usó el conjunto de consultas de prueba usadas en el CLEF CL-SR 2007. Este corpus de prueba esta compuesto de 33 consultas con un total de 2449 documentos relevantes. La tabla 4.17 muestra los resultados obtenidos con nuestro método.

	Combinación
MAP	0.0795
P@10	17.3%
Rel. Rec.	1319

Tabla 4.17 Combinación Texto y Soundex con los datos de prueba

Estos resultados son muy alentadores, ya que al observar los resultados reportados por el CLEF en el 2007 (tabla 3.5) nos coloca en la segunda posición.

4.3 Discusión

En este capítulo se mostró el método propuesto para realizar la recuperación de información en transcripciones de habla. Se mostraron los pasos que se siguieron para generar otra representación que pudiera solventar los errores generados por el reconocedor de habla. El método se enfoca en generar una representación de códigos fonéticos y agregarlos al documento original, así como en modificar la consulta y asignarle un peso mayor a las palabras, ya que estas aportan mayor precisión.

Como se mencionaba anteriormente, la codificación Soundex, genera muchas palabras que aunque apoyan, también generan confusión. En el ejemplo de la tabla 4.18 se puede observar un caso donde la codificación fonética puede generar confusiones.

Transcripción automática	our local rabbis came from Syracuse
Codificación fonética	O6000 L24000 R12000 C50000 F65000 S62200
Confusión	Rabbis = Refugees = R12000

Tabla 4.18 Ejemplo de confusiones con la codificación

La palabra *rabbis* se codifica igual que la palabra *refugees*, esta es una razón más de por que la codificación Soundex necesita ir acompañada de la transcripción automática y realizar una combinación de ambas para mejorar los resultados, además de la complementariedad que se observa en las tablas 4.10 y 4.11. Otro ejemplo de cómo un solo código representa varias palabras se encuentra en la tabla 4.1 y una mayor cantidad de códigos aparecen en el apéndice E, donde se muestra la tabla de los códigos más frecuentes.

Debido a este tipo de colisiones que ocurren, se trató de realizar una prueba con otro tipo de codificación, para observar el comportamiento que se obtenía. La codificación fonética que se utilizó fue el algoritmo llamado Daitch-Mokotoff Soundex (D-M Soundex) que fue el que le siguió al Soundex de Rusell y Odell. Este algoritmo genera una mayor cantidad de códigos que el Soundex. Soundex para la transcripción automática, colapsa todo el vocabulario en 7765 códigos, DM-Soundex colapsa todo el vocabulario en 9526 códigos. Esto haría suponer que hubiera menos colisiones y generara menos ruido, sin embargo en la tabla 4.19 se muestran los resultados obtenidos con este método y resultaron ser más bajos que la recuperación solo en texto.

	Recuperación tradicional (texto)	**Combinación > 4000**
MAP	0.1086	.1056
P@10	25.4%	25.7%
Rel. Rec.	2389	2251

Tabla 4.19 Combinación Texto y D-M Soundex

En la tabla 4.15 se muestra el resultado de quitar los códigos cuya frecuencia es mayor a 4000. Al eliminar código cuya frecuencia es mayor a 4000 en esté método los resultados no bajan demasiado, pero no proporcionan una mejora, que es lo que se hubiera esperado. Sería interesante realizar más experimentos, primero quitando cantidades diferentes del código DM-Soundex y también con otra codificación fonética.

A continuación en la tabla 4.20 se puede observar la comparación de utilizar nuestro método con respecto a otros en la tarea del CLEF CL-SR 2007. Cabe notar que nuestro método queda en segundo lugar y como aplica una representación diferente de la colección, puede ser implementado junto a algún otro método de los usados por otros equipos.

Universidad	**Campos**	**MAP**
Ottawa	AK1,AK2,ASR04	.0855
Nuestro método	*AK1,AK2,ASR06*	*.0795*
Dublín	AK1,AK2,ASR06	.0787
Brown	AK1,AK2,ASR06	.0785
Chicago	AK1,AK2,ASR06	.0571
Ámsterdam	AK2,ASR06	.0444

Tabla 4.20 Comparación de nuestros resultados contra métodos en CL-SR 2007

Capítulo 5

Conclusiones y Trabajo Futuro

En la actualidad los métodos propuestos para recuperar información en documentos orales no toman en cuenta los problemas generados por el reconocedor automático de habla. En esta tesis, se propuso un método para abordar este problema. La idea consistió en enriquecer la representación de documentos orales utilizando la codificación fonética de la transcripción automática. La codificación fonética busca reducir el impacto de los errores generados en la transcripción, representando aquellas palabras con pronunciación similar a través del mismo código fonético.

Las conclusiones principales de este trabajo se resumen en los siguientes puntos:

- La codificación fonética es útil en la tarea de recuperación de información en documentos orales.
- Se aplicaron dos métodos de codificación fonética, Soundex y DM-Soundex, donde los resultados alcanzados en la recuperación de información fueron superiores con el primer método.
- La codificación fonética al emplearse en conjunto con otras técnicas de RI mejora los resultados de la recuperación.
- El método propuesto es sencillo de implementar para realizar RI en transcripciones automáticas.

Como se menciona en el capítulo de resultados, un problema que se presenta en la codificación fonética es que existen demasiadas colisiones, esto se debe a que el algoritmo Soundex es muy sencillo y no cuenta con reglas para el

manejo de excepciones en la pronunciación. Sin embargo, al realizar pruebas con el algoritmo D-M Soundex el cual tiene reglas que “mejoran” la codificación –generando menos colisiones– los resultados no mostraron una mejora. De ahí la utilización del algoritmo Soundex sea suficiente para el tratamiento de los errores de substitución en las transcripciones.

La eliminación de códigos frecuentes incrementa la discriminación entre los documentos y mejora los resultados de la recuperación. La eliminación de estos códigos frecuentes también contribuyó en un mejor comportamiento al aplicar la técnica de retroalimentación de pseudo relevancia.

Finalmente, los resultados experimentales obtenidos sobre la colección del CLEF CL-SR demostraron ser alentadores. Nuestro método alcanza resultados que demuestran la utilidad de la codificación fonética en una tarea tan exigente como la del foro CL-SR.

5.1 Trabajo futuro

Como se demostró el método es adecuado para la tarea. Aún así existen algunos puntos a estudiar y experimentar. Entre ellos la granularidad del código fonético a utilizar. En esta primera versión el método utilizó una codificación fonética a 6 posiciones, no obstante, el método Soundex mejorado permite generar una codificación de entre 4 y 10 posiciones. Es muy probable que a mayores tasas de error (WER) del RAH corresponda una codificación con menores posiciones. Por supuesto, es necesario realizar más experimentos para concluir al respecto.

Otro camino podría consistir en el enriquecimiento de la descripción de los documentos con distintos tamaños de codificación; o inclusive con diferentes algoritmos de codificación fonética. De esta manera, otras codificaciones

fonéticas como D-M Soundex, NYSIIS, Phonix, Metaphone, o Double Metaphone, podrían ser usadas para enriquecer la descripción del documento; y de igual forma concluir sobre cual codificación fonética es la más conveniente para esta tarea.

Pruebas realizadas con el algoritmo D-M Soundex, el cual tiene reglas que "mejoran" la codificación –generando menos colisiones–no mostraron mejora en los resultados. De ahí que debe realizarse un análisis más profundo para determinar el tipo de reglas adecuado para la generación de los códigos fonéticos.

La eliminación de códigos frecuentes mejora el alcance de la recuperación, ahora bien, el criterio propuesto para determinar el conjunto de códigos a eliminar puede refinarse, además de que sebe experimentarse en otras colecciones para obtener conclusiones generales.

Otro aspecto que debe determinarse es el método de expansión, hasta el momento se utilizó un método la retroalimentación propio de la máquina de recuperación de información. Sin embargo, un método con mayor libertad para reformular la consulta nos lleve a obtener mejores resultados, entre otras cosas, saber las palabras que se están agregando nos permitiría hacer un análisis de cómo está ayudando la retroalimentación y nos permita reformular la consulta al expandir únicamente con códigos o con palabras, o estudiar sobre la mejor proporción entre ambos.

Finalmente, existe un enfoque para la recuperación en transcripciones automáticas, que intenta hacer la recuperación a nivel de caracteres y no de palabras. Para ello, se calculan los n-gramas a nivel de caracteres, de esta forma el agrupamiento de los fonemas impuesto por el RAH para formar las palabras se rompe. No obstante este enfoque no se realiza a nivel fonético, lo cual mantiene parte de los errores generados por el reconocedor de habla. De

ahí que un idea a explorar como trabajo futuro sería la utilización de la codificación fonética a nivel de n-gramas. Esto nos permitiría atacar el problema de la inserción y el borrado de palabras. Por ejemplo, la siguiente tabla muestra un ejemplo de segmentación de la palabra “december” con su respectiva codificación fonética. Como se puede observar los diferentes trigramas de “december” pueden asociarse a otras palabras a través del código fonético.

3-grama	Codificación Soundex	Palabras con el mismo código fonético
dec	D20	dice
ece	E20	each, ease
cem	C50	come, came
emb	E50	???
mbe	M00	???
ber	B60	bear, beer

Tabla 5.1 Segmentación de la palabra “december”

Incluso esta codificación utilizando 3-gramas permite obtener una similitud con las palabras como “decima, decent, decant” y “embargo, embark, ember, embarras”, etc. Por otro lado, también habría de tener en cuenta que este tipo de representación multiplicaría los códigos dificultando la tarea de recuperación. De ahí la importancia de proponer métodos para identificación de las códigos más frecuentes a eliminar.

BIBLIOGRAFIA

[1] Allan J. Perspectives on Information Retrieval and Speech. Workshop on Information Retrieval Techniques for Speech Applications, New Orleans LA, Lecture Notes in Computer Science, Vol. 2273, pp. 1-10,.2002.

[2] Huang X., Acero A. and Hon. H. Spoken Language Processing: A guide to Theory, Algorithm and system Development, Prentice-Hall 2001.

[3] Bernal J. Bobadilla J. y Gómez P. Reconocimiento de Voz y Fonética acústica, Alfaomega 2000

[4] Juang B. Pattern Recognition in Speech and Language Processing, CRC Press, 2003

[5] Meadow C. T. Text Information retrieval Systems. San Diego: Academic Press, 1993.

[6] Grossman, D. A. and Frieder, O. Information retrieval: algorithms and heuristics. Kluwer Academic Publishers, 1998.

[7] Greengrass, E. Information Retrieval: A Survey. 2000. http://www.cs.umbc.edu/cadip/readings/IR.report.120600.book.pdf

[8] Manning C. D., Raghavan P. and Schütze H., Introduction to Information Retrieval, 2007. http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

[9] Chowdhury G. G., Introduction to Modern Information Retrieval, 2nd ed. London Facet Publishing, 2004.

[10] Salton G. Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer. Reading, MA: Addison Wesley, 1989.

[11] Strohman T., Metzler D., Turtle H. and Croft W.B. Indri: A Language-Model based Search Engine for Complex Queries. Proceedings of the International Conference on Intelligence Analysis, McLean, VA. (Poster). May 2-6, 2005.

[12] Baeza-Yates R. and Ribeiro-Neto B., Modern Information Retrieval, New York : ACM Press, Addison-Wesley, 1999.

[13] UzZaman N. and Khan M. A Bangla Phonetic Encoding for Better Spelling Suggestions. Proc. 7th International Conference on Computer and Information Technology, Dhaka, December, 2004.

[14] Galvéz C. Identificación de Nombres Personales por Medio de Sistemas de Codificación Fonética. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação 2 semestre 2006(22):pp. 105-116.

[15] Creativyst, Inc. 2002 – 2007 Understanding Classic SoundEx Algorithms http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm

[16] Mokotoff G. Soundexing and Genealogy 1997-2007 http://www.avotaynu.com/soundex.html

[17] Renals S. and Abberley D. The THISL SDR system at TREC-9. Proceedings of the Ninth Text Retrieval Conference (TREC-9), Voorhees E. M. and Harman D. (Eds), pp 627-634, NIST Special Publication 500-249, 2000

[18] Gauvain J-L., Lamel L., Barras C., Adda G., and Kercardio Y., The LIMSI SDR System for TREC-9, Proceedings of the Ninth Text Retrieval Conference (TREC-9), Voorhees E. M. and Harman D. (Eds), pp 335-359, NIST Special Publication 500-249, 2000

[19] Johnson S.E., Jourlin P., K. Jones S. and Woodland P.C., Spoken Document Retrieval for TREC-9 at Cambridge University, Proceedings of the Ninth Text Retrieval Conference (TREC-9), Voorhees E. M. and Harman D. (Eds), pp 335-359, NIST Special Publication 117-126, 2000

[20] Shuaixiang Dai, Qian Diao and Changle Zhou. Performance comparison of language models for information retrieval, Artificial Intelligence Applications and Innovations II, pp 721-730, Edited by Daoliang Li and Baoji Wang, 2006.

[21] Garafolo, J. S., Auzanne, C. G. P., and Voorhees, E. The TREC Spoken Document Retrieval Track: A Success Story. In Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access. Paris, France, pp 107-130, 2000.

[22] Lam-Adesina A. and Jones. G. J. F., Exeter at CLEF 2003: Cross-Language Spoken Document Retrieval Experiments, Proceedings of the CLEF 2003:

Workshop on Cross-Language Information Retrieval and Evaluation, Trondheim, Norway, pp 653-657, 2004.

[23] Lam-Adesina A. and Jones. G. J. F., Exeter at CLEF 2002: Cross-Language Spoken Document Retrieval Experiments, Proceedings of the CLEF 2002: Workshop on Cross-Language Information Retrieval and Evaluation, Rome, Italy, pp458-475, 2003.

[24] Llopis F. and Martínez-Barco P., Spoken Document Retrieval experiments with IR-n system, Workshop of the Cross-Language Evaluation Forum, Trondheim, pp 446-457, Agosto 2003.

[25] Levow G., and Matveeva I., University of Chicago at CLEF2004: Cross-Language Text and Spoken Document Retrieval, In Proceedings of CLEF'2004. pp.170-179, 15-17 September, Bath, UK 2004.

[26] White R., Oard D., Jones G., Soergel D. Huang X. Overview of the CLEF-2005 Cross-Language Speech Retrieval Track. Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005). Vienna, Austria, pp 744-759, 21-23 September 2005.

[27] Holmes D. and McCabe M. C., Improving Precision and Recall for Soundex Retrieval, ITCC Proceedings of the International Conference on Information Technology: Coding and Computing, pp 22-26, 2002

[28] Raghavan H. and Allan J. Using Soundex Codes for Indexing Names in ASR documents. In Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at Human Language Technology Conference and North American chapter of Association of Computational Linguistics, pp 22–27, Boston, MA, USA, 2004.

[29] Zobel J. and Dart P., Phonetic String Matching: Lessons from Information Retrieval, Sigir Forum, Association for Computing Machinery, pp. 166-172, New York, 1996

[30] Jones G. Zhang K. and Lam-Adesina A. Dublin City University at CLEF 2007: Cross-Language Speech Retrieval (CL-SR) Experiments. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, pp 794-802, 19-21 September 2007

[31] Alzghool M. and Inkpek D. Model Fusion Experiments for the Cross Language Speech Retrieval Task at CLEF 2007. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007

[32] Lease M. and Charniak E. Brown at CL-SR'07: Retrieving Conversational Speech in English and Czech. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007

[33] Pecina P., Hoffmannová P. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007

[34] Kessler B. Phonetic Comparison Algorithms. *Transactions of the Philological Society Volume 103:2*, pp 243-260, 2005.

[35] Witbrock M. and Hauptmann A.G., Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents, Proceedings of the 1997 DARPA Speech Recognition Workshop, February 2-5, 1997

[36] Wechsler M., Munteanu E. and Schäuble P., New Approaches to Spoken Document Retrieval. *Information Retrieval, volume 3*, pp 173-188, Octubre 2000.

[37] Coden A. R., Brown E. and Srinivasan S., Information Retrieval Techniques for Speech Applications, *ACM SIGIR Workshop*, London Uk, Springer-Verlag publishers, pp 23-77, 2001

[38] James, D.A., The application of classical information retrieval techniques to spoken documents, Ph.D. thesis, University of Cambridge, UK 1995.

[39] Callan J.P., Croft W.B., and Harding S.M., The INQUERY Retrieval System, In Proceedings of the Third International Conference on Database and Expert Systems Applications, pp 78-83, Valencia, Spain, 1992.

[40] Strohman, T., Metzler, D., Turtle, H., and Croft, W.B., Indri: A language-model based search engine for complex queries, CIIR Technical Report, 2005.

[41] Metzler, D. and Croft, W.B., Combining the Language Model and Inference Network Approaches to Retrieval, Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), 735-750, 2004.

[42] Voorhees, E., Garofolo, J., and Jones, K. The TREC-6 Spoken Document Retrieval Track. Proceedings of the Sixth Text Retrieval Conference (TREC-6). Gaithersburg, Maryland, pp 83-92, November 19–21, 1997.

[43] Odell, M. K., Russell, R. C. U. S. Patent Numbers 1261167 (1918) and 1435663 (1922). Washington, D.C.: U.S. Patent Office, 1918.

[44] Lavrenko, V. and Croft, W.B., Relevance-Based Language Models, Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '01), pp 120-127, 2001.

APÉNDICES

APÉNDICE A

A continuación se muestra la tabla de codificación del método Daitch-Mokotoff Soundex, se pueden observar las reglas que se usan para codificar las diferentes letras, cómo se puede modificar la ortografía de acuerdo a cual sea la secuencia de letras y la variación si es al inicio de la palabras, antes de una vocal o en algún otra situación.

Letra	Alternar Ortografía	Inicio de un nombre	Antes de una vocal	Cualquier otra situación
NC = no codificado				
AI	AJ, AY	0	1	NC
AU		0	7	NC
Ą	(Polish a-ogonek)	NC	NC	6 or NC
A		0	NC	NC
B		7	7	7
CHS		5	54	54
CH	Try KH (5) and TCH (4)			
CK	Try K (5) and TSK (45)			
CZ	CS, CSZ, CZS	4	4	4
C	Try K (5) and TZ (4)			
DRZ	DRS	4	4	4
DS	DSH, DSZ	4	4	4
DZ	DZH, DZS	4	4	4
D	DT	3	3	3
EI	EJ, EY	0	1	NC
EU		1	1	NC
Ę	(Polish e-ogonek)	NC	NC	6 or NC
E		0	NC	NC
FB		7	7	7
F		7	7	7
G		5	5	5
H		5	5	NC
IA	IE, IO, IU	1	NC	NC
I		0	NC	NC
J	Try Y (1) and DZH (4)			
KS		5	54	54
KH		5	5	5
K		5	5	5

L		8	8	8
MN			66	66
M		6	6	6
NM			66	66
N		6	6	6
OI	OJ, OY	0	1	NC
O		0	NC	NC
P	PF, PH	7	7	7
Q		5	5	5
RZ, RS	Try RTZ (94) and ZH (4)			
R		9	9	9
SCHTSCH	SCHTSH, SCHTCH	2	4	4
SCH		4	4	4
SHTCH	SHCH, SHTSH	2	4	4
SHT	SCHT, SCHD	2	43	43
SH		4	4	4
STCH	STSCH, SC	2	4	4
STRZ	STRS, STSH	2	4	4
ST		2	43	43
SZCZ	SZCS	2	4	4
SZT	SHD, SZD, SD	2	43	43
SZ		4	4	4
S		4	4	4
TCH	TTCH, TTSCH	4	4	4
TH		3	3	3
TRZ	TRS	4	4	4
TSCH	TSH	4	4	4
TS	TTS, TTSZ, TC	4	4	4
TZ	TTZ, TZS, TSZ	4	4	4
Ţ	(Romanian t-cedilla)	3 or 4	3 or 4	3 or 4
T		3	3	3
UI	UJ, UY	0	1	NC
U	UE	0	NC	NC
V		7	7	7
W		7	7	7
X		5	54	54
Y		1	NC	NC
ZDZ	ZDZH, ZHDZH	2	4	4
ZD	ZHD	2	43	43
ZH	ZS, ZSCH, ZSH	4	4	4
Z		4	4	4
Letra	**Alternar Ortografía**	**Inicio de un nombre**	**Antes de una vocal**	**Cualquier otra situación**

APÉNDICE B

Estructura de un documento en la colección CL-SR. Tiene un identificador del número del documento, que incluye un código numérico, un segmento de identificación y un número de secuencia. Incluye nombres de personas o lugares, palabras clave extraídas manual y automáticamente, además de las transcripciones realizadas por los reconocedores automáticos.

<DOC>
<DOCNO>VHF[Code]-[SegId].[SequenceNum]</DOCNO>
<INTERVIEWDATA>Interviewee name(s) and birthdate</INTERVIEWDATA>
<NAME>Full name of every person mentioned</NAME>
<MANUALKEYWORD>Thesaurus keywords assigned to the segment</MANUALKEYWORD>
<SUMMARY>3-sentence segment summary</SUMMARY>
<ASRTEXT2003A>ASR transcript produced in 2003</ASRTEXT2003A>
<ASRTEXT2004A>ASR transcript produced in 2004</ASRTEXT2004A>
<AUTOKEYWORD2004A1>Thesaurus keywords from a kNN classifier</AUTOKEYWORD2004A1>
<AUTOKEYWORD2004A2>Thesaurus keywords from a second kNN classifier</AUTOKEYWORD2004A2>
</DOC>

APÉNDICE C

Resultados obtenidos durante los años 2005 al 2007 que se efectuó la tarea CL-SR. Las tablas no son comparables entre si, pero se ponen como referencia, para ver los resultados generales.

2005

Universidad	Campos	MAP
Ottawa	ASR04 AK1,AK2	0.1653
Maryland	ASR04 AK2	0.1288
Waterloo	ASR03,ASR04	0.1121
UNED	ASR04	0.0934
Alicante	ASR04	0.0768
Pitt	ASR04	0.0757
Dublín	ASR03,ASR04 AK2	0.0654

2006

Univer.	Campos	MAP
Dublín	ASR06B AK1,AK2	0.0733
Ottawa	ASR04,ASR06B AK1,AK2	0.0565
UMD	ASR04,ASR06B AK1,AK2	0.0543
UT	ASR04	0.0381
UNED	ASR06B	0.0376
Alicante	ASR06B	0.0375

2007

Univer.	Campos	MAP
Ottawa	ASR04 AK1,AK2	0.0855
Dublín	ASR06B AK1,AK2	0.0787
Brown	ASR06B AK1,AK2	0.0785
UC	ASR06B AK1,AK2	0.0571
UVA	ASR06B AK2	0.0444

APÉNDICE D

Ejemplos de tópicos a consultar en las competencias del 2006 y 2007. En los siguientes se muestra tanto el Título como la descripción de lo que se desea obtener. Estos campos son los únicos que se podían utilizar para reportar una corrida estándar. Estas consultas también cuentan con un número que las identifica y un campo donde se realiza la narración a detalle de lo que se pretende recuperar con la consulta. Sin embargo la narración no se puede tomar si se quiere comparar con otros sistemas.

<num>1133</num>
<title>Varian Fry</title>
<desc>The story of Varian Fry and the Emergency Rescue Committee who saved thousands in Marseille</desc>
<narr>Varian Fry, a young American journalist, created an underground operation that smuggled more than 2,000 refugees (including Marc Chagall, Max Ernst, and Andre Breton) out of Vichy France in 1940-1941. The relevant material should contain information about this operation. Any first-hand information of people who have been rescued by Fry is highly relevant</narr>

<num>1159</num>
<title>Child survivors in Sweden</title>
<desc>Describe survival mechanisms of children born in 1930-1933 who spend the war in concentration camps or in hiding and who presently live in Sweden.</desc>
<narr>The relevant material should describe the circumstances and inner resources of the surviving children. The relevant material also describes how the war-time experience affected their post-war adult life.</narr>

<num>1166</num>
<title>Hasidism</title>
<desc>Hasidim and their unquestioning faith</desc>
<narr>The relevant material should talk about Hasidism before, during, and after the Holocaust. The information about Hasidic dynasties and geographic localities that were established and destroyed.</narr>

<num>1173</num>
<title>Children's art in Terezin</title>

<desc>We are looking for the description of the art-related activities of children in Terezin such as music, plays, paintings, writings and poetry.</desc>
<narr>The relevant material should include discussions of such activities and how they influenced the survival and following life of the children. Any episodes where the interviewee demonstrates examples of such an art are highly relevant.</narr>

<num>1179</num>
<title>Bulgaria saved its Jews?</title>
<desc>We are looking for material that will support or rebuff the claim that Bulgaria saved its Jews from Nazism</desc>
<narr>Stories of Jews who were rescued or perished in Bulgaria and any related information about Jews in the Bulgarian society. Includes, but is by no means limited to, the following topics: how Bulgarian Jews define themselves with respect to Bulgarian society; how the fate of the Bulgarian Jewish community was different from other Jewish communities in Europe.</narr>

<num>1181</num>
<title>Sonderkommando in Auschwitz</title>
<desc>Stories of all people who came into direct or indirect contact with Sonderkommando in Auschwitz</desc>
<narr>There were two groups of Sonderkommando who maintained their dignity: historians who documented and preserved the diary of the tragedy and fighters who succeeded in revolting on 7 October of 1944. The material that deals with one of these groups of Sonderkommando is relevant.</narr>

<num>1185</num>
<title>Doctors and Nurses in the Holocaust</title>
<desc>The Good Doctors and Nurses in the Holocaust. Ethical dilemmas that confronted health-care professionals seeking to do good in the midst of evil.</desc>
<narr>These could be Jewish or non-Jewish doctors providing care in different environments: prisons, concentration camps, resistance, hiding places. Care could be formal or informal.</narr>

APÉNDICE E

Tabla con 80 palabras vacías (sólo una porción) que fueron eliminadas en el segundo paso de la codificación fonética de los documentos, usando Soundex.

Codificación Fonética	Texto
A00000	a,ah,au,aha,awe,away
A14000	awful,appel,apple,apply,apollo,appeal,appell,awfully,avail,able
A16300	afford,apart,abroad,appeared,afraid,aboard,abort
A16355	apartment,apartments
A20000	ac,ag,ak,as,ax,a's,aeg,age,ago,asa,ash,ask,ass,aug,aux,axe,aische,ache,akcio,asi,asks,auch,awake,awash
A23200	actious,assets,auschwitz,acts
A23400	actually,actual
A40000	al,allah,alley,allow,ale,ali,all,aliyah,awhile,alla,ally
A41000	alibi,alive
A51300	anybody
A53000	annuity,ahmed,aimed,ain't,and,annoyed,aneta,anita,amid,andy,anti,auntie,aunt,Anette
A55600	anymore
A60000	air,are,array,arrow,area,aria,arre,aware
A61300	arrived,arafat,arbeit,arbite
A65000	aaron,arm,ayran,arena,armée,armia,aryan,army,aron
B00000	b,be,by,baia,bay,bea,bee,bei,bow,boy,buy,bye
B20000	bacau,backs,badge,baeck,baggy,beach,becky,batch,bach,back,bags,bake,base,bash,bass,baus,bays,biecz,beck,bees,bias,bike,boca,book,boss,bows,boxy,boys,bitch,buck,bugs,bush,busk,busy,buys,bucks,buicks,budge,buggy,busch,bushy,b's,bag,bbc,beg,big,box,bug,bus,bychawa,bessie,boise,books,boy's,bojowa
B23000	baked,botched,beast,based,bassett,best,bogota,beset,backseat,backed,backside,bust,booked,begged,bashed,basket,basset,bicsad,beside,bucket,boost,bayside,buzzed,biscuit,biased,boycott,budget
B25200	bashing,baking,begins,basing,basins,bushings,biking,bygones,business,boxing,backing,begging,becomes,bagging,backswing,busing,beijing
B25520	begining,beginning,becoming,beginnings
B30000	bathe,bath,beat,beet,beta,beth,betty,bite,boat,body,boot,both,bout,buddy,bad,bat,bed,bet,bid,bit,but,buyout,bought,beauty
B36000	betar,battery,butter,bother,batory,better,beater,bitter
B41000	bellevue,balf,bulb,bolivia,behalf,belief,beloff,bluff,believe
B42000	bales,balls,bells,balk,biology,bills,bloc,blacks,bulk,bleach,belushi,bulge,bulky,bulls,bielski,bielsko,bullish,belsky,boils,bilcze,bill's,bowls,biloxi,bailey's,bellies,billy's,blocks,black,blaze,bleak,bless,blouse,bliss,bayless,block,blows,blues
B43520	bleeding,balloting,buildings,building,billeting

62000	berries,bears,barrage,barricks,beers,barak,barco,barcy,barge,barks,bruise,bark,bars,baruch,barry's,birch,broz,bereza,burke,barshaw,borrows,brace,bragg,brass,buyers,break,brick,brisk,brock,broke,bruce,bruck,brush,brooks,bryce,burgau,burris,boris,bourg,bearish,breach,breaks,breaux,breeze,bricks,bridge,barracks,bayerische
B63000	b'rith,beard,barrett,brought,bard,bart,beret,berta,breathe,barred,bird,birth,brat,brit,brod,burt,bertha,bertie,bayreuth,brady,bread,breed,brett,bride,borrowed,broad,brody,buried,board,bordo,bored,burrito,breath,bright,brewed,broadway
B63600	barter,brighter,breather,brother,broader,border
B63620	brothers,borders,brother's
B65000	baron,barn,barney,bern,barren,born,brno,burn,bernie,brougham,burma,byrne,byron,brain,brian,bromo,broom,brown,bruno,bronia,borne
B65200	,barns,barring,brunch,barnes,barons,brainwash,burns,borrowing,brens,bring,brink,bronx,burmese,bronco,bronze,brooms,browns,bearings,brownish,burying,baranowicze,brewing,brains,branch,brings,boring
C14000	chapelle,couple,covel,coppell,cabell,coppola,cable,chapel,civil
C20000	cashew,czech,cage,cake,casa,case,cash,chic,ciuc,cox,coax,coca,cock,coke,cook,cows,cuss,coach,cuckoo,cocoa,choice,cokie,cooks,choose,couch,chaos,chase,check,cheek,chess,chock,choke,chose,cease,checks,cheeks,cheese,cheesy,casey,catch,cause,cookie,cissy
C25000	casino,cushion,chisinau,chechen,cosmo,chosen,cousin,cocaine,chicken,chasm,chechnya,chazen,ciechanow
C30000	coughed,co
C42000	,colleague,calais,calloway's,celica,cluj,collage,college,coals,cools,clocks,chalk,clash,class,claus,claws,click,cloak,clock,clogs,close,clues,calls,challahs,cells,chelsea,chiles,chills,coalesce,clicks
C43000	claudio,called,clot,cold,collide,could,cloudy,child,colette,cloth,cloud,clout,cooled,chilled,claude
C43200	celtics,celtic,colds,colts,cloths,clouds,childish,clothes,childs,child's,cleats
C43530	couldn't
C43650	coltrane,children
C50000	cuomo,chimney,came,cane,chin,chun,can,com,con,cohn,coin,coma,comb,come,cone,cohen,coney,chaim,chain,china,connie,chuni,cheine,cheney,chiune
C51000	camp,comp,comfy,compu,champ,convey,convoy,canopy
C52000	chimneys,chunks,cynic,cans,comics,cons,cynics,china's,chinese,coins,combs,comes,comic,commish,congo,cuomo's,connick,coughing,chiemsee,chums,chung,chewing,chanukah,canoes,chains,cames,camps,canes,champs,chance,chancy,change
C53000	committee,combat,cent,comedy,commit,cynthia,comet,conde,conte,county,count,commute,chant,canada,canned,cannot,canoed,cambodia,cyanide,camed,can't,candy,cindy,chained
C53600	counter,country,contier,contour,csendor,cemetery,commodore,century,commuter,center,centre,cinder
C55200	combines,canyons,canning,cannons,cummings,coming,commons,commence,combing,cinemax,canons,cummins
C55300	combined,command,commend,comment,commando,cement,community
C60000	curry,care,char,cher,car,cry,core,crew,crow,cure,corey,chair,cheer,choir,cairo,carei,carry,cherry
C62000	cherries,curse,church,curries,crack,craig,crash,crass,crazy,creek,crews,chorzow,cries,cherokee,croak,crock,crook,cross,crows,crotch,crush,coerce,carries,carriage,cars,cork,cherish,crux,cruz,chores,chorus,course,cracks,cracow,creche,creeks,chriscurious,chairs,cruise,car's,cares,cargo,cheers,charge,crutch,courage

C63000	cheered,cured,cohort,croatia,crate,creed,cried,charade,croat,crowd,charity,crude,carried,card,cart,court,create,chart,cruddy,cared,carrot,chaired
D00000	d,da,de,di,do,day,ddt,dia,die,doe,dow,due
D16530	different,deformed
D20000	dc,dauchau,docks,dodge,d's,dej,des,dig,doc,dog,dos,dug,deutch,daisy,decks,dash,days,dachau,deck,desk,day's,diaz,dice,dick,dies,disc,dish,dicey,diego,doch,dock,does,dogs,duck,dues,duke,dyke,ditch,dixie,dizzy,ducks,dutch
D23300	dictate,decided
D23600	doctor,destroy
D30000	dwight,dad,det,did,dot,daddy,death,data,date,dati,dead,deed,died,diet,dude,duty,ditto
D35300	detained,didn't
D52000	diagnose,dummies,damage,damask,dean's,doing,domes,downs,demise,danish,dance,danes,dams,denies,denise,demos,dennis,dense,dimes,dingy,dunk,dinesh,dumps,dunes,dying
D53000	d'amato,don't,damed,denied,dent,doomed,dined,dawned,dignity
D60000	dr,dowry,der,dry,dairy,darrow,dare,dear,diary,dior,dire,diarrhea,door,dora,dory,doorway,draw,drew
E00000	e,eh,eye
E16130	everybody
E21652	experience's,experiencing,experience,experienced,experiences
E30000	ed,et,eddie,eat,eight,eighth,eighty,eddy,eyed
E52420	enclose,english,
E53000	end,embody,empty
F13000	fifth,fifty
F23000	facade,faced,facet,facto,fact,fast,faked,feast,fist,fixed,fished,fetched,faucet,fiesta,
F23600	factory,factor,foster,fixture,faster
F30000	faith,fade,fate,feat,feed,feet,feud,fiat,fatti,fatty,food,foot,fiato,fight,fat,fed,fit,phd,fythe,fought,photo
F36000	footwear,future,fodder,feather,feature,father,fighter
F36200	father's,fathers,feathers,fighters,features
F40000	fail,fall,feel,fell,file,fill,flea,flee,flew,flow,flue,fella,fool,foul,fuel,full,follow,fully,fly,fellow,phil,foley,folly
F43000	fleet,flood,floyd,fluid,fooled,felt,fault,flat,fled,fueled,fold,field,flight,filed,filet,filth,followed,failed,filled,filthy,faulty,fallout
F54000	finale,final,finally,family,funnily,female
F60000	faire,fairy,fair,fare,fear,fire,fore,four,free,frye,fiery,ferry,fury,fewer,freeway,far,for,fro,fry,fur,faraway,foyer
F62300	freaked,first,freezed,foresight,forrest,frost,phrased,frigid,forced,forest,forged,forget,forgot
F63000	firewood,fared,ford,fort,fred,freight,fired,fraught,fredy,freed,fried,frito,fruit,fürth,feared,forehead,fourth,freddy,forte,forth,forty,freida,frayed,friday,fright,freddie
F65000	foreign,farm,faron,firm,form,from,frum,firma,frame,freon,frown,forum
F65200	faring,frying,ferenc,farms,formica,freeing,firing,firms,ferrence,frankie,frank,franz,frames,france,franco,francs,franka,franks,forms,furnace,pharmacy,french,frenzy,furnish,fearing,farmhouse,fringe,forums,fairness