Universidad de Valladolid

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
DEPARTAMENTO DE INFORMÁTICA

TESIS DOCTORAL:

Estrategias para el Acceso a Contenidos Web Mediante Habla

Presentada por D. César González Ferreras para optar al grado de doctor por la Universidad de Valladolid

Dirigida por:
Dr. Valentín Cardeñoso Payo

A mis padres.

Agradecimientos

Esta tesis ha sido realizada en el grupo de investigación ECA-SIMM del Departamento de Informática de la Escuela Técnica Superior de Ingeniería Informática de la Universidad de Valladolid.

La Consejería de Educación de la Junta de Castilla y León y el Ministerio de Ciencia y Tecnología han financiado parcialmente la investigación realizada.

En primer lugar quiero agradecer a Valentín Cardeñoso por su esfuerzo y dedicación. En todo momento ha sabido orientarme y motivarme para seguir adelante. También le agradezco todas las tareas de gestión que se ha encargado de hacer y que han servido para poner a mi disposición todos los recursos necesarios.

A David Escudero, que ha sido un gran apoyo en los momentos difíciles y que siempre ha tenido la palabra de ánimo necesaria.

A Rubén San-Segundo por su ayuda en el desarrollo del sistema de diálogo hablado. Su experiencia y sus consejos hicieron que el sistema mejorara notablemente.

A Ronald Cole por acogerme en la Universidad de Colorado en Boulder. A Bryan Pellom por dejarme SONIC, un software con el que se han hecho parte de los experimentos de esta tesis.

A Pablo de la Fuente y a Joaquín Adiego, por su ayuda en los experimentos de recuperación de información y por dejarme su motor de recuperación de información.

A Emilio Sanchis por acogerme en Valencia, donde la tranquilidad de la que disfruté me dio las fuerzas necesarias para finalizar este trabajo.

A todos los compañeros del Departamento de Informática, por sus ánimos, su apoyo y sus consejos. Especialmente a Diego Llanos, Arturo González, Carlos Enrique Vivaracho y Alejandra Martínez.

Abstract

The objective of the thesis is to design and evaluate different strategies to access web contents using speech. The work has focused on reusing existing web contents and on providing a spoken interaction that allows the user to access the contents quickly and in a user friendly way. The work has been divided in three parts. In the first phase the problem of generic conversion of web contents for their access using a vocal browser has been analyzed and two approaches have been proposed: automatic conversion and semiautomatic conversion. In both cases, the way of access is limited by the way in which the original web contents are structured. In the second phase the use of a spoken dialog system has been proposed to access web contents in restricted domains. The approach is based on an information model, which describes how web contents must be processed and structured, and on an interaction model, which describes how the system dialogs with the user using browse and search. A system that allows access to a web site of a digital newspaper using speech has been built. However, some limitations have been detected in the search strategy, caused by speech recognition errors. This has motivated a third and last phase in which several experiments have been carried out with a speech driven information retrieval system. Some improvements have been proposed in order to increase system performance: dynamic adaptation of vocabulary and language model; the use of pseudo-relevance feedback in the information retrieval engine; and the inclusion of pronunciation of English words. The final experiments have shown the feasibility of building speech driven information retrieval systems, although the performance is not as good as using text in the input.

Resumen

El objetivo de la tesis es diseñar y evaluar diferentes estrategias para el acceso a contenidos web empleando habla. El trabajo se ha centrado en la reutilización de los contenidos web existentes y en plantear la interacción hablada de manera que el usuario pueda acceder a los contenidos de manera rápida y amigable. El trabajo realizado se ha dividido en tres partes. En la primera fase se ha analizado el problema de la conversión genérica de contenidos web para su acceso a través de un navegador vocal y se han propuesto dos alternativas: la conversión automática y la conversión semiautomática. En ambos casos, la forma de acceso está condicionada por la manera en la que están estructurados los contenidos web originales. En la segunda fase se ha planteado la utilización de un sistema de diálogo hablado para el acceso a contenidos web en dominios restringidos. La propuesta está basada en un modelo de información, que describe cómo se deben procesar y estructurar los contenidos web, y en un modelo de interacción, que describe cómo el sistema dialoga con el usuario empleando navegación y búsqueda. Se ha construido un sistema que permite acceder al sitio web de un periódico digital empleando habla. Sin embargo, se han detectado ciertas limitaciones en la estrategia de búsqueda, provocadas por los errores de reconocimiento del habla. Esto ha motivado una tercera y última fase en la que se han realizado diversos experimentos con un sistema de recuperación de información dirigida por habla. Se han propuesto varias mejoras que permiten incrementar el rendimiento del sistema: la adaptación dinámica del vocabulario y del modelo de lenguaje; la utilización de realimentación por pseudo-relevancia en el motor de recuperación de información; y la inclusión de la pronunciación de palabras en inglés. Los experimentos finales han demostrado la viabilidad de construir sistemas de recuperación de información dirigida por habla, aunque el rendimiento no es tan bueno como el obtenido al emplear texto como entrada.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

- HMM: modelo oculto de Markov (*hidden Markov model*).
- IDF: frecuencia inversa de documento (*inverse document frequency*).
- IR: recuperación de información (*information retrieval*).
- LM: modelo de lenguaje (*language model*).
- MAP: precisión media promediada (*mean average precision*).
- MFCC: coeficientes cepstrales en las frecuencias de mel (*mel frequency cepstral coefficients*).
- NE: entidades nombradas (*named entities*).
- OOV: fuera del vocabulario (*out of vocabulary*).
- RTF: factor de tiempo real (*real time factor*).
- SDR: recuperación de documentos hablados (*spoken document retrieval*).
- TF: frecuencia de término (*term frequency*).
- TF-IDF: frecuencia de término-frecuencia inversa de documento (*term frequency-inverse document frequency*).
- WER: tasa de error de palabra (*word error rate*).

Parte I

Introducción

Capítulo 1

Introducción

1.1. Motivación

El crecimiento y popularidad de la web como repositorio universal de información se incrementa día a día. Hay una gran cantidad de información disponible en línea y una diversidad de contenidos tan grande que se pueden encontrar documentos sobre cualquier dominio. La web se ha convertido, sin duda, en el almacén de información más importante y más usado de cuantos existen. Por tanto, emplear la web para acceder a la información es ya algo habitual, tanto para trabajar como por placer o entretenimiento. Por ejemplo, un estudio realizado a finales de 2003 en Estados Unidos [41] reveló que el 88 % de los usuarios de Internet afirma que Internet juega un papel importante en sus rutinas diarias y el 92 % aseguran que Internet es un buen lugar para obtener información del día a día.

Sin embargo, la ingente cantidad de información, su variedad y su dinamismo plantean barreras para una explotación adecuada, principalmente en dos direcciones importantes: la representación y el acceso. En el primer caso, las limitaciones surgen debido a que la mayoría de los contenidos web están diseñados para ser visualizados empleando un navegador, y no para ser manipulados por programas de ordenador. La posibilidad de procesar automáticamente los contenidos web facilitaría la utilización de la información en distintos contextos. Las soluciones propuestas hasta ahora van en la dirección de lo que se ha denominado la web semántica, que se propone como una extensión de la web actual en la que la información tiene un significado bien definido [13]. Aunque todavía no se ha producido la adopción a gran escala, se han realizado avances significativos [122]. En el segundo caso, existe un amplio número de usuarios que encuentra serias dificultades en el acceso a la web. Esto ha motivado que en los últimos años se haya realizado un gran esfuerzo en accesibilidad web, para intentar favorecer que los sitios web puedan ser visitados y utilizados de forma satisfactoria por el mayor número posible de personas, independientemente de las limitaciones personales que tengan. Además, la accesibilidad web beneficia a todos los usuarios, puesto que favorece que los sitos web sean más usables. La solución más extendida es emplear las

pautas de accesibilidad desarrolladas por el W3C [21].

Por otro lado, la gran mayoría de la información disponible en la web se ha creado para ser accedida empleando una interfaz gráfica de usuario, mediante un navegador web y, por ello, el énfasis está en la apariencia visual, ignorando los beneficios potenciales de otras modalidades. El habla tiene ciertas ventajas frente a las interfaces gráficas de usuario tradicionales porque es una forma natural de interacción para la mayoría de las personas. La gente está acostumbrada a emplear este medio, que sigue siendo el más usado para la comunicación del día a día. Por tanto, el uso del habla para acceder a contenidos web puede enriquecer la experiencia del usuario. Como beneficio adicional, puede ser apropiado para ciertos entornos y situaciones en las que no se puede emplear la interacción visual, como por ejemplo conduciendo. Se ha comprobado que la conducción es más segura si se utiliza el habla para controlar los dispositivos de un vehículo en lugar de entrada manual [95]. Asimismo, el habla resulta adecuada para los usuarios invidentes, puesto que existen estudios que han puesto de manifiesto las dificultades que tienen este tipo de usuarios para acceder a los sitios web [128]. Por todo ello, el acceso a los contenidos web empleando habla abre nuevas oportunidades de acceso y explotación de la información, y por tanto, es uno de los retos que deben ser abordados para incrementar la usabilidad de la web.

Otro aspecto cada vez más relevante es el uso de dispositivos móviles para el acceso a la web. El empleo de habla para buscar contenidos web es especialmente útil en estos pequeños dispositivos, debido a que resulta incómodo introducir las consultas empleando el teclado y a que sus pequeñas pantallas dificultan la búsqueda de información [52]. El habla puede usarse sola en un sistema de diálogo hablado o combinada con una interfaz gráfica de usuario en un sistema de diálogo multimodal. Aunque actualmente el mercado de las tecnologías del habla se centra en los centros de atención telefónica a usuarios, dentro de poco los sistemas multimodales en dispositivos móviles dominarán el mercado [84].

En cuanto a la tecnología del habla, cabe destacar los avances realizados a los largo de los últimos años. En un principio el mayor esfuerzo se centró en la mejora de las prestaciones de los módulos de reconocimiento y síntesis de habla. Posteriormente, el interés se ha ido desplazando hacia la forma de gestionar el diálogo con el usuario. Actualmente, existen sistemas de diálogo hablado capaces de interactuar con los usuarios de manera natural y flexible, empleando lenguaje natural y alcanzando altas tasas de rendimiento [81, 125, 140]. Estos sistemas están diseñados para tareas específicas y permiten el acceso a información estructurada a través de la línea telefónica. Al mismo tiempo, la industria de los sistemas de diálogo hablado ha alcanzado un alto grado de madurez, caracterizado por una estructura vertical de empresas que venden tecnología, integradores de plataformas, desarrolladores de aplicaciones y compañías de alojamiento de aplicaciones [104]. El éxito de los sistemas automáticos de diálogo hablado está propiciando su progresiva implantación en centros de atención telefónica a usuarios, donde cada vez se utilizan más. Un informe de DMG Consulting [35] estima que los ingresos en todo el mundo de la tecnología de repuesta telefónica interactiva alcanzaron los

1.867 millones de dólares en 2007 y predice que superará los 2.400 millones de dólares en 2010.

En resumen, todo esto hace pensar que el acceso a la web mediante habla será una realidad en un futuro cercano. Existen ya iniciativas en esta dirección, aunque las soluciones propuestas son simples y la interacción resultante dista mucho de ser intuitiva y amigable. Además, en algunos casos, suelen estar restringidas a sitios concretos con una estructura previamente fijada. El problema principal a superar es el cambio de modalidad, que obliga a plantear la interacción de otra manera, debido a las diferencias entre la interfaz gráfica de usuario y la interfaz hablada.

1.2. Objetivos

El objetivo general de la tesis es:

Proponer diferentes estrategias para el acceso a contenidos web empleando habla y analizar los resultados obtenidos con cada una de ellas.

Como objetivos específicos planteamos los siguientes:

- Favorecer la reutilización de los contenidos web existentes.
- Procesar automáticamente los contenidos web para adecuarlos al canal de comunicación hablado.
- Proponer un método general para la navegación de contenidos web que permita el acceso mediante habla a cualquier página web.
- Construir un sistema de diálogo que permita el acceso a los contenidos de un sitio web concreto.
- Experimentar las posibilidades de la recuperación de información dirigida por habla.
- Evaluar los sistemas con usuarios reales para medir la usabilidad y comprobar la validez de las propuestas realizadas.

1.3. Problemática

El acceso a contenidos web empleando habla no es una tarea sencilla, ya que requiere un cambio profundo en la forma de plantear el diseño de la interfaz de usuario. La interacción visual y la interacción vocal son dos modalidades con características muy dispares, por lo que es imprescindible una reestructuración de los contenidos para adaptarlos a la nueva modalidad, muy distinta de la visual. El problema es verdaderamente complejo, puesto que se deben superar las limitaciones

impuestas tanto por la naturaleza de los documentos web, como por las dificultades de construir sistemas de diálogo hablado que interactúen adecuadamente con los usuarios. Desde el punto de vista de la información, las limitaciones surgen debido a la gran variabilidad de páginas web existente y debido a la ausencia de metadatos que describan la información y que faciliten el procesamiento automático de la misma. Desde el punto de vista de los sistemas de diálogo hablado, las características del canal hablado son muy diferentes de las de uno visual, lo que provoca que la interacción deba plantearse de manera totalmente diferente.

Para poder establecer un diálogo con el usuario es preciso procesar la información de las páginas web y extraer la estructura conceptual. Esto es una tarea complicada, debido a que los autores de contenidos web ponen el énfasis en la presentación visual en lugar de en la correcta estructuración de la información. Aunque en un principio HTML se diseñó para describir contenidos, el uso y la evolución han puesto más énfasis en la parte visual. Como resultado, HTML se ha convertido en un lenguaje para la presentación visual de contenidos en el que se ignora en muchos casos la correcta estructuración de la información. La ausencia de metainformación que describa los contenidos complica el procesamiento automático de la misma por parte del ordenador. El problema se agrava debido a la variedad de páginas web, que complica el tratamiento y la extracción de la información. Cada diseñador web emplea estrategias diferentes, y es complejo establecer mecanismos genéricos que funcionen para todas las páginas.

Por otra parte, la información textual tiende a ser masiva, y mostrarla en una navegador web tradicional no es un problema, porque toda la información se presenta de una vez y el usuario elige la parte que le interesa. Sin embargo, la interfaz hablada es secuencial y no persistente, y por tanto, hay que limitar la información que se envía al usuario y proporcionar mecanismos que faciliten el acceso a la misma. Por tanto, no es posible una mera presentación de contenidos, es preciso encontrar una manera eficiente y natural de plantear la interacción.

Por último, para permitir el acceso mediante habla a un sitio web será preciso construir un sistema de diálogo hablado que interactúe con el usuario. El problema surge debido a que la información de la web es, en la mayoría de los casos, texto libre que carece de la estructura necesaria. Los sistemas de diálogo hablado tradicionales están diseñados para acceder a información estructurada almacenada en una base de datos y, por tanto, deben ser adaptados para acceder a información poco estructurada, texto en lenguaje natural en la mayoría de los casos.

1.4. Planteamiento General

En esta tesis se estudian diferentes alternativas que permiten el acceso a contenidos web empleando habla. El enfoque que se ha elegido pretende dar una visión global del problema desde distintos puntos de vista. De esta manera, se ha planteado el trabajo de lo general a lo particular, de modo que nuestros primeros esfuerzos fueran exploratorios de la problemática en su conjunto y a continuación

se fuera centrando la tarea en problemas más específicos y particulares.

El trabajo realizado se ha dividido en tres fases. En la primera fase se analiza el problema de la conversión de contenidos web de manera general, permitiendo el acceso mediante habla a cualquier página web. Este enfoque está fuertemente limitado por la estructura del sitio web original. En la segunda fase, con el fin de superar esta limitación y conseguir mejores resultados, se plantea la utilización de un modelo de información que permita una mayor libertad a la hora de plantear la interacción hablada con el usuario. La idea es utilizar un sistema de diálogo hablado para el acceso a contenidos web en dominios restringidos. Para ello, además del modelo de información, se utiliza un modelo de interacción que describe cómo el sistema dialoga con el usuario empleando navegación y búsqueda. Los resultados obtenidos mostraron ciertas limitaciones en la estrategia de búsqueda, debido principalmente a errores de reconocimiento del habla. Esto ha motivado la tercera y última fase, donde se plantean una serie de experimentos con un sistema de recuperación de información dirigida por habla.

1.5. Organización de la memoria

Hemos dividido la presente memoria en cuatro partes:

- En la parte I, capítulo 1, se ha presentado la introducción de la tesis, que incluye la motivación, los objetivos, la problemática y el planteamiento general del trabajo desarrollado.

- En la parte II se presenta un estudio del estado de la cuestión.

  En el capítulo 2 se describen en detalle los sistemas de diálogo hablado. Se hace una descripción de los componentes básicos de estos sistemas, de las diversas metodologías y herramientas para su desarrollo y de las distintas técnicas de evaluación que se pueden emplear para medir su rendimiento.

  En el capítulo 3 nos centramos en la recuperación y extracción de información. Por un lado, se describe en detalle el modelo vectorial de recuperación de información. Por otro lado, se presentan diversas técnicas de procesamiento automático de contenidos web, como son la conversión de contenidos web y la extracción de información de páginas web.

  En el capítulo 4 se hace una revisión de los distintos trabajos directamente relacionados con el tema de la tesis, el acceso a contenidos web mediante habla. En unos casos se proponen soluciones generales, válidas para cualquier página web. En otros casos las soluciones se plantean para dominios restringidos. También es posible plantear el problema como una búsqueda de información, lo que da lugar a la recuperación de información dirigida por habla.

- En la parte III se presentan las distintas soluciones propuestas para permitir el acceso a contenidos web empleando habla.

En el capítulo 5 se hace una propuesta para la conversión de contenidos web que permite el acceso mediante habla a cualquier página web. La idea es realizar la conversión de contenidos web escritos en HTML a diálogos escritos en VoiceXML. Se plantean dos enfoques diferentes: un enfoque automático, en el que el sistema trata de inferir la estructura inherente de las páginas HTML y en base a ella plantear la interacción hablada con el usuario; y un enfoque semiautomático, en el que es preciso que un desarrollador construya una aplicación vocal en la que se especifica cómo llevar a cabo la conversión de contenidos web.

En el capítulo 6 se propone el empleo de un sistema de diálogo hablado para el acceso a contenidos web en dominios restringidos. La propuesta está basada en un modelo de interacción, que describe cómo el sistema dialoga con el usuario, y en un modelo de información, que describe cómo se deben procesar y estructurar los contenidos web. Para demostrar la validez de la propuesta se presenta un sistema que permite acceder a la versión digital de un periódico empleando habla. Se presentan también los resultados obtenidos en la evaluación de usabilidad del sistema, donde se midió el rendimiento y la satisfacción de los usuarios.

En el capítulo 7 se describe un sistema de recuperación de información dirigida por habla, que permite emplear el habla como la entrada de un motor de recuperación de información. El objetivo es permitir a los usuarios buscar información en una colección de documentos de texto empleando preguntas habladas en lenguaje natural. Se estudian cuáles son los factores que más inciden en la degradación del sistema y se plantean una serie de propuestas de mejora que permiten incrementar el rendimiento del sistema.

- En la parte IV, capítulo 8, se presentan las conclusiones y el trabajo futuro.

Parte II

Estado de la cuestión

Capítulo 2

Sistemas de diálogo hablado

2.1. Introducción

El habla es un medio de comunicación natural para los seres humanos, de ahí que la comunicación con las máquinas empleando habla natural haya sido una meta y un reto recurrente a lo largo de varias décadas. Para ello, se han construido los sistemas de diálogo hablado, que permiten la interacción con los sistemas informáticos empleando lenguaje natural hablado.

La construcción de sistemas de diálogo hablado es una tarea compleja, debido a que es preciso resolver problemas de muy diversa índole: tratamiento de señal vocal (síntesis y reconocimiento del habla), procesamiento del lenguaje natural (comprensión y generación de lenguaje natural), y gestión y planificación del diálogo con el usuario.

Los sistemas de diálogo hablado han experimentado un gran desarrollo en los últimos años. En un primer momento el énfasis se encontraba en mejorar los módulos de síntesis y reconocimiento del habla. A medida que la tecnología ha evolucionado y estos módulos ofrecían unos resultados aceptables, se ha ido desplazando el interés hacia la forma de gestionar el diálogo con el usuario. Un buen resumen de esta evolución histórica puede encontrarse en [94], donde se realiza una revisión del estado de la cuestión de los sistemas de diálogo hablado. Actualmente sigue siendo un problema abierto, y es preciso mejorar los sistemas existentes para obtener una interacción más natural, y de esta manera conseguir la aceptación de este tipo de sistemas por parte de los usuarios. El usuario no percibe como natural un sistema que sólo responde a ciertas palabras clave, y que obliga a que el diálogo se ajuste a una serie de patrones predefinidos. Se pretende, por tanto, que el sistema sea capaz de entender todo lo que le diga el usuario y sea este último el que lleve la iniciativa, guiado adecuadamente por el sistema.

De entre todos los sistemas de diálogo que permiten una interacción en lenguaje natural podemos destacar los más representativos: JUPITER, un sistema de diálogo hablado que proporciona información meteorológica y que ha sido desarrollado en el MIT [140]; ARISE, que proporciona información sobre horarios de tren

e información relacionada [81]; el sistema ELVIS para la consulta del correo electrónico a través del teléfono [134]; SENECA, un sistema de diálogo para coches que permite controlar el sistema de navegación, el teléfono, el CD, el aire acondicionado y otros dispositivos [95].

En España también se han llevado a cabo esfuerzos para desarrollar sistemas de diálogo hablado, como por ejemplo BASURDE [130] y DIHANA [65] para la información de horarios de tren, y E-MATTER [11] y TelCorreo[108] para acceso al correo electrónico.

Por último, cabe destacar la aparición de estándar VoiceXML [93], cuyo objetivo es facilitar la construcción de aplicaciones habladas de respuesta telefónica. Se trata de un lenguaje de marcas que pretende ser el equivalente vocal de las páginas HTML. En las páginas VoiceXML se describe cómo debe llevarse a cabo el diálogo con el usuario para la consecución de las tareas. La aparición de este estándar ha sido el punto de partida para la proliferación de sistemas comerciales que permiten la automatización de ciertas tareas en los centros de atención telefónica de usuarios.

En esta tesis se plantea la utilización de sistemas de diálogo hablado para permitir el acceso a contenidos web. Aunque los sistema de diálogo hablado se han utilizado con éxito para otro tipo de tareas, será necesario adaptarlos a las características de la información disponible en la web.

En este capítulo se describen con detalle los sistemas de diálogo hablado. En primer lugar se analiza cuáles son los componentes básicos y se describe la función de cada uno de ellos. A continuación se presentan las metodologías empleadas para el desarrollo de sistemas de diálogo hablado, las arquitecturas más representativas y las herramientas disponibles para el desarrollo de este tipo de sistemas. Se enumeran también los dominios de aplicación más habituales. Posteriormente, se introduce brevemente el estándar VoiceXML, cuyo objetivo es facilitar la construcción de sistemas interactivos de respuesta telefónica. Por último, se presentan las técnicas de evaluación empleadas para medir el rendimiento de los sistemas de diálogo hablado.

2.2. Componentes de un sistema de diálogo hablado

Un sistema de diálogo hablado es un sistema informático que permite a los usuarios interactuar con una aplicación software mediante el uso de lenguaje natural hablado. En general, los sistemas de diálogo hablado están orientados a la consecución de una tarea y permiten la interacción hablada con los usuarios en dominios restringidos. Este tipo de sistemas se organizan en torno a los siguientes bloques funcionales [73] (ver figura 2.1):

- **Reconocedor de habla**: se encarga de convertir la entrada hablada del usuario en una cadena de texto. La tarea no es sencilla y debe enfrentarse a la variabilidad lingüística, la variabilidad de hablantes y la variabilidad de me-

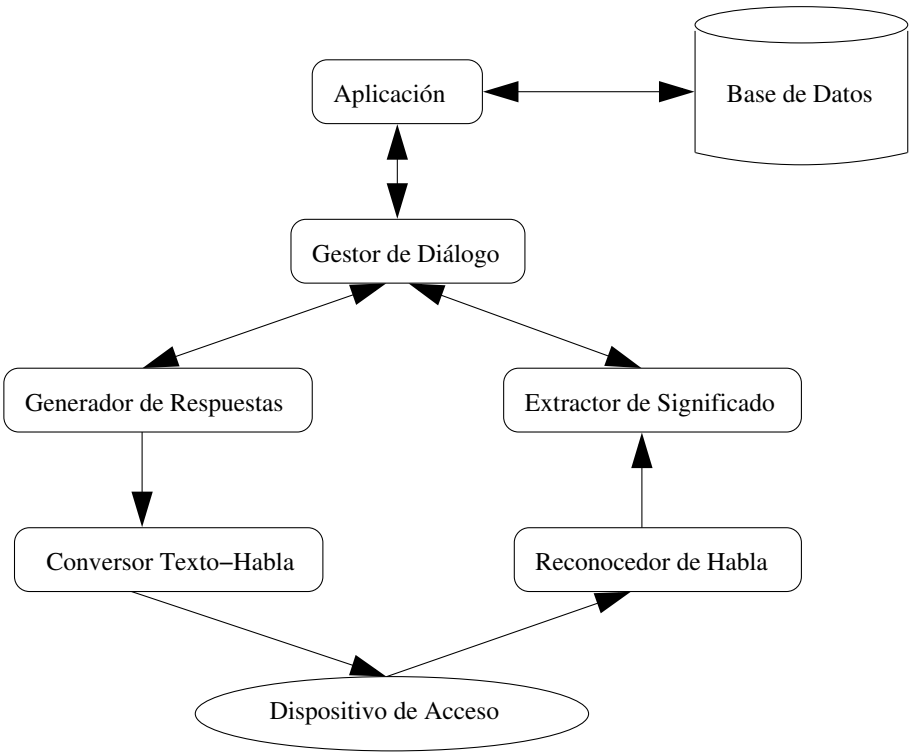


Figura 2.1: Componentes de un sistema de diálogo hablado.

dios de comunicación o canales. El reconocimiento del habla se describe en detalle en el apartado 2.2.1.

- **Extractor de significado**: analiza la semántica de la cadena dada por el reconocedor del habla, extrayendo el significado y representándolo de manera que pueda ser utilizado por el gestor de diálogo. Para ello, es preciso disponer de una representación interna del significado que permita un adecuado procesamiento de la información. Una de las representaciones más populares es la de marcos (*frames*). En esta representación, cada concepto se representa mediante un conjunto de atributos con unos determinados valores.

  La forma de llevar a cabo la extracción de significado es muy variada, y en general, dependiente del sistema de diálogo y del dominio elegido. Sin embargo, en todos los casos suele ser necesario un análisis sintáctico y un análisis semántico.

- **Base de datos o aplicación**: el objetivo de un sistema de diálogo hablado es establecer un diálogo con el usuario con el fin de llevar a cabo una tarea. Para ello, el sistema debe comunicarse con una fuente de información externa. En unos casos, la salida del extractor de significado determina que se realice una consulta o actualización sobre un repositorio de información, que en la mayoría de los casos será una base de datos. En otras ocasiones será una aplicación la que intervendrá en el proceso.

- **Gestor de diálogo**: controla la interacción o diálogo entre el sistema y el usuario, y coordina al resto de componentes del sistema. El gestor de diálogo decide el siguiente paso en la secuencia de la conversación, empleando para ello la información proporcionada por los otros módulos. En la sección 2.2.2 se describe en detalle la gestión del diálogo.

- **Generador de respuestas**: este módulo es el encargado de construir el mensaje que después de pasar por el conversor texto-habla será enviado al usuario. Para ello utiliza una representación semántica de la salida y las instrucciones oportunas del gestor de diálogo. En base a ello debe decidir qué información hay que enviar al usuario, cómo debe ser estructurada y cuál debe ser la forma del mensaje.

  Generalmente se suelen emplear plantillas predefinidas, en las cuales hay una serie de huecos donde se inserta la información. Las ventajas de los sistemas basados en plantillas son que permiten la edición directa de los mensajes y que la implementación resulta más sencilla. En este tipo de sistemas existe la posibilidad de usar mensajes hablados pregrabados, que permiten obtener una mayor calidad. El inconveniente de estos sistemas es que son muy específicos, y por tanto, dependientes de la aplicación.

  Un enfoque más general consiste en usar técnicas de procesamiento de lenguaje natural para generar las frases, a partir de algún tipo de la representación semántica proporcionada por el gestor de diálogo. Aunque el resultado

final es más flexible, se incrementa notablemente la complejidad, puesto que es necesario organizar la información a transmitir, seleccionar las palabras para expresar los conceptos, generar pronombres y otro tipo de expresiones de referencia, emplear reglas para la realización sintáctica y morfológica, generar la puntuación, etcétera.

- **Conversor texto-habla**: es el encargado de producir la salida en forma de sonido al usuario a partir de la cadena proporcionada por el generador de respuestas.

  Aunque existe la posibilidad de emplear muestras de voz pregrabadas, éstas únicamente sirven para mensajes fijos e invariables. Para poder emitir cualquier tipo de mensaje es necesario el uso de síntesis del habla.

  En la síntesis del habla se distinguen dos etapas: la conversión texto a fonema y la conversión fonema a habla. En la primera, se realiza un análisis y procesado del texto y en la segunda se genera la señal sonora propiamente.

  En general, además de la cadena de texto, se suele indicar información adicional (prosódica y de otros tipos) para mejorar la calidad de la señal sonora resultante. De esta manera se consigue un resultado más natural y agradable para el usuario.

  El análisis detallado del problema de la conversión texto-habla excede los límites de este trabajo. El lector interesado podrá encontrar una excelente revisión sobre conversión texto-habla en el libro de Taylor [129] y sobre el tratamiento de los aspectos prosódicos en el trabajo de Escudero [40].

A continuación vamos a describir con más detalle los módulos de reconocimiento del habla y de gestión del diálogo, que son los más relevantes para el contenido del presente trabajo.

2.2.1. Reconocimiento del habla

El reconocimiento del habla es la conversión de la entrada hablada del usuario en una secuencia de palabras. El funcionamiento básico es similar a cualquier sistema de reconocimiento de patrones: se emplean una serie de modelos que son entrenados previamente y que posteriormente se utilizan para reconocer el habla. Es preciso distinguir dos tipos de modelos:

- Los **modelos acústicos**, que modelan la representación sonora de las palabras.
- Los **modelos de lenguaje**, que modelan cómo se combinan los anteriores para formar frases.

De forma general, los modelos acústicos se pueden considerar independientes de la tarea, mientras que los modelos de lenguaje tienen una fuerte dependencia de la tarea.

En los siguientes apartados vamos a describir con mayor detalle el proceso de reconocimiento del habla. En primer lugar, vamos a presentar la formulación matemática del problema. A continuación describiremos la fase inicial de extracción de características, el modelado acústico y el modelado de lenguaje. Se presentan también los algoritmos de búsqueda de hipótesis que permiten reducir la complejidad computacional del proceso y de esta manera obtener resultados en un tiempo razonable. Posteriormente, se detallan las medidas de rendimiento empleadas para evaluar los reconocedores del habla. Por último, se describen las técnicas de adaptación del modelo de lenguaje que permiten adaptar el modelo inicial a la tarea concreta de reconocimiento.

2.2.1.1. Formulación matemática

Dada una observación acústica concreta $\mathbf{O} = O_1 O_2 ... O_n$, el objetivo del reconocimiento del habla es encontrar la secuencia de palabras $\mathbf{W}^* = w_1 w_2 ... w_m$ para la que la probabilidad *a posteriori* $P(\mathbf{W}|\mathbf{O})$ alcanza un máximo. Para ello, se emplea la regla de Bayes [76]:

$$\mathbf{W}^* = \arg \max_W P(\mathbf{W}|\mathbf{O}) = \arg \max_W \frac{P(\mathbf{W})P(\mathbf{O}|\mathbf{W})}{P(\mathbf{O})} \tag{2.1}$$

Dado que la observación **O** es fija, la fórmula es equivalente a:

$$\mathbf{W}^* = \arg \max_W P(\mathbf{W})P(\mathbf{O}|\mathbf{W}) \tag{2.2}$$

La probabilidad $P(\mathbf{W})$ es la probabilidad de la secuencia de palabras **W**, y se calcula mediante el modelo de lenguaje. La probabilidad $P(\mathbf{O}|\mathbf{W})$ es la probabilidad de que la secuencia de palabras **W** produzca una observación **O** y se calcula mediante los modelos acústicos.

2.2.1.2. Extracción de características

El primer paso para el reconocimiento del habla es la adquisición de la señal hablada. A continuación se lleva a cabo el procesado acústico, en el que se realiza la extracción de características. El resultado es una secuencia de vectores que representan las características acústicas de la señal hablada (la observación **O** de la fórmula 2.2).

Para tareas de reconocimiento del habla se obtienen mejores resultados empleando las características en el dominio de la frecuencia, en lugar de las del dominio del tiempo. Una de las representaciones más usadas es la que emplea coeficientes cepstrales en las frecuencias de mel (*mel frequency cepstral coefficients, MFCC*). Además de usar estos coeficientes, se suelen incorporar también la primera y segunda derivadas, puesto que los cambios temporales en el espectro juegan un papel importante. También se suele incluir el valor de la energía.

2.2.1.3. Modelado acústico

El modelado acústico se utiliza para determinar el valor de la probabilidad $P(\mathbf{O}|\mathbf{W})$ de la fórmula 2.2. Aunque es posible emplear modelos acústicos a nivel de palabra, para tareas de gran vocabulario se utilizan modelos a nivel subléxico, debido a que proporcionan una mayor flexibilidad. En este caso, la unidad de modelado empleada suele ser el fonema, de modo que cada palabra se modela mediante una combinación de modelos de fonema.

Hay dos ventajas fundamentales en el uso de modelos acústicos a nivel fonema. En primer lugar, el número de fonemas es reducido, lo que facilita el proceso de entrenamiento de los modelos. En segundo lugar, los modelos de fonema son independientes del vocabulario, lo que permite emplear los mismos modelos para diferentes tareas de reconocimiento.

Lo más habitual es que el modelado acústico se lleve a cabo empleando modelos ocultos de Markov (*hidden Markov models, HMM*) [106]. De esta manera, se entrena un HMM para cada uno de los fonemas del lenguaje. Para entrenar los modelos acústicos se suele emplear el algoritmo de Baum-Welch, también conocido como algoritmo *forward-backward*. La ventaja de este algoritmo es que no precisa una segmentación de las muestras de voz de entrenamiento, únicamente es necesario conocer la transcripción.

Asimismo, es necesario disponer de la pronunciación de cada una de las palabras del vocabulario. En base a ella, se concatenan los modelos fonéticos necesarios para representar adecuadamente cada una de las palabras a reconocer. En el caso del castellano se puede realizar la transcripción ortográfico-fonética de forma automática, empleando un sistema basado en reglas.

Para obtener un mayor rendimiento del sistema, se suelen emplear modelos dependientes del contexto en lugar de simplemente fonemas. Al tener en cuenta el contexto del fonema es posible obtener unos modelos más precisos. Lo más usado son los trifonemas, que tienen en cuenta el fonema anterior y el posterior. El inconveniente de estos modelos es que precisan de gran cantidad de datos de entrenamiento, ya que es necesario entrenar un gran número de parámetros. Se pueden utilizar diversas técnicas para reducir el número de parámetros a entrenar, como por ejemplo técnicas de clustering y árboles de decisión.

2.2.1.4. Modelado de lenguaje

El modelado de lenguaje consiste en expresar las restricciones en la manera en la que se combinan las palabras para formar frases. La idea es capturar el subconjunto del lenguaje que va a ser capaz de procesar el sistema. Se pueden distinguir dos tipos de modelos: los basados en especificaciones formales y los modelos de lenguaje probabilísticos. En el primer caso, lo más habitual es emplear una gramática para indicar todas las posibles secuencias de palabras de un lenguaje. En el segundo caso, se asigna una probabilidad a las posibles secuencias de palabras.

Para tareas simples de reconocimiento de habla, es suficiente con emplear una

gramática, puesto que es posible predecir anticipadamente todas las frases de entrada. Sin embargo, para tareas de gran vocabulario, la construcción de gramáticas se complica enormemente. En este caso, se emplean modelos estocásticos que se construyen a partir de grandes corpus de texto.

Los modelos de lenguaje probabilísticos pretenden determinar el valor de la probabilidad $P(\mathbf{W})$ para una determinada secuencia de palabras $\mathbf{W} = w_1, w_2, ..., w_m$. Esta probabilidad se calcula como el producto de las probabilidades de cada palabra, asumiendo que la ocurrencia de cada palabra está determinada por las palabras precedentes. En concreto, la ocurrencia la palabra w_i está determinada por las $i-1$ palabras precedentes. Es decir, $P(\mathbf{W})$ se puede descomponer como:

$$\begin{aligned} P(\mathbf{W}) &= P(w_1, w_2, ..., w_m) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_m|w_1, w_2, ..., w_{m-1}) \\ &= \prod_{i=1}^{m} P(w_i|w_1, w_2, ..., w_{i-1}) \end{aligned} \tag{2.3}$$

donde $P(w_i|w_1, w_2, ..., w_{i-1})$ es la probabilidad de que w_i venga a continuación, sabiendo que la secuencia de palabras $w_1, w_2, ..., w_{i-1}$ ha aparecido previamente. La secuencia de palabras $w_1, w_2, ..., w_{i-1}$ se suele llamar historia.

En realidad, dado que la mayoría de historias $w_1, w_2, ..., w_{i-1}$ son únicas u ocurren pocas veces, las probabilidades $P(w_i|w_1, w_2, ..., w_{i-1})$ son imposibles de estimar, incluso para valores moderados de i. Una solución práctica consiste en limitar la historia a un número determinado de palabras $w_{i-n+1}, w_{i-n+2}, ..., w_{i-1}$. Esto es lo que se conoce como modelos de n-grama, en los que únicamente se tiene en cuenta las $n-1$ palabras anteriores. El caso de $n = 3$ se denomina trigrama, y las palabras dependen únicamente de las dos palabras previas. Los trigramas son muy utilizados en reconocimiento de habla debido a sus buenos resultados. Además, se pueden estimar relativamente bien a partir de un corpus de texto. En el caso de modelos de lenguaje trigrama la fórmula 2.3 se puede escribir:

$$P(\mathbf{W}) = \prod_{i=1}^{m} P(w_i|w_{i-2}, w_{i-1}) \tag{2.4}$$

Para estimar $P(w_i|w_{i-2}, w_{i-1})$ se cuenta el número de veces que la secuencia w_{i-2}, w_{i-1}, w_i ocurre en el corpus, $C(w_{i-2}, w_{i-1}, w_i)$, y se normaliza por el número de veces que la secuencia w_{i-2}, w_{i-1} aparece en dicho corpus, $C(w_{i-2}, w_{i-1})$:

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \tag{2.5}$$

Para la estimación de modelos de lenguaje n-grama se emplean grandes corpus de texto. Sin embargo, por muy grande que sea el corpus, siempre habrá problemas

de escasez de datos, en el sentido de que alguna secuencia de palabras puede no ser correctamente observada. El caso más grave se produce con secuencias **W** que no aparecen en el corpus de entrenamiento, ya que en ese caso $P(\mathbf{W}) = 0$, y por tanto, nunca podrán aparecer como resultado del proceso de reconocimiento del habla. Por ello, es importante asignar un valor no cero a las secuencias no vistas. Esto se consigue aplicando alguna de las múltiples técnicas de suavizado para proporcionar robustez a las estimaciones del modelo en presencia de secuencias no vistas [26].

2.2.1.5. Algoritmos de búsqueda de hipótesis

En reconocimiento de habla, el proceso para encontrar la mejor secuencia de palabras **W** dada una señal hablada de entrada **O** se denomina decodificación. Este proceso de decodificación se plantea como un problema de búsqueda, en el que hay que encontrar la secuencia de palabras cuyos correspondientes modelos acústicos y de lenguaje se ajusten mejor a la señal hablada de entrada, según se describe en la fórmula 2.2.

Una forma obvia de realizar la búsqueda es mediante fuerza bruta, analizando todas las posibles secuencias de palabras. El inconveniente es la enorme complejidad computacional que supondría este enfoque. Para conseguir una solución en un tiempo razonable hay que utilizar otras estrategias de búsqueda. Lo más habitual es emplear algoritmos de búsqueda en grafos.

El algoritmo básico que se emplea en reconocimiento del habla es el algoritmo de Viterbi, aunque se vuelve impracticable cuando el número de palabras posibles crece. En ese caso, se suele recurrir al algoritmo de Viterbi con búsqueda en haz (*beam search*). También es habitual emplear un algoritmo A* para obtener la lista de los n-mejores resultados, la red de palabras y los valores de confianza.

Para estos algoritmos, la probabilidad $P(\mathbf{W}|\mathbf{O})$ se utiliza como función de coste F. Dado que la búsqueda se realiza para encontrar el mínimo coste, se emplea el inverso de dicha probabilidad y se emplean logaritmos para simplificar los cálculos:

$$F(\mathbf{W}|\mathbf{O}) = \log\left[\frac{1}{P(\mathbf{W})P(\mathbf{O}|\mathbf{W})}\right] = -\log\left[P(\mathbf{W})P(\mathbf{O}|\mathbf{W})\right] \tag{2.6}$$

2.2.1.6. Medidas de rendimiento

Una vez obtenida una hipótesis para la secuencia de palabras **W**, un problema importante es determinar la calidad de dicha elección. En esencia, se trata de diseñar un mecanismo de evaluación que permita valorar comparativamente cuál de dos hipótesis se ajusta mejor a la realidad, cuando se dispone de la misma como referencia en el proceso de entrenamiento de los modelos.

La tasa de error de palabra (*word error rate, WER*) es la medida más empleada para medir el rendimiento de los sistemas de reconocimiento del habla [73]. En primer lugar se debe alinear la sentencia reconocida con la frase de referencia,

empleando alineamiento aproximado de cadenas (usando un algoritmo de programación dinámica). Para evaluar la distancia entre la frase de referencia y la sentencia reconocida a lo largo del camino de alineamiento, se consideran tres tipos de errores de palabra:

- Sustitución (S): la palabra correcta ha sido sustituida por una palabra incorrecta.
- Borrado (B): la palabra correcta ha sido omitida.
- Inserción (I): una palabra extra ha sido añadida en la sentencia reconocida.

En el siguiente ejemplo se muestran los tres tipos posibles de errores:

```
Referencia: dime la COMUNIDAD autónoma *** de MAYOR extensión
Reconocida: dime la COLINA    autónoma CON de ***** extensión
                    S                  I      B
```

Finalmente, se cuentan el número de sustituciones, borrados e inserciones y se calcula la tasa de error de palabra empleando la fórmula:

$$WER = \frac{S + B + I}{N} \tag{2.7}$$

donde S es el número de sustituciones, B es el número de borrados, I es el número de inserciones y N es el número de palabras en la frase de referencia.

También es interesante calcular la tasa de palabras fuera del vocabulario (*out of vocabulary word rate, OOV word rate*). Esta tasa indica si el vocabulario del reconocedor se adecúa a la tarea concreta de reconocimiento. Para calcular la tasa de palabras fuera del vocabulario se emplean una serie de frases de entrada, y se calcula el número de palabras de estas frases que no están en el vocabulario, dividido por el número total de palabras en las frases.

Por último, para medir la velocidad de los sistemas de reconocimiento del habla, se suele utilizar el factor de tiempo real (*real time factor, RTF*):

$$RTF = \frac{P}{E} \tag{2.8}$$

donde P es el tiempo empleado para procesar una entrada de duración E.

2.2.1.7. Adaptación del modelo de lenguaje

De lo expuesto hasta ahora, queda claro que un modelo de lenguaje adecuado a la tarea es una componente clave del éxito del reconocedor de habla de un sistema. La obtención del modelo de lenguaje pasa por disponer de un corpus textual extenso y adecuado a la tarea de reconocimiento que se pretenda resolver. Esto no es, en general, posible por razones de coste de elaboración y procesamiento. Una

solución muy empleada es, entonces, recurrir a técnicas de adaptación de modelo de lenguaje que permitan disponer de modelos adecuados partiendo de corpus generales de gran extensión y de corpus dependientes de tarea más reducidos.

El procedimiento a seguir obedece habitualmente al siguiente esquema: primero, se entrena un modelo de lenguaje inicial genérico usando un corpus genérico y suficientemente variado y extenso; en un segundo paso, se refina el modelo de lenguaje adaptándolo a un corpus específico de tarea pero de talla más reducida.

Entre los múltiples métodos de adaptación existentes, el basado en interpolación lineal [12] es, por su sencillez y robustez, uno de los más empleados. Este método propone entrenar un modelo de lenguaje general (usando el corpus general) y un modelo de lenguaje específico (empleando el corpus de adaptación), y a continuación combinarlos. Como resultado, la probabilidad de la palabra w_i dada la historia h_i se obtiene a partir de la estimación del modelo general $P_G(w_i|h_i)$ y del modelo específico $P_E(w_i|h_i)$:

$$P(w_i|h_i) = (1-\lambda)P_G(w_i|h_i) + \lambda P_E(w_i|h_i) \tag{2.9}$$

donde $h_i = w_{i-n+1}, w_{i-n+2}, ..., w_{i-1}$ es la historia de palabra y λ es el coeficiente de interpolación, que es un valor entre 0 y 1.

El coeficiente de interpolación se puede estimar empleando un pequeño corpus y el algoritmo expectación-maximización.

2.2.2. Gestión del diálogo

El módulo de gestión del diálogo es el encargado de controlar la interacción entre el usuario y el sistema. El objetivo es ayudar a los usuarios a completar las tareas de la manera más eficiente posible. El gestor de diálogo obtiene la información necesaria del usuario, mantiene la historia del diálogo, se comunica con la aplicación externa y decide la respuesta a enviar al usuario.

Podemos distinguir tres alternativas a la hora de plantear el diálogo, en función de quién lleva la iniciativa:

- Control del diálogo **dirigido por el sistema**: el sistema realiza preguntas al usuario para obtener la información necesaria para completar la tarea. El orden en el que el sistema recopila los datos necesarios del usuario ha sido preestablecido por el diseñador del sistema.

- Control del diálogo **dirigido por el usuario**: el usuario es el que define el curso del diálogo.

- **Iniciativa mixta**: el control del diálogo es compartido entre el usuario y el sistema. El usuario dispone de cierta iniciativa para cambiar el flujo del diálogo.

En definitiva, el gestor de diálogo controla el flujo de la interacción y su tarea principal es decidir qué hacer a continuación, a partir de la información proporcionada por el extractor de significado, la información de la aplicación y de la historia del diálogo. Para realizar el control del diálogo existen tres estrategias básicas [94]:

- Sistemas **basados en estados finitos**: la estructura del diálogo se expresa mediante un diagrama de estados en el cual los nodos representan las preguntas del sistema y las transiciones representan todos los posibles caminos alternativos que puede seguir el diálogo.

  Estos sistemas, en general, restringen lo que puede decir el usuario en cada estado, especificando un vocabulario y una gramática para cada uno de ellos. Esto permite obtener tasas más altas de reconocimiento y comprensión.

  Esta estrategia es sencilla y apropiada para tareas bien definidas y estructuradas, en las que está clara la información a intercambiar. Sin embargo los diálogos resultantes son poco flexibles, ya que la forma de interacción está predefinida y no se puede alterar. Esto hace que el sistema resulte poco natural para el usuario, aunque se suelen obtener altas tasas de éxito de la tarea.

- Sistemas **basados en marcos**: se basan en el relleno de formularios, donde es preciso recopilar una serie de elementos de información. El flujo del diálogo no está predeterminado y depende de la información proporcionada por el usuario. La ventaja es que la información puede ser recogida en cualquier orden, dando lugar a diálogos flexibles. Un ejemplo de este tipo de sistemas es ARISE, que proporciona información sobre horarios de tren [81].

- Sistemas **basados en agentes**: estos sistemas modelan el diálogo como una comunicación entre agentes, y emplean técnicas de inteligencia artificial para gestionar la interacción. Se modelan explícitamente el conocimiento del dominio y cuáles son las creencias, los deseos y las intenciones de cada agente. El sistema *the Circuit-Fix-It Shop* para ayudar a arreglar circuitos electrónicos [124] es un ejemplo de este tipo de sistemas.

  Este enfoque incorpora un modelo de comportamiento más rico y racional, dando lugar a diálogos más naturales. Sin embargo, este tipo de sistemas resultan difíciles de construir en la práctica. Además, requieren un esfuerzo significativo de expertos para construir los modelos necesarios.

Es también tarea del gestor de diálogo guardar un registro con la información que se ha intercambiado con el usuario. Es lo que se denomina historia de la interacción. Esta información de contexto es necesaria tanto para poder llevar a cabo la tarea, como para el funcionamiento de ciertos módulos del sistema (por ejemplo para la resolución de referencias). En sistemas avanzados se incluye información adicional que permite adaptar la forma de interacción en función de cómo se va desarrollando el diálogo con el usuario.

Por último, es preciso disponer de estrategias de confirmación adecuadas, dado que pueden producirse errores en el reconocimiento del habla y en la extracción de significado. Existen dos posibilidades: confirmación explícita o implícita. La confirmación es explícita si se emplea un turno de diálogo para confirmar lo que el sistema ha entendido. Esta solución asegura la confirmación del dato pero los diálogos resultantes son más largos. La confirmación es implícita si el sistema informa al usuario de los datos comprendidos y continúa con el diálogo esperando que el usuario rectifique los datos si alguno es incorrecto. Es habitual emplear medidas de confianza del reconocedor del habla para elegir el tipo de confirmación más adecuado en cada momento: confirmación explícita para información con baja confianza y confirmación implícita para datos con alta confianza.

2.3. Desarrollo de sistemas de diálogo hablado

El desarrollo de un sistema de diálogo hablado es una tarea compleja y es recomendable abordar su construcción empleando los métodos y las técnicas de la ingeniería de software. Aunque es posible la utilización de metodologías genéricas para el desarrollo de sistemas computacionales, es recomendable la utilización de metodologías especializadas que permitan describir adecuadamente las particularidades de un sistema de diálogo hablado.

Cabe destacar en este punto las contribuciones realizadas por dos proyectos de investigación de especial relevancia: EAGLES [51] y DISC [15]. En ambos casos se realizaron recomendaciones y se propusieron mejores prácticas para el desarrollo y evaluación de sistemas de diálogo hablado.

El objetivo de EAGLES (*Expert Advisory Group on Language Engineering Standards*) fue recopilar y unificar la información disponible para proporcionar información actualizada a los investigadores y desarrolladores. Se realizó un estudio de los distintos sistemas existentes, de los estándares y de los recursos disponibles. Como resultado, se generaron una serie de especificaciones y recomendaciones comunes para sistemas que utilizan lenguaje hablado, tanto para el diseño y recopilación de corpus, como para el diseño y evaluación de sistemas de lenguaje hablado.

En el marco del proyecto europeo DISC se desarrolló una metodología para el desarrollo y evaluación de sistemas de diálogo hablado. Esta metodología propone una lista de propiedades para la caracterización de sistemas de diálogo hablado (*DISC grid*) y un modelo de ciclo de vida para el desarrollo y evaluación de sistemas de diálogo hablado y sus componentes. El trabajo desarrollado en el proyecto DISC se basó en el análisis detallado de diversos sistemas de diálogo hablado.

Parte del trabajo de DISC se basó en la metodología empleada para el desarrollo del sistema de diálogo hablado *Danish Dialogue System*, que se describe en [14]. La propuesta detalla de manera rigurosa los pasos a seguir para la construcción de este tipo de sistemas. Durante la fase de análisis y diseño, se debe construir el modelo de interacción. Para ello, es recomendable disponer de una serie de di-

rectrices que guíen el proceso. Un punto clave para el éxito del sistema es asegurar que su comportamiento durante la interacción será cooperativo. Esta metodología propone 13 directrices genéricas y 11 directrices específicas para obtener una interacción del sistema cooperativa.

Una contribución importante, desde el punto de vista metodológico, es la de San Segundo y colaboradores [115], que propusieron una metodología específica para el diseño de gestores de diálogo. Esta metodología fue empleada con éxito para la construcción de un sistema de diálogo hablado que proporciona información sobre viajes en tren a través del teléfono.

En todos los casos, para construir un sistema de diálogo hablado es preciso determinar la forma en la que los usuarios interactuarán con él, es decir, cómo hablarán los usuarios al sistema. Además, la manera de hablar de los usuarios estará condicionada por la funcionalidad que proporcione el sistema. Básicamente, existen tres maneras de plantear el desarrollo de un sistema de diálogo hablado:

- **Diseño por inspiración**: se analiza la funcionalidad que se desea proporcionar y se construye el sistema de diálogo hablado utilizando la intuición de los desarrolladores. Uno de los inconvenientes de este método es que puede haber situaciones del diálogo que los desarrolladores no hayan previsto.

- **Diseño por observación**: se basa en estudiar cómo los usuarios resuelven las mismas tareas en diálogos con otras personas. En este caso, lo que se hace es recopilar y analizar un corpus de diálogos humano-humano. Este tipo de diálogos ayuda a comprender cómo dialogan los humanos para la consecución de la tarea. La recopilación de este tipo de diálogos es una tarea costosa. Sin embargo, hay que ser consciente de que no se puede generalizar a partir de diálogos humano-humano a diálogos humano-computador que son más restringidos, principalmente debido a las limitaciones de la tecnología del habla actual.

- **Diseño por simulación:** se emplea la técnica de mago de Oz, que consiste en simular un sistema de diálogo hablado con el fin de estudiar la forma de interacción del usuario [44]. Esta técnica consiste en realizar una interacción hombre-máquina simulada en la que una persona (normalmente llamada mago) se hace pasar por el sistema. El usuario que interactúa con el sistema no sabe que se trata de una simulación. Es especialmente importante hacer creer a la otra persona que efectivamente se está interactuando con una máquina, de esta forma los diálogos resultantes reflejarán en lo posible el tipo de interacción que se pretende estudiar y analizar con este tipo de simulaciones. Además, el mago debe imitar de la manera más fiel posible el comportamiento esperado del sistema. Para facilitar el proceso es recomendable emplear escenarios previamente diseñados.

  Esta técnica permite también la adquisición de corpus hablados con las intervenciones de los usuarios en el diálogo con el sistema, que serán de gran utilidad en etapas posteriores para entrenar los modelos del sistema.

Por último, dada la naturaleza altamente interactiva de estos sistemas, es recomendable realizar pruebas con usuarios reales antes de construir la versión definitiva del sistema. De esta manera, el desarrollo se llevará a cabo en varias iteraciones. Una vez se dispone de la primera versión del sistema, se evalúa con usuarios reales y a continuación se revisan y corrigen aquellos aspectos necesarios. Pueden realizarse varias iteraciones a este ciclo.

2.4. Arquitecturas para sistemas de diálogo hablado

En la sección 2.2 se han descrito los componentes básicos de un sistema de diálogo hablado. La forma de interacción entre los distintos módulos es lo que determina la arquitectura del sistema. En los sistema de diálogo hablado de la bibliografía podemos encontrar diversas maneras de plantear la arquitectura.

Las soluciones arquitectónicas más sencillas son aquellas en las que el sistema está implementado como una aplicación monolítica que reside en una única máquina. Este tipo de sistemas han ido evolucionando hacia soluciones más flexibles que permiten que cada componente se ejecute como un proceso independiente y, por tanto, pueda ejecutarse en una máquina diferente. Dentro de este tipo de sistemas, podemos distinguir aquellos en los que hay un control centralizado, llevado a cabo por un módulo especializado, y aquellos en los que el control está distribuido entre todos los módulos del sistema. En este último caso, los módulos tienen una mayor autonomía y se deben habilitar los mecanismos de comunicación necesarios para permitir su coordinación.

En esta sección se describen dos ejemplos de arquitecturas empleadas para desarrollar sistemas de diálogo hablado. En primer lugar se presenta la arquitectura Galaxy, que se ha convertido en la arquitectura de referencia para la evaluación DARPA. En segundo lugar, se presenta la arquitectura OAA pensada para desarrollar sistemas multimodales basados en sistemas multiagentes.

2.4.1. Arquitectura Galaxy

La arquitectura Galaxy [120] es una arquitectura diseñada específicamente para la implementación de sistemas de diálogo. Su objetivo primordial es la integración de componentes distribuidos. La filosofía es cliente/servidor y el núcleo central del sistema es el *Hub*, que es un módulo que se encarga de comunicar a los clientes y a los servidores. Una configuración típica es la mostrada en la figura 2.2.

El sistema funciona mediante el intercambio de mensajes. Cuando llega un mensaje al Hub, éste consulta una serie de reglas que son las que le indican qué debe hacer con el mensaje. Generalmente el mensaje se distribuirá a otro componente del sistema. El formato de los mensajes es bastante sencillo, ya que está basado en marcos. Cada mensaje tiene un tipo asociado y el resto son pares atributo/valor.

Galaxy fue la arquitectura de referencia para la evaluación DARPA Communicator [135]. Además, esta arquitectura ha sido empleada por diversos grupos de

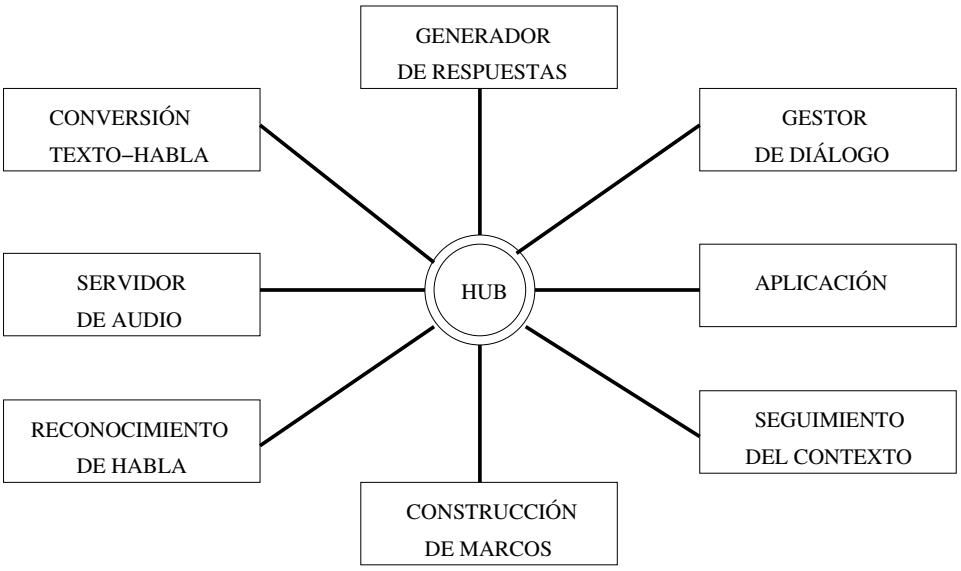


Figura 2.2: Arquitectura Galaxy.

investigación.

2.4.2. Arquitectura OAA

La arquitectura OAA (*Open Agent Architecture*) permite construir aplicaciones multiagente que presentan una interfaz de usuario multimodal [27].

El modelo de agentes propone tener una serie de componentes independientes que interactúan entre ellos mediante paso de mensajes. De esta manera, el flujo de información que se genera es totalmente descentralizado. El conocimiento que se maneja es generalmente declarativo: un agente tiene creencias acerca del mundo que le rodea, deseos e intenciones que guían su comportamiento.

En la figura 2.3 podemos ver un ejemplo de la arquitectura OAA. Existe un agente especial, *facilitator*, que se encarga de gestionar las peticiones de servicio, ya que conoce las capacidades de cada agente en el sistema. Cuando un agente se conecta al sistema, debe registrarse indicando al facilitator qué funcionalidad proporciona. Cuando un agente necesita un servicio, manda su petición al facilitator que la reenvía al agente correspondiente.

2.5. Herramientas para el desarrollo de sistemas de diálogo hablado

La construcción de sistemas de diálogo hablado es una tarea costosa y que requiere gran cantidad de recursos. En los últimos años han ido apareciendo diversos

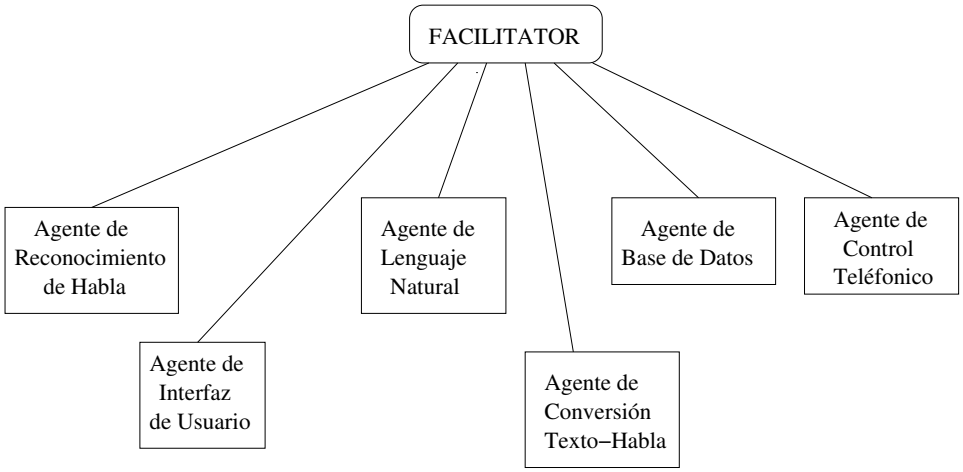


Figura 2.3: Arquitectura OAA.

tipos de herramientas encaminadas a facilitar el desarrollo de este tipo de sistemas. Existen gran cantidad de herramientas y en este apartado citaremos únicamente algunas de las más representativas.

Una de las primeras herramientas en aparecer fue el CSLU toolkit [127]. Este toolkit proporciona un entorno potente y flexible para la creación de sistemas interactivos que usan las tecnologías del habla. Su objetivo es permitir la experimentación con las tecnologías del habla y los sistemas de diálogo hablado. Ha sido muy empleado debido a que sus herramientas de autor gráficas son fáciles de usar y aislan de los detalles de implementación de la tecnología del habla subyacente.

SPEECHBUILDER [53] se creó con el objetivo de permitir el prototipado rápido y simplificar el desarrollo de sistemas de diálogo hablado. Se trata de un entorno de desarrollo de sistemas de diálogo hablado con iniciativa mixta. El usuario construye la aplicación especificando toda la información necesaria para describir la forma de interacción y el sistema se encarga de generar todos los recursos necesarios para construir el sistema de diálogo.

El sistema GEMINI [66] facilita el desarrollo de sistemas de diálogo interactivos multilingües y multimodales para acceder a información almacenada en bases de datos. Dispone de varios asistentes que ayudan al usuario a generar las aplicaciones de manera semiautomática. El sistema se basa en la estructura de la base de datos para implementar el flujo de diálogo que especifica el desarrollador. La especificación final del sistema de diálogo se hace empleando los estándares XHTML y VoiceXML.

El framework Olympus [16] fue desarrollado con el objetivo de simplificar la construcción de sistemas orientados a la investigación que incluyen interfaces conversacionales. Este framework está basado en la arquitectura Galaxy descrita en la sección 2.4.1.

Por último, cabe destacar la existencia de multitud de entornos para desarrollar sistemas de respuesta telefónica con VoiceXML. Por ejemplo, Tellme Studio[1], que dispone de un conjunto de herramientas de desarrollo que son accesibles a través de la web y que permiten la construcción y prueba de aplicaciones de respuesta telefónica en VoiceXML.

2.6. Dominios de aplicación

El dominio de un sistema de diálogo hablado determina los aspectos del mundo sobre los que el sistema puede dialogar. Los sistemas de diálogo hablado se han empleado en una gran variedad de dominios. Sin embargo, cada sistema se centra en una tarea concreta, lo que permite obtener sistemas muy especializados que consiguen un alto rendimiento. Entre los sistemas encontrados en la bibliografía podemos citar los siguientes dominios como los más representativos:

- Información de horarios de tren [34, 81].
- Información meteorológica [140].
- Información sobre vuelos de avión [121].
- Planificación de rutas [103, 109].
- Acceso al correo electrónico [134, 136].
- Redirección de llamadas [37, 64].
- Registro en una conferencia [107].
- Manejo de un robot [9].

2.7. El estándar VoiceXML

La construcción de sistemas de diálogo hablado es un proceso complejo. Esto ha motivado que en los últimos años se haya producido un gran esfuerzo de estandarización para simplificar la tarea de diseño y construcción de sistemas de diálogo hablado. El estándar más aceptado, en el entorno de los sistemas interactivos de respuesta telefónica es sin duda VoiceXML [93] (*Voice eXtensible Markup Language*). El objetivo de este estándar fue facilitar la creación de sistemas comerciales que pudieran ser usados en entornos de explotación reales. Para ello, el estándar permite independizar la aplicación vocal de la plataforma tecnológica empleada, lo que permite que las aplicaciones sean portables entre plataformas de diferentes fabricantes.

[1] `http://studio.tellme.com`

La aparición de la versión 1.0 de VoiceXML en el año 2000 convirtió el diseño de interfaces vocales en una tarea de producción, al eliminar la necesidad de un diseño a bajo nivel que requería un gran dominio de las tecnologías de síntesis y reconocimiento de habla. VoiceXML facilita el desarrollo de aplicaciones vocales siguiendo un modelo declarativo, al separar el modelado del diálogo, que se realiza mediante un lenguaje de marcas, de la implementación del mismo, llevada a cabo por una plataforma de interpretación externa. El modo de funcionamiento es muy similar al de la web, lo que permite emplear las técnicas de modelado y de desarrollo de aplicaciones web. Además, la existencia de entornos de desarrollo integrados facilita la producción industrial de aplicaciones vocales.

Una descripción más detallada del estándar VoiceXML puede encontrarse en el apéndice A.

2.8. Evaluación de sistemas de diálogo hablado

La evaluación de los sistemas de diálogo hablado es crucial para medir su rendimiento y la aceptación por parte de los usuarios. Además, la evaluación permite analizar la calidad de los diálogos que se establecen entre el sistema y el usuario.

En los siguientes apartados se describen en detalle las diversas iniciativas llevadas a cabo para la evaluación de sistemas de diálogo hablado. En primer lugar, se presentan las metodologías y recomendaciones encontradas en la bibliografía. A continuación nos centramos en la evaluación de usabilidad mediante pruebas con usuarios reales, que incluyen una evaluación objetiva y una evaluación subjetiva. Por último se describe la técnica de evaluación mediante usuarios simulados.

2.8.1. Metodologías y recomendaciones

Como ya se ha comentado anteriormente, en los proyectos EAGLES y DISC se realizaron una serie de recomendaciones para el desarrollo y evaluación de sistemas de diálogo hablado.

En lo que respecta a la evaluación, en el proyecto EAGLES [51] se señala la importancia del entorno de experimentación, y se distingue entre las evaluaciones llevadas a cabo en el laboratorio frente a las realizadas en entornos reales. El propósito de la evaluación es el que determina la elección del entorno concreto. Por otro lado, se distingue entre las pruebas de caja blanca y las pruebas de caja negra. En el primer caso, la evaluación mide el rendimiento de uno o varios componentes del sistema de diálogo. El objetivo es evaluar el componente en el contexto de un sistema de diálogo completo y diagnosticar la contribución de cada parte al éxito o fracaso global del sistema. En el segundo caso, la evaluación considera el rendimiento global del sistema de diálogo, sin analizar el comportamiento de los componentes internos del mismo.

La metodología de evaluación propuesta por el proyecto DISC [15] describe en detalle la terminología a emplear y las plantillas que deben rellenarse para cada

propiedad del sistema que se desea evaluar.

Otro hito importante en la evaluación de sistemas de diálogo hablado fue la aparición del marco de trabajo PARADISE (*PARAdigm for DIalogue System Evaluation*) [138]. Esta herramienta calcula una función de rendimiento que predice la satisfacción del usuario a partir del éxito de la tarea y de los costes del diálogo. La formulación tiene en cuenta la complejidad de la tarea y permite comparar agentes incluso aunque éstos realicen tareas distintas. También es posible medir el rendimiento para subdiálogos de la misma forma que para diálogos completos. El marco de evaluación PARADISE fue utilizado en el proyecto *Communicator* [137], financiado por DARPA, que llevó a cabo la evaluación de sistemas de diálogo hablado a gran escala. En el proyecto participaron nueve centros de investigación y todos utilizaron la arquitectura Galaxy (ver sección 2.4.1). La tarea fue la misma para todos: planificación de viajes, incluyendo reserva de avión, reserva de hotel y alquiler de coche. La forma de realizar la evaluación fue igual para todos, permitiendo de esta manera la comparación de los resultados de los diferentes grupos.

Por otro lado, la organización *International Telecommunication Union* (ITU), a través de su sector de normalización de las telecomunicaciones (ITU-T), realizó la recomendación P.851, que explica cómo llevar a cabo una evaluación de servicios telefónicos basados en tecnología del habla [75]. Primero se presentan los aspectos y factores que afectan a la calidad. A continuación se describe cómo diseñar el entorno de experimentación, cómo crear los escenarios y cómo seleccionar a los usuarios de prueba. Además, se propone el uso de tres tipos de cuestionarios: uno al principio de la evaluación con preguntas acerca del historial del usuario, otro para cada interacción individual y otro al final con preguntas sobre el sistema en su conjunto. Por último, se describe cómo se debe analizar e interpretar la información recolectada.

También se han propuesto otras metodologías para la evaluación de sistemas de diálogo hablado, como por ejemplo [25]. En esta propuesta se describen en detalle las métricas, los ficheros de bitácora y las herramientas a emplear. La metodología se basa en la anotación de los ficheros de bitácora del sistema. Para ello, se emplea un formato XML propio para dichos ficheros.

Por último, en la bibliografía podemos encontrar varios estudios que recopilan y analizan las diversas propuestas que se han venido utilizando para la evaluación de sistemas de diálogo hablado. En [83] se presenta una revisión de las distintas técnicas para la evaluación de sistemas de diálogo hablado. En [38] se presenta una revisión exhaustiva de las distintas iniciativas llevadas a cabo en los últimos años en evaluación de sistemas de diálogo, principalmente en el marco de grandes proyectos de investigación. También se describen las extensiones necesarias para evaluar sistemas de diálogo multimodales. En [100] se presenta una visión general de los métodos de evaluación y predicción de la calidad de los sistemas de diálogo hablado a través de teléfono.

2.8.2. Usabilidad de sistemas de diálogo hablado

La evaluación de los sistemas de diálogo hablado debe centrarse en el usuario, que es a fin de cuentas el que va a utilizar el sistema. El objetivo de la evaluación es obtener la información necesaria para poder mejorar la experiencia de los usuarios al interactuar con un sistema de diálogo hablado. Es decir, el objetivo es conseguir sistemas con una mayor usabilidad.

El estándar ISO 9241 [74] define la usabilidad como: *grado en el que un producto puede ser usado por usuarios específicos para conseguir objetivos específicos con efectividad, eficiencia y satisfacción en un contexto de uso específico*. La efectividad hace referencia a la precisión y completitud con la que los usuarios consiguen sus objetivos. La eficiencia es el coste de obtener esos objetivos. La satisfacción está relacionada con el confort y la aceptación de los usuarios.

Para evaluar la usabilidad de un sistema se realizan pruebas con usuarios reales. Como resultado de dichas pruebas, se obtienen una serie de medidas objetivas y subjetivas. Las medidas objetivas nos sirven para medir la efectividad y eficiencia del sistema, mientras que las medidas subjetivas nos sirven para medir la satisfacción de los usuarios.

En el caso de los sistemas de diálogo hablado, la evaluación objetiva está bastante bien establecida y se han propuesto numerosas medidas, como se describe por ejemplo en [51]. Entre todas las medidas propuestas, podemos citar entre las más utilizadas las siguientes: tiempo para completar la tarea, tasa de éxito de la tarea, tasa de reconocimiento de habla, tasa de comprensión de habla, porcentaje de turnos de corrección, número de peticiones de ayuda, número de interrupciones del usuario, número de turnos por tarea y tiempo por turno.

En cuanto a las medidas subjetivas, éstas no se pueden observar directamente y se obtienen preguntando a los usuarios después de que hayan usado el sistema. Suelen medirse mediante cuestionarios o entrevistas. Lo más utilizado son los cuestionarios, en los que hay una serie de afirmaciones sobre la percepción de los usuarios acerca del sistema y los usuarios tienen que indicar su grado de conformidad con cada afirmación. Para que los resultados obtenidos sean significativos y reflejen las verdaderas actitudes de los usuarios, es preciso que el cuestionario se diseñe cuidadosamente, se documente y se valide. Sin embargo, en algunos casos se han seguido enfoques poco rigurosos y se han empleado cuestionarios que no han sido correctamente desarrollados, sino que contenían lo que los que realizaron la evaluación creyeron más oportuno. Es decir, se han utilizado cuestionarios que no estaban previamente validados para garantizar su fiabilidad [82]. Desde el punto de vista de la interacción persona-ordenador podemos concluir que no se han usado las técnicas adecuadas.

2.8.3. Cuestionarios para evaluación subjetiva

El desarrollo y validación de un cuestionario para la evaluación subjetiva de sistemas de diálogo hablado es un proceso costoso. En la bibliografía podemos

encontrar dos cuestionarios específicamente desarrollados para evaluar este tipo de sistemas: SASSI y CCIR-BT.

El objetivo de SASSI (*Subjective Assessment of Speech System Interfaces*) es conseguir un cuestionario que sea fiable y válido para medir la experiencia del usuario utilizando una aplicación hablada. El proceso de desarrollo de SASSI se describe en [69]. Inicialmente se propuso un cuestionario de 50 afirmaciones, que fue utilizado para evaluar ocho sistemas distintos. Para analizar los resultados se emplearon técnicas estadísticas, lo que resultó en la eliminación de varias preguntas, pasando a tener el cuestionario 34 preguntas. A continuación, mediante análisis de componentes principales se identificaron seis factores, donde cada uno de ellos mide un aspecto de la percepción de los usuarios sobre el sistema: corrección en la respuesta del sistema, afabilidad, demanda cognitiva, molestia, habitabilidad y rapidez. Se entiende lo que mide cada factor por su nombre, con la excepción de la habitabilidad, que hace referencia a si el usuario sabe lo que debe hacer y lo que el sistema está haciendo. Puede entenderse también como la adecuación entre el modelo conceptual del usuario acerca del sistema de diálogo como agente conversacional y el propio sistema de diálogo.

Se empleó un proceso similar para desarrollar el cuestionario CCIR-BT [87]. La diferencia principal es que los usuarios no interactuaron con un sistema real, sino con una simulación de mago de Oz. El cuestionario estaba formado por 22 afirmaciones y se identificaron 5 factores mediante un análisis de componentes principales: calidad del rendimiento de la interfaz, esfuerzo cognitivo y estrés experimentado por el usuario, modelo conversacional del usuario, fluidez de la experiencia y transparencia de la interfaz.

2.8.4. Evaluación mediante usuarios simulados

La evaluación de un sistema de diálogo hablado empleando usuarios reales es un proceso costoso. Por ello, resulta interesante poder realizar la evaluación empleando usuarios simulados. Este enfoque es particularmente útil en las etapas iniciales de desarrollo, ya que es posible evaluar aspectos concretos del sistema y permite evaluar el impacto de las diversas decisiones de diseño.

En [88] se propone la construcción de un simulador que representa a un usuario interactuando con un sistema de diálogo hablado. Para simular la entrada del usuario se utiliza voz real de varias personas y, para ello, se utiliza un corpus hablado de diálogos. De este corpus se van seleccionando las frases a utilizar en cada turno del diálogo. Esto permite evaluar distintas estrategias en el sistema de diálogo de manera sencilla. Además, es posible realizar pruebas sistemáticas de toda la funcionalidad del sistema, lo que permite detectar problemas en los distintos módulos.

2.9. Resumen

En este capítulo se han presentado los sistemas de diálogo hablado, que permiten a los usuarios interactuar con una aplicación software mediante el uso de lenguaje natural hablado. El objetivo de la tesis es emplear este tipo de sistemas para acceder a contenidos web, lo que obligará a adaptarlos a las características específicas de este tipo de contenidos.

En primer lugar, se han descrito los distintos componentes que forman parte de este tipo de sistemas, a saber: reconocedor de habla, extractor de significado, base de datos o aplicación, gestor de diálogo, generador de respuestas y conversor texto-habla. Por su especial interés para el presente trabajo, nos hemos centrado en el reconocimiento del habla y en la gestión del diálogo.

A continuación se ha hecho una revisión de las diferentes propuestas para la construcción de sistemas de diálogo hablado y se han presentado las soluciones arquitectónicas más habituales y las diversas herramientas de desarrollo existentes.

Posteriormente, se ha presentado el estándar VoiceXML, cuyo objetivo es permitir el desarrollo de aplicaciones habladas de respuesta telefónica que sean portables entre plataformas de diferentes fabricantes. Este estándar ha permitido la automatización de parte de las tareas de los centros de atención telefónica a usuarios.

Por último, un tema de especial importancia y que ha cobrado interés en los últimos años es la evaluación de sistemas de diálogo hablado. En primer lugar se ha hecho una revisión de las distintas metodologías y recomendaciones para la evaluación de este tipo de sistemas. Posteriormente, se ha descrito la evaluación centrada en el usuario, que pretende evaluar la usabilidad mediante la realización de pruebas con usuarios reales.

Capítulo 3

Recuperación y extracción de información

3.1. Introducción

A medida que la cantidad y la disponibilidad de información en Internet crecen, se hace indispensable diseñar mecanismos eficaces de acceso y explotación de la misma. Desde el punto de vista del usuario, surge la necesidad de proporcionar mecanismos de búsqueda y recuperación de información que permitan la adecuada localización y explotación de los documentos relevantes relacionados con los temas de interés de la consulta.

La recuperación de información (*information retrieval, IR*) reúne el conjunto de modelos, técnicas y herramientas útiles para seleccionar entre un conjunto de documentos aquéllos que se ajustan a los contenidos o características especificadas por el usuario a través de un documento de entrada denominado genéricamente consulta. La recuperación de información se está convirtiendo en infraestructura básica de las vías de acceso dominante a la información en Internet. En el marco de esta tesis, las técnicas de IR serán necesarias para abordar el desarrollo de los sistemas de acceso hablado a la web que siguen un paradigma de pregunta-respuesta y que serán presentados y discutidos en el capítulo 7.

Por otro lado, la mayoría de la información de la web está diseñada para ser accedida mediante un navegador de un ordenador de sobremesa. Esto limita enormemente el uso de este tipo de contenidos. El problema reside en que los contenidos han sido diseñados para ser leídos por personas y no están preparados para la manipulación automática, debido a que no incluyen información extra que los describa. Esto ha motivado la aparición de diversas técnicas para el procesamiento automático de contenidos web.

En primer lugar, las técnicas de conversión de los contenidos web permiten adaptar los contenidos a las características de los distintos dispositivos de acceso. El objetivo es proporcionar acceso a la web empleando dispositivos tales como agendas electrónicas o teléfonos móviles. Estos dispositivos tienen unas caracterís-

ticas que limitan el acceso a determinadas páginas, debido a sus pequeñas pantallas, su escasa capacidad de cómputo y sus dispositivos de entrada de datos reducidos. Dado que los diseñadores de páginas web no suelen prestar atención a estos dispositivos móviles, es preciso realizar una conversión de contenidos para permitir un acceso amigable a la web desde este tipo de dispositivos.

En segundo lugar, resulta interesante poder procesar la información de las páginas web desde una aplicación. De esta manera, sería posible utilizar los contenidos web para otros usos que no haya previsto su creador. Para ello, es necesario extraer la información de las páginas web y transformarla a un formato adecuado que permita el procesamiento automático de la misma.

En esta tesis se propone la utilización de técnicas de procesamiento automático de contenidos web para adaptar los contenidos al canal de comunicación hablado. Además, una de las estrategias que se propone en el presente trabajo es el acceso a la información empleando búsquedas, lo que requiere la utilización de técnicas de recuperación de información.

En este capítulo se describen técnicas de recuperación de información y de procesamiento automático de contenidos web. En el primer caso, nos centraremos en el modelo vectorial, que es uno de los más utilizados. Se presenta la formulación básica del modelo, así como los distintos esquemas de pesado que pueden utilizarse. También se describe la realimentación por pseudo-relevancia y los métodos de evaluación de sistemas de recuperación de información. En cuanto a las técnicas de procesamiento automático de contenidos web, se analiza la conversión de contenidos web y la extracción de información de páginas web. Se presenta una taxonomía de los sistemas de extracción de información existentes, así como una descripción de los basados en la estructura HTML de las páginas web.

3.2. Recuperación de información

Los sistemas de recuperación de información se emplean para seleccionar los ejemplares de un universo de documentos que contienen los elementos de información elegidos por el usuario. El proceso de recuperación de información comienza cuando un usuario introduce una consulta en el sistema. Las consultas especifican la información que el usuario quiere recuperar y se suelen expresar mediante sentencias en lenguaje natural. La búsqueda se realiza sobre una colección de documentos. En general, las consultas no identifican un único documento de la colección, sino que varios documentos se ajustan a la consulta. Estos documentos que el sistema estima que son relevantes para los requisitos del usuario se le presentan a éste a través de una interfaz. El número de documentos recuperados de la colección puede ser muy grande y, por tanto, es necesario disponer de algún mecanismo para establecer el orden de presentación de los documentos que refleje de manera conveniente y natural la adecuación de cada documento a los términos buscados. La mayoría de sistemas calculan un valor numérico que indica el nivel de correspondencia entre la consulta y cada documento de la colección.

Existen varios modelos para realizar la recuperación de información. En el presente trabajo se ha empleado el modelo vectorial. Este modelo es uno de los más utilizados en la práctica, principalmente debido a su sencillez conceptual y a que proporciona buenos resultados.

En los siguientes apartados se describe en detalle el modelo vectorial de recuperación de información. En primer lugar se presenta la formulación básica del modelo y los distintos esquemas de pesado que pueden utilizarse. A continuación se describe la realimentación por pseudo-relevancia que permite mejorar los resultados obtenidos inicialmente. Por último, se describe la manera de evaluar los sistemas de recuperación de información.

3.2.1. Modelo vectorial

El modelo vectorial representa a los documentos de la colección de documentos mediante un vector [114]. Cada dimensión del espacio se corresponde con un término de la colección de documentos. Si k es el número de términos en la colección y el conjunto de términos es $\{t_1, t_2, \ldots, t_k\}$, entonces, el documento **D** se representa mediante:

$$\mathbf{D} = (w_{1,d}, w_{2,d}, \ldots, w_{k,d}) \tag{3.1}$$

donde cada coordenada del vector o peso, $w_{r,d}$, se calcula a partir de las ocurrencias del término t_r en el documento **D**.

De igual manera, el modelo vectorial representa las preguntas mediante un vector. Por tanto, la pregunta **Q** se representa mediante:

$$\mathbf{Q} = (w_{1,q}, w_{2,q}, \ldots, w_{k,q}) \tag{3.2}$$

Para calcular la similitud entre un documento **D** y una pregunta **Q** se utiliza el coseno del ángulo que forman ambos vectores:

$$sim(\mathbf{D}, \mathbf{Q}) = cos(\mathbf{D}, \mathbf{Q}) = \frac{\mathbf{D} \cdot \mathbf{Q}}{|\mathbf{D}||\mathbf{Q}|} \tag{3.3}$$

En el caso de vectores normalizados, la fórmula básica utilizada es:

$$sim(\mathbf{D}, \mathbf{Q}) = \sum_{r=1}^{k} w_{r,d} \times w_{r,q} \tag{3.4}$$

Una vez disponemos de una medida de similitud, es posible realizar búsquedas en la colección de documentos. Para ello, se calcula la similitud entre la consulta y cada documento de la colección de documentos, se ordena el resultado y muestra al usuario.

3.2.2. Esquemas de pesado

Existen diversas alternativas para calcular los pesos tanto de los documentos como de las preguntas en el modelo vectorial [112]. La manera más inmediata es contar el número de veces que aparece el término en el documento[1]. Sin embargo, hay métodos más eficientes.

En general, para calcular los pesos se suelen utilizar tres componentes:

1. **Frecuencia de término**: refleja la importancia del término dentro del documento. La forma más sencilla de calcularlo es contando el número de veces que aparece el término en el documento. Es lo que se denomina *TF, term frequency*. Para una consulta concreta, cuantas más veces aparezcan los términos buscados en un documento, más relevante es el documento.

2. **Frecuencia de colección**: indica la importancia de la palabra en la colección de documentos. El esquema que más se utiliza se llama *IDF* (*inverse document frequency*), y se calcula como $log\left(\frac{N}{n}\right)$, siendo n el número de documentos en los que aparece el término y N el número total de documentos de la colección. Las palabras que aparecen en muchos documentos son poco informativas, y las palabras que aparecen en pocos documentos tienen gran valor informativo.

3. **Normalización**: intenta dar una relevancia similar a los documentos grandes y a los pequeños. Si no se emplea normalización se favorecen los documentos grandes, ya que en ellos los términos aparecen un mayor número de veces.

Se suele utilizar una letra para designar la forma de calcular cada componente, según se detalla en la figura 3.1. Por tanto, el esquema de pesado se representa mediante 3 letras. En función de cómo se calcule cada componente existen diversos esquemas de pesado, según se muestra en la figura 3.2. El mismo enfoque se emplea para calcular los pesos en los documentos y en las preguntas. Por tanto, el esquema completo de pesado se representa mediante 6 letras. Por ejemplo, $ltc.ltn$ representa que se ha utilizado un pesado ltc para documentos y un pesado ltn para las preguntas.

3.2.3. Realimentación por pseudo-relevancia

El proceso de búsqueda de información en una colección de documentos puede plantearse de manera iterativa, incorporando cierta realimentación por parte del usuario. La idea es analizar los documentos obtenidos en una recuperación de información inicial y seleccionar una serie de términos para incorporarlos en la pregunta empleada, con el fin de obtener mejores resultados.

[1] En este apartado usaremos documento para referirnos tanto a documentos como a preguntas.

Frecuencia de término		
	n: natural	$tf_{r,i}$
	l: logarítmica	$1 + log(tf_{r,i})$
Frecuencia de colección		
	n: ninguna	no incluirla
	t: idf	$log\left(\frac{N}{df_r}\right)$
Normalización		
	n: ninguna	no usar normalización
	c: coseno	$\frac{1}{\sqrt{\sum_{s=1}^{k} {w'}_{s,i}^2}}$

Figura 3.1: Componentes del esquema de pesado para el peso $w_{r,i}$. ($tf_{r,i}$ representa el número de veces que el término t_r aparece en el documento i; df_r es el número de documentos en la colección en los que el término t_r aparece; N es el número total de documentos de la colección; ${w'}_{s,i}$ es el peso del término t_s en el documento i antes de realizar la normalización; k es el número de términos en la colección).

Este proceso de reformulación de la pregunta se denomina realimentación por relevancia (*relevance feedback*). Para poder llevarlo a cabo es preciso que el usuario determine qué documentos son relevantes y cuáles no, de entre los documentos inicialmente recuperados. A continuación se incrementa la contribución de los términos que se encuentran en los documentos recuperados identificados como relevantes y se reduce la contribución de los términos que se encuentran en los documentos recuperados identificados como no relevantes. De esta manera, al utilizar la nueva pregunta se pretende obtener más documentos relevantes y menos documentos no relevantes.

Uno de los métodos de realimentación por relevancia más empleados es el de Rocchio [113]. Este método emplea la siguiente fórmula para mejorar la pregunta inicial:

$$\mathbf{Q}_m = \alpha\mathbf{Q}_i + \beta\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{R}_i - \gamma\frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{S}_i \tag{3.5}$$

donde $\mathbf{Q}_m$ es la pregunta mejorada; $\mathbf{Q}_i$ es la pregunta inicial; n_1 es el número de documentos relevantes; $\mathbf{R}$ es el conjunto de documentos relevantes; n_2 es el número de documentos no relevantes; $\mathbf{S}$ es el conjunto de documentos no relevantes; α, β y γ son pesos que determinan la influencia de cada componente.

En general, la pregunta mejorada $\mathbf{Q}_m$ no se suele emplear tal cual, ya que contiene demasiados términos. Lo más habitual es seleccionar los términos de $\mathbf{Q}_m$ con mayor peso y añadirlos a los de la pregunta original. En este caso, el peso

nnn	$w_{r,i} = tf_{r,i}$
lnn	$w_{r,i} = 1 + log(tf_{r,i})$
ntn	$w_{r,i} = tf_{r,i} \times log\left(\frac{N}{df_r}\right)$
ltn	$w_{r,i} = (1 + log(tf_{r,i})) \times log\left(\frac{N}{df_r}\right)$
nnc	$w_{r,i} = \frac{tf_{r,i}}{\sqrt{\sum_{s=1}^{k} tf_{s,i}^2}}$
lnc	$w_{r,i} = \frac{1+log(tf_{r,i})}{\sqrt{\sum_{s=1}^{k}(1+log(tf_{s,i}))^2}}$
ntc	$w_{r,i} = \frac{tf_{r,i} \times log\left(\frac{N}{df_r}\right)}{\sqrt{\sum_{s=1}^{k}\left(tf_{s,i} \times log\left(\frac{N}{df_s}\right)\right)^2}}$
ltc	$w_{r,i} = \frac{(1+log(tf_{r,i})) \times log\left(\frac{N}{df_r}\right)}{\sqrt{\sum_{s=1}^{k}\left((1+log(tf_{s,i})) \times log\left(\frac{N}{df_s}\right)\right)^2}}$

Figura 3.2: Distintos esquemas de pesado. ($w_{r,i}$ es el peso del término t_r en el documento i; $tf_{r,i}$ representa el número de veces que el término t_r aparece en el documento i; df_r es el número de documentos en la colección en los que el término t_r aparece; N es el número total de documentos de la colección; k es el número de términos en la colección).

asignado a cada término es el que determina la fórmula 3.5.

Existe un método que permite automatizar el proceso completamente y evitar la intervención del usuario. Es lo que se denomina realimentación por pseudo-relevancia (*pseudo-relenvance feedback*) o realimentación por relevancia ciega (*blind relevance feedback*). Este método consiste en realizar una primera recuperación para obtener una lista de documentos ordenada por relevancia. Entonces, se asume que los primeros documentos son relevantes y los últimos documentos son no relevantes. De esta manera se evita la intervención del usuario.

3.2.4. Evaluación de sistemas de recuperación de información

Para evaluar un sistema de recuperación de información es necesario utilizar tres elementos: una colección de documentos, un conjunto de preguntas y un conjunto de juicios de relevancia que indiquen qué documentos son relevantes para cada pregunta [89].

Las dos medidas más utilizadas son la precisión (*precision*) y la cobertura (*recall*). La precisión, p, indica cuántos documentos recuperados son relevantes y la cobertura, c, indica cuántos documentos relevantes se han recuperado:

$$p = \frac{|\text{documentos relevantes recuperados}|}{|\text{documentos recuperados}|} \tag{3.6}$$

$$c = \frac{|\text{documentos relevantes recuperados}|}{|\text{documentos relevantes}|} \tag{3.7}$$

Ambas medidas se calculan sobre conjuntos, pero en general, un sistema de recuperación de información asigna un orden al conjunto de documentos recuperados. Se define entonces la precisión y la cobertura para los k primeros documentos recuperados como:

$$p_k = \frac{|D_k^{rel}|}{|D_k|} \tag{3.8}$$

$$c_k = \frac{|D_k^{rel}|}{|D^{rel}|} \tag{3.9}$$

donde D_k es el conjunto de los k primeros documentos recuperados; D^{rel} es el conjunto de todos los documentos relevantes; D_k^{rel} es el conjunto de documentos relevantes que hay en los k primeros documentos recuperados.

Si determinamos el valor de la precisión y la cobertura para cada posible tamaño del conjunto de documentos recuperados, podemos dibujar una curva precisión/cobertura. Dicho gráfico tendrá forma de dientes de sierra, debido a que si el documento recuperado $k + 1$ no es relevante, entonces el valor de la cobertura es igual que para k documentos recuperados, pero el valor de la precisión decrece; en cambio, si es relevante, tanto la cobertura como la precisión se incrementan.

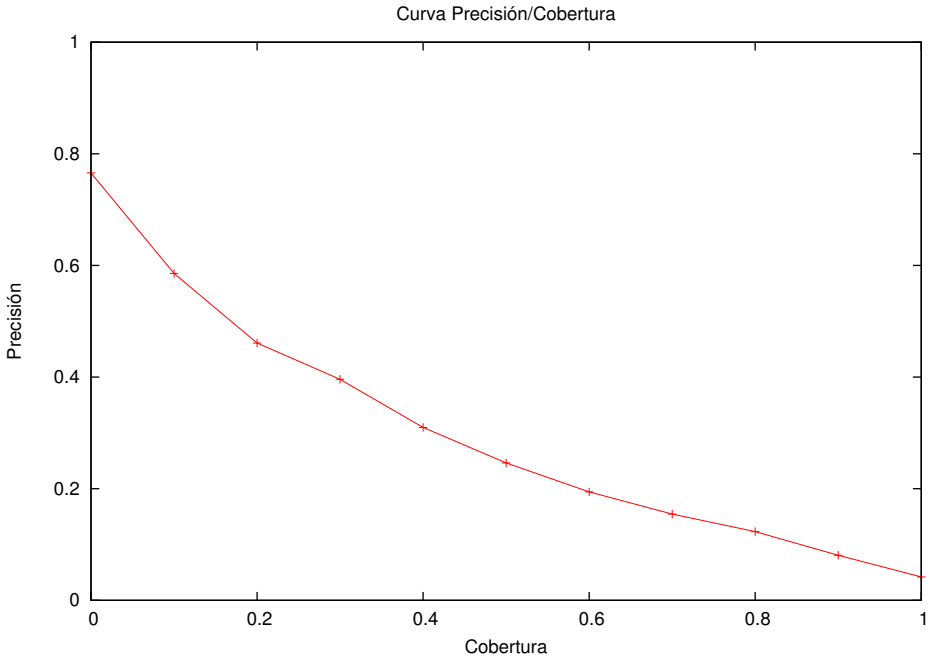


Figura 3.3: Ejemplo de curva precisión/cobertura.

Para dibujar las curvas precisión/cobertura se suele emplear la precisión interpolada para evitar los dientes de sierra. La precisión interpolada en un punto de cobertura c se define como el mayor valor de precisión encontrado para cualquier valor de cobertura $c' \geq c$:

$$p_{interp}(c) = \max_{c' \geq c} p(c') \tag{3.10}$$

Cuando se realiza la evaluación de un sistema, se suele utilizar una colección de prueba que incluye un conjunto de preguntas. En ese caso, la curva precisión/cobertura se suele generar empleando la media obtenida sobre todas las preguntas. Además, el cálculo se suele reducir a la media de la precisión interpolada en 11 valores de cobertura: 0,0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9 y 1.0. En la figura 3.3 se muestra un ejemplo de curva precisión/cobertura empleando la precisión interpolada y una colección de prueba de 49 preguntas.

La curva precisión/cobertura es bastante informativa, pero en ocasiones es deseable obtener un único valor que permita realizar comparaciones fácilmente. La medida más utilizada es la precisión media. Para calcularla, se calcula el valor de la precisión (sin interpolar) después de cada documento relevante recuperado y a continuación se hace la media de todos esos valores de precisión. La precisión media enfatiza recuperar más documentos relevantes antes.

Cuando se calcula el rendimiento del sistema empleando un conjunto de pre-

guntas se emplea la precisión media promediada (*mean average precision, MAP*), que es la media de las precisiones medias calculadas para cada una de las preguntas de la colección de prueba. Si Q es el conjunto de preguntas, si el conjunto de documentos relevantes para una pregunta $q_j \in Q$ es $\{d_1, d_2, \ldots, d_{m_j}\}$ y si R_{jl} es el conjunto de documentos recuperados hasta que obtenemos el documento d_l. Entonces la precisión media promediada se calcula mediante:

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{l=1}^{m_j} \frac{l}{|R_{jl}|} \tag{3.11}$$

3.3. Procesamiento automático de contenidos web

La información disponible en la web ha sido diseñada, en la mayoría de los casos, para ser accedida empleando un navegador web de un ordenador de sobremesa. Además, en general, las páginas web han sido construidas en base a criterios que pretenden mejorar la apariencia visual. El problema principal es que se tiende a descuidar la correcta estructuración de los contenidos. Empleando el lenguaje HTML es sencillo formatear documentos de cara a su visualización en un navegador web, pero no es posible utilizar directamente los contenidos para otros usos que no haya previsto su creador. Por ello, han aparecido una serie de técnicas que tratan de realizar un procesamiento automático de los contenidos web y que persiguen bien reconvertir los contenidos para adecuarlos a un dispositivo de acceso diferente, bien reestructurarlos para su uso por otras aplicaciones.

En el presente trabajo se utilizan este tipo de técnicas para permitir el acceso a contenidos web mediante habla. Por un lado, en el capítulo 5 se plantea la conversión de contenidos web, escritos en HTML, a diálogos hablados escritos en VoiceXML. Por otro lado, en el capítulo 6 se propone una forma de solución basada en construir un modelo de información utilizando técnicas de extracción de información de páginas web.

En los siguientes apartados se presentan las técnicas de procesamiento automático de contenidos web. En primer lugar la conversión de contenidos web, que permite acceder a la información desde diferentes dispositivos. En segundo lugar las técnicas de extracción de información, que permiten acceder a los contenidos web en un formato adecuado para el procesamiento desde una aplicación.

3.3.1. Conversión de contenidos web

La tecnología de conversión de contenidos permite traducir los contenidos web dinámicamente de un formato origen a un formato destino. El objetivo de esta conversión es optimizar la presentación según las características del dispositivo a utilizar para el acceso a la información. La situación más frecuente es la conversión de contenidos diseñados para un ordenador de sobremesa para adaptarlos a un dispositivo móvil. De esta manera se pueden superar las limitaciones en el acceso

a los contenidos que se producen debido a las características especiales de estos dispositivos: pantallas de reducidas dimensiones, poca memoria, poca potencia de procesamiento, dispositivos de entrada de datos reducidos, etc.

Los sistemas de conversión más generales tratan de hacer la conversión de las páginas sin centrarse en un sitio web o dominio concreto [118]. Estos sistemas, en primer lugar, realizan un análisis de la página y procesan tanto la estructura como el contenido. A continuación, presentan la información de manera adaptada a las características del nuevo dispositivo y proporcionan los mecanismos adecuados para la navegación. La arquitectura suele incluir un intermediario (*proxy*) que es capaz de realizar la conversión bajo demanda de manera dinámica. Uno de los puntos importantes es conseguir reducir el contenido que se muestra al usuario, y para ello, es necesario emplear estrategias de resumen automático adaptadas a contenidos web [20].

Una alternativa para mejorar el resultado final consiste en la anotación de las páginas HTML originales, de manera que el conversor de contenidos disponga de información adicional que facilite la conversión [71]. Otros sistemas permiten la personalización y la creación de diferentes vistas para diferentes dispositivos, como WebViews [45]. En ambos casos es preciso especificar la conversión necesaria para cada sitio web, lo que supone un esfuerzo adicional. Sin embargo el resultado final es mucho mejor.

3.3.2. Extracción de información de páginas web

Las técnicas de extracción de información tienen como objetivo la estructuración de la información. Partiendo de información con poca o ninguna estructura, se extraen ciertas porciones y se organizan adecuadamente para su posterior tratamiento.

Podemos distinguir tres tipos de documentos: texto libre, semiestructurado y estructurado [1]. Para el caso de texto libre, se han utilizado tradicionalmente técnicas de procesamiento de lenguaje natural. En el caso de texto semiestructurado (las páginas web se suelen definir como semiestructuradas), las técnicas clásicas de procesamiento de lenguaje natural no proporcionan buenos resultados, y se suelen emplear reglas basadas en delimitadores. Para texto estructurado el problema se reduce a un cambio de formato y por ello se emplean reglas de transformación sencillas.

Para la extracción de información de páginas web se emplean programas que transforman la información de una representación poco estructurada a otra con una estructura bien definida [79]. La idea es traducir esos documentos a un formato que se pueda manipular de una manera más sencilla desde una aplicación. Sin embargo, la mayoría de sistemas funciona solamente para un conjunto pequeño de documentos y no se adapta bien a cambios en la estructura. Como resultado, todavía se requiere la intervención de expertos para extraer información de páginas web.

A estos programas se les denomina *wrappers* y se suelen diseñar específica-

mente para una fuente de información particular. Su tarea consiste en identificar la información de interés y convertirla al formato adecuado. La extracción de la información se realiza en base a unas reglas o patrones de extracción. Hay dos enfoques principales para la construcción de wrappers. El primero está basado en construir manualmente las reglas de extracción, con la ayuda de herramientas de desarrollo específicas, que simplifican la construcción de wrappers. El segundo se basa en obtener las reglas de forma automática o semiautomática a partir de una serie de ejemplos anotados. En ambos casos se suele emplear una plantilla que describe cómo está dispuesta la información en la página, detallando las etiquetas que delimitan cada uno de los elementos de información.

Una vez realizado el proceso de extracción, el resultado es un árbol que contiene la información deseada adecuadamente estructurada. Es habitual emplear lenguajes de marcado para describir el resultado, siendo XML el más utilizado.

Como inconvenientes principales de los wrappers podemos citar dos: dificultad de construcción y coste de mantenimiento. En primer lugar, resulta inviable la obtención de información de manera masiva de la web. En segundo lugar, es necesario revisar y corregir los wrappers cada cierto tiempo, debido a la naturaleza dinámica de la web.

3.3.2.1. Taxonomía de sistemas

Existe una gran cantidad de herramientas de extracción de información de la web, puesto que el problema se puede abordar mediante diferentes técnicas, que podemos clasificar en [79]:

- Lenguajes para el desarrollo de *wrappers*: se trata de lenguajes específicamente desarrollados para permitir a los usuarios realizar la extracción de información de páginas web. Se plantean como alternativa a los lenguajes de programación de propósito general.

- Herramientas basadas en la estructura HTML: estas herramientas aprovechan la estructura inherente de las páginas. Para ello, en primer lugar, se realiza un análisis de la página para obtener el árbol que representa la jerarquía de etiquetas HTML. Posteriormente se aplican las reglas de extracción.

- Herramientas basadas en procesamiento de lenguaje natural: estas herramientas están dirigidas a extraer información que tiene una estructura muy clara y generalmente afrontan el problema directamente sobre el texto libre. Se suelen utilizar técnicas de filtrado, etiquetado gramatical y etiquetado léxico-semántico para obtener relaciones entre las frases y los elementos de las sentencias. A continuación se aplican las reglas de extracción, que se basan en restricciones sintácticas y semánticas.

- Herramientas de inducción de *wrappers*: a partir de una serie de ejemplos son capaces de generar un wrapper para extraer la información. Las reglas

de extracción suelen basarse en delimitadores e información de formato del documento.

- Herramientas basadas en modelado: partiendo de una descripción de la estructura de los objetos que se buscan, tratan de localizar en las páginas porciones de información que se adecúen a dicha estructura.

- Herramientas basadas en ontología: se basan exclusivamente en el contenido y no utilizan la información de formato o presentación de los documentos. Dado un dominio específico de información, se utiliza una ontología para localizar los objetos de interés.

3.3.2.2. Sistemas basados en la estructura HTML

Las herramientas basadas en la estructura HTML tratan de inferir la estructura inherente de las páginas web para realizar la extracción de información. Las etiquetas HTML están orientadas a la presentación de documentos, y por tanto, la estructura del documento, si existe, está definida implícitamente por estas etiquetas. En este apartado vamos a describir dos ejemplos relevantes de este tipo de sistemas: W4F y XWRAP.

W4F (*World Wide Web Wrapper Factory*) es un entorno de desarrollo que permite construir wrappers a partir de una especificación en un lenguaje declarativo [111]. El primer paso es analizar la página y construir el árbol que representa la jerarquía de etiquetas HTML. A continuación se aplican las reglas de extracción para extraer los pedazos de información de interés. La especificación de la extracción se hace mediante un lenguaje que permite hacer referencia a cualquier elemento del árbol empleando expresiones de camino. El sistema permite recuperar tanto texto como atributos de las etiquetas HTML. Asimismo, se pueden emplear expresiones regulares para transformar la información extraída de un nodo. Por último, la información extraída se almacena en un documento XML o en una estructura utilizable desde una aplicación. También se incluyen herramientas visuales para facilitar el desarrollo de wrappers.

XWRAP es un entorno de desarrollo que permite generar programas para extraer información de páginas web de manera automática [86]. Los programas estructuran la información y generan la salida en formato XML. El primer paso es obtener la página web y realizar el análisis para construir el árbol que representa la página HTML. A continuación el sistema interactúa con el usuario para encontrar los objetos de interés dentro del código HTML, identificar los tokens semánticos importantes y descubrir la estructura inherente del documento. Con esa información se generan las reglas de extracción y el código del programa wrapper.

3.4. Resumen

En este capítulo se han presentado las técnicas de recuperación de información, que permiten a los usuarios realizar búsquedas en una colección de documentos, y las técnicas de procesamiento automático de contenidos web, que permiten la conversión de contenidos web y la extracción de información de páginas web.

En el presente trabajo se utilizan estas técnicas para permitir a los usuarios el acceso a contenidos web empleando habla. En primer lugar, se plantea la conversión de contenidos web, escritos en HTML, a diálogos hablados escritos en VoiceXML, según se presenta en el capítulo 5. En segundo lugar, se propone una forma de solución basada en construir un modelo de información utilizando técnicas de extracción de información de páginas web, tal y como se describe en el capítulo 6. En tercer lugar, las técnicas de IR son necesarias para abordar el desarrollo de sistemas de acceso hablado a la web siguiendo un paradigma de pregunta-respuesta, según se describe en el capítulo 7.

Capítulo 4

Acceso a contenidos web mediante habla

4.1. Introducción

En los capítulos anteriores se han resumido los aspectos fundamentales de los sistemas de diálogo hablado y de la recuperación y extracción de información, dos pilares sobre los que se apoyarán las propuestas que presentaremos en la tercera parte de la tesis. En este capítulo se presentan diferentes alternativas para hacer accesibles los contenidos web mediante el empleo de habla.

En primer lugar, alineadas con nuestra propuesta de conversión de un portal web genérico en un portal hablado (ver capítulo 5), revisaremos algunas propuestas que tratan también de resolver el problema de forma general para cualquier tipo de página web. Dado lo ambicioso del enfoque, no resulta sencillo obtener un resultado final realmente amigable y útil para los usuarios.

Un segundo grupo de alternativas, alineadas con nuestra segunda propuesta (descrita en el capítulo 6), incluye soluciones asociadas específicamente a determinados sitios web, que pretenden mejorar los resultados obtenidos por enfoques más generales en términos de usabilidad y rendimiento en el uso de los sistemas.

Finalmente, y en consonancia con nuestra tercera propuesta (presentada en el capítulo 7), revisaremos aquellos trabajos que plantean el problema como una búsqueda de información, sustituyendo la entrada textual al sistema de recuperación de información por el resultado de la conversión a texto de una entrada hablada.

4.2. Soluciones generales

Las soluciones generales engloban las propuestas diseñadas para permitir el acceso mediante habla a cualquier tipo de página web, con independencia tanto de la estructura y naturaleza de la información contenida en la misma como del tipo de escenario de acceso y explotación de información al que se pretenda aplicar la solución.

Abordar el problema con este grado de generalidad ha motivado el desarrollo de propuestas que obedecen a alguno de los dos esquemas fundamentales siguientes: sustituir las acciones de navegación basadas en teclado y ratón por comandos hablados o transformar los contenidos de la página en una versión navegable mediante habla usando lenguajes de representación de contenidos adecuados para un determinado modelo de gestor de diálogo.

La aproximación más simple, explorada desde mediados de los noventa, consiste en extender un navegador web existente añadiendo una interfaz hablada. De esta manera, se emplean comandos vocales en lugar del ratón y el teclado, como en el sistema SAM [68] y en el sistema SLAM [72]. Sin embargo, la interacción resultante es muy pobre, porque los contenidos no se adaptan a las características de la nueva modalidad. Con el fin de tener una salida hablada, se puede usar un lector de pantalla para presentar los contenidos de la página web al usuario [132]. Algunos sistemas añaden etiquetas al código HTML para indicar a la plataforma vocal subyacente como se organiza la información [36]. Aunque los resultados de este enfoque son buenos, es necesario la modificación de los contenidos originales para añadir la información adicional a los sitios web.

El sistema PhoneBrowser desarrollado por Lucent propone realizar la interacción exclusivamente a través del canal hablado [18]. El sistema realiza un análisis de los contenidos de la página web y envía al usuario una descripción de los mismos. El usuario controla la interacción mediante comandos y realiza la navegación de los contenidos de manera similar a como lo haría con un navegador visual. Por tanto, el resultado obtenido está limitado por la forma en la que se han diseñado las páginas web originales.

El sistema WIRE [62] ha sido desarrollado para permitir el acceso a contenidos web mediante habla en el coche, empleando la metáfora de radio de coche para la interacción. El sistema analiza la estructura de la página para comprender el contenido y proporciona distintos modos de interacción, según el tipo de página y la información a la que quiera acceder el usuario.

En un segundo grupo de contribuciones, se opta por convertir los contenidos HTML a un formato intermedio de representación, apto para su navegación mediante habla empleando algún sistema de gestión de diálogo hablado, como VoiceXML. IBM dispone de un sistema para la conversión genérica de páginas HTML a VoiceXML [80]. Un enfoque similar ha sido empleado en [63]. Ambos sistemas tratan de inferir la estructura de los documentos web, y basándose en ésta, realizar el diálogo con el usuario. La solución deja de ser válida en cuanto aumenta la complejidad de las páginas. Sin embargo, dadas las características del canal hablado, puede ser una mejor solución hacer la conversión semiautomática de versiones simplificadas de las páginas HTML, empleando anotaciones manuales [45]. En [8] se emplea un enfoque más adecuado, que consiste en traducir los contenidos previamente a un formato estructurado en XML, y plantear la conversión a VoiceXML en un segundo paso.

Todas estas propuestas abordan el problema individualmente para cada página HTML y no aportan un modelo que permita diseñar soluciones globales. El rendi-

miento de estos sistemas es bajo, ya que plantean diálogos sencillos con el usuario, debido a que únicamente realizan un procesado sintáctico de la información.

4.3. Soluciones en dominios restringidos

Con el objetivo de conseguir una interacción más amigable y eficiente, hay sistemas que plantean la interacción solamente para un determinado sitio web. Al restringir el dominio de aplicación del sistema, los desarrolladores pueden diseñar diálogos específicos que están adaptados al contenido concreto de las páginas web. Además, la información puede ser almacenada en una base de datos, lo que permitiría el uso de sistemas de diálogo hablado tradicionales. El único inconveniente es que esta solución es poco flexible, y no puede usarse para acceder a sitios web para los cuales no ha sido diseñada. Ejemplos de este tipo de sistemas son [85, 105].

4.4. Recuperación de información dirigida por habla

La recuperación de información dirigida por habla (*speech driven infomation retrieval*) consiste en usar habla en lugar de texto, como entrada a un motor de recuperación de información. El usuario formula su consulta y el sistema realiza la recuperación de documentos, iniciando un diálogo con el usuario para que éste seleccione los documentos relevantes.

En la bibliografía se encuentran diversos experimentos realizados con este tipo de sistemas. En la mayoría de los casos se utiliza un reconocedor de habla continua de gran vocabulario, con vocabularios que oscilan entre las 20.000 y las 60.000 palabras. Para la recuperación de información se suelen utilizar motores de recuperación basados en el modelo vectorial o en el modelo probabilístico. Respecto a las preguntas utilizadas, encontramos experimentos usando preguntas de distinta longitud, desde preguntas cortas que incluyen pocos términos, hasta preguntas con más de 50 términos.

Para la evaluación de estos sistemas es habitual utilizar conjuntos de prueba estándar de recuperación de información. Estos conjuntos de prueba han sido diseñados para evaluar sistemas de recuperación de información empleando preguntas textuales y, por tanto, ha sido preciso extenderlos mediante la grabación de varios locutores leyendo las preguntas. La ventaja de utilizar conjuntos de prueba estándar es que ya han sido previamente validados. De esta manera, es posible evaluar el rendimiento de distintos sistemas en condiciones comparables. También es posible comparar los resultados obtenidos empleando preguntas habladas con los que se obtienen cuando se emplean preguntas textuales. Este procedimiento de evaluación ha sido utilizado en diversos experimentos llevados a cabo en inglés, chino y japonés. En el caso de inglés se ha usado la colección TREC-2 y la colección Boston Globe [10, 30, 31]; en el caso de chino se ha utilizado la colección TREC-5 y TREC-6 [24]; y en el caso de japonés las colecciones NTCIR-1, NTCIR-2, NTCIR-3 y NTCIR-4 [5, 6, 46, 47, 90, 91].

Los primeros trabajos en el tema han estado enfocados a estudiar el impacto de la tasa de error del reconocedor de habla (WER) y de la longitud de las preguntas en la precisión de la recuperación [10]. Los resultados mostraron que incrementando el WER se reduce la precisión de la recuperación y que las preguntas de gran longitud (más de 50 términos) son más robustas ante errores de reconocimiento de habla que las preguntas cortas (unos pocos términos). También se han estudiado las limitaciones que impone la tecnología de síntesis de habla, comparando la presentación de documentos en tres situaciones: en pantalla, mediante un locutor y mediante síntesis de habla [31]. La conclusión fue que la presentación de documentos al usuario empleando habla es posible y efectiva, incluso usando síntesis de habla a través del teléfono. Por otro lado, se ha comprobado que la utilización de técnicas de realimentación por relevancia en el motor de recuperación de información contribuye a mejorar los resultados del sistema [30]. Por último, se han realizado experimentos para analizar cómo influye el canal en el rendimiento del sistema [24]. Para ello, se utilizaron tres canales distintos: auriculares con micrófono, micrófono de agenda electrónica y teléfono móvil. Los resultados mostraron que la precisión de la recuperación en dispositivos móviles con micrófonos de alta calidad (como una agenda electrónica) es satisfactoria, aunque el rendimiento para teléfonos móviles es significativamente peor.

Con el fin de obtener unos mejores resultados, se han propuesto mejoras en el reconocimiento de habla, principalmente encaminadas a adaptar el reconocedor de habla a la tarea de recuperación de información. En [47] los resultados mostraron que usando la colección de documentos objetivo para construir el modelo de lenguaje del reconocedor contribuye a una mejora significativa de los resultados. Esto fue confirmado en [46], donde se comprobó además que el empleo de un vocabulario más grande también contribuye a una mejora significativa de los resultados. Los mismos resultados se han obtenido en [90], donde también obtuvieron mejores resultados empleando un LM con un mayor tamaño de vocabulario. Otra posible mejora de los modelos de lenguaje consiste en incluir en el vocabulario colocaciones, además de palabras [43].

También se han realizado diversas propuestas para conseguir una mayor integración entre el reconocimiento de habla y la recuperación de información. En [99] se presenta un método que utiliza la red de palabras completa que genera el reconocedor de habla como entrada al motor de recuperación de información, en lugar de usar únicamente la mejor hipótesis que es lo que se hace en la mayoría de casos. Además, se utilizan los valores de confianza de cada palabra de la red de palabras para pesar los términos en la recuperación de información. Otra posibilidad es emplear técnicas para combinar la salida de múltiples reconocedores de habla [90, 91]. Para ello, se deben usar diferentes motores de reconocimiento del habla con diferentes configuraciones. En los experimentos descritos la técnica de combinación que mejores resultados proporciona es SVM. Por último, en [77] se propone realizar un post-procesamiento de los resultados de reconocimiento de habla para corregir errores. La propuesta combina un modelo de canal ruidoso basado en sílaba, a nivel léxico, con un enfoque basado en conocimiento lingüístico

de alto nivel, a nivel semántico.

La recuperación de información dirigida por habla es una tarea con un vocabulario abierto, y por tanto, uno de los mayores problemas son las palabras fuera de vocabulario (*OOV words*). Para superar este problema en [48] se propone un método en dos pasos: en primer lugar se detectan las palabras OOV en la pregunta, se eliminan y se recuperan los documentos relevantes de la colección de documentos; en segundo lugar, se busca entre las palabras de los documentos recuperados aquellas que son similares fonéticamente a las palabras OOV detectadas, y de esta manera se consigue completar la transcripción de la pregunta.

Con el objetivo de superar las limitaciones de este tipo de sistemas se puede establecer un diálogo con el usuario para tratar de conseguir información adicional. En [96] se describe una estrategia de diálogo que pretende afrontar dos de las principales limitaciones: los errores del reconocedor de habla y la vaguedad de las preguntas de los usuarios. En el primer caso, se plantea un método de confirmación que utiliza dos medidas para identificar las partes que es preciso confirmar: el valor de relevancia y el valor de significatividad. Los experimentos demostraron que el método propuesto realiza la confirmación de manera más eficiente que empleando los valores de confianza del reconocedor de habla. En el segundo caso, se propone un método para generar preguntas de clarificación para los casos en los que la necesidad de información no está especificada completamente. De entre todas las preguntas generadas, se selecciona la pregunta a realizar empleando un criterio de ganancia de información, es decir, se elige aquella pregunta cuya respuesta permita una mayor reducción del número de documentos recuperados. Los experimentos demuestran que el método mejora la tasa de éxito de la recuperación.

En cuanto a la utilización de habla espontánea, los experimentos realizados muestran una degradación significativa de los resultados, lo que indica que todavía es necesario mejorar las técnicas de reconocimiento del habla para poder emplear habla espontánea en un sistema de recuperación de información dirigida por habla [5, 6].

4.5. Búsqueda de respuestas dirigida por habla

Los sistemas de búsqueda de respuestas dirigida por habla permiten obtener la respuesta concreta a una pregunta hablada realizada en lenguaje natural. La diferencia básica con los sistemas presentados en el apartado anterior, es que los sistemas de recuperación de información devuelven una lista de documentos relevantes, mientras que los sistemas de búsqueda de respuestas devuelven una respuesta concreta, formada por una o varias frases.

Los sistemas de búsqueda de respuestas dirigida por habla se basan en la integración de sistemas de reconocimiento de habla y de búsqueda de respuestas. El principal problema son los errores de reconocimiento de habla introducidos en el proceso de transcripción de la pregunta hablada. Es preciso, por tanto, encontrar los mecanismos adecuados para reducir el efecto de estos errores en el rendimiento

del sistema de respuesta a preguntas.

La utilización de un modelo de lenguaje específico para la tarea permite mejorar el resultado del sistema. El modelo de lenguaje debe reflejar la manera en la que los usuarios realizan preguntas al sistema. Una manera de conseguirlo es mediante la adaptación de un modelo de lenguaje más general [7]. El método consiste en construir un modelo de lenguaje basado en n-gramas usando la colección de documentos objetivo. A continuación, se adapta dicho modelo de lenguaje enfatizando aquellos n-gramas pertenecientes a las estructuras fijas que aparecen en las sentencias interrogativas (por ejemplo *"Cuál era el nombre...?"*). Otra solución consiste en utilizar un modelo de lenguaje para cada dominio [78]. En este trabajo se entrenan cinco modelos de lenguaje dependientes de dominio usando un corpus de noticias, y a continuación se interpolan con un modelo de preguntas.

También se puede plantear la solución en dos pasos, como en [4]. En un primer paso se transcribe la pregunta a texto, se obtienen los pasajes relacionados y se revalúa la lista de los n-mejores candidatos del reconocedor de habla empleando el valor de similitud que proporciona la recuperación de pasajes. En un segundo paso, la mejor hipótesis es usada por el módulo de búsqueda de respuestas para obtener la respuesta final.

Otra alternativa es realizar un filtrado del resultado del reconocimiento de habla antes de hacer la búsqueda de respuestas. En [70] se realiza un filtrado para eliminar errores e información redundante. En [67] se emplea un mecanismo de filtrado que elimina las palabras que no pueden ser procesadas por un sistema de respuesta a preguntas debido a inconsistencias sintácticas, semánticas o pragmáticas.

En algunos sistemas se establece un diálogo con el usuario para corregir errores de reconocimiento de habla y para obtener información adicional mediante preguntas de clarificación. Esto permite mejorar el rendimiento, ya que se refina el resultado obtenido por la búsqueda de respuestas inicial. En la bibliografía se encuentran ejemplos de sistemas de diálogo exclusivamente hablados [42, 131] y de sistemas multimodales [119].

Por último, en [116] se destaca la importancia de las entidades nombradas (*named entities, NE*) en la búsqueda de respuestas dirigida por habla: buenos resultados en el reconocimiento de NE permitiría a un sistema de búsqueda de respuestas dirigida por habla obtener un rendimiento comparable a los sistemas de respuesta a preguntas cuya entrada es texto.

4.6. Resumen

En este capítulo se han analizado las distintas estrategias que permiten el acceso a contenidos web empleando habla. De la revisión realizada se derivan líneas de trabajo que se han explorado en el resto de la tesis.

En primer lugar, se han presentado las propuestas diseñadas para permitir el acceso de forma general a cualquier tipo de página web. Se pueden distinguir dos tipos de enfoques, los completamente automáticos y los semiautomáticos. Las so-

luciones automáticas son demasiado generales y no proporcionan unos resultados satisfactorios. En cuanto a las soluciones semiautomáticas, en esta tesis se introduce el concepto de aplicación vocal, que permite a los desarrolladores diseñar cómo debe realizarse la interacción para permitir al usuario acceder a la información de un determinado sitio web.

En segundo lugar, se plantea restringir el dominio del sistema para obtener mejores resultados. En esta tesis se presenta una propuesta basada en utilizar dos modelos: un modelo de información y un modelo de interacción.

Por último, otro tipo de propuestas plantean la integración de sistemas de reconocimiento del habla con motores de recuperación de información. En esta tesis se analizan los factores que más influyen en el rendimiento de los sistemas de recuperación de información dirigida por habla. Además, se realizan varias propuestas para reducir el impacto de los errores de reconocimiento del habla en el rendimiento del sistema.

Parte III

Propuestas para el acceso a contenidos web empleando habla

Capítulo 5

Conversión de un portal web en un portal hablado

5.1. Introducción

En este capítulo presentamos nuestra propuesta para la conversión de contenidos web, escritos en HTML, a diálogos hablados descritos en VoiceXML. Se presentan dos enfoques para realizar la conversión, uno automático y otro semiautomático. Ambos permiten realizar la conversión de manera que las páginas resultantes puedan ser accedidas empleando cualquier navegador VoiceXML estándar. En ambos casos, la conversión se realiza dinámicamente, de manera que si los contenidos cambian, se proporcionará la conversión de los contenidos actualizados. En primer lugar se presenta un sistema de conversión automática de contenidos web. El sistema trata de inferir la estructura inherente de las páginas HTML y en base a ella plantear la interacción hablada con el usuario. En segundo lugar se presenta un sistema que permite realizar la conversión de contenidos web de manera semiautomática. Primero es necesario desarrollar una aplicación vocal que especifica cómo llevar a cabo la conversión para cada página HTML. A continuación esa aplicación se despliega en el servidor, lo que permite que los usuarios puedan acceder a los contenidos convertidos.

5.2. Conversión automática de contenidos web

La conversión automática de contenidos web se centra en permitir el acceso a páginas web usando habla de manera que no sea necesario indicar cómo hay que realizar la conversión de los contenidos.

En los siguientes apartados se describe en detalle la propuesta. En primer lugar se describe el procedimiento que es preciso aplicar para realizar la conversión. A continuación se describe el sistema desarrollado que implementa dicho procedimiento. Posteriormente, se presenta un caso de estudio, que muestra el funcionamiento de la propuesta. Finalmente, se analizan las limitaciones del enfoque de

conversión automática de contenidos web.

5.2.1. Procedimiento de conversión

Para realizar la conversión automática de contenidos web proponemos un procedimiento basado en emplear diagramas de estados, tal y como se ha descrito en [60]. Los diagramas de estados se han venido utilizando para modelar diálogos. Empleando este formalismo se pueden especificar los estados del diálogo y las transiciones entre ellos [33]. En cada estado hay un mensaje del sistema, que se enviará al usuario, y un modelo de lenguaje (*language model*, LM) para describir la entrada esperada del usuario. Cada transición tiene asociada una condición que es empleada para elegir el siguiente estado. Esta es una aproximación sencilla, donde el sistema tiene la iniciativa y guía al usuario a través de los contenidos del portal. Hay ciertas limitaciones en la flexibilidad, aunque este modelo es muy adecuado para la navegación, debido a que recuerda el de la web. Otra ventaja adicional de este modelo es que se puede implementar en VoiceXML de manera directa.

Para hacer corresponder los contenidos de la web con un diagrama de estados que represente la manera en la que el sistema plantea el acceso a la información, se han propuesto un conjunto de reglas:

1. La estructura del sitio web debe ser capturada empleando un grafo. Este grafo se puede emplear como primer diagrama de estados, que debe ser mejorado para conseguir una interacción amigable con el usuario.

2. Cada estado del diagrama de estados que contenga gran cantidad de información a enviar al usuario debe ser modificado para permitir un acceso adecuado a sus contenidos. Hay dos maneras de hacer esto:

   a) Dividir el estado en varios, añadiendo las transiciones adecuadas entre ellos.

   b) Habilitar un mecanismo de búsqueda a través de los contenidos del estado.

Finalmente, se debe asociar un modelo de lenguaje a cada estado. Para ello, se construye una gramática para cada estado empleando la información asociada a las transiciones que salen de dicho estado. Para los estados de búsqueda se construye una gramática que refleje las posibles preguntas que puede realizar el usuario.

5.2.2. Descripción del sistema

Partiendo del procedimiento descrito en el apartado anterior, se ha construido un sistema que automatiza la conversión. El sistema realiza la conversión de contenidos web de manera dinámica, estableciendo una correspondencia uno a uno entre las páginas HTML y las páginas VoiceXML: cada página HTML se convierte en una página VoiceXML.

La arquitectura propuesta puede verse en la figura 5.1, y es una extensión de la arquitectura propuesta en el estándar VoiceXML (ver apéndice A). La solución adoptada añade un intermediario entre el intérprete VoiceXML y el servidor de documentos que es el encargado de realizar la conversión de contenidos. La conversión se realiza mediante un *Servlet Java*, de manera que la comunicación con el navegador VoiceXML se realiza de manera estándar, empleando el protocolo HTTP. Se ha utilizado *Apache Tomcat* como servidor de aplicaciones.

Desde un punto de vista funcional, la conversión de contenidos se puede dividir en dos partes: (1) la extracción de información de páginas web y la estructuración de la misma; (2) la generación de los diálogos VoiceXML a partir de la información estructurada.

La fase de extracción de información es compleja, y para llevarla a cabo, se ha construido un módulo software que transforma una página HTML en una página XML correctamente estructurada. El proceso se muestra en la figura 5.2. Primero se corrigen todos los errores sintácticos empleando *Tidy*[1], que es un programa específicamente diseñado para ello. También se eliminan todas las etiquetas que contienen solamente información de formato. A continuación se construye el árbol DOM, que representa la jerarquía de etiquetas HTML. Finalmente se realiza la extracción de información, que identifica los elementos principales de la página (título, cuerpo y barra de navegación) y se genera el fichero estructurado en XML.

La fase de filtrado de las etiquetas HTML de formato pretende eliminar todas aquellas etiquetas que no aportan ninguna información acerca de la estructura de la página, sino que están orientadas a la correcta visualización de la misma. Las etiquetas eliminadas son las siguientes: `<font>`, `<comment>`, `<meta>`, `<p>`, `<div>`, `<br>`, `<center>` y `<script>`.

La fase de extracción de información obtiene tres elementos básicos dentro de una página:

1. **El título** de la página, que se obtiene del texto de la etiqueta `<title>`.

2. **El cuerpo** del documento, que está compuesto por secciones. Se organizará mediante un árbol, que se obtiene de la siguiente manera:

   - Las etiquetas `<h1>`, `<h2>`, `<h3>` y `<h4>` son las que marcan un inicio de sección.
   - Cuando encontramos alguna de ellas empezamos una sección nueva, en la que incluiremos todo lo que encontremos en el árbol DOM como hermano del nodo, hasta que encontremos una nueva etiqueta de inicio de sección.
   - Si la nueva etiqueta es de una sección de tipo más bajo la incluimos en la actual como hija y si es del mismo tipo la añadimos como hermana. Si es de un tipo mayor la incluimos en la sección superior.

[1] `http://www.w3.org/People/Raggett/tidy/`

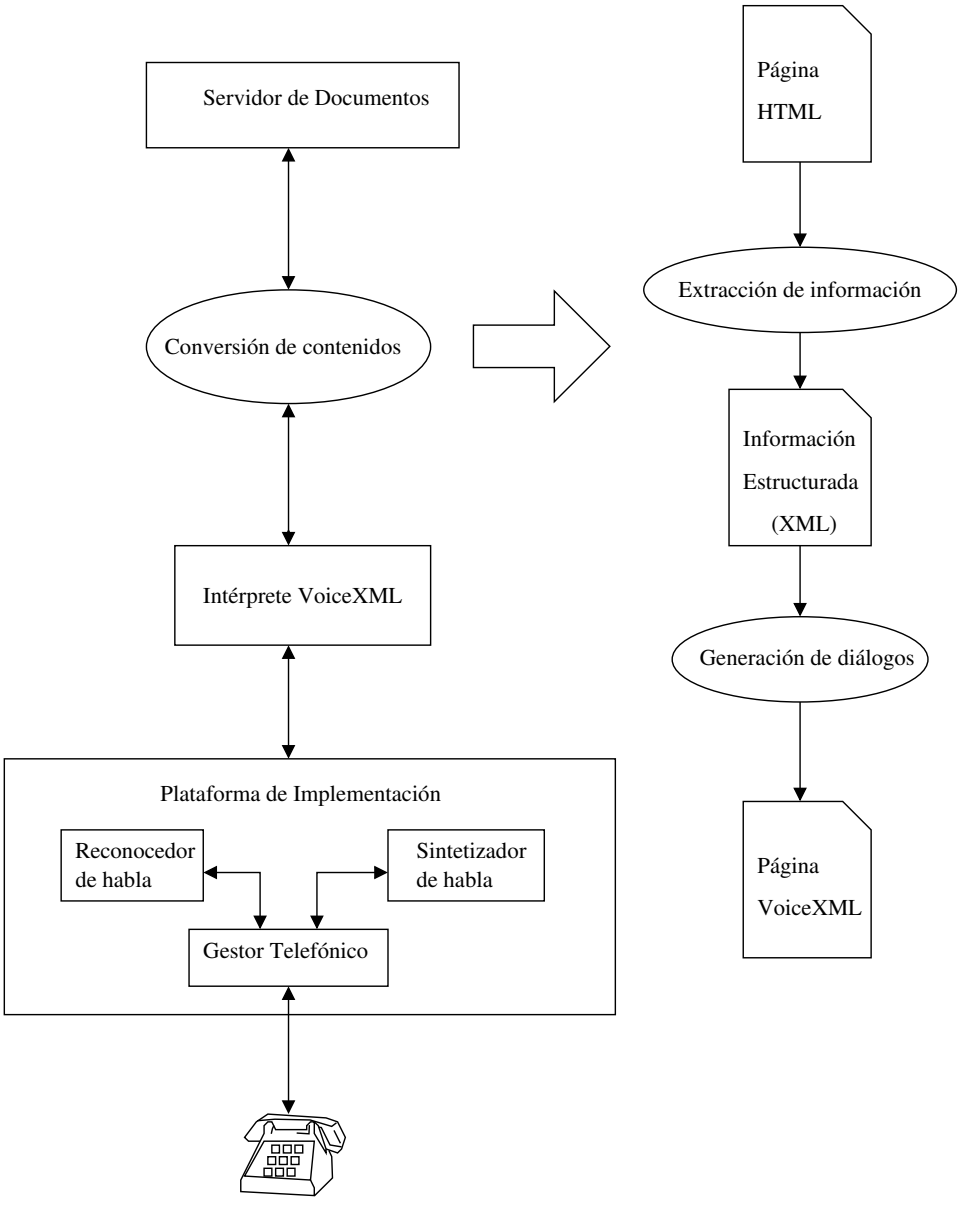


Figura 5.1: Arquitectura del sistema de conversión automática de contenidos web.

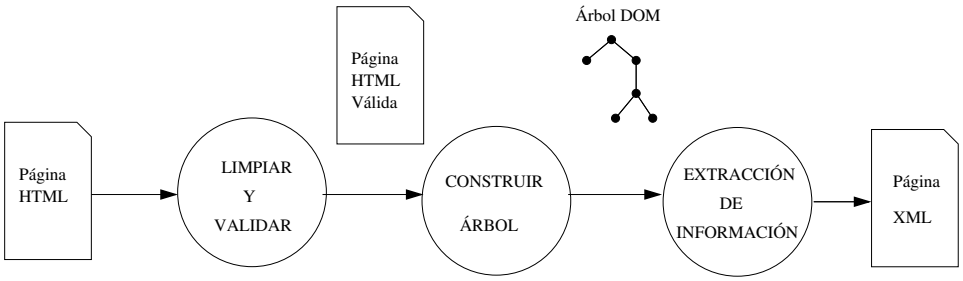


Figura 5.2: Módulo de extracción de información.

3. **Las barras de navegación**. Para localizar la barra de navegación nos centraremos en las tablas de la página (etiqueta `<table>`). Si todos los elementos que forman parte de una tabla son enlaces a otras páginas, entonces esa tabla será una barra de navegación. Cada uno de los enlaces será un ítem de navegación. Una de las formas más comunes para los ítem de navegación es que sean un nodo de tipo `<a>` y que tengan un hijo que sea de tipo `<img>`.

   Para cada ítem de navegación, almacenamos la dirección de destino y un nombre para el enlace. El nombre del enlace se elige entre tres posibilidades:

   a) Del texto que hay entre los elementos `<a>` y `</a>`.

   b) De la propiedad `alt` de los elementos `<img>` que hay entre `<a>` y `</a>`.

   c) Del título de la página destino.

Partiendo de la información estructurada, el sistema genera los diálogos VoiceXML correspondientes. La interacción planteada en dichos diálogos permite al usuario navegar dentro de las secciones de la página HTML y también permite acceder a otras páginas, de acuerdo con los siguientes pasos:

- El sistema presenta todas las secciones disponibles y el usuario elige una. El sistema envía al usuario la información de dicha sección.
- El procedimiento se repite y el usuario va accediendo a todas las secciones de la página que desee.
- Cuando el usuario ya ha obtenido toda la información que quería de la página, puede saltar a otra página empleando un menú presentado por el sistema a partir de la barra de navegación.

En aquellas páginas en las que la información del cuerpo está muy estructurada, como por ejemplo una tabla, se puede proporcionar un mecanismo de búsqueda que facilitará el acceso a la información. Para ello es preciso construir un modelo de lenguaje que contenga todos los ítems por los que el usuario puede realizar la búsqueda.

5.2.3. Caso de estudio

Para ilustrar el funcionamiento del sistema se ha planteado un caso de estudio, empleando la web del Departamento de Informática de la Universidad de Valladolid[2]. El sistema partió de los contenidos HTML de la página web del departamento, donde se puede encontrar información general, información de docencia, una descripción de los grupos de investigación y la información de contacto de los miembros del departamento. En cada página se han identificado tres partes: el título, la barra de navegación y el cuerpo, que está formado por secciones. La estructura de la web puede verse en la figura 5.3.

El sistema realiza la conversión de cada una de las páginas del sitio web. Vamos a mostrar cómo se realiza la conversión para la página de información general[3]. En la figura 5.4 puede verse la página web original y cada una de las partes que ha identificado el sistema automáticamente: el título, la barra de navegación y las tres secciones que componen el cuerpo. El diagrama de estados resultante que modela el diálogo para esta página se muestra en la figura 5.5.

El diagrama de estados asociado al sitio web completo puede verse en la figura 5.6 y es el resultado de realizar la conversión para todas las páginas que componen el sitio web.

En la figura 5.7 se muestra un ejemplo de interacción en el que el usuario accede a la página de información general. Otro ejemplo de interacción se muestra en la figura 5.8, y en este caso el usuario accede a la información sobre miembros del departamento. La página HTML original contiene información que puede ser accedida mediante una búsqueda, al tratarse de información con una estructura clara.

5.2.4. Limitaciones

La conversión automática de contenidos web tiene dos limitaciones principales:

- La gran variabilidad de páginas web hace complejo conseguir unas reglas para extraer y estructurar correctamente la información que funcionen para todo tipo de páginas. Este problema se complica dado que cada desarrollador emplea distintas estrategias para organizar la información, lo que hace imposible la construcción de un sistema suficientemente general.

- La interacción planteada al usuario puede resultar monótona y poco agradable. Por un lado, las páginas web suelen contener demasiada información, y cuando se realiza una interacción empleando habla se deben seleccionar los contenidos más relevantes, ya que si no la información puede abrumar al usuario. Por otro lado, en un sistema hablado, el diseño de la forma de interacción más adecuada para cada contenido es crucial, y difícil de llevar a cabo por un sistema automático. Ambos problemas se ven agravados por

[2] http://www.infor.uva.es

[3] http://www.infor.uva.es/informacion.htm

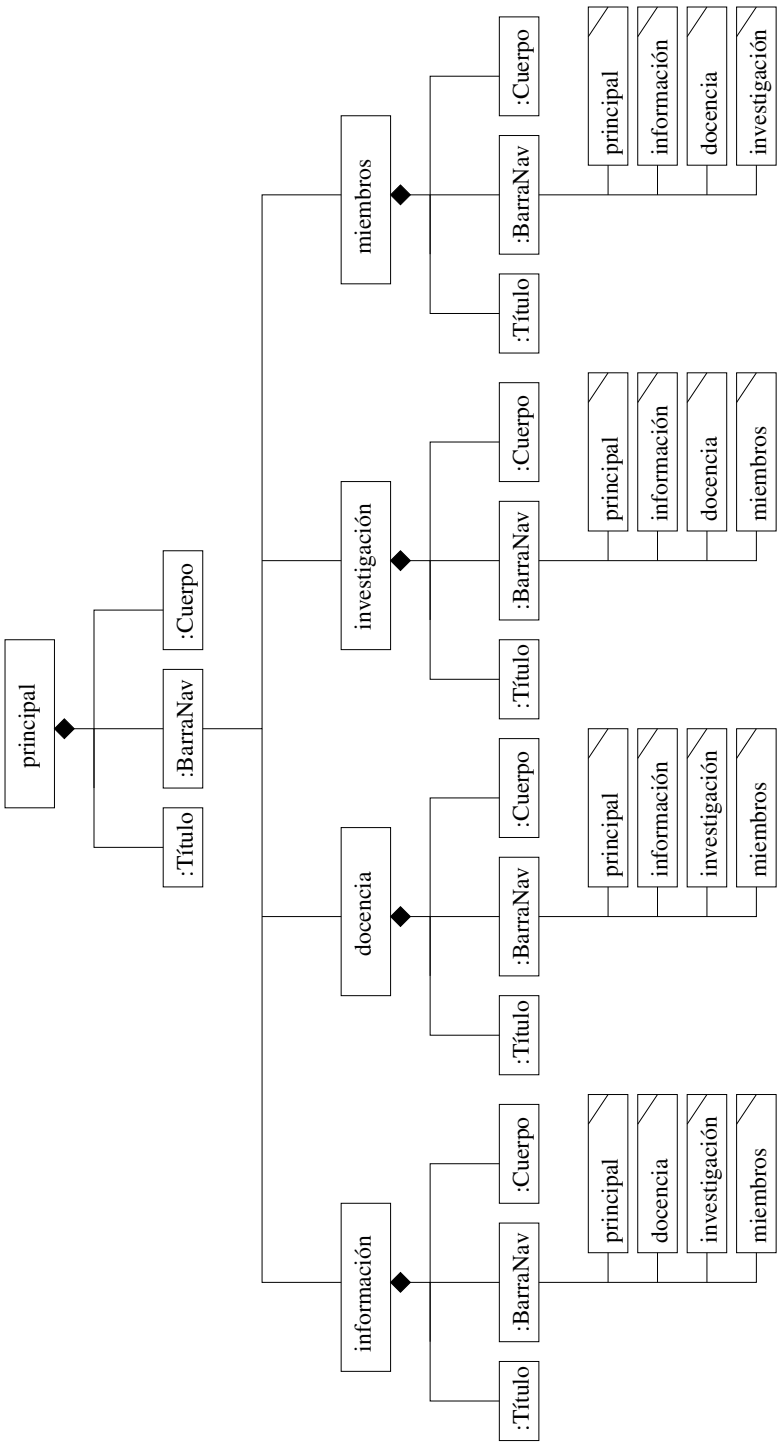


Figura 5.3: Estructura del sitio web del Departamento de Informática.

Figura 5.4: Página de información.

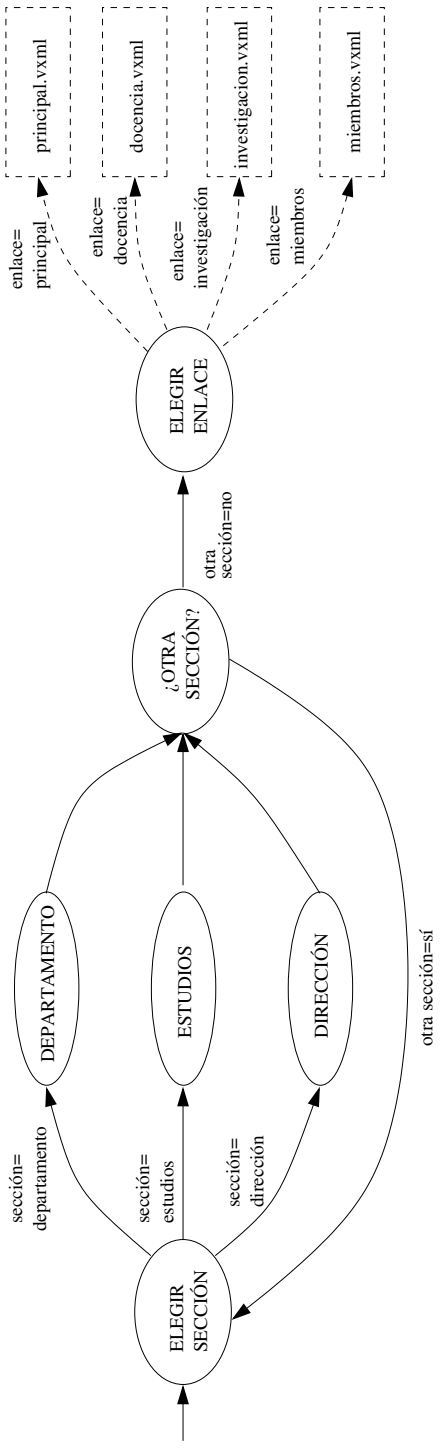


Figura 5.5: Diagrama de estados correspondiente a la página de información.

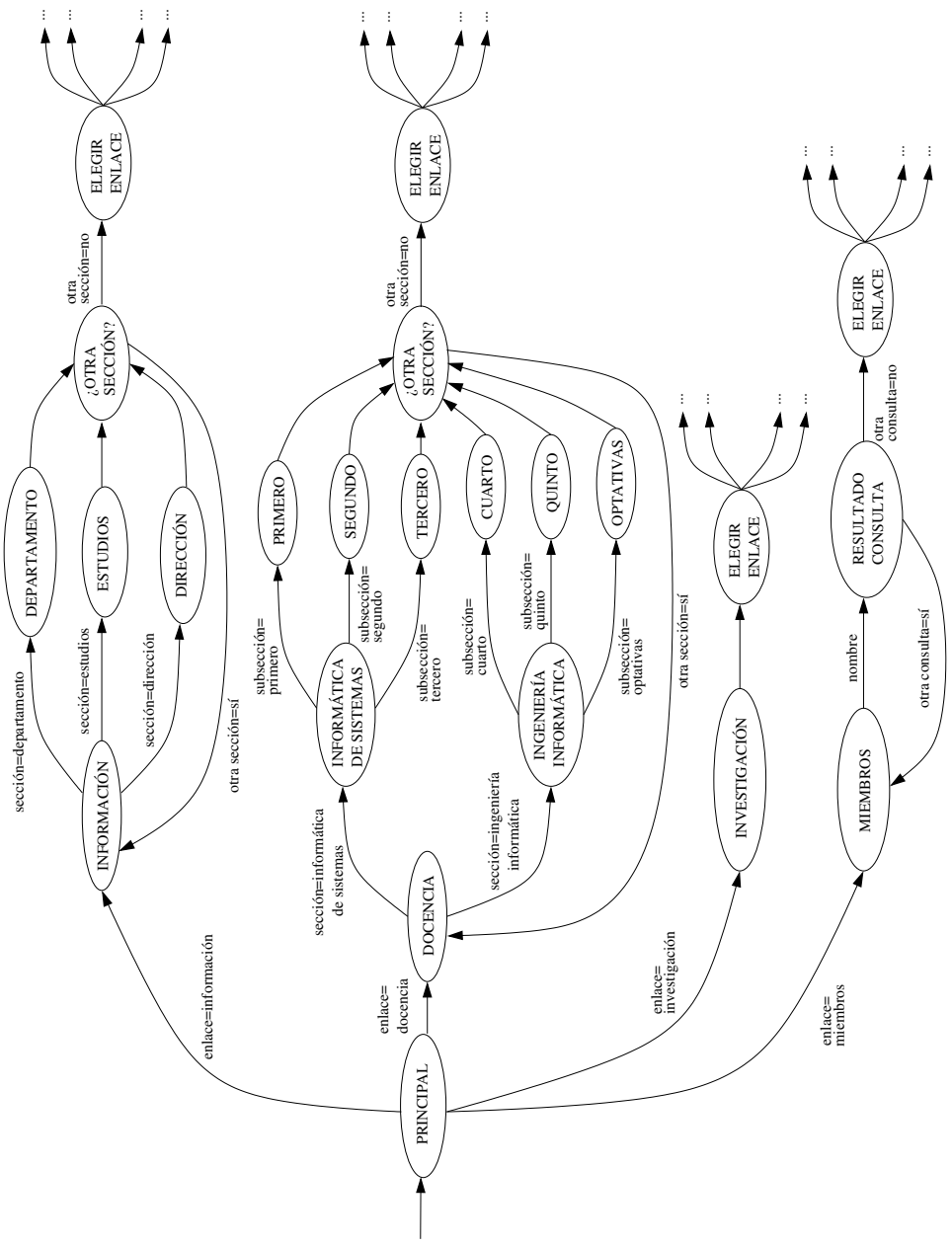


Figura 5.6: Diagrama de estados correspondiente al sitio web completo.

Sistema: Bienvenido al Departamento de Informática de la Universidad de Valladolid. Elija uno de los siguientes enlaces: información, docencia, investigación, miembros.
Usuario: Información.
Sistema: Elija una de las siguientes secciones: el departamento, los estudios, dirección.
Usuario: Dirección.
Sistema: Departamento de Informática. Edificio de Tecnología de la Información y de las Telecomunicaciones. Campus Miguel Delibes. 47011 Valladolid. España. Teléfono +34 983 423670, Fax +34 983 423671. ¿Desea acceder a otra sección?
Usuario: No.
Sistema: Elija uno de los siguientes enlaces: principal, docencia, investigación, miembros.
...

Figura 5.7: Ejemplo de interacción para acceder a la página de información.

Sistema: Bienvenido al Departamento de Informática de la Universidad de Valladolid. Elija uno de los siguientes enlaces: información, docencia, investigación, miembros.
Usuario: Miembros
Sistema: Servicio de Directorio del Departamento de Informática. Diga el nombre o apellido de la persona que busca:
Usuario: Pérez.
Sistema: Juan Pérez Crespo, teléfono: 3456. Luisa Martín Pérez, teléfono: 3452. ¿Otra consulta?
Usuario: No.
Sistema: Elija uno de los siguientes enlaces: principal, docencia, investigación, miembros.
...

Figura 5.8: Ejemplo de interacción para acceder a la página de miembros.

la ausencia de metainformación o información semántica que describa los contenidos.

Con el objetivo de superar estas limitaciones se ha planteado un enfoque semiautomático, que se describe a continuación.

5.3. Conversión semiautomática de contenidos web

La conversión semiautomática de contenidos web consiste en que un desarrollador indica al sistema cómo se deben procesar los contenidos para realizar la interacción hablada con el usuario. El objetivo es reutilizar la información HTML, convirtiéndola en páginas VoiceXML, pero de una manera controlada, de modo que la interacción resulte más agradable al usuario.

Primero es preciso crear una aplicación vocal que describa cómo debe hacerse la conversión para cada página HTML. Para ello se utilizan plantillas VoiceXML y reglas de extracción escritas en lenguaje XSLT. Se ha creado una herramienta de desarrollo para simplificar y facilitar la construcción de aplicaciones vocales. Una vez se ha construido la aplicación vocal, se emplea un servidor de conversión para acceder a la información empleando habla.

Puesto que la manera de organizar la información es similar para los distintos sitios web, se han identificado cinco patrones HTML que se utilizan frecuentemente en las páginas web. Para cada uno de ellos se ha proporcionado una manera de acceder a la información empleando habla.

En las siguientes secciones se presenta en detalle nuestra propuesta para la conversión semiautomática de contenidos web, que se encuentra también descrita en [54]. En primer lugar se hace una descripción del sistema, de sus componentes y de las aplicaciones vocales. A continuación se explica cómo realiza el sistema la conversión de páginas HTML en páginas VoiceXML, se presenta la herramienta de desarrollo diseñada para facilitar la creación de aplicaciones vocales, y se describe el funcionamiento del servidor de conversión. Posteriormente se enumeran los patrones web comunes y se muestra cómo acceder a sus contenidos empleando habla. Después se presenta un caso de estudio para ilustrar el funcionamiento del sistema. En último lugar, se analizan las limitaciones del enfoque semiautomático de conversión de contenidos web.

5.3.1. Descripción del sistema

El sistema está formado por dos componentes principales, una herramienta de desarrollo y un servidor de conversión, como puede verse en la figura 5.9. El sistema usa un enfoque semiautomático para realizar la conversión. Primero, un desarrollador debe crear una aplicación vocal, empleando la *herramienta de conversión*, especificando cómo debe hacerse la conversión para cada página HTML. A continuación, la aplicación se despliega en el *servidor de conversión*, donde los usuarios pueden acceder a la aplicación.

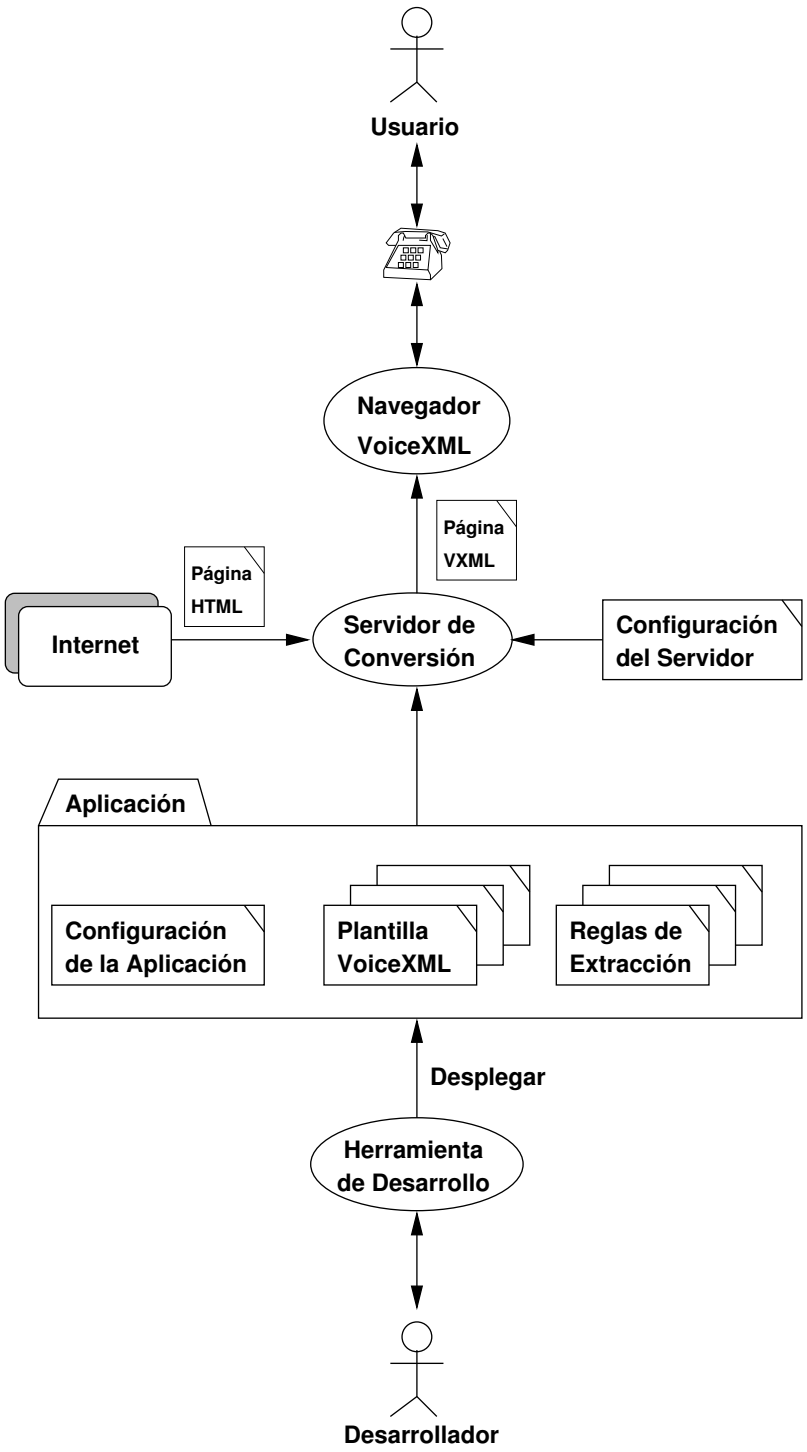


Figura 5.9: Arquitectura del sistema de conversión semiautomática de contenidos web.

Debido a las limitaciones del canal hablado, no es recomendable usar toda la información de los contenidos web originales. El desarrollador debe primero seleccionar qué elementos emplear y después debe diseñar la manera de convertirlos a VoiceXML. Una vez la aplicación ha sido construida, puede ser usada múltiples veces, incluso si la información cambia, siempre y cuando no haya cambios en la estructura de la página. Esto es posible ya que la conversión de contenidos se realiza de manera dinámica cuando el usuario accede a la aplicación.

Una aplicación vocal especifica cómo debe convertirse cada página HTML a VoiceXML. Está formada por tres elementos: un conjunto de ficheros con plantillas VoiceXML, que contienen el esqueleto de los ficheros VoiceXML resultantes; un conjunto de ficheros con reglas de extracción, que describen cómo extraer y transformar la información de las páginas HTML; y un fichero de configuración de la aplicación, que define para cada dirección URL qué plantilla y fichero de reglas deben ser usados.

5.3.2. Plantillas y reglas de conversión

El núcleo del sistema es la conversión de páginas HTML en páginas VoiceXML. El propósito de la conversión es doble: seleccionar los elementos de la página HTML original que van a ser usados en la aplicación vocal y describir cómo transformar el código HTML a VoiceXML.

La conversión se lleva a cabo en varios pasos, tal y como se muestra en la figura 5.10. Se emplea un esquema de conversión semiautomático, en el que el desarrollador tiene que especificar cómo se debe hacer la conversión, proporcionando dos ficheros XML:

- **Plantilla VoiceXML**: contiene la estructura de la página VoiceXML resultante. Tiene referencias a las reglas de extracción, que proporcionarán los contenidos finales de la página.
- **Reglas de extracción**: estas reglas se usan para seleccionar los elementos de la página HTML y transformarlos en VoiceXML. Estas reglas se escriben usando XSLT [28] y XPath [29].

Primero, la página HTML se transforma en una página XML, empleando *Tidy*[4]. A continuación, la plantilla VoiceXML y las reglas de extracción se unen, generando una página XSLT. Por último, la página XML con la información original se transforma, según indica la página XSLT, usando un procesador XSLT para producir la página VoiceXML resultante.

5.3.3. Herramienta de desarrollo

La herramienta de desarrollo permite a los desarrolladores construir aplicaciones vocales a partir de contenidos web y desplegarlas en el servidor de conversión.

[4] `http://www.w3.org/People/Raggett/tidy/`

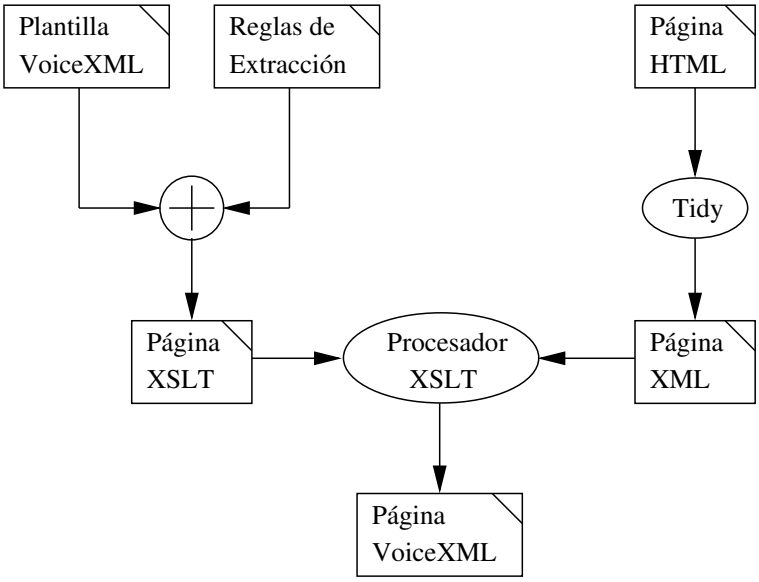


Figura 5.10: Transformación de una página HTML en una página VoiceXML.

Primero hay que seleccionar las páginas web a procesar. Para cada nueva dirección URL que se añade a la aplicación, hay que crear una plantilla VoiceXML y un conjunto de reglas de extracción. Esto puede hacerse usando un asistente, si la estructura de los contenidos web coincide con uno de los patrones predefinidos (ver sección 5.3.5), o manualmente en cualquier otro caso.

El sistema facilita la creación de plantillas VoiceXML mediante un menú en el que se pueden seleccionar cada una de las etiquetas necesarias. Al crear las reglas de extracción, es preciso construir expresiones en XPath para seleccionar los elementos de la página web original. Empleando la herramienta, los desarrolladores pueden construir dichas expresiones pinchando con el ratón en el elemento deseado de la página. Una vez se han creado la plantilla y las reglas de extracción, el sistema las aplica a la página web original, permitiendo a los desarrolladores comprobar el resultado obtenido. Esto les permite probar y depurar las aplicaciones. Por último, la aplicación puede desplegarse en el servidor de conversión.

5.3.4. Servidor de conversión

El servidor de conversión convierte las páginas HTML en páginas VoiceXML. La conversión se realiza de manera dinámica, al recibir una petición del navegador VoiceXML. Cuando el servidor recibe una petición, también recibe como parámetro la dirección URL que hay que convertir. El fichero de configuración de la aplicación especifica qué plantilla y qué reglas se tienen que usar para convertir esa URL. Empleándolas, la página web original se convierte a VoiceXML, tal y como se describe en la sección 5.3.2.

El servidor de conversión se ha implementado como un *Servlet Java*, de manera que la comunicación con el navegador VoiceXML se realiza de manera estándar, empleando el protocolo HTTP. Se ha usado *Apache Tomcat* como servidor de aplicaciones. El servidor de conversión usa un fichero de configuración para describir en que directorio deben desplegarse las aplicaciones.

5.3.5. Acceso hablado a patrones web comunes

Aunque existe una gran diversidad y variabilidad en los contenidos web, algunos patrones se usan frecuentemente para estructurar la información. En nuestro trabajo, se han identificado cinco patrones HTML típicos. Para cada uno de esos patrones se ha diseñado una manera de acceder a sus contenidos a través de una aplicación vocal. El uso de estos patrones ayuda en la automatización del desarrollo de aplicaciones vocales para acceder a información HTML en línea.

Los patrones seleccionados pueden verse en la tabla 5.1. Esta tabla muestra el nombre de cada patrón, sus características y la manera que se propone para acceder a sus contenidos usando habla. Se han seleccionado estos cinco patrones porque son los que se encuentran en las páginas HTML con mayor frecuencia.

5.3.6. Caso de estudio

Para mostrar cómo funciona el sistema para una página HTML dada, se ha incluido un ejemplo. Se ha utilizado una página web de *Yahoo!* que proporciona información sobre el índice Dow Jones Industrial Average [5]. En la figura 5.11 se muestra la página HTML original, la página VoiceXML generada por el sistema y un ejemplo de interacción con un usuario.

5.3.7. Limitaciones

La conversión semiautomática de contenidos web tiene dos limitaciones principales:

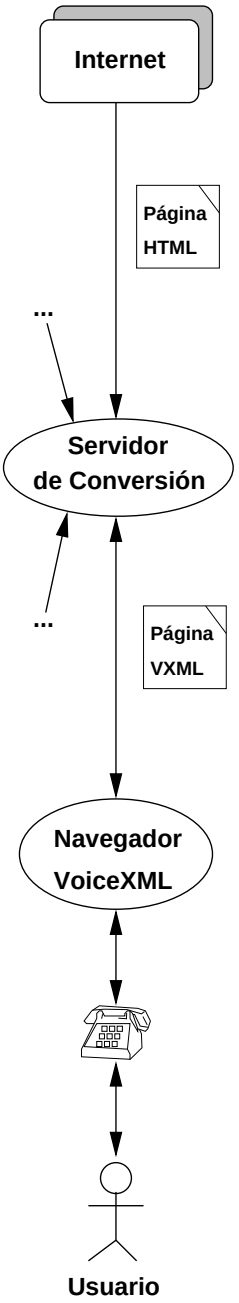
- Para cada sitio web al que se quiere acceder es necesario construir una aplicación vocal que describa cómo realizar la conversión. Esta solución permite conseguir una interacción hablada más amigable, puesto que la interacción está diseñada específicamente para cada contenido. Sin embargo, esto impone un alto coste, ya que requiere el desarrollo de múltiples aplicaciones vocales.

- La forma de acceder a la información está limitada por la estructura de los contenidos web originales. Aunque el diseñador de la aplicación vocal trate de adaptar los contenidos a las características del canal hablado, la información se obtiene por conversión de las páginas HTML originales, y por tanto, no dispone de completa libertad para plantear la interacción.

[5] http://finance.yahoo.com/q/cp?s=^DJI

Patrón	Características	Interacción Hablada
Texto	Información textual dividida en secciones. Cada sección tiene un título y un cuerpo.	Generar un mensaje describiendo el contenido de todas las secciones. Los títulos de todas las secciones se enumeran y el usuario selecciona una. Entonces, el cuerpo de esa sección es enviado al usuario.
Tabla	Información estructurada en filas y columnas. Se usa típicamente para describir objetos: cada objeto es una fila de la tabla y las columnas son sus propiedades.	Generar un mensaje describiendo el contenido de la tabla. El usuario selecciona un objeto de la tabla (una fila) empleando una de sus propiedades (columna). Enviar al usuario toda la información asociada a dicho objeto.
Formulario	Varios campos de información que deben ser rellenados por el usuario. Cada campo tiene un texto que describe ese campo. Hay un botón de envío, que manda toda la información al servidor web.	Generar un mensaje describiendo el contenido del formulario. La información para cada uno de los campos se solicita al usuario. Cuando el formulario está completo, toda la información es enviada al servidor web.
Lista	Un conjunto de elementos de texto, presentado cada uno en una línea diferente y empezando con un símbolo o un número. Existe la posibilidad de listas anidadas.	Generar un mensaje describiendo el contenido de la lista. Se presentan los elementos de la lista, en el mismo orden que la lista original. Si hay un elemento anidado, el usuario puede elegir entre navegarlo o continuar con el siguiente elemento del mismo nivel.
Barra de Navegación	Un conjunto de enlaces agrupados en forma de menú. Se suele encontrar en el lado superior o en el lado izquierdo de la página. Cada enlace tiene un texto que describe el contenido de la página a la que apunta.	Generar un mensaje describiendo la barra de navegación. Todos los elementos de la barra de navegación son enviados al usuario y éste elige uno de ellos. La interacción continúa en la URL del elemento seleccionado.

Tabla 5.1: Patrones HTML más frecuentes y su interacción hablada asociada.



YAHOO! FINANCE Sign In New User? Sign Up

Thursday, March 4, 2004, 11:38am ET - U.S. Markets close

1 - 30 of 30 | **First** | Previous | Next | **Last**

COMPONENTS FOR ^DJI

Symbol	Name	Last Trade	Change	Volume
AA	ALCOA INC	**36.84** 11:18AM ET	↓0.18 (0.49%)	1,476,600
AXP	AMER EXPRESS CO	**53.53** 11:18AM ET	↑0.04 (0.07%)	702,600
BA	BOEING CO	**42.60** 11:18AM ET	↓0.46 (1.07%)	977,900
C	CITIGROUP	**49.75** 11:18AM ET	↓0.07 (0.14%)	3,172,700
CAT	CATERPILLAR INC	**76.75** 11:18AM ET	↓0.57 (0.74%)	583,800
DD	DU PONT CO	**44.90** 11:18AM ET	↓0.17 (0.38%)	1,094,700
DIS	WALT DISNEY CO	**26.85** 11:18AM ET	↑0.20 (0.75%)	3,178,900
EK	EASTMAN KODAK	**28.05** 11:18AM ET	↑0.21 (0.75%)	535,400

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version="2.0">
<form id="start">
  <block>Bienvenido al índice Dow Jones Industrial Average.
  <goto next="#main"/></block>
</form>
<menu id="main">
  <prompt>Por favor, seleccione una empresa.</prompt>
  <choice next="#ALCOA_INC">ALCOA INC</choice>
  <choice next="#AMER_EXPRESS_CO">AMER EXPRESS CO</choice>
  <choice next="#BOEING_CO">BOEING CO</choice>
  <choice next="#CITIGROUP">CITIGROUP</choice>
  ...
</menu>
<form id="ALCOA_INC">
  <block>Última Cotización: 36.84 11:18AM ET<goto next="#main"/></block>
</form>
<form id="AMER_EXPRESS_CO">
  <block>Última Cotización: 53.53 11:18AM ET<goto next="#main"/></block>
</form>
<form id="BOEING_CO">
  <block>Última Cotización: 42.60 11:18AM ET<goto next="#main"/></block>
</form>
<form id="CITIGROUP">
  <block>Última Cotización: 49.75 11:18AM ET<goto next="#main"/></block>
</form>
...
</vxml>
```

Sistema:	Bienvenido al índice Dow Jones Industrial Average. Por favor, seleccione una empresa.
Usuario:	CITIGROUP
Sistema:	Última Cotización: 49.75 11:18AM ET. Por favor, seleccione una empresa.
	...

Figura 5.11: Conversión de la página de Yahoo! sobre el índice Dow Jones Industrial Average y un ejemplo de interacción.

5.4. Conclusiones

En este capítulo se han presentado dos enfoques para permitir el acceso a contenidos web empleando habla basados en la conversión de los contenidos originales. El enfoque propuesto en ambos casos está basado en la conversión de las páginas HTML originales en páginas VoiceXML, que pueden ser accedidas usando un navegador vocal.

La primera propuesta planteada realiza la conversión de manera automática. Para ello el sistema trata de inferir la estructura inherente en cada página, identificando los elementos que la componen. Una vez disponemos de la información estructurada, se generan las páginas VoiceXML que permiten el acceso a la información mediante habla.

Este enfoque automático pretende conseguir el acceso a los contenidos web de manera general. No obstante, el problema es complejo, y en la práctica existen contenidos para los que el sistema no es capaz de obtener la estructura interna, y por tanto, no es capaz de plantear una forma de acceso empleando habla. Por otro lado, en algunos casos la forma de interacción planteada no resulta agradable al usuario. El principal problema es la enorme cantidad de información disponible en las páginas web. Mostrar toda la información en una interfaz gráfica no es un problema, puesto que el usuario puede elegir la que más le interesa. Sin embargo, en una interfaz hablada proporcionar toda la información desborda al usuario.

El segundo enfoque está basado en realizar la conversión de manera semiautomática. En primer lugar se desarrolla una aplicación vocal que describe cómo realizar la conversión. A continuación, empleando el servidor de conversión, el usuario puede acceder a los contenidos mediante un navegador vocal. Para facilitar la creación de aplicaciones vocales se ha construido una herramienta de desarrollo y se han identificado cinco patrones HTML para los que se ha propuesto una forma de acceder a ellos empleando interacción hablada.

Este enfoque semiautomático trata de superar las limitaciones del enfoque automático. Para ello se introduce el concepto de aplicación vocal, que permite especificar cómo debe llevarse a cabo la conversión de los contenidos. En este caso, la identificación de la estructura de los contenidos web y el diseño de la interacción hablada es llevada a cabo por el desarrollador de la aplicación vocal. Esto permite conseguir una interacción más amigable, ya que un experto ha seleccionado los contenidos de interés y ha diseñado la forma más apropiada de acceder a ellos. Por tanto, los resultados obtenidos son mejores. Sin embargo, es necesario desarrollar una aplicación vocal para cada uno de los sitios web que se desea acceder.

Desde un punto de vista práctico, el enfoque semiautomático resulta muy interesante, ya que como modelo general es muy natural, aunque la limitación esencial está ligada a la necesidad de construir de forma manual tanto las plantillas como las reglas. Desafortunadamente, la obtención de las mismas de forma automática se muestra como un problema de muy elevada complejidad en el caso general, en el que habría que disponer de sistemas avanzados de procesamiento de lenguaje natural que permitiesen detectar los patrones lingüísticos que actúan

como indicadores de estructura.

Por último, cabe comentar que ambos enfoques están limitados por la estructura del sitio web original, dado que en ambos casos se realiza una conversión directa de páginas HTML en páginas VoiceXML. Para superar esta limitación y conseguir mejores resultados será preciso construir un modelo de información que permita una mayor libertad a la hora de plantear la interacción hablada con el usuario, tal y como se describe en el siguiente capítulo.

Capítulo 6

Sistema de diálogo hablado para el acceso a un sitio web

6.1. Introducción

En este capítulo se propone una forma de solución para el acceso a contenidos web mediante el empleo de un sistema de diálogo hablado. Partiendo de la información disponible en un sitio web, se pretende elaborar un sistema de diálogo hablado que permita al usuario acceder a la mayor parte de los contenidos de ese sitio, que podría visualizar con un navegador web estándar, pero en un formato que resulte adecuado a las características del canal hablado. En primer lugar se expone el planteamiento general de la propuesta. Seguidamente se justifica la idoneidad del dominio seleccionado. Posteriormente se presenta el sistema desarrollado, describiendo en detalle el modelo de interacción, el modelo de información y la arquitectura del sistema. Por último, se describe la evaluación del sistema llevada a cabo y se presentan los resultados obtenidos.

6.2. Planteamiento general de la propuesta

Debido a las diferencias existentes entre la interacción visual y la interacción hablada, no es posible realizar un mero cambio de los dispositivos de entrada y salida, sino que el sistema debe analizar la información y elaborar una forma de interacción adecuada que se adapte a las características del canal hablado. Para ello, elaboraremos un modelo de interacción y un modelo de información.

En los siguientes apartados se describe con detalle qué es cada uno de los modelos y cómo van elaborarse en nuestra propuesta.

6.2.1. Modelo de interacción

El modelo de interacción nos permite describir cómo interactúa el sistema con el usuario. El objetivo es conseguir una interacción lo más amigable posible y que

permita acceder a la información de manera rápida y sencilla. Es preciso tener en cuenta las características del canal hablado y las diferencias con el modo de interacción visual, que es como se encontraba la información inicialmente.

Para acceder a la información, el usuario puede emplear dos estrategias distintas: navegación y búsqueda. La navegación permite al usuario ver qué información está disponible antes de acceder a ella. Sin embargo, si el usuario tiene una necesidad de información específica, puede usar una pregunta para acceder directamente a ella. Además, debido al gran volumen de la información textual, la información deberá presentarse de manera gradual, a diferentes niveles de detalle.

El modo básico de interacción será la navegación de los contenidos: el sistema presenta varias opciones y el usuario elige una de ellas; el sistema le presenta la información seleccionada y vuelve a plantearle opciones al usuario. De este modo, el usuario irá navegando los contenidos del sitio web que le interesen.

Adicionalmente, el sistema permitirá que el usuario acceda a información concreta mediante la estrategia de búsqueda. El usuario realizará una consulta y a continuación podrá acceder a los contenidos relacionados con dicha consulta.

También será preciso disponer de estrategias de confirmación, para los casos en los que el reconocedor de habla no proporcione resultados fiables y estrategias de recuperación de errores.

6.2.2. Modelo de información

El modelo de información es el soporte que emplearemos para almacenar la información que obtendremos del sito web y se construirá automáticamente a partir de los contenidos de las páginas HTML.

Utilizaremos como modelo de información un grafo en el cual los nodos serán los elementos de información y los arcos serán las relaciones existentes entre dichos elementos de información. Trataremos que se refleje de la manera más fiel posible la estructura de la información presente en el sitio web original.

El modelo de información debe facilitarnos el acceso posterior a la información por parte del usuario que interactúa con el sistema. Por tanto, deberemos procesar los contenidos para incluir la información adicional necesaria. En el caso de la búsqueda de información, incluiremos una serie de índices que nos permitirán llevarla a cabo de manera eficiente.

6.3. Selección del dominio

El acceso a contenidos web mediante habla es una tarea difícil, principalmente debido a que la naturaleza de los documentos accesibles en Internet es muy variable. Por ello, para poder abordar en la práctica la tarea, hemos decidido restringir el sistema a un dominio que nos permita probar y demostrar la validez de nuestra propuesta.

Se han identificado una serie de requisitos que sería deseable que el dominio seleccionado cumpliera:

- Bien estructurado.
- Documentos en los cuales el texto predomine sobre el contenido gráfico.
- El objetivo sea más informacional que actuacional.
- Presentado en documentos web de estructura lo más invariable posible en el tiempo.
- Variedad de contenidos informativos, fáciles de catalogar.
- Contenidos redactados con arreglo a una guía de estilo fija.

Un dominio adecuado es la versión digital de un periódico, puesto que la información está bien estructurada y los contenidos textuales son de calidad. En concreto, se ha seleccionado el sitio web de *El Norte de Castilla*[1], que es un periódico local. En la figura 6.1 se puede observar la página de entrada del periódico seleccionado.

6.4. Descripción del sistema

Desde un punto de vista funcional, el problema se puede dividir en dos partes. Por un lado, la extracción y estructuración de la información presente en las páginas web. Por otro, la generación de los diálogos para interactuar con el usuario. Por tanto, se usan dos modelos para describir el sistema: modelo de información y modelo de interacción. En cuanto al sistema de diálogo hablado, se usa un enfoque basado en marcos para controlar el flujo del diálogo, lo que proporciona una mayor flexibilidad. La implementación del sistema ha sido presentada en [55, 61].

En los siguientes apartados se describe en detalle el sistema desarrollado. En primer lugar se presenta el modelo de interacción, en el que se explican las estrategias de navegación y búsqueda. A continuación se describe el modelo de información, compuesto por un árbol de navegación y varios índices de búsqueda. Posteriormente, se describe la arquitectura del sistema, donde se dan detalles acerca de la implementación. Por último, se incluye un caso de estudio en el que se ilustra el funcionamiento del sistema empleando varios ejemplos de interacción entre el usuario y el sistema.

6.4.1. Modelo de interacción

Vamos a utilizar diagramas de estados para describir las dos estrategias de acceso a la información. Se han elegido este tipo de diagramas únicamente para describir en este documento de manera sencilla la forma de interacción, puesto que el

[1] `http://www.nortecastilla.es`

Figura 6.1: Versión digital de *El Norte de Castilla*.

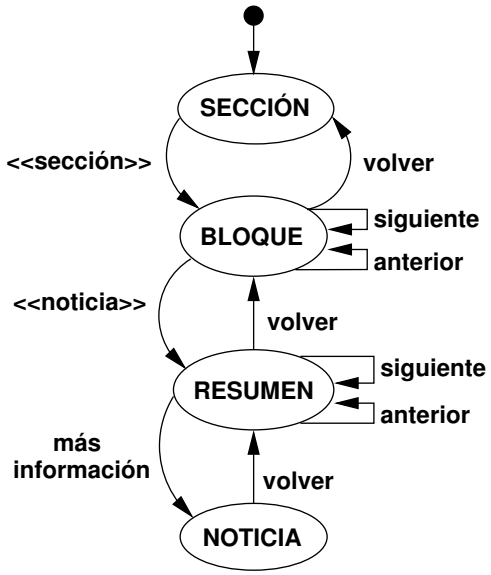


Figura 6.2: Diagrama de estados para navegación.

sistema de diálogo está basado en marcos, y permite una forma de interacción más flexible que la que describen los diagramas, tal y como se detalla en el apartado 6.4.3.2.

6.4.1.1. Navegación

El mecanismo de navegación es útil cuando el usuario no tiene una necesidad de información específica, y quiere saber qué información está disponible.

La información se presenta de forma gradual, en distintos niveles de detalle. El diagrama de estados usado para navegación puede verse en la figura 6.2. Cuando el usuario navega por la información, la interacción es muy similar a moverse a través del árbol que contiene la información (ver figura 6.4).

Primero, el usuario selecciona la sección del periódico a la que quiere acceder. A continuación, el sistema presenta los titulares de todas las noticias de esa sección. Si hay más de cinco noticias, se agrupan en bloques. Entonces, el usuario selecciona una noticia y el sistema presenta un breve resumen de esa noticia, Por último, si el usuario quiere más información, puede acceder a la noticia completa. En caso contrario, puede volver atrás y seleccionar otra noticia.

Las noticias en cada sección están agrupadas en bloques de 5, para evitar presentar todas las noticias a la vez. En cada bloque, el usuario puede elegir una noticia o ir al siguiente bloque.

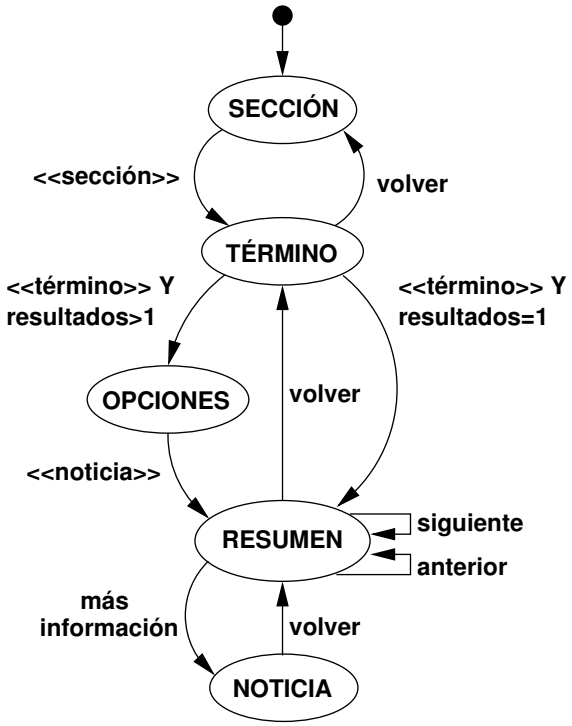


Figura 6.3: Diagrama de estados para búsqueda.

6.4.1.2. Búsqueda

El mecanismo de búsqueda es útil cuando el usuario tiene una necesidad de información específica que puede expresar mediante una consulta. La complejidad de la consulta depende del sistema, desde términos aislados hasta preguntas en lenguaje natural, en función de las capacidades de la tecnología de reconocimiento y comprensión de habla disponible.

El diagrama de estados que describe la estrategia de búsqueda puede verse en la figura 6.3. El usuario primero selecciona la sección del periódico a la que quiere acceder y después hace una pregunta. Entonces, el sistema busca la información solicitada en el índice invertido de esa sección (ver figura 6.5). El sistema presenta los titulares de las noticias que están relacionadas con la pregunta y el usuario selecciona una. El sistema presenta el resumen de esa noticia y el usuario puede acceder a la noticia completa para obtener más información.

6.4.2. Modelo de información

El modelo de información contiene toda la información del dominio y se construye automáticamente a partir de los contenidos del sitio web. Una vez elegido el dominio de aplicación, es preciso definir cuál es la información a tratar. Para el do-

minio seleccionado en nuestro caso, la información a manejar serán las secciones y las noticias del periódico digital:

- Cada sección contiene una serie de noticias.
- Las noticias están compuestas por varios elementos: titular, sección, autor, resumen y cuerpo.

Tal y como se describe a continuación, el modelo de información consta de dos elementos: un árbol de navegación e índices de búsqueda. Esto permite al sistema implementar las estrategias de navegación y búsqueda de manera eficiente.

6.4.2.1. Árbol de navegación

La estructura de datos más adecuada para la navegación de contenidos es un árbol, en el que los elementos de información están en las hojas y en los nodos internos el usuario decide qué información le interesa. Cuando nos movemos por el árbol, en cada nodo el sistema realiza una pregunta al usuario y la respuesta nos dice cuál es el siguiente nodo que tenemos que visitar a continuación.

La figura 6.4 muestra la estructura del árbol. Los contenidos del periódico están organizados en 15 secciones. Cada sección tiene varias noticias, agrupadas en bloques de cinco. Cada noticia está estructurada en tres niveles diferentes de detalle: titular, resumen y cuerpo. Usando esos elementos estructurales se construye el árbol. El cuerpo de las noticias está en las hojas del árbol y los resúmenes en el nivel superior. Por último, una sección está compuesta por varios bloques. Acceder a la información usando navegación consiste en moverse a lo largo de los nodos del árbol.

6.4.2.2. Índices de búsqueda

Un índice es una estructura de datos que permite la búsqueda de información de manera eficiente. Para cada término almacena todos los documentos en los que dicho término aparece, tal y como se muestra en la figura 6.5. Para construir el índice, se usa el modelo de espacio vectorial [114]. Cada documento se representa mediante un vector en el espacio de documentos. Cada dimensión del espacio corresponde con un término en la colección de documentos (se puede usar un algoritmo de lematización para reducir la dimensionalidad del espacio). Dado un documento, hay varios métodos de calcular el valor de cada coordenada del vector. Se ha utilizado el llamado TF-IDF (*term frequency-inverse document frequency*). Se usa la siguiente fórmula para calcular el peso, w, de cada término en el documento:

$$w = (1 + log(tf)) * log\frac{N}{df} \tag{6.1}$$

donde tf es el número de veces que el término aparece en el documento; df es el número de documentos en los que dicho término aparece; y N es el número de documentos de la colección.

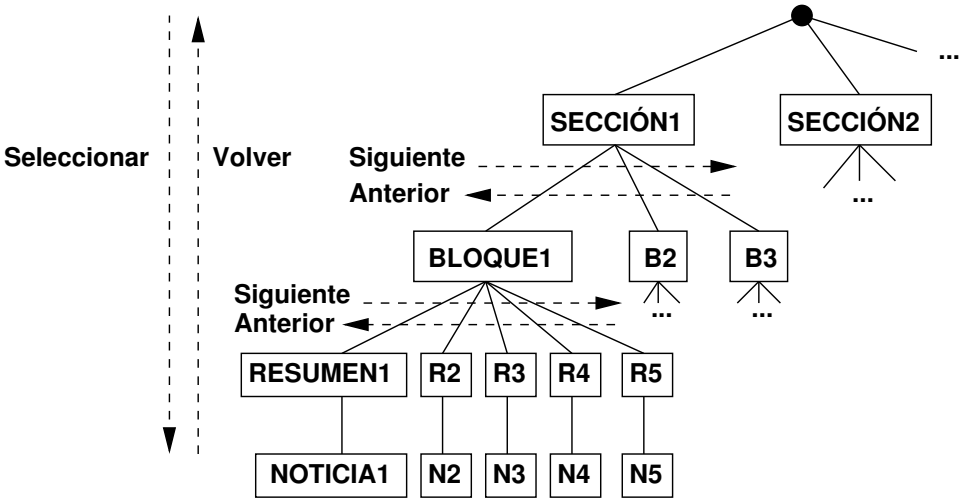


Figura 6.4: Árbol de navegación.

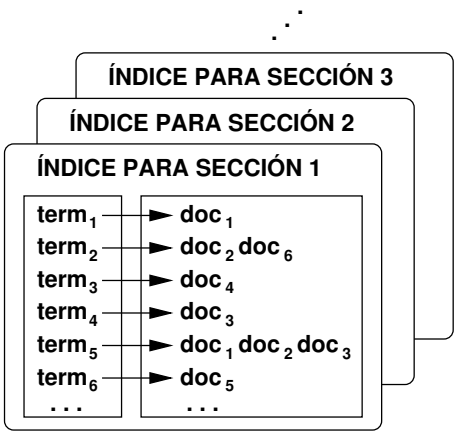


Figura 6.5: Índices de búsqueda.

6.4.3. Arquitectura del sistema

La arquitectura del sistema puede verse en la figura 6.6. El sistema tiene dos partes principales. La primera procesa toda la información y construye el modelo de información. La segunda dialoga con el usuario utilizando el modelo de interacción.

6.4.3.1. Procesamiento de la información

La información se obtiene y se procesa mediante programas desarrollados específicamente para el sitio web elegido. Para la construcción de los programas, partimos del conocimiento de cómo se organizan las páginas HTML del periódico digital.

En primer lugar, se descargan las páginas HTML del sitio web del periódico usando un recolector de páginas web (*web crawler*) y se almacenan en un repositorio local para su uso posterior. Hay dos tipos de páginas útiles para nuestro sistema: secciones y noticias. A continuación, las páginas HTML se convierten a XML. Primero, hay que limpiar y validar el código HTML, ya que algunas de las páginas contienen errores sintácticos. Entonces, se selecciona la información de interés y se genera un documento XML con la información estructurada adecuadamente. Para ello empleamos *Tidy*[2] y páginas XSLT.

Por último, el gestor de información construye el modelo de información. Empleando los contenidos estructurados en XML se construye el árbol de navegación directamente. Para cada sección del periódico se construye un índice. Primero, cada noticia se convierte a un vector: se extraen todos los términos y se usa el lematizador *Snowball stemmer*[3]. A continuación, se calcula el peso de cada término usando TF-IDF. Por último, se construye el índice con los 25 componentes más relevantes de cada noticia. Esto nos permite obtener un vocabulario más pequeño, que es necesario dadas las limitaciones del motor de reconocimiento de habla utilizado.

Para usar el esquema de pesado TF-IDF se necesita una colección de documentos. Se han recopilado noticias de la página web del periódico durante más de un año (71.141 noticias). Con todas esas noticias se han construido diccionarios que proporcionan la frecuencia de documento para cada término, es decir, en cuántos documentos de la colección aparece. Se construye un diccionario distinto para cada sección del periódico, con el objetivo de conseguir resultados más exactos.

6.4.3.2. Diálogo con el usuario

Para gestionar el diálogo se usa un enfoque basado en marcos. El usuario debe proporcionar ciertos ítems de información para acceder a las noticias. Los elementos de información que se utilizan en el sistema pueden verse en la figura 6.7. La

[2] `http://www.w3.org/People/Raggett/tidy/`
[3] `http://snowball.tartarus.org`

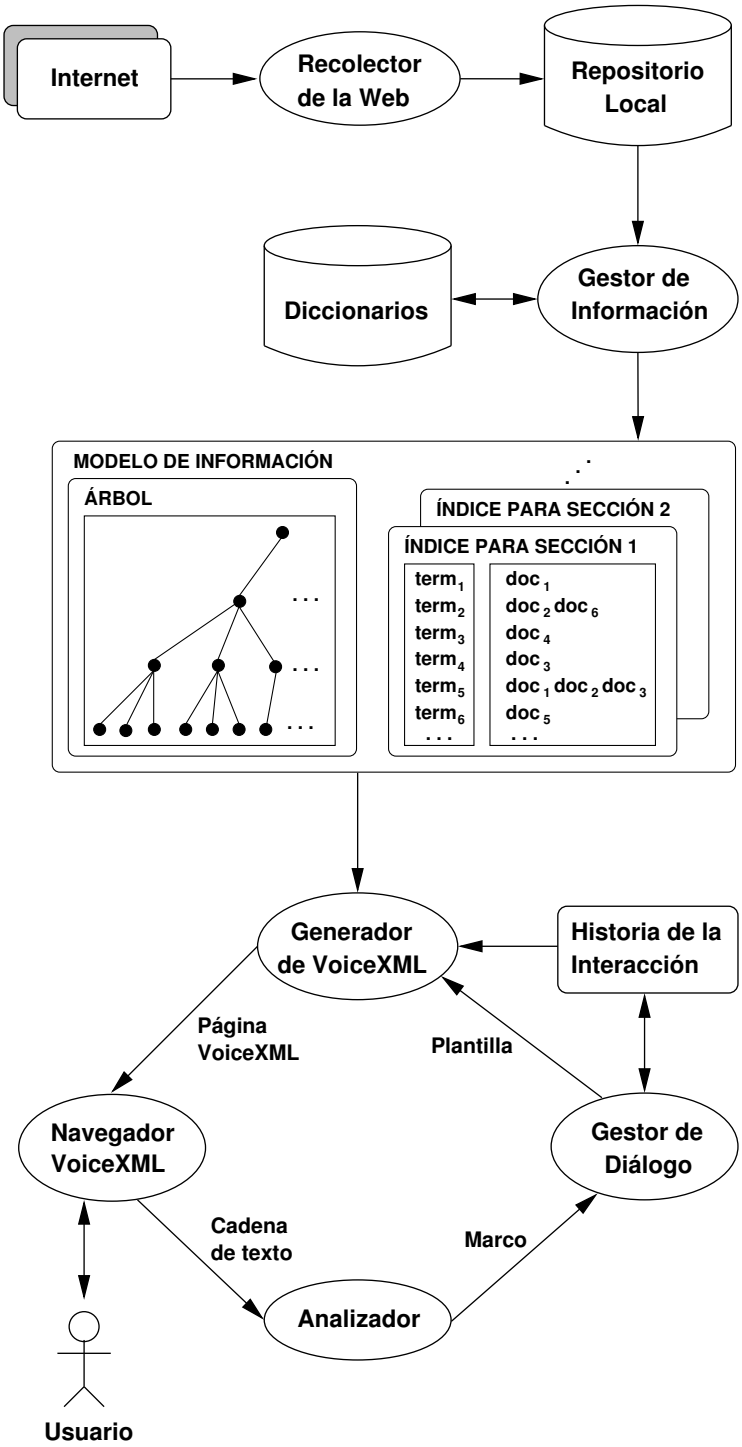


Figura 6.6: Arquitectura del sistema.

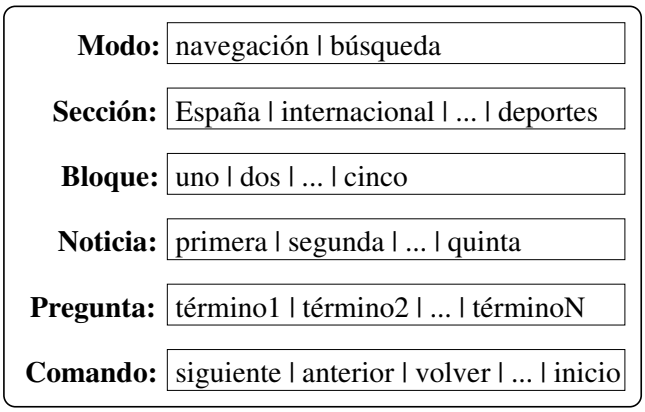


Figura 6.7: Marco de información.

ventaja principal de usar un enfoque basado en marcos es la flexibilidad, porque la interacción no está limitada por un flujo de control predefinido.

La interacción comienza con una página VoiceXML de bienvenida. Entonces, la entrada proporcionada por el usuario es procesada por el reconocedor de habla. El resultado se manda al analizador, que extrae los ítems de información y construye un marco. A continuación, el gestor de diálogo actualiza la historia de la interacción y selecciona una plantilla para generar dinámicamente la próxima página VoiceXML. El generador de VoiceXML usa esa plantilla, la historia de la interacción y el modelo de información para construir la siguiente página VoiceXML. El navegador VoiceXML interpreta esa página, manda un mensaje al usuario usando conversión texto-habla y obtiene la entrada del usuario usando reconocimiento de habla. El resultado del reconocedor del habla se envía al analizador y el bucle continúa hasta el final del diálogo.

Para describir la interacción con el usuario se usa el lenguaje VoiceXML. Cada página VoiceXML especifica el siguiente mensaje del sistema y qué es lo que puede decir el usuario. Para construir las páginas VoiceXML se usan 6 plantillas distintas. Cada una de ellas describe cómo construir una página VoiceXML usando el modelo de información y la historia de la interacción. La ventaja principal de usar VoiceXML es que se trata de una tecnología estándar, y se puede usar cualquier navegador VoiceXML para acceder a la información. Se ha probado el sistema usando nuestra plataforma VoiceXML, que se describe en la sección A.6 del apéndice A.

Para describir el conjunto de sentencias que puede utilizar el usuario como entrada al sistema se usan gramáticas en formato ABNF, que es uno de los tipos de gramáticas soportados por VoiceXML. Hay una gramática general para la interacción y gramáticas específicas para la búsqueda. Las gramáticas de búsqueda se generan dinámicamente en función de los contenidos de las noticias disponibles en el periódico, a partir de los índices del modelo de información. Para analizar las

cadenas de texto proporcionadas por el reconocedor de habla se ha construido un analizador LALR.

El sistema usa dos estrategias de confirmación distintas, dependiendo del tamaño del vocabulario del estado actual. Para tamaños de vocabulario pequeños se usa confirmación implícita, que es más rápida. Para vocabularios grandes se usa confirmación explícita, porque la probabilidad de que ocurra un error de reconocimiento es mayor. También se usan medidas de confianza del reconocedor de habla para rechazar hipótesis de reconocimiento con baja probabilidad. Por otro lado, con el objetivo de incrementar la velocidad del sistema para usuarios expertos, se ha habilitado la posibilidad de que el usuario interrumpa al sistema (*barge-in*).

Para mejorar la calidad de la síntesis de habla se ha etiquetado el texto a sintetizar para marcar pausas y enfatizar ciertas palabras. Esto permite conseguir una mayor naturalidad y evitar, en cierta medida, la monotonía. Todos los mensajes del sistema tienen dos versiones: la más descriptiva, empleada la primera vez que el usuario la escucha y una más breve usada en el resto de ocasiones.

En la historia de la interacción se almacenan dos tipos de información distintos: los ítems de información proporcionados por el usuario y la información de turnos de diálogo anteriores, tanto del sistema como del usuario.

El gestor de diálogo, el generador de VoiceXML y el analizador se han implementado usando la tecnología *Servlet Java*. Esto permite conectarlos de manera estándar con el navegador VoiceXML, usando el protocolo HTTP. Se ha utilizado *Apache Tomcat* como servidor de aplicaciones.

6.4.4. Caso de estudio

En la figura 6.8 se puede ver un ejemplo de interacción que ilustra el funcionamiento del sistema empleando la estrategia de navegación. El usuario selecciona en primer lugar la sección de cultura y a continuación el sistema enumera todos los titulares de dicha sección. De entre todos, el usuario elige el segundo y accede a la información de la noticia.

En la figura 6.9 se muestra una interacción de ejemplo en la que el usuario realiza una búsqueda de noticias relacionadas con “Cuba” en la sección internacional. El usuario elige una de las dos noticias encontradas y accede a la información de la noticia.

En los dos anteriores ejemplos el sistema ha utilizado una forma de interacción básica, indicada para usuarios novatos, en la que se sigue un flujo de control predefinido y el usuario proporciona los ítems de información necesarios de uno en uno. Sin embargo, también es posible una forma de interacción avanzada, indicada para usuarios expertos, en la que el flujo del diálogo no está predefinido y además es posible proporcionar varios ítems de información en una única locución, tal y como se muestra en el ejemplo de la figura 6.10.

Sistema: Bienvenido a la edición digital de El Norte de Castilla. Recuerde que en cualquier momento puede interrumpir al sistema. Diga ayuda siempre que lo desee para solicitar ayuda. Empleando navegación accederá a todas las noticias del periódico. Empleando búsqueda podrá elegir una palabra y localizar todas las noticias que la contienen. Diga navegación o búsqueda.
Usuario: Navegación.
Sistema: Elija una de las siguientes secciones: local, Castilla, España, internacional, economía, deportes, cultura. Si desea volver al menú principal diga inicio.
Usuario: Cultura.
Sistema: La sección cultura contiene trece noticias. Titulares del bloque de noticias uno. Primera noticia: Caballero Bonald recibe el Reina Sofía como un estímulo inolvidable. Segunda: Marian Izaguirre gana el premio de novela Ciudad de Salamanca. Tercera: La Junta potenciará el atractivo de la región para el aprendizaje del español. Cuarta: Tusquets crea un nuevo galardón dotado con 20.000 euros. Quinta: Rubens repintado por Rubens. Elija una noticia de la primera a la quinta, diga siguiente para ir al bloque siguiente o inicio para volver al menú principal.
Usuario: Segunda.
Sistema: Noticia de la sección cultura. Titular: Marian Izaguirre gana el premio de novela Ciudad de Salamanca. Resumen: La escritora bilbaína Marian Izaguirre y el cordobés Alejandro López Andrada se alzaron ayer con la novena y octava edición de los premios de novela y poesía Ciudad de Salamanca, respectivamente. Diga más información, siguiente, anterior, volver o inicio.
Usuario: Más información.
...

Figura 6.8: Ejemplo de interacción empleando la estrategia de *navegación*.

Sistema: Bienvenido a la edición digital de El Norte de Castilla. Recuerde que en cualquier momento puede interrumpir al sistema. Diga ayuda siempre que lo desee para solicitar ayuda. Empleando navegación accederá a todas las noticias del periódico. Empleando búsqueda podrá elegir una palabra y localizar todas las noticias que la contienen. Diga navegación o búsqueda.
Usuario: Búsqueda.
Sistema: Elija una de las siguientes secciones: local, Castilla, España, internacional, economía, deportes, cultura. Si desea volver al menú principal diga inicio.
Usuario: Internacional.
Sistema: Búsqueda en la sección internacional. Diga una palabra para buscar las noticias que la contienen. Diga inicio para volver al menú principal.
Usuario: Cuba.
Sistema: He entendido Cuba ¿Es correcto?
Usuario: Sí.
Sistema: En la sección internacional hay dos noticias con la palabra Cuba. Primera noticia: Cuba excarcela a seis disidentes que habían sido trasladados a hospitales. Segunda: Gestiones y gestos. Elija una noticia de la primera a la segunda, diga búsqueda para cambiar la palabra de búsqueda o inicio para volver al menú principal.
Usuario: Primera.
Sistema: Noticia de la sección internacional. Titular: Cuba excarcela a seis disidentes que habían sido trasladados a hospitales. Resumen: Los condenados forman parte de 75 presos políticos que el pasado año fueron enviados a prisión. Diga más información, siguiente, volver, búsqueda o inicio.
Usuario: Más información.
...

Figura 6.9: Ejemplo de interacción empleando la estrategia de *búsqueda*.

Sistema: Bienvenido a la edición digital de El Norte de Castilla. Recuerde que en cualquier momento puede interrumpir al sistema. Diga ayuda siempre que lo desee para solicitar ayuda. Empleando navegación accederá a todas las noticias del periódico. Empleando búsqueda podrá elegir una palabra y localizar todas las noticias que la contienen. Diga navegación o búsqueda.
Usuario: Acceso mediante navegación en la sección de Valladolid.
Sistema: La sección Valladolid contiene treinta y tres noticias. Titulares del bloque de noticias uno. Primera noticia: Reanudada la vista oral por la tramitación de falsos permisos de trabajo. Segunda: Rescatan de las aguas del Pisuerga el cuerpo de un hombre ahogado. Tercera: El sector tecnológico ha crecido en los años noventa y ejerce mayor atracción. Cuarta: Los médicos impulsan nuevas estrategias para mejorar las consultas. Quinta: Renault duplica en tres años la fabricación de motores en Valladolid. Elija una noticia de la primera a la quinta, diga siguiente para ir al bloque siguiente o inicio para volver al menú principal.
Usuario: Pasa a la segunda noticia.
Sistema: Noticia de la sección Valladolid. Titular: Rescatan de las aguas del Pisuerga el cuerpo de un hombre ahogado. Resumen: El cuerpo sin vida de un hombre de 29 años, cuya identidad responda a las iniciales S. A. S. B, fue encontrado en el río Pisuerga después de que se precipitara al agua en la capital vallisoletana, según informan fuentes policiales. Diga más información, siguiente, anterior, volver o inicio.
Usuario: Acceso mediante búsqueda en la sección internacional.
Sistema: Búsqueda en la sección internacional. Diga una palabra para buscar las noticias que la contienen. Diga inicio para volver al menú principal.
...

Figura 6.10: Ejemplo de interacción de un usuario experto.

6.5. Evaluación del sistema

Se ha llevado a cabo una evaluación del sistema para medir su usabilidad y obtener realimentación de los usuarios finales, tal y como se describe en [55]. Para realizar la evaluación se han diseñado varios escenarios y se ha empleado un conjunto de usuarios que utilizaron el sistema para completar dichos escenarios. Cada escenario consiste en una descripción de cuál es la información a obtener. Al final de la evaluación cada usuario rellenó un cuestionario con su percepción sobre el sistema. Con todos los datos obtenidos se midió la usabilidad del sistema, empleando métricas objetivas para medir el rendimiento del sistema y métricas subjetivas para medir la satisfacción de los usuarios.

6.5.1. Descripción del proceso de evaluación

El sistema ha sido evaluado por 22 usuarios, que en su totalidad eran alumnos de la Universidad de Valladolid. La evaluación se realizó a través de la línea telefónica y tuvo lugar en un entorno controlado, en un despacho en el que estaba el usuario junto con el evaluador.

Todos los usuarios resolvieron 5 escenarios (ver sección B.1 del apéndice B). Para cada uno de ellos, se realizaba una llamada distinta en la que se intentaba obtener la información solicitada. Si con la primera llamada no se conseguía obtener toda la información, se permitía realizar otra llamada para resolver el escenario. Si con la segunda llamada tampoco se conseguía, se consideraba que el escenario no había sido resuelto satisfactoriamente y se pasaba al siguiente escenario. Una vez resuelto el escenario, se introducía la información obtenida en un formulario web. Las respuestas a cada escenario se usaron para calcular la tasa de éxito de la tarea.

Para cada llamada, el sistema de diálogo genera un fichero de bitácora (*log*) que registra todos los detalles de la interacción entre el usuario y el sistema. Este fichero se utiliza posteriormente para calcular las medidas objetivas. También se guarda cada intervención del usuario en un fichero de sonido, que una vez transcrito manualmente, se utiliza para calcular la tasa de error de palabra del reconocedor de habla.

Para saber la opinión de los usuarios sobre el sistema, cada usuario rellenó un cuestionario después de completar los cinco escenarios. Los datos del cuestionario se utilizaron para medir la satisfacción de usuario. Además, una vez finalizado el cuestionario, se ofreció la posibilidad a los usuarios de hacer un comentario en formato libre sobre su percepción del sistema.

6.5.2. Medidas empleadas en la evaluación

A continuación vamos a describir las medidas de rendimiento del sistema y las medidas de satisfacción de usuario utilizadas para medir la usabilidad del sistema.

6.5.2.1. Medidas de rendimiento

Para medir el rendimiento del sistema se han utilizado las siguientes medidas objetivas [51]:

- Tasa de éxito de la tarea: porcentaje de escenarios que han sido completados con éxito, sobre el total de escenarios.
- Duración media de las llamadas: media de la duración de las llamadas, medida en segundos.
- Turnos de usuario por llamada: número promedio de turnos de usuario por llamada.
- Tasa de error de palabra: tasa de error de palabra del reconocedor de habla (*WER*).
- Locuciones de usuario rechazadas: porcentaje de locuciones de usuario rechazadas debido a que el reconocedor de habla asigna un nivel bajo de confianza, sobre el total de locuciones de usuario.
- Turnos de sistema interrumpidos: porcentaje de turnos de sistema interrumpidos por el usuario, sobre el total de turnos de sistema.
- Falsas interrupciones: porcentaje de interrupciones que han sido mal detectadas (el sistema confunde ruido con habla del usuario), sobre el número total de turnos de sistema interrumpidos.
- Turnos de usuario sin entrada: porcentaje de turnos de usuario en los que el usuario no dice nada, sobre el total de turnos de usuario.
- Turnos de usuario de ayuda: porcentaje de turnos de usuario en los que el usuario solicita ayuda, sobre el total de turnos de usuario.

6.5.2.2. Medidas de satisfacción de usuario

Para medir la satisfacción de los usuarios, éstos completaron un cuestionario al finalizar el proceso de evaluación. Se ha utilizado el cuestionario SASSI (*Subjective Assessment of Speech System Interfaces*), que consiste en 34 afirmaciones que los usuarios deben puntuar [69] (ver sección 2.8.3). La lista de afirmaciones del cuestionario puede verse en la sección B.2 del apéndice B.

Las afirmaciones se ordenaron de manera aleatoria y distinta para cada usuario, para evitar que los resultados dependieran del orden de presentación. Algunas puntuaciones se invirtieron de manera que valores altos en todas las categorías fueran considerados buenos.

Los usuarios puntuaron cada afirmación empleando una escala de Likert de 7 puntos:

1. Totalmente en desacuerdo.
2. Desacuerdo.
3. Ligeramente en desacuerdo.
4. Neutro.
5. Ligeramente de acuerdo.
6. De acuerdo.
7. Totalmente de acuerdo.

A partir de esas afirmaciones se calculan 6 factores, cada uno de ellos mide un aspecto de la percepción de los usuarios sobre el sistema:

- Corrección en la respuesta del sistema: hace referencia a la percepción del usuario sobre si el sistema es preciso y sobre si éste hizo lo que esperaba. Este factor está relacionado con la habilidad del sistema para reconocer lo que dice el usuario, interpretarlo y actuar apropiadamente.
- Afabilidad: hace referencia a si el usuario percibe el sistema como útil, agradable y amigable.
- Demanda cognitiva: esfuerzo requerido para interactuar con el sistema.
- Molestia: hace referencia a si los usuarios perciben el sistema como repetitivo, aburrido, irritante y frustrante.
- Habitabilidad: hace referencia a si el usuario sabe lo que debe hacer y lo que el sistema está haciendo. Puede entenderse como la adecuación del modelo conceptual del usuario acerca del sistema de diálogo como agente conversacional.
- Rapidez: velocidad de respuesta del sistema.

6.5.3. Resultados de la evaluación

La media sobre todos los usuarios de las medidas objetivas de rendimiento obtenidas en la evaluación se muestran a continuación. El valor para cada usuario se puede consultar en la sección B.3 del apéndice B.

- Tasa de éxito de la tarea: 92 %.
- Duración media de las llamadas: 214,5 segundos.
- Turnos de usuario por llamada: 9,6.
- Tasa de error de palabra: 18,1 %.

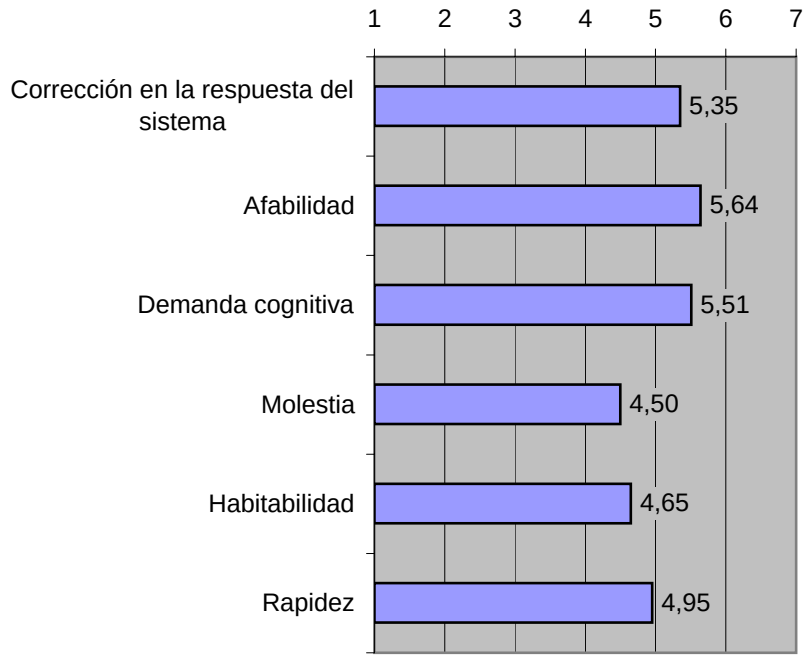


Figura 6.11: Factores de satisfacción de usuario.

- Locuciones de usuario rechazadas: 6,0 %.
- Turnos de sistema interrumpidos: 59,1 %.
- Falsas interrupciones: 9,0 %.
- Turnos de usuario sin entrada: 0,7 %.
- Turnos de usuario de ayuda: 0,2 %.

En cuanto a las medidas de satisfacción de usuario, los factores obtenidos se muestran en la figura 6.11. Las respuestas de los usuarios a cada pregunta concreta pueden verse en la sección B.4 del apéndice B.

6.5.4. Análisis de resultados

El sistema tiene una alta tasa de éxito de la tarea, lo que permite a los usuarios obtener la información deseada. El bajo número de turnos de usuario sin entrada o de ayuda muestra que los usuarios tienen claro qué decir en cada punto de la interacción.

La mayoría de errores de reconocimiento de habla tuvieron lugar en la estrategia de búsqueda, y muchos de ellos debido a que el usuario usó una palabra que

no estaba en el vocabulario. Esto es debido a que el dominio de noticias tiene un vocabulario con un número de palabras grande, y es difícil limitar al usuario a decir únicamente las palabras que están en el vocabulario del reconocedor de habla. También aparecieron algunos problemas con acrónimos y con palabras en otro idioma que el reconocedor de habla no era capaz de reconocer.

Los usuarios aprendieron rápidamente a interrumpir al sistema. Sin embargo, algunos de ellos tuvieron problemas, ya que el sistema interpretó su respiración de manera errónea como interrupción. Esto fue crítico para 3 usuarios con una tasa de falsas interrupciones superior al 20 %.

Se ha obtenido un resultado positivo de la satisfacción de usuario, ya que todos los factores son mayores de 4. Los factores con valores más altos son afabilidad y demanda cognitiva, que indican que a los usuarios les gusta el sistema, piensan que es útil, fácil de usar y que no requiere gran esfuerzo para interactuar con él. Factores con valores bajos son molestia y habitabilidad, lo que indica que los usuarios piensan que el sistema es repetitivo y aburrido, y algunas veces no supieron qué hacer o qué estaba haciendo el sistema.

El sistema permite usar dos estrategias para acceder a la información: navegación y búsqueda. Los usuarios prefieren usar búsqueda cuando quieren acceder a información específica, porque es más rápido. Sin embargo, cuando usan palabras que el reconocedor de habla no puede entender (palabras fuera del vocabulario) se sienten confusos, porque no saben qué es lo que está funcionando mal. Los usuarios se sienten más seguros empleando la estrategia de navegación porque se sienten en control de la interacción, aunque la interacción lleva más tiempo.

Se ha observado durante la evaluación que los usuarios tienden a cambiar de una estrategia a la otra cuando ocurren errores. Cuando los usuarios están usando una estrategia normalmente lo intentan dos o tres veces, y si no lo consiguen, cambian a la otra estrategia para acceder a la misma información. Se ha comprobado que algunos usuarios cambian de una estrategia a la otra dos o tres veces, hasta que encuentran la información que querían o hasta que se dan por vencidos y renuncian. Por tanto, disponer de dos estrategias de interacción diferentes puede servir como mecanismo de recuperación de errores.

6.6. Conclusiones

La construcción de un sistema de diálogo que permita a los usuarios acceder a contenidos web usando habla plantea varios retos a los que hay que hacer frente. Primero, hay que encontrar cierta estructura en los contenidos web para permitir el uso de un sistema de diálogo hablado. Segundo, es preciso plantear la interacción de manera amigable a los usuarios, utilizando las estrategias más convenientes según la información a la que el usuario quiere acceder.

En este capítulo se ha presentado un sistema de diálogo hablado que proporciona acceso al sitio web de un periódico. El sistema está basado en un modelo de información, en el que se han organizado las noticias en secciones y en bloques

dentro de cada sección. Cada noticia está a su vez dividida en tres niveles de detalle distintos: titular, resumen y cuerpo. El modelo de interacción está basado en las estrategias de navegación y búsqueda. La primera permite a los usuarios ver qué información está disponible y la segunda permite encontrar información específica. Se ha comprobado que ambas estrategias se complementan cuando se produce un error.

Se ha evaluado el sistema para medir su usabilidad. Los resultados muestran una alta tasa de éxito de la tarea. La mayoría de los errores de reconocimiento de habla tuvieron lugar en la estrategia de búsqueda, debido a palabras fuera del vocabulario. Los usuarios aprendieron rápidamente a interrumpir al sistema para ir más rápido. El estudio de satisfacción de usuario mostró que a los usuarios les gusta el sistema y piensan que es útil. Sin embargo, también piensan que es un poco aburrido y repetitivo.

Con el fin de superar las limitaciones en la estrategia de búsqueda, provocadas por los errores de reconocimiento del habla, se han planteado una serie de experimentos con un sistema de recuperación de información dirigida por habla, tal y como se describe en el siguiente capítulo.

Capítulo 7

Recuperación de información dirigida por habla

7.1. Introducción

Para acceder a contenidos web empleando habla se han propuesto diversos enfoques y uno de los más naturales y efectivos es emplear el habla como la entrada a un sistema de recuperación de información. De hecho, la búsqueda es una de las formas más empleadas de acceso a la web, y en muchos casos suele ser el punto de partida de los usuarios. Por otro lado, la búsqueda puede también ayudar a superar las limitaciones del canal hablado, que no permite el envío de grandes cantidades de información. Del mismo modo, el uso de habla como entrada a un motor de recuperación de información puede agilizar la búsqueda de información en dispositivos móviles.

El objetivo de la recuperación de información dirigida por habla es buscar información en una colección de documentos de texto empleando una pregunta hablada. Un área de trabajo relacionada es la recuperación de documentos hablados (*spoken document retrieval, SDR*), cuyo propósito es inverso: indexar y recuperar elementos relevantes de una colección de grabaciones sonoras de habla en respuesta a una pregunta de texto. Se ha invertido mucho esfuerzo en SDR y se han obtenido muy buenos resultados [50]. Sin embargo, la recuperación de información dirigida por habla es una tarea más difícil, porque las preguntas habladas contienen menos redundancia para superar los errores de reconocimiento del habla.

En el sistema descrito en el capítulo anterior el usuario podía acceder a la información empleando la estrategia de búsqueda, lo que permitía localizar los documentos de interés mediante una consulta. Sin embargo, se detectaron ciertas limitaciones en esta estrategia provocadas por los errores de reconocimiento del habla. Esto ha motivado la tercera fase de nuestro trabajo, que se presenta en este capítulo y en la que se han realizado diversos experimentos con un sistema de recuperación de información dirigida por habla, con el objetivo de analizar los factores que más influyen en el rendimiento de este tipo de sistemas y de proponer mejoras para re-

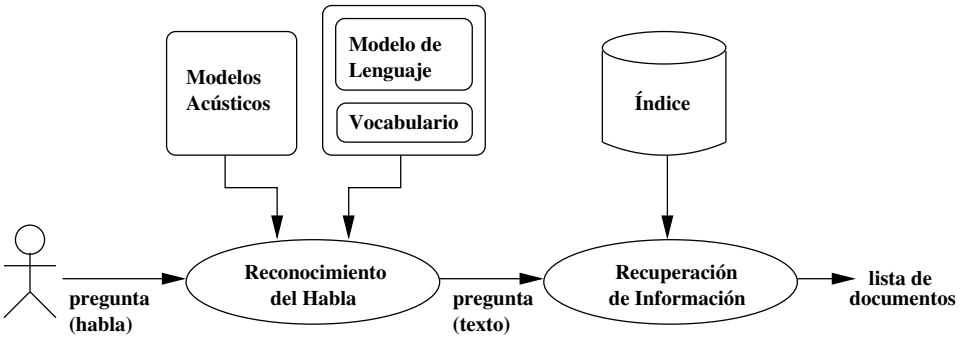


Figura 7.1: Arquitectura del sistema.

ducir el impacto de los errores de reconocimiento del habla en el rendimiento final del sistema.

En este capítulo se describe un sistema que permite realizar búsquedas en una colección de documentos empleando habla. En primer lugar se hace una descripción del sistema en la que se describen los detalles del reconocedor del habla y del motor de recuperación de información. A continuación se presenta la metodología de evaluación que vamos a utilizar. Seguidamente, se presentan los experimentos iniciales que nos han servido para establecer el sistema de referencia. Posteriormente, se proponen y evalúan varias mejoras sobre el sistema de referencia. Por último, se realizan una serie de experimentos finales, que incorporan todas las mejoras descritas anteriormente, y se comparan los resultados obtenidos con sistemas similares documentados en la bibliografía.

7.2. Descripción del sistema

El objetivo del sistema es recuperar todos los documentos relevantes para una pregunta hablada dada. El sistema consta de dos componentes: un reconocedor de habla continua de gran vocabulario y un motor de recuperación de información. La arquitectura del sistema se muestra en la figura 7.1. Primero, el usuario hace una pregunta empleando habla. Entonces, el reconocedor del habla transcribe la pregunta hablada a texto. Finalmente, el motor de recuperación de información obtiene la lista de documentos relevantes para esa pregunta.

En los siguientes apartados se describen en detalle el sistema de reconocimiento del habla y el sistema de recuperación de información.

7.2.1. Reconocimiento del habla

Para reconocimiento del habla se ha empleado SONIC, un reconocedor de habla continua de gran vocabulario de la Universidad de Colorado [102]. SONIC está basado en modelos ocultos de Markov (*hidden Markov models, HMM*) continuos e

implementa una estrategia de búsqueda en dos pasadas: la primera pasada consiste en una búsqueda de Viterbi con paso de testigo y la segunda pasada emplea un algoritmo A*.

Se han construido los modelos acústicos, el vocabulario y el modelo lenguaje para castellano, tal y como se describe en los siguientes apartados.

7.2.1.1. Modelos acústicos

Los modelos acústicos empleados por el reconocedor de habla son modelos ocultos de Markov de densidad continua. Además, se modela la duración de los estados mediante funciones de densidad de probabilidad gamma.

Para representar las unidades fonéticas empleadas por el reconocedor de habla se ha empleado el alfabeto fonético SAMPA [139]. Aunque SAMPA distingue 29 unidades fonéticas para castellano, es conveniente reducir el número de unidades a utilizar por dos motivos: por un lado, la realización de algunas unidades es muy similar en la práctica, y por otro, no se dispone de suficientes muestras de entrenamiento para ciertas unidades, lo que hace que los modelos no puedan ser entrenados correctamente. Se han empleado 23 unidades fonéticas para el reconocedor del habla, tal y como se muestra en la tabla 7.1.

Se ha utilizado una parametrización estándar, empleando coeficientes cepstrales en las frecuencias de mel (MFCC) y un vector de características de 39 dimensiones: 12 MFCCs y el logaritmo de la energía normalizada, junto con las derivadas de primer y segundo orden. Los vectores de parámetros se calculan con una ventana de 20 ms, que se va desplazando 10 ms.

Se han empleado modelos acústicos trifonema independientes de género. Los modelos se entrenaron empleando la base de datos oral Albayzin. Se ha utilizado el corpus fonético [98], compuesto por elocuciones de frases equilibradas fonéticamente y el corpus geográfico [39], compuesto por elocuciones de frases correspondientes a la consulta de una base de datos geográfica española. En ambos casos se ha utilizado tanto la parte de entrenamiento como la de test. En total se emplearon 13.600 frases leídas por 304 locutores, mitad hombres y mitad mujeres. La duración promedio de las frases es de unos 4 segundos.

7.2.1.2. Vocabulario y modelo de lenguaje

El corpus de entrenamiento empleado es EFE94, la colección de documentos objetivo sobre la que realizaremos las búsquedas. Esta colección se utilizó como material de entrenamiento porque esto permite adaptar el reconocedor del habla a la tarea y proporciona un rendimiento mejor del sistema [46]. La colección de documentos EFE94 está compuesta por todas las noticias de la agencia EFE del año 1994, lo que supone un tamaño de 511 Mb.

El material utilizado para entrenar los modelos es escrito, y es conveniente realizar una normalización del texto para acondicionarlo de manera que sea más representativo de una realización oral. Este proceso de normalización permite ob-

Vocales

	Anterior	Central	Posterior
cerrada	i		u
media	e		o
abierta		a	

Consonantes

	Bilabial	Labio-dental	Dental	Inter-dental	Alveolar	Palatal	Velar
Oclusiva	p, b		t, d				k, g
Fricativa		f		T	s	jj	x
Africada						tS	
Nasal	m				n	J	
Lateral					l		
Vibrante simple					r		
Vibrante múltiple					rr		

Tabla 7.1: Unidades fonéticas empleadas para reconocimiento del habla.

tener unos mejores resultados [2, 110]. Para llevar a cabo el acondicionamiento de los textos se ha construido un programa que realiza las siguientes tareas:

- Detectar correctamente el principio y el fin de las sentencias.
- Convertir las mayúsculas en minúsculas.
- Eliminar todos los signos de puntuación.
- Procesar los números para adecuarlos a la forma en la que aparecerán en las consultas habladas de los usuarios.
- Eliminar información que no forma parte del cuerpo de la noticia. En las noticias suele incluirse información adicional acerca del lugar, la fecha, el nombre de la agencia, el autor, etcétera. Por ejemplo, la noticia puede incluir una información al principio como `Nueva York, 31 dic (EFE)` y una nota al final como `PD/FMR 01/01/00-34/94`.

El vocabulario se construyó seleccionando las palabras más frecuentes encontradas en los documentos. La colección de documentos tiene 406.762 palabras diferentes. En los experimentos realizados se han empleado tres tamaños diferentes de vocabulario: 20.000, 40.000 y 60.000 palabras.

Para obtener la pronunciación de todas las palabras del vocabulario se utilizó el programa *ort2fon* [23], que realiza la transcripción ortográfico-fonética para castellano y está basado en reglas. La salida del transcriptor ha sido modificada para reducir el número de unidades fonéticas empleadas a 23, que son las que utiliza el sistema.

Como modelo de lenguaje se utilizan trigramas basados en palabras. Para entrenar estos modelos se utilizó la herramienta de modelado estadístico de lenguaje SRILM [126], con un suavizado *Katz backoff*.

7.2.2. Recuperación de información

Para hacer la recuperación de información se ha utilizado un motor de recuperación de información desarrollado para castellano que está basado en el modelo booleano [3]. Hemos extendido este sistema para usar el modelo vectorial con distintos esquemas de pesado [114]. Se ha utilizado un algoritmo de lematización [1] para reducir la dimensionalidad del espacio. También se ha usado una lista de parada para eliminar las palabras función [2]. La lista de parada utilizada se describe en [117].

Se han realizado una serie de experimentos para seleccionar el esquema de pesado a utilizar por el motor de recuperación de información. Para ello se ha empleado una colección de prueba estándar para recuperación de información desarrollada por *Cross-Language Evaluation Forum (CLEF)* [17]. CLEF organiza campañas de evaluación de manera similar a TREC. Su objetivo es desarrollar una infraestructura para la prueba y evaluación de sistemas de recuperación de información en idiomas europeos, bajo condiciones estándar y comparables, tanto en contextos monolingües como multilingües.

En los siguientes apartados se describen la colección de prueba utilizada y los experimentos llevados a cabo.

7.2.2.1. La colección de prueba CLEF 2001

Se ha utilizado la colección de prueba para IR monolingüe para español del año 2001 (*CLEF 2001 Spanish monolingual IR test-suite*). La colección de prueba incluye una colección de documentos, un conjunto de temas y un conjunto de juicios de relevancia.

La colección de documentos EFE94 está formada por noticias del año 1994 de la agencia de noticias EFE. La colección contiene 215.738 documentos, lo que supone un tamaño de 511 Mb.

Los temas simulan necesidades de información y se utilizan para construir las preguntas. Cada uno de ellos tiene tres partes: un breve título, una descripción de una frase y una narración más extensa que especifica el criterio para la evaluación de relevancia. Un ejemplo de tema puede verse en la figura 7.2. La colección de

[1] `http://snowball.tartarus.org`

[2] `http://members.unine.ch/jacques.savoy/clef/index.html`

Título:	Colisiones navales.
Descripción:	Encontrar información sobre el número de personas heridas o muertas en colisiones entre barcos.
Narrativa:	Los documentos relevantes deben informar sobre el número de víctimas (muertos o heridos) en colisiones entre barcos o vehículos navales de cualquier tipo. Los documentos que hablan de las víctimas sin proporcionar datos no son relevantes.

Figura 7.2: Tema número 67 de la colección de prueba CLEF 2001.

	Título	Descripción	Narración
Longitud máxima	8	33	74
Longitud mínima	1	5	16
Longitud media	4,2	16,0	39,2

Tabla 7.2: Longitud máxima, mínima y media, medida en número de palabras, para cada una de las partes de todos los temas de la colección de prueba CLEF 2001.

prueba CLEF 2001 incluye 49 temas, numerados del 41 al 90 (el tema 61 ha sido eliminando). La longitud de cada parte de todos los temas de la colección de prueba CLEF 2001 puede verse en la tabla 7.2.

Los juicios de relevancia determinan el conjunto de documentos relevantes para cada tema. Para crear los juicios de relevancia es preciso que una serie de personas comprueben la relevancia de los documentos de la colección de documentos para cada uno de los temas. Para evitar que estos evaluadores tengan que comprobar todos los documentos de la colección, lo que resultaría excesivamente costoso, se emplea la técnica de *pooling*, según se describe en [17].

7.2.2.2. Elección del esquema de pesado

Se han realizado una serie de experimentos encaminados a determinar cuál es el mejor esquema de pesado del motor de recuperación de información empleando consultas en texto. Para ello, se midió el rendimiento para diferentes esquemas de pesado empleando la metodología de evaluación utilizada en las evaluaciones CLEF [17]. Las consultas con las que se evalúa el sistema se construyen empleando los temas de la colección de prueba CLEF 2001. Se van a utilizar preguntas de longitud media, construidas empleando el campo descripción de cada tema y preguntas cortas, construidas empleando el campo título de cada tema.

Como esquemas de pesado para los documentos, se han realizado experimentos con los 8 esquemas de pesado descritos en la sección 3.2.2. En el caso de las consultas, solamente presentamos los resultados empleando 2 esquemas de pesado distintos, ya que los demás proporcionan resultados similares. Esto es debido, por un lado, a que la normalización de las consultas afecta igual a todos los docu-

	Descripción	**Título**
nnn.nnn	0,1523	0,1943
lnn.nnn	0,3034	0,3816
ntn.nnn	0,2828	0,2625
ltn.nnn	0,4304	0,4427
nnc.nnn	0,2865	0,2445
lnc.nnn	0,3125	0,2929
ntc.nnn	0,3608	0,3012
ltc.nnn	0,3830	0,3341
nnn.ntn	0,2828	0,2625
lnn.ntn	0,4304	0,4427
ntn.ntn	0,3262	0,2953
ltn.ntn	0,4497	0,4419
nnc.ntn	0,3958	0,3150
lnc.ntn	0,4240	0,3726
ntc.ntn	0,3865	0,3444
ltc.ntn	0,3958	0,3588

Tabla 7.3: Precisión media promediada (MAP) para distintos esquemas de pesado y para preguntas de longitud media (descripción) y preguntas cortas (título). La notación utilizada ha sido presentada en la sección 3.2.2.

mentos y, por tanto, se obtiene la misma lista ordenada de documentos. Por otro lado, en las consultas los términos aparecen una única vez, lo que provoca que los resultados sean iguales usando la frecuencia de término natural y la logarítmica.

Los resultados obtenidos para distintos esquemas de pesado pueden verse en la tabla 7.3, en la que aparece el MAP para cada esquema de pesado. En las figuras 7.3 y 7.4 pueden verse las curvas precisión/cobertura. Analizando los resultados observamos que los mejores esquemas de pesado son $ltn.nnn$ (que es igual a $lnn.ntn$) y $ltn.ntn$. Para el resto de experimentos de este capítulo usaremos el esquema de pesado $ltn.ntn$.

7.3. Metodología de evaluación

Para evaluar el rendimiento del sistema se ha empleado una colección de prueba estándar para recuperación de información extendida para incluir preguntas habladas. Se ha utilizado la colección de prueba para IR monolingüe para español del año 2001 (*CLEF 2001 Spanish monolingual IR test-suite*).

En los siguientes apartados vamos a describir cómo se ha realizado la grabación de las consultas y cómo se evalúan los resultados del sistema.

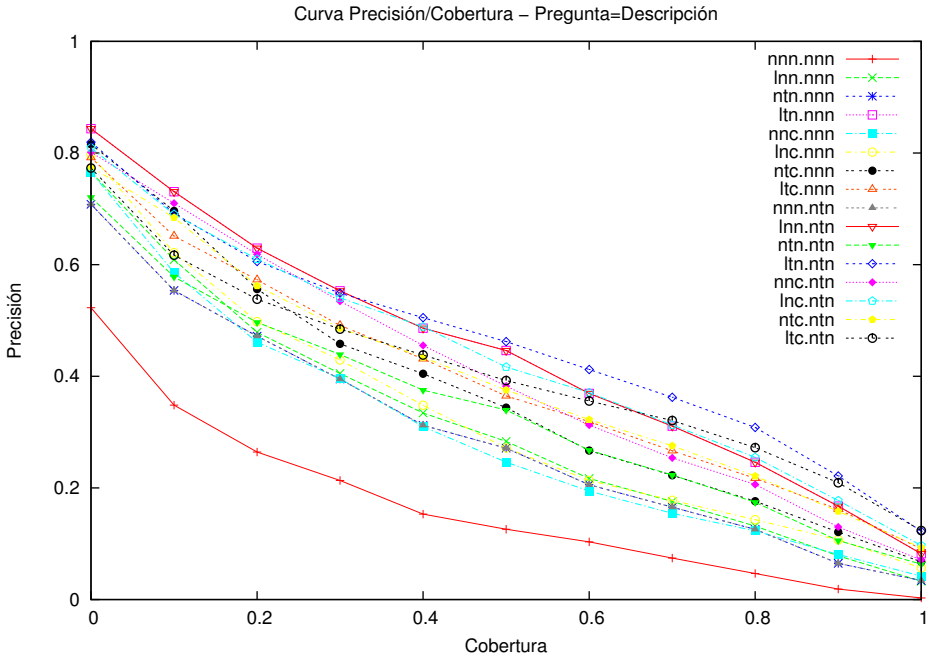


Figura 7.3: Curva precisión/cobertura para preguntas de longitud media.

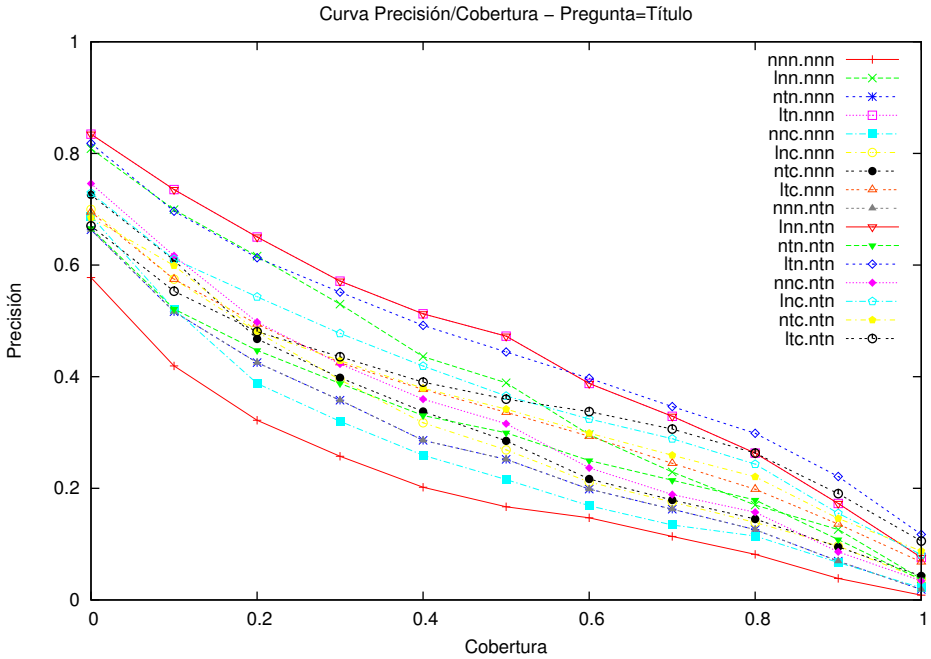


Figura 7.4: Curva precisión/cobertura para preguntas cortas.

7.3.1. Grabación de las preguntas

Para medir el rendimiento de sistemas de recuperación de información dirigida por habla es preciso disponer de una serie de preguntas habladas. Para ello, se ha grabado a varios locutores leyendo las preguntas de la colección CLEF. En concreto, se han grabado 10 locutores diferentes leyendo las preguntas (5 masculinos y 5 femeninos). Se emplearon unos auriculares con micrófono incorporado, modelo *Plantronics DSP-400*, en condiciones de oficina. Se usó una resolución de 16 bits y una frecuencia de muestreo de 16 kHz. El software utilizado para realizar la grabación fue *Praat*[3].

Partiendo de los temas de la colección de pruebas, se grabaron dos tipos de preguntas:

- Preguntas de longitud media, construidas empleando el campo descripción de cada tema. Este tipo de consultas se corresponde con frases correctamente enunciadas en lenguaje natural, con una longitud media de 16 palabras. En la colección de prueba hay tanto frases enunciativas como interrogativas.

- Preguntas cortas, construidas empleando el campo título de cada tema. Se trata de consultas cortas, con una longitud media de 4,2 palabras, y que en la mayoría de los casos se corresponde con un conjunto de palabras clave.

La lista de preguntas empleadas se encuentra en la sección C.1 del apéndice C.

7.3.2. Evaluación de resultados

Para evaluar el rendimiento del sistema se ha utilizado la misma metodología de CLEF [17]: para cada pregunta se recuperan los 1000 documentos más relevantes, ordenados por relevancia, y se calcula la precisión media promediada o MAP utilizando los juicios de relevancia.

En nuestro caso, nos interesa también medir el rendimiento del reconocedor del habla. Para ello, se calcula la tasa de error de palabra o WER y la tasa de palabras fuera del vocabulario o tasa de palabras OOV.

La forma de calcular la precisión media promediada se describe en la sección 3.2.4. Para realizar estos cálculos se ha empleado la herramienta *trec_eval* [4]. La manera de calcular la tasa de error de palabra del reconocedor se describe en la sección 2.2.1.6. Para calcular dicha tasa se ha utilizado la herramienta *NIST Scoring Toolkit (SCTK)* [5]. La forma de calcular la tasa de palabras fuera del vocabulario se describe en la sección 2.2.1.6.

[3] `http://www.fon.hum.uva.nl/praat/`
[4] `http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz`
[5] `http://www.nist.gov/speech/tools/`

7.4. Experimentos iniciales

En primer lugar se han realizado una serie de experimentos para evaluar el rendimiento del sistema con la configuración básica. Posteriormente, en base al análisis de los resultados obtenidos, se proponen ciertas modificaciones sobre la configuración inicial con el objetivo de mejorar el rendimiento del sistema. De esta manera, el resultado de los experimentos iniciales sirve de referencia y permite comprobar la mejora real introducida por cada una de las mejoras propuestas. Los experimentos iniciales se han presentado en [56].

7.4.1. Sistema de referencia

El funcionamiento del sistema de referencia se muestra en la figura 7.1. En primer lugar, la consulta hablada del usuario es transcrita por el reconocedor de habla. A continuación, la mejor hipótesis es utilizada por el motor de IR para obtener la lista de documentos relevantes para dicha pregunta.

Para medir el rendimiento del sistema se emplean las consultas habladas de la colección de prueba CLEF 2001, tal y como se ha descrito en el apartado 7.3. En primer lugar, para estudiar cómo afecta el tamaño del vocabulario al resultado final, se han realizado diferentes experimentos con un vocabulario de 20.000, 40.000 y 60.000 palabras. En segundo lugar, para estudiar cómo afecta la longitud de la consulta del usuario, se han empleado consultas de longitud media y consultas cortas. Las primeras se corresponden con el campo descripción de cada tema de la colección de pruebas CLEF y las segundas con el campo título de cada tema.

Los resultados obtenidos se muestran en las tablas 7.4 y 7.5. En cada tabla se muestra la tasa de palabras fuera de vocabulario (OOV), la tasa de error de palabra (WER) y la precisión media promediada (MAP). Se incluyen también los resultados obtenidos con las preguntas textuales originales con el fin de poder compararlos. Para cada experimento se muestra la media para los 10 locutores. Los resultados obtenidos para cada locutor se encuentran en la sección C.2 del apéndice C.

Es importante destacar que el empleo de vocabularios de tamaño más grande produce un incremento del tiempo computacional del reconocedor del habla. En una máquina *AMD Opteron Dual Core* a 1,8 GHz, los factores de tiempo real (ver sección 2.2.1.6) obtenidos fueron: 0,76 para un vocabulario de 20.000 palabras, 1,00 para un vocabulario de 40.000 palabras y 1,16 para un vocabulario de 60.000 palabras.

7.4.2. Análisis de errores

Estudiando el comportamiento del sistema para distintos tamaños de vocabulario se comprueba que los peores resultados se obtienen para un vocabulario de 20.000 palabras. Al aumentar el tamaño del vocabulario los resultados mejoran. Esto es así debido a que la tarea de recuperación de información dirigida por habla

	OOV	WER	MAP
Texto	–	–	0,4497
20k	7,15 %	24,3 %	0,3029
40k	2,81 %	18,8 %	0,3390
60k	2,17 %	18,3 %	0,3540

Tabla 7.4: Resultados del sistema de referencia para preguntas de longitud media. Se muestra la media sobre los 10 locutores de los resultados obtenidos empleando vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra el resultado obtenido usando las preguntas textuales originales. (OOV: tasa de palabras fuera del vocabulario; WER: tasa de error de palabra; MAP: precisión media promediada).

	OOV	WER	MAP
Texto	–	–	0,4419
20k	8,70 %	25,2 %	0,2997
40k	2,42 %	16,1 %	0,3357
60k	1,45 %	15,1 %	0,3441

Tabla 7.5: Resultados del sistema de referencia para preguntas cortas. Se muestra la media sobre los 10 locutores de los resultados obtenidos empleando vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra el resultado obtenido usando las preguntas textuales originales. (OOV: tasa de palabras fuera del vocabulario; WER: tasa de error de palabra; MAP: precisión media promediada).

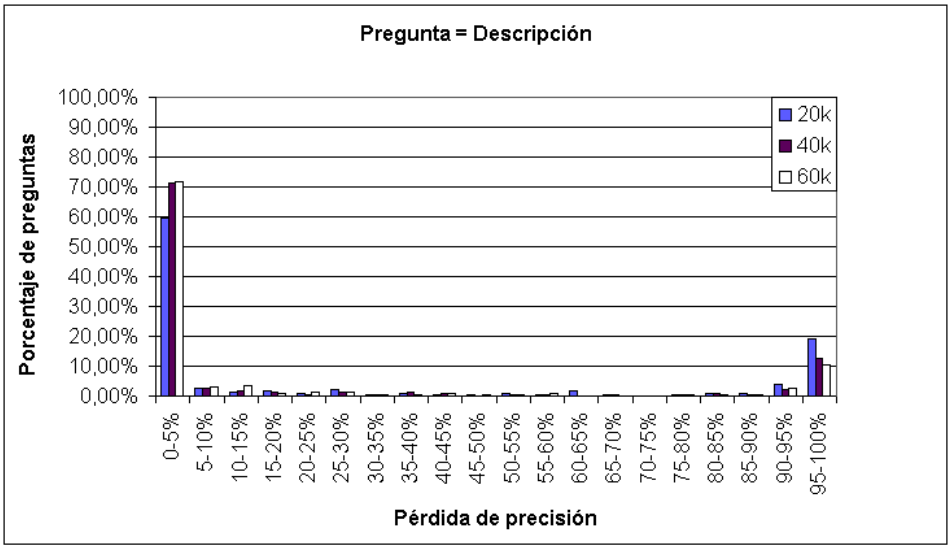


Figura 7.5: Porcentaje de preguntas en función de la pérdida de precisión, para preguntas de longitud media y para vocabularios de 20.000, 40.000 y 60.000 palabras.

tiene un vocabulario abierto, y por tanto, se obtendrán mejores resultados cuanto mayor sea el vocabulario.

Analizando los resultados para consultas de distinta longitud se observa que se obtienen mejores resultados de precisión empleando preguntas de longitud media frente a preguntas cortas, tanto en texto como en habla. Sin embargo, hay una pérdida de precisión similar de las consultas habladas frente a las consultas textuales, tanto empleando preguntas de longitud media como preguntas cortas.

Si se analizan los resultados obtenidos para cada pregunta, se comprueba que la pérdida de precisión de las preguntas habladas frente a las preguntas textuales no se distribuye de manera uniforme entre las preguntas. En la figuras 7.5 y 7.6 se muestra el porcentaje de preguntas en función de la pérdida de precisión, usando el campo descripción y el campo título para construir las preguntas respectivamente. En ambos casos se observa que la mayoría de las preguntas tiene una pérdida de precisión pequeña (menor del 5 %), mientras que unas pocas preguntas tienen una gran pérdida de precisión (mayor del 90 %). Esto quiere decir que en general las preguntas funcionaron bien, pero unas pocas funcionaron muy mal.

A continuación se han investigado cuáles son los factores que provocan la degradación de los resultados al emplear consultas habladas. Para ello, se han estudiado los resultados de cada pregunta, comparando el MAP de las preguntas habladas con el MAP de las preguntas textuales. Se han considerado erróneas aquellas preguntas con más de un 25 % de pérdida relativa de precisión. Se han estudiado los motivos que provocan esa degradación en el rendimiento y se han identificado tres tipos de errores:

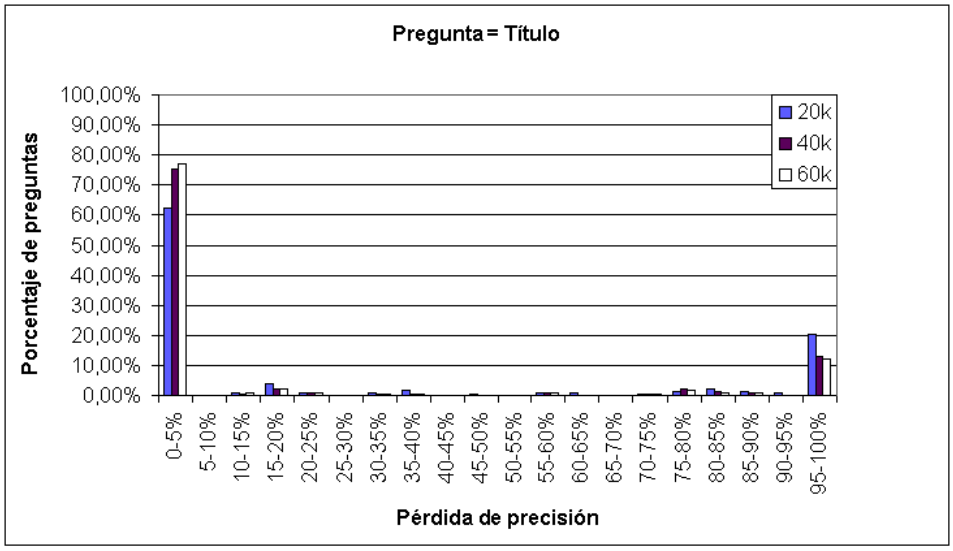


Figura 7.6: Porcentaje de preguntas en función de la pérdida de precisión, para preguntas cortas y para vocabularios de 20.000, 40.000 y 60.000 palabras.

- **Tipo I**: error causado por una palabra fuera de vocabulario.
- **Tipo II**: error provocado por una palabra en otro idioma.
- **Tipo III**: cualquier otro error de reconocimiento del habla.

Se han clasificado las preguntas erróneas en función del tipo de error. El número y porcentaje para cada tipo de error se muestra en las tablas 7.6 y 7.7, usando la descripción y el título para construir las preguntas, respectivamente. El número total de preguntas es de 490 (49 preguntas y 10 locutores). Para un vocabulario de 20.000 palabras la mayoría de los errores fueron causados por palabras fuera del vocabulario. Para vocabularios de 40.000 y 60.000 palabras, los errores debidos a palabras fuera del vocabulario se reducen y los errores tipo II y tipo III se incrementan, debido a una mejor cobertura del vocabulario.

Finalmente, se ha observado que cada tipo de error tiene un impacto distinto en la precisión. Por ejemplo, en el caso de preguntas formadas por el campo descripción y con un vocabulario de 60.000 palabras, los errores tipo I son catastróficos (pérdida de precisión del 99,7 % de media), los errores tipo II tiene un grave impacto (pérdida de precisión del 85,0 % de media), y los errores tipo III son los que producen una menor degradación de la precisión (pérdida de precisión del 72,6 % de media).

	Tipo I	Tipo II	Tipo III	Total
20k	115 (68,86 %)	25 (14,97 %)	27 (16,17 %)	167
40k	39 (35,14 %)	30 (27,03 %)	42 (37,84 %)	111
60k	20 (20,41 %)	28 (28,57 %)	50 (51,02 %)	98

Tabla 7.6: Distribución de errores del sistema de referencia para preguntas de longitud media. Se muestra el número y porcentaje de cada tipo de error para vocabularios de 20.000, 40.000 y 60.000 palabras. (Tipo I: causado por una palabra OOV; Tipo II: producido por una palabra en otro idioma; Tipo III: cualquier otro error de reconocimiento del habla).

	Tipo I	Tipo II	Tipo III	Total
20k	115 (73,72 %)	12 (7,69 %)	29 (18,59 %)	156
40k	40 (38,83 %)	19 (18,45 %)	44 (42,72 %)	103
60k	20 (21,51 %)	19 (20,43 %)	54 (58,06 %)	93

Tabla 7.7: Distribución de errores del sistema de referencia para preguntas de cortas. Se muestra el número y porcentaje de cada tipo de error para vocabularios de 20.000, 40.000 y 60.000 palabras. (Tipo I: causado por una palabra OOV; Tipo II: producido por una palabra en otro idioma; Tipo III: cualquier otro error de reconocimiento del habla).

7.5. Mejoras sobre el sistema de referencia

En vista de los resultados obtenidos en el apartado anterior, se han planteado varias mejoras sobre el sistema de referencia:

- Adaptación del vocabulario y del modelo de lenguaje, con el objetivo de reducir los errores provocados por palabras fuera del vocabulario.
- Empleo de la técnica de realimentación por pseudo-relevancia, para mejorar los resultados del motor de recuperación de información.
- Introducción de pronunciaciones alternativas en el vocabulario del reconocedor de habla, para reducir los errores provocados por palabras en otro idioma.

En el resto de experimentos de este capítulo únicamente se han utilizado preguntas de longitud media. El motivo es que en el sistema de referencia se ha comprobado que el comportamiento es similar para preguntas de longitud media y preguntas cortas. Por tanto, es suficiente realizar los experimentos con un tipo de preguntas. Se han elegido las preguntas de longitud media porque son frases en lenguaje natural que reflejan mejor las preguntas que realizan los usuario al sistema. Por contra, las preguntas cortas suelen ser una serie de términos clave, en lugar de frases en lenguaje natural.

Parte de los experimentos realizados han sido presentados en [57, 58].

7.5.1. Adaptación del vocabulario y del modelo de lenguaje

La propuesta consiste en adaptar dinámicamente el vocabulario y el modelo de lenguaje a la pregunta realizada por el usuario. Se emplea una estrategia basada en dos pasos, como se muestra en la figura 7.7. En el primer paso, el reconocedor del habla transcribe a texto la pregunta hablada del usuario, empleando el vocabulario y el LM general. A continuación, el motor de recuperación de información recupera los 1000 documentos más relevantes para la pregunta. Entonces, se realiza la adaptación dinámica del vocabulario y del LM usando esos documentos. En el segundo paso, el reconocedor del habla usa el vocabulario y el LM adaptados en lugar del vocabulario y el LM general. Por último, se obtiene la lista de documentos relevantes a la pregunta y se presenta al usuario.

El objetivo perseguido con la adaptación del vocabulario es reducir la tasa de palabras fuera del vocabulario. Para ello, se modifica el vocabulario inicial para incorporar las palabras de los documentos devueltos por el motor de IR en el primer paso. Las palabras incorporadas tienen una relación con la pregunta realizada y, por tanto, es de esperar que el vocabulario se ajuste mejor al dominio de la pregunta.

Para realizar la adaptación del vocabulario se crea una lista con todas las palabras que aparecen en los documentos recuperados en el primer paso. Entonces, se añaden las palabras más frecuentes del vocabulario general hasta que se alcanza el tamaño de vocabulario deseado. En el vocabulario adaptado, el número de palabras que proviene de cada fuente depende de la pregunta. Los valores que se han obtenido en los experimentos realizados, haciendo la media para todas las preguntas y según el tamaño del vocabulario fueron:

- Para un tamaño de vocabulario de 20.000 palabras, todas las palabras provenían de los documentos.
- Para vocabularios de 40.000 y 60.000 palabras, alrededor de 27.000 palabras procedían de los documentos y el resto del vocabulario general.

El objetivo de la adaptación del modelo de lenguaje es conseguir un LM que modele mejor la pregunta realizada por el usuario. Esto se consigue debido a que la adaptación se realiza utilizando los documentos devueltos por el motor de IR en el primer paso, y que por tanto, son del mismo dominio que la pregunta.

Para realizar la adaptación del modelo de lenguaje, primero entrenamos un nuevo LM con los documentos obtenidos en el primer paso. A continuación se interpola el nuevo LM con el LM general. Se ha utilizado interpolación lineal con un coeficiente de 0,5. El método de interpolación lineal se describe en el apartado 2.2.1.7.

Para estudiar la contribución individual de cada adaptación se han realizado tres experimentos distintos: realizando únicamente adaptación del modelo de lenguaje; realizando únicamente adaptación del vocabulario; realizando adaptación del modelo de lenguaje y del vocabulario.

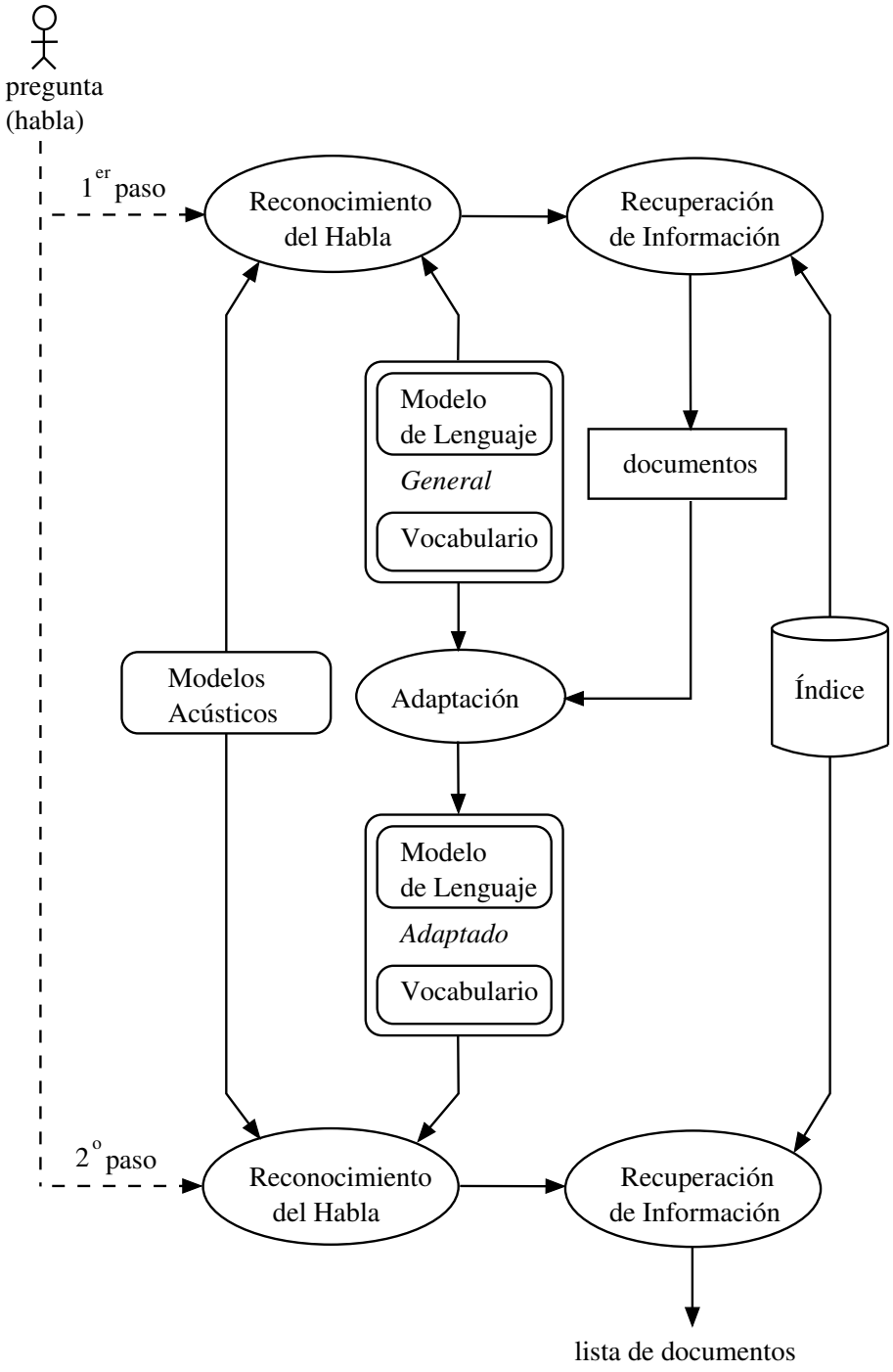


Figura 7.7: Arquitectura del sistema, basada en una estrategia de dos pasos.

		OOV	WER	MAP
20k	Referencia	7,15 %	24,3 %	0,3029
	Ad. LM	7,15 %	24,1 %	0,3060
	Ad. Voc.	3,38 %	18,6 %	0,3467
	Ad. LM y Voc.	3,38 %	19,1 %	0,3563
40k	Referencia	2,81 %	18,8 %	0,3390
	Ad. LM	2,81 %	18,1 %	0,3411
	Ad. Voc.	1,88 %	17,7 %	0,3567
	Ad. LM y Voc.	1,88 %	17,1 %	0,3643
60k	Referencia	2,17 %	18,3 %	0,3540
	Ad. LM	2,17 %	17,3 %	0,3579
	Ad. Voc.	1,62 %	17,9 %	0,3649
	Ad. LM y Voc.	1,62 %	16,7 %	0,3721

Tabla 7.8: Resultados obtenidos al realizar adaptación del LM, adaptación de vocabulario y adaptación del LM y vocabulario, para preguntas de longitud media. Se muestra la media sobre los 10 locutores de los resultados obtenidos empleando vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra el resultado del sistema de referencia. (OOV: tasa de palabras fuera del vocabulario; WER: tasa de error de palabra; MAP: precisión media promediada).

Los resultados obtenidos se muestran en la tabla 7.8. Si se comparan los resultados con el sistema de referencia se observa una reducción en la tasa de palabras OOV, una reducción en WER y un incremento en MAP. Los modelos adaptados proporcionan mejores resultados porque los documentos usados para la adaptación tenían una relación semántica con la pregunta. Cuando se adapta únicamente el modelo de lenguaje, la mejora es muy pequeña. Sin embargo, cuando se adapta únicamente el vocabulario, se consigue una mejora significativa. Los mejores resultados se obtienen adaptando el vocabulario y el modelo de lenguaje.

En la tabla 7.9 se muestran los tipos de errores que se han producido. Cuando únicamente se adapta el modelo de lenguaje, la mejora es muy pequeña. Sin embargo, la adaptación de vocabulario produce una reducción significativa de los errores tipo I, debido a una mejor cobertura del vocabulario. Cuando se adapta el vocabulario y el modelo de lenguaje, además de la reducción en errores tipo I, se consigue también una reducción de errores tipo III, debido a un mejor modelado del lenguaje. Los resultados obtenidos para cada locutor se encuentran en la sección C.2 del apéndice C.

7.5.2. Realimentación por pseudo-relevancia

La realimentación por pseudo-relevancia permite mejorar los resultados obtenidos por el motor de recuperación de información. De entre los diversos métodos existentes, se ha utilizado el método de Rocchio [113]. En este apartado se de-

		Tipo I	Tipo II	Tipo III	Total
20k	Referencia	115 (68,86 %)	25 (14,97 %)	27 (16,17 %)	167
	Ad. LM	115 (69,70 %)	24 (14,55 %)	26 (15,76 %)	165
	Ad. Voc.	34 (33,66 %)	33 (32,67 %)	34 (33,66 %)	101
	Ad. LM y Voc.	34 (35,79 %)	32 (33,68 %)	29 (30,53 %)	95
40k	Referencia	39 (35,14 %)	30 (27,03 %)	42 (37,84 %)	111
	Ad. LM	40 (37,38 %)	30 (28,04 %)	37 (34,58 %)	107
	Ad. Voc.	20 (20,41 %)	32 (32,65 %)	46 (46,94 %)	98
	Ad. LM y Voc.	20 (22,22 %)	33 (36,67 %)	37 (41,11 %)	90
60k	Referencia	20 (20,41 %)	28 (28,57 %)	50 (51,02 %)	98
	Ad. LM	20 (21,28 %)	28 (29,79 %)	46 (48,94 %)	94
	Ad. Voc.	15 (16,13 %)	30 (32,26 %)	48 (51,61 %)	93
	Ad. LM y Voc.	15 (17,05 %)	31 (35,23 %)	42 (47,73 %)	88

Tabla 7.9: Distribución de errores al realizar adaptación del LM, adaptación de vocabulario y adaptación del LM y vocabulario, para preguntas de longitud media. Se muestra el número y porcentaje de cada tipo de error para vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra la distribución de errores del sistema de referencia. (Tipo I: causado por una palabra OOV; Tipo II: producido por una palabra en otro idioma; Tipo III: cualquier otro error de reconocimiento del habla).

terminan los parámetros óptimos para la colección de pruebas CLEF, empleando preguntas de texto. Una vez fijados los parámetros, se estudia si la mejora de resultados también se produce cuando la consulta es hablada en lugar de textual.

El método de Rocchio consiste en utilizar la pregunta original para recuperar una lista preliminar de documentos. De entre esos documentos, se determina cuáles son relevantes y cuáles no lo son. En base a ello se construye una pregunta mejorada añadiendo una serie de términos nuevos y recalculando el peso de todos los términos. Por último, la pregunta mejorada es usada para obtener la lista final de documentos. Una descripción detallada del método se encuentra en la sección 3.2.3.

Para emplear el método de realimentación por pseudo-relevancia de Rocchio es preciso determinar el valor de una serie de parámetros:

- El número de documentos relevantes, n_1. De entre la lista preliminar de documentos recuperados, se considera que los n_1 primeros documentos son relevantes.

- El número de términos que se añaden de cada documento relevante. En lugar de añadir todos los términos de los documentos considerados relevantes, únicamente se añaden un determinado número de ellos, que son aquellos con mayor peso.

- Los parámetros α, β y γ de la fórmula de pesado 3.5 (ver sección 3.2.3). Dado que los tres están relacionados, fijamos $\alpha = 1$ y ajustaremos β y γ.

En cuanto a los documentos no relevantes, consideramos que los documentos recuperados entre la posición 501 y 1000 son no relevantes.

Para determinar los parámetros óptimos se han realizado varios experimentos empleando preguntas textuales y distintos valores de los parámetros, de forma similar a otros trabajos llevados a cabo con la colección de pruebas CLEF [117] y con la colección de pruebas TREC [22, 97, 123].

En la tabla 7.10 se muestran los resultados de los experimentos realizados para determinar el valor óptimo del número de términos, del número de documentos relevantes (n_1) y de β. Una vez fijados estos tres parámetros, se han realizado varios experimentos para determinar el mejor valor de γ, según se muestra en la tabla 7.11. Los parámetros óptimos obtenidos son: $n_1 = 2$; número de términos $= 10$; $\alpha = 1$; $\beta = 0,15$; y $\gamma = 0,15$.

Una vez fijados los parámetros, se ha evaluado el comportamiento del sistema empleando realimentación por pseudo-relevancia con preguntas habladas. Los resultados obtenidos pueden verse en la tabla 7.12. También se han identificado los tipos de errores que se producen, que se muestran en la tabla 7.13. Los resultados muestran que emplear realimentación por pseudo-relevancia mejora el resultado del sistema tanto para preguntas textuales como para preguntas habladas. Los resultados obtenidos para cada locutor se encuentran en la sección C.2 del apéndice C.

7.5.3. Modelado de palabras en otro idioma

Al analizar los resultados del sistema de referencia se han detectado errores causados por palabras en otro idioma. El problema surge debido a que el reconocedor del habla no está preparado para reconocer estas palabras, puesto que funciona exclusivamente para castellano. En general, las palabras extranjeras más frecuentes en castellano son las inglesas. Por ello, en este apartado se presenta una propuesta para extender el reconocedor del habla de modo que sea capaz de comprender a locutores castellanos diciendo palabras inglesas.

Hay una limitación importante en los datos de entrenamiento, porque no disponemos de un corpus de locutores castellanos diciendo palabras inglesas. Por este motivo se ha utilizado un enfoque similar al que usan los hablantes castellanos: establecer una correspondencia entre los fonemas ingleses y los fonemas castellanos. De esta manera, se pueden usar los modelos acústicos castellanos. Se ha desarrollado la correspondencia manualmente, en base a los sonidos similares de ambos idiomas (empleando el Alfabeto Fonético Internacional como referencia). La correspondencia puede verse en la tabla 7.14.

Una vez establecida la correspondencia, se ha añadido al diccionario de pronunciación la pronunciación de las palabras inglesas. Es útil mantener la pronunciación en castellano porque algunos hablantes pronuncian las palabras inglesas

#términos = 10					$\alpha = 1,\ \gamma = 0$
	$n_1 = 1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 4$	$n_1 = 5$
$\beta = 0,05$	0,4757	0,4738	0,4681	0,4658	0,4598
$\beta = 0,15$	0,4788	**0,4844**	0,4676	0,4670	0,4643
$\beta = 0,25$	0,4662	0,4808	0,4640	0,4653	0,4620
$\beta = 0,35$	0,4570	0,4730	0,4596	0,4623	0,4570
$\beta = 0,45$	0,4488	0,4573	0,4439	0,4594	0,4446

#términos = 20					$\alpha = 1,\ \gamma = 0$
	$n_1 = 1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 4$	$n_1 = 5$
$\beta = 0,05$	**0,4806**	0,4756	0,4684	0,4671	0,4593
$\beta = 0,15$	0,4803	0,4800	0,4651	0,4670	0,4643
$\beta = 0,25$	0,4672	0,4751	0,4577	0,4635	0,4612
$\beta = 0,35$	0,4551	0,4693	0,4375	0,4488	0,4445
$\beta = 0,45$	0,4467	0,4507	0,4316	0,4447	0,4405

#términos = 30					$\alpha = 1,\ \gamma = 0$
	$n_1 = 1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 4$	$n_1 = 5$
$\beta = 0,05$	**0,4811**	0,4763	0,4699	0,4656	0,4621
$\beta = 0,15$	0,4802	0,4795	0,4655	0,4686	0,4651
$\beta = 0,25$	0,4637	0,4728	0,4465	0,4639	0,4622
$\beta = 0,35$	0,4525	0,4670	0,4383	0,4470	0,4434
$\beta = 0,45$	0,4435	0,4511	0,4324	0,4434	0,4400

#términos = 40					$\alpha = 1,\ \gamma = 0$
	$n_1 = 1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 4$	$n_1 = 5$
$\beta = 0,05$	0,4822	0,4780	0,4709	0,4670	0,4617
$\beta = 0,15$	0,4768	**0,4828**	0,4657	0,4691	0,4649
$\beta = 0,25$	0,4595	0,4741	0,4479	0,4631	0,4612
$\beta = 0,35$	0,4489	0,4689	0,4376	0,4464	0,4427
$\beta = 0,45$	0,4401	0,4532	0,4313	0,4420	0,4389

Tabla 7.10: Experimentos para determinar el valor óptimo del número de términos, del número de documentos relevantes (n_1) y de β.

#términos = 10, $n_1 = 2$, $\alpha = 1$, $\beta = 0,15$				
$\gamma = 0,05$	$\gamma = 0,15$	$\gamma = 0,25$	$\gamma = 0,35$	$\gamma = 0,45$
0,4833	**0,4849**	0,4829	0,4795	0,4745

Tabla 7.11: Experimentos para determinar el valor óptimo de γ.

		OOV	WER	MAP
20k	Referencia	7,15 %	24,3 %	0,3029
	Con realimenta.	7,15 %	24,3 %	0,3384
40k	Referencia	2,81 %	18,8 %	0,3390
	Con realimenta.	2,81 %	18,8 %	0,3753
60k	Referencia	2,17 %	18,3 %	0,3540
	Con realimenta.	2,17 %	18,3 %	0,3892

Tabla 7.12: Resultados obtenidos al emplear realimentación por pseudo-relevancia, para preguntas de longitud media. Se muestra la media sobre los 10 locutores de los resultados obtenidos empleando vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra el resultado del sistema de referencia. (OOV: tasa de palabras fuera del vocabulario; WER: tasa de error de palabra; MAP: precisión media promediada).

		Tipo I	Tipo II	Tipo III	Total
20k	Referencia	115 (68,86 %)	25 (14,97 %)	27 (16,17 %)	167
	Con realimenta.	118 (69,41 %)	22 (12,94 %)	30 (17,65 %)	170
40k	Referencia	39 (35,14 %)	30 (27,03 %)	42 (37,84 %)	111
	Con realimenta.	39 (36,45 %)	24 (22,43 %)	44 (41,12 %)	107
60k	Referencia	20 (20,41 %)	28 (28,57 %)	50 (51,02 %)	98
	Con realimenta.	20 (21,05 %)	23 (24,21 %)	52 (54,74 %)	95

Tabla 7.13: Distribución de errores al emplear realimentación por pseudo-relevancia, para preguntas de longitud media. Se muestra el número y porcentaje de cada tipo de error para vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra la distribución de errores del sistema de referencia. (Tipo I: causado por una palabra OOV; Tipo II: producido por una palabra en otro idioma; Tipo III: cualquier otro error de reconocimiento del habla).

Inglés	Castellano	Inglés	Castellano	Inglés	Castellano
AA	a	F	f	P	p
AE	a	G	g	R	r
AH	a	HH	x	S	s
AO	o	IH	i	SH	tS
AW	a u	IY	i	T	t
AY	a i	JH	jj	TH	T
B	b	K	k	UH	u
CH	tS	L	l	UW	u
D	d	M	m	V	b
DH	d	N	n	W	u
EH	e	NG	n	Y	i
ER	e r	OW	o u	Z	s
EY	e i	OY	o i	ZH	s

Tabla 7.14: Correspondencia entre los fonemas ingleses y los fonemas castellanos. Los fonemas ingleses se representan mediante el alfabeto fonético Arpabet (que es el utilizado por el diccionario de pronunciación de CMU) y los fonemas castellanos mediante el alfabeto fonético SAMPA.

como si fueran palabras castellanas. Se ha utilizado el diccionario de pronunciación de CMU[6] para identificar las palabras inglesas y para obtener su pronunciación. Para cada palabra del vocabulario inicial que aparece en el diccionario de CMU, se ha añadido su pronunciación inglesa como alternativa.

El número de pronunciaciones de palabras inglesas que se han añadido ha sido: 3.272 pronunciaciones para el vocabulario de 20.000 palabras; 6.042 pronunciaciones para el vocabulario de 40.000 palabras; y 8.456 pronunciaciones para el vocabulario de 60.000 palabras.

Se ha evaluado el rendimiento del sistema empleando el diccionario de pronunciación extendido con pronunciaciones alternativas. Se observa una reducción significativa de los errores tipo II. Sin embargo, nuevos errores tipo I y tipo III han aparecido (ver tabla 7.16). Por otro lado, la mejora en la precisión de la recuperación respecto al sistema de referencia es pequeña (ver tabla 7.15). Los resultados obtenidos para cada locutor se encuentran en la sección C.2 del apéndice C.

Estos resultados demuestran que es necesario realizar un tratamiento especial de las palabras en otro idioma, debido a que tienen una influencia clara en el rendimiento del sistema. Sin embargo, es necesario mejorar el enfoque propuesto. Por un lado, hay que incorporar palabras en otros idiomas además de inglesas. Por otro lado, se hace imprescindible disponer de un corpus de entrenamiento y pruebas adecuado, en el que haya locuciones de hablantes castellanos diciendo palabras en otros idiomas.

[6] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

		OOV	WER	MAP
20k	Referencia	7,15 %	24,3 %	0,3029
	Pron. inglés	7,15 %	24,5 %	0,3050
40k	Referencia	2,81 %	18,8 %	0,3390
	Pron. inglés	2,81 %	18,9 %	0,3416
60k	Referencia	2,17 %	18,3 %	0,3540
	Pron. inglés	2,17 %	18,2 %	0,3549

Tabla 7.15: Resultados obtenidos al añadir la pronunciación de palabras inglesas al diccionario de pronunciación, para preguntas de longitud media. Se muestra la media sobre los 10 locutores de los resultados obtenidos empleando vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra el resultado del sistema de referencia. (OOV: tasa de palabras fuera del vocabulario; WER: tasa de error de palabra; MAP: precisión media promediada).

		Tipo I	Tipo II	Tipo III	Total
20k	Referencia	115 (68,86 %)	25 (14,97 %)	27 (16,17 %)	167
	Pron. inglés	119 (72,56 %)	12 (7,32 %)	33 (20,12 %)	164
40k	Referencia	39 (35,14 %)	30 (27,03 %)	42 (37,84 %)	111
	Pron. inglés	40 (39,60 %)	14 (13,86 %)	47 (46,53 %)	101
60k	Referencia	20 (20,41 %)	28 (28,57 %)	50 (51,02 %)	98
	Pron. inglés	20 (21,28 %)	14 (14,89 %)	60 (63,83 %)	94

Tabla 7.16: Distribución de errores al añadir la pronunciación de palabras inglesas al diccionario de pronunciación, para preguntas de longitud media. Se muestra el número y porcentaje de cada tipo de error para vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra la distribución de errores del sistema de referencia. (Tipo I: causado por una palabra OOV; Tipo II: producido por una palabra en otro idioma; Tipo III: cualquier otro error de reconocimiento del habla).

		OOV	WER	MAP
20k	Referencia	7,15 %	24,3 %	0,3029
	Sistema final	3,52 %	19,6 %	0,3966
40k	Referencia	2,81 %	18,8 %	0,3390
	Sistema final	2,13 %	17,1 %	0,4066
60k	Referencia	2,17 %	18,3 %	0,3540
	Sistema final	1,63 %	16,5 %	0,4122

Tabla 7.17: Resultados del sistema final, para preguntas de longitud media. Se muestra la media sobre los 10 locutores de los resultados obtenidos empleando vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra el resultado del sistema de referencia. (OOV: tasa de palabras fuera del vocabulario; WER: tasa de error de palabra; MAP: precisión media promediada).

7.6. Experimentos finales

Una vez evaluadas las mejoras propuestas individualmente, se han realizado una serie de experimentos finales que incorporan todas las mejoras descritas anteriormente. Para ello, se ha combinado la estrategia de dos pasos de adaptación de vocabulario y LM con el modelado de palabras extranjeras, usando el motor de IR con realimentación por pseudo-relevancia.

En primer lugar, se han obtenido los 1000 documentos más relevantes para la pregunta, empleando el diccionario de pronunciación que contenía la pronunciación de palabras en inglés y el motor de IR con realimentación por pseudo-relevancia. Entonces, se ha adaptado el vocabulario y el LM, empleando los documentos recuperados. A continuación, se ha expandido el diccionario de pronunciación para incluir la pronunciación de las palabras en inglés. Para cada pregunta se emplea un vocabulario distinto, por tanto, el número de pronunciaciones de palabras inglesas añadidas es variable. Por último, se ha recuperado la lista final de documentos relevantes, empleando el motor de IR con realimentación por pseudo-relevancia.

Los resultados muestran que hubo un aumento en la precisión de la recuperación, como se muestra en la tabla 7.17. También hubo una reducción en el número de preguntas erróneas, como se muestra en la tabla 7.18. Para finalizar, se ha comparado el sistema final con el sistema de referencia. Para el caso de vocabulario de 60.000 palabras, hubo un ganancia relativa en MAP de 16,4 %, una reducción relativa de 24,9 % en la tasa de palabras OOV, y una reducción relativa de 9,8 % en WER. Los resultados obtenidos para cada locutor se encuentran en la sección C.2 del apéndice C. También se incluyen dos tablas resumen de todos los experimentos realizados.

		Tipo I	Tipo II	Tipo III	Total
20k	Referencia	115 (68,86 %)	25 (14,97 %)	27 (16,17 %)	167
	Sistema final	33 (36,67 %)	18 (20,00 %)	39 (43,33 %)	90
40k	Referencia	39 (35,14 %)	30 (27,03 %)	42 (37,84 %)	111
	Sistema final	18 (23,68 %)	16 (21,05 %)	42 (55,26 %)	76
60k	Referencia	20 (20,41 %)	28 (28,57 %)	50 (51,02 %)	98
	Sistema final	13 (16,88 %)	16 (20,78 %)	48 (62,34 %)	77

Tabla 7.18: Distribución de errores del sistema final, para preguntas de longitud media. Se muestra el número y porcentaje de cada tipo de error para vocabularios de 20.000, 40.000 y 60.000 palabras. También se muestra la distribución de errores del sistema de referencia. (Tipo I: causado por una palabra OOV; Tipo II: producido por una palabra en otro idioma; Tipo III: cualquier otro error de reconocimiento del habla).

7.7. Comparación con otros sistemas

Resulta interesante comparar nuestros resultados con los obtenidos por otros sistemas similares documentados en la bibliografía. Sin embargo, realizar esta comparación no es una tarea sencilla. La mayor dificultad reside en que cada sistema funciona para un idioma diferente, y por tanto, no es posible usar el mismo conjunto de prueba, requisito imprescindible para una comparación justa. Además, existen también otros factores que afectan al rendimiento obtenido por los diferentes sistemas: longitud de las preguntas, tamaño del vocabulario del reconocedor, modelo de recuperación usado por el motor de IR, etcétera.

En la tabla 7.19 se presenta una comparación de los sistemas descritos en la sección 4.4. Para cada sistema, se calcula la pérdida de MAP usando preguntas habladas comparado con usar preguntas textuales. Aunque la comparación no es concluyente, podemos decir que nuestro sistema tiene un rendimiento equiparable al de esas otras contribuciones.

Cabe destacar también que existen algunas diferencias importantes entre nuestros experimentos y los experimentos descritos por otros investigadores. Barnett et al. [10] emplearon preguntas largas (50-60 palabras) para el experimento, que son más robustas frente a errores de reconocimiento del habla. Chang et al. [24] usaron un reconocedor del habla dependiente de género, con adaptación de locutor y de canal. Fujii et al. [46] y Matsushita et al. [90] emplearon una colección de documentos mucho mayor (100 Gb, sobre 10 millones de documentos) lo que hace la tarea más difícil.

	Colección	MAP-T	MAP-H	Pérdida
Barnett	TIPSTER	0,3465	0,3020	12,8 %
Chang	TREC-5	0,3580	0,2570	28,2 %
	TREC-6	0,4890	0,4630	5,3 %
Fujii	NTCIR-3	0,1257	0,0766	39,1 %
Matsushita	NTCIR-3	0,1181	0,0820	30,6 %
Nuestro sistema	CLEF01	0,4849	0,4122	15,0 %

Tabla 7.19: Resultados de los sistemas descritos en la bibliografía y de nuestro sistema. (MAP-T: precisión media promediada empleando preguntas textuales; MAP-H: precisión media promediada empleando preguntas habladas; Pérdida: pérdida relativa de MAP de las preguntas habladas comparadas con las preguntas de texto).

7.8. Conclusiones

La integración de las tecnologías de reconocimiento del habla y de recuperación de información no es sencilla. La razón es que cada tecnología emplea un enfoque distinto: el reconocimiento del habla favorece las palabras frecuentes, porque es más probable que sean dichas (mayor probabilidad en el modelo de lenguaje); mientras que la recuperación de información favorece las palabras infrecuentes, porque tienen mayor contenido semántico (empleando el esquema de pesado TF-IDF). El peor caso se produce con los nombres propios: son términos de búsqueda muy eficientes, pero debido a su baja frecuencia de aparición tienen baja probabilidad en el modelo de lenguaje, o incluso, no están incluidos en el vocabulario del reconocedor del habla.

En este capítulo se han realizado diversos experimentos encaminados a la construcción de un sistema de recuperación de información dirigida por habla. El objetivo del sistema es que el usuario pueda recuperar todos los documentos relevantes a una consulta hablada en lenguaje natural.

En primer lugar, se ha analizado el impacto del tamaño del vocabulario en el rendimiento del sistema. Los mejores resultados se obtuvieron empleando un vocabulario de 60.000 palabras, porque hubo menos errores de reconocimiento del habla. La recuperación de información dirigida por habla es una tarea con un vocabulario abierto, y se obtendrán mejores resultados usando modelos de lenguaje con vocabularios más grandes, debido a su mejor cobertura del vocabulario. Sin embargo, incrementar indefinidamente el tamaño del vocabulario no es posible. Primero, hay algunas limitaciones al estimar modelos de lenguaje con grandes vocabularios, debido a la inherente escasez de datos del corpus de entrenamiento. Segundo, hay restricciones computacionales, relacionadas con las limitaciones de tiempo y memoria.

En segundo lugar, se ha estudiado el impacto de la longitud de las preguntas. Las consultas de longitud media tuvieron mejores resultados en precisión que las

preguntas cortas, tanto en texto como en habla. Por este motivo, en el resto de experimentos de este capítulo se han utilizado las preguntas de longitud media. Además, este tipo de consultas son frases en lenguaje natural que reflejan mejor las preguntas que realizan los usuarios al sistema.

También se han analizado los resultados de cada pregunta individualmente y se ha comprobado que la pérdida de precisión de las preguntas habladas frente a las preguntas textuales no está uniformemente distribuida: la mayoría de las preguntas funcionaron bien (pequeña pérdida de precisión), mientras que unas pocas preguntas funcionaron muy mal (pérdida de precisión alta). Esto es debido a que cada error de reconocimiento del habla tuvo un gran impacto en la precisión de la recuperación. Por un lado, palabras clave con importante contenido semántico pueden perderse y algunos documentos relevantes no serán recuperados. Por otro lado, palabras no relacionadas con la pregunta pueden ser incorporadas, haciendo que el sistema recupere documentos que no están relacionados con la pregunta. Sorprendentemente, hubo algunas preguntas que mejoraron cuando ocurrieron errores de reconocimiento del habla.

Se han identificado las tres fuentes de errores principales: palabras fuera de vocabulario, palabras en otro idioma y otros errores de reconocimiento del habla. El primero ocurre porque palabras poco frecuentes no son incluidas en el vocabulario. El segundo sucede debido a que el reconocedor del habla no está preparado para reconocer palabras en otro idioma. En base a estos tipos de errores, se han realizado tres propuestas para mejorar los resultados del sistema de referencia.

Se ha presentado una propuesta basada en la adaptación dinámica del vocabulario y del modelo de lenguaje que consigue una reducción de la tasa de palabras OOV y del WER del reconocedor del habla. El vocabulario y el modelo de lenguaje adaptados proporcionan mejores resultados porque han sido entrenados empleando documentos con una relación semántica con la pregunta.

La utilización de la técnica de realimentación por pseudo-relevancia en el motor de IR contribuye a mejorar el rendimiento del sistema, tanto para preguntas textuales como para preguntas habladas.

El objetivo del modelado de palabras en otro idioma es conseguir que el reconocedor del habla sea capaz de comprender a locutores castellanos diciendo palabras en otros idiomas. En nuestro caso nos hemos centrado en las palabras inglesas, ya que son las que más aparecen en castellano. Se han utilizado los modelos acústicos castellanos y se ha establecido una correspondencia entre los fonemas ingleses y los fonemas castellanos. Se observa una reducción significativa de los errores provocados por palabras en otro idioma. Sin embargo, la disponibilidad de un corpus de entrenamiento adecuado permitiría mejorar el enfoque propuesto.

En los experimentos finales se ha comprobado que los resultados del sistema de referencia fueron mejorados al incorporar las propuestas realizadas. Empleando un vocabulario de 60.000 palabras, hubo una mejora relativa de 16,4 % en MAP. También hubo una reducción relativa de la tasa de palabras OOV de 24,9 % y una reducción relativa de 9,8 % en WER.

Como conclusión, los resultados obtenidos son esperanzadores y muestran la

viabilidad de construir sistemas de recuperación de información dirigida por habla. Aunque el rendimiento no es tan bueno como empleando texto como entrada, el sistema puede ser útil para superar las limitaciones de los dispositivos móviles, y en situaciones en las que el habla es la única modalidad disponible (por ejemplo cuando se conduce un coche).

Parte IV

Conclusiones

Capítulo 8

Conclusiones

8.1. Conclusiones

En esta memoria se han presentado diferentes estrategias para el acceso a contenidos web empleando habla. El trabajo se ha centrado en la reutilización de los contenidos web existentes y en plantear la interacción hablada de manera que el usuario pueda acceder a los contenidos de manera rápida y amigable. Para ello, ha sido necesario adaptar los contenidos web a las características de la nueva modalidad. El trabajo realizado se ha dividido en tres partes, que se describen a continuación.

En primer lugar se ha propuesto la conversión de contenidos web de manera general, con el objetivo de permitir el acceso a cualquier página web. Se han propuesto dos maneras de realizar la conversión, una automática y otra semiautomática. En el caso de la conversión automática, es preciso inferir la estructura inherente de cada página web, identificando los elementos que la componen. Una vez disponemos de la información estructurada, se plantea la interacción hablada con el usuario. Se ha construido un sistema que lleva a cabo la conversión automática y se ha realizado un caso de estudio. Sin embargo, el enfoque de conversión automática de contenidos web presenta ciertas limitaciones. Por un lado, la gran variabilidad de páginas web hace que sea complicado construir un sistema para extraer y estructurar la información que funcione para cualquier página web. Por otro lado, la interacción resultante resulta poco amigable, debido a que las páginas web contienen mucha información, y es difícil automatizar el proceso de selección y procesamiento necesario para que pueda ser accedida a través de una interfaz hablada. Para superar estas limitaciones, se ha propuesto un enfoque de conversión semiautomática, donde se utiliza una aplicación vocal para indicar al sistema cómo debe llevarse a cabo la conversión. Esta forma de solución requiere que un desarrollador de aplicaciones se encargue de construir una aplicación vocal para cada sitio web que se desea acceder. Se ha construido el sistema, que incluye una herramienta de desarrollo para facilitar la construcción de aplicaciones de conversión. Además, se han identificado los patrones típicos que más se utilizan al diseñar las páginas

web, proponiendo para cada uno de ellos la forma más adecuada para acceder a la información empleando el canal habado. La principal limitación de este enfoque semiautomático es que hay que construir una aplicación vocal para cada sitio web al que se quiere acceder. Por último, cabe resaltar que tanto en la conversión automática como en la semiautomática, la forma de acceso está limitada por la forma en la que están estructurados los contenidos web originales.

En segundo lugar se ha propuesto una forma de solución basada en el empleo de un sistema de diálogo hablado para el acceso a un sitio web concreto. La propuesta está basada en el empleo de dos modelos, un modelo de información y un modelo de interacción. El objetivo es poder procesar la información de manera global, a través de un modelo de información que permita procesar los contenidos para adaptarlos a la nueva modalidad, y de esta manera conseguir una interacción más amigable con el usuario. Desde el punto de vista de la interacción se plantean dos estrategias, navegación y búsqueda. La primera permite al usuario ver que información está disponible, mientras que la segunda permite encontrar información específica. Para validar nuestra propuesta hemos elegido el dominio de noticias, centrándonos en el sitio web de un periódico digital. Una vez construido el sistema se ha llevado a cabo una evaluación del mismo para medir su usabilidad y obtener realimentación de los usuarios finales. El sistema ha sido evaluado por 22 usuarios, que resolvieron 5 escenarios cada uno. Para medir el rendimiento del sistema se han utilizado medidas objetivas, y para medir la satisfacción de los usuarios se ha utilizado un cuestionario que los usuarios completaban al finalizar el proceso de evaluación. Los resultados mostraron una alta tasa de éxito de la tarea, lo que permite a los usuarios obtener la información deseada. Sin embargo, los errores de reconocimiento del habla imponen ciertas limitaciones, principalmente al emplear la estrategia de búsqueda. El estudio de satisfacción de usuario mostró que a los usuarios les gusta el sistema y piensan que es útil. Sin embargo, también piensan que es un poco aburrido y repetitivo.

En tercer lugar se han realizado diversos experimentos encaminados a la construcción de un sistema de recuperación de información dirigida por habla. El objetivo del sistema es permitir a los usuarios realizar búsquedas empleando lenguaje natural hablado. Se han realizado una serie de experimentos iniciales y se ha utilizado una colección de prueba estándar para medir el rendimiento. Una vez fijado el entorno de experimentación se han analizado los factores que más inciden en la degradación del sistema. Los experimentos mostraron que el tamaño del vocabulario del reconocedor del habla tiene gran impacto en el rendimiento del sistema, debido a que la recuperación de información dirigida por habla es una tarea con un vocabulario abierto. Posteriormente, se han propuesto varias mejoras que permiten incrementar el rendimiento del sistema. En primer lugar se ha propuesto la adaptación dinámica del vocabulario y del modelo de lenguaje, lo que ha permitido reducir la tasa de palabras fuera del vocabulario y la tasa de error de palabra del reconocedor del habla. En segundo lugar se ha utilizado la técnica de realimentación por pseudo-relevancia en el motor de recuperación de información, lo que ha contribuido a mejorar el rendimiento del sistema, tanto al emplear preguntas tex-

tuales como al emplear preguntas habladas. En tercer lugar se ha propuesto incluir la pronunciación de palabras inglesas, con el objetivo de que el reconocedor del habla sea capaz de comprender a locutores castellanos diciendo palabras en inglés. Al incorporar estas mejoras, los experimentos finales han demostrado la viabilidad de construir sistemas de recuperación de información dirigida por habla. Aunque el rendimiento no es tan bueno como el obtenido al emplear texto como entrada, el sistema resulta de utilidad en situaciones donde el habla es la única modalidad disponible, por ejemplo cuando se conduce un coche, o en aquellas situaciones en las que el habla es la modalidad más cómoda, por ejemplo en dispositivos móviles en los que resulta laborioso introducir texto.

8.2. Trabajo futuro

A pesar de que los resultados del trabajo desarrollado son satisfactorios, hay aspectos que pueden mejorarse y que abren la puerta hacia una serie de trabajos futuros.

La continuación natural del trabajo desarrollado consistiría en extender el sistema de diálogo hablado para el acceso a un sitio web descrito en el capítulo 6 incorporando el sistema de búsqueda de información dirigida por habla descrito en el capítulo 7. Por un lado se superarían las limitaciones encontradas en la estrategia de búsqueda del sistema de diálogo. Por otro lado la interacción con el usuario proporcionaría una realimentación valiosa para refinar la búsqueda y mejorar el proceso de recuperación de información. Para que esta búsqueda interactiva de información fuera posible, sería necesario que el sistema fuera capaz de generar preguntas al usuario con el fin de concretar progresivamente su necesidad de información. Esto abre una línea de trabajo en la generación de preguntas de clarificación.

Por otro lado, es posible mejorar los módulos que se encargan de generar la salida hablada del sistema de diálogo. En la parte de generación de respuestas, la utilización de técnicas de resumen automático puede ayudar a reducir el flujo de información que se envía al usuario. En la parte de conversión texto-habla, un correcto modelado prosódico permitiría obtener una salida hablada más natural, ya que los propios usuarios durante el proceso de evaluación han comentado que el sistema resulta aburrido y monótono.

La utilización de un sistema de búsqueda de respuestas es también una posible vía de trabajo futuro. Este tipo de sistemas obtienen la respuesta a preguntas formuladas en lenguaje natural. La incorporación de un sistema de este tipo puede ser útil para el acceso mediante habla a contenidos web que sean en su mayoría texto libre, y que por tanto, tengan poca estructura. Sin embargo, es necesario evaluar el impacto de los errores de reconocimiento de habla en este tipo de sistemas. En [59] se describen nuestros primeros trabajos en esta dirección.

Otra línea de trabajo interesante es el desarrollo de un sistema de diálogo multimodal que permita el acceso a contenidos web a través de dispositivos móviles.

Se trata de un tema de gran interés, debido a la proliferación de este tipo de dispositivos. La idea es utilizar el habla para superar las limitaciones de entrada de datos de estos dispositivos. Además, la combinación de modalidades complementarias daría lugar a un sistema más robusto, en el que sería más fácil la recuperación de errores. Sin embargo, debido a las limitaciones de cómputo de los dispositivos móviles, sería necesario utilizar una arquitectura cliente/servidor, lo que complicaría la implementación final del sistema. Nuestros primeros trabajos en esta dirección se describen en [32], donde se presenta un sistema de navegación web multimodal que combina una interfaz gráfica con una interfaz hablada y que permite acceder a los servicios ofrecidos por *Google Search* [1].

[1] `http://www.google.es`

Apéndices

Apéndice A

El estándar Voice eXtensible Markup Language (VoiceXML)

A.1. Introducción

VoiceXML es un lenguaje de marcas diseñado para especificar diálogos hablados entre una persona y un ordenador. Además, permite desarrollar aplicaciones vocales de una manera similar a como se desarrollan las aplicaciones web. De igual manera que los documentos HTML son interpretados por un navegador web, los documentos VoiceXML son interpretados por un navegador vocal.

El origen de VoiceXML se remonta a 1999, cuando AT&T, IBM, Lucent y Motorola crearon el VoiceXML Forum con el objetivo de desarrollar y promocionar un lenguaje de marcas estándar para especificar diálogos hablados. Aunque ya existían otros lenguajes de marcas similares (por ejemplo VoxML de Motorola y SpeechML de IBM), se consideró interesante la creación de un lenguaje estándar que permitiera independizar la aplicación de la plataforma tecnológica empleada. De esta manera es posible desarrollar aplicaciones que se pueden ejecutar en plataformas de diferentes fabricantes. En marzo de 2000 apareció VoiceXML 1.0, desarrollado por el VoiceXML Forum [133]. Posteriormente fue remitido al W3C, que a partir de entonces es el encargado del desarrollo del estándar. VoiceXML 2.0 alcanzó el estado de *recomendación del W3C* en marzo de 2004 [93]. VoiceXML 2.1 incorporó pocas características adicionales a VoiceXML 2.0 y alcanzó el estado de *recomendación del W3C* en junio de 2007 [101].

Sin duda, el punto fuerte de VoiceXML es la portabilidad, ya que es posible independizar la aplicación de la plataforma que hay por debajo. Por un lado, las empresas pueden crear su propia plataforma de implementación con características particulares. Por otro lado, los desarrolladores de aplicaciones sólo se dedican a la construcción de portales de voz, sin preocuparse de los detalles de bajo nivel.

Por último, el modo de funcionamiento es muy similar al de la web, y por tanto, permite trasladar la potencia del desarrollo web a las aplicaciones guiadas por habla. Esto facilita la integración de servicios de voz con servicios de datos usan-

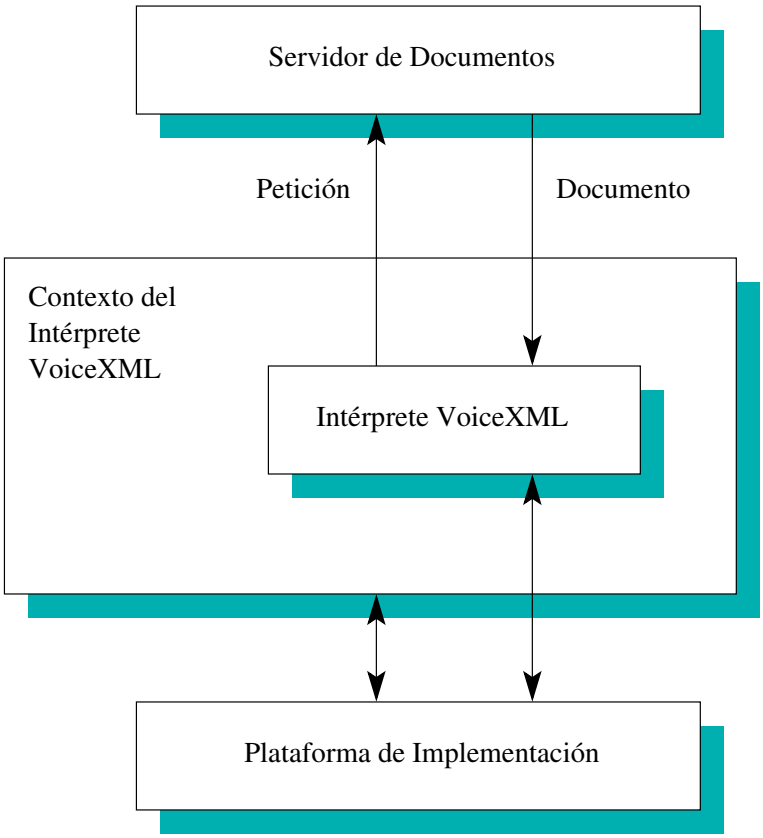


Figura A.1: Modelo arquitectónico de VoiceXML.

do la arquitectura cliente/servidor tradicional. Además, permite la separación de la lógica de aplicación, que se ejecuta en un servidor web estándar, de los diálogos hablados, que se ejecuta en un servidor de telefonía.

A.2. Modelo arquitectónico

Para el acceso a los contenidos escritos en VoiceXML se utiliza un navegador vocal. Al igual que un navegador web interpreta las páginas HTML y permite acceder a la información de manera gráfica, un navegador vocal interpreta las páginas VoiceXML y proporciona acceso a la información empleando habla.

El modelo arquitectónico propuesto en el estándar VoiceXML puede verse en la figura A.1. La arquitectura está compuesta por tres elementos principales: el servidor de documentos, el intérprete VoiceXML y la plataforma de implementación.

A.2.1. Servidor de documentos

El servidor de documentos es un servidor web que se encarga de proporcionar al intérprete las páginas VoiceXML que contienen la aplicación vocal. Este modelo permite utilizar páginas dinámicas, facilitando de esta manera la creación de todo tipo de aplicaciones vocales. También permite de manera sencilla la integración de bases de datos.

A.2.2. Intérprete VoiceXML

El intérprete es el encargado de procesar las páginas VoiceXML que recibe del servidor de documentos y en función del contenido de éstas gestionar el diálogo con el usuario. Para ello, utiliza la plataforma de implementación, a través de la cual se realiza la interacción con el usuario. Para interpretar las páginas VoiceXML se emplea el algoritmo de interpretación de formularios, según se describe en la sección A.4.

A.2.3. Plataforma de implementación

La plataforma de implementación es la encargada de abstraer el sistema hardware y software responsable de la interacción con el usuario del intérprete VoiceXML. La plataforma de implementación debe ser capaz de gestionar la línea telefónica, reproducir ficheros de audio, hacer la conversión texto-habla, grabar audio, reconocer tonos DTMF y hacer reconocimiento de habla.

A.3. Construcción de aplicaciones

La construcción de una aplicación VoiceXML comprende tanto la preparación de los contenidos como la elaboración del flujo de control. Por un lado, hay que indicar qué información se enviará al usuario y qué información se recogerá de éste. Por otro lado, hay que especificar cómo se van a navegar dichas páginas y en qué orden se envía y recibe la información del usuario.

A.3.1. Concepto de aplicación

Una aplicación está formada por un conjunto de documentos VoiceXML que comparten un documento raíz. En dicho documento raíz se colocan los elementos que queremos que sean accesibles en cualquier momento. Básicamente, una aplicación no es más que una máquina de estados, en la que para cada estado se especifica la información a enviar al usuario, la información que debe proporcionar el usuario y cuál será el siguiente estado.

Con este modelo, la ejecución de aplicaciones es muy sencilla. En primer lugar se envía el mensaje al usuario empleando conversión texto habla. A continuación se realiza la captura de datos mediante reconocimiento del habla. Por último, en función de la captura se decide el siguiente estado y se repite todo el proceso.

A.3.2. Diálogos

Cada uno de los estados de una aplicación es lo que VoiceXML denomina *diálogo*, que es la unidad mínima de ejecución. En una misma página VoiceXML podemos tener más de un diálogo. Se distinguen dos tipos de diálogos: formularios y menús.

A.3.2.1. Formularios

Un formulario sirve para recolectar un conjunto de ítems de información, de manera similar a un formulario web. Para cada uno de los campos del formulario se utiliza una gramática para describir qué información se debe recopilar. Se especifica igualmente el orden de recogida de los datos y la transición posterior.

Los formularios se describen con la etiqueta `<form>`, y se les asigna un nombre, que servirá para poder seleccionarlo como destino de una transición. Los campos del formulario más comunes son los de tipo `<field>`, que sirven para realizar una captura simple de datos. A cada campo se le asigna un nombre, que será la variable en la que se almacenará dicha captura, y una gramática, que indica qué puede decir el usuario.

También existe la posibilidad de ejecutar código ECMAScript dentro de un formulario.

A.3.2.2. Menús

Un menú sirve para controlar el flujo de ejecución y permite seleccionar cuál es el diálogo siguiente. Para el menú se emplea la etiqueta `<menu>`, con el nombre del diálogo, que servirá para seleccionarlo como destino de una transición. Cada una de las posibles opciones se especifica con la etiqueta `<choice>`. Para cada opción hay que indicar el mensaje a enviar al usuario y cuál será el diálogo destino. La etiqueta `<enumerate>` sirve para indicar al usuario las posibles alternativas.

En este tipo de diálogo no es necesario especificar la gramática, ya que ésta se genera automáticamente a partir de las posibilidades del menú.

A.3.3. Especificación de salidas

Se utiliza la etiqueta `<prompt>` para especificar los mensajes a enviar al usuario. El texto proporcionado será convertido a habla. Dicho texto puede ser una cadena de texto o un valor calculado mediante una expresión ECMAScript.

Además, se puede utilizar el lenguaje de marcado *W3C Speech Synthesis Markup Language* [19] para proporcionar información adicional al motor de conversión texto-habla. Básicamente se pueden incluir pausas, énfasis e información prosódica. También se puede indicar cómo pronunciar ciertos elementos (fechas, números, etc.) y la voz a emplear.

También existe la posibilidad de que la salida sea un fichero de sonido o un flujo de audio a través de Internet (etiqueta `<audio>`).

A.3.4. Especificación de entradas

Para especificar lo que el usuario puede decir en un punto determinado de la interacción se emplean gramáticas (etiqueta `<grammar>`), que pueden ser de entrada hablada o de tonos telefónicos (DTMF). Hay una serie de gramáticas de uso frecuente que están incluidas en VoiceXML para utilizarse directamente: sí/no, fecha, hora, dígitos, números, número de teléfono y cantidad de dinero.

El ámbito de las gramáticas es variable, pueden estar asociadas a un campo, a un diálogo, a una página o a una aplicación completa. Además, puede haber varias gramáticas activas al mismo tiempo.

También existe la posibilidad de que la entrada se grabe directamente a un fichero de audio (etiqueta `<record>`).

A.3.5. Especificación de transiciones

Para indicar cuál será el siguiente diálogo a ejecutar se emplea la etiqueta `<goto>`. Para seleccionar el diálogo destino se indica el nombre asociado al diálogo, si está en la página actual, o el nombre y la URL, si el diálogo está en otra página VoiceXML.

También es posible enviar la información recolectada a un programa CGI, en cuyo caso se emplea la etiqueta `<submit>`.

Existe la posibilidad de redirigir la llamada a otro número de teléfono, lo cual finaliza la ejecución de la aplicación (etiqueta `<transfer>`).

A.3.6. Subdiálogos

Un subdiálogo es similar a una llamada a función en un lenguaje de programación. Mediante la etiqueta `<subdialog>` se realiza la invocación del subdiálogo correspondiente. Se pueden pasar argumentos (etiqueta `<param>`) y como resultado se puede devolver un objeto ECMAScript (etiqueta `<return>`).

La utilización de subdiálogos permite estructurar mejor las aplicaciones y construir bibliotecas de diálogos reutilizables.

A.3.7. Condiciones de error

En cualquier punto de la interacción pueden darse circunstancias que rompen el flujo normal de ejecución. Cuando esto ocurre se lanza un evento y esto provoca que se ejecute el código VoiceXML asociado a dicho evento. Los gestores de eventos se especifican con la etiqueta `<catch>`.

Los eventos más habituales son:

- *Exit*: el usuario dice “salir”.
- *Nomatch*: el usuario dice algo pero no se reconoce correctamente.
- *Noinput*: el usuario no dice nada.

- *Error*: se produce un error al cargar la página o al procesarla.
- *Help*: el usuario dice "ayuda".

A.3.8. Soporte de iniciativa mixta

El modelo básico de interacción que proporciona VoiceXML es dirigido por el sistema. Esto hace que sea sencillo el desarrollo de aplicaciones, pero que la interacción resultante sea poco flexible. Sin embargo, VoiceXML dispone de mecanismos para permitir que la iniciativa esté compartida entre el sistema y el usuario, lo que se denomina iniciativa mixta.

Básicamente hay dos posibilidades. La primera consiste en emplear gramáticas cuyo alcance sea todo un formulario. Esto permite rellenar varios campos a la vez o en un orden diferente al establecido. La segunda consiste en utilizar gramáticas que estén activas al mismo tiempo que las gramáticas de cada diálogo. De este modo es posible alterar el flujo de control en función de lo que diga el usuario.

A.3.9. Código ejecutable

VoiceXML permite el uso de variables, de sentencias condicionales y de saltos incondicionales. Si es necesario hacer algo más complejo se puede incluir código ejecutable expresado en ECMAScript mediante la etiqueta `<script>`, al igual que en HTML.

También existe la posibilidad de ejecutar objetos incrustados dependientes de plataforma (etiqueta `<object>`).

A.4. Ejecución de aplicaciones

La ejecución de las aplicaciones se realiza mediante la interpretación de los diálogos. Se empieza con el primer diálogo, y su interpretación determina cuál será el siguiente diálogo. De esta manera, se van ejecutando sucesivamente diálogos hasta que, o bien hay un diálogo que finaliza la aplicación, o bien el usuario cuelga el teléfono. La ejecución de diálogos se realiza mediante el algoritmo de interpretación de formularios (*Form Interpretation Algorithm*, FIA). El funcionamiento del FIA se modela como una máquina de estados. Podemos distinguir cuatro fases: inicialización, selección, recolección y procesamiento. La fase de inicialización se ejecuta únicamente una vez al comenzar la interpretación del formulario, sin embargo las otras tres se ejecutan repetidamente una detrás de otra hasta que salimos del formulario. El esquema de ejecución de aplicaciones se muestra en la figura A.2. En la fase de *inicialización* se inicializan las variables y los contadores. En la fase de *selección* se elige el siguiente elemento del formulario a visitar, en función de lo indicado por el último elemento visitado o en función de las condiciones de guarda de los elementos. En la fase de *recolección* se envía al usuario el mensaje

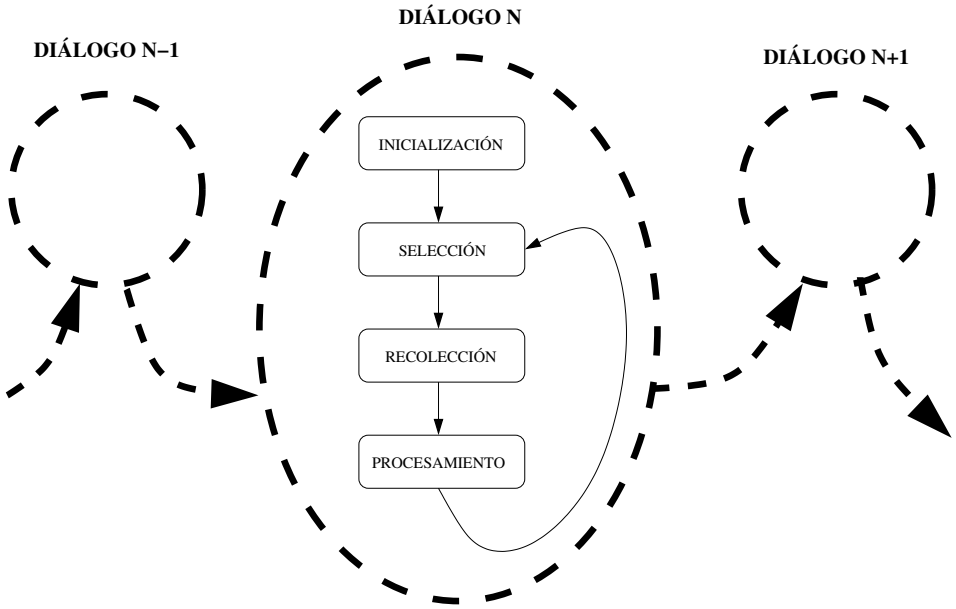


Figura A.2: Ejecución de aplicaciones VoiceXML.

que corresponda y se espera por una entrada (o evento) del usuario. En la fase de *procesamiento* se procesa la entrada o evento que se obtuvo en la fase anterior.

A.5. Ejemplos

A continuación se describen dos ejemplos con el fin de ilustrar las posibilidades de VoiceXML.

A.5.1. Ejemplo de formulario

En este ejemplo se utiliza un formulario que permite al usuario hacer un pedido a una pizzería. El usuario debe proporcionar todos los campos del formulario: la cantidad de pizzas, el tamaño, si quiere extra de queso, los ingredientes vegetales y los ingredientes de carne. El formulario se muestra en la figura A.3 y el código VoiceXML en la figura A.4. En la figura A.5 se muestra un ejemplo de interacción con un usuario.

A.5.2. Ejemplo de menú

En este ejemplo se utiliza un menú que permite al usuario seleccionar una de las cuatro posibilidades siguientes: información, docencia, investigación o miembros. El diagrama de estados asociado al diálogo se muestra en la figura A.6. y el código

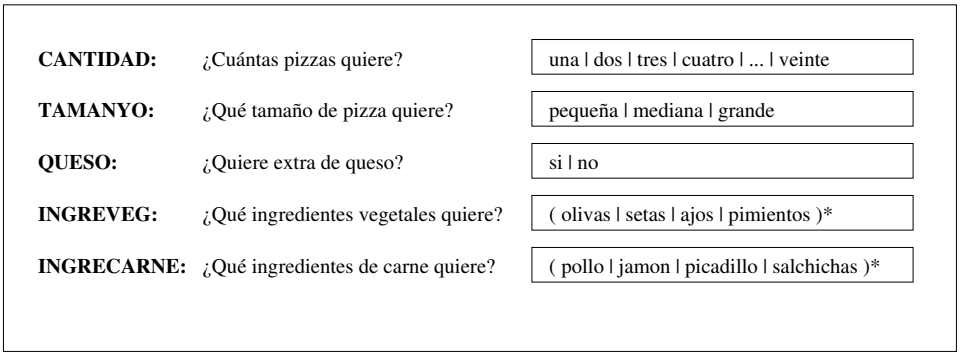


Figura A.3: Ejemplo de formulario.

fuente en la figura A.7. En la figura A.8 se muestra un ejemplo de interacción con un usuario.

A.6. Plataforma VoiceXML del grupo ECA-SIMM

En el presente trabajo de tesis hemos utilizado la plataforma VoiceXML del grupo ECA-SIMM. Esta plataforma ha sido desarrollada partiendo del resultado de varios proyectos fin de carrera [49, 92]. La plataforma está compuesta por un intérprete VoiceXML de desarrollo propio; motores de síntesis de habla y reconocimiento de habla desarrollados en la *Universidad Politécnica de Cataluña*[1]; y una tarjeta telefónica *Dialogic D/120JCT-LS*.

[1] `http://www.verbio.com`

```
<?xml version="1.0"?>
<!DOCTYPE vxml PUBLIC "-//W3C//DTD VOICEXML 2.0//EN" "vxml.dtd">
<vxml version="2.0" xmlns="http://www.w3.org/2001/vxml">
 <form id="pedido">
  <field name="cantidad">
   <prompt>¿Cuántas pizzas quiere?</prompt>
   <grammar> una | dos | tres | cuatro | cinco | seis | siete |
             ocho | nueve | diez | once | doce | trece | catorce |
             quince | dieciseis | diecisiete | dieciocho |
             diecinueve | veinte </grammar>
   <catch event="help nomatch noinput">Diga un número entre una
                                       y veinte.</catch>
  </field>
  <field name="tamanyo">
   <prompt>¿Qué tamaño de pizza quiere?</prompt>
   <grammar> pequeña | mediana | grande </grammar>
   <catch event="help nomatch noinput">Diga pequeña mediana o
                                       grande.</catch>
  </field>
  <field name="queso" type="boolean">
   <prompt>¿Quiere extra de queso?</prompt>
   <catch event="help nomatch noinput">Diga sí o no.</catch>
  </field>
  <field name="ingreVeg">
   <prompt>¿Qué ingredientes vegetales quiere?</prompt>
   <grammar> (olivas | setas | ajos | pimientos)* </grammar>
   <catch event="help nomatch noinput">Diga olivas, setas, ajos
                                       o pimientos.</catch>
  </field>
  <field name="ingreCarne">
   <prompt>¿Qué ingredientes de carne quiere?</prompt>
   <grammar> (pollo | picadillo | jamón | salchichas )* </grammar>
   <catch event="help nomatch noinput">Diga pollo, picadillo,
                                       jamón o salchichas.</catch>
  </field>
  <filled>
   <prompt> Ha elegido: <value expr="cantidad"/> pizzas</prompt>
   <prompt> de tamaño <value expr="tamanyo"/>.</prompt>
   <if cond="queso==’si’">
    <prompt>Ha elegido extra de queso.</prompt>
   </if>
   <prompt>Con ingredientes <value expr="ingreVeg"/></prompt>
   <prompt> e ingredientes <value expr="ingreCarne"/></prompt>
   <submit next="procesa.php" method="get" namelist=
           "cantidad tamanyo queso ingreVeg ingreCarne"/>
  </filled>
 </form>
</vxml>
```

Figura A.4: Código VoiceXML correspondiente al formulario de la figura A.3.

Sistema: ¿Cuántas pizzas quiere?
Usuario: una
Sistema: ¿Qué tamaño de pizza quiere?
Usuario: mediana
Sistema: ¿Quiere extra de queso?
Usuario: sí
Sistema: ¿Qué ingredientes vegetales quiere?
Usuario: olivas ajos
Sistema: ¿Qué ingredientes de carne quiere?
Usuario: pollo
Sistema: Ha elegido una pizzas de tamaño mediana. Ha elegido extra de queso. Con ingredientes olivas ajos e ingredientes pollo.
(... *el sistema envía la información a procesa.php* ...)

Figura A.5: Ejemplo de interacción para la página VoiceXML de la figura A.4.

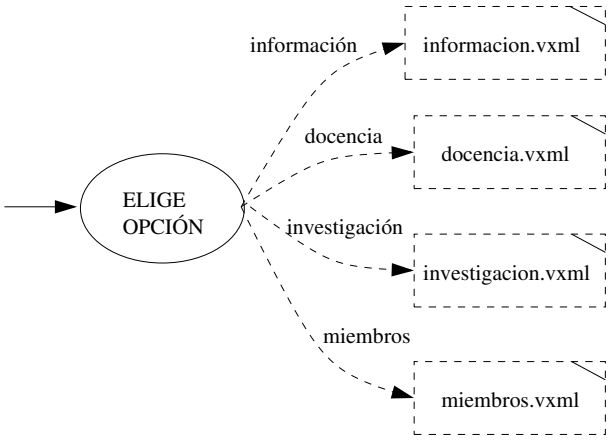


Figura A.6: Diagrama de estados de un menú.

```
<?xml version="1.0"?>
<!DOCTYPE vxml PUBLIC "-//W3C//DTD VOICEXML 2.0//EN" "vxml.dtd">
<vxml version="2.0" xmlns="http://www.w3.org/2001/vxml">
 <menu id="seccion">
  <prompt>Por favor, elija una de las siguientes opciones:
          <enumerate/> </prompt>
  <choice next="informacion.vxml">información</choice>
  <choice next="docencia.vxml">docencia</choice>
  <choice next="investigacion.vxml">investigación</choice>
  <choice next="miembros.vxml">miembros</choice>
  <catch event="help nomatch noinput">Por favor diga uno de
                                      <enumerate/> </catch>
 </menu>
</vxml>
```

Figura A.7: Código VoiceXML correspondiente al menú de la figura A.6.

Sistema: Por favor, elija una de las siguientes opciones: información, docencia, investigación, miembros.
Usuario: miembros
(*... el sistema pasa a la página miembros.vxml ...*)

Figura A.8: Ejemplo de interacción para la página VoiceXML de la figura A.7.

Apéndice B

Evaluación del sistema de diálogo hablado para el acceso al sitio web de "El Norte de Castilla"

En este apéndice se incluye información extra del proceso de evaluación del sistema de diálogo hablado descrito en el capítulo 6, que permite el acceso a los contenidos del periódico "El Norte de Castilla" [1].

La evaluación del sistema se realizó con los contenidos del periódico correspondientes al día 30 de noviembre de 2004.

En primer lugar se presentan los escenarios utilizados en la evaluación. A continuación se enumeran las preguntas del cuestionario de satisfacción de usuario. También se muestran las medidas de rendimiento y satisfacción de usuario obtenidas para cada usuario. Por último, se incluyen todos los comentarios realizados por los usuarios tras finalizar el proceso de evaluación.

B.1. Escenarios de la evaluación

Para evaluar el sistema se utilizaron los 5 escenarios siguientes:

Escenario 1 Acceda al sistema de diálogo empleando navegación o búsqueda. Acceda a la primera noticia que llame su atención. Describa brevemente dicha noticia.

Escenario 2 ¿Cuál es el suceso ocurrido en España que considera más importante de los que aparecen en el periódico? Obtenga una breve descripción.

Escenario 3 ¿Hay alguna actividad cultural a la que pueda acudir próximamente? En ese caso, obtenga dónde tiene lugar y en qué fecha.

[1] `http://www.nortecastilla.es`

Escenario 4 En su acceso al sistema debe informarse sobre el número de disidentes cubanos que han sido liberados recientemente.

Escenario 5 Busque la noticia donde se informa sobre las razones que han llevado a la coordinadora de Pajarillos a exigir una rectificación del alcalde.

B.2. Preguntas del cuestionario

A continuación se muestra la lista de afirmaciones del cuestionario empleado para medir la satisfacción de usuario. Para cada afirmación se muestra la versión original en inglés junto con la traducción que se ha utilizado para la evaluación.

1. The system is accurate.
   El sistema es preciso.
2. The system is unreliable.
   El sistema es poco fidedigno.
3. The interaction with the system is unpredictable.
   La interacción con el sistema es impredecible.
4. The system didn't always do what I wanted.
   El sistema no hizo siempre lo que yo quería.
5. The system didn't always do what I expected.
   El sistema no hizo siempre lo que yo esperaba.
6. The system is dependable.
   El sistema es fiable.
7. The system makes few errors.
   El sistema comete pocos errores.
8. The interaction with the system is consistent.
   La interacción con el sistema es consistente.
9. The interaction with the system is efficient.
   La interacción con el sistema es eficiente.
10. The system is useful.
    El sistema es útil.
11. The system is pleasant.
    El sistema es agradable.

12. The system is friendly.
    El sistema es amable.
13. I was able to recover easily from errors.
    Fui capaz de recuperarme fácilmente de los errores.
14. I enjoyed using the system.
    Disfruté usando el sistema.
15. It is clear how to speak to the system.
    Está claro cómo hay que hablar al sistema.
16. It is easy to learn to use the system.
    Es fácil aprender a usar el sistema.
17. I would use this system.
    Yo usaría este sistema.
18. I felt in control of the interaction with the system.
    Sentí que controlaba la interacción con el sistema.
19. I felt confident using the system.
    Me sentí seguro usando el sistema.
20. I felt tense using the system.
    Me sentí tenso usando el sistema.
21. I felt calm using the system.
    Me sentí tranquilo usando el sistema.
22. A high level of concentration is required when using the system.
    Se requiere un alto grado de concentración cuando se usa el sistema.
23. The system is easy to use.
    El sistema es fácil de usar.
24. The interaction with the system is repetitive.
    La interacción con el sistema es repetitiva.
25. The interaction with the system is boring.
    La interacción con el sistema es aburrida.
26. The interaction with the system is irritating.
    La interacción con el sistema es irritante.

27. The interaction with the system is frustrating.

    La interacción con el sistema es frustrante.

28. The system is too inflexible.

    El sistema es demasiado inflexible.

29. I sometimes wondered if I was using the right word.

    Algunas veces me pregunté si estaba usando la palabra adecuada.

30. I always knew what to say to the system.

    Siempre supe qué decir al sistema.

31. I was not always sure what the system was doing.

    No estuve siempre seguro de qué estaba haciendo el sistema.

32. It is easy to lose track of where you are in an interaction with the system.

    Es fácil perder el hilo de dónde te encuentras en la interacción con el sistema.

33. The interaction with the system is fast.

    La interacción con el sistema es rápida.

34. The system responds too slowly.

    El sistema responde demasiado despacio.

B.3. Medidas de rendimiento

Para medir el rendimiento del sistema se han utilizado una serie de medidas objetivas. Las medidas obtenidas para cada usuario se muestran en la tabla B.1.

B.4. Medidas de satisfacción de usuario

Para calcular la satisfacción de usuario se ha empleado un cuestionario. La media de cada pregunta puede verse en la tabla B.2. Se han normalizado los valores para simplificar la manipulación de los datos. Para ello, algunas puntuaciones se invirtieron de manera que valores altos en todas las categorías fueran considerados buenos. En la figura B.1 se muestra un gráfico con las preguntas normalizadas (las respuestas invertidas están marcadas en blanco). Las respuestas de todos los usuarios a cada pregunta pueden verse en la tabla B.3 y B.4.

Usuario	TET	DML	TUL	WER	LUR	TSI	FI	TUSE	TUA
1	100 %	230,2	12,8	18,6 %	3,1 %	49,3 %	20,6 %	0,0 %	1,6 %
2	100 %	245,6	13,2	13,0 %	7,6 %	81,4 %	0,0 %	0,0 %	1,5 %
3	100 %	264,3	10,7	13,8 %	6,3 %	32,9 %	8,7 %	0,0 %	0,0 %
4	100 %	150,6	6,2	3,0 %	0,0 %	69,4 %	0,0 %	0,0 %	0,0 %
5	100 %	154,6	9,2	26,9 %	8,7 %	84,3 %	0,0 %	0,0 %	0,0 %
6	100 %	184,4	7,0	0,0 %	0,0 %	62,5 %	0,0 %	0,0 %	0,0 %
7	100 %	150,4	9,0	31,8 %	11,1 %	60,0 %	0,0 %	0,0 %	0,0 %
8	100 %	164,6	5,6	14,7 %	0,0 %	66,7 %	0,0 %	0,0 %	0,0 %
9	100 %	147,6	6,8	19,4 %	11,8 %	71,8 %	0,0 %	0,0 %	0,0 %
10	80 %	236,6	10,4	22,8 %	13,5 %	57,9 %	0,0 %	0,0 %	0,0 %
11	100 %	311,8	16,6	8,4 %	2,4 %	79,5 %	1,4 %	1,2 %	0,0 %
12	100 %	188,6	9,0	0,0 %	0,0 %	84,0 %	9,5 %	0,0 %	0,0 %
13	80 %	171,6	7,4	15,4 %	10,8 %	69,0 %	3,4 %	0,0 %	0,0 %
14	80 %	176,3	10,7	35,2 %	7,8 %	61,4 %	30,2 %	0,0 %	0,0 %
15	60 %	219,8	13,5	38,7 %	4,9 %	60,9 %	1,9 %	2,5 %	0,0 %
16	100 %	234,4	8,2	16,3 %	12,2 %	73,9 %	2,9 %	2,4 %	0,0 %
17	80 %	213,0	7,3	12,0 %	9,1 %	50,0 %	8,0 %	2,3 %	0,0 %
18	100 %	221,2	8,3	23,1 %	2,0 %	25,0 %	14,3 %	0,0 %	0,0 %
19	80 %	336,4	8,2	6,7 %	0,0 %	4,3 %	50,0 %	0,0 %	0,0 %
20	100 %	241,8	12,2	24,5 %	4,9 %	57,6 %	36,8 %	0,0 %	0,0 %
21	80 %	300,8	10,2	26,4 %	7,8 %	16,1 %	11,1 %	5,9 %	2,0 %
22	80 %	175,0	9,0	27,3 %	7,4 %	81,7 %	0,0 %	1,9 %	0,0 %
Media	**92 %**	**214,5**	**9,6**	**18,1 %**	**6,0 %**	**59,1 %**	**9,0 %**	**0,7 %**	**0,2 %**

Tabla B.1: Medidas objetivas de rendimiento para cada usuario. (TET: tasa de éxito de la tarea; DML: duración media de las llamadas; TUL: turnos de usuario por llamada; WER: tasa de error de palabra; LUR: locuciones de usuario rechazadas; TSI: turnos de sistema interrumpidos; FI: falsas interrupciones; TUSE: turnos de usuario sin entrada; TUA: turnos de usuario de ayuda).

	Pregunta	Media	
1	El sistema es preciso	5,45	
2	El sistema es poco fidedigno	2,18	(5,82)
3	La interacción con el sistema es impredecible	2,36	(5,64)
4	El sistema no hizo siempre lo que yo quería	3,09	(4,91)
5	El sistema no hizo siempre lo que yo esperaba	3,41	(4,59)
6	El sistema es fiable	5,82	
7	El sistema comete pocos errores	5,41	
8	La interacción con el sistema es consistente	5,23	
9	La interacción con el sistema es eficiente	5,32	
10	El sistema es útil	6,32	
11	El sistema es agradable	4,86	
12	El sistema es amable	5,45	
13	Fui capaz de recuperarme fácilmente de los errores	5,73	
14	Disfruté usando el sistema	5,23	
15	Está claro cómo hay que hablar al sistema	5,77	
16	Es fácil aprender a usar el sistema	6,55	
17	Yo usaría este sistema	5,09	
18	Sentí que controlaba la interacción con el sistema	5,73	
19	Me sentí seguro usando el sistema	5,55	
20	Me sentí tenso usando el sistema	2,41	(5,59)
21	Me sentí tranquilo usando el sistema	5,55	
22	Se requiere un alto grado de concentración cuando se usa el sistema	3,55	(4,45)
23	El sistema es fácil de usar	6,41	
24	La interacción con el sistema es repetitiva	4,82	(3,18)
25	La interacción con el sistema es aburrida	3,73	(4,27)
26	La interacción con el sistema es irritante	3,18	(4,82)
27	La interacción con el sistema es frustrante	2,27	(5,73)
28	El sistema es demasiado inflexible	3,50	(4,50)
29	Algunas veces me pregunté si estaba usando la palabra adecuada	5,27	(2,73)
30	Siempre supe qué decir al sistema	4,82	
31	No estuve siempre seguro de qué estaba haciendo el sistema	2,45	(5,55)
32	Es fácil perder el hilo de dónde te encuentras en la interacción con el sistema	2,50	(5,50)
33	La interacción con el sistema es rápida	5,23	
34	El sistema responde demasiado despacio	3,32	(4,68)

Tabla B.2: Media de las respuestas originales de todos los usuarios al cuestionario. Para las preguntas que precisan normalización el valor invertido se muestra entre paréntesis al lado del valor original.

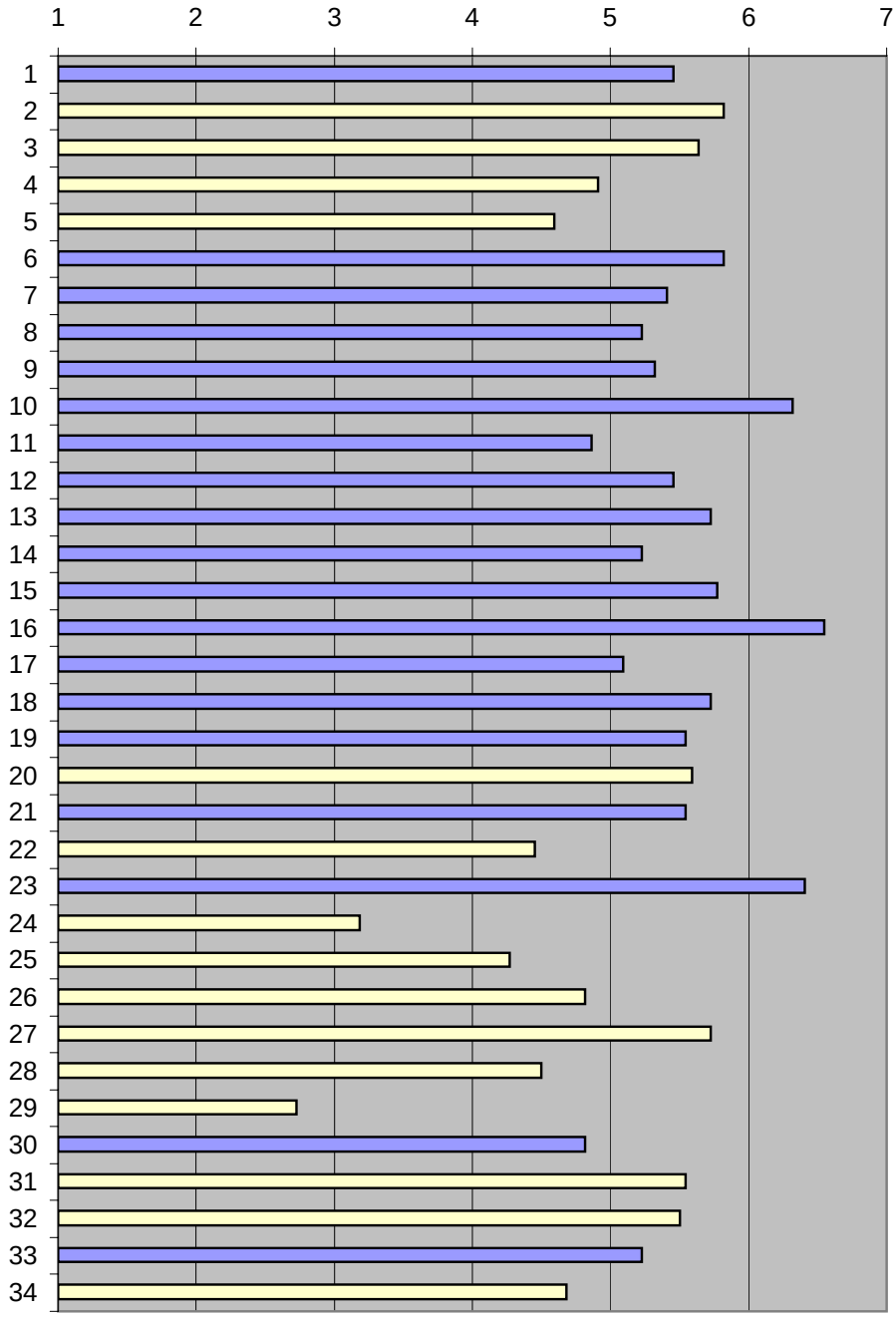


Figura B.1: Media de las respuestas de todos los usuarios al cuestionario. Las preguntas invertidas se muestran en blanco.

	1	2	3	4	5	6	7	8	9	10	11
1	6	6	6	6	3	6	5	6	6	4	6
2	2	1	3	1	4	1	4	2	2	3	2
3	2	3	2	2	2	1	6	2	2	2	2
4	6	1	2	2	1	1	2	3	2	4	4
5	5	3	2	1	6	1	2	5	2	4	2
6	6	7	3	6	6	6	6	5	7	5	4
7	6	7	6	6	7	7	6	6	5	4	2
8	6	4	4	6	4	6	6	5	6	6	6
9	4	7	5	6	7	6	5	5	3	5	6
10	6	6	5	6	7	7	7	6	7	6	6
11	4	4	3	6	4	5	3	4	3	5	7
12	7	7	4	6	5	4	4	5	4	5	6
13	6	7	5	7	5	4	6	6	6	6	7
14	4	6	2	4	4	6	6	4	4	5	6
15	6	6	6	6	5	6	7	4	3	5	6
16	7	7	6	7	6	7	7	6	7	6	7
17	4	6	2	4	6	6	6	4	5	5	6
18	6	7	4	6	3	6	6	5	6	6	6
19	4	4	4	7	5	7	6	5	5	6	7
20	4	5	2	2	1	1	3	2	2	3	2
21	5	3	3	7	5	7	6	5	5	7	7
22	2	6	6	1	3	5	4	3	4	2	5
23	7	7	5	7	5	7	7	6	6	7	7
24	5	6	7	3	4	5	4	5	6	7	5
25	4	5	6	2	4	3	3	5	4	5	4
26	5	5	6	3	3	2	4	2	2	3	2
27	5	2	4	1	2	1	3	2	2	3	2
28	5	4	6	2	4	2	4	4	5	4	4
29	6	7	5	1	7	4	5	6	5	7	5
30	3	6	2	7	3	6	7	3	6	3	6
31	5	1	1	3	3	2	2	2	2	2	1
32	2	2	5	1	3	1	4	3	2	5	3
33	6	5	2	6	2	6	7	6	2	4	6
34	4	2	5	4	6	2	6	3	6	5	2

Tabla B.3: Respuestas al cuestionario de los usuarios 1 al 11.

	12	13	14	15	16	17	18	19	20	21	22
1	6	6	3	5	6	7	5	6	5	5	6
2	2	2	2	1	4	2	2	1	1	4	2
3	1	2	2	7	1	2	5	1	1	2	2
4	2	5	5	6	2	3	5	1	5	3	3
5	2	5	5	5	2	2	7	2	5	4	3
6	6	6	6	5	6	7	6	7	7	6	5
7	6	6	3	2	6	7	5	7	4	5	6
8	6	4	6	4	5	4	5	6	6	5	5
9	6	6	5	4	6	6	3	7	3	6	6
10	5	7	7	6	7	7	6	7	6	6	6
11	5	6	6	4	6	6	4	5	6	5	6
12	5	6	7	4	6	6	7	5	7	6	4
13	6	6	6	3	6	6	3	7	7	6	5
14	4	7	6	6	6	4	6	7	6	6	6
15	6	7	3	7	6	6	6	7	6	6	7
16	7	7	7	6	5	7	6	7	7	7	5
17	5	6	6	5	6	4	6	6	3	5	6
18	6	7	6	4	6	6	6	7	6	5	6
19	6	6	6	3	6	5	6	7	7	4	6
20	1	5	2	5	2	2	1	1	2	2	3
21	6	3	6	3	6	6	6	7	7	5	7
22	2	5	2	1	4	6	6	1	1	4	5
23	6	7	7	6	6	6	6	7	7	6	6
24	5	4	2	4	7	4	4	4	7	5	3
25	5	1	2	3	2	2	4	6	6	2	4
26	2	3	3	2	2	2	4	5	6	3	1
27	2	3	2	2	1	2	2	1	4	2	2
28	5	5	3	2	2	2	1	4	3	3	3
29	2	5	6	7	6	4	5	6	5	6	6
30	6	6	6	2	3	5	6	7	3	4	6
31	2	3	2	2	3	2	4	1	5	2	4
32	2	5	1	2	3	4	1	1	1	2	2
33	6	6	6	7	6	6	6	5	6	4	5
34	2	2	1	1	3	2	4	5	2	5	1

Tabla B.4: Respuestas al cuestionario de los usuarios 12 al 22.

B.5. Comentarios de los usuarios

Después de rellenar el cuestionario de satisfacción, los usuarios tenían la posibilidad de escribir un comentario libre acerca de sus impresiones al utilizar el sistema. A continuación se muestran los comentarios que realizaron los usuarios:

1. Tal vez habría que añadir alguna sección de espectáculos (¿?) y usar varias voces para no aburrir al que escucha.
2. Ninguno.
3. Ninguno.
4. El sistema me ha parecido bastante eficiente, sobre todo por el hecho de no tener que escuchar todas las opciones para poder elegir, y que las opciones sean siempre las mismas para no tener que pensar qué decir en cada momento. Por otro lado, no siempre se entienden todos los titulares, sobre todo alguna vez junta palabras, es decir, se salta los espacios, haciendo difícil la comprensión del titular. Por lo demás lo considero un sistema rápido y con un funcionamiento correcto.
5. Cuando pides más información sobre una noticia, que te repita el titular, o tengas la posibilidad de repetir la información recibida.
6. Me ha parecido que el sistema ha reconocido perfectamente las palabras a buscar sin tener que decírselas excesivamente despacio ni vocalizar demasiado.
7. Me ha parecido un sistema normal como los que te puedes encontrar actualmente, me parece útil ya que normalmente te interesa un noticia, y no leer todas.
8. Ninguno.
9. El lenguaje debería ser un poco más fluido. También debería tener en cuenta la pronunciación de los nombres extranjeros. Las opciones que se dan son suficientes para acceder a las noticias pero no excesivas como para perder al usuario. El reconocimiento de voz a veces falla.
10. Durante la búsqueda de la respuesta para el escenario 3 no dí con la palabra exacta para encontrar la respuesta o la que decía no la reconocía. En cuanto al manejo del sistema se hace un poco repetitivo si estás un rato consultando noticias, aunque cuando empiezas da toda la información necesaria. También se hecha de menos una opción como “salir” para terminar de emplear el sistema.
11. La interacción con el sistema y su manejo es bastante intuitivo y el usuario rápidamente se familiariza con él. Quizás a la hora de realizar las búsquedas y la propia navegación estaría bien poder acceder a más secciones.

12. En ocasiones al respirar fuerte cerca del auricular el sistema se confundía.

13. Ninguno.

14. Me ha parecido un buen sistema, aunque todavía no reconoce bien palabras sueltas. Aún así, para noticias navegación va muy bien aunque en este otro campo falle un poco.

15. Me ha parecido un sistema en desarrollo muy interesante. He podido apreciar que posee algunas deficiencias a nivel de reconocimiento de voz. En el transcurso de la prueba he nombrado muchas palabras que no me ha cogido. Sobre la navegación se podría depurar más añadiendo por ejemplo siempre vínculos a las palabras navegación o búsqueda porque son las que más se usan y haciendo también accesos a niveles superiores de una manera más rápida. Además tendría que tener una forma de parar en el momento en el que te interesa la noticia que está citando para poder escucharla y no tragarte todas las noticias y luego tener que decir el número. Pero en global me ha parecido un programa muy bueno.

16. Quizás al principio me resultó un poco difícil pero en la tercera noticia ya me sentí más seguro y lo utilicé de forma más rápida y segura. Quizás me causaba algunas dudas en el momento que seleccionaba alguna noticia y le pedía más información y me saltaba antes de recibir la información un menú sobre las posibilidades que tenía, es decir al saltarme el menú de opciones después de haber elegido una noticia dudaba si me había reconocido la opción de conseguir más información, quizás el sistema podía avisar que la opción de obtener más información estaba seleccionada. Por lo demás bastante bien ya que el menú de opciones es muy intuitivo y el menú de búsqueda por palabras funcionaba bastante bien.

17. Es algo novedoso, puede ser rápido si buscas un titular en particular. Me parece bien.

18. Creo que el sistema es fiable y salvo que no responde bien con la interacción del usuario, ya que en alguna ocasión no me entendió con claridad. Importante destacar que en el momento de la lectura de alguna noticia el sistema ha fallado y me ha dado una noticia fuera de la sección, por ejemplo cuando quería escuchar más información sobre una noticia el sistema ha ido a las noticias locales de Zamora.

19. El sistema es correcto pero debería permitir elegir la opción sin tener que escuchar todas. Porque si quieres elegir local desde el principio no tienes porque escuchar todo.

20. Sí que resulta algo irritante sobre todo cuando no entiende lo que le dices y te dice todo el rato que se lo repitas, pero una vez que le coges el juego creo

que es fácil de usar y te sabes adelantar al sistema y llegar antes a la noticia que deseas.

21. Ninguno.

22. El sistema me ha parecido interesante y útil, menos en el tercer escenario que no sé qué ocurrió que el sistema terminó colgándome sin saber el por qué; pero por lo demás está muy bien.

Apéndice C

Evaluación del sistema de recuperación de información dirigida por habla

En este apéndice se incluye información adicional del proceso de evaluación realizado al sistema de recuperación de información dirigida por habla descrito en el capítulo 7.

En primer lugar se presenta la lista de preguntas utilizadas para evaluar el sistema. A continuación se muestran los resultados obtenidos para cada locutor en los distintos experimentos realizados. Por último, se presentan dos tablas resumen de todos los experimentos.

C.1. Preguntas del corpus CLEF 2001

En este apartado se presentan las preguntas utilizadas para evaluar el sistema. Las consultas han sido confeccionadas a partir de la colección de prueba para IR monolingüe para español desarrollada por CLEF en el año 2001 (*CLEF 2001 Spanish monolingual IR test-suite*).

C.1.1. Preguntas de longitud media

Estas consultas se han creado empleando el campo descripción de cada tema de la colección CLEF 2001.

41. Encontrar noticias sobre pesticidas en alimentos para bebés.

42. Encontrar documentos sobre la invasión de Haití por los soldados de la ONU y de los Estados Unidos.

43. Encontrar noticias que expliquen el fenómeno de El Niño y su repercusión en el clima del planeta, incluidos los efectos que tiene sobre la temperatura,

presión atmosférica, precipitaciones, etcétera.

44. Reacciones al cuarto Tour de Francia ganado por Miguel Indurain.

45. Encontrar noticias que mencionen los nombres de los principales negociadores del tratado de paz en el Medio Oriente entre Israel y Jordania, y también documentos que den una información detallada sobre el tratado.

46. ¿Qué efectos ha tenido el embargo de la ONU en la vida del pueblo iraquí?

47. ¿Cuáles son las causas de la intervención militar de Rusia en Chechenia?

48. Razones del retiro de las fuerzas de paz de las Naciones Unidas, ONU, de Bosnia.

49. Los documentos informarán sobre el decrecimiento de las exportaciones de automóviles en Japón.

50. Encontrar noticias sobre el levantamiento de campesinos indígenas en Chiapas, México.

51. Encontrar documentos sobre la final del Mundial de fútbol de 1994.

52. Encontrar documentos que describan las razones y los efectos de la devaluación de la moneda china.

53. ¿Qué genes causan o contribuyen a la aparición de enfermedades o trastornos de desarrollo en seres humanos?

54. Encontrar documentos que informen sobre el resultado de la Final Four europea de baloncesto.

55. Encontrar documentos que informen sobre la iniciativa suiza destinada a regular el tráfico a través de los Alpes.

56. Encontrar documentos que hablen sobre campañas contra el racismo en Europa.

57. Encontrar toda la información sobre los juicios sobre sangre infectada en Francia, incluyendo las sentencias emitidas y los nombres de las personas que se encontraron culpables.

58. Los documentos describirán casos de eutanasia entendidos como muerte digna o derecho a morir.

59. Encontrar documentos sobre virus informáticos.

60. Encontrar documentos sobre corrupción política en Francia, en particular con referencia a la financiación ilegal de los partidos políticos franceses.

61. Encontrar información sobre la avería en un oleoducto en Siberia.

62. Encontrar documentos que informen sobre el terremoto en la costa este de Hokkaido, en el norte de Japón, en 1994.

63. Encontrar documentos sobre la reserva de la Antártida en la que está prohibido cazar ballenas.

64. Encontrar documentos que informen sobre RSI, repetitive strain injuries, o enfermedad del periodista, producidas por el uso del ratón del ordenador.

65. Encontrar documentos sobre buscadores de tesoros y actividades de búsqueda de tesoros.

66. Encontrar noticias y debates sobre la retirada de las tropas rusas de Letonia.

67. Encontrar información sobre el número de personas heridas o muertas en colisiones entre barcos.

68. Encontrar documentos que describan actos de terrorismo o vandalismo contra sinagogas europeas desde el fin de la segunda guerra mundial.

69. ¿Cuáles son las aplicaciones prácticas de la clonación y qué argumentos existen en contra de ella?

70. Encontrar documentos que den información biográfica sobre Kim Il Sung, presidente de Corea del Norte, que murió en 1994.

71. Encontrar documentos que relacionen la ingestión de verduras y frutas con el cáncer.

72. ¿Qué papel jugó Rusia en la cumbre del G7 en Nápoles, en 1994?

73. ¿Cuáles fueron las reacciones en Europa a los resultados negativos del referéndum Noruego sobre su posible incorporación a la Unión Europea?

74. Encontrar documentos que describan la inauguración del Eurotúnel y de los nombres de los representantes del Reino Unido y Francia presentes en la ceremonia.

75. Siete personas murieron en la masacre que tuvo lugar en el juzgado de Euskirchen, Alemania.

76. ¿Para qué aplicaciones ha sido usada o se piensa usar la energía solar?

77. ¿De qué información se dispone acerca de los suicidios entre adolescentes?

78. ¿Qué película o películas ganaron el León de Oro en la 51 edición del Festival de Cine de Venecia en Septiembre de 1994?

79. Encontrar documentos que describan la misión de la sonda espacial europea Ulises o discutan sus objetivos.

80. Los documentos proporcionarán cualquier información relacionada con huelgas de hambre organizadas con el fin de atraer la atención hacia una causa.

81. Encontrar toda la información concerniente al papel de un grupo islámico armado en el secuestro de un Airbus de Air France.

82. Encontrar documentos que describan actos terroristas cometidos por el Ejército Republicano Irlandés, IRA, en aeropuertos europeos.

83. Encontrar subastas públicas de objetos de John Lennon.

84. Los documentos contienen cualquier información relacionada con los ataques de tiburones a humanos.

85. Encontrar información detallada sobre la operación Turquesa, un programa francés de ayuda humanitaria en Ruanda.

86. Encontrar documentos que describan la utilización o políticas relacionadas con energías ecológicas, es decir, energía generada a partir de fuentes renovables.

87. Encontrar documentos que analicen la influencia del Plan Real contra la inflación en las elecciones brasileñas.

88. Encontrar documentos que mencionen casos de Encefalopatía Espongiforme Bovina, el mal de las vacas locas, en Europa.

89. Encontrar documentos sobre la bancarrota del agente inmobiliario alemán Schneider.

90. ¿Qué países son exportadores de verduras frescas, secas o congeladas?

C.1.2. Preguntas cortas

Estas consultas se han creado empleando el campo título de cada tema de la colección CLEF 2001.

41. Pesticidas en alimentos para bebés.

42. Naciones Unidas y Estados Unidos invaden Haití.

43. El Niño y el tiempo.

44. Indurain gana el Tour.

45. El tratado de paz entre Israel y Jordania.

46. Embargo sobre Irak.
47. Intervención rusa en Chechenia.
48. Fuerzas de paz en Bosnia.
49. Caída de las exportaciones de coches en Japón.
50. Levantamiento en Chiapas.
51. Mundial de fútbol.
52. Devaluación de la moneda china.
53. Genes y enfermedades.
54. Resultados de la Final Four.
55. Iniciativa Suiza para los Alpes.
56. Campañas europeas contra el racismo.
57. Juicio sobre sangre infectada.
58. Eutanasia.
59. Virus informáticos.
60. Corrupción política en Francia.
61. Ruptura de oleoducto en Siberia.
62. Terremoto en el norte de Japón.
63. Reserva de ballenas.
64. Síndrome RSI y ratones de ordenador.
65. Búsqueda de tesoros.
66. Retirada de tropas rusas de Letonia.
67. Colisiones navales.
68. Ataques a sinagogas europeas.
69. Clonación y ética.
70. Muerte de Kim Il Sung.
71. Verduras, frutas y cáncer.
72. Cumbre del G7 en Nápoles.

73. Referéndum en Noruega sobre la Unión Europea.
74. Inauguración del Eurotúnel.
75. La matanza en el juzgado de Euskirchen.
76. Energía solar.
77. Suicidios entre adolescentes.
78. Festival de Cine de Venecia.
79. Sonda espacial Ulises.
80. Huelgas de hambre.
81. Secuestro de un Airbus francés.
82. El IRA ataca aeropuertos.
83. Subasta de objetos de Lennon.
84. Ataques de tiburones.
85. Programa Turquesa en Ruanda.
86. Energía renovable.
87. Inflación y elecciones en Brasil.
88. Vacas locas en Europa.
89. Schneider en quiebra.
90. Exportadores de verduras.

C.2. Resultados de los experimentos

En este apartado se incluyen los resultados obtenidos para cada uno de los 10 locutores empleados en la evaluación. Los resultados del sistema de referencia se muestran en las tablas C.1, C.2 y C.3. Los resultados obtenidos al realizar adaptación del LM, adaptación de vocabulario y adaptación del LM y vocabulario se muestran en las tablas C.4, C.5 y C.6. Los resultados obtenidos al emplear realimentación por pseudo-relevancia se muestran en la tabla C.7. Los resultados obtenidos al añadir la pronunciación de palabras inglesas al diccionario de pronunciación se muestran en la tabla C.8. Los resultados del sistema final se muestran en la tabla C.9.

Por último, se muestran dos tablas resumen de todos los experimentos realizados, con el fin de ver la evolución de los resultados en función de las mejoras introducidas. En la tabla C.10 se muestran los resultados obtenidos y en la tabla C.11 se muestra la distribución de errores.

20k	descripción		título	
	WER	**MAP**	**WER**	**MAP**
Locutor 1	23,8 %	0,3123	35,7 %	0,2848
Locutor 2	18,4 %	0,3286	20,8 %	0,3029
Locutor 3	22,9 %	0,2930	18,8 %	0,3077
Locutor 4	25,0 %	0,2892	24,2 %	0,3031
Locutor 5	26,3 %	0,3248	26,1 %	0,3079
Locutor 6	35,6 %	0,2393	29,5 %	0,3076
Locutor 7	16,5 %	0,3217	19,3 %	0,3145
Locutor 8	23,2 %	0,2888	26,6 %	0,2931
Locutor 9	24,3 %	0,3003	23,7 %	0,2935
Locutor 10	27,5 %	0,3313	27,5 %	0,2822
Media	24,3 %	0,3029	25,2 %	0,2997

Tabla C.1: Resultados del sistema de referencia para preguntas de longitud media (descripción) y preguntas cortas (título). Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando un vocabulario de 20.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

40k	descripción		título	
	WER	**MAP**	**WER**	**MAP**
Locutor 1	18,5 %	0,3472	23,2 %	0,3273
Locutor 2	12,4 %	0,3580	10,6 %	0,3439
Locutor 3	17,4 %	0,3394	11,1 %	0,3471
Locutor 4	18,9 %	0,3243	15,9 %	0,3399
Locutor 5	21,6 %	0,3623	17,4 %	0,3324
Locutor 6	29,0 %	0,2852	19,8 %	0,3571
Locutor 7	9,2 %	0,3660	11,6 %	0,3466
Locutor 8	17,6 %	0,3173	15,5 %	0,3246
Locutor 9	20,1 %	0,3327	15,5 %	0,3207
Locutor 10	23,4 %	0,3577	20,8 %	0,3172
Media	18,8 %	0,3390	16,1 %	0,3357

Tabla C.2: Resultados del sistema de referencia para preguntas de longitud media (descripción) y preguntas cortas (título). Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando un vocabulario de 40.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

60k	descripción		título	
	WER	MAP	WER	MAP
Locutor 1	17,5 %	0,3699	23,2 %	0,3266
Locutor 2	13,0 %	0,3751	10,1 %	0,3501
Locutor 3	16,6 %	0,3603	10,6 %	0,3540
Locutor 4	19,0 %	0,3385	15,0 %	0,3459
Locutor 5	20,7 %	0,3636	15,9 %	0,3406
Locutor 6	28,0 %	0,3084	17,9 %	0,3694
Locutor 7	8,8 %	0,3671	10,1 %	0,3515
Locutor 8	17,2 %	0,3388	15,0 %	0,3310
Locutor 9	20,3 %	0,3330	14,0 %	0,3276
Locutor 10	22,1 %	0,3848	18,8 %	0,3443
Media	18,3 %	0,3540	15,1 %	0,3441

Tabla C.3: Resultados del sistema de referencia para preguntas de longitud media (descripción) y preguntas cortas (título). Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando un vocabulario de 60.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

20k	Ad. LM		Ad. Voc.		Ad. LM y Voc.	
	WER	MAP	WER	MAP	WER	MAP
Locutor 1	23,6 %	0,3082	17,0 %	0,3778	18,4 %	0,3783
Locutor 2	18,0 %	0,3276	12,8 %	0,3479	11,9 %	0,3809
Locutor 3	24,0 %	0,3055	16,2 %	0,3742	18,3 %	0,3854
Locutor 4	23,1 %	0,3042	19,2 %	0,3330	19,0 %	0,3484
Locutor 5	26,2 %	0,3230	20,9 %	0,3815	20,6 %	0,3785
Locutor 6	35,9 %	0,2401	30,5 %	0,2752	33,0 %	0,2714
Locutor 7	15,2 %	0,3213	9,8 %	0,3651	9,3 %	0,3717
Locutor 8	23,9 %	0,3007	16,6 %	0,3227	18,4 %	0,3299
Locutor 9	23,2 %	0,2997	19,7 %	0,3295	19,2 %	0,3397
Locutor 10	27,7 %	0,3301	23,0 %	0,3597	23,5 %	0,3783
Media	24,1 %	0,3060	18,6 %	0,3467	19,1 %	0,3563

Tabla C.4: Resultados obtenidos al realizar adaptación del LM, adaptación de vocabulario y adaptación del LM y vocabulario, para preguntas de longitud media. Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando un vocabulario de 20.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

40k	Ad. LM		Ad. Voc.		Ad. LM y Voc.	
	WER	MAP	WER	MAP	WER	MAP
Locutor 1	17,5 %	0,3487	16,3 %	0,3799	16,3 %	0,3729
Locutor 2	11,5 %	0,3539	11,6 %	0,3759	10,2 %	0,3991
Locutor 3	19,0 %	0,3507	16,0 %	0,3632	17,2 %	0,3746
Locutor 4	16,5 %	0,3262	17,6 %	0,3505	15,6 %	0,3621
Locutor 5	19,8 %	0,3576	19,7 %	0,3856	17,8 %	0,3866
Locutor 6	28,7 %	0,2768	28,5 %	0,2934	28,1 %	0,2882
Locutor 7	8,2 %	0,3720	8,4 %	0,3805	7,2 %	0,3903
Locutor 8	18,3 %	0,3258	16,7 %	0,3249	17,5 %	0,3347
Locutor 9	18,9 %	0,3356	20,1 %	0,3258	19,0 %	0,3409
Locutor 10	22,9 %	0,3641	21,7 %	0,3874	22,2 %	0,3937
Media	18,1 %	0,3411	17,7 %	0,3567	17,1 %	0,3643

Tabla C.5: Resultados obtenidos al realizar adaptación del LM, adaptación de vocabulario y adaptación del LM y vocabulario, para preguntas de longitud media. Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando un vocabulario de 40.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

60k	Ad. LM		Ad. Voc.		Ad. LM y Voc.	
	WER	MAP	WER	MAP	WER	MAP
Locutor 1	16,5 %	0,3712	16,5 %	0,3935	15,8 %	0,3920
Locutor 2	10,6 %	0,3769	12,5 %	0,3762	10,5 %	0,3984
Locutor 3	17,4 %	0,3716	16,0 %	0,3814	16,2 %	0,3931
Locutor 4	15,1 %	0,3405	18,5 %	0,3525	14,6 %	0,3679
Locutor 5	19,9 %	0,3649	19,7 %	0,3823	18,6 %	0,3858
Locutor 6	27,5 %	0,3022	28,0 %	0,3109	27,2 %	0,3061
Locutor 7	7,9 %	0,3731	8,3 %	0,3777	7,2 %	0,3746
Locutor 8	17,6 %	0,3508	17,5 %	0,3431	16,9 %	0,3564
Locutor 9	19,0 %	0,3374	20,6 %	0,3247	18,9 %	0,3349
Locutor 10	22,0 %	0,3906	21,3 %	0,4062	21,5 %	0,4122
Media	17,3 %	0,3579	17,9 %	0,3649	16,7 %	0,3721

Tabla C.6: Resultados obtenidos al realizar adaptación del LM, adaptación de vocabulario y adaptación del LM y vocabulario, para preguntas de longitud media. Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando un vocabulario de 60.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

	20k		40k		60k	
	WER	**MAP**	**WER**	**MAP**	**WER**	**MAP**
Locutor 1	23,8 %	0,3527	18,5 %	0,3755	17,5 %	0,3872
Locutor 2	18,4 %	0,3620	12,4 %	0,3935	13,0 %	0,4132
Locutor 3	22,9 %	0,3266	17,4 %	0,3762	16,6 %	0,3958
Locutor 4	25,0 %	0,3217	18,9 %	0,3599	19,0 %	0,3728
Locutor 5	26,3 %	0,3548	21,6 %	0,3944	20,7 %	0,3946
Locutor 6	35,6 %	0,2860	29,0 %	0,3194	28,0 %	0,3437
Locutor 7	16,5 %	0,3526	9,2 %	0,4001	8,8 %	0,3918
Locutor 8	23,2 %	0,3241	17,6 %	0,3648	17,2 %	0,3889
Locutor 9	24,3 %	0,3286	20,1 %	0,3643	20,3 %	0,3729
Locutor 10	27,5 %	0,3753	23,4 %	0,4045	22,1 %	0,4308
Media	24,3 %	0,3384	18,8 %	0,3753	18,3 %	0,3892

Tabla C.7: Resultados obtenidos al emplear realimentación por pseudo-relevancia, para preguntas de longitud media. Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando vocabularios de 20.000, 40.000 y 60.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

	20k		40k		60k	
	WER	**MAP**	**WER**	**MAP**	**WER**	**MAP**
Locutor 1	23,0 %	0,3179	18,4 %	0,3393	17,1 %	0,3625
Locutor 2	18,3 %	0,3253	11,7 %	0,3667	11,6 %	0,3873
Locutor 3	23,4 %	0,2954	18,0 %	0,3419	17,1 %	0,3644
Locutor 4	25,2 %	0,2971	18,5 %	0,3307	18,6 %	0,3394
Locutor 5	27,2 %	0,3232	21,8 %	0,3613	20,8 %	0,3585
Locutor 6	36,4 %	0,2494	30,0 %	0,2871	28,5 %	0,3091
Locutor 7	16,5 %	0,3155	9,1 %	0,3602	8,4 %	0,3613
Locutor 8	23,0 %	0,2869	17,8 %	0,3161	16,9 %	0,3354
Locutor 9	24,5 %	0,3036	19,8 %	0,3450	20,7 %	0,3460
Locutor 10	27,8 %	0,3360	23,6 %	0,3677	22,5 %	0,3855
Media	24,5 %	0,3050	18,9 %	0,3416	18,2 %	0,3549

Tabla C.8: Resultados obtenidos al añadir la pronunciación de palabras inglesas al diccionario de pronunciación, para preguntas de longitud media. Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando vocabularios de 20.000, 40.000 y 60.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

	20k		**40k**		**60k**	
	WER	**MAP**	**WER**	**MAP**	**WER**	**MAP**
Locutor 1	18,5 %	0,3770	16,3 %	0,3725	14,9 %	0,3833
Locutor 2	11,2 %	0,4354	8,4 %	0,4605	7,3 %	0,4614
Locutor 3	18,5 %	0,4094	15,8 %	0,4194	16,1 %	0,4357
Locutor 4	18,6 %	0,4029	14,9 %	0,4163	15,1 %	0,4004
Locutor 5	21,6 %	0,4080	19,7 %	0,4123	19,0 %	0,4109
Locutor 6	34,0 %	0,3170	29,0 %	0,3257	28,0 %	0,3418
Locutor 7	9,7 %	0,4147	8,2 %	0,4315	7,5 %	0,4306
Locutor 8	19,0 %	0,3928	18,0 %	0,4070	17,0 %	0,4248
Locutor 9	19,4 %	0,3931	19,5 %	0,3768	19,2 %	0,3743
Locutor 10	25,2 %	0,4160	21,2 %	0,4441	21,1 %	0,4591
Media	19,6 %	0,3966	17,1 %	0,4066	16,5 %	0,4122

Tabla C.9: Resultados del sistema final, para preguntas de longitud media. Se muestra los resultados obtenidos para cada uno de los 10 locutores empleando vocabularios de 20.000, 40.000 y 60.000 palabras. (WER: tasa de error de palabra; MAP: precisión media promediada).

		OOV	WER	MAP
20k	Referencia	7,15 %	24,3 %	0,3029
	Ad. LM	7,15 %	24,1 %	0,3060
	Ad. Voc.	3,38 %	18,6 %	0,3467
	Ad. LM y Voc.	3,38 %	19,1 %	0,3563
	Con realimenta.	7,15 %	24,3 %	0,3384
	Pron. inglés	7,15 %	24,5 %	0,3050
	Sistema final	3,52 %	19,6 %	0,3966
40k	Referencia	2,81 %	18,8 %	0,3390
	Ad. LM	2,81 %	18,1 %	0,3411
	Ad. Voc.	1,88 %	17,7 %	0,3567
	Ad. LM y Voc.	1,88 %	17,1 %	0,3643
	Con realimenta.	2,81 %	18,8 %	0,3753
	Pron. inglés	2,81 %	18,9 %	0,3416
	Sistema final	2,13 %	17,1 %	0,4066
60k	Referencia	2,17 %	18,3 %	0,3540
	Ad. LM	2,17 %	17,3 %	0,3579
	Ad. Voc.	1,62 %	17,9 %	0,3649
	Ad. LM y Voc.	1,62 %	16,7 %	0,3721
	Con realimenta.	2,17 %	18,3 %	0,3892
	Pron. inglés	2,17 %	18,2 %	0,3549
	Sistema final	1,63 %	16,5 %	0,4122

Tabla C.10: Resultados de todos los experimentos realizados, para preguntas de longitud media. Se muestra la media sobre los 10 locutores de los resultados obtenidos empleando vocabularios de 20.000, 40.000 y 60.000 palabras. (OOV: tasa de palabras fuera del vocabulario; WER: tasa de error de palabra; MAP: precisión media promediada).

		Tipo I	**Tipo II**	**Tipo III**	**Total**
20k	Referencia	115 (68,86 %)	25 (14,97 %)	27 (16,17 %)	167
	Ad. LM	115 (69,70 %)	24 (14,55 %)	26 (15,76 %)	165
	Ad. Voc.	34 (33,66 %)	33 (32,67 %)	34 (33,66 %)	101
	Ad. LM y Voc.	34 (35,79 %)	32 (33,68 %)	29 (30,53 %)	95
	Con realimenta.	118 (69,41 %)	22 (12,94 %)	30 (17,65 %)	170
	Pron. inglés	119 (72,56 %)	12 (7,32 %)	33 (20,12 %)	164
	Sistema final	33 (36,67 %)	18 (20,00 %)	39 (43,33 %)	90
40k	Referencia	39 (35,14 %)	30 (27,03 %)	42 (37,84 %)	111
	Ad. LM	40 (37,38 %)	30 (28,04 %)	37 (34,58 %)	107
	Ad. Voc.	20 (20,41 %)	32 (32,65 %)	46 (46,94 %)	98
	Ad. LM y Voc.	20 (22,22 %)	33 (36,67 %)	37 (41,11 %)	90
	Con realimenta.	39 (36,45 %)	24 (22,43 %)	44 (41,12 %)	107
	Pron. inglés	40 (39,60 %)	14 (13,86 %)	47 (46,53 %)	101
	Sistema final	18 (23,68 %)	16 (21,05 %)	42 (55,26 %)	76
60k	Referencia	20 (20,41 %)	28 (28,57 %)	50 (51,02 %)	98
	Ad. LM	20 (21,28 %)	28 (29,79 %)	46 (48,94 %)	94
	Ad. Voc.	15 (16,13 %)	30 (32,26 %)	48 (51,61 %)	93
	Ad. LM y Voc.	15 (17,05 %)	31 (35,23 %)	42 (47,73 %)	88
	Con realimenta.	20 (21,05 %)	23 (24,21 %)	52 (54,74 %)	95
	Pron. inglés	20 (21,28 %)	14 (14,89 %)	60 (63,83 %)	94
	Sistema final	13 (16,88 %)	16 (20,78 %)	48 (62,34 %)	77

Tabla C.11: Distribución de errores de todos los experimentos realizados, para preguntas de longitud media. Se muestra el número y porcentaje de cada tipo de error para vocabularios de 20.000, 40.000 y 60.000 palabras. (Tipo I: causado por una palabra OOV; Tipo II: producido por una palabra en otro idioma; Tipo III: cualquier otro error de reconocimiento del habla).

Bibliografía

[1] S. Abiteboul. Querying Semi-Structured Data. En *International Conference on Database Theory (ICDT)*, páginas 1–18. 1997.

[2] G. Adda, M. Adda-Decker, J.-L. Gauvain, y L. Lamel. Text Normalization and Speech Recognition in French. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 2711–2714. 1997.

[3] J. Adiego, P. de la Fuente, J. Vegas, y M. A. Villarroel. System for Compressing and Retrieving Structured Documents. *UPGRADE - The European Journal for the Informatics Professional*, 3(3):62–69, junio 2002.

[4] T. Akiba y H. Abe. Exploiting Passage Retrieval for N-best Rescoring of Spoken Questions. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 65–68. 2005.

[5] T. Akiba, A. Fujii, T. Ishikawa, y K. Itou. Experiments on Web Retrieval Driven by Spontaneously Spoken Queries. En *NTCIR Workshop*. 2004.

[6] T. Akiba, A. Fujii, y K. Itou. Collecting Spontaneously Spoken Queries for Information Retrieval. En *International Conference on Language Resources and Evaluation (LREC)*, páginas 1439–1442. 2004.

[7] T. Akiba, A. Fujii, y K. Itou. Effects of Language Modeling on Speech-driven Question Answering. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 1053–1056. 2004.

[8] M. Araki, T. Ono, K. Ueda, T. Nishimoto, y Y. Nimi. An Automatic Dialogue System Generator from the Internet Information Contents. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1743–1746. 2001.

[9] H. Asoh, T. Matsui, J. Fry, F. Asano, y S. Hayamizu. A Spoken Dialog System for a Mobile Office Robot. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1139–1142. 1999.

[10] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, y S. W. Kuo. Experiments in Spoken Queries for Document Retrieval. En *European Con-*

ference on Speech Communication and Technology (Eurospeech), páginas 1323–1326. 1997.

[11] N. Bel, J. Caminero, L. Hernández, M. Marimón, J. F. Morlesín, J. M. Otero, J. Relaño, M. C. Rodríguez, P. M. Ruz, y D. Tapias. Design and Evaluation of a SLDS for E-Mail Access Through the Telephone. En *International Conference on Language Resources and Evaluation (LREC)*, páginas 537–544. 2002.

[12] J. R. Bellegarda. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93–108, enero 2004.

[13] T. Berners-Lee, J. Hendler, y O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, mayo 2001.

[14] N. O. Bernsen, H. Dybkjær, y L. Dybkjær. *Designing Interactive Speech Systems. From First Ideas to User Testing*. Springer Verlag, 1998.

[15] N. O. Bernsen y L. Dybkjær. A Methodology for Evaluating Spoken Language Dialogue Systems and Their Components. En *International Conference on Language Resources and Evaluation (LREC)*, páginas 183–188. 2000.

[16] D. Bohus, A. Raux, T. K. Harris, M. Eskenazi, y A. I. Rudnicky. Olympus: an open-source framework for conversational spoken language interface research. En *Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL)*, páginas 32–39. 2007.

[17] M. Braschler y C. Peters. CLEF Methodology and Metrics. En *Workshop of the Cross-Language Evaluation Forum (CLEF)*, páginas 394–404. 2001.

[18] M. K. Brown, S. C. Glinski, B. P. Goldman, y B. C. Schmult. PhoneBrowser: A Web-Content-Programmable Speech Processing Platform. En *W3C Workshop on Voice Browsers*. 1998.

[19] D. C. Burnett, M. R. Walker, y A. Hunt. Speech Synthesis Markup Language (SSML) Version 1.0. W3C Recommendation, septiembre 2004. <http://www.w3.org/TR/2004/REC-speech-synthesis-20040907/> [12/07/2005].

[20] O. Buyukkokten, O. Kaljuvee, H. Garcia-Molina, A. Paepcke, y T. Winograd. Efficient Web Browsing on Handheld Devices Using Page and Form Summarization. *ACM Transactions on Information Systems*, 20(1):82–115, enero 2002.

[21] B. Caldwell, M. Cooper, L. G. Reid, y G. Vanderheiden. Web Content Accessibility Guidelines (WCAG) 2.0. W3C Recommendation, diciembre 2008. <http://www.w3.org/TR/2008/REC-WCAG20-20081211/> [17/01/2009].

[22] C. Carpineto, G. Romano, y V. Giannini. Improving Retrieval Feedback with Multiple Term-Ranking Function Combination. *ACM Transactions on Information Systems*, 20(3):259–290, julio 2002.

[23] M. J. Castro, S. España, A. Marzal, y I. Salvador. Transcriptor ortográfico-fonético para el castellano. En *Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, páginas 241–245. 2001.

[24] E. Chang, F. Seide, H. M. Meng, Z. Chen, Y. Shi, y Y.-C. Li. A System for Spoken Query Information Retrieval on Mobile Devices. *IEEE Transactions on Speech and Audio Processing*, 10(8):531–541, noviembre 2002.

[25] A. Charfuelán Oliva. *Técnicas de Evaluación de Sistemas de Diálogo*. Tesis Doctoral, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 2003.

[26] S. Chen y J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Informe Técnico TR-10-98, Harvard University, 1998.

[27] A. Cheyer y D. Martin. The Open Agent Architecture. *Autonomous Agents and Multi-Agent Systems*, 4(1-2):143–148, marzo 2001.

[28] J. Clark. XSL Transformations (XSLT) Version 1.0. W3C Recommendation, noviembre 1999. <http://www.w3.org/TR/1999/REC-xslt-19991116> [5/09/2004].

[29] J. Clark y S. DeRose. XML Path Language (XPath) Version 1.0. W3C Recommendation, noviembre 1999. <http://www.w3.org/TR/1999/REC-xpath-19991116> [5/09/2004].

[30] F. Crestani. Spoken query processing for interactive information retrieval. *Data & Knowledge Engineering*, 41(1):105–124, abril 2002.

[31] F. Crestani. Vocal Access to a Newspaper Archive: Assessing the Limitations of Current Voice Information Access Technology. *Journal of Intelligent Information Systems*, 20(2):161–180, marzo 2003.

[32] E. Dacosta Guisado y I. Gómez Renedo. *diGoo: Interacción multimodal con Google*. Proyecto Fin de Carrera, Departamento de Informática, Universidad de Valladolid, septiembre 2007.

[33] R. De Mori, editor. *Spoken Dialogues with Computers*. Academic Press, 1998.

[34] E. den Os, L. Boves, L. Lamel, y P. Baggia. Overview of the ARISE Project. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1527–1530. 1999.

[35] DMG Consulting LLC. 2008 Worldwide Interactive Voice Response Trends and Market Share Report, 2008.

[36] S. Dobrisek, J. Gros, B. Vesnicer, F. Mihelic, y N. Pavesic. A Voice-Driven Web Browser for Blind People. En *International Conference on Text, Speech and Dialogue (TSD)*, páginas 453–459. 2002.

[37] P. J. Durston, M. Farrell, D. Attwater, J. Allen, H.-K. J. Kuo, M. Afify, E. Fosler-Lussier, y C.-H. Lee. OASIS Natural Language Call Steering Trial. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1323–1326. 2001.

[38] L. Dybkjær, N. O. Bernsen, y W. Minker. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1-2):33–54, junio 2004.

[39] J. Díaz-Verdejo, A. M. Peinado, A. J. Rubio, E. Segarra, N. Prieto, y F. Casacuberta. ALBAYZIN: a Task-Oriented Spanish Speech Corpus. En *International Conference on Language Resources and Evaluation (LREC)*, páginas 497–501. 1998.

[40] D. Escudero-Mancebo y V. Cardeñoso-Payo. Applying data mining techniques to corpus based prosodic modeling. *Speech Communication*, 49(3):213–229, marzo 2007.

[41] D. Fallows. The Internet and Daily Life. Pew Internet and American Life Project, 2004. <`http://www.pewinternet.org/pdfs/PIP_Internet_and_Daily_Life.pdf`> [17/09/2008].

[42] J. Feng, S. Reddy, y M. Saraçlar. WebTalk: Mining Websites for Interactively Answering Questions. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 2485–2488. 2005.

[43] A. Franz y B. Milch. Searching the Web by Voice. En *International Conference on Computational Linguistics (COLING)*, páginas 1213–1217. 2002.

[44] N. M. Fraser y G. N. Gilbert. Simulating speech systems. *Computer Speech and Language*, 5(1):81–99, enero 1991.

[45] J. Freire, B. Kumar, y D. Lieuwen. WebViews: Accessing Personalized Web Content and Services. En *International World Wide Web Conference*, páginas 576–586. 2001.

[46] A. Fujii y K. Itou. Building a Test Collection for Speech-Driven Web Retrieval. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1153–1156. 2003.

[47] A. Fujii, K. Itou, y T. Ishikawa. Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition. En *SIGIR Workshop on Information Retrieval Techniques for Speech Applications*, páginas 94–104. 2001.

[48] A. Fujii, K. Itou, y T. Ishikawa. A Method for Open-Vocabulary Speech-Driven Text Retrieval. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 188–195. 2002.

[49] L. García Muñoz y L. Rodero Merino. *Desarrollo de Aplicaciones Vocales en VoiceXML*. Proyecto Fin de Carrera, Departamento de Informática, Universidad de Valladolid, septiembre 2000.

[50] J. S. Garofolo, C. G. P. Auzanne, y E. M. Voorhees. The TREC Spoken Document Retrieval Track: A Success Story. En *Text Retrieval Conference (TREC)*, páginas 107–129. 1999.

[51] D. Gibbon, R. Moore, y R. Winski, editores. *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, 1997.

[52] M. Gilbert y J. Feng. Speech and Language Processing over the Web. *IEEE Signal Processing Magazine*, 25(3):18–28, mayo 2008.

[53] J. Glass, E. Weinstein, S. Cyphers, J. Polifroni, G. Chung, y M. Nakano. A Framework for Developing Conversational User Interfaces. En *International Conference on Computer-Aided Design of User Interfaces (CADUI)*, páginas 354–365. 2004.

[54] C. González-Ferreras y V. Cardeñoso-Payo. Building Voice Applications from Web Content. En *International Conference on Text, Speech and Dialogue (TSD)*, páginas 587–594. 2004.

[55] C. González-Ferreras y V. Cardeñoso-Payo. Development and Evaluation of a Spoken Dialog System to Access a Newspaper Web Site. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 857–860. 2005.

[56] C. González-Ferreras y V. Cardeñoso-Payo. Experiments in Speech Driven Information Retrieval for Spanish Language. En *Jornadas en Tecnología del Habla*, páginas 149–153. 2006.

[57] C. González-Ferreras y V. Cardeñoso-Payo. A System for Speech Driven Information Retrieval. En *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, páginas 624–628. 2007.

[58] C. González-Ferreras y V. Cardeñoso-Payo. Dynamic Adaptation of Language Models in Speech Driven Information Retrieval. En *International Conference on Text, Speech and Dialogue (TSD)*, páginas 214–221. 2007.

[59] C. González-Ferreras, V. Cardeñoso-Payo, y E. Sanchis Arnal. Experiments in Speech Driven Question Answering. En *IEEE Workshop on Spoken Language Technology (SLT)*, páginas 85–88. 2008.

[60] C. González-Ferreras, D. Escudero-Mancebo, y V. Cardeñoso-Payo. From HTML to VoiceXML: A First Approach. En *International Conference on Text, Speech and Dialogue (TSD)*, páginas 441–444. 2002.

[61] C. González-Ferreras, R. San-Segundo-Hernández, y V. Cardeñoso-Payo. A Spoken Dialog System to Access a Newspaper Web Site. En *Berliner XML Tage*, páginas 266–275. 2004.

[62] S. Goose y S. Djennane. WIRE3: Driving Around the Information Super-Highway. *Personal and Ubiquitous Computing*, 6(3):164–175, mayo 2002.

[63] S. Goose, M. Newman, C. Schmidt, y L. Hue. Enhancing Web Accessibility Via the Vox Portal and a Web Hosted Dynamic HTML<->VoxML Converter. En *International World Wide Web Conference*, páginas 583–592. 2000.

[64] A. L. Gorin, G. Riccardi, y J. H. Wright. How may I help you? *Speech Communication*, 23(1-2):113–127, octubre 1997.

[65] D. Griol, L. F. Hurtado, E. Segarra, y E. Sanchis. A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, 50(8-9):666–682, agosto-septiembre 2008.

[66] S. W. Hamerich, R. de Córdoba, V. Schless, L. F. d'Haro, B. Kladis, V. Schubert, O. Kocsis, S. Igel, y J. M. Pardo. The GEMINI Platform: Semi-Automatic Generation of Dialogue Applications. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 2629–2632. 2004.

[67] S. Harabagiu, D. Moldovan, y J. Picone. Open-Domain Voice-Activated Question Answering. En *International Conference on Computational Linguistics (COLING)*, páginas 321–327. 2002.

[68] C. T. Hemphill y P. R. Thrift. Surfing the Web by Voice. En *ACM International Conference on Multimedia*, páginas 215–222. 1995.

[69] K. S. Hone y R. Graham. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering*, 6(3-4):287–303, septiembre 2000.

[70] C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, y S. Furui. Deriving Disambiguous Queries in a Spoken Interactive ODQA System. En *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, tomo 1, páginas 624–627. 2003.

[71] M. Hori, G. Kondoh, K. Ono, S. Hirose, y S. Singhal. Annotation-based Web content transcoding. *Computer Networks*, 33(1-6):197–211, junio 2000.

[72] D. House. *Spoken-Language Access to Multimedia (SLAM): a Multimodal Interface to the World-Wide-Web*. Proyecto Fin de Carrera, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, abril 1995.

[73] X. Huang, A. Acero, y H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[74] International Organization for Standardization (ISO). International Standard ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, 1998.

[75] International Telecommunication Union (ITU). ITU-T Recommendation P.851: Subjective quality evaluation of telephone services based on spoken dialogue systems, 2003.

[76] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, 1998.

[77] M. Jeong, B. Kim, y G. G. Lee. Using Higher-level Linguistic Knowledge for Speech Recognition Error Correction in a Spoken Q/A Dialog. En *Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL)*, páginas 48–55. 2004.

[78] D. Kim, S. Furui, y H. Isozaki. Language Models and Dialogue Strategy for a Voice QA System. En *International Congress on Acoustics (ICA)*, páginas 3705–3708. 2004.

[79] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, y J. S. Teixeira. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record*, 31(2):84–93, junio 2002.

[80] M. Lamb y B. Horowitz. Guidelines for a VoiceXML Solution Using WebSphere Transcoding Publisher. IBM White Paper, 2001. <ftp://ftp.software.ibm.com/software/wtp/info/VxmlTranscodingGuide.pdf> [20/06/2003].

[81] L. Lamel, S. Rosset, J. L. Gauvain, S. Bennacef, M. Garnier-Rizet, y B. Prouts. The LIMSI ARISE System. *Speech Communication*, 31(4):339–353, agosto 2000.

[82] L. B. Larsen. Assessment of Spoken Dialogue System Usability - What are We really Measuring? En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1945–1948. 2003.

[83] L. B. Larsen. *On the Usability of Spoken Dialogue Systems*. Tesis Doctoral, The Faculty of Engineering and Science, Alborg University, 2003.

[84] J. A. Larson. VoiceXML: Industry Perspectives and Business Opportunities. En *Applied Spoken Language Interaction in Distributed Environments (ASIDE)*. 2005.

[85] R. Lau, G. Flammia, C. Pao, y V. Zue. WebGALAXY - Integrating Spoken Language and Hypertext Navigation. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 883–886. 1997.

[86] L. Liu, C. Pu, y W. Han. XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources. En *International Conference on Data Engineering (ICDE)*, páginas 611–621. 2000.

[87] S. Love, R. T. Dutton, J. C. Foster, M. A. Jack, y F. W. M. Stentiford. Identifying Salient Usability Attributes for Automated Telephone Services. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 1307–1310. 1994.

[88] R. López-Cózar, A. de la Torre, J. C. Segura, y A. J. Rubio. Assessment of dialogue systems by means of a new simulation technique. *Speech Communication*, 40(3):387–407, mayo 2003.

[89] C. D. Manning, P. Raghavan, y H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[90] M. Matsushita, H. Nishizaki, S. Nakagawa, y T. Utsuro. Keyword Recognition and Extraction by Multiple-LVCSRs with 60,000 Words in Speech-driven WEB Retrieval Task. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 1625–1628. 2004.

[91] M. Matsushita, H. Nishizaki, T. Utsuro, Y. Kodama, y S. Nakagawa. Evaluating Multiple LVCSR Model Combination in NTCIR-3 Speech-Driven Web Retrieval Task. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1205–1208. 2003.

[92] C. Mayo Martín. *Estudio e Implementación de un Intérprete de VoiceXML*. Proyecto Fin de Carrera, Departamento de Informática, Universidad de Valladolid, septiembre 2001.

[93] S. McGlashan, D. C. Burnett, J. Carter, P. Danielsen, J. Ferrans, A. Hunt, B. Lucas, B. Porter, K. Rehor, y S. Tryphonas. Voice Extensible Markup Language (VoiceXML) Version 2.0. W3C Recommendation, marzo 2004. <http://www.w3.org/TR/2004/REC-voicexml20-20040316/> [10/07/2004].

[94] M. F. McTear. Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Computing Surveys*, 34(1):90–169, marzo 2002.

[95] W. Minker, U. Haiber, P. Heisterkamp, y S. Scheible. The SENECA spoken language dialogue system. *Speech Communication*, 43(1-2):89–102, junio 2004.

[96] T. Misu y T. Kawahara. Dialogue strategy to clarify user's queries for document retrieval system with speech interface. *Speech Communication*, 48(9):1137–1150, septiembre 2006.

[97] M. Mitra, A. Singhal, y C. Buckley. Improving Automatic Query Expansion. En *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, páginas 206–214. 1998.

[98] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, y C. Nadeu. ALBAYZIN Speech Database: Design of the Phonetic Corpus. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 175–178. 1993.

[99] A. Moreno-Daniel, S. Parthasarathy, B. H. Juang, y J. G. Wilpon. Spoken Query Processing for Information Retrieval. En *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, tomo 4, páginas 121–124. 2007.

[100] S. Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, 2005.

[101] M. Oshry, R. J. Auburn, P. Baggia, M. Bodell, D. Burke, D. C. Burnett, E. Candell, J. Carter, S. McGlashan, A. Lee, B. Porter, y K. Rehor. Voice Extensible Markup Language (VoiceXML) 2.1. W3C Recommendation, junio 2007. <http://www.w3.org/TR/2007/REC-voicexml21-20070619/> [17/11/2007].

[102] B. Pellom y K. Hacioglu. Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task. En *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, tomo 1, páginas 4–7. 2003.

[103] B. Pellom, W. Ward, J. Hansen, R. Cole, K. Hacioglu, J. Zhang, X. Yu, y S. Pradhan. University of Colorado Dialog Systems for Travel and Navigation. En *Human Language Technology Conference (HLT)*. 2001.

[104] R. Pieraccini y J. Huerta. Where do we go from here? Research and commercial spoken dialog systems. En *SIGdial Workshop on Discourse and Dialogue*, páginas 1–10. 2005.

[105] J. Polifroni, G. Chung, y S. Seneff. Towards the Automatic Generation of Mixed-Initiative Dialogue Systems from Web Content. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 193–196. 2003.

[106] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, febrero 1989.

[107] M. Rahim, G. Di Fabbrizio, C. Kamm, M. Walker, A. Pokrovsky, P. Ruscitti, E. Levin, S. Lee, A. Syrdal, y K. Schlosser. Voice-IF: A Mixed-Initiative Spoken Dialogue System for AT&T Conference Services. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1339–1342. 2001.

[108] L. Rodríguez Liñares, A. Cardenal López, C. García Mateo, D. Pérez-Piñar López, E. Rodríguez Banga, y X. Fernández Salgado. TelCorreo: A Bilingual E-mail Client over the Telephone. En *International Conference on Text, Speech and Dialogue (TSD)*, páginas 381–386. 2000.

[109] A. I. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, y A. Oh. Creating natural dialogs in the Carnegie Mellon Communicator system. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1531–1534. 1999.

[110] M. Ruiz Costa-jussà, J.-L. Gauvain, y O. Galibert. Normalización de textos y selección del vocabulario para estimar el modelo de lenguaje de un sistema de transcripción de noticias. En *Jornadas en Tecnología del Habla*, páginas 153–158. 2004.

[111] A. Sahuguet y F. Azavant. Building intelligent Web applications using lightweight wrappers. *Data & Knowledge Engineering*, 36(3):283–316, marzo 2001.

[112] G. Salton y C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[113] G. Salton y C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[114] G. Salton, A. Wong, y C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, noviembre 1975.

[115] R. San-Segundo, J. M. Montero, J. Colás, J. Gutiérrez, J. M. Ramos, y J. M. Pardo. Methodology for Dialogue Design in Telephone-Based Spoken Dialogue Systems: a Spanish Train Information System. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 2165–2168. 2001.

[116] E. Sanchis, D. Buscaldi, S. Grau, L. Hurtado, y D. Griol. Spoken QA based on a Passage Retrieval Engine. En *IEEE Workshop on Spoken Language Technology (SLT)*, páginas 62–65. 2006.

[117] J. Savoy. Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. En *Workshop of the Cross-Language Evaluation Forum (CLEF)*, páginas 27–43. 2001.

[118] B. N. Schilit, J. Trevor, D. M. Hilbert, y T. K. Koh. Web Interaction Using Very Small Internet Devices. *Computer*, 35(10):37–45, octubre 2002.

[119] E. Schofield y Z. Zheng. A speech interface for open-domain question-answering. En *Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 177–180. 2003.

[120] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, y V. Zue. GALAXY-II: A Reference Architecture for Conversational System Development. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 931–934. 1998.

[121] S. Seneff y J. Polifroni. Dialogue Management in the Mercury Flight Reservation System. En *ANLP-NAACL Workshop on Conversational Systems*, páginas 11–16. 2000.

[122] N. Shadbolt, T. Berners-Lee, y W. Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96–101, mayo-junio 2006.

[123] A. Singhal, J. Choi, D. Hindle, D. D. Lewis, y F. Pereira. AT&T at TREC-7. En *Text Retrieval Conference (TREC)*, páginas 239–251. 1999.

[124] R. W. Smith y D. R. Hipp. *Spoken Natural Language Dialog Systems: A Practical Approach*. Oxford University Press, 1994.

[125] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, y F. Seide. The Thoughtful Elephant: Strategies for Spoken Dialog Systems. *IEEE Transactions on Speech and Audio Processing*, 8(1):51–62, enero 2000.

[126] A. Stolcke. SRILM – An Extensible Language Modeling Toolkit. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 901–904. 2002.

[127] S. Sutton, R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. Massaro, y M. Cohen. Universal Speech Tools: the CSLU Toolkit. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 3221–3224. 1998.

[128] H. Takagi, S. Saito, K. Fukuda, y C. Asakawa. Analysis of Navigability of Web Applications for Improving Blind Usability. *ACM Transactions on Computer-Human Interaction*, 14(3), septiembre 2007.

[129] P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[130] R. Trias-Sanz y J. B. Mariño. Basurde[lite], a machine-driven dialogue system for accessing railway timetable information. En *International Conference on Spoken Language Processing (ICSLP)*, páginas 2685–2688. 2002.

[131] B. van Schooten, S. Rosset, O. Galibert, A. Max, R. op den Akker, y G. Illouz. Handling speech input in the Ritel QA dialogue system. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 126–129. 2007.

[132] B. Vesnicer, J. Zibert, S. Dobrisek, N. Pavesic, y F. Mihelic. A Voice-driven Web Browser for Blind People. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1301–1304. 2003.

[133] VoiceXML Forum. Voice eXtensible Markup Language (VoiceXML) Version 1.0. marzo 2000. <`http://www.voicexml.org/specs/VoiceXML-100.pdf`> [21/03/2002].

[134] M. Walker. An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email. *Journal of Artificial Intelligence Research*, 12:387–416, 2000.

[135] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, y S. Whittaker. DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 1371–1374. 2001.

[136] M. Walker, D. Hindle, J. Fromer, G. Di Fabbrizio, y C. Mestel. Evaluating Competing Agent Strategies for a Voice Email Agent. En *European Conference on Speech Communication and Technology (Eurospeech)*, páginas 2219–2222. 1997.

[137] M. Walker, L. Hirschman, y J. Aberdeen. Evaluation for DARPA Communicator Spoken Dialogue Systems. En *International Conference on Language Resources and Evaluation (LREC)*, páginas 735–741. 2000.

[138] M. Walker, D. J. Litman, C. A. Kamm, y A. Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. En *Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 271–280. 1997.

[139] J. C. Wells. SAMPA computer readable phonetic alphabet. En D. Gibbon, R. Moore, y R. Winski, editores, *Handbook of Standards and Resources for Spoken Language Systems*, páginas 684–732. Mouton de Gruyter, 1997. <http://www.phon.ucl.ac.uk/home/sampa> [16/02/2006].

[140] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, y L. Hetherington. JUPITER: A Telephone-Based Conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96, enero 2000.