Universitat Politècnica de València
Departament de Sistemes Informàtics i Computació

Modelado de lenguaje basado en categorías para Búsqueda de Respuesta dirigida por la Voz

Trabajo Fin de Máster
Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Presentada por: Joan Pastor Pellicer

Dirigida por: Dr. Lluís-Felip Hurtado
Departament de Sistemes Informàtics i Computació

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

La interacción conversacional hombre-máquina requiere de la utilización de sistemas avanzados con un alto grado de integración de diferentes tecnologías para poder manejar así una creciente cantidad de información. En general, los sistemas resultantes de la integración de sistemas de voz con sistemas de procesado de información no estructurada permiten un mejor acceso y uso posterior de dicha información. Para el desarrollo de aplicaciones para el mundo real, resulta interesante diseñar sistemas dirigidos por voz que proporcionen el acceso a la información a través de teléfonos móviles y tabletas digitales y otras interfaces con acceso a voz.

Los sistemas de Búsqueda de Respuesta (*Question Answering, QA*) permiten a los usuarios realizar preguntas formuladas en lenguaje natural y obtener la respuesta correcta sobre una colección de documentos no estructurada. La mayoría de los sistemas de QA aceptan sentencias (preguntas o consultas) escritas y correctamente formuladas como entrada, pero recientemente se ha incrementado el interés en la utilización de voz para la realización de consultas [Rosso et al., 2010, Mishra and Bangalore, 2010], así como el acceso a grandes repositorios de audio [Wang et al., 2008, Chelba et al., 2008].

Búsqueda de Respuesta Dirigida por la Voz (*Voice Activated Question Answering, VAQA*) es una de las diferentes formas de acceso a grandes repositorios de información para obtener información específica sobre algún dato en concreto.

Debido al interés de estas aplicaciones, se ha incluido la tarea sobre VAQA sobre diferentes idiomas en conferencias de evaluación de sistemas de recuperación de información

como son la competición CLEF (Cross-Language Evaluation Forum) [1] y TREC (Text REtrieval Conference) [2] [Turmo et al., 2009].

Para la realización de Búsqueda de Respuesta dirigida por la Voz sobre grandes repositorios, es necesario la utilización de un Sistema Automático de Reconocimiento de Voz (*Automatic Speech Recognizer, ASR*) capaz de trabajar con grandes vocabularios y proveer un modelado de lenguaje adecuado que caracterice los tipos de preguntas que se pueden realizar sobre el sistema. Además, es importante que el ASR pueda reconocer correctamente aquellas palabras específicas que tienen una gran influencia en el rendimiento global del sistema. En particular, el reconocimiento de Entidades Nombradas (NE) es una de los principales problemas a tener en cuenta en sistemas de extracción de información y aplicaciones de QA.

Las entidades nombradas son elementos claves en el proceso de búsqueda de respuesta [Chu-Carroll and Prager, 2007, Kubala et al., 1998], por lo tanto un mal reconocimiento de estas entidades afecta de forma importante en los resultados obtenidos por el sistema encargado de encontrar la respuesta. Diferentes estudios muestran que mejores resultados de reconocimiento de las Entidades Nombradas [Chu-Carroll and Prager, 2007] implican un mejor rendimiento del sistema VAQA. Por tanto, se precisa de una alta tasa de reconocimiento sobre este tipo de entidades. No obstante, esta tarea se vuelve realmente difícil cuando las NE provienen de un idioma diferente al del usuario, por ejemplo, si el idioma nativo del usuario es el castellano y quiere realizar una consulta sobre un personaje célebre americano utilizará una fonética diferente a la del castellano: "¿Quién fue George Washington?". Para tratar este tipo de problema el Sistema de Reconocimiento de Voz deberá incorporar mecanismos para considerar pronunciaciones específicas y alternativas sobre una misma palabra o acrónimo [Fujii et al., 2002, Stoyanchev et al., 2008, Wang et al., 2010].

En cuanto al modelado de lenguaje, los modelos utilizados deberán adaptarse tanto a las preguntas realizadas como al contenido del repositorio de documentos donde se encuentran las respuestas a las preguntas formuladas. Por tanto, deberán tomarse en cuenta las restricciones sintácticas que hay en ambos casos. Si se utiliza únicamente un corpus de propósito general, de gran tamaño y gran vocabulario, para el entrenamiento del modelo de lenguaje, entonces las ventajas de conocer las restricciones sintácticas específi-

[1] `http://www.clef-campaign.org`
[2] `http://trec.nist.gov`

cas de las posibles preguntas del usuario se pierden [Akiba et al., 2007, Kim et al., 2004, Harabagiu et al., 2002].

Otro aspecto importante a considerar, es conocer qué Entidades Nombradas se encuentran en la colección de documentos, esta tarea no es trivial dado que estas no están anotadas de ninguna forma, por tanto es necesario la utilización de herramientas y mecanismos capaces de anotar automáticamente dichas entidades.

En el presente trabajo se presenta una aproximación para la realización de QA con voz, tomando consultas de voz en vez de preguntas correctamente escritas como entrada del sistema. Se propone la utilización de un modelo de lenguaje aprendido a partir de un conjunto de preguntas de entrenamiento, que contienen unas restricciones sintácticas diferentes de las frases puramente declarativas que se encuentran en los repositorios de información. Parte del trabajo se ha centrado principalmente en el impacto del incremento de la cantidad de Entidades Nombradas en el vocabulario del modelo de lenguaje, para poder así incrementar la cobertura de las posibles entidades demandadas por el usuario.

1.1. Búsqueda de Respuesta dirigida por la Voz

Un sistema Búsqueda de Respuesta dirigida por la Voz consiste en un sistema QA donde la interacción con el usuario se realiza mediante la utilización de un ASR como entrada del mismo. Además, es posible utilizar un sintetizador de voz (*Text-to-Speech, TTS*) como salida del sistema en vez de texto.

La pregunta es formulada por el usuario y posteriormente es convertida a texto mediante el sistema ASR. La pregunta obtenida por el reconocedor se envía entonces al módulo de análisis de pregunta, cuyo propósito es extraer el tipo de pregunta o cuestión (por ejemplo, el usuario pregunta por un nombre o una fecha), así como otras características que serán usadas en la fase de extracción de la pregunta. La fase de recuperación de pasajes extrae la información demandada de una colección textual de pasajes que el sistema considera relevantes para la pregunta. El módulo de extracción de la pregunta examina dichos pasajes y, usando la información obtenida durante la fase de análisis de la pregunta, selecciona la pregunta que será presentada finalmente al usuario.

La figura 4.1 muestra los principales módulos de un sistema de Búsqueda de Respuesta dirigida por la Voz.

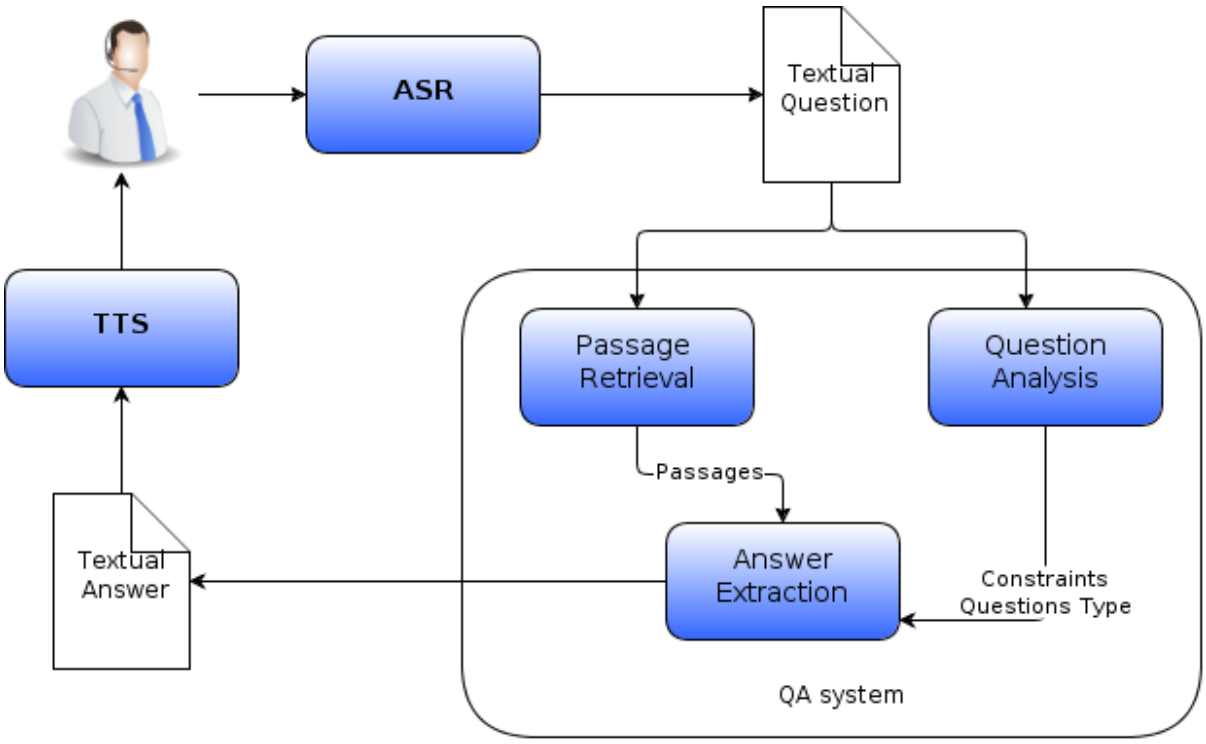

Figura 1.1: Módulos del Sistema de Búsqueda de respuesta dirigida por la Voz

La fase de análisis de la pregunta, es una parte crítica del sistema QA. De hecho, se estima que alrededor del 40 % de los errores en QA derivan directamente de un error en esta fase, donde aproximadamente un 33 % de este error es debido a la identificación del tipo de pregunta. La fase de de extracción de la respuesta es también una fuente importante de errores, con un 20 % de los errores en QA [Moldovan et al., 2003].

Por esta razón, se ha focalizado el trabajo realizado, en el estudio de los efectos en términos de Recuperación de Pasajes, donde los efectos de un mal reconocimiento de las cuestiones se reflejan directamente en el ranking de los pasajes obtenidos. En cambio, estos efectos no se pueden detectar usando el sistema de QA completo, donde la mayor parte del error obtenido proviene de las fases de análisis de la pregunta y extracción de la respuesta.

La combinación de un sistema de reconocimiento automático del habla conjuntamente con un sistema de QA degrada las prestaciones de este. En [Harabagiu et al., 2002] se muestra la utilización de un ASR para las frases utilizadas en TREC y se muestra la notable diferencia de prestaciones al utilizar una entrada reconocida automáticamente.

Además, en estudios realizados anteriormente [Sanchis et al., 2006, González et al., 2008] se muestra la importancia del reconocimiento de Entidades Nombradas, donde las presta-

ciones del sistema se mantienen incluso hasta con un 25 % de error de reconocimiento (WER) siempre y cuando se sigan reconociendo bien estas entidades. Con un error de reconocimiento mayor al 30 % las prestaciones del sistema se degradan de forma muy notable.

Como en cualquier sistema dirigido por la voz existen ciertos factores que no están presentes en el lenguaje escrito: pausas, dubitaciones, tos, etc... Tampoco existen símbolos de puntuación, lo cual puede ser un problema para el módulo de recuperación de información que necesita tener sentencias delimitadas. No se encuentran tampoco, fronteras claramente delimitadas entre chunks, es el caso de los números que pueden ser transcritos de diferentes formas y existe cierta ambigüedad que no se presenta en el caso del lenguaje escrito (por ejemplo: "*mil veinte*" se puede transcribir como 1,020 o 1,000 20).

En cuanto a los sistemas de reconocimiento automático de voz, existen varios elementos que son altamente dependientes de la tarea. Uno de ellos es el vocabulario necesario para realizar la tarea así como el modelo de lenguaje utilizado. El vocabulario utilizado para la tarea de QA (reconocimiento de preguntas sobre un determinado corpus) difiere del vocabulario utilizado para un tarea de reconocimiento de propósito general. No obstante, existe un problema al reconocer palabras en las preguntas del usuario que no están incluidas en el vocabulario específico de la tarea (Out-of-Vocabulary words, OOV).

1.2. Búsqueda de respuesta

En esta sección se muestran las características principales de los sistemas de Búsqueda de Respuesta. Como se ha comentado en el apartado anterior, la funcionalidad básica del los sistemas de QA es dar la capacidad al usuario de realizar preguntas en lenguaje natural sobre una colección de documentos no estructurados. A continuación se describen las fases principales, previamente introducidas, en las que se divide la tarea de QA.

1.2.1. Fase de Análisis y Clasificación de la pregunta

Esta fase se encarga de clasificar la pregunta. Dependiendo del tipo de pregunta la respuesta esperada variará radicalmente. Por ejemplo, la respuesta esperada para la pregunta *"¿Quién ascendió por primera vez el Everest?"* no es la misma que para la pregunta *"¿Qué es el Parlamento Europeo?"*. En la primera se espera un nombre concreto y en la

segunda una definición, el tipo de respuesta difiere entre otros aspectos en la longitud de la respuesta esperada. La estrategia y algoritmos utilizados para la extracción de la respuesta en fases posteriores dependerá de la correcta clasificación de la pregunta.

A raíz del ejemplo mostrado se puede ver que existen diferentes tipos de pregunta. Las más simples son las conocidas como preguntas factuales. Estas preguntas esperan respuestas formadas por una sola palabra o una sola expresión compuesta de pocas palabras, usualmente se corresponde con una Entidad Nombrada. Existen preguntas que pueden esperar una lista de elementos: pueden ser listas cerradas, donde se específica el número de elementos: *"¿Quiénes fueron los tres mosqueteros?"*; o abiertas, donde no se indica cuántos elementos se esperan: *"Nombre los presidentes de la democracia española"*.

Otro tipo de preguntas son las que esperan una definición: *"¿Quién es Felipe González?"* o la anteriormente mostrada: *"¿Qué es el Parlamento Europeo?"*. A parte de estas preguntas existen otro tipo de preguntas de diferente complejidad, son preguntas como *"¿Cómo...?,¿Porqué...?*", o finalmente preguntas cuya respuesta puede ser una afirmación: *"¿Fue Al Gore presidente de los Estados Unidos?"* (respuesta: *No.*).

1.2.2. Recuperación de pasajes

En esta fase se extraen los pasajes del repositorio de documentos donde es posible encontrar la respuesta. Los sistemas de recuperación de pasajes devuelven un conjunto ordenado de documentos a partir de una función de disimilitud entre la pregunta o consulta y el pasaje devuelto. Existen diferentes métodos de *ranking* para la ponderación de los pasajes: basados en frecuencias de términos o métodos estadísticos.

La principal diferencia entre los sistemas de recuperación de información clásicos y los orientados a QA, es que mientras los primeros se centran en la recuperación de pasajes relevantes para la consulta y/o el tópico deseado, los orientados a QA se centran en la extracción de pasajes que puedan contener la respuesta a la pregunta formulada. Otra diferencia es que los sistemas clásicos no están orientados a la búsqueda de pasajes a partir de preguntas o consultas en lenguaje natural, es decir, para sistemas de QA. Si se observa la oración *"¿Quién es Pau Gasol?"* los términos que deberían pasarse al buscador de pasajes es *"Pau Gasol"* y no la pregunta completa. Por tanto, existen términos de la pregunta que no están relacionados con la recuperación de pasajes. Otro detalle a tener en cuenta es el tamaño de los pasajes devueltos, puesto que si el pasaje es muy pequeño puede no llegar a contener la respuesta o la totalidad de ella y si es demasiado extenso se complica el procedimiento de recuperación de la respuesta.

1.2.3. Extracción de la respuesta

La fase de extracción de la respuesta, como su nombre indica, se encarga de recuperar la respuesta a la pregunta formulada a partir de los pasajes devueltos en la fase anterior. Para ello, se utiliza la información extraída en la fase de análisis de pregunta. El principal problema de esta fase es decidir de todas las respuestas posibles cuál es la que debe devolver el sistema, en algunos casos puede haber una o más correctas con lo que el sistema debe elegir cuál es la más adecuada. También se puede dar el caso de que se creen redundancias entre las respuestas, lo cual puede beneficiar al sistema puesto si una respuesta aparece varias veces pero de diferentes maneras es un indicativo de que debe ser correcta.

1.3. Estructura del trabajo

El siguiente documento se estructura como sigue:

- En el capítulo 2 se hará una revisión del estado del arte en el campo de la Búsqueda de Respuesta dirigida por Voz.
- La tarea presentada para la evaluación del sistema se presenta en el capítulo 3.
- En el capítulo 4 se presenta la aproximación utilizada y diferentes aspectos a tener en cuenta al tratar con sistemas de Búsqueda de Respuesta dirigida por Voz.
- Para conocer mejor cómo se ha realizado el trabajo presentado, en el capítulo 5 se describen algunas herramientas utilizadas y sus principales características.
- El capítulo 6 introduce las métricas que se evaluarán sobre la tarea propuesta, así como los modelos utilizados finalmente en la experimentación.
- Los resultados extraídos de la experimentación propuesta y una discusión sobre ellos se muestran en el capítulo 7.
- Finalmente se presentan las conclusiones y trabajo futuro en el capítulo 8.
- En el apéndice A.1 se incluye la evaluación y resultados de la tarea QAst 2009 de forma adicional a la tarea principal evaluada en el trabajo (CLEF).

Capítulo 2

Estado del arte

En este capítulo se muestra el estado del arte para la tarea de Búsqueda de Respuesta dirigida por la Voz.

Competiciones como Text REtrieval Conference (TREC) y Cross Language Evaluation (CLEF) han sido creadas para evaluar y comparar los sistemas de QA, en las últimas ediciones se ha añadido una sesión especial para QAst (Question Answering in Speech Transcripts).

Se han realizado diversos experimentos sobre el impacto de la combinación de sistemas de reconocimiento automático del habla con sistemas de QA. En [Harabagiu et al., 2002] los autores muestran un experimento donde se utiliza el mejor sistema de QA en TREC [Moldovan et al., 2003] junto a un sistema ASR. Se muestra que el impacto del ASR influye considerablemente en una tarea de dominio general como es el caso. El sistema original (sin la entrada de voz) conseguía un 76 % de rendimiento en el sistema QA, mientras que al añadir el reconocedor de voz (la salida del reconocedor operaba con un WER de alrededor del 30 %) los resultados se degradaban hasta obtener solo un 7 % de precisión.

Uno de los principales problemas del uso de los reconocedores en tareas de QA es debido a las palabras fuera del vocabulario del reconocedor, en concreto las entidades nombradas, en esta línea se proponen varias mejoras. En [Fujii et al., 2002] se propone una técnica para intentar reconocer aquellas secuencias de voz que se corresponden con palabras fuera del vocabulario, por ello se propone un modelo de lenguaje formado por las 20,000 palabras más frecuentes combinado con un conjunto de sílabas. En la fase de reconocimiento el resultado es un conjunto de palabras y sílabas. A partir de las

palabras fuera del vocabulario es decir, las que están formadas por sílabas, se realiza una búsqueda en los documentos utilizando como indexado aquellos términos que se asemejan fonéticamente a las sílabas. La precisión de detección de OOV con está técnica es del 26 %.

En [Allauzen and Gauvain, 2005] se utiliza un modelo de lenguaje estático que permite la adición de nuevas palabras en el vocabulario sin necesidad de reentrenamiento del modelo ni adaptación a la tarea. Para añadir una nueva palabra al ASR se debe incluir en el vocabulario de éste, añadir la transcripción (o transcripciones) fonética y añadirse al modelo del lenguaje con una distribución de probabilidad mayor que 0. Para la incorporación de una palabra sin reentrenar el modelo del lenguaje los autores proponen el uso especial de un modelo de clases al que llaman back-off word classes. Durante el entrenamiento del modelo de lenguaje estático, estas clases reemplazan una o más palabras cuya probabilidad se descuenta a partir de la masa de probabilidad destinada a las palabras fuera del vocabulario. Por tanto cuando se añade una nueva palabra debe indicarse la back-off word class, permitiendo así que el reconocedor asigne una probabilidad a esta nueva palabra. Los experimentos realizados muestran un 30 % de mejora de en el reconocimiento de OOV, con un 80 % de las palabras introducidas posteriormente reconocidas.

Otro tipo de técnicas para aumentar la precisión del reconocimiento para tareas de QA consiste en el uso de un sistema que permita la interacción con el usuario en la fase del reconocimiento de la pregunta como se presenta en [Stoyanchev et al., 2008]. En el sistema interactivo el usuario en una primera instancia específica la Entidad Nombrada de interés (el nombre de una persona u organización). Para ello, se utiliza una gramática a partir de las Entidades Nombradas existentes en el corpus de documentos. Una vez la NE es reconocida se extrae un conjunto de documentos que coinciden con la entidad reconocida, a partir de los documentos recuperados se entrena un modelo de lenguaje específico. El modelo de lenguaje utilizado para el reconocimiento se basa en la combinación lineal de un modelo entrenado a partir de preguntas tipo y el modelo creado de forma dinámica. Los resultados obtenidos muestran una considerable reducción del WER (un 32 % en la mejor experimentación).

En la literatura se proponen diferentes mejoras basadas en la forma que debe estimarse el modelo de lenguaje para sistemas de VAQA. Un estudio interesante en la utilización de modelos de lenguajes específicos para QA se presenta en [Schofield, 2003]. Este trabajo compara la perplejidad del uso de modelos de lenguaje basados en n-gramas que han

sido entrenados a partir de un conjunto de 280,000 preguntas únicas formuladas por usuarios extraídas de diferentes fuentes de información en la red. Además se utilizan diferentes tallas de vocabulario, técnicas de descuento y otros factores para la construcción de dichos modelos. A partir de los resultados se llega a la conclusión que la complejidad del modelo de lenguaje es suficiente simple para ser usados en un escenario de reconocimiento predictivo.

En [Kim et al., 2004] se trata el problema de la elaboración de un modelo de lenguaje preciso para la tarea. Se entrenan diferentes modelos de lenguaje para el reconocimiento de consultas de voz a partir de conjuntos de artículos de prensa. Se definen diferentes dominios (economía, entretenimiento, asuntos internacionales, sociedad y deportes). Se entrena cada modelo del lenguaje basados en dichos dominios y se interpola con un modelo de lenguaje de categorías a partir de consultas reconocidas. Los experimentos demuestran una mejora del 5 % de WER tras el uso de modelos de lenguaje dependientes del dominio que el uso de un modelo de lenguaje general para todos los dominios. Incluso el uso de un modelo de lenguaje específico para una tarea desconocida seleccionado a partir del criterio de máxima similitud mejora los resultados frente al uso del modelo de lenguaje genérico.

La estructura de la mayoría de preguntas en sistemas de QA se pueden dividir en dos partes: a) el tópico por el que se pregunta. Y b) la estructura interrogativa de la pregunta. En [Akiba et al., 2007], se estudia como aprender un modelo de lenguaje que aúne varias partes de forma eficiente. El método propuesto consiste la ampliación de un modelo de lenguaje basados en n-gramas a partir de un conjunto de expresiones interrogativas definidas manualmente. La tasa de reconocimiento de expresiones interrogativas mejora con esta técnica frente al uso de un modelo de lenguaje entrenado a partir de corpus de documentos.

Otros estudios proponen la interacción con el usuario para mejorar las prestaciones del reconocedor de voz. Es el caso de [Kim et al., 2004], donde tras reconocer la consulta del usuario, el sistema extrae las palabras clave y las muestra gráficamente al usuario. Éste es el responsable de indicar que palabras han sido mal reconocidas y de seleccionar la correcta a partir de una lista de n-bests obtenida por el ASR.

Un sistema más complejo se presenta en [Hori et al., 2003], se utiliza un filtro que extrae la información significativa de la sentencia reconocida. A partir de las medidas de

confianza aportadas por el reconocedor para cada palabra, se eliminan aquellas que no superen cierto umbral. De la sentencia filtrada se obtiene una oración con sentido a partir de la utilización de una técnica de resumen de voz. Cuando el sistema de QA no puede extraer la respuesta apropiada a la pregunta es entonces cuando se considera ambigua y el mecanismo de interacción se activa demandándole más información al usuario.

Durante las últimas ediciones del CLEF (2007, 2008 y 2009) [Turmo et al., 2007, Turmo et al., 2008, Turmo et al., 2009] se ha añadido la tarea QAst (Question Answering in Speech Transcripts), este apartado dentro del CLEF se engloba en la tarea de Spoken-Question Answering [Gokhan and De Mori, 2011]. Esto implica realizar la búsqueda de información sobre datos de audio (o sus transcripciones automáticas) y/o a partir de preguntas con voz. Sólo en la edición de 2009 se incluyen un conjunto de preguntas con voz. Realizar QA sobre conjuntos de información con voz implica que la respuesta a la pregunta debe encontrarse sobre datos de audio: noticias de radio/televisión, conferencias y seminarios, etc ...

Es interesante remarcar los trabajos realizados en el tratamiento de los errores cometidos por el ASR en este tipo de tareas, en [Comas and Turmo, 2008, Comas and Turmo, 2009] se utiliza una aproximación de búsqueda a partir del uso de fonemas. Se basan en la idea que si el error se ha producido sobre una palabra clave (por ejemplo, entidades nombradas) la transcripción fonética de estos errores se debe asemejar a la del verdadero contenido de la señal de audio. Para ello se utiliza un sistema específico de recuperación de información: Phonetic Alignment Search tool [Comas and Turmo, 2008] , para recuperar aquellas secuencias de fonemas recuperados sobre la pregunta para la búsqueda en los documentos a partir de medidas de similitud fonéticas.

Capítulo 3

La tarea

3.1. CLEF contest

Para evaluar las prestaciones del sistema propuesto se ha utilizado un conjunto de evaluación QA estándar, modificado en este caso para incluir consultas con voz. Se han usado las sentencias provenientes de test de la edición CLEF 2005 (Spanish monolingual) [Vallin et al., 2005]. El conjunto de prueba está formado por un conjunto de 200 preguntas y las correspondientes respuestas.

Además se tiene el repositorio de documentos de donde se extraerá la información correspondiente a las preguntas. Esta colección de documentos esta formada por $454,045$ documentos correspondientes a las todas las noticias de los años 1994 y 1995 de la agencia EFE. El conjunto de documentos tiene una extensión aproximada de 1.06 GB.

El conjunto de preguntas de test se distribuyen en diferentes tipos de preguntas (descritas en la sección 1.2.1): 118 factuales, 50 de definición y 32 de restricciones temporales.

La información necesaria para entrenar el sistema propuesto para abordar la tarea se ha extraído a partir de la colección de preguntas de entrenamiento de las ediciones 2005 y 2006 del CLEF Contest [Vallin et al., 2005, Buscaldi et al., 2006]. En total se ha usado un conjunto de 1,600 preguntas.

Para la obtención de las preguntas con voz se ha realizado un proceso de adquisición del conjunto de las 200 preguntas de test. De esta forma se puede apreciar los efectos de utilizar consultas con voz a diferencia de preguntas escritas correctamente formuladas sin errores de reconocimiento.

El proceso de adquisición se ha realizado mediante un conjunto de micrófonos y auriculares, en condiciones de oficina (sin ruidos externos), a una resolución de 16 bits y una frecuencia de muestreo de 16KHz. El conjunto de 200 preguntas ha sido adquirido por un solo locutor, no obstante, los modelos acústicos utilizados son de propósito general y no han sido previamente adaptados a un locutor específico.

3.2. QAst contest

La tarea *Question-Answering on Speech Transcription evaluation campaign (QAst)* fue creada en 2007 para investigar el problema de realizar tareas de QA sobre voz. Aunque el interés de esta tarea reside en la búsqueda de información sobre repositorios de audio, en la edición 2009 se añadió a la tarea un conjunto de preguntas adquiridas por voz. Adicionalmente a los resultados mostrados para la tarea presentada en el apartado anterior, se ha utilizado este conjunto de preguntas adquiridas para evaluar las prestaciones de la aproximación presentada en este trabajo. En el anexo A.1 se muestran la experimentación y resultados obtenidos para las oraciones adquiridas del QAst 2009 Contest.

En la edición del año 2009 [Turmo et al., 2009], en la subtarea utilizada para la experimentación (la referente al castellano), las preguntas se realizaban sobre el corpus *TC-STAR05 EPPS Spanish Corpus*. Este corpus está formado por tres horas de de grabaciones en castellano del Parlamento Europeo. Está información se usó para evaluar los reconocedores desarrollados en el proyecto TC-STAR. Junto a las grabaciones existen tres salidas de tres reconocedores diferentes con diferentes tasas de reconocimiento (WER): 11,5, 12,7 y 13,7.

La información extraída del TC-STAR se ha realizado sobre las transcripciones incluidas en el corpus. Para mejorar la precisión de los modelos de lenguaje y evitar errores al aprender modelos de lenguaje sobre las transcripciones, se ha utilizado también el corpus formado por las Actas del Parlamento Europeo EUROPARL[Summit, 2005, europarl, 2009]. Las actas en español están formadas por $1,942,761$ oraciones y un total de $55,105,479$ palabras.

El conjunto de preguntas analizadas está formado por 100 preguntas en castellano de la tarea y sus correspondientes transcripciones, además se han incluido las 50 preguntas adquiridas para el training.

Capítulo 4

Modelado de Lenguaje

El proceso de reconocimiento automático del habla consiste en la obtención de una secuencia de palabras a partir de una señal de audio como entrada [Jurafsky and Martin, 2009]. La formulación matemática de este proceso desde un punto de vista estadístico puede verse como [Bahl et al., 1983]:

$$\widehat{W} = \underset{W \in \Sigma^*}{\operatorname{argmax}} P(W|X) \tag{4.1}$$

donde $\widehat{W}$ es la mejor secuencia de todas las posibles unidades lingüísticas en el vocabulario Σ conforme una secuencia de características acústicas (X) extraídas de la señal acústica a partir de un proceso de parametrización.

Aplicando el teorema de Bayes y asumiendo que la probabilidad de que se pronuncie la secuencia acústica $P(X)$ es independiente de la sentencia W, el proceso de maximización se formula como:

$$\widehat{W} = \underset{W \in \Sigma^*}{\operatorname{argmax}} \frac{P(X|W)P(W)}{P(X)} = \underset{W \in \Sigma^*}{\operatorname{argmax}} P(X|W)P(W) \tag{4.2}$$

Por tanto, se puede resumir el proceso de reconocimiento como encontrar la secuencia $\widehat{W}$ que maximice el producto de las dos probabilidades:

- $P(W)$: es la probabilidad a priori de la secuencia W y la proporciona el modelo de lenguaje.
- $P(X|W)$: esta probabilidad la proporciona el modelo acústico.

El modelo de lenguaje restringe las secuencias de unidades lingüísticas que serán reconocidas, además representa restricciones sintácticas para tareas de gran vocabulario con lo que permite reducir la complejidad computacional y mejorar las prestaciones del sistema. Los modelos de lenguaje basados en n-gramas son los más conocidos y los más extendidos en la actualidad. Este sistema modela la aparición de secuencias de n unidades lingüísticas. Esta frecuencia se estima a partir de un conjunto de entrenamiento de oraciones escritas.

En un modelo de n-gramas, la probabilidad de una palabra depende de las $n-1$ palabras anteriores. Por tanto, la probabilidad del modelo de lenguaje asignada a una palabra se expresa como:

$$P(W) = \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1}) \tag{4.3}$$

Una vez se han estimado los modelos de lenguaje y acústico a partir de un corpus de entrenamiento, el proceso consiste en la obtención de la secuencia que maximice la ecuación 4.2, para ello se utilizan algoritmos de búsqueda. El más conocido y utilizado para esta tarea es el algoritmo de Viterbi.

En esta sección se muestra con detalle la aproximación utilizada para la modelización específica del lenguaje para la tarea descrita. Se describen varios métodos de modelización que ilustran la problemática de la tarea y los pasos seguidos para la obtención del modelo de categorías. En primer lugar mostramos la modelización del lenguaje para la tarea sin usar categorías: utilizando para entrenar el modelo de lenguaje las preguntas, el repositorio de documentos o la combinación de ambos. Posteriormente se describe la modelización basada en categorías y algunas mejoras fonéticas propuestas para la tarea en cuestión.

4.1. Modelado sin categorías

En este apartado se muestran aproximaciones para entrenar el modelo de lenguaje para la tarea dada, sin utilizar técnicas basadas en el uso de categorías. Se presenta así, la problemática de la incorporación de nueva información al modelo de lenguaje pero a su vez mantener la estructura propia de las preguntas.

4.1.1. Modelado usando preguntas

El paso más simple para entrenar el modelo de lenguaje consiste en utilizar las preguntas de entrenamiento y esperar que el sistema funcione bien con las preguntas de test. Esta aproximación puede resultar adecuada cuando el conjunto de entrenamiento es muy amplio y cubre gran parte de las posibles preguntas que se pueden presentar al sistema. Sin embargo, dado el conjunto reducido de preguntas y que los conjuntos de test y entrenamiento son conjuntos disjuntos, el número de palabras fuera de vocabulario es crítico. Si una palabra, o lo que es peor, una entidad nombrada no ha sido pronunciada en el conjunto de entrenamiento es prácticamente imposible que la consulta se resuelva con éxito.

Por tanto, la aproximación es bastante pobre, por lo que hay que buscar una técnica de modelado de lenguaje que contenga un conjunto más general de información.

4.1.2. Modelado usando el corpus de documentos

Toda la información que pueda ser preguntada está en el conjunto de documentos. Pero principalmente en forma declarativa o en redacción pura, por lo que la ventaja de tener la estructura de las preguntas se pierde si se usan exclusivamente los documentos para aprender el modelo de lenguaje. Además se ha visto que el repositorio de documentos en el caso de la agencia EFE es muy extenso, con lo que la complejidad del reconocedor al acceder el modelo de lenguaje es mayor. Incluso el cálculo del modelo de lenguaje se hace costoso. Esta aproximación, aunque tiene sus inconvenientes, tiene la ventaja que cubre un gran número de palabras del vocabulario y si la acústica del sistema de reconocimiento funciona bien se pueden recuperar gran parte de las preguntas que se realicen al sistema.

4.1.3. Modelo interpolado de preguntas y corpus de documentos

Otra aproximación consistiría en utilizar un modelo de lenguaje híbrido que incorporara las preguntas y el conjunto de documentos. Dado que el tamaño de los documentos es mucho mayor al conjunto de preguntas, es necesario realizar algún tipo de combinación de los modelos de lenguaje por separado. Se pueden combinar de forma lineal como:

$$P(W) = \lambda \prod_{k=1}^{n} P_{QM}(w_k|w_{k-N+1}^{k-1}) + (1-\lambda) \prod_{k=1}^{n} P_{CM}(w_k|w_{k-N+1}^{k-1}) \tag{4.4}$$

dónde P_{QM} es la probabilidad del modelo de lenguaje formado por las preguntas, y P_{CM} es la probabilidad devuelta por el modelo de los documentos.

El problema de esta combinación lineal es que se le da el mismo peso a las partículas de las preguntas (*Quién, Dónde, etc...*) que a las entidades que aparecen en este conjunto, si el valor de λ es muy alto las entidades que aparecen en las preguntas tendrán un peso superior que las entidades que aparecen en el documento, y eso para el cometido de generalización del modelo de lenguaje y por extensión del reconocedor, no es deseable.

Otra manera de combinar ambos modelos de lenguaje es replicar manualmente las preguntas y crear un corpus con los documentos y el conjunto replicado de preguntas. Así al calcular el modelo de lenguaje por conteo, las estructuras de las preguntas adquieren más peso. Si sólo se replican las partículas interrogativas se resuelve el problema de ponderado de entidades, pero se pierde la expresividad de los n-gramas puesto que las partículas se replican de forma aislada al resto de las pregunta.

4.2. Modelado de categorías

Tanto en Búsqueda de Respuesta y como es Búsqueda de Respuesta dirigido por la Voz las preguntas (la entrada del sistema de QA) están formuladas en lenguaje natural, por tanto las preguntas pueden no formularse siempre en forma interrogativa (*¿Quien fue el primer premio nobel de literatura?*), es decir, pueden aparecer sentencias expresadas de forma declarativa (*Nombre jugadores de baloncesto.*), por tanto el sistema debe tener en cuenta este tipo de restricciones.

La Figura 4.1 muestra como el modelo de lenguaje para el ASR ha sido construido. Como se puede ver se han usado dos conjuntos para el entrenamiento del modelo de lenguaje. Uno se corresponde con el conjunto de preguntas de entrenamiento, de las cuales el sistema estimará la estructura sintáctica de las preguntas; el otro conjunto se corresponde con la colección de documentos donde se realizará la búsqueda, del que se extraerá la información adicional para ampliar el modelo de lenguaje generado.

La idea principal es construir un modelo de lenguaje capaz de incluir la mayor cantidad de Entidades Nombradas y otras palabras relevantes contenidas en el repositorio de información y que son importantes para la consulta de información, pero que a la vez sea capaz de representar la estructura sintáctica de las preguntas extraídas del conjunto de test (por ejemplo, *¿Quién es...?, ¿Cuando fue...?,¿Qué es...?*). Para poder llevar

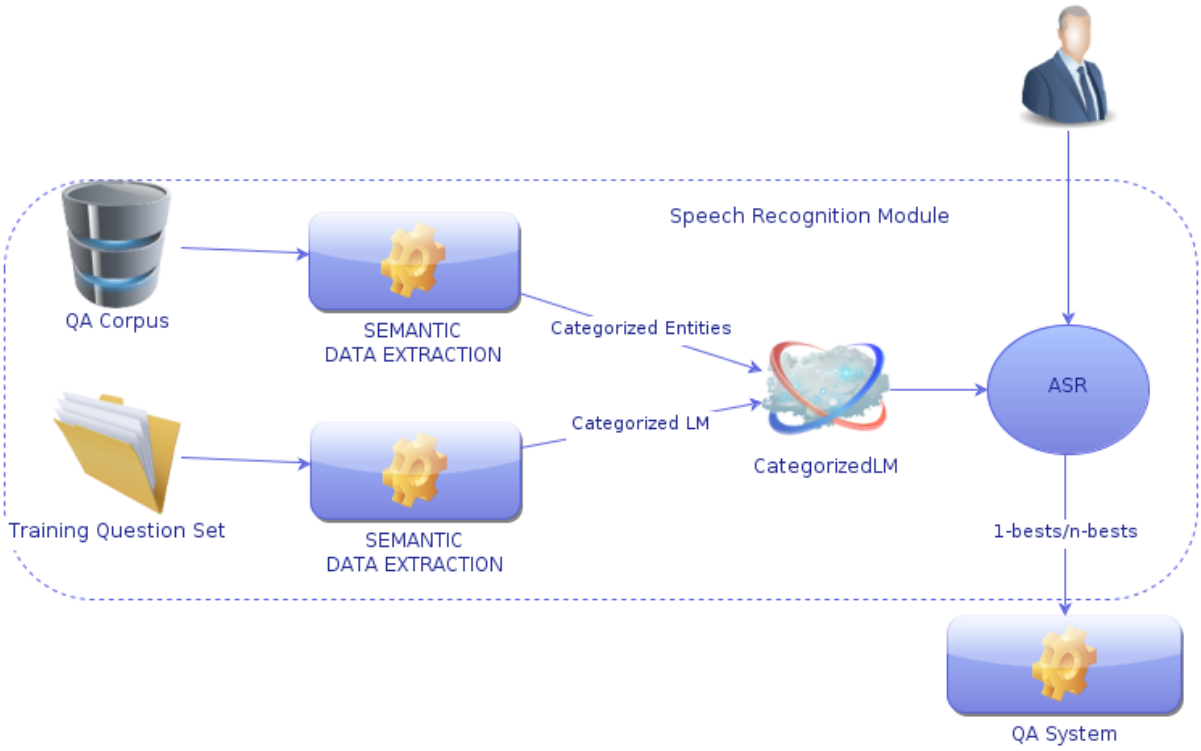

Figura 4.1: Módulo de Reconocimiento de Voz

a cabo tal cometido es necesario construir un modelo de lenguaje basado en categorías (o clases), donde las categorías están relacionadas con los posibles conceptos que serán preguntados por el usuario.

Estás categorías no tienen por que corresponderse con las categorías gramaticales propiamente del lenguaje, por ejemplo, en un modelo de lenguaje formado por fechas, es mucho más flexible categorizar el modelo de lenguaje con frases del estilo: "*$DÍA de $MES de $AÑO*" donde *$DÍA, $MES, $AÑO* son las posibles categorías que incluyen los días, meses y años posibles; que tener un modelo de lenguaje plano entrenado a partir de sentencias en las que no aparecen todos los meses o incluso la combinación de ciertas fechas, haciendo que la probabilidad de que aparezca una determinada fecha sea muy baja.

El mismo problema sucede con los conceptos que pueden ser preguntados, si algo no ha sido preguntado en el conjunto de entrenamiento no aparecerá en el vocabulario del ASR y no podrá ser reconocido por este. Otra opción es que la palabra sí que esté en el modelo de lenguaje como unigrama pero que no aparezca en ninguna de las consultas. Por ejemplo, si queremos formular la pregunta "*¿Quién es Nelson Mandela?*" y en nuestro modelo de lenguaje aparece "*¿Quien es Morgan Freeman?*" y el nombre "*Nelson*

Mandela" en un contexto diferente, el modelo de lenguaje podrá recuperar la oración deseada (dependiendo del descuento utilizado), pero el sistema sería mucho más robusto si el modelo de lenguaje estuviera constituido de la forma "*¿Quien es $Entidad_Nombrada*" donde *$Entidad_Nombrada* es una categoría que incluye los nombres para este ejemplo *(Nelson Mandela, Morgan Freeman)*. De esta forma, es posible ampliar fácilmente la potencia del modelo del lenguaje incluyendo nuevos elementos a las categorías. En este caso, la categoría principal utilizada será la de Entidad Nombrada, y por tanto una parte de la experimentación llevada a cabo se refiere directamente a la cantidad de palabras relevantes que serán incluidas en tal categoría.

Un modelo de categorías se utiliza principalmente para reducir el número de parámetros a estimar, con lo que se necesitan un número menor de muestras de aprendizaje. En este caso, se puede entrenar el modelo de categorías con un conjunto relativamente pequeño de preguntas.

De la estimación del modelo de lenguaje basado en n-gramas tenemos que la probabilidad de una palabra depende de las $n-1$ palabras anteriores:

$$P(w_i|w_{i-1},\ldots w_{i-n}) \tag{4.5}$$

Con la aproximación basada en categorías, podemos decir que una palabra se estima a partir de las categorías de las $n-1$ palabras anteriores:

$$P(w_i|w_{i-1},\ldots,w_{i-n}) \approx P(c_i|c_{i-1}\ldots c_{i-n})P(w_i|c_i) \tag{4.6}$$

donde:

- $P(w_i|w_{i-1},\ldots,w_{i-n})$ es la probabilidad condicional de que la palabra w_i aparezca después de la secuencia $w_{i-1},\ldots,w_{i-n}$.
- c_i es la categoría asociada a la palabra w_i.
- $P(w_i|c_i)$ es la probabilidad de pertenencia de la palabra w_i a la clase c_i.
- $P(c_i|c_{i-1},\ldots,c_{i-n})$ es la probabilidad condicional de que una palabra que pertenece a la categoría c_i aparezca después de una secuencia de palabras pertenecientes a las categorías $c_{i-1},\ldots,c_{i-n}$.

La probabilidad de pertenencia de una palabra a una clase puede estimarse por criterio de máxima verosimilitud como el cociente entre el número de ocurrencias de una palabra en una categoría y el número total de palabras (contando repeticiones) que pertenecen a la categoría $C(c_i)$:

$$P(w|c) = \frac{C(w)}{C(c_i)} \tag{4.7}$$

Para nuestro dominio y viendo que la intención es la de ampliar posteriormente las palabras correspondientes a cada categoría, se utiliza una aproximación basada en que la probabilidad de pertenencia de una palabra a una categoría es equiprobable a todas las palabras de la categoría. Se define como:

$$\forall w \in c_i, P(w|c_i) = \frac{1}{|c_i|} \tag{4.8}$$

Donde el $|c_i|$ es el número de elementos en la categoría. Al tener todas las palabras de una categoría la misma probabilidad de aparición se asume que la acústica será la encargada de discernir entre las diferentes pronunciaciones. Por tanto, un gran número de elementos en una categoría puede afectar al rendimiento del proceso de búsqueda del ASR así como a las prestaciones de este, puesto que existe más variabilidad.

Para la estimación de estas probabilidades es necesario tener un texto debidamente etiquetado y a su vez un diccionario que permita la agrupación de las palabras con la clase correspondiente.
Existen diferentes métodos para la definición de las categorías y las palabras que pertenecen a dichas categorías:

- Conocimiento lingüístico: En este caso se suele utilizar un etiquetado POS (Parts of Speech). Este método introduce el error en la fase de etiquetado POS, tanto manual como automático, debido sobretodo a problemas de ambigüedad.

- Conocimiento específico del dominio: Se definen específicamente las categorías. Usualmente suele encargarse un experto de definir las categorías, tarea que puede resultar compleja dependiendo del dominio.

- Agrupamiento automático de clases: Se utilizan generalmente algoritmos de agrupamiento basados en modelos estadísticos y otros métodos basados en el principio

de Información Mutua [P.F. et al., 1992]. En [Niesler et al., 1998] se muestran diferencias de utilizar métodos de agrupamiento automático y etiquetamiento POS.

4.2.1. Selección de categorías

Para poder incorporar información relevante al modelo de lenguaje basado en categorías, deben seleccionarse qué palabras relevantes deben ser categorizadas. La flexibilidad del modelo de lenguaje dependerá del tipo de categoría seleccionada.

A continuación se describen aquellas categorías que se han utilizado para la elaboración del modelo de lenguaje:

- Entidades Nombradas: normalmente se trata del concepto principal que puede ser preguntado por el usuario.
- Fechas y números: como se ha comentado, no todas las posibles combinaciones de días, meses y años se encuentran en el conjunto de frases de entrenamiento.
- Nombres comunes: en algunos casos las entidades nombradas no aportan toda la información sobre el concepto que se ha preguntado, o también puede ser que no aparezcan en la pregunta. En casos como estos algunos nombres comunes pueden proveer información relevante para encontrar la respuesta (¿Quién es el *presidente* de los Estados Unidos).
- El resto de palabras que no se han incluido en las categorías anteriores, formalmente se puede decir que una palabra pertenece a su propia categoría. Se trata de partículas interrogativas, preposiciones, artículos, y otros elementos que no se etiquetan.

4.2.2. Extracción de Entidades Nombradas

Para la obtención de las palabras que formarán parte de las categorías previamente introducidas, al tratarse de grandes conjuntos de datos, se precisan métodos automáticos. En esta experimentación se ha utilizado la herramienta FreeLing, en concreto la característica de etiquetado POS [Padró et al., 2010, Atserias et al., 2006, Carreras et al., 2004].

Como indica la figura 4.1, este proceso se realiza en dos partes. En la primera referente al conjunto de cuestiones, se realiza un etiquetado POS del cual se extraen las palabras

que pertenecen a las categorías, a su vez las frases son sustituidas por las etiquetas de clase de las palabras extraídas. Por ejemplo, la pregunta *"¿Cual es la montaña más alta del Nepal?"* tras este proceso quedaría como *"¿Cual es la $NOMBRE_COMÚN más alta del $ENTIDAD_NOMBRADA?"*. Estas frases etiquetadas son las que se usarán para la creación y entrenamiento del modelo de lenguaje basado en categorías.

En el caso de la colección de documentos, se realiza un proceso de etiquetado POS de la misma forma que se realizó con el conjunto de preguntas. A partir de este proceso se extraen dos listas ordenadas por frecuencia de aparición de las entidades nombradas y nombres comunes. Del conjunto de documento se extrae sólo la información de estas dos categorías porque son los dos conceptos que se quieren usar para generalizar la información del reconocedor a partir de la información existente en el conjunto de documentos.

4.2.3. Filtro de las Entidades Nombradas

El conjunto de Entidades Nombradas tras la extracción del conjunto de documentos obtenido está formado por más de un millón de entidades, se trata de un conjunto relativamente alto para la utilización en un sistema de reconocimiento automático. De hecho, puede haber cierto ruido y datos no deseados en este conjunto que pueden afectar a las prestaciones del sistema de reconocimiento. El objetivo de esta fase consiste en la obtención de un conjunto de Entidades Nombradas más reducido y que incorpore, dentro de unos límites, la información que maximice las prestaciones del ASR para la tarea en concreto. En este apartado se exponen los métodos de selección y filtrado de los elementos que se incluirán en las categorías propuestas. En especial las entidades nombradas que son la categoría que aporta más información para el funcionamiento de la tarea QA.

Analizando el conjunto de entidades obtenido tras el etiquetado del repositorio de documentos, se ve que gran parte de las entidades tienen una frecuencia de aparición muy baja, en algunas ocasiones debido a faltas de ortografía y problemas de codificación. Si realizamos un primer filtro basado en la frecuencia de aparición y eliminamos las que aparecen menos de 10 veces el conjunto remanente se reduce a 80,000 y si el umbral se fija en 20 el conjunto obtenido ronda las 48,000 entidades.

Tras analizar estos resultados, se ha utilizado el conjunto de frases de entrenamiento para comprobar la cobertura de Entidades Nombradas y Nombres Comunes del proceso realizado y para comprobar que la eficacia del filtrado.

Otro criterio a tener en cuenta para filtrar las Entidades Nombradas son los valores de confianza aportados por la herramienta POS, en este caso cada palabra etiquetada como Entidad Nombrada o Nombre Común posee un valor de confianza normalizado entre 0 y 1 que indica la confianza de pertenencia a esa categoría.

La figura 4.2 muestra la cobertura, porcentaje de palabras del conjunto de training que han sido encontradas y extraídas correctamente, en el proceso sobre la colección de documentos. Las gráficas muestran los resultados a partir de la utilización de un umbral de confianza y del tamaño del conjunto seleccionado para cada categoría.

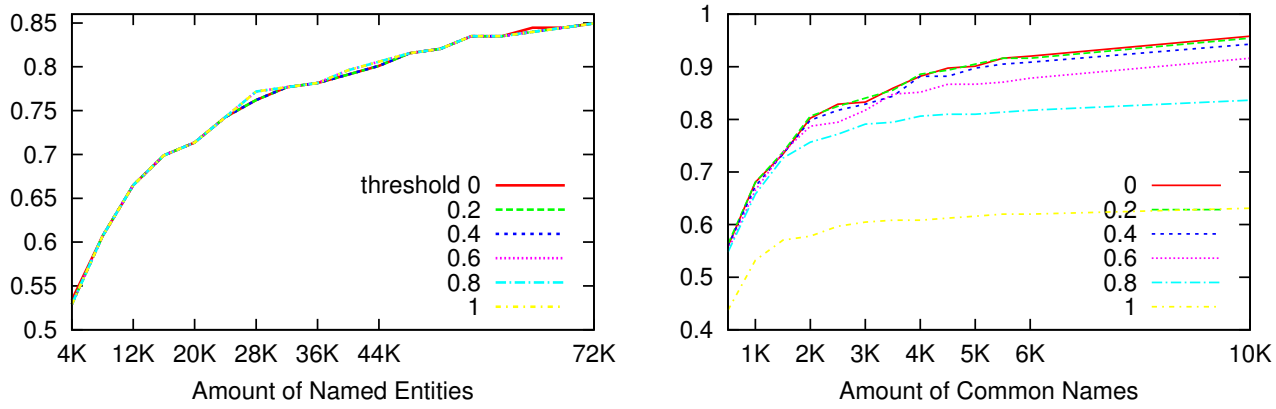

Figura 4.2: Cobertura de NE/CN test

Se ve que para las dos categorías la cobertura aumenta conforme se añaden más elementos a cada categoría pero que, a valores más altos la tasa de crecimiento se reduce (más o menos de forma logarítmica) haciendo que sea necesario añadir muchos más elementos a la categoría para poder así cubrir algunos elementos del conjunto de entrenamiento.

En el caso de las entidades nombradas se puede ver que no existe diferencia en la utilización de un umbral más alto o más bajo, esto se debe principalmente a dos factores: a) no existe un conjunto significativo de entidades con valores de confianza bajo puesto que si el valor de confianza es muy bajo es muy probable que el sistema POS adjudique estas palabras a otra categoría. Y b) los valores de confianza más bajos se debe también a palabras ambiguas y errores (ortográficos o de la misma redacción de los documentos) por lo que tienen una frecuencia de aparición muy baja y se eliminan del conjunto tras realizar el filtro de frecuencia.

Para los nombre comunes parece que hay más diferencia en la utilización de un umbral de confianza, sin embargo se ve que finalmente cuanto más permisivo se es, mayor cobertura se obtiene en el sistema tras realizar el filtrado.

4.3. Mejoras fonéticas

Se ha visto que uno de los problemas de Búsqueda de Respuesta dirigida por la Voz es el reconocimiento de entidades nombradas que están formuladas en un idioma que no es el idioma nativo del usuario del sistema de QA. Esto plantea diferentes formas de abordar el problema, en principio el sistema deberá disponer de ciertos mecanismos que permitan pronunciaciones alternativas de este tipo de palabras o la posibilidad de combinar modelos acústicos y de lenguaje provenientes de otro idioma. En este apartado se muestran algunas técnicas utilizadas en este campo para mejorar las prestaciones del sistema frente a este tipo de situaciones.

4.3.1. Utilización de pronunciaciones alternativas en el lenguaje nativo

Cuando se pronuncian entidades nombradas provenientes de un idioma diferente al que se está formulando la pregunta, por ejemplo *"¿Quién es George Washington?"*, se puede pensar que la entidad nombrada no se pronuncia realmente como se pronunciaría en el idioma original. Sino que se utilizan fonemas propios del idioma original (/llorch guasinton/). Es decir se usa una fonética más cercana al usuario. También depende de parámetros intrínsecos al tipo de usuario que realiza la consulta, puesto que cada persona puede pronunciar este tipo de entidades de forma diferente.

Este tipo de pronunciaciones presentan una ventaja en la fase de reconocimiento puesto que utilizando los modelos de lenguaje y acústicos del idioma de la tarea se pueden tratar este tipo de entidades sin necesidad de reentrenar los modelos acústicos con nuevos fonemas. El problema reside en la obtención de las pronunciaciones de este tipo de palabras utilizando la fonética del idioma original.

En la tarea propuesta se abarca este problema para las pronunciaciones de entidades nombradas, en concreto para el inglés utilizando la fonética propia del castellano. En principio se utiliza un transcriptor ortográfico-fonético para obtener la fonética de las palabras del vocabulario del reconocedor, la idea es obtener aquellas entidades que puedan considerarse inglesas y obtener la pronunciación de esta en inglés para posteriormente realizar una equivalencia a fonemas del castellano, evidentemente este último paso no es trivial y no contempla con exactitud todas las transcripciones. El problema se agrava además, si tenemos en cuenta la dificultad añadida de la transcripción ortográfica-fonética del inglés, por tanto se propone utilizar un diccionario de pronunciaciones como es el Carnegie

Mellon University pronuncing dictionary [1] (CMU dictionary). para obtener las transcripciones de la fonética inglesa de las entidades. Además se utiliza el mismo diccionario para obtener que entidades pueden ser susceptibles de tener pronunciaciones en inglés.

El algoritmo para esta aproximación se resumiría de la siguiente forma:

1. A partir de la lista de entidades nombradas, se extraen sus transcripciones ortográfico-fonéticas utilizando un transcriptor del castellano.
2. De la lista de entidades nombradas, se seleccionan aquellas entidades que se incluyen en el Carnegie Mellon University pronuncing dictionary, es decir, aquellas que se pueden pronunciar en inglés.
3. Se extrae la transcripción obtenida por el diccionario, utilizando la fonética del inglés, y se realiza una conversión a la fonética del español.
4. Se añaden las pronunciaciones adaptadas como pronunciaciones alternativas a las que se tenían en un principio.

4.3.2. Utilización de modelos acústicos híbridos o combinados

Otra alternativa consiste en incluir en el mismo reconocedor la posibilidad de reconocer entidades de diferentes idiomas utilizando las fonéticas de la combinación de estos idiomas.

Una aproximación se basa en la utilización de modelos acústicos híbridos que contemplen en la fonética los fonemas de los idiomas que se quiera tratar, entonces no existiría la necesidad de obtener una equivalencia entre fonéticas, pero a su vez necesita del entrenamiento de los modelos acústicos utilizando la fonética combinada.

También, es posible utilizar los valores de confianza devueltos por el reconocedor para detectar aquellas secuencias de audio que no han sido debidamente reconocidas, o cuyos valores no superen un cierto umbral, y utilizar estos segmentos de audio en un reconocedor entrenado con la fonética del idioma alternativo para ver si la acústica se asemeja más a las entidades en dicho idioma. El algoritmo utilizado se mostraría de la siguiente forma:

1. Se entrena un modelo de lenguaje sólo con las entidades nombradas que aparezcan en el Carnegie Mellon University pronuncing dictionary. Las transcripciones ortográfico-fonéticas utilizadas son las que proporciona el diccionario.

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

2. Se reconoce la pregunta utilizando como base el reconocedor configurado para el idioma original de la pregunta (el castellano en este caso).

3. Se toman las secuencias de audio que no sobrepasan cierto umbral. Este umbral puede fijarse de forma arbitraria, siguiendo alguna aproximación de corte basada en histogramas o calculando los valores medios de los valores de reconocimiento y aplicar algún corte en torno a esta medida (por ejemplo, la media menos la desviación típica).

4. Estás secuencias de audio se reconocen usando el modelo de entidades del nuevo idioma, y si los valores de confianza acústicos superan los del idioma original se sustituyen los segmentos por los nuevos.

En este caso no se contempla que la fonética utilizada para pronunciar estas entidades puede diferir ligeramente (por ser una pronunciación de una palabra extranjera para el usuario) de la de los modelos acústicos aprendidos.

4.3.3. Unificación fonética de Entidades

Un problema que deriva del modelado propuesto se presenta cuando varias entidades nombradas comparten la misma transcripción fonética. Si la probabilidad del modelo del lenguaje es equiprobable para todos aquellos elementos de la misma categoría, es la acústica la que se encarga de decidir que entidad es la pronunciada por el usuario. Pero si dos palabras comparten la misma fonética debe definirse un procedimiento para la selección de la entidad adecuada.

Para abordar el problema de pronunciaciones idénticas, se pueden unir aquellas entidades con la misma fonética manteniendo la transcripción ortográfica para la entidad cuya frecuencia de aparición en el corpus es mayor. Para hacer esto se ha utilizado la herramienta de transcripción ortográfico-fonético Ort2Fon [Castro-Bleda et al., 2001].

Debido a que el corpus de documentos consta de una colección heterogénea de documentos que incluyen todas las noticias publicadas por la agencia EFE a lo largo de dos años, existen entidades nombradas que aparecen escritas de diferente forma. Muchas veces estas entidades tienen diferentes transcripciones ortográficas pero la misma transcripción fonética. (por ejemplo, *Korea y Corea*, *Qatar y Catar* y *Tokio y Tokyo*). También existen casos especiales donde dos entidades diferentes tienen la misma transcripción fonética

(por ejemplo, *Baldi y Valdi*). En este caso, debido al hecho de que el reconocedor es incapaz de discriminar entre las entidades, se usa la entidad más frecuente.

De las aproximaciones presentadas, en la experimentación solo se ha llevado a cabo esta última.

Capítulo 5

Herramientas Utilizadas

El presente trabajo aúna tareas de reconocimiento de voz con tareas de extracción de información y búsqueda de respuesta, es por ello que se precisan de herramientas específicas para cada una de las tareas presentes en el trabajo. En esta sección se muestran las principales herramientas utilizadas, así como sus características más destacables y detalles de funcionamiento que son importantes para comprender la calidad de los resultados obtenidos.

5.1. Reconocedor Automático del Habla Loquendo

Los experimentos llevados a cabo se han realizado con el reconocedor automático del habla Loquendo ASR ®[1] [Loquendo, a] . Se trata de un reconocedor comercial disponible para diversos idiomas. En el presente trabajo se ha utilizado el castellano.

El reconocedor automático de Loquendo puede funcionar mediante gramáticas o mediante modelos estadísticos basados son los n-gramas. Además, permite la utilización de modelos de lenguaje categorizados.

Para la definición del modelado de lenguaje, Loquendo incluye su propia herramienta de entrenamiento del modelo de lenguaje de categorías: Loquendo SATCA [Loquendo, b]. Los modelos de lenguaje entrenados con Loquendo utilizan un modelo de trigramas y suavizado lineal.

[1] `www.loquendo.com`

5.2. Passage Retrieval JIRS

Para la fase de extracción de pasajes se ha usado el sistema de recuperación de pasajes JIRS [2] (Java Information Retrieval System) [Buscaldi et al., 2010], este sistema se basa en la coincidencia de n-gramas entre la pregunta y los documentos recuperados.

En JIRS se considera un n-grama como una secuencia de n términos adyacentes extraídos de la consulta. JIRS se basa en la premisa de que en un una colección bastante extensa de documentos, los n-gramas que aparecen en la pregunta aparecerán próximos a la respuesta en la colección de documentos.

En un primer paso, JIRS extrae los pasajes que contienen los términos de la pregunta en la colección de documentos usando un esquema estándar $tf.idf$ [Buscaldi et al., 2006]. El sistema recibe la pregunta formulada por el usuario y a partir de esta devuelve una lista ordenada de frases conforme el esquema de pesos tf.idf. A estas oraciones se les añaden las k oraciones contiguas para crear el pasaje de tamaño $m = 2k + 1$ que será devuelto.

Una vez obtenidos los pasajes se calcula la similitud entre los pasajes y la cuestión a partir de la ecuación 5.1. La arquitectura de JIRS se muestra en la figura 5.1.

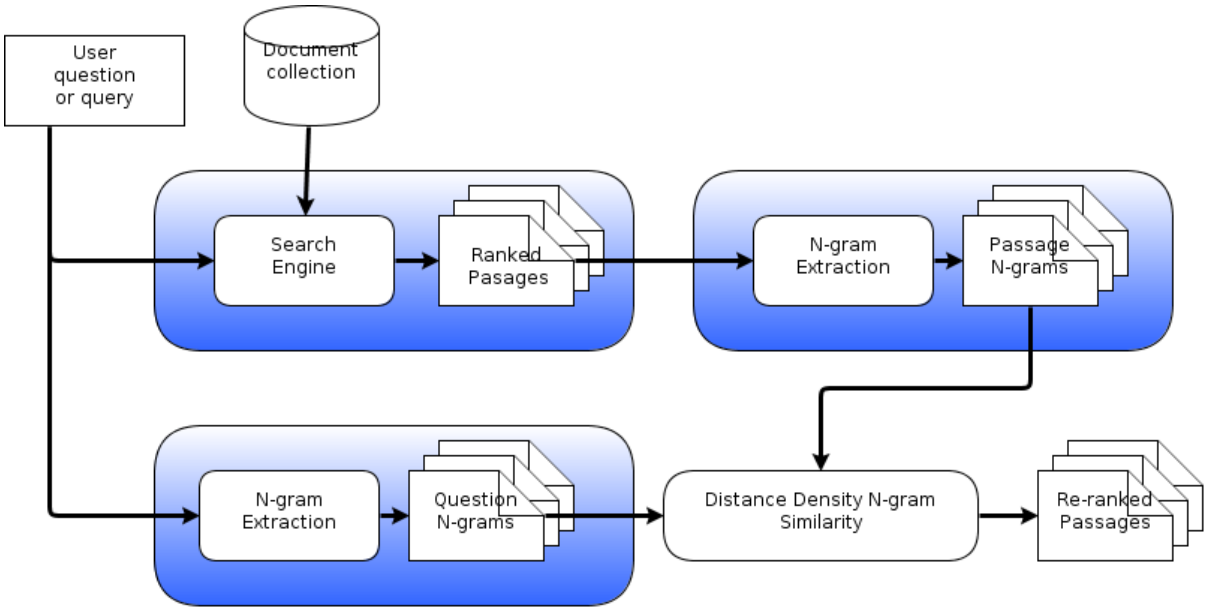

Figura 5.1: Arquitectura de JIRS

[2] Las herramienta de recuperación de información puede obtenerse en la dirección URL `http://www.sourceforge.net/project/jirs`

$$Sim(p,q) = \frac{\sum_{\forall x \in (P \cap Q)} \frac{h(x)}{1 + \alpha \cdot\ ln(1 + d(x, x_{max}))}}{\sum_{i=1}^{n} w(t_{q_i})} \tag{5.1}$$

donde:

- P es el conjunto de k-gramas ($1 \leq k \leq n$) en el pasaje p.
- Q es el conjunto de k-gramas en la cuestión q.
- n es el número total de términos extraídos en la pregunta,
- $h(x)$ indica el peso de cada k-grama y se define a partir de la función:

$$h(x) = \sum_{j=1}^{k} w(t_{x_j}) \tag{5.2}$$

- $w(t)$ es el peso del término t y se determina como:

$$w(t) = 1 - \frac{log(n_t)}{1 + log(N)} \tag{5.3}$$

  donde:

  - n_t es el numero de sentencias donde el término t aparece.
  - N es el número de oraciones en la colección.

- $d(x, x_{max})$ se calcula como el número de palabras entre cualquier k-grama x y aquel que tenga el máximo peso (x_{max}), existe un factor α que determina la importancia de la distancia en la medida de similitud:

$$d(x, x_{max}) = 1 + \alpha \cdot ln(1 + L) \tag{5.4}$$

Si el término t_k aparece sólo una vez en la colección su peso será 1 (peso máximo). Por otro lado si un termino aparece en cada una de las sentencias se le asignará el mínimo peso, es el caso de las palabras que no aportan información como las stopwords.

5.3. FreeLing

FreeLing[3] proporciona una librería gratuita de análisis lingüístico. Está desarrollado por el Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP) de la Universitat Politècnica de Catalunya (UPC). Permite múltiples funciones, como división en oraciones, lematización, etc., para español, catalán, gallego, italiano, inglés, galés, portugués y bable (asturiano) (posee diccionarios específicos para cada lengua).

De las características que proporciona Freeling, se ha utilizado el etiquetado de categorías morfosintácticas (*Part Of Speech Tagging,PoS*), que a su vez es capaz de detectar las entidades nombradas como otras categorías interesantes para la tarea.

Es importante destacar que el diccionario utilizado para el etiquetado POS para el castellano, que es el que se ha utilizado en esta tarea, contiene más de 550,000 elementos que se corresponden con $76,000$ combinaciones de lemas y PoS.

Se puede encontrar documentación sobre las características de la librería y funcionamiento de estas en [manfreeling, 2010, Padró et al., 2010].

[3] `http://nlp.lsi.upc.edu/freeling/index.php`

Capítulo 6

Evaluación

En este capítulo se introduce la experimentación llevada a cabo en el trabajo realizado. Se definen las métricas utilizadas para la evaluación de los resultados y el diseño de los experimentos llevados a cabo.

6.1. Definición de métricas

Las métricas utilizadas se pueden dividir en tres grupos: uno referente a los resultados de reconocimiento, otros inherentes al modelado de lenguaje, y finalmente los valores evaluados tras la fase de recuperación de información.

6.1.1. Métricas relacionadas con el ASR

Para medir las prestaciones del modelo de lenguaje utilizado en la tarea se han tomado diferentes medidas relacionadas con la salida del reconocedor. Normalmente la evaluación del proceso de reconocimiento se evalúa mediante el Word Error Rate (WER) que se define a partir del número de substituciones, inserciones y borrados de palabras que se necesitan para transformar la frase reconocida en la frase original. Se formula de la siguiente forma:

$$WER = \frac{I + S + B}{N} \tag{6.1}$$

donde:

- I indica el número de inserciones realizadas.

- S indica el número de sustituciones.
- B indica el número de eliminaciones.
- N es el número de palabras del conjunto de referencia.

Debido a que el conjunto de entrenamiento se divide en oraciones (preguntas o consultas) es posible medir el WER de todas las oraciones como un conjunto o calcular el WER medio de todas las oraciones.

En los resultados cuando se habla de WER se considera el WER tomando todas las oraciones como un conjunto, de todos modos ya que la longitud de las preguntas no difiere mucho, no existen diferencias significativas entre ambos valores de reconocimiento.

Otro factor importante para calcular las prestaciones del sistema para la tarea de VAQA es calcular el número de Entidades Nombradas que han sido reconocidas correctamente. En los resultados nos referimos a este factor como Cobertura de Entidades Nombradas.

6.1.2. Métricas para el modelo de lenguaje

La mejor manera de evaluar las prestaciones del modelo de lenguaje es incluirlo en el sistema de reconocimiento (o el de VAQA) y analizar las prestaciones de toda la aplicación. No obstante, existen métricas para modelo de lenguaje que evalúan, de forma independiente de la salida del reconocedor, las prestaciones del modelo. Aunque estas medidas no garantizan mejoras en el reconocimiento, sí que suelen tener una relación directa en los valores de reconocimiento y los posteriores resultados sobre la fase de QA.

La perplejidad, en modelado de lenguaje, se puede utilizar como índice de la capacidad que tiene un modelo de lenguaje para, una vez determinada una secuencia inicial de palabras, predecir la continuación de esta secuencia. Además de su interés como parámetro de diseño a tener en cuenta para evaluar las prestaciones de un sistema, permite comparar diferentes modelos en función de su adecuación a un determinado lenguaje objetivo.

La perplejidad del modelo de lenguaje para un conjunto de frases de prueba (T) se define como:

$$PPL(T) = 2^{-\frac{1}{|t|}\sum_{W_i \in T} P(W_i)} \tag{6.2}$$

donde $P(W_i)$ es la probabilidad asignada a la frase W_i.

Para evaluar el impacto del modelado de lenguaje en la tarea que estamos tratando es importante conocer las palabras fuera del vocabulario del conjunto de test, es decir, aquellas palabras que el modelo de lenguaje no ha sido capaz de reconocer *(Out-of-vocabulary Words, OOV)*.

Para saber el impacto de las entidades nombradas que han sido reconocidas correctamente también es interesante tener un indicador de aquellas que formaban parte del vocabulario del modelo de lenguaje. Este valor es indicativo de cuántas entidades nombradas pueden ser recuperadas por el reconocedor. Si la cobertura de entidades nombradas (sobre el test) del modelo de lenguaje se acerca a la cobertura de entidades nombradas de la salida del reconocedor, es un indicativo de que el reconocedor funciona bien, independientemente que el valor de cobertura de NE tras el reconocimiento sea bajo. No obstante, una baja cobertura del modelo de lenguaje sobre las entidades nombradas significa una mala modelización para la tarea.

En los resultados nos referiremos a este valor como la inversa de la cobertura, es decir, las entidades del conjunto de test que están fuera del vocabulario: *NE Out-of-Vocabulary (NE_OOV)*.

6.1.3. Métricas relacionadas con la Búsqueda de Respuesta

Se puede evaluar la calidad de la fase de extracción de pasajes a partir del conjunto ordenado de pasajes devueltos. Por tanto se utilizan métricas de evaluación usadas de forma común en tareas de recuperación de información.

Cobertura de Pasajes Recuperados. Se trata de un factor clave para la evaluación de los pasajes devueltos, este factor indica cuántas veces el sistema ha sido capaz de recuperar al menos un pasaje que contenga la respuesta a la pregunta formulada. Se calcula de la siguiente forma:

$$\frac{1}{|Q|} \sum_{q \in Q} r(P_q) \tag{6.3}$$

donde q es una pregunta del conjunto de test Q, P_q son los pasajes devueltos por el sistema para la pregunta q, y r es una función definida como:

$$r(P_q) = \begin{cases} 1 & \text{si } \exists p \in P_q \text{ es decir, } p \text{ contiene la respuesta a la pregunta } q \\ 0 & \text{si no} \end{cases} \tag{6.4}$$

Como se puede observar, este valor no se ve influido por los errores de clasificación de la pregunta y de la búsqueda de respuesta, si no que depende únicamente del motor de búsqueda de pasajes y de la entrada de este, es decir, la salida del reconocedor.

Discounted Cumulative Gain y Normalized-Discounted Cumulative Gain. Otra medida que se puede utilizar para medir las prestaciones del reconocedor en la posterior fase de búsqueda de pasajes es el DCG (Discounted Cumulative Gain) y el nDCG (Normalized Discounted Cumulative Gain)

DCG se basa en la relevancia de la información obtenida basándose en la posición que ocupa en la lista de resultados.

Para este caso es necesario conocer previamente un conjunto de documentos que se consideren relevantes para la consulta. En esta tarea, el conjunto de documentos se ha construido utilizando patrones realizados a mano y expresiones regulares para detectar si el pasaje contiene la respuesta o no.

El valor DCG para un ranking de π pasajes se calcula como.

$$DCG_\pi = rel_1 + \sum_{i=2}^{\pi} \frac{rel_i}{\log_2 i} \tag{6.5}$$

donde rel_i es el grado de relevancia del pasaje en la posición i. Para los experimentos realizados el valor utilizado ha sido de $\pi = 30$.

Debido a que este valor depende del tipo de pregunta y el número de pasajes devuelto por diferentes sistemas, para poder comparar prestaciones se utiliza el valor DCG Normalizado (nDCG) y se calcula cómo:

$$nDCG_\pi = \frac{DCG_\pi}{IDCG_\pi} \tag{6.6}$$

donde $IDCG_\pi$ es el DCG "ideal" obtenido tras reordenar todos los pasajes relevantes y considerándolos los primeros en la lista devuelta de pasajes en orden de relevancia.

De la misma forma que la cobertura de Pasajes este parámetro es independiente de la respuesta extraída por el sistema de QA, ambas métricas dependen en gran medida de los errores de la fase de reconocimiento.

Mean Reciprocal Rank. Se basa en la posición de la primera respuesta correcta en la lista ordenada de respuestas. En el caso de que la respuesta no se halle en la lista de pasajes recuperados el valor de la posición se considera infinito, y al realizar la inversa $\frac{1}{rank_i}$ toma el valor 0.

Se fórmula como:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{6.7}$$

dónde $|Q|$ es el número de cuestiones y para cada pregunta i se tiene la posición de la primera respuesta como $rank_i$. En el mejor de los casos esta posición es 1.

En la evaluación realizada se evalúa la respuesta correcta a partir de los pasajes devueltos. Se considera la respuesta correcta si el pasaje incluye uno de los patrones que representan la respuesta. En otras palabras, estamos suponiendo un sistema de extracción de respuesta que es capaz de extraer la respuesta correcta si se encuentra en los pasajes devueltos.

Mean Average precision. Este valor devuelve la media de la average precision ($Avg.P$) para un conjunto de consultas $|Q|$:

$$MAP = \frac{\sum_{i=1}^{|Q|} AvgP(i)}{|Q|} \tag{6.8}$$

El término Average Precision, indica para el conjunto de respuestas de tamaño devuelto N, la relevancia de los documentos devueltos para la respuesta correcta. Se calcula como:

$$AvgP = \frac{\sum_{r=1}^{N} (P(r)rel(r))}{|N_r|} \tag{6.9}$$

donde $P(r)$ indica el número de elementos relevantes tomados en el rango r, y $rel(r)$ indica si el documento r obtenido es relevante para la pregunta o no.

De las métricas evaluadas sólo Mean Reciprocal Rank depende de la respuesta extraída por el sistema. Sin embargo, como se considera que una respuesta es correcta si está

incluida en un pasaje, ninguna de las métricas mostradas están influidas por el sistema de extracción de respuestas.

6.2. Diseño de experimentos

Para evaluar el rendimiento del modelado de lenguaje basado en categorías, se han evaluado diferentes modelos de lenguaje sobre la tarea descrita en el apartado 3.1 (CLEF Contest). A continuación se muestran los modelos de lenguaje evaluados:

Modelo entrenado a partir de las preguntas de entrenamiento. Este modelo ha sido entrenado a partir las preguntas de entrenamiento como se describe en la sección 4.1.1. El objetivo es poder comparar la aproximación expuesta en este trabajo con este tipo de modelos.

Modelo entrenado con el conjunto de documentos. De la misma forma que se ha experimentado con el modelo de preguntas, se muestran resultados en base al uso repositorio de documentos para entrenar el modelo de lenguaje como se ha presentado en 4.1.2.

Modelo Simple de Entidades Nombradas. Se presenta un modelo de categorías construido a partir de la aproximación mostrada en 4.2.

Modelo Cerrado de Entidades Nombradas. De la misma forma que el modelo anterior, se trata de un modelo categorizado. Pero con la diferencia de que se han incluido todas las entidades nombradas del conjunto de test en la lista de entidades. Debido a que se está incluyendo información del conjunto de test en el entrenamiento del modelo de lenguaje, este modelo se ha utilizado solamente para comprobar el impacto de ampliar la categoría de las entidades nombradas aún teniendo todas las entidades de test.

Modelo de Entidades Modificado. Este modelo se ha construido de la misma forma que el modelo anterior pero se ha añadido la mejora de la Unión fonética de Entidades presentada en 4.3.3.

Modelo de Entidades Nombradas y Nombres Comunes Modificado. Este modelo utiliza las misma categoría de Entidades Nombradas que el modelo anterior (incluyendo la Unión fonética de Entidades), pero se categorizan también los nombres comunes.

Debido a que el interés de la aproximación de generalización de entidades nombradas reside en el tamaño de entidades incluidas en el modelo, los resultados para los modelos categorizados se muestran en base al tamaño de entidades. Para el caso de la categoría de nombres comunes el tamaño de la categoría permanece fijo para toda la experimentación.

Capítulo 7

Resultados

En este apartado se muestran los resultados obtenidos tras realizar los experimentos de reconocimiento y posterior fase de recuperación de pasajes. Dependiendo del experimento realizado y del aspecto en el que se quiera focalizar, se muestran diferentes métricas para cada uno de los modelos propuestos. De forma paralela se presenta una discusión de los resultados para cada uno de los experimentos realizados.

7.1. Perplejidad del modelo de lenguaje

Antes de empezar con la evaluación de las prestaciones del reconocedor y los resultados obtenidos de la búsqueda de pasajes, se ha evaluado la adaptación del modelado de lenguaje a la tarea propuesta. La tabla 7.1 muestra los valores de perplejidad de los diferentes modelos de lenguaje utilizados para el conjunto de test. También se muestra la relación de palabras de test que quedan fuera del vocabulario (*Out of Vocabulary words, OOV*) con cada modelización. El conjunto de pruebas consta de 200 frases, y $1,622$ palabras (651 sin contar repeticiones).

Se observa que utilizar el modelo aprendido con el conjunto de documentos tiene una perplejidad mucho mayor que utilizar estructuras de preguntas que se asemejan a las del conjunto de test. No obstante el problema de palabras fuera del vocabulario es el principal problema que se tiene en los modelos de lenguaje basados en preguntas.

	Perplejidad				OOV
	1-grama	**2-gramas**	**3-gramas**	**4-gramas**	
Modelo Preguntas	180,560	37,930	33,062	33,176	0,141
Modelo Documentos	1351,190	332,597	249,957	239,299	0,022
Modelo NE	112,048	34,691	31,789	31,627	0,082
Modelo NE/CN	71,886	30,636	28,957	29,010	0,075

Tabla 7.1: Perplejidad de los Modelos del Lenguaje

7.2. Modelos de referencia, categorizados y sin categorizar

En la siguiente experimentación se muestra una tabla resumen (tabla 7.2) de los resultados obtenidos para los modelos no categorizados: modelo entrenado con las preguntas, modelo entrenado con el corpus de documentos; y los modelos de categorías presentados en el apartado anterior. Paralelamente se muestran los resultados obtenidos utilizando las transcripciones correctas, es decir con WER 0 % y debidamente puntuadas y delimitadas.

Modelo	**WER**	**NE Cov**	**NE OOV**	**PR Cov**	**nDCG**	**MAP**
Referencia	0,000	1,000	0,000	0,815	0,584	0,457
ML preguntas	0,329	0,495	0,461	0,505	0,335	0,256
ML documentos	0,374	0,519	0,000	0,550	0,353	0,265
ML NE Simple	0,288	0,665	0,189	0,610	0,409	0,308
ML NE Mod	0,280	0,680	0,189	0,605	0,406	0,306
ML NE/NC Mod	0,273	0,621	0,189	0,575	0,392	0,300

Tabla 7.2: Resumen de los diferentes modelos

Se aprecia un claro empeoramiento de las prestaciones al utilizar voz como entrada del sistema. En este caso la utilización de modelos categorizados mejora las prestaciones que el resto de modelos pero aún así distan bastante de los resultados obtenidos utilizando preguntas escritas y correctamente formuladas.

7.3. Modelo Abierto y Cerrado de Entidades Nombradas

En este experimento se muestran las prestaciones del sistema al incluirse todas las entidades del conjunto de test en el modelo de lenguaje. Se debe tener en cuenta que este experimento se realiza sólo para comprobar el impacto en las prestaciones del sistema si se amplia el conjunto de entidades pese a seguir teniendo todas las entidades en el modelo de lenguaje. En el resto de experimentos no se ha tomado ningún dato de test para el entrenamiento de los modelos de lenguaje.

La tabla 7.3 muestra los resultados obtenidos para el modelo de entidades cerrado (el modelo que incluye todas las entidades de test). En este experimento el objetivo es mostrar como influyen las entidades añadidas al modelo de lenguaje, es decir, al añadir más entidades al modelo de lenguaje cuánto influye negativamente en el reconocimiento de las entidades que ya se tenían tras la ampliación del vocabulario. A su vez ayuda a comprender en el caso del modelo de Entidades Abierto el impacto de la ampliación del vocabulario en las entidades reconocidas utilizando vocabularios de menor tamaño.

#NE	WER	NE Cov	PR Cov	nDCG	MAP	MRR
4.000	0,253	0,791	0,635	0,440	0,343	0,380
8.000	0,251	0,786	0,640	0,444	0,345	0,384
12.000	0,249	0,782	0,635	0,443	0,345	0,385
16.000	0,247	0,782	0,635	0,445	0,347	0,392
20.000	0,245	0,782	0,640	0,444	0,344	0,388
24.000	0,243	0,782	0,650	0,449	0,348	0,391
28.000	0,241	0,782	0,655	0,452	0,349	0,394
32.000	0,239	0,782	0,655	0,454	0,351	0,398
36.000	0,237	0,782	0,660	0,458	0,355	0,403
40.000	0,236	0,782	0,660	0,458	0,355	0,403
44.000	0,235	0,782	0,660	0,462	0,360	0,408
48.000	0,235	0,782	0,660	0,465	0,364	0,412

Tabla 7.3: Resultados del Modelo Modificado de Entidades Cerrado

A su vez, la tabla 7.4 muestra los mismos resultados pero esta vez incluyendo las entidades extraídas del corpus de documentos. Evidentemente, tanto el WER como la

cobertura de NE son peores que en el caso anterior. Las diferencias se reducen conforme se amplía el vocabulario en el modelo abierto de entidades, aún así las diferencias son significativas.

#NE	WER	NE Cov	NE OOV	PR Cov	nDCG	MAP	MRR
4.000	0,319	0,558	0,461	0,500	0,321	0,237	0,277
8.000	0,312	0,617	0,359	0,545	0,349	0,255	0,305
12.000	0,307	0,636	0,301	0,560	0,366	0,271	0,321
16.000	0,303	0,636	0,277	0,565	0,375	0,281	0,330
20.000	0,301	0,641	0,267	0,575	0,379	0,282	0,334
24.000	0,299	0,636	0,218	0,585	0,385	0,288	0,341
28.000	0,296	0,646	0,199	0,590	0,391	0,293	0,344
32.000	0,293	0,665	0,199	0,605	0,401	0,301	0,354
36.000	0,291	0,660	0,194	0,605	0,401	0,300	0,354
40.000	0,288	0,665	0,189	0,610	0,409	0,308	0,363
44.000	0,287	0,665	0,180	0,615	0,410	0,309	0,363
48.000	0,285	0,675	0,175	0,630	0,417	0,312	0,365

Tabla 7.4: Resultados del Modelo de Entidades Abierto

En cuanto a los resultados del conjunto cerrado, se ve como la adición de más entidades al modelo de lenguaje no afecta prácticamente a las entidades previamente reconocidas. Es más, hay un conjunto de entidades, alrededor del 20 %, que no se reconocen en ningún caso, a pesar de estar incluidas en el modelo de lenguaje. Esto se debe a varios factores, uno de ellos es la existencia de entidades pronunciadas en otro idioma, como se ha explicado en 4.3.1. Por lo que el transcriptor ortográfico-fonético no es capaz de relacionar la fonética pronunciada por el usuario con la entidad. Otro factor es la confusión de entidades con otras palabras (pueden ser otra entidad o no) que tiene una fonética idéntica o similar (4.3.3). Otros casos, se deben a errores intrínsecos del reconocedor o incluso mala pronunciación o incluso, ruido en la señal acústica.

Al aumentar el conjunto de entidades nombradas, aún teniendo todas las entidades del test en el modelo de lenguaje, el WER del reconocedor continúa mejorando. Este fenómeno sucede porque se añaden al vocabulario palabras del test que no son entidades nombradas que se reconocen correctamente. Aunque la diferencia no es muy notable, se aprecia una mejora en la fase de Recuperación de Pasajes.

7.4. Modelo de Entidades Modificado y de Nombres Comunes

A continuación se muestran la misma experimentación realizada en el apartado anterior pero en este caso se introducen dos nuevos modelos: Modelo de Entidades Modificado y el Modelo de Entidades Nombradas y Nombres Comunes.

La tabla 7.5 muestra las métricas descritas para el modelo de entidades creado utilizando la unificación fonética presentado en el apartado 4.3.3 (Modelo de Entidades Modificado). Mientras que la tabla 7.6 utiliza un modelo con una nueva categoría que son los nombres comunes.

#NE	WER	NE Cov	NE OOV	PR Cov	nDCG	MAP	MRR
4.000	0,313	0,553	0,461	0,500	0,316	0,231	0,271
8.000	0,305	0,612	0,359	0,550	0,349	0,255	0,301
12.000	0,301	0,631	0,301	0,555	0,363	0,269	0,317
16.000	0,298	0,641	0,277	0,575	0,381	0,285	0,334
20.000	0,295	0,646	0,267	0,580	0,382	0,285	0,337
24.000	0,292	0,660	0,218	0,590	0,391	0,294	0,343
28.000	0,289	0,660	0,199	0,590	0,393	0,296	0,346
32.000	0,285	0,670	0,199	0,600	0,402	0,303	0,356
36.000	0,282	0,675	0,194	0,600	0,401	0,302	0,356
40.000	0,280	0,680	0,189	0,605	0,406	0,306	0,360
44.000	0,278	0,684	0,180	0,615	0,413	0,312	0,366
48.000	0,276	0,684	0,175	0,620	0,416	0,315	0,367

Tabla 7.5: Resultados del Modelo de Entidades Modificado

Si comparamos los resultados obtenidos mediante el modelo de Entidades Modificado con el caso anterior se observa una leve mejora, alrededor de un punto, en los valores de reconocimiento (WER y cobertura de entidades). Esta mejora no se presenta en los resultados posteriores de recuperación de información, que se mantienen prácticamente iguales. Se debe a que se han recuperado unas pocas NE pero que debido al contexto en el que se encuentran y naturaleza de la pregunta en sí, el sistema de recuperación de pasajes no ha sido capaz de encontrar pasajes relevantes y por ende la respuesta a la pregunta afectada por esta mejora.

#NE	WER	NE Cov	NE OOV	PR Cov	nDCG	MAP	MRR
4.000	0,297	0,500	0,461	0,490	0,312	0,229	0,262
8.000	0,291	0,549	0,359	0,510	0,333	0,247	0,287
12.000	0,286	0,573	0,301	0,525	0,352	0,265	0,307
16.000	0,283	0,587	0,277	0,535	0,365	0,278	0,320
20.000	0,281	0,597	0,267	0,535	0,366	0,280	0,320
24.000	0,279	0,607	0,218	0,545	0,375	0,288	0,330
28.000	0,277	0,612	0,199	0,545	0,376	0,288	0,330
32.000	0,276	0,612	0,199	0,555	0,381	0,292	0,335
36.000	0,275	0,617	0,194	0,565	0,386	0,295	0,337
40.000	0,273	0,621	0,189	0,575	0,392	0,300	0,342
44.000	0,272	0,631	0,180	0,580	0,398	0,306	0,347
48.000	0,271	0,636	0,175	0,585	0,402	0,309	0,349

Tabla 7.6: Resultados del Modelo Modificado de Entidades Nombradas y Nombres Comunes

Mucho más significativos son los cambios obtenidos en el modelo de entidades y nombres comunes. Se observa una pequeña mejora en el error de reconocimiento. No obstante, la cobertura de entidades nombradas se ha visto afectada en 5 puntos. La cobertura de pasajes (PR Cov.) es la que más se ha visto perjudicada con un descenso de 3 puntos, recordemos que esta métrica indica si se ha recuperado un pasaje relevante y no tiene en cuenta su posición, ni la respuesta extraída. Un descenso proporcional se encuentra en los valores de MRR, MAP y nDCG, por tanto se ve que existe una relación fuerte entre el reconocimiento de Entidades Nombradas aún habiendo mejorado el WER.

7.5. Resultados utilizando las n-bests del reconocedor

En este apartado se muestran los resultados tomando varias hipótesis como salida del reconocedor. Se trata de las n-mejores hipótesis devueltas por el reconocedor de voz. Los resultados mostrados (tabla 7.7) se han obtenido utilizando un oráculo que es capaz de seleccionar entre las n-bests la mejor hipótesis. El valor de n utilizado en los resultados mostrados es de 10.

	1-best		10-best		Diferencia	
#NE	**WER**	**NE Cov**	**WER**	**NE Cov**	**WER**	**NE Cov**
4.000	0,297	0,500	0,237	0,553	0,060	0,053
8.000	0,291	0,549	0,232	0,612	0,059	0,063
12.000	0,286	0,573	0,229	0,631	0,058	0,058
16.000	0,283	0,587	0,227	0,641	0,056	0,053
20.000	0,281	0,597	0,227	0,631	0,054	0,034
24.000	0,279	0,607	0,227	0,650	0,052	0,044
28.000	0,277	0,612	0,227	0,646	0,050	0,034
32.000	0,276	0,612	0,227	0,646	0,049	0,034
36.000	0,275	0,617	0,226	0,650	0,048	0,034
40.000	0,273	0,621	0,227	0,641	0,047	0,019
44.000	0,235	0,782	0,227	0,650	0,008	-0,131
48.000	0,235	0,782	0,227	0,655	0,008	-0,126

Tabla 7.7: Resultados tomando las 10-best del Modelo Simple de Entidades

Las mejoras son significativas, y los resultados se acercan a los obtenidos con el conjunto cerrado. Esto presenta la problemática añadida de ser capaz de crear un oráculo capaz de seleccionar de entre todas las salidas del reconocedor la mejor para la tarea. O en un caso más práctico, utilizar un sistema combinado: un sistema de reconocimiento automático del habla y un sistema de recuperación de información capaz de explotar la información de las n-bests hipótesis para mejorar así las prestaciones del sistema. Los resultados obtenidos son una cota superior de las prestaciones del sistema propuesto.

7.6. Tamaño del conjunto de Entidades Nombradas

En la experimentación presentada hasta el momento se han mostrado los resultados obtenidos variando el conjunto de entidades nombradas entre valores de 4,000 y 48,000 entidades nombradas. La figura 7.1 muestra la evolución de las prestaciones del sistema para un conjunto de entidades más grande. En la figura se muestra la evolución de la cobertura de Entidades Nombradas y WER de los modelos cerrado y abierto y también se muestran las entidades fuera del vocabulario para el conjunto abierto.

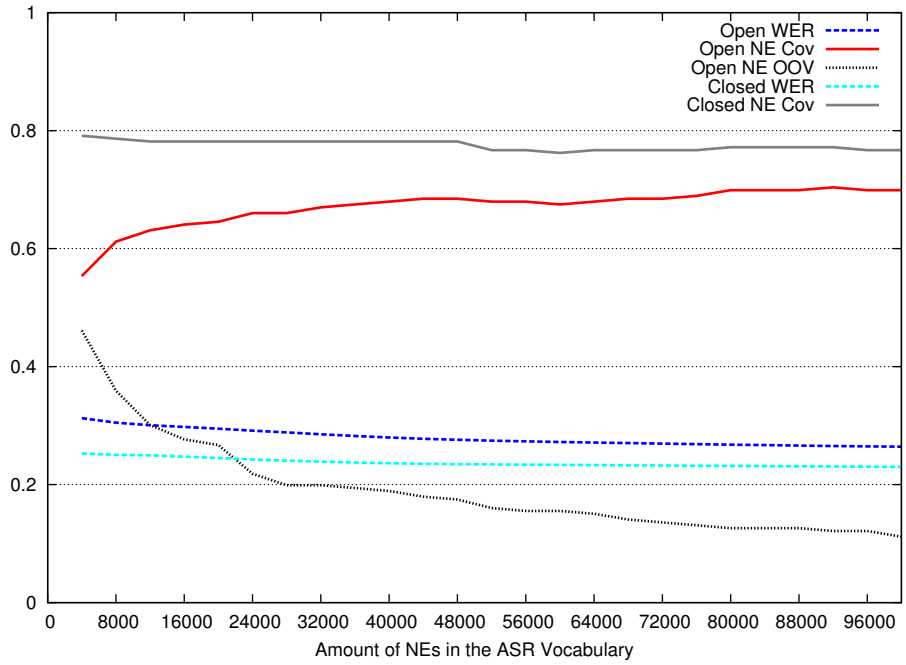

Figura 7.1: Evolución de las prestaciones del reconocedor para el modelo abierto y cerrado

En ambos modelos (abierto y cerrado) el WER se reduce hasta conjuntos de entidades de entre 40,000 y 50,000 donde se mantiene prácticamente constante.

Por otro lado, la cobertura de entidades nombradas se mantiene constante para el conjunto cerrado y crece en el modelo abierto de NE con una tendencia similar conforme se añaden entidades nombradas al vocabulario. Entre 16,000 y 24,000 entidades se observa una desaceleración de la mejora de la cobertura y a partir de 50,000 el crecimiento es muy leve, en otras palabras, conforme se añaden más entidades en el modelo de lenguaje hacen falta conjuntos mucho más grandes de entidades para reconocer correctamente entidades que están en el conjunto de test y no habían sido previamente reconocidas, tal como se aprecia en linea de cobertura de Entidades sobre el conjunto de test.

Otro detalle importante, es ver que el crecimiento de las entidades nombradas de test incluidas en el vocabulario (inversa de las entidades fuera del vocabulario) y por extensión, el correcto reconocimiento de estas tras usar el ASR sigue una tendencia similar a las realizadas para el cálculo de umbrales de la herramienta de etiquetamiento POS en la sección 4.2.3.

7.7. Impacto del WER y Cobertura de Entidades

Como ya se ha comentado, hay diferentes estudios que muestran la relación entre el WER y el reconocimiento de las entidades nombradas; y los resultados obtenidos en la fase de búsqueda de pasajes. En este apartado se muestra gráficamente esta relación para los resultados obtenidos en la experimentación.

En la figura 7.2 se muestran las diferentes métricas de recuperación de pasajes (Cobertura de pasajes, nDCG, MAP y MRR) respecto el WER obtenido en la experimentación. Mientras que la figura 7.3 muestra los mismos resultados pero tomando la cobertura de entidades. Los resultados mostrados se refieren a los modelos categorizados de entidades nombradas y nombres comunes.

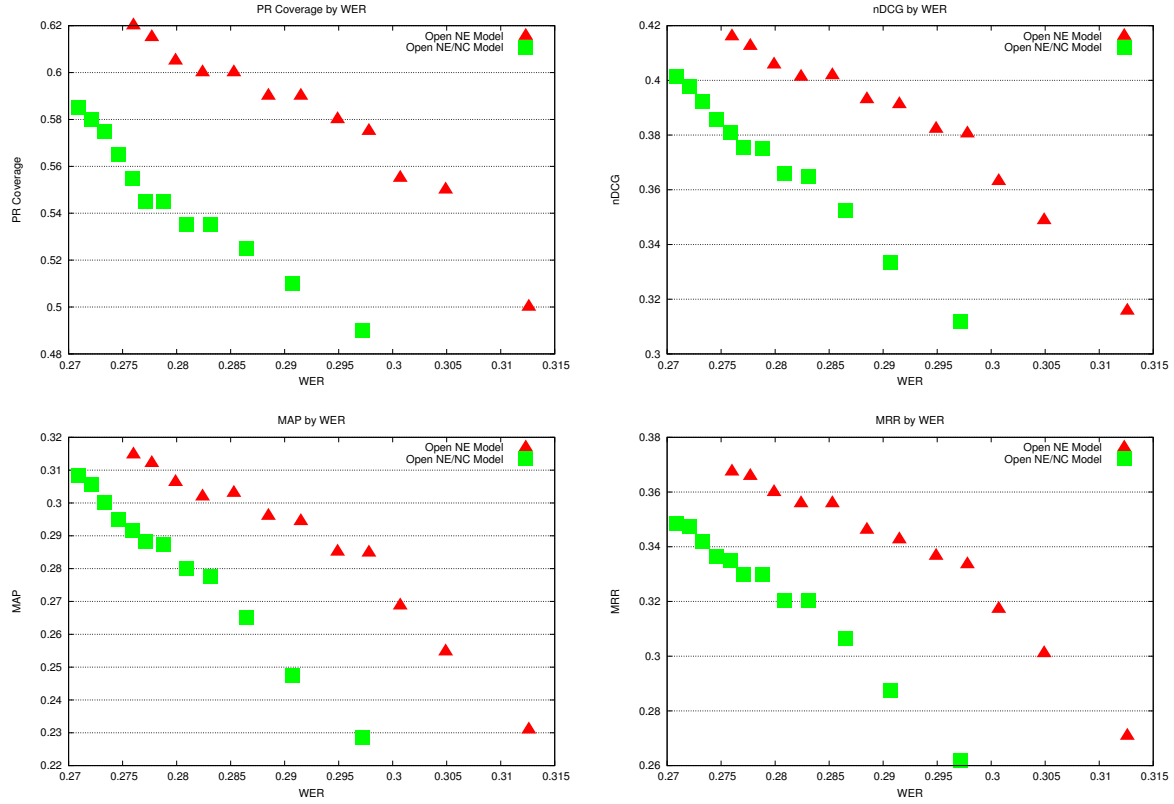

Figura 7.2: Relación entre el WER y los resultados en la fase de PR

En ambas figuras se observa una relación directa entre las métricas evaluadas, aunque existen diferencias entre los diferentes modelos utilizados. No obstante, se ve una relación más clara entre el reconocimiento de entidades nombradas y las métricas analizadas puesto que existe menor diferencia entre ambos modelos. En el caso de la de cobertura de pasajes, cuyos valores dependen más directamente de la fase de reconocimiento que el resto de métricas, se observa una relación de deterioro directa en ambos modelos analizados.

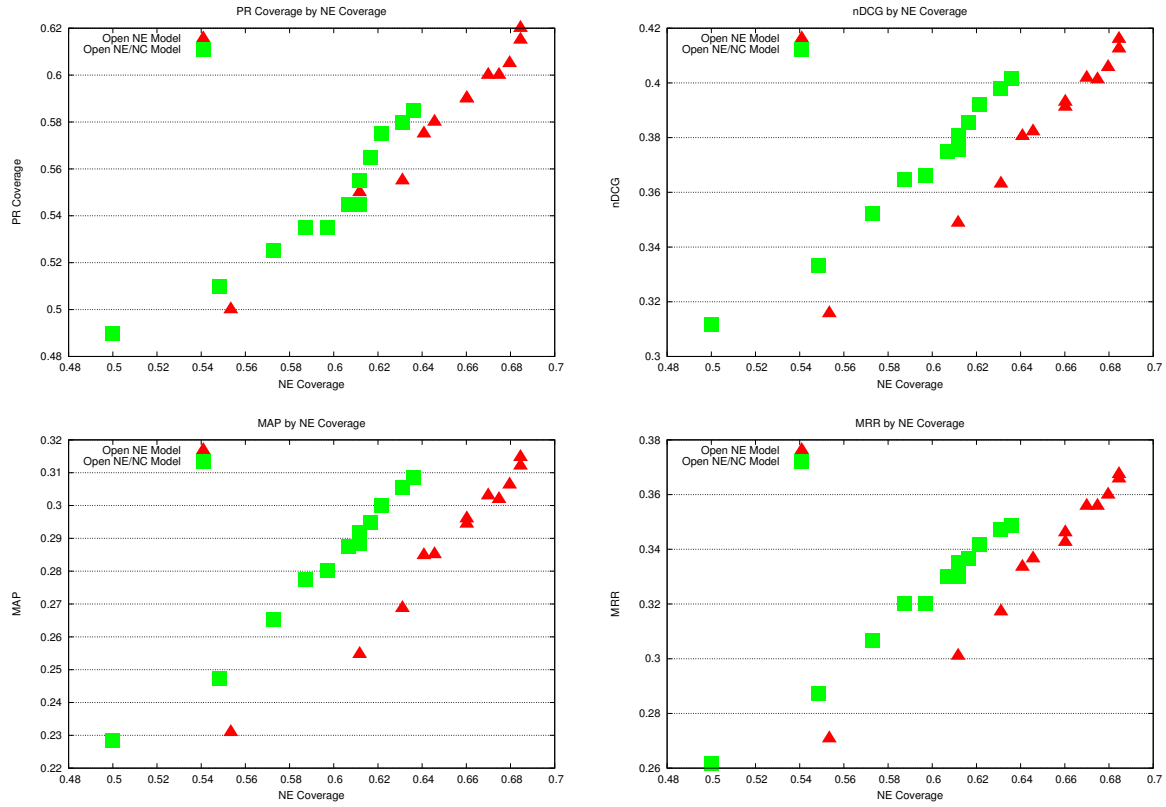

Figura 7.3: Relación entre la cobertura de entidades y los resultados en la fase de PR

7.8. Estudio del reconocimiento de Entidades Nombradas

Las figuras 7.4 y 7.5 muestran con más detalle como varia el reconocimiento de las entidades nombradas del conjunto de test conforme el incremento del conjunto de entidades en el modelo de lenguaje categorizado. Para cada incremento se puede observar el número de entidades que pasan a ser reconocidas nuevamente (In) y las que se han perdido a causa del incremento en el modelo (Out). Las gráficas muestran también, la cobertura total de las entidades para cada modelo que varía conforme las entidades nuevamente reconocidas y perdidas que están representadas conjuntamente en las gráficas.

La figura 7.4 muestra la evolución para el conjunto abierto de entidades, mientras que la figura 7.5 muestra la evolución en el conjunto cerrado de entidades. A parte, en el conjunto abierto de entidades se muestra el conjunto de entidades nombradas fuera del vocabulario.

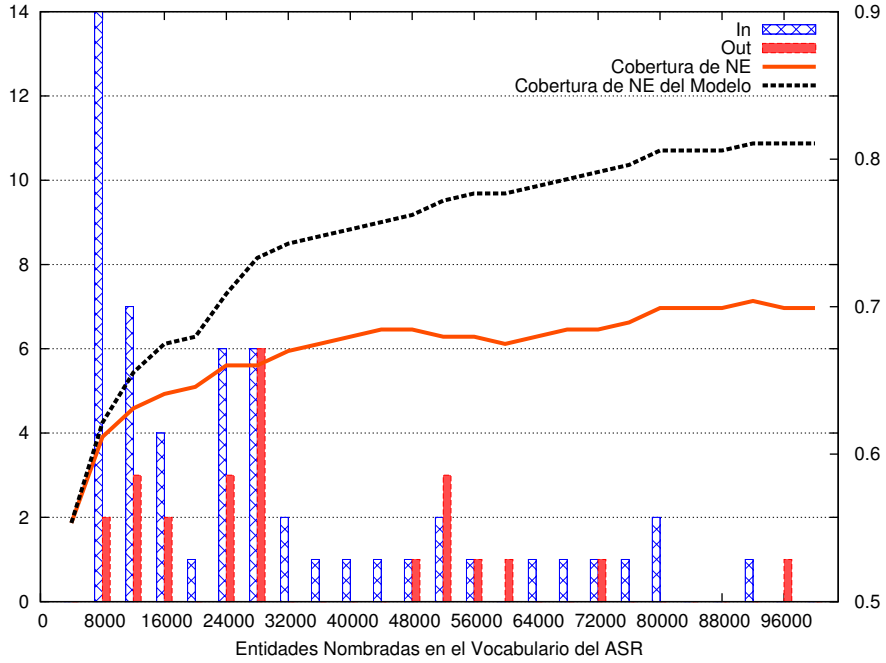

Figura 7.4: Evolución de las entidades del reconocedor para el modelo abierto de entidades

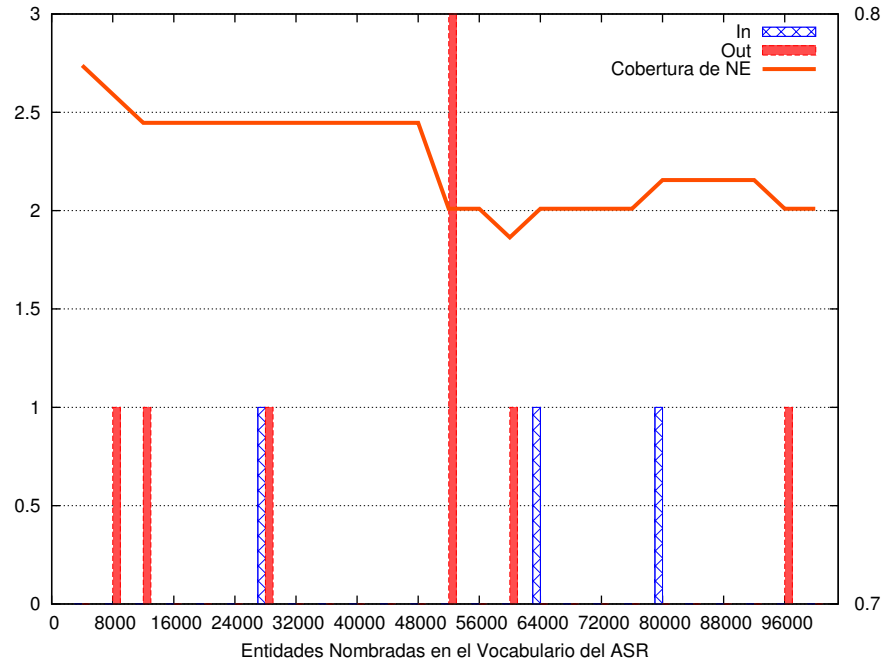

Figura 7.5: Evolución de las entidades del reconocedor para el modelo cerrado de entidades

Del modelo abierto de entidades cabe remarcar que conforme se van incorporando nuevas entidades al modelo de lenguaje se van reconociendo en la fase de reconocimiento automático de voz, aunque en algunos casos se pierden ciertas entidades y conforme se aumenta el número de entidades la diferencia entre las entidades incluidas en el modelo de lenguaje y las reconocidas por el ASR es mayor.

La figura 7.6 muestra con más detalle el reconocimiento de las entidades nombradas en el modelo categorizado de NE. El gráfico muestra diferentes cuadriculas (en total 6), cada una de esta representa las entidades nombradas del conjunto de test para el modelo utilizando conjuntos de diferentes tamaños en la categoría de entidades. Dentro de cada cuadrícula, cada celda representa una entidad nombrada en concreto y el color utilizado para representarla indica el estado de dicha entidad en cada modelo. El código de colores se describe como sigue:

- Azul: Representa a una NE que ha sido reconocida correctamente. Se marcan sólo en azul aquellas entidades que no se reconocen en modelos con menor número de entidades nombradas.
- Rojo: Representa aquellas entidades del modelo que no han sido reconocidas, pero que podían ser reconocidas por el modelo de lenguaje. Es decir, aquellas que están en el vocabulario de entidades nombradas. A diferencia de las entidades mostradas en color naranja estas entidades no han sido reconocidas utilizando un conjunto más pequeño de entidades.
- Verde: Se muestran en este color aquellas entidades bien reconocidas pero que se pueden reconocer utilizando un modelo menor de entidades.
- Naranja: Son entidades mal reconocidas pero que habían sido bien reconocidas utilizando un modelo de menor tamaño de entidades. Representan errores derivados de aumentar el número de elementos en la categoría de entidades nombradas.
- Negro: Son aquellas entidades no reconocidas que no se encuentran en el vocabulario del modelo de lenguaje y por tanto es muy difícil de que se puedan reconocer correctamente.

Este tipo de gráfico es interesante para estudiar con más detalle las causas que provocan errores en las entidades nombradas. Se puede ver que gran parte de las entidades

nombradas se reconocen correctamente al aumentar el conjunto de entidades. Sin embargo, aparece un número considerable de entidades fuera del vocabulario debido a la modelización utilizada. También aparecen entidades mal reconocidas debido al aumento del tamaño de la categoría (naranja) con una frecuencia similar a las mal reconocidas (rojo).

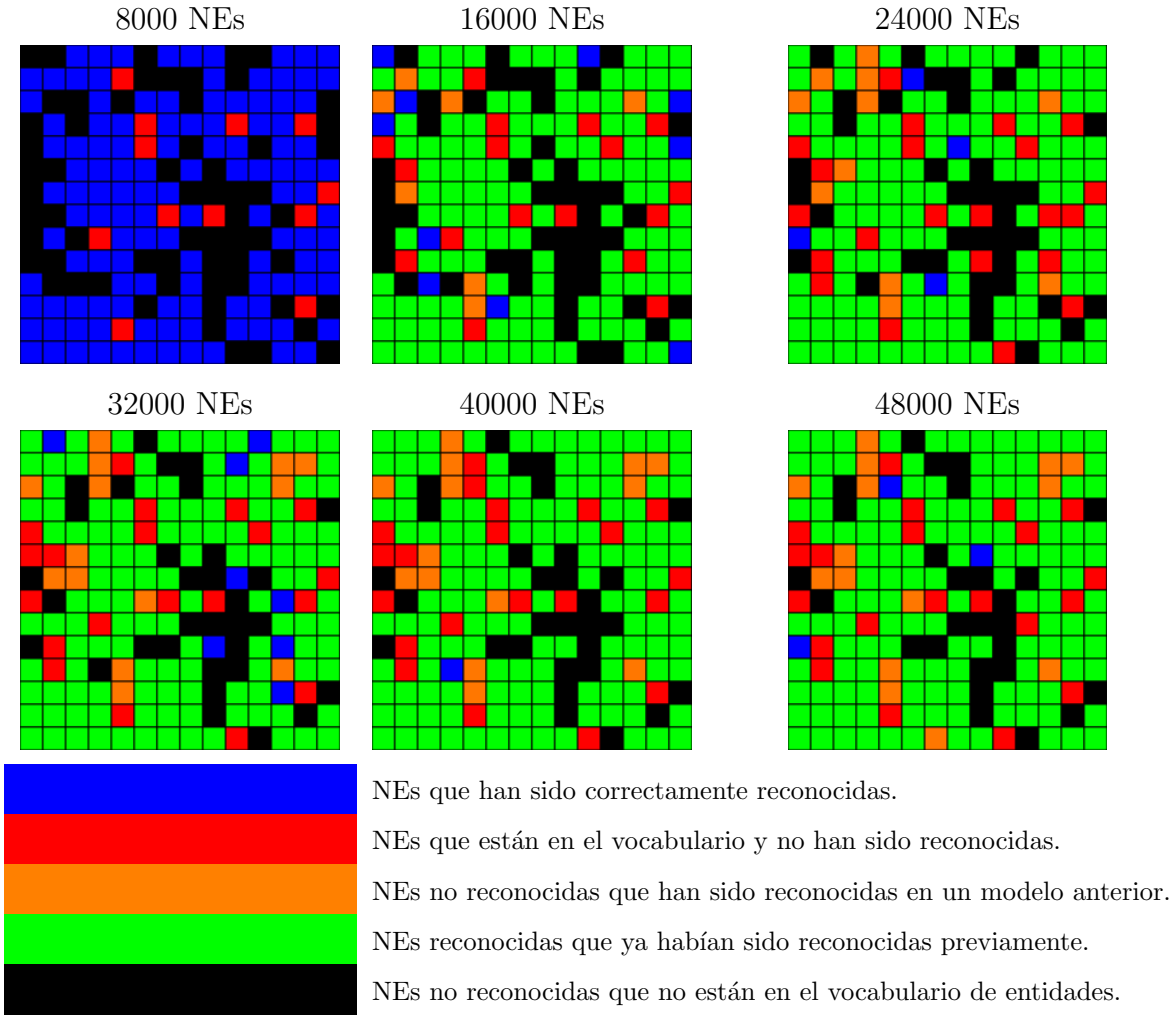

Figura 7.6: Reconocimiento de las entidades del conjunto de test

Capítulo 8

Conclusiones y trabajo futuro

8.1. Conclusiones

En este trabajo se ha presentado una aproximación para la tarea de Búsqueda de Respuesta dirigida por la Voz, es decir, la entrada del sistema de Búsqueda de Respuesta se toma a través de pronunciaciones en vez de la utilización de frases correctamente escritas. Se ha mostrado la importancia de utilizar modelos de lenguaje específicos que se adapten a la tarea de QA y la importancia del correcto reconocimiento de las preguntas y el posterior impacto en las prestaciones del sistema global de VAQA. Se ha visto que con la combinación de un conjunto relativamente pequeño de preguntas (1,600) y una lista de entidades nombradas se es capaz de mejorar las prestaciones del sistema respecto al entrenamiento con grandes conjuntos de documentos, con la consiguiente mejora de rendimiento que esto supone.

Además, se ha puesto especial interés en estudiar como el incremento de las entidades nombradas presentes en el vocabulario del sistema de reconocimiento automático, para reducir el número de entidades fuera del vocabulario de la tarea, afecta a los resultados del reconocimiento y rendimiento global del sistema.

Se ha estudiado el comportamiento del sistema considerando diferentes métricas de reconocimiento (WER, Cobertura de Entidades) y de recuperación de información (Cobertura de recuperación de pasajes, nDCG, MAP y MRR). Se ha visto que al disminuir el número de Entidades Nombradas fuera del vocabulario del conjunto del test se ha mejorado notablemente la prestaciones del sistema.

En los experimentos realizados se muestra la importancia de un correcto reconocimiento de la consulta independientemente de la calidad del sistema de recuperación de pasajes y extracción de la respuesta. Se ha visto una relación entre el error de reconocimiento de la pregunta (WER) y el posterior análisis y extracción de la respuesta a la pregunta formulada. Pero también una relación con el reconocimiento de palabras clave como son las entidades nombradas, se ha visto que es posible recuperar la respuesta (o al menos pasajes relevantes) de una pregunta si se reconocen correctamente las palabras clave aún teniendo un tasa de WER relativamente alto (entre el 25 % y 30 %). Y también el caso contrario, tasas de WER más bajas que no han reconocido correctamente entidades nombradas (como por ejemplo, el modelo que utiliza nombres comunes) presentan peores resultados que teniendo una mayor cobertura de entidades nombradas.
También, se ha visto que es posible mejorar las prestaciones del sistema si se es capaz de explotar la información obtenida de las n-best hipótesis devueltas del sistema de reconocimiento. Para realizar esto, es necesario un sistema de recuperación de pasajes que sea capaz de trabajar con este tipo de salidas.

8.2. Trabajo Futuro

Como trabajo futuro se contemplan diversas mejoras tanto en la fase de modelado del lenguaje, como en la combinación del ASR y el sistema de recuperación de pasajes.

Se han visto diferentes aproximaciones para el reconocimiento de entidades nombradas en otros idiomas, basados en modelos acústicos híbridos o aproximaciones de transcripciones fonéticas específicas para este tipo de lenguaje.

En cuanto al modelado del lenguaje la utilización de un conjunto de preguntas conjuntamente con una lista de entidades nombradas funciona bien, pero deja fuera del vocabulario algunas palabras que no son entidades nombradas ni nombres propios, que el sistema es incapaz de recuperar. Se deben explotar formas de recuperar este tipo de palabras o construcciones utilizando modelos híbridos entre el modelo categorizado y uno de n-gramas sin categorizar, evitando perder el conocimiento sintáctico de la construcción de las consultas al sistema.

También sería interesante hacer hincapié en la extracción de entidades nombradas para tener listas más precisas de entidades y evitar así ruido introducido por el sistema de extracción de entidades.

Otro punto importante a destacar, es el desarrollo de un sistema de recuperación de pasajes que sea capaz de trabajar con varias hipótesis con sus correspondientes valores de confianza y ser capaz de obtener un ranking de pasajes basados en las múltiples salidas del reconocedor y sus valores de confianza. Además, para el sistema de VAQA es muy importante trabajar con un modelo de error, que sea capaz de omitir y recuperar errores de reconocimiento (a partir de medidas de confianza) o recuperarse de estos errores mediante medidas de similitud fonéticas.

Publicaciones relacionadas

Durante la elaboración del Trabajo Final de Máster se han presentado diversos trabajos en conferencias internacionales relacionadas con tecnologías del habla:

- Joan Pastor, Lluís-F. Hurtado, Encarna Segarra and Emilio Sanchis, "Language Modelization and Categorization for Voice-Activated QA", CIARP 2011: 16th Iberoamerican Congress on Pattern Recognition, Pucón, Chile, 2011.
- E. Segarra, L. Hurtado, J.A. Gómez, F. García, J. Planells, J. Pastor, L. Ortega, M. Calvo, E. Sanchis. A prototype of a spoken dialog system based on statistical models. *Proceedings of FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, pages 243- 246.
- Emilio Sanchis, Lluís-Felip Hurtado, Fernando García, Joan Pastor, Joaquin Planells, Encarna Segarra. Sistema de diálogo multimodal basado en modelos estadísticos. XXVII edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), 2011.

Apéndice A

A.1. Modelado de Lenguaje para la tarea QAst-09

En este apéndice se muestra la adecuación del modelo de lenguaje a las frases utilizadas para la evaluación de la tarea QAst incluido en el CLEF Contest 2009.

Se ha realizado la experimentación con las preguntas del QAst-09, en primer lugar, para probar si la aproximación utilizada se adecuaba a una tarea deferente a la analizada en el resto de experimentos de la memoria. En segundo lugar, comprobar si es factible utilizar los modelos de lenguaje entrenados con las preguntas de una tarea para otra tarea diferente, simplemente cambiando las entidades nombradas y nombres comunes para la tarea propuesta. Para está tarea se muestran sólo las métricas relativas a los modelos de lenguaje utilizados, es decir: perplejidad sobre la tarea, palabras fuera del vocabulario (OOV) y entidades nombradas fuera del vocabulario (NE_OOV).

Los modelos de lenguaje analizados son los siguientes:

Modelo entrenado con el corpus TC-STAR: Modelo de lenguaje entrenado con las transcripciones automáticas del conjunto TC-STAR. En este caso no se ha utilizado ningún tipo de categorización.

Modelo entrenado con el corpus EUROPARL: Modelo entrenado con el conjunto de actas del Parlamento Europeo (EUROPARL). Las frases están correctamente transcritas y el contenido es mucho mayor que el corpus TC-STAR.

Modelo categorizado de Entidades Nombradas: Este modelo de lenguaje ha sido entrenado con las $1,600$ preguntas de entrenamientos utilizadas para entrenar el modelo de lenguaje categorizado de la tarea del CLEF (sección 3.1). Pero a diferencia de este, se

incluyen las entidades nombradas extraídas del corpus EUROPARL, tras la extracción y filtrado de entidades sobre dicho corpus.

Modelo categorizado de Entidades Nombradas y Nombres Comunes: Este modelo ha sido entrenado de la misma forma que el modelo anterior pero incluye además, la categoría de Nombres Comunes con los nombres comunes extraídos sobre el corpus EUROPARL.

La figura A.1 muestra las métricas analizadas sobre los modelos descritos. El número de palabras del conjunto de preguntas es de $1,930$.

		TC-STAR	**EUROPARL**	**Modelo NE**	**Modelo NE/NC**
Perplejidad	**1-grama**	900,748	1667,280	124,166	79,220
	2-gramas	247,843	633,613	44,501	32,772
	3-gramas	212,333	508,111	44,527	33,176
	4-gramas	208,076	467,414	45,834	33,413
OOV		0,047	0,015	0,090	0,071
NEOOV		0	0	0,168	0,168

Tabla A.1: Métricas de los modelos de lenguaje para la tarea QAst2009.

Se puede ver que de la misma forma que sucede con la tarea del CLEF, la perplejidad de los modelos de lenguaje basados en preguntas es mucho menor que los entrenados a partir del corpus. Pero el número de palabras fuera del vocabulario es mucho más reducido para este caso.

La figura A.1 muestra el conjunto de entidades de test que quedan fuera del vocabulario (NE_OOV) si se toman diferentes tamaños de la categoría de entidades nombradas. Se observa que al principio se incluyen muchas entidades del test en el vocabulario, y conforme se aumenta el tamaño de entidades de la categoría se necesitan más elementos para cubrir nuevas entidades de test. Al realizar el filtrado de la categoría de Entidades Nombradas, se han cogido aquellas palabras que tienen una frecuencia de aparición mayor a 8, es decir, aproximadamente, un total de $20,000$ entidades (con lo que quedaría un 0,168 de entidades de test fuera del vocabulario). Para el caso de los nombres comunes, si se sigue este criterio el número de elementos en la categoría es poco mayor que $12,000$.

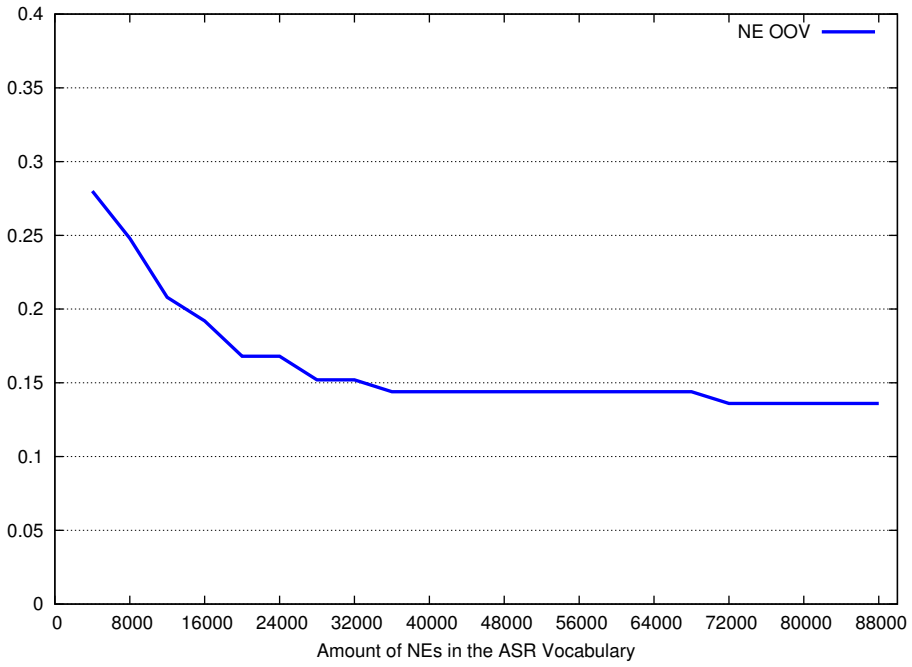

Figura A.1: Entidades Nombradas fuera del vocabulario para la tarea QAst.

Bibliografía

[Akiba et al., 2007] Akiba, T., Itou, K., and Fujii, A. (2007). Language model adaptation for fixed phrases by amplifying partial n-gram sequences. *Systems and Computers in Japan*, 38(4):63–73.

[Allauzen and Gauvain, 2005] Allauzen, A. and Gauvain, J. (2005). Open Vocabulary ASR for Audiovisual Document Indexation. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*, pages 1013–1016. IEEE.

[Atserias et al., 2006] Atserias, J., Casas, B., Comelles, E., Gónzalez, M., Padró, L., and Padró, M. (2006). Freeling 1.3: Five years of open-source language processing tools. In *Proceedings of the 5th internacional conference on Language Resources and Evaluation*.

[Bahl et al., 1983] Bahl, L., Jelinek, F., and Mercer, R. (1983). A maximum likelihood approach to continuos speech recognition. In *IEEE-J. Pattern Anal. Mach. Iintell.*, volume PAMI-5, pages 179–190.

[Buscaldi et al., 2006] Buscaldi, D., Gómez, J. M., Rosso, P., and Sanchis, E. (2006). N-Gram vs. Keyword-Based Passage Retrieval for Question Answering. In *Proceedings of CLEF 2006*, pages 377–384.

[Buscaldi et al., 2010] Buscaldi, D., Rosso, P., Soriano, J. M. G., and Sanchis, E. (2010). Answering questions with an -gram based passage retrieval engine. *J. Intell. Inf. Syst.*, 34(2):113–134.

[Carreras et al., 2004] Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th Language Resources and Evaluation Conference*.

[Castro-Bleda et al., 2001] Castro-Bleda, M. J., España-Boquera, S., Marzal, A., and Salvador, I. (2001). Grapheme-to-phoneme conversion for the spanish language. In *Pattern Recognition and Image Analysis. Proceedings of the IX Spanish Symposium on Pattern*

Recognition and Image Analysis, pages 397–402, Benicàssim (Spain). Asociación Española de Reconocimiento de Formas y Análisis de Imágenes.

[Chelba et al., 2008] Chelba, C., Hazen, T. J., and Saraclar, M. (2008). Retrieval and browsing of spoken content. *Signal Processing Magazine, IEEE*, 25(3):39–49.

[Chu-Carroll and Prager, 2007] Chu-Carroll, J. and Prager, J. (2007). An experimental study of the impact of information extraction accuracy on semantic search performance. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 505–514. ACM.

[Comas and Turmo, 2008] Comas, P. and Turmo, J. (2008). Phast: Spoken document retrievel based on sequence alignment. In *Report de recerca del LS: LSI-08-2-R.*

[Comas and Turmo, 2009] Comas, P. and Turmo, J. (2009). Robust question answering for speech transcripts: Upc experience in qast 2009. In *Working Notes of CLEF2009.*

[europarl, 2009] europarl (1996-2009). Europarl: European parliament proceedings parallel corpus.

[Fujii et al., 2002] Fujii, A., Itou, K., and Ishikawa, T. (2002). Speech-driven text retrieval: Using target ir collections for statistical language model adaptation in speech recognition. In Coden, A., Brown, E., and Srinivasan, S., editors, *Information Retrieval Techniques for Speech Applications*, volume 2273 of *Lecture Notes in Computer Science*, pages 331–334. Springer.

[Gokhan and De Mori, 2011] Gokhan, T. and De Mori, R. (2011). *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech.* WILEY.

[González et al., 2008] González, C., Cardeñoso-Payo, V., and Sanchis, E. (2008). Experiments in speech driven question answering. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 85–88.

[Harabagiu et al., 2002] Harabagiu, S., Moldovan, D., and Picone, J. (2002). Open-domain voice-activated question answering. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7. Association for Computational Linguistics.

[Hori et al., 2003] Hori, C., T.Hori, Isozaki, H., Maeda, E., S.Katagiri, and S.Furui (2003). Study on spoken interactive open domain question answering. In *ISCA/IEEE Workshop Spontaneus Speech Process Recognition.*, pages 111–114.

[Jurafsky and Martin, 2009] Jurafsky, D. and Martin, J. (2009). *Speech and Language Processing*. Pearson International Edition, second edition.

[Kim et al., 2004] Kim, D., Furui, S., and Isozaki, H. (2004). Language models and dialogue strategy for a voice QA system. In *18th International Congress on Acoustics*, pages 3705–3708, Kyoto, Japan.

[Kubala et al., 1998] Kubala, F., Schwartz, R., Stone, R., and Weischedel, R. (1998). Named entity extraction from speech. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*.

[Loquendo, a] Loquendo. *Loquendo ASR ®, Programmer's Guide*. 2001-2009 Loquendo - All rights reserved, 7.8 edition.

[Loquendo, b] Loquendo. *Loquendo ASR ®, SATCA User's Guide*. 2001-2009 Loquendo - All rights reserved, 7.8 edition.

[manfreeling, 2010] manfreeling (2010). *FreeLing User Manual*. Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP), Universitat Politècnica de Catalunya.

[Mishra and Bangalore, 2010] Mishra, T. and Bangalore, S. (2010). Speech-driven query retrieval for question-answering. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5318–5321. IEEE.

[Moldovan et al., 2003] Moldovan, D., Pasca, M., Harabagiu, S., and Surdeanu, M. (2003). Performance Issues and Error Analysis in an Open-Domain Question Answering System. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 133–154, New York, USA.

[Niesler et al., 1998] Niesler, T., Whittaker, E., and Woodland, P. (1998). Comparison of part-of-speech and automatically derived category-based langauga models for speech recognition. In *Acoustics Speech and Signal Processing (ICASSP) IEEE International Conference on,1998*.

[Padró et al., 2010] Padró, L., Collado, M., Reese, S., Lloberes, M., and Castellón, I. (2010). Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference*.

[P.F. et al., 1992] P.F., B., Della Pietra, V. J., de Souza, P. V., J.C, L., and Mercer, R. L. (1992). Class-based n-gram models of natural language. In *Computational Linguistics*, volume 18, pages 467–479.

[Rosso et al., 2010] Rosso, P., Hurtado, L.-F., Segarra, E., and Sanchis, E. (2010). On the voice-activated question answering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, PP(99):1 –11.

[Sanchis et al., 2006] Sanchis, E., Buscaldi, D., Grau, S., Hurtado, L., and Griol, D. (2006). Spoken QA based on a Passage Retrieval engine. In *IEEE-ACL Workshop on Spoken Language Technology*, pages 62–65, Aruba.

[Schofield, 2003] Schofield, E. (2003). Language-models for questions. In *Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*, TextEntry '03, pages 17–24. Association for Computational Linguistics.

[Stoyanchev et al., 2008] Stoyanchev, S., Tur, G., and Hakkani-Tur, D. (2008). Name-aware speech recognition for interactive question answering. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, pages 5113–5116. IEEE.

[Summit, 2005] Summit, M., editor (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*.

[Turmo et al., 2007] Turmo, J., Comas, P., Ayache, C., Mostefa, D., Rosset, S., and Lame, L. (2007). Overview of qast 2007.

[Turmo et al., 2009] Turmo, J., Comas, P., Rosset, S., Galibert, O., Moreau, N., Mostefa, D., Rosso, P., and Buscaldi, D. (2009). Overview of QAST 2009. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, volume 6241 of *Lecture Notes in Computer Science*, pages 197–211. Springer.

[Turmo et al., 2008] Turmo, J., Comas, P., Rosset, S., Lamel, L., Moreau, N., and and, D. M. (2008). Overview of QAST 2008.

[Vallin et al., 2005] Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., and Sutcliffe, R. (2005). Overview of the clef 2005 multilingual question answering track. In *Proceedings of CLEF 2005*.

[Wang et al., 2010] Wang, D., King, S., Evans, N., and Troncy, R. (2010). Crf-based stochastic pronunciation modeling for out-of-vocabulary spoken term detection. In *Proc. of InterSpeech 2010*, pages 1668–1671, Makuhari, Chiba, Japan.

[Wang et al., 2008] Wang, Y.-Y., Yu, D., Ju, Y.-C., and Acero, A. (2008). An introduction to voice search. *Signal Processing Magazine, IEEE*, 25(3):28 –38.