Universidad Politécnica de Valencia

Departamento de Sistemas Informáticos y Computación

Trabajo Fin de Máster

Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Título:

Un método de aprendizaje semi-supervisado para comprensión del habla

Presentada por: Lucía Ortega Álvarez

Supervisores: Dr. Mª Isabel Galiano
Departamento de Sistemas Informáticos y Computación
Dr. Emilio Sanchis
Departamento de Sistemas Informáticos y Computación

Resumen En este trabajo presentamos un algoritmo para el aprendizaje estadístico de modelos semánticos, a partir de un corpus no alineado de pares de frases y su representación semántica en términos de frames. El objetivo final es poder asociar automáticamente segmentos de longitud variable con sus correspondientes unidades semánticas para ser usados en tareas de comprensión de habla. Una de las ventajas de esta aproximación consiste en evitar el costoso trabajo de segmentar y etiquetar todo el corpus de aprendizaje, como necesitan la mayor parte de los métodos basado en corpus. Por otra parte, resulta de especial interés la capacidad de aprendizaje discriminativo que presenta este método. Hemos aplicado este algoritmo al desarrollo del módulo de comprensión de un sistema de diálogo hablado, cuya tarea es el acceso a información sobre trenes. Se presentan experimentos que confirman lo adecuado del método, dado el ahorro de esfuerzo en la preparación del corpus.

Además, nos proponemos utilizar este método automático para construir un modelo inicial a partir de un porcentaje de frases de entrada más pequeño, para posteriormente, mediante un proceso de aprendizaje activo ser capaces de tener la misma capacidad de segmentación que cuando hemos entrenado el modelo con un corpus mayor. Esto supone un enorme ahorro en el coste que supone etiquetar y segmentar a mano un corpus de gran tamaño.

Resum En aquest treball presentem un algorisme per a l'aprenentatge estadístic de models semàntics, a partir d'un corpus no alineat de parells de frases i la seva representació semàntica en termes de frames. L'objectiu final és poder associar automàticament segments de longitud variable amb les seves corresponents unitats semàntiques per a ser usats en tasques de comprensió de parla. Un dels avantatges d'aquesta aproximació consisteix a evitar el costós treball de segmentar i etiquetar tot el corpus d'aprenentatge, com necessiten la major part dels mètodes basat en corpus. D'altra banda, resulta d'especial interès la capacitat d'aprenentatge discriminatiu que presenta aquest mètode. Hem aplicat aquest algoritme al desenvolupament del mòdul de comprensió d'un sistema de diàleg parlat, la tasca és l'accés a informació sobre trens. Es presenten experiments que confirmen el adequat del mètode, donat l'estalvi d'esforç en la preparació del corpus.

A més, ens proposem utilitzar aquest mètode automàtic per a construir un model inicial a partir d'un percentatge de frases d'entrada més petit, per posteriorment, mitjançant un procés d'aprenentatge actiu ser capaços de tenir la mateixa capacitat de segmentació que quan hem entrenat el model amb un corpus més gran. Això suposa un enorme estalvi en el cost que suposa etiquetar i segmentar a mà un corpus de grans dimensions.

Abstract In this work we present an algorithm for learning statistical semantic models, based on a corpus of sentence pairs non-aligned and their semantic representation in terms of frames. The ultimate goal is to automatically associate variable-length segments with their corresponding semantic units for use in speech understanding tasks. One advantage of this approach is to avoid the expensive work of segmentation and tagging of the entire corpus like most corpus-based methods need. On the other hand, is of particular interest discriminative learning ability of this method. We applied this algorithm to the development of the understanding module of a spoken dialogue system, whose task is access to information about trains. Experiments are presented which confirm the suitability of the method, since the economy of effort in preparing the corpus.

In addition, we propose to use this automated method to build an initial model based on a percentage of smaller input sentences, later, through a process of active learning to be able to have the same ability to segment when we trained the model with a larger corpus. This is a huge savings in the cost of hand-labeling and segmenting a large corpus.

Índice general

Parte I

Introducción

Motivación

Desde los inicios de la informática ha existido la necesidad de una comunicación sencilla entre hombre y máquina. Representar el lenguaje humano de forma que pueda ser entendido por el sistema ha sido y sigue siendo uno de los grandes retos de la informática. Los sistemas de comprensión de lenguaje hablado (Spoken Language Understanding) están diseñados para extraer el significado de frases en lenguaje natural obteniendo una representación conceptual de la oración, lo que permite al sistema poder interpretar el lenguaje humano. Las aplicaciones de estos sistemas son muy extensas, desde búsqueda por voz en dispositivos móviles a resúmenes de reuniones, lo que supone de gran interés para los sectores comercial y académico. En los sistemas de diálogo hablado, el módulo de comprensión adquiere, por lo tanto, una gran importancia, ya que tiene la delicada tarea de extraer tanto la intención como la información que proporciona el usuario.

Uno de los casos más simples en los que se pueden utilizar este tipo de aplicaciones son los sistemas de acceso telefónico a información. Los más sencillos se basan en que el usuario, mediante la pronunciación de dígitos o palabras clave, indique al sistema la información que requiere. Este tipo de aplicaciones están absolutamente dirigidas por el sistema y el usuario sólo tiene opciones para indicar mediante un número o un nombre la opción elegida, proporcionar su número de identificación o DNI, y responder palabras como "sí" o "no". Aunque estos sistemas pueden ser útiles para algunas aplicaciones concretas, en las que la información solicitada está claramente predeterminada, no existe ambigüedad, y la variedad en cuanto al tipo de las informaciones es pequeña, no están exentos de las dificultades del reconocimiento robusto de voz en ambientes ruidosos, ya que su utilidad se basa en que sean

accesibles desde cualquier lugar, por ejemplo mediante teléfono móvil, dentro de un coche, o desde una oficina en la que hay ruido ambiente. El siguiente tipo de sistemas de interacción hombre-máquina, que es el que más interés despierta en la actualidad, es aquél que permite la comunicación oral, y permite a su vez establecer un diálogo que ayude a la consecución de los objetivos planteados por el usuario.Los sistemas de diálogo de este tipo, requieren tener las siguientes características:

- El modo de acceso telefónico: Cada vez es más importante el acceso a través de teléfono móvil, con las dificultades añadidas que conlleva.
- Debe ser independiente del locutor: Lógicamente los accesos serán de múltiples usuarios.
- Uso de lenguaje natural: Los usuarios deben poder hablar de forma natural, sin una sintaxis impuesta por el sistema y usando el léxico más amplio posible. Se deben además aceptar las incorrecciones léxicas y sintácticas propias del habla espontánea, y aunque el léxico esté limitado por el ámbito de la tarea, se ha de permitir que el usuario utilice un amplio vocabulario.
- Aceptar habla continua: Para una interacción fluida se debe permitir habla continua, sin pausas entre palabras.
- Gestión mixta del diálogo: El diálogo no estará totalmente dirigido por el sistema. El usuario podrá tomar la iniciativa y orientar con sus preguntas el curso de la interacción.

La modelización estadística de algunos de los componentes de los sistemas de interacción oral hombre-máquina ha permitido grandes avances en el comportamiento de estos sistemas. Uno de los aspectos fundamentales que ha servido para el éxito de estos modelos es la existencia de buenos algoritmos de aprendizaje automático a partir de corpus, y por tanto, la capacidad de representar la variabilidad (acústica, lingüística, semántica) propia del lenguaje humano. Sin embargo, estos métodos tienen por contrapartida la necesidad de adquirir y

etiquetar los corpus de datos que, cuanto más grandes, mejores modelos pueden proporcionar. Además los corpus de aprendizaje son un elemento estático que confiere unas características a los modelos muy dependientes del propio corpus. Incluso a veces puede llegar a producirse un efecto hiper-aprendizaje de modo que se representan muy bien las características del corpus mientras quedan pobremente recogidas aquellas situaciones no vistas en el corpus, es decir hay una pérdida de cobertura. Por otra parte, en determinadas aplicaciones puede ser conveniente disponer de modelos que se adapten a nuevos entornos, o dominios, de modo que los modelos tienen que tener la capacidad de modificarse para representar nuevos eventos. Todas estas circunstancias han llevado a que en los últimos años se hayan hecho grandes esfuerzos en el desarrollo de técnicas para reducir el esfuerzo en el etiquetado de corpus, lo que por una parte permite evitar los errores inducidos por un mal etiquetado, y por otra posibilita el uso de corpus más grandes (o más representativos de la variabilidad que se quiere representar). Algunas de estas técnicas son las basadas en el aprendizaje activo [1], [2], [3], [4] y [5], o en el aprendizaje no supervisado, o semi-supervisado [6].

Las técnicas de aprendizaje activo tienen por objetivo la adaptabilidad de los modelos a partir de las interacciones con los usuarios, mediante un etiquetado manual muy reducido, mientras que el aprendizaje semi-supervisado se basa en algoritmos que no requieren que el corpus esté totalmente etiquetado manualmente.

Si nos centramos en el ámbito de comprensión del habla vemos que estas técnicas de modelización estadística y de aprendizaje, han dado buenos resultados cuando se trata de tareas de dominios restringidos. En este trabajo presentamos una propuesta para la modelización semántica en un sistema de diálogo hablado. Aunque la tarea sobre la que se presentan resultados es una tarea para acceso a información sobre trenes, las características del proceso de aprendizaje permite una fácil adaptación a nuevos dominios. Una de las características de la aproximación propuesta es que el algoritmo de aprendizaje sólo requiere un etiquetado semántico global de las frases, no necesariamente secuencial con el texto, y por tanto, no es necesario el esfuerzo de la segmentación manual del texto. Se basa en un mecanismo discriminativo que asigna segmentos de diversas longitudes a las unidades

semánticas definidas, basándose en la frecuencia de co-ocurrencias entre segmentos y unidades semánticas [7]. También se ha desarrollado un mecanismo de aprendizaje activo que, a partir de las medidas de confianza proporcionadas por los modelos aprendidos inicialmente, identifica aquellas frases susceptibles de ser analizadas manualmente, para ser incorporadas con un correcto etiquetado al corpus de aprendizaje. De este modo se espera que los modelos vayan adaptándose dinámicamente a las nuevas interacciones con los usuarios, minimizando el esfuerzo de análisis y etiquetado manual.

Objetivos

En este apartado se van a detallar aquellos aspectos más significativos que este trabajo fin de máster pretende abordar. El principal objetivo es el desarrollo de un algoritmo de aprendizaje que permita asociar automáticamente segmentos de longitud variable con sus correspondientes unidades semánticas para ser usados en tareas de comprensión de habla. Esto nos permitirá evitar el etiquetado y la segmentación manual del corpus.

Para ello se han propuesto diversas aproximaciones que han sido evaluadas sobre el corpus de datos, en nuestro caso diálogos hablados, adquirido dentro del proyecto DIHANA.

Las principales tareas que se han desarrollado son:

- Definición de los modelos estadísticos a partir de un conjunto de pares de frases y conceptos no alineados ni secuenciales con la frase. Con lo cual nos evitamos el costoso proceso de definir un lenguaje intermedio.

- Desarrollo de un algoritmo de aprendizaje que a partir de los modelos es capaz de asociar un conjunto de segmentos de longitud variable a cada uno de los conceptos que aparecen en el corpus.

- Definición de unos criterios de desambiguación y poda de segmentos que permiten que, a partir de un corpus con pocas muestras, se puedan conseguir resultados con una alta cobertura y precisión. Por otra parte, se han aplicado técnicas de categorización léxicas y semánticas que permiten mejorar las probabilidades de los segmentos más significativos dentro de un conjunto.

- Los algoritmos desarrollados se han utilizado a modo de herramienta de segmentado de corpus y se han demostrado sus buenos resultados mediante diferentes experimentos.

- Se han desarrollado algoritmos de aprendizaje activo, por los cuales se pretende obtener buenos resultados a partir de un pequeño porcentaje del corpus.

Estructura

El resto de este documento se organiza de la siguiente manera. En la Parte II, se define el estado del arte en los sistemas de comprensión del habla. En la Parte III, definiremos el método de aprendizaje propuesto. Los siguientes capítulos estarán dedicados a definir las estructuras usadas para guardar la información obtenida del algoritmo de aprendizaje y de la información lingüística que hemos utilizado para refinar nuestro método. En concreto habrá un capítulo dedicado a exponer las diferentes herramientas y diccionarios utilizados. Los últimos capítulos de esta parte definen los experimentos que han sido llevados a cabo para demostrar la utilidad del método, en concreto se define una batería de experimentos donde se definen una serie de medidas y se comparan con la segmentación manual.

La parte V está dedicada a las conclusiones obtenidas de la experimentación llevada a cabo y se describen una serie de trabajos que serán llevados a cabo en el futuro, asi como las publicaciones relacionadas con este trabajo.

Parte II

Estado del arte

Capítulo 1

Modelos para comprensión del lenguaje hablado

En los modelos para comprensión del habla se han estudiado dos enfoques principales para determinar la correlación entre las palabras y conceptos en las tareas:

1. los modelos de generadores, cuyos parámetros se refieren a la probabilidad conjunta de conceptos y componentes semánticos.

2. los modelos discriminativos, que aprenden una clasificación basada en función de las probabilidades condicionales de conceptos dadas palabras.

Suponiendo un problema de etiquetado donde a las observaciones X se les debe asignar una etiqueta y_i de un conjunto Y, dada una muestra de datos que proporcionan ejemplos de etiquetado, los modelos generativos asumen que la muestra se generó a partir de una fuente estocástica a raíz de una distribución de probabilidad conjunta $P(X, Y)$ y estiman la probabilidad con los datos disponibles, ya sea directamente o a través de la descomposición en una probabilidad condicional $P(X|Y)$ y una distribución a priori de $P(Y)$. Por el contrario, los modelos discriminativos estiman directamente la probabilidad a posteriori $P(Y|X)$ a partir de datos, sin ningún intento de modelizar la distribución de probabilidad subyacente.

Estos dos enfoques conducen a formulaciones muy diferentes y complejas, dependiendo de si se trata de problemas de clasificación simple, por ejemplo, Clasificadores de Naive Bayes o de Regresión Logística [8], o problemas de etiquetado de secuencias, por ejemplo, Modelos Ocultos de Markov o Conditional Random Fields [9] [10].

1.1. Modelos generativos

Un ejemplo de modelo generativo son los Hidden Vector State model (HVS) [11]. Este enfoque se extiende a los modelos discretos de Markov codificando el contexto de cada estado como un vector. Las transiciones de estado se realizan como operaciones de cambio de pila, mediante una etiqueta de categoría semántica como por un tree parser. De esta manera el modelo puede capturar estructuras jerárquicas semánticas, sin la necesidad de anotación de árbol (con un conjunto de datos de arranque con un anotado manual mínimo).

Otro modelo generativo es el basado en Stochastic Finite State Transducers (SFST), que realiza SLU como un proceso de traducción de palabras a conceptos. Este modelo ha demostrado una gran precisión a pesar de su simplicidad [12]. Otro aspecto interesante es su fácil integrabilidad en los sistemas de reconocimiento de voz, donde la salida puede ser una red de palabras, que a su vez es codificada como un FST estocástico.

Un modelo generativo más reciente para SLU se basa en las Redes Bayesianas Dinámicas (DBN). Se han aplicado a muchas tareas de modelado de datos secuenciales, por ejemplo el reconocimiento automático del habla [13], part-of-speech tagging [14], dialog-act tagging [15] o el análisis de secuencias de ADN. Las DBN han demostrado que proporciona una gran flexibilidad para la representación de sistemas estocásticos complejos con un buen rendimiento en comparación con otros métodos estocásticos.

1.2. Modelos discriminativos

Un ejemplo de modelo discriminativo utilizado para SLU es el basado en Máquinas de Soporte Vectorial (SVM) [16], como se muestra en [12]. En esta aproximación, los datos se asignan en un espacio vectorial y el SLU se realiza como una secuencia de problemas de clasificación utilizando Maximal Margin Classifiers.

Un enfoque relativamente más reciente para SLU se basa en Conditional Random Fields (CRF) [10]. Los CRFs entrenan probabilidades condicionales teniendo en cuenta muchas de las características de la frase de entrada. Puesto que son entrenados condicionalmente, no necesitan representar explícitamente las dependencias de las características, la dependencia condicional se captura con las funciones características. Los CRF pertenecen a la familia de modelos log-lineales, la diferencia de otros modelos de esta familia está en el factor utilizado para la normalización de probabilidad. La elección de una normalización a nivel secuencia conduce a la cadena lineal del CRF, la elección de una normalización a nivel de posición conduce a el modelo de máxima entropía [17]. Este modelo ha sido aplicado a SLU en [18] y [19] mostrando buenos resultados.

Capítulo 2

Algunos sistemas de diálogo hablado

En los últimos años se han hecho grandes esfuerzos en el desarrollo de sistemas de diálogo, lo que ha impulsado también los trabajos en el área de comprensión de habla. Son muchos los laboratorios que han dedicado grandes esfuerzos a la obtención de sistemas de diálogo. Aunque las aplicaciones escogidas son variadas, la mayoría de ellos se ha centrado en sistemas de acceso a información.

Una de las primeras tareas que se planteó fue la tarea ATIS (Air Travel Information services) [20] patrocinado por la organización ARPA (Advanced Research Projects Agency) que consiste en obtener información sobre vuelos. En torno a esta tarea se desarrollaron múltiples proyectos y se convirtió, junto a la de información sobre trenes en un tipo de aplicación ampliamente estudiada en otras lenguas. Un ejemplo de ello son los proyectos SUNDIAL [21], ARISE [22], MASK [23].

A continuación se muestra una breve descripción de algunos de los proyectos y laboratorios más representativos en el desarrollo de sistemas de diálogo:

ATT : Desde los primeros proyectos de desarrollo de sistemas de diálogo, los laboratorios de ATT han trabajado en esta línea, haciendo especial énfasis en la utilización de modelos estocásticos, tanto para comprensión como para diálogo. Desarrollaron diversas aplicaciones para la tarea ATIS, como fue el proyecto AMICA [24]. Actualmente desarrollan

proyectos como “How May I Help you?” [25], consistente en una tarea de “callrouting”.

MIT : Además de trabajar con la tarea ATIS, en el MIT se desarrolló un sistema de diálogo, GALAXY [26] [27], con el objetivo de ser un sistema conversacional válido para distintos dominios. Un ejemplo de los diferentes sub-dominios, es el WHEELS que es un sistema de acceso a información sobre ventas de coches, el VOYAGER, cuyo objetivo es proporcionar información típicamente relacionadas con viajes, como distancias entre ciudades, hoteles, direcciones o números de teléfono. El JUPITER [28] es un sistema de información sobre el tiempo.

CMU : Uno de los principales proyectos desarrollados en la CMU es el Communicator Travel Planning system, cuya tarea es la de planificación de viajes: aviones, hoteles o reservas de coche [29]. La talla del vocabulario es de 2.500 palabras. Otro proyecto es el CMU-VODIS [30], orientado al desarrollo de aplicaciones de interfaz oral hombre-máquina en los automóviles.

SUNDIAL : Entre los primeros proyectos desarrollados en Europa se encuentra el SUNDIAL [21]. Se desarrollaron cuatro prototipos en cuatro lenguas distintas para las consultas de horarios de trenes en Alemán e Italiano y de vuelos en Inglés y Francés. El objetivo del proyecto era construir sistemas de diálogo integrados en tiempo real capaces de mantener diálogos cooperativos con los usuarios.

LIMSI : A partir de los trabajos desarrollados sobre la versión francesa de ATIS [31], se desarrolló el proyecto ARISE [22], y el MASK [23]. El ARISE fue un proyecto europeo para desarrollar un prototipo automático de consulta de horarios y servicios para trenes que permita manejar la gran mayoría de las rutinarias consultas telefónicas. Se construyó un sistema para los operadores alemanes e italianos y dos para el francés. El proyecto predecesor RAILTEL [32], definió la estructura para el desarrollo de los servicios interactivos de voz que proporcionan los horarios y planificación en diversos lenguajes (Alemán, Inglés, Francés y Italiano) a través del teléfono. El proyecto MASK desarrolló un servicio de quiosco multimodal y multimedia para ser colocado en las es-

taciones de tren. Se desarrolló un prototipo de quiosco de información que se instaló en la estación de "St. Lazare" en París. El quiosco pretende mejorar la eficacia de tales servicios permitiendo la interacción con el uso coordinado de entradas multimodales (discurso y tacto) y salidas multimedia (sonido, vídeo, texto y gráficos) creando así una nueva modalidad de servicios al público.

TRAINS Universidad de Rochester: [33] es un sistema de diálogo en lenguaje natural para la planificación de la ruta de trenes desarrollado en la Universidad de Rochester. La motivación es obtener el conjunto de rutas más eficiente entre dos ciudades. Un análisis bottom-up para Context Free Grammars produce una secuencia de actos de diálogo a la vez de llevar a cabo un exacto análisis sintáctico. El rendimiento de la tarea de TRAINS fue evaluada en términos de dos métricas: la cantidad de tiempo que se necesita para obtener la información del itinerario y la calidad de la solución, medida por la cantidad de tiempo necesaria para cubrir las rutas.

Capítulo 3

Definición de un sistema de diálogo hablado

En la Figura 3.1 se muestra un esquema general de un sistema de diálogo hablado. Como puede verse existen múltiples fuentes de conocimiento que deben tenerse en consideración para su desarrollo. Podemos establecer tres bloques:

- El bloque correspondiente al tratamiento del turno del usuario. Comprende la adquisición, preproceso, reconocimiento y comprensión. Al final de este bloque se espera que el sistema haya comprendido la pronunciación del usuario, lo cual significa que mediante algún tipo de representación se conozca el objetivo o función del turno (llamado acto de diálogo) y la información (o datos) proporcionados.

- El gestor de diálogo, que debería tomar una decisión para generar un turno de respuesta. La actuación del gestor de diálogo se basaría en tres factores: la información proporcionada por el usuario en el ultimo turno; la información almacenada por el propio gestor a lo largo del diálogo hasta este momento; y la información del contexto de la aplicación, que básicamente estaría representada por la base de datos sobre la que se está preguntando. Las principales acciones que suele realizar un gestor de diálogo son: confirmar datos, recuperar errores, dirigir el diálogo hacia el objetivo solicitando nue-

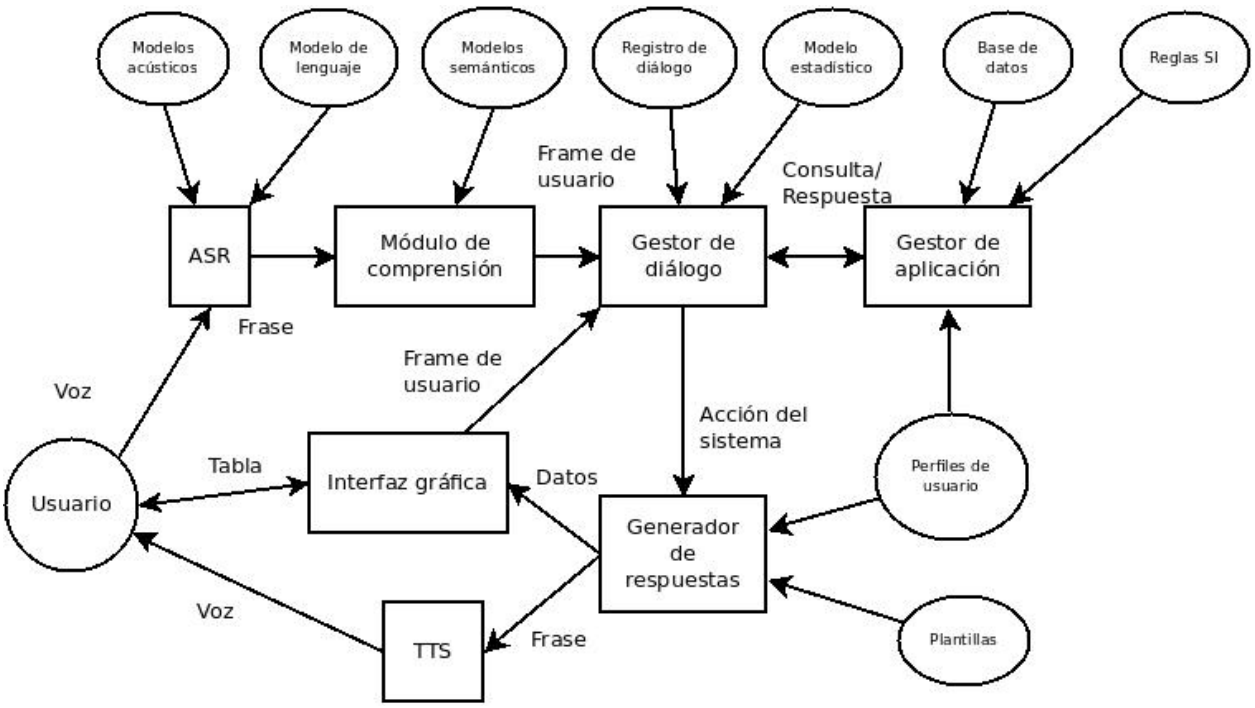

Figura 3.1:

vos datos, proporcionar la información solicitada, y acciones propias del metalenguaje de diálogo como cortesía apertura de diálogo, o frases del tipo "espere un momento por favor".

- El bloque de generación de respuesta, al que el gestor de diálogo enviaría una representación del mensaje que se debe emitir al usuario y que se ha de convertir en una frase en lenguaje natural, y posteriormente sería sintetizada y emitida.

Como se ha dicho anteriormente el desarrollo de sistemas de diálogo hablado ha sido posible gracias a las prestaciones alcanzadas en las distintas áreas implicadas. En particular, el reconocimiento automático del habla, que es el primer eslabón del sistema, sin el cual no podría pensarse en la existencia de sistemas de diálogo.

a) Reconocimiento del habla.

Los sistemas de reconocimiento del habla empezaron a dar buenos resultados a partir

de los años 80 en que se generalizó el uso de Modelos Ocultos de Markov (HMM) como forma de representar las características acústicas de las unidades del habla. El éxito de los HMM se basa principalmente en la existencia de algoritmos de aprendizaje automático de los parámetros del modelo (Baum-Welch) [34], así como en su capacidad para representar el habla como un fenómeno secuencial en el tiempo. Se han estudiado múltiples aproximaciones, como son los modelos discretos, semicontinuos o continuos, así como diversas topologías de los modelos. Un factor importantísimo es la elección de las unidades acústicas básicas. Aunque inicialmente es el fonema la unidad más elemental del habla, para la implementación de sistemas de reconocimiento es necesario el uso de fonemas con contexto (u otras unidades mayores como las sílabas o semisílabas), como unidad básica, ya que de esta forma se pueden representar los efectos acústicos debido a la coarticulación de fonemas y palabras. El inconveniente que tiene la proliferación de unidades básicas, así como el de modelos más precisos (como son los modelos continuos con múltiples gaussianas) es que se requiere mayor computación en el proceso de aprendizaje, y por tanto mayor número de muestras.

b) Comprensión del habla.

En primer lugar podemos establecer dos tipos de aproximaciones al proceso de comprensión: las basadas en reglas [31] [35] [21] [36] y las basadas en modelos estocásticos [37] [38] [39] [40] [41].

En el caso de las basadas en reglas la información semántica se extrae a partir del análisis sintáctico-semántico de las frases, utilizando gramáticas definidas para la tarea, o a partir de la detección de palabras (o secuencias de palabras) clave, con significado semántico. En el caso de los métodos estocásticos el proceso se basa en la definición de unidades lingüísticas con contenido semántico y en la obtención de modelos a partir de muestras etiquetadas. El proceso de comprensión se realiza de forma similar al del reconocimiento del habla, mediante el algoritmo de Viterbi, puede interpretarse como un proceso de traducción de una frase de entrada (secuencia de palabras) en una frase de salida (secuencia de unidades semánticas). Un aspecto importante a considerar es

la forma de transmitir la información entre los módulos de reconocimiento y de comprensión. Con el objetivo de que los errores de la etapa de reconocimiento puedan ser recuperados en posteriores etapas hay múltiples propuestas para proporcionar más de una sola frase, como es el caso de las N mejores frases (N-best) [42], o de los grafos de palabras [43]. De esta forma la etapa de comprensión puede tener en consideración múltiples hipótesis del reconocedor. Esta misma idea podría ser aplicada a la comunicación entre el módulo de comprensión y el gestor de diálogo.

c) Gestión del diálogo.

Así como en los módulos anteriormente descritos es habitual encontrar aproximaciones basadas en métodos estocásticos, en el caso de los gestores de diálogo hay mayores dificultades para estas modelizaciones, principalmente debido a la falta de muestras de aprendizaje y a la gran cantidad de situaciones, o estados del diálogo, que habría que representar. Por ello la mayoría de sistemas de diálogo tienen representada la estrategia del diálogo en forma de reglas. Se han desarrollado en los últimos años, algunas herramientas "toolkit" para el desarrollo de sistemas de diálogo que permiten al diseñador que defina el comportamiento del gestor de diálogo (CMU Communicator [44], VOICEXML [45], CSLU [46]). Sin embargo, también hay aproximaciones basadas en métodos estocásticos [47] [48] donde las unidades básicas son los "actos de diálogo", y se puede modelizar el comportamiento del diálogo como una secuencia de éstos.

d) Síntesis de voz.

Finalmente en el caso de la síntesis del habla, existen en la actualidad buenos sistemas [49]. Para producir voz, pueden utilizarse mecanismos diversos, dependiendo de la complejidad de los recursos que se disponga. Existen sistemas que limitan a unos pocos los mensajes que puede generar la máquina (cita previa de la ITV, información bursátil). En estos sistemas la generación de voz puede realizarse mediante la reproducción de mensajes grabados, o concatenando grabaciones de palabras o frases. Sin embargo existen aplicaciones donde la información es tan grande que es muy posible

que nunca se llegue a escuchar toda (noticias, lectura de correo electrónico, etc.) en este contexto se contemplan los sistemas de conversión de texto en habla, capaces de producir voz a partir de un representación escrita.

Parte III

Método de aprendizaje semi-supervisado

Capítulo 4

Corpus DIHANA

Para realizar la modelización semántica de este trabajo fin de máster, se ha elegido el corpus DIHANA [50] [51] [52]. En el marco del proyecto DIHANA, se presenta el proceso de adquisición de un corpus de diálogo de habla espontáneo en español. La aplicación seleccionada consiste en la recuperación de información de trenes por teléfono.

Existen varios tipos de escenarios se definidos con el fin de controlar la interacción del usuario con el sistema. Un escenario se define por:

- un objetivo: la información que necesita el usuario,
- una situación: las circunstancias específicas en relación con la solicitud de viaje, y
- los requisitos específicos del viaje: tipo de viaje, la ciudad de salida, ciudad de destino, y una o varias restricciones.

Se definen tres tipos de escenarios basados en el objetivo que se quiere obtener para obtener los horarios de los viajes de ida, para obtener los horarios (y, opcionalmente, los precios) para viajes de ida, y para obtener los horarios (y, opcionalmente, los precios) para los viajes de regreso. También se definieron diferentes situaciones con diferentes restricciones. Por lo tanto, 300 escenarios diferentes fueron diseñados, que fueron adquiridos por 225 usuarios que

realizaron cuatro escenarios de cada uno. Un total de 900 diálogos fueron desarrollados. Las características del corpus transcrito se muestran en la Tabla 4.1

Número de turnos	6 226
Número de palabras	47 222
Talla del vocabulario	811
Media de palabras por turno de usuario	7.6

Tabla 4.1: *Características del corpus transcrito.*

A continuación se muestra un fragmento de un diálogo real de la tarea. La primera columna indica el ponente: la máquina (M) o turno de usuario (U):

M: Bienvenido al sistema de información para trenes, ¿qué información le gustaría?
U: Me gustaría saber los horarios de los trenes Euromed de Barcelona a Valencia.
M: ¿Desea viajar desde Barcelona a Valencia?
U: Sí.
M: ¿Desea viajar hoy?
U: No, el próximo jueves.
M: Estoy buscando los horarios de Barcelona a Valencia para el 15 de julio. Un momento, por favor.
M: Hay dos trenes. El primero sale a las siete y media de la mañana, el segundo a las tres de la tarde. ¿Desea algo más?

4.1. La estrategia de Mago de Oz

La tarea del mago es ayudar al usuario a obtener la información requerida mediante la interacción con el usuario siguiendo una estrategia determinada. Para ello, se utiliza una estructura de blackboard para proporcionar toda la información al mago. La salida de los servidores de reconocimiento y comprensión del habla están escritos en el blackboard, y

su contenido se actualiza en cada turno del diálogo. El mago hace uso de un servidor de generación de respuesta oral y un servidor de conversión de texto a voz para interactuar con el usuario. El mago supervisa la información generada durante el proceso de diálogo, modificando si fuese necesario. Puede interactuar con el usuario para hacer lo siguiente:

- para completar la información necesaria para dar una respuesta,
- para confirmar la información o aclarar malentendidos si es necesario,
- para validar la información, y
- consultar la base de datos para dar una respuesta.

La selección de uno de estos modos se relaciona con la medida de confianza dada por el servidor de reconocimiento y comprensión de voz y la propia experiencia del mago. Los errores semánticos del módulo semántico automático son evaluados por el mago. Si la información no es suficiente, el mago le pide al usuario información nueva hasta que el mago considera que existe suficiente información fiable para dar la respuesta necesaria al usuario.

4.2. Proceso de adquisición

La forma en que el usuario interactúa con el sistema en el proceso de adquisición de los diálogos es el siguiente:

1. El usuario realiza una llamada para interactuar con el sistema. La señal de audio se redirige hacia el servidor de reconocimiento de voz y al servidor de Mago de Oz, para que el mago escuche la persona que llama.
2. El servidor de reconocimiento de voz envía la cadena de palabras reconocida y la medida de confianza al servidor de comprensión del habla.
3. El servidor de comprensión del habla extrae la información relevante, llena el frame que se asocia al escenario, y envía el frame al servidor de Mago de Oz.

4. Si el frame requiere más información, el mago la pide al usuario en un turno de diálogo nuevo. El proceso se repite hasta que la información es suficiente para realizar una consulta a la base de datos.

En otros trabajos previos del grupo en comprensión del habla [7] se llevó a cabo un etiquetado y una segmentación del corpus en unidades semánticas ad-hoc; sin embargo, para la aproximación que aquí presentamos sólo se necesita la representación en frames. Por simplificación,ya hemos comentado que llamamos concepto a cada unidad semántica, bien sea un concepto o un par atributo-valor.

Se definieron 17 conceptos para la tarea DIHANA, agrupados en tres conjuntos: Conceptos generales para cualquier sistema de diálogo (GENERAL), conceptos que representan consultas a la información del sistema (CONSULTA), y conceptos que tienen asociados valores y que representan restricciones de la consulta (ATRIBUTO).

En la Tabla 4.2 se muestran estos conceptos.

GENERAL	CONSULTA	ATRIBUTO
Afirmacion	*Hora*	*Tipo-tren*
Negacion	*Fecha*	*Tipo-viaje*
No-entendido	*Duración*	*Origen*
	Precio	*Destino*
		Ciudad
		Salida
		Llegada
		Clase
		Numero-orden
		Servicios

Tabla 4.2: *Lista de conceptos para la tarea DIHANA.*

Hola quiero saber el horario de ida de Palencia a Oviedo el viernes dieciocho de junio.

```
hola:cortesia
quiero saber:consulta
el horario de:hora
ida:tipo-viaje
de palencia:ciudad-origen
a oviedo:ciudad-destino
el viernes dieciocho de junio:fecha

TIPO-VIAJE:ida
(HORA)
CIUDAD-ORIGEN:palencia
CIUDAD-DESTINO:oviedo
```

Figura 4.1: *Ejemplo de lenguaje intermedio del corpus DIHANA.*

En el Ejemplo 4.1 se puede ver como se segmenta la frase en lenguaje intermedio y como es representada en forma de frames. Nuestro objetivo es saltarnos el paso de la segmentación manual intermedia que se muestra en el ejemplo.

Capítulo 5

Modelización semántica

En los modelos semánticos se representan las secuencias posibles de unidades que describen el contenido conceptual del lenguaje. El objetivo de estos modelos es representar el significado de las frases.

Para ello, en primer lugar, se propone una modelización semántica que permita un proceso de decodificación semántica (comprensión del habla) por el que obtenemos una secuencia de unidades semánticas (que a partir de ahora llamaremos conceptos) c_1^M, asociado a una secuencia de palabras w_1^N. Esta secuencia de conceptos será la que maximice la probabilidad condicional:

$$P(c_1^M|w_1^N) = \text{máx } P(w_1^N|c_1^M) \cdot P(c_1^M) \tag{5.1}$$

donde $P(c_1^M)$ es la probabilidad a priori de la secuencia de conceptos c y $P(w_1^N|c_1^M)$ es la probabilidad de la secuencia de palabras w, dada la secuencia de conceptos c, es decir, existe una asociación entre segmentos de palabras y conceptos.

Esta aproximación estadística a la comprensión del habla requiere un método para el aprendizaje de las probabilidades $P(c_1^M)$ y $P(w_1^N|c_1^M)$, además de un algoritmo de búsqueda para obtener c_1^M entre todas las posibles secuencias de conceptos.

En el siguiente ejemplo se muestra una posible secuencia de palabras y su secuencia de conceptos asociado:

w = quiero los horarios de trenes para el veinticuatro de abril de Valencia a Madrid

c = TIPO-VIAJE (HORA) CIUDAD-ORIGEN CIUDAD-DESTINO FECHA

5.1. Modelización estadística

Para el proceso de modelización estadística consideraremos las siguientes características:

- El modelo se aprende automáticamente a partir de un conjunto de pares (secuencia de palabras, secuencia de conceptos), que nombramos (c_1^M, w_1^N), los cuáles no tienen por qué estar alineados. Es decir, aparte de definir las unidades semánticas que se van a considerar, no hay que hacer un etiquetado explícito a nivel de palabra, sino sólo a nivel de frase. De hecho la representación semántica utilizada es el frame, donde los conceptos y atributos se presentan en una forma canónica, y, por tanto, la secuencia de unidades semánticas no es necesariamente secuencial con la frase de entrada.

- Asumiendo que el orden en que se proporciona la información semántica en una frase no es relevante para determinar qué conceptos se han observado, estimaremos la probabilidad a priori $P(c)$ como la probabilidad de los unigramas de los conceptos:

$$P(c_1^M) = \prod_{i=1}^{M} P(c_i) \tag{5.2}$$

  donde $P(c_i)$ es la estimación del unigrama c_i en el corpus de entrenamiento. Con esta asunción, dos secuencias de conceptos diferentes son equivalentes si el conjunto de conceptos de que están compuestas es el mismo.

- Dada la ausencia de alineamiento explícito en el corpus de aprendizaje, las probabilidades de qué palabras, o secuencias de palabras, están asociadas a los conceptos se

estiman a partir de la co-ocurrencias de segmentos y conceptos. Para calcular $P(c_1^M|w_1^N)$ proponemos un proceso de clasificación en el cual segmentos de longitud variable se asocian a conceptos según criterios discriminativos.

- A partir de esta definición de la función objetivo, el proceso de decodificación semántica podría realizarse mediante un algoritmo de programación dinámica clásico, con las modificaciones necesarias para tratar con segmentos de longitud variable. Considerando que $P(c_1^M|w_1^N)$ puede aproximarse por la probabilidad de la mejor de las segmentaciones de w_1^N en un conjunto de M segmentos de longitud variable, se tiene:

$$P(w_1^N|c_1^M) = \max_{\forall l_1, l_2, ... l_{M-1}} \{P(w_1, ..., w_{l_1}|c_1) \cdot P(w_{l_1+1}, ..., w_{l_2}|c_2) \cdot ... \cdot P(w_{l_{M-1}+1}, ..., w_N|c_M)\} \quad (5.3)$$

5.2. Propuesta inicial de aprendizaje

La mayoría de los sistemas SLU que se basan en métodos de clasificación o en modelos estocásticos necesitan ser entrenados a partir de un corpus de frases etiquetadas semánticamente. Este tipo de etiquetado debe asociar las palabras o las secuencias de palabras con su significado correspondiente, que generalmente es representado por una etiqueta semántica. Esto implica un trabajo muy pesado de etiquetado a nivel de palabras, y por añadidura, una posible fuente de errores. Sin embargo, en muchos casos es fácil dar una representación semántica de una frase, si bien es más difícil determinar la asociación entre las palabras y conceptos. Además es menos costoso ofrecer una representación semántica exclusivamente a nivel de frase. Podría ser un logro interesante tener la posibilidad de detectar automáticamente las secuencias de palabras representativas asociadas a cada concepto.

El proceso de aprendizaje propuesto en este trabajo tiene la intención de abordar este problema, es decir, a partir de un corpus de entrenamiento de frases con su representación semántica pretendemos obtener las probabilidades de que las secuencias de palabras sean generadas por un concepto. Estas probabilidades son los $P(w_{i-t+1}^i|c_j)$ que se utilizarán en el

Me gustaría conocer los horarios, precios y tipos de trenes de Valencia a Madrid

```
(HORA)
(PRECIO)
(TIPO-TREN)
CIUDAD-ORIGEN:Valencia
CIUDAD-DESTINO:Madrid
```

Sí, está bien, gracias

```
(AFIRMACION)
```

Figura 5.1: Ejemplos de etiquetado del corpus DIHANA

proceso de comprensión.

El corpus de aprendizaje está compuesto por secuencias de pares (frase, conceptos). Un ejemplo de la representación semántica del corpus se muestra en la figura 5.1, en el que a una frase a se le pueden asignar uno o más conceptos y atributos.

El orden en que se proporciona la información, es decir se estructuran los conceptos, no es un factor relevante, ya que es posible trasmitir el mismo mensaje semántico en diferente orden. En la 5.2 se muestra un ejemplo de dos frases que tienen el mismo significado en las que no sólo las palabras que contienen son diferentes, sino también el orden en que se dan los conceptos.

El proceso de aprendizaje consiste en un procedimiento iterativo que obtiene segmentos de palabras de longitudes desde 1 a $l_{\text{máx}}$ asociados a las clases semánticas, utilizando para ello criterios discriminativos.

El proceso de aprendizaje propuesto inicialmente está compuesto de dos etapas, que se realizan para cada una de las sucesivas longitudes de segmento, $\forall l \in 1...l_{\text{máx}}$:

1. Como no hay información sobre la correlación explícita entre segmentos y conceptos,

Sentence 1: "Sí, me gustaría volver el viernes, ¿cuál es el horario?"

```
(AFIRMACION)
TIPO-VIAJE
(HORA)
FECHA:Viernes
```

Sentence 2: *"Sí, querría saber los horarios para volver el viernes"*

```
(AFIRMACION)
TIPO-VIAJE
(HORA)
FECHA:Viernes
```

Figura 5.2: Ejemplo de frases diferentes con el mismo etiquetado semántico en el corpus DIHANA

en la primera etapa se asigna cada segmento a todos los conceptos que aparecen en el etiquetado semántico de la frase. A partir de esta información se obtiene un conjunto de segmentos asociados a cada concepto.

2. Con el objeto de aumentar la capacidad de discriminar entre conceptos, basándonos en los segmentos que los representan, se hace un proceso de refinamiento y podado de estos conjuntos. Se considera que un segmento es representativo de un concepto si aparece frecuentemente asociado a ese concepto, y no aparece frecuentemente en otros conceptos. A partir de esta información se realiza un podado de los conjuntos utilizando un umbral de pertenencia: sólo los segmentos que tienen alta probabilidad en un conjunto y baja probabilidad en otros se mantienen en el conjunto de segmentos asociados a dicho concepto, y son eliminados de los otros.

   Para ello, se calcula $\forall s_l, c_i$ los valores $P(c_i|s_l)$, donde c_i es un concepto, s_l es un segmento de longitud l y $P(c_i|s_l)$ es la probabilidad de que al observar el segmento s_l

el concepto que se le asocie sea c_i. A partir de estas probabilidades, sólo se permite que permanezcan en el conjunto asociado a cada clase c_i todos los s_l tal que $P(c_i|s_l) >$ umbral.

5.3. Resultados iniciales con umbrales 1 y 0,8

Para los experimentos iniciales dividimos el corpus en dos particiones, una de entrenamiento con el 80 % de las frases y otra de test con el 20 %. Los resultados de los experimentos se muestran en términos de Precisión y Cobertura.

$$\text{Precisión} = \frac{\text{conceptos de referencia detectados correctamente}}{\text{total de conceptos detectados}}$$

$$\text{Cobertura} = \frac{\text{conceptos de referencia detectados correctamente}}{\text{total de conceptos de referencia}}$$

Los experimentos se dividen en dos, en unos se ha tomado como entrada al decodificador semántico las frases de prueba correctamente transcritas ("Transcripción" en la tabla), mientras que en otros la entrada ha sido el resultado del proceso de reconocimiento de las mismas frases de prueba ("Voz" en la tabla).

La Tabla 5.1 contiene los resultados de los experimentos. Además se muestra como varían los resultados en función a la longitud de los segmentos utilizados en los modelos. Como se puede observar el aumento en la longitud de los segmentos influye en una subida de cobertura de los resultados, sobre todo al pasar de longitud 1 a 2. Además se observa una leve pérdida de precisión, que en el caso de los resultados de voz es todavía mayor.

En los experimentos iniciales el umbral que utilizamos para decidir si asociamos un segmento a un concepto es de 1 ($P(c_i|s_l) >$ umbral). En la Tabla 5.2 se muestran los resultados de los mismos experimentos pero utilizando un umbral más bajo, del 0,8. Así conseguimos que el número de segmentos que se asocian a cada concepto sea menos restrictivo. Esto, como era de esperar, influye en un aumento de la cobertura, sobre todo en el caso de los segmentos de longitud 1. Aunque también se produce una pérdida importante de la precisión.

Experimento		Cobertura	Precisión
Longitud 1	Voz	38,1	90,4
	Transcripción	37,8	99,4
Longitud 2	Voz	71,5	88,2
	Transcripción	80,3	98,3
Longitud 3	Voz	75,3	89,8
	Transcripción	84,6	97,6

Tabla 5.1: *Resultados de la experimentación sin categorización con umbral 1*

Experimento		Cobertura	Precisión
Longitud 1	Voz	72,0	85,1
	Transcripción	78,2	92,7
Longitud 2	Voz	83,7	82,3
	Transcripción	89,1	91,0
Longitud 3	Voz	84,1	81,9
	Transcripción	89,3	90,9

Tabla 5.2: *Resultados de la experimentación sin categorización con umbral 0,8*

Capítulo 6

Mejoras en el proceso de aprendizaje

Viendo los resultados al aplicar la propuesta inicial de aprendizaje, se han estudiado varias opciones para mejorar los resultados de cobertura. La opción de aumentar la longitud de los segmentos, es decir, dar un paso más en de iteración en el algoritmo, Como se puede ver ya en la Tabla 5.1 implica una subida en la cobertura del experimento pero no en gran cantidad a partir de longitud 3. Además también hay una pérdida de precisión. Por otra parte, bajar el umbral de pertenencia a un concepto genera más segmentos asociados a cada concepto, como se ha podido observar en los experimentos, pero también implica una pérdida considerable en la precisión.

La opción que hemos escogido es recurrir a una **categorización** que permita generalizar para encontrar en test palabras que no han sido encontradas en entrenamiento pero que sabemos que pertenecen a la misma clase. Por ejemplo, en el caso de las ciudades, utilizar una categoría "nombre-ciudad" que englobe todas las ciudades del corpus.

6.1. Generalización basada en diccionarios

Un aspecto se que hace necesario en el proceso de aprendizaje es la capacidad de generalización cuando se trabaja con corpus de entrenamiento pequeños. Para poder abarcar este

problema hemos aplicado métodos de generalización basados en conocimiento lingüístico.

Un ejemplo de este tipo de categorización definido en nuestra propuesta es considerar el conjunto de días de la semana, meses, o números, como una información a priori que detecta si una palabra pertenece a estas categorías. Sin embargo se mantiene la salvaguarda de que si el mecanismo de categorización genera algún tipo de ambigüedad en alguna situación concreta, entonces no se aplica y se mantiene la palabra. También son utilizados los lemas en lugar de las palabras siempre que no produzcan ambigüedad.

El conocimiento lingüístico que hemos utilizado proviene de varias fuentes distintas de conocimiento. Por un lado podemos obtener información de tipo genérico como números, meses, días de las semana. Por otra parte, hemos recurrido a herramientas online lingüísticas, para obtener nuevas etiquetas semánticas que engloban conceptos con el mismo significado, que pueden ser sinónimos, o que engloban expresiones que se utilizan en el mismo contexto, como es el caso de la etiqueta [time-top] que englobaría todas las expresiones para definir períodos de tiempo (noche, el mediodía, por la tarde, ...). Además de esta información, nos guardamos lemas verbales, plurales y de sexo.

A continuación se muestra un breve resumen de las fuentes de conocimiento lingüístico utilizadas para este trabajo.

6.1.1. Conocimiento a priori

Nos referimos como conocimiento a priori de la tarea a aquella información que disponemos sobre el corpus y que nos ofrece, en algunas casos, información disponible sobre el vocabulario de la tarea cuando ésta comprende un ámbito restringido.

- Conocimiento a priori de tipo genérico. Dentro del conocimiento genérico a cualquier tarea podemos incluir información sobre:
  - Listados de números: ya sea definiendo únicamente la etiqueta número o separando en unidades, decenas, centenas, etc. Además de números de orden.

- Listados sobre ubicaciones físicas: países y ciudades.
- Listados sobre fechas: días de la semana, meses, nombres de festividades.

- Conocimiento de la tarea. En el corpus DIHANA, al tratarse de una tarea sobre información de trenes. disponemos de listados sobre los tipos de trenes, los tipos de billetes y servicios ofrecidos por los trenes.

6.1.2. Herramientas lingüísticas

Además de la información que tenemos a priori del corpus, podemos beneficiarnos del herramientas de uso libre disponibles en la Web, para recapitular información lingüística sobre la tarea.

Del gran abanico de herramientas que podemos encontrar en internet, utilizaremos las dos que se muestran a continuación.

Analizador morfológico STILUS

La herramienta de análisis morfológico STILUS [1], en un primera fase para el tratamiento texto real, antes de proceder a su análisis, procesa el texto a fin de detectar y marcar adecuadamente sus partes básicas. Estas tareas incluyen, entre otras: la descomposición del texto en párrafos y éstos en frases; la detección de abreviaturas, siglas y acrónimos; la detección de nombres propios, topónimos, extranjerismos, arcaísmos, etc.; la detección de citas textuales (comillas, paréntesis, guiones) y la detección e interpretación de cifras, numerales y ordinales.

Ejemplo de etiquetado obtenido por STILUS.

Frase de entrada: quiero salir por la mañana

Salida del analizador:

[1] http://stilus.daedalus.es/stilus.php

- quiero: Verbo léxico transitivo 1ª persona singular presente indicativo (VI-S1PTL-N6). Lema querer
- salir: Verbo léxico intransitivo infinitivo (VN—OIL-N6)
- por-la-mañana: Adverbio (E-X-N5). Lema por la mañana

De toda la información que nos proporciona STILUS, utilizaremos para nuestros diccionarios las etiquetas sintácticas y semánticas, y los lemas.

Herramienta Wordnet

MultiWordNet [2] es una base de datos léxica multilingüe, la cual está estrictamente alineada con la base de datos WordNet [3].

WordNet es una base de datos léxica originalmente en inglés. Los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos (synsets), cada uno expresando un concepto distinto. Estos a su vez están vinculados entre sí por medio de relaciones conceptuales semánticas y léxicas. La red resultante de palabras y conceptos relacionados por significado, está constituida de forma que es fácilmente accesible a través de un navegador.

Ejemplo de sinónimos del corpus DIHANA encontrados en MultiWordNet:

barato = económico = [barato]
vuelta = regreso = [volver]
salir = partir = [salir]

6.1.3. Cómo usamos el conocimiento lingüístico

Para cada uno de los segmentos que cumplen $P(c_i|s_l) >$ umbral guardamos una entrada en nuestros diccionarios.

[2] http://multiwordnet.fbk.eu
[3] http://wordnet.princeton.edu

Las entradas del diccionario contienen la siguiente información sobre un segmento:

- Lema sintáctico
- Lema semántico
- Etiqueta sintáctica
- Etiqueta semántica
- Números de segmentos iguales en el corpus
- Número de etiquetas (sintácticas o semánticas) iguales en el corpus
- Número de apariciones del concepto

Ejemplo de bolsa HORA en el diccionario:

noche : : [time-top] : [nombre-hora] : [time-top] : 60 : 83 : 645
cincuenta : [num] : [num] : : [num-hora] : 9 : 17 : 645
pronto : : : : : 5 : 5 : 645
media : : : : : 35 : 35 : 645
pudiera : poder : : [verbo-hora] : : 2 : 9 : 645
antes-de : : antes-de : : : 109 : 109 : 645
las-once : las-[num] : -el-[num] : [numeral-hora] : [num-hora] : 37 : 489 : 645
por-la : por-la : por-'el : : : 221 : 284 : 645
la-tarde : la-[time-top] : el-[time-top] : [nombre-hora] : [time-top] : 221 : 504 : 645
las-siete : las-[num] : el-[num] : [numeral-hora] : [num-hora] : 47 : 489 : 645
catorce-horas : : [num]-horas : [nombre-hora] : : 5 : 44 : 645
del-mediod'ia : del-[time-top] : del-[time-top] : [nombre-hora] : [time-top] : 12 : 504 : 645
y-media : y-media : y-media : : : 34 : 34 : 645

El proceso de aprendizaje y de comprensión intenta en primer lugar utilizar la palabra, pero en caso de que no la encuentre utilizará el lema o la categoría. Debe tenerse en cuenta que este mecanismo de generalización está fuertemente restringido para mantener la capacidad discriminatoria de los segmentos asociados a las unidades semánticas. Esta generalización ayuda, además de a encontrar segmentos que no están explícitamente en el corpus de aprendizaje, a agrupar las probabilidades por conceptos, lo que ayuda a compensar las probabilidades entre segmentos de longitud pequeña con los de longitud mayor.

Algunas de las características de nuestra categorización son las siguientes:

- No se sustituyen los segmentos por sus lemas si no que se guarda la información de ambos en el diccionario.
- En la primera iteración del algoritmo se decide cuáles serán las categorías sintácticas que pertenecen a cada uno de los conceptos.
- En iteraciones posteriores sólo se utilizarán aquellas categorías sintácticas que han aparecido en la primera iteración.

6.2. Criterios de desambiguación y poda de segmentos

6.2.1. Desambiguación de segmentos

Dado un segmento s de longitud l donde $P(c_i|s)$ es la probabilidad de que al observar el segmento s el concepto que se le asocie sea c_i. Decimos que es un segmento ambiguo cuando existe más de un c que cumple $P(c_i|s_l) == 1$.

Gracias al uso de información lingüística podemos definir criterios por los que estos segmentos pueden ser desambiguados. Algunos de los criterios que podemos utilizar son los siguientes:

- Por cada ci en la que s es ambiguo, si lema sintáctico o lema semántico de s se encuentra

en ci y sólo en ci, entonces s pertenece a ci.

- Por cada ci en la que s es ambiguo, si s está compuesto por un conjunto de varios subsegmentos, decimos s que pertenece a ci, si cualquier subsegmento sufijo de s fue asociado a ci en iteraciones anteriores y sólo a ci.

6.2.2. Poda de segmentos no correspondientes a un concepto

Además de los criterios de desambiguación mencionados, podemos definir, por el contrario, unos criterios de poda de segmentos. Se elimina un segmento cuando:

- Un concepto está asociada siempre a una categoría semántica:
  - CIUDAD-ORIGEN: [nom-ciudad-o], CIUDAD-DESTINO: [nom-ciudad-d], CIUDAD: [nom-ciudad-c]
  - CLASE-BILLETE: [clase-billete]
  - TIPO-TREN: [tipo-tren]
- En segmentos de longitud 2 o superiores:
  - Los subsegmentos de los que se compone el segmento aparecen asociados a conceptos diferentes. (Frontera entre bolsas). Esto es, si s aparece asociado a un c_i de longitud l y s se puede descomponer en un conjunto subsegmentos sc de longitud menor a l, entonces decimos que s es un segmento frontera, si y sólo si, cualquier sc_j ha aparecido en iteraciones anteriores asociado un concepto c_k diferente a c_i.

6.2.3. Añadir símbolo inicial y final

Una de las cuestiones más recurrentes cuando se trabaja con corpus de frases es si incluir el símbolo inicial y final de frase. Existen segmentos que en frases de longitudes cortas tienen un significado que difiere de cuando se encuentra es frases más largas, y utilizar los

símbolos inicial y final nos ayuda a encontrar estas singularidades. Es decir, porque contiene un segmento y un sólo concepto.

Añadir estos símbolos a los modelos implica en nuestro caso un incremento en el número de segmentos asociados a cada concepto. En lugar de esta solución, hemos optado por añadirlos sólo a las frases cuyo número de palabras coincide con la longitud de segmento de la iteración del algoritmo, y que sólo contiene un segmento asociado. Es decir, en la iteración 1; en frases de longitud 1 y un concepto, en la segunda iteración; longitud 2 y un concepto, etc.

Un ejemplo se muestra a continuación:
<no>: (NEGACION)
<Valencia>:CIUDAD
<sí>: (AFIRMACION)
<eso es>: (AFIRMACION)
<la hora>: (HORA)
<el precio>: (PRECIO)

El uso de estos símbolos mejora la precisión de los experimentos sin afectar a la cobertura de los mismos.

6.3. Algoritmo de aprendizaje mejorado

A partir de las mejoras nombradas nos proponemos hacer un refinamiento del algoritmo de aprendizaje inicial presentado en 5.2.

El proceso de aprendizaje consiste, nuevamente, en un procedimiento iterativo que obtiene segmentos de palabras de longitudes desde 1 a $l_{\text{máx}}$ asociados a las clases semánticas, utilizando criterios discriminativos.

El proceso de aprendizaje se compone en esta ocasión de cuatro etapas, que se realizan para cada una de las sucesivas longitudes de segmento, $\forall l \in 1...l_{\text{máx}}$. Este algoritmo es el

mismo que el del Apartado 5.2, añadiendo el paso 3.

1. Al no haber información sobre la correlación entre segmentos y conceptos, en la primera etapa se asigna cada segmento a todos los conceptos que aparecen en el etiquetado semántico de la frase. A partir de esta información se obtiene un conjunto de segmentos asociados a cada concepto.

2. Con el objeto de aumentar la capacidad de discriminar entre conceptos, basándonos en los segmentos que los representan, se hace un proceso de refinamiento y podado de estos conjuntos. Se considera que un segmento es representativo de un concepto si aparece frecuentemente asociado a ese concepto, y no aparece frecuentemente en otros conceptos. A partir de esta información se realiza un podado de los conjuntos utilizando un umbral de pertenencia: sólo los segmentos que tienen alta probabilidad en un conjunto y baja probabilidad en otros se mantienen en el conjunto de segmentos asociados a dicho concepto, y son eliminados de los otros.

   Para ello, se calcula $\forall s_l, c_i$ los valores $P(c_i|s_l)$, donde c_i es un concepto, s_l es un segmento de longitud l y $P(c_i|s_l)$ es la probabilidad de que al observar el segmento s_l el concepto que se le asocie sea c_i. A partir de estas probabilidades, sólo se permite que permanezcan en el conjunto asociado a cada clase c_i todos los s_l tal que $P(c_i|s_l) >$ umbral.

3. En la tercera etapa, para aumentar la cobertura del modelo y poder tratar con las realizaciones léxicas de ciertas categorías naturales que son bien conocidas a priori aunque posiblemente no hayan sido observadas en el entrenamiento, se realiza el proceso de categorización basado en criterios lingüísticos y en diccionarios, explicado en la Sección 6.

4. Para que el proceso incremental de ir construyendo segmentos de longitudes cada vez mayores mantenga un criterio discriminativo que elimine al máximo las ambigüedades, en cada conjunto obtenido en el paso anterior se podan los segmentos, que están compuestos por otros más cortos que pertenecen a algún otro diccionario. Así por ejemplo,

cuando se construyen segmentos de longitud dos se comprueba que las dos palabras que los componen, o las categoría ellas asociadas, no sean las palabras representativas de otras clases antes de asignarlos a una nueva clase. El objetivo de este procedimiento discriminativo es encontrar segmentos de diversas longitudes que caracterizan las unidades semánticas. Hay muchos casos en los que cuando se aumenta la longitud de los segmentos se puede discriminar mejor entre palabras que son semánticamente ambiguas si se las considera aisladamente. Este es el caso, por ejemplo, de la palabra "Valencia" que puede asociarse a Ciudad-Origen o Ciudad-Destino, pero cuando se consideran segmentos de longitud dos la secuencia "a Valencia" se debe asignar claramente a Ciudad-Destino.

6.3.1. Experimentos con el algoritmo mejorado

En los resultados de los experimentos que se muestran en la Tabla 6.1, se busca mostrar las diferencias entre el algoritmo de aprendizaje inicial, que mostrábamos en la Sección 5.2 y las diferentes mejoras que se han propuesto a lo largo de este capítulo.

Para realizar los experimentos hemos dividido el corpus en 5 particiones donde cada una contiene el 20 % de las frases del corpus. De este modo nos proponemos realizar una Validación Cruzada. Por lo tanto, se realizarán 5 experimentos donde se entrenará con 4 de las particiones (el 80 % del corpus) y se hará el test con la restante.

Para todos los experimentos se ha mantenido el umbral de pertenencia a un conjunto (o concepto) con probabilidad igual a 1. Los resultados de los experimentos se muestran en términos de Precisión y Cobertura, como anteriormente, además en este caso se ha calculado el Concept Accuracy (que es el equivalente al Word Accuracy pero usando conceptos como unidades), con el fin de poder comparar nuestros experimentos.

$$\text{Concept Accuracy} = 1 - \frac{\text{Conceptos insertados/borrados/substituidos}}{\text{Conceptos de referencia}}$$

En el primer experimento ("Sin categorización" en la tabla) utilizamos el corpus sin ningún preprocesado, al igual que hacíamos en la sección 5.3. En el resto de experimentos se

ha utilizado la categorización mostrada en la Sección 6.

Para mostrar las diferentes opciones de clasificación de segmentos que disponemos, se muestran 4 tipos de experimentos de categorización. En el primero ("Categorización SD SP" en la tabla) se utilizaron categorías pero no se hizo ninguna limpieza ni desambiguación posterior de los segmentos incluidos en las bolsas. En el experimento "Categorización D P" de la tabla se muestran los resultados aplicando las técnicas de desambiguación y limpieza de segmentos. Los otros 2 experimentos corresponden a la combinación de usar sólo desambiguación, o sólo poda.

El uso de información lingüística extra mejora el comportamiento del sistema cuando se usan transcripciones correctas como entrada. Esto es coherente con el hecho de que la estructura lingüística es más correcta. Además, se puede observar como la desambiguación y poda mejoran la precisión del método.

Con el fin de poder comparar con los resultados de comprensión cuando existe segmentación explícita y anotación del corpus de entrenamiento, se ha realizado otro experimento. En este caso entrenamos modelos estocásticos a partir corpus segmentado y etiquetado, como en [7]. Los resultados correspondientes del proceso de decodificación se presenta también en la Tabla 6.1 (Segmentados manualmente en la tabla). Se utilizó el CMU-sphinx2 como reconocedor y obtuvo una precisión de Concept Accuracy de 82 %.

Los resultados muestran un mejor comportamiento del modelo cuando se utilizaron las transcripciones correctas, como se esperaba. Comparando el primer y el segundo experimento se puede apreciar que el uso de la categorización mejora el comportamiento del sistema cuando se utilizan transcripciones correctas. Sin embargo, los resultados no son mejores cuando la entrada corresponde a voz. Podría ser debido al hecho de que el proceso de reconocimiento genera errores que se propagan al proceso de categorización. Por otro lado, los resultados de los modelos que se han aprendido de corpus de entrenamiento segmentado superar el enfoque que utiliza menos información en la formación. Sin embargo, nuestro enfoque tiene la ventaja de la reducción del esfuerzo en la preparación del corpus.

Experimento	% Cobertura	% Precisión	% Concept Accuracy
Sin categorización	83,9	97,7	82,4
Categorización SD SP	95,7	94,2	89,8
Categorización SD P	95,5	95,7	91,7
Categorización D SP	95,8	95,0	91,1
Categorización D P	95,6	95,7	91,7
Segmentación manual	-	-	96,0

Tabla 6.1: *Comparación de resultados de los experimentos.*

Parte IV

Aprendizaje activo

Capítulo 7

Experimentos de aprendizaje activo

Para demostrar la capacidad de generalización de nuestros modelos y teniendo en cuenta que la finalidad de este método es reducir el número de frases que han de ser segmentadas manualmente, se han propuesto dos experimentos de aprendizaje activo, en los cuáles, partiremos entrenando los modelos con pequeño porcentaje del corpus, que a partir de unas correcciones posteriores mostraremos que nos permite crear unos modelos tan robustos como los entrenados con todo el corpus.

A continuación se muestran dos experimentos de aprendizaje activo. En el primer experimento se utilizará el proceso de aprendizaje descrito en 6.3. En el segundo el proceso de aprendizaje está basado en modelos estocásticos de estados finitos.

7.1. Experimento 1

El proceso de aprendizaje activo que proponemos intenta reducir el número de frases que hay que etiquetar semánticamente, ya que este es un proceso manual de un alto coste temporal. Intentaremos, por lo tanto, que en la fase de construcción de los modelos, aquellas frases que se etiqueten sean lo más representativas posibles de los frames que puedan aparecer, así como de las diversas formas en que se pueden expresas esas frames. Esto lo conseguimos

gracias a generalización basada en conocimiento lingüístico adquirido durante la fase de aprendizaje.

El algoritmo para el proceso de aprendizaje activo parte de un primer modelo semántico aprendido a partir de un porcentaje pequeño del corpus. Posteriormente, mediante un proceso incremental, en cada iteración se analiza un nuevo conjunto de frases a partir del cual se adaptan los modelos.

Este proceso incremental consiste en realizar el proceso de decodificación semántica de las nuevas frases generando dos conjuntos: uno de ellos formados por las frases que han dado un alto valor de confianza, y otro formado por aquellas frases que han proporcionado bajo valor de confianza. Estas últimas frases son analizadas y, si procede, etiquetadas manualmente. De esta forma algunas de ellas se pueden descartar y otras pueden aportar nueva información a los modelos.

Para el proceso incremental se ha escogido como medida de confianza la probabilidad de decodificación semántica de la frase, dados los modelos.

7.1.1. Resultados

Para realizar estos experimentos hemos dividido el corpus en 5 particiones. De forma que cada una contiene el 20 % de las frases del corpus. Cuatro de las particiones se utilizarán para entrenamiento, y una para el test.

En la Tabla 7.1 se muestran los resultados de los experimentos en términos de precisión y cobertura correspondientes al proceso de aprender los modelos a partir de las muestras etiquetadas manualmente en término de frames. Esto se ha realizado para los diferentes subconjuntos de training, siempre reconociendo el mismo conjunto de test (partición 1). Como puede verse hay una clara mejora cuando se pasa del 20 % al 40 % del corpus de aprendizaje, que luego se estabiliza o incluso empeora. Ello podría deberse a que haya una fuerte correlación en estos dos primeros conjuntos de aprendizaje con el conjunto de test, por lo que las características de test están muy bien representadas en estos dos conjuntos. Sin embargo al

añadir más información a los modelos se introduce más ruido por lo que empeoran algo los resultados.

Para el proceso de aprendizaje activo se ha escogido el segundo subconjunto de training. Es decir, partiendo de los modelos aprendidos con el primer 20 % se ha realizado el proceso de comprensión del segundo 20 %. Usando como medida de confianza la probabilidad que proporciona el algoritmo de comprensión para cada frase, se ha ordenado y posteriormente dividido este segundo conjunto de aprendizaje en dos partes: el 80 % de frases comprendidas con mayor probabilidad, y el restante 20 % de frases comprendidas con menor probabilidad. El primer subconjunto se ha incluido directamente en el proceso de reentrenamiento de los modelos y el segundo subconjunto se ha revisado manualmente para corregir, si procede, los errores de comprensión. Una vez realizado el proceso de corrección se ha usado también para el reentrenamiento de los modelos. Tras este proceso de han aplicado los nuevos modelos al proceso de comprensión del conjunto de test, dando un 96,5 % de cobertura y un 95,8 % de precisión. Es decir, con un esfuerzo mucho menor que el proceso original que requiere el etiquetado de todas las frases, se han obtenido los mismos resultados, que en este caso son casi iguales que los que se obtienen entrenando con el 80 % del corpus. Aunque estos resultados son preliminares, dadas las características de este corpus, en el que la información semántica se puede capturar con un pequeño conjunto de muestras, indican que la aproximación es prometedora y que puede reducir, sin empeoramiento de los resultados, el esfuerzo de etiquetado del corpus.

7.2. Experimento 2

7.2.1. Proceso de aprendizaje

Aplicamos una aproximación para la comprensión del lenguaje basada en el aprendizaje automático de los modelos estocásticos de estados finitos [7]. El conocimiento sintáctico y semántico involucrado en el proceso de comprensión es modelado por una representación de

Experimento		Sin Categorizar	Original	Active Learning
20 %	Cobertura	78,5	93,7	-
	Precisión	95,2	94,4	-
40 %	Cobertura	81,2	94,5	96,5
	Precisión	96,2	95,4	95,8
60 %	Cobertura	82,7	94,7	-
	Precisión	97,2	95,3	-
80 %	Cobertura	83,9	95,5	-
	Precisión	97,7	95,7	-

Tabla 7.1: *Resultados de la experimentación*

dos niveles. Los modelos son aprendidas de forma automática e integrados en un autómata que se utiliza en el proceso de comprensión. Desde este punto de vista, el objetivo del proceso de comprensión consiste en la obtención de la secuencia de unidades semánticas (unidades de frame) asociadas a la frase de entrada, así como los segmentos de palabras asociados a las unidades semánticas. En el enfoque estocástico que se aplica, deben ser aprendidos dos tipos de modelos: un modelo semántico que representa las concatenaciones de unidades semánticas, y un modelo para cada unidad semántica que representa el conjunto de secuencias de palabras asociadas a esa unidad semántica.

Con el fin de aprender estos modelos estocásticos, debe estar disponible un conjunto de secuencias de unidades semánticas asociadas a las frases de entrada, así como la segmentación correspondiente. En otras palabras, digamos que W es el vocabulario de la tarea, y sea V el alfabeto de unidades semánticas, el conjunto de entrenamiento es un conjunto de pares (u, v) donde:

$$u = u_1, u_2 \dots u_{\mathrm{n}},\ u_{\mathrm{i}} = w_{\mathrm{i}_1} w_{\mathrm{i}_2} \dots w_{\mathrm{i}_{|u_{\mathrm{i}}|}},\ w_{\mathrm{i}_{\mathrm{j}}} \in W,\ i = 1, \dots, n,\ j = 1, \dots, |u_{\mathrm{i}}|,$$
$$v = v_1 v_2 \dots v_{\mathrm{n}},\ v_{\mathrm{i}} \in V,\ i = 1, \dots, n.$$

El enfoque que aplicamos a la comprensión del lenguaje consiste en aprender dos tipos

de modelos de estados finitos de un conjunto de pares de entrenamiento (u, v). Se estima un modelo de A_{s} para el *lenguaje semántico* a partir de las secuencias de unidades semánticas asociadas a las frases entrada. Se estima un conjunto de modelos, *modelos de unidades semánticas* $A_{\mathrm{v_i}}$ (uno para cada unidad semántica $v_{\mathrm{i}} \in V$), de todos los segmentos de palabras asociados a esa unidad semántica. El modelo semántico A_{s} representa la información semántica proporcionada por los datos de entrenamiento, y cada modelo de unidad semántica $A_{\mathrm{v_i}}$ representa la información sintáctica para la unidad semántica correspondiente v_{i}, que también es proporcionada por los datos de entrenamiento. Todas estas estimaciones se hacen a través de técnicas de aprendizaje automático.

Para el proceso de comprensión, todos los modelos deben ser combinados con el fin de aprovechar de todas las restricciones sintácticas y semánticas. Para ello, los estados del autómata estocástico A_{s} son sustituidos por los autómatas estocásticos correspondientes $A_{\mathrm{v_i}}$.

Una vez que este autómata integrado A_{t} se construye, el proceso de comprensión consiste en encontrar el mejor camino en este autómata dada la frase de entrada.

Es decir, dada la frase de entrada $w = w_1 w_2 \ldots w_n \in W^*$, el proceso consiste en encontrar la secuencia de unidades semánticas $v = v_1 v_2 \ldots v_k \in V^*$ que maximiza la probabilidad:

$$\widehat{v} = \underset{v}{\operatorname{argmax}}\, P(w|v)P(v)$$

El término $P(w|v)$ es la probabilidad de la secuencia de palabras w dada la secuencia de unidades semánticas v y $P(v)$ es la probabilidad de la secuencia de unidades semánticas.

El proceso de comprensión se realiza utilizando el algoritmo de Viterbi, que proporciona el mejor camino en el modelo integrado.

Este camino no sólo da la secuencia de unidades semánticas, sino que también da la segmentación asociada a ella.

7.2.2. Algoritmo de aprendizaje activo

Con el fin de reducir el esfuerzo de etiquetar un gran número de muestras de entrenamiento, y tener la posibilidad de adaptar dinámicamente los modelos cuando los usuarios reales están interactuando con el sistema, hemos propuesto un proceso de dos pasos. De esta manera, sólo un pequeño conjunto de muestras de entrenamiento deberá estar inicialmente etiquetado (pero no segmentado) en términos de frames. Durante la interacción posterior con el usuario, y siguiendo un enfoque de aprendizaje activo, las nuevas frases observadas por el sistema se incorporan directamente al aprendizaje iterativo de los modelos o son analizados manualmente (en función de sus medidas de confianza). El proceso de aprendizaje activo para cada nuevo conjunto de oraciones es el siguiente:

- Hacer la comprensión de un nuevo conjunto de frases.
- Seleccionar un subconjunto de estas oraciones basándose en una medida de confianza, las de menos confianza.
- Corregir, en su caso, la decodificación semántica de las muestras seleccionadas.
- Adaptar el modelo de comprensión teniendo en cuenta que este nuevo conjunto de muestras no sólo contiene las frases con confianza alta, sino también las frases que se corrigen manualmente.

Un punto importante en este proceso es el criterio para seleccionar las frases que se analizaron manualmente. En nuestro caso, como el proceso de comprensión se basa en estadísticas de aparición de las secuencias de palabras cuando una unidad semántica es encontrada, la medida de confianza se basa en la probabilidad de este evento. Para cada par $(u_i v_i)$, una combinación lineal de dos medidas se utiliza para determinar si la asignación del segmento de u_i para la unidad semántica v_i se ha hecho correctamente durante el proceso de decodificación:

- $\frac{logP(u_i|v_i)}{|u_i|}$ es la probabilidad del segmento u_i dentro de la unidad semántica v_i normalizada en función del número de palabras en el segmento. Esta medida es más sensible

a las variaciones sintácticas.

- $\frac{log \prod_{w_j \in u_i} P(w_j|v_i)}{|u_i|}$ es la misma probabilidad, pero considerando sólo la probabilidad del unigrama. Esta medida es más sensible a las palabras fuera de vocabulario.

Las oraciones que contengan uno o más segmentos con un valor bajo para la combinación lineal de estas medidas son revisadas manualmente. Estas frases suelen contener palabras o las relaciones sintácticas entre palabras que no se han visto en el corpus de entrenamiento.

7.2.3. Resultados

Se realizaron los siguientes experimentos para evaluar las técnicas propuestas. Elegimos 80 % del corpus con fines de entrenamiento y desarrollo, y el 20 % para test. En todos los casos, se presentan los resultados teniendo en cuenta la transcripción correcta de las frases de test (TEST.txt) y la salida del proceso de reconocimiento de las frases de test (TEST.speech). La precisión a nivel palabra del reconocimiento de voz fue 76,0 %. Los resultados se dan en términos de:

- el porcentaje de frames correcto (% cf), es decir, el porcentaje de frames que son exactamente los mismos que el sistema de referencia correspondiente.
- el porcentaje de unidades de frames correcto (conceptos y sus atributos) (% ufc).

Se llevaron a cabo tres tipos de experimentos. Exp1 es el experimento de referencia, en el que hemos entrenado el modelo estadístico de dos niveles, con el corpus segmentado y etiquetado de forma manual. Exp2 es una evaluación de la capacidad de segmentación y etiquetado del algoritmo de aprendizaje semi-supervisado, el corpus de entrenamiento consiste en la transcripción de las frases y sus frames correspondientes, pero sin ningún tipo de asociación de unidades de frame a los segmentos. Exp3 es una evaluación del proceso de aprendizaje activo.

En el proceso de aprendizaje activo, el corpus de entrenamiento se dividió en cuatro subgrupos (TR20.1, TR20.2, TR20.3 y TR20.4) de modo que cada subconjunto contenía 20 % del número total de frases en el corpus. Los modelos de comprensión diferente obtenidos en cada paso son los modelos estadísticos de dos niveles. El proceso fue el siguiente:

- Considerando el primer 20 % de las frases (TR20.1) y sus frames asociados correspondientes. Se ha realizado una segmentación automática y etiquetado mediante el algoritmo semi-supervisado. De estos datos segmentados y etiquetados, se entrenó el modelo de comprensión base.
- Con este modelo de comprensión, se llevó a cabo un proceso de comprensión del subconjunto siguiente (TR20.2).
- Utilizando las medidas de confianza generadas en este proceso de comprensión, una parte de las frases procesadas es seleccionada para analizar de forma manual y, si fuese necesario, corregir la segmentación o etiquetado. En lugar de encontrar un umbral de las medidas de confianza, se seleccionaron el 20 % de los segmentos con la menor medida de confianza.
- Después de este proceso de corrección, se genera un nuevo corpus de entrenamiento. Este corpus incluye el original y el nuevo corpus TR20.2 segmentado y etiquetado. El corpus TR20.2, a su vez, se compone de las frases que fueron etiquetados automáticamente por el proceso de comprensión, donde se han corregido una pequeña parte de ellas de forma manual. Usando este nuevo corpus de entrenamiento se aprende un modelo nuevo.
- Este proceso se repite para los subconjuntos (TR20.3) y (TR20.4).

Se presentan los resultados del proceso de comprensión sobre el test para todos los pasos, es decir, teniendo en cuenta los diferentes modelos aprendidos de forma incremental.

La Tabla 7.2 muestra los resultados del Exp1. Se puede ver, como se esperaba, que los valores de cfu son siempre superiores a los valores de cf, porque cf es una medida más estricta:

un error en una unidad de frame implica un error en todo el frame. Como era de esperar también, los resultados para la entrada de voz son peores que los de la transcripción correcta de las frases de prueba.

	cf	cfu
TEST.txt	91,7	96,1
TEST.speech	65,9	79,8

Tabla 7.2: *Exp1 Experimento de referencia.*

La Tabla 7.3 muestra los resultados del Exp2. Un incremento en la cantidad de los datos de entrenamiento implica una mejora en todas las medidas. Utilizando el conjunto completo de datos de entrenamiento los resultados no mejoran los del Exp1, pero hay que tener en cuenta que en Exp2 sólo se utilizaron la datos de entrenamiento etiquetados (Exp1 utiliza segmentado y etiquetado los datos de entrenamiento).

	TEST.txt		TEST.speech	
	cf	cfu	cf	cfu
20 %	78,7	89,5	60,2	76,6
40 %	80,4	91,4	60,7	77,0
60 %	81,5	91,4	61,5	78,3
80 %	84,9	92,9	63,6	78,5

Tabla 7.3: *Exp2 Proceso de etiquetado semi-supervisado.*

La Tabla 7.4 muestra los resultados del Exp3. Un incremento en la cantidad de los datos de entrenamiento implica una mejora en todas las medidas, sin segmentación y etiquetado de los datos de entrenamiento. Sólo el 20 % de cada nuevo conjunto de entrenamiento ha sido segmentado y etiquetado. Los resultados del Exp3 superan los del Exp2. Los resultados en la última fila de Exp3, es decir, considerando todos los datos de entrenamiento, son muy similares a los del Exp1, con una mucho menor segmentación manual y esfuerzo de etiquetado.

	TEST.txt		TEST.speech	
	cf	cfu	cf	cfu
TR20.1	78,7	89,5	60,2	76,6
TR20.2	83,4	92,3	62,4	78,1
TR20.3	85,4	93,4	63,6	79,0
TR20.4	87,0	94,1	64,0	79,7

Tabla 7.4: *Exp3 Evaluación del proceso de aprendizaje activo.*

Parte V

Conclusiones y trabajo futuro

Conclusiones

En este trabajo fin de máster hemos presentado una aproximación al desarrollo del módulo de comprensión de un sistema de diálogo hablado. Los modelos semánticos se aprenden automáticamente a partir de un corpus de entrenamiento en el cual el proceso de etiquetado ha sido simplificado. El uso de este método permite que sólo se requiera de una anotación global de la frase para entrenar los modelos, en lugar de un etiquetado detallado de palabra o segmento de palabras a concepto, como era antes necesario.

Resulta prometedora la capacidad que ha demostrado el método propuesto de encontrar los segmentos de palabras apropiados que pueden ser asociados a los diferentes conceptos. Los experimentos muestran que esta aproximación ofrece buenos resultados, requiriendo menos esfuerzo en el etiquetado y evitando los errores propios de una segmentación manual.

Además, nuestro segmentador ha sido utilizado para un proceso de aprendizaje activo que mediante un pequeño porcentaje de las muestras del corpus es capaz de obtener resultados similares a aquellos obtenidos haciendo un proceso de entrenamiento mucho más costoso. Los modelos aprendidos mediante este método muestran por lo tanto una gran capacidad de generalización. Esto se debe sobre todo al uso de herramientas lingüísticas, que permiten obtener buenos resultados con pocas muestras. Además, gracias a un pequeño reaprendizaje del corpus conseguimos refinar los modelos con un mínimo esfuerzo de forma que los resultados igualan a aquellos obtenidos entrenando los modelos con el corpus completo. Por los buenos resultados que ofrece este proceso se espera que al aplicar este método a tareas más complejas y de mayor ambigüedad se puedan dar también buenos resultados.

Trabajo futuro

A raíz de los resultados obtenidos en los experimentos y viendo lo adecuado del método para la segmentación de corpus, las líneas de trabajo futuro relacionadas, están orientadas en gran parte a usar el algoritmo como herramienta de segmentación de corpus.

Además se proponen algunas mejoras en la aproximación presentada:

- En primer lugar se propone un estudio detallado de como afecta la elección del umbral de aceptación de palabras a las prestaciones del sistema.
- También existe la posibilidad de utilización de otros métodos de clasificación o clustering para la estimación de las probabilidades $P(c_i|s_l)$.
- A pesar de que la mayoría de los segmentos significativos han sido encontrados entre longitudes de entre 1 y 3, sería conveniente probar con segmentos de longitud mayor.
- Una de mejoras que se ha planteado en este trabajo es la poda de segmentos no significativos. Una de las propuestas de mejora del método consistiría en encontrar los segmentos no significativos con métodos basados en criterios de máxima verosimilitud.
- El método desarrollado para este trabajo fin de máster, es totalmente independiente del corpus, sin embargo sería conveniente aplicar este método a tareas más complejas, donde las realizaciones léxicas de los conceptos presenten más intersecciones y ambigüedad.

Publicaciones relacionadas

Las publicaciones relacionadas con este trabajo son las que se muestran a continuación, en las dos primeras se presentó el método de etiquetado automático y los experimentos relacionados. En el siguiente se ha utilizado el método para segmentado de texto, y se han realizado experimentos de aprendizaje activo.

- L. Ortega, I. Galiano, Lluís-F. Hurtado, E. Sanchis, E. Segarra. .ª Statistical Segment-Based Approach for Spoken Language Understanding", The Eleven Annual Conference of the International Speech Communication Association. (INTERSPEECH'10). Proc. pp.1836-1839, Makuhari (JAPAN). Oct. 2010 (CORE A)

- L. Ortega, I. Galiano, L. F. Hurtado, E. Sanchis, E. Segarra. Ün método de aprendizaje semi-supervisado para la modelización semántica en comprensión del habla", Procesamiento del Lenguaje Natural. Número 45 pp.199-205, 2010

- Lucía Ortega, Isabel Galiano, Emilio Sanchis, "Minimizando el etiquetado manual en la modelización estadística para la comprensión del habla". Procesamiento del Lenguaje Natural. 2011.

Agradecimientos

Este trabajo es apoyado por el MEC y FEDER bajo contrato TIN2008-06856-C05-02. También me gustaría dar las gracias a todos los que han trabajado conmigo durante este tiempo por su asesoramiento y ayuda durante el proceso de este Trabajo Fin de Máster. Gracias.

Bibliografía

[1] Min Tang, Xiaoqiang Luo, and Salim Roukos, "Active learning for statistical natural language parsing," in *In Proceedings of ACL 2002*, 2002, pp. 120–127.

[2] Giuseppe Riccardi and Dilek Hakkani-Tür, "Active learning: theory and applications to automatic speech recognition," in *Ieee Transactions On Speech And Audio Processing*, 2005, vol. 15, pp. 2105–5113.

[3] Gokhan Tur, Dilek Hakkani-Tür, and Robert E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," in *Speech Communication*, 2005, vol. 45, pp. 171–186.

[4] Pierre Gotab, Frederic Bechet, and Geraldine Damnati, "Active learning for rule-based and corpus-based spoken labguage understanding moldes," in *IEEE Workshop Automatic Speech Recognition and Understanding (ASRU'09)*, 2009, pp. 444–449.

[5] Ye-Yi Wang, "Strategies for statistical spoken language understanding with small amount of data - an empirical study," in *Proc. of InterSpeech 2010*, Makuhari, Chiba, Japan, 2010, pp. 2498–2501.

[6] Lucía Ortega, Isabel Galiano, Lluís-F. Hurtado, Emilio Sanchis, and Encarna Segarra, "A statistical segment-based approach for spoken language understanding," in *Proc. of InterSpeech 2010*, Makuhari, Chiba, Japan, 2010, pp. 1836–1839.

[7] E. Segarra, E. Sanchis, M. Galiano, F. García, and L. Hurtado, "Extracting Semantic Information Through Automatic Learning Techniques," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 3, pp. 301–307, 2002.

[8] Tom Mitchell, *Machine Learning*, McGraw Hill, 1997.

[9] Lawrence R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE, 77(2)*, 1989, p. 257–286.

[10] A. McCallum J. Lafferty and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, Williamstown, MA, USA, 2001, p. 282–289.

[11] Y. He and S. Young, *Semantic processing using the hidden vector state model*, p. 19:85–106, 2005.

[12] Christian Raymond and Giuseppe Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Interspeech*, Antwerp, Belgium, 2007, p. 1605–1608.

[13] Geoffrey Zweig and Stuart Russell, *Speech recognition with dynamic bayesian networks*, Technical report, 1998.

[14] Christopher D. Manning Kristina Toutanova, Dan Klein and Yoram Singer, "Feature rich part-of-speech tagging with a cyclic dependency network," in *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Morristown, NJ, USA, 2003, p. 173–180.

[15] Gang Ji and Jeff Bilmes, "Backoff model training using partially observed data: Application to dialog act tagging," in *Proceedings of the Human Language Technology Conference of the NAACL*, New York City, USA, 2006, p. 280–287.

[16] Vladimir N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.

[17] Franz-Josef Och Oliver Bender, Klaus Macherey and Hermann Ney, "Comparison of alignment templates and maximum entropy models for natural language understanding," in *Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003, p. 11–18.

[18] Patrick Lehnen Stefan Hahn and Hermann Ney, "System combination for spoken language understanding," in *Interspeech*, Brisbane, Australia, 2008, p. 236–239.

[19] Christian Raymond Stefan Hahn, Patrick Lehnen and Hermann Ney, "A comparison of various methods for concept tagging for spoken language understanding," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

[20] J.G. Fiscus W.M. Fisher J.S. Garofolo B.S. Lund A. Martin y M.A. Przybocki Pallet, D.S., "The 1994 benchmark tests for the arpa spoken language program," in *Proceedings of ARPA Workshop on Spoken Language Technology*, 1995.

[21] J. Peckham, "A new generation of spoken dialogue systems: results and lessons from the SUNDIAL project," in *Proc. of 3rd Eurospeech*, Berlin (Germany), 1993, vol. 1, pp. 33–42.

[22] L. Lamel, S. Rosset, J. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts, "The LIMSI ARISE system," in *Speech Communication*, 2000, vol. 31 (2000), pp. 339–353.

[23] J.L. Gauvain, S. Bennacef, L. Devillers, and L. Lamel, "Spoken language system development for the Mask Kiosk," in *Proc. of IEEE Workshop on Automatic Speech Recognition*, Salt Lake City, USA, 1995.

[24] R. Pieraccini, E. Levin, and W. Eckert, "AMICA: The AT&T mixed initiative conversational architecture," in *Proc. of 5th Eurospeech*, Rhodes (Greece), 2000, pp. 1875–1878.

[25] A.L. Gorin, G. Riccardi, and J.H. Wright, "How may I help you?," in *Speech Communication*, 1997, pp. (23):113–127.

[26] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," in *In Proc. ICSLP'98*, Sydney (Australia), 1998, pp. 931–934.

[27] S. Seneff, R. Lau, and J.Polifroni, "Organization, communication, and control in the galaxy-II conversational system," in *In Proc. Eurospeech*, 1999.

[28] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," in *IEEE Transactions on Speech and Audio Processing*, 2000, pp. vol. 8 (1), 85–96.

[29] A. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, and A. Oh, "Creating natural dialogs in the Carnegie Mellon Communicator system," *Proceedings of Eurospeech*, vol. 1, no. 4, pp. 1531–1534, 1999.

[30] M. Denecke U. Meier M. Westphal y A. Waibel Geutner, P., "Conversational speech systems for on-board car navigation and assistance," in *Proceedings of the ICSLP*, Adelaide, Australia.

[31] S.K. Bennacef, H. Bonneau-Maynard, J.L. Gauvian, L. Lamel, and W. Minker, "A Spoken Language System For Information Retrieval," in *Proc. ICSLP-94*, 1994, pp. S22–8.1, 8.4.

[32] R. Billi and L.F. Lamel, "Railtel: Railway telephone services," in *Speech Communication*, 1997, vol. 23, pp. 63–82.

[33] James Allen, George Ferguson, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski Zollo, "Dialogue systems: From theory to practice in trains-96," in *Handbook of Natural Language Processing*, 2000, pp. 347–376.

[34] Biing-Hwang Juang Rabiner, Lawrence R., *Fundamentals of Speech Recognition*, 1993.

[35] S. Seneff, "TINA: A natural language system for spoken language applications," in *Computational Linguistics*, 1992, vol. 18(1), pp. 61–86.

[36] W. Ward and S. Issar, "Recent improvements in the CMU spoken language understanding system," in *Proc. of the ARPA Human Language Technology Workshop*, 1994, pp. 213–216.

[37] Fernando García, *Una aproximación estocástica a la comprensión del lenguaje*, Ph.D. thesis, DSIC - UPV, Valencia, España, 2003.

[38] E Segarra, E. Sanchis, F.García, and L.F. Hurtado, "Extracting semantic information through automatic learning," in *Pattern Recognition and Image Analysis. Proceedings of IX Spanish Symposium on Pattern Recognition and Image Analysis (SNRFAI'01)*, Benicàssim (Spain), 2001.

[39] R. Schwartz, S. Miller, D. Stallard, and J. Makhoul, "Language understanding using hidden understanding models," in *Proc. ICSLP'96*, Philadelphia (USA), Oct. 1996, pp. 997–1000.

[40] W. Minker, "Stochastic versus Rule-Based Speech Understanding for Information Retrieval," in *Speech Communication*, 1998, pp. 25(4):223–247.

[41] E. Levin and P. Pieraccini, "Concept-based spontaneous speech understanding system," in *Proc. of EuroSpeech'95*, 1995, pp. 555–558.

[42] F. y E. Huang Soong, "A tree-treliss based fast search for finding the n best sentence hypoteses in continuous speech recognition," in *Proceedings of ICASSP'91*, 1991, p. 537–540.

[43] Xavier Aubert and Hermann Ney, "Large vocabulary continuous speech recognition using word graphs," in *Proc. of ICASSP 95*, 1995.

[44] Wei Xu and Alex Rudnicky, "Can Artificial Neural Networks Learn Language Models?," in *Proc. ICSLP'00*, Beijing (China), 2000.

[45] S. McGlashan, D.C. Burnett, J. Carter, P. Danielsen, J. Ferrans, A. Hunt, B. Lucas, B. Porter, K. Rehor, and S. Tryphonas, "Voice Extensible Markup Language (VoiceXML) Version 2.0," in *Recomendación del W3C. www.w3.org/TR/voicexml20/* , 2004.

[46] M. McTear, "Modelling Spoken Dialogues with State Transition Diagrams: Experiences with the CSLU Toolkit," in *Proc. of 5th International Conference on Spoken Language Processing*, Sydney, (Australia), 1998.

[47] Lluís-F. Hurtado, Joaquin Planells, Encarna Segarra, Emilio Sanchis, and David Griol, "A stochastic finite-state transducer approach to spoken dialog management," in *Proc. of InterSpeech 2010*, Makuhari, Chiba, Japan, 2010, pp. 3002–3005.

[48] F. Neel y H. Bonneau-Maynard Bennacef, S., "An oral dialogue model based on speech acts categorization," in *ESCA Workshop of Spoken Dialog System*, 1995.

[49] A. Bonafonte, P. Aibar, E. Castell, E. Lleida, J.B. Mariño, E. Sanchís, and M. I. Torres, "Desarrollo de un sistema de diálogo oral en dominios restringidos," in *I Jornadas en Tecnología del Habla*, Sevilla (Spain), 2000.

[50] José-Miguel Benedí, Eduardo Lleida, Amparo Varona, María-José Castro, Isabel Galiano, Raquel Justo, Iñigo López de Letona, and Antonio Miguel, "Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA," in *Proceedings of LREC 2006*, Genoa (Italy), May 2006, pp. 1636–1639.

[51] J.M. Benedí, A. A. Varona, and E. Lleida, "DIHANA: Sistema de diálogo para el acceso a la información en habla espontánea en diferentes entornos," in *Actas de las III Jornadas en Tecnología del Habla*, Valencia (España), 2004, pp. 141–146.

[52] D. Griol, F. Torres, L.F Hurtado, S. Grau, E. Sanchis, and E. Segarra, "Development and evaluation of the dihana project dialog system," in *Proc. of InterSpeech'06 - ICSLP*, Pittsburgh, USA, 2006.