Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 31-38 recibido 27-03-2015 revisado 27-04-2015 aceptado 18-05-2015

Diseño y comparación de varias aproximaciones estadísticas a la comprensión del habla en dos tareas e idiomas distintos*

Design and comparison of several statistical approaches to Speech Understanding in two different tasks and languages

Fernando García, Marcos Calvo, Lluís-F. Hurtado, Emilio Sanchis, Encarna Segarra
Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{fgarcia,mcalvo,lhurtado,esanchis,esegarra}@dsic.upv.es

Resumen: En este artículo se presenta un estudio de diversas aproximaciones al problema de la comprensión del habla en dominios semánticos restringidos. Se proponen dos sistemas basados en modelos generativos y se comparan con un sistema basado en un método discriminativo. La experimentación se ha realizado sobre dos tareas diferentes, DIHANA y MEDIA, que a su vez están en dos idiomas diferentes. El uso de las dos tareas tiene interés no sólo por las diferencias en la forma de expresar los conceptos en los dos idiomas, sino también por las diferencias en la forma de representar la semántica. Los resultados muestran la capacidad de los modelos estadísticos aprendidos automáticamente para representar la semántica, incluso cuando se trata con voz, que introduce errores generados en el proceso de reconocimiento.
Palabras clave: comprensión del habla, modelos estocásticos, modelos generativos, modelos discriminativos

Abstract: In this paper, a study of different approaches to the problem of speech understanding in restricted semantic domains is presented. Two systems based on generative models are proposed and they are compared with a system based on discriminative methods. The experiments were conducted on two different tasks, DIHANA and MEDIA, which are in two different languages. The use of the two tasks is of interest not only because of the differences in how concepts are expressed in both languages, but also because of the differences in the way of representing the semantics. The results show the ability of automatically learned statistical models to represent the semantics, even when dealing with voice input, which introduces errors that are generated in the recognition process.
Keywords: spoken language understanding, stochastic models, generative models, discriminative models

1 Introducción

La componente de comprensión del habla tiene una importancia capital en muchos de los sistemas de interacción con los ordenadores, bien oral o escrita. Si bien está todavía muy lejos la posibilidad de disponer de componentes que proporcionen una interpretación semántica de un texto en un universo semántico no restringido, su uso para tareas acotadas semánticamente proporciona resultados razonables.

Uno de los ámbitos de aplicación de estos componentes son los sistemas de diálogo hablado para dominios restringidos. En un gran número de sistemas de este tipo se abordan tareas cuyo objetivo final es la obtención de una plantilla con las informaciones necesarias para realizar una consulta a un sistema de información. Esto se realiza a lo largo de diversos turnos, por lo que para cada turno es necesario obtener la información semántica proporcionada por el usuario, es decir, los datos concretos que se han proporcionado y también información sobre la intención subyacente en el turno.

* This work is partially supported by the Spanish MEC under contract TIN2014-54288-C4-3-R and FPU Grant AP2010-4193

ISSN 1135-5948

Entre las aproximaciones desarrolladas para la obtención de los componentes de comprensión del habla podemos hablar de las basadas en reglas que determinan los patrones sintácticos de ciertos significados (Ward y Issar, 1994; Seneff, 1992) y de las basadas en modelos estadísticos (Hahn et al., 2010; Raymond y Riccardi, 2007) que pueden ser estimados a partir de conjuntos de muestras etiquetadas semánticamente. Los modelos estadísticos presentan diversas ventajas. Una de ellas es que representan adecuadamente la secuencialidad de las frases y los conceptos. Disponen de mecanismos para ampliar la cobertura de forma que pueden tratar frases con errores (habituales cuando su entrada es la salida de un reconocedor). En contrapartida necesitan que los corpus de entrenamiento estén segmentados y etiquetados.

Algunos de estos modelos estadísticos son los llamados generativos basados en Modelos de Markov o Gramáticas estocásticas (Servan et al., 2010; Ortega et al., 2010; Hurtado et al., 2004; Esteve et al., 2003; He y Young, 2003; Segarra et al., 2002). También se han utilizado modelos discriminativos, como son clasificadores bayesianos, SVM o CRF (Dinarelli, Moschitti, y Riccardi, 2009; Lefèvre, 2007; Lafferty, McCallum, y Pereira, 2001). En este tipo de modelos uno de los principales problemas que se tiene que abordar es el de la segmentación de la frase de entrada, ya que el objetivo no es solamente obtener una o más clases asociadas a una frase sino también el segmento de texto que se corresponde con cada significado semántico encontrado, considerando el contexto de toda la frase.

En este artículo presentamos tres aproximaciones estadísticas al problema de la comprensión del lenguaje, explorando su comportamiento sobre diferentes corpus de diálogo. Se presentan dos aproximaciones que modelizan la semántica con autómatas finitos, desarrolladas anteriormente para comprensión multilingüe (García et al., 2012; Calvo et al., 2013), y una aproximación que utiliza CRF. Se ha hecho una experimentación en la que se han utilizado dos corpus con tareas e idiomas diferentes. Estas aproximaciones funcionan en tiempo real y se han implementado en un prototipo de comprensión de habla multilingüe (Laguna et al., 2014).

Los corpus utilizados son el corpus DIHANA, en castellano, una tarea de acceso a información sobre trenes, y el corpus MEDIA, en francés, una tarea de información y reserva turística. Estos son dos de los principales corpus etiquetados semánticamente que existen en la actualidad.

2 El corpus DIHANA

La tarea del corpus DIHANA (Benedí et al., 2006) consiste en un sistema de información sobre horarios, precios y servicios de trenes de larga distancia españoles en castellano. El corpus consiste en 900 diálogos que se adquirieron por 225 hablantes empleando la técnica del Mago de Oz, que permite que estén presentes la mayor parte de las características del habla espontánea en la adquisición de los diálogos.

El corpus DIHANA se ha transcrito y etiquetado manualmente a nivel de concepto, para ello se definieron 30 etiquetas semánticas. Estas etiquetas semánticas se pueden agrupar en: independientes de la tarea como "cortesía", "nada" ..., otras cuyo segmento asociado contiene un valor que es relevante para la comprensión p.e. "ciudad_origen", "ciudad_destino", "hora", "tipo_tren" ..., y otras que son relevantes para la tarea e identifican el tipo de concepto del cuál se está hablando: "<hora>", "<hora_salida>", "<hora_llegada>", "<tipo_tren>" ..., (estas últimas vienen parentizadas por las marcas < y > para distinguirlas de las anteriores). A continuación se muestra un ejemplo de representación semántica de la frase: *"me podría decir horarios de tren intercity de Zamora a Valladolid"*.

Ejemplo:
"Me podría decir horarios de tren intercity de Zamora a Valladolid por favor"

Palabras	Concepto
me podría decir	consulta
horarios de tren	<hora>
intercity	tipo_tren
de Zamora	ciudad_origen
a Valladolid	ciudad_destino
por favor	cortesía

Algunas características del corpus DIHANA etiquetado semánticamente se muestra en la Tabla 1.

3 El corpus MEDIA

La tarea del corpus MEDIA (Bonneau-Maynard et al., 2005) consiste en un sistema de información turística y reservas de

Número de turnos de usuario:	6.229
Total de palabras:	47.222
Talla del vocabulario:	811
Media de palabras por turno de usuario:	7,6
Número total de segmentos semánticos:	18.588
Media de palabras por segmento semántico:	2,5
Media de segmentos por turno de usuario:	3,0
Media de muestras por unidad semántica:	599,6

Tabla 1: Características del corpus DIHANA etiquetado semánticamente.

Número de turnos de usuario:	16.279
Total de palabras:	114.969
Talla del vocabulario:	2.357
Media de palabras por turno de usuario:	7,1
Número total de segmentos semánticos:	53.942
Media de palabras por segmento semántico:	2,1
Media de segmentos por turno de usuario:	3,3
Media de muestras por unidad semántica:	709,8

Tabla 2: Características del corpus MEDIA etiquetado semánticamente.

hoteles en francés. Al igual que DIHANA se adquirió empleando la técnica del Mago de Oz. Se definieron ocho categorías de escenarios con diferentes niveles de complejidad.

El corpus adquirido está compuesto por 1.257 diálogos de 250 hablantes y contiene aproximadamente 70 horas de diálogos. Cada hablante grabó cinco escenarios diferentes de reservas de hotel. El corpus francés MEDIA se ha transcrito y etiquetado manualmente a nivel de concepto. Para realizar el etiquetado se definieron más de 80 conceptos semánticos de los cuales solo 72 están presentes en el corpus.

Un ejemplo de etiquetado semántico para la frase: *"je souhaiterais réserver pour la deuxième semaine de janvier à Paris à côté de l' arc de triomphe"* se muestra a continuación.

Ejemplo:
"Je souhaiterais réserver pour la deuxième semaine de janvier à Paris à côté de l' arc de triomphe"

Palabras	Concepto
je souhaiterais réserver	command-tache
pour la deuxième	rang-temps
semaine	temps-unite
de janvier	temps-mois
à Paris	localisation-ville
à côté de	localisation-distanceRelative
l' arc de triomphe	localisation-lieuRelatif

Algunas características del corpus MEDIA etiquetado semánticamente se muestra en la Tabla 2.

4 El sistema de comprensión

El problema de la comprensión del habla puede abordarse como la búsqueda de la secuencia de conceptos que se corresponden con el significado de una frase de entrada. Cada concepto representa el significado de una secuencia de palabras (un segmento) de la frase. Por ejemplo, el concepto "consulta" del corpus DIHANA puede ser asociada con los segmentos "Me podría decir", "Por favor dígame", "Cuál es", etc. De esta forma, el sistema de comprensión es capaz de asociar a cada frase de entrada una secuencia semántica (secuencia de conceptos), así como los segmentos de palabras consecutivas asociados a los conceptos, es decir, la segmentación subyacente.

En la Figura 1 se presenta un esquema del proceso de comprensión, tanto de la fase de entrenamiento como de la fase de test. En la fase de entrenamiento se dispone del corpus de entrenamiento que deberá presentar las frases etiquetadas y segmentadas en términos de conceptos, como se explica a continuación. En la fase de test, dada una frase de entrada $W = w_1, w_2, \dots w_{\mathrm{N}}$, obtenida a partir de la salida de un reconocedor del habla o bien proporcionada directamente como texto, el módulo de comprensión obtiene la secuencia de pares (*segmento*,*concepto*).

Con el fin de aprender los modelos de comprensión, se dispone de un conjunto de secuencias de conceptos asociadas a las frases de entrada, así como la asociación de segmentos de palabras correspondientes a cada concepto. En otras palabras, sea $\mathscr{W}$ el vocabulario de la tarea, y sea $\mathscr{C}$ el alfabeto de conceptos; el conjunto de entrenamiento es un conjunto de pares (s, c) donde:

$$s = s_1 s_2 \dots s_{\mathrm{n}}\ ,\ \ s_{\mathrm{i}} = w_{\mathrm{i}_1} w_{\mathrm{i}_2} \dots w_{\mathrm{i}_{|s_{\mathrm{i}}|}},$$
$$w_{\mathrm{i}_{\mathrm{j}}} \in \mathscr{W}\ ,\ i = 1, \dots, n\ ,\ j = 1, \dots, |s_{\mathrm{i}}|;$$
$$c = c_1 c_2 \dots c_{\mathrm{n}}\ ,\ c_{\mathrm{k}} \in \mathscr{C}\ ,\ k = 1, \dots, n.$$

cada frase de entrada en $W \in \mathscr{W}^*$ tiene un par (s,c) asociado, donde s es una secuencia de segmentos de palabras y c es una secuencia de unidades semánticas. Un ejemplo de par de entrenamiento para el

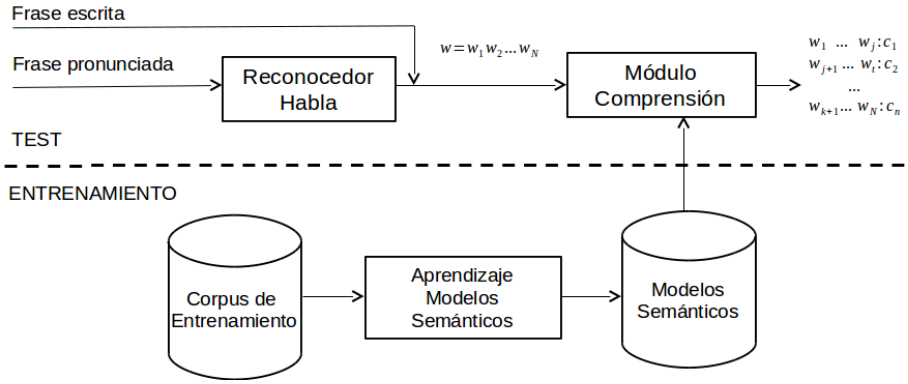

Figura 1: Esquema del proceso de comprensión

corpus DIHANA sería:

Frase de entrada W:

"me podría decir el que sale a las ocho qué tipo de tren es"

Par de entrada: $(s,c)=(s_1 s_2 s_3, c_1 c_2 c_3)$ donde:

s_1: me podría decir	c_1: *consulta*
s_2: el que sale a las ocho	c_2: *hora_salida*
s_3: qué tipo de tren es	c_3: $< tipo\_tren >$

La secuencia semántica s para el modelo semántico sería:

$$consulta\ hora\_salida < tipo\_tren >$$

Dado un conjunto de entrenamiento de este tipo, el problema de aprender el modelo de comprensión del lenguaje puede resolverse aplicando diferentes aproximaciones. En este trabajo presentamos tres aproximaciones estadísticas al problema de la comprensión del lenguaje: dos aproximaciones que modelizan la semántica con autómatas finitos y una aproximación que utiliza CRF.

Las dos aproximaciones basadas en autómatas finitos estiman dos tipos de modelos a partir de un conjunto de entrenamiento como el descrito anteriormente: un modelo semántico que representa el lenguaje de las concatenaciones de conceptos, y un modelo para cada concepto que representa el lenguaje de secuencias de palabras asociadas a ese concepto.

En el caso de la aproximación basada en CRF se establece para cada una de las palabras su concepto asociado empleando la notación IOB. Para ello se etiqueta el principio de un segmento con el prefijo "$B\_$" sobre el concepto, y para el resto de palabras del segmento se utiliza el segmento "$I\_$". Este etiquetado permite proporcionar información de los segmentos asociados a cada concepto al modelo CRF.

Palabra	Etiqueta
me	$B\_consulta$
podría	$I\_consulta$
decir	$I\_consulta$
el	B_hora_salida
que	I_hora_salida
sale	I_hora_salida
a	I_hora_salida
las	I_hora_salida
ocho	I_hora_salida
qué	$B\_ < tipo\_tren >$
tipo	$I\_ < tipo\_tren >$
de	$I\_ < tipo\_tren >$
tren	$I\_ < tipo\_tren >$
es	$I\_ < tipo\_tren >$

4.1 La aproximación 2-niveles

En esta aproximación, a partir de un conjunto de pares de entrenamiento (s,c) se estiman dos tipos de modelos (autómatas finitos): un modelo para el lenguaje de las concatenaciones de conceptos que llamaremos el lenguaje semántico $L_c \subseteq \mathscr{C}^*$, y un conjunto de modelos, uno por concepto $c_i \in \mathscr{C}$. El autómata finito A_c para el lenguaje semántico L_c se estima a partir de las cadenas semánticas $c \in \mathscr{C}^*$ del conjunto de entrenamiento. Un autómata finito A_{c_i} es estimado para cada concepto $c_i \in \mathscr{C}$ a partir del conjunto de segmentos s_i obtenido del conjunto de entrenamiento asociado a cada una de estas unidades semánticas c_i. Estas estimaciones son llevadas a cabo a través de técnicas de aprendizaje automático.

El modelo de comprensión A es un autómata finito que se obtiene con la susti-

tución en el autómata A_c de cada estado que representa un concepto $c_i \in \mathscr{C}$ por el autómata A_{c_i} correspondiente.

Una de las ventajas de esta aproximación, es que podemos escoger la técnica de aprendizaje más adecuada para la estimación de cada modelo: el modelo para lenguaje semántico y los modelos para los conceptos. La única restricción es que la representación de estos modelos debe ser dada en forma de un autómata finito. En este trabajo se han utilizado modelos de n-gramas.

Dada una frase de entrada, su análisis en base al algoritmo de Viterbi en el autómata finito A devuelve el mejor camino. Este camino proporciona, no solo la secuencia de conceptos sino también al segmentación de la frase analizada.

4.2 Aproximación basada en grafos

Esta aproximación se basa en la idea de construir un *grafo de conceptos* de forma que las distintas posibles interpretaciones semánticas de la frase de entrada (de test) estén codificadas como caminos en dicho grafo. En él cada arco estará etiquetado con una secuencia de palabras y el concepto que representa. La construcción de este grafo se lleva a cabo por medio de un algoritmo de programación dinámica que utiliza la información de modelos que representan las distintas formas de expresar léxicamente cada concepto. Estos modelos junto con un modelo semántico que representa las posibles concatenaciones de conceptos, se estiman a partir del conjunto de entrenamiento descrito al inicio de la sección. Al igual que en la aproximación 2-niveles, estos modelos pueden entrenarse siguiendo distintas técnicas de aprendizaje automático, en nuestro caso n-gramas.

Una vez construido el grafo de conceptos se realiza la búsqueda del mejor camino combinando la información de este grafo con la del modelo semántico. Esta búsqueda del mejor camino en el grafo de conceptos da como resultado la mejor secuencia de conceptos junto con la segmentación de la frase de entrada. iberspeechdemo2014 El grafo de conceptos tendrá $N+1$ nodos, donde N es el número de palabras de la cadena de entrada W. Para construir el grafo de conceptos se consideran todas las subcadenas $w_i...w_j$ contenidas en W, y para cada concepto c_k se calcula la probabilidad t que el modelo asociado a c_k asigna a ese segmento. Si la probabilidad t es no nula, entonces se crea un arco entre los nodos i y $j+1$ etiquetado con el par $(w_i...w_j,\ c_k)$ y con peso t^α. Este método se muestra en el Algoritmo 1. La función de α es escalar las probabilidades de los modelos que intervienen en la construcción del grafo de conceptos para su posterior combinación con la del modelo semántico durante la búsqueda del mejor camino en el grafo.

Algoritmo 1 Método para la construcción de un grafo de conceptos.

Entrada: Frase de entrada $W = w_1 w_2 ... w_N$, conjunto de modelos $M = \{M(c_1) \dots M(c_{|\mathscr{C}|})\}$ que permitan estimar las probabilidades $p(w_i...w_j|c_k)$, factor de escala α
Salida: Grafo de conceptos GC

```
Crear N + 1 nodos para GC y numerarlos comenzando en 1
Para i = 1 hasta N − 1 hacer
    Para j = i + 1 hasta N hacer
        Para k = 1 hasta |C| hacer
            t = p(w_i...w_j|c_k)
            Si t > 0 entonces
                Añadir a GC un arco con origen el nodo i y destino el j+1 etiquetado con el par (w_i...w_j, c_k) y con peso t^α
            Fin Si
        Fin Para
    Fin Para
Fin Para
Devolver GC
```

4.3 CRF

Los "Conditional Random Fields" (Lafferty, McCallum, y Pereira, 2001) son modelos "log-lineal" en los que se realiza una normalización a nivel de toda la frase. Los CRF reúnen algunas de las ventajas de los modelos generativos y discriminativos. Con ellos se pueden tener en cuenta muchas características de las entradas que se han aprendido de forma discriminativa, pero también tienen en cuenta las decisiones previas para escoger la mejor etiqueta semántica en cada momento. En estos modelos se representa la probabilidad condicional de una secuencia de etiquetas de conceptos $c_1 \dots c_N$ dada una secuencia de palabras $w_1 \dots w_N$ como:

$$p(c_1^n|w_1^n) = \frac{1}{Z} \quad \prod_{m=1}^{N} \lambda_m \cdot h_m(c_{n-1}, c_n, w_{n-k}^{n+k}) \Bigg)$$

donde λ_m es el vector de parámetros que se aprende a partir de un corpus etiquetado, y $h_m(c_{n-1}, c_n, w_{n-k}^{n+k})$ representa las dependencias entre las entradas (palabras u otras características que pueden aparecer en una ventana alrededor de la palabra a etiquetar) y los conceptos de salida. Por otra parte, el factor de normalización Z viene dado por

$$Z = \sum_{\tilde{c}_1^N} \prod_{n=1}^{N} (\tilde{c}_{n-1}, \tilde{c}_n, w_{n-k}^{n+k})$$

donde $\tilde{c}_{n-1}$ y $\tilde{c}_n$ son los conceptos predichos por las palabras previa y actual.

En nuestro caso hemos usado un conjunto de características básicas como son la información léxica y el etiquetado semántico. Para ello se ha definido una ventana $k = 2$ que incorpora las dos palabras (y su concepto asociado) previas y posteriores.

5 Experimentos

Para evaluar los diferentes sistemas de comprensión expuestos en este trabajo se ha realizado una serie de experimentos con el corpus en castellano DIHANA y el corpus en francés MEDIA. Para ello se definieron dos medidas: el CA (Concept Accuracy), que es el porcentaje de conceptos correctos; y el CSS (Correct Semantic Sequence) que es el porcentaje de secuencias completas semánticas correctas. Además se han incluido el porcentaje de sustituciones (S), borrados (B) e inserciones (I) entre la secuencia semántica correcta y la hipótesis dada por cada sistema para cada una de las frases. En los dos corpus se han evaluado las transcripciones correctas de los diálogos y en el caso del corpus DIHANA también las proporcionadas por un reconocedor de habla basado en HTK.

En las tablas se presentan los resultados para los tres sistemas: 2-niveles, Grafos y CRF. En todas las tablas de resultados se ha añadido la fila Oráculo en la que se elige para cada frase de test la interpretación semántica que más se ajusta a la de referencia, considerando las salidas de los tres sistemas. Estos valores constituyen una cota superior de los resultados alcanzables usando los tres sistemas a la vez.

Los 6.280 turnos del corpus de DIHANA se dividieron en un conjunto de entrenamiento de 4.887 turnos, un conjunto de desarrollo de 340 turnos y un conjunto de test de 1000 turnos. Los resultados obtenidos para las transcripciones correctas se muestran en la Tabla 3.

Sistema	**S**	**B**	**I**	**CSS**	**CA**
2-niveles	4,6	2,1	5,3	76,0	87,9
Grafos	2,1	1,5	1,0	88,2	95,4
CRF	3,9	3,8	2,0	79,6	90,4
Oráculo	1,5	1,2	0,7	91,5	96,6

Tabla 3: Resultados con DIHANA empleando como test las transcripciones correctas.

Los resultados obtenidos para la salida del reconocedor HTK con el corpus DIHANA se muestran en la Tabla 4. Como puede verse los resultados sobre las transcripciones correctas son muy altos, principalmente en el caso de los Grafos. Sin embargo, cuando se utiliza la salida del reconocedor, lógicamente disminuyen aunque en ese caso son los CRFs los que mejor resultado proporcionan. Esto puede deberse a que el método de aprendizaje de los modelos del sistema de Grafos genera un modelo muy adaptado a las muestras de entrenamiento y desarrollo. De este modo si hay una fuerte correlación entre el conjunto de entrenamiento y el de test el sistema funciona muy bien, pero conforme las estructuras sintácticas de las muestras de test se alejan de las de entrenamiento (como es el caso de la salida del reconocedor, y del corpus MEDIA que se verá más adelante) bajan las prestaciones de este sistema en comparación al resto.

Sistema	**S**	**B**	**I**	**CSS**	**CA**
2-niveles	8,0	2,9	11,3	64,0	77,9
Grafos	7,0	2,4	11,1	67,0	79,6
CRF	6,6	4,9	5,3	68,9	83,2
Oráculo	5,6	2,6	5,5	72,9	85,6

Tabla 4: Resultados con DIHANA empleando como test la salida del reconocedor HTK.

A la hora de mostrar los resultados de la segmentación y etiquetado semántico del corpus muchas veces no se evalúan etiquetas que no tienen significado semántico para la tarea como "*nada*", "*cortesía*", etc. Los resultados obtenidos para las transcripciones correctas y la salida del reconocedor HTK sin evaluar este tipo de etiquetas se muestran en la Tabla 5 y Tabla 6.

Como puede verse se mantiene la misma tendencia que en las tablas anteriores, aunque los resultados mejoran, ya que se eliminan los errores producidos por esas etiquetas sin significado semántico.

Sistema	S	B	I	CSS	CA
2-niveles	1,0	2,8	6,1	81,2	90,0
Grafos	0,2	1,6	1,7	91,5	96,5
CRF	1,3	3,3	3,5	83,5	92,0
Oráculo	0,2	0,9	1,4	94,3	97,5

Tabla 5: Resultados con las transcripciones correctas de DIHANA sin evaluar etiquetas sin significado semántico.

Sistema	S	B	I	CSS	CA
2-niveles	4,6	3,4	11,8	68,2	80,2
Grafos	4,3	2,8	11,6	70,2	81,3
CRF	4.2	4.3	7.2	72,2	84,3
Oráculo	3,6	2,5	7,2	75,9	86,7

Tabla 6: Resultados con DIHANA empleando como test la salida del reconocedor HTK sin evaluar etiquetas sin significado semántico.

Los 16.279 turnos de usuario del corpus MEDIA se dividieron en, 12.000 turnos para entrenamiento, 1.279 turnos para desarrollo y por último 3.000 turnos para test. En la Tabla 7 se muestran los resultados obtenidos para las transcripciones correctas del MEDIA.

Sistema	S	B	I	CSS	CA
2-niveles	5,7	7,8	6,9	69,7	79,6
Grafos	4,8	5,0	7,3	73,1	82,9
CRF	2,5	7,7	3,1	76,9	86,6
Oráculo	8,1	9,8	5,6	82,0	90,8

Tabla 7: Resultados con MEDIA empleando como test las transcripciones correctas.

Los resultados obtenidos para las transcripciones correctas sin evaluar etiquetas sin significado semántico como la etiqueta "*null*" se muestran en la Tabla 8.

En general los resultado para el corpus MEDIA son peores que los del corpus DIHANA. Esto puede deberse a que es un corpus con muchas más etiquetas semánticas, además de que algunas de ellas tienen pocas muestras en el corpus de entrenamiento. Por otra parte, ciertas etiquetas son iguales en cuanto a su realización léxica y sólo son distinguibles con el contexto semántico, como números en las etiquetas "*nombre_chambre*" (número de habitaciones), "*nombre_reservation*" (número de reservas) o "*nombre_hotel*" (número de hoteles). Es por tanto una tarea más difícil, aunque los resultados son adecuados a la tarea.

No existen otros trabajos comparables sobre el corpus DIHANA. Sin embargo si que

Sistema	S	B	I	CSS	CA
2-niveles	5.1	7.4	6.9	73,9	80,6
Grafos	4.2	5.6	6.1	77,7	84,1
CRF	2,8	7,8	2,0	80,9	87,3
Oráculo	2,3	4,9	1,8	85,4	91,0

Tabla 8: Resultados con MEDIA empleando como test las transcripciones correctas sin evaluar etiquetas sin significado semántico.

hay sobre el corpus MEDIA (Hahn et al., 2010) donde los mejores resultados reportados hasta el momento son del 89,4 % de CA, lo cual indica que nuestros resultados son competitivos.

6 Conclusiones

Hemos presentado en este artículo tres propuestas para la comprensión del habla, basadas todas ellas en el aprendizaje automático de modelos estadísticos. Los resultados experimentales muestran que este tipo de modelos es adecuado para las tareas de semántica restringida. Aunque existe un cierto deterioro en los resultados cuando se considera la salida de un reconocedor de voz, la capacidad de generalización y suavizado intrínseca a los modelos estadísticos permite mantener un buen resultado de comprensión.

Se ha comprobado que para estas dos tareas, DIHANA y MEDIA, los sistemas que devuelven mejores resultados en general son los basados en el modelo discriminativo CRF. En una comparación entre los dos métodos basados en modelos de autómatas finitos, los resultados muestran que el sistema de grafos funciona mejor que el de 2-niveles. Posiblemente esto ocurre porque en el sistema de grafos se realizan dos etapas. En la primera de seleccionan determinadas estructuras sintácticas asociadas a conceptos y en la segunda se elige la mejor concatenación de estas estructuras. Este proceso es más selectivo que la búsqueda del mejor camino en el modelo integrado del sistema de 2-niveles, que incorpora los autómatas de cada concepto al autómata que representa las concatenaciones de conceptos, y constituye un espacio de búsqueda mucho mayor. Esto supone un aumento de la cobertura del modelo en relación al modelo de grafos, pero puede introducir más confusión.

Bibliografía

Benedí, J.M., E. Lleida, A. Varona, M.J. Castro, I. Galiano, R. Justo, I. López de Letona, y A. Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. En *LREC*, páginas 1636–1639.

Bonneau-Maynard, H., S. Rosset, C. Ayache, A. Kuhn, y D. Mostefa. 2005. Semantic annotation of the French MEDIA dialog corpus. En *Proc. of InterSpeech 2005*, páginas 3457–3460, Portugal.

Calvo, M., F. García, L.-F. Hurtado, S. Jiménez, y E. Sanchis. 2013. Exploiting multiple hypotheses for multilingual spoken language understanding. En *Proc. of the CoNLL-2013*, páginas 193–201.

Dinarelli, M., A. Moschitti, y G. Riccardi. 2009. Concept Segmentation And Labeling For Conversational Speech. En *Interspeech*, Brighton.

Esteve, Y., C. Raymond, F. Bechet, y R. De Mori. 2003. Conceptual Decoding for Spoken Dialog systems. En *Proc. of EuroSpeech'03*, páginas 617–620.

García, F., L.-F. Hurtado, E. Segarra, E. Sanchis, y G. Riccardi. 2012. Combining multiple translation systems for Spoken Language Understanding portability. En *Proc. of IEEE Workshop on Spoken Language Technology (SLT 2012)*, páginas 282–289, Miami.

Hahn, S., M. Dinarelli, C. Raymond, F. Lefèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, y G. Riccardi. 2010. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 6(99):1569–1583.

He, Y. y S. Young. 2003. A data-driven spoken language understanding system. En *Proc. of ASRU'03*, páginas 583–588.

Hurtado, L., E. Segarra, F. García, y E. Sanchis. 2004. Language understanding using n-multigram models. En *Advances in Natural Language Processing, Proceedings of 4th International Conference EsTAL*, volumen 3230 de *Lecture Notes in Computer Science*. Springer-Verlag, páginas 207–219.

Lafferty, J., A. McCallum, y F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. En *International Conference on Machine Learning*, páginas 282–289. Citeseer.

Laguna, S., M. Giménez, M. Calvo, F. García, E. Segarra, E. Sanchis, y L.-F. Hurtado. 2014. A Multilingual Spoken Language Understanding System. En *Proc. of the Iberspeech*, páginas 348–353, Las Palmas de Gran Canaria.

Lefèvre, F. 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. En *ICASSP 2007*, volumen 4, páginas 13–16.

Ortega, L., I. Galiano, L.-F. Hurtado, E. Sanchis, y E. Segarra. 2010. A statistical segment-based approach for spoken language understanding. En *Proc. of InterSpeech 2010*, páginas 1836–1839, Makuhari, Chiba, Japan.

Raymond, C. y G. Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. *Proc. of Interspeech 2007*, páginas 1605–1608.

Segarra, E., E. Sanchis, M. Galiano, F. García, y L. Hurtado. 2002. Extracting Semantic Information Through Automatic Learning Techniques. *IJPRAI*, 16(3):301–307.

Seneff, S. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 1(18):61–86.

Servan, C., N. Camelin, C. Raymond, F. Béchet, y R. De Mori. 2010. On the use of Machine Translation for Spoken Language Understanding portability. En *Procs. of ICASSP'10*, páginas 5330–5333.

Ward, W. y S. Issar. 1994. Recent improvements in the CMU spoken language understanding system. En *Proc. of the ARPA Human Language Technology Workshop*, páginas 213–216.