Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 105-112 recibido 30-01-12 revisado 09-03-12 aceptado 18-03-12

Un algoritmo para la comprensión automática del habla sobre grafos de palabras*

An algorithm for automatic speech understanding over word graphs

Marcos Calvo, Jon Ander Gómez, Emilio Sanchis, Lluís-F. Hurtado
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n - 46022 València
{mcalvo, jon, esanchis, lhurtado}@dsic.upv.es

Resumen: En este trabajo se propone un algoritmo para la comprensión automática del habla que toma como entrada un grafo de palabras. Este grafo es procesado en primer lugar mediante un algoritmo de programación dinámica, obteniendo como resultado un segundo grafo enriquecido con información semántica. El cálculo del mejor camino sobre este segundo grafo permite obtener la secuencia de conceptos más verosímil de acuerdo con la evidencia acústica reflejada en el grafo de palabras. También como resultado de la decodificación semántica se obtiene la secuencia de palabras asociada a dicha secuencia de conceptos, así como la segmentación semántica de la secuencia de palabras.
Palabras clave: Comprensión automática del habla, grafos de palabras, programación dinámica, algoritmos.

Abstract: In this work we propose an algorithm for automatic speech understanding that takes a word graph as its input. First, this word graph is processed by means of a dynamic programming algorithm which gives as a result a second graph that includes semantic information. Computing the best path over this second graph allows us to obtain the most likely concept sequence, given the acoustic evidence reflected on the input word graph. As a result of the semantic decoding, the word sequence attached to the concept sequence as well as its semantic segmentation are also obtained.
Keywords: Spoken language understanding, word graphs, dynamic programming, algorithms.

1. *Introducción*

La comprensión automática del habla es el proceso por el que, dada una pronunciación emitida por un locutor, se extrae una interpretación semántica de la información contenida en ésta basada en un conjunto de conceptos (o etiquetas semánticas) definido a priori. El ámbito donde los sistemas de comprensión del habla tienen mayor aplicación práctica es el de los sistemas de diálogo hablado, en los cuales es crucial que el sistema pueda extraer la información asociada a la pronunciación de entrada (la "comprenda") para devolver una respuesta coherente con lo que ha dicho el usuario. En la estructura típica de un sistema de diálogo hablado, el módulo de comprensión toma como entrada la salida del módulo de reconocimiento de voz (ASR) y pasa su salida al módulo gestor del diálogo, el cual utilizará esta información para decidir la siguiente acción a realizar. Por tanto, teniendo en cuenta que al módulo de comprensión le llega una estructura de datos en la que está representada la información que ha extraído el ASR, la comprensión del habla consiste de dos subtareas, que son:

1. La identificación de la secuencia de conceptos y su asignación a secuencias de palabras.

2. La extracción de la información relativa a estos conceptos codificada en las secuencias de palabras asignadas y la construcción de estructuras de datos que representen dicha información.

* Este trabajo ha sido subvencionado por el Ministerio de Economía y Competitividad en el marco del proyecto TIN2011-28169-C05-01 y por el Vicerrectorat d'Investigació, Desenvolupament i Innovació de la Universitat Politècnica de València con proyecto 20110897.

ISSN 1135-5948

Atendiendo a la entrada que toma el sistema de comprensión, podemos distinguir entre los que trabajan con la mejor transcripción proporcionada por el ASR (1-*best*) y los que toman una representación de las n mejores decodificaciones (Hakkani-Tür et al., 2006; Tur et al., 2002). Sin embargo, el hecho de emplear algún tipo de grafo de palabras como entrada al módulo de comprensión hace esta tarea más complicada, ya que hace el espacio de búsqueda de la decodificación semántica correcta todavía más grande. La ventaja, por el contrario, es que entre las posibles frases representadas en el grafo podría estar la correcta, lo que podría ayudar a recuperar errores cometidos durante el reconocimiento.

En los últimos años se han propuesto varias aproximaciones para la comprensión automática del habla (Hahn et al., 2010; Raymond y Riccardi, 2007) basadas muchas de ellas en métodos ya aplicados con éxito para otras tareas relacionadas. Por ejemplo, entre los modelos log-lineales, caben destacar los *maximum entropy Markov models* (McCallum, Freitag, y Pereira, 2000) y los *conditional random fields* (CRF) (Lafferty, McCallum, y Pereira, 2001). De hecho, en la experimentación presentada en (Hahn et al., 2010) los mejores resultados se obtienen utilizando CRFs. Otra aproximación es la que utiliza modelos similares a los típicos de la Traducción Automática (Macherey, Och, y Ney, 2001). También las Redes Bayesianas Dinámicas (*Dynamic Bayesian Networks*), que fueron aplicadas con éxito a tareas relacionadas como el etiquetado de actos de diálogo (Ji y Bilmes, 2006), pueden ser aplicadas a la comprensión automática del habla, tal y como puede verse en (Hahn et al., 2010; Lefèvre, 2006; Lefèvre, 2007).

Un último enfoque es el basado en transductores estocásticos (Hahn et al., 2010), donde el objetivo es maximizar la probabilidad conjunta de la secuencia de palabras y la secuencia de conceptos. Esta maximización se lleva a cabo efectuando la composición de un cierto número de transductores, los cuales se corresponden con diversos niveles de conocimiento acústico, léxico y semántico. Por último se efectúa una búsqueda al estilo de Viterbi sobre el transductor resultante. Una forma fácil de llevar a cabo esta composición de transductores es utilizando la biblioteca AT&T FSM/GRM (Mohri y Pereira Michael, 2002). Como se apunta en (Hahn et al., 2010), para evitar la complejidad introducida por el tratamiento de una representación de las n-*best* decodificaciones proporcionadas por el ASR, suele fijarse la secuencia de palabras a la mejor transcripción proporcionada por éste. Sin embargo, el modelo estadístico general no asume conocida a priori la secuencia de palabras a analizar.

En este trabajo nuestro objetivo será abordar el problema de la decodificación semántica tomando como entrada una representación en forma de grafo de palabras de la salida del módulo reconocedor utilizando el paradigma basado en transductores. Para ello, presentamos un algoritmo que efectúa la composición de los transductores a la vez que se lleva a cabo la decodificación, la cual a su vez admite la utilización de diversos mecanismos de poda. La única restricción que deberá cumplir el grafo de palabras que se tome como entrada es que posea ciertas características en su topología.

El grafo de palabras de entrada lo convertiremos en primer lugar en otro grafo que asocie segmentos de palabras del grafo original a unidades semánticas. Esta asociación se lleva a cabo mediante una modelización de los conceptos en términos de secuencias de palabras, como puede ser el caso de modelos de lenguaje locales a cada uno de los conceptos. Al grafo resultado de la primera etapa lo denominaremos grafo de conceptos, dado que ya posee información semántica. A partir de este grafo puede obtenerse la mejor secuencia de conceptos, utilizando un modelo de lenguaje cuyas unidades básicas serán las propias unidades semánticas.

Por tanto, el resto de este documento se estructura de la siguiente manera. En la sección 2 se expone brevemente el paradigma de la comprensión del habla basado en transductores. Seguidamente, presentamos con detalle el algoritmo que proponemos para la decodificación semántica tomando como entrada un grafo de palabras. En la sección 4 exponemos la experimentación que se ha llevado a cabo para comprobar la potencia del algoritmo. Finalmente, en la sección 5 se presentan las conclusiones a las que se ha llegado.

2. El paradigma de la comprensión del habla basado en transductores

El paradigma de comprensión del habla basado en transductores (Hahn et al., 2010; Raymond y Riccardi, 2007) se fundamenta en la utilización de este formalismo para calcular la secuencia de conceptos $\hat{C}$ tal que cumple la ecuación 1, en la que A representa la pronunciación de entrada al sistema.

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, p(C|A) \tag{1}$$

Introduciendo como una variable aleatoria más la secuencia de palabras W y aplicando la regla de Bayes tenemos que:

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, \frac{\sum_W p(A|W,C) \cdot p(W,C)}{p(A)} \tag{2}$$

Dado que no depende del término de la maximización, podemos eliminar el denominador de la fracción de la ecuación 2. Además, es posible efectuar la consideración usual de que la suma para toda secuencia de palabras W queda convenientemente aproximada por su máximo. Una última suposición razonable es considerar que la acústica sí depende de la palabra, pero no de la categoría semántica a la que ésta pertenezca. Estas simplificaciones nos llevan a la ecuación 3.

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, \underset{W}{\operatorname{máx}}\, p(A|W) \cdot p(W,C) \tag{3}$$

Para calcular la mejor secuencia de conceptos expresada de la forma de la ecuación 3 este paradigma propone buscar el mejor camino en un transductor λ_{SLU} resultado de componer 4 transductores:

$$\lambda_{SLU} = \lambda_G \circ \lambda_{gen} \circ \lambda_{W2C} \circ \lambda_{SLM} \tag{4}$$

Donde

- λ_G es una representación de la pronunciación de entrada en forma de máquina de estados finitos en la que están especificadas las probabilidades acústicas $p(A|W)$ calculadas por el módulo reconocedor.
- λ_{gen} sustituye algunas de las palabras por sus categorías léxicas (por ejemplo ciudades, horas o números). Representa el conocimiento que se tiene a priori de la tarea a abordar y permite generalizar a partir de los datos de entrenamiento.
- λ_{W2C} convierte secuencias de palabras a conceptos.
- λ_{SLM} codifica un modelo de lenguaje de categorías semánticas, el cual nos permite finalmente estimar la probabilidad conjunta $p(W,C)$.

La búsqueda del mejor camino en el transductor permite obtener la secuencia de conceptos que maximiza la ecuación 3, ya que la probabilidad $p(A|W)$ ya viene codificada en λ_G y la composición de λ_{W2C} y λ_{SLM} nos permite calcular $p(W,C)$.

3. Presentación del algoritmo

La aproximación que aquí se presenta está basada en la idea de la aplicación de transformaciones sucesivas que propone el paradigma basado en transductores para obtener la secuencia de conceptos que maximiza la ecuación 3. La entrada para nuestro método será un grafo de palabras, que deberá ser suministrado por el módulo de reconocimiento. A partir de él se obtiene un grafo de conceptos, en el que se asocia información semántica a secuencias de palabras. Decodificando este segundo grafo se obtiene como salida la mejor secuencia de conceptos $\hat{C}$. Como consecuencia de la búsqueda de $\hat{C}$ también se consigue la secuencia de palabras $\tilde{W}$, la cual debe interpretarse como la secuencia de palabras que nos permite encontrar la secuencia de conceptos de mayor probabilidad. Hay que tener en cuenta que esta secuencia de palabras no tiene por qué coincidir con la decodificación 1-*best* que proporcionaría un ASR.

A continuación se expondrá en primer lugar las características que deberán tener los grafos de palabras para poderse utilizar en el sistema que aquí se propone, junto con la explicación del significado de sus nodos y arcos. Seguidamente detallaremos el procedimiento de construcción del grafo de conceptos. Por último, explicaremos cómo obtener la mejor secuencia de conceptos a partir del grafo que se acaba de crear.

3.1. Topología y semántica del grafo de entrada

Los grafos de entrada al algoritmo tendrán las siguientes características:

- Sus nodos representan instantes de tiempo y deben estar etiquetados con ellos.
- No deben existir nodos aislados, es decir, ningún nodo tendrá grados de entrada y salida iguales a 0.
- Dados dos nodos con identificadores i y j con $i < j - 1$, existirá un arco del nodo i al nodo j etiquetado con la palabra w si el sistema de reconocimiento ha detectado que dicha palabra ha podido ser pronunciada comenzando en el instante de tiempo i y terminando en el $j - 1$. Además, el peso asignado a estos arcos se corresponderá con el *score* acústico que el ASR haya dado a esta ocurrencia de w.
- Introducimos además un suavizado consistente en permitir la existencia de una λ-transición (debidamente penalizada) entre cada par de nodos consecutivos.

Esta definición del grafo de entrada al algoritmo de decodificación semántica permite modelar la distribución de probabilidad $p(A|w)$, siendo A el conjunto de *frames* acústicos comprendidos entre los instantes de tiempo inicial y final del arco y w la palabra asociada a éste. Esta probabilidad es la que en la ecuación 4 viene dada por el modelo λ_G.

3.2. Construcción del grafo de conceptos

El grafo de conceptos que obtendremos a partir del grafo de palabras tiene las siguientes características:

- El conjunto de nodos es el mismo que el del grafo de palabras de entrada y mantienen sus identificadores.
- Al igual que en el grafo de palabras, desde un nodo sólo podrán partir arcos que alcancen nodos posteriores en el tiempo.
- Cada arco entre dos nodos i y j estará etiquetado con un par (W, c), donde W es una secuencia de palabras y c es el concepto que éstas representan. El peso asociado a cada arco será $\max_W(p(A_i^j|W) \cdot p(W|c))$, donde A_i^j es el fragmento de audio correspondiente al intervalo temporal $[i, j[$. La secuencia de palabras W que deberá tener asignado el par será la que maximice dicha probabilidad para el concepto c.
- Dados dos nodos del grafo de segmentos i y j con $i < j$, sólo se permite que exista un arco etiquetado con el concepto c cuyo origen sea i y su destino j.

En esta especificación del grafo de conceptos aparece la distribución de probabilidad $p(W|c)$, cuya estimación puede efectuarse por medio de diversos métodos. Una opción, que es la que hemos utilizado en la experimentación, es la estimación de un modelo de lenguaje de n-gramas local a cada concepto. Para que esto sea posible se necesita que las frases de entrenamiento estén segmentadas en fragmentos que denoten conceptos y se haya adjuntado a éstos la etiqueta del concepto correspondiente.

El grafo de conceptos que pretendemos obtener puede generarse mediante un algoritmo de programación dinámica que encuentre, para cada concepto c y cada par de nodos i y j con $i < j$, el mejor camino en el grafo de palabras de entrada con origen en i y destino en j. En este caso, con mejor camino en el grafo entendemos el camino que maximiza la probabilidad resultado de combinar las probabilidades acústicas $p(A_i^j|w)$ expresadas en el grafo de palabras y las probabilidades $p(W|c)$ del modelo asociado al concepto c, donde $W = w_1 \dots w_n$ es la secuencia de palabras resultante de concatenar las etiquetas w_k de los arcos del grafo de palabras que forman el camino. Cada uno de estos "mejores caminos" se convertirá en un arco en el grafo de conceptos, etiquetado con el par (W, c) y con un peso asociado igual a la probabilidad $p(A_i^j|W) \cdot p(W|c)$.

El algoritmo 1 muestra el método de generación del grafo de conceptos. En él denotamos un camino h en el grafo de palabras por la 4-tupla $h = (inicio, fin, secuencia, peso)$, donde *inicio* es el nodo origen del camino, fin su nodo final, *secuencia* la secuencia de palabras asociada y *peso* el coste del camino. Dado un camino h accederemos a su información mediante las funciones $b(h)$, $e(h)$, $W(h)$ y $s(h)$, las cuales devolverán respectivamente los atributos *inicio*, fin, *secuencia* y *peso* de h. Además, en dicho algoritmo aparece la función "iguales", la cual determinará, dados

Algoritmo 1 Método para la construcción del grafo de conceptos

Entrada: Grafo de palabras GP, conjunto de modelos $M = \{M(c_1) \dots M(c_k)\}$ que permitan estimar las probabilidades $p(W|c_i)$.

Salida: Grafo de conceptos GC.

1: Nodos(GC) = Nodos(GP)
2: Crear para todo nodo $n \in GP$ y para cada concepto c una lista $L(n, c)$ donde guardaremos los caminos h que lleguen a n por medio del concepto c.
3: **Para todo** nodo $n \in GP$ tomados en orden temporal **hacer**
4: **Para todo** concepto c **hacer**
5: **Para todo** nodo $n' < n$ **hacer**
6: $h = argmax_{h' \in L(n,c)}\{s(h') \mid b(h') = n'\}$
7: Añadir a GC un arco con origen n' y destino n etiquetado con el par $(W(h), c)$ y con un peso igual a $s(h)$.
8: **Fin Para**
9: $L(n, c) = L(n, c) \cup \{(n, n, \lambda, 1)\}$
10: **Para todo** camino $h \in L(n, c)$ **hacer**
11: **Para todo** arco $a \in$ Arcos(GP) $\mid$ origen(a) $= n$ **hacer**
12: $nd =$ destino (a)
13: $wa =$ palabra asociada a a
14: $wn =$ concatenar $(W(h), wa)$
15: $pa =$ probabilidad $p(A_n^{nd}|wa)$ asociada a a
16: $pwn = p(wa \mid W(h), M(c))$
17: $hnueva = (b(h), nd, wn, s(h) \cdot pa \cdot pwn)$
18: **Si** $\exists h' \in L(nd, c) \mid$ iguales $(h', hnueva)$ **entonces**
19: **Si** $s(hnueva) > s(h')$ **entonces**
20: $L(nd, c) = (L(nd, c) - \{h'\}) \cup \{hnueva\}$
21: **Fin Si**
22: **Si no**
23: $L(nd, c) = L(nd, c) \cup \{hnueva\}$
24: **Fin Si**
25: **Fin Para**
26: **Fin Para**
27: **Fin Para**
28: **Fin Para**
29: **Devolver** GC

dos caminos y de acuerdo a sus nodos iniciales, finales y sus secuencias de palabras (o un fragmento de éstas), si ambas se consideran iguales o no.

Este algoritmo de construcción explota la topología izquierda - derecha del grafo de palabras sobre el que se basa. De este modo las hipótesis que el algoritmo de programación dinámica va generando sólo dependen de nodos que se han procesado ya y que no se van a volver a considerar. Además, para cada nodo cada uno de los conceptos se considera por separado, utilizando por tanto un modelo especifico para cada concepto.

Puede observarse que este proceso de construcción del grafo de conceptos tiene una correspondencia con el transductor λ_{W2C} descrito anteriormente, ya que nuestro objetivo aquí es determinar secuencias de palabras válidas en el grafo de entrada que estén asociadas a alguno de los conceptos disponibles. De este modo obtenemos correspondencias entre secuencias de palabras y conceptos enriquecidas con datos sobre sus instantes de inicio y final, lo que nos permite seguir representando la información en forma de grafo.

Destacar también que previamente al proceso de construcción del grafo de conceptos podría utilizarse una generalización por categorías léxicas (números, meses, etc.) al estilo de lo especificado por el transductor λ_{gen}.

3.3. Decodificación sobre el grafo de conceptos

Una vez obtenido el grafo de conceptos, el objetivo de esta segunda etapa es buscar la

secuencia de conceptos $\hat{C}$ que maximiza la ecuación 3. Para ello utilizaremos un algoritmo de programación dinámica que encuentre el mejor camino en el grafo de conceptos, combinando la información contenida en éste con la de un modelo de lenguaje de conceptos (por ejemplo, un modelo de lenguaje de n-gramas de conceptos). Este modelo de lenguaje puede estimarse a partir de la segmentación en categorías semánticas de las frases de entrenamiento. El hecho de que para cada par de nodos i y j sólo hay un arco etiquetado con cada concepto simplifica el algoritmo de decodificación.

Como resultado de la aplicación del algoritmo de decodificación sobre el grafo de conceptos, no sólamente obtenemos $\hat{C}$, sino también la secuencia de palabras $\tilde{W}$ asociada a dicha secuencia de conceptos. Adicionalmente, y dado que cada arco del grafo de conceptos posee información sobre un concepto y su secuencia de palabras asociada, es posible obtener una segmentación de $\tilde{W}$ según los conceptos contenidos en $\hat{C}$.

4. Experimentación y resultados

Para realizar una experimentación que nos permita evaluar el método que se ha presentado, hemos utilizado el corpus DIHANA (Benedí et al., 2006). Este es un corpus de diálogo de habla espontánea en castellano en el que todos sus diálogos son telefónicos y tratan sobre información sobre trenes. Los experimentos fueron llevados a cabo utilizando los turnos de usuario de los 900 diálogos adquiridos por 225 usuarios mediante la técnica del Mago de Oz. Estos turnos de usuario se han dividido según una partición del corpus, utilizando 1 340 turnos de diálogo para test y el resto para entrenamiento. Algunos datos interesantes al respecto de las transcripciones de este corpus se muestran en la tabla 1.

Número de turnos	6 229
Número de palabras	47 222
Talla del vocabulario	811
Media de palabras por turno de usuario	7,6

Tabla 1: Características del corpus DIHANA.

Por otro lado, las transcripciones ortográficas de DIHANA están segmentadas y etiquetadas de forma semiautomática en términos de unidades semánticas. Esta segmentación es necesaria para nuestros experimentos, ya que necesitamos entrenar modelos de lenguaje locales a cada concepto. Algunas características sobre esta parte del corpus se muestran en la tabla 2.

Número de conceptos	30
Número de segmentos semánticos	18 588
Número medio de palabras por segmento	2,5
Número medio de segmentos por turno	3,0
Número medio de muestras por concepto	599,6

Tabla 2: Características del corpus DIHANA (segmentación semántica).

La entrada al sistema de comprensión ha estado constituida por tres conjuntos de grafos de palabras correspondientes a las frases de test más los modelos de lenguaje de palabras y conceptos necesarios para el sistema. Los conjuntos de grafos de palabras se han obtenido de la siguiente manera:

- El conjunto de grafos G_1 se ha generado utilizando las pronunciaciones de los usuarios y aplicándoles una extensión a palabras del método de obtención de grafos de fonemas expuesto en (Gómez Adrián, Calvo Lance, y Sanchis Arnal, 2010). El *Word Error Rate* (WER) del oráculo obtenido en estos grafos es del 4,10 % (por oráculo se entiende considerar la secuencia de palabras S contenida en el grafo más cercana a la de referencia).
- El conjunto de grafos G_2 se ha obtenido al considerar para cada grafo de G_1 la secuencia de palabras S más cercana a la frase de referencia y construyendo para cada una de ellas un grafo en el que sólo esté contenida esa frase.
- El conjunto de grafos G_3 se ha calculado generando un grafo por cada frase de referencia en el que sólo esté contenida ella. Todos los arcos tienen el mismo peso. La experimentación sobre los grafos de este conjunto es equivalente a realizar comprensión sobre texto.

Por otro lado, todos los modelos de lenguaje que se han entrenado han sido modelos de bigramas, utilizando el suavizado de Witten-Bell con interpolación lineal. Destacar también que, aunque el modelo lo permite, en este caso no se ha empleado ningún tipo de categorización léxica.

Se han tomado dos medidas de error para evaluar experimentalmente el algoritmo presentado. La primera de ellas ha sido el CER (*Concept Error Rate*), medido sobre la secuencia de conceptos proporcionada por el algoritmo. La definición del CER es análoga a la del WER pero tomando como unidad fundamental el concepto en lugar de la palabra. La segunda medida es el error a nivel de valor del concepto (EV), que es análogo al CER pero eliminando los segmentos semánticamente no relevantes (por ejemplo las cortesías) e instanciando las secuencias de palabras de cada segmento a un valor canónico (por ejemplo, si el segmento es "antes de las ocho de la tarde", se convertiría en "HORA:<=(08.00.PM)"). Los resultados obtenidos en los experimentos se muestran en la tabla 3.

Grafos de entrada	CER	EV
G_1	31,794 %	35,392 %
G_2	11,230 %	9,104 %
G_3	9,933 %	5,321 %

Tabla 3: CER y error a nivel de valor obtenidos en los experimentos utilizando la partición 1 del corpus DIHANA

De los resultados de la tabla anterior pueden extraerse varias conclusiones. En primer lugar, llama la atención que para el conjunto G_1 el EV sea mayor que el CER, mientras que para los otros dos conjuntos el EV es menor. Esto se explica considerando que los grafos obtenidos a partir de la pronunciación tienen más confusión léxica, por lo que es posible que se acierte un concepto determinado pero que la secuencia de palabras que se le haya asignado no sea la correcta y de ahí que no pueda extraerse el valor correcto. Por otro lado, y como parece lógico, puede verse que cuanto menor es la confusión y el error de los grafos de entrada, mejores resultados se obtienen. En concreto los valores alcanzados para G_2 constituyen una cota inferior del error que podemos alcanzar al decodificar semánticamente los grafos de entrada. Los resultados mostrados en la tabla podrían mejorarse utilizando categorías léxicas para agrupar palabras semánticamente similares (como meses, números, ciudades). De este modo se generalizarían las secuencias de palabras del conjunto de entrenamiento y los modelos locales a cada concepto podrían estimar mejor la probabilidad $p(W|c)$. La utilización de modelos categorizados, así como la profundización en el estudio de la relación entre la sintaxis y la semántica (Fernández, de la Clergerie, y Vilares, 2008), quedan como trabajo futuro.

5. Conclusiones

En este trabajo se ha expuesto un método para la comprensión automática del habla inspirado en el paradigma basado en transductores. Este método toma como entrada un grafo de palabras con una topología determinada. A partir de este grafo se genera un grafo de conceptos mediante un algoritmo de programación dinámica. Este algoritmo consiste en la formación de secuencias de palabras correspondientes a caminos en el grafo de palabras y en la asignación de información semántica a dichas secuencias. Aplicando otro algoritmo de programación dinámica sobre este grafo es posible obtener la secuencia de conceptos más verosímil dada la evidencia acústica de entrada. Además, es posible obtener la secuencia de palabras asociada a dicha secuencia de conceptos, así como su segmentación semántica.

Para evaluar este algoritmo, se ha llevado a cabo una evaluación experimental con el corpus DIHANA, utilizando como entrada grafos de palabras obtenidos de tres formas distintas. Los resultados obtenidos son bastante satisfactorios. Además, se han identificado algunos puntos donde podrían aplicarse futuras mejoras, y se han propuesto algunas líneas de trabajo futuro que podrían permitir mejorar los resultados aquí presentados.

Bibliografía

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

Fernández, Milagros, Eric de la Clergerie, y Manuel Vilares. 2008. Mining conceptual graphs for knowledge acquisition. En *Proceedings of the 2nd ACM workshop on Improving non english web searching*, iNEWS '08, páginas 25–32, New York, NY, USA. ACM.

Gómez Adrián, J.A., M. Calvo Lance, y E. Sanchis Arnal. 2010. Localización

de palabras basada en grafos de fonemas. *Procesamiento del lenguaje natural*, 44:59–66.

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. *Audio, Speech, and Language Processing, IEEE Transactions on*, 6(99):1569–1583.

Hakkani-Tür, D., F. Béchet, G. Riccardi, y G. Tur. 2006. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4):495–514.

Ji, G. y J. Bilmes. 2006. Backoff model training using partially observed data: Application to dialog act tagging. En *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, páginas 280–287. Association for Computational Linguistics.

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.

Lefèvre, F. 2006. A DBN-based multi-level stochastic spoken language understanding system. En *Spoken Language Technology Workshop, 2006. IEEE*, páginas 78–81. IEEE.

Lefèvre, F. 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. En *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volumen 4, páginas 13–16. IEEE.

Macherey, K., F.J. Och, y H. Ney. 2001. Natural language understanding using statistical machine translation. En *European Conf. on Speech Communication and Technology*, páginas 2205–2208. Citeseer.

McCallum, A., D. Freitag, y F. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. En *Proceedings of the Seventeenth International Conference on Machine Learning*, páginas 591–598. Citeseer.

Mohri, M. y F. Pereira Michael. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Raymond, C. y G. Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. *Proceedings of Interspeech2007, Antwerp, Belgium*, páginas 1605–1608.

Tur, G., J. Wright, A. Gorin, G. Riccardi, y D. Hakkani-Tür. 2002. Improving spoken language understanding using word confusion networks. En *Proceedings of the ICSLP*. Citeseer.