Un método de aprendizaje semi-supervisado para la modelización semántica en comprensión del habla *

A semi-supervised learning method for semantic modeling in language understanding

L. Ortega, I. Galiano, L.F. Hurtado, E. Sanchis, E. Segarra
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera sn, 46022 València
{lortega, mgaliano, lhurtado, esanchis, esegarra}@dsic.upv.es

Resumen: En este artículo 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.
Palabras clave: Comprensión del habla, Clasificación semántica, Modelización Estadística

Abstract: In this paper we present a algorithm for the statistical learning of semantic models, based on a corpus of unaligned pairs of sentences and semantic representations in terms of frames. The objective is automatically associate variable-length segments with their corresponding semantic labels to be used in speech understanding tasks. One advantage of this approach is to avoid the expensive work of segmenting and labeling the whole training corpus, process which is needed by almost all the corpus based methods. Moreover, the discrimination learning ability of this method is specially interesting. We have applied this algorithm to the development of the understanding module of a spoken dialog system, whose task is the access to information about trains. We present experiments that confirm the appropriateness of the methodology.
Keywords: Spoken Language Understanding, Semantic Classification, Statistical Modelization

1. Introducción

En muchas de las aplicaciones relacionadas con la interacción oral hombre máquina, el proceso de comprensión del habla tiene una importancia muy relevante. Tal es el caso de los sistemas de diálogo hablado, en que el módulo de comprensión debe extraer tanto la intención como la información proporcinada por el usuario. 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 (De Mori et al., 2008). Habitualmente estas aplicaciones se limitan a dominios restringidos dado que las capacidades de representación y modelización están

* Trabajo parcialmente subvencionado por el gobierno español con el proyecto TIN2008-06856-C05-02

muy condicionadas por la cantidad y variabilidad de campo semántico que se quiere modelizar. Ejemplo típicos de aplicaciones de diálogo y comprensión son los sistemas de acceso a un servicio de información considerando habla espontánea, iniciativa mixta y un vocabulario de talla media. Al igual que en otras modelizaciones en el campo del reconocimiento del habla, como son la modelización acústica o los modelos de lenguaje, los métodos estadísticos han sido ampliamente usados en comprensión del habla (Minker, 1999). Entre estas aproximaciones se encuentran aquellas basadas en clasificadores aprendidos con criterios discriminativos (Hahn et al., 2009), (Dinarelli, Moschitti, y Riccardi, 2009), las basadas en modelos estocásticos (He y Young, 2006), (Segarra et al., 2002), y las basadas en traductores (Raymond et al., 2006). La representación semántica escogida en la mayoría de las propuestas es la de frame, consistentes en un conjunto de conceptos y pares atributo-valor. Esta representación es especialmente útil cuando la aplicación consiste en un servicio de información que requiere el acceso a una base de datos y por tanto hay que completar una serie de campos para realizar un requerimiento (slot-filling). Sin embargo, en los últimos años se ha empezado a estudiar la posibilidad de trabajar con estructuras semánticas jerárquicas más complejas (Quarteroni, Riccardi, y Dinarelli, 2009), (Pérez et al., 2006).

Algunas de las ventajas de utilizar métodos estadísticos son que pueden representar la variabilidad léxica a la hora de expresar la semántica y que pueden ser aprendidos automáticamente a partir de corpus. Sin embargo, el trabajo de obtener y preparar un amplio y representativo corpus de aprendizaje es uno de sus principales inconvenientes. En toda modelización semántica hay un paso inicial en que se define la semántica que se quiere representar y el tipo de representación: su estructura y su lenguaje (conceptos, marcadores, atributos,..). Posteriormente, cuando se hace un aprendizaje basado en corpus, se debe hacer un etiquetado semántico de las frases (tanto para los clasificadores basados en criterios discriminativos como para los modelos basados en máxima verosimilitud) a nivel de palabra o secuencia de palabras. Este proceso se realiza mediante varios etiquetadores y, además de costoso, puede ser origen de errores debido a la disparidad de criterios que pueden aplicar los etiquetadores. Además no siempre es posible detectar de forma manual cuáles son los segmentos de palabras más representativos para las unidades semánticas definidas.

El trabajo que presentamos en este artículo propone un método de aprendizaje en el que no sea necesario un etiquetado manual, a nivel de las palabras, del corpus de aprendizaje y que sea capaz de obtener de forma automática los conjuntos de segmentos representativos de las unidades semánticas definidas para la aplicación. Para ello se realiza un proceso iterativo de clasificación. El corpus de aprendizaje consiste en un conjunto de frases y sus correspondientes significados (no necesariamente secuenciales con la frase), y por lo tanto no se conoce la asociación de palabras a unidades semánticas. El método propuesto parte de una asignación inicial de cada palabra o segmento de palabras a todas las unidades semánticas que aparecen en su representación semántica, y a partir del análisis estadístico de las correlaciones que existen entre segmentos y unidades detecta aquellos segmentos de longitud variable que pueden ser representativos. Este método se ha aplicado al módulo de comprensión de un sistema de diálogo cuya tarea es la consulta a un sistema de información de trenes.

En la Sección 2 se presenta la modelización semántica y los algoritmos utilizados para el aprendizaje y la decodificación. En la Sección 3 se presenta una evaluación de los métodos propuestos sobre la tarea de información sobre trenes, y en la Sección 4 se presentan las conclusiones.

2. Modelización semántica

El proceso de decodificación semántica (comprensión del habla) puede entenderse como la obtención de la secuencia de unidades semánticas (que a partir de ahora llamaremos conceptos) $c = c_1^M$, dada una secuencia de palabras $w = w_1^N$ que maximiza la probabilidad condicional $P(c|w)$. Usando la regla de Bayes, se tiene:

$$\hat{c} = \arg\max_{c} P(w|c) \cdot P(c)$$

donde $P(c)$ es la probabilidad a priori de la secuencia de conceptos c y $P(w|c)$ es la probabilidad de la secuencia de palabras w, dada la secuencia 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 $P(c)$ y $P(w|c)$, además de un algoritmo de búsqueda para obtener $\hat{c}$ entre todas las posibles secuencias de conceptos.

2.1. La propuesta de modelización estadística

Las características principales de la aproximación que presentamos en este artículo para la modelización semántica son las siguientes:

- El modelo se aprende a partir de un conjunto de pares secuencia-de-palabras w, secuencia-de-conceptos c, que no tienen por qué estar alineados. 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. 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.
- 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 $P(c_1^M) = \prod_{i=1..M} P(c_i)$, 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 no se pueden estimar por conteo, en contraste con otras aproximaciones de aprendizaje basadas en Expectation-Maximization (Dempster, Laird, y Rubin, 1977). Por tanto, para calcular $P(w|c)$ proponemos un proceso iterativo de clasificación en el cual segmentos de longitud variable se asocian a conceptos según criterios discriminativos. Considerando que $P(w|c)$ 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:

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

- A partir de esta definición de la función objetivo, el proceso de decodificación semántica puede realizarse mediante un algoritmo de programación dinámica clásico, con las modificaciones necesarias para tratar con segmentos de longitud variable.

2.2. Proceso de aprendizaje

El corpus de aprendizaje está compuesto por secuencias de pares (frase, conceptos). En la Figura 1 se muestra un ejemplo de este tipo de etiquetado semántico, en el que a una frase se le pueden asignar uno o más conceptos y atributos.

"Me gustaría informarme sobre los horarios, precios y tipos de trenes para Madrid, saliendo desde Valencia"

```
(Horario)
(Precio)
(Tipo-tren)
Origen=Valencia
Destino=Madrid
```

"Sí, gracias"

```
(Afirmación)
```

Figura 1: Ejemplos del corpus DIHANA

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 Figura 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_{max}

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

```
(Afirmación)
Tipo-viaje
(Horario)
Fecha=Viernes
```

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

```
(Afirmación)
Tipo-viaje
(Horario)
Fecha=Viernes
```

Figura 2: Ejemplo de frases diferentes con el mismo etiquetado semántico en el corpus DIHANA

asociados a las clases semánticas, utilizando para ello criterios discriminativos.

El algoritmo de aprendizaje está compuesto de cuatro etapas, que se realizan para cada una de las sucesivas longitudes de segmento, $\forall l \in 1 \cdots l_{max}$:

1. Como no hay 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. Por otra parte, 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 un proceso de categorización basado en criterios lingüísticos y en diccionarios. 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.

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 estén 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 categorías a ellas asociadas, no sean 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 CiudadOrigen o CiudadDestino, pero cuando se consideran segmentos de longitud dos la secuencia "*a Valencia*" se debe asignar claramente a CiudadDestino.

3. Resultados experimentales

En este apartado nos proponemos describir los resultados de la experimentación de nuestro sistema de comprensión con el corpus DIHANA (Benedí et al., 2006).

3.1. El corpus DIHANA

El corpus DIHANA consiste en diálogos en habla espontánea de interrogación a un sistema de información sobre trenes por teléfono.

En el proyecto DIHANA se adquirieron, mediante la técnica del Mago de Oz (Fraser y Gilbert, 1991), 900 diálogos. Se definieron tres tipos de escenarios:

- Horarios para viajes de ida.
- Horarios para viajes de ida y vuelta.
- Horarios, precios y servicios.

El número de usuarios fue de 225 con 4 diálogos por usuario. Las características del corpus transcrito se muestran en la Tabla 1

Número de turnos	6 226
Número de palabras	47 222
Talla del vocabulario	811
Media de palabras por turno	7,6

Tabla 1: *Características del corpus transcrito.*

Después de transcribir el corpus, se definieron un conjunto de frames (secuencias de conceptos y pares atributo-valor), y se etiquetó el corpus en términos de estos frames. Por simplificación,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 2 se muestran estos conceptos.

3.2. Resultados

Para realizar los experimentos hemos definido un conjunto de entrenamiento y un conjunto de test. Para ello se ha dividido el corpus de forma que el conjunto de test contiene el 20 % de las frases del corpus.

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

Tabla 2: *Lista de conceptos para la tarea DIHANA.*

En las tablas 3 y 4 se muestran los resultados de los diferentes experimentos. 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 ("Reco" en la tabla) usando CMU-Sphinx2 que proporcionó un Word Accuracy del 82 %. Los resultados de los experimentos se muestran en términos de Precisión y Cobertura.

Se han realizado tres experimentos. En el primer experimento ("Palabras" en la tabla) utilizamos el corpus sin ningún preprocesado. Con el objetivo de mejorar la cobertura de los modelos, se han propuesto otros dos experimentos, uno utilizando sólo unas categorías básicas definidas a priori ("Categorías básicas" en la tabla) y otro añadiendo categorías extendidas con información morfosintáctica y lematización ("Categorías extendidas" en la tabla).

Las categorías definidas para el etiquetado son las siguientes:

- Categorización: Las categorías básicas definidas a priori son: *ciudades, servicios, números, días* y *meses*. La figura 3 muestra un ejemplo de categorización.
- Lematización: nombres, verbos,...
- Lexicalización de algunos segmentos. Por ejemplo, el segmento "*para ir a*" se considera una única palabra (*"para_ir_a"*).

Además, se han realizado los experimentos considerando un umbral de 1 (Tabla 3),

Frase original: *"Me gustaría saber los horarios desde Valencia a Madrid para este domingo"*
Frase categorizada: *"Me gustaría saber los horarios desde [nom-ciudad] a [nom-ciudad] este [dia-semana]"*

Figura 3: Ejemplo de categorización

o de 0,8 (Tabla 4) a la hora de permitir asociar los segmentos a los conceptos $P(c_i|s_i)$. Al comparar los resultados de ambas tablas se aprecia como la cobertura mejora para los experimentos con un umbral de aceptación de palabras de 0,8 pero empeora la precisión.

En los tres experimentos, la tabla muestra la diferencia de resultados cuando se utiliza como entrada la frase transcrita correctamente o el resultado de la fase de reconocimiento, siendo los primeros siempre los mejores, como cabía esperar.

Al comparar el primer y el segundo experimento no se aprecia mejora alguna, pero es cuando usamos las categorías extendidas donde vemos claramente que 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 en las secuencias transcritas. Sin embargo, los resultados no son mejores cuando se considera como entrada la salida del reconocedor (que contiene errores de reconocimiento en algunos casos).

Experimento		Cobertura	Precisión
Palabras	Reco	0,82	0,83
	Transcripción	0,85	0,96
Categorías básicas	Reco	0,79	0,85
	Transcripción	0,85	0,96
Categorías extendidas	Reco	0,84	0,83
	Transcripción	0,90	0,95

Tabla 3: *Resultados de la experimentación con un umbral para $P(c_i|s_i)$ de 1*

4. Conclusiones

En este artículo 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. Sólo se requiere

Experimento		Cobertura	Precisión
Palabras	Reco	0.89	0.78
	Transcripción	0.93	0.87
Categorías básicas	Reco	0,88	0,77
	Transcripción	0,93	0,87
Categorías extendidas	Reco	0,90	0,78
	Transcripción	0,95	0,87

Tabla 4: *Resultados de la experimentación con un umbral para $P(c_i|s_i)$ de 0,8*

la anotación global de la frase, en vez de un etiquetado detallado de palabra o segmento de palabras a concepto. 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. Algunas mejoras en la aproximación presentada consistirían en un estudio detallado de cómo afecta la elección del umbral de aceptación de palabras a las prestaciones del sistema y en explorar la posibilidad de utilización de otros métodos de clasificación o clustering para la estimación de las probabilidades $P(c_i|s_l)$. Además sería conveniente aplicar este método a tareas más complejas, donde las realizaciones léxicas de los conceptos presenten más intersecciones y ambigüedad.

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.

De Mori, R., F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, y G. Tur. 2008. Spoken language understanding: A survey. *IEEE Signal Processing magazine*, 25(3):50–58.

Dempster, A. P., N. M. Laird, y D. B. Rubin. 1977. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society*, 39:1–38.

Dinarelli, Marco, Alessandro Moschitti, y Giuseppe Riccardi. 2009. Concept Seg-

mentation And Labeling For Conversational Speech. En *Interspeech 2009*, Brighton, U.K.

Fraser, M. y G. Gilbert. 1991. Simulating speech systems. En *Computer Speech and Language*, volumen 5, páginas 81–99.

Hahn, Stefan, Patrick Lehnen, Georg Heigold, y Hermann Ney. 2009. Optimizing CRFs for SLU Tasks in Various Languages Using Modified Training Criteria. En *Interspeech 2009*, Brighton, U.K.

He, Yulan y Steve Young. 2006. Spoken language understanding using the hidden vector state model. *Speech Communication*, 48:262–275.

Minker, W. 1999. Stocastically-based semantic analysis. En *Kluwer Academic Publishers*, Boston, USA.

Pérez, Guillermo, Gabriel Amores, Pilar Manchón, Fernando Gómez, y Jesús González. 2006. Integrating OWL Ontologies with a Dialogue Manager. *Procesamiento del Lenguaje Natural*, 37:153–160.

Quarteroni, S., G. Riccardi, y M. Dinarelli. 2009. What's In An Ontology For Spoken Language Understanding. En *Interspeech 2009*, Brighton, U.K.

Raymond, C., F. Bechet, R. De Mori, y G. Damnati. 2006. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48:288–304.

Segarra, E., E. Sanchis, M. Galiano, F. García, y L. Hurtado. 2002. Extracting Semantic Information Through Automatic Learning Techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(3):301–307.