MODELOS DE LENGUAJE BASADOS EN CATEGORÍAS SEMÁNTICAS EN UN SISTEMA DE DIÁLOGO DE HABLA ESPONTÁNEA EN CASTELLANO

*Raquel Justo, M. Inés Torres**

Universidad del País Vasco
Dpto. de Electricidad y Electrónica
48080 Leioa, España
webjublr@lg.ehu.es, manes@we.lc.ehu.es

Lluis Hurtado

Universidad Politécnica de Valencia
Dpto. de Sistemas Informáticos y Computación
46071 Valencia, España
lhurtado@dsic.upv.es

RESUMEN

El principal objetivo de este trabajo es comprobar si un modelo de lenguaje basado en categorías semánticas puede mejorar el rendimiento de un sistema de Reconocimiento Automático del Habla integrado en un sistema de diálogo de habla espontánea en castellano. Los experimentos se han llevado a cabo generando diferentes modelos de lenguaje basados en categorías semánticas. Se ha tratado de que estos modelos incorporen conocimiento relativo a las relaciones entre palabras además de la información asociada a las categorías. Para ello se han utilizado diferentes aproximaciones y los mejores resultados se han obtenido para un modelo híbrido que es una combinación lineal de un modelo basado en palabras y un modelo basado en categorías.

1. INTRODUCCIÓN

Los sistemas de diálogo se encuentran entre las aplicaciones de mayor interés en el ámbito de las tecnologías del habla. La finalidad de estos sistemas es interaccionar oralmente con los usuarios a fin de proporcionar determinados servicios como acceso a la información o control de funcionalidad de una máquina. Diversos ejemplos de este tipo de sistemas y las descripciones de los mismos pueden encontrarse en [1, 2, 3]

En un sistema de diálogo intervienen principalmente los módulos que se enumeran a continuación: Módulo de Reconocimiento Automático del Habla (RAH), Módulo de comprensión, Módulo de gestión de diálogo, Módulo de generación de respuestas y Módulo de Síntesis de Voz. Estos módulos cooperan para generar las intervenciones del sistema con el usuario [4].

El módulo de reconocimiento captura la señal acústica pronunciada por el usuario y proporciona la transcripción ortográfica al resto de los módulos del sistema de diálogo que se encargarán de interpretarla y generar la repuesta adecuada. De esta forma, un módulo de reconocimiento con bajo rendimiento, incapacita la evaluación global del sistema de diálogo, puesto que no es posible que el sistema de diálogo genere una respuesta adecuada a partir de una frase mal reconocida. Es por ello que este trabajo se centra en el módulo de RAH y más concretamente en el Modelo de Lenguaje (ML) utilizado en él. El módulo de comprensión, por su parte, recoge la salida del módulo de RAH y extrae el significado de las palabras reconocidas expresándolo en términos de un lenguaje semántico especificado para la tarea [5]. Si la salida del módulo de RAH proporciona una frase reconocida que aporte información semántica además de la transcripción ortográfica, el proceso de reconocimiento y comprensión podría integrarse en un solo paso.

El lenguaje natural entre personas se basa en una gran diversidad de conocimiento a priori que permite realizar presuposiciones y simplificar el lenguaje que se utiliza. Este hecho dificulta en gran medida el trabajo de los sistemas que tratan con el habla natural y carecen de esa información adicional, como por ejemplo el módulo de RAH. Por este motivo la utilización de un ML apropiado que se adapte a los requisitos de la tarea y que sea capaz de capturar la estructura de las frases pronunciadas por los locutores, cobra una gran importancia. Actualmente, se utilizan ML estadísticos en el proceso de reconocimiento. Para estimar de forma robusta los parámetros de este tipo de modelos se necesitan grandes cantidades de datos de entrenamiento. Sin embargo, en el caso de los sistemas de diálogo no existe mucho material de entrenamiento disponible, debido en gran medida a la dificultad de adquisición del mismo. Una de las vías para compensar la escasez o dispersión de los datos de entrenamiento se basa en la categorización del vocabulario de la aplicación en un conjunto más pequeño de clases [6].

En este trabajo la categorización se ha realizado utilizando un criterio semántico. De esta forma, se podría recuperar la información semántica asociada a las categorías en el proceso de reconocimiento, integrando en un solo paso la obtención de la transcripción ortográfica de la frase y la interpretación semántica de la misma. Este tipo de categorías semánticas están estrechamente relacionadas con la tarea que nos ocupa. En este caso la tarea consiste en consultas telefónicas acerca de horarios y precios

*Este trabajo ha sido parcialmente subvencionado por el proyecto CICYT TIN2005-08660-C04-03 y la Universidad del País Vasco 9/UPV 00224.310-15900/2004

de trenes de largo recorrido, pronunciadas en castellano por usuarios potenciales del sistema.

El principal objetivo de este trabajo es estudiar la influencia de esa categorización semántica en un ML que se encuentra integrado en el módulo de RAH de un sistema de diálogo. A partir de esta categorización se generan diferentes ML-s basados en las categorías obtenidas. Estos modelos se integran en un sistema de RAH y se evaluan en términos de WER.

En la sección 2 se describen los diferentes modelos basados en categorías que se han utilizado en este trabajo. La sección 3 detalla las características de la tarea y el corpus. En la sección 4 se describen los experimentos realizados y se presenta la evaluación de las propuestas descritas en la secciones previas. Finalmente, en la sección 5 se presentan las conclusiones principales extraídas de los experimentos, así como las propuestas para trabajos futuros

2. MODELOS DE LENGUAJE BASADOS EN CATEGORÍAS SEMÁNTICAS

Los ML basados en categorías recogen la información asociada a las relaciones entre grupos de palabras "olvidando" las relaciones entre palabras particulares. Esta pérdida de información deriva habitualmente en una reducción del rendimiento del sistema de RAH. Por este motivo, existen diversos trabajos en los que se intenta generar modelos que combinen ambas fuentes de conocimiento, obteniendo así un modelo enriquecido [7]. En este trabajo se ha realizado esta tarea mediante tres métodos diferentes que se detallan más abajo.

2.1. Modelos basados en k-gramas de categorías

En primer lugar se han generado ML basados en k-gramas de categorías. Este tipo de modelos se genera en dos pasos. En primer lugar, a partir de un corpus categorizado, se aprende un modelo de lenguaje basado exclusivamente en las relaciones entre categorías. Para ello se utilizan gramáticas k-Explorables en sentido estricto (k-EEE) [8] que son la aproximación gramatical de los bien conocidos modelos de k-gramas. Las categorías utilizadas están constituídas por palabras o secuencias de palabras que cumplen una función semántica concreta. Así, de la misma manera que se aprenden las relaciones entre categorías, se aprenden también modelos de lenguaje basados en palabras para cada una de las categorías definidas, de forma que no se pierde toda la información asociada a las relaciones entre palabras.

Los ML k-EEE vienen representados por Autómatas de Estados Finitos Estocásticos (AEFE). La integración, en un sistema de RAH, de este tipo de ML k-EEE basados en categorías se realiza en dos niveles: por un lado se considera el autómata que modela las relaciones existentes entre las categorías y por otro los autómatas correspondientes a cada una de las categorías. En el primer nivel, el más genérico, se transita mediante las etiquetas de las categorías de un estado a otro del autómata. Cada estado de este autómata se corresponde con una categoría específica, para la cual existe un autómata de palabras asociado. En el segundo nivel, el más específico, se transita por estos autómatas de palabras. En el proceso de reconocimiento se integran los autómatas de ambos niveles en un único autómata que se va construyendo al vuelo, en función de los caminos que se mantienen activos cuando se realiza la búsqueda de la hipótesis más probable.

En este trabajo se han generado dos ML de estas características. En el primero se consideran bigramas de categorías y bigramas de palabras dentro de las categorías (bi-bi). En el segundo se consideran trigramas de categorías y bigramas de palabras dentro de las categorías (tri-bi).

2.2. Modelo híbrido

Por otra parte se ha considerado un modelo híbrido. Este modelo trata de aprovechar la información relativa a las relaciones entre palabras particulares y la asociada a las relaciones entre categorías, mediante una combinación lineal de ambos modelos [7]. Es decir, se trata de una combinación lineal de un modelo de categorías como el descrito en el apartado previo y un modelo de k-gramas de palabras convencional. Se han generado dos ML híbridos:

1. Una combinación lineal de un modelo de trigramas de palabras con un modelo de bigramas de categorías (hib-bi).
2. Una combinación lineal de un modelo de trigramas de palabras con un modelo de trigramas de categorías (hib-tri).

En este caso, los dos modelos que combina el modelo híbrido se toman en cuenta de forma independiente, otorgándole a cada uno el peso que viene dado por el parámetro de la combinación lineal.

2.3. Modelos basados en MGGI

La metodología Morphic Generator Grammatical Inference (MGGI) es una metodología de inferencia gramatical que se ha utilizado con éxito en el campo del modelado de lenguaje [9]. Básicamente consiste en lo siguiente:

- Las muestras de entrenamiento son etiquetadas utilizando una función de etiquetado g.
- Se aprende un lenguaje regular 2-Explorable en sentido estricto utilizando las muestras etiquetadas. El resultado se puede representar como un autómata finito determinista. El alfabeto de este modelo es el alfabeto etiquetado, no el original.
- Aplicando un homomorfismo se procede a desetiquetar el modelo aprendido en el paso anterior. En

nuestro caso el proceso de desetiquetado se consiguió definiendo cada palabra etiquetada como una categoría cuyo único miembro era la propia palabra desetiquetada.

El resultado final es un autómata no determinista que admite cadenas sobre el alfabeto original. Los modelos MGGI son modelos más grandes que los 2-EEE para las mismas muestras de entrenamiento y el lenguaje que aceptan es un subconjunto del lenguaje aceptado por los 2-EEE. En realidad, el lenguaje inferido utilizando la metodología MGGI fluctúa entre el lenguaje 2-Explorable en sentido estricto inferido a partir de las muestras de entrenamiento sin etiquetar y el aceptor de prefijos aprendido con esas mismas muestras, dependiendo de la función de etiquetado g utilizada. En el caso que nos ocupa la función de etiquetado consiste en añadir a cada palabra la etiqueta de la categoría semántica a la que pertenecía en esa frase. Una variante de la metodología MGGI es la denominada tri-MGGI. La única diferencia respecto a la original consiste en utilizar modelos 3-Explorables en sentido estricto en lugar de los 2-Explorables de la metodología original. En este trabajo se han generado dos modelos basados en esta técnica: mggi y trimggi (utilizando la técnica tri-MGGI)

3. CARACTERÍSTICAS DE LA TAREA Y CORPUS

En este trabajo se ha utilizado una base de datos hombre-máquina en castellano. Este corpus fue adquirido dentro del marco del proyecto DIHANA [10] y consta de diferentes intervenciones en las que los usuarios solicitan información sobre horarios y precios de trenes de largo recorrido mediante el teléfono. Las características de este corpus se detallan en la tabla 1

DIHANA	
diálogos	900
locutores	225
nº frases total	6229
nº frases entrenamiento	4888
nº frases test	1341
vocabulario	718
OOV (palabras no vistas)	95
nº palabras	47219

Tabla 1. Características del corpus DIHANA

4. RESULTADOS EXPERIMENTALES

En este trabajo se han llevado a cabo un conjunto de experimentos preliminares en los que se han generado ML que tratan de aprovechar la información asociada a las relaciones entre palabras y entre categorías. Para generar un ML basado en categorías, en primer lugar se requiere un conjunto de categorías seleccionado a partir de un criterio de categorización previamente estipulado. En este caso, las categorías se han obtenido a partir de un criterio semántico y coinciden con los conceptos semánticos especificados para la tarea que utiliza el módulo de comprensión. Se han utilizado, en concreto, 31 categorías en las que se han clasificado todas las palabras del vocabulario de la tarea. A continuación se enumeran algunas de las categorías seleccionadas:

- **afirmación:** “sí”, “vale”, “está bien”, . . .
- **negación:** “no”, “nada”, “nada más”, . . .
- **hora_llegada:** “a qué hora llega”, . . .
- **hora_salida:** “tengo que salir”, “a qué hora sale”, “de salida”, . . .
- **ciudad_origen:** “de madrid”, “desde bilbao”, . . .
- **ciudad_destino:** “a girona”, “a valladolid”, . . .

Los experimentos se han llevado a cabo utilizando el corpus descrito en el punto anterior. Los modelos generados se representan mediante autómatas que han sido integrados en un sistema de RAH. Este sistema consiste finalmente en un único autómata estocástico en el que se integran los modelos acústicos, el modelo léxico y los modelos de lenguaje basados en palabras y categorías. Mediante el algoritmo de Viterbi se realiza una búsqueda en haz de la hipótesis más probable a través de este autómata general.

En la tabla 2 se muestran los resultados de WER obtenidos para los diferentes ML generados. En primer lugar se han considerado dos ML basados en palabras como referencia. Por otra parte se han generado dos ML basados en k-gramas de categorías (bi-bi y tri-bi) y otros dos ML híbridos. Estos modelos híbridos se generan como una combinación lineal de un modelo basado en trigramas de palabras y un modelo basado en bigramas (hib-bi) o trigramas (hib-tri) de categorías. Finalmente se han generado ML basados en MGGI (mggi y trimggi)

WER (%)		
pal.	**bigramas**	20.34
	trigramas	19,05
categorías	**bi-bi**	20.35
	tri-bi	20.44
	hib-bi	18.78
	hib-tri	18.77
	mggi	19.01
	trimggi	18.96

Tabla 2. Resultados de WER para los diferentes modelos basados en palabras y categorías. El conjunto de categorías utilizado es el mismo en todos los casos y los modelos que se han utilizado son los descritos en el apartado 2

La tabla 2 muestra que los mejores resultados se obtienen para el modelo híbrido. Los valores de WER obtenidos tanto para el hib-bi como para el hib-tri son mejores que el resto de los modelos basados en categorías y además mejoran el rendimiento de los ML basados en palabras. Los modelos basados en MGGI proporcionan resultados de WER ligeramente mejores que los modelos basados en palabras y se encuentran en un punto intermedio entre éstos últimos modelos y los modelos híbridos. Finalmente, los modelos basados en k-gramas de categorías (bi-bi y tri-bi) ofrecen resultados peores que el resto de los modelos basados en categorías, y muy similares a los modelos basados en palabras. En cambio, según se aprecia en trabajos previos [11], cuando se utiliza este tipo de modelos considerando unigramas de palabras dentro de las categorías, los valores de WER son considerablemente mayores que en los ML basados en palabras. Es decir, el hecho de considerar bigramas de palabras dentro de las categorías produce que los resultados de WER mejoren y sean muy próximos a los obtenidos mediante ML basados en palabras.

5. CONCLUSIONES Y TRABAJO FUTURO

Los experimentos realizados muestran que incluyendo información semántica en los modelos se puede mejorar ligeramente el rendimiento del reconocedor.

Por otra parte, mediante los experimentos realizados se ha tratado de evaluar diferentes ML que integran información relativa a dos fuentes de conocimiento. Es decir, información relativa a las relaciones entre palabras y a las relaciones entre categorías. Los experimentos muestran que la mejor manera de combinar ambas fuentes de conocimiento es la combinación lineal de las mismas, es decir, la aproximación proporcionada por el modelo híbrido. Sin embargo, dado que las diferencias entre los resultados obtenidos no son muy significativas habría que realizar una experimentación más extensa para estudiar la potencia de los modelos propuestos. Se podría ampliar la experimentación utilizando diferentes combinaciones de k-gramas en los modelos basados en categorías, e incluso experimentar con valores de k mayores que 3.

Por otra parte, la utilización de ML basados en categorías semánticas nos permite extraer la categoría semántica de las palabras reconocidas mediante el módulo de RAH. De esta forma los procesos de reconocimiento y comprensión podrían quedar integrados en un solo paso. En un trabajo futuro se quiere obtener tasas de error de las frases reconocidas en lo que a la categoría semántica se refiere. Por otra parte sería interesante comparar la información semántica proporcionada por el reconocedor con la obtenida mediante un módulo de comprensión cuando el proceso de reconocimiento y comprensión se encuentran desacoplados.

6. BIBLIOGRAFÍA

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

[2] L. Lamel, S. Rosset, J.L. Gauvain, S. Bennacef, M. Garnier-Rizet, y B. Prouts, "The LIMSI ARISE System," vol. 31, no. 4, pp. 339–354, Aug 2000.

[3] S. Seneff y J. Polifroni, "Dialogue management in the mercury flight reservation system," in *Proc. ANLP-NAACL 2000 Satellite Workshop*, 2000, pp. 1–6.

[4] E. Gianchin y S. McGlashan, "Corpus-Based Methods in Speech Processing," *Kluwer Academic, ch. Spoken Language Dyalogue Systems*, pp. 67–117, 1997.

[5] D. Vilar, M. J. Castro, y E. Sanchis, "Connectionist classification and specific stochastic models in the understanding process of a dialogue system," in *03*, Geneva (Switzerland), Sept. 2003, vol. 1, pp. 645–648.

[6] T. R. Niesler y P. C. Woodland, "A variable-length category-based n-gram language model," in *IEEE ICASSP-96*, Atlanta, GA, 1996, IEEE, vol. I, pp. 164–167.

[7] J.M. Benedí y J.A. Sánchez, "Estimation of stochastic context-free grammars and their use as language models," *Computer Speech and Language*, vol. 19, no. 3, pp. 249–274, 2005.

[8] P. Garcia y E. Vidal, "Inference of k-testable languages in the strict sense and application to syntactic pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 9, pp. 920–925, 1990.

[9] E. Segarra y L. Hurtado, "Construction of Language Models using Morfic Generator Grammatical Inference MGGI Methodology," in *97*, Rhodes (Greece), Sept. 1997, pp. 2695–2698.

[10] DIHANA project, "Dialogue System for Information Access Using Spontaneous Speech in Different Environments," Comisión Interministerial de Ciencia y Tecnología TIC2002-04103-C03-03, 2005, http://www.dihana.upv.es.

[11] R. Justo, I. Torres, y J.M. Benedí, "Category-based language models in a spanish sponken dialogue system," in *XXII Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural*, 2006, pp. 19–24.