Restricciones de Funcionamiento en Tiempo Real de un Sistema Automático de Diálogo

R. López-Cózar, A. J. Rubio, M. C. Benítez, D. H. Milone
Dpto. Electrónica y Tecnología de Computadores, Universidad de Granada, 18071, España
Tel.: +34 958243271, FAX: +34 958243230 E-mail: {ramon,rubio,carmen,dmilone}@hal.ugr.es

Resumen

En este artículo se tratan algunos de los problemas relacionados con el funcionamiento en tiempo real de los sistemas automáticos de diálogo. El trabajo se centra en una metodología utilizada para intentar que el tiempo de respuesta de un sistema de diálogo sea aceptable por los usuarios. En primer lugar, se realiza una breve descripción del sistema y de las diversas tareas de reconocimiento consideradas. Posteriormente, se presentan los resultados de una evaluación con respecto al tiempo de reconocimiento, exactitud de palabras, recuperación implícita y comprensión de frases. A partir de los resultados obtenidos se puede concluir que la estrategia utilizada es aceptable para la mayoría de las tareas de reconocimiento consideradas. Asimismo, los resultados muestran que es necesario realizar cambios en la estrategia usada para otras tareas, pues la tasa de comprensión es insuficiente y el tiempo de reconocimiento es excesivo para un sistema interactivo. Finalmente, se indican algunas líneas de trabajo futuro.

1. Introducción

En trabajos recientes se puede encontrar un considerable número de sistemas automáticos destinados a proporcionar diversos tipos de servicios a los usuarios haciendo uso del habla [1], [2], [3]. Estos sistemas utilizan principalmente las tecnologías de reconocimiento y comprensión del habla, control del diálogo, y generación de voz. Inicialmente, la interacción hombre-máquina se llevaba a cabo haciendo uso de mensajes escritos. Sin embargo, los sistemas automáticos que interactúan a partir de la voz (*sistemas automáticos de diálogo*) deben tratar diversos fenómenos no presentes en los mensajes escritos, como por ejemplo, diferencias en la voz de los usuarios, falso comienzo de las frases, ruido ambiental, cruce de conversaciones (*crosstalk*), palabras no incluidas en el vocabulario, etc.

Los corpora de grabaciones de diálogos, palabras y frases son esenciales para el desarrollo de estos sistemas. Por una parte, es necesario determinar el conjunto de palabras y expresiones lingüísticas relacionadas con el dominio de aplicación del sistema a desarrollar. Por otra parte, es necesario obtener los modelos acústicos independientes del locutor que utilizará el módulo de reconocimiento del sistema, así como estimar los modelos del lenguaje [4], [5]. Si bien los avances logrados en los últimos años son considerables, aún existen diversos problemas por resolver. Asimismo, son numerosas las líneas de investigación abiertas, las cuales tienen como finalidad lograr una interacción más cómoda y natural, logrando así una mayor aceptación por parte de los usuarios.

2. El sistema de diálogo

Hemos desarrollado un sistema automático de diálogo, denominado SAPLEN (Sistema Automático de Pedidos en LEnguaje Natural), cuya finalidad es atender telefónicamente a los clientes de restaurantes de comida rápida [6]. Creemos que este sistema podría ser útil pues permitiría ofrecer un servicio automático a los clientes durante las 24 horas del día. En particular, el sistema podría ser útil durante las horas de cierre de estos comercios, pues podría proporcionar información telefónicamente a los usuarios. Asimismo, podría mantener un registro de productos encargados telefónicamente, los cuales podrían ser enviados posteriormente a domicilio. Actualmente, el sistema puede usarse experimentalmente en nuestro laboratorio. Está implementado utilizando dos máquinas interconectadas mediante sockets: una máquina Sun SparcStation5 se usa para la adquisición de las muestras de la señal de voz, y una máquina Sun UltraEnterprise3000 se usa para realizar el reconocimiento y la conversión de texto a voz (*text-to-speech*).

Inicialmente, analizamos un corpus de 520 diálogos grabados en un restaurante de comida rápida. A partir de este análisis, determinamos el vocabulario, las estructuras sintácticas y semánticas que los clientes suelen usar, los objetivos que tanto los clientes como los encargados del restaurante pretenden conseguir, y el conjunto de frases que el sistema debe generar a modo de respuestas.

2.1. Reconocimiento del habla

El sistema de diálogo usa un reconocedor de voz continua desarrollado en nuestro laboratorio, basado en unidades fonéticas independientes del contexto modeladas mediante SCHMM (modelos ocultos de Markov semicontinuos) [7], [8]. Este reconocedor se usa también actualmente en otro sistema de diálogo desarrollado en nuestro laboratorio, llamado STACC, cuya finalidad es informar telefónicamente acerca de las calificaciones obtenidas por los estudiantes de diversas titulaciones de la Universidad de Granada [9]. El tamaño del vocabulario del sistema SAPLEN es de 2.000 palabras aproximadamente, incluyendo nombres de productos de restaurante, números, nombres de calles, avenidas, plazas, etc. El lenguaje se modela mediante bigramáticas basadas en 53 clases de palabras [10], [11].

La siguiente tabla muestra las diversas tareas de reconocimiento consideradas, así como el número de palabras que pueden combinarse para formar las frases de cada tarea.

Tarea	Nº palabras diferentes
(1) Confirmación guiada	6
(2) Confirmación libre	24
(3) Petición de información	8
(4) Corrección	16
(5) Código postal	116
(6) Pedido de productos	158
(7) Número de teléfono	116
(8) Dirección	1723
(9) Consulta	220

Tabla 1. Tareas de reconocimiento consideradas

Denominamos *confirmaciones guiadas* a aquéllas utilizadas para confirmar datos críticos en el dominio de aplicación del sistema: número de teléfono, código postal, dirección, productos encargados, precio a pagar, y tiempo estimado de entrega de los productos. Para estos datos, el sistema solicita confirmaciones del tipo "*sí/no*" a los usuarios, a fin de obtener la mayor tasa posible de reconocimiento. Por ejemplo, para confirmar el número de teléfono, el sistema genera frases como la siguiente: *"¿Has dicho 9, 5, 8, 17, 13, 28?, Por favor, responde sí o no"*. Denominamos *confirmaciones libres* a aquéllas utilizadas para confirmar datos no críticos en la aplicación. En este caso, los usuarios pueden utilizar las palabras o expresiones que deseen (como por ejemplo, "*vale*", "*claro*", "*por supuesto*", "*de acuerdo*", "*en absoluto*", etc.) sin que el sistema les indique qué palabras deben utilizar.

2.2. Comprensión del lenguaje

En trabajos recientes pueden encontrarse diversas técnicas para realizar la comprensión del lenguaje [12], [13], [14]. Algunas están basadas en palabras clave (*keywords*) proporcionadas por el reconocedor, y en reglas semánticas que tienden a ignorar las palabras no significativas desde un punto de vista semántico. El sistema de diálogo que hemos desarrollado sigue esta aproximación, y utiliza reglas de análisis basadas en conceptos semánticos clave. Las ventajas principales de este tipo de análisis son dos. Por una parte, pueden evitarse ambigüedades sintácticas si las interpretaciones correspondientes no tienen significado semántico. Por otra parte, durante el análisis pueden ignorarse los detalles sintácticos que no afectan a la interpretación semántica [15]. El módulo de análisis lingüístico del sistema utiliza 45 reglas, y puede resolver anáforas, elipsis, ambigüedades, y tautologías. A fin de contar con cierta robustez frente a errores de reconocimiento, el sistema utiliza una estrategia de recuperación implícita de errores (*Implicit Recovery*) [16]. Dicha estrategia permite, en ocasiones, obtener la interpretación semántica correcta a pesar de que algunas palabras no hayan sido reconocidas correctamente.

2.3. Gestión del diálogo

En publicaciones recientes pueden encontrarse diversas técnicas para llevar a cabo la gestión del diálogo [17], [18]. Desde un punto de vista general, estas técnicas pueden agruparse en tres categorías: dirigidas por el usuario, dirigidas por el sistema y mixtas. En el sistema de diálogo que hemos desarrollado hemos establecido una estrategia de gestión del diálogo mixta. Dicha estrategia está basada en un conjunto de objetivos que el sistema debe tratar de lograr, y diversos subobjetivos que los usuarios pueden generar durante la conversación con el sistema. Por ejemplo, un objetivo del sistema consiste en intentar vender productos del restaurante, otro consiste en obtener el número de teléfono del usuario actual, etc.

Decimos que una determinada interacción de un usuario contiene *información parcial* si la interacción no proporciona suficiente información al sistema. Una interacción de este tipo da lugar a uno o más subobjetivos que el sistema debe tratar de cubrir para completar la información parcial. Una vez logrados estos subobjetivos, el sistema puede realizar alguna acción, por ejemplo, registrar el pedido de algún producto, o responder a alguna pregunta de los usuarios. Tras lograr los subobjetivos generados por el usuario, el sistema vuelve a sus objetivos originales. Por consiguiente, la gestión del diálogo se lleva a cabo en base a una estrategia de control mixta, considerando tanto los propios objetivos del sistema como los subobjetivos generados por los usuarios.

2.4. Generación de respuestas

El sistema SAPLEN utiliza 41 patrones para la generación de respuestas en forma de texto. Utiliza varias reglas gramaticales para determinar el género, el número y el uso de pronombres. Cada patrón consiste en una serie de conceptos, expresiones y huecos. Durante la generación de las respuestas, el sistema expande los conceptos y las expresiones, y rellena los huecos con las palabras correspondientes. La generación de respuestas en modo de texto se realiza en décimas de segundo, por lo que apenas representa demora en el tiempo total de respuesta del sistema. Para efectuar la transformación de texto a voz se utiliza la plataforma de síntesis de voz denominada FESTIVAL, desarrollada en la Universidad de Edimburgo [19].

3. Evaluación

A fin de realizar una evaluación del sistema, 6 locutores varones familiarizados con el uso de sistemas de dialogo, grabaron en laboratorio 100 frases de test por cada una de las tareas de reconocimiento consideradas (ver Tabla 1). Se crearon 126 bigramáticas diferentes, cuyas probabilidades se estimaron a partir de conjuntos de frases representativas de las diversas tareas de reconocimiento. Para crear estas frases se realizó una combinación automática de todas las palabras en cada clase, siguiendo las estructuras sintácticas y semánticas derivadas del análisis del corpus de diálogos inicial, obtenido en un restaurante de comida rápida. Durante los experimentos, se utilizaron 109 bigramáticas para el reconocimiento de los datos de los usuarios (número de teléfono, código postal y dirección) y las restantes 17 bigramáticas se usaron para el reconocimiento de los pedidos de productos, peticiones de información, confirmaciones, consultas y correcciones de los usuarios.

3.1 Uso de umbrales de poda

La tasa de error y el tiempo de reconocimiento son dos medidas en mutua oposición. Generalmente, para lograr una tasa de error suficientemente baja se requiere explorar un número considerable de frases candidatas, lo cual consume un tiempo de reconocimiento alto. El tiempo de respuesta es un factor primordial para los sistemas de diálogo. En la práctica, el funcionamiento de los mismos requiere llegar a un compromiso entre ambas medidas, a fin de contar con tasas de error suficientemente reducidas y tiempos de reconocimiento aceptables por los usuarios. El umbral de poda (Up) puede utilizarse para descartar frases poco probables. Así, dicho umbral permite reducir el número de frases posibles a explorar y el tiempo total de reconocimiento.

En el sistema de diálogo que hemos desarrollado se puede asociar un umbral de poda a cada tarea de reconocimiento considerada. De esta forma, el gestor del diálogo del sistema puede proporcionar al reconocedor tanto la bigramática como el umbral de poda más apropiado, según el estado del diálogo. A fin de intentar determinar el valor idóneo para cada umbral, se utilizaron 6 valores diferentes y se obtuvieron resultados correspondientes a tiempo de reconocimiento, exactitud de palabras (*Word Accuracy, WA*), recuperación implícita, y porcentaje de comprensión de frases. Para ello, se utilizó el corpus de frases de test grabadas en el laboratorio.

3.2 Tiempo de reconocimiento

La Tabla 2 muestra los resultados obtenidos con respecto al tiempo promedio de reconocimiento de las frases de cada tarea considerada (ver Tabla 1). En la tabla se puede observar claramente la relación entre el umbral de poda y el tiempo de reconocimiento. En término medio, el tiempo de reconocimiento de las tareas (1)-(6) requirió menos de 5 segundos. El reconocimiento de los números de teléfono (7) requirió 5.22 segundos, el reconocimiento de las direcciones (8) requirió 7.20 segundos, y el reconocimiento de las consultas (9) requirió 16.45 segundos.

Tarea	Up=10	Up=20	Up=30	Up=40	Up=50	Up=60
(1)	3.04	3.05	3.05	3.05	3.05	3.06
(2)	3.17	3.18	3.20	3.21	3.23	3.24
(3)	3.68	3.70	3.79	4.02	4.33	4.54
(4)	3.65	3.64	3.71	3.76	3.94	4.24
(5)	3.50	3.55	3.65	3.81	4.0	4.31
(6)	3.65	3.65	3.75	3.92	4.05	4.32
(7)	4.02	4.12	4.91	4.93	5.84	7.05
(8)	5.35	5.82	6.25	7.14	8.35	10.30
(9)	4.35	6.35	10.14	15.55	24.15	38.20

Tabla 2. Tiempo de reconocimiento (segundos)

3.3 Exactitud de palabras

La siguiente tabla muestra los resultados obtenidos con respecto a la exactitud de palabras para el mismo conjunto de frases.

Tarea	Up=10	Up=20	Up=30	Up=40	Up=50	Up=60
(1)	60.0	85.0	93.0	100	100	100
(2)	20.0	48.0	55.0	60.0	61.0	61.0
(3)	5.0	10.0	12.0	41.0	58.0	65.0
(4)	48.0	72.0	82.0	85.0	90.0	91.0
(5)	76.63	92.42	98.62	98.62	98.96	98.96
(6)	67.16	79.10	83.58	85.07	86.56	93.41
(7)	27.53	52.17	80.67	90.33	90.82	91.30
(8)	51.35	80.99	81.08	90.99	91.20	93.72
(9)	45.97	57.47	80.45	81.60	83.90	85.05

Tabla 3. Exactitud de palabras

En la tabla puede observarse que la exactitud de palabras aumentaba conforme lo hacía el umbral de poda. La mayor tasa de exactitud de palabras correspondió a las *confirmaciones guiadas* (1), pues el vocabulario usado en esta tarea era muy reducido y la gramática utilizada era muy simple. Los valores obtenidos para las confirmaciones libres (2) y para las peticiones de información (3) fueron muy bajos. Con respecto a las *confirmaciones libres*, el reconocedor cambió muchas palabras por otras acústicamente similares. Además, se produjo un considerable número de inserciones de palabras dubitativas (como por ejemplo, *"eh"*) presentes en las bigramáticas. No obstante, muchos de estos errores fueron recuperados implícitamente durante el análisis semántico.

No se definió ninguna bigramática específica para el reconocimiento de las *peticiones de información*, pues se permite que los usuarios puedan solicitar información en cualquier momento de la conversación. El resultado mostrado en la tabla fue obtenido utilizando la bigramática correspondiente al pedido de productos. Como consecuencia, se produjo un considerable número de palabras insertadas, ya que las salidas del reconocedor tendían a seguir las estructuras sintácticas correspondientes a los pedidos de productos.

3.4 Recuperación implícita

La Tabla 4 muestra los resultados obtenidos con respecto a la tasa de recuperación implícita.

Tarea	Up=10	Up=20	Up=30	Up=40	Up=50	Up=60
(1)	17.50	26.66	42.85	0	0	0
(2)	43.54	65.0	83.78	78.12	81.81	81.81
(3)	-	-	-	-	-	-
(4)	51.85	33.33	0	0	0	0
(5)	-	-	-	-	-	-
(6)	36.36	42.85	50.0	40.0	50.0	36.36
(7)	-	-	-	-	-	-
(8)	6.25	12.50	16.66	33.33	33.33	33.33
(9)	13.33	20.0	55.55	66.66	75.0	66.0

Tabla 4. Recuperación implícita

Como puede observarse en la tabla, no hubo recuperación implícita durante la comprensión de los códigos postales (5) ni durante la comprensión de los números de teléfono (7). Ello se debió a que un error de reconocimiento en cualquier dígito (o par de dígitos) provocaba la construcción de una interpretación semántica válida, pero que no se correspondía con la frase del usuario. Tampoco existió recuperación implícita para las peticiones de información (3), pues las salidas del reconocedor correspondientes a esta tarea (en caso de ser correctas) únicamente podían ser de la forma *"necesito información"*, *"quiero información"*, etc., las cuales siempre eran comprendidas correctamente; y cuando las salidas eran erróneas, siempre se obtenían representaciones semánticas incorrectas. Las confirmaciones libres (2) obtuvieron el mayor porcentaje de recuperación implícita cuando Up=30, pues aproximadamente un 84% de las frases con algún error de reconocimiento fueron comprendidas correctamente.

3.5 Comprensión de frases

La siguiente tabla muestra los resultados obtenidos con respecto al porcentaje de comprensión de frases (*Sentence Understanding, SU*).

Tarea	Up=10	Up=20	Up=30	Up=40	Up=50	Up=60
(1)	67.0	89.0	96.0	100	100	100
(2)	65.0	86.0	93.0	94.0	94.0	94.0
(3)	48.0	55.0	75.0	81.0	86.0	90.0
(4)	87.0	90.0	91.0	94.0	95.0	96.0
(5)	48.0	80.0	88.0	88.0	96.0	96.0
(6)	42.0	65.0	83.0	87.0	90.0	91.0
(7)	4.0	34.0	51.0	64.0	67.0	67.0
(8)	25.0	65.0	75.0	90.0	90.0	91.0
(9)	60.0	65.0	80.0	85.0	88.0	90.0

Tabla 5. Comprensión de frases

Como puede observarse en la tabla, el menor porcentaje de comprensión se obtuvo para los números de teléfono (7), ya que la bigramática permitía que cualquier dígito o par de dígitos fuera seguido por cien palabras (dígitos o pares de dígitos). En consecuencia, incluso en el caso de que la exactitud de palabras fuera muy alta, un error en una palabra (dígito o par de dígitos) provocaba que el número de teléfono fuera incorrectamente comprendido.

4. Conclusiones

A la vista de los resultados expuestos, la siguiente tabla muestra los valores del umbral de poda que pueden ser considerados más apropiados para cada tarea. Además, la tabla muestra los valores de tiempo de reconocimiento, exactitud de palabras y tasa de comprensión correspondientes.

Tarea	Up	Tiempo	WA	SU
(1)	50	3.05	100	100
(2)	60	3.24	61.0	94.0
(3)	60	4.54	65.0	90.0
(4)	60	4.24	91.0	96.0
(5)	60	4.31	98.96	96.0
(6)	60	4.32	93.41	91.0
(7)	50	5.84	90.82	67.0
(8)	50	8.35	90.99	90.0
(9)	40	15.55	81.60	85.0

Tabla 6. Up seleccionado para cada tarea de reconocimiento

Como comentamos previamente, el tiempo de respuesta del sistema debe ser aceptable por los usuarios, y la tasa de exactitud de palabras debe proporcionar un porcentaje de comprensión suficientemente alto. Los resultados mostrados en la tabla representan un compromiso entre ambas medidas.

Consideramos que los resultados obtenidos para las tareas (1)-(6) son aceptables. Con respecto a los números de teléfono (7), hubiéramos obtenido mejores resultados si nos hubiéramos limitado al reconocimiento de los diez dígitos. No obstante, inicialmente decidimos utilizar también pares de dígitos para proporcionar una mayor comodidad a los usuarios. Con respecto a las direcciones (8), serían deseables mejoras en cuanto a tasa de comprensión y tiempo de reconocimiento (hay que tener en cuenta que son necesarios algunos segundos adicionales para realizar la transformación de texto a voz). Claramente, los resultados obtenidos con respecto a las consultas no son aceptables, ya que el porcentaje de comprensión obtenido es demasiado bajo y el tiempo requerido no es admisible en un sistema interactivo. Por consiguiente, es necesario establecer una estrategia diferente para esta tarea.

5. Futuro Trabajo

Son varios los aspectos del sistema en los cuales debemos seguir trabajando. Por una parte, podríamos dividir la tarea relacionada con las *consultas* en varias subtareas (cada una correspondiente a los diversos estados del diálogo) y crear las bigramáticas correspondientes. De esta forma, el sistema podría utilizar una bigramática de tamaño más reducido, acorde con el contexto de la conversación en cada momento. Así, se obtendrían reducciones en el tiempo de reconocimiento y mejoras en la tasa de comprensión de las frases.

Los resultados presentados en este trabajo se obtuvieron a partir de frases grabadas en laboratorio. Claramente, debe prestarse especial atención al funcionamiento del sistema en condiciones reales. En este sentido, tenemos previsto realizar próximamente un trabajo para evaluar el funcionamiento del sistema en condiciones de ruido.

Finalmente, sería posible mejorar el modo de interacción entre los usuarios y el sistema. Actualmente, los usuarios no pueden comunicarse con el sistema de nuevo hasta que la respuesta de éste haya finalizado. Podría lograrse una interacción más natural, y más rápida, si la respuesta del sistema pudiera cancelarse cuando se detecte que los usuarios comienzan a hablar de nuevo.

6. Referencias

[1] Gustafson J., Lindberg N., Lundeberg M., “The August Spoken Dialogue System”, Eurospeech ’99, pág. 1151-1154

[2] Asoh H., Matsui T., Fry J., Asano F. Hayamizu S., “A Spoken Dialog System for a Mobile Office Robot”, Eurospeech ’99, pág. 1139-1142

[3] Chao H., Xu P., Zhang X., Zhao S., Huang T., Xu B., “LODESTAR: A Mandarin Spoken Dialogue System for Travel Information Retrieval”, Eurospeech ’99, pág. 1159-1162

[4] Pfitzinger Hartmut R., “The Collection of Spoken Language Resources in Car Environment”, First International Conference on Language Resources and Evaluation, pág. 1097-1100

[5] Ziegenhain U., Harengel S., Kaiser J., Wilhem R. “Creating Large Pronunciation Lexica for Speech Applications”, First International Conference on Language Resources and Evaluation, pág. 1039-1043

[6] López-Cózar R., Rubio A. J., García P., Segura J. C., ”A New Word-Confidence Threshold Technique to Enhance the Performance of Spoken Dialogue Systems”, Eurospeech ’99, pág. 1395-1398

[7] Rabiner L. R., Juang B. H., “An Introduction to Hidden Markov Models”, IEEE ASSP Magazine, Enero 1986.

[8] Rabiner L. R., Juang B. H., “Fundamentals of Speech Recognition”, Prentice-Hall, 1993.

[9] Rubio A.J., García P., De la Torre A., Segura J.C., Díaz-Verdejo J.E., Benítez M.C., Sánchez V., Peinado A.M., López-Soler J.M., Pérez-Córdoba J.L., ”STACC: An Automatic Service for Information Access Using Continuous Speech Recognition Through Telephone Line”, Eurospeech ’97, pág. 1779-1782

[10] Nasr A., Esteve Y., Béchet F., Spriet T., de Mori R., “A Language Model Combining N-grams and Stochastic Finite State Automata”, Eurospeech ’99, pág. 2175-2178

[11] Jelinek F., “Statistical Methods for Speech Recognition”, MA: MIT Press, 1999

[12] Boros M., Heisterkamp P., “Linguistic Phrase Spotting in a Single Application Spoken Dialogue System”, Eurospeech ’99, pág. 1983-1986

[13] Noeth E., Boros M., Haas J., Warnke V., Gallwitz F., “A Hybrid Approach to Spoken Dialogue Understanding: Prosody, Statistics and Partial Parsing”, Eurospeech ’99, pág. 2019-2022

[14] Schadle I., Antoine J. Y., Memmi D., “Connectionist Language Models for Speech Understanding: The Problem of Word Order Variation”, Eurospeech ’99, pág. 2035-2038

[15] J. Allen, “Natural language Understanding”, Benjamin/Cummings Publishing Company Inc., 1995

[16] M. Danieli, E. Gerbino, “Metrics for evaluating dialogue strategies in a spoken language system”, AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, 1995, pág. 34-37

[17] Rosset S., Bennacef S., Lamel L., “Design Strategies for Spoken Language Dialog Systems”, Eurospeech ’99, pág. 1535-1538

[18] Relaño Gil J., Tapias D., Villar J. M., Gancedo M. C., Hernández L. A., “Flexible Mixed-Initiative Dialogue for Telephone Services”, Eurospeech ’99, pág. 1179-1182

[19] http://www.cstr.ed.ac.uk/projects/festival/festival.html