EVALUACIÓN DE UN CORPUS DE DIÁLOGOS MEDIANTE UN GESTOR DE DIÁLOGO ESTOCÁSTICO

David Griol, Lluís F. Hurtado, Emilio Sanchis, Encarna Segarra

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València. E-46022 València, Spain
{dgriol,lhurtado,esanchis,esegarra}@dsic.upv.es

RESUMEN

El principal objetivo de este artículo es evaluar las características fundamentales de un corpus de diálogos mediante su uso en una aproximación estocástica para realizar la gestión del diálogo. El gestor se basa en la generación de la nueva respuesta del sistema a través de un proceso de clasificación mediante redes neuronales. En dicho proceso, se tiene en cuenta la información suministrada por el usuario a lo largo del diálogo y la última respuesta proporcionada por el sistema. El modelo de gestión se aprende automáticamente a partir de un corpus de aprendizaje etiquetado en términos de actos de diálogo.

Esta aproximación se ha aplicado para la elaboración de un gestor de diálogo estocástico en el marco del proyecto DIHANA, cuya tarea es proporcionar información sobre trayectos en tren en el territorio español.

En el artículo se describe la metodología propuesta y se presentan las diferentes experimentaciones llevadas a cabo para evaluar la influencia de las características más relevantes del corpus DIHANA en el funcionamiento del gestor.

1. INTRODUCCIÓN

Un tema de especial interés, y de especial dificultad, en el campo de las Tecnologías del Habla es el desarrollo de sistemas de diálogo entre las computadoras y sus usuarios humanos. Un sistema de diálogo es un sistema automático capaz de emular a un ser humano en un diálogo con otra persona, con el objetivo de que el sistema cumpla con una cierta tarea (dar una cierta información, o proporcionar ciertos servicios).

Usualmente, diferentes módulos suelen interactuar para llevar a cabo este objetivo final: deben reconocer las palabras pronunciadas por el usuario, comprender su significado, gestionar el diálogo, realizar el tratamiento de errores, acceder a las bases de datos y generar la respuesta oral.

En cuanto a la gestión del diálogo, a medida que los procesos de diseño, implementación y evaluación de las estrategias de gestión del diálogo se hacen cada vez más complejos, el interés de la comunidad científica se centra en la utilización de métodos estocásticos basados en el aprendizaje de un modelo a partir de un corpus de datos correctamente etiquetado. Los modelos estadísticos permiten modelar la variabilidad de comportamientos del usuario, presentando el inconveniente de su dependencia con respecto a la calidad y cantidad de las muestras disponibles en los corpus adquiridos. En este contexto, pueden referenciarse diferentes metodologías estocásticas que modelan el comportamiento del gestor de diálogo [1] [2] [3].

Recientemente, hemos presentado una aproximación estocástica para el desarrollo de un gestor del diálogo [4]. Esta aproximación se basa principalmente en la estimación de un modelo estocástico a partir de las secuencias de actos de diálogo de usuario y de sistema obtenidos a partir de un corpus de entrenamiento. Dado un turno de usuario, el gestor de diálogo asigna la nueva respuesta del sistema como resultado de un proceso de clasificación. Para realizar esta clasificación hemos utilizado redes neuronales.

En este artículo, realizamos diferentes estudios para evaluar, mediante el uso del gestor de diálogo, diferentes aspectos representativos de un corpus de datos adquirido en el marco del proyecto DIHANA [5]. Este proyecto tiene como uno de sus principales propósitos el diseño y desarrollo de un sistema de diálogo adecuado para el acceso a la información mediante habla espontánea. La tarea del proyecto es la consulta en castellano a un sistema de información sobre trayectos de trenes de largo recorrido.

La sección 2 del artículo describe las principales características del corpus adquirido para el proyecto DIHANA, detallándose el proceso de adquisición y de etiquetado del corpus. La sección 3 presenta el gestor de diálogo estocástico utilizado para la evaluación del corpus DIHANA. En esta sección se describe la metodología de gestión propuesta, la representación utilizada para describir la historia del diálogo y el proceso de clasificación para obtener la respuesta del sistema. La sección 4 muestra los resultados de la evaluación del corpus teniendo en cuenta diferentes características representativas del mismo. Finalmente, en la sección 5 se exponen las conclusiones y trabajo futuro.

Este trabajo se ha desarrollado en el marco del proyecto EDECÁN subvencionado por la CICYT número TIN2005-08660-C04-02.

2. EL CORPUS DIHANA

La adquisición de un corpus específico de diálogos usuario-sistema plantea una gran dificultad; ya que, para que esta adquisición se realice de una manera natural se precisa un sistema de diálogo que funcione eficientemente, pero para desarrollar un sistema de diálogo eficiente es necesario una gran cantidad de datos (diálogos naturales) para el entrenamiento de sus modelos. Una de las posibles soluciones para este problema consiste en la utilización de la técnica de Mago de Oz, en la que una persona asume el papel del gestor de diálogo y ayuda al usuario a obtener respuestas a sus consultas siguiendo una estrategia definida.

En el proyecto DIHANA se adquirió un corpus de 900 diálogos utilizando la técnica del Mago de Oz siguiendo una estrategia prefijada. En esta estrategia, el gestor interacciona con el usuario en base a los niveles de confianza suministrados por el sistema, la información proporcionada por el usuario en el turno correspondiente y el estado de una estructura de datos que denominamos registro del diálogo (*dialog register* (DR)). El registro del diálogo se define como una estructura de datos que contiene la información sobre los valores de los conceptos y atributos suministrados por el usuario a través de la historia previa del diálogo.

Si todos los datos del DR disponen de una medida de confianza asociada mayor que el umbral fijado (estado seguro), el gestor elige una de las tres interacciones siguientes:

- Confirmación Implícita y Consulta a la base de datos si el DR dispone del valor de uno de los conceptos y, al menos, de los valores de sus atributos mínimos. (Ej. *Le consulto horarios de trenes con salida en Valencia destino Zaragoza para el viernes 22 de septiembre.*)
- Socitud de información al usuario si el DR no dispone de ningún concepto y/o de alguno de sus atributos mínimos (sin valor por defecto).
- Confirmación Mixta. (Ej. *Quiere horarios a Barcelona , ¿saliendo desde Granada?*). En las confirmaciones mixtas se incorporan referencias no únicamente a valores marcados como poco fiables (*Granada* en el ejemplo anterior), sino también a uno o más conceptos con suficiente fiabilidad (*solicitud de horarios* y *Barcelona*), favoreciéndose una mayor naturalidad en el diálogo. Se realiza sobre el 30 % de turnos seguros en lugar de una Confirmación Implícita-Consulta.

Si el estado es inseguro (aquel en el que uno o más datos del DR poseen una confianza menor que el umbral), el gestor selecciona una de las dos interacciones siguientes:

- Confirmación Explícita del primero de los ítems inciertos que aparezca en el DR. (Ej. *¿Quiere viajar en Talgo?*)
- Confirmación Mixta para darle naturalidad al diálogo. Se realiza sobre el 30 % de turnos de diálogo inseguros en lugar de una Confirmación Explícita.

La adquisición del corpus se llevó a cabo en tres sedes (Bilbao, Zaragoza y Valencia), disponiéndose de un Mago de Oz distinto en cada una de las sedes y utilizándose el mismo sistema y estrategia para la adquisición de diálogos. En la adquisición participaron 225 usuarios (153 hombres y 72 mujeres; 75 usuarios por sede), el número total de turnos de usuario que se adquirieron es 6280 (una media de siete turnos de usuario por diálogo). El vocabulario contiene 823 palabras.

2.1. Etiquetado del corpus

La representación de los turnos de usuario y sistema se realiza en términos de actos de diálogo. Para el caso de los turnos de usuario, los actos de diálogo se corresponden con la interpretación semántica de la intervención del usuario en base a frames (conceptos y atributos). Para la tarea se definieron ocho conceptos y diez atributos (Figura 1).

Conceptos	**Atributos**
Hora	*Origen*
Precio	*Destino*
Tipo-Tren	*Fecha-salida*
Tiempo-Recorrido	*Fecha-Llegada*
Servicios	*Hora-Salida*
Afirmación	*Hora-Llegada*
Negación	*Clase*
No-Entendido	*Tipo-tren*
	Número-Orden
	Servicios

Figura 1. Listado de conceptos y atributos definidos en la tarea DIHANA.

La Figura 2 muestra un ejemplo de la interpretación semántica de un turno de usuario realizada por el módulo de comprensión.

Turno de usuario:
Quisiera conocer los horarios y precios desde Madrid.
Interpretación semántica:
(Hora) Origen: Madrid (Precio) Origen: Madrid

Figura 2. Interpretación semántica de un turno de usuario.

En el caso de los turnos de sistema se definió un etiquetado a tres niveles. El primer nivel contiene las etiquetas referentes a un diálogo general independientemen-

te de la tarea. El segundo nivel representa los conceptos existentes en el turno, siendo dependiente de la tarea. El tercer nivel representa los valores de los atributos dados en el turno. La Figura 3 muestra las etiquetas definidas para cada uno de los niveles del etiquetado.

Primer nivel
Apertura, *Cierre*, *Indefinido*, *No-Entendido*, *Espera*, *Nueva-Consulta*, *Afirmación*, *Negación*, *Pregunta*, *Confirmación* y *Respuesta*
Segundo y tercer nivel
Hora-salida, *Hora-Llegada*, *Precio*, *Tipo-tren*, *Origen*, *Destino*, *Fecha*, *Número-orden*, *Número-Trenes*, *Servicios*, *Clase*, *Tipo-Viaje*, *Tiempo-Recorrido* y *Nil*

Figura 3. Listado de etiquetas definidas para el etiquetado de los turnos de sistema.

En la Figura 4 se muestra un ejemplo del etiquetado de un turno de sistema.

Turno de Sistema:
Le consulto horarios de Valencia a Zaragoza para el viernes 22 de septiembre.
Etiquetado:
(M:Confirmacion:Hora_salida:Destino,Dia,Origen)

Figura 4. Etiquetado de un turno del sistema.

3. EL GESTOR DE DIÁLOGO ESTOCÁSTICO

El gestor de diálogo desarrollado se basa en la modelización estocástica de las secuencias de actos de diálogo, del usuario y del sistema [4]. El gestor genera turnos de sistema basándose únicamente en la información suministrada por los turnos de usuario y la información contenida en el modelo.

Una descripción formal del modelo estocástico propuesto es la siguiente:

Sea A_i la salida del sistema de diálogo (turno de sistema) en el instante i, expresada en términos de actos de diálogo. Sea U_i la representación semántica del turno de usuario (la salida generada por el módulo de comprensión para la intervención del usuario) en el instante i, expresada en términos de frames. Un diálogo comienza con un turno de sistema que da la bienvenida al usuario y le ofrece sus servicios; llamamos a este turno A_1. Consideramos que un diálogo es una secuencia de pares (*turno de sistema, turno de usuario*):

$$(A_1, U_1), \cdots, (A_i, U_i), \cdots, (A_n, U_n)$$

donde A_1 es el turno de bienvenida del sistema, y U_n es el último turno de usuario. De ahora en adelante, denotaremos el par (A_i, U_i) como S_i, el estado de la secuencia del diálogo en el instante i.

En este contexto, consideramos que, en el instante i, el objetivo del gestor de diálogo es encontrar la mejor respuesta de sistema A_i. Esta selección es un proceso local para cada instante i y tiene en cuenta la secuencia de estados de diálogo que preceden a dicho instante. La selección se realiza mediante la siguiente maximización:

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i | S_1, \cdots, S_{i-1})$$

donde el conjunto $\mathcal{A}$ contiene todas las posibles respuestas del sistema. Como el número de posibles secuencias de estados es muy grande, establecemos una partición en el espacio de las secuencias de estados, es decir, en la historia del diálogo precediendo el instante i). Para ello, utilizamos el concepto de DR.

Para una secuencia de estados de un diálogo, existe su correspondiente secuencia de DR, donde el primer registro del diálogo (DR_1) contiene la información inicial por defecto del gestor del diálogo (*Origen* y *Clase*), y los siguientes registros DR_i se actualizan teniendo en cuenta la información suministrada durante la evolución del diálogo. Toda la información almacenada en el registro del diálogo en un instante i (DR_i) es un resumen de la información suministrada por la secuencia $S_1, \cdots, S_{i-1}$. Cabe destacar que diferentes secuencias de estados pueden conducir al mismo DR.

De este modo, la partición en el espacio de secuencias de estados se basa en considerar que dos secuencias de estados diferentes son equivalentes si conducen al mismo DR_i. Mediante esta partición obtenemos una gran reducción en el número de historias diferentes en los diálogos a expensas de la pérdida del orden cronológico en el que se suministró la información. Este factor no es un relevante para determinar la próxima respuesta del sistema A_i.

Tras aplicar las consideraciones anteriores y establecer la relación de equivalencia en las historias de los diálogos, la selección de la mejor A_i viene dada por:

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i | DR_{i-1}, S_{i-1})$$

Cada turno de usuario suministra al sistema información relativa a la tarea, es decir, el usuario solicita información sobre un concepto específico y/o suministra los valores de determinados atributos. No obstante, un turno de usuario puede además aportar otros tipos de información, como por ejemplo información independiente de la tarea. Éste es el caso de los turnos correspondientes a los actos de diálogo *Afirmación*, *Negación* y *No-Entendido*. Este tipo de información implica una toma de decisiones diferente a una mera actualización del registro DR_{i-1}. Por esta razón, para la selección de la mejor respuesta del sistema A_i, debemos tener en cuenta el DR generado desde el turno 1 al turno $i-2$, y considerar explícitamente el último estado S_{i-1}.

3.1. Representación del Registro del Diálogo

Para la tarea DIHANA, el DR se ha definido como una secuencia de 15 campos, cada uno de ellos asociado a un determinado concepto o atributo.

Para que el gestor de diálogo determine la siguiente respuesta, asumimos que no son significativos los valores exactos de los atributos. Estos valores son importantes para acceder a la base de datos y construir la respuesta del sistema en lenguaje natural. Sin embargo, la única información necesaria para determinar la siguiente acción del sistema es la presencia o no de conceptos y atributos. Por tanto, la información que almacena el DR es una codificación de cada uno de sus campos en términos de tres valores, $\{0, 1, 2\}$, de acuerdo con el siguiente criterio:

- **0**: El usuario no ha suministrado el concepto o valor del atributo correspondiente.
- **1**: El concepto o atributo está presente con una medida de confianza superior a un umbral prefijado. Las medidas de confianza se generan durante los procesos de reconocimiento y comprensión [6].
- **2**: El concepto o atributo está presente con una medida de confianza inferior al umbral.

3.2. Clasificación mediante perceptrones multicapa

Los perceptrones multicapa (MLPs) [7] son las redes neuronales más comúnmente utilizadas en tareas de clasificación. Utilizando un perceptrón multicapa, la capa de entrada se definió teniendo en cuenta la codificación utilizada para el par de entrada (DR_{i-1}, S_{i-1}). La capa de salida se definió de acuerdo con el número de posibles respuestas del sistema (51 en nuestro caso), seleccionándose como salida la respuesta de sistema A_i con mayor probabilidad asociada.

La representación definida para representar el par de entrada (DR_{i-1}, S_{i-1}) es la siguiente:

- Los dos primeros niveles del etiquetado de la última respuesta dada por el sistema (A_{i-1}): Esta información se modela mediante una variable, que posee tantos bits como posibles combinaciones de estos dos niveles (51).

$$\vec{x}_1 = (x_{1_1}, x_{1_2}, x_{1_3}, \cdots, x_{1_{51}}) \in \{0,1\}^{51}$$

- Registro del diálogo (DR): Tal y como se ha comentado previamente, el DR almacena un total de quince características (5 conceptos y 10 atributos). Cada una de estas características pueden tomar los valores $\{0, 1, 2\}$. De este modo, cada uno de los atributos del DR puede modelarse utilizando una variable con tres bits.

$$\vec{x}_i = (x_{i_1}, x_{i_2}, x_{i_3}) \in \{0,1\}^3 \, i = 2, ..., 10$$

  Dado que el sistema está orientado a completar los datos mínimos para hacer una única consulta a la base de datos, la estrategia definida para el Mago de Oz limita que en el DR únicamente puede existir un único concepto (consulta) activo simultáneamente. De esta forma, los diferentes valores que pueden tomar el conjunto de los conceptos se codifican mediante una única variable de 11 bits.

$$\vec{x}_{11} = (x_{i_1}, x_{i_2}, ..., x_{i_{11}}) \in \{0,1\}^{11}$$

- Información independiente de la tarea (actos de diálogo *Afirmación*, *Negación*, y *No-Entendido*): Estos tres actos de diálogo se han codificado de forma idéntica a los atributos almacenados en el DR. De esta forma, cada uno de estos tres actos de diálogo puede tomar los valores $\{0, 1, 2\}$ y modelarse utilizando una variable con tres bits.

$$\vec{x}_i = (x_{i_1}, x_{i_2}, x_{i_3}) \in \{0,1\}^3 \, i = 12, ..., 14$$

De este modo, la variable (DR_{i-1}, S_{i-1}) puede representarse mediante el vector de 14 características:

$$(DR_{i-1}, S_{i-1}) = (\vec{x}_1, \vec{x}_2, \vec{x}_3, \cdots, \vec{x}_{14})$$

4. EVALUACIÓN

La evaluación de las características del corpus se llevó a cabo mediante un proceso de validación cruzada. En cada una de las experimentaciones que se presentan, el corpus se dividió aleatoriamente en cinco subconjuntos. Cada evaluación, de este modo, consistió en cinco experimentaciones. En cada una de ellas se utilizó un subconjunto diferente de los cinco definidos como muestras de test, y el 80 % del corpus restante se utilizó como partición de entrenamiento.

Los clasificadores mediante MLP se programaron utilizando un software desarrollado por nuestro grupo de investigación. Los MLPs se entrenaron utilizando el algoritmo de Backpropagation con momento [7]. La topología y parámetros del algoritmo (como la tasa de aprendizaje y el momento) se estimaron mediante una búsqueda exhaustiva, utilizando el error cuadrático medio obtenido (MSE) como criterio de parada para el conjunto de validación (20 % de las muestras de entrenamiento). La mejor topología fue dos capas ocultas con 100 y 10 neuronas respectivamente.

Para realizar la evaluación se definieron cuatro medidas. La primera es el porcentaje de respuestas que coinciden con aquella generada por el Mago de Oz ($\%exacta$). Teniendo en cuenta que la estrategia del Mago de Oz posibilita un conjunto de respuestas válidas dado un determinado estado del diálogo, la segunda medida indica el porcentaje de respuestas que siguen dicha estrategia ($\%estrategia$). La tercera medida es el porcentaje de respuestas que son coherentes con el estado actual del diálogo

($\%correcta$), aunque no necesariamente sigan la estrategia fijada para el Mago de Oz. Finalmente, la cuarta medida es el porcentaje de respuestas que no son compatibles con el estado actual del diálogo ($\%error$), provocando el fallo del diálogo. Estas tres últimas medidas han sido obtenidas tras un proceso de revisión manual. Los resultados que son relevantes en el ámbito de la gestión del diálogo son $\%estrategia$, $\%correcta$ y $\%error$.

4.1. Evaluación del gestor de diálogo

Para realizar una evaluación del funcionamiento de la metodología estocástica, se realizaron unas particiones de entrenamiento y test que equilibran el número de turnos de usuario en cuanto a su localización, sexo, duración temporal de las intervenciones y número de palabras. Los conjuntos de entrenamiento contenían 4420 muestras y los de test, 1105 muestras.

Los resultados obtenidos para cada una de las medidas definidas aparecen en la Figura 5. Dichos resultados muestran el funcionamiento satisfactorio del gestor de diálogo desarrollado. La codificación propuesta para representar el estado del diálogo y el buen funcionamiento del clasificador MLP hacen posible que la respuesta generada por el sistema coincida con una de las respuestas definidas en la estrategia en un porcentaje del 97,73 %. Dicha respuesta, además, coincide exactamente con la seleccionada por el Mago de Oz en un 78,62 % de los casos. Finalmente, el número de respuestas que pueden generar el fallo del sistema es sólo un 0,18 %. Teniendo en cuenta que la media de turnos de sistema por diálogo es aproximadamente 7, a partir del porcentaje $\%error$ obtenido puede estimarse que la probabilidad de que un diálogo sea fallido es del 1,25 %.

$\%exacta$	78,62 %
$\%estrategia$	97,73 %
$\%correcta$	99,68 %
$\%error$	0,18 %

Figura 5. Evaluación del modelo estocástico.

4.2. Evaluación de la influencia de la talla del corpus

Para realizar una evaluación del funcionamiento del modelo de diálogo frente a la talla del corpus de entrenamiento, se utilizaron las mismas particiones que en el apartado anterior, descartándose muestras de forma aleatoria en los conjuntos de entrenamiento para reducir el tamaño de esta partición. Se llevaron a cabo tres experimentaciones, utilizándose un 75 % de las muestras de entrenamiento (3315 muestras), un 50 % de dicho conjunto (2210 muestras) y un 25 % (1105 muestras). Los conjuntos de test son idénticos a los del apartado anterior. La Figura 6 muestra los resultados de las experimentaciones llevadas a cabo.

	100 %	75 %	50 %	25 %
$\%exacta$	78,62 %	73,77 %	71,17 %	66,21 %
$\%estrategia$	97,73 %	94,88 %	92,56 %	90,61 %
$\%correcta$	99,68 %	98,62 %	97,33 %	96,10 %
$\%error$	0,18 %	0,52 %	0,69 %	0,95 %

Figura 6. Evaluación del modelo estocástico frente a diversas tallas del corpus de entrenamiento.

La evaluación realizada muestra el correcto funcionamiento del gestor incluso si se utiliza únicamente un 25 % del corpus de entrenamiento para construir el modelo estocástico. De esta forma, los resultados obtenidos para la medida $\%correcta$ son muy similares en las tres experimentaciones. No obstante, si utilizamos únicamente un 25 % del corpus de entrenamiento, el porcentaje de respuestas del sistema que pueden causar el fallo del diálogo es del 0,95 %, lo que nos indica que aproximadamente una de cada cien respuestas dadas por el sistema ocasionará el fallo del diálogo (6,46 % de probabilidad de fallo del diálogo).

El alto porcentaje de acierto del gestor de diálogo (incluso para un tamaño del corpus de entrenamiento reducido) puede explicarse porque existen muchos estados del diálogo que, o bien son muy frecuentes en el corpus, o bien son similares a otros estados y, por tanto, fácilmente clasificables por el MLP.

4.3. Evaluación de la influencia del sexo

Para evaluar el comportamiento del gestor teniendo en cuenta el sexo de los usuarios, se realizaron particiones del corpus con igual número de muestras de mujeres y de hombres. Los resultados de esta experimentación aparecen en la Figura 7, indicándose qué partición se utilizó como entrenamiento y como test (Entrenamiento / Test).

De la observación de los resultados, puede destacarse que prácticamente no existen diferencias remarcables en los resultados obtenidos si se realiza el aprendizaje del modelo utilizando únicamente muestras de hombres o de mujeres (primeras dos columnas de resultados en la Figura 7). Mayores diferencias se observan en las experimentaciones llevadas a cabo evaluando el modelo teniendo en cuenta el sexo de los usuarios en las particiones de test (tercera y cuarta columna de dicha figura). La diferencias obtenidas en esta última experimentación parecen indicar una mayor similitud en las muestras de hombres.

4.4. Evaluación de la procedencia de los diálogos

La evaluación del gestor de diálogo en este contexto se llevó a cabo partiendo de las mismas particiones que en el primer apartado de la evaluación, entrenando el modelo de diálogo únicamente con las muestras procedentes de la sede a evaluar y utilizando las mismas particiones de test que en dicho apartado (muestras procedentes de las tres

	Mujeres / Ambos	Hombres / Ambos	Ambos / Mujeres	Ambos / Hombres
$\%exacta$	70,55 %	69,97 %	71,72 %	76,09 %
$\%estrategia$	92,56 %	95,92 %	94,67 %	97,79 %
$\%correcta$	96,93 %	97,96 %	97,75 %	99,71 %
$\%error$	0,73 %	0,58 %	0,49 %	0,14 %

Figura 7. Evaluación del modelo estocástico teniendo en cuenta el sexo de los usuarios.

sedes). Los resultados de la experimentación aparecen reflejados en la Figura 8.

	Sede1	Sede2	Sede3
$\%exacta$	71,94 %	77,24 %	70,39 %
$\%estrategia$	95,55 %	95,51 %	90,03 %
$\%correcta$	99,16 %	97,11 %	92,15 %
$\%error$	0,42 %	1,44 %	3,92 %

Figura 8. Evaluación del modelo estocástico teniendo en cuenta la procedencia de los diálogos.

De la observación de los resultados de las diferentes experimentaciones, se aprecia un mejor funcionamiento de la metodología realizando el aprendizaje del modelo con los diálogos adquiridos en la Sede1. Aprendiendo el modelo únicamente con los diálogos de la Sede2, se obtiene un porcentaje de respuestas que siguen la estrategia ($\%estrategia$) equivalente al de la Sede1. Sin embargo, el porcentaje de respuestas erróneas ($\%error$) es tres veces mayor. En cuanto al corpus de diálogos adquirido en la Sede3, a pesar de obtener un porcentaje de respuestas correctas del 92,15 %, el número de respuestas de sistema que pueden causar el fallo del diálogo es muy elevado en comparación con el obtenido en las dos sedes restantes (probabilidad del 24,42 % de que el diálogo sea fallido). Por tanto, teniendo en cuenta la medida $\%error$, existe una diferencia significativa entre los diálogos adquiridos en cada una de las sedes.

5. CONCLUSIONES

En este artículo, hemos presentado una aproximación para el desarrollo de un gestor de diálogo estocástico, aprendida a partir de un conjunto de muestras de entrenamiento y basada en un proceso de clasificación para obtener la siguiente respuesta del sistema. La metodología propuesta es aplicable a otras tareas que compartan la filosofía de solicitar información al usuario para completar un determinado objetivo (tareas $slot - filling$).

Se ha realizado una evaluación para conocer tanto el funcionamiento de la metodología utilizada como las características del corpus adquirido para nuestra tarea. Esta experimentación nos ha permitido conocer el buen funcionamiento del sistema, así como evaluar la influencia de las características más relevantes del corpus en el funcionamiento del modelo estocástico. Los diferentes trabajos llevados a cabo para realizar la elección más adecuada de la codificación de la historia del diálogo, la metodología de clasificación propuesta y la calidad y coherencia del corpus, son las principales causas del buen funcionamiento del sistema. Como trabajo futuro, cabe destacar la utilización de un simulador de usuarios desarrollado en nuestro grupo de investigación con los objetivos principales de ampliar la evaluación presentada y mejorar el modelo de gestión estocástico.

6. BIBLIOGRAFÍA

[1] J. Williams, P. Poupart, y S. Young, "Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management," in *Recent Trends in Discourse and Dialogue. Eds L. Dybkjaer and W. Minker, Springer*, 2006.

[2] E. Levin, R. Pieraccini, y W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies," in *IEEE Transactions on Speech and Audio Processing*, 2000, pp. 8(1):11–23.

[3] F. Torres, L.F. Hurtado, F. García, E. Sanchis, y E. Segarra, "Error handling in a stochastic dialog system through confidence measures," in *Speech Communication*, 2005, pp. (45):211–229.

[4] L.F Hurtado, D. Griol, E. Sanchis, y E. Segarra, "A Stochastic Approach for Dialog Management based on Neural Networks," in *Proc. Interspeech*, 2006.

[5] J.M. Benedí, A. Varona, y E. Lleida, "DIHANA: Sistema de diálogo para el acceso a la información en habla espontánea en diferentes entornos," in *Actas de las III Jornadas en Tecnología del Habla*, Valencia (España), 2004, pp. 141–146.

[6] F. García, L. Hurtado, E. Sanchis, y E. Segarra, "The incorporation of Confidence Measures to Language Understanding," in *International Conference on Text Speech and Dialogue (TSD 2003). Lecture Notes in Artificial Intelligence series 2807*, Ceské Budejovice (Czech Republic), 2003, pp. 165–172.

[7] D. E. Rumelhart, G. E. Hinton, y R. J. Williams, *PDP: Computational models of cognition and perception, I*, chapter Learning internal representations by error propagation, pp. 319–362, MIT Press, 1986.