

Uso de valores de confianza y expectativas en el sistema de diálogo *SAPLEN*

R. López-Cózar, A.J. Rubio, P. García, J.C. Segura

{ramon,rubio,pedro,segura}@hal.ugr.es

Dpto. Electrónica y Tecnología de Computadores

Universidad de Granada

Resumen

El sistema **SAPLEN** (**S**istema **A**utomático de **P**edidos en **L**enguaje **N**atural) es un sistema de diálogo en lenguaje natural capaz de atender consultas y peticiones de productos de los clientes de los restaurantes de comida rápida [1], [2], [3], [4]. En este trabajo presentamos en primer lugar la metodología empleada para realizar el reconocimiento de las palabras de las frases de los usuarios durante el desarrollo y test del sistema. A continuación comentamos los tipos de fenómenos relacionados con el reconocimiento de palabras tratados por el sistema. Posteriormente describimos el uso de valores de confianza y expectativas para tratar apropiadamente algunos errores de reconocimiento. Finalmente, comentamos algunas líneas de trabajo futuro y conclusiones.

1. Introducción

Durante el desarrollo y test del sistema **SAPLEN** la comunicación entre éste y el usuario se ha llevado a cabo de forma escrita, usando el lenguaje natural. El sistema debe trabajar a partir de las palabras introducidas por el usuario, y debe encargarse de realizar la comprensión de las frases, realizar las acciones derivadas de dicha comprensión, mantener el diálogo con el usuario, y proporcionarle a éste la información deseada. Estas tareas entrañan una gran dificultad debido a la complejidad inherente del lenguaje natural y al uso continuo de información contextual por parte del usuario.

Sin embargo, un sistema de diálogo que pretenda ser usado en condiciones reales no puede limitarse a trabajar con la seguridad de contar siempre con las palabras introducidas

por el usuario. Debido a las condiciones ambientales, en la mayoría de las situaciones el módulo de reconocimiento no estará seguro de proporcionar a los módulos superiores de análisis la secuencia exacta de palabras pronunciadas por el usuario. El módulo de reconocimiento puede proporcionar una secuencia de palabras en la cual se hayan suprimido, añadido o cambiado palabras, respecto a la secuencia de palabras original producida por el usuario [5]. Por tanto, es necesario que los módulos superiores de análisis cuenten con la existencia de dichas situaciones anómalas a la hora de comprender las frases y gestionar el diálogo con el usuario.

2. Simulación del sistema de reconocimiento

Para poder trabajar con condiciones “reales” durante el desarrollo del sistema sin contar con un módulo de reconocimiento real, hemos usado un módulo denominado Módulo de Reconocimiento Simulado, el cual simula los efectos negativos comentados anteriormente. Describimos a continuación con más detalle los efectos a los que hacemos referencia.

2.1 Distorsiones

Para simular los efectos negativos provocados por las condiciones ambientales, el Módulo de Reconocimiento Simulado toma la secuencia de palabras real producida por el usuario y proporciona a los módulos encargados de obtener la interpretación semántica una versión de la misma que puede estar alterada por varios tipos de distorsiones, según una determinada *probabilidad de distorsión*, la cual que se fija a un valor arbitrario por el administrador del sistema al inicio de la sesión de diálogos. Tras reconocer cada palabra, el Módulo de Reconocimiento Simulado decide si debe distorsionar o no la palabra reconocida. En caso afirmativo, decide el tipo de distorsión que debe realizar (insertar cualquier otra palabra del diccionario, cambiar la palabra reconocida por cualquier otra palabra del diccionario, o bien, suprimir la palabra reconocida). Los tres tipos de distorsiones son equiprobables en nuestro sistema.

Si por ejemplo se fija la probabilidad de distorsión con el valor 0.15, se introduce una distorsión con una probabilidad de 0.15 tras reconocer cada palabra, y el tipo de distorsión a aplicar, en su caso, se decide con una probabilidad de 0.3. Esto supone que cuanto más palabras tenga una frase, más probable es que se haya producido alguna distorsión en la

misma. Por ejemplo, el usuario ha podido introducir la frase: “*HOLA QUIERO UN BOCADILLO CANTÁBRICO Y UNA CERVEZA*”, y el Módulo de Reconocimiento Simulado ha podido reconocer la frase: “*HOLA ALCOHOL UN BOCADILLO CANTÁBRICO Y CERVEZA POSTRE*”, debido a las tres distorsiones producidas: cambio de la palabra “*QUIERO*” por la palabra “*ALCOHOL*”, supresión de la palabra “*UNA*”, e inserción de la palabra “*POSTRE*”.

2.2 Valores de confianza

El Módulo de Reconocimiento Simulado asigna un valor de confianza a cada palabra reconocida. Dicho valor representa la confianza en el reconocimiento correcto de la palabra. A partir de estos valores de confianza asignados a las palabras, se obtienen otros valores de confianza que se asocian a los slots del módulo de memoria del sistema.

El valor de confianza asociado a una palabra w_o de la secuencia distorsionada, $conf(w_o)$, es un número real perteneciente al intervalo (0,1). Dicho valor se obtiene teniendo en cuenta dos factores. Uno de ellos es la *confianza en el reconocimiento* correcto de la palabra, el cual se obtiene mediante de la expresión:

$$conf_{reconocimiento}(w_o) = 1 - ruido(w_o)$$

donde $ruido(w_o)$ es una función que simula el ruido existente a la hora del reconocimiento de la palabra w_o . Dicha función devuelve un número real aleatorio perteneciente al intervalo (0,1).

El segundo factor a tener en cuenta es la *confianza en el lenguaje*, el cual se obtiene a partir de las expectativas proporcionadas por los niveles superiores de análisis [6]. Dichas expectativas suponen una información respecto al tipo de palabras que probablemente utilizará el usuario en su próxima interacción con el sistema. Por ejemplo, es de esperar que si el sistema formula al usuario la pregunta: “¿*DE QUE TAMAÑO QUIERES LA CERVEZA?*” en un determinado momento m de la conversación, el usuario responda con una palabra perteneciente a la clase “*tamaños*”, como por ejemplo, “*GRANDE*”. Para calcular la confianza en el lenguaje utilizamos la siguiente expresión:

$$conf_{lenguaje}(w_o) = \begin{cases} 1/w_i & \text{si } C_j \neq EXPECT(m) \\ 0 & \text{si } C_j = EXPECT(m) \end{cases}$$

donde $w_i \in C_j$, y $C_j \in EXPECT(m)$.

En la anterior expresión $EXPECT(m)$ es una función que proporciona el conjunto de clases de palabras C_j esperadas en un momento m , y w_i representa el conjunto de las palabras pertenecientes a cada clase C_j .

La idea es suponer que el usuario utilizará en su próxima interacción con el sistema las palabras pertenecientes a las clases indicadas por $EXPECT(m)$, con lo cual, podemos lograr que el sistema de reconocimiento pueda centrarse en esos tipos de palabras, y por tanto, pueda reconocerlas mejor. Si el usuario introduce una palabra perteneciente a una clase no esperada, es decir, una clase no contenida en $EXPECT(m)$, la *confianza en el lenguaje* es cero, y por tanto, el valor de confianza en la palabra dependerá únicamente de la *confianza en el reconocimiento* de la misma.

Una vez calculados ambos factores de confianza, el valor de confianza asociado a la palabra reconocida, $conf(w_o)$, se obtiene sumando los valores de confianza obtenidos previamente, es decir:

$$conf(w_o) = conf_{reconocimiento}(w_o) + conf_{lenguaje}(w_o)$$

Dado que $conf(w_o)$ debe estar en el intervalo (0,1), en caso de que el resultado de la suma de la expresión anterior resulte ser igual o superior al valor 1, se toma como valor para $conf(w_o)$ el valor 0.99, considerado valor de confianza máximo.

A partir de los valores de confianza asignados a las palabras de la secuencia distorsionada se determinan los valores de confianza asignados a los slots del módulo de memoria del sistema. Si en un determinado slot se almacena una única palabra, el valor de confianza asociado al slot será el valor de confianza asociado a la palabra introducida en el mismo. En caso de que en el slot se almacenen varias palabras, el valor de confianza asociado al slot será el menor de los valores de confianza asociados a las palabras introducidas en el slot.

2.3 Umbral de confianza

El sistema SAPLEN utiliza un umbral de confianza t que se fija por el administrador del sistema al inicio de la sesión de diálogos. Dicho umbral permite decidir cuándo se debe entender que una determinada palabra ha sido correctamente reconocida. Una palabra cualquiera w_o se considera correctamente reconocida si $conf(w_o) = t$.

El sistema comprueba los valores de confianza asociados a los slots inmediatamente después de que se éstos se crean. En caso de encontrar un valor de confianza asociado al slot por debajo del umbral, se considera(n) reconocida(s) incorrectamente la(s) palabra(s) introducida(s) en el mismo, en cuyo caso, se solicita al usuario la introducción de la(s) misma(s) de nuevo.

Por ejemplo, supongamos que el usuario genera la siguiente frase: “*UN BOCADILLO DE LOMO POR FAVOR*”. Supongamos además que debido a las distorsiones realizadas por el Módulo de Reconocimiento Simulado se ha producido la supresión de la palabra “*BOCADILLO*”, con lo cual, la secuencia de palabras con la que trabajaría el sistema a partir de ahora sería: “*UN DE LOMO POR FAVOR*”. Los valores de confianza asociados a las palabras podrían ser:

$$\text{conf}(\text{"UN"}) = 0.12$$

$$\text{conf}(\text{"DE"}) = 0.54$$

$$\text{conf}(\text{"LOMO"}) = 0.72$$

$$\text{conf}(\text{"POR"}) = 0.39$$

$$\text{conf}(\text{"FAVOR"}) = 0.64$$

Si se ha fijado el umbral de confianza $t=0.2$ entonces la *cantidad de producto* solicitado se considera incorrectamente reconocida, por estar por debajo del umbral. En este caso el sistema preguntará al usuario: “¿*CUÁNTOS BOCADILLOS HAS DICHO QUE QUIERES?*” a fin de intentar conseguir un valor de confianza por encima del umbral para la *cantidad del producto* solicitado. Si todos los valores de confianza están por debajo del umbral, el sistema no puede confiar en las palabras que ha reconocido. En tal caso, podría generar la frase: “*DISCULPA, NO HE COMPRENDIDO LO QUE HAS DICHO,*”.

En el anterior ejemplo podemos observar que el sistema ha sido capaz de comprender correctamente la frase del usuario y generar una respuesta apropiada, a pesar de la supresión provocada por el Módulo de Reconocimiento Simulado a causa de una distorsión. Dicha capacidad del sistema para recuperar errores provenientes del módulo de reconocimiento se denomina Recuperación Implícita [7].

2.4 Uso de expectativas para descartar palabras reconocidas incorrectamente

Como hemos visto anteriormente, es posible que se introduzcan palabras no producidas por el usuario en la frase reconocida debido al posible reconocimiento incorrecto. Sin embargo, es de esperar que una parte de dichas palabras introducidas incorrectamente cuente con un valor de confianza asociado bajo.

Como hemos comentado previamente, los valores de confianza de las palabras pertenecientes a clases no esperadas dependen únicamente de las condiciones ambientales, así que es de esperar que dichos valores sean menores que los correspondientes a las palabras pertenecientes a clases esperadas. Por tanto, podemos descartar (eliminar) las palabras no esperadas que tengan un valor de confianza por debajo del umbral, al suponer que provienen de una distorsión (inserción o cambio de palabra). De esta forma evitamos que las mencionadas palabras puedan llegar a niveles superiores de análisis.

Por ejemplo, supongamos que el sistema formula al usuario la pregunta: “¿*DE QUE QUIERES EL BOCADILLO?*” en un determinado momento de la conversación. Supongamos además, que el usuario responde de forma esperada, e introduce “*DE LOMO*”, por ejemplo. Pensemos que debido a las distorsiones (inserciones, en este caso) se ha reconocido la frase: “*DE LOMO CON CERVEZA*”. Los valores de confianza asignados a las palabras reconocidas podrían ser los siguientes:

$$\text{conf}(\text{“DE”}) = 0.35$$

$$\text{conf}(\text{“LOMO”}) = 0.85$$

$$\text{conf}(\text{“CON”}) = 0.05$$

$$\text{conf}(\text{“CERVEZA”}) = 0.18$$

Si suponemos que previamente hemos fijado el umbral de confianza al valor $t=0.2$, entonces podemos descartar las palabras “*CON*” y “*CERVEZA*” pues son palabras no esperadas cuyos valores de confianza están por debajo del umbral. Si bien puede considerarse que las palabras “*CON*” y “*DE*” son palabra esperadas en este momento del diálogo, pueden descartarse sin mayor problema pues no aportan ningún contenido semántico.

No obstante, este método presenta el inconveniente de descartar las palabras no esperadas introducidas realmente por el usuario en caso de tener un valor de confianza por debajo del umbral. Por ejemplo, si en la situación del ejemplo anterior el usuario introduce la

frase: “*DE LOMO Y TAMBIÉN ME PONES UNA CERVEZA GRANDE*”, es probable que el sistema descarte las palabras correspondientes al pedido de la bebida, por tratarse de palabras no esperadas. Las palabras no esperadas podrán llegar a los módulos superiores de análisis únicamente en el caso de que el sistema esté muy seguro de haberlas reconocido correctamente (valor del factor *confianza en el reconocimiento* alto)

3. Trabajo futuro

El trabajo futuro se orienta en dos direcciones diferentes. Por una parte, en el uso de un umbral de confianza dinámico que permita al sistema adaptarse automáticamente a las diversas condiciones ambientales en cada momento. Por otra parte, en el uso de un umbral de confianza doble que permita considerar las palabras como *reconocidas incorrectamente*, *reconocidas con reservas*, o bien, *reconocidas correctamente*. Comentamos a continuación con más detalle ambas líneas de investigación.

3.1 Umbral de confianza dinámico

Como hemos comentado a lo largo de este trabajo, el sistema SAPLEN trabaja con dos parámetros fijados arbitrariamente por el administrador del sistema al inicio de la sesión de diálogos. Uno de ellos es el *umbral de confianza*, el cual se utiliza para decidir si las palabras producidas por el usuario han sido correctamente reconocidas. El otro parámetro arbitrario es la *probabilidad de distorsión*, el cual permite al sistema trabajar con una versión alterada de la frase del usuario. Esta metodología permite que el sistema pueda trabajar con errores de reconocimiento simulados, pero presenta el inconveniente de no poder adaptarse de forma automática a las diversas condiciones ambientales en cada momento.

Supongamos que el sistema se encuentra trabajando en condiciones reales, con lo cual, ya no sería necesario el parámetro *probabilidad de distorsión* por contar con un módulo de reconocimiento real. Pensemos por ejemplo, que en un momento dado el usuario realiza sus pedidos o consultas en un ambiente muy poco ruidoso. En este caso, es de esperar que los valores de confianza asociados a las palabras sean altos en general, incluso cuando los valores de confianza en el lenguaje sean muy bajos o nulos, ya que el sistema tendrá una gran confianza en el reconocimiento. Aún en condiciones ambientales favorables pueden producirse errores de reconocimiento, inserciones de palabras por ejemplo, las cuales tendrían asociado

un valor de confianza alto por ser alta la confianza en el reconocimiento. Por tanto, sería probable que ante un umbral de confianza demasiado bajo dichas palabras insertadas incorrectamente lleguen indebidamente a los módulos superiores del análisis. Este efecto no deseable se puede producir por estar fijado el umbral de confianza a un valor demasiado bajo, dadas las condiciones ambientales favorables.

Por otra parte, pensemos que en un momento dado el usuario realiza sus pedidos o consultas en un ambiente muy ruidoso. En este caso, es de esperar que los valores de confianza asociados a las palabras sean bajos en general, incluso cuando los valores de confianza en el lenguaje sean relativamente altos, ya que el sistema tendrá muy poca confianza en el reconocimiento. En esta situación quizás las palabras con un valor de confianza relativamente alto deberían ser aceptadas, debido a la confianza en el lenguaje. Pues bien, si el umbral de confianza está fijado a un valor demasiado alto, incluso estas palabras serán rechazadas si no cuentan con un valor de confianza lo suficientemente alto. En este caso, el efecto no deseable se puede producir por estar fijado el umbral de confianza a un valor demasiado alto, teniendo en cuenta las condiciones ambientales adversas.

En vista de lo anterior, parece apropiado usar un mecanismo que permita cambiar el umbral de confianza dinámicamente. Inicialmente, el umbral podría estar situado a un valor intermedio, y podría incrementarse o decrementarse automáticamente en función de las condiciones ambientales en que se produzcan los diálogos con los usuarios. Dicho valor inicial podría determinarse a partir del número medio de errores del módulo de reconocimiento en condiciones reales. Por ejemplo, supongamos que en media se produce un 15% de errores. En este caso podríamos fijar el umbral de confianza al valor 0.15.

3.2 Uso de un umbral de confianza doble

Como hemos visto anteriormente, en nuestro sistema se usa un único valor de confianza t que permite aceptar o rechazar las palabras. Una alternativa más elaborada consiste en usar un umbral de confianza doble, o dicho de otra forma, utilizar dos umbrales de confianza en lugar de sólo uno. La idea en este caso es poder considerar una palabra como *no reconocida*, *reconocida con reservas*, o *reconocida correctamente*. Por ejemplo, podríamos usar dos umbrales t_1 y t_2 ? $(0,1)$, $t_1 < t_2$, de forma que si $conf(w_o) < t_1$ entonces w_o

se consideraría *no reconocida*, si $\text{conf}(w_o)=t_1$ y $\text{conf}(w_o)<t_2$ w_o se consideraría *reconocida con reservas*, y finalmente, si $\text{conf}(w_o)=t_2$ w_o se consideraría *reconocida correctamente*.

Por ejemplo, supongamos que ante la pregunta del sistema: “¿DE QUE QUIERES EL BOCADILLO?” el usuario responde “CANTÁBRICO”. Por simplificar, supongamos que no se ha producido ninguna distorsión de la entrada. El valor de confianza asociado a las palabras de la frase podría ser por ejemplo $\text{conf}(\text{“CANTÁBRICO”})=0.15$. Si supuestamente los umbrales de confianza son $t_1=0.1$ y $t_2=0.2$, entonces la respuesta del usuario se considera reconocida con reservas, y el sistema debería generar una pregunta para intentar confirmar la anterior respuesta del usuario, por ejemplo: “¿HAS DICHO QUE LO QUIERES CANTÁBRICO?”.

Sea cual sea el número de umbrales de confianza que usemos en nuestro sistema, no se reduce el número medio de interacciones en los diálogos, ya que en cualquier caso, sería necesario como mínimo una interacción adicional para repetir o confirmar (según sea el caso) la palabra no reconocida correctamente. Sin embargo, la confirmación de la entrada presenta una ventaja respecto a la repetición de la misma. En el caso de la confirmación, el número de palabras esperadas puede ser menor en general, y por tanto, la confianza en la palabra reconocida mayor. Por ejemplo, en el caso del anterior ejemplo, las expectativas ante la repetición de la entrada generarían en nuestro sistema una clase con 35 palabras, con lo cual, la confianza en el lenguaje sería de $1/35 = 0.02$. Por contra, las expectativas ante la confirmación de la entrada generarían una clase con 8 palabras, con lo cual, la confianza en el lenguaje sería de $1/8 = 0.12$. Por tanto, es más probable que el sistema reconozca correctamente la confirmación que la repetición de la entrada.

4. Conclusiones

En este trabajo hemos comentado algunos de los problemas a los que deben enfrentarse los sistemas de diálogo mediante voz. Además de los problemas propios del lenguaje natural (anáforas, elipsis, ambigüedad, etc.) este tipo de sistemas debe tratar apropiadamente problemas originados en los niveles inferiores de análisis (reconocimiento de palabras).

El sistema SAPLEN ha sido desarrollado en ausencia de un módulo de reconocimiento real. Para lograr condiciones semejantes a las reales, hemos dotado al sistema

de un Módulo de Reconocimiento Simulado que permite transformar la frase producida por el usuario en una versión distorsionada de la misma, la cual se utiliza por los módulos superiores del análisis encargados de obtener la interpretación semántica.

Hemos mostrado que el uso de valores de confianza asociados a las palabras constituye una información muy valiosa para los módulos del sistema encargados de la gestión del diálogo. A su vez, las expectativas proporcionadas por dichos módulos constituyen una información muy eficaz para el módulo del sistema encargado del reconocimiento de palabras.

Finalmente, hemos mencionado dos posibles mejoras a realizar en el sistema. El uso de un umbral de confianza dinámico permitiría una adaptación automática del sistema a las diversas condiciones ambientales. El uso de umbral de confianza doble conllevaría una mayor sensación de comprensión, y probablemente, un menor número de interacciones por parte de los usuarios.

5. Referencias

- [1] "A Voice Activated Dialog System for Fast-Food Restaurant Applications". Ramón López-Cózar, Pedro García, J. Díaz, Antonio J. Rubio. EUROSPEECH '97.
- [2] "A Knowledge Representation Model for a Voiced Dialogue System". R. López-Cózar, A. J. Rubio, P. García, J. Díaz. SPECOM '97.
- [3] "Una Introducción al Mecanismo de Generación de Lenguaje Natural utilizado por el Sistema SAPLEN". Ramón López-Cózar Delgado, Antonio J. Rubio Ayuso. Magazine no. 21. Sociedad Española para el Procesamiento del Lenguaje Natural.
- [4] "SAPLEN: Un Sistema de Diálogo en Lenguaje Natural para una Aplicación Comercial". Ramón López-Cózar Delgado, Antonio J. Rubio Ayuso. III Jornadas de Informática '97.
- [5] "Fundamentals of Speech Recognition". L. Rabiner, B.H. Juang. Prentice-Hall, 1993.
- [6] "On the use of expectations for detecting and repairing human-machine miscommunication". Morena Danieli. Working notes of the AAI-96.
- [7] "Metrics for evaluating dialogue strategies in a spoken language system". Morena Danieli y Elisabetta Gerbino. Proceedings of the 1995 AAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.