Medidas de confianza en sistemas de diálogo

R. San-Segundo, J. Macías-Guarasa, J.M. Montero, J. Ferreiros, R. Córdoba, J.M. Pardo
Grupo de Tecnología del Habla. UPM.
Ciudad Universitaria s/n, 28040 Madrid, Spain,
{lapiz,macias,juancho,jfl,cordoba,pardo}@die.upm.es

Resumen: Este artículo trata el cálculo de medidas de confianza en sistemas de diálogo a tres niveles diferentes: palabra, concepto y frase. Las medidas de confianza se utilizan para detectar errores de reconocimiento, conceptos semánticos incorrectos y frases fuera del dominio de la tarea (o frases mal comprendidas) en el entorno del sistema DARPA Communicator desarrollado en la Universidad de Colorado (Estados Unidos). En este artículo se proponen nuevos parámetros extraídos del analizador semántico, para el cálculo de la confianza a nivel de concepto semántico y frase completa. En nuestro caso hemos considerado una red neuronal para combinar todos los parámetros utilizados y generar una única medida de confianza. Se ha conseguido detectar un 53,2% de palabras erróneas, un 50,1% de conceptos semánticos erróneos y un 76,1% de frases erróneas (fuera del dominio de la tarea o más interpretadas por el sistema) considerando un rechazo incorrecto del 5%.
Palabras clave: Medidas de confianza en sistemas de diálogo, parámetros obtenidos del analizador semántico, redes neuronales.

Abstract: This paper investigates improved confidence assessment for spoken dialogue systems at three levels: word, concept and utterance levels. The confidence scores are used to detect word-level speech recognition errors, incorrect concepts and out of domain or miss-understood utterances in the CU Communicator system. New measures from the speech understanding component are proposed for confidence annotation at concept and utterance levels. We have considered a neural network to combine all measures in order to provide a unique confidence score. Using the data collected from a live telephony system, it is shown that 53.2% of incorrectly recognized words, 50.1% of incorrect concepts and 76.1% of out of domain or miss-understood utterances are detected at a 5% false rejection rate.
Keywords: confidence measures, understanding confidence annotation, neural networks.

1 Introducción

Debido a que el reconocimiento automático del habla dista mucho de ser perfecto, cuando se utilice un sistema de reconocimiento automático, se debe analizar la calidad de las palabras reconocidas o de los conceptos comprendidos por el sistema con el fin de detectar posibles errores o zonas de gran ambigüedad. Esta necesidad es aún más importante en los sistemas de diálogo (SDs) donde una mala interpretación de la frase pronunciada puede llevar al sistema a realizar un comportamiento erróneo. Típicamente en los SDs existen dos módulos anteriores al módulo de gestión de diálogo: reconocimiento y comprensión. Dichos módulos se encargan de extraer la información semántica de la frase pronunciada por el usuario. Esta información es utilizada por el gestor de diálogo para avanzar en su interacción con él. Las medidas de confianza obtenidas en estos dos módulos tienen como objetivo evaluar su comportamiento de forma que el gestor de diálogo pueda medir la calidad de la información recibida y en consecuencia, elegir la acción concreta a realizar: rechazar la frase, preguntar otra vez, o pedir confirmación de alguno de los datos obtenidos (Sturm et al, 1999; San-Segundo et al, 2001b). Según la resolución de las medidas de confianza, podemos clasificarlas en 3 niveles diferentes:

- *Nivel de palabra:* en este caso el objetivo es detectar palabras mal reconocidas. Para ello utilizaremos parámetros obtenidos del módulo de reconocimiento de voz: tanto del

proceso de descodificación como del modelo de lenguaje.

- *Nivel de concepto:* en este caso se pretende detectar conceptos erróneos dentro de una frase determinada. Las medidas de confianza en este caso son muy importantes para la gestión de diálogo puesto que es la información semántica, la que se utiliza para realizar esta labor de gestión y decidir cuales van a ser las acciones del sistema en su interacción con el usuario. En este caso utilizaremos parámetros obtenidos del reconocedor de voz y del analizador semántico.
- *Nivel de frase:* en este nivel, el objetivo es detectar, por un lado, frases fuera del dominio de la aplicación, y por otro, frases del dominio con problemas en el reconocimiento que no tienen ninguna información semántica o concepto correcto. Se pretende por tanto, detectar frases que no van a ser correctamente reconocidas y comprendidas por nuestro sistema, evitando realizar interpretaciones erróneas.

El sistema de diálogo utilizado en nuestros experimentos ha sido desarrollado por el CSLR (The Center for Spoken Language Research) de la Universidad de Colorado, dentro del proyecto DARPA Communicator. El sistema combina reconocimiento de habla continua, compresión de lenguaje natural y control de diálogo flexible para ofrecer, mediante una interacción natural con el usuario a través del teléfono, información y reserva de billetes de avión, hoteles y coches de alquiler. El sistema se conecta a la página web de una agencia de viajes, de donde extrae la información actualizada (Ward y Pellom, 1999).

2 Base de datos

La base de datos utilizada para los experimentos realizados sobre este sistema, se obtuvo durante la recogida de datos realizada en el CSLR desde noviembre de 1999 hasta mayo de 2000 (Pellom et at, 2000). Durante este período se recogieron más de 900 llamadas telefónicas obteniendo alrededor de 11.500 frases con más de 30.000 palabras en total.

A la hora de realizar los experimentos, hemos dividido aleatoriamente el conjunto de frases en tres subconjuntos: 66% de las frases para el entrenamiento de la red neuronal utilizada en la combinación de los parámetros (ver apartado 8), 17% para su validación y el 17% para evaluación. Esta división se ha repetido 6 veces realizando un proceso Round-Robin, de forma que cada vez, se van utilizando unos datos diferentes para entrenar, validar o evaluar la Red Neuronal, consiguiendo que se usen todos los datos para evaluar las medidas de confianza al menos una vez. Los resultados presentados en esta sección son la media de los valores obtenidos en todos los experimentos.

3 Etiquetado automático de ejemplos

Para poder realizar la experimentación sobre medidas de confianza es necesario clasificar cada ejemplo como correcto (el sistema debe aceptarlo) o incorrecto (el sistema lo debe rechazar). Al nivel de palabra, este etiquetado se realiza mediante un alineamiento dinámico entre la hipótesis del reconocedor y la frase de referencia, obteniéndose las palabras correctas, insertadas, borradas y sustituidas. En un sistema real, a la salida de reconocedor se dispone únicamente de las palabras correctas, insertadas y sustituidas, por esta razón, trabajaremos siempre con este tipo de palabras, no asignando ningún valor de confianza a las palabras borradas.

Al igual que en el caso anterior, necesitamos etiquetar cada concepto semántico de la base de datos como correcto o incorrecto. Este etiquetado se ha realizado de forma automática de la siguiente forma. Al analizar la frase de referencia obtenemos una secuencia de conceptos determinada. Esta secuencia se utiliza como base con la que comparar la secuencia de conceptos obtenida de la hipótesis, y así poder calcular los casos de conceptos correctos, sustituidos, borrados e insertados. El proceso es análogo al seguido para el caso de las secuencias de palabras.

A la hora de decidir un criterio para etiquetar las frases, debemos pensar en qué información utiliza el gestor de diálogo para interactuar con el usuario. Esta información es la obtenida a la salida del analizador semántico. En nuestro caso consideraremos como frase a rechazar aquellas que no pertenezcan al dominio de aplicación, o perteneciendo, no tengan ningún concepto semántico correcto. Este criterio permite la aceptación de frases con algún error, lo que dificulta en mayor medida las estrategias de rechazo. Si en una misma frase existen conceptos correctos e incorrectos, estos deberán ser discriminados con las medidas de confianza al nivel de concepto.

4 Evaluación de las medidas de confianza

Antes de proceder a la presentación de los resultados, definiremos algunos términos importantes y explicaremos el proceso de evaluación utilizado. Veamos las siguientes definiciones:

- Rechazo Correcto (RC): porcentaje de ejemplos a rechazar que han sido rechazados correctamente.
- Rechazo Incorrecto (RI): porcentaje de ejemplos a aceptar que han sido rechazados incorrectamente.
- Error de Clasificación (EC): porcentaje de ejemplos que se etiquetan incorrectamente, ya sean ejemplos a rechazar o ejemplos que se deben aceptar. El complementario al error de clasificación lo denominaremos Tasa de Clasificación (TC): TC=100%–EC.
- Error de Referencia (ER): el error de referencia queda definido por la distribución inicial de los ejemplos para el estudio de las medidas de confianza (en el caso del estudio al nivel de palabra, esta distribución viene definida por la calidad del sistema de reconocimiento).

Al nivel de palabra dispondremos de palabras correctas, insertadas y sustituidas, siendo el Error de Referencia el porcentaje correspondiente a las palabras insertadas y sustituidas. De igual forma el Error de Referencia al nivel de concepto es el porcentaje de conceptos insertados y sustituidos en la secuencia. En el caso de las frases, el Error de Referencia lo define el conjunto de casos etiquetados como frases a rechazar.

Para evaluar las medidas de confianza procedemos de la siguiente forma: a cada uno de los ejemplos del conjunto de test le aplicamos la Red Neuronal, obteniendo un valor de confianza comprendido en el intervalo [0, 1]. Para cada tipo de ejemplo (ejemplo a rechazar o a aceptar) se representa la distribución de casos según la medida de confianza obtenida de la Red Neuronal. Para realizar esta representación se divide el intervalo [0, 1] en 100 segmentos de anchura 0,01. En la figura 1 podemos ver un ejemplo de esta representación.

Sobre el eje x podemos definir un umbral de forma que los casos que obtengan un valor de confianza por debajo de este umbral se etiqueten como ejemplos a rechazar y los que obtengan un valor de confianza mayor se considerarán como ejemplos a aceptar. Para cualquier umbral considerado, podemos calcular el Rechazo Correcto (RC) y el Rechazo Incorrecto (RI).

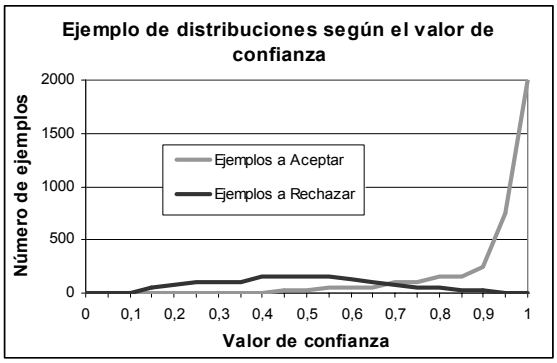

Figura 1: Ejemplo de representación del número de casos según el valor de confianza.

Las fórmulas son las siguientes:

$$RC\ (\%) = 100 \times \frac{N_{ERRORES\ \ DEBAJO\ \ UMBRAL}}{N_{ERRORES}}$$

$$RI\ (\%) = 100 \times \frac{N_{ACIERTOS\ \ DEBAJO\ \ UMBRAL}}{N_{ACIERTOS}}$$

donde:

- $N_{ACIERTOS\ DEBAJO\ UMBRAL}$: nº de casos a aceptar con confianza menor que el umbral.
- $N_{ERRORES\ ENCIMA\ UMBRAL}$: nº de casos a rechazar con confianza mayor que el umbral.
- N_{TOTAL}: nº total de ejemplos.

Como se puede ver en la figura 1, generalmente las dos distribuciones se solapan, lo que impide la definición de un umbral que permita separar ambas distribuciones sin error.

A la hora de evaluar las medidas de confianza calcularemos la evolución del RC según el RI, al ir modificando el umbral de confianza considerado, y el Error Mínimo de Clasificación obtenido a lo largo de esta variación del umbral. Especial énfasis haremos sobre la tasa de Rechazo Correcto para tasas de Rechazo Incorrecto del 2,5% y del 5,0%. Estos valores corresponden con el margen sobre el que oscilará el punto de trabajo en el que queremos que funcione nuestro sistema. Generalmente no es recomendable que el sistema rechace más del 5% de casos correctos (que deberían haber sido aceptados), porque podría ocasionar al usuario cierta sensación de frustración en su interacción. Para estos límites de Rechazo Incorrecto deseamos detectar y

rechazar la mayor cantidad posible de errores. Por último como ya hemos comentado, el Error de Referencia vendrá determinado por la distribución inicial de ejemplos.

5 Nivel de palabra

En este nivel se pretende el etiquetado de cada palabra con un valor de confianza que nos ofrezca una idea de la certeza con la que se ha reconocido dicha palabra. Los parámetros considerados del proceso de decodificación son los siguientes (Chase, 1997; Kamppari y Hazen, 2000; Macías-Guarasa et al, 2000, Zhang, R., Rudnicky, A., 2001;Hazen et al, 2002):

- *Verosimilitud normalizada:* es el logaritmo de la verosimilitud acumulada a lo largo de la palabra (durante el proceso de reconocimiento), dividido por el número de tramas.
- *Homogeneidad de la palabra en la lista de las 100 mejores hipótesis:* porcentaje de veces que una misma palabra aparece en posición análoga (mismo segmento de voz) en las 100 mejores hipótesis de reconocimiento.
- *Densidad del grafo de palabras:* número de enlaces o transiciones, desde cualquier palabra hasta la palabra considerada, calculadas sobre el grafo de palabras obtenido durante la primera etapa de reconocimiento.
- *Perplejidad de fonemas:* número medio de modelos de alófono activos (sobreviven a la poda introducida por el Beam Search) a lo largo de las tramas en las que permanece activa la palabra bajo estudio.

Los parámetros considerados provenientes del modelo de lenguaje son los siguientes (San-Segundo et al, 2000):

- *Comportamiento Back-Off del modelo de lenguaje:* comportamiento del modelo de lenguaje utilizado para calcular la probabilidad de la palabra en la secuencia (P(Wj) como función de las palabras anteriores: Wj-1 y Wj-2). En la tabla 1, se pueden ver los valores de confianza (dentro del intervalo 0-1) asignados a cada tipo de comportamiento.
- *Probabilidad de la palabra en la secuencia P(Wj), obtenida del modelo de lenguaje:* este parámetro nos ofrece información complementaria puesto que palabras con comportamientos iguales pueden tener probabilidades diferentes.

Valor	Comportamiento
1,0	$P(W_i)$ como sucesión trigram: $P(W_j,W_{j-1},W_{j-2})$
0,8	$P(W_i)$ como sucesión bigram-bigram: $P(W_j,W_{j-1})$ y $P(W_{j-1},W_{j-2})$
0,6	$P(W_i)$ como sucesión bigram: $P(W_j,W_{j-1})$
0,4	$P(W_i)$ como sucesión unigram-bigram: $P(W_j)$ y $P(W_{j-1},W_{j-2})$
0,3	$P(W_i)$ como sucesión unigram-unigram: $P(W_j)$ y $P(W_{j-1})$
0,2	$P(W_j)$ como unigram: $P(W_j)$
0,1	Palabra desconocida. Nunca se da en la salida del reconocedor.

Tabla 1: Asignación del valor de confianza según el comportamiento utilizado en el cálculo de la probabilidad de la palabra en la secuencia.

A la hora de analizar una palabra concreta, se considerarán como parámetros de confianza tanto el comportamiento y la probabilidad de la palabra considerada, como el de las dos palabras anteriores y las dos posteriores. La razón de utilizar un contexto de dos es porque el modelo de lenguaje utilizado es 3-gram con lo que errores de palabras contextuales pueden hacer que los parámetros de la palabra analizada revelen baja confianza sin que esta sea una palabra incorrecta.

6 Nivel de concepto

En este nivel, trabajaremos con parámetros del decodificador, del modelo de lenguaje y del sistema de compresión o analizador semántico. Los dos primeros parámetros pretenden la incorporación de la confianza de cada palabra en el cálculo de la confianza al nivel de concepto semántico. De esta forma se incorpora el conocimiento del decodificador y del modelo de lenguaje en el etiquetado de confianza a este nivel (San-Segundo et al, 2001a):

- *Confianza Media de las palabras pertenecientes a la Regla utilizada en la obtención del Concepto (CMRC)*. Este parámetro se obtiene realizando la media de la confianza obtenida en el nivel anterior, a

lo largo de las palabras utilizadas en la aplicación de la regla que generó el concepto analizado.

- *Confianza Media de las palabras pertenecientes al Valor del Concepto (CMVC)*. De forma análoga al anterior, este parámetro se calcula realizando la media de la confianza obtenida en el nivel anterior, a lo largo de las palabras que forman el valor del concepto analizado.

  La razón de hacer esta diferenciación es porque el conjunto de palabras pertenecientes a la regla aplicada puede ser mayor que las que forman el valor del concepto propiamente dicho.

Además se proponen un conjunto de parámetros obtenidos exclusivamente del módulo de compresión. Son los siguientes:

- *Número de Palabras contenidas en la Regla aplicada para obtener el concepto (NPR)*. A medida que se aplican reglas que involucran mayor número de palabras, se pone de manifiesto un mayor ajuste entre la frase y la gramática, lo que redunda en una mayor confianza del concepto obtenido.
- *Número de Palabras contenidas en el Valor del concepto obtenido (NPV)*. Siguiendo un razonamiento similar al anterior, el valor de un concepto suele ser una secuencia de palabras características con alta probabilidad en el modelo de lenguaje, lo que refleja también una mayor confianza cuando estas secuencias son más largas.
- *Homogeneidad del concepto en las 100 mejores hipótesis (HV)*. De forma análoga al caso de las palabras, este parámetro es el porcentaje de veces que un concepto aparece en las 100 mejores hipótesis de reconocimiento.
- *Homogeneidad del concepto y su valor en las 100 mejores hipótesis (HCV)*. En este caso exigimos que aparezca el concepto con el mismo valor.

  Estos dos últimos parámetros son muy útiles cuando tenemos palabras que tienen el mismo comportamiento semántico (ej: dos nombres de ciudad), y además, tienen un gran parecido acústico como por ejemplo las ciudades Boston y Austin (en Inglés). En estos casos se producen patrones característicos en los que el concepto semántico aparece muchas veces, pero su valor fluctúa bastante.

Los dos últimos parámetros que describimos a continuación se obtienen a partir de la definición de un modelo de lenguaje al nivel de concepto. En nuestro caso, hemos entrenado un modelo de lenguaje conceptual 3-gram, utilizando las secuencias de conceptos obtenidas como resultado de analizar las frases de referencia (transcripciones manuales) del conjunto de entrenamiento. De forma análoga al caso del nivel de palabra, podemos definir los siguientes dos parámetros:

- *Comportamiento Back-Off del modelo de lenguaje (CML)*: comportamiento del modelo de lenguaje utilizado para calcular la probabilidad del concepto en la secuencia ($P(C_j)$ como función de los conceptos anteriores: C_{j-1} y C_{j-2}).
- *Probabilidad del concepto en la secuencia $P(C_j)$, obtenida del modelo de lenguaje (PML)*.

  De igual forma que en el apartado anterior, se utilizarán los parámetros del concepto analizado, junto con los parámetros de los dos conceptos anteriores y posteriores (contexto de 5 conceptos).

Un detalle importante a la hora de entrenar el modelo de lenguaje conceptual es que no se deben considerar los valores de los conceptos. Es decir, el concepto Ciudad Destino, por ejemplo, debe ser considerado como la misma unidad independientemente del valor que tenga asociado.

7 Nivel de frase

En este nivel utilizaremos medidas obtenidas del proceso de decodificación, del modelo de lenguaje y del módulo de comprensión. Estas medidas básicamente son las mismas que las propuestas al nivel de concepto pero extendidas a toda la frase.

- *Confianza Media al nivel de Palabra:* es la media de los valores de confianza obtenidos para cada una de las palabras que componen la frase.
- *Confianza Media al nivel de Concepto*: es la media de los valores de confianza obtenidos para cada uno de los conceptos que componen la frase.
- *Porcentaje de Palabras Analizadas Semánticamente*: número de palabras que pertenecen a algún concepto o a alguna regla utilizada para obtener algún concepto, dividido por el número de palabras de la frase.

- *Porcentaje de Palabras pertenecientes a la Tarea:* número de palabras que pertenecen a algún concepto o a alguna regla definida en la tarea (aunque no haya sido utilizada en la frase actual), dividido por el número de palabras de la frase.
- *Porcentaje de Conceptos*: número de conceptos extraídos divido por el número de palabras que componen la frase. Cuando el número de conceptos es muy bajo comparado con el número de palabras contenidas en una frase, este hecho nos revela una baja confianza de la frase.
- *Porcentaje de frases en las 100 Mejores hipótesis con algún Concepto*: porcentaje de hipótesis de reconocimiento en las que se obtiene algún concepto al ser analizadas semánticamente.

8 Combinación de parámetros

Para todos los niveles hemos utilizado un Perceptrón MultiCapa para combinar los diferentes parámetros y obtener una única medida de confianza.

Al nivel de palabra, realizamos una cuantificación de los parámetros de entrada de la red. Para cada una de las entradas consideramos 10 bits, excepto para el comportamiento del modelo de lenguaje en el que únicamente son necesarios 6 bits para codificar las 6 posibles situaciones. La codificación se ha realizado utilizando intervalos de tamaño variable, de forma que se permita una mayor resolución en rangos con mayor cantidad de datos. Esta distribución se ha realizado teniendo en cuenta los datos del conjunto de entrenamiento. La capa oculta está formada por 30 neuronas y la capa de salida por una única neurona. En el entrenamiento se utiliza el algoritmo de retropropagación para calcular los pesos de la red. En esta fase se fija un valor objetivo de 1 para el caso de una palabra correcta, y 0 para los casos de palabras incorrectas (sustituciones e inserciones).

Al realizar la codificación de los parámetros, las entradas, y por tanto el número de pesos a entrenar, se incrementa considerablemente pero puede ofrecer mejores resultados si se utiliza un número de bits elevado y se dispone de suficientes datos para entrenar correctamente los pesos de la red. En el caso de nivel de palabra, disponemos de 3.480 pesos a entrenar y alrededor de 20.000 ejemplos para entrenarlos. Al nivel de concepto y frase, el número de entradas sería mayor puesto que tenemos un mayor número de parámetros, y la cantidad de datos para entrenar es menor: 10.000 ejemplos para el nivel de concepto y alrededor 6.000 para el caso de frase. Por esta razón, para estos dos últimos niveles no realizaremos la codificación de los parámetros y los aplicaremos directamente a las entradas de la red. Con esta solución, es necesario realizar un preproceso de estos parámetros con el fin de limitar su rango dinámico al intervalo [0,1]. En este caso, el preproceso consiste en un reescalado del parámetro, teniendo en cuenta los valores máximo y mínimo obtenidos del conjunto de entrenamiento.

En este punto es importante comentar la siguiente consideración: en los datos de entrenamiento se pueden encontrar ejemplos, que debido a un mal funcionamiento hagan que el valor mínimo o máximo se aleje mucho del resto de valores del parámetro. Este hecho puede provocar que el reescalado haga perder cierta resolución en aquellos rangos con mayor cantidad de ejemplos. Para evitar este problema, los máximos y mínimos utilizados para reescalar los parámetros se obtienen haciendo la media con el 5% de los mayores (cálculo del máximo) o menores (cálculo del mínimo) valores.

9 Resultados obtenidos

A la hora de evaluar las medidas de confianza calcularemos la evolución del Rechazo Correcto (RC) según el Rechazo Incorrecto (RI), al ir modificando el umbral de confianza, y el Error Mínimo de Clasificación obtenido a lo largo de esta variación del umbral. Especial énfasis haremos sobre la tasa de Rechazo Correcto para tasas de Rechazo Incorrecto del 2,5% y del 5,0%. Estos valores corresponden con el margen sobre el que oscilará el punto de trabajo en el que queremos que funcione nuestro sistema.

9.1 Nivel de Palabra

Como se puede observar en la figura 2, los parámetros del modelo de lenguaje (ML) son mejores indicadores de la confianza al nivel de palabra que los obtenidos del proceso de decodificación (PD). Utilizando únicamente los parámetros del modelo de lenguaje, el 42,0% de errores de reconocimiento se detectan para un RI del 5%. Estos resultados son similares a los

obtenidos anteriormente (San-Segundo et al, 2000;Moreau y Jouvet 1999).

Utilizando únicamente parámetros del proceso de decodificación sólo podemos detectar el 28,5% para el mismo RI. Los mejores resultados se obtuvieron combinando todos los parámetros. En este caso podemos rechazar más de la mitad de los errores con un Rechazo Incorrecto del 5%.

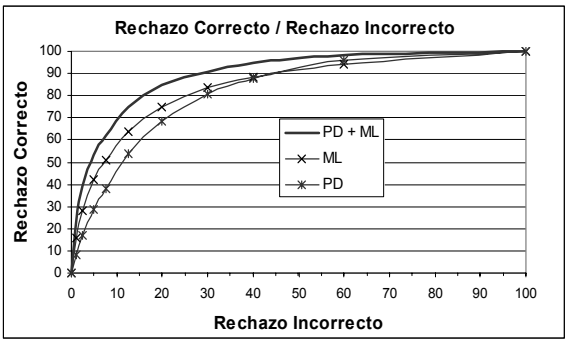

Figura 2: Rechazo Correcto (RC) según el Rechazo Incorrecto (RI) para los parámetros del proceso de decodificación (PD) del modelo de lenguaje (ML) o todos combinados (PD + ML)

Estos parámetros permiten reducir 6,2 puntos el Error de Clasificación (de 19,0% a 12,8%) lo que supone una reducción relativa del 32,6%.

9.2 Nivel de Concepto

Como ya hemos comentado, el objetivo de este nivel es analizar cada concepto de forma independiente y asignarle un valor de confianza entre 0 y 1. En la figura 3 se muestra la evolución del Rechazo Correcto con el Rechazo Incorrecto.

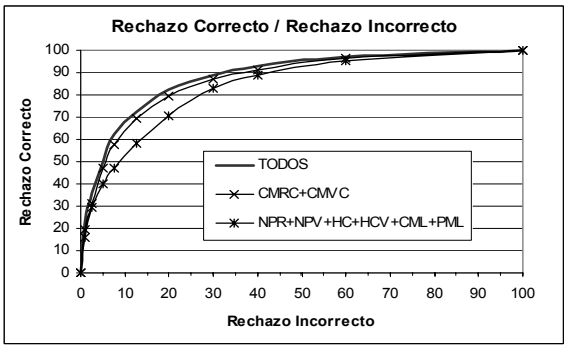

Figura 3: RC vs. RI para las confianzas medias al nivel de palabra (CMRC + CMRV), el resto de parámetros (NPR+NPV+HC+HCV+ CML+PML) y combinándolos todos juntos.

Como se puede deducir de los resultados presentados, las confianzas medias de las palabras en la regla o en el valor del concepto, CMRV y CMVC, son los mejores parámetros considerados en este nivel. Por ejemplo, el 47,1% de los conceptos erróneos fueron detectados para un RI del 5%, valor muy cercano al 50,1% conseguido cuando se combinan todos los parámetros propuestos. De los parámetros que utilizan exclusivamente información del analizador semántico, los que mejores resultados ofrecieron fueron los derivados del modelo de lenguaje conceptual: CML y PML. El Error de Clasificación total se ha reducido de un 16,5% a un 12,0%.

9.3 Nivel de Frase

La figura 4 muestra cómo los parámetros provenientes del analizador semántico son mucho mejores que el parámetro CMP (confianza media de las palabras de la frase) para la cálculo de confianza a nivel de frase. Para un Rechazo Incorrecto del 5% el CMP consigue detectar más del 53% de las frases erróneas pero los parámetros semánticos superan el 68%. Como en situaciones anteriores, los mejores resultados se obtienen cuando se combinan todos los parámetros. En este caso, se detecta más del 76% de las frases para un RI del 5%. Este aumento tan importante al combinar todos los parámetros pone de manifiesto la complementariedad de los mismos. En cuanto a los parámetros considerados, cabe comentar que la Confianza Media de los Conceptos de la frase (CMC) es, junto con el Porcentaje de Palabras Analizadas Semánticamente (PPAS), los dos mejores parámetros.

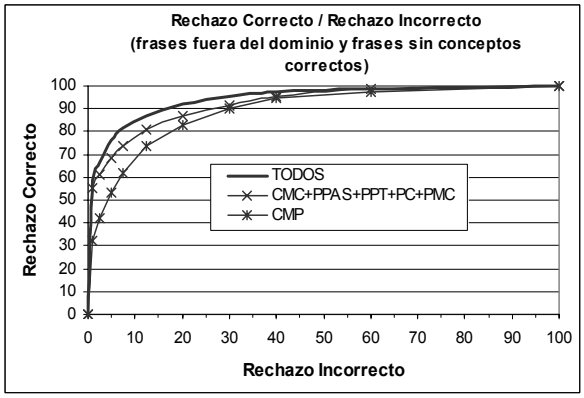

Figura 4: RC vs. RI para la detección de frases fuera del dominio de aplicación y frases sin ningún concepto correcto.

Considerando una distribución inicial en la que el 22,1% de las frases deben ser rechazadas, el poder de discriminación de estos parámetros permite reducir 12,1 puntos el error de clasificación (54,8% relativo).

10 Conclusiones

En este trabajo se ha realizado un estudio importante de diferentes parámetros con el fin de proporcionar medidas de confianza tanto para el sistema de reconocimiento como el sistema de comprensión. De los resultados presentados podemos resumir que considerando como punto de trabajo un Rechazo Incorrecto (RI) del 5%, hemos conseguido rechazar más del 50% de palabras erróneas y conceptos incorrectos, y más del 76% de frases mal interpretadas semánticamente por el sistema.

Al nivel de palabra, los parámetros obtenidos del modelo de lenguaje funcionan mejor para tasas de RI bajas. Combinando los parámetros del proceso de decodificación y del modelo de lenguaje se consiguen resultados bastante mejores que utilizando cada grupo de parámetros de forma independiente, lo que pone de manifiesto la complementariedad de ambas fuentes de información.

En los niveles de concepto y frase, cabe comentar que las medidas obtenidas al nivel de palabra y de concepto respectivamente, son muy útiles para predecir la confianza en niveles superiores.

Agradecimientos

Nuestro más sincero agradecimiento a B.Pellom y W.Ward por permitirnos trabajar con el sistema CU DARPA Communicator.

Bibliografía

Chase, L., 1997. Word and acoustic confidence annotation for large vocabulary speech recognition. Proc. EUROSPEECH, Rodas, Grecia. pp 815-818. 1997.

DARPA Communicator. Página web del proyecto (http://fofoca.mitre.org).

Hazen, T,. Seneff, S., Polifroni, J., 2002. Recognition confidence scoring and its use in speech understanding systems. Computer Speech and Language (2002) 16, 49-67.

Kamppari, S., Hazen, T., 2000, Word and phone level acoustic confidence scoring. Proc. ICASSP, Estambul, Turquía. pp III-1799, III-1802. 2000.

Macías-Guarasa, J., Ferreiros, J., San-Segundo, R., Montero, JM., Pardo, JM., 2000. Acoustical and Lexical Based Confidence Measures for a Very Large Vocabulary Telephone Speech Hypothesis-Verification System. Proc. ICSLP. Pekín, China. Vol. IV, pp. 446-449. 2000.

Moreau, N., Jouvet, D., H., 1999. Use of a confidence measure based on frame level likelihood ratios for the rejection of incorrect data. Proc. EUROSPEECH, Budapest, Hungría. pp 291-294, 1999.

Pellom, B., Ward, W., Sameer Pradhan, 2000. The CU Communicator: An Architecture for Dialogue Systems. Proc. ICSLP, Pekín, China. Vol II. pp723-726. 2000.

San-Segundo, R., Pellom, B., Ward, W., and Pardo, JM., 2000. Confidence measures for dialogue management in the CU communicator system. Proc. ICASSP, Estambul, Turquía. Vol III, pp1237-1240. 2000.

San-Segundo, R., Pellom, B., Hacioglu, K., Ward, W., and Pardo, JM., 2001a. Confidence measures for Spoken Dialogue Systems. Proc. ICASSP, Salt-Lake-City, USA. 2001.

San-Segundo, R., Montero, JM., Ferreiros, J., Córdoba, R., Pardo. JM., 2001b. Designing Confirmation Mechanisms and Error Recover Techniques in a Railway Information System for Spanish. SIGDIAL 2001 WORKSHOP. Sep. 1-2,. Aalborg (Dinamarca), 2001.

Sturm, J., den Os, E., and Boves, L., 1999. Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE System. Proceedings of ESCA Workshop on Interactive Dialogue in MultiModal Systems. Kloter Irsee, Germany, 1-4.

Ward W., Pellom B. 1999. The CU Communicator System. Proc. IEEE Workshop on Automatic speech Recognition and Understanding (ASRU), Keystone Colorado.

Zhang, R., Rudnicky, A., 2001. Word level confidence annotation using combinations of features. Proc. EUROSPEECH. Aalborg, Dinamarca. Vol III, pp 2105-2109, 2001.