Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 229-234 recibido 29-04-2005; aceptado 01-06-2005

Utilización de medidas de confianza en sistemas de comprensión del habla

Valentín Sama Rojo, Javier Ferreiros, Fernando Fernández, Rubén San Segundo, José M. Pardo
Grupo de Tecnología del Habla, Depto. de Ingeniería Electrónica,
ETSI Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria s/n Madrid 28040
{vsama,jfl,efhes,lapiz}@die.upm.es

Resumen: La utilización de medidas de confianza en sistemas de comprensión de habla nos permite tener información sobre la calidad del resultado de nuestro sistema. Utilizando filtros basados en las medidas de confianza del reconocedor, se consiguen mejorar las tasas de *Concept Accuracy*.
Palabras clave: Comprensión, reconocimiento de habla, medidas de confianza

Abstract: The use of confidence measures at speech understanding systems allows us to have information on the quality of our system. Using filters based on the confidence measures of the speech recognizer, we can improve the rates of Concept Accuracy.
Keywords: Understanding, speech recognition, confidence measures

1. Introducción

En este artículo presentamos una estrategia para la utilización de las medidas de confianza que nos proporciona un reconocedor de habla en un sistema de comprensión para extracción de información de las conversaciones de los controladores aéreos.

El uso de medidas de confianza en el módulo de comprensión nos permite discriminar qué resultados pueden ser erróneos y cuales son correctos.

Los resultados que aquí presentamos proceden del proyecto INVOCA (Sama Rojo et al., 2003; Fernández Martínez et al., 2003), que tenía como objetivo desarrollar un sistema de extracción y comprensión de información de las conversaciones de los controladores aéreos en la torre de control.

El sistema está basado en reglas dependientes de contexto con las que se extrae la información relevante de las intervenciones de los controladores.

La información extraída de cada una de las frases se presenta en pares concepto=valor, por ejemplo un indicativo de un avión aparecería en la forma `indicativo=[iberia3456]`; a cada uno de estos pares lo llamaremos *Slot de información* o simplemente *Slot*.

En la torre de control del Aeropuerto de Barajas hay cinco posiciones de control sobre las que se desarrolló este proyecto:

- Arribadas, autoriza el aterrizaje y proporciona instrucciones para esta maniobra
- Autorizaciones, autorización de los planes de vuelo
- Despegues, autoriza el despegue y proporciona instrucciones para esta maniobra
- Norte, control del tráfico rodado en la zona norte del aeropuerto
- Sur, control del tráfico rodado en la zona sur del aeropuerto

Las intervenciones de los controladores pueden ser en español o en inglés, por lo que el sistema cuenta con un módulo de identificación de idioma que permite diferenciar entre ambos idiomas para utilizar los modelos acústicos correspondientes al idioma identificado y posteriormente aplicar a la frase reconocida el módulo de comprensión del idioma reconocido (Fernández et al., 2004).

ISSN: 1135-5948

Hay que resaltar que las expresiones utilizadas por los controladores aéreos están sujetas a una fraseología que indica cómo debe ser cada frase, el orden de los elementos e incluso la pronunciación correcta de determinadas palabras, y a su vez en cada frecuencia nos encontramos con normas diferentes, lo que implica que cada frecuencia debía tener un sistema de comprensión diferente.

Otro aspecto importante es que esta fraseología no suele ser respetada por los usuarios, por lo que no era suficiente con ceñirse a la documentación oficial, y fue necesario realizar un estudio detallado de cada frecuencia, realizándose grabaciones y una base de datos compuesta por las intervenciones del controlador en audio y su transcripción en formato `SAM`.

Para evaluar el módulo de comprensión del sistema se construyó manualmente una referencia para gran parte de los ficheros de nuestra base de datos partiendo de la salida del módulo de comprensión, en estas referencias estaba el resultado correcto para cada una de las frases, de modo que luego se podría calcular el *Concept Accuracy* para cada una de las frecuencias. Posteriormente en el apartado 5 se tratará en más profundidad este tema.

En el sistema que se desarrolló para el proyecto INVOCA no se utilizaban medidas de confianza, y estamos experimentando su utilización en diferentes niveles. Una de las posibilidades que ya se ha utilizado en otros sistemas con cierto éxito es el uso de medidas de confianza en el módulo de comprensión (García et al., 2003; Hazen, Seneff, y Polifroni, 2002; Pao, Schmid, y Glass, 1998; San-Segundo et al., 2004) y veremos que en el sistema INVOCA su utilización ha aportado mejoras en *Concept Accuracy*.

Como mejora del sistema original se han introducido una serie de filtros de formato que comprueban que determinados *slots* tienen el formato correcto que le correspondería a dicho *slot*; y otra de las mejoras son los filtros de confianza que eliminan aquellos resultados que presenten un nivel de confianza por debajo de un determinado umbral, a continuación presentaremos todas estas mejoras y las compararemos con el sistema original.

2. Descripción del sistema

2.1. Front-End

El sistema consta de un primer módulo o front-end que convierte la señal acústica en un conjunto de vectores. Este front-end está compuesto a su vez por un detector, cuyo cometido es la detección de voz / no voz a la entrada del sistema, y el parametrizador, que lleva a cabo la parametrización de la información acústica segmentada.

2.2. Módulo de reconocimiento

La salida del front-end pasa al módulo de reconocimiento compuesto por dos reconocedores, uno para inglés y otro para castellano. Se obtienen las frases reconocidas en el idioma correspondiente en ambos reconocedores y se pasan como entradas al módulo de detección de idioma.

2.3. Módulo de detección de idioma

Este módulo decide el idioma al que corresponde la frase que está siendo procesada. La decisión puede tomarse aplicando modelos de lenguaje en base a medidas de perplejidad sobre los resultados de ambos reconocedores. En nuestro caso, debido al gran peso dado al modelado de lenguaje durante el proceso de reconocimiento, la decisión puede tomarse directamente en base a la diferencia de *scores* entre ambos reconocedores.

2.4. Módulo de comprensión

Para este módulo se diseñaron una serie de primitivas que nos permiten escribir reglas dependientes de contexto para extraer la información relevante de las intervenciones de los locutores. Se utilizan diccionarios etiquetados semánticamente en función de la información que se desea extraer y a las palabras que no aportan información a la tarea, se les asigna la categoría "basura", al igual que a las palabras fuera de vocabulario que pudiesen aparecer, aunque esto pueda suponer una posible perdida de información. A su vez una palabra puede tener varias categorías incluyendo entre ellas la categoría "basura".

Un grupo de reglas se ejecutarán antes de eliminar las *basuras*, de modo que puede resultar relativamente sencillo separar muchos

conceptos presentes en la frase. Y posteriormente se ejecutará otro grupo de reglas, tras eliminar las palabras categorizadas como "basura", con el fin de resolver aquellos conceptos que no hayan sido resueltos en la fase anterior, o para la utilización de elementos construidos o extraídos durante la fase en la que las *basuras* estaban presentes.

Consideramos importante la utilización de los elementos categorizados como "basura" dentro de nuestras reglas, ya que a pesar de que no contienen información semántica de interés para la aplicación, pueden ser útiles en determinadas situaciones para aislar o extraer la información que realmente nos interesa. Estos elementos "basura" pueden estar separando diferentes conceptos, diferentes números, etc; por lo que podremos utilizarlos a modo de separador entre los diferentes conceptos que pueda haber en una intervención del locutor.

2.4.1. Proceso de ejecución de las reglas

El sistema de comprensión recibe una frase y se siguen una serie de pasos para procesarla, se podrían resumir en los siguientes pasos:

1. preprocesado de la secuencia de palabras procedente del reconocedor (incluye asignación de categorías, procesado de dígitos, eliminación de posibles interjecciones, etc.), este paso es común para todas las reglas
2. reglas antes de eliminar las palabras que tienen la categoría "basura"
3. eliminación de las palabras con categoría "basura"
4. reglas situadas tras la eliminación de *basuras*
5. escritura de los conceptos resultantes del proceso de comprensión

Las filtros de formato y los filtros de confianza que introducimos como mejora se sitúan antes del paso 5 de modo que ya han terminado todas las reglas de comprensión y se dispone del resultado completo.

3. Utilización de medidas de confianza

Como comentábamos, estamos trabajando en la integración y utilización de medidas de confianza a nivel de la comprensión, siendo posible determinar la confianza de los elementos utilizados en las reglas de comprensión, así como la confianza del resultado de cada regla. Posteriormente en base a estos valores se pueden detectar posibles errores y el gestor de diálogo podría realizar diferentes acciones como volver a preguntar un elemento en el que no se confía, alterar el curso del diálogo, etc.

En INVOCA, al tratarse de un sistema de extracción de información, hemos optado por la utilización de dos tipos de filtros, uno eli-mina el *slot* menos fiable si hay dos *slots* referidos al mismo concepto, y el otro elimina los *slots* que tengan un valor de confianza menor que un determinado umbral, estos filtros los trataremos más en profundidad en el apartado 4.2.

3.1. Medidas de confianza utilizadas

Nuestro reconocedor nos proporciona diferentes medidas de confianza, hemos utilizado la pureza de palabra, que es la fracción de hipótesis que contienen la palabra en la misma posición, frente al número total de hipótesis, es decir el número de veces que una misma palabra aparece en la misma posición dentro de todas las hipótesis obtenidas en la fase de reconocimiento.

Posteriormente la pureza de palabra pasa a una red neuronal que integra otras medidas de confianza y como salida tendremos el valor de confianza que utilizaremos.

Las medidas que se integran en la red neuronal son:

- pureza de palabra
- *score* de palabra, es la puntuación acústica que presenta la palabra. Dicho *score* surge de la aplicación del modelo acústico, sobre el segmento de datos de voz que se corresponde a la palabra.
- *score* medio de palabra, es la puntuación acústica dividida entre el número de tramas que componen la palabra.

- varianza del *score*, es la varianza de las puntuaciones de trama, dentro de la palabra.
- *score* de la peor trama, es la puntuación más alta entre las tramas pertenecientes a la palabra.
- *score* de la mejor trama, es la puntuación más baja entre las tramas pertenecientes a la palabra.
- diferencia media del mejor *score* de trama, es la media de la diferencia entre la puntuación de cada trama, y el menor *score* de trama.
- pureza de *N-best*, es el porcentaje de hipótesis en la lista *N-best*, que tienen esa palabra en la misma posición.
- *score* de *N-best*, es similar a la pureza, pero empleando los *scores* de las hipótesis en la lista *N-best*.
- pureza de grafo, la pureza obtenida a partir del grafo.

También se utilizará la confianza a nivel de frase para ponderar las confianzas de palabra, estos cálculos se realizan dentro de las primitivas del módulo de comprensión.

3.2. Integración de las medidas en las primitivas

De cada primitiva que se utilizará para escribir las reglas del módulo de comprensión se obtiene la confianza del elemento resultante de la regla, calculando la media tomando el valor de cada elemento del contexto presente en la regla.

Posteriormente este valor puede ser utilizado para realizar otro cálculo si dicho elemento es utilizado en otra regla.

Al ejecutarse las reglas del módulo de comprensión obtendremos un marco semántico con una serie de slots de información y cada uno de estos slots tendrá un valor de confianza situado entre 0 y 1.

De este modo todo resultado del proceso de comprensión tiene un valor de confianza que lo caracteriza como más o menos fiable.

Es importante remarcar que no es necesario volver a escribir las reglas o modificarlas, ya que los cambios realizados sólo afectan a la primitiva de cada tipo de regla.

4. Mejoras introducidas en el sistema

4.1. Filtros de formato

Estos filtros limitan la aparición de resultados de comprensión que tengan una forma que no se corresponda con la que debería tener determinado concepto.

En INVOCA teníamos que trabajar con frecuencias de radio y en la mayoría de las ocasiones los locutores usan abreviaturas para ahorrar tiempo, de modo que los conceptos referentes a las frecuencias de radio debían ser convertidos a los formatos completos de dichas frecuencias, por ejemplo:

El locutor podía decir *veintiuno ochenta y cinco* para referirse a *ciento veintiuno punto ochenta y cinco*, de modo que las reglas de comprensión transforman los números que se refieren a una frecuencia de radio formateándolo de modo que aparezca con la forma *121.85*.

Posteriormente, justo después de que se han ejecutado todas las reglas de comprensión, un filtro comprueba que estos elementos tienen el formato adecuado y que se encuentran dentro del rango de frecuencias de radio del aeropuerto.

4.2. Filtros basados en la confianza

Hemos podido comprobar que al aplicar filtros basados en los valores de confianza resultantes de las reglas de comprensión se consigue una mejora en cuanto a la calidad del resultado ya que se eliminan conceptos erróneos que están penalizando la evaluación del sistema.

Principalmente hemos trabajado con dos tipos de filtros que pasamos a tratar con algo más de detalle.

4.2.1. Eliminación del menos fiable

Este filtro elimina el *slot* con un valor de confianza más bajo, es decir el menos fiable, en el caso de que aparezcan como resultado de la comprensión dos *slots* referentes al mismo concepto.

En este sistema no existe la posibilidad de que determinados elementos aparezcan más de una vez a la hora de obtener el resultado final de la comprensión, por ejemplo no suele ser normal que aparezcan dos indicativos de diferentes aeronaves en la misma frase, de modo que al aplicar estos filtros eliminamos el resultado menos fiable.

4.2.2. Eliminación según umbral

El otro tipo de filtros que hemos utilizado elimina aquellos elementos cuyo valor de confianza sea menor que un determinado umbral.

En nuestro caso hemos establecido manualmente el umbral en 0.1 de modo que estamos seguros que sólo eliminaremos aquellos *slots* en los que no es conveniente fiarse dada su baja confianza.

5. Método de Evaluación

Para evaluar el sistema de comprensión utilizaremos el *Concept Accuracy* (Boros et al., 1996), es similar al *Word Accuracy* normalmente utilizado para evaluar sistemas de reconocimiento de habla, donde se compara el resultado del sistema con una referencia con el resultado correcto.

En algunos casos nos encontramos con el problema de que teníamos frases de las que no se podía generar una referencia, por ejemplo por falta de contexto no se podía asegurar que ese resultado era el correcto, obviamente dicha frase sin referencia no se puede evaluar.

Nuestros sistemas de comprensión devuelven como resultado pares concepto=valor, y cada uno de estos pares es un *slot* de información. Para calcular el *Concept Accuracy* utilizaremos las inserciones, borrados y sustituciones de *slots* pero dentro de cada *slot* tendremos que tener en cuenta si el concepto es correcto y si el valor es correcto. El número total de sustituciones será la suma de las sustituciones de concepto (cuando el valor es correcto pero el concepto no lo es) y las sustituciones de valor (cuando el concepto es acertado pero el valor es erróneo) que se produzcan en los *slots* resultantes del proceso de comprensión.

Para calcular el *Concept Accuracy* utilizamos:

$$CA = 100(1 - \frac{SU_S + SU_I + SU_D}{SU})\,\% \quad (1)$$

Donde SU_S son el número de sustituciones en los conceptos, SU_I son el número de inserciones, SU_D es el número de borrados y finalmente SU el números total de conceptos en nuestra referencia.

6. Resultados

En estos experimentos hemos implementado tres sistemas diferentes:

- El sistema base. Lo denominaremos A
- Al sistema base se le han añadido filtros de formato. Lo denominaremos B
- Se le añaden filtros basados en las medidas de confianza al sistema B . Lo denominaremos C

En los resultados medios presentados en la tabla 1 se puede observar que el uso de ambos tipos de filtro supone una mejora considerable respecto al sistema base.

Castellano	**CA Medio**
A	63.64 %
B	65.83 %
C	67.60 %
Inglés	**CA Medio**
A	44.97 %
B	48.87 %
C	52.47 %

Cuadro 1: Resultados Concept Accuracy Medio

Para entender o explicar dichos cambios es necesario tener en cuenta los datos de borrados, inserciones y sustituciones totales de *slots* que presentamos en la tabla 2.

Si se observa la tabla 2 podemos comprobar como el número de inserciones y sustituciones disminuye al introducir filtros de formato, y al añadir filtros de confianza se produce otra reducción, al mismo tiempo aumenta el número de borrados. A pesar de estos

Castellano	**Ins**	**Borr**	**Sust**
A	335	373	606
B	255	413	567
C	180	428	563
Inglés	**Ins**	**Borr**	**Sust**
A	279	201	423
B	215	232	392
C	151	242	387

Cuadro 2: Insersiones, Borrados y Sustituciones de slots totales

borrados la tasa de *Concept Accuracy* aumenta, probablemente se debe a que se eliminan slots que eran correctos pero que por la razón que sea tienen una confianza baja, y por el otro lado, se dejan de producir una cantidad considerable de inserciones o sustituciones erróneas que se producían con el sistema A, ya que no tenía restricción alguna.

7. Conclusiones

La utilización de medidas de confianza en sistemas de comprensión es un buen método para detectar posibles errores que deberán ser tratados por el gestor de diálogo o por otro módulo. La eliminación de los *slots* con bajo nivel de confianza nos ha reportado mejoras en el comportamiento del sistema, concretamente en la tasa de *Concept Accuracy* aumentando el número de borrados pero disminuyendo el número de inserciones y sustituciones sin que esto suponga un empeoramiento del comportamiento del sistema.

Utilizar estas medidas de confianza proporciona mucha información para poder tomar decisiones que afectarán al proceso del diálogo con el usuario, de modo que creemos que es una buena vía para mejorar los sistemas de comprensión de habla.

Bibliografía

Boros, Manuela, Wieland Eckert, Florian Gallwitz, Günther Görz, Gerhard Hanrieder, y Heinrich Niemann. 1996. Towards understanding spontaneous speech: Word accuracy vs. Concept accuracy. En *Proceedings of International Conference on Spoken Language (ICSLP96)*, volumen 2, páginas 1005–1008, Philadelphia, October. IEEE Computer Society Press.

Fernández, Fernando, Ricardo De Córdoba, Javier Ferreiros, Valentín Sama, L. F. D'Haro, y Javier Macías-Guarasa. 2004. Language identification techniques based on full recognition in an air traffic control task. En *Proceedings of International Conference on Spoken Language (ICSLP2004)*, páginas II–1565–1568, Jeju, October.

Fernández Martínez, Fernando, Valentín Sama Rojo, Javier Ferreiros López, Javier Macias-Guarasa, Ricardo De Córdoba, J. M. Montero Martínez, J. Colas Pasamontes, E. Campos Palarea, y J. M Pardo Muñoz. 2003. Demostración del sistema de comprensión de comunicaciones habladas para control de tráfico aéreo del proyecto INVOCA. *Procesamiento del Lenguaje Natural*, 31:337–338.

García, F., L.F. Hurtado, E. Sanchis, y E. Segarra. 2003. The incorporation of confidence measures to language understanding. En Pavel Mautner Vaclav Matousek, editor, *Proceedings of the Sixth Conference on Text Speech and Dialogue (TSD)*, LNAI 2807, páginas 165–172. Springer, September.

Hazen, T. J., S. Seneff, y J. Polifroni. 2002. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*, 16:49–67.

Pao, C., P. Schmid, y J. Glass. 1998. Confidence scoring for speech understanding systems. En *Proceedings of the 5th. International Conference of Spoken Language Processing (ICSLP'98)*, páginas 815–818.

Sama Rojo, Valentín, Fernando Fernández Martínez, Javier Ferreiros López, Javier Macias-Guarasa, Ricardo De Córdoba, J. M. Montero Martínez, J. Colas Pasamontes, E. Campos Palarea, y J. M Pardo Muñoz. 2003. Sistema de comprensión de comunicaciones habladas para el control de tráfico aéreo del proyecto invoca. *Procesamiento del Lenguaje Natural*, (31):313–314.

San-Segundo, Rubén, Javier Macías-Guarasa, J. M. Montero, Javier Ferreiros, Ricardo De Córdoba, y J.M. Pardo. 2004. Medidas de confianza en sistemas de diálogo. *Procesamiento del Lenguaje Natural*, 33:95–102.