Sistema de Reconocimiento Automático de Habla basado en Máxima Entropía

Claudio F. Estienne
Facultad de Ingeniería Inst. de Ingeniería Biomédica, Universidad de Buenos Aires
Buenos Aires, Argentina
cestien@fi.uba.ar

and

Alberto Sanchis
Depto. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia
Valencia, España
asanchis@iti.upv.es

Abstract

In this work we present the application of confidence measures in automatic speech recognition in order to obtain a reliable measure of recognized words performed by a speech recognizer. This measure is then used to detect wrong words in order to accept or reject the whole utterance. The main goal of such techniques is to increase the reliability of automatic speech recognition systems. This work focus on the implementation of a word verification system based on scores which give truthful information in testing recognized words. Those scores are embedded inside a statistical classifier based on the maximum entropy paradigm. Maximum entropy models have the ability to combine different sources of information. This property is applied to combine mentioned scores with other words properties. As a consequence, significant improvement on the reliability of the verification system is obtained. Evaluation of the whole system is performed on a automatic voice driven hotel book system.

Keywords: maximum entropy, speech processing, languaje modelling, confidence measures.

Resumen

El presente trabajo describe la aplicación de técnicas de medida de confianza en reconocimiento automático del habla. Las mismas tienen por objeto medir la confiabilidad de las palabras reconocidas por el sistema de reconocimiento y detectar aquellas que puedan tener errores con el fin de aceptar como válida, o rechazar una frase reconocida. El objetivo final de dichas técnicas, es mejorar la confiabilidad de los sistemas de reconocimiento automático del habla. Concretamente el trabajo se centra en la implementación de un sistema de verificación de palabras basado en características que aportan información útil para la corrección de palabras reconocidas. Dichas características son utilizadas dentro de un clasificador estadístico basado en el modelo de máxima entropía. La posibilidad de combinar diferentes fuentes de información que permiten los modelos de máxima entropía es utilizada en este trabajo para combinar las características mencionadas con otras propiedades de las palabras, logrando un aumento significativo en el rendimiento del sistema de verificación. La evaluación del sistema completo se realiza en el marco de un sistema automático de reserva y consulta de disponibilidades en un hotel por medio de la voz.

Palabras claves: máxima entropía, procesamiento de habla, modelos de lenguaje, medidas de confianza.

1. INTRODUCCION

La investigación en tecnologías del habla data de los años 50 y desde entonces ha sido un área activa de estudio. El presente trabajo se enfoca en una de las áreas del procesamiento del habla, el reconocimiento automático reconocimiento del habla. Es decir, en la extracción de la secuencia de palabras que componen una frase a partir de la emisión acústica de dicha frase. El esquema propuesto consta básicamente de dos etapas, el sistema de reconocimiento automático del habla (SRAH) propiamente dicho, el cual emite la frase mas probable que corresponde a una dada emisión acústica (llamada hipótesis), y el sistema de verificación de hipótesis (VH) a la salida del sistema de reconocimiento. Este último tiene por objeto asignar a cada una de las palabras que componen la salida, una medida de confianza que indique con que grado de certeza se puede aceptar que la misma fue correctamente reconocida. Esta última etapa agrega un importante grado de confiabilidad al sistema total, ya que un potencial usuario del mismo, en base a ésta puede aceptar la hipótesis del SRAH como verdadera, o pedir una repetición de la emisión a fin de que este genere una hipótesis mas confiable.

Varios esquemas de VH así como varias medidas de confianza han sido propuestas en el pasado [1], [2], y la determinación de un esquema de verificación de hipótesis óptimo es todavía objeto de investigación. El objetivo del presente trabajo es el planteo de un modelo estadístico de verificación de hipótesis basado en el método de máxima entropía. La modelización estadística mediante máxima entropía ha sido utilizada en diversos campos de la ciencia incluido el procesamiento del habla, [3]. La gran ventaja de éste método es que permite la incorporación de diferentes fuentes de información a un mismo modelo, sin asumir ninguna clase de hipótesis a priori exceptuando las propias mediciones experimentales. Con el fin de evaluar y contrastar el sistema de verificación de hipótesis propuesto, utilizaremos el SRAH y el VH usado en trabajos anteriores por uno de los autores [4], [5]. También haremos las evaluaciones de perfomance usando la misma base de datos de entrenamiento y testeo utilizadas en dicho sistema.

El resto del trabajo se divide como sigue: En la sección 2 se describe brevemente la aproximación estadística al reconocimiento automático del habla y a la verificación de hipótesis usada por los sistemas actuales. En la sección 3 se describe un verificador de hipótesis basado en características del habla, el cual servirá de base para el verificador basado en máxima entropía. En la sección 4 se describe brevemente el principio de ME y el modelo de verificación de hipótesis de máxima entropía, que constituye el principal aporte de éste trabajo. En la sección 5 se muestran los resultados experimentales así como la unidad de medida usada para evaluar dichos resultados y el corpus de datos usados . También se discuten los resultados obtenidos. Finalmente en la sección 6 se extraen las correspondientes conclusiones del trabajo.

2. RECONOCIMIENTO AUTOMATICO DEL HABLA

2.1. El problema del reconocimiento del habla

El problema de reconocimiento del habla es visto generalmente como un problema de codificación en el cual se asume que la señal de habla acústica lleva implícito un mensaje codificado que consiste en una secuencia de símbolos. Dichos símbolos pueden representar fonos (es decir los sonidos que conforman el habla), sílabas, palabras o cualquier otra unidad fonética. La resolución del problema consistiría entonces en encontrar a partir de la señal acústica dicha secuencia de símbolos. A fin de disminuir la enorme variabilidad de la señal acústica y permitir un tratamiento estadístico, la misma es parametrizada convirtiéndola en una secuencia de vectores equiespaciados llamados vectores acústicos [6]. El reconocimiento será entonces un problema de decodificación en el cual se busca aquella

secuencia de símbolos $\hat{W}$ tal que:

$$\hat{W} = \underset{W}{argmax}\ P(W/Y) \tag{1}$$

donde $Y = y_1 \dots y_T$ es la secuencia de vectores acústicos también llamada secuencia de observación y W la secuencia de símbolos correspondiente a dicha secuencia de observación. En un sistema de reconocimiento típico la secuencia de símbolos será una secuencia de palabras es decir una frase, y se tratará de encontrar sobre todas las frases posibles aquella que maximice la probabilidad de observar la secuencia de vectores acústicos Y. Tendremos entonces dos problemas claramente diferenciados a resolver. Por un lado la determinación del $P(W/Y)$ es decir, la determinación del modelo estadístico. Esto se realiza a partir de grandes bases de datos de emisiones de hablantes las cuales, previamente convertidas en vectores acústicos, son usadas para el entrenamiento de modelos estadísticos. El modelo estadístico mas aceptado en las últimas décadas es el de los modelos ocultos de Markov (HMM), [6], el cual con diversas variantes ha probado ser por lejos el mas eficiente modelo en procesamiento del habla. El segundo problema es el reconocimiento propiamente dicho, en el cual deberá hallarse la frase óptima $\hat{W}$ que maximiza la probabilidad del modelo. Este proceso se realiza mediante la técnica de búsqueda de Viterbi [6] que permite encontrar en forma muy eficiente la frase mas probable de acuerdo a la secuencia de vectores acústicos observados.

2.2. Verificación de hipótesis

El grado de confiabilidad de los sistemas actuales aun no suele ser suficiente para muchas aplicaciones prácticas, sobre todo en condiciones adversas de ruido, variedad de hablantes, etc.. Por dicho motivo se suele implementar una segunda etapa a la salida del reconocedor de habla cuya función es eliminar aquellas frases que por alguna razón se cree que fueron reconocidas incorrectamente. Dicha etapa constituye lo que se llama un verificador de hipótesis. La hipótesis está representada por la secuencia de palabras que el sistema de reconocimiento da como mas probable. Esta podría ser una sola (la mejor hipótesis) o varias (las N mejores hipótesis) (figura 1). El objetivo del verificador de hipótesis es asignar una "medida de confianza" a cada una de las unidades (usualmente palabras) que forman parte de la hipótesis obtenida tras el proceso de reconocimiento. En general la medida de confianza se puede definir como una función que mide el grado de verosimilitud entre la observación acústica y el modelo proporcionado por el reconocedor. En la figura 1 se muestra el esquema completo de un sistema de reconocimiento de habla y el verificador de hipótesis sobre el que basaremos el presente trabajo. En este caso el módulo de verificación de hipótesis utiliza una medida de confianza binaria, asignando a cada palabra w_i (salida del SARH) la etiqueta c_i *correcta* o *incorrecta*.

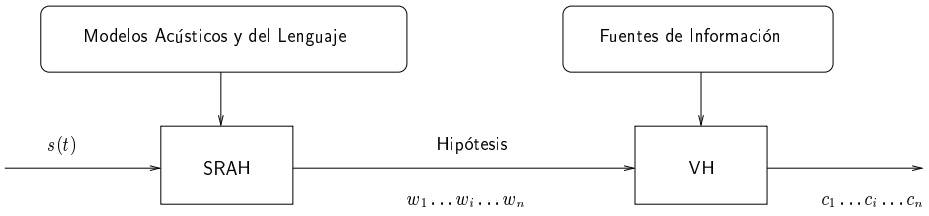

Figura 1: Sistema completo de reconocimiento y verificación de hipótesis

3. SISTEMA DE VERIFICACION DE HIPOTESIS

La implementación del verificador de hipótesis de la figura 1 requiere la resolución de dos problemas. En primer lugar se deben definir las fuentes de información, llamadas características o scores, que aporten información útil sobre la corrección de las palabras reconocidas. En segundo término se debe definir un modelo probabilístico que obtenga resultados satisfactorios en la estimación de la medida de confianza. En el presente trabajo nos centraremos en el segundo problema, es decir en la obtención del modelo probabilístico, para lo cual usaremos la técnica de máxima entropía. No abordaremos el primero de los problemas y nos limitaremos a usar dos de las mejores características que fueron usadas en [4] y [5], llamadas estabilidad acústica (AS), y Word Trellis Stability (WTS). Dichas características han probado aportar información útil en la detección de palabras erróneas. Antes de definir el modelo probabilístico de máxima entropía describiremos el modelo probabilístico usado en [4] y [5], ya que los problemas asociados al mismo también aparecerán en el modelo de máxima entropía.

3.1. Modelo probabilístico del verificador de hipótesis

Supongamos para simplificar que solo tenemos una única característica x. El objetivo sería encontrar la distribución $P(c/w,x)$. Donde $c=0$ significa que el VH determinó que en función del valor de x la palabra reconocida w tiene una probabilidad $P(c/x,w)$ de ser *incorrecta*. Análogamente, si $c=1$ significa que tiene una probabilidad $P(c/x,w)$ de ser *correcta*. Asumiendo que x es una variable discreta la probabilidad puede ser estimada como: $P(c/x,w)=\frac{N(c,x,w)}{N(x,w)}$, donde N representa la cantidad de veces que ocurrieron los triples (c,x,w) y los pares (x,w) en el conjunto de entrenamiento. Este estimador de probabilidad tiene dos problemas conocidos [7], [8]. En primer lugar puede ocurrir que no se tengan ejemplos de muchos triples (c,x,w) en el conjunto de entrenamiento, por lo que habrá un gran número de estimaciones que dan probabilidad nula. Esto no es correcto ya que un triple (c,x,w) puede no aparecer en el conjunto de entrenamiento y sin embargo ocurrir en una emisión real. En segundo lugar, aun teniendo algunos ejemplos estos pueden ser muy pocos, por lo que se tendrá una mala estimación. Ambos problemas son muy conocidos en el área de modelos de lenguaje y las técnicas de resolución se conocen como suavizado de modelos [9], [8]. Una de dichas técnicas es la de descuento absoluto [8]. La idea es ganar una masa de probabilidad que se pueda distribuir entre los eventos no vistos a base de descontar una constante $0<b<1$ a cada evento visto en el conjunto de entrenamiento. La masa de probabilidad que se gana es luego distribuida entre los eventos no vistos siguiendo otra distribución de probabilidad de menor orden. En caso del verificador de hipótesis $P(c/x,w)$ se podría calcular siguiendo esta técnica del siguiente modo:

$$P(c/x,w)=\begin{cases}\dfrac{N(c,x,w)-b}{N(x,w)} & si\ \ N(c,x,w)>k\\ \gamma(x,w)P(c/x) & si\ \ N(c,x,w)\leq k\end{cases} \tag{2}$$

k es un valor umbral que se fija empíricamente. La probabilidad $P(c/x)$ se calcula como: $P(c/x)=\frac{N(c,x)}{N(x)}$. El factor $\gamma(x,w)$ es un factor de normalización de modo que $P(c/x)$ pueda ser una función de probabilidad. Para el caso en que $N(c,w,x)<k\ \ \forall c$, no habrá masa de probabilidad a descontar por lo que directamente se asume:

$$P(c/x,w)=P(c/x) \tag{3}$$

En el caso de tener mas de una característica tendremos un vector de dimensión D $\mathbf{x}=(x_1,\cdots,x_D)$. y se deberá estimar $P(c/\mathbf{x},w)$.

El modelo también puede ser implementado mediante la regla de Bayes:

$$P(c/\mathbf{x}, w) = \frac{P(\mathbf{x}/c, w)P(c/w)}{\sum_{c'} P(\mathbf{x}/c', w)P(c'/w)} \tag{4}$$

Donde en lugar de plantear un único modelo $P(c/\mathbf{x})$ se modelizan $P(\mathbf{x}/c, w)$ y (c/w). En este caso se pueden aplicar suavizados sobre cada modelo por separado, lo cual da lugar a un modelo total mas preciso.

4. VERIFICADOR DE HIPOTESIS DE MAXIMA ENTROPIA

4.1. El principio de Máxima Entropía

4.1.1. Formulación general

El principio de Máxima Entropía (ME) fue formulado por Jaynes [10] en el marco de la mecánica estadística, y desde entonces fue aplicado en gran cantidad de áreas científicas y de ingeniería, incluidas varias especialidades del procesamiento del habla [11], [3], [12]. Supongamos una variable aleatoria x que puede adoptar valores $(x_1, x_2, \cdots, x_n)$ (n puede ser finito o infinito). También se dispone de la media de un conjunto de funciones $f_1(x),\ f_2(x), \cdots, f_m(x)$ (con $m < n$). Dichas medias vienen dadas como resultado de mediciones experimentales por los números $F_1,\ F_2, \cdots, F_m$. El problema es encontrar la distribución de probabilidades que satisfaga:

$$\sum_{x=x_1}^{x_n} P(x) f_k(x) = F_k, \quad k = 1, 2, \cdots, m$$

y que maximice la entropía de la distribución $P(x)$: $S = -\sum_{x=x_1}^{x_n} P(x)\log(P(x))$

Usando multiplicadores de Lagrange es fácil mostrar ([10]) que la distribución que satisface dichos requerimientos tiene la forma:

$$P(x) = \frac{e^{\sum_{i=1}^{m} \lambda_i f_i(x)}}{Z}$$

con: $Z = \sum_{y=x_1}^{x_n} e^{\sum_{i=1}^{m} \lambda_i f_i(y)}$. Donde los λ_i son los parámetros de la distribución. Existe una técnica muy eficiente para hallar dichos parámetros llamada algoritmo GIS [3], para un conjunto de funciones de restricción f_k y sus correspondientes medias F_k (con $k = 1, \cdots, m$) . De ésta manera el problema quedaría completamente formulado.

4.1.2. Máxima entropía en distribuciones condicionales

Es posible formular el problema en términos de máxima entropía en distribuciones condicionales. Supongamos que x e y son dos variables aleatorias discretas, y se desea hallar $P(y/x)$ la probabilidad condicional de y dado x de máxima entropía para un conjunto de restricciones dado. Es fácil probar que la distribución de máxima entropía tendrá la forma [3]:

$$P(y/x) = \frac{e^{\sum_i \lambda_i f_i(x,y)}}{Z(x)} \tag{5}$$

donde: $Z(x) = \sum_{y'} e^{\sum_i \lambda_i f_i(x,y')}$. Con las ecuaciones de restricción dadas por:

$$\sum_{x,y} P(x)P(y/x)f_k(y,x) = F_k \quad k = 1, \cdots, m \tag{6}$$

En general la estimación de $P(x)$ se aproxima mediante: $P(x) = \frac{N(x)}{N}$, donde N es la cantidad total de datos de entrenamiento.

4.1.3. Funciones de restricción para el caso de conteos

Si se quieren introducir conteos en un modelo de máxima entropía se pueden definir funciones de restricción binarias [11]. Supongamos por ejemplo que el resultado de un experimento arroja que $N(x_3, y_7) = 39$ (es decir, que el evento x_3, y_7 ocurrió 39 veces). En ese caso se podría definir una función de restricción:

$$f_{x_3,y_7}(x,y) = \begin{cases} 1 & si \;\; x = x_3, \;\; y = y_7 \\ 0 & otro\, caso \end{cases}$$

En ese caso la ecuación (6) se podría expresar como una probabilidad empírica $\hat{P}(x,y) = \frac{N(x,y)}{N}$ cuya media respecto de la función de restricción es: $F_{x_3,y_7} = \sum_{x,y} \hat{P}(x,y) f_{x_3,y_7}(x,y) = 39$. Análogamente, si otro conteo arroja $N(x_{56}) = 1230$ se podrá definir:

$$f_{x_{56}} = \begin{cases} 1 & si \;\; x = x_{56} \\ 0 & otro\, caso \end{cases}$$

con $F_{x_{56}} = \sum_{x,y} \hat{P}(x,y) f_{x_{56}}(x,y) = \sum_y \hat{P}(x_{56}, y) = 1230$. Es factible de éste modo introducir en el modelo de máxima entropía cualquier conteo de eventos que se considere relevante al modelo. Si la probabilidad empírica $\hat{P}(x,y)$ se define usando la ecuación (2) se estará introduciendo la técnica de suavizado de descuento absoluto en el modelo de máxima entropía. De esta manera es posible combinar en el modelo diferentes conteos y tipos de suavizado con solo definir adecuadamente las funciones de restricción.

4.2. Verificación de hipótesis mediante máxima entropía

Como vimos en la sección anterior, el modelo probabilístico utilizado, ya sea aplicando la regla de Bayes, o calculando directamente la probabilidad a posteriori $P(c/x,w)$, se determina en base al conteo de los datos de entrenamiento, suavizando con la técnica de descuento absoluto. Siguiendo esta misma línea, impondremos al modelo de máxima entropía restricciones basadas en la misma clase de conteos. Esta aproximación fue utilizada con éxito en varias aplicaciones del principio de máxima entropía a modelos de lenguaje [12], [11], por lo que analizaremos su aplicación en nuestro verificador de hipótesis.

El modelo de máxima entropía para la distribución $P(c/x,w)$ tendrá la forma:

$$P(c/x,w) = \frac{e^{\sum_i \lambda_i f_i(c,x,w)}}{Z(x,w)} \tag{7}$$

donde

$$Z(x,w) = \sum_{c'} e^{\sum_i \lambda_i f_i(c',x,w)} \tag{8}$$

Las ecuaciones de restricción vendrán dadas por:

$$\sum_{c,x,w} P(x,w)P(c/w,x)f_k(c,x,w) = F_k \quad k = 1, \cdots, m \tag{9}$$

Donde $P(x,w) = \frac{N(x,w)}{N}$, siendo N es la cantidad total de datos de entrenamiento.

En el caso de tener mas de una característica x se deberá reemplazar por el vector de características $\mathbf{x}$. Como se dijo, en el presente trabajo solo usaremos las características: estabilidad acústica (AS) que llamaremos x_1, y Word Trellis Stability (WTS) que llamaremos x_2. Vamos a definir tres modelos de verificación de hipótesis: los modelos de una característica $P_1(c/x_1,w)$ y $P_2(c/x_2,w)$ y el modelo combinado $P(c/x_1,x_2,w)$. El objetivo es verificar si las dos características AS y WTS son estadísticamente independientes. Si esto fuera así el modelo combinado se podría calcular a partir de P_1 y P_2. Además, a diferencia del modelo planteado en [4] y [5], solo modelizaremos probabilidades a posteriori, es decir, no se aplicará la regla de Bayes. Para cada uno de los tres modelos definiremos dos posibles conjuntos de funciones de restricción. El primer conjunto que llamaremos conjunto de base, utiliza restricciones definidas de modo que el modelo resultante sea lo mas similar posible al de descuento absoluto. El segundo conjunto que llamaremos conjunto optimizado, resulta del mejor modelo obtenido con una gran cantidad de experimentos en los cuales se testearon numerosas combinaciones de restricciones. En la siguiente sección se describen las funciones de restricción usadas, así como la performance de los modelos resultantes.

5. RESULTADOS EXPERIMENTALES

5.1. Medida de evaluación de los resultados

Una medida muy frecuente en la evaluación de clasificadores bayesianos de tipo *aceptación-rechazo* como es nuestro caso, es la llamada curva ROC (Receiving Operating Characteristic) [1]. La curva ROC grafica el porcentaje de palabras mal reconocidas por el SRAH y detectadas como incorrectas por el verificador de hipótesis (TRR), en función del porcentaje de palabras bien reconocidas por el SRAH pero detectadas como incorrectas por el verificador de hipótesis (FRR). Dichos cocientes se determinan variando entre cero y uno el umbral de decisión contra el cual se contrasta la probabilidad. Dependiendo el tipo de curva que se obtenga se puede evaluar el rendimiento del sistema. El mejor caso correspondería a una curva cuyo TRR sea siempre uno ante cualquier valor de FRR, y el peor a una recta en 45 grados. El caso habitual es una curva comprendida entre estos dos casos extremos. En la figura 2 se muestra la curva ROC correspondiente al modelo de máxima entropía para la característica WTS, también puede verse la recta de peor caso. La medida de bondad del sistema que utilizaremos en el presente trabajo no es exactamente la curva ROC, sino un desprendimiento de la misma llamada AROC. La medida AROC se define como el cociente entre el área por debajo de la curva ROC y el área por debajo de la curva ROC de peor caso (recta de 45 grados). Cuanto mas se aproxime el valor AROC a 2 mas se acercará la curva ROC al caso ideal.

5.2. Corpus de datos utilizado

Los datos usados en el presente trabajo corresponden a la llamada “Tarea del Turista”. La misma comprende un corpus de voz y texto en castellano adquirido en el marco del proyecto Eu–Trans [5], compuesto por frases que típicamente se pronunciarían en el mostrador de un hotel. Concretamente: información sobre habitaciones, reservas, precio, solicitudes sobre la factura, petición de servicios, quejas, etc. El corpus está formado por 490.000 frases y un vocabulario de 683 palabras. El SRAH tiene un porcentaje de palabras reconocidas cercano al 95 %. A los efectos del sistema de verificación de hipótesis, objeto del presente trabajo, los datos de entrenamiento se presentan como $N = 13720$ n–uplas $(c, w, x_1, \cdots, x_d)$ que corresponden al etiquetado como correcto o incorrecto de cada una de las palabras reconocidas w por el SRAH y el correspondiente valor de x_i. También se dispone de 3365 n–uplas similares que no fueron usadas en el entrenamiento, para la evaluación del sistema.

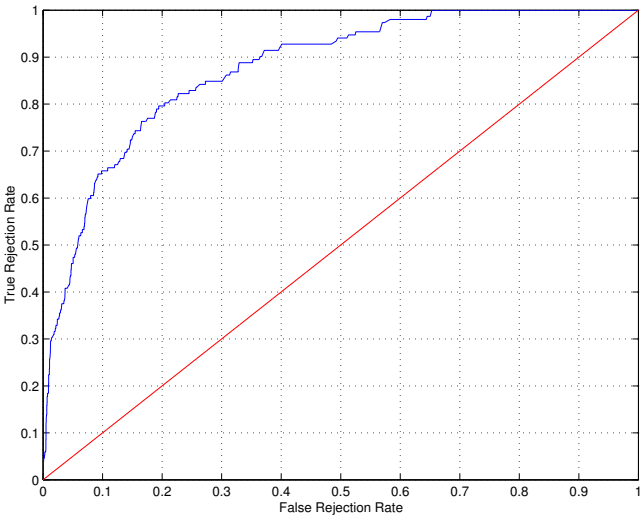

Figura 2: Curva ROC y recta de peor caso para el caso WTS

5.3. Evaluación de los modelos

5.3.1. Evaluación de los modelos de una sola característica (x_1 y x_2)

En ambos modelos (P_1 y P_2) se utilizaron los mismos tipos de funciones de restricción, por lo que llamamos genéricamente x a x_1 y x_2.

El objetivo del conjunto base es aproximar el modelo de máxima entropía al de descuento absoluto. Por lo tanto se introducen funciones de restricción que modelizan todos los conteos de triples (c_{ti}, x_{ti}, w_{ti}) que ocurrieron en los datos de entrenamiento análogamente a lo visto en 4.1.3. Esto origina un conjunto de funciones (una para cada triple) de la forma:

$$f_{c_{ti},x_{ti},w_{ti}}(c,x,w) = \begin{cases} 1 & si\ \ c = c_{ti},\ x = x_{ti},\ w = w_{ti} \\ 0 & otro\,caso \end{cases} \tag{10}$$

Para el caso en que un triple no ocurra en el conjunto de entrenamiento, ninguna de estas funciones será activada. Por lo tanto se define otro conjunto de restricciones que cumplirá la función de suavizado del modelo.

$$f_{c_i,x_i}(c,x,w) = \begin{cases} 1 & si\ \ c = c_i,\ x = x_i, \\ 0 & otro\,caso \end{cases} \tag{11}$$

donde c_i y x_i son todos los posibles valores que pueden adoptar las variables c y x. (Eg.: $(c = 0, x = 0{,}25)$, $(c = 1, x = 0{,}3)$, etc.). Dichas funciones se activarán con la restricción menos severa de que exista la dupla (c_i, x_i), independientemente del valor de w.

El conjunto optimizado fue definido empíricamente bajo la hipótesis de agrupar las palabras por su frecuencia de ocurrencia en los datos de entrenamiento. Se definieron dos conjuntos de funciones que se activan para cada posible frecuencia de palabra w: aquellas que se activan cuando $c = 1$ y x supera un cierto umbral um, y aquellas que se activan cuando $c = 0$ y x se mantiene por debajo de un umbral uM. Es decir,

$$f_{c_1,x>um,Fw}(c,x,w) = \begin{cases} 1 & si\ \ c = 1,\ x > um,\ N(w) = Fw \\ 0 & otro\,caso \end{cases} \tag{12}$$

$$f_{c_0,x_i<uM,Fw}(c,x,w) = \begin{cases} 1 & si\ \ c=0,\ x<uM,\ N(w)=Fw \\ 0 & otro\,caso \end{cases} \tag{13}$$

Los valores de los umbrales usados fueron $um = 0{,}25$ y $uM = 0{,}35$ para la característica x_1. Para la característica x_2 se determinó $um = uM = 9$. $N(w)$ es la cantidad de veces que ocurrió la palabra w en el conjunto de entrenamiento.

También se definieron las siguientes funciones con el objeto de suavizar el modelo:

$$f_{c_0,x_i}(c,x,w) = \begin{cases} 1 & si\ \ c=0,\ x=x_i \\ 0 & otro\,caso \end{cases} \tag{14}$$

$$f_{c_0,x_i<}(c,x,w) = \begin{cases} 1 & si\ \ c=0,\ x<x_i \\ 0 & otro\,caso \end{cases} \tag{15}$$

$$f_{N(c_0,x_i,Fw)}(c,x,w) = \begin{cases} 1 & si\ \ N(c=0,x=x_i,N(w)=Fw)=0 \\ 0 & otro\,caso \end{cases} \tag{16}$$

En el cuadro 1 se dan los resultados de AROC para las características x_1 y x_2 correspondientes al conjunto base y al conjunto optimizado. También se dan los resultados para las mismas usando el modelo de descuento absoluto, y los resultados obtenidos en [4]. Los valores entre paréntesis indican las mejoras relativas de los conjunto base y optimizado respecto del modelo de descuento absoluto.

AROC	Descuento Absoluto	Conjunto base	Conjunto optimizado	[4]
x_1 (AS)	1.64	1.68 (2.44 %)	1.72 (4.88 %)	1.73
x_2(WTS)	1.61	1.70 (5.6 %)	1.75 (8.7 %)	1.76

Cuadro 1: AROC para las características x_1 y x_2

5.3.2. Evaluación de los modelos para las características combinadas

Para el conjunto base se utilizaron las mismas funciones de restricción que para el modelo de una sola característica, es decir, las funciones (10) y (11) aplicadas a x_1 y x_2. Además se agregó un conjunto de funciones que modeliza los conteos de cuádruples $(c_{ti}, x_1^{ti}, x_2^{ti}, w_{ti})$ que ocurrieron en el conjunto de entrenamiento.

$$f_{c_{ti},x_1^{ti},x_2^{ti},w^{ti}}(c,x_1,x_2,w) = \begin{cases} 1 & si\ \ c=c_{ti},\ x_1=x_1^{ti},\ \ x_2=x_2^{ti},\ w=w_{ti} \\ 0 & otro\,caso \end{cases}$$

Para el conjunto optimizado se utilizaron las funciones de restricción usadas en el caso de una sola característica (14) a (16), aplicadas a cada una de las características, y las siguientes restricciones que pretenden modelizar las dependencias entre x_1 y x_2 (los valores de los umbrales son los mismos que en el caso de una característica):

$$f_{c_1,x_1>u1m,x_2>u2m,Fw}(c,x_1,x_2,w) = \begin{cases} 1 & si\ \ c=1,\ x_1>u1m,\ x_2>u2m,\ N(w)=Fw \\ 0 & otro\,caso \end{cases}$$

$$f_{c_0,x_1<u1M,x_2<u2M,Fw}(c,x_1,x_2,w) = \begin{cases} 1 & si\ \ c=0,\ x_1 < u1M,\ x_2 < u2M,\ N(w) = Fw \\ 0 & otro\, caso \end{cases}$$

En el cuadro 2 se dan los resultados de AROC para los conjuntos de características base y optimizados para la distribución conjunta. En la primera fila la misma se determina con el el modelo $P(c/x_1, x_2, w)$, mientras que en el segunda se asume independencia de las características y se determina con la siguiente ecuación:

$$P_{12}(c/x_1, x_2, w) = \frac{P_1(c/x_1, w)P_2(c/x_2, w)}{P(c/w)}$$

AROC	Conjunto base	Conjunto optimizado
Distribución conjunta	1.69	1.76
Producto de distribuciones	1.70	1.75

Cuadro 2: AROC de las distribuciones conjuntas e independientes

5.4. Análisis de resultados

5.4.1. Caso de una sola característica

En el cuadro 1 se puede ver en primer lugar que tanto el modelo basado en el conjunto base, como el modelo optimizado producen una importante mejora respecto del modelo de descuento absoluto tomado como base. En el modelo de máxima entropía presentado, los conteos asociados a las funciones de restricción definidas, han sido realizados basados en la ecuación (2), es decir, se ha utilizado la técnica de descuento absoluto, por lo que deberían esperarse resultados similares. Sin embargo, la diferencia radica en el caso en que $N(c, x, w) < k\ \forall\, c$. En ese caso como se dijo en la sección 3, el modelo de descuento absoluto, utiliza la distribución (3), mientras que el modelo de máxima entropía automáticamente mantiene desactivadas todas las funciones de restricción relacionadas con el conteo $N(c, x, w)$. Esto destaca una de las ventajas del modelo de máxima entropía, éste solo incorpora las fuentes de información disponibles, no haciendo ninguna suposición sobre lo que no dispone información. El modelo de descuento absoluto, por el contrario, está asumiendo la distribución (3) que no necesariamente es la mejor.

La segunda cuestión a analizar, radica en la mejora producida por el conjunto optimizado. Aquí se puede ver otra de las ventajas del método de máxima entropía, la posibilidad de incorporación de diferentes fuentes de conocimiento al modelo. En particular las características (12) y (13), agrupan las palabras del vocabulario por su frecuencia de ocurrencia. Es un hecho conocido en modelos de lenguaje [7], que el agrupamiento de palabras por su frecuencia puede producir buenos estimadores de probabilidad. Como vemos, la incorporación de esta característica a nuestro modelo produce una importante mejora respecto del conjunto base, en ambas características. Una ventaja adicional resultante del agrupamiento de palabras por su frecuencia, radica en la enorme reducción del número de parámetros a estimar. Por ejemplo para el conjunto base, con las funciones de restricción dadas, el número de parámetros para el modelo con característica x_2 (WTS) es de 6290, mientras que para el

conjunto optimizado es de 249, o sea, una reducción de mas de un orden de magnitud en la cantidad de parámetros a estimar. Esta reducción sin duda da lugar a modelos mas robustos con menor cantidad de datos de entrenamiento.

Finalmente, se han agregado en el cuadro 1 los resultados correspondientes al modelo planteado en [4] por uno de los autores. Vemos que los valores de AROC obtenidos son ligeramente superiores al modelo de máxima entropía. Si bien dicho modelo se basa en la técnica de descuento absoluto, no se modelizó la probabilidad a posteriori sino que se aplicó la regla de Bayes. Esto como dijimos, permite la implementación de un modelo total mas detallado que el de descuento absoluto con probabilidades a posteriori. Por lo tanto dichos resultados no son comparables con el modelo planteado de máxima entropía. En su lugar la comparación debe hacerse con el modelo de descuento absoluto a posteriori (columna 1 del cuadro 1).

5.4.2. Caso de múltiples características

Como vemos el modelo conjunto produce resultados similares al modelo independiente (las diferencias relativas son menores al $1\,\%$ por lo que no las consideramos significativas). Podemos concluir que al menos las características AS y WTS pueden considerarse estadísticamente independientes.

6. CONCLUSIONES

En el presente trabajo se ha implementado un sistema de verificación de hipótesis para un sistema de reconocimiento de habla. El mismo ha sido implementado utilizando el paradigma de máxima entropía, y los resultados han sido contrastados con el esquema basado en la técnica de descuento absoluto. Las principales ventajas del método de máxima entropía verificadas en este trabajo fueron:

- El modelo de máxima entropía realiza un suavizado automático cuando no posee información suficiente para estimar la probabilidad. Dicho suavizado utiliza toda la información disponible pero no hace ninguna suposición sobre la distribución. Como resultado se vio que el modelo resultante era significativamente mejor al de descuento absoluto.

- El modelo resultante permite la incorporación de fuentes diversas de conocimiento. En este caso se utilizaron los conteos de frecuencias de palabras que, se sabe, producen buenas estimaciones, especialmente cuando se tienen pocos ejemplos de entrenamiento. Los resultados también arrojaron incrementos significativos en la performance del verificador de hipótesis.

- La elección adecuada de las fuentes mas importantes de información a través de las funciones de restricción puede originar una reducción muy grande de los parámetros del modelo. Esto resulta en una mejor estimación cuando se disponen de pocos ejemplos, y en un aumento de la eficiencia del algoritmo de estimación de los parámetros.

Como desventaja principal cabe mencionar el tiempo de estimación de los parámetros. Si bien el algoritmo GIS asegura la convergencia si los modelos son consistentes, el entrenamiento puede requerir al menos veinte iteraciones del algoritmo. Esto suele ser muy costoso en términos computacionales cuando el número de parámetros es muy elevado. Por lo que se hace fundamental una adecuada elección de las restricciones del modelo.

REFERENCIAS

[1] M. Siu and H. Gish. Evaluation of word confidence for speech recognition systems. In *Computer, Speech and Language*, volume 13, pages 299–318, 1999.

[2] F. Wessel, R. Schuter, K. Macherey, and Hermann Ney. Confidence measures for large vocabulary continuous speech recognition. In *IEEE, Transactions on Speech and Audio Processing*, volume 9, pages 288–298. IEEE Press, 1999.

[3] V. D. Pietra S. D. Pietra and J. Lafferty. Inducing features of random fields. *IEEE Trans. Pattern Anal. Machine Intell.*, 1997.

[4] A. Sanchis. *Estimación y aplicación de medidas de confianza en reconocimiento automático del habla*. PhD thesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, May 2004.

[5] A. Sanchis, A. Juan, and E. Vidal. Improving utterance verification using a smoothed naive bayes model. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, volume 1, pages 592–595. IEEE Press, April 2003.

[6] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for speech recognition*. Edinburg University Press, 1992.

[7] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264, 1953.

[8] Reinhard Kneser and Hermann Ney. Improved backing–off for m–gram language modeling. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1995.

[9] Slava M. Katz. Estimation of probabilities from sparse data for language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401, 1987.

[10] E. T. Jaynes. Information theory and statistical mechanics, I and II. *Physical Reviews*, 106 and 108:620–630 and 171–190, 1957.

[11] Sven C. Martin, Hermann Ney, and Joerg Zaplo. Smoothing methods in maximum entropy language modeling. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, volume 1, pages 545–548, 1999.

[12] Juan Pablo Piantanida and Claudio Estienne. Maximum entropy good–turing estimator for language modeling. In *8th European Conference on Speech communication and technology (EUROSPEECH 2003)*, Geneva, Switzerland, Sep 1-4 2003.