UNA NUEVA FAMILIA DE ESTIMADORES DE PROBABILIDADES DE N-GRAMAS EN MODELOS COERSITIVOS

J. P. PIANTANIDA†, N. BARRAZA† y C. F. ESTIENNE†

†*Facultad de Ingenieria*
Universidad de Buenos Aires, Argentina
jpianta,nbarraz,cestien@fi.uba.ar

***Abstract*— Se propone una nueva familia de estimadores de probabilidades de n-gramas en modelos coersitivos. Estos estimadores se formulan sobre un modelo de probabilidad que considera el mecanismo coersitivo que poseen los n-gramas. Estudiando cuatro casos particulares del modelo probabilístico se obtienen los estimadores mas populares y utilizados en modelos del lenguaje, el estimador de Good-Turing, el estimador de descuento absoluto, un nuevo estimador de descuento generalizado y una nueva ley generalizada de sucesiones que contiene a la ley de Laplace y Lidstone. También se realiza un estudio de la performance de la ley de sucesión y el estimador propuesto. Finalmente se encuentra experimentalmente que estos estimadores conducen a una considerable reducción en la perplejidad evaluada en varios textos de diferentes vocabularios.**

***Keywords*— Probabilidad, estimación, inferencia, modelos coersitivos, ignorabilidad, suficiencia.**

I. INTRODUCCION

Existe un gran número de problemas prácticos que puede reducirse al problema de estimar distribuciones conjuntas de probabilidades para palabras o n-gramas en general (Shannon, 1948). Los n-gramas son un grupo de n palabras adyacentes, por ejemplo "las ciencias" es un bigrama y "las ciencias duras" un trigrama, etc. Podemos encontrar aplicaciones en compresión de datos, modelos de Markov, obtención de información, biología, reconocimiento de habla y muchas otras. Para estimar dichas distribuciones se utilizan como datos de entrenamiento las frecuencias de n-gramas que se extraen de textos. Generalmente el análisis de datos como estos está basado en asumir implícita o explícitamente, que el proceso que causa la dispersión en los datos puede ser ignorado (Rubin, 1976). Pero hay muchos casos donde la estadística de frecuencias no permite obtener una inferencia de la probabilidad que tienen esos eventos. Esto se debe a la naturaleza aleatoria que poseen las frecuencias de los diferentes eventos (Lindsey, 1998). Por otro lado aunque el tamaño del texto sea suficientemente grande, en la realidad hay muchos eventos que no ocurren, lo que no significa que tengan probabilidad nula. Cuando estimamos la probabilidad utilizando el estimador de máxima verosimilitud r/n en donde r es la frecuencia de un evento y n el número total de eventos, obtenemos una sobre-estimación. Además éste asigna una probabilidad nula a los eventos que tienen frecuencia nula, lo que finalmente conduce a estimaciones erroneas (Lindsey, 1998). El problema aquí planteado hace que sea necesario utilizar estimadores de probabilidad que ajusten la frecuencia de los eventos. Los más populares son el estimador de Good-Turing (1953), junto con otros estimadores empíricos, el estimador de Katz (1987) y los estimadores de descuento de Kneser y Ney (1995). Posteriores estudios estadísticos realizados por Church y Gale (2000) sobre la frecuencia, sugieren la existencia de mecanismos coersitivos en el proceso. Ignorar tales mecanismos nos conduce a cometer errores en la inferencia de los parámetros que definen el proceso (Heitjan, 1997).

En este trabajo proponemos una dirección diferente del clásico enfoque, el cual ignora el proceso de dispersión de las frecuencias y el mecanismo coersitivo. Sin embargo veremos que aquí se lo contiene como un caso particular del mismo. Para considerar el mecanismo que causa dispersión en las frecuencias es necesario poseer un modelo del mismo. Vamos a presentar un modelo que contemple el mecanismo de dispersión, así como también el mecanismo implícito de coersión. El que permite explicar la imposibilidad de inferir mediante máxima verosimilitud, la probabilidad de un evento desde la estadística de la frecuencia, cuando ignoramos estos mecanismos. En el nuevo modelo la probabilidad que deseamos estimar es una variable aleatoria. Por otro lado una distribución modela cómo se relacionan las observaciones (la frecuencia) con la probabilidad de cada evento. Utilizando el modelo y buscando el mejor estimador lineal de la probabilidad,

obtenemos una nueva familia de estimadores. Luego por medio de un riguroso tratamiento matemático encontramos dos estimadores generalizados que contienen a los estimadores más populares. Utilizamos estos dos estimadores generalizado para estimar las probabilidades en modelos del lenguaje y evaluar su performance.

En la siguiente sección describimos los fundamentos del modelo probabilístico. En la sección III encontramos una nueva familia de estimadores. En la sección IV desarrollamos dos estimadores generalizados que contienen a los más utilizados. Luego en la sección V demostramos las propiedades asintónticas de estos estimadores encontrados. En la sección VI presentamos los resultados experimentales. Finalmente en la sección VII extraemos algunas conclusiones.

II. FORMULACION DEL MODELO

Para comenzar formulamos cuidadosamente el experimento desde el cuál vamos a asumir que es diseñado el texto de donde extraemos las frecuencias. Supongamos que disponemos de S fuentes que definen los eventos, es decir n-gramas (unigramas, bigramas, etc.). La elección de una fuente esta representada por la variable aleatoria I que toma valores sobre el espacio muestral $\mathcal{S}_I$ y posee una distribución $P_\phi^I(i)$. Extraemos un n-grama de esa fuente con una probabilidad dada por la variable aleatoria Θ, que representa la probabilidad de un n-grama y toma valores sobre un espacio muestral $\mathcal{S}_\Theta$. Continuamos este proceso hasta realizar n extracciones, donde la probabilidad de obtener r éxitos de un n-grama de la fuente i, con una probabilidad θ, esta dada por $P^{R|\Theta,I}(r|\theta,i)$. Si las fuentes son independientes e idénticamente distribuidas, estamos en el caso del clasico estimador de Good-Turing. La variable Θ, en la fuente i, tiene una distribución $P_\gamma^{\Theta|I}(\theta|i)$. Finalmente para este modelo la probabilidad de tener r exitos de la fuente i, es dada por la distribución de R dado $I = i$ es

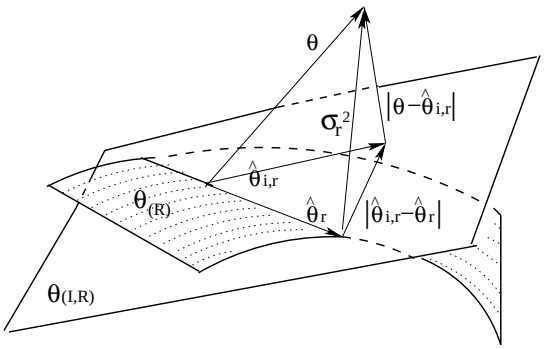
$$P_{\gamma,\phi}^{R|I}(r|i) = \frac{\displaystyle\int_\Theta P^{R|\Theta,I}(r|\theta,i) P_\phi^I(i)\, dU_\gamma(\theta|i)}{\displaystyle\sum_{u\in\mathcal{S}_R(i)} \int_\Theta P^{R|\Theta,I}(u|\theta,i) P_\phi^I(i)\, dU_\gamma(\theta|i)} \tag{1}$$

donde los parámetros γ y ϕ pueden ser o no diferentes. El espacio conjunto de parámetros es el producto cartesiano de los dos espacios (γ,ϕ). En este modelo de datos coersitivos R es un subconjunto del dato completo e I indica el tipo y grado de coersión. Es importante destacar que no podemos observar Θ, pero observamos $R = R(\Theta, I)$. Donde R es el subconjunto más pequeño de $\mathcal{S}_\Theta$, que contiene la verdadera Θ. Vamos a hacer una suposición adicional, necesaria para poder asegurar la identificabilidad del modelo, y es que cada valor de I induce una partición de $\mathcal{S}_\Theta$. Dado $\mathcal{S}_R(i) = \{r : \exists\, \theta \,/\, r = R(\theta, i)\}$, asumimos que si $u_1, u_2 \in \mathcal{S}_R(I)$ y $u_1 \neq u_2$ implica que $u_1 \cap u_2 = \emptyset$ y además $\cup\{u : u \in \mathcal{S}_R(I)\} = \mathcal{S}_\Theta$.

La probabilidad $P^{R|\Theta,I}(r|\theta,i)$ es el modelo de observaciones del proceso y representa un indicador de la dispesión en la frecuencia observada r, validando su valor. En el modelo aquí propuesto, el par de datos observado es (R, I), donde R es un subconjunto del espacio muestral del dato completo. El mecanismo de coersión se ignora si la distribución $P_{\gamma,\phi}^{R|I}(r|i)$ (1) es equivalente a una incorrecta distribución calculada sin considerar la distribución $P_\gamma^{\Theta|I}(\theta|i)$. Es fácil ver que si para un valor fijo de las observaciones r e i, y para cada γ, $P_\gamma^{\Theta|I}(\theta|i)$ toma el mismo valor para todo $\theta \in r$ entonces R es observada L-suficiente para θ. En este sentido R contiene toda la información necesaria a saber sobre θ.

III. FORMULACION DEL ESTIMADOR

El objetivo del estimador es obtener una estimación de la probabilidad θ de un n-grama, conociendo su frecuencia r en el texto. Para ello contamos con una estadística[1] conjunta $\theta(I, R)$ de la correspondiente frecuencia R de cada n-grama I, y la estadística marginal $\theta(R)$ de frecuencias. Deseamos encontrar el mejor estimador lineal, de la probabilidad θ, que contemple el proceso de coersión descripto en la sección anterior. Para ello nos proponemos buscar el estimador $\hat{\theta}_r$ que minimice la varianza de estimación σ_r^2. Podemos hacer



una interpretación geométrica del problema de estimación, como la observada en la figura. El estimador de mínima varianza esta dado por

$$\hat{\theta}_r = \inf_{\theta_r \in \theta(r)} \mid \theta - \theta_r \mid^2 \tag{2}$$

Para minimizar la varianza del estimador σ_r^2, es necesario descomponerla en sus componentes ortogonales

$$\sigma_r^2 = \mid \theta - \hat{\theta}_r \mid^2 = \mid \theta - \hat{\theta}_{i,r} \mid^2 + \mid \hat{\theta}_{i,r} - \hat{\theta}_r \mid^2 \tag{3}$$

ya que deseamos poner de manifiesto el proceso coersitivo. Entonces la solución buscada (2), será la

[1] Estadística se denomina a la distribución de probabilidad de θ condicionada a las variables observadas que definen la estadística.

proyección ortogonal de θ sobre el subespacio de la estadística conjunta $\theta(I,R)$ que notamos $\hat{\theta}_{i,r}$, y minimiza el primer término de la ecuación (3). Luego proyectamos este resultado $\hat{\theta}_{i,r}$ sobre el subespacio generado por la estadística marginal $\theta(R)$ que notamos $\hat{\theta}_r$, minimizando así el segundo término de la ecuación (3). Es decir la primer proyección $\hat{\theta}_{i,r} =< \theta, \theta(I,R) >_I$, y entonces el estimador óptimo (2) será $\hat{\theta}_r =< \hat{\theta}_{i,r}, \theta(R) >_\Theta$. Mediante esta interpretación geométrica del problema, no es difícil demostrar que la matríz de correlación de la probabilidad deseada θ y la estimada $\hat{\theta}_r$ van a ser coincidentes $R_{\theta,I,R} = R_{\hat{\theta},I,R}$.

Ahora si reescribimos las proyecciones, que definen el estimador, en términos probabilísticos finalmente resulta

$$\hat{\theta}_{r,\gamma,\phi} = E^{\Theta}\{E^{I}_{\gamma,\phi}\{\theta|r,i\}|r\} \tag{4}$$

Utilizando la distribución conjunta de probabilidad $P^{R,\Theta,I}_{\gamma,\phi}(r,\theta,i)$ podemos escribir la expresión final del estimador (4) como

$$\hat{\theta}_{r,\gamma,\phi} = \frac{\sum_{i\in\mathcal{S}_I}\int_\Theta \theta P^{R|\Theta,I}(r|\theta,i)\,P^I_\phi(i)\,dU_\gamma(\theta|i)}{\sum_{i\in\mathcal{S}_I}\int_\Theta P^{R|\Theta,I}(r|\theta,i)\,P^I_\phi(i)\,dU_\gamma(\theta|i)} \tag{5}$$

En la mayoría de los casos puede que la probabilidad $P^{R|\Theta,I}(r|\theta,i)$ no dependa de la fuente i. Es natural pensar que el mecanismo mediante el cual se relacionan las observaciones con el proceso oculto θ, es igual para todas las fuentes. En este caso la ecuación (5) resulta

$$\hat{\theta}_{r,\gamma,\phi} = \frac{\int_\Theta \theta P^{R|\Theta}(r|\theta)\,dS_{\gamma,\phi}(\theta)}{\int_\Theta P^{R|\Theta}(r|\theta)\,dS_{\gamma,\phi}(\theta)} \tag{6}$$

donde $dS_{\gamma,\phi}(\theta)$ esta dado por una mezcla de probabilidades

$$dS_{\gamma,\phi}(\theta) = \sum_{i\in\mathcal{S}_I} P^I_\phi(i)\,dU_\gamma(\theta|i) \tag{7}$$

Little (1993) encuentra que las mezclas de probabilidades son las distribuciones más adecuadas para trabajar en presencia de datos dispersos. En la siguiente sección estudiamos cuatro casos particulares de la expresión del estimador (6).

IV. RELACION CON OTROS ESTIMADORES

A. Estimador de Good-Turing

Si consideramos que los eventos de la misma fuente son marginalmente binomiales; que la distribución de probabilidad para I es uniforme; y que la probabilidad de cada evento θ es determinística. Es decir, la probabilidad de obtener r éxitos de n extracciones posee una distribución binomial

$$P^{R|\Theta,I}_n(r|\theta,i) = \binom{n}{r}\theta^r(1-\theta)^{n-r} \tag{8}$$

y $P^I_\phi(i) = \frac{1}{s}$ y la distribución de probabilidad $U_\gamma(\theta|i) = \delta(\theta - p_i)$. La ecuación (6) resulta

$$\hat{\theta}_{r,n} = \frac{r+1}{n+1}\frac{\sum_{i=0}^{S}\binom{n+1}{r+1}p_i^{r+1}(1-p_i)^{n-r}}{\sum_{i=0}^{S}\binom{n}{r}p_i^{r}(1-p_i)^{n-r}} \tag{9}$$

que es el clásico estimador de Good-Turing (1953).

B. Ley generalizada de sucesiones

En este caso vamos a considerar que (7) es caracterizada por una distribución Beta. Sabemos por el teorema de descomposición de Lebesgue que existen 2^{S-1} posibles mezclas que estamos representando mediante esta distribución

$$S_{\gamma,\phi}(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{b-1}(1-\theta)^{a-1}, \quad a,b>0 \tag{10}$$

donde $\gamma = (a,b)$. Las observaciones siguen siendo marginalmente binomiales como en (8). El estimador que resulta es

$$\hat{\theta}_{r,\gamma,n} = \frac{\Gamma(b+r+1)}{\Gamma(b+r)}\frac{\Gamma(a+b+n)}{\Gamma(a+b+n+1)} \tag{11}$$

Si tomamos $b=1$ en (11) y hacemos que $a+1 = s$, donde s es la cantidad de fuentes. Obtenemos la conocida ley de sucesión de Laplace (Ristad, 1995).

$$\hat{\theta}_{r,\gamma,n} = \frac{r+1}{n+s} \tag{12}$$

Por otro lado si tomamos (11) y hacemos $a = b(s-1)$, resulta

$$\hat{\theta}_{r,\gamma,n} = \frac{r+b}{n+sb} \tag{13}$$

que es la conocida ley de sucesión de Lidstone (Ristad, 1995). Para la estimación de los parámetros, es conveniente reescribir (11) como

$$\hat{\theta}_{r,\gamma,n} = \mu\frac{r}{n} + (1-\mu)\frac{1}{a/b+1}, \quad \mu = \frac{1}{(a+b)/n+1} \tag{14}$$

donde el valor del cociente a/b es optimizado para minimizar la perplejidad, y luego el parámetro μ es estimado mediante el algoritmo EM.

C. Estimador generalizado de descuento

Supongamos que la densidad de probabilidad (7) es Gamma, es decir

$$s_{\gamma,\phi}(\lambda) = \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)}\lambda^\alpha e^{-\beta\lambda} \quad \alpha>-1,\ \beta>0 \tag{15}$$

donde $\gamma = (\alpha, \beta)$. El proceso que genera las observaciones es un proceso Poisson, con una media dada por la variable λ que se relaciona con la probabilidad por medio de $\lambda = \theta\, n$

$$P^{R|\Lambda,I}(r|\lambda, i) = \frac{\lambda^r e^{-\lambda}}{r\,!}, \qquad \lambda > 0 \tag{16}$$

entonces la ecuación (4) para el cálculo del estimador se convierte en

$$\hat{\theta}_{r,n,\gamma,\phi} = E^{\Lambda}\{E^{I}_{\gamma,\phi}\{\Lambda|r, i\}/n|r\} \tag{17}$$

Calculamos el estimador definido por (17) y renombrando a

$$(1-k) = \frac{1}{\beta+1} \qquad b = \alpha + 1$$

resulta

$$\hat{\theta}_{r,n,k,b} = (1-k)\frac{r+b}{n} \tag{18}$$

donde $0 < k < 1$ y $b > 0$. En la ecuación (18) está contenido el estimador de descuento lineal propuesto por Ney y Kneser (1995). Para la estimación de los parámetros, utilizamos dos criterios. El parámetro b es optimizado para disminiuir la perplejidad por el método usual y k se obtiene maximizando el likelihood con el método de leaving-one-out (Ney y Kneser, 1995), resultando

$$k = \frac{C_1}{n} - \frac{bS}{n}\left(1 - \frac{C_1}{n}\right) \tag{19}$$

donde C_1 es la cantidad de eventos con frecuencia 1 y S el tamaño del vocabulario.

D. Estimador de descuento absoluto

Supongamos que la distribución de probabilidad de los eventos es Pareto

$$s_{\gamma,\phi}(\lambda) = b\lambda^{-(b+1)} \qquad b > 0,\ \lambda > 0 \tag{20}$$

Las observaciones provienen de un proceso Poisson, con una media dada por la variable λ que se relaciona con la probabilidad por medio de $\lambda = \theta\, n$, y utilizando la ecuación del estimador (17) resulta

$$\hat{\theta}_{r,n,b} = \frac{r-b}{n} \tag{21}$$

V. ANALISIS ASINTOTICO

A. Leyes de sucesiones

Cuando se trabaja con leyes de sucesiones es importante estudiar lo que sucede con la probabilidad asignada a una cadena de n-gramas a medida que la longitud de dicha cadena se incrementa. Sea una cadena $\mathcal{X}^n$ generada de s fuentes de símbolos y de longitud n; donde indicamos por q la cantidad de n-gramas diferentes que estan presentes en la cadena. Llamamos $P_L(\mathcal{X}^n/n)$ a la probabilidad de la cadena $\mathcal{X}^n$ estimada mediante la ley de Lidstone, y $P_\lambda(\mathcal{X}^n/n)$ la probabilidad utilizando la ley de Laplace. Podemos ver que cuando n se incrementa estas probabilidades poseen las siguentes relaciones asintóticas (Ristad, 1995)

$$P_L(\mathcal{X}^n/n) = \Theta(\left(\frac{1}{n-1}\right)^{\lambda(s-q)}) \tag{22}$$

$$P_\lambda(\mathcal{X}^n/n) < \left(\frac{s-1}{n+s-1}\right)^{s-q} \tag{23}$$

en donde a medida que incrementamos n, como $q < s$, las probabilidades de ámbas cadenas tienden rápidamente a cero. Lo que suele ser un punto importante ya que en general en sucesiones las cadenas de datos son largas y se desea mantener la probabilidad constante de n. Ahora estudiemos la probabilidad que asigna la ley generalizada de sucesiones que encontramos en (11). La probabilidad total asignada a una cadena $\mathcal{X}^n$, donde r_i es la frecuencia de cada símbolo estará dada por

$$P_{a,b}(\mathcal{X}^n/n) = \frac{\prod\limits_{i\,\in\, q}(\Gamma(r_i+b)/\Gamma(b))}{\Gamma(n+a+b)/\Gamma(a+b)} \tag{24}$$

utilizando la aproximación de Stirling

$$\begin{aligned} P_{a,b}(\mathcal{X}^n/n) &= \prod_{i\,\in\, q}^{n} \frac{i+bq}{i+a} \\ &\approx \frac{\Gamma(a)}{\Gamma(bq)}(\left(\frac{1}{n-1}\right)^{b(a/b-q)}) \\ &= \Theta(\left(\frac{1}{n-1}\right)^{b(a/b-q)}) \end{aligned} \tag{25}$$

si tomamos $a/b \sim q$ entonces estaremos logrando una asignación de probabilidad que posea una cuasi independencia de la longitud de la cadena.

B. Estimadores de descuento

Una forma de comparar dos estimadores es mediante el cociente de la probabilidad que estos asignan a la misma cadena, y analizar que sucede cuando incrementamos n (Ristad, 1995). Podemos encontrar que la probabilidad asignada por el estimador de descuento absoluto (21) a una cadena $\mathcal{X}^n$, esta dada por

$$P_\delta(\mathcal{X}^n/n) = \frac{\delta^{q-1}(q-1)!(s-q)!}{s(s-1)!(n-1)!}\prod_{i\,\in\, q}\frac{\Gamma(r_i-\delta)}{\Gamma(1-\delta)} \tag{26}$$

y la probabilidad aignada a la misma cadena por el estimador de descuento lineal (Ney y Kneser, 1995)

$$P_\alpha(\mathcal{X}^n/n) = \frac{\alpha^{q-1}(1-\alpha)^{n-q}(s-q)!}{s(s-1)!(n-1)!}\prod_{i\,\in\, q}\Gamma(r_i) \tag{27}$$

Para comparar estas dos probabilidades realizamos el cociente y se demuestra que el órden asintótico con n (Ristad, 1995)

$$P_\delta(\mathcal{X}^n/n)/P_\alpha(\mathcal{X}^n/n) = \Theta(2^{n-q}) \tag{28}$$

Entonces podemos observar que la performance del estimador de descuento lineal es muy inferior al de descuento absoluto. Debido a que será mucho mas fácil codificar una fuente estimada con el de descuento absoluto que con el otro estimador. Un caso mas interesante, es comparar el estimador generalizado de descuento (18), con el de mejor performance (descuento absoluto). Para ello expresemos (18) como

$$\hat{\theta}_{k,b,r,n} = \begin{cases} (r+b)(1-k)/n & r > 0 \\ \dfrac{k(bq+n)}{n(s-q)} & r = 0 \end{cases} \quad (29)$$

y es fácil demostrar que la probabilidad asignada a la cadena

$$P_{k,b}(\mathcal{X}^n/n) =$$

$$= \frac{(s-q)!(bq+n)!(1-k)^{n-q}}{s(s-1)!(n-1)!(bq)!k^{1-q}} \prod_{i \in q} \frac{\Gamma(r_i+b)}{\Gamma(1+b)} \quad (30)$$

Luego calculamos el cociente entre (26) y (30) para encontrar el órden asintótico de este con n

$$P_\delta(\mathcal{X}^n/n)/P_{k,b}(\mathcal{X}^n/n) =$$

$$= \frac{\delta^{q-1}(s-1)!(bq)!}{k^{q-1}(bq+n)!(1-k)^{n-q}} \prod_{i \in q} \frac{\Gamma(r_i-\delta)\Gamma(r_i+b)}{\Gamma(1-\delta)\Gamma(1+b)}$$

$$= \Theta\left(\left((bq+n)!\,(1-k)^{n-q} \prod_{i \in q} (r_i-\delta)^\delta (r_i+b)^b \right)^{-1} \right)$$

$$= \Theta\left(\left(\frac{(1-k)^{n-q}}{(n+bq)^{bq}} \frac{\prod_{i \in q} r_i^{1+\delta}}{\prod_{i \in q} r_i^{1-b}} \right)^{-1} \right) \quad (31)$$

donde

$$n^{\delta+b} \leq \prod_{i \in q} r_i^{1+\delta} / \prod_{i \in q} r_i^{1-b} < n^{q(\delta+b)} \quad (32)$$

y entonces

$$\Theta\left(\frac{1}{n^{q\delta}(1-k)^{n-q}} \right) \leq P_\delta(\mathcal{X}^n/n)/P_{k,b}(\mathcal{X}^n/n) \quad (33)$$
$$\leq \Theta\left(\frac{n^{bq-b-\delta}}{(1-k)^{n-q}} \right)$$

Finalmente la performance relativa del modelo de descuento absoluto y el modelo generalizado de descuento es función de la entropía empírica $\mathcal{H}(r_i/n)$. Cuando la entropía empírica es alta, entonces el modelo generalizado de descuento será mejor estimador; cuando la entropía empírica es menor, entonces el modelo de descuento absoluto será mejor.

VI. RESULTADOS EXPERIMENTALES

A. Descripción de datos

Los experimentos se realizan en tres cuerpos de datos: una base de datos en Inglés, una fase de switchboard, y dos bases de datos en Español, Latino 40 (disponible en LDC) y Latin-American generada por SRI International. También utilizamos textos extraídos de diarios. La perplejidad es medida sobre parte de estas bases de datos. Utilizamos modelos de bigramas para Latino40 y modelos de trigramas en switchboard y Latin-American. Los textos se dividen en tres clases:

- Texto A: Consiste de texto tomado de las trascripciones de Latino40, utilizamos 32k palabras para entrenamiento y 8k palabras para testeo.
- Texto B: Consiste de texto tomado de las trascripciones de Latin-American y texto de diarios, combinando los dos textos resultan 752k palabras de entrenamiento y 33k palabras para testeo.
- Texto C: Consiste de 3M palabras tomadas de las trscripciones de switchboard utilizada para entrenamiento, y 59k palabras tomadas de la evaluación de hub-5 2001 utilizadas para testeo.

B. Resultados

La calidad de un modelo del lenguaje se mide mediante la perplejidad tomada de un texto de evaluación, cuanto más baja sea esta mejor es el modelo. Esta se define como

$$H = -\lim_{m \longrightarrow \infty} \sum_{w_1, w_2, \ldots, w_m} \log \hat{P}(w_k | \mathbf{W}_{k-n+1}^{k-1}) \quad (34)$$

donde $\hat{P}(w_k|\mathbf{W}_{k-n+1}^{k-1})$ son las probabilidades condicionales de cada n-grama, estimadas mediante un estimador de probabilidad.

Tabla 1: *Perplejidades de distintos estimadores con diferentes vocabularios*

Estimator	Text A (bi-grama)	Text B (tri-grama)	Text C (tri-grama)
CGT	219	739	534
ADE	149	251	160
LDE	138	693	176
KATZ	156	232	155
GDE	**131**	**233**	**153**
GLS	**142**	**231**	**151**

Las perplejidades están medidas sobre el Good-Turing estimator (1953) (GT), el estimador de Katz (1987) (KATZ), el estimador de descuento absoluto (ADE) y descuento lineal Ney and Kneser (1995), el estimador de descuento generalizado (GDE) y la ley generalizada de sucesiones (GLS). Podemos observar los resultados en la tabla 1.

C. Discusión

Podemos ver en los resultados que el estimador GDE introduce una mejora en la perplejidad del texto A con respecto a los otros estimadores. Mientras que la ley generalizada de sucesiones GLS presenta una performance inferior a éste. Sin embargo la performance de GLS es superior en los textos B y C. Dichos textos presentan una cantidad superior de eventos al texto A, debido al incremento del vocabulario. Además las frases de éstos son considerablemente mas largas, lo que justifica la mejora de este estimador. Podemos verlo como una consecuencia de lo estudiado en la sección V, donde demostramos que probabilidad estimada decae muy poco con la longitud de la frase. Finalmente debemos concluir que si el texto posee frases largas o gran vocabulario conviene utilizar la ley generalizada de sucesiones GLS, y en caso contrario el estimador GDE.

VII. CONCLUSIONES

Sobre la base de un modelo probabilístico de n-gramas que poseen un mecanismo de coersión, encontramos una familia de estimadores de la probabilidad de ocurrencia. Esta familia representa los mejores estimadores lineales en un modelo coersitivo. Como casos de particular iterés presentamos una ley de sucesiones generalizada que contiene a la ley de Laplace y Lidstone, el estimador de Good-Turing, los estimadores de descuento de Ney y un estimador generalizado de descuento. Es importante destacar que los conocidos estimadores de descuento hasta el momento no poseían una formulación formal, ya que eran empíricos. Demostramos que la ley generalizada de sucesiones puede mantener constante la probabilidad de una cadena sin importar la longitud de la misma. Estudiamos en qué casos el estimador generalizado de descuento es superior al de descuento absoluto. Por último evaluamos los dos estimadores propuestos en distintos textos y éstos presentan una reducción en la perplejidad. Pero fundamentalmente mantienen su performance, para los tres textos, mientras que otros estimadores presentan diferencias considerables.

VIII. AGRADECIMIENTOS

Queremos agradecer al Star-Lab de SRI International y especialmente al Dr. Horacio Franco por permitirnos utilizar su base de datos Latin-American Spanish.

REFERENCIAS

Church, W. K., and Gale, W. A., "Poisson mixtures" *AT&T Bell Labs-Research.* (2000)

Heitjan, D.F. "Ignorability, Suffiency and Ancillarity" *Journal of the Royal Statistical Society*, **59**, 375-381 (1997).

Donald B. Rubin, "Inference and missing data" *Biometrika.*, **63(3)**, 581-592 (1976).

Feller, W., *An Introduction to Probability Theory and its Applications*, John Wiley & Sons Inc., New York., **II**, (1966)

Good, I. J., "The population frequencies of species and the estimation of population parameters", *Biometrika.*, **40**, 237-264 (1953).

Katz, S. M., "Estimation of probabilities from sparse data for language model component of a speech recognizer", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, **35(3)**, 400-401 (1987).

Lindsey, J. K., and Denne, J. S., "Missing data: a fundamental frequentist problem" *Report Biostatistics, Limburgs University, Diepenbeek, Belgium.*, (1998)

Little, R.J.A. "Pattern-mixture models for multivariate incomplete data" *Journal of American Statistical Association*, **88**, 125-134 (1993).

Ney, H., Essen, U., and Kneser, R., "On the Estimation of 'Small'Probabilities by Leaving-One-Out" *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, **17**, 1202-1212 (1995).

Ristad Eric S., "A Natural Law of Succession", *Research Report CS-TR-495-95.*, July (1995).

Shannon C., "The Mathematical Theory of Communication" *Bell System Technical Journal.*, (1948)