Un Nuevo Modelo Para la Estimación de Bi-gramas en Reconocimiento del Habla

Claudio Estienne

Instituto de Ingeniería Biomédica, Facultad de Ingeniería,
Universidad de Buenos Aires, Argentina
cestien@fi.uba.ar

Resumen Se presenta un nuevo método para el suavizado de N-gramas utilizando regularización en un modelo de máxima entropía. Dicha regularización se efectúa introduciendo un término en la función objetivo al estilo de las máquinas de soporte vectorial. Relacionado con dicho término se incluye una variable que actúa como descuento de probabilidades en el estimador, similar al usado en otros métodos de suavizado de modelos de lenguaje, pero considerando dicho descuento como otra variable a optimizar. El modelo fue evaluado en una tarea de reconocimiento de habla usando modelos de lenguaje de bi-gramas. Los resultados se testaron usando la base de datos Latino-40 midiendo perplejidad y porcentaje de palabras reconocidas. Los resultados fueron significativamente superiores a un modelo que es estado del arte.
Palabras claves: modelos de lenguaje, máxima entropía, regularización

1. Introducción

[1] Los modelos probabilísticos de lenguaje son componentes claves en muchas áreas de procesamiento de la información tales como reconocimiento de textos manuscritos, traducción automática y reconocimiento automático del habla [14], [1]. Una gran ventaja que presentan dichos modelos es su naturaleza probabilística, la cual les permite hacer predicciones acerca de datos futuros basados en los datos actuales usando el sólido marco de la teoría de la probabilidad. Uno de los modelos mas difundidos es el llamado modelo de N-gramas [10]. Supongamos una secuencia de palabras, por ejemplo: *La probabilidad es baja*, la probabilidad de ocurrencia de dicha secuencia se aproxima como:

$$\begin{aligned} P(la, probabilidad, es, baja) &= P(baja/es, probabilidad, la) \\ &\quad P(es/probabilidad, la)P(probabilidad/la)P(la) \\ &\approxeq P(baja/es)P(es/probabilidad) \\ &\quad P(probabilidad/la)P(la). \end{aligned} \tag{1}$$

Es decir, aplicamos repetidamente el teorema de Bayes y limitamos el modelo de modo que solo tenga *memoria* de un evento anterior. En general cuando se

[1] Este trabajo fue financiado por la Universidad de Buenos Aires en el marco del proyecto UBACYT-2008-2011 código I003

limita a $n-1$ eventos anteriores, decimos que es un modelo de N-gramas. En el presente trabajo solo usaremos modelos de bi-gramas, aunque las conclusiones que se obtienen son perfectamente extrapolables a modelos de N-gramas. La estimación de las probabilidades de cada bi-grama se realiza tomando un texto de entrenamiento "suficientemente grande" en el cual se cuenta el número de ocurrencias de cada bi-grama y se divide por el número total de bi-gramas en dicho texto. Sin embargo, debido a la forma en que las palabras se distribuyen en los lenguajes esta aproximación tiene al menos dos problemas. En primer lugar, aún cuando se disponga de una gran cantidad de texto de entrenamiento la cantidad de ejemplos de cada bi-grama es insuficiente para tener un estimador adecuado. En segundo lugar, muchas palabras o bi-gramas pueden no haber ocurrido nunca en el texto de entrenamiento, sin embargo esto no significa que tengan probabilidad nula de ocurrencia, por lo tanto no puede asignárseles probabilidad nula en el modelo. El problema a resolver consiste entonces en interpolar probabilidades o "suavizar" el modelo de modo de obtener estimaciones de probabilidad no nulas cuando se tienen muy pocos o ningún ejemplo de determinados eventos [10]. Dicho problema de suavizado de modelos estadísticos es muy conocido en modelización del lenguaje y ha sido enfocado por diversos autores usando técnicas de back-off, [10], descuento de probabilidades [14], y máxima entropía [13].

El principio de Máxima Entropía fue formulado por Jaynes [9] en el marco de la mecánica estadística, (ver por ejemplo [1] para una revisión), y desde entonces fue aplicado en gran cantidad de áreas científicas y de ingeniería, incluidas varias ramas del procesamiento del habla como modelos de lenguaje [5], [8], [3],traducción automática [1], procesamiento natural del lenguaje [11] y estimación de medidas de confianza [7] . El suavizado de modelos de N-gramas usando máxima entropía ha sido tratado en [13], [5] y mas recientemente en [4]. Lamentablemente aún cuando los modelos de máxima entropía se combinan con modelos tradicionales de suavizado de lenguaje, los problemas mencionados como la falta de datos para estimar eventos raros, y el sobre ajuste (over-fitting) de datos, no se han podido solucionar. Una técnica que ha probado ser exitosa es la regularización de los parámetros resultantes del modelo de máxima entropía. Esto se ha realizado de varias maneras. Chen y Rosenfeld [5], utilizaron regularización en el suavizado de modelos de lenguaje incluyendo una distribución a priori de tipo Gausiana de los parámetros, en el modelo de máxima entropía condicional. Goodman [8] obtuvo mejoras cambiando dicha Gausiana por una distribución de tipo Laplace. Kazama y Tsujii [11] en un problema de procesamiento de lenguaje natural usaron un método diferente. Incluyeron regularización relajando las restricciones del modelo usando desigualdades en lugar de igualdades. Dudik [6] propuso un esquema general en el cual tanto el modelo de Rosenfeld como el de Goodman y el de Kasama resultaron ser casos particulares de dicho esquema. Mas recientemente, Chen [3] mediante una técnica de *angostamiento* del tamaño del modelo logró mejoras en la perfomance de un modelo de máxima entropía regularizada.

El modelo de máxima entropía aplicado a modelos de N-gramas se puede plantear del siguiente modo: tenemos eventos de la forma (x, y), donde intenta-

mos estimar la palabra y dadas las palabras previas x. Supongamos que tenemos una estadística en la cual estamos interesados, generalmente dada por frecuencias de ocurrencia de N-gramas en el conjunto de entrenamiento. Es posible incluir dicha estadística en el modelo por medio de funciones indicadoras $f_i(x, y)$ cuyo valor es igual a uno cuando el evento (x_i, y_i) ocurre y cero cuando no ocurre. La media de f_i con respecto a la distribución empírica $\tilde{p}(x, y)$ sería la característica en la cual estamos interesados. El objetivo es estimar la probabilidad condicional $p(y/x)$ con la mayor entropía $H(p) = -\sum_{x,y} \tilde{p}(x)p(y/x) \log p(y/x)$, donde $\tilde{p}(x)$ es la distribución empírica de x en la muestra de entrenamiento. También se requiere que la media de f_i con respecto a la distribución conjunta $p(x, y) = \tilde{p}(x)p(y/x)$ sea igual a la media de f_i con respecto a la distribución empírica, lo cual se expresa como: $E_p\{f_i\} = E_{\tilde{p}}\{f_i\}$. Dadas estas condiciones, hemos restringido el conjunto de distribuciones de máxima entropía posibles solamente a aquellos que satisfacen las estadísticas que consideramos importantes. El problema se puede plantear en el marco de un problema de optimización restringida, en el cual se debe halla la distribución $p(y/x)$ que maximice la función objetivo $H(p)$ sujeta a las restricciones $E_p\{f_i\} = E_{\tilde{p}}\{f_i\}$. Se puede demostrar que dicha distribución corresponde a un modelo exponencial de la forma [1]:

$$p_\Lambda(y/x) = \frac{exp(\sum_i \lambda_i f_i(x, y))}{Z_\lambda(x)}. \tag{2}$$

donde $Z_\Lambda(x)$ es un factor de normalización. El conjunto de parámetros Λ igual al número de restricciones f_i se estima como la solución del correspondiente problema dual, el cual se puede demostrar que es igual al logaritmo de la verosimilitud del conjunto de entrenamiento [1]. Eligiendo adecuadamente las restricciones f_i es fácil combinar el modelo de máxima entropía con otros modelos que han resultado ser eficientes. Por ejemplo el modelo de Kneser-Ney para el suavizado de N-gramas [12] así como otros métodos de suavizado pueden ser fácilmente incluidos en el modelo de máxima entropía [5], [13].

En este trabajo también veremos el problema de suavizado de modelos de bigramas usando máxima entropía enfocándonos en la regularización como forma de mejorar las partes débiles de dicho modelo. Sin embargo, a diferencia de los esquemas mencionados mas arriba, no incluiremos la regularización en los parámetros del modelo. En su lugar definiremos un nuevo conjunto de variables de regularización (slack variables) las cuales serán introducidas de dos modos diferentes. En primer lugar modificando la función objetivo. Para ello definiremos una nueva función objetivo que no solo incluya la entropía, sino que agregue un término proporcional a la suma de las variables de regularización, de un modo similar al utilizado en las máquinas de soporte vectorial [2]. En segundo lugar cada una de las variables de regularización será sumada a la correspondiente distribución empírica $\tilde{p}(x, y)$, con el objetivo de permitir que la distribución de máxima entropía $p(y/x)$ no satisfaga exactamente los datos de entrenamiento, sino que se permita un cierto grado de libertad con la esperanza de que esto disminuya el over-fitting del modelo.

El resto del trabajo se organiza del siguiente modo: En la sección 2 formularemos el modelo, en la sección 3 evaluaremos el modelo comparando la perplejidad

y la tasa de reconocimiento de error (WER) con el modelo de Kneser-Ney. En la sección 4 analizaremos el comportamiento del modelo cuando se varían los principales parámetros del mismo con el fin de establecer criterios de optimización. Finalmente en la sección 5 daremos algunas conclusiones.

2. Máxima Entropía Regularizada

Aunque este modelo puede ser generalizado a N-gramas, por razones claridad en la explicación y de costo computacional en las simulaciones, limitaremos el análisis siguiente al caso de bi-gramas. Tomamos de un texto de N palabras de entrenamiento, un conjunto de bi-gramas $\{(x_1, y_1), \cdots, (x_m, y_m), \cdots, (x_M, y_M)\}$ de M bi-gramas diferentes, y $\{y_1, \cdots, y_r, \cdots, y_R\}$ de R uni-gramas diferentes. Nos interesa tener estadísticas de frecuencias de uni-gramas y bi-gramas, por lo tanto definimos las siguientes funciones indicadoras

$$f_m^B(x, y) = \begin{cases} 1 & \text{para} \quad x = x_m,\ y = y_m \\ 0 & \text{en otro caso} \end{cases} \tag{3}$$

para el caso de bi-gramas

$$f_r^U(x, y) = \begin{cases} 1 & \text{para} \quad y = y_r \\ 0 & \text{en otro caso} \end{cases} \tag{4}$$

Para el caso de uni-gramas. También definimos la distribución empírica de un modo similar al que se utiliza en los modelos tradicionales de backing-off [12]:

$$\tilde{p}(x, y) = \begin{cases} \tilde{P}_{ml}(x, y) + \xi(x, y) & \text{para } N(x, y) > 0 \\ \delta(x)\tilde{P}_{ml}(y) & \text{para } N(x, y) = 0 \end{cases} \tag{5}$$

donde $\tilde{P}_{ml}(x, y) = N(x, y)/N$ es el estimador de máxima verosimilitud de probabilidad para el bi-grama (x, y), $N(x, y)$ es la frecuencia de ocurrencia de dicho bi-grama, $\delta(x)$ es un factor de normalización, y $\tilde{P}_{ml}(y)$ es la distribución menos específica (es decir un estimador de uni-grama para el caso de bi-gramas). En este trabajo usamos el estimador de máxima verosimilitud, entonces $\tilde{P}_{ml}(y) = N(y)/N$, pero se podrían usar distribuciones mas complejas. El Termino $\xi(x, y)$ se define por:

$$\xi(x, y) = \begin{cases} \xi_m & \text{para} \quad x = x_m,\ y = y_m \\ 0 & \text{en otro caso} \end{cases} \tag{6}$$

Este término $\xi(x, y)$ se puede interpretar como la masa de probabilidad en la cual el estimador de probabilidad empírico $\tilde{p}(x, y)$ difiere del estimador de máxima verosimilitud. Si $\tilde{p}(x, y)$ se interpreta como en el modelo clásico de backing-off, entonces $-N \cdot \xi(x, y)$ puede ser interpretado como un descuento variable que se resta a cada bi-grama que ocurre en el texto de entrenamiento. Entonces el conjunto de M bi-gramas diferentes del texto de entrenamiento definirá un conjunto de M variables $\Xi = \{\xi_1, \cdots, \xi_m, \cdots, \xi_M\}$. Llamaremos a estas variables

variables de regularización ξ_m. Se pueden interpretar como la cantidad en la que difieren $\tilde{P}(x_m, y_m)$ y $\tilde{P}_{ml}(x_m, y_m)$ (ver ecuación (5)). En lugar de tomar $\tilde{P}_{ml}(x_m, y_m)$ como un estimador de la verdadera probabilidad para el bi-grama m, le damos al modelo algo de libertad para elegir un mejor estimador, esperando minimizar el over-fitting. También introducimos las variables de regularización ξ_m en la función objetivo $H(p)$ redefiniéndola como:

$$H(p, \Xi) = -\sum_{x,y} \tilde{p}(x)p(y/x)\log p(y/x) - \gamma \sum_{m=1}^{M} \xi_m. \tag{7}$$

El primer término representa la entropía condicional utilizada en el modelo tradicional. El segundo término, como dijimos, está inspirado en el modelo de máquinas de soporte vectorial (SVM) [2], éste término impone una especie de penalización a la función objetivo si la suma de sus variables de regularización ξ_m tiende a crecer en exceso. El modelo debe satisfacer además las siguientes restricciones de bi-gramas y uni-gramas:

$$E_p\{f_m^B\} = E_{\tilde{p}}\{f_m^B\} \text{ para } m = 1, \cdots, M \tag{8}$$

$$E_p\{f_r^U\} = E_{\tilde{p}}\{f_r^U\} \text{ para } r = 1, \cdots, R \tag{9}$$

donde $E_x\{f\}$, como se dijo en la sección 1 es la media de la función f con respecto a la distribución x. También imponemos restricciones sobre los valores mínimos y máximos de ξ_m

$$\varepsilon \leq \xi_m \leq \theta \text{ para } m = 1, \cdots, M \tag{10}$$

Donde $-1 \leq \varepsilon, \theta \leq 1$. Finalmente como $p(y/x)$ es una distribución de probabilidades, $p(y/x)$ debe satisfacer:

$$\sum_y p(y/x) = 1. \; \forall x \tag{11}$$

Como dijimos, el problema se puede resolver en el marco del paradigma de optimización restringida. La solución al problema primal es la distribución:

$$p^*(y/x) = \frac{exp\left(-\sum_{m=1}^{M} \lambda_m^B f_m^B(x,y) - \sum_r \lambda_r^U f_r^U(x,y)\right).}{Z_\Lambda(x)} \tag{12}$$

y el conjunto de ecuaciones para $m = 1, \cdots, M$

$$-\gamma = -\lambda_m^B - \sum_{r=1}^{R} \lambda_r^U C(r,m) + \lambda_m^S - \lambda_m^I. \tag{13}$$

La matriz $C \in \mathbb{R}^{MxR}$ se puede demostrar que vendrá dada por:

$$C(r,m) = \begin{cases} 1 & \text{para } N(x(m), y_r) > 0 \\ -\dfrac{\tilde{P}_{ml}(y_r)}{\sum\limits_{y/N(x(m),y)=0} \tilde{P}_{ml}(y)} & \text{para } N(x(m), y_r) = 0 \end{cases} \tag{14}$$

Los parámetros λ^S y λ^I están asociados a las restricciones impuestas por las variables de regularización ξ_m. Los parámetros λ^B y λ^U están asociados a las restricciones de bi-gramas y uni-gramas. $Z_\Lambda(x)$ es un factor de normalización dado por:

$$Z_\Lambda(x) = \sum_x exp\left(-\sum_{m=1}^{M} \lambda_m^B f_m^B(x,y) - \sum_r \lambda_r^U f_r^U(x,y)\right) \tag{15}$$

Los parámetros se hallan maximizando la función dual:

$$\mathbb{L}(\lambda^B, \lambda^U, \lambda^S, \lambda^I) = H(p^*, \xi_1^*, \cdots, \xi_m^*, \cdots, \xi_M^*) \tag{16}$$

donde p^* y $\xi_1^*, \cdots, \xi_m^*, \cdots, \xi_M^*$ son los valores óptimos encontrados como solución del problema de optimización restringida, es decir el problema primal. Reemplazado (12) y (13) en (7)

$$\begin{aligned}\mathbb{L}(\lambda^B, \lambda^U, \lambda^S, \lambda^I) = &-\sum_x \tilde{p}(x)p^*(y/x)\log Z_\Lambda(x) - \\ &-\sum_m \lambda_m^B \tilde{P}_m - \sum_r \lambda_r^U \tilde{P}_r - \theta\sum_m \lambda_m^S + \varepsilon\sum_m \lambda_m^I\end{aligned} \tag{17}$$

Con $\tilde{P}_m = \tilde{P}_{ml}(x_m, y_m)$ y $\tilde{P}_r = \tilde{P}_{ml}(y_r)$ como se definió arriba, y las restricciones:

$$\begin{aligned}&\lambda_m^S \geq 0 \text{ para } m = 1, \cdots, M \\ &\lambda_m^I \geq 0 \text{ para } m = 1, \cdots, M \\ &-\gamma = -\lambda_m^B - \textstyle\sum_{r=1}^{R} \lambda_r^U C(r,m) + \lambda_m^S - \lambda_m^I\end{aligned} \tag{18}$$

3. Resultados experimentales

El problema algorítmico consiste en maximizar la función (17) respecto de los parámetros λ sujeta a las restricciones (18). El problema es un problema de optimización no lineal, esto puede verse en el primer término de la ecuación (17), ya que $Z_\Lambda(x)$ es una suma de exponenciales en los parámetros mas el agregado del logaritmo. Sin embargo, aún con una función objetivo altamente no lineal, el problema es convexo. Debido a que cada bigrama y cada unigrama originan un parámetro λ_m y λ_r respectivamente, la cantidad de parámetros del problema crece con la cantidad de bigramas y unigramas que se quieran modelizar.

Aún no hemos desarrollado algoritmos específicos para este modelo. La implementación del mismo se hizo usando una herramienta de libre distribución para optimización convexa desarrollada en pyton llamada CVXOPT [2]. Este paquete permite el desarrollo de modelos convexos generales en una forma muy directa pero a un costo computacional extremadamente alto. Sin embargo, aunque esta opción nos obligó a probar el modelo con una base de datos pequeña, nos permitió tener una mayor versatilidad para modificar y optimizar fácilmente los parámetros del modelo.

[2] disponible en el sitio "http://abel.ee.ucla.edu/cvxopt'

Una medida frecuentemente usada para determinar la bondad de un modelo de lenguaje es la perplejidad. Dado un texto cualquiera $X = (x_1, x_2, \cdots, x_N)$ donde x_i son las palabras que componen el mismo, la perplejidad se define como: $PP = 2^{-\sum_{x=1}^{N} \frac{1}{N} log(P(x))}$. Una perplejidad baja indica que el modelo es capaz de predecir con mayor exactitud las palabras de un texto dado. Si bien en algunas aplicaciones como reconocimiento automático de habla una baja perplejidad del modelo de lenguaje no se corresponde necesariamente con una disminución en el índice de reconocimiento de palabras del sistema, por dicho motivo también se utiliza como medida de la bondad del modelo el porcentaje de palabras reconocidas correctamente por un sistema de reconocimiento de habla WER.

Los experimentos fueron realizados usando el texto extraido de la base de datos en español Latino-40 disponible a través de LDC [3]. También se construyó usando dicha base de datos un sistema de reconocimiento de habla (ASR) usado para evaluar el WER. Ambas medidas, el WER y la perplejidad, fueron evaluadas usando diferentes tamaños del texto de entrenamiento y vocabulario. El tamaño del corpus no nos permitió hacer una validación cruzada, sin embargo se reservó una parte de los datos de test para el ajuste de los parámetros (development set), con lo que los resultados mostrados a continuación corresponden a datos que nunca fueron utilizados ni en el entrenamiento ni en el ajuste de parámetros. La tabla 1 muestra resultados de perplejidad usando un conjunto de test de 744 palabras y la tabla 2 muestra el WER. En ambos casos usamos el modelo de Kneser-Ney modificado [5] como linea de base para contrastar nuestro sistema. Como vemos en la tabla1 obtuvimos mejoras en todos los tamaños

Cuadro 1. Perplejidad para el modelo de Kneser-Ney modificado y Máxima Entropía. Tamaño del texto de test: 744 palabras

Tamaño-texto	Tamaño-vocab	Modif. KN	Max-Ent
2242	950	96.8001	86.2985
4492	1642	129.479	114.783
8988	2724	130.061	108.057
17967	3923	105.67	90.4583

Cuadro 2. WER para los modelos de Kneser-Ney modificado y Máxima Entropía

Tamaño-texto	Tamaño-vocab.	Modif. KN	Max-Ent
2242	950	19.13	19.08
4492	1642	29.23	32.17
8988	2724	49.51	50.28
17967	3923	60.51	61.72

[3] Linguistic Data Consortium "http://www.ldc.upenn.edu/"

del texto. Para el mejor de los casos (17967 palabras) mejoramos la perplejidad en aproximadamente un 14,4 por ciento. En la tabla 2 la única mejora significativa en el WER fué para un tamaño de la base de datos 17967 palabras. En este caso obtuvimos una mejora cercana al 2 por ciento. En la siguiente sección discutiremos como varía la perfomance con la elección de los parámetros.

4. Discusión

Con la idea de ganar conocimiento sobre nuestro modelo, mostraremos el comportamiento de las variables de regularización ξ_m con los parámetros γ, ε y θ. Primeramente fijaremos $\gamma = 0$, y veremos las variaciones de ξ_m cuando los límites ε y θ son modificados. Luego analizaremos el comportamiento de los mismos cuando γ es modificado. Puesto que las variables de regularización tienen valores muy bajos, las normalizaremos multiplicando dichas variables y los límites ε y θ por N. Entonces al referirnos a ξ_m, ε y θ implícitamente asumimos que nos referimos a $N \cdot \xi$, $N \cdot \varepsilon$ y $N \cdot \theta$ respectivamente. Esto no solo facilita la lectura de las figuras, sino que permite una comparación directa entre las variables ξ_m y los descuentos usados en [12].

4.1. Análisis de los límites ε y θ

Como se puede ver de (10), ε y θ restringen los valores mínimos y máximos respectivamente de las variables ξ_m. En el caso particular en que $\varepsilon = \theta$ forzamos a las variables de regularización ξ_m a ser iguales. En este caso tendremos un descuento constante (5), y estaremos en el caso del modelo de backing-off tradicional. Por ejemplo si elegimos $\tilde{P}_m(y)$ como en [12] y $\varepsilon = \theta = -\left(\frac{n_1}{n_1+2n_2}\right)$ donde n_1 y n_2 son el número de palabras en el texto de entrenamiento con frecuencia uno y dos respectivamente, obtenemos el modelo de Kneser y Ney como un caso particular de nuestro modelo. La Fig. 1 muestra un histograma de ξ_m donde hemos seteado $\varepsilon = -1$ y $\theta = 1$ para el caso $N = 2242$. Esto corresponde al caso cuando las variables de regularización ξ_m son dejadas completamente libres. En este caso el histograma muestra que los valores de ξ_m tienden a agruparse alrededor de 1, lo cual difiere de $\frac{n_1}{n_1+2n_2} = 0{,}91$ hallado en el modelo de Kneser-Ney . Vemos también que no se encuentran valores significativos de ξ_m por debajo de 0,5, y que tampoco hay datos significativos por encima de 2,1.

Veremos ahora el efecto de decrementar los valores de los límites de ξ_m. Fijamos el valor de $\varepsilon = 0$ ya que como vimos en la Fig. 1 no tenemos valores de $\xi_m \leq 0$. También limitamos el análisis de $\theta \leq 5$ por la misma razón. La Fig. 2 muestra la variación de θ con la perplejidad. Como vemos, el mínimo valor de perplejidad se alcanza para un valor de casi 0,7. Este valor no es ni el máximo mostrado en el histograma de la Fig. 1 ni el valor óptimo $\frac{n_1}{n_1+2n_2}$ usado en [12]. La Fig. 3 muestra los histogramas de ξ_m para θ variando desde 0,6 (al principio de la figura) hasta 0,9 (final de la figura). Dichos casos corresponden como se puede ver en la figura Fig. 2, a los valores mas bajos de perplejidad. En todos los casos los valores de ξ_m tienden a agruparse cerca del límite θ, y prácticamente no

hay dispersión. Si θ se varía desde 1,0 (el histograma no se muestra por razones de espacio) los valores de ξ_m tienden a agruparse cerca de 1,0, y tenemos considerablemente mas variabilidad que en la Fig. 3 3. Nótese que como puede verse en la Fig. 2 en este caso, la perplejidad crece considerablemente. Podemos concluir entonces que si se quieren obtener valores de perplejidad bajos deberíamos limitar ξ_m a valores menores que 1,0. En cuyo caso, casi todos los ξ_m tenderán a agruparse en el valor límite de θ.

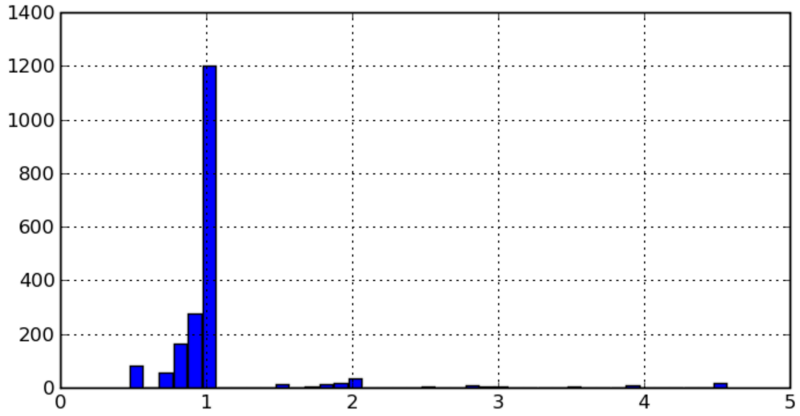

Figura 1. Histograma de ξ_m.

4.2. Análisis de γ

Veremos a continuación qué variaciones del parámetro γ pueden decrementar aún mas la perplejidad. Como dijimos el parámetro γ en la ecuación (7) que multiplica a $\sum_m \xi_m$ está inspirado en el modelo de SVM descripto en [2]. Fijando γ también estamos fijando el valor máximo que puede tomar $\sum_m \xi_m$. A medida que γ tienda a infinito, esta suma tenderá a cero. Esto significa que el conjunto de las variables de regularización ξ_m también deberá tender a cero ya que como vimos son todas positivas. Esto a su vez hará que no haya ningún descuento en la ecuación (5) y que la probabilidad empírica sea igual al estimador de máxima verosimilitud. En este caso la perplejidad tenderá a incrementarse hasta infinito. Sin embargo, para ciertos valores de γ hemos visto experimentalmente que el segundo término de la ecuación (7) *empuja* la perplejidad a valores mas bajos que cuando este término no está presente ($\gamma = 0$). Esto se puede ver en la Fig. 4 donde hemos graficado la perplejidad para diferentes valores del límite superior θ, para valores de γ variando de -5 a 5. La figura muestra que para valores de θ variando desde 0,75 hay un pico negativo que siempre ocurre para valores de γ

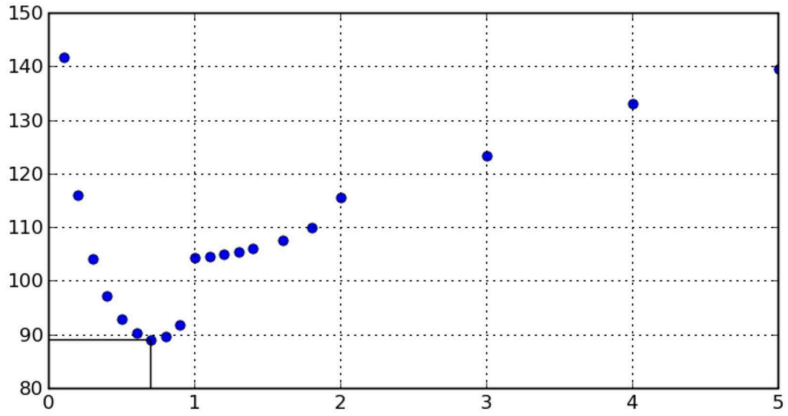

Figura 2. Perplejidad en función de θ.

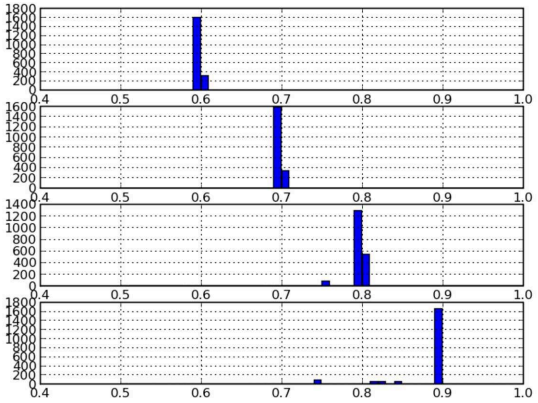

Figura 3. Histograma de ξ_m para $0{,}6 \leq \theta \leq 0{,}9$

distinto de cero, y para γ variando desde $-1{,}5$ a -2. Para valores de θ mayores que 0,85, aún tenemos un pico negativo en la perplejidad, pero esto ya no nos es útil ya que como se puede ver en la Fig. 2 la perplejidad crece rápidamente y ningún valor de γ parece compensar tal incremento.

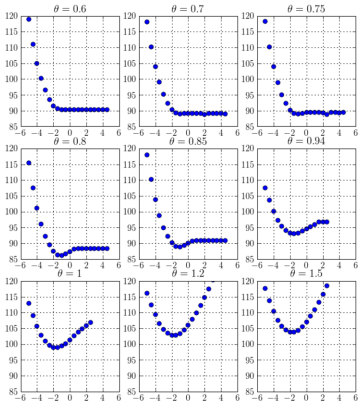

Figura 4. Perplejidad en función de γ para valores diferentes de θ

5. Conclusiones

Hemos presentado un nuevo modelo para el suavizado de N-gramas. Por razones de claridad y computacionales nos hemos concentrado en el caso de bi-gramas pero es posible una extensión al caso de N-gramas. El modelo está basado en el paradigma de optimización restringida de máxima entropía con dos modificaciones que introducen regularización de dos maneras diferentes. Por un lado modificamos la función objetivo a maximizar. En lugar de maximizar la entropía como es lo habitual, maximizamos una función objetivo que es la suma de la entropía mas un término de penalización proporcional a un conjunto de variables de regularización. Por otro lado introdujimos el mismo conjunto de variables de regularización en las ecuaciones de restricción, asumiendo que la densidad de probabilidad empírica es igual a la suma de la densidad estimada de máxima verosimilitud y una variable de regularización ξ_m. Cada una de estas variables puede ser interpretada como la diferencia entre la probabilidad de máxima verosimilitud de un bi-grama y el valor que efectivamente usamos en la estimación empírica de esa probabilidad. Además ξ_m puede ser interpretado como un parámetro de descuento generalizado del tipo de los usados en los modelos de backing-off con diferentes valores para cada bi-grama.

Los resultados mostraron una mejora significativa de 14,4 por ciento en la perplejidad y 2 por ciento en el WER para la base de datos Latino-40. Dichas mejoras son importantes ya que el contraste se realizó contra el mejor modelo de suavizado conocido. El próximo paso sería el desarrollo de un algoritmo específico para este modelo que permita la implementación de modelos de trigramas y el el testeo del modelo con bases de datos de mayor tamaño.

Referencias

1. Berger, A.L., Pietra, S.A.D., Pietra, V.J.D.: A maximum entropy approach to natural language processing. Computational Linguistics 22, 39–71 (1996)
2. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2(2), 121–167 (1998)
3. Chen, S.F.: Scaling shrinkage-based language models. In: In Proceedings of ASRU (2009)
4. Chen., S.F.: Shrinking exponential language models. In Proceedings of NAACL HLT (2009)
5. Chen, S.F., Rosenfeld, R.: A survey of smoothing techniques for me models. IEEE Transctions on speech and Audio Processing, Vol. 8(1), January 2000 8-1, 37–50 (2000)
6. Dudik, M., Phillips, S.J., Schapire, R.E.: Performance guarantees for regularized maximum entropy density estimation. In: Proceedings of the 17th Annual Conference on Computational Learning Theory (2004)
7. Estienne, C., Sanchis, A., Juan, A., Vidal, E.: Maximum entropy models for speech confidence estimation. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2008). pp. 4421–4424. Las Vegas, EEUU (Apr 1-4 2008), iSSN: 1520-6149
8. Goodman, J.: Exponential prios for maximum entropy models. North American ACL (2004)
9. Jaynes, E.T.: Information theory and statistical mechanics, I and II. Physical Reviews 106 and 108, 620–630 and 171–190 (1957)
10. Katz, S.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. Acoustics, Speech and Signal Processing, IEEE Transactions on 35(3), 400 – 401 (mar 1987)
11. Kazama, J., Tsujii, J.: Maximum entropy models with inequality constraints: A case study on text categorization. Machine Learning 60, 159–194 (2005)
12. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. vol. 1, pp. 181 –184 vol.1 (may 1995)
13. Martin, S., Ney, H., Hamacher, C.: Maximum entropy language modeling and the smoothing problem. IEEE Trans. on Speech and Audio Proc., VOL. 8, NO. 5, SEPTEMBER 2000 8-5, 626–632 (2000)
14. Ney, H., Essen, U., Kneser, R.: On the estimation of 'small' probabilities by leaving-one-out. Pattern Analysis and Machine Intelligence, IEEE Transactions on 17(12), 1202 –1212 (dec 1995)