PEDECIBA Informática
Instituto de Computación – Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay

Reporte Técnico RT 08-04

Evaluación de modelos de ngramas construidos de derecha a izquierda

Guillermo Moncecchi

2008

Evaluación de modelos de ngramas construidos de derecha a izquierda
Moncecchi, Guillermo
ISSN 0797-6410
Reporte Técnico RT 08-04
PEDECIBA
Instituto de Computación – Facultad de Ingeniería
Universidad de la República

Montevideo, Uruguay, 2008

Evaluación de modelos de ngramas construidos de derecha a izquierda

Guillermo Moncecchi
gmonce@fing.edu.uy

Instituto de Computación
Facultad de Ingeniería, Universidad de la República
Uruguay

Mayo de 2008

Resumen En este trabajo se presenta la evaluación de varios modelos de ngramas en dos escenarios simétricos: en el primero, los ngramas del modelo se construyen leyendo el corpus de izquierda a derecha, y en el segundo lo hacen de derecha a izquierda. En cada caso, se estudia su rendimiento, utilizando la medida de perplejidad, considerando diferentes opciones de cut-off, reducción de vocabulario e interpolación con modelos de clases. Los resultados, aunque no concluyentes, parecen indicar que los valores de perplejidad son menores para el segundo escenario.

Yo estoy al derecho, dado vuelta estás vos...
L. Prodan, "El cieguito volador"

1. Introducción

La utilización de enfoques probabilísticos para el modelado del lenguaje natural ha tomado un nuevo impulso en los últimos años, sin duda provocado por la disponibilidad de grandes cantidades de información en formato electrónico. Los modelos del lenguaje natural[3,1] tienen por objetivo calcular, dada una secuencia de palabras (que podrían ser una oración o una frase) su probabilidad de aparición en un lenguaje. Para resolver este problema, los enfoques probabilistas no utilizan, a diferencia de modelos más tradicionales, conocimiento lingüístico profundo, sino que parten de la información obtenida a partir de corpus de gran tamaño que suponen representativos del lenguaje y, utilizando el principio de máxima verosimilitud[4], asignan las probabilidades correspondientes según las frecuencias de aparición en el corpus. Por supuesto, sobre esta base existen diversas elaboraciones que pueden incorporar en mayor o menor medida información lingüística o de otro tipo para afinar los resultados.
El curso *Construcción de modelos probabilistas del lenguaje natural*, dictado por el Prof. Gustavo Crispino en abril de 2007, presentó los fundamentos y técnicas para la construcción de estos modelos. En este trabajo, parte del curso, se construyen varios modelos originados en corpus del

idioma español de México, y se evalúa su entropía sobre dos corpus indepedientes. Para cada modelo construido, se observa cómo afecta (desde el punto de vista del rendimiento y del tamaño) la aplicación de *cut-offs* sobre los ngramas obtenidos a partir del corpus, la limitación del vocabulario utilizado para construir el modelo, y su interpolación con modelos de clases, construidos utilizando métodos estadísticos de *clustering* sobre el mismo corpus de entrenamiento.
Los métodos basados en ngramas (como los que se estudian en este documento), asumen que solamente las $n-1$ palabras anteriores a la actual tienen efecto sobre su probabilidad de aparición (hipótesis markoviana). En un lenguaje *ergódico*[1], esto se cumple si n es suficientemente grande, y es equivalente utilizar contextos izquierdos o derechos para estimar la probabilidad de aparición de una palabra. Pero, para los ngramas utilizados en la práctica, por lo general n no toma valores mayores a 4 o 5. Por lo tanto, vale hacerse la pregunta: para un lenguaje dado, ¿es igual utilizar las 3 o 4 palabras a la izquierda de cada palabra para estimar su probabilidad, que utilizar las 3 o 4 a la derecha? En este trabajo se pretende estudiar cómo se comportan dos modelos que son iguales excepto por el hecho de que uno de ellos se basa en ngramas construidos leyendo el corpus de izquierda a derecha, y el otro lo hace de derecha a izquierda.

2. Escenarios

Para realizar el trabajo se construyen dos escenarios idénticos, incluyendo cada uno varios procesos con el objetivo de construir los modelos de lenguaje y evaluar su perplejidad sobre el corpus de prueba. En el primer escenario, los ngramas se construyen sobre el corpus de prueba original (preprocesado para marcar oraciones, eliminar puntuaciones y pasar todas las palabras a mayúsculas), mientras que en el segundo se utiliza para entrenar y evaluar el mismo corpus, pero donde las palabras de los textos se han invertido.
En cada escenario se plantea construir los siguientes modelos, evaluarlos sobre el corpus correspondiente, y comparar los resultados.

- Construir un modelo de trigramas sobre el corpus de prueba, utilizando todas las palabras y todos los ngramas que aparezcan en el corpus.
- Construir modelos cortando aquellos ngramas que aparezcan menos de un determinado número de veces(*cutting-off*)
- Eliminar de los ngramas aquellas palabras que aparezcan menos de un número determinado de veces en el corpus de prueba. Mapear todas esas palabras a una clase al obtener los ngramas y construir el modelo.
- Generar modelos n-clases, utilizando una clasificación rígida y un criterio contextual y estadístico[3], con diferente número de clases, e interpolarlos con los obtenidos anteriormente.

Finalmente, se utiliza un corpus de evaluación completamente diferente al original, repitiendo el proceso de evaluación, para validar los resultados.

3. Implementación y Resultados.

Para la implementación de los escenarios se utiliza un corpus de sesiones del senado mexicano, de aproximadamente 72 millones de palabras y un tamaño de 400 Mb, el cual se divide en un corpus de entrenamiento de 60 millones de palabras y 12 millones de palabras para evaluación. Para la validación de los resultados, se utiliza un corpus periodístico, también de México, de 9 millones de palabras.
En cuanto a herramientas, el escenario se construye sobre una plataforma unix, utilizando la herramienta HTK[6] para la construcción de modelos y evaluación de perplejidad, y el lenguaje de programación Perl para el procesamiento de los archivos de texto.
A continuación, se muestran los resultados para cada una de las etapas:

3.1. Escenario 1: al derecho

Como ya se mencionó, en este escenario se utilizan en todos los casos los contextos izquierdos de las palabras para estimar su probabilidad de aparición.

Modelos utilizando *cut-offs* Los primeros modelos construidos (luego de generar la base de ngramas a partir del corpus de entrenamiento), son modelos ngramas (utilizan las apariciones de palabras para estimar probabilidades) donde se han especificado diferente umbrales de aparición de ngramas para considerar su uso en el modelo. En el cuadro 1 se muestran los tamaños en Mb y las perplejidades obtenidas al evaluar su comportamiento sobre el corpus de evaluación. Los índices númericos en el nombre indican los cut-offs aplicados para trigramas y bigramas, respectivamente.

Cuadro 1. Modelos de trigramas con diferentes cut-offs

Modelo	Tam (Mb)	Perplejidad
tg-1-1	112.8	60.85
tg-2-2	68.3	69.37
tg-3-3	50.2	75.00

Limitación del vocabulario En la segunda etapa, se eliminan las palabras menos frecuentemente utilizadas. En este caso, se decide quitar las que aparecen menos de treinta veces en el corpus original.[1] Esto produce una reducción del vocabulario, de 169.213 palabras a 28.883.

[1] La idea era original era limitar menos el vocabulario pero limitaciones técnicas de la herramienta no permitían trabajar con vocabularios demasiado grandes

Los resultados obtenidos para un cut-off de 1-1 se muestran en el cuadro 2. Para el caso del vocabulario reducido, se muestran los resultados de perplejidad cuando se utilizan los contextos con palabras fuera del vocabulario para la estimación, y cuando esto no se hace.

Cuadro 2. Modelos de trigramas con vocabulario completo y vocabulario reducido

Modelo	Tam (Mb)	Perplejidad	Varianza	Palabras OOV
tg-1-1	112.8	60.85	12.65	0.11 %
tg-b-1-1	71.0	53.63	11.20	1.35 %
tg-b-1-1(oov)	71.0	55.15	11.22	1.35 %

Lo primero que puede observarse es que la perplejidad en todos los casos es muy baja. Segun los resultados presentados por Rouko[2], la perplejidad del idioma inglés es de 247, bajando a valores cercanos al 60 en áreas específicas, como la medicina. En este caso, al tratarse de sesiones del senado, una baja perplejidad parece razonable, ya que la temática se supone que es bastante uniforme. Al validar los resultados sobre un corpus periodístico, encontraremos perplejidades mucho más altas.

Por otra parte, un resultado que puede parecer sorprendente es que la perplejidad *baja* al reducir el vocabulario. Esto probablemente se deba a que, al aparecer tan pocas veces, las palabras menos frecuentes participan en trigramas que tendrán baja probabilidad asignada. Al agruparlas en una sola clase, la probabilidad asignada a los trigramas en que aparecen es mayor, y entonces la pérdida de información que se produce al utilizar la clase (y no la palabra) para predecir se ve compensada. Si se observa la tabla 3, donde aparecen las estadísticas de acceso al modelo —cantidad de trigramas que existían en el modelo, cantidad que debió estimarse utilizando bigramas *backoff*, etc—, puede verse que hay más casos en los que el trigrama ya estaba "visto" en el corpus de prueba. Esto es, hay casos donde el trigrama no aparecía, pero la triclase (considerando la clase de las palabras OOV) sí, y por lo tanto no es necesario hacer *backoff*.

Cuadro 3. Acceso a trigramas con vocabulario completo y vocabulario reducido

Modelo	solicitados	exactos	backed	no disp
tg-1-1	3282492	78.8 %	20.5 %	0.7 %
tg-b-1-1	2956665	82.3 %	17.7 %	0.0 %
tg-b-1-1(oov)	3025879	82.4 %	17.6 %	0.0 %

Interpolación con modelos n-clase En esta fase, se construyen modelos n-clase estadísticos[3,5] agrupando las palabras que aparecen en contextos iguales en la misma clase. Estos modelos tienen un tamaño mucho menor a los modelos ngramas, pero al calcular la entropía cruzada se obtienen valores mucho más altos de perplejidad. Sin embargo, si estos modelos son interpolados con los modelos obtenidos en la etapa anterior, puede verse en el cuadro 4 que las perplejidades obtenidas son menores. En ese cuadro se muestran las perplejidades obtenidas utilizando el modelo de la parte anterior, un modelo de 450 clases (valor recomendado dada la cantidad de palabras en el documento de HTK[6]), y modelos de 225 y 900 clases.

Cuadro 4. Modelos de triclase con vvocabulario reducido

Modelo	Tam (Mb)	Perplejidad	Varianza	Palabras OOV
tg-b-1-1	71.0	53.63	11.20	1.35 %
tg-c-450-1-1	20.2	113.83	10.08	1.35 %
tg-1c-450i-1-1	20.2+71.0	53.13	9.14	1.35 %
tg-1c-225i-1-1	11.3+71.0	54.61	8.97	1.35 %
tg-1c-900i-1-1	32.2+71.0	52.16	9.41	1.35 %

Puede verse que a medida que crece el número de clases (y por tanto el tamaño del modelo), la perplejidad va bajando, y que con 225 clases la perplejidad es mayor que la obtenida directamente por el modelo de ngramas. Sería interesante investigar la curva de descenso en relación al crecimiento de la cantidad de clases.
Sobre la implementación, vale observar que el proceso de generación de las clases es costoso en tiempo de cálculo, y crece en forma muy importante cuando crece el número de clases. Por ejemplo, si para generar 450 clases el proceso lleva 3 horas, para generar 900 toma 8 horas.

3.2. Escenario 2: al revés

Como se dijo, en el segundo escenario se consideran los mismos modelos, sobre el mismo corpus, pero invirtiendo el orden de las palabras. Por lo tanto, si en el corpus origen aparecen la oración

```
<s> EL SEÑOR DIPUTADO SE ENCUENTRA ENFERMO </s>
```

Los modelos se construirán sobre un corpus que tiene las oraciones invertidas:

```
<s> ENFERMO ENCUENTRA SE DIPUTADO SEÑOR EL </s>
```

y donde además el orden de las oraciones se invirtió. Por supuesto, el corpus sobre el que se evalúa también tiene los textos invertidos.
No se va a entrar en detalle de cada una de las secciones mostradas en la sección anterior, pero el comportamiento "relativo" de los modelos se mantiene. Esto es, las variaciones en perplejidad y tamaños al realizar cut-offs, modificar vocabularios y generar diferente número de clases se mantiene invariable respecto a lo observado en la sección anterior. Sí interesa *comparar* los resultados obtenidos entre los modelos simétricos de cada uno de los escenarios. La tabla 5 muestra un resumen, cuyos valores se explican a continuación.

Cuadro 5. Comparación de perplejidades en los dos escenarios

Modelo	Esc.	Tam (Mb)	Perplejidad	Varianza	Palabras OOV
tg-1-1	1	112.8	60.85	12.65	0.11 %
	2	112.9	60.40	15.63	0.11 %
tg-b-1-1	1	71.0	53.63	11.20	1.35 %
	2	71.0	52.64	14.15	1.35 %
tg-1c-225i-1-1	1	11.3+71.0	54.61	8.97	1.35 %
	2	11.2+71.0	53.86	11.98	1.35 %
tg-1c-450i-1-1	1	20.1+71.0	53.13	9.14	1.35 %
	2	20.1+71.0	52.37	12.16	1.35 %
tg-1c-900i-1-1	1	32.2+71.0	52.16	9.41	1.35 %
	2	32.2+71.0	51.36	12.44	1.35 %

En la primer columna aparecen los modelos evaluados: se consideró el modelo original, el modelo con el vocabulario recortado, y la interpolación de éste con los modelos de clases. En la segunda se muestra el escenario sobre el que se calculó la perplejidad del corpus de prueba, y luego los valores usuales de tamaño, perplejidad y varianza.
Viendo los resultados, pueden hacerse algunas observaciones

- Los tamaños de los modelos generados son prácticamente iguales. Analizándolo un momento, este comportamiento es obvio: los trigramas aparecen igual, pero invertidos. Las diferencias de tamaño probablemente se deban a cuestiones de redondeo de la herramienta en el cálculo de las probabilidades
- La perplejidad es ligeramente más baja cuando se considera los contextos derechos, pero la varianza es también consistentemente mayor

Con esos valores es muy difícil sacar alguna conclusión. Para intentar agregar información, se prueban los modelos contra un nuevo corpus de evaluación, esta vez sobre una temática totalmente diferente (periodística), por lo que se espera que los valores de perplejidad sean diferentes (y bastante mayores). Efectivamente, en la tabla 6 pueden verse los resultados.
El cambio de corpus ha traído algunas variaciones en las observaciones. Puede verse que la perplejidad es mucho mayor, lo cual se justifica por

Cuadro 6. Comparación de perplejidades en los dos escenarios, segundo corpus

Modelo	Esc.	Tam (Mb)	Perplejidad	Varianza	Palabras OOV
tg-1-1	1	112.8	555.63	22.42	1.93 %
	2	112.9	530.99	25.32	1.93 %
tg-b-1-1	1	71.0	331.97	17.40	6.57 %
	2	71.0	314.5	20.83	6.57 %
tg-1c-225i-1-1	1	11.3+71.0	253.56	13.37	6.57 %
	2	11.2+71.0	244.50	16.88	6.57 %
tg-1c-450i-1-1	1	20.1+71.0	253.20	13.67	6.57 %
	2	20.1+71.0	243.14	15.28	6.57 %
tg-1c-900i-1-1	1	32.2+71.0	258.40	14.14	6.57 %
	2	32.2+71.0	247.63	17.71	6.57 %

la diferencia de temática. Análogamente, la cantidad de palabras fuera del vocabulario ha crecido mucho, llegando a un 6.57 % luego de recortar el vocabulario. Además, la perplejidad es mucho más sensible a la interpolación con modelos de clases, cayendo abruptamente cuando se interpola.
Sin embargo, el comportamiento en ambos escenarios se confirma: la perplejidad es menor en todos los casos en el segundo escenario.

4. Conclusiones

En este trabajo se presentaron diferentes modelos de ngramas generados sobre un mismo corpus, y se evaluó su entropía cruzada respecto a un corpus de evaluación, utilizando la medida de perplejidad. El mismo trabajo fue realizado en dos escenarios, con el objetivo de comparar el comportamiento de los modelos cuando los ngramas se construyen de derecha a izquierda, en lugar de la estrategia usual de hacerlo de izquierda a derecha.
Del trabajo puede concluirse que los modelos probabilistas (en sus diferentes sabores) deben ajustarse cuidadosamente (tamaño de los corpus, tamaño de los modelos, tamaño del vocabulario, precisión deseada en la estimación), para que se adapten óptimamente a la realidad sobre la que trabajan
Si bien los resultados no son concluyentes, de la evaluación de estos modelos (con diferentes variantes) sobre dos corpus de evaluación, uno de ellos en la misma temática que el de entrenamiento y el otro en un área completamente diferente, surge que las perplejidades obtenidas son ligeramente menores cuando se generan los ngramas y se evalúa de derecha a izquierda que en el caso inverso. Este comportamiento no parece fácil de justificar "teóricamente" y caería dentro del tipo de regularidades que se detectan estadísticamente, que es el objetivo de los enfoques probabilistas. Parece interesante investigar un poco más en esta dirección, ya que de confirmarse esta regularidad, podría lograrse una mejora en el comportamiento de los modelos, sin un aumento importante en su tamaño.

Referencias

1. Eugene Charniak. *Statistical Language Learning (Language, Speech, and Communication)*. The MIT Press, September 1996.
2. R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue. Survey of the state of the art in human language technology, 1995.
3. G. Crispino. Apuntes del curso sobre modelos probabilistas del lenguaje natural, 2007.
4. Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
5. G.L. Moore. Adaptative statistical class-based models. 2001.
6. S.Young, G.Everman, and Marc Gales. *The HTK Book (for HTK Version 3.4)*. 2006.