Modelos de Lenguaje Estadísticos

Víctor Peinado

victor@lsi.uned.es

Grupo de PLN y Recup. de Información
Dept. LSI - UNED

3 de febrero de 2004

Índice

- Introducción: probabilidad, Regla de Bayes, Modelo de Lenguaje, *Noisy channel*.
- Modelos de Lenguaje: definición, construcción, evaluación (entropía y perplejidad).
- Mejoras en MLs: suavizado, *skipping*, *clustering*, otras técnicas.
- MLs en traducción automática: *IBM Models*, traducción como reescritura, parámetros.
- MLs en Recuperación de Información.
- Referencias.

Índice

- **Introducción: probabilidad, Regla de Bayes, Modelo de Lenguaje, *Noisy channel*.**
- Modelos de Lenguaje: definición, construcción, evaluación (entropía y perplejidad).
- Mejoras en MLs: suavizado, *skipping*, *clustering*, otras técnicas.
- MLs en traducción automática: *IBM Models*, traducción como reescritura, parámetros.
- MLs en Recuperación de Información.
- Referencias.

Probabilidad

- P(e) Probabilidad *a priori*: probabilidad de que aparezca *e* en nuestro corpus.
- P(e,f) Probabilidad conjunta: probabilidad de que se den tanto *e* como *f*.
- P(f|e) Probabilidad condicional: probabilidad de que, dado *e*, se dé *f*.

Regla de Bayes

Calcular P(e|f) es complicado. Existen infinitos *e* y *f*. Normalmente, lo calculamos en términos de *e* y usamos la regla de Bayes para transcribir:

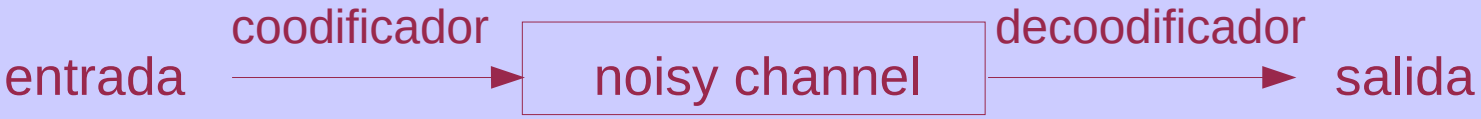
$$P(e|f) = \frac{P(e) \cdot P(f|e)}{P(f)}$$

Si estamos interesados en el evento más probable, podemos ignorar el denominador, entendiéndolo como una constante de normalización y reescribir:

$$\arg\max_e P(e|f) = \arg\max_e P(e) \cdot P(f|e)$$

Noisy Channel

- Un canal de comunicación es un sistema en el que la salida depende estadísticamente de la entrada: P(salida|entrada).



- Objetivo: optimizar la comunicación en un canal con ruido de manera que ocupen el mínimo espacio y conteniendo la redundancia suficiente que permita decodificar correctamente.

Índice

- Introducción: probabilidad, Regla de Bayes, Modelo de Lenguaje, *Noisy channel*.
- **Modelos de Lenguaje: definición, construcción, evaluación (entropía y perplejidad).**
- Mejoras en MLs: suavizado, *skipping*, *clustering*, otras técnicas.
- MLs en traducción automática: *IBM Models*, traducción como reescritura, parámetros.
- MLs en Recuperación de Información.
- Referencias.

Modelos de Lenguaje

- Los MLs se utilizan ampliamente en muchos ámbitos: reconocimiento de habla, OCR, reconocimiento de escritura, corrección ortográfica y traducción automática.
- Normalmente, se modelan lenguajes basados en ngramas (cadenas de Markov), ya que son fáciles de construir.
- Se formulan como una distribución de probabilidades *P(x)*, que reflejan la frecuencia de aparición de la cadena *x* en un corpus de entrenamiento.

Modelo de Lenguaje (II)

- Distribución de las probabilidades de secuencias de palabras. Si en un corpus de 10.000 oraciones 'I like snakes' aparece 3 veces:

$$P(\text{'I like snakes'}) = \frac{3}{10.000}$$

- ngrama: secuencia de palabras de longitud *n*.

$$P(w_i | w_{i-2}\ w_{i-1}) \approx \frac{C(w_{i-2}\ w_{i-1}\ w_i)}{C(w_{i-2}\ w_{i-1})}$$

- Para determinar la probabilidad de una secuencia de palabras (ngrama), descomponemos las probabilidades de los componentes:

$$P(w_1 \ldots w_n) = P(w_1) \cdot P(w_2|w_1) \cdot \ldots P(w_n|w_1 \ldots w_{n-1})$$

Modelos de Lenguaje (y III)

- Se construyen MLs basados en bigramas o trigramas a partir de un corpus de entrenamiento.
- ¿Cuál es la probabilidad de este modelo, dado el corpus de prueba?

$$P(\text{modelo}|\text{corpus\_de\_prueba}) = P(\text{modelo}) \cdot P(\text{corpus\_de\_prueba}|\text{modelo})$$

entropía

perplejidad

- La entropía/perplejidad de un ML con respecto a un corpus de prueba mide la probabilidad del modelo de generar dicho corpus.

Entropía

- Cantidad de información, en bits, de una variable aleatoria.
- Incertidumbre media de una variable aleatoria.
- Sorpresa o imprevisibilidad media asociada a un evento.
- Tamaño del espacio de búsqueda consistente en los valores posibles de una variable aleatoria y en sus probabilidades.
- Se corresponde con la longitud media de un mensaje necesaria para transmitir el valor de una variable.

Entropía (y II)

$$H(X) = -\sum P(X) \log_2 P(X)$$

$$H(\text{dado de 8 caras}) = -\sum \frac{1}{8} \log_2 \frac{1}{8} = -\log_2 \frac{1}{8} = \log_2 8 = 3\ bits$$

1	2	3	4	5	6	7	8
001	010	011	100	101	110	111	000

P(p)=1/8 P(t)=1/4 P(k)=1/8 P(a)=1/4 P(i)=1/8 P(u)=1/4

$$H(\text{Polinesio}) = -\left[4 \cdot \frac{1}{8} \log_2 \frac{1}{8} + 2 \cdot \frac{1}{4} \log_2 \frac{1}{4}\right] = 2\frac{1}{2}\ bits$$

p	t	k	a	i	u
100	00	101	01	110	111

Perplejidad

- Mide la incertidumbre de un evento. El número de opciones posibles en un determinado punto de decisión.
- Al evaluar la perplejidad de un ML, se calcula la probabilidad media que el modelo asigna a cada palabra del corpus de prueba.
- Media geométrica de la probabilidad inversa de las palabras del corpus de prueba que ocurren en el de entrenamiento.

$$pp = \sqrt[N]{\prod \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

$$pp = 2^H$$

Perplejidad

- La perplejidad de un modelo con respecto a una palabra *p1* representa el número de palabras que podrían seguir a *p1*.
- La perplejidad de un lenguaje depende, por definición, del propio corpus de entrenamiento.
  - Lengua coloquial: 247; Periodismo: 105; Radiología: 60
- Menor perplejidad indica que un lenguaje es más predecible.
- Cuanto menor es la perplejidad, entendemos que el ML es mejor.

Índice

- Introducción: probabilidad, Regla de Bayes, Modelo de Lenguaje, *Noisy channel*.
- Modelos de Lenguaje: definición, construcción, evaluación (entropía y perplejidad).
- **Mejoras en MLs: suavizado, *skipping*, *clustering*, otras técnicas.**
- MLs en traducción automática: *IBM Models*, traducción como reescritura, parámetros.
- MLs en Recuperación de Información.
- Referencias.

Suavizado

- Siempre existe la posibilidad de que haya ngramas perfectamente válidos no reconocidos cuya probabilidad sea igual a 0.
- Podemos utilizar suavizado (*smoothing*): redistribuimos la masa de probabilidad utilizando **coeficientes de descuento** para aumentar la estimación de los eventos no vistos.
- Si $P(z|xy)=0$ y $P(z|y) > 0$, podemos aventurar que *xyz* es una cadena válida.

$$P_{interp}(z|xy) = \lambda P_{3gram}(z|xy) + (1- \lambda) [\mu \cdot P_{2gram}(z|y) + (1-\mu) P_{1gram}(z)]$$

donde, $0 \leq \lambda, \mu \leq 1$

Skipping

- Cuanto más grandes sean los ngramas utilizados más difícil es ver el contexto completo en el corpus de entrenamiento.
- Podemos tener en cuenta contextos «similares» más amplios sin que sean contiguos.
- Un modelo con *backing-off* calcula:

$$\lambda\, P(z|wxy) + \mu\, P(z|xy) + (1-\lambda-\mu)\, P(z|y)$$

- Mientras que, un modelo con skipping calcula:

$$\lambda\, P(z|vwxy) + \mu\, P(z|vw\_y) + (1-\lambda-\mu)\, P(z|v\_xy)$$

Clustering

- Imaginemos que *P(*Tuesday|party on*)=0*, mientras que 'party on Wednesday' o 'party on Friday' sí aparecen en el corpus de entrenamiento.
- Podemos crear clases de palabras e incluir 'Tuesday' o 'Friday' por la clase WEEKDAY.

$$P(\text{Tuesday}|\text{party on}) =$$
$$= P(\text{WEEKDAY}|\text{party on}) \cdot P(\text{Tuesday}|\text{party on WEEKDAY})$$

Otras técnicas

- *Backing-off*: si *P(wxyz)=0*, podemos estimar *P(z|wxy)* a partir de *P(z|xy)*.
- *Cutoffs*: para reducir el tamaño de los LMs, podemos establecer umbrales para pasar por alto los ngramas poco frecuentes.
- Utilizar ngramas de orden superior.
- Utilizar una *caché* con los ngramas ya reconocidos: «Si un usuario usa una palabra, es probable que la vuelva a utilizar en un futuro cercano».
- Construir LM más específicos, uno por dominio o restringiendo el tipo de oración.

Índice

- Introducción: probabilidad, Regla de Bayes, Modelo de Lenguaje, *Noisy channel*.
- Modelos de Lenguaje: definición, construcción, evaluación (entropía y perplejidad).
- Mejoras en MLs: suavizado, *skipping*, *clustering*, otras técnicas.
- **MLs en traducción automática: *IBM Models*, traducción como reescritura, parámetros.**
- MLs en Recuperación de Información.
- Referencias.

Traducción A → B

- Transformar la oración A en predicados lógicos o en conjunción de aserciones que posteriormente son traducidos a la lengua B.
- Analizar sintáticamente la oración A y traducir el árbol de análisis con las modificaciones pertinentes a la lengua B.
- Tomar las palabras de la oración A, traducirlas y formar una *bolsa de palabras* no necesariamente correcta en la lengua B.

Traducción basada en reescritura

Mary did not slap the green witch.

- Asignamos a cada palabra su *fertilidad* y reescribimos tantas veces como valor tenga φ.

Mary not slap slap slap the the green witch.

- Reemplazamos las palabras en inglés por sus equivalentes en español.

Mary no daba una bofetada a la verde bruja.

- Reordenamos correctamente las palabras.

Mary no daba una bofetada a la bruja verde.

Parámetros

- Probabilidad de **aparición**: P('I like snakes').
- **Traducibilidad**: probabilidad de que *witch* se traduzca por *bruja*: t(bruja|witch)
- **Fertilidad**: probabilidad de que *witch* genere una sola palabra en español: n(1|witch).
- **Distorsión**: probabilidad de que una palabra en pos. 7 genere otra en pos. 8: d(8|7). Podemos tener en cuenta longitud de la oración: d(8|7,7,9)

Palabras «espurias»

- Palabras en la lengua de destino no generadas por ninguna palabra en la lengua de origen.
- Así pues, en sucesivas revisiones del Modelo de IBM, tenemos parámetros del tipo:
  - t(a|NULL): prob. de generar la palabra espuria *a*.
  - n(2|NULL): prob. de que existan 2 palabras espurias.
  - d(5|0,4,6): prob. de que una palabra NULL genere una palabra espuria en la pos. 5.

Índice

- Introducción: probabilidad, Regla de Bayes, Modelo de Lenguaje, *Noisy channel*.
- Modelos de Lenguaje: definición, construcción, evaluación (entropía y perplejidad).
- Mejoras en MLs: suavizado, *skipping*, *clustering*, otras técnicas.
- MLs en traducción automática: *IBM Models*, traducción como reescritura, parámetros.
- **MLs en Recuperación de Información.**
- Referencias.

MLs en Recup. de Información

- ¿Es el documento D relevante para la consulta Q?
- $P(Q|D) = P(D) \cdot P(D|Q)$
- Entrenamos un ML con los documentos.
- Calculamos la perplejidad del modelo con respecto al consulta.
- Aquellos docs relevantes serán mejor modelo que los no relevantes.

Índice

- Introducción: probabilidad, Regla de Bayes, Modelo de Lenguaje, *Noisy channel*.
- Modelos de Lenguaje: definición, construcción, evaluación (entropía y perplejidad).
- Mejoras en MLs: suavizado, *skipping*, *clustering*, otras técnicas.
- MLs en traducción automática: *IBM Models*, traducción como reescritura, parámetros.
- MLs en Recuperación de Información.
- **Referencias.**

Referencias

- (Goodman & Chen, 1998), “An empirical study of smoothing techniques for Language Modeling”.
- (Goodman, 2001),”A bit of progress in Language Modeling”.
- (Goodman, 2003), “The State of Art in Language Modeling”.
- (Knight, 1999), “A Statistical MT Tutorial Workbook”.
- (Manning & Schüzte, ****), *Foundations of Statistical NLP*.
- Herramientas
  - CMU-Cambridge Statistical Toolkit:
    `http://mi.eng.cam.ac.uk/~prc14/toolkit.html`
  - SRI Language Modeling Toolkit:
    `http://www.speech.sri.com/projects/srilm/`

Modelos de Lenguaje Estadísticos

Víctor Peinado

victor@lsi.uned.es

Grupo de PLN y Recup. de Información
Dept. LSI - UNED

3 de febrero de 2004