Reconocimiento Automático del Habla
2023-2024

Fundamentos en RAH:
Modelos de Lenguaje

MIARFID-RAH mcastro@dsic.upv.es

El "problema fundamental"

El denominado **problema fundamental del reconocimiento del habla** se formula así:

> Dada una observación acústica $O = o_1 o_2 \dots o_n$ se desea hallar una secuencia de palabras $\hat{W} = \hat{w}_1 \hat{w}_2 \dots \hat{w}_m$ pertenecientes a un vocabulario $V = \{v_1, v_2, \dots, v_N\}$ tal que

$$\hat{W} = \arg\max_W \frac{P(O \mid W) \cdot P(W)}{P(O)} = \arg\max_W P(O \mid W) \cdot P(W).$$

Ese problema se desglosa en cuatro:

- ¿Qué es O?: El problema de la parametrización.

  Lo abordamos en el tema dedicado a preproceso y parametrización de la voz.

- ¿Qué es $P(O \mid W)$?: El problema del modelado acústico.

  Lo hemos abordado al estudiar entrenamiento de HMMs y cálculo de probabilidades (y la puntuación de Viterbi) con HMM.

- ¿Qué es $P(W)$?: **El problema del modelado de lenguaje**.

El modelo de lenguaje

Por Bayes:

$$P(W) = P(w_1 w_2 \dots w_m) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2) \dots P(w_m \mid w_1 w_2 \dots w_{m-1})$$

"La historia ayuda ayuda a predecir el futuro."

¿Podemos "recordar" $P(w_i \mid \text{historia})$ para toda historia posible y buscar luego el W de mayor probabilidad?

Supongamos $N = 70000$, $n = 10$: ¡hay cerca de 10^{48} historias diferentes!

Una simplificación: historias equivalentes

La solución adoptada pasa por definir **clases de equivalencia para las historias**.

Una historia $w_1 w_2 \dots w_i$ pertenece a la clase de equivalencia $\Phi(w_1 w_2 \dots w_i)$ y lo que memorizamos es $P(w_{i+1} \mid \Phi(w_1 w_2 \dots w_i))$.

¿Qué tipos de clases de equivalencia se usan?

Marcas POS (Part of speech) Podemos representar cada palabra por la categoría sintáctica a la que pertenece.

Ejemplo: $\Phi(\texttt{el niño pequeño}) = \Phi(\texttt{el coche rojo}) = \texttt{ART NOM ADJ}$.

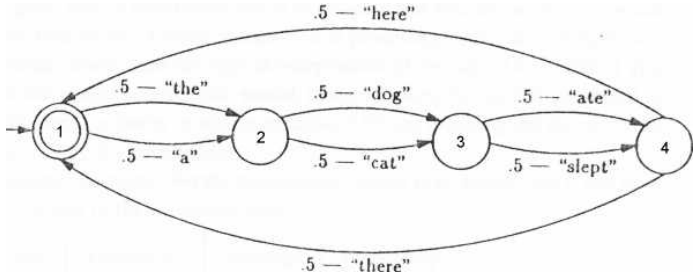
Marcas de similitud de contextos Algunas palabras pueden marcarse con información contextual: días de la semana, nombres de persona, etc.

Ejemplo: $\Phi(\texttt{quedamos el martes}) = \Phi(\texttt{quedamos el miércoles}) = \texttt{quedamos el DIASEM}$

Marcas obtenidas mediantes procedimientos de agrupación aglomerativa o divisiva Se puede encontrar familias de palabras por procedimientos automáticos. Las palabras suelen clasificarse utilizando información contextual.

Estados de un autómata Puede utilizarse un modelo de lenguaje regular representado con un autómata de estados finito estocástico. La función Φ asigna a una secuencia de palabras un estado (o conjunto de estados si el autómata es no determinista).

Ejemplo:



Autómata finito determinista estocástico.

$$\Phi(\texttt{the cat}) = \Phi(\texttt{a dog}) = 3$$

Modelos de n-gramas Son el estado del arte en los reconocedores de discurso continuo.

Un n-grama recuerda únicamente las $n-1$ últimas palabras:

$$\Phi(w_1 w_2 \dots w_i) = w_{i-(n-1)} w_{i-(n-2)} \dots w_i$$

- **unigramas**: $P(w_{i+1} \mid w_1 w_2 \dots w_i) = P(w_{i+1})$.
- **bigramas**: $P(w_{i+1} \mid w_1 w_2 \dots w_i) = P(w_{i+1} \mid w_i)$.

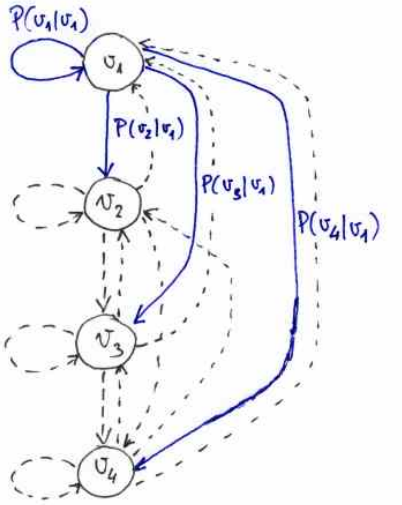
- **trigramas**: $P(w_{i+1} \mid w_1 w_2 \ldots w_i) = P(w_{i+1} \mid w_{i-1} w_i)$.

Ejemplo: En un bigrama tenemos

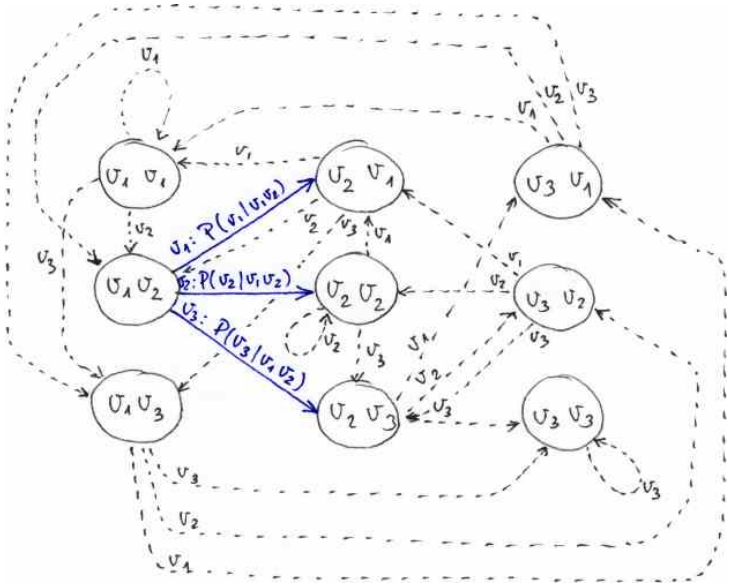
$$P(\texttt{un ejemplo de frase}) = P(\texttt{un} \mid \texttt{INICIO}) \cdot P(\texttt{ejemplo} \mid \texttt{un}) \cdot P(\texttt{de} \mid \texttt{ejemplo}) \cdot P(\texttt{frase} \mid \texttt{de})$$

(Nota: ha convenido introducir una marca especial de `INICIO`.)

Un n-grama es un caso particular de autómata estocástico con una estructura fija.



Bigrama.



Trigrama.

El número de posibles trigramas sigue siendo un número enorme. Por ejemplo, si tenemos 70000 palabras en el vocabulario, hay más de 10^{14} posibles trigramas.

Pero en una lengua no todos son válidos.

n-gramas

No podemos modelar la sintaxis de un lenguaje natural con un AF (los lenguajes naturales no son regulares —ni siquiera son incontextuales—).

Tampoco deseamos **generar** frases del lenguaje, sino **evaluar** hasta qué punto una frase dada sigue las reglas del lenguaje.

Los modelos de n-gramas permiten hacer participar información "futura" en decisiones "actuales". Si hemos de elegir entre

$$\texttt{el} \left\{ \begin{array}{c} \texttt{niño} \\ \\ \texttt{coche} \end{array} \right\} \texttt{estudioso}$$

$$P(\texttt{el niño estudioso}) = P(\texttt{el}) \cdot P(\texttt{niño} \mid \texttt{el}) \cdot P(\texttt{estudioso} \mid \texttt{el niño})$$

$$P(\texttt{el coche estudioso}) = P(\texttt{el}) \cdot P(\texttt{coche} \mid \texttt{el}) \cdot P(\texttt{estudioso} \mid \texttt{el coche})$$

Si $P(\texttt{niño} \mid \texttt{el})$ es igual a $P(\texttt{coche} \mid \texttt{el})$, entonces la elección depende de la siguiente palabra. El valor de n determina el intervalo de "tiempo" que participa en la decisión.

Aunque el propósito no es usar el modelo para generar frases, sino medir su probabilidad de pertemecer a un lenguaje, aquí tienes un ejemplo de texto generado al azar con un modelo de trigramas aprendido con el primer libro de El Quijote:

así debe de guardar una manada de puercos que sin que él pondría remedio cor
estos señores guardianes y comisarios sean servidos de informalle y decille
fuerte brazo y mi persona . destas lágrimas y suspiros así pasándoseme aquel
cautela se podía tener por verdaderas tantas falsedades pero no me dejo esti
estime y así con muestras de verdadero amor no se pudo escudar tan bien y aú
excepto aquellas que le habían hablado en su imaginación un buen espacio de
de escarlata con que fácilmente se sanase y tomando de la mancha i capítulo
hidalgo don quijote . sancho asimismo callaba y comía bellotas y visitaba mu
segundo que trata de la misma envidia ni debe de haber sido su autor arábigo
propio de los veros azules ni endiablados . don quijote sin que nos ha depar
que sin que yo desease jamás había de decir de esta tan tenebrosa aventura a
que anochecía . lloráralas yo dijo el cabrero mas nunca lo bueno que aquel s

me acuerdes y representes lo que vas diciendo no acabarás de venir a verlas
cosas que en efecto que este rey no murió sino que por grandes maestros que
puño . causó risa al licenciado ni al barbero la aventura y las espaldas los
venta donde estuviesen porque se advierta cuan sin culpa le habéis dado con
del mundo estaban repartidas . si no era muy buen parecer señor maese nicolá
playas desnudas de contrato humano o adonde el caballero platir dijo el gale
acertado el que tenía la pesadilla y comenzó a decir sancho que no daban pur
calamidades de sus enemigos .

Los n-gramas funcionan sorprendentemente bien modelando cada palabra en función, únicamente, de un bajo número de palabras anteriores.

Ventajas de los n-gramas:

- Implementación sencilla.
- Integración fácil con modelos acústicos (no son más que autómatas finitos).
- Hay algoritmos eficientes para reconocimiento (no son más que autómatas finitos).
- La estructura es fija, sólo hay que estimar probabilidades.

Problemas de los n-gramas:

- Modelan deficientemente dependencias a largo término.
- No garantizan la concordancia en número, persona, género, etc.
- En lenguajes fuertemente flexionados explota el número de unidades elementales y ello impacta negativamente en el número de n-gramas (¿Es diferente "si yo estuviera" de "si yo estuviese"?).
- Necesitan una gran cantidad de datos para efectuar buenas estimaciones de probabilidad.
- Asignan probabilidad nula a los n-gramas que no se han visto previamente.
- Aceptan frases sin sentido al no capturar el significado del texto.

Estimación de probabilidades

Sea h una historia (las $n-1$ últimas palabras). Deseamos estimar $P(w_i \mid h)$ y disponemos de un corpus con ejemplos de frases del lenguaje.

- Sea $C(h)$ el número de veces que apareció h en el corpus.
- Sea $C(hw_i)$ el número de veces que apareció h seguido de w_i en el corpus.

$$P(w_i \mid h) = \frac{C(hw_i)}{C(h)} = \frac{C(hw_i)}{\sum_{v \in \mathcal{V}} C(hv)} = f(w_i \mid h)$$

donde f representa la frecuencia.

Unigrama
$$P(w_i) = \frac{C(w_i)}{\mid \mathcal{V} \mid} = f(w_i)$$
Bigrama
$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} = f(w_i \mid w_{i-1})$$
Trigrama
$$P(w_i \mid w_{i-1}w_{i-2}) = \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})} = f(w_i \mid w_{i-1}w_{i-2})$$

El problema de la falta de datos

Un experimento. Aprendizaje con 1500000 palabras (*running words*) de un vocabulario de 1000 palabras. Aplicación del modelo a un texto con 300000 palabras: el 23 % de los trigramas en las frases de test no fueron vistos en el corpus de aprendizaje, así que tienen probabilidad cero.

¡Si en una frase aparece un trigrama con probabilidad cero, la frase entera tiene probabilidad cero! ¡Pero es una frase válida en el lenguaje y no se le puede asignar una probabilidad nula!

¿Cómo estimamos la probabilidad de un trigrama que no apareció en el corpus de aprendizaje?

Es necesario suavizar/interpolar las frecuencias.

Otros modelos de lenguaje

- Modelos de lenguaje con cache.
- Modelos de lenguaje con disparador (trigger). Modelan relaciones a larga distancia.
- Modelos de lenguaje multinivel.
- Modelos de lenguaje basados en morfemas.
- Modelos de lenguaje basados en árboles de decisión.
- ...

Bibliografía

- Frederick Jelinek: `Statistical Methods for Speech Recognition`. The MIT Press. 1998.
- Daniel Jurafsky and James H. Martin: *SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-hall, 2000.
- Chris Manning and Hinrich Schütze: *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
- Lawrence Rabiner, Biing-Hwang Juang: *Fundamentals of speech recognition*. Prentice Hall, 1993.