FUNDAMENTOS DEL RECONOCIMIENTO AUTOMÁTICO DE LA VOZ

"Modelos de lengua"

Agustín Álvarez Marquina

Introducción. Modelos de lengua (I)

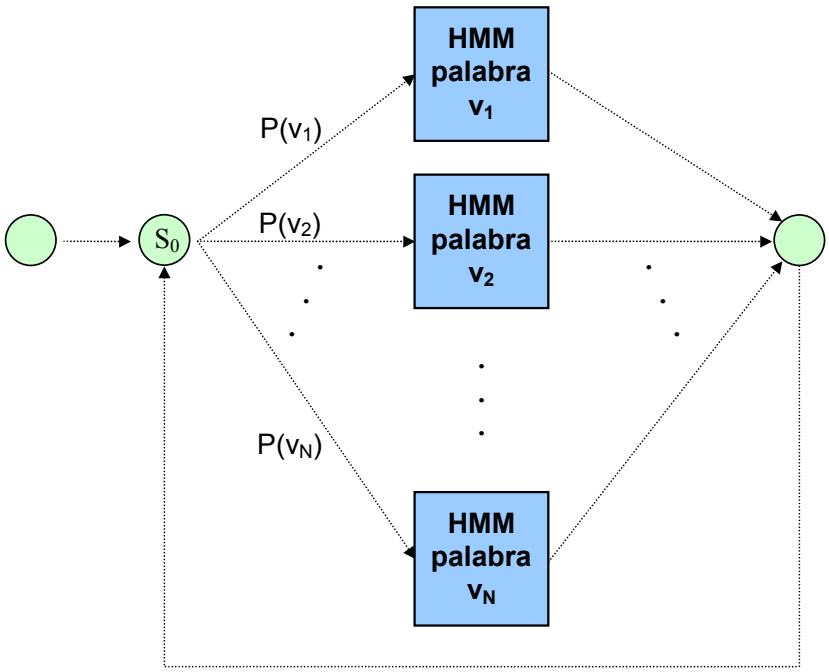

Figura 1. Modelo compuesto de producción de discurso cuando la generación de palabras no sigue ninguna gramática.

Introducción. Modelos de lengua (II)

- **La tarea que debe realizar un sistema automático de reconocimiento de voz es encontrar la cadena de palabras que satisfaga:**

$$\hat{W} = \arg\max_{W} P(A \mid W) P(W)$$

donde *A* son los datos acústicos y $W = w_1, w_1, \ldots, w_n$, con $w_i \in V$, denota la cadena de *n* palabras de entre un vocabulario de tamaño fijo *V*.

Definición de modelos de lengua (I)

- **Un modelo de lengua es el mecanismo que permite asignar a cada posible secuencia de palabras *W*, la probabilidad asociada *P(W)*.**
- **En este caso una palabra queda definida por su pronunciación.**
  - Si una palabra tiene varias pronunciaciones se considerará que son dos entidades diferentes.
  - Los vocablos homófonos se considerarán una única palabra.

Definición de modelos de lengua (II)

- **Si empleamos la regla de Bayes, el valor de *P(W)* puede descomponerse de manera formal en:**

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_1,\ldots, w_{i-1})$$

donde $P(w_i|w_1, ..., w_{i-1})$ es la probabilidad de que w_i sea realizada teniendo en cuenta que las palabras $w_1, ..., w_{i-1}$ fueron pronunciadas previamente.

  - Estas *n* últimas palabras se conocen como historia y suelen denotarse por h_i.

Modelos de lengua. Aproximación práctica (I)

- **En la práctica sin embargo, no es posible estimar los valores $P(w_i|w_1, ..., w_{i-1})$.**
  - Suponiendo que tengamos un vocabulario de tamaño $|V|$, podremos tener $|V|^{i-1}$ historias diferentes, lo cual para vocabularios de varios miles de palabras es inabordable.

- **La solución a este problema pasa por agrupar todas las posibles historias en un número manejable de clases de equivalencia [JEL92].**
  - La idea que sigue este planteamiento es que incluso para valores pequeños de *i*, muchas combinaciones de esas *i* palabras no se producen en la realidad.

Modelos de lengua. Aproximación práctica (II)

- Definimos la función *F* como la aplicación de un conjunto de historias en alguna de las clases de equivalencia anteriormente citadas.

- Si $F(w_1, w_2, w_{i-1})$ denota la clase de equivalencia de la cadena w_1, w_2, w_{i-1} entonces la probabilidad $P(W)$ puede aproximarse por medio de la siguiente expresión:

$$P(W) = \prod_{i=1}^{n} P(w_i \mid \Phi(w_1,\ldots, w_{i-1}))$$

Modelos de lengua. Aproximación práctica (III)

- Viendo el conjunto de clases como un autómata de estados finitos, podemos considerar que la introducción de una nueva palabra hace evolucionar a la gramática del estado Φ_{i-1}, (instante *i-1)* al estado Φ_i.

- De esta forma la ecuación anterior puede expresarse como:

$$P(W) = \prod_{i=1}^{n} P(w_i \mid \Phi_{i-1})$$

Modelos de lengua. Aproximación práctica (IV)

- **Con objeto de poder evaluar los términos $P(w_i | \Phi_{i-1})$, definimos $C(w, \Phi)$ como el número de veces que la palabra *w* alimenta al autómata de estados inmediatamente después de que éste se encuentre en el estado Φ.**
  - De igual manera $C(\Phi)$ denota el número de veces que el autómata alcanza el estado Φ:

$$C(\Phi) = \sum_{w} C(w, \Phi)$$

Modelos de lengua. Aproximación práctica (V)

- **De esta forma la estimación de la probabilidad buscada queda:**

$$P(w_i \mid \Phi_i = \Phi) = \frac{C(w, \Phi)}{C(\Phi)}$$

N-gramas (I)

- En la práctica los valores $C(w, \Phi)$ se obtienen a partir de grandes bases de datos con documentos de texto.
- Usualmente el criterio para determinar las clases de equivalencia suele ser lo que se conoce como N-gramas (*N-grams*).
- Dos historias son equivalentes si las *N-1* palabras coinciden.

N-gramas (II)

- Los casos más comunes son N=2 o bigramas (*bigrams*) [NEY92] y N=3 o trigramas (*trigrams*) [JEL85].
- Ejemplo. Fórmula para el caso de los trigramas:

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1})$$

$$P(w_3 \mid w_1, w_2) = f(w_3 \mid w_1, w_2) \approx \frac{C(w_1, w_1, w_3)}{C(w_1, w_2)}$$

Bigramas

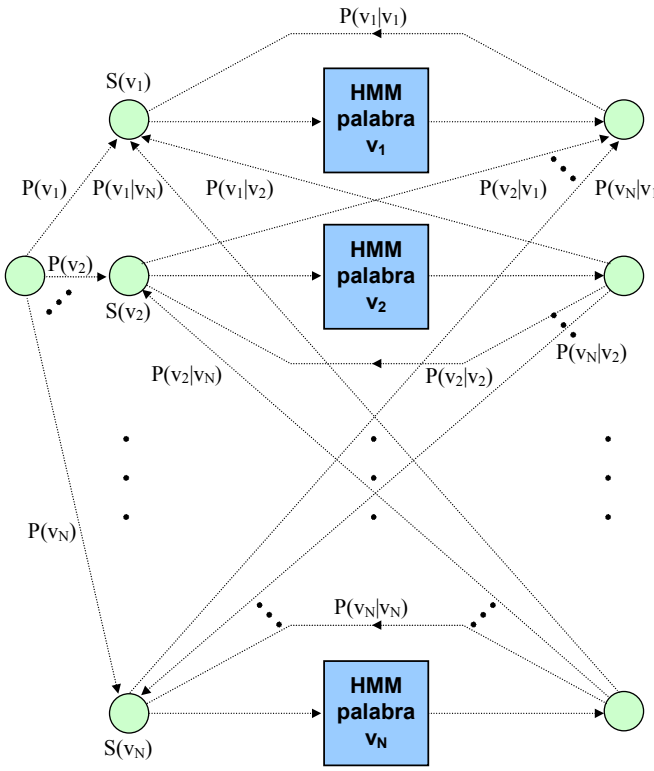

Figura 2. Modelo compuesto de producción de discurso cuando la generación de palabras depende solamente de la identidad de la palabra precedente (modelo de lengua de bigramas).

Trigramas (I)

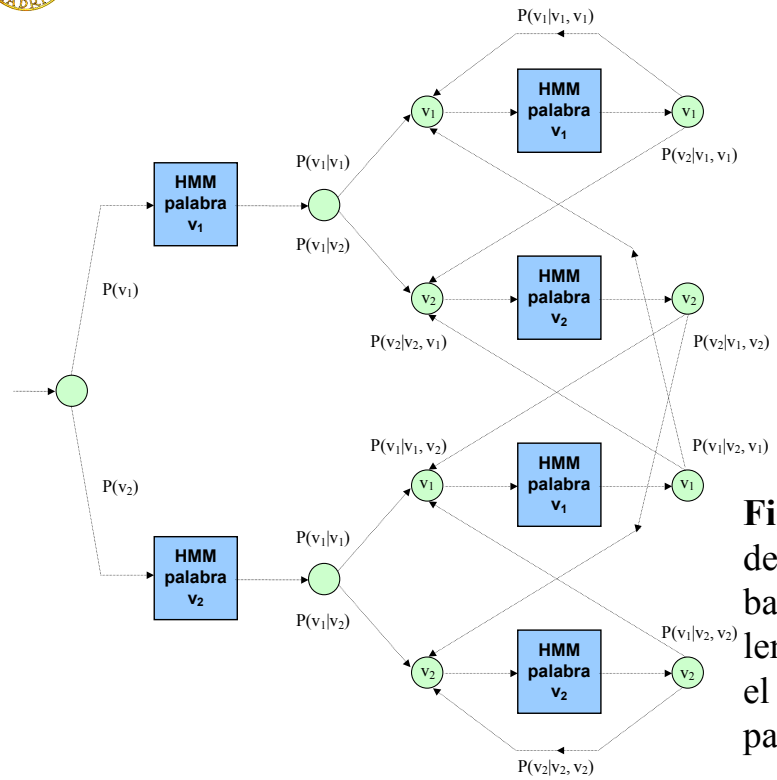

Figura 3. Modelo compuesto de producción de discurso basado en un modelo de lengua con trigramas cuando el vocabulario consta de dos palabras.

Trigramas (II)

- **Uno de los problemas con los trigramas es que para sistemas con vocabularios muy grandes (200.000 palabras), la red resultante puede no resultar factible.**
- **Una solución a este problema aparece en [DER86].**
  - En este caso se emplea como modelo de lengua uno que denominan tri-POS (*Parts Of Speech*) y un conjunto de restricciones sintácticas globales empleando un analizador o *parser* de oraciones.

Trigramas (III)

- La gran ventaja de este método es la importante disminución en el espacio de memoria requerido para almacenar la red fonética, así como, la necesidad de una menor cantidad de datos de entrenamiento.
- Sin embargo, el cálculo de la probabilidad condicionada a partes de la red de las 2 palabras anteriores proporciona unas restricciones lingüísticas mucho más débiles que las del modelo de trigramas.

Problemas de estimación

- La falta de suficientes datos en forma de documentos de texto producirá que muchas de las combinaciones de N-palabras no aparecerán en el conjunto de textos disponibles.
- Consecuencia: probabilidad valdrá *0*.
- Se hace necesario algún método que permita suavizar la estimación de probabilidad para eventos no presentes [ESS92].

Método de interpolación lineal

- Una forma de hacerlo es mediante la interpolación lineal de las frecuencias de trigramas, bigramas y unigramas (*deleted interpolation*) [JEL92]:

$$P(w_3 \mid w_1, w_2) = \lambda_3 f(w_3 \mid w_1, w_2) + \lambda_2 f(w_3 \mid w_2) + \lambda_1 f(w_3)$$

  - Los pesos de la interpolación lineal λ_i se estiman haciendo máxima la probabilidad para un conjunto de datos (textos) diferente de los empleados para calcular la frecuencia de aparición de los n-gramas.
    - El algoritmo de Baum-Welch puede emplearse perfectamente para hallar la solución de este problema de máxima semejanza.

Método de retroceso (I)

- **Otra técnica que sirve para suplir la falta de datos de entrenamiento es el método de retroceso *(backing-off)* [KAT87].**

$$\hat{P}(w_3 \mid w_1, w_2) = \begin{cases} f(w_3 \mid w_1, w_2) w_2 & si\ C(w_2, w_3) \geq K \\ \alpha Q_T(w_3 \mid w_1, w_2) & si\ 1 \leq C(w_2, w_3) \geq K \\ \beta(w_1, w_w) \hat{P}(w_3, w_2) & resto\ de\ casos \end{cases}$$

  - Los valores α y β se escogen, de forma que la probabilidad quede normalizada de manera adecuada.
  - $Q_t(w_3 \mid w_1, w_2)$ es una función del tipo Good-Turing [GOO53].

Método de retroceso (II)

- La estimación de probabilidad de $P(w_3 \mid w_2)$ es la estimación de probabilidad del bigrama que se realiza de la misma forma que la estimación de $P(w_3 \mid w_1, w_2)$:

$$\hat{P}(w_3 \mid w_2) = \begin{cases} f(w_3 \mid w_2) & si\ C(w_2, w_3) \geq L \\ \alpha Q_T(w_3 \mid w_2) & si\ 1 \leq C(w_2, w_3) \geq L \\ \beta(w_2) f(w_3) & resto\ de\ casos \end{cases}$$

  - siendo los valores de *L* y *M* umbrales determinados de forma intuitiva.

Método de retroceso (III)

- **Este método es el que prevalece en muchos reconocedores actuales.**
- **La idea detrás de este método, es que si hay suficiente número de casos, la frecuencia relativa es una buena estimación de la probabilidad.**
  - Si no, se debe retroceder y calcular estas probabilidades tomando las frecuencias de los bigramas y llegado el caso a partir de los unigramas.

N-gramas con N>3 (I)

- **En general no resulta práctico emplear n-gramas para valores de n mayores de 3.**
  - La ligera ganancia en la complejidad de decisión de la gramática, se ve claramente contrarrestada por la mayor dificultad que presenta la estructura de decisión.
  - Algunos trabajos además intentan añadir criterios sintácticos, como la categoría que ocupan las palabras en la oración (sustantivos, verbos, adjetivos, etc.).
    - Objetivo: disminuir la complejidad de decisión y con el propósito de mejorar los procedimientos de suavizado [MAL92].

N-gramas con N>3 (II)

- Incluso puede permitirse la existencia de palabras que aparezcan en diversas categorías, como consecuencia de las distintas funciones gramaticales o significados radicalmente diferentes que pueden presentar éstas [JAR96].

- Algunas mejoras en el modelado tienen en cuenta la historia de ocurrencias en las palabras, modelando la mayor o menor concentración en determinados textos o partes de un texto de manera dinámica [LAU93].

N-gramas con N>3 (III)

- En otros casos, en vez de emplear la identidad de las palabras más recientes para definir las clases de equivalencia de una historia, se usa el estado de un parser gramatical con objeto de definir el evento condicionante [GOD92].

- Por otra parte, si además queremos disponer de conocimiento semántico, el proceso requerirá etapas adicionales para aprender y extraer de forma automática dichas estructuras semánticas [KUH94].

Bibliografía (I)

[DER86] A. M. Derouault and B. Merialdo, "Natural Language Modeling for Phoneme-to-Text Transcription", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, Nº. 6, noviembre 1986, pp. 742-749.

[ESS92] U. Essen and V. Steinbiss, "Cooccurrence Smoothing for Stochastic Language Modeling", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 161-164.

[GOD92] D. Goddeau and V. Zue, "Integrating Probabilistic LR Parsing Into Speech Understanding Systems", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 181-184.

[GOO53] I. J. Good, "The Population Frequencies of Species and the Estimation of Population Parameters", *Biometrika*, Vol. 40, Partes 3 y 4, diciembre 1953, pp. 237-264.

[JAR96] M. Jardino, "Multilingual Stochastic N-Gram Class Language Models", *Proc. of ICASSP'96*, Atlanta, Estados Unidos, 7-10 mayo 1996, pp. 161-163.

[JEL85] F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer", *Proceedings of the IEEE*, Vol. 73, Nº. 11, noviembre 1985, pp. 1616-1623.

Bibliografía (II)

[JEL92] F. Jelinek, R. L. Mercer and S. Roukos, "Principles of Lexical Language Modeling for Speech Recognition", *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi editores, Marcel Dekker Inc., 1992, pp. 651-699.

[KAT87] S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-35, Nº. 3, marzo 1987, pp. 400-401.

[KUH94] R. Kuhn, R. De Mori and E. Millien, "Learning Consistent Semantics from Training Data" *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. II, pp. 237-240.

[LAU93] R. Lau, R. Rosenfeld and S. Roukos, "Trigger-Based Language Models: A Maximum Entropy Approach", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 45-48.

[MAL92] G. Maltese and F. Mancini, "An Automatic Technique to Include Grammatical and Morphological Informatical and Morphological Information in a Trigram-Based Statistical Language Model", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 157-160.

[NEY92] H. Ney, D. Mergel and A. Noll, "Data Driven Search Organization for Continuous Speech Recognition", *IEEE Transactions on Signal Processing*, Vol. 40, Nº. 2, febrero 1992, pp. 272-281.