MODELIZACIÓN DEL LENGUAJE MEDIANTE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Germán Bordel*

Departamento de Electricidad y Electrónica

Universidad del País Vasco / Euskal Herriko Univertsitatea

INTRODUCCIÓN

La aparición y la amplia difusión de los equipos de procesamiento de la información que se ha dado en las últimas décadas, ha supuesto la disponibilidad de una magnífica herramienta para el estudio del lenguaje así como el surgimiento de un nuevo área de aplicación de los conocimientos adquiridos al respecto.

En relación con las aplicaciones que precisan del conocimiento relativo al lenguaje, cabe distinguir dos tendencias a la hora de ser abordadas. Por un lado, la que suele denominarse como "basada en el conocimiento" y que, en lineas generales, consiste en plasmar en el sistema los conocimientos ya adquiridos para que éste se encargue de gestionarlos adecuadamente y así pueda hacer frente a los requerimientos que se le presenten. Por otro lado se encuentra el enfoque del "aprendizaje automático" que pretende que el conocimiento necesario sea adquirido por el mismo sistema.

Se trata de dos tendencias por cuanto que las investigaciones que se realizan en procesamiento del lenguaje adoptan una u otra vía de forma prácticamente excluyente, pero no quiere decir esto que sean realmente excluyentes entre sí. Una posible vía de evolución para estos sistemas podría llevarnos a la utilización de técnicas de aprendizaje automático que llegaran a representar la información obtenida de forma similar a como se hace en el enfoque "basado en el conocimiento", y de esta forma a aplicar también las técnicas que le son propias. También puede pensarse para un tal sistema en que no todo el conocimiento fuese extraído de manera automática sino que una parte se aportara por expertos humanos.

En lo que sigue se pretende dar una visión general del enfoque basado en el aprendizaje automático. Se comienza por ver una definición de lenguaje, de modelo de lenguaje y la necesidad de establecer modelos estocásticos (es decir con asignación de probabilidades) para, a continuación, pasar a ver los tipos de modelos más estudiados. Dada la amplitud del tema, el enfoque consiste en presentar en lineas generales estas técnicas, en ocasiones haciendo referencia a casos concretos, y en todo momento proporcionando la bibliografía más notable de forma que el lector interesado en iniciarse más a fondo en alguno de los puntos encuentre una vía de acceso adecuada.

MODELO DE LENGUAJE ESTOCÁSTICO.

En sentido amplio, un **lenguaje** no es más que un conjunto cuyos elementos tienen una determinada estructura. Esta estructura consiste en la concatenación lineal de unas determinadas unidades básicas[1]. Podemos convenir en denominar **frases** a los elementos del lenguaje y **palabras** a las unidades que las forman. El conjunto de palabras será el **vocabulario** del lenguaje en cuestión. Esta nomenclatura corresponde a la utilizada en el lenguaje natural, hablado o escrito, pero la definición permite englobar a otras muchas entidades.

---

* German@we.lc.ehu.es

Expreso mi agradecimiento a Natividad Prieto del Departamento de Sistemas Informáticos y Computación de la Universidad Politécnica de Valencia por su colaboración en la preparación de este trabajo. Muchos de los datos que aquí se aportan han sido proporcionados por ella y figuran en su Tesis Doctoral [Prieto 94].

[1]Más adelante se verá un caso concreto en que se considera un lenguaje cuyos elementos presentan una estructura más compleja, pero este planteamiento es suficiente habitualmente.

Las técnicas de procesamiento de lenguajes han demostrado ser de gran utilidad en áreas más amplias que la que nos ocupa. Por ejemplo, dentro del reconocimiento de patrones existen multitud de problemas que pueden tratarse bajo esta óptica. Un claro ejemplo lo tenemos en el trabajo desarrollado por Ledley que K. S. Fu explica en [Fu 74]. Se trata de la clasificación de cromosomas en función de su geometría, para lo cual se establece una descripción de sus perímetros mediante códigos que expresan la curvatura de pequeños segmentos contiguos de los mismos. De esta forma los distintos perímetros son las frases del lenguaje y cada tipo de segmento una de sus palabras.

Por tanto, con independencia de si tratamos con lenguajes naturales o de otro tipo, cabe plantearse qué es un modelo de lenguaje. Los lenguajes que presentan algún interés sólo contienen determinadas concatenaciones de palabras de un vocabulario dado. Existen ciertas reglas que determinan la pertenencia o no de una de tales frases al conjunto. Un **modelo de lenguaje** es un sistema que incluye el conocimiento de esas reglas (de forma explícita o implícitamente en su estructura) por lo que permite determinar si una frase dada pertenece al lenguaje de que se trate.

En el caso de los lenguajes naturales el conjunto de reglas es extremadamente complejo, a lo que debemos añadir que habitualmente el interés no se centra en el lenguaje que se ciñe estrictamente a las reglas establecidas, sino en el lenguaje real en el que es frecuente la transgresión de algunas de sus normas. Por tanto se suele plantear el problema de la pertenencia de una frase al lenguaje como un problema de grado, es decir, hasta que punto se puede considerar a la frase perteneciente al lenguaje aun cuando no cumpla estrictamente con alguna de sus normas. Un buen modelo determinara como claramente pertenecientes al lenguaje aquellas frases que cumplan todas sus normas y asignara una probabilidad también considerable a aquellas que se aparten poco de las mismas y puedan ser por tanto aceptadas. Hablamos por tanto de modelos que atribuyen probabilidades a todas las frases posibles (todas las concatenaciones posibles de palabras pertenecientes al vocabulario), es decir **modelos estocásticos**. → (también podríamos hablar de [illegible])

Los métodos de obtención de modelos estocásticos mediante aprendizaje automático son variados. Podemos enumerar como principales (en principio como métodos diferentes, si bien algunos están relacionados) los N-gramas, la inferencia gramatical (IG), los Modelos de Markov (MM), las redes neuronales (RN) y los arboles de decisión. Lo que es común a todos ellos es la necesidad de disponer de un subconjunto del lenguaje como muestra a partir de la cual se aprenderán las características del mismo. Una característica importante de cada método es su capacidad de inferencia de las reglas del lenguaje en función del tamaño de la muestra que se le proporciona.

En caso de proporcionar utilizar como muestra de aprendizaje un conjunto de frases del lenguaje se habla de aprendizaje con muestras positivas (R+). En ocasiones es posible proporcionar también muestras no pertenecientes al lenguaje (muestras negativas R-) y en tal caso se habla de aprendizaje mediante muestras positivas y negativas o bien de una "muestra completa". En el capítulo referente a inferencia gramatical mencionaremos algunas características del aprendizaje con ambos tipos de muestras.

De entre los métodos mencionados, hablaremos de N-gramas , de inferencia gramatical y de árboles de decisión ya que son los que, por diversos motivos, presentan mayor interés a la hora de hacer una revisión sobre modelos estocásticos. La utilización de Modelos de Markov se refiere principalmente a el uso de las técnicas de entrenamiento que les son propias (como los algoritmos de Viterbi y Baum-Welch [Forney 73] [Baum 72]), ya que la topología de los modelos suele ser semejante a la de un Autómata de Estados Finito (AEFs) [Jelinek 85]. Lo mismo sucede con las Redes Neuronales ya que es común el uso de Redes Neuronales Recurrentes (RNR) que son en cierto modo equivalentes a los MM, variando las técnicas de estimación de los parámetros. Los modelos de N-gramas serán una constante referencia a lo largo de la exposición, viéndose cómo encajan dentro de las demás metodologías.

Evaluación de modelos de lenguaje.

La formación automática de un modelo de un lenguaje supone la "generalización" del conjunto de muestras, es decir, la extracción de las regularidades que se detecten, de forma que el modelo pueda entender como pertenecientes al lenguaje frases que no le han sido aportadas como ejemplo. Se tiene por tanto que la "calidad" del modelo (del método de obtención de modelos) depende en gran manera del tipo de generalización que realice. Una buena generalización conseguirá incluir prácticamente todas las frases del lenguaje (buena cobertura) y no considerara validas otras que no pertenecen a él (no sobregeneralizará) (véase figura 1).

La evaluación última del modelo debe venir dada en términos de su efectividad en la ejecución de la tarea para la que ha sido construido. Este enfoque tiene como dificultad la necesidad de disponer del resto del sistema y de la integración del modelo en él. También supone una dependencia de los resultados obtenidos en función del resto del sistema, con lo que la evaluación presenta una dificultad añadida. Es por ello que se ha buscado la manera de evaluar los modelos de lenguaje de una forma aislada.

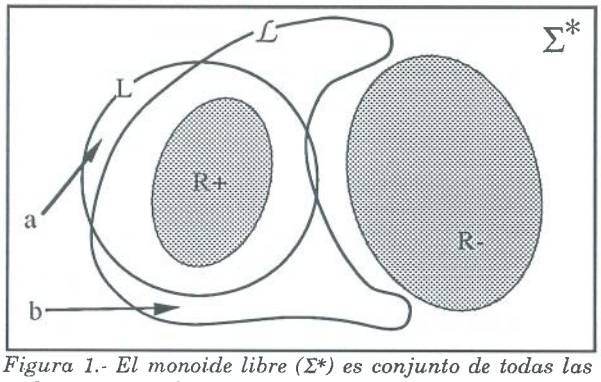


Figura 1.- El monoide libre (Σ) es conjunto de todas las cadenas que pueden formarse con elementos de un alfabeto Σ. En él se representa un lenguaje dado L a modelizar. Para ello se dispone de un conjunto de muestras positivas R+ pertenecientes al lenguaje, y posiblemente un conjunto de muestras negativas R- no pertenecientes al mismo. El modelo obtenido representa a un lenguaje ℒ que coincidirá mejor o peor con el L en función de múltiples factores. Las diferencias son de dos tipos: a) falta de cobertura b) sobregeneralización.*

Para determinar este método de evaluación se recurre a la Teoría de la Información [Shannon 64]. Según esta teoría, puede aplicarse la expresión de la entropía (proveniente de la mecánica estadística) para medir la información que proporciona un lenguaje o un sistema (una fuente) generador de un lenguaje P:

$$H(P) \triangleq -\sum_{i=1}^{L} p_i \log_2(p_i) \tag{1}$$

donde L es el número elementos del lenguaje y p_i la probabilidad de cada uno de ellos.

Dado que el objetivo es la comparación entre la distribución del lenguaje (P) y la obtenida por el modelo (Q), puede utilizarse la entropía relativa o distancia de Kullback-Lieber[2] :

$$\chi(P,Q) \triangleq \sum_{i=1}^{L} p_i \log_2\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{L} p_i \log_2(p_i) - \sum_{i=1}^{L} p_i \log_2(q_i) \tag{2}$$

donde p_i y q_i son las probabilidades asignadas a cada elemento por P y Q respectivamente.

El método de evaluación se vale de esta medida de la siguiente manera: por un lado tendremos el modelo de lenguaje (Q), y por otro lado tomaremos un conjunto de muestras (T) del lenguaje P diferente del utilizado para la formación del modelo. La entropía relativa nos dará una medida de la disimilitud entre ambos lenguajes (T y el determinado por Q), de esta forma para otro modelo Q´ se puede obtener otra medida de disimilitud frente a T y determinar cual de los dos modelos es el que se le representa mejor.

En realidad la expresión (2) no se ajusta bien a las posibilidades de evaluación relativa de dos modelos de lenguaje, por lo que se aplica según las siguientes consideraciones:

i) El primer término se anula al evaluar un modelo frente a otro;

$$\chi(T,Q) - \chi(T,Q') = \sum_{i=1}^{L} t_i \log_2(q_i') - \sum_{i=1}^{L} t_i \log_2(q_i)$$

por lo que se considera sólo el segundo como representativo de la calidad del modelo:

$$H^*(Q) = -\sum_{i=1}^{L} t_i \log_2(q_i) \tag{3}$$

Dados varios modelos, el cálculo de esta cantidad frente a T para cada uno de ellos, nos permite asegurar que aquel que presenta el valor más bajo es el mejor ajustado a la muestra. La expresión (3) es lo que se conoce como entropía cruzada ya que supone una estimación de la entropía de una distribución respecto a otra.

ii) El sumatorio se extiende a todos los elementos de P, pero como de lo que se dispone es de una muestra T del lenguaje, suponiendo que esta es suficientemente representativa de las frecuencias

[2] Que en realidad no es una distancia. Basta para ello el hecho de no ser una función simétrica respecto a sus argumentos.

de aparición de sus elementos, puede extenderse a todas las muestras normalizando posteriormente[3]:

$$H^*(Q) = \frac{-1}{M} \sum_{i=1}^{M} \log_2(q_i)$$

donde M es el tamaño de la muestra.

iii) La "perplejidad" es un valor que se relaciona con la entropía según:

$$\mathrm{PP(P)} \triangleq 2^{\mathrm{H(P)}}$$

de forma que, dada la semejanza de (3) con una entropía, suele utilizarse la expresión [Jelinek 85/86/90]:

$$\mathrm{PP}(Q) = 2^{H^*(Q)} = 2^{\frac{-1}{M}\sum_{i=1}^{M}\log_2(q_i)} \tag{4}$$

para dar la medida de la calidad del modelo.

Este valor resulta ser el inverso de la media geométrica de las probabilidades que el modelo asigna a los elementos del conjunto de muestra:

$$\mathrm{PP}(Q) = 2^{\frac{-1}{M}\sum_{i=1}^{M}\log_2(q_i)} = 2^{\frac{-1}{M}\log_2\left(\prod_{i=1}^{M} q_i\right)} = 2^{\log_2\left(\prod_{i=1}^{M} q_i\right)^{\frac{-1}{M}}} = \frac{1}{\sqrt[M]{\prod_{i=1}^{M} q_i}}$$

Tenemos en definitiva, que el término empleado para medir la "calidad" de los modelos en experimentos comparativos es la Perplejidad (4), un valor que nos da una idea del "factor de ramificación medio". En caso de que todos los elementos sean equiprobables presentará el valor máximo, que coincide con la talla del vocabulario (en cada punto del análisis cualquier palabra es igualmente probable y por tanto las ramificaciones "mas probables" son todas).

N-GRAMAS

Si bien en el campo del reconocimiento de patrones puede considerarse la técnica de los árboles de decisión (en versión muy simplificada) como una de las primeras en haber sido utilizada para resolver problemas con estructura de lenguaje, en el procesamiento del lenguaje natural las primeras técnicas en ser utilizadas fueron las de los N-gramas.

Expresaremos la probabilidad que un modelo estocástico Q asigna a una frase W como $P_Q(W)$. Esta probabilidad puede descomponerse en el producto de la probabilidad de cada palabra una vez que se han dado las que le preceden en la frase:

$$P_Q(W) = \prod_{i=1}^{l} P(\omega_i | \omega_1^{i-1}) \tag{5}$$

donde $W = \omega_1 \ldots \omega_l$ y ω_i^j representa a un segmento $\omega_i \ldots \omega_j$.

En consecuencia se tiene que la determinación de la probabilidad de cada frase del lenguaje, puede basarse en la estimación de la probabilidad que cada palabra del vocabulario tiene de aparecer detrás de otras. Se trata por tanto de establecer una distribución de probabilidad para cada combinación de palabras posible ω_i^j. Esto en la práctica resulta imposible ya que el número de estas combinaciones crece exponencialmente con su longitud. Basta con ver lo que sucede con un lenguaje cuyo vocabulario sea de 100 palabras: cada distribución constara de 100 valores, y habrá 100 distribuciones relacionadas con cadenas previas de longitud 1, 10.000 relacionadas con cadenas de longitud 2, 1.000.000 con cadenas de longitud 3, etc.

Por ello la aplicación de estos modelos suele llevarse a cabo sobre lenguajes restringidos de alguna forma. Lo más habitual suele ser la restricción de la temática de manera que las frases se refieran sólo a una determinada tarea, y en consecuencia el vocabulario se encuentre suficientemente acotado. En estos casos suele hablarse de lenguajes pseudo-naturales

Por otro lado suele reducirse el número de distribuciones asociando la misma distribución a todas las cadenas previas que tengan iguales un cierto número de palabras más recientes. De esta forma la fórmula (5) pasa a ser:

---

[3] Es habitual en la literatura denominar erroneamente a este valor como "entropía" del lenguaje determinado por el modelo (H(Q)). Como se ha visto se trata de la entropía cruzada y solo es válido a la hora de realizar comparaciones de modelos frente a un mismo conjunto de muestra. Si el modelo describe exactamente a la muestra, este valor será la entropía de esta, y si la muestra describe exactamente al lenguaje tendremos la entropía del mismo.

$$P_Q(W) = \prod_{i=1}^{I} P(\omega_i | \omega_{i-N+1}^{i-1}) \quad (6)$$

donde N determina que la longitud de las cadenas previas sea N-1. De aquí que se hable de N-gramas. Habitualmente el valor de N se toma como 2 (y se habla de bigramas) o todo lo más como 3 (y se habla de trigramas). Para valores de N más altos no se han conseguido mejoras significativas en los modelos teniendo en cuenta el aumento de su tamaño [Jelinek 91a].

El modelo de 1-gramas consiste en una sola distribución de probabilidad ya que no se considera ninguna cadena previa a cada palabra (la longitud de la cadena previa es N-1=0).

El establecimiento de los valores para estas distribuciones de probabilidad puede llevarse a cabo mediante el simple recuento del número de veces que se da cada situación en el conjunto de muestra y tomando como valor de la probabilidad el de la frecuencia.

$$P(\omega_i | \omega_{i-N+1}^{i-1}) = \frac{c(\omega_i | \omega_{i-N+1}^{i-1})}{c(\omega_{i-N+1}^{i-1})} \quad (7)$$

donde c(x) representa la cuenta del numero de veces que se da x en la muestra. Esto obedece al criterio de máxima verosimilitud [Duda 73].

Esta demostrado que si bien el modelo de N-gramas se trata de un modelo muy sencillo, es capaz de recoger importantes restricciones de un lenguaje pseudo-natural. No obstante, los N-gramas establecen un modelo de las relaciones entre palabras a corta distancia, de forma que las posibles reglas del lenguaje que actúen sobre palabras que distan entre si más de N posiciones no pueden ser aprendidas.

Suavizado de modelos de N-gramas.

Un importante problema que presentan los N-gramas es el de la falta de cobertura, es decir, el porcentaje de situaciones con que se encuentra el sistema en su aplicación que no han sido aprendidas. Cuanto más alto es el valor de N más situaciones quedan sin ser vistas en la fase de aprendizaje, por lo cual se tiene una peor cobertura. El modelo de 1-gramas presenta una cobertura del 100%. Por otro lado, el aumento del valor de N supone la obtención de un modelo con mayor número de distribuciones de probabilidad, por lo tanto más preciso, de forma que cobertura y precisión son intereses contrapuestos.

En consecuencia para resolver este problema, se adoptan diferentes estrategias que básicamente consisten en utilizar modelos de N-gramas para varios valores de N conjuntamente, de forma que se aprovechen la capacidad de discriminación de los modelos con mayores valores de N y la cobertura de los de N más baja. A continuación enumeramos algunas de éstas técnicas:

- Estimación de Jeffrey [Lehmann 83]. Es el criterio más simple y surge de una aplicación estricta de la teoría Bayesiana de la probabilidad. Consiste en sumar 1 a todas las cuentas, de forma que todas las situaciones no vistas en la fase de aprendizaje se consideran como vistas una vez. Para obtener la probabilidad se calculan las frecuencias de aparición teniendo en cuenta que la cantidad de unidades utilizadas en el aprendizaje se considerará aumentada en un número igual al número de N-gramas posibles. Los resultados experimentales han demostrado que este criterio sobrestima las probabilidades bajas.
- Interpolación lineal [Jelinek 85/86/90]. La probabilidad del modelo se calcula como una interpolación lineal de la dada por los diferentes N-gramas. Para el caso de trigramas tenemos:

  $$P_M(\omega_i | \omega_{i-2}^{i-1}) = \alpha_3 P_{M3}(\omega_i | \omega_{i-2}^{i-1}) + \alpha_2 P_{M2}(\omega_i | \omega_{i-1}) + \alpha_1 P_{M1}(\omega_i) + \alpha_0 \quad (8)$$

  donde los coeficientes α_i pueden ser estimados mediante una técnica de tipo "validación cruzada" [Raudis 91] denominada "deleted interpolation" [Jelinek 80].
- Interpolación no lineal [Ney 91a]. En la fórmula (8) queda patente cómo al modelo más preciso (el de N=3) se le reduce su estimación para permitir que los otros modelos aporten también su información. Esta reducción se realiza independientemente del valor dado por el modelo de trigramas. Es razonable pensar que si este valor es alto, se debe a un aprendizaje suficiente y por tanto no es tan necesario descansar en otros modelos como en otros casos. Por ello se establece una fórmula en que se descuenta una cantidad fija a cada N-grama, siendo por tanto mayor su influencia porcentual para los valores bajos de probabilidad. El valor a descontar se obtiene mediante estimación por algún método de tipo validación cruzada.
- Back-off [Katz 87]. Desde su presentación, de éste método ha sido considerado como muy eficaz ya que conduce a mejores resultados que los anteriores con una complejidad inferior y sin la necesidad de estimación de parámetros de interpolación. La idea en que descansa es la de confiar en la estimación proporcionada por el modelo con valor de N más alto. Para ello se establece el descuento para los elementos de cada modelo y, en caso de que el más preciso no proporcione la probabilidad del N-grama en cuestión, se recurre a asignarle la parte correspondiente del total

descontado de acuerdo con la distribución que proporcione un modelo inferior. La presentación de éste método fue realizada con la utilización de la fórmula de Turing [Good 53] como alternativa a la estimación de máxima verosimilitud que proporciona un descuento a cada N-grama visto en el aprendizaje. No obstante, esta fórmula presenta algunas desventajas por lo que posteriormente se ha reformulado la técnica de back-off con otros métodos de descuento [Ney 93] [Dagan 94].

- Coocurrencias [Essen 92]. La idea aplicada en este método difiere de las anteriores. En este caso no se dispone de modelos menos precisos. Se estudian las semejanzas en la distribuciones que presenta cada palabra frente a sus precedentes de forma que se establecen unos valores de semejanza entre ellas. Cuando un N-grama no ha sido visto en el aprendizaje, se le asigna una probabilidad extraída de las distribuciones de las palabras que se asemejan a ella según el estudio previo. Al ser éste un método no excluyente con los demás, puede aplicarse en paralelo, como se hace en [Dagan 94].

En [Ney 94] se presenta una visión "estructurada" de los problemas a resolver en los modelos de N-gramas con algunas soluciones.

N-POS

Otro importante problema de los modelos de N-gramas es la cantidad de muestras necesarias para el aprendizaje. Considerando que las reglas que ligan las palabras no se establecen en relación con la identidad concreta de cada palabra, sino con su función gramatical, pueden establecerse agrupamientos de palabras. De esta forma no se modelizan las distribuciones frente a secuencias de palabras sino en función de las secuencias de grupos [Jelinek 85/86/90]. Estos grupos se suelen denominar "Parts Of Speech" (POS).

En todo caso, cuanto menor sea el número de grupos formados más sencillo resulta el modelo al reducirse su vocabulario y por tanto el número de parámetros a ser aprendidos. No obstante, aunque menos agravada, existe la misma problemática que en los modelos de N-gramas en lo referente a la cobertura y por ello son también necesarios los métodos de suavizado.

Como ejemplo podemos poner el trabajo referido en [Shikano 87] donde se muestra un lenguaje con un vocabulario de 422 palabras y se definen 89 categorías. En consecuencia, al tratarse de un modelo de 3-gramas/3-POS, el número de probabilidades que debe establecer en la fase de aprendizaje pasa de 75.151.448 (422**3) a 704.969 (89**3).

Es interesante apreciar cómo una asignación de una clase para cada palabra lleva a un modelo equivalente al modelo de N-gramas y, en el extremo opuesto, una asignación de todas las palabras a la misma clase lleva a un modelo equivalente al de 1-gramas.

Una evaluación de estos modelos frente a los de N-gramas la encontramos en [Paeseler 89] donde se describe el uso de un modelo de biPOS y otro de triPOS con las clases, definidas manualmente y atiendo a criterios sintácticos y semánticos. El vocabulario del lenguaje era de 917 palabras y las clases definidas 63. Estos modelos se aplicaron en una tarea de reconocimiento del habla dependiente del locutor haciendo descender las tasas de error de un 21.8% a un 13.7% con biPOS y a un 9,6 con triPOS.

Veamos ahora como son las expresiones de probabilidad en estos casos. Si $g(\omega_i)$ denota la clase a la que pertenece la palabra ω_i tendremos:

$$P\left(\omega_i|\omega_{i-N+1}^{i-1}\right) \triangleq P\left(\omega_i|g(\omega_{i-N+1})\ldots g(\omega_{i-1})\right) \tag{9}$$

Puede también argumentarse que la secuencia de clases determina la siguiente clase, pero no la identidad concreta de la palabra. De esta forma puede establecerse una probabilidad de pertenencia de cada palabra a su clase y tenemos:

$$P\left(\omega_i|\omega_{i-N+1}^{i-1}\right) \triangleq P\left(\omega_i \in g(\omega_i)\right) P\left(g(\omega_i)|g(\omega_{i-N+1})\ldots g(\omega_{i-1})\right) \tag{10}$$

y generalizando esto de forma que una misma palabra pueda pertenecer a varias clases:

$$P\left(\omega_i|\omega_{i-N+1}^{i-1}\right) \triangleq \sum_{\forall g_j \in G} P\left(\omega_i \in g_j\right) P\left(g_j|g(\omega_{i-N+1})\ldots g(\omega_{i-1})\right) \tag{11}$$

Esta última aproximación es la más utilizada en los trabajos que pueden verse en la bibliografía.

Habitualmente, como se ha visto en el caso de [Paeseler 89], la definición de las clases se hace de forma manual. El preprocesado de los textos de aprendizaje para convertirlos en cadenas de POS, se realiza según diversas alternativas que permiten automatizar hasta cierto punto el proceso [Jelinek 85] [Church 89] [Merialdo 91] [Maltese 91]. Desde hace algún tiempo se viene estudiando la efectividad de algunos métodos para obtener las clases de forma automática [Kneser 91] [Kneser 93].

De esta forma, mediante técnicas de agrupamiento, se obtienen no sólo las clases, sino el grado de pertenencia de cada palabra a las mismas y, en algunos casos, la capacidad de realizar la conversión de palabras a clases de forma totalmente automática. Estas agrupaciones tienen la virtud de ser más "naturales" que las que pueden realizarse manualmente ya que, como puede observarse en las experimentaciones referidas en la bibliografía, las clases obedecen en cada caso a diferentes criterios (sintácticos, semánticos, o variantes diversas).

En [Dumouchel 88] se estudian diferentes modelos de N-gramas y N-POS, llegando, entre otras, a la conclusión de que las ventajas de los modelos N-POS se hacen patentes en su uso con lenguajes de temática muy amplia (en este caso concreto el corpus Brown[4]). Para lenguajes de temática restringida (en este caso se utilizo el corpus Hansard[5]) puede darse el caso de que el modelo de N-gramas sea preferible.

Modelos dinámicos

Es conocido que en el lenguaje natural, muchas palabras presentan un comportamiento de agrupamiento local, es decir, a lo largo de un texto hay zonas donde no aparecen y otras zonas donde su frecuencia puede ser elevada. Una de las causas de este comportamiento que puede verse claramente esta en la temática a que se refiere el texto. Podemos poner como ejemplo éste mismo escrito: la palabra "gramatical" no ha aparecido hasta este preciso instante a lo largo de todo el apartado; sin embargo su frecuencia de aparición en el siguiente apartado es elevada.

Esta característica puede ser tenida en cuenta para mejorar los modelos de lenguaje, haciendo que las distribuciones de probabilidad se adapten de acuerdo a cierta información que se obtiene a la par que se procesa un texto. A este tipo de modelos se les denomina "dinámicos" y, por contraposición, a los que no adaptan sus distribuciones de probabilidad se les denomina "estáticos".

Una primera propuesta de un modelo de este tipo fue realizada en [Kuhn 88] y posteriormente desarrollada en [Kupiec 89] y [Kuhn 90]. Caracteriza a estos modelos una componente "cache" que mantiene información de un segmento de L palabras inmediatamente anteriores a la que se procesa. Por el momento la bibliografía relata pocas experiencias con modelos de este tipo y en todo caso los enfoques son muy similares. En [Jelinek 91b] dos modelos de trigramas suavizados se combinan mediante una interpolación lineal para obtener un modelo conjunto. Uno de los dos modelos es estático, obteniéndose de acuerdo a la técnica clásica de los N-gramas. El otro es dinámico: un modelo representando en cada momento las estadísticas recogidas sobre las últimas L palabras del texto que se analiza. El parámetro de interpolación se determina mediante el algoritmo "forward-backward". En [Essen 91] se analiza la aplicación de la componente "cache" de dos maneras: como modificador de la probabilidad de pertenencia de palabras a clases en un modelo de biPOS, y como modificador de la estimación de unigramas en un modelo de bigramas.

ENFOQUE GRAMATICAL DE LA MODELIZACIÓN ESTOCÁSTICA DEL LENGUAJE.

La modelización de lenguajes con N-gramas (o N-POS) tiene una limitación evidente en el hecho de que la información que adquiere el modelo hace referencia a segmentos locales de palabras. Esto hace imposible modelizar relaciones a larga distancia. Este tipo de relaciones no presentan ninguna dificultad en la especificación de lenguajes formales mediante gramáticas, por lo que resulta interesante aplicar los conocimientos de la Teoría de Lenguajes Formales a la modelización automática de lenguajes. Además, el interés también se debe a que la Teoría de Lenguajes Formales es una disciplina bien asentada, con una formulación rigurosa que permite desarrollos en un marco puramente teórico. La conexión de esta disciplina con las técnicas de modelización del lenguaje a partir de muestras se encuentra en la Inferencia Gramatical (IG).

Como ya se ha visto anteriormente, también será necesario un enfoque estocástico al plantear la modelización del lenguaje desde el punto de vista gramatical. En [Gonzalez 78] o [Fu 82] pueden encontrarse las bases de la teoría de lenguajes en su extensión estocástica. El problema de la determinación de las reglas de la gramática estocástica asociada al lenguaje a partir de las muestras se encuentra formalizado en [Gold 67].

La clasificación de lenguajes dada por Chomsky en función del tipo de reglas de sus gramáticas generadoras, representa también una clasificación de la dificultad de su aprendizaje automático. Los lenguajes más restrictivos de dicha clasificación, los Lenguajes Regulares, son los que presentan una menor dificultad.

---

[4] Textos en inglés-americano extraidos de fuentes muy diversas en cuanto a temática y estilo, totalizando 1.1 millones de palabras.

[5] Textos extraidos de las actas del Parlamento Canadiense con un total de 15 millones de palabras.

En el anteriormente mencionado trabajo de Gold [Gold 67], se encuentra un estudio de los límites del aprendizaje automático. Estos límites son diferentes en función de si se dispone de una muestra completa o únicamente positiva. Para el caso de una muestra completa, cualquier clase de lenguaje (dependiente de contexto o más simple) puede ser identificado en el límite (cuando el tamaño de la muestra crece). Pero es habitual no disponer de muestras negativas y en este caso surge la imposibilidad de identificación (ni siquiera en el límite) de ninguna clase superfinita[6] de lenguajes. En consecuencia, dado que la clase de los Lenguajes Regulares es superfinita [Hopcroft 79], ni tan siquiera ésta es identificable.

Por tanto, dada la limitación anterior, la inferencia de lenguajes se aborda en dos direcciones diferentes: por una parte se intenta caracterizar clases de lenguajes no superfinitos que puedan presentar algún interés, y por otro se establecen heurísticos basados en el conocimiento a priori de la tarea a abordar y que dan lugar a lenguajes no caracterizados (se desconocen sus propiedades).

El desarrollo de la Inferencia Gramatical para clases dentro de los Lenguajes Regulares es bastante amplio debido a su utilidad en muchos problemas prácticos. Incluso lenguajes que no son Regulares pueden aproximarse con arbitraria precisión mediante gramáticas regulares estocásticas [Fu 82]. Los métodos de inferencia desarrollados son numerosos ya que cada uno de ellos determina una subclase en particular de los Lenguajes Regulares.

El siguiente nivel en dificultad es el de los Lenguajes Incontextuales. Estos suponen un fuerte incremento en la complejidad de los métodos a aplicar y en los requerimientos computacionales necesarios frente a los anteriores. También en este caso los métodos desarrollados infieren, en general, subclases particulares.

Una revisión de las técnicas de aprendizaje sintáctico puede encontrarse en [Vidal 93].

Gramáticas regulares

A continuación se presenta una referencia de algunos de los métodos de inferencia de subclases de Lenguajes Regulares. Como puede verse tan solo uno de ellos es caracterizable siendo el resto basados en heurísticos. Un bloque de heurísticos está basado en criterios de "fusión de estados", mientras que otros obedecen a estrategias particulares.

- [Bierman 72]. Método de las k-colas (heurístico). Basado en la construcción de un árbol aceptor de prefijos[7] de las frases de aprendizaje y la fusión en uno todos aquellos estados que tengan una misma secuencia de palabras hasta un estado final.
- [Levine 82] (heurístico). Una variante del algoritmo de las k-colas en el que se establece un parámetro que mide el grado de similitud entre las secuencias que siguen a dos estados de manera que puede establecerse un valor umbral para la fusión de estados.
- [Miclet 80] (heurístico). Método de fusión de estados basado en técnicas de agrupamiento estadísticas. También se parte de el árbol aceptor de prefijos.
- [Angluin 82]. Inferencia de lenguajes k-reversibles (caracterizable). Estos lenguajes tienen la propiedad de sus autómatas asociados son "deterministas con adelanto k"[8] y los reversos de sus autómatas son deterministas.
- [Muggleton 84]. Inferencia de lenguajes k-contextuales (heurístico). Estos lenguajes son similares a los k-reversibles con la diferencia que las condiciones exigidas se aplican al revés: su grafo es determinista y su reverso es "determinista con adelanto k". En el mismo artículo se presentan los anteriores métodos así como un algoritmo para su "aplicación" a partir de heurísticos.
- [Prieto 91] (heurístico). Método de corrección de errores (Error Correcting Grammatical Inference - ECGI). Este método comienza construyendo el aceptor para la primera frase de entrenamiento y posteriormente busca el mejor alineamiento de las frases que van llegando con las aceptadas por el grafo. Este alineamiento origina la adición de nuevos estados y transiciones.
- [García 87] (heurístico). Método de los generadores mórficos (Morfic Generator Grammatical Inference - MGGI). Consiste en definir una función de renombrado que al aplicarla sobre las muestras permita la inferencia de un lenguaje local (2-explorable) y posteriormente aplicar un morfismo alfabético (típicamente la función de renombrado inversa) sobre el grafo construido. Si bien este método requiere una intervención específica para cada caso en la búsqueda de las transformaciones, tiene la importante característica de ser capar de inferir cualquier lenguaje

---

[6] Las clases superfinitas son aquellas que contienen todos los lenguajes finitos y al menos uno infinito.

[7] Arbol formado por todas las ramificaciones, asociadas a cada palabra, necesarias para dar cuenta de las frases de entrenamiento a partir de un único estado inicial.

[8] Es decir, una secuencia de k palabras sólo tiene un camino posible a través del grafo.

regular ya que puede demostrarse que para todo lenguaje regular existe un homomorfismo alfabético que transforma a un lenguaje local en él.

Una técnica aplicable a la inferencia de la clase completa de los modelos de Lenguajes Regulares puede consistir en establecer un autómata de estados finitos con un número de estados adecuado [Ziv 92] y todas las transiciones establecidas con una inicialización de sus probabilidades según algún criterio. Posteriormente mediante la aplicación de los algoritmos de Viterbi o Baum-Welch pueden irse reestimando los valores de las probabilidades de transición. Esta es un alternativa poco explorada que tendría la virtud de, a diferencia de los demás métodos, proporcionar cualquier lenguaje regular. Su gran problema estriba en que los métodos de reestimación son fuertemente dependientes de la inicialización, de forma que la estimación final de la distribución de probabilidades óptima no está resuelta.

Los N-gramas como gramáticas regulares

Recientemente se ha demostrado que los lenguajes definidos por los modelos de N-gramas son equivalentes a una subclase de los Lenguajes Regulares [Segarra 93]. Estos son los llamados Lenguajes K-Explorables en Sentido Estricto (K-EE). El interés de esta clase de lenguajes dentro de los Regulares estriba en que es caracterizable. La definición formal de estos lenguajes puede darse de la siguiente manera:

Dado un alfabeto Σ, un lenguaje K-EE es un subconjunto de Σ* formado por todas las cadenas con segmentos iniciales pertenecientes a un determinado conjunto I, segmentos finales pertenecientes a otro conjunto F y que no contienen ningún segmento de un tercer conjunto T. Los elementos de los conjuntos F e I son cadenas de elementos de Σ con longitud menor o igual a K, y los elementos del conjunto T son cadenas de elementos de Σ con longitud K.

$$L_{K-EE} = \Gamma(\Sigma, I, F, T) \triangleq I\Sigma^* \cap \Sigma^* F \; - \; \Sigma^* T \Sigma^*$$

$$donde \quad \Sigma \equiv alfabeto, \quad I, F \subset \left\{ \bigcup_{i=1}^{k-1} \Sigma^i \right\}, \quad T \subset \Sigma^k$$

Un caso especial de estos lenguajes es el subconjunto de los llamados "lenguajes locales" para los que K toma el valor 2.

Los lenguajes K-EE no son cerrados bajo operaciones booleanas, de forma que, por ejemplo, la unión de dos lenguajes K-EE definidos sobre un mismo alfabeto puede ser no K-EE. Al cierre de los lenguajes K-EE respecto a las operaciones booleanas se le denomina clase de los lenguajes K-explorables (no en sentido estricto), y su inferencia no está resulta.

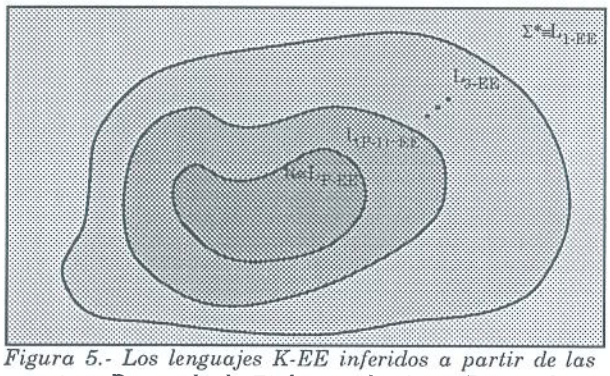


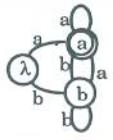
Figura 5.- Los lenguajes K-EE inferidos a partir de las muestras R van desde Σ hasta el mismo R cuando se aumenta K.*

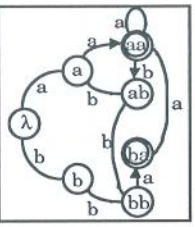
Al inferir lenguajes K-EE a partir de muestras, se tiene que el lenguaje inferido para una determinada K está incluido en el inferido para K-1, obteniendo así una serie de lenguajes que irían desde el dado para K=1 (y que coincide con el monoide libre Σ*) hasta un cierto valor de K para el cual el lenguaje inferido coincide con las frases de entrenamiento (y en consecuencia valores de K mayores generan el mismo lenguaje).

En [García 90] se presenta un método de construcción del grafo correspondiente a la inferencia de lenguajes K-EE que consiste básicamente en, para toda frase, partir de un estado inicial λ y generar una transición para cada símbolo y un nuevo estado si es necesario, donde los estados se etiquetan con la cadena de los K-1 últimos símbolos vistos (o menos en los comienzos de frase). A continuación se desarrolla un ejemplo de inferencia de lenguajes K-EE a partir de un conjunto de muestras positivas R+:

R+≡{abba,aaabba,bbaaa,bba}

Para K=1 el lenguaje inferido es Σ*: I, segmentos iniciales de longitud 0: ϕ F, segmentos finales de longitud 0: ϕ T, segmentos prohibidos de longitud 1: ϕ $L_{1\text{-}EE\,(R+)} \equiv (\{a,b\},\phi,\phi,\phi) \equiv (a \mid b)^*$	

Para K=2: I, segmentos iniciales de longitud 1: {a,b} F, segmentos finales de longitud 1: {a} T, segmentos prohibidos de longitud 2: ϕ $L_{2\text{-}EE} \equiv (\{a,b\},\{a,b\},\{a\},\phi) \equiv (a \mid b)^* a$	

Para K=3: I, segmentos iniciales de longitud 1 y 2: {aa,ab,bb} F, segmentos finales de longitud 1y 2: {aa,ba} T, segmentos prohibidos de longitud 3: {aba,bab,bbb} $L_{3\text{-}EE} \equiv (\{a,b\},\{aa,ab,bb\},\{aa,ba\},\{aba,bab,bbb\}) \equiv (a \mid abb \mid bb)a^+(bba^+)^*$	

De esta forma vemos como cada modelo reconoce un lenguaje cada vez más reducido y en el caso mostrado tendremos que para K=5 el modelo identifica exclusivamente las cuatro frases del conjunto de entrenamiento R+.

En [Bordel 93] se encuentran diversas experiencias llevadas a cabo con lenguajes K-EE sobre el corpus "bdgeo" de la base de datos "Albaycin" [Diaz 93].

Gramáticas de contexto libre.

Como se ha mencionado con anterioridad, los Lenguajes Regulares Estocásticos son capaces de aproximarse con precisión a lenguajes que no lo son. Naturalmente esto es posible en muchas ocasiones estableciendo modelos aparatosos y poco eficaces. Por ello es interesante disponer de técnicas de inferencia de lenguajes más amplios. La siguiente clase en la jerarquía de Chomsky es la de los Lenguajes Incontextuales.

El análisis de textos pertenecientes a Lenguajes Incontextuales a partir de sus gramáticas se lleva a cabo mediante los algoritmos de Earley y de Cocke-Younger-Kasami (CYK) [Aho 86]. Para trasladarlos al área de la modelización del lenguaje es necesaria su extensión estocástica. En algunos trabajos de Reconocimiento Automático de Habla se han planteado versiones estocásticas del algoritmo CYK [Levinson 85] [Ney 87] [Miller 88] [Ney 91b] así como del de Earley [Paeseler 88]. El problema de la estimación de las probabilidades de las reglas puede ser obtenido según el criterio de máxima verosimilitud sólo en caso de gramáticas no ambiguas [Chanduri 86]. Para gramáticas ambiguas se dispone del algoritmo "inside-outside" [Lari 91] que sólo lleva a estimaciones localmente óptimas debido a su dependencia del estado inicial.

Una posible vía de modelización de Lenguajes Incontextuales viene dada por el hecho de que el conjunto de los arboles de derivación de las sentencias pertenecientes al lenguaje es un "lenguaje de arboles" de tipo Regular. De esta forma, la adaptación de los métodos conocidos en inferencia de Lenguajes Regulares para tratar con árboles en lugar de con cadenas, daría lugar a sus métodos asociados para el tratamiento de Lenguajes Incontextuales [Vidal 93].

ARBOLES DE DECISIÓN.

Al comienzo del apartado sobre N-gramas se hacia una referencia a cómo en el campo del reconocimiento de patrones, puede considerarse la técnica de los árboles de decisión como una de las primeras en haber sido utilizada para resolver problemas con estructura de lenguaje. Básicamente consiste en disponer de las distribuciones de probabilidad a aplicar en las "hojas" de una estructura de árbol. Los nodos de este árbol tienen asociadas cuestiones relativas al elemento que se pretende caracterizar, y una ramificación hacia un nodo "hijo" para cada posible respuesta. El procedimiento consiste en situarse en el nodo raíz y plantear su cuestión asociada. En función de la respuesta a la misma se pasará al nodo "hijo" correspondiente. El proceso se repite hasta llegar a una de las "hojas", momento en el que queda determinada la distribución a aplicar.

El establecimiento de las cuestiones asociadas a los nodos se realizaba de acuerdo con el conocimiento que se tenia de la tarea a llevar a cabo, siendo el resultado dependiente del acierto que se tuviera en la selección y organización de dichas cuestiones.

Actualmente esta técnica ha resurgido gracias a que el progreso en la capacidad de cálculo de los sistemas de computación permite abordar la cuestión de la ordenación de las cuestiones de forma automática [Gelfand 91a]. Algunos algoritmos de este tipo pueden verse en [Breiman 84] [Chou 91] y [Gelfand 91b]. El proceso se lleva a cabo de la siguiente manera: se comienza con un solo nodo —el raiz— y un conjunto de cuestiones sin que ninguna de ellas este asociada al nodo. Se analizan todos

los árboles formados al asociar al nodo cada una de las cuestiones y establecer los nodos hijos. Cada uno de ellos se aplica sobre el conjunto de aprendizaje (como se indicaba en el procedimiento clásico) para determinar las distribuciones de las hojas y posteriormente se obtiene una medida de la calidad del modelo (normalmente aplicándolo sobre un conjunto de test). De esta forma el árbol que presenta la máxima calidad es seleccionado. Este proceso se repite para todo nuevo nodo hasta que se cumpla una cierta condición de parada (por ejemplo que ninguna cuestión aumente la calidad del modelo).

Pese a la capacidad de esta metodología para generar una clase muy amplia de modelos y de establecer un conjunto de distribuciones de probabilidad adecuado, los resultados obtenidos hasta el momento sólo han conducido a mejoras poco significativas [Bahl 90]. A esto debe añadirse que el procedimiento de formación de los arboles tiene un muy elevado coste computacional de forma que, así como hace no mucho tiempo era en la práctica inabordable, actualmente supone tiempos de ejecución extremadamente largos que dificultan la experimentación.

N-gramas y arboles de decisión.

La capacidad de generación de modelos mediante arboles de decisión es muy amplia. A modo de ejemplo podemos ver cómo los modelos de N-gramas (de igual forma N-POS) son un caso particular de los árboles de decisión y cómo la mayor flexibilidad de éstos permite atenuar ciertas características negativas de los N-gramas.

Como se veía en el apartado referente a N-gramas, uno de tales modelos queda definido por un conjunto de distribuciones de probabilidad donde cada una de ellas se asocia a una cadena de palabras de longitud N-1 (ω_{i-N+1}^{i-1}). Por tanto un árbol de decisión equivalente dispondrá de una hoja para cada ω_{i-N+1}^{i-1} con la distribución adecuada. Para alcanzar una hoja a partir del nodo raíz, las cuestiones harán referencia a la identidad de las palabras ω_j con $i-N+1 \le j \le i-1$. Pongamos por caso un modelo de bigramas. Podemos proponer un árbol binario en cuyo nodo raíz se cuestione si la palabra ω_{i-1} pertenece a la primera mitad del alfabeto o a la segunda, y en cada uno de los subnodos se siga dividiendo el alfabeto en dos mitades hasta llegar al nivel de profundidad en que la subdivisión determine la identidad concreta de ω_{i-1}. En ese nodo hoja se encontrara la distribución de probabilidad que se deberá utilizar para obtener la probabilidad correspondiente a ω_i. Un modelo de trigramas podría ser igual hasta el momento de determinar ω_{i-1} y continuar del mismo modo determinando la identidad de ω_{i-2}.

Una de las ventajas del procedimiento automático de asignación de cuestiones a nodos estriba en que las cuestiones con mayor poder de discriminación en cada caso son seleccionadas de forma que, llegados a un punto, puede darse por terminada la formación del árbol sin que se disponga de hojas para todo ω_{i-N+1}^{i-1}. Esto implica dejar en una misma hoja, es decir con una misma distribución de probabilidad, varias situaciones diferentes al considerar que su distinción no aporta información interesante. En un modelo de N-gramas se construyen todas las distribuciones que tienen representación en el conjunto de muestra independientemente de las semejanzas entre unas y otras, por lo que el modelo es mayor necesitando más aprendizaje.

MODELOS COMBINADOS

Con la intención de explotar las ventajas de cada tipo de modelo y atenuar sus desventajas, se han planteado en ocasiones modelos combinados. En [Rohlicek 88], para una tarea restringida y un conjunto de aprendizaje reducido, se combinan gramáticas regulares que describen las frases con un modelo de bigramas para especificar las probabilidades de secuencias de ellas. En [Derouault 86] se combinan, a modo de producto, las probabilidades que proporcionan un modelo de triPOS y una gramática regular. En [El-Bèze 90] se combina también un modelo de triPOS con información sobre la morfología del francés. En [Wright 92] la combinación se lleva a cabo entre bigramas, gramáticas incontextuales estocásticas e información sobre la dependencia de primer orden entre las reglas de la gramática.

REFERENCIAS

[Aho 86] A. V. Aho, R. Sethi, J. D. Ullman, "**Compilers. Principles, Techniques and Tools,**" Addison-Wesley, 1986.

[Angluin 82] D. Angluin, **"Inference of Reversible Languages,"** *Journal of the Association for Computing Machinery*, vol.29, nº.3, 1982.

[Bahl 90] L. R. Bahl, P. F. Brown, P. V. de Sonza, R. L. Mercer, "**A tree-based statistical language model for natural language speech recognition,**" en A. Waibel and K.F. Lee editores, *Readings in Speech Recognition*, pp. 507-514, Morgan Kaufmann, San Mateo, CA, 1990.

[Baum 72] L. E. Baum "**An inequality and associated maximization technique in statistical esimation for probabilistic functions of Markov processes,**" *Inequalities*, vol. 3, pp. 1-8, 1972.

[Bordel 93] G. Bordel, "**Modelización del lenguaje: una visión general desde el análisis de los lenguajes K-explorables en sentido estricto (N-gramas),**" *Research report DSIC-II/40/93 DEE-1/93*, Universidad Politécnica de Valencia, Universidad del País Vasco, 1993.

[Bierman 72] A. W. Biermann, J.A.Feldman, **"On the Synthesis of Finite-State Machines from Samples of Their Behaviour,"** *IEEE Transactions on Computers*, C-21, pp. 592-597, 1972.

[Breiman 84] L. Breiman, J, H, Friedman, R. A. Olshen, C. J. Stone, **"Classification and Regression Trees,"** Wadsworth Ic., Belmont CA, 1984.

[Chandhuri 86] R. Chandhuri and T. L. Booth, **"Approximating Grammar Probabilities: Solution of a Conjecture,"** *Journal ACM*, vol 33, nº 4, pp. 702-705, 1986.

[Chou 91] P.A. Chou, **"Optimal partitioning for classification and regression trees,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-13, nº 4, pp. 340-354, 1991

[Church 89] K. W. Church, **"A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text,"** *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 695-698, Glasgow, May. 23-26, 1989.

[Dagan 94] I. Dagan, F. Pereira, L. Lee, **"Similarity-Based Estimation of Word Coocurrence Probabilities,"** a aparecer en *Proceedings if the 32nd Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, junio 1994.

[Derouault 86] A. M. Derouault, B. Merialdo, **"Natural language Modelling for Phoneme-to-Text Transcription,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, nº 6, pp. 742-749, nov. 1986.

[Diaz 93] J. E. Diaz, A. J. Rubio, A. M. Peinado, E. Segarra, N. Prieto and F. Casacuberta, **"Development of Task Oriented Spanish Speech Corpora,"** *Proceedings of EUROSPEECH 93* , 1993.

[Duda 73] R. O. Duda, P. E. Hart, **"Pattern Classification and Scene Analysis,"** John Wiley & Sons, 1973.

[Dumouchel 88] P. Dumouchel, V. Gupta, M. Lenning, P. Mermelstein, **"Three probabilistic Languaje Models for a Large-Vocabulary Speech Recognizer,"** *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 513-516, 1988.

[El-Bèze 90] M. El-Bèze, A. M. Derouault, **"A Morphological Model for Large Vocabulary Speech Recognition,"** *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 577-580, 1990.

[Essen 91] U.Essen, H.Ney, **"Statistical Language Modelling Using a Cache Memory,"** *Proceedings of QUALICO*, sep.1991.

[Essen 92] U. Essen and V. Steinbiss, **"Coocurrence Smoothing for Stochastic Language Modelling,"** *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 161-164, 1992.

[Forney 72] G. D. Forney, **"The Viterbi Algorithm,"** *Proceedings of the IEEE*, vol. 61, pp.268-278, mar. 1973.

[Fu 74] K. S. Fu, **"Syntactic Methods in Pattern Recognition,"** serie Mathematics in Science and Egineering, Academic Press, N. Y., 1974

[Fu 82] K. S. Fu, **"Syntactic Pattern Recognition and Applications,"** Prentice Hall, 1982.

[García 87] P.García, E. Vidal, F. Casacuberta, **"Local Languajes, the Succesor Method and a Step towards a General Methodology for the Inference of Regular Grammars,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, nº. 6, pp. 841-877, 1987.

[García 90] P. García and E. Vidal, **"Inference of k-testable languages in the strict sense and application to syntactic pattern recognition,"** *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol 12, nº 9, pp. 920-925, sep. 1990.

[Gelfand 91a] S. B. Gelfand, E. J. Delp, **"On tree Structured Classifiers,"** en I. K. Sethi y A. K. Jain editores, *Artficial Neural Networks and Statistical Pattern Recognition: Old and new connections*, pp. 51-70, North Holland, Amsterdam, 1991.

[Gelfand 91b] S. B. Gelfand, C. Ravishankar, E. J. Delp, **"An iterative growing and pruning algorithm for classification tree design,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-13, nº. 2, pp.163-174, 1991.

[Gold 67] E. M. Gold, **"Languaje Identification in the Limit,"** *Information and Control*, vol. 10, pp. 447-474, 1967.

[Gonzalez 78] R. C. Gonzalez, M. G. Thomason, **"Syntactic Pattern Recognition. An Introduction,"** Addisson-Wesley, 1978.

[Good 53] I. J. Good, **"The Population Frequencies of Species and the estimation of population parameters,"** *Biometrika*, vol. 40, parts 3 and 4, pp. 237-264, 1953.

[Hopcroft 79] J. E. Hopcroft, J. D. Ullman, **"Introduction to Automata Theory, Languages, and Computation,"** Addison-Wesley, 1979.

[Jelinek 80] F. Jelinek, R. L. Mercer, **"Interpolated Estimation of Markov Source Parameters from Sparse Data,"** *Proceedings of the Workshop on Pattern Recognition in Practice*, E. S. Gelsema and L. N. Kanal editores, Amsterdam, pp. 381-402, 1980.

[Jelinek 85] F. Jelinek, **"Markov Source Modelling of Text Generation,"** *The impact of processing techniques on communication.*, Ed. J.K.Skwirzynski, Nijhoff, Dordrecht, The Netherlands, 1985.

[Jelinek 85/86/90] F. Jelinek, **"Self-Organized Languaje Modeling for Speech Recognition,"** IBM Research Report, NY, 1985. //tambien en// IBM Europe Institute, *Advances in Speech Processing*, Oberlech, Austria, jul. 14-18, 1986. //tambien en// en A. Waibel and K.F. Lee editores, *Readings in Speech Recognition*, pp. 450-506, Morgan Kaufmann, San Mateo, CA, 1990.

[Jelinek 91a] F. Jelinek, "**Up from trigrams: the struggle for improved language models,**" *proc. of the Eurospeech 91*; Genova, Italy, pp.1037-1039, sep. 24-26, 1991.

[Jelinek 91b] F. Jelinek, B. Merialdo, S. Roukos, M. Strauss, "**A dynamic language model for speech recognition,**" *Proceedings of DARPA Speech and Natural Language Workshop*, pp. 293-295, Feb. 1991.

[Jelinek 92] F. Jelinek, J. D. Lafferty and R. L. Mercer, "**Basic Methods of Probabilistic Context Free Grammars,**" *Speech Recognition and Understanding*, Springer-Verlag, 1992.

[Katz 87] S. M. Katz, "**Estimation of Probabilities from Sparse Data for The Language Model Component of a Speech Recognizer,**" *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-35, nº.3, pp.400-401, march 1987.

[Kneser 91] R. Kneser, H. Ney, "**Forming Word Classes by Statistical Clustering for Statistical Language Modelling,**" *Proceedings of the QUALICO*, 1991.

[Kneser 93] R. Kneser, H. Ney, "**Improved Clustering Techniques for class-based statistical language modelling,**" *European Conference on Speech Communication and Technology*, pp. 973-976, Berlin, Alemania, sep. 1993.

[Kuhn 88] R. Kuhn, "**Speech recognition and the frequecy of recently used words: A modified Markov model for natural language,**" *International Conference of Computational Linguistics*, pp. 348-350, Budapest, Aug. 1988.

[Kuhn 90] R. Kuhn, R. de Mori, "**A cache-based natural language model for speech recognition,**" *IEEE Transactions on Patern Analysis and Machine Intelligence*, vol. PAMI-12, nº. 6, pp. 570-583, 1990.

[Kupiec 89] J. Kupiec, "**Probabilistic models of short and long distance word dependencies in running text,**" *Proceedings of DARPA Speech and Natural Language Workshop*, pp. 290-295, Feb. 1989.

[Lari 90] K. Lari, S. J. Young, "**The Estimation of Stochastic Context-free Grammars Using the Inside-Outside Algorithm,**" *Computer Speech and Language*, nº 4, pp.35-36, 1990.

[Lehmann 83] E. L. Lehmann, "**Theory of Point Estimation,**" Wiley, New York, 1983.

[Levine 82] B. Levine, "**The Use of Tree Derivatives and Sample Support Parameter for Inferring Tree Systems,**" *IEEE Transactions on Patern Analysis and Machine Intelligence*, vol. PAMI-4, pp. 25-34, 1982.

[Levinson 85] S. E. Levinson, "**Structural Methods in Automatic Speech Recognition,**" *Proceedings of the IEEE*, vol. 75, nº 11, pp. 1625-1650, 1985.

[Maltese 91] G. Maltese, F. Mancini, "**A thecnique to automatically assign parts-of-speech to words taking into account word-ending information through a probabilistic model,**" *Proceedings of the Eurospeech 91*, pp. 753-756, 1991.

[Merialdo 91] B. Merialdo, "**Using Probabilistic Models for tagging english text,**" *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, May. 14-17, 1991.

[Miclet 80] L. Miclet, "**Regular inference with a tail clustering method,**" *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-10, pp. 737-743, 1980.

[Miller 88] L. G. Miller, S. E. Levinson, "**Syntactic Analysis for Large Vocabulary Speech Recognition Using a Context Free Covering Grammar,**" *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 271-274, 1988.

[Muggleton 84] S. Muggleton, "**Induction of Regular Languages from Positive Examples,**" TIRM-84-009, The Turing Institute, 1984.

[Ney 87] H. Ney, "**Dynamic Programming Speech Recognition Using a Context Free Grammar,**" *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 69-72, 1987

[Ney 91a] H. Ney, U. Essen, "**On Smoothing Techniques for Bigram-Based Natural Language Modelling,**" *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 825-828, 1991.

[Ney 91b] H. Ney, "**Dynamic Programming Parsing for Context-free Grammars in Continuous Speech Recognition,**" *IEEE Transactions on Signal Processing*, vol. SP-39, nº 2, pp. 336-340, 1991.

[Ney 93] H. Ney, "**Architecture and Search Strategies for Large-Vocabulary Continuous-Speech Recognition,**" *NATO ASI. New Advances and Trends in Speech Recognition and Coding*, Lectures, pp. 59-84. Bubión, 1993.

[Ney 94] H. Ney, U. Essen and R. Kneser, "**On Structuring Probabilistic Dependencies in Stochastic Language Modelling,**" *Computer Speech and Language* 8, pp 1-38, 1994.

[Paeseler 88] A. Paeseler, "**Modification of Earley´s Algorithm for Speech Recognition,**" Recent Advances in Speech Understanding and Dialog Systems, H. Nieman et al. editores, NATO ASI series, vol. F-46, pp. 465-472, Springer-Verlag, 1988

[Paeseler 89] A. Paeseler and H. Ney, "**Continuous-Speech recognition using a Stochastic Language Model,**" *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 719-722, 1989.

[Prieto 91] N. Prieto, E. Vidal, "**Learning Languaje Models Throughg ECGI Method,**" *Proceedings of the Eurospeech*, vol. 2, pp. 395-398, 1991.

[Prieto 94] N. Prieto, Tesis Doctoral (en preparación), DISC, Universidad Politécnica de Valencia, 1994.

[Raudys 91] S. J. Raudys and A. K. Jain, "**Small Sample Effects in Statistical Pattern Recognition: Recommendations for Practitioners and Open Problems**," *IEEE Trans. on PAMI*, vol. 13, n3, pp. 252-263, 1991.

[Rohlicek 88] J. R. Rohlicek, Y. L. Chow, S. Roucos, "**Statistical Language Modelling Using a Small Corpus from an Application Domain**," *Proceedings of the EEE International Conference on Acoustics, Speech and Signal Processing*, pp. 267-270, 1988.

[Segarra 93] E. Segarra, "**Una Aproximación Inductiva a la Comprensión del Discurso Continuo**," Tesis Doctoral, DISC, Universidad Politécnica de Valencia, 1993.

[Shannon 64] C. E. Shannon, W. Weaver, "**The Mathematical Theory of Communication**," The University of Illinois Press, Urbana, Il, 1964.

[Vidal 93] E. Vidal, F. Casacuberta and P. García, "**Syntactic Learning Techniques for Language Modelling and Acoustic-Phonetic Decoding**," *Lectures of the NATO ASI* (New Advances and Trends in Speech Recognition and Coding), pp. 95-201, Bubión jun-jul 1993.

[Wright 92] J. H. Wright, G. J. F Jones, E. N. Wrigley, "**Hybrid grammar-bigram speech recognition system with first-order dependence model**," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 169-172, 1992

[Ziv 92] J. Ziv, N. Merhav, "**Estimating the Number of States of a Finite-State Source**," *IEEE Transactions on Information Theory*, vol. 38, nº. 1, jan 1992.