MIARFID Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

TEMA 1. MODELOS DE LENGUAJE

1. Modelos de lenguaje y evaluación
2. Modelos estadísticos y técnicas de suavizado
3. Modelos gramaticales y otros
4. Herramientas de modelización del lenguaje

Modelos de Lenguaje

Objetivo principal: capturar las regularidades del lenguaje (natural) para mejorar las prestaciones de aplicaciones en tecnologías del lenguaje.

Aplicaciones:

Reconocimiento automático del habla

Traducción automática

Clasificación de documentos

Reconocimiento óptico de caracteres

Recuperación de información

Reconocimiento de escritura manuscrita

Corrección ortográfica

1.1 MODELOS DE LENGUAJE Y EVALUACIÓN

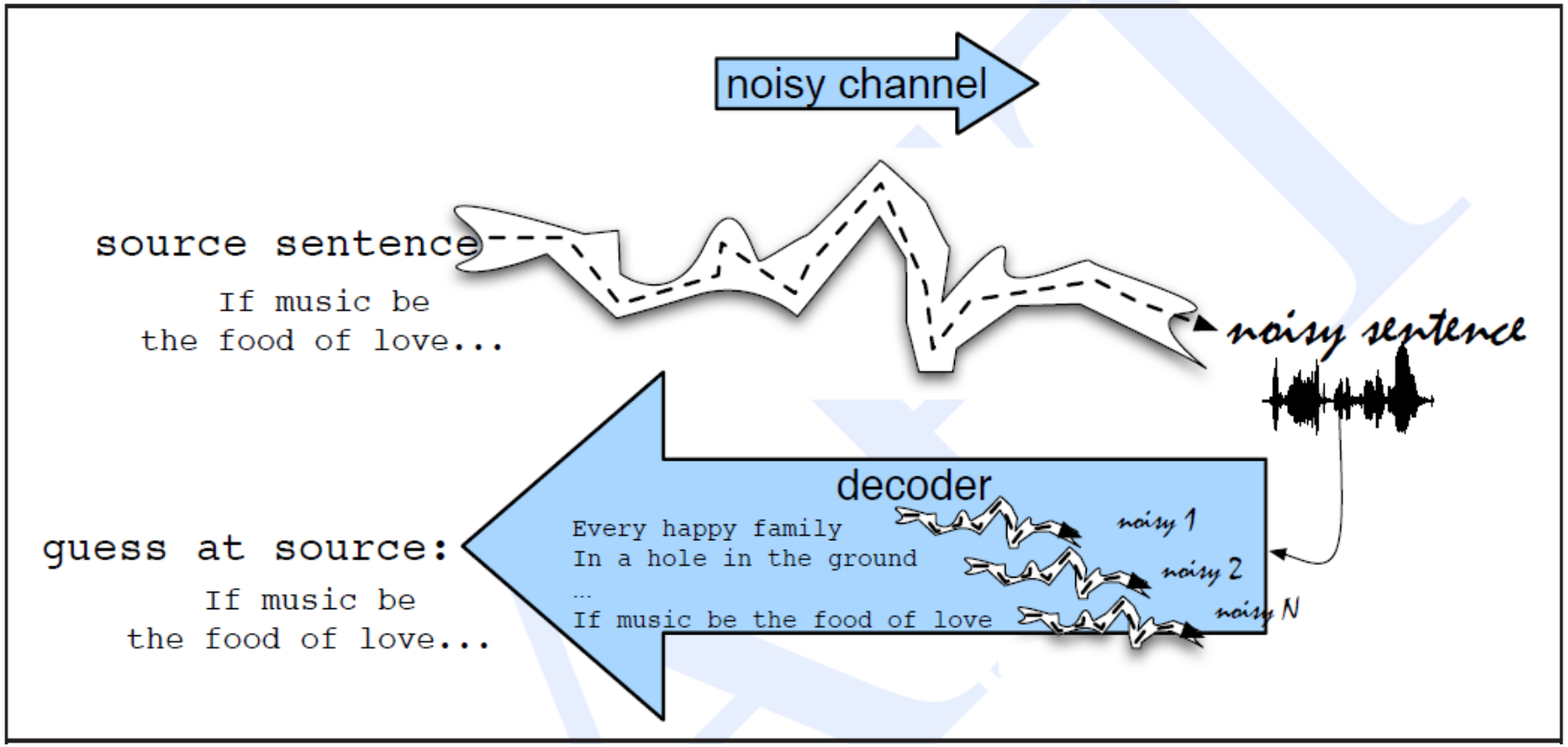
Modelos de lenguaje y evaluación

- **Planteamiento general del reconocimiento automático del habla**
- **Modelos de lenguaje**
- **Clasificación de modelos de lenguaje**
- **Evaluación de modelos de lenguaje**

Modelos de lenguaje y evaluación

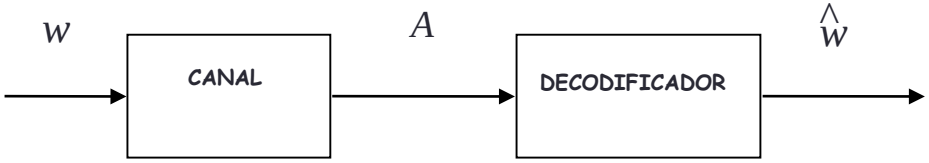
- **Planteamiento general del reconocimiento automático del habla**
- **Modelos de lenguaje**
- **Clasificación de modelos de lenguaje**
- **Evaluación de modelos de lenguaje**

Planteamiento general del RAH



Planteamiento general del RAH

El modelo del canal ruidoso en lingüística computacional



Dados:

Un conjunto unidades lingüísticas $W=\{w_1,w_2,\ldots,w_{|W|}\}$,

Una secuencia acústica $A=A_1A_2..A_{|A|}$

El objetivo de un sistema de RAH es encontrar la secuencia de unidades lingüísticas $\hat{w}= w_1,w_2,\ldots,w_{|w|}$ con $w_i \in W$ y $1<=|w|<=|A|$, que mejor se adapta a la secuencia acústica pronunciada.

$$\hat{w} = \arg\max_{w \in W^+} P(w|A)$$

Planteamiento general del RAH

Utilizando la regla de Bayes:

$$\hat{w} = \arg\max_{w \in W^+} \left(P(w) P(A|w) / P(A) \right)$$

Como *P(A)* es independiente de *w*:

$$\hat{w} = \arg\max_{w \in W^+} P(w) P(A|w)$$

Aparecen dos distribuciones de probabilidad a considerar:

- *P(w)* es la probabilidad a priori de la secuencia *w*, lo que llamaremos ***modelo de lenguaje***.
- *P(A/w)* es la probabilidad del canal, el ***modelo acústico***.

Sistema de Reconocimiento Automático del Habla

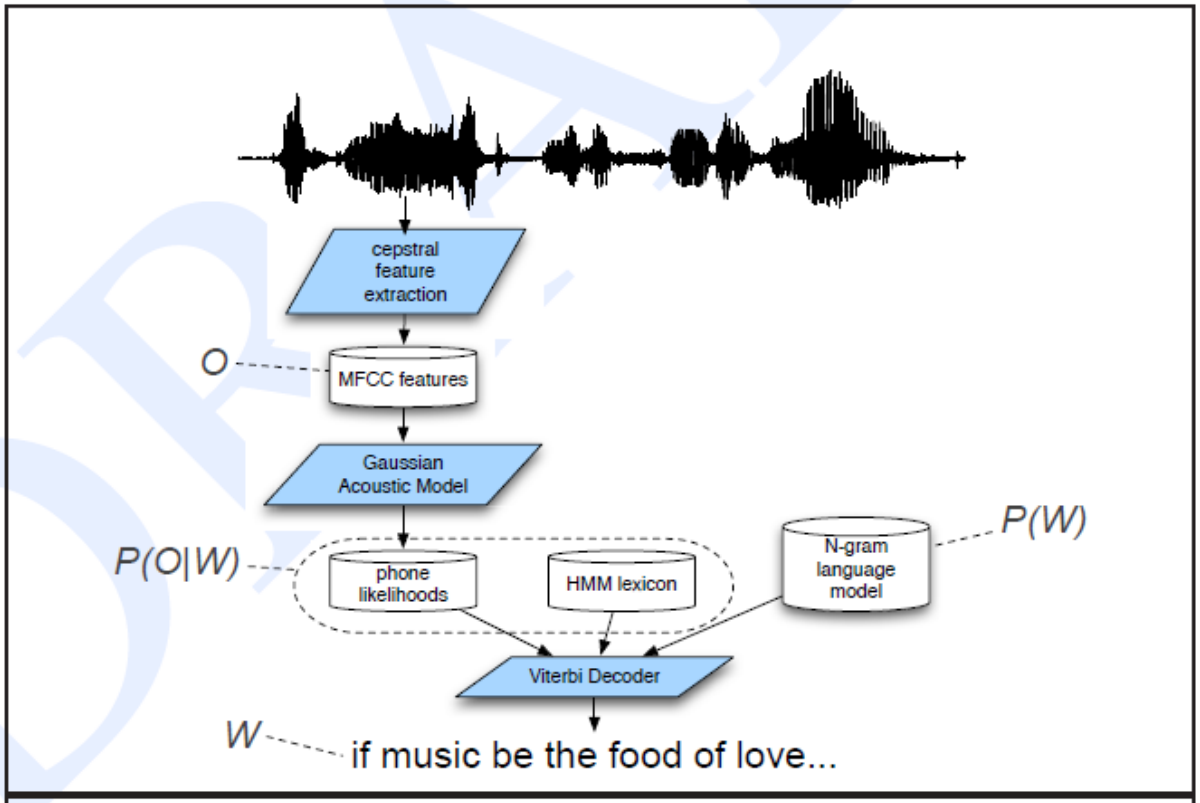
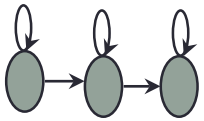


Figure 9.3 Schematic architecture for a (simplified) speech recognizer decoding a single sentence. A real recognizer is more complex since various kinds of pruning and fast matches are needed for efficiency. This architecture is only for decoding; we also need a separate architecture for training parameters.

Sistema de Reconocimiento Automático del Habla



Modelos acústico-fonéticos

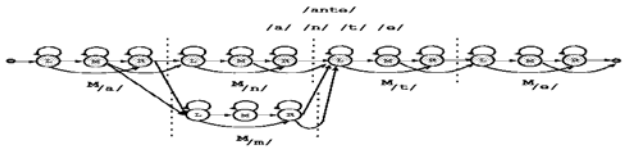
DECODIFICACIÓN

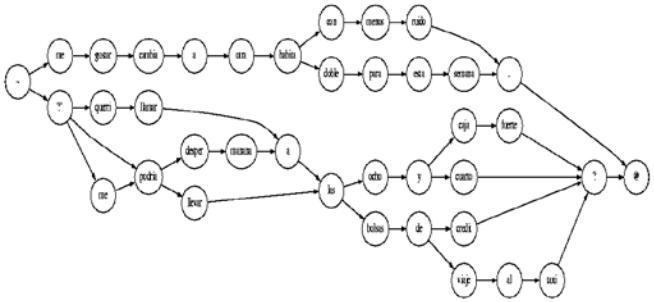
señal vocal

Frase escrita

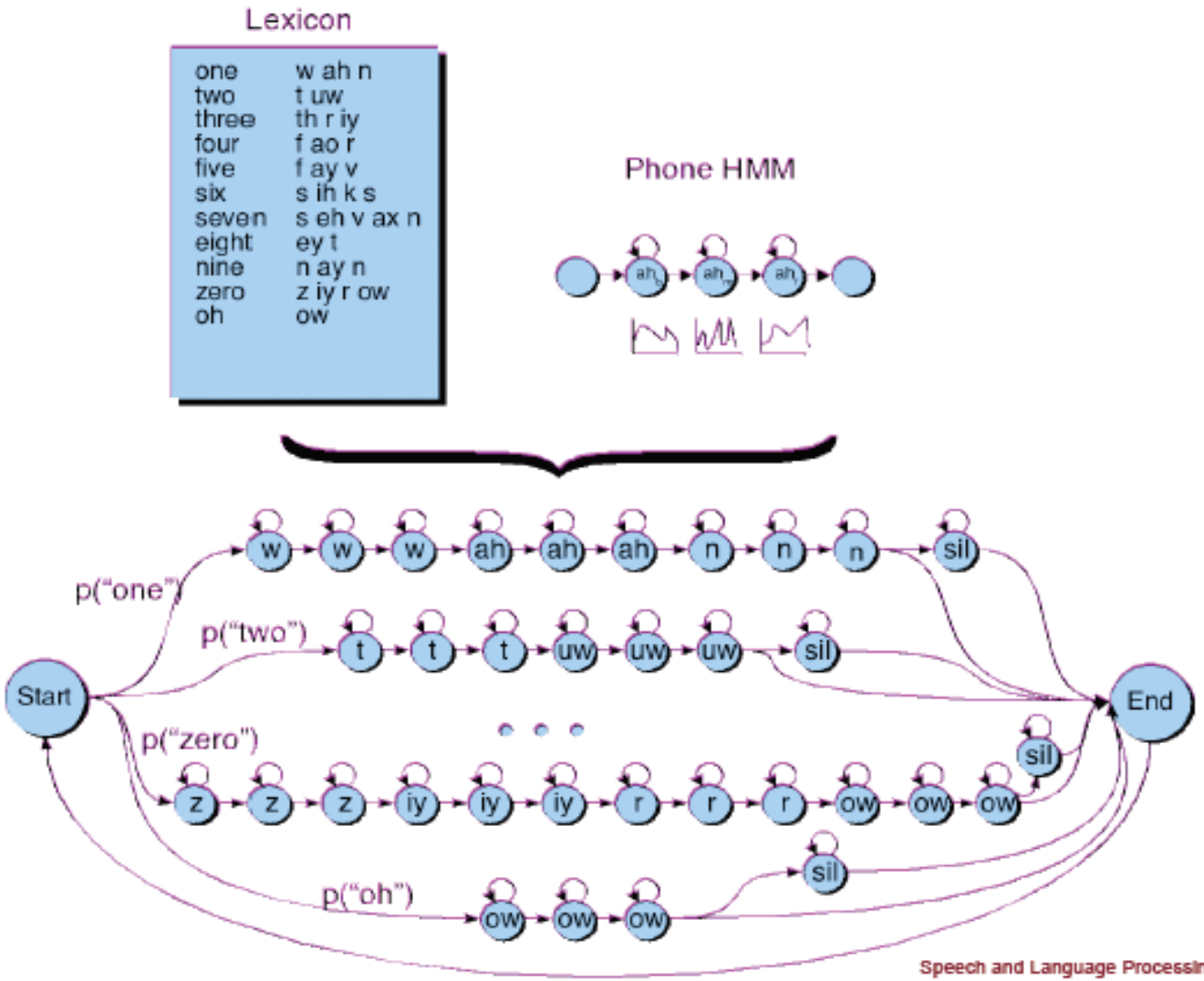
Léxico

Sintaxis





Grafo de búsqueda sin ML



Modelos de lenguaje y evaluación

- **Planteamiento general del reconocimiento automático del habla**
- **Modelos de lenguaje**
- **Clasificación de modelos de lenguaje**
- **Evaluación de modelos de lenguaje**

Modelos de lenguaje

QUÉ ES UN MODELO DE LENGUAJE?

Es el conjunto de mecanismos que se emplean para definir la estructura de un lenguaje de una determinada aplicación. El modelo de lenguaje restringe las secuencias de unidades lingüísticas permitidas en el lenguaje

La utilización de un ML en un sistema de RHA:

- Permite dirigir la acción de búsqueda de la cadena
- Mejora las prestaciones del sistema
- Permite reducir su complejidad computacional

Modelos de lenguaje

Objetivo: Asignar probabilidades a secuencias de palabras o dar una predicción probabilística de la siguiente palabra dada una secuencia.

- **Modelo básico**: cada palabra presenta la misma probabilidad de seguir a cualquier otra. En un vocabulario de 100.000 palabras la probabilidad es 0.00001.
- **Modelo 1**: cada palabra puede seguir a cualquier otra, pero en proporción a la frecuencia relativa de ocurrencia.

  Brown Corpus (1.000.000 palabras):

  frecuencia “the” 0.07,

  frecuencia “rabbit” 0.00001.

  Secuencia “Just then the white” aunque “rabbit” parece más razonable para seguir la secuencia, la más probable según el modelo de frecuencia relativa de ocurrencia sería “the”.

Modelos de lenguaje

- En lugar de usar las frecuencias individuales de cada palabra, se debería calcular la probabilidad condicional de una palabra dadas las anteriores en la secuencia.

- Si se considera que la ocurrencia de cada palabra en su posición en una secuencia es un evento independiente, se puede usar la ***regla de la cadena*** para descomponer la probabilidad a priori como:

$$P(w) = \prod_{i=1 \ldots n} P(w_i | w_1 \cdots w_{i-1})$$

Modelos de lenguaje y evaluación

- **Planteamiento general del reconocimiento automático del habla**
- **Modelos de lenguaje**
- **Clasificación de modelos de lenguaje**
- **Evaluación de modelos de lenguaje**

Clasificación de ML

Según la ***unidad lingüística*** elegida:

- ML ***sintácticos***: describen las concatenaciones de palabras.
- ML ***semánticos***: representan las restricciones semánticas de la tarea. Las unidades básicas son conceptos o categorías relacionados con el significado de la intervención. Se suelen combinar con modelos sintácticos.

Según el formalismo usado para expresar el modelo:

- Modelos basados en ***gramáticas***. Se asume que el lenguaje a modelar es un subconjunto del lenguaje natural y se define un formalismo gramatical que determina las posibles secuencias de palabras en ese subconjunto. Suelen definir la estructura completa de las frases.
- Modelos basados en ***N-gramas***. Modela las concatenaciones de palabras a través de las probabilidades de ocurrencia de secuencias de palabras de longitud fija (N).

Clasificación de ML

Factores a considerar en la elección de un ML:

- Potencia expresiva del modelo
- Algoritmos de análisis eficientes
- Integración con los modelos acústicos
- Posibilidad de estimación automática de los modelos
- El modelo debe proporcionar la probabilidad a priori *P(w)*, por lo cual debe asignar una probabilidad a cada secuencia de palabras.

Clasificación de ML

ML gramaticales versus ML de N-gramas

ML gramaticales:

- Representan las restricciones de manera natural
- La definición de estos modelos es manual (lingüistas) y requieren bastante tiempo y esfuerzo
- Las estructuras gramaticales que rigen las frases en lenguaje escrito suelen resultar demasiado rígidas para su aplicación al lenguaje hablado.

ML de N-gramas:

- Se estiman automáticamente a partir de un conjunto de frases
- Se conserva la flexibilidad del lenguaje hablado
- Sólo captan restricciones a corto plazo (N=1,2,3)

Ejemplos

1) Deficiente modelado de las relaciones a larga distancia entre términos en modelos de bigramas o trigramas.

Lenguaje de los números comprendidos entre cero y el millón en castellano:

dos cientos dos cientos

doce mil dos cientos mil

2) Deficiente modelado de la concordancia entre el sujeto y el verbo cuando hay una cierta distancia entre ellos.

Estos errores pueden ser evitados incorporando restricciones de forma sencilla a través de un modelo gramatical (incluso regular)

Ejemplos

3) Flexibilidad de los modelos de n-gramas. En [Shanon,64] se presentan diferentes lenguajes artificiales de aproximación al inglés basados en n-gramas.

Texto generado con un modelo de bigramas:

"The head and in frontal attack on an english writer that the character of this point is therefore another method for the letters that the time of who ever told the problem for an unexpected".

Se observan construcciones habituales de frases en inglés:

“attack on an english writer”

“therefore another method for the letters”

Modelos de lenguaje y evaluación

- **Planteamiento general del reconocimiento automático del habla**
- **Modelos de lenguaje**
- **Clasificación de modelos de lenguaje**
- **Evaluación de modelos de lenguaje**

Evaluación de ML

Cómo de bueno es nuestro modelo de lenguaje?

- Prefiere las oraciones “buenas” a las “malas”?
  - Asigna una probabilidad mayor a las oraciones “observadas frecuentemente” que a las “ no gramaticales” o “raramente observadas”
- Entrenamos los parámetros del modelo usando un conjunto de entrenamiento (**training set**).
- Probamos las prestaciones del modelo usando datos que no han sido observados en entrenamiento.
  - Un conjunto de prueba (**test set**) es un conjunto de datos no observados en entrenamiento (diferentes de los que contiene el training set).
  - Una métrica de evaluación nos dirá cómo de bueno es nuestro modelo cuando lo probamos con el test set.

Evaluación de ML

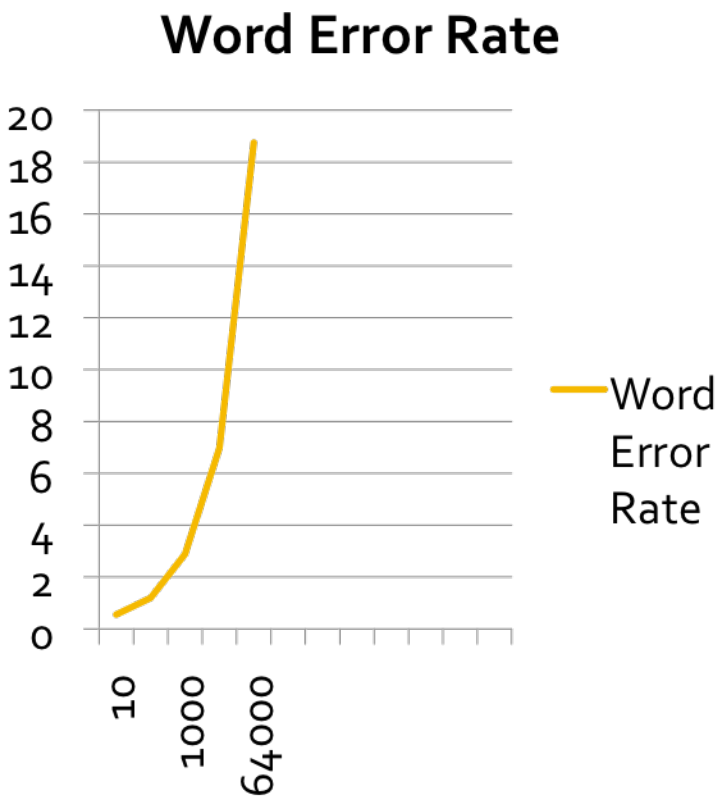
La complejidad de un sistema de RAH depende de diversos factores:

- La dificultad intrínseca del lenguaje a reconocer: fonética, morfología, sintaxis y semántica.
- La similitud acústica entre las palabras del vocabulario
- Las restricciones impuestas en la forma de hablar
- El número de locutores que van a usar el sistema
- El tipo de entorno del sistema y el nivel de ruido
- ...

Tamaño del vocabulario

Vocabulary	Sphinx4 WER
Digits 0-9	.549%
100 Word	1.192%
1,000 Word	2.88%
5,000 Word	6.97%
64,000 Word	18.756%

**If you have noisy audio input multiply expected error rate x 2*



Evaluación de ML

Entropía de un lenguaje

- El concepto de ***Entropía*** o ***Perplejidad***, extraído de la Teoría de la Información, se puede utilizar como índice de la capacidad que tiene un modelo de lenguaje para, una vez determinada una secuencia inicial de palabras, predecir la continuación de esta secuencia.

- Además de su interés como parámetro de diseño a tener en cuenta para evaluar las prestaciones de un sistema, permite comparar diferentes ML en función de su adecuación a un determinado lenguaje objetivo.

Evaluación de ML

Entropía

Sea X una variable aleatoria que toma valores en un espacio X y que tiene asociada una función de probabilidad p(x). La entropía de esta variable es:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

Esta medida puede interpretarse intuitivamente como el número medio ponderado de elecciones que una variable aleatoria debe hacer.

Ejemplo: Si una variable puede tomar 8 valores con una elección equitativa, su entropía es 3 bits (su perplejidad es 8, $2^3=8$). Si, en cambio, su distribución para los diferentes valores es:

V1= 1/2	V2= 1/4	V3= 1/8	V4= 1/16	V5= 1/64	V6= 1/64	V7= 1/64	V8= 1/64

Entonces su entropía es 2 bits (su perplejidad es 4, $2^2=4$). La entropía se puede interpretar como la longitud media del mensaje para transmitir un resultado de la variable.

Evaluación de ML

Fórmulas de cálculo de la Entropía de un lenguaje

1. Se aproxima el lenguaje a través de la salida de una fuente de información discreta de Markov de orden N, considerando bloques de N símbolos.

$$H(L) = -\left(\frac{1}{N}\right)\sum p(w_1 w_2 ... w_N)\log p(w_1 w_2 ... w_N)$$

El sumatorio se extiende a todas las secuencias de N símbolos

Evaluación de ML

Fórmulas de cálculo de la Entropía de un lenguaje

La entropía del inglés (Shannon, 1951)

- Requirió a un grupo de personas que formularan hipótesis sobre la siguiente letra proporcionándoles una secuencia (de diferente longitud) de las letras anteriores. Observando cuántas hipótesis necesitaban formular para acertar estimó la probabilidad de la letra y la entropía de la secuencia. El resultado de la entropía por letra del inglés fue de 1,3.
- Estimó a partir de un gran corpus de textos modelos de N-gramas de símbolos (26 letras más el espacio, descartando mayúsculas y signos de puntuación).

Modelo	Entropía cruzada (bits)
Orden 0	4.76
Orden 1	4.03
Orden 2	2.8

Evaluación de ML

Fórmulas de cálculo de la Entropía de un lenguaje

2. A partir de un modelo generador del lenguaje (en bits por símbolo)

$$H(L) = \sum_{i=1}^{|Q|} p(q_i) H(q_i)$$

$$p(q_i) = \frac{N(q_i)}{\sum_{j=1}^{|Q|} N(q_i)}$$

$$H(q_i) = -\sum_{j=1}^{|W|} p(w_j / q_i) \log p(w_j / q_i)$$

Evaluación de ML

Fórmulas de cálculo de la Entropía de un lenguaje

3. A partir de la distribución de probabilidades asociada a una secuencia suficientemente larga del lenguaje $w_1 w_2 \ldots w_n$, y sabiendo que para medir la verdadera entropía de un lenguaje hay que considerar secuencias de longitud infinita

$$H(L) = \lim_{n\to\infty} \frac{1}{n} H(w_1 w_2 \ldots w_n) = -\lim_{n\to\infty} \frac{1}{n} \sum_{w_1 w_2 \ldots w_n \in L} p(w_1 w_2 \ldots w_n) \log p(w_1 w_2 \ldots w_n)$$

Hay un teorema que establece que si el lenguaje presenta ciertas características, entonces

$$H(L) = -\lim_{n\to\infty} \left(\frac{1}{n}\right) \log p(w_1 w_2 \ldots w_n)$$

Suponiendo un n suficientemente grande

$$H(L) = -\left(\frac{1}{n}\right) \log p(w_1 w_2 \ldots w_n)$$

Evaluación de ML

Perplejidad de un conjunto de prueba

Un buen *ML*:

- Debería asignar una probabilidad alta (aceptar) las secuencias de palabras que pertenecen al lenguaje objetivo.
- Debería asignar una probabilidad muy baja (rechazar) aquellas secuencias que no sean del lenguaje.

Un modelo de lenguaje *ML* será adecuado para modelar el lenguaje *L* si:

- Representa una adecuada cobertura de *L*
- No presenta una excesiva sobregeneralización de *L*

Evaluación de ML

Perplejidad de un conjunto de prueba

Una medida probabilística adecuada de las diferencias entre un modelo de lenguaje *ML* y el lenguaje objetivo *L* es la Entropía Cruzada por símbolo entre el modelo *ML* y el lenguaje *L*, es decir *H(L,ML)*.

Una aproximación a este valor es la entropía del conjunto de prueba *T*:

$$H(L,ML) \equiv H(T,ML) = -\frac{1}{\sum_{\forall t \in T} |t|} \sum_{\forall t \in T} \log p(t|ML)$$

Donde,

p(t|ML) es la probabilidad que el modelo *ML* asigna a la secuencia *t*

$|t|$ es su longitud.

Evaluación de ML

Perplejidad de un conjunto de prueba

Esta medida es válida siempre que *ML* sea un modelo consistente.

$$\sum_{\forall x \in W^*} p(x/ML) = 1$$

La ***Perplejidad del conjunto de prueba***, *Q(T,ML)*, se define como

$$Q(T, ML) = 2^{H(T, ML)}$$

El mejor modelo será aquel que proporcione una menor perplejidad sobre el conjunto de prueba.

Evaluación de ML

Perplejidad de un conjunto de prueba

La Perplejidad del conjunto de prueba:

- Proporciona la dificultad intrínseca de una tarea.
- Proporciona una medida para la comparación de diferentes modelos para la misma tarea.
- Mantiene cierta relación con las prestaciones del sistema completo.

Evaluación de ML

Perplejidad de un conjunto de prueba

Uso de *Q* como el parámetro de comparación de la dificultad relativa de dos tareas con el mismo sistema SPICOS.

|V|	ML	Q(T,ML)	%ERROR
917	NO	917	21.8
9686	NO	9686	43.1
9686	BIPOS	1003	23.5

Evaluación de ML

Perplejidad de un conjunto de prueba

- $|W|$ no es una medida adecuada para medir la dificultad de una tarea.
- El sistema se comporta de forma parecida para tareas de perplejidad similar.
- Un incremento en la perplejidad comporta un incremento en la tasa de error del sistema.