Evolución del PLN

Ferran Pla
VRAIN - Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València
fpla@dsic.upv.es

experto en procesamiento del lenguaje natural

ÍNDICE

Problemas NLP abordados mediane técnicas de Machine Learning

1. **POS TAGGING**. Seguiremos con otros paradigmas de Machine Learning (a parte de HMM, ya visto) para su resolución (CRF,MEMM, …)
2. Introducción al **Shallow Parsing** y su resolución como etiquetado de secuencias. Medidas de evaluación.
3. Detección de Entidades Nombradas (**NER**)
4. **WSD**: definición, recursos, y métodos de resolución.

5. Ejemplos de resolución de estos problemas usando **NLTK** y **Python**.

experto en procesamiento del lenguaje natural

Log-Linear models (CRF, MEMM, …)

- Clasificador exponencial o log-linear

$$p(c|x) = \frac{1}{Z}\exp(\sum_i w_i f_i)$$

- Dada una entrada (x), el objetivo es asignarle una clase (c), donde $\mathbf{f_i}$ son un conjunto de características , $\mathbf{w_i}$ un conjunto de pesos y **Z** un factor de normalización
- Asumimos que las características sólo pueden tomar los valores **0** o **1**, y que tanto estas como sus pesos dependen de cada clase.

$$p(c|x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i(c,x)\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i(c',x)\right)}$$

donde ***C*** es el conjunto de clases en las que se puede clasificar **x**

CLASIFICACIÓN:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|x)$$

experto en procesamiento del lenguaje natural

Ejemplo

Secretariat/NNP is/BEZ expected/VBN to/TO race/?? tomorrow/

Problema: clasificar/etiquetar la palabra “race”

Conjunto de características

$$f_1(c,x) = \begin{cases} 1 & \text{if } word_i = \text{“race”} \ \& \ c = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(c,x) = \begin{cases} 1 & \text{if } t_{i-1} = \text{TO} \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(c,x) = \begin{cases} 1 & \text{if suffix}(word_i) = \text{“ing”} \ \& \ c = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$f_4(c,x) = \begin{cases} 1 & \text{if is\_lower\_case}(word_i) \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_5(c,x) = \begin{cases} 1 & \text{if } word_i = \text{”race”} \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_6(c,x) = \begin{cases} 1 & \text{if } t_{i-1} = \text{TO} \ \& \ c = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

Pesos de las características

		f1	f2	f3	f4	f5	f6
VB	f	0	1	0	1	1	0
VB	w		.8		.01	.1	
NN	f	1	0	0	0	0	1
NN	w	.8					-1.3

$$p(c|x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i(c,x)\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i(c',x)\right)}$$

$$P(NN|x) = \frac{e^{.8} e^{-1.3}}{e^{.8} e^{-1.3} + e^{.8} e^{.01} e^{.1}} = .20$$

$$P(VB|x) = \frac{e^{.8} e^{.01} e^{.1}}{e^{.8} e^{-1.3} + e^{.8} e^{.01} e^{.1}} = .80$$

$$\hat{c} = \underset{c \in C}{\text{argmax}} P(c|x)$$

experto en procesamiento del lenguaje natural

Cómo se aprenden estos modelos (CRF,MEMM, ..)?

MODELOS DISCRIMINATIVOS: maximización de la probabilidad condicionada

- Definición de las características relevantes para el problema.
- Estimación de parámetros: consiste maximizar la función de probabilidad sobre el conjunto de entrenamiento.
- Existen varios métodos:
  - Iterative Scaling
  - Gradient Descent
  - Newton's method
  - Quasi-Newton methods
  -

$$\text{MEMM:} \quad p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} \frac{1}{Z(\mathbf{X}, t, y_{t-1})} \exp\left(\sum_a \lambda_a f_a(y_t, y_{t-1}, \mathbf{X}, t)\right)$$

experto en procesamiento del lenguaje natural

Reglas de Transformación (Brill, 1992)

- Método supervisado (necesita corpus anotado)
- Aprende reglas contextuales correctoras a partir de plantillas predefinidas
- **Aprendizaje:**
  - Paso 1: Asignar a cada palabra su etiqueta POS más probable.
  - Paso 2: Se compara el corpus etiquetado con el corpus de referencia y se instancian las reglas que mejoran el etiquetado inicial.
  - Paso 3: Reetiquetar el corpus aplicando la mejor transformación.
  - Paso 4: Repetir pasos 1, 2 y 3 hasta que no se observen mejoras.
  - Resultado: lista ordenada de transformaciones
- **Etiquetado:**
  - Paso 1: Asignar a cada palabra su etiqueta POS más probable.
  - Paso 2: Reetiquetar el corpus aplicando la lista ordenada de transformaciones obtenida en el aprendizaje.

experto en procesamiento del lenguaje natural

Reglas de Transformación (Brill, 1992)

- Ejemplo:
  - They are expected to **race** tomorrow.
  - The **race** for outer space.
- Plantilla:
  - T1 por T2, si la palabra precedente está etiquetada con T3
- Regla de transformación:
  - NN por VB, si la palabra precedente está etiquetada con TO
- Algoritmo de etiquetado:
  1. Asignar la etiqueta más probable. En el corpus Brown la etiqueta más probable para "race" es NN.
     - They are expected to **race/NN** tomorrow
     - The **race/NN** for outer space
  2. Aplicar la regla de transformación instanciada:
     - They are expected to **race/VB** tomorrow
     - The **race/NN** for outer space

experto en procesamiento del lenguaje natural

Reglas de Transformación (Brill, 1992)

- Ejemplos de plantillas:

The preceding (following) word is tagged **z**.
The word two before (after) is tagged **z**.
One of the two preceding (following) words is tagged **z**.
One of the three preceding (following) words is tagged **z**.
The preceding word is tagged **z** and the following word is tagged **w**.
The preceding (following) word is tagged **z** and the word two before (after) is tagged **w**.

experto en procesamiento del lenguaje natural

Reglas de Transformación (Brill, 1992)

- Ejemplos de reglas aprendidas:

#	Change tags From	To	Condition	Example
1	NN	VB	Previous tag is TO	to/TO race/NN → VB
2	VBP	VB	One of the previous 3 tags is MD	might/MD vanish/VBP → VB
3	NN	VB	One of the previous 2 tags is MD	might/MD not reply/NN → VB
4	VB	NN	One of the previous 2 tags is DT	
5	VBD	VBN	One of the previous 3 tags is VBZ	

experto en procesamiento del lenguaje natural

Reglas de Transformación (Brill, 1992)

- Reglas contextuales aprendidas (20 primeras):

	Change Tag		
#	From	To	Condition
1	NN	VB	Previous tag is *TO*
2	VBP	VB	One of the previous three tags is *MD*
3	NN	VB	One of the previous two tags is *MD*
4	VB	NN	One of the previous two tags is *DT*
5	VBD	VBN	One of the previous three tags is *VBZ*
6	VBN	VBD	Previous tag is *PRP*
7	VBN	VBD	Previous tag is *NNP*
8	VBD	VBN	Previous tag is *VBD*
9	VBP	VB	Previous tag is *TO*
10	POS	VBZ	Previous tag is *PRP*
11	VB	VBP	Previous tag is *NNS*
12	VBD	VBN	One of previous three tags is *VBP*
13	IN	WDT	One of next two tags is *VB*
14	VBD	VBN	One of previous two tags is *VB*
15	VB	VBP	Previous tag is *PRP*
16	IN	WDT	Next tag is *VBZ*
17	IN	DT	Next tag is *NN*
18	JJ	NNP	Next tag is *NNP*
19	IN	WDT	Next tag is *VBD*
20	JJR	RBR	Next tag is *JJ*

experto en procesamiento del lenguaje natural

Reglas de Transformación (Brill, 1992)

- Reglas léxicas aprendidas (20 primeras).
- Tratamiento de palabras desconocidas.

	Change Tag		
#	From	To	Condition
1	NN	NNS	Has suffix -s
2	NN	CD	Has character .
3	NN	JJ	Has character -
4	NN	VBN	Has suffix **-ed**
5	NN	VBG	Has suffix **-ing**
6	??	RB	Has suffix **-ly**
7	??	JJ	Adding suffix **-ly** results in a word.
8	NN	CD	The word **$** can appear to the left.
9	NN	JJ	Has suffix **-al**
10	NN	VB	The word **would** can appear to the left.
11	NN	CD	Has character 0
12	NN	JJ	The word **be** can appear to the left.
13	NNS	JJ	Has suffix **-us**
14	NNS	VBZ	The word **it** can appear to the left.
15	NN	JJ	Has suffix **-ble**
16	NN	JJ	Has suffix **-ic**
17	NN	CD	Has character **1**
18	NNS	NN	Has suffix **-ss**
19	??	JJ	Deleting the prefix **un-** results in a word
20	NN	JJ	Has suffix **-ive**

experto en procesamiento del lenguaje natural

Memory-Based Learning (MBL) (Daelemans 1996)

Se almacena en memoria el conjunto de entrenamiento estructurado como vectores de características y la categoría asociada al vector.

IB1-IG: "Overlap Metric" + "Gain Ratio"

$$\mathrm{D}(X,Y) = \sum_{i=1}^{n} \omega_i \sigma(x_i, y_i)$$

Para asignar pesos a las características

$$\sigma(x_i, y_i) = \begin{cases} 0 & \text{si } x_i = y_i \\ 1 & \text{otro caso} \end{cases}$$

Etiquetado: problema de clasificación, una muestra se clasifica en términos de su similitud con los ejemplos almacenados

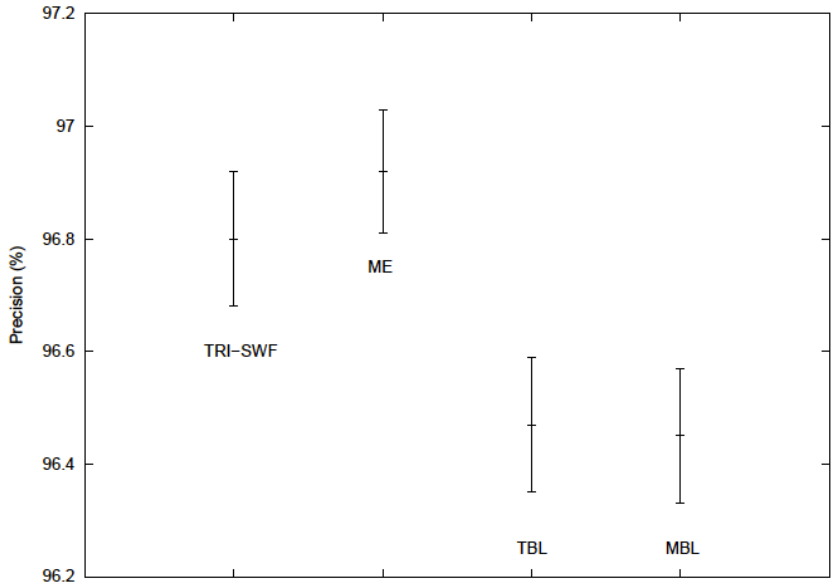
experto en procesamiento del lenguaje natural

Evaluación (Pla and Molina, 2004)

COMPARACIÓN CON OTRAS APROXIMACIONES

Ordenador: Pentium 266Mhz, 256MB RAM

Tagger	Precision	Training	Testing
TRI-SFW (Brants,00)+ Lex	96.80 %	20 sec.	18,000 w/s
ME (Ratnaparkhi,96)	96.92 %	1 day	70 w/s
TBL(Brill,95)	96.47 %	9 days	750 w/s
MBL(Zavrel & Daelemans,99)	96.45 %	4.5 min.	11,200 w/s



10-fold cross-validation: TRI-SFW(96.58%) ME(96.63%)

experto en procesamiento del lenguaje natural

Introducción al Análisis Sintáctico

- El **Análisis Sintáctico** completo consiste en recuperar la estructura sintáctica o árbol sintáctico asociado a una oración.

Aproximación "clásica"

  - Una **gramática** que describe o representa la estructura sintáctica del lenguaje.
  - Un **algoritmo** que determina cuál es el árbol sintáctico de la oración mediante una estrategia de búsqueda (dirección del análisis, orden de aplicación de las reglas, etc.)

experto en procesamiento del lenguaje natural

Ejemplo de Análisis Sintáctico

G = {N, T, S, P}
N = {S, NP, VP}
T = {nprop, n, v, det, adj}
(categorías morfológicas)

(1) S → NP VP
(2) NP → nprop
(3) NP → det n
(4) NP → det n adj
(5) VP → v
(6) VP → VP NP

Gramática CFG

La	det
niña	n
baja	v
la	det
persiana	n

POS TAGGING

(3) NP v det n
(5) NP VP det n
(3) NP VP NP
(6) NP VP
(1) S

Análisis ascendente

Más información en: [Allen, 1995],[Moreno et al., 1999], ...

experto en procesamiento del lenguaje natural

Alternativas: Aprendizaje Automático

Aprender las gramáticas a partir de una colección de ejemplos analizados

- Necesidad de Treebanks. Ej: *Penn Treebank* para el inglés. [Marcus et al., 1993]

Inferencia Gramatical

- [González and Thomason, 1978], [Fu and Booth, 1975]

Aprendizaje estadístico de analizadores completos

- [Charniak, 1997]
- [Collins, 1996]
- [Ratnaparkhi, 1997]

... un tema de investigación abierto

experto en procesamiento del lenguaje natural

Análisis Superficial

Partial Parsing
Chunking
Shallow Parsing
[Abney, 1991]

Obtener información sintáctica de la oración en **unidades sintácticas de interés** de forma **eficiente** y fiable **sacrificando la completitud** y profundidad del análisis global.

CARACTERÍSTICAS

- Análisis robusto para textos no restringidos.
- Algoritmos computacionalmente más eficientes.
- Combinación de diferentes técnicas.
- Posibilidad de conseguir distintos grados de profundidad del análisis.

experto en procesamiento del lenguaje natural

Chunking: máquinas de estados finitos en cascada

[Abney, 1991, Abney, 1996, Abney, 1997]

- Secuencia de **niveles**. Cada nivel reconoce un conjunto de estructuras sintácticas no solapados (chunks) que se describen mediante **expresiones regulares o patrones**.

  Ejemplo: [Molina et al., 1999]

  Nivel 1 // núcleos nominales y verbales

  NSN → (NC | NP)+

  NSV → (VM | VA VMP)

  Nivel 2 // sintagmas nominales

  SN → TD? AQ* NSN AQ*

  Nivel 3 // sintagmas preposicionales

  SPR → SP SN

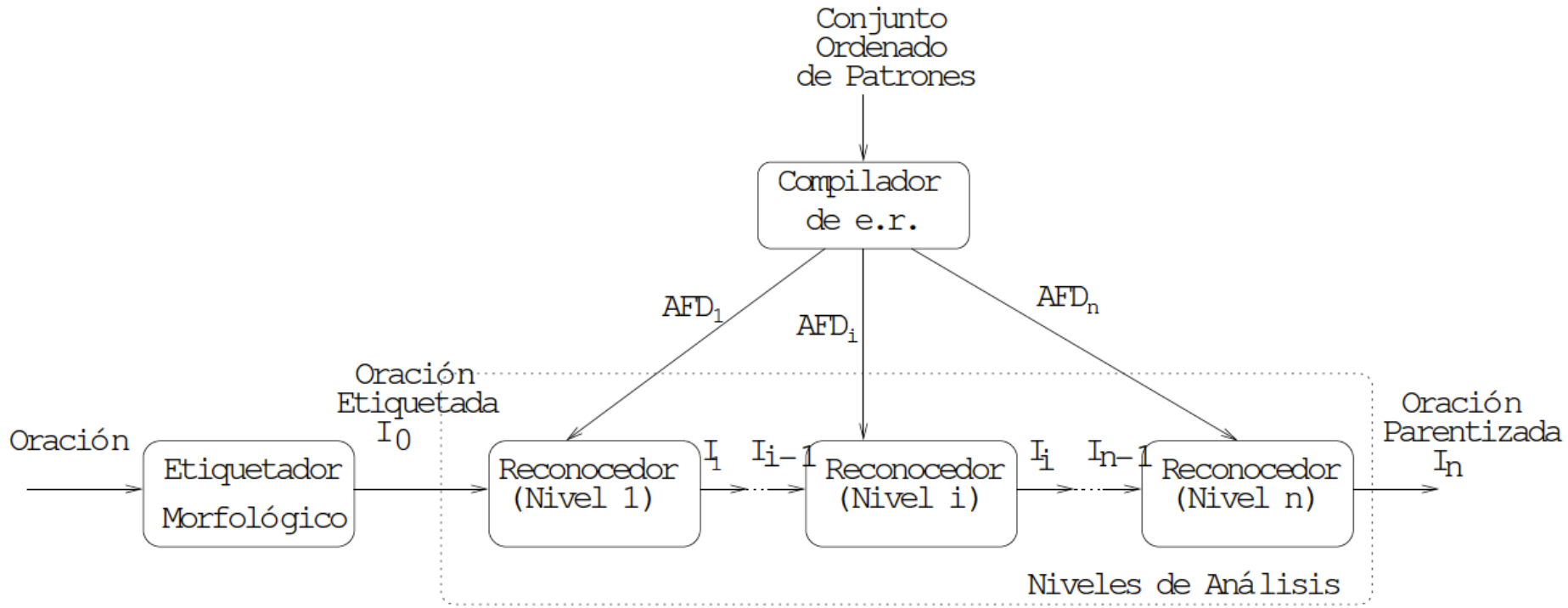
  Nivel 4 // sintagmas verbales

  SV → NSV (SN | SPR)*

- Los patrones de un nivel se definen en función de las etiquetas morfosintácticas o los sintagmas definidos en niveles anteriores.
- Las expresiones regulares de un nivel se traducen a un **autómata de estados finitos determinista.**
- En caso de ambigüedad se toma el sintagma más largo (Longest matching)
- Gran eficiencia computacional debido a la arquitectura.

experto en procesamiento del lenguaje natural

Esquema de un analizador parcial de estados finitos



experto en procesamiento del lenguaje natural

Aproximaciones Inductivas: Inicios

NP-Chunking

- Se siguen aproximaciones similares a las de "POS tagging"
  [Church, 1988]
- [Ramshaw and Marcus, 1995]
  - El chunking se puede plantear como un problema de etiquetado, definiendo un conjunto de etiquetas para marcar los chunks.
  - Además, proponen un conjunto de aprendizaje y de prueba estándar, extraído del Penn Treebank Corpus, para contrastar diferentes aproximaciones inductivas.
    (Aprendizaje: secciones 15–18, Prueba:seccio´n 20).

- PROBLEMAS
  - ¿Cómo obtener/definir los chunks?
  - ¿Cómo definir las etiquetas?
  - ¿Se pueden definir distintos conjuntos de etiquetas equivalentes?
  - ¿Influye el conjunto de etiquetas elegido en las prestaciones?

experto en procesamiento del lenguaje natural

Ejemplo de conversión de árboles a chunks

PENN TREEBANK

```
( (S
    (NP-SBJ-1 (PRP You) )
    (VP (MD will)
     (VP (VB start)
      (S
       (NP-SBJ (-NONE- *-1) )
       (VP (TO to)
        (VP (VB see)
         (NP
          (NP (NNS shows) )
          (SBAR
           (WHADVP-2 (WRB where) )
           (S
            (NP-SBJ (NNS viewers) )
            (VP (VBP program)
             (NP (DT the) (NN program) )
             (ADVP-LOC (-NONE- *T*-2) ))))))))))
    (. .) ))
```

CHUNKS

```
[ NP You ]
[ VP will
start
to
see ]
[ NP shows ]
[ ADVP where ]
[ NP viewers ]
[ VP program ]
[ NP the
program ]
.
```

experto en procesamiento del lenguaje natural

Representación de los chunks

Objetivo: buscar un conjunto de etiquetas equivalente a la segmentación en chunks. [Ramshaw and Marcus, 1995] [Tjong Kim Sang and Veenstra, 1999]

CHUNKS		PALABRAS	IOB	IOE
[NP You]		You	B-NP	E-NP
[VP will		will	B-VP	I-VP
start		start	I-VP	I-VP
to		to	I-VP	I-VP
see]		see	I-VP	E-VP
[NP shows]	⟹	shows	B-NP	E-VP
[ADVP where]		where	B-ADVP	E-ADVP
[NP viewers]		viewers	B-NP	E-NP
[VP program]		program	B-VP	E-VP
[NP the		the	B-NP	I-NP
program]		program	I-NP	E-NP
.		.	O	O

experto en procesamiento del lenguaje natural

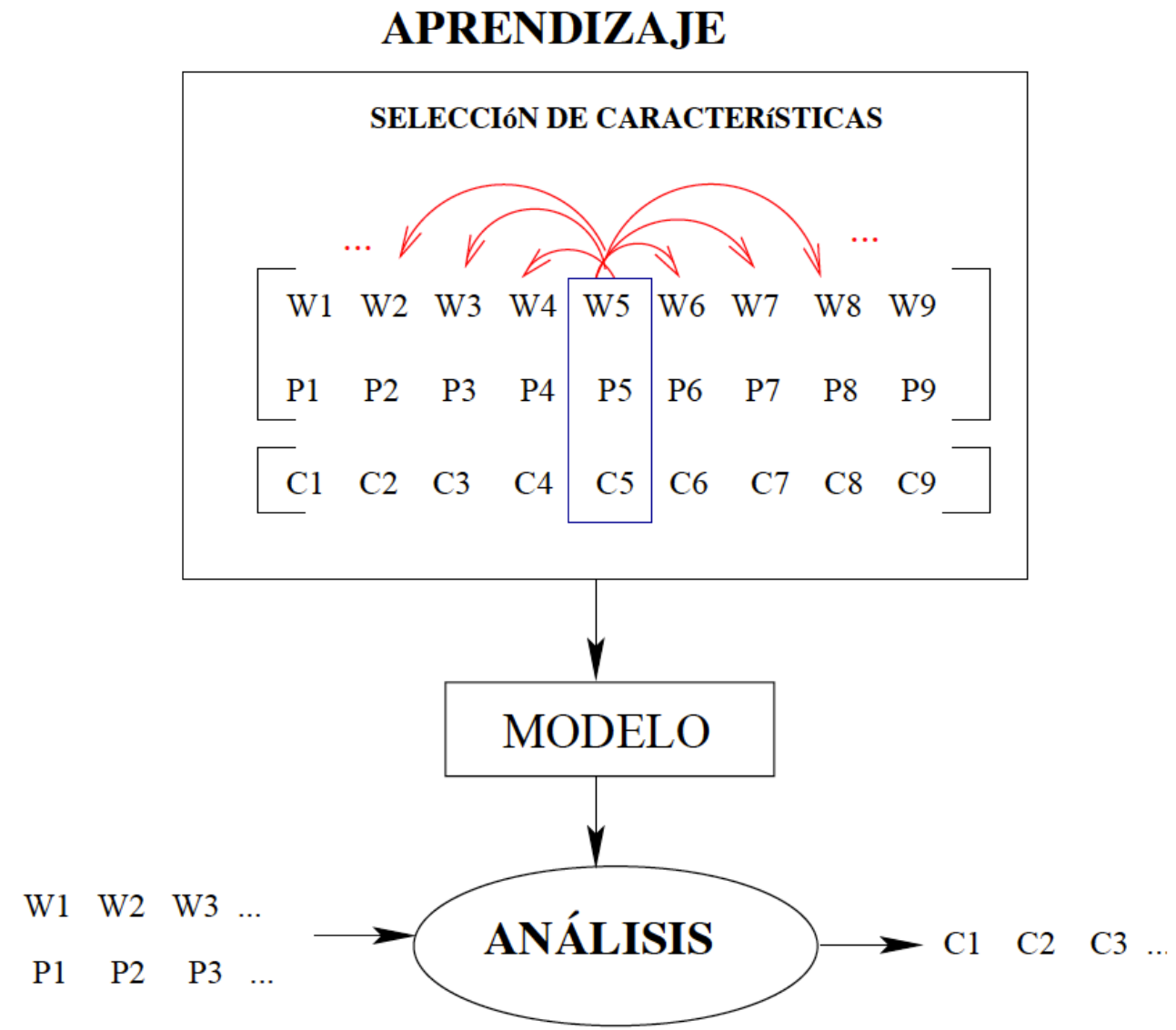
Otro ejemplo

[$_{SN}$ El cartero] [$_{SV}$ da] [$_{SP}$ al] [$_{SN}$ hombre] [$_{SN}$ una carta] .

	IOB1	**IOB2**	IOE1	IOE2
El	I-SN	**B-SN**	I-SN	I-SN
cartero	I-SN	**I-SN**	I-SN	E-SN
da	I-SV	**B-SV**	I-SV	E-SV
al	I-SP	**B-SP**	I-SP	E-SP
hombre	I-SN	**B-SN**	E-SN	E-SN
una	B-SN	**B-SN**	I-SN	I-SN
carta	I-SN	**I-SN**	I-SN	E-SN
.	O	**O**	O	O

experto en procesamiento del lenguaje natural

Aproximaciones basadas en corpus



experto en procesamiento del lenguaje natural

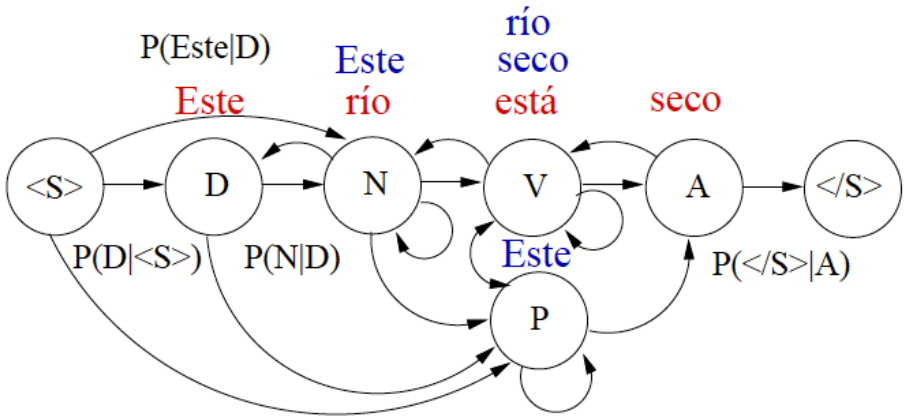
Análisis Superficial: representación en HMM

POS TAGGING

ENTRADA

Este río está seco

D V V V
P N A
N



SALIDA

Este río está seco

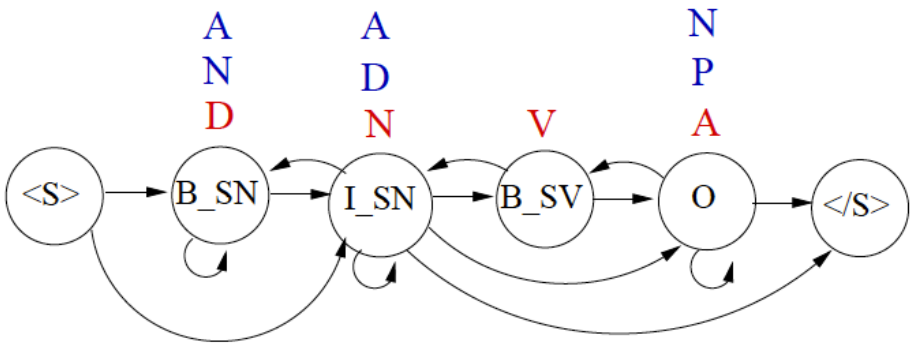
D N V A

CHUNKING

ENTRADA

Este río está seco

D N V A



SALIDA

Este río está seco

B_SN I_SN B_SV O

experto en procesamiento del lenguaje natural

Análisis superficial: etiquetado de secuencias

Técnicas similares para problemas distintos.

Formalismo	POS Tagging	Chunking
HMM	[Church, 1988] [Cutting et al., 1992], [Merialdo, 1994], [Brants, 2000]	[Church, 1988] [Brants, 1999] [Molina and Pla, 2002]
ME	[Ratnaparkhi, 1996]	[Skut and Brants, 1998] [Osborne, 2000],[Koeling, 2000]
TBL	[Brill, 1992, Brill, 1995]	[Ramshaw and Marcus, 1995]
MBL	[Daelemans et al., 1996]	[Veenstra, 1998], [Argamon et al., 1998], [Daelemans et al., 1999]
Comb.	[Van Halteren et al., 1998], [Brill and Wu, 1998]	[Tjong Kim Sang, 2000a]
Winnow		[Li and Roth, 2001] [Zhang et al., 2001]
SVMs		[Kudo and Matsumoto, 2000]
	. . .	. . .

experto en procesamiento del lenguaje natural

Medidas de Evaluación

$$Accuracy(A) = \frac{\#\, etiquetas\ correctas\ en\ el\ análisis\ propuesto}{\#\, etiquetas\ en\ el\ análisis\ de\ referencia}$$

Medidas usuales: Precisión y Cobertura

$$Precisión(P) = \frac{\#\, constituyentes\ correctos\ en\ el\ análisis\ propuesto}{\#\, constituyentes\ en\ el\ análisis\ propuesto}$$

$$Cobertura(C) = \frac{\#\, constituyentes\ correctos\ en\ el\ análisis\ propuesto}{\#\, constituyentes\ en\ el\ análisis\ de\ referencia}$$

$$F_\beta = \frac{(\beta^2 + 1) \times P \times C}{\beta^2 \times P + C}$$

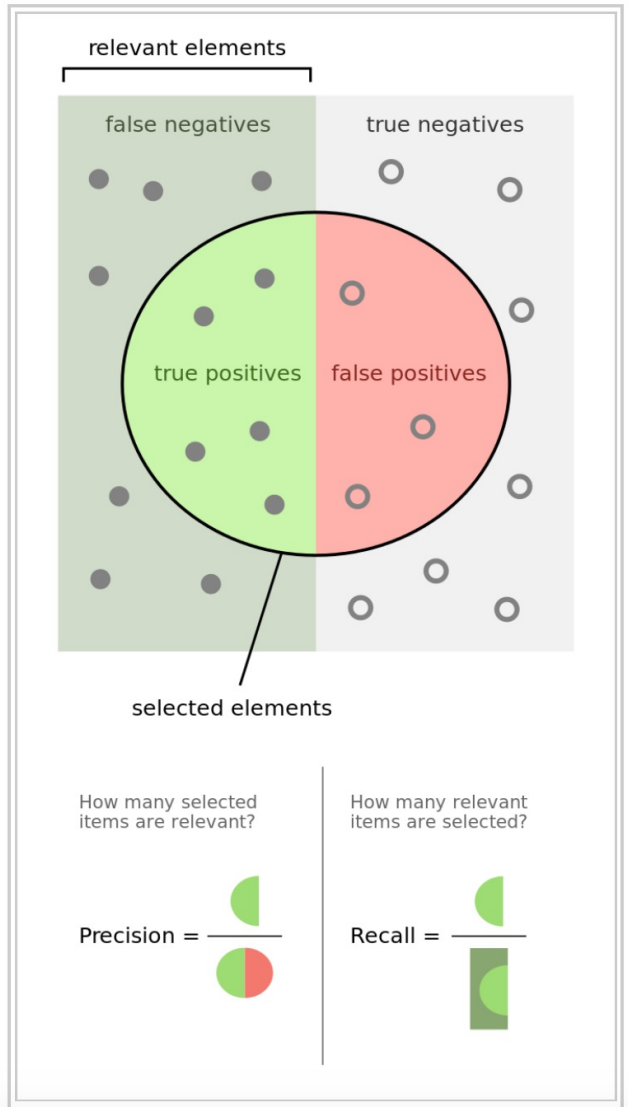
Ejemplo:

	Frase	W1	W2	W3	W4	W5	W6	W7
	Referencia	B-NP	I-NP	B-NP	I-NP	B-VP	B-NP	I-NP
	Salida	B-NP	I-NP	I-NP	I-NP	B-VP	B-NP	I-NP

Evaluación: A=6/7 P=2/3 C=2/4

experto en procesamiento del lenguaje natural

Evaluation at class level



$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

https://en.wikipedia.org/wiki/Precision_and_recall

experto en procesamiento del lenguaje natural

Medidas usuales Multiclass/Multilabel

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. ,34(1):1–47, March.

	Microaveraging	**Macroaveraging**
Precision (π)	$\pi = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i}{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i + FP_i}$	$\pi = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \pi_i}{\lvert\mathcal{C}\rvert} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \frac{TP_i}{TP_i + FP_i}}{\lvert\mathcal{C}\rvert}$
Recall (ρ)	$\rho = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i}{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i + FN_i}$	$\rho = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \rho_i}{\lvert\mathcal{C}\rvert} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \frac{TP_i}{TP_i + FN_i}}{\lvert\mathcal{C}\rvert}$

- **Microaveraging**: “las categorías cuentan en el resultado final proporcionalmente al número de instancias en el corpus”
- **Macroaveraging**: “todas las categorías cuentan lo mismo en el resultado final independientemente del número de instancias en el corpus”
  - **TPi**: número de aciertos para la categoría i
  - **FPi**: número de fallos para la categoría i
  - **FNi**: número de instancias de la categoría i que no han sido asignados por el sistema.

experto en procesamiento del lenguaje natural

Medidas usuales Multiclass/Multilabel

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. ,34(1):1–47, March.

	Microaveraging	**Macroaveraging**
Precision (π)	$\pi = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i}{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i + FP_i}$	$\pi = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \pi_i}{\lvert\mathcal{C}\rvert} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \frac{TP_i}{TP_i + FP_i}}{\lvert\mathcal{C}\rvert}$
Recall (ρ)	$\rho = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i}{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i + FN_i}$	$\rho = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \rho_i}{\lvert\mathcal{C}\rvert} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \frac{TP_i}{TP_i + FN_i}}{\lvert\mathcal{C}\rvert}$

$$\text{micro\_F1} = \frac{2\sum_{i=1}^{|T|} \text{TP}_i}{2\sum_{i=1}^{|T|} \text{TP}_i + \sum_{i=1}^{|T|} \text{FN}_i + \sum_{i=1}^{|T|} \text{FP}_i}$$

$$\text{macro\_F1} = \frac{\sum_{i=1}^{|T|} \frac{2\,\text{TP}_i}{2\,\text{TP}_i + \text{FN}_i + \text{FP}_i}}{|T|}$$

experto en procesamiento del lenguaje natural

Example: for SA in TASS competition

```
Results for LinearSVC()+Lexicon DEV
             precision    recall  f1-score   support

          N       0.56      0.70      0.62       219
        NEU       0.22      0.16      0.18        69
       NONE       0.21      0.11      0.15        62
          P       0.53      0.51      0.52       156

avg / total       0.46      0.49      0.47       506

Accuracy=  0.54347826087
macro=  (0.37036828534232624, 0.38008248883818063, 0.36106862025494307, None)
micro=  (0.54347826086956519, 0.54347826086956519, 0.54347826086956519, None)
```

```
Altres
             precision    recall  f1-score   support

          N       0.47      0.94      0.62       767
        NEU       0.00      0.00      0.00       216
       NONE       0.00      0.00      0.00       274
          P       0.67      0.36      0.47       642

avg / total       0.41      0.50      0.41      1899

Accuracy=  0.501843075303
macro=  (0.28366412723870621, 0.32541266089103882, 0.27318432707201601, None)
micro=  (0.50184307530279093, 0.50184307530279093, 0.50184307530279093, None)
```

experto en procesamiento del lenguaje natural

Detección de Entidades Nombradas (NER)

- Las Entidades son segmentos de la oración que contienen nombres de persona, organizaciones, lugares, fechas, cantidades, etc.

- **Problema**: detección y clasificación.
- **Ejemplo:**

  [PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid]

INFO: http://www.clips.ua.ac.be/conll2002/ner/

experto en procesamiento del lenguaje natural

Entidades biomédicas

DNA, RNA, protein, cell_line, cell_type, ...

PROBLEMAS:

- No existe un diccionario completo para muchos tipos de entidades biológicas nombradas, por lo que un simple algoritmo de 'matching' no es suficiente.
- La misma palabra o frase se puede referir a diferentes entidades dependiendo del contexto (sinónimos), un término varios significados (homónimos),variantes tipográficas o léxicas,
- Rápida evolución de la terminología
- Muchas entidades están compuestas por palabras compuestas ('multiwords')
- Detección de acrónimos

INFO: http://biocreative.sourceforge.net/

experto en procesamiento del lenguaje natural

WSD

- WSD es el problema de determinar computacionalmente cuál es el sentido correcto de una palabra en un contexto.
- Es un problema de clasificación (o etiquetado).
  - Asignar a una palabra un sentido (clase).
  - Cada palabra tiene un núm. finito de sentidos (recogidos en un diccionario)
- Gran complejidad ¿Por qué?
  - P.e. Las 120 palabras inglesas más frecuentes → 7,8 sentidos de media en WordNet
  - Sentido?, Contexto? Recursos? Evaluación? …
- WSD es un problema abierto en PLN
- WSD explícita: módulo genérico de desambiguación.
- WSD implícita: particular de cada aplicación.
- ¡ Pero, no se ha demostrado suficientemente su utilidad en aplicaciones reales !
  - Escasez de recursos adaptados a dominios (diccionarios o corpora) y de desarrollo de técnicas que faciliten esa adaptación.
  - WSD explícita no ofrece tasas de acierto suficientemente altas.

experto en procesamiento del lenguaje natural

Recursos Lingüísticos

- Corpora
  - Semcor (corpus supervisado)
- Diccionarios
- WordNet (English) y EWN
  - WordNet: BD léxica (ontología) N, V, Adj, Adv
  - Las palabras sinónimas se agrupan para formar conjuntos de sinónimos o SYNSETS
  - Cada synset tiene asociado una definición o glosa
  - Los synsets están conectados entre sí a través de relaciones semánticas explícitas que están definidas en WordNet
  - Estas relaciones conectan principalmente sentidos de palabras que pertenecen a la misma categoría gramatical (POS tag), aunque existen algunas relaciones entre distintas catgegorías N-Adj, V-Adj

```
Overview of noun tree

The noun tree has 2 senses (first 1 from tagged texts)

1. {09396070} <noun.plant> tree --
   (a tall perennial woody plant having a main trunk
    and branches forming a distinct elevated crown;
    includes both gymnosperms and angiosperms)

2. {10025462} <noun.shape> tree, tree diagram --
   (a figure that branches from a single root;
    "genealogical tree")
```

experto en procesamiento del lenguaje natural

Recursos Semánticos

- WordNet EWN

http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl

- BabelNet

https://babelnet.org/

- ConceptNet

http://conceptnet.io/

- Dbpedia

https://wiki.dbpedia.org/

- Word Embeddings

https://projector.tensorflow.org/

experto en procesamiento del lenguaje natural

Aproximaciones

- Basadas en conocimiento (tesauros, diccionarios, reglas):
  - Solapamiento del contexto de una palabra con glosas.
  - Medidas de similitud semánticas.
  - Heurísticas.
  - Preferencias seleccionales adquiridas (semi) automáticamente.
- Basadas en corpus, NO supervisadas:
  - Agrupamiento de palabras en diferentes contextos.
  - Corpus paralelos alineados (traducción automática).
- Basadas en corpus, Supervisadas:
  - Aprendizaje automático o estadístico sobre corpus anotados.
  - Técnicas de bootstrapping o semisupervisadas.
- Otras:
  - Métodos combinados
  - Conocimiento del dominio

experto en procesamiento del lenguaje natural

Aproximaciones WSD (Knowledge-based)

- Se usa información de léxicos o bases de conocimiento
  - Machine-readable dictionary (Longman Dictionary of Contemporary English)
  - Tesaurus (Rodget's Thesaurus)
  - Bases de datos de concocimiento léxico (WordNet)
  - …
- Ejemplos:
  - (Lesk 1986) (Yarowsky 1992) (Voorhee 1993) (Resnik 1995) (Agirre & Rigau 1996) (Stevenson & Wilks 2001) …

experto en procesamiento del lenguaje natural

Aproximaciones WSD (Knowledge-based)

Ejemplo (Lesk) :

BANCO:

- Asiento, con respaldo o sin él, en que pueden **sentarse** varias personas.
- Madero grueso escuadrado que se coloca horizontalmente sobre cuatro pies y sirve como de mesa para muchas labores de los carpinteros, cerrajeros, herradores y otros artesanos.
- Conjunto de **peces** que van juntos en gran número.
- Establecimiento público de crédito, constituido en sociedad por acciones.

Contextos:

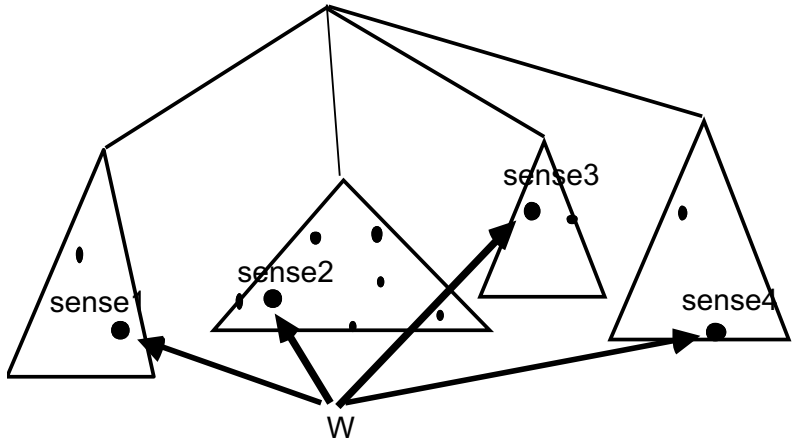
a) El pescador está **sentado** en un banco sobre cubierta.
Solapamiento(1)=1
Solapamiento(2)=Solapamiento(3)=Solapamiento(4)=0

b) El pescador divisó un banco de **peces**.
Solapamiento(3)=1
Solapamiento(1)=Solapamiento(3)=Solapamiento(4)=0

experto en procesamiento del lenguaje natural

Basadas en WordNet

Densidad Conceptual (Agirre & Rigau 96)



1. Se identifican todos los nodos que corresponden a los distintos sentidos de la palabra a desambiguar *W* y de las palabras del contexto.
2. Se mide la densidad conceptual alrededor de cada sentido de *W*
3. Se elige el sentido con la mayor densidad

experto en procesamiento del lenguaje natural

Machine Learning Supervised Approaches

- Definir un vector de características ($\vec{f}$) para predecir el correcto sentido de una palabra
- Usualmente se usan 2 clases de características
  - Collocations: importa la posición de la característica respecto a la palabra a desambiguar
    - Ej: Wi-2, Pi-2, Wi-1, Pi-1, wi , Wi+1,Pi+1, Wi+2,Pi+2
  - Bag-of_Words: no importa la posición
    - Ej: (w1,w2,w3,w4,…,wn-1,wN) →(0, 1, 1,0, …, 1, 0)

experto en procesamiento del lenguaje natural

Clasificador Naïve Bayes

La aproximación de Naïve Bayes para WSD consiste en encontrar el mejor sentido $\hat{s}$ del conjunto de sentidos S para un vector de características $\vec{f}$

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s|\vec{f})$$

Para estimar el vector $\vec{f}$ hay que hacer simplificaciones.
Ejemplo: Si definimos un vector de 20 palabras, tendríamos 2^{20} posibles vectores de características, considerando características binarias

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \frac{P(\vec{f}|s)P(s)}{P(\vec{f})}$$

Simplificación: Las características son independientes unas de otras

$$P(\vec{f}|s) \approx \prod_{j=1}^{n} P(f_j|s)$$

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s) \prod_{j=1}^{n} P(f_j|s)$$

$$P(s_i) = \frac{\text{count}(s_i, w_j)}{\text{count}(w_j)}$$

$$P(f_j|s) = \frac{\text{count}(f_j, s)}{\text{count}(s)}$$

experto en procesamiento del lenguaje natural

WSD -- HMM

Planteamiento general del problema

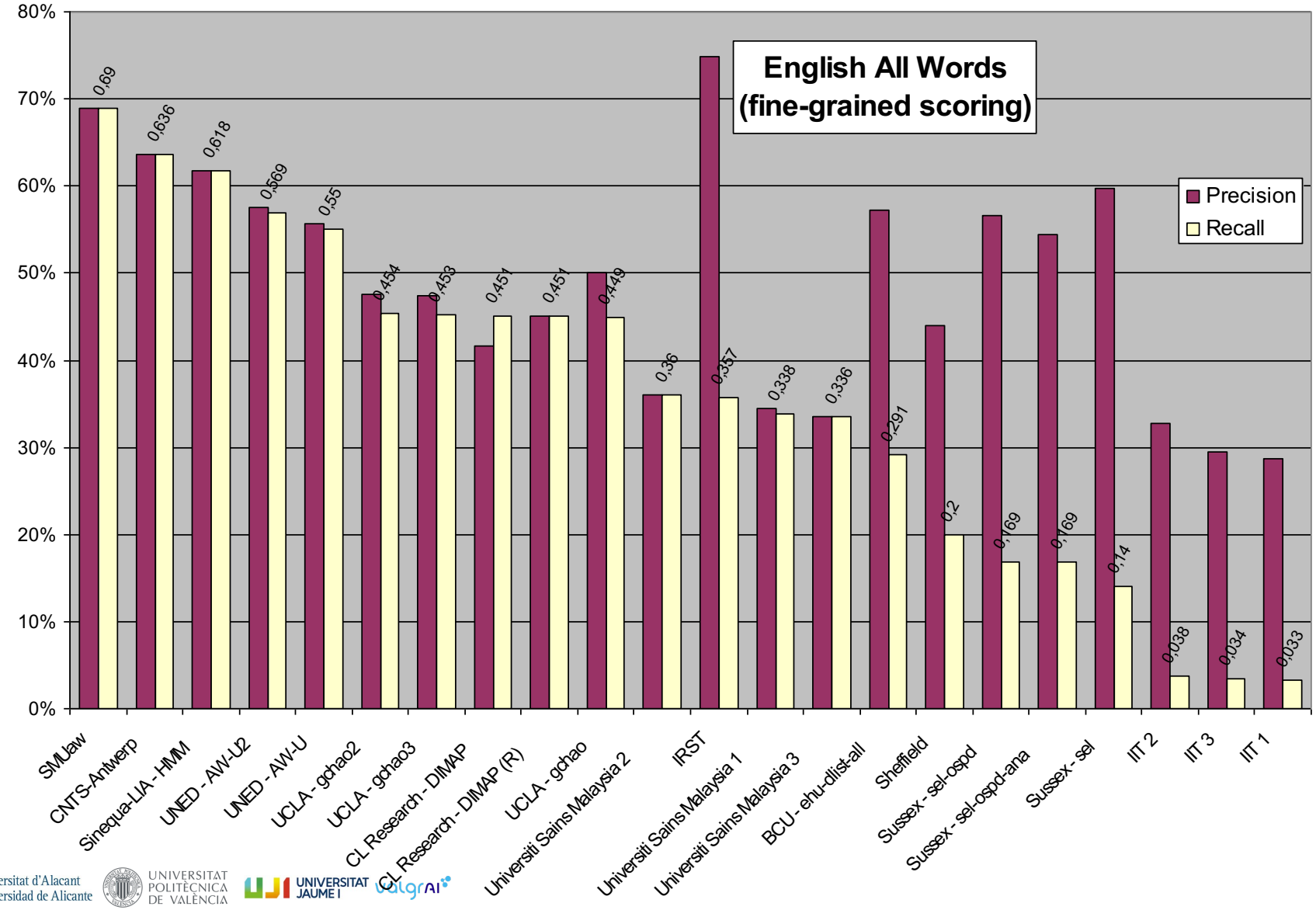
$$\widehat{S} = \arg\max_{S} P(S|W)$$

$$= \arg\max_{S} \left(\frac{P(S) \cdot P(W|S)}{P(W)} \right) ; S \in \mathcal{S}^T$$

$$\arg\max_{S} \left(\prod_{i:1\ldots T} P(s_i|s_{i-1}) \cdot P(w_i|s_i) \right)$$

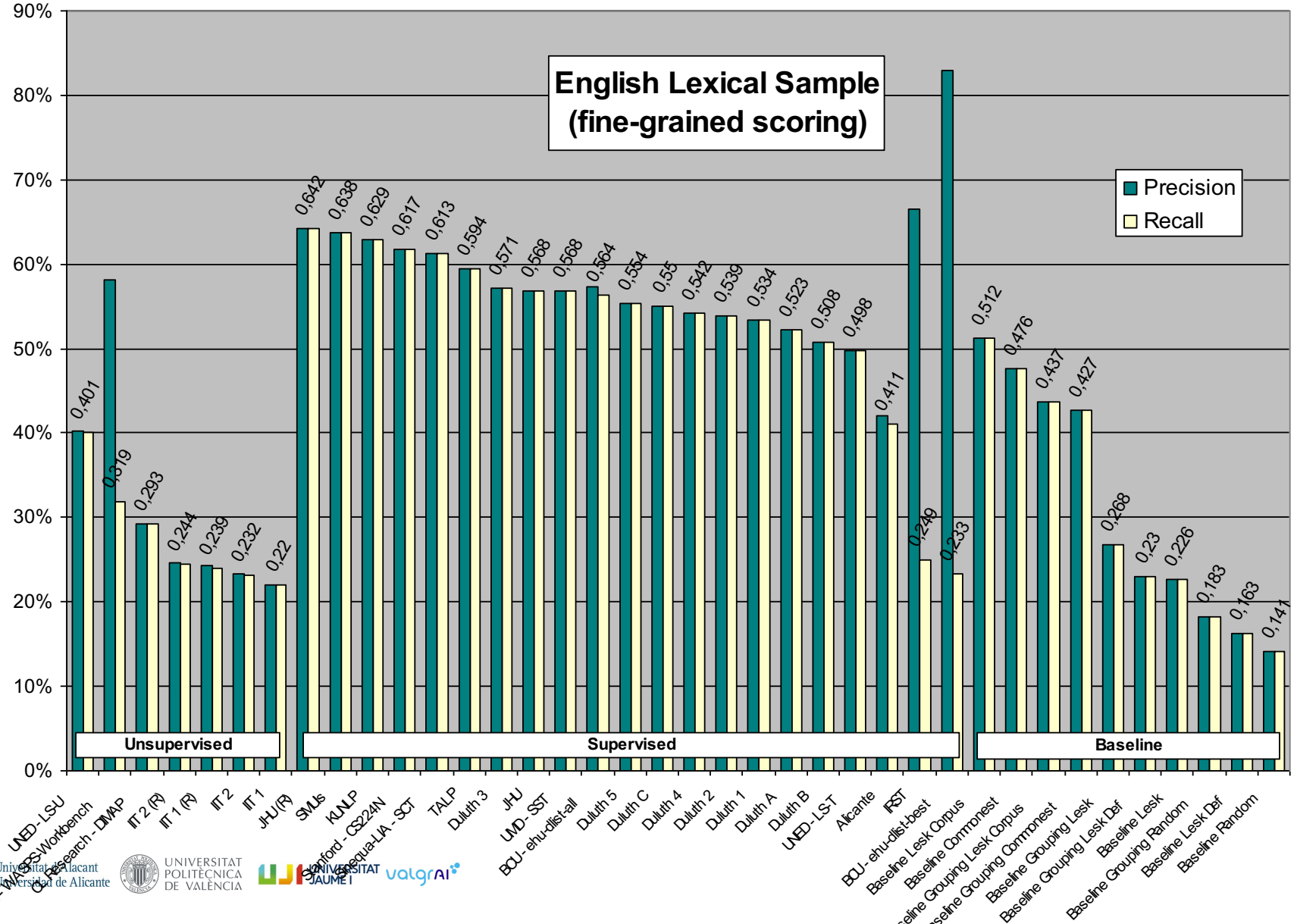
experto en procesamiento del lenguaje natural

Senseval2 (English All Words)



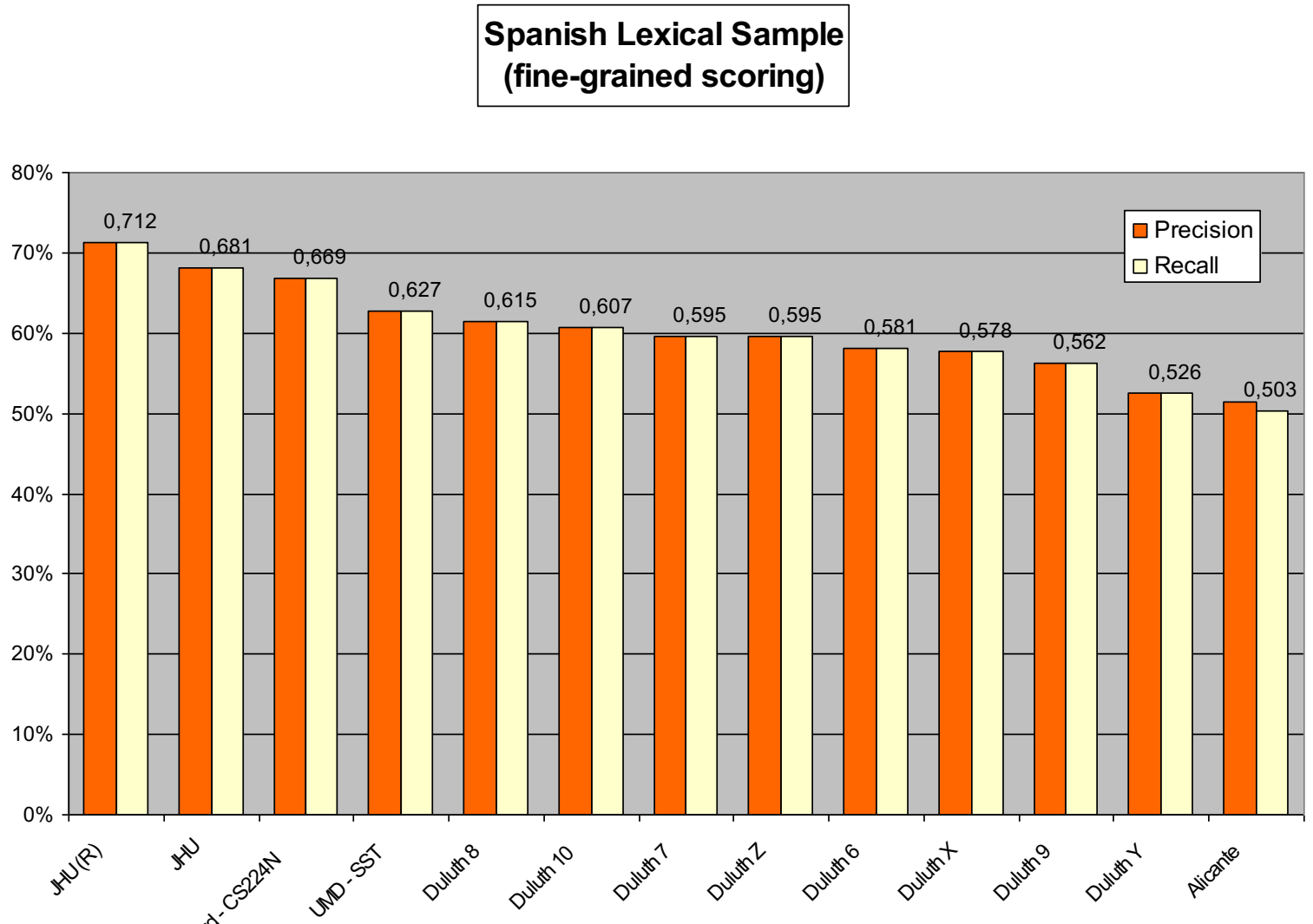
experto en procesamiento del lenguaje natural

Senseval2 (English Lexical Sample)



experto en procesamiento del lenguaje natural

Senseval2 (Spanish)



experto en procesamiento del lenguaje natural

Evaluación sistemas WSD

- Senseval3 (2004) www.senseval.org:

Language	Task[a]	Systems	Lemmas	Instances	ITA[b]	Baseline[c]	Best score
English	AW	26	–	2,081	62%	62%/–[d]	65%/58%
Basque	LS	8	40	7,362	78	59	70
Catalan	LS	7	27	6,721	93	66	85
English	LS	47	57	–	67	55/–	73/66
Italian	LS	6	45	7,584	89	18	53
Romanian	LS	7	39	11,532	–	58	73
Spanish	LS	9	46	12,625	83–90	67	84
Hindi	TM	8	41	11,984	–	56	67
English	GL	10	–	42,491	–	–	68

Copyright © 2004, Association for Computational Linguistics. Reproduced with permission of the Association for Computational Linguistics and Mihalcea and Edmonds.

[a] *AW* all-words, *LS* lexical sample, *TM* translation memory, *GL* gloss task.

[b] ITA is inter-tagger agreement.

[c] The baseline is most-frequent sense.

[d] Scores separated by a slash are supervised/unsupervised methods; supervised when there is no slash.

experto en procesamiento del lenguaje natural

WSD Evaluación

Senseval3 (2004)

System	Precision	Recall
GAMBL-AW-S	.651	.651
SenseLearner-S	.651	.642
Koc University-S	.648	.639
R2D2: English-all-words	.626	.626
Meaning-allwords-S	.625	.623
Meaning-simple-S	.611	.610
LCCaw	.614	.606
upv-shmm-eaw-S	.616	.605
UJAEN-S	.601	.588
IRST-DDD-00-U	.583	.582
University of Sussex-Prob5	.585	.568
University of Sussex-Prob4	.575	.550
University of Sussex-Prob3	.573	.547
DFA-Unsup-AW-U	.557	.546
KUNLP-Eng-All-U	.510	.496
IRST-DDD-LSI-U	.661	.496
upv-unige-CIAOSENSO-eaw-U	.581	.480
merl.system3	.467	.456
upv-unige-CIAOSENSO2-eaw-U	.608	.451
merl.system1	.459	.447
IRST-DDD-09-U	.729	.441
autoPS-U	.490	.433
clr04-aw	.506	.431
autoPSNVs-U	.563	.354
merl.system2	.480	.352
DLSI-UA-all-Nosu	.343	.275

experto en procesamiento del lenguaje natural