PROYECTO FIN CARRERA

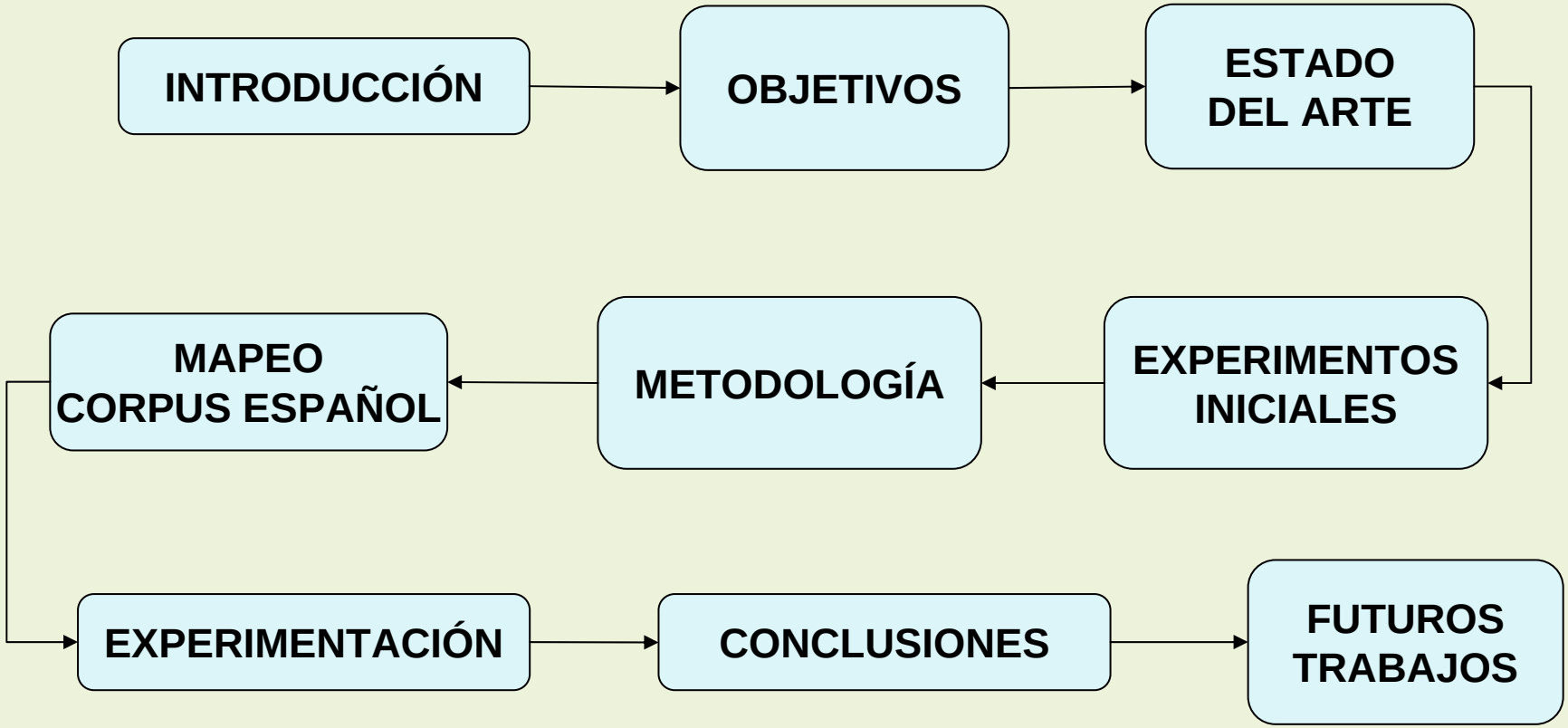
TÉCNICAS DE ETIQUETADO Y DESAMBIGUACIÓN MORFOLÓGICA DEL CASTELLANO CON REDUCIDA INFORMACIÓN CONTEXTUAL

Tutores:
Valentín Moreno Pelayo
Sonia Sánchez-Cuadrado

Alumna:
Patricia González Bodega

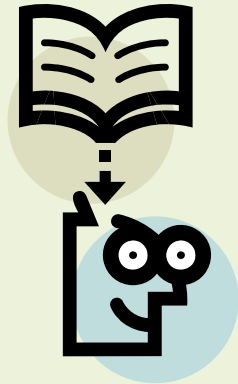
Febrero 2009

ÍNDICE


INTRODUCCIÓN
OBJETIVOS
ESTADO DEL ARTE
EXPERIMENTOS INICIALES
METODOLOGÍA
MAPEO CORPUS ESPAÑOL
EXPERIMENTACIÓN
CONCLUSIONES
FUTUROS TRABAJOS

INTRODUCCIÓN

- ¿En que consiste el proyecto?
- ¿ Que es etiquetar?
- ¿Qué es desambiguar?
- ¿Para que sirve?



OBJETIVOS

- Etiquetar morfológicamente textos en español, incluso en casos en la que la información contextual es escasa o nula.
- Integrar bajo el sistema de codificación del grupo KR los corpus anotados.
- Evaluar cual es el rendimiento de la herramienta

ESTADO DEL ARTE

- Análisis morfológico en procesamiento del lenguaje natural
- Etiquetado morfológico manual, automático y mixto
- Etiquetas del corpus CESS-ESP y Conll2002
- Herramienta de minería de datos : Weka

METODOLOGÍA

1. Se mapean dos corpus
   - CESS-ESP
   - Conll 2002

2. Experimentación: experimentos con los dos corpus individual y conjuntamente. Con palabras que contienen caracteres alfabéticos y con el total de las palabras.
   - Primera Fase. Desambiguación sin contexto.
     1. Se generan las reglas de desambiguación morfológica
     2. Computar el grado de acierto para cada regla

   - Segunda Fase. Desambiguación con contexto. Se parte de los resultados de la primera fase.
     1. Se generan las reglas de desambiguación morfológica
     2. Computar el grado de acierto para cada regla

CORRESPONDENCIA ENTRE CATEGORÍA Y ETIQUETA

- La correspondencia entre la categoría general y la etiqueta del grupo KR es la siguiente:

MAPEAR CORPUS CASTELLANO

- Cambiar las etiquetas del corpus CESS-ESP/Conll2002 por las etiquetas creadas por el grupo KR.
- CESS-ESP esta anotado morfológica y sintácticamente, se han eliminado algunas etiquetas del corpus.
- Comprobación de la correspondencia de categorías gramaticales.
- Existen algunas diferencias entre las asignaciones de las diferentes etiquetas para los dos corpus.

REALIZACIÓN DE LOS EXPERIMENTOS

- Generación de reglas de desambiguación de palabras de los corpus etiquetados anteriormente.

- La generación de dichas reglas, se realiza en dos fases:

1. PRIMERA FASE: las reglas se crean en función del término (sin contexto).

2. SEGUNDA FASE: las reglas se crean en función del contexto de la palabra y de las reglas generadas en la primera fase.

EXPERIMENTACIÓN:
GENERACIÓN DE REGLAS DE DESAMBIGUACIÓN

- Para la generación de reglas de desambiguación se utiliza la herramienta Weka, en la que se introduce una lista con los datos, la información que te proporciona está dividida en tres partes:

1ª parte: Información de los datos de entrada y las opciones de la ejecución

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     morfologia
Instances:    492404
Attributes:   3
              declinar
              longitud
              clase
Test mode:    10-fold cross-validation
```

EXPERIMENTACIÓN:
GENERACIÓN DE REGLAS DE DESAMBIGUACIÓN

2ª parte : Reglas para la etiquetación de las palabras

```
=== Classifier model (full training set) ===

PART decision list
------------------

longitud > 8 AND
declinar <= 120224 AND
declinar > 119884: 57 (2371.0/1.0)

declinar <= 124855 AND
declinar > 124573 AND
longitud > 11: 57 (222.0/2.0)

longitud > 8 AND
declinar > 118848 AND
declinar <= 118924: 57 (680.0)

longitud > 8 AND
declinar <= 124504 AND
declinar > 124392: 57 (261.0)

longitud > 8 AND
declinar <= 124668 AND
declinar > 124613: 57 (290.0/10.0)

longitud > 8 AND
declinar <= 116918 AND
declinar > 116763: 57 (552.0/1.0)
```

EXPERIMENTACIÓN:
GENERACIÓN DE REGLAS DE DESAMBIGUACIÓN

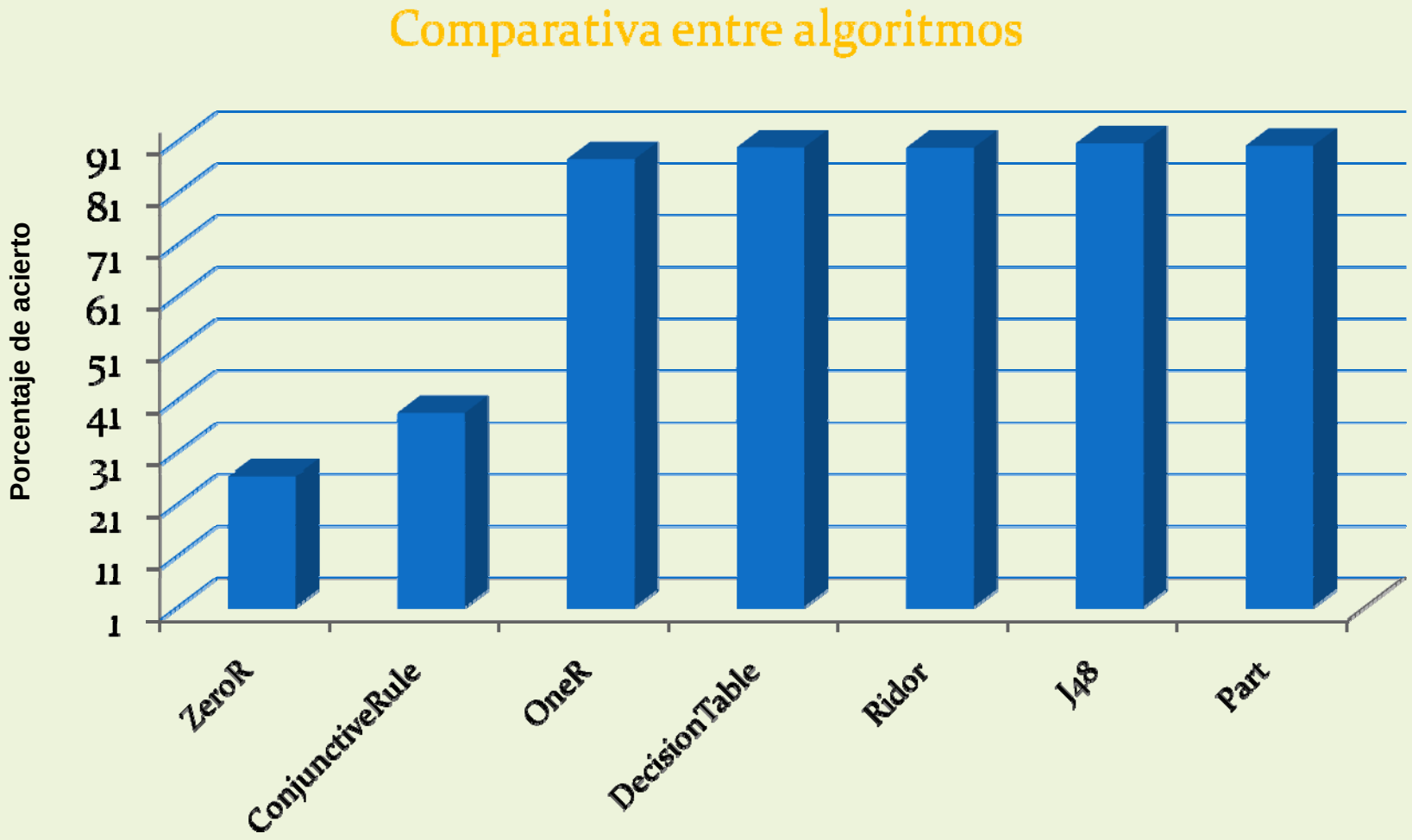
3ª parte :

Porcentajes de acierto

```
Time taken to build model: 8307.76 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      460376               93.4956 %
Incorrectly Classified Instances     32028                6.5044 %
Kappa statistic                          0.9225
Mean absolute error                      0.0195
Root mean squared error                  0.1029
Relative absolute error                 10.4344 %
Root relative squared error             33.6462 %
Total Number of Instances           492404

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.998     0.003      0.985     0.998      0.991    1
  0.924     0.013      0.76      0.924      0.834    10
  0.99      0.003      0.981     0.99       0.986    20
  0.885     0.008      0.928     0.885      0.906    30
  0.958     0.001      0.967     0.958      0.962    46
  0.755     0.015      0.823     0.755      0.788    52
  0.945     0.035      0.91      0.945      0.927    57
  0.75      0          0.991     0.75       0.854    72
  1         0          1         1          1        95

=== Confusion Matrix ===

     a      b      c      d      e      f      g      h      i   <-- classified as
 82342     33     35      1      2      5      3     71      0 |      a = 1
   415  19701    958      4     91     47     89     25      0 |      b = 10
   505     58  63071      0     47     23     10      0      0 |      c = 20
    15      8      1  46235      9   1735   4217      0      0 |      d = 30
    82     64     46     25  10655     66    157     29      0 |      e = 46
    25     37    130   1685     40  30780   8056      1      1 |      f = 52
    79    477     29   1890    176   4698 127307     37     16 |      g = 57
   156   5550      0      0      4     27     33  17325      0 |      h = 72
     0      0      0      0      0      0      5      0  62960 |      i = 95
```

EXPERIMENTACIÓN:
GENERACIÓN DE REGLAS DE DESAMBIGUACIÓN

- Elección del algoritmo de Clasificación:

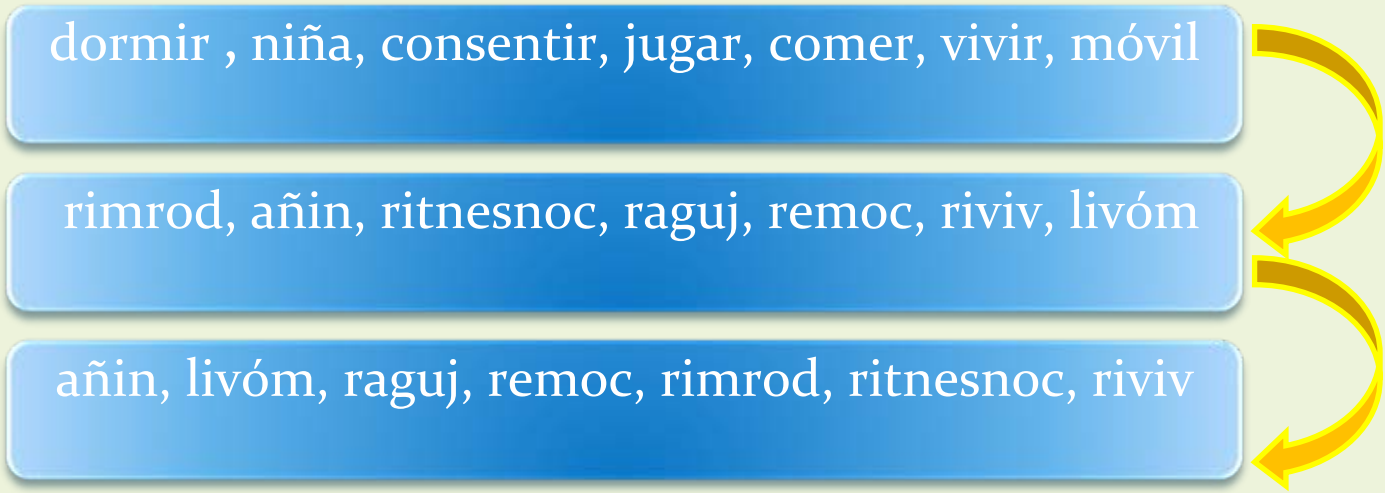


EXPERIMENTACIÓN (PRIMERA FASE): GENERACIÓN DEL DICCIONARIO

- Los términos objeto de estudio deben procesarse de forma sistemática. Con este objetivo se han organizado en un listado (que denominaremos diccionario). Etapas:

1. Se utilizó una lista muy amplia de palabras en castellano
2. Se escribieron al revés y
3. Se ordenaron alfabéticamente, para poder tenerlas ordenadas según su terminación

- Un posible ejemplo del diccionario que se ha creado es el siguiente:



EXPERIMENTACIÓN (PRIMERA FASE): INFORMACIÓN ANALIZADA Y EXPERIMENTOS

- En la lista introducida en Weka, para cada palabra, se dispone de la siguiente información:

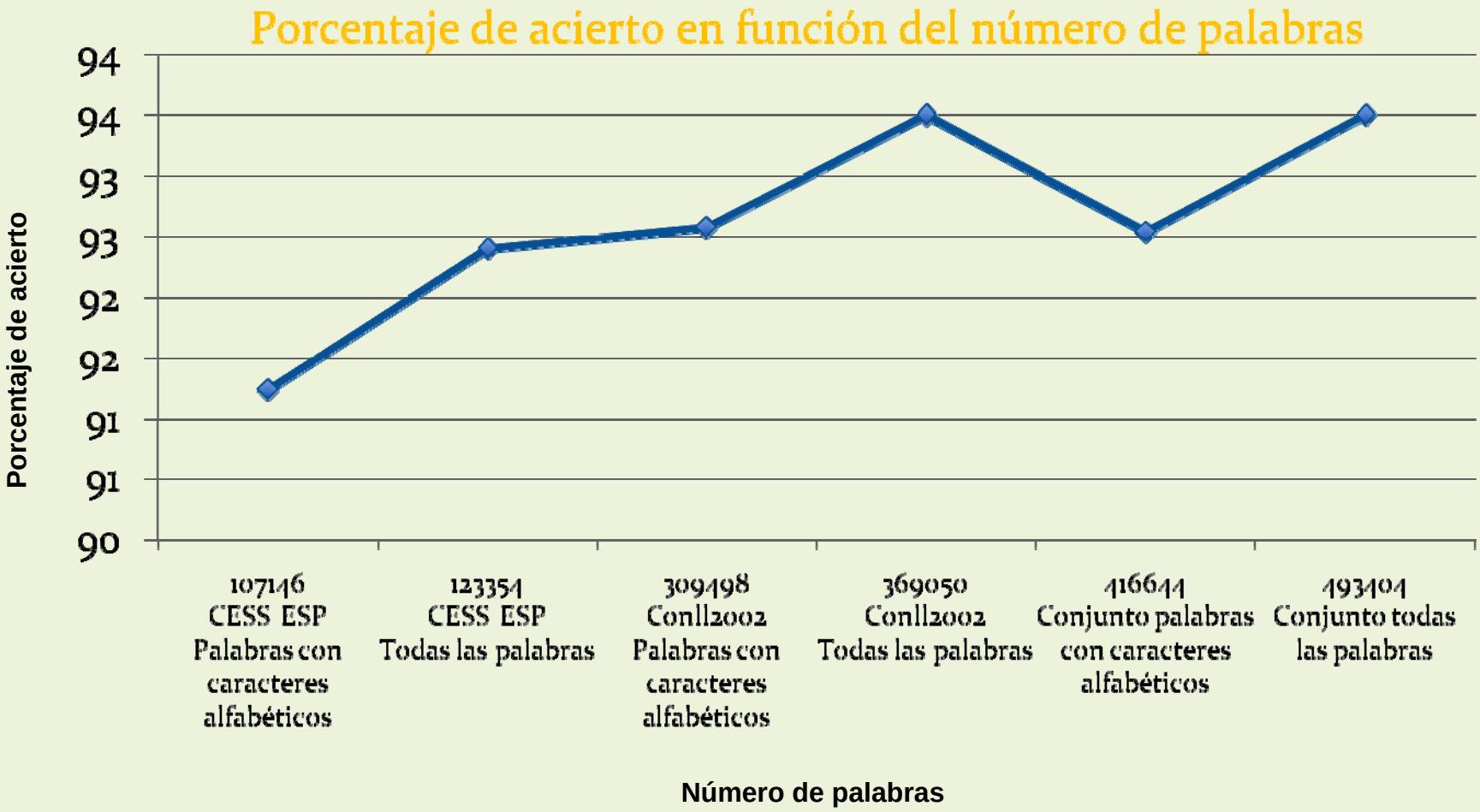
Posición en el diccionario, longitud, categoría general de palabra

- Para esta fase se han realizado 6 experimentos:
  1. CESS y términos solo alfabéticos
  2. CESS y términos alfabéticos, alfanuméricos y caracteres especiales
  3. CONLL y términos solo alfabéticos
  4. CONLL y términos alfabéticos, alfanuméricos y caracteres especiales
  5. CESS más CONLL y términos solo alfabéticos
  6. CESS más CONLL y términos alfabéticos, alfanuméricos y caracteres especiales

EXPERIMENTACIÓN (PRIMERA FASE): EJEMPLO DE REGLA PRODUCIDA

```
longitud > 8 AND
declinar <= 124668 AND
declinar > 124613: 57 (290.0/10.0)
```

EXPERIMENTACIÓN (PRIMERA FASE): RESULTADOS DE LA PRIMERA FASE



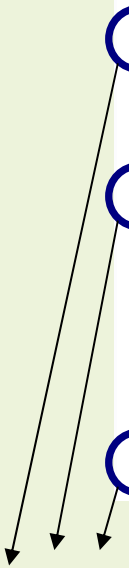
EXPERIMENTACIÓN (SEGUNDA FASE): DISEÑO

- Esta segunda fase parte de la salida de la primera fase.
- A diferencia de la anterior fase se tiene en cuenta el contexto: Se generan reglas de desambiguación comparando y analizando las categorías de las palabras que tenga delante y/o detrás la palabra a etiquetar.
- Se han realizado 43 experimentos diferentes, bajo los siguientes parámetros

VENTANA CONFIGURABLE		
NÚMERO DE PALABRAS	POSICIÓN DE LA PALABRA	CÓDIGO DE NÚMERO DE REGLA

EXPERIMENTACIÓN (SEGUNDA FASE): VENTANA PARA CONFIGURAR LOS PARÁMETROS

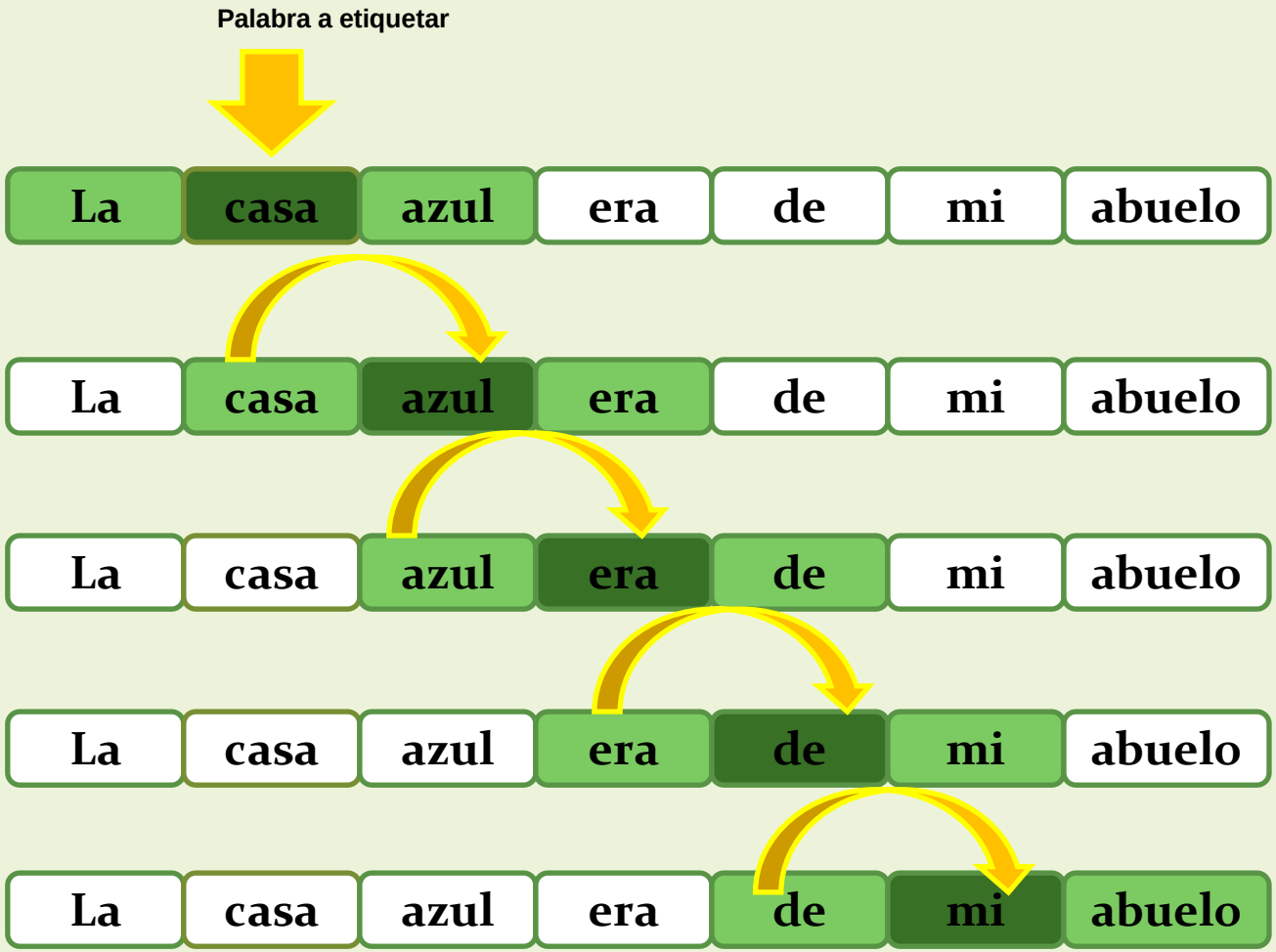
```
Configuración de la ventana

Introduzca el número de palabras que contendrá dicha ventana

3
Introduzca la posición de la palabra en la ventana

2
Introduzca 0, 1 o 2 si desea que no se escriba ningún número de regla,
 solo el número de regla de la palabra a evaluar
o todos los números de regla correspondientemente.

2
```



Parámetros de configuración: 3
2 2

EXPERIMENTACIÓN (SEGUNDA FASE): EJEMPLO

- Ejemplo de cómo sería etiquetar todas las palabras de una frase, en la que la ventana es de 3 y la posición de la palabra a evaluar es la segunda:



EXPERIMENTACIÓN (SEGUNDA FASE): WEKA

```
@relation       morfologia2
@attribute      1clase{1,10,20,30,46,52,57,72,95}
@attribute       1numero          integer
@attribute      2clase{1,10,20,30,46,52,57,72,95}
@attribute       2numero          integer
@attribute      3clase{1,10,20,30,46,52,57,72,95}
@attribute       3numero          integer
@attribute       2clasecorpus  {1,10,20,30,46,52,57,72,95}
@data
57,2692,30,1770,95,179,57
30,1770,95,179,57,1265,95
95,179,57,1265,95,179,57
57,1265,95,179,95,179,95
95,179,95,179,57,792,95
95,179,57,792,57,2442,57
57,792,57,2442,95,179,57
```

```
3clase = 57 AND
1numero <= 1354 AND
1numero > 931 AND
1clase = 10: 20 (4.0 /1.0)
```

EXPERIMENTACIÓN (SEGUNDA FASE): RESULTADOS SEGÚN CORPUS ANALIZADOS, TÉRMINOS Y PARÁMETROS

Corpus	Palabras de la 2ª fase	1ª Fase	Parámetros de la ventana	Número de experimento	Número instancias	Porcentaje de acierto
Cess-esp	Con todas las Palabras	El mismo 92,399%	3 1 2	1	123351	94,7564
			3 2 1	2	123351	95,8452
			3 2 0	3	123351	95,1415
			3 2 2	4	123351	96,1095
			4 2 2	5	123350	95,7835
			2 1 2	6	123352	95,0799
			2 2 2	7	123352	95,8323
		1º Conll2002 93,499%	3 2 2	12	123351	94,1046
			4 2 2	13	123350	93,6214
			2 1 2	14	123352	92,8003
			2 2 2	15	123352	93,9263
	Solo palabras con caracteres alfabéticos	El mismo 91,237%	3 2 2	8	107143	95,2008
			4 2 2	9	107142	94,8666
			2 1 2	10	107144	94,1677
			2 2 2	11	107144	94,9974
		1º Conll2002 92,57%	3 2 2	16	107143	92,7023
			4 2 2	17	107142	92,1767
			2 1 2	18	107144	91,4974
			2 2 2	19	107144	92,7574

EXPERIMENTACIÓN (SEGUNDA FASE): RESULTADOS SEGÚN CORPUS ANALIZADOS, TÉRMINOS Y PARÁMETROS

Corpus	Palabras de La 2ª fase	1ª Fase	Parámetros de la ventana	Número de experimento	Número instancias	Porcentaje de acierto
Conll2002	Con todas las palabras	El mismo 93,499%	3 2 2	20	369047	96,2539
			4 2 2	21	369046	96,0317
			2 1 2	22	369048	95,4887
			2 2 2	23	369048	96,0081
		1º cess-esp 92,399%	3 2 2	28	369047	94,3303
			4 2 2	29	369046	93,9994
			2 1 2	30	369048	92,5348
			2 2 2	31	369048	94,0360
	Solo palabras con caracteres Alfabéticos	El mismo 92,57%	3 2 2	24	309495	95,4141
			4 2 2	25	309494	95,1165
			2 1 2	26	309496	94,7492
			2 2 2	27	309496	95,1899
		1º cess-esp 91,237%	3 2 2	32	309495	92,7876
			4 2 2	33	309494	92,4144
			2 1 2	34	309496	91,0694
			2 2 2	35	309496	92,7531

EXPERIMENTACIÓN (SEGUNDA FASE): RESULTADOS SEGÚN CORPUS ANALIZADOS, TÉRMINOS Y PARÁMETROS

Corpus	Palabras de la 2ª fase	1ª Fase	Parámetros de la ventana	Número de experimento	Número instancias	Porcentaje de acierto
Conjunto	Con todas las Palabras	El mismo 93,4956%	3 2 2	36	492401	95,9245
			4 2 2	37	492400	95,6956
			2 1 2	38	492402	95,3189
			2 2 2	39	492402	95,7330
	Solo palabras con caracteres Alfabéticos	El mismo 92,5392%	3 2 2	40	416641	95,1555
			4 2 2	41	416640	94,8673
			2 1 2	42	416642	94,5500
			2 2 2	43	416642	95,0847

```
Que desea realizar:
 1 - Mapear
 2 - 1ª Fase de desambiguación
 3 - 2ª Fase de desambiguación
 4 - Eliminar las etiquetas del corpus
 5 - Crear lista de palabras ordenadas
```

CONCLUSIONES

- Etiquetado y desambiguación morfológica del castellano con reducida información contextual realizado con éxito.
- Se han mapeado correctamente los corpus escogidos.
- Cuanto mayor sea el texto a etiquetar, se obtienen mejores resultados
- Mejores resultados cuando se etiquetan todas las palabras.
- En la segunda fase , la mejor configuración de ventana es: 3 2 2
- Para textos en ingles, mejor no abreviar
- Mejor algoritmo rules –part

TRABAJOS FUTUROS

- Realizar las comprobaciones hechas para los corpus en ingles, para los hechos en castellano.
- Realizar los experimentos existentes con corpus más amplios.
- Realizar la segunda fase de experimentación teniendo en cuenta más palabras del contexto.
- Realizar la experimentación para corpus en inglés.
- Crear un texto en castellano etiquetado morfológicamente a partir de las reglas generadas.

¿DUDAS Y SUGERENCIAS?