

## 1. Experimentación

### 1.1. Segundo experimento: Utilizar la nueva fuente de información generada para entrenar y etiquetar el corpus WSJ

El segundo experimento realizado tiene como objetivo evaluar la nueva fuente de información obtenida (NFI) como corpus de entrenamiento. Para esto se entrenarán 2 etiquetadores gramaticales (Stanford Tagger y TnT) y se etiquetará con ellos el corpus Wall Street Journal (WSJ). Posteriormente se realizarán mediciones de desempeño pertinentes.

### 1.2. Tercer experimento: Utilizar la nueva fuente de información generada para entrenar y etiquetar el corpus BNC

El tercer experimento realizado tiene como objetivo evaluar la nueva fuente de información obtenida (NFI) como corpus de entrenamiento. Para esto se entrenarán 2 etiquetadores gramaticales (Stanford Tagger y TnT) y se etiquetará con ellos el British National Corpus (BNC).

#### 1.2.1. Etiquetar el corpus BNC con Stanford Tagger

La segunda evaluación de este experimento consiste en entrenar el etiquetador gramatical Stanford Tagger con WSJ como corpus de entrenamiento y con WSJ + NFI. Luego se procede a etiquetar el BNC plano (sin etiquetas gramaticales) con estos dos modelos. Por último se contruye la matriz de confusión:

**Cuadro 1:** *BNC original contra BNC etiquetado con MaxEnt (entrenado con WSJ)*

MaxEnt(BNC) BNC	NNP	NN1	AJ0	JJ	DT0	NN	AV0	VBN	NNS	NN0
<b>NNP</b>	-	-	-	-	-	-	-	-	-	-
<b>NN1</b>	<b>26141</b>	-	-	<b>4045</b>	-	-	-	115	533	-
<b>AJ0</b>	<b>8675</b>	-	-	-	-	<b>2860</b>	-	<b>3276</b>	30	-
<b>JJ</b>	-	-	-	-	-	-	-	-	-	-
<b>DT0</b>	119	-	-	<b>8132</b>	-	192	-	-	5	-
<b>NN</b>	-	-	-	-	-	-	-	-	-	-
<b>AV0</b>	982	-	-	2314	-	<b>3753</b>	-	159	236	-
<b>VBN</b>	-	-	-	-	-	-	-	-	-	-
<b>NNS</b>	-	-	-	-	-	-	-	-	-	-
<b>NN0</b>	567	-	-	1115	-	-	-	19	<b>3132</b>	-

Porcentaje de aciertos: 92,86 %

**Cuadro 2:** *BNC original contra BNC etiquetado con MaxEnt (entrenado con WSJ + NFI)*

MaxEnt(WSJ+NFI) BNC	NNP	NN1	AJ0	JJ	DT0	NN	AV0	CD	NN0	NNS
<b>NNP</b>	-	-	-	-	-	-	-	-	-	-
<b>NN1</b>	<b>26206</b>	-	-	<b>3864</b>	-	-	-	22	-	663
<b>AJ0</b>	<b>8263</b>	-	-	-	-	2099	-	4	-	25
<b>JJ</b>	-	-	-	-	-	-	-	-	-	-
<b>DT0</b>	109	-	-	<b>7836</b>	-	188	-	-	-	4
<b>NN</b>	-	-	-	-	-	-	-	-	-	-
<b>AV0</b>	840	-	-	2210	-	<b>3640</b>	-	195	-	279
<b>CD</b>	-	-	-	-	-	-	-	-	-	-
<b>NN0</b>	469	-	-	863	-	-	-	<b>3222</b>	-	<b>3146</b>
<b>NNS</b>	-	-	-	-	-	-	-	-	-	-

Porcentaje de aciertos: 93,01 %

Se puede observar que el rendimiento del etiquetador entrenado con WSJ es un poco mejor (93,01 %) que cuando es entrenado con WSJ + NFI (92,86 %). La mayoría de los errores para Stanford Tagger entrenado con WSJ se da en etiquetas NN del gold standard cuando son etiquetadas como JJ y NNP. Para Stanford Tagger entrenado con WSJ + NFI la mayoría de los errores se da en las mismas etiquetas, pero con cantidad de errores mayor, sobre todo para NN etiquetado como JJ.

La segunda evaluación de este experimento consiste en entrenar Stanford Tagger con la mitad de WSJ y con la mitad de WSJ + NFI. Posteriormente con estos dos modelos se etiqueta BNC y se construye la matriz de confusión. Se realiza la misma operación para cada mitad:

**Cuadro 3:** *BNC original contra BNC etiquetado con MaxEnt (entrenado con 2 mitad de WSJ)*

MaxEnt(2WSJ) BNC	NNP	NN1	AJ0	JJ	DT0	NN	AV0	NNS	NN0	VCN
<b>NNP</b>	-	-	-	-	-	-	-	-	-	-
<b>NN1</b>	<b>26061</b>	-	-	<b>4689</b>	-	-	-	620	-	97
<b>AJ0</b>	<b>8570</b>	-	-	-	-	<b>3574</b>	-	49	-	<b>3001</b>
<b>JJ</b>	-	-	-	-	-	-	-	-	-	-
<b>DT0</b>	133	-	-	<b>8068</b>	-	224	-	1	-	-
<b>NN</b>	-	-	-	-	-	-	-	-	-	-
<b>AV0</b>	985	-	-	2428	-	<b>3611</b>	-	475	-	143
<b>NNS</b>	-	-	-	-	-	-	-	-	-	-
<b>NN0</b>	554	-	-	1404	-	-	-	<b>3128</b>	-	7
<b>VCN</b>	4	-	-	-	-	-	-	-	-	-

Porcentaje de aciertos: 92,6 %

**Cuadro 4:** *BNC original contra BNC etiquetado con MaxEnt (entrenado con 2 mitad de WSJ+NFI)*

<div>MaxEnt(2WSJ+NFI)</div> <div>BNC</div>	NNP	NN1	AJ0	JJ	DT0	NN	AV0	CD	NN0	NNS
<b>NNP</b>	-	-	-	-	-	-	-	-	-	-
<b>NN1</b>	<b>26036</b>	-	-	<b>3867</b>	-	-	-	17	-	657
<b>AJ0</b>	<b>8130</b>	-	-	-	-	2151	-	5	-	24
<b>JJ</b>	-	-	-	-	-	-	-	-	-	-
<b>DT0</b>	108	-	-	<b>7742</b>	-	215	-	-	-	3
<b>NN</b>	-	-	-	-	-	-	-	-	-	-
<b>AV0</b>	875	-	-	2220	-	<b>3538</b>	-	203	-	276
<b>CD</b>	-	-	-	-	-	-	-	-	-	-
<b>NN0</b>	452	-	-	866	-	-	-	<b>3240</b>	-	<b>3138</b>
<b>NNS</b>	-	-	-	-	-	-	-	-	-	-

Porcentaje de aciertos: 93,01 %

**Cuadro 5:** *BNC original contra BNC etiquetado con MaxEnt (entrenado con 1 mitad de WSJ)*

<div>MaxEnt(1WSJ)</div> <div>BNC</div>	NNP	NN1	AJ0	JJ	DT0	NN	VBN	AV0	NN2	NNS
<b>NNP</b>	-	-	-	-	-	-	-	-	-	-
<b>NN1</b>	<b>27101</b>	-	-	<b>4519</b>	-	-	146	-	-	550
<b>AJ0</b>	<b>9043</b>	-	-	-	-	<b>3838</b>	<b>3837</b>	-	-	43
<b>JJ</b>	-	-	-	-	-	-	-	-	-	-
<b>DT0</b>	128	-	-	<b>8324</b>	-	185	-	-	-	1
<b>NN</b>	-	-	-	-	-	-	-	-	-	-
<b>VBN</b>	-	-	-	-	-	-	-	-	-	-
<b>AV0</b>	1071	-	-	2583	-	<b>3445</b>	141	-	-	311
<b>NN2</b>	<b>3415</b>	-	-	83	-	885	-	-	-	-
<b>NNS</b>	-	-	-	-	-	-	-	-	-	-

Porcentaje de aciertos: 92,45 %

**Cuadro 6:** *BNC original contra BNC etiquetado con MaxEnt (entrenado con 1 mitad de WSJ+NFI)*

<div>MaxEnt(1WSJ+NFI)</div> <div>BNC</div>	NNP	NN1	AJ0	JJ	DT0	NN	AV0	CD	NN0	VBN
<b>NNP</b>	-	-	-	-	-	-	-	-	-	-
<b>NN1</b>	<b>26776</b>	-	-	<b>3786</b>	-	-	-	20	-	109

**Cuadro 6:** *BNC original contra BNC etiquetado con MaxEnt (entrenado con 1 mitad de WSJ+NFI)*

MaxEnt(1WSJ+NFI) BNC	NNP	NN1	AJ0	JJ	DT0	NN	AV0	CD	NN0	VBN
<b>AJ0</b>	<b>8368</b>	-	-	-	-	2113	-	3	-	<b>3249</b>
<b>JJ</b>	-	-	-	-	-	-	-	-	-	-
<b>DT0</b>	108	-	-	<b>7768</b>	-	192	-	-	-	-
<b>NN</b>	-	-	-	-	-	-	-	-	-	-
<b>AV0</b>	950	-	-	2360	-	<b>3475</b>	-	194	-	90
<b>CD</b>	-	-	-	-	-	-	-	-	-	-
<b>NN0</b>	473	-	-	752	-	-	-	<b>3302</b>	-	2
<b>VBN</b>	-	-	-	-	-	-	-	-	-	-

Porcentaje de aciertos: 92,91 %

Se puede apreciar una leve mejoría en el porcentaje de etiquetas acertadas; 92,6 % contra 93,01 % y 92,45 % contra 92,91 % para cada modelo respectivamente. Los errores más comunes son producidos en etiquetas NN del gold standard cuando son etiquetadas como JJ y NNP por , para las dos mitades entrenadas tanto con WSJ como con WSJ + NFI. Se puede notar que el porcentaje de error al etiquetar JJ cuando era NN es menor en la evaluación realizada sobre Stanford Tagger entrenado con WSJ + NFI.

A continuación se presentan las matrices de confusión para BNC etiquetado con Stanford Tagger entrenado con la mitad de WSJ con y sin NFI.

**Cuadro 7:** *BNC etiquetado por MaxEnt (entrenado con 2 mitad WSJ) vs BNC etiquetado con MaxEnt (entrenado con 2 mitad de WSJ + NFI)*

MaxEnt(2WSJ+NFI) MaxEnt(2WSJ)	NN	IN	DT	JJ	NNP	NNS	,	
<b>NN</b>	<b>258306</b>	107	7	4691	2413	392	-	
<b>IN</b>	247	<b>213508</b>	552	146	203	32	-	
<b>DT</b>	8	152	<b>188537</b>	-	3	-	-	
<b>JJ</b>	3864	133	2	<b>131409</b>	1168	85	-	
<b>NNP</b>	2727	93	243	1688	<b>128525</b>	938	-	
<b>NNS</b>	331	-	-	82	433	<b>95875</b>	-	
<b>,</b>	-	-	-	-	-	-	<b>95405</b>	
<b>.</b>	-	-	-	-	-	-	-	<b>80</b>
<b>RB</b>	525	298	10	1115	271	24	-	
<b>PRP</b>	1	-	-	12	-	-	-	

Porcentaje de aciertos: 9,51 %

**Cuadro 8:** BNC etiquetado por MaxEnt (entrenado con 1 mitad WSJ) vs BNC etiquetado con MaxEnt (entrenado con 1 mitad de WSJ + NFI)

MaxEnt(1WSJ+NFI) MaxEnt(1WSJ)	NN	IN	DT	NNP	JJ	NNS	,	
<b>NN</b>	<b>257540</b>	74	4	2370	5058	414	-	
<b>IN</b>	302	<b>213004</b>	640	246	169	49	-	
<b>DT</b>	8	214	<b>188559</b>	9	-	-	-	
<b>NNP</b>	2670	45	100	<b>130600</b>	2055	957	-	
<b>JJ</b>	3812	81	3	1138	<b>129382</b>	80	-	
<b>NNS</b>	208	-	-	561	64	<b>96407</b>	-	
,	-	-	-	-	-	-	<b>95405</b>	
.	-	-	-	-	-	-	-	<b>80</b>
<b>RB</b>	453	278	100	209	1218	34	-	
<b>PRP</b>	3	-	-	-	4	-	-	

Porcentaje de aciertos: 9,56 %

La tercer evaluación de este experimento consiste en entrenar Stanford Tagger con un cuarto de WSJ y con un cuarto de WSJ + NFI. Posteriormente con estos dos modelos se etiqueta BNC y se construye la matriz de confusión. Se realiza la misma operación para cada uno de los cuartos:

**Cuadro 9:** Rendimiento de Stanford Tagger entrenado con cuartos de WSJ con y sin NFI

Evaluación	Porcentaje de aciertos
Stanford Tagger entrenado con el primer 1/4 de WSJ	92.09 %
Stanford Tagger entrenado con el primer 1/4 de WSJ + NFI	92.92 %
Stanford Tagger entrenado con el segundo 1/4 de WSJ	92.10 %
Stanford Tagger entrenado con el segundo 1/4 de WSJ + NFI	92.91 %
Stanford Tagger entrenado con el tercer 1/4 de WSJ	92.14 %
Stanford Tagger entrenado con el tercer 1/4 de WSJ + NFI	92.89 %
Stanford Tagger entrenado con el cuarto 1/4 de WSJ	91.98 %
Stanford Tagger entrenado con el cuarto 1/4 de WSJ + NFI	92.83 %

En todos los casos se puede apreciar una mejora en el acierto de etiquetas para el corpus de entrenamiento WSJ + NFI contra WSJ.

La cuarta evaluación de este experimento consiste en entrenar Stanford Tagger con un décimo de WSJ y con un décimo de WSJ + NFI. Posteriormente con estos dos modelos se etiqueta BNC y se presentan los resultados:

- 91.25 % de acierto de etiquetas para el etiquetado de BNC con Stanford Tagger entrenado con 1/10 WSJ
- 92.81 % de acierto de etiquetas para el etiquetado de BNC con Stanford Tagger entrenado con 1/10 WSJ+NFI

Se puede apreciar un aumento del porcentaje de aciertos en el corpus de entrenamiento que incorpora NFL.

## **2. Conclusiones**