

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.

1.1.1. Etiquetar el corpus WSJ con Stanford Tagger

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

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

MaxEnt(WSJ) WSJ	JJ	NN	IN	RB	NNP	VBN	NNPS	VBG	VBD	RP
JJ	-	301	13	117	164	166	-	63	10	-
NN	482	-	5	3	277	7	-	77	6	-
IN	29	2	-	211	9	-	-	1	-	141
RB	160	29	371	-	33	2	-	-	-	56
NNP	134	111	10	4	-	3	89	-	-	-
VBN	259	7	-	-	5	-	-	-	209	-
NNPS	16	-	-	-	253	-	-	-	-	-
VBG	43	221	-	-	7	1	-	-	-	-
VBD	9	5	-	-	1	199	-	-	-	-
RP	1	-	159	53	-	-	-	-	-	-

Porcentaje de aciertos: 97,9 %

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

MaxEnt(WSJ+NFI) WSJ	JJ	NN	IN	RB	NNP	VBD	VBN	RP	NNPS	VBG
JJ	-	1260	68	583	685	45	851	2	-	237
NN	1918	-	16	73	1314	21	19	3	-	440
IN	107	3	-	1062	29	-	-	1005	-	2
RB	742	177	1403	-	141	-	3	527	-	1
NNP	555	458	29	21	-	4	9	1	546	5
VBD	34	24	-	-	10	-	923	-	-	-

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

MaxEnt(WSJ+NFI) WSJ	JJ	NN	IN	RB	NNP	VBD	VCN	RP	NNPS	VBG
VCN	980	49	-	-	29	1049	-	-	-	-
RP	3	-	427	165	1	-	-	-	-	-
NNPS	39	-	-	-	935	-	-	-	-	-
VBG	294	817	-	-	25	-	1	-	-	-

Porcentaje de aciertos: 97,73 %

Se puede observar que el rendimiento del etiquetador entrenado con WSJ es un poco mejor (97,9 %) que cuando es entrenado con WSJ + NFI (97,73 %). 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 la mitad restante de WSJ y se construye la matriz de confusión. Se realiza la misma operación para cada mitad:

Cuadro 3: *1 mitad WSJ original contra 1 mitad WSJ etiquetado con MaxEnt (entrenado con 2 mitad de WSJ)*

MaxEnt(2WSJ) WSJ	JJ	NN	NNP	IN	RB	NNPS	VCN	VBD	VBG	RP
JJ	-	1309	606	32	299	-	746	39	263	1
NN	1558	-	1027	6	38	-	22	15	403	-
NNP	401	542	-	23	38	250	8	10	37	-
IN	72	5	26	-	489	-	1	-	2	304
RB	512	104	31	989	-	-	2	1	1	123
NNPS	31	-	943	-	-	-	-	-	-	-
VCN	545	28	36	-	-	-	-	722	-	-
VBD	41	26	9	-	-	-	604	-	-	-
VBG	192	614	22	-	-	-	1	-	-	-
RP	2	3	1	449	179	-	-	-	-	-

Porcentaje de aciertos: 96,59 %

Cuadro 4: 1 mitad WSJ original contra 1 mitad WSJ etiquetado con MaxEnt (entrenado con 2 mitad de WSJ + NFI)

MaxEnt(2WSJ+NFI) WSJ	JJ	NN	NNP	IN	RB	NNPS	VBN	VBD	RP	VBG
JJ	-	1119	571	31	334	-	625	31	1	176
NN	1434	-	1033	6	41	-	13	14	1	343
NNP	445	497	-	19	24	257	3	6	-	28
IN	80	7	27	-	531	-	-	-	573	2
RB	470	81	73	851	-	-	2	1	251	-
NNPS	27	3	834	-	-	-	-	-	-	-
VBN	564	26	34	-	-	-	-	642	-	-
VBD	35	16	14	-	-	-	748	-	-	-
RP	-	1	1	221	106	-	-	-	-	-
VBG	262	531	26	-	-	-	1	-	-	-

Porcentaje de aciertos: 96,76 %

Cuadro 5: 2 mitad WSJ original contra 2 mitad WSJ etiquetado con MaxEnt (entrenado con 1 mitad de WSJ)

MaxEnt(1WSJ) WSJ	JJ	NN	IN	RB	NNP	VBN	VBD	NNPS	VBG	RP
JJ	-	1197	45	381	522	483	37	-	202	1
NN	1604	-	12	36	916	21	16	-	381	1
IN	50	3	-	698	26	-	-	-	2	294
RB	466	168	944	-	62	1	1	-	1	196
NNP	462	503	18	19	-	19	2	518	16	1
VBN	863	29	-	1	32	-	779	-	-	-
VBD	66	43	2	-	16	570	-	-	-	-
NNPS	16	-	-	-	651	-	-	-	-	-
VBG	198	572	-	-	19	-	-	-	-	-
RP	3	1	426	129	-	-	-	-	-	-

Porcentaje de aciertos: 96,64 %

Cuadro 6: 2 mitad WSJ original contra 2 mitad WSJ etiquetado con MaxEnt (entrenado con 1 mitad de WSJ)

MaxEnt(1WSJ+NFI) WSJ	JJ	NN	NNP	IN	RB	VBN	VBD	NNPS	RP	VBG
JJ	-	997	522	43	382	444	26	1	8	126
NN	1482	-	1011	10	42	18	12	-	2	341
NNP	475	467	-	16	17	8	3	483	1	10
IN	50	3	29	-	727	-	-	-	586	2

Cuadro 6: 2 mitad WSJ original contra 2 mitad WSJ etiquetado con MaxEnt (entrenado con 1 mitad de WSJ)

MaxEnt(1WSJ+NFI)	JJ	NN	NNP	IN	RB	VBN	VBD	NNPS	RP	VBG
WSJ										
RB	438	141	105	819	-	1	-	-	344	1
VBN	810	28	27	-	-	-	706	-	-	-
VBD	40	20	11	-	-	742	-	-	-	-
NNPS	13	-	597	-	-	-	-	-	-	-
RP	2	-	-	237	74	-	-	-	-	-
VBG	256	530	21	-	-	-	-	-	-	-

Porcentaje de aciertos: 96,73 %

Se puede apreciar una leve mejoría en el porcentaje de etiquetas acertadas; 96,23 % contra 96,46 % y 96,20 % contra 96,36 % para cada mitad respectivamente. Los errores más comunes son producidos en etiquetas NN del gold standard cuando son etiquetadas como JJ y NNP por TnT, 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 TnT entrenado con WSJ + NFI.

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

Cuadro 7: 1 mitad WSJ etiquetado por MaxEnt (entrenado con 2 mitad WSJ) vs 1 mitad WSJ etiquetado con MaxEnt (entrenado con 2 mitad de WSJ + NFI)

MaxEnt(2WSJ+NFI)	JJ	NN	RP	IN	VBN	VBD	RB	NNP	VB	WDT
MaxEnt(2WSJ)										
JJ	-	596	1	15	175	19	202	183	48	-
NN	686	-	2	-	13	21	46	291	276	-
RP	-	-	-	6	-	-	7	-	-	-
IN	17	6	507	-	-	-	318	8	-	41
VBN	309	18	-	-	-	248	-	9	2	-
VBD	29	9	-	-	460	-	-	11	11	-
RB	128	17	206	141	1	-	-	65	14	-
NNP	268	242	-	4	5	-	8	-	14	-
VB	21	113	-	-	12	15	3	10	-	-
WDT	-	-	-	252	-	-	-	-	-	-

Porcentaje de aciertos: 98,5 %

Cuadro 8: 2 mitad WSJ etiquetado por MaxEnt (entrenado con 1 mitad WSJ) vs 2 mitad WSJ etiquetado con MaxEnt (entrenado con 1 mitad de WSJ + NFI)

MaxEnt(1WSJ+NFI) MaxEnt(1WSJ)	JJ	NN	VBN	VBD	RP	IN	NNP	RB	VB	VBP
JJ	-	513	254	37	-	9	196	195	48	13
NN	611	-	25	27	1	5	344	32	244	64
VBN	246	18	-	257	-	-	7	-	9	2
VBD	28	12	494	-	-	-	3	1	2	3
RP	-	-	-	-	-	13	-	11	-	-
IN	9	3	-	2	494	-	18	326	-	2
NNP	211	215	8	4	-	5	-	26	24	6
RB	149	13	1	-	222	166	70	-	19	3
VB	22	97	16	10	-	5	25	7	-	100
VBP	17	132	6	15	-	2	4	-	283	-

Porcentaje de aciertos: 98,51 %

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 los 3/4 restantes de WSJ y se construye la matriz de confusión. Se realiza la misma operación para cada uno de los cuartos:

Cuadro 9: Rendimiento de TnT entrenado con cuartos de WSJ con y sin NFI

Evaluación	Porcentaje de aciertos
Stanford Tagger entrenado con el primer 1/4 de WSJ	96.30 %
Stanford Tagger entrenado con el primer 1/4 de WSJ + NFI	96.57 %
Stanford Tagger entrenado con el segundo 1/4 de WSJ	96.30 %
Stanford Tagger entrenado con el segundo 1/4 de WSJ + NFI	96.52 %
Stanford Tagger entrenado con el tercer 1/4 de WSJ	96.28 %
Stanford Tagger entrenado con el tercer 1/4 de WSJ + NFI	96.57 %
Stanford Tagger entrenado con el cuarto 1/4 de WSJ	96.24 %
Stanford Tagger entrenado con el cuarto 1/4 de WSJ + NFI	96.53 %

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 los 9/10 restantes de WSJ y se presentan los resultados:

- 95.67 % de acierto de etiquetas para el etiquetado de 9/10 de WSJ con Stanford Tagger entrenado con 1/10 WSJ
- 96.27 % de acierto de etiquetas para el etiquetado de 9/10 de WSJ 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