UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Nuevas fuentes de información para entrenamiento de etiquetadores gramaticales

Tesista: Fernando Jorge Rodriguez
Director: Dr. José Castaño

Buenos Aires, Marzo de 2012.

0.1. Experimentación

0.1.1. Primer experimento

El primer experimento consiste en medir (generando una matriz de confusión) la información extraída de COBUILD contra la misma información generada a partir de un etiquetador automático (TnT). Es decir, la información extraída de COBUILD, como se mencionó anteriormente, es la unión de definiciones y ejemplos, con la información gramatical correspondiente a la palabra definida. A continuación se presenta un pequeño extracto:

A	are
cat NN	often
is	kept
a	as
small	pets
furry	.
animal	She
with	put
a	out
tail	a
,	hand
whiskers	and
,	stroked
and	the
sharp	cat NN
claws	softly
that	...
kills	...
smaller	domestic
animals	animals
such	such
as	as
mice	dogs
and	and
birds	cats NNS
.	.
Cats NNS	

Esta es la información extraída de COBUILD para la palabra *cat*; la unión de la definición:

A cat is a small furry animal with a tail, whiskers, and sharp claws that kills smaller animals such as mice and birds. Cats are often kept as pets.

y los ejemplos

She put out a hand and stroked the cat softly...
...domestic animals such as dogs and cats.

Se puede notar la información gramatical expresada mediante las etiquetas NN y NNS para las palabras *cat* y *cats* respectivamente. La idea de este experimento será comparar estas etiquetas contra las etiquetas asignadas por el etiquetador automático TnT. Entonces se tomará este corpus plano (sin etiquetas), se lo etiquetará utilizando TnT entrenado con el corpus de entrenamiento Wall Street Journal (de ahora en más WSJ) [1] y luego se realizará la comparación.
La matriz de confusión[2] generada a partir de dicha comparación es la siguiente:

Cuadro 1: *Matriz de confusión para etiquetas extraídas de COBUILD vs generadas por TnT*

COBUILD \ TnT	**NN**	**VB**	**JJ**	**VBN**	**RB**	**VBG**	**NNP**	**IN**	**VBZ**	**NNS**
NN	-	**556**	**1953**	52	86	276	-	8	-	-
VB	**2616**	-	**614**	-	42	-	77	15	-	5
JJ	**1577**	96	-	**1361**	**634**	**555**	**281**	30	-	16
VBN	-	-	-	-	-	-	-	-	-	-
RB	219	23	**408**	10	-	9	34	249	-	11
VBG	-	-	-	-	-	-	-	-	-	-
NNP	-	-	-	-	-	-	-	-	-	-
IN	-	-	-	-	-	-	-	-	-	-
VBZ	-	-	-	-	-	-	-	-	-	-
NNS	83	1	17	-	1	2	104	3	192	-

Porcentaje de aciertos: 99,16 %
Cantidad de errores: 13082

Se puede apreciar un alto porcentaje de aciertos entre las etiquetas extraídas de COBUILD (99,16 %) y las etiquetas asignadas por TnT. Este porcentaje indica que la información de etiquetas extraídas de COBUILD es consistente con las producidas por TnT. La mayoría de los errores se da en etiquetas VB, NN y JJ de COBUILD cuando son etiquetadas como NN, JJ y NN por TnT respectivamente.

0.1.2. Segundo experimento: entrenamiento de TnT con la nueva fuente de información generada

El segundo experimento realizado tiene como objetivo evaluar la nueva fuente de información obtenida (NFI) como corpus de entrenamiento. Para esto se utilizará el Wall Street Journal (WSJ), parte de Penn Tree Bank, como corpus objetivo.

La primer evaluación de este segundo experimento consiste en entrenar el etiquetador gramatical con WSJ como corpus de entrenamiento y con WSJ +

[1] Wall Street Journal es un corpus anotado, parte del Penn Treebank
[2] Las matrices de confusión presentadas de aquí en adelante contienen las primeras 10 etiquetas de mayor error

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 2: *WSJ original contra WSJ etiquetado con TnT (entrenado con WSJ)*

WSJ \ TnT(WSJ)	**JJ**	**NN**	**NNP**	**VBN**	**VBD**	**IN**	**RB**	**RP**	**NNPS**	**VBG**
JJ	-	909	717	**932**	51	71	569	23	1	229
NN	**2537**	-	**1680**	30	40	19	72	7	-	622
NNP	454	358	-	4	3	26	47	1	**1262**	7
VBN	**1215**	37	22	-	**1151**	-	-	-	-	-
VBD	60	40	6	**1522**	-	-	-	-	-	-
IN	109	4	28	-	-	-	**1429**	**1353**	-	2
RB	752	203	46	4	1	**1500**	-	842	-	1
RP	2	-	1	-	-	371	144	-	-	-
NNPS	34	-	418	-	-	-	-	-	-	-
VBG	411	879	19	-	-	-	-	-	-	-

Porcentaje de aciertos: 97,38 %

Cuadro 3: *WSJ original contra WSJ etiquetado con TnT (entrenado con WSJ + NFI)*

WSJ \ TnT(WSJ+NFI)	**JJ**	**NN**	**NNP**	**VBN**	**VBD**	**IN**	**RB**	**RP**	**NNPS**	**VB**
JJ	-	895	1032	908	53	83	620	28	1	77
NN	**2776**	-	**2087**	48	43	16	83	7	-	337
NNP	361	256	-	4	2	19	45	1	**1255**	11
VBN	**1326**	35	27	-	**1236**	-	-	-	-	43
VBD	69	24	10	**1760**	-	-	-	-	-	42
IN	107	4	46	-	-	-	**1163**	**1584**	-	3
RB	834	193	63	4	1	**1705**	-	**1138**	-	50
RP	2	1	1	-	-	348	88	-	-	-
NNPS	33	-	448	-	-	-	-	-	-	-
VB	89	408	34	48	42	13	11	-	-	-

Porcentaje de aciertos: 97,12 %

Se puede observar que el rendimiento del etiquetador TnT entrenado con WSJ es un poco mejor (97,38 %) que el rendimiento de TnT entrenado con WSJ + NFI (97,12 %). La mayoría de los errores para TnT entrenado con WSJ se da en etiquetas NN del gold standard cuando son etiquetadas como JJ y NNP por TnT. Para TnT 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 TnT 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 4: *Primer mitad WSJ original vs primer mitad WSJ etiquetado con TnT (entrenado con la segunda mitad de WSJ)*

TnT(2WSJ) / WSJ	**JJ**	**NN**	**NNP**	**VBN**	**VBD**	**IN**	**RB**	**VB**	**VBP**	**RP**
JJ	-	**1042**	547	**810**	54	20	312	61	6	6
NN	**1968**	-	**1167**	25	24	6	55	272	72	2
NNP	422	520	-	9	11	19	43	21	3	-
VBN	605	25	24	-	**831**	-	-	38	2	-
VBD	75	21	12	**1140**	-	-	1	33	8	-
IN	70	4	24	1	-	-	**617**	1	4	**616**
RB	437	91	26	3	2	**844**	-	19	3	316
VB	77	414	41	25	13	6	8	-	343	-
VBP	26	296	19	19	33	5	4	**623**	-	-
RP	-	-	2	-	-	230	122	-	-	-

Porcentaje de aciertos: 96,23 %

Cuadro 5: *Primer mitad WSJ original vs primer mitad WSJ etiquetado con TnT (entrenado con segunda mitad de WSJ + NFI)*

TnT(2WSJ+NFI) / WSJ	**JJ**	**NN**	**NNP**	**VBN**	**VBD**	**IN**	**RB**	**RP**	**VBG**	**NNPS**
JJ	-	**879**	**674**	**616**	29	37	337	7	210	-
NN	**1742**	-	**1293**	29	27	6	65	2	560	1
NNP	336	354	-	6	4	15	42	-	22	557
VBN	**636**	15	25	-	**716**	-	-	-	-	-
VBD	56	10	17	**1073**	-	-	-	-	-	-
IN	66	3	32	-	-	-	506	**750**	2	-
RB	448	79	31	2	1	**894**	-	521	-	-
RP	-	-	2	-	-	182	53	-	-	-
VBG	309	467	23	-	-	-	-	-	-	-
NNPS	22	1	521	-	-	-	-	-	-	-

Porcentaje de aciertos: 96,46 %

Cuadro 6: *Segunda mitad de WSJ original vs segunda mitad WSJ etiquetado con TnT (entrenado con la primera mitad de WSJ)*

TnT(1WSJ) / WSJ	**JJ**	**NN**	**VBN**	**VBD**	**NNP**	**RB**	**IN**	**NNPS**	**RP**	**VB**
JJ	-	**1084**	610	79	564	352	44	2	18	83
NN	**1828**	-	36	30	**1102**	51	13	-	6	263
VBN	**829**	22	-	**874**	30	-	-	-	-	34
VBD	70	38	**1107**	-	13	-	-	-	-	21
NNP	454	458	18	5	-	40	19	**850**	2	18
RB	384	164	3	-	47	-	**753**	-	512	21
IN	43	5	-	-	19	**851**	-	-	**693**	4
NNPS	11	-	-	-	351	-	-	-	-	-
RP	2	1	-	-	1	85	227	-	-	-
VB	76	410	27	21	50	10	7	-	-	-

Porcentaje de aciertos: 96,20 %

Cuadro 7: *Segunda mitad de WSJ original vs segunda mitad WSJ etiquetado con TnT (entrenado con primera mitad de WSJ)*

TnT(1WSJ+NFI) / WSJ	**JJ**	**NN**	**NNP**	**VBN**	**VBD**	**IN**	**RB**	**RP**	**NNPS**	**VB**
JJ	-	**842**	661	504	35	51	353	23	2	64
NN	**1808**	-	**1258**	34	30	10	40	4	-	217
NNP	334	318	-	14	8	11	27	1	**783**	18
VBN	**808**	21	26	-	**749**	-	-	-	-	26
VBD	46	16	13	**1072**	-	-	-	-	-	19
IN	43	3	23	-	-	-	652	**875**	-	3
RB	426	137	56	2	-	**878**	-	**692**	-	36
RP	2	1	-	-	-	188	45	-	-	-
NNPS	15	-	319	-	-	-	-	-	-	-
VB	58	272	37	24	29	7	8	-	-	-

Porcentaje de aciertos: 96,36 %

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 la primer y segunda mitad de WSJ etiquetado con TnT entrenado con la mitad restante

contra la misma mitad de WSJ etiquetado con TnT entrenado con la mitad restante + NFI.

Cuadro 8: *Primer mitad de WSJ etiquetado por TnT (entrenado con la segunda mitad WSJ) vs primer mitad WSJ etiquetado con TnT (entrenado con la segunda mitad de WSJ + NFI)*

TnT(2WSJ) \ TnT(2WSJ+NFI)	**NN**	**JJ**	**VBN**	**VBD**	**NNP**	**VBG**	**VBP**	**VB**	**RP**	**RB**
NN	-	**567**	32	19	**458**	**334**	108	**319**	-	43
JJ	**633**	-	143	39	**350**	63	12	59	2	162
VBN	15	**376**	-	**443**	7	-	5	8	-	-
VBD	10	36	**499**	-	13	-	5	8	-	-
NNP	159	142	1	2	-	6	3	18	-	5
VBG	175	216	-	1	36	-	-	1	-	2
VBP	94	2	-	7	2	-	-	233	-	3
VB	198	56	24	23	12	-	**326**	-	-	1
RP	-	-	-	-	-	-	-	-	-	6
RB	26	146	1	1	13	-	-	7	302	-

Porcentaje de aciertos: 98,39 %

Cuadro 9: *Segunda mitad WSJ etiquetado por TnT (entrenado con la primera mitad WSJ) vs segunda mitad WSJ etiquetado con TnT (entrenado con la primera mitad de WSJ + NFI)*

TnT(1WSJ) \ TnT(1WSJ+NFI)	**JJ**	**NN**	**VBN**	**VBD**	**NNP**	**RP**	**RB**	**IN**	**VB**	**VBP**
JJ	-	**510**	195	43	**319**	-	125	9	47	19
NN	**729**	-	27	25	**426**	1	41	4	**291**	69
VBN	**282**	20	-	**438**	4	-	1	-	11	5
VBD	56	12	**522**	-	3	-	1	-	8	3
NNP	117	133	5	3	-	-	10	9	25	4
RP	-	2	-	-	1	-	35	32	-	-
RB	160	35	-	-	35	**353**	-	**300**	16	2
IN	2	2	-	-	14	159	73	-	-	1
VB	43	183	18	17	17	1	4	1	-	275
VBP	16	102	3	7	1	-	1	1	201	-

Porcentaje de aciertos: 98,45 %

La tercer evaluación de este experimento consiste en entrenar TnT 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 10: *Rendimiento de TnT entrenado con cuartos de WSJ con y sin NFI*

Evaluación	**Porcentaje de aciertos**
TnT entrenado con el primer 1/4 de WSJ	95.92 %
TnT entrenado con el primer 1/4 de WSJ + NFI	96.25 %
TnT entrenado con el segundo 1/4 de WSJ	95.88 %
TnT entrenado con el segundo 1/4 de WSJ + NFI	96.25 %
TnT entrenado con el tercer 1/4 de WSJ	95.90 %
TnT entrenado con el tercer 1/4 de WSJ + NFI	96.28 %
TnT entrenado con el cuarto 1/4 de WSJ	95.89 %
TnT entrenado con el cuarto 1/4 de WSJ + NFI	96.29 %

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 TnT 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:

- 95.31 % de acierto de etiquetas para el etiquetado de 9/10 de WSJ con TnT entrenado con 1/10 WSJ
- 96.09 % de acierto de etiquetas para el etiquetado de 9/10 de WSJ con TnT entrenado con 1/10 WSJ+NFI

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

0.1.3. Tercer experimento