

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 confusión² generada a partir de dicha comparación es la siguiente:

Cuadro 1: Matriz de confusión para etiquetas extraídas de COBUILD vs WSJ

WSJ COBUILD	VBD	VBN	VBP	VB	NN	JJ	VBZ	NNS	CC	NNP
VBD	-	.1145	.0009	.0009	.0003	.0094	-	-	-	.0001
VBN	.0023	-	-	-	-	.0008	-	-	-	-
VBP	.0298	.0134	-	.0978	.0514	.0141	.0005	.0001	-	-
VB	.0020	.0022	.0113	-	.0261	.0042	.0000	.0001	-	.0022
NN	.0020	.0029	.0068	.0193	-	.0591	-	.0114	.0000	.0371
JJ	.0042	.0525	.0004	.0033	.0463	-	-	.0005	.0001	.0085
VBZ	-	-	.0022	.0042	.0017	.0004	-	.0520	-	.0001
NNS	-	-	-	.0001	.0020	.0005	.0068	-	.0000	.0029
CC	.0022	.0023	.0082	.0077	.0433	.0258	.0000	.0005	-	.0049
NNP	-	-	-	-	.0004	.0000	-	-	-	-

Porcentaje de aciertos: 98,09 %

Donde las etiquetas de las filas representan las etiquetas asignadas por TnT mientras que las etiquetas de las columnas representan las etiquetas extraídas de COBUILD. Se puede apreciar un alto porcentaje de aciertos entre las etiquetas extraídas de COBUILD (98,09 %) 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 VBN y VB de COBUILD cuando son etiquetadas como VBD y VBP 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 + NFI. Luego se procede a etiquetar el WSJ plano (sin etiquetas gramaticales) con estos dos modelos. Por último se contruye la matriz de confusión:

¹Wall Street Journal es un corpus anotado, parte del Penn Treebank

²Las matrices de confusión presentadas de aquí en adelante contienen las primeras 15 etiquetas de mayor error

Cuadro 2: Matriz de confusión para WSJ etiquetado con TnT (entrenado con WSJ)

	JJ	NN	NNP	VCN	VBD	IN	RB	RP	NNPS	VBG
JJ	-	.0766	.0137	.0367	.0018	.0033	.0227	.0001	.0010	.0124
NN	.0274	-	.0108	.0011	.0012	.0001	.0061	-	-	.0265
NNP	.0216	.0507	-	.0007	.0002	.0008	.0014	.0000	.0126	.0006
VCN	.0281	.0009	.0001	-	.0459	-	.0001	-	-	-
VBD	.0015	.0012	.0001	.0347	-	-	.0000	-	-	-
IN	.0021	.0006	.0008	-	-	-	.0453	.0112	-	-
RB	.0172	.0022	.0014	-	-	.0431	-	.0043	-	-
RP	.0007	.0002	.0000	-	-	.0408	.0254	-	-	-
NNPS	.0000	-	.0381	-	-	-	-	-	-	-
VBG	.0069	.0188	.0002	-	-	.0001	.0000	-	-	-

Porcentaje de aciertos: 97,38 %

Cuadro 3: Matriz de confusión para WSJ etiquetado con TnT (entrenado con WSJ + NFI)

	JJ	NN	NNP	IN	RB	RP	VCN	VBD	NNPS	VBG
JJ	-	.0752	.0096	.0030	.0226	.0001	.0362	.0018	.0009	.0136
NN	.0254	-	.0084	.0001	.0052	.0000	.0011	.0008	-	.0223
NNP	.0287	.0559	-	.0013	.0015	.0000	.0007	.0003	.0120	.0008
IN	.0024	.0005	.0005	-	.0471	.0097	-	-	-	.0001
RB	.0173	.0022	.0012	.0308	-	.0022	-	-	-	-
RP	.0008	.0002	.0000	.0425	.0308	-	-	-	-	-
VCN	.0240	.0011	.0001	-	.0001	-	-	.0425	-	-
VBD	.0017	.0011	.0001	-	.0000	-	.0388	-	-	-
NNPS	.0000	-	.0342	-	-	-	-	-	-	-
VBG	.0060	.0225	.0001	.0001	.0000	-	-	-	-	-

Porcentaje de aciertos: 97,07 %

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 (97,07 %). La mayoría de los errores para TnT entrenado con WSJ se da en etiquetas JJ y NNP del gold standard cuando son etiquetadas como NN por TnT. Para TnT entrenado con WSJ + NFI la mayoría de los errores se da en las mismas etiquetas, pero con porcentaje de errores menor para JJ etiquetado como NN.

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: Matriz de confusión para la 2 mitad de WSJ etiquetado con TnT (entrenado con la primer mitad de WSJ)

	JJ	NN	NNP	VDN	VBD	IN	RB	VB	VBP	RP
JJ	-	.0827	.0177	.0254	.0032	.0029	.0184	.0032	.0011	-
NN	.0438	-	.0219	.0011	.0009	.0002	.0038	.0174	.0124	-
NNP	.0230	.0490	-	.0010	.0005	.0010	.0011	.0017	.0008	.0001
VDN	.0340	.0011	.0004	-	.0479	.0000	.0001	.0011	.0008	-
VBD	.0023	.0010	.0005	.0349	-	-	.0001	.0005	.0014	-
IN	.0008	.0003	.0008	-	-	-	.0355	.0003	.0002	.0097
RB	.0131	.0023	.0018	-	.0000	.0259	-	.0003	.0002	.0051
VB	.0026	.0114	.0009	.0016	.0014	.0000	.0008	-	.0262	-
VBP	.0003	.0030	.0001	.0001	.0003	.0002	.0001	.0144	-	-
RP	.0003	.0001	-	-	-	.0259	.0133	-	-	-

Porcentaje de aciertos: 96,23 %

Cuadro 5: Matriz de confusión para la 2 mitad de WSJ etiquetado con TnT (entrenado con la primer mitad de WSJ + NFI)

	JJ	NN	NNP	VBD	VDN	IN	RB	RP	VBG	NNPS
JJ	-	.0747	.0142	.0022	.0272	.0030	.0193	-	.0131	.0011
NN	.0389	-	.0238	.0006	.0009	.0001	.0033	-	.0207	.0001
NNP	.0302	.0537	-	.0007	.0014	.0014	.0012	.0001	.0011	.0220
VBD	.0026	.0010	.0002	-	.0435	-	.0000	-	-	-
VDN	.0253	.0011	.0004	.0394	-	-	.0001	-	-	-
IN	.0017	.0005	.0008	-	-	-	.0393	.0080	.0003	-
RB	.0146	.0034	.0019	-	-	.0209	-	.0023	-	-
RP	.0003	.0001	-	-	-	.0322	.0225	-	-	-
VBG	.0092	.0254	.0009	-	-	.0001	-	-	-	-
NNPS	-	.0000	.0236	-	-	-	-	-	-	-

Porcentaje de aciertos: 96,31 %

Cuadro 6: Matriz de confusión para la 1 mitad de WSJ etiquetado con TnT (entrenado con la 2 mitad de WSJ)

	JJ	NN	VDN	VBD	NNP	RB	IN	NNPS	RP	VB
JJ	-	.0761	.0345	.0029	.0189	.0160	.0018	.0005	.0001	.0032
NN	.0451	-	.0009	.0016	.0191	.0068	.0002	-	.0000	.0171
VDN	.0254	.0015	-	.0461	.0007	.0001	-	-	-	.0011
VBD	.0033	.0012	.0364	-	.0002	-	-	-	-	.0009
NNP	.0235	.0459	.0012	.0005	-	.0020	.0008	.0146	.0000	.0021
RB	.0147	.0021	-	-	.0017	-	.0354	-	.0035	.0004
IN	.0018	.0005	-	-	.0008	.0314	-	-	.0095	.0003

Cuadro 6: Matriz de confusión para la 1 mitad de WSJ etiquetado con TnT (entrenado con la 2 mitad de WSJ)

	JJ	NN	VDN	VBD	NNP	RB	IN	NNPS	RP	VB
NNPS	.0001	-	-	-	.0354	-	-	-	-	-
RP	.0007	.0002	-	-	.0001	.0213	.0289	-	-	-
VB	.0035	.0110	.0014	.0009	.0007	.0009	.0002	-	-	-

Porcentaje de aciertos: 96,20 %

Cuadro 7: Matriz de confusión para la 1 mitad de WSJ etiquetado con TnT (entrenado con la 2 mitad de WSJ + NFI)

	JJ	NN	NNP	VBD	VDN	IN	RB	RP	NNPS	VBG
JJ	-	.0751	.0140	.0018	.0347	.0018	.0178	.0001	.0006	.0128
NN	.0372	-	.0218	.0007	.0010	.0001	.0057	.0000	-	.0194
NNP	.0285	.0520	-	.0005	.0011	.0010	.0021	-	.0131	.0014
VBD	.0024	.0012	.0004	-	.0438	-	-	-	-	-
VDN	.0204	.0013	.0005	.0373	-	-	.0001	-	-	-
IN	.0022	.0005	.0005	-	-	-	.0376	.0080	-	.0003
RB	.0150	.0017	.0010	-	-	.0259	-	.0018	-	.0000
RP	.0010	.0002	.0000	-	-	.0373	.0309	-	-	-
NNPS	.0001	-	.0327	-	-	-	-	-	-	-
VBG	.0058	.0233	.0005	-	-	.0000	.0000	-	-	-

Porcentaje de aciertos: 96,22 %

Se puede apreciar una leve mejoría en el porcentaje de etiquetas acertadas; 96,23 % contra 96,31 % y 96,20 % contra 96,22 % para cada mitad respectivamente. Los errores más comunes son producidos en etiquetas JJ y NNP del gold standard cuando son etiquetadas como NN 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.

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 8: 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	96.06 %

Cuadro 8: Rendimiento de TnT entrenado con cuartos de WSJ con y sin NFI

Evaluación	Porcentaje de aciertos
TnT entrenado con el segundo 1/4 de WSJ	95.88 %
TnT entrenado con el segundo 1/4 de WSJ + NFI	96.06 %
TnT entrenado con el tercer 1/4 de WSJ	95.90 %
TnT entrenado con el tercer 1/4 de WSJ + NFI	96.08 %
TnT entrenado con el cuarto 1/4 de WSJ	95.89 %
TnT entrenado con el cuarto 1/4 de WSJ + NFI	96.09 %

En todos los casos se puede apreciar una mejoría en el acierto de etiquetas para el corpus de entrenamiento WSJ + NFI contra WSJ de alrededor del 18 %

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
- 95.81 % 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 de .50 % en el corpus de entrenamiento que incorpora NFI.

0.1.3. Tercer experimento