Capítulo 1

Desarrollo

Las entradas que conforman el diccionario Cobuild y que constituyen el conjunto de datos principal sobre el cual se basa este trabajo fueron cuidadosamente procesadas y refinadas intentando mantener toda la información disponible, explícita e implícita.

Se desarrollaron algoritmos para extraer de estas entradas los ejemplos junto a sus rótulos asociados. Se tradujeron las etiquetas gramaticales provistas por Cobuild en etiquetas gramaticales standard. Se analizaron todos estos procesos y se ajustaron hasta aplacar sus fallas.

El resultado obtenido fué un corpus parcialmente anotado correspondiente a la concatenación de los ejemplos extraídos.

Luego se completó la anotación utilizando etiquetación automática, obteniendo como resultado final un corpus completamente anotado.

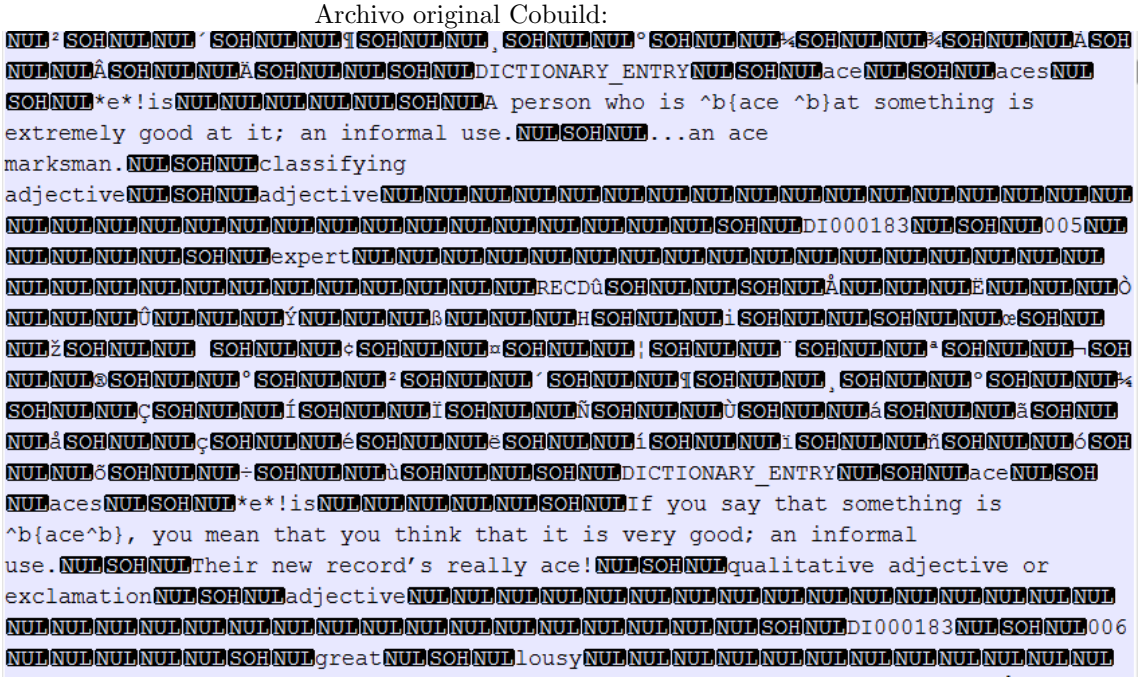
1.1. Extracción de la información

Cobuild guarda su información en un archivo de texto difícilmente legible con un formato carente de documentación conocida. El primer desafío de este trabajo consistió en identificar y obtener las entradas del archivo mencionado.

En ese sentido se crearon los algoritmos de extracción necesarios que consistieron en la eliminación de caracteres no ascii para poder obtener un archivo legible y en identificar y separar cada entrada.

A continuación se presentan extractos a modo de ejemplo:

Archivo original Cobuild:



Entradas extraídas correspondientes al fragmento anterior:

```
DICTIONARY_ENTRY
ace
aces
*e*!is
A person who is ace at something is extremely good at it; an informal use.
...an ace marksman.
classifying adjective
adjective

DICTIONARY_ENTRY
ace
aces
*e*!is
If you say that something is ace, you mean that you think that it is very good;
an informal use.
Their new records really ace!
qualitative adjective or exclamation
adjective
```

Las entradas de *Cobuild* se caracterizan por poseer una cantidad variable de campos difícilmente identificables. Sin embargo contienen algunos rasgos comu-

nes: la palabra, sus formas, la pronunciación, su definición y uno o más ejemplos donde se indica como se emplea (mediante una etiqueta gramatical).

Por ejemplo, en la primer entrada se pueden distinguir estos campos:

`DICTIONARY_ENTRY`
`ace` ⟶ *palabra*
`aces` ⟶ *formas flexionadas*
`*e*!is` ⟶ *pronunciación*
`A person who is ace at something is extremely good at it; an informal use.` ⟶ *definición*
`...an ace marksman.` ⟶ *ejemplo*
`classifying adjective` ⟶ *etiqueta específica*
`adjective` ⟶ *etiqueta general*

Se procesó cada entrada identificando la palabra que se está definiendo, las formas flexionadas de la misma y los ejemplos junto a su etiqueta gramatical asociada.

Como consecuencia se determinaron algunas características particulares para las entradas de Cobuild.

1. Algunas entradas presentan la pronunciación mientras que otras presentan detalles de la misma descriptos en lenguaje natural. Ejemplo:

```
DICTIONARY_ENTRY
abstract
abstracts, abstracting, abstracted
An idea, argument, or way of thinking that is abstract is based on general
ideas and principles rather than on particular things and events.
The arguments of contemporary science are so abstract that they are no longer intelligible...
...our capacity for abstract reasoning.
qualitative adjective
adjective
The word abstract is pronounced /*!abstr!akt/ when it is an adjective or a
noun, and /%e3bstr*!akt/ when it is a verb.
```

2. En la mayoría de los casos las palabras se definen con una oración, sin embargo existen entradas que presentan definiciones utilizando más de una oración como se muestra abajo:

```
DICTIONARY_ENTRY
account
accounts, accounting, accounted
%ek*a*!unt
The word account is also used in the following expressions. If you say that
something is the case by all accounts or from all accounts, you mean that everyone
you talk to about it, or everyone who writes about it, says that it is so.
From all accounts she was a clever girl.
phrase: used as an adjunct
phrase
```

Para el caso 1, las frases explicativas sobre pronunciación de las palabras fueron identificadas y descartadas, para no confundirlas con ejemplos.

Ante la imposibilidad de distinguir si una oración es parte de la definición o es parte de un ejemplo (caso 2), se asumió que la primera oración de la entrada corresponde a la definición (esto es cierto en la mayoría de los casos).

De esta manera se evita la pérdida de ejemplos por confundirlos con la definición. No obstante puede introducirse falsa información al identificar una oración perteneciente a la definición como un ejemplo.

Sin embargo este hecho no es tan grave: debido a las características de Cobuild, la palabra es generalmente definida utilizando el mismo sentido que exhibe el ejemplo.

Por lo tanto se puede concluir que inclusive para los pocos casos en que las oraciones pertenecientes a una definición se identifican como un ejemplo, éstas aportan información gramatical válida.

1.2. Traducción de etiquetas

La información gramatical que Cobuild presenta en cada uno de sus ejemplos no posee un formato conocido ni pertenece a ningún conjunto de etiquetas documentado.

Por ejemplo en la siguiente entrada:

`DICTIONARY_ENTRY`
`acid` ⟶ *palabra*
`acids` ⟶ *formas flexionadas*
`*!as!id` ⟶ *pronunciación*
`An acid fruit or drink has a sour or sharp taste.` ⟶ *definición*
`These oranges are very acid.` ⟶ *ejemplo*
`qualitative adjective` ⟶ *etiqueta específica*
`adjective` ⟶ *etiqueta general*

Podemos apreciar que la palabra *acid* está anotada como *qualitative adjective*. Notemos también la presencia de la anotación *adjective*. Ésta es una etiqueta general mientras que *qualitative adjective* es una etiqueta específica que brinda mayor información gramatical y sintáctica.

Como la idea de este trabajo es producir un corpus anotado a partir de este diccionario para utilizar como fuente de entrenamiento de etiquetadores gramaticales, es necesario que el conjunto de etiquetas empleado sea el mismo que emplea el *Gold Standard* para posteriormente poder medir los resultados. En ese sentido se realizó la traducción de etiquetas *Cobuild* en *Penn Treebank*.

Para llevar a cabo este proceso se construyó una tabla de conversión. Se realizó un análisis sobre las entradas de *Cobuild* arrojando como resultado la existencia de más de 4000 etiquetas diferentes.

A partir de este hecho se identificaron las etiquetas *Cobuild* que ocurren con mayor frecuencia y se seleccionó la etiqueta *Penn Treebank* equivalente para cada una de ellas, obteniendo una tabla de traducción ad hoc que se presenta a continuación:

Cuadro 1.1: *Tabla de traducción de etiquetas*

Etiqueta Cobuild	**Etiqueta Penn Treebank**
coordinating conjunction	CC
number	CD
determiner	DT
determiner + countable noun in singular	DT

Cuadro 1.1: *Tabla de traducción de etiquetas*

Etiqueta Cobuild	**Etiqueta Penn Treebank**
preposition	IN
subordinating conjunction	IN
preposition, or adverb after verb	IN
preposition after noun	IN
adjective	JJ
classifying adjective	JJ
qualitative adjective	JJ
adjective colour	JJ
ordinal	JJ
adjective after noun	JJ
modal	MD
adverb	RB
noun	NN
uncountable noun	NN
noun singular	NN
countable or uncountable noun	NN
countable noun with supporter	NN
uncountable or countable noun	NN
noun singular with determiner	NN
mass noun	NN
uncountable noun with supporter	NN
partitive noun	NN
noun singular with determiner with supporter	NN
countable noun + of	NN
countable noun, or by + noun	NN
countable noun or partitive noun	NN
count or uncountable noun	NN
countable noun or vocative	NN
partitive noun + uncountable noun	NN
noun singular with determiner + of	NN
noun in titles	NN
noun vocative	NN
uncountable noun + of	NN
indefinite pronoun	NN
uncountable noun, or noun singular	NN
countable noun, or in + noun	NN
partitive noun + noun in plural	NN
countable or uncountable noun with supporter	NN
uncountable noun, or noun before noun	NN
uncountable or countable noun with supporter	NN
noun before noun	NN
noun plural with supporter	NNP
noun in names	NNP
proper noun or vocative	NNP
proper noun	NNP
noun plural	NNS
predeterminer	PDT

Cuadro 1.1: *Tabla de traducción de etiquetas*

Etiqueta Cobuild	**Etiqueta Penn Treebank**
pronoun	PP
possessive	PPS
adverb with verb	RB
adverb after verb	RB
sentence adverb	RB
adverb + adjective or adverb	RB
adverb + adjective	RB
preposition or adverb	RB
adverb after verb, or classifying adjective	RB
adverb or sentence adverb	RB
adverb with verb, or sentence adverb	RB
exclamation	UH
exclam	UH
verb	VB
verb + object	VB
verb or verb + object	VB
ergative verb	VB
verb + adjunct	VB
verb + object + adjunct	VB
verb + object (noun group or reflexive)	VB
verb + object or reporting clause	VB
verb + object (reflexive)	VB
verb + object, or phrasal verb	VB
verb + to-infinitive	VB
ergative verb + adjunct	VB
verb + object + adjunct (to)	VB
verb + object, or verb + adjunct	VB
verb + object + adjunct (with)	VB
verb + adjunct (with)	VB
verb + complement	VB
verb + object, or verb	VB
verb + object + to-infinitive	VB
verb + reporting clause	VB
verb or ergative verb	VB
verb + adjunct (from)	VB
wh: used as determiner	WDT
wh: used as relative pronoun	WP
wh: used as pronoun	WP
wh: used as adverb	WRB

El proceso de traducción de etiquetas consiste en intentar encontrar en la tabla la etiqueta *Penn Treebank* correspondiente a la etiqueta específica de *Cobuild* (en el ejemplo anterior*qualitative adjective*), si no fuera el caso se busca la etiqueta *Penn Treebank* correspondiente a la etiqueta general *Cobuild*(*adjective* para el ejemplo).

Utilizando este método se logró traducir aproximadamente el 99.26 % de las etiquetas.

Finalizado este proceso se verificó que el etiquetado haya sido correcto: se realizó una rutina que etiquetara automáticamente todos los ejemplos de Cobuild (utilizando el etiquetador TnT) y se generó una matriz de confusión comparando las etiquetas extraídas y traducidas provenientes de Cobuild contra las asignadas por TnT.

El resultado fué de 71 % de aciertos. Se analizaron las etiquetas que diferían, luego se corrigieron las traducciones y se ajustaron los algoritmos de extracción y traducción de etiqueteas hasta alcanzar un grado de error mínimo.

Los mayores focos de error que no pudieron ser aplacados corresponden a palabras etiquetadas como VBD cuando son VBN y viceversa. Estas etiquetas son difícilmente desambigüables automáticamente.

1.2.1. Recuperación de precisión gramatical

Una vez obtenidos los ejemplos con las etiquetas traducidas a Penn Treebank, se realiza un nuevo proceso para aportar o recuperar precisión gramatical perdida en la traducción: se comparan las etiquetas traducidas contra las etiquetas asignadas automáticamente por TnT. En caso de coincidir y en caso de que la etiqueta TnT aporte mayor precisión, se asigna esta última. Se utilizó el siguiente algoritmo:

Algoritmo 1 Obtener la etiqueta de mayor detalle gramatical

Entrada: Etiqueta obtenida de Cobuild, etiqueta asignada por TnT

Si se obtuvo de Cobuild la etiqueta:

NN y TnT asignó **NNS, NNP o NNPS**: Asignar la etiqueta TnT

NNS y TnT asignó **NNPS**: Asignar la etiqueta NNPS

VB y TnT asignó **VBN, VBD, VBZ, VBP o VBG**: Asignar la etiqueta TnT

JJ y TnT asignó **JJR o JJS**: Asignar la etiqueta TnT

RB y TnT asignó **RBR o RBS**: Asignar la etiqueta TnT

WP y TnT asignó **WP$**: Asignar la etiqueta WP$

PRP y TnT asignó **PRP$**: Asignar la etiqueta PRP$

A continuación se exhiben ejemplos de este proceso:

Etiqueta extraída de Cobuild	
It	
cannot	
produce	
enough	
heat	
to	
activate	***VB***
the	
electrons	
.	

$\longrightarrow$

Etiquetado automático	
It	
cannot	
produce	
enough	
heat	
to	
activate	***VBP***
the	
electrons	
.	

Etiqueta extraída de Cobuild	
Armed	
with	
this	
information	
,	
parents	
will	
be	
better	***RB***
able	
to	
cater	
for	
their	
childrens	
needs	
...	

$\longrightarrow$

Etiquetado automático	
Armed	
with	
this	
information	
,	
parents	
will	
be	
better	***RBR***
able	
to	
cater	
for	
their	
childrens	
needs	
...	

Etiqueta extraída de Cobuild	
They	
have	
their	
first	
degrees	
and	
are	
studying	
for	
higher	***JJ***
degrees	
...	

$\longrightarrow$

Etiquetado automático	
They	
have	
their	
first	
degrees	
and	
are	
studying	
for	
higher	***JJR***
degrees	
...	

En los ejemplos presentados se puede observar como el etiquetado automático aporta información gramatical a las palabras anotadas durante el proceso de extracción y traducción.

Luego de ejecutado este proceso el análisis de comparación contra las etiquetas asignadas por TnT dió un 86 % de aciertos .

Cabe destacar que más allá de los esfuerzos por preservar la información gramatical, inevitablemente se generó una pérdida de información semántica durante el proceso de traducción ya que las etiquetas *Penn Treebank* son menos específicas que las etiquetas de Cobuild. En otras palabras, las etiquetas *Penn Treebank* carecen del rico detalle lingüístico que las etiquetas Cobuild poseen, originando una pérdida natural de información en la traducción.

1.2.2. Reconocimiento de formas flexionadas

Las entradas de Cobuild exponen formas flexionadas de la palabra: plurales, pasados, etc. En muchas de ellas ocurre la palabra que se está definiendo y uno o más ejemplos en donde aparecen formas flexionadas de la misma junto con sus rótulos.

El objetivo entonces es reconocer y registrar esta información gramatical implícita.

Tomemos la siguiente entrada:

`DICTIONARY_ENTRY`
`celebrate` ⟶ *palabra*
`celebrates, celebrating, celebrated` ⟶ *formas flexionadas*
`s*!el%e1bre!it` ⟶ *pronunciación*
`If you celebrate someone or something, you praise them for their good qualities; a fairly formal use.` ⟶ *definición*
`People were celebrating him as a bright alternative to Nixon.` ⟶ *ejemplo*
`verb + object, or verb + object + adjunct (as/for)` ⟶ *etiqueta específica*
`verb` ⟶ *etiqueta general*

Aquí arriba se puede observar una entrada del diccionario para la palabra *celebrate*, que contiene la definición y un ejemplo en donde aparece la forma derivada *celebrating* junto a una etiqueta gramatical asociada:

> *People were **celebrating** him as a bright alternative to Nixon.*

En la entrada presentada la palabra definida es *celebrate* y las formas derivadas son *celebrates, celebrating* y *celebrated*. Con esta información y la etiqueta asignada por *Cobuild* (*verb + object, or verb + object + adjunct (as/for)*) se pueden inferir y generar etiquetas *Penn Treebank* para dichas formas.

En lugar de guardar la etiqueta *Penn Treebank* VB correspondiente a *verb* para la palabra *celebrating*, guardaríamos la etiqueta VBG[1] que contiene más información gramatical.

La tarea aquí será reconocer *celebrating* como verbo gerundio o presente participio a partir de que está etiquetada como verbo y que deriva de *celebrate*. Es decir, inferir el tipo de la forma derivada a partir de la palabra y la etiqueta asignada por Cobuild.

En este sentido se desarrollaron reglas y métodos para aprovechar toda la información presente. Entonces, a partir de la palabra, la forma en que ocurre y la etiqueta asignada se aplican las siguientes reglas para reconocer etiquetas gramaticales para formas flexionadas:

[1] verbo gerundio o presente participio

Algoritmo 2 Reconocimiento de formas flexionadas

Entrada: Etiqueta asignada por Cobuild, forma flexionada

Traducir la etiqueta asignada por Cobuild a PenTreeBank
Si la etiqueta obtenida es

JJ:
Si la forma termina en *er* o empieza en *more* o *less* aplicar **JJR**
Si la forma termina en *est* o empieza en *most* o *least* aplicar **JJS**

RB:
Si la forma termina en *er* o empieza en *more* o *less* aplicar **RBR**
Si la forma termina en *est* o empieza en *most* o *least* aplicar **RBS**

NN:
Si la forma termina en *s* aplicar **NNS**

VB:
Si la forma termina en *ed* aplicar **VBD|VBN**
Si la forma termina en *ing* aplicar **VBG**
Si la forma termina en *s* aplicar **VBZ**

A continuación se presentan algunos ejemplos del resultado de aplicar este proceso:

1. Forma derivada *celebrating* inferida como VBG a partir de la palabra *celebrate* y la etiqueta VB:

```
DICTIONARY_ENTRY
celebrate
celebrates, celebrating, celebrated
s*!el%e1bre!it
If you celebrate someone or something, you praise them for their good qualities;
a fairly formal use.
People were celebrating him as a bright alternative to Nixon.
verb + object, or verb + object + adjunct (as/for)
verb
```

Resultado: *People were celebrating/**VBG** him as a bright alternative to Nixon.*

2. Forma derivada *faults* inferida como NNS a partir de la palabra *fault* y la etiqueta NN:

```
DICTIONARY_ENTRY
fault
faults, faulting, faulted
f*!o*:lt
A fault on a machine or in a structure is a broken part or a mistake in
the way it was made.
Send it back to the manufacturer if the machine develops the same fault...
Technicians laboriously tried to find and remedy faults.
countable noun
noun
```

Resultado: *Technicians laboriously tried to find and remedy faults/**NNS**.*

3. Forma derivada *larger* inferida como JJR a partir de la palabra *large* y la etiqueta JJ:

```
DICTIONARY_ENTRY
large
larger, largest
l*%a*:d!z
If you say that someone or something is larger than life, you mean that they
appear or behave in a way that seems more important or exaggerated than usual.
The central character is a larger than life, cantankerous New Englander...
...a larger-than-life version of our present society.
classifying adjective
adjective
```

Resultado: *The central character is a larger/**JJR** than life, cantankerous New Englander...*

1.3. Nuevo Corpus generado

A partir del corpus parcialmente anotado generado en la etapa anterior, se completarán las anotaciones con un etiquetador automático (TnT) preservando las etiquetas obtenidas a partir de la información gramatical proveniente del diccionario *Cobuild*.

A continuación se exhibe un ejemplo de este proceso:

Entrada *Cobuild* para la palabra *abide*

```
DICTIONARY_ENTRY
abide
abides, abiding, abided
%eb*a*!id
If something abides, it continues to happen or exist for a long time.
We feel the need to lean on something that abides.
verb
verb
```

Resultado de extracción, reconocimiento y traducción de etiquetas y formas flexionadas correspondiente a la entrada anterior:

We
feel
the
need
to
lean
on
something
that
abides *VBZ*
.

Se puede apreciar que en el ejemplo se ha reconocido *abides* como verbo en tercera persona a partir de *abide* y el rótulo *verb*, asignando la etiqueta gramatical traducida correspondiente: VBZ (obtenida por inferencia).

El próximo paso será el de completar las anotaciones gramaticales para las palabras restantes. Este proceso se realiza anotando el corpus con el etiquetador

gramatical automático TnT, como puede verse en 2). Luego se une el corpus anotado parcialmente procedente de Cobuild 1) con 2), preservando todas las etiquetas del diccionario.

El resultado es un nuevo corpus obtenido a partir de Cobuild, con las anotaciones que este provee y completado con anotaciones automáticas 3).