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.

Índice

1. Introducción

1.1. Motivación

El etiquetado o anotado gramatical, también conocido como Part-of-speech tagging, POS tagging o simplemente POST, es el proceso de asignar una etiqueta gramatical a cada una de las palabras de un texto según su categoría léxica. Por ejemplo tomemos la oración siguiente:

(1) *There is no asbestos in our products now.*

El resultado de etiquetarla gramaticalmente es:

(2) *There/EX is/VBZ no/DT asbestos/NN in/IN our/PRP products/NNS now/RB ./.*

donde cada palabra está sucedida por una barra oblicua seguida de la etiqueta gramataical asignada. Se puede apreciar por ejemplo que la palabra *is* fué etiquetada como VBZ (verbo de tiempo presente en tercera persona singular), *products* fué etiquetada como NNS (sustantivo plural), étc. Es decir que a cada palabra se le asignó un código que se corresponde con una función gramatical.

A simple vista el etiquetado gramatical parece una tarea trivial o al menos sencilla, sin embargo no es así. La complejidad de este proceso reside en la ambigüedad gramatical inherente al lenguaje. Por ejemplo, la palabra *premio* puede funcionar como sustantivo:

(1) *Gané un premio*

o como verbo

(2) *Por tu esfuerzo te premio*

En (1), *premio* tendría que recibir la etiqueta gramatical NN (sustantivo común) mientras que en (2) tendría que recibir la etiqueta gramatical VB (verbo). Ahora bien, sería interesante conocer que factor indica cual es la etiqueta correspondiente a una palabra ambigüa. En (1) se puede deducir que *premio* es sustantivo porque está precedido por la palabra *un* mientras que en (2) *premio* está precedido por la palabra *te* y a partir de este hecho se puede deducir que en esta caso *premio* funciona como verbo. En fin, las palabras circundantes brindan información vital para deducir el sentido gramatical en palabras ambigüas.

Como se menciona más adelante, el etiquetado gramatical juega un papel importante en áreas de la lingüistica computacional como síntesis del habla, reconocimiento del habla y recuperación de la información. El etiquetado gramatical es realizado manualmente por lingüistas; especialistas en el lenguaje que se ocupan de determinar una etiqueta gramatical para cada palabra. Desde luego que también es realizado automáticamente por computadoras, mediante programas conocidos como etiquetadores gramaticales. Algunas implementaciones actuales de estos programas están basadas en el aprendizaje; toman un corpus [1]

[1] Colección de textos escritos y/o transcripciones del lenguaje oral para cierto idioma

anotado correctamente con el cual se entrenan y luego emplean el conocimiento adquirido para etiquetar el corpus objetivo. En esa primer etapa conocida como entrenamiento, el etiquetador gramatical obtiene, procesa y retiene información sobre cada palabra, su etiqueta asignada y su contexto. Posteriormente el etiquetador determina una etiqueta para cada palabra del corpus objetivo, analizando su ubicación y contexto y utilizando el conocimiento adquirido en la etapa previa.

Uno de los grandes problemas del etiquetado gramatical reside en la falta de corpus anotados para utilizar como corpus de entrenamiento. Los corpus de entrenamiento son etiquetados manualmente por lingüistas especializados. Es un trabajo profundamente meticuloso y tedioso ya que el lingüista debe dar una etiqueta gramatical palabra por palabra en corpus del orden del millón de palabras. Además de la laboriosidad del trabajo, el tiempo empleado para etiquetar un corpus es sumamente extenso y como consecuencia el valor económico es significativo, ya que intervienen grupos de trabajo altamente capacitados durante períodos prolongados. El resultado de este complejo proceso es una tabla de palabras con su correspondiente etiqueta gramatical, como se muestra a continuación:

A	DT
form	NN
of	IN
asbestos	NN
once	RB
used	VBN
to	TO
make	VB
Kent	NNP
cigarette	NN
filters	NNS
has	VBZ
caused	VBN
a	DT
high	JJ
percentage	NN
of	IN
cancer	NN
deaths	NNS
among	IN
a	DT
group	NN
of	IN
workers	NNS
exposed	VBN
to	TO
it	PRP
more	RBR
than	IN
30	CD
years	NNS
ago	RB
,	,
researchers	NNS
reported	VBD
.	.

Ante la importancia que adquieren los corpora etiquetados es inevitable pensar en algún otro tipo de texto que posea información de etiquetas. Por ejemplo algunos diccionarios contienen una palabra, su definición y algunos ejemplos en donde ésta aparece con cada uno de sus sentidos. Es decir que de alguna manera un diccionario contiene por cada palabra uno o más contextos en donde ésta aparece etiquetada. Entonces si tomamos todos los ejemplos de cada palabra de un diccionario y su etiqueta podemos construir un corpus parcialmente anotado. Esta es la idea central de este trabajo.

1.2. Trabajo realizado

Como se mencionó en la sección anterior, la idea de este trabajo es suplir la falta de corpus de enternamiento utilizando la información de etiquetado que

posee un diccionario, generando una nueva fuente de información que servirá para entrenar etiquetadores automáticos. Este trabajo menciona detalladamente la forma de extraer la información relevante sobre etiquetas gramaticales a partir de un diccionario y las decisiones que fueron aplicadas. Esta nueva fuente de información se utiliza para entrenar etiquetadores gramaticales automáticos. Una vez entrenados dichos etiquetadores se emplean para etiquetar un corpus objetivo y se analiza el resultado obtenido. Se realiza el mismo procedimiento, pero ahora combinando la nueva fuente de información generada con un corpus de entrenamiento clásico. Se vuelve a etiquetar un corpus objetivo y se analiza el resultado obtenido. Por último se realizan mediciones sobre el rendimiento de los etiquetadores gramaticales entrenados con esta nueva fuente de información y con los corpora clásicos de entrenamiento y se presentan las conclusiones.

1.3. Etiquetado gramatical

Como se mencionó anteriormente, el etiquetado gramatical es el proceso de asignar una etiqueta a cada una de las palabras de un texto según su categoría léxica. Este proceso se realiza en base a la definición de la palabra y la de sus palabras vecinas, es decir, el contexto en que ésta aparece.

Por ejemplo en:

Does that flight serve dinner

dinner es un sustantivo y por lo tanto recibe la etiqueta para sustantivos NN.

El etiquetado gramatical brinda una gran cantidad de información sobre una palabra y sus vecinas. Por ejemplo, las etiquetas distinguen entre pronombres posesivos (mi, tu, su, étc.) y pronombres personales (Yo, Tú, Él, étc.). Saber si una palabra es un pronombre posesivo o personal nos brinda información sobre las palabras que pueden ocurrir a continuación: los pronombres posesivos generalmente son sucedidos por un sustantivo (como en *Mi comida*) mientras que los personales son sucedios por un verbo (como en *Yo duermo*).

Utilizando esta deducción podemos aseverar que si una palabra fué etiquetada como pronombre personal, es muy probable que la próxima palabra sea un verbo. Este conocimiento puede ser de útil aplicación en modelos linguísticos para reconocimiento del habla (voz a texto). Pero esta no es la única información que una etiqueta gramatical puede ofrecer.

Una etiqueta gramatical también nos puede acercar información relacionada con la pronunciación de la palabra. En inglés la palabra *content* puede ser un sustantivo o un adjetivo y su pronunciación varía dependiendo de este hecho. Utilizando estas ideas se pueden producir pronunciaciones más naturales en un sistema de síntesis del habla (texto a voz) o también se puede obtener más exactitud en un sistema de reconocimiento del habla (voz a texto).

Otra aplicación importante del etiquetado gramatical en sistemas de recuperación de la información es el reconocimiento de sustantivos u otro tipo de palabras importantes dentro de un documento, para guardar y utilizar esta información en búsquedas posteriores.

Por último, la asignación automática de etiquetas gramaticales juega un papel importante en algoritmos de desambigüación del sentido de la palabra y en

modelos lingüísticos basados en n-gramas utilizados en sistemas de reconocimiento del habla.

2. Definiciones y marco teórico

A continuación se presentan definiciones y teorías que ayudan a comprender el trabajo realizado. Se presenta el concepto de etiqueta gramatical, es decir, una etiqueta que identifica el rol que cumple una palabra dentro de cierto contexto. Se muestran los tipos de etiquetas que han sido utilizados intentando abarcar los distintos significados que pueden tener las palabras. Hasta el dia de hoy no se ha llegado a un consenso sobre un conjunto de etiquetas adecuado y se siguen explorando distintas alternativas. Se explica el concepto de etiquetado gramatical, es decir, la tarea de asignar a cada palabra una etiqueta gramatical adecuada según el contexto en donde ésta aparece. Se muestran ejemplos de que esta tarea está muy lejos de ser trivial, introduciendo el concepto de ambigüedad gramatical. Esto ocurre cuando una palabra puede tener muchos significados (y por lo tanto distintas etiquetas gramaticales) dependiendo del contexto en dónde aparece.

Se exhibe la importancia del etiquetado gramatical dentro de distintas áreas como la computación lingüistica, reconocimiento y síntesis del habla. Se muestra como se maneja este proceso utilizando programas que lo realizan automáticamente, es decir, etiquetadores gramaticales automáticos. Se explica en profundidad como funcionan estos etiquetadores gramaticales automáticos, mostrando como las implementaciones actuales utilizan un proceso de entrenamiento. Este proceso ocurre a partir de un corpus previamente anotado que el etiquetador automático toma como ejemplo para reproducir el etiquetado.

Se presenta el concepto de corpus y corpus anotados gramaticalmente como conjuntos de información extremadamente valiosos para todo estas tareas. Se muestra la forma de medir, evaluar y comparar el rendimiento de los etiquetadores gramaticales, introduciendo los conceptos de corpus de entrenamiento y corpus de verificación. Se muestran técnicas de análisis de error para el proceso de etiquetación automática. Se exhibe también el manejo de ciertos casos especiales dentro del proceso de etiquetación automático; las palabras desconocidas. Y por último se explican en detalle los etiquetadores automáticos utilizados en el presente trabajo.

2.1. Etiquetas

Tradicionalmente la deficinición de POS o etiqueta gramatical se ha basadado en funciones sintácticas y morfológicas, es decir que se agrupan en clases las palabras que funcionan similarmente con respecto a lo que puede ocurrir a su alrededor (sus propiedades de distribución sintáctica) o con respecto a los afijos que poseen (sus propiedades morfológicas). Mientras que las clases de palabras tienen tendencia hacia la coherencia semántica (por ejemplo los sustantivos generalmente describen gente, lugares o cosas y los adjetivos generalmente describen propiedades), este no es necesariamente el caso y en general no se utiliza coherencia semántica como criterio para la definición de POS o etiqueta gramatical.

Las etiquetas gramaticales pueden ser divididas en dos grandes categorías: clases cerradas y clases abiertas. Las clases cerradas son aquellas que tienen miembros relativamente fijos. Por ejemplo, las preposiciones son una clase cerrada porque hay un conjunto cerrado de ellas, es decir que son un grupo de palabras que raramente varía ya que raramente se agregan nuevas preposiciones. En contraste, los sustantivos y los verbos son clases abiertas ya que continuamente se introducen y eliminan nuevos verbos y sustantivos al lenguaje. Es probable que cualquier hablante o corpus tenga una clase abierta de palabras diferente, pero todos los hablantes de un lenguaje y corpora suficientemente grandes, seguramente van a compartir el conjunto de clases de palabras cerradas. Las clases de palabras cerradas también son generalmente palabras funcionales como *de*, *y* o *tu*, que tienden a ser muy cortas, ocurrir frecuentemente y generalmente tienen usos estructurales en gramática.

Hay cuatro clases abiertas principales:

- **Sustantivos** Es el nombre dado a la clase sintáctica que denota personas, lugares o cosas. Pero desde que las clases sintácticas como sustantivos son definidas sintáctica y morfológicamente en vez que semánticamente, algunas palabras para personas, lugares y cosas pueden no ser sustantivos y a la inversa, algunos sustantivos pueden no ser palabras para personas, lugares o cosas. Por lo tanto los sustantivos incluyen términos concretos como *barco* y *silla*, abstracciones como *banda ancha* y *relación*. Se puede definir a una palabra como sustantivo basándose en características como la capacidad de ocurrir con determinantes (una *cabra*, su *banda ancha*), tomar posesivos (los ingresos anuales de *IBM*) y para la mayoría pero no todos los sustantivos, ocurrir en la forma plural (*cabras*, *teléfonos*). Los sustantivos tradicionalmente son agrupados en sustantivos propios y sustantivos comunes.

  - **Sustantivos propios:** Son nombres de personas específicas o entidades y usualmente son escritos en mayúscula.
  - **Sustantivos comunes:** En algunos lenguajes se dividen en sustantivos contables e incontables.
    - **Sustantivos contables:** Son aquellos que permiten establecer su número en unidades. En general esta clase posee forma singular y plural (silla/s, dedo/s).
    - **Sustantivos incontables:** Se refieren a sustantivos que no se puede determinar su número en unidades (harina, nieve, azúcar).

- **Verbos:** Los verbos son una clase de palabras que incluye a la mayoría de las palabras referidas a acciones y procesos. Tienen ciertas formas morfológicas como tiempo, modo, persona, regularidad, étc. Además, el verbo puede concordar en género, persona y número con algunos de sus argumentos o complementos (a los que normalmente se conoce como sujeto, objeto, étc.). En español concuerda con el sujeto siempre en número y casi siempre en persona (la excepción es el caso del llamado sujeto inclusivo: *Los españoles somos así*).

  Algunos ejemplos:

  Marisol *canta* una ópera.

La comida *está* caliente.

- **Adjetivos:** Las palabras pertenecientes a esta clase expresan propiedades o cualidades. Por ejemplo *Ese hombre es* ***alto***. Los adjetivos tienen género y número al igual que los sustantivos. El género y el número de los adjetivos depende del sustantivo al que acompañan. Hay adjetivos que presentan una sola forma para el masculino y para el femenino. Son adjetivos de una sola terminación (verde, especial, amable, grande, étc.). Por el otro lado, los adjetivos de dos terminaciones presentan distintas formas para el masculino y el femenino (feo-fea, pequeño-pequeña, blanco-blanca, étc.) Se clasifican en:

  - **Determinativos:** Preceden al sustantivo, lo concretan y lo presentan
    - **Demostrativos:** *Esta* niña
    - **Posesivos:** *Mi* niña
    - **Numerales:** *Tres* niñas
    - **Indefinidos:** *Algunas* niñas
    - **Exclamativos:** ¡*Qué* niña!
    - **Interrogativos:** ¿*Qué* niña?
  - **Calificativos:** Califican al sustantivo, es decir, añaden cualidades al sustantivo. Los adejtivos calificativos se dividen en especificativos y explicativos o epítetos.
    - **Adjetivos calificativos especificativos:** Son aquellos que concretan el significado del sustantivo. Suelen aparecer detrás del sustantivo.
      Ej: Quiero una corbata *azul*.
    - **Adjetivos calificativos explicativos o epítetos:** indican cualidades que ya de por sí lleva el sustantivo. Suelen ir delante del sustantivo.
      Ej: *Blanca* nieve, *Verde* hierba.

- **Adverbios:** Los adverbios son otro ejemplo de clase abierta de palabras: se definen como modificadores del verbo, adjetivo o de otro adverbio. Tradicionalmente se dividen en:

  - **Adverbios de lugar**: aquí, allí, ahí, allá, acá, arriba, abajo, cerca, lejos, delante, detrás, encima, debajo, enfrente, atrás, alrededor, etc.
  - **Adverbios de tiempo absoluto**: pronto, tarde, temprano, todavía, aún, ya, ayer, hoy, mañana, siempre, nunca, jamás, próximamente, prontamente, anoche, enseguida, ahora, mientras.
  - **Adverbios de modo**: bien, mal, regular, despacio, deprisa, así, tal, como, aprisa, adrede, peor, mejor, fielmente, estupendamente, fácilmente - todas las que se formen con las terminaciones "mente".
  - **Adverbios de cantidad o grado**: muy, poco, muy poco, mucho, bastante, más, menos, algo, demasiado, casi, sólo, solamente, tan, tanto, todo, nada, aproximadamente.

Por otro lado tenemos las clases cerradas de palabras que detallamos a continuación:

- **Preposiciones:** Las preposiciones son enlaces que relacionan los componentes de una oración para brindarles sentido. La unión se lleva a cabo con una o varias palabras. La significación que dan las preposiciones responde a circunstancias de movimiento, lugar, tiempo, modo, causa, posesión, pertenencia, materia y procedencia.
  Algunos ejemplos:

  Me levanté de la cama ***a*** *las ocho de la mañana.*
  Dejé mis cuadernos ***sobre*** *el sillón.*
  Corrí apresurado ***hacia*** *la calle pero no logré divisarte.*
  Lucía se divierte ***con*** *sus muñecas.*

- **Determinantes:** Los determinantes son clases cerradas de palabras que ocurren con sustantivos, generalmente marcando el principio de una frase sustantiva. Un pequeño subtipo de determinantes es el artículo: *a*, *el*. Otros determinantes incluyen *ese* (como en *el libro ese*).

- **Pronombres:** Los pronombres son formas que generalmente actúan como una clase de atajo para referirse a alguna frase sustantiva, entidad o evento. Se dividen en:

  - **Pronombres personales:** Hacen referencia a personas o entidades (Yo, tú, él, ella, nosotros, ellos, étc.)
  - **Pronombres posesivos:** Son formas de pronombres personales que indican una posesión actual o mas generalmente solo una relacion abstracta entre la persona y algun objeto (mío, tuyo, suyo, mi, nuestro, étc.)

- **Conjunciones:** Las conjunciones son utilizadas para unir dos frases, cláusulas o sentencias. Las conjunciones coordinantes como *y*, *o* unen dos elementos de igual estado. Las conjunciones subordinativas son utilizadas cuando uno de los elementos es de algún tipo de estado integrado. Por ejemplo *Me molestó* ***que*** *no me lo dijeras.*

- **Verbos auxiliares:** Los verbos auxiliares son verbos que proporcionan información gramatical y semántica adicional a un verbo de significado completo. Dichos verbos auxiliares brindan la información gramatical de modo, tiempo, persona y número y las formas no personales. Por ejemplo *¿por qué no* ***has*** *llegado a la hora prevista?* o también *La avenida principal de la ciudad* ***fue*** *clausurada por obras de refacción.*

- **Numerales:** Los determinantes numerales o simplemente numerales son los que expresan de modo preciso y exacto la cantidad de objetos designados por el sustantivo al que acompañan, delimitan o designan. Limitan el significado general del sustantivo, precisando con exactitud la cantidad de objetos que aquel designa o el lugar de orden que ocupan. Los numerales pueden ser de varias clases. Los más importantes son:

- **Cardinales**: informan una cantidad exacta:
  Quiero *cuatro* libros.
- **Ordinales**: informan del orden de colocación:
  Quiero el *cuarto* libro.
- **Fraccionarios**: informan de particiones de la unidad:
  Quiero la *cuarta* parte.
- **Multiplicativos**: informan de múltiplos:
  Quiero *doble* ración.

2.2. Conjuntos de etiquetas

La sección anterior dió una descripción general de los tipos de clases sintácticas a las que pertenecen las palabras. Esta sección presenta los conjuntos de etiquetas actuales utilizados en la etiquetación gramatical. Es decir, las etiquetas que se corresponden con cada una de estas clases sintácticas. Todavía no existe un consenso sobre el conjunto de etiquetas o tagset más adecuado. Generalmente los conjuntos de etiquetas grandes ofrecen una decripción sintáctica más específica mientras que los conjuntos de etiquetas más pequeños usualmente brindan una información linguística más acotada. Una de las características clave para decidir que conjunto de etiquetas es el más adecuado justamente depende del nivel de detalle linguístico que se esté buscando. Otro hecho notable referido a los conjuntos de etiquetas es que los más pequeños generalmente están contenidos en los conjuntos mayores. Ya que las etiquetas más específicas que se encuentran en los conjuntos mayores pueden ser convertidas en etiquetas de menor especificidad con la consecuente pérdida de detalle linguístico. Por el otro lado, también se pueden convertir las etiquetas pertenecientes a un conjuntos pequeños en etiquetas de mayor especificidad que pertenecen a conjuntos más grandes, ya que generalmente existen etiquetas equivalentes en los conjuntos de mayor tamaño.

Cuadro 1: *Conjunto de Etiquetas Penn Tree Bank*

Etiqueta	**Descripción**	**Ejemplo**
CC	Coordinating conjunction	*and*
CD	Cardinal number	*1, third*
DT	Determiner	*the*
EX	Existential	*there there is*
FW	Foreign word	*d'hoevre*
IN	Preposition/subordinating conjunction	*in, of, like*
JJ	Adjective	*green*
JJR	Adjective, comparative	*greener*
JJS	Adjective, superlative	*greenest*
LS	List marker	*1)*
MD	Modal	*could, will*
NN	Noun, singular or mass	*table*
NNS	Noun plural	*tables*
NNP	Proper noun, singular	*John*
NNPS	Proper noun, plural	*Vikings*
PDT	Predeterminer both	*the boys*

Cuadro 1: *Conjunto de Etiquetas Penn Tree Bank*

Etiqueta	**Descripción**	**Ejemplo**
POS	Possessive ending	*friend's*
PRP	Personal pronoun	*I, he, it*
PRP$	Possessive pronoun	*my, his*
RB	Adverb	*however, usually, naturally, here, good*
RBR	Adverb, comparative	*better*
RBS	Adverb, superlative	*best*
RP	Particle	*give up*
SYM	Symbol	*+, %,&*
TO	To	*to go, to him*
UH	Interjection	*uhhuhhuhh*
VB	Verb, base form	*take*
VBD	Verb, past tense	*took*
VBG	Verb, gerund/present participle	*taking*
VBN	Verb, past participle	*taken*
VBP	Verb, sing. present, non-3d	*take*
VBZ	Verb, 3rd person sing. present	*takes*
WDT	Wh-determiner	*which*
WP	Wh-pronoun	*who, what*
WP$	Possessive wh-pronoun	*whose*
WRB	Wh-abverb	*where, when*
$	Dollar sign	*$*
#	Pound sign	*#*
”	Left quote	(’ or ”)
”	Right quote	(’ or ”)
(	Left parenthesis	([, (, {, ¡)
)	Right parenthesis	( ],), }, ¿)
,	Comma	,
.	Sentence-final punc	(. ! ?)
:	Mid-sentence punc	(: ; ... -)

Más allá de que no exista aún un consenso sobre que conjunto de etiquetas utilizar, hay un pequeño número de conjuntos de etiquetas o tagsets populares para el idioma inglés, muchos de los cuales evolucionaron a partir del conjunto de etiquetas utilizado para etiquetar el corpus Brown. Este conjunto de etiquetas se conoció como el Brown Corpus Tag-set, un conjunto de 87 etiquetas que se utilizó para etiquetar el corpus Brown: un corpus de 1 millón de palabras construído a partir de ejemplos provenientes de 500 textos de diferentes géneros (diarios, novelas, no ficción, académico, étc.) que fué ensamblado en la Universidad Brown entre 1963 y 1964. Este corpus fué etiquetado gramaticalmente aplicando en primera instancia un etiquetador automático, el programa TAGGIT, y luego corregido manualmente.

Al lado del conjunto de etiquetas Brown se encuentran dos de los conjuntos de etiquetas más utilizados: el conjunto de etiquetas reducido Pen Treebank de 45 etiquetas y el conjunto de etiquetas CLAWS C5 de tamaño medio con 62 etiquetas que fué utilizado para etiquetar el corpus British National Corpus

(BNC).

El conjunto de etiquetas Penn Treebank mostrado anteriormente también fué utilizado para etiquetar el corpus Brown, el corpus Wall Street Journal y el corpus Switchboard entre otros. En realidad, quizás en parte por su pequeño tamaño es uno de los conjuntos de etiquetas más utilizado. A continuación se exhiben algunos ejemplos de oraciones del corpus Brown etiquetadas con el conjunto de etiquetas Penn Treebank. Representaremos una palabra etiquetada mediante la colocación de una barra oblicua seguida de su etiqueta:

1. The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

2. **There/EX** are/VBP 70/CD children/NNS **there/RB**

3. Although/IN preliminary/JJ findings/NNS were/VBD **reported/VBN** more/RBR than/IN a/DT year/NN aho/IN ,/, the/DT latest/JJS results/NNS appear/VBP in/IN today/NN **'s/POS** New/NNP England/NNP Journal/NNP of/IN Medicine/NNP ./.

El primer ejemplo exhibe los determinantes *the* y a, los adjetivos *grand* y *other*, los sustantivos comunes *jury*, *number* y *topics* y el verbo en tiempo pasado *commented*. El segundo ejemplo muestra el uso de la etiqueta ET para marcar la construcción existencial *there* y otro uso de *there* que es etiquetado como un adverbio (RB). El tercer ejemplo muestra la segmentación del morfema posesivo *'s* y un ejemplo de la construcción pasiva 'were reported', en la cual el verbo *reported* está marcado como un pasado participio (VBN) en vez de como un pasado simple (VBD). También es interesante notar que el sustantivo propio *New England* está etiquetado como NNP. Finalmente, se puede observar que como *New England Journal of Medicine* es un sustantivo propio, el etiquetado de Treebank elige marcar cada sustantivo separado como NNP, incluyendo *journal* y *medicine*, que en otro caso serían etiquetados como sustantivos comunes (NN).

2.2.1. Especificidad de etiquetas: Treebank vs C5 y Brown

El conjunto de etiquetas Penn Treebank es una selección de 45 etiquetas del conjunto de etiquetas Brown (de 87 etiquetas). Este conjunto reducido deja afuera información que puede ser recuperada desde la identidad del item léxico. Por ejemplo los conjuntos de etiquetas Brown y C5 incluyen una etiqueta para cada una de las diferentes formas de los verbos *do*, *be* y *have*(C5 propone la etiqueta VDD para *did* y VDG para *doing*). Estas etiquetas fueron omitidas en el conjunto Treebank.

Ciertas distinciones sintácticas no fueron marcadas en el conjunto de etiquetas Penn Treebank. Por ejempo, la etiqueta IN es utilizada para preposiciones como para conjunciones subordinadas. El conjunto del Penn Treebank no es suficientemente específico en ciertos casos. Los conjuntos de etiquetas de Brown y C5 , por ejemplo, distinguen preposiciones (IN) de conjunciones subordinadas (CS), como en los siguiente ejemplos:

1. **after/CS** spending/VBG a/AT few/AP days/NNS at/IN the/AT Brown/NP Palace/NN Hotel/NN

2. **after/IN** a/AT wedding/NN trip/NN to/IN Corpus/NP Christi/NP ./.

También tienen dos etiquetas para la palabra *to*; en Brown el uso del infinitivo es etiquetado como TO, mientras que las preposiciones son etiquetadas como IN:

1. **to/TO** give/VB priority/NN **to/IN** teacher/NN pay/NN raises/NNS

El conjunto de etiquetas Brown también posee la etiqueta NR para sustantivos adverbiales como *home*, *west*, *Monday* y *tomorrow*. Como Treebank carece de esta etiqueta, hay una política mucho menos consciente para sustantivos adverbiales; *Monday*, *Tuesday* y otros días de la semana son marcados como NNP, *tomorrow*, *west* y *home* son marcados algunas veces como NN y algunas veces como RB. Esto hace al conjunto de etiquetas Treebank menos útil para tareas de alto nivel linguístico como la detección del tiempo de frases. Sin embargo, el conjunto de etiquetas de Treebank ha sido el más utilizado para la evaluación de algoritmos de etiquetación automática. Esta es la razón por la cual elegimos este conjunto de etiquetas para utilizar en el desarrollo del presente trabajo.

2.3. Corpus

Un corpus es una colección de textos escritos y/o transcripciones del lenguaje oral para cierto idioma que generalmente se utiliza para el estudio del lenguaje. La palabra corpus significa cuerpo en latín, su plural es corpora. Habitualmente el tamaño de un corpus es superior al millón de palabras. Para construir un corpus se reúne una cantidad considerable de textos escritos y/o transcripciones orales para luego ser preservado en algún formato (generalmente electrónico).

Los corpora son utilizados por lingüistas para describir naturalmente el lenguaje basados en la evidencia obtenida de sus observaciones. En su trabajo generalmente utilizan operaciones estadísticas sobre los corpora para medir la frecuencia de algún aspecto léxico. Los corpora, grandes cantidades de ocurrencia natural del lenguaje, han ayudado a realizar progresos en diferentes campos del lenguaje como el estudio de fraseología, análisis crítico del discurso, estilismos, lingüística forense, traducciones y enseñanza del lenguaje entre otros.

Diferentes tipos de corpora permiten el análisis de distintas clases de discuros para hallar evidencia cuantitativa sobre la existencia de patrones en el lenguaje o para verificar teorías. Los primeros estudios sobre un corpus se enfocaron en palabras; su frecuencia y ocurencia. Con el desarrollo de la tecnología y de motores de búsqueda más precisos y eficientes, las posibilidades crecieron ampliamente. Hoy en día es posible realizar búsquedas para una palabra perteneciente a cierta clase sintáctica o patrones completos como por ejemplo:

- preposición + sustantivo
- determinante + sustantivo
- una palabra particular + clase de palabra específica sucediéndola.

Cuando corpora escritos y hablados se hicieron disponibles, los lingüistas comenzaron a analizarlos para verificar patrones o diferencias entre el lenguaje hablado y el lenguaje escrito. Parece que aparte de algunas características obvias como salidas en falso y vacilaciones que se producen en el habla, la utilización de un gran número de expresiones deícticas es más frecuente en los discursos orales. Probablemente esto es debido a los signos linguísticos extra en

los que el lenguaje hablado es más vago. Adicionalmente ciertas características gramaticales manifestadas en el habla deben ser consideradas agramaticales en la escritura.

Otra área importante de estudio lingüístico de corpora es el cambio histórico de los significados de las palabras y la gramática. Y aunque la cantidad de viejos textos disponibles en formato electrónico es mucho más pequeña que la cantidad de textos contemporáneos, el trabajo es factible. En efecto fueron establecidas las diferencias en los aspectos gramaticales concernientes a la voz pasiva.

Por otro lado, en las traducciones es habitual utilizar corpora paralelos que permiten una mejor elección de equivalencias y estructuras gramaticales que podrían reflejar el significado deseado. Estudios adicionales sobre corpora revelaron que los traductores no traducen palabra por palabra sino unidades más grandes (cláusulas o sentencias).

Los estudios de corpora probablemente han tenido una gran influencia en la enseñanza del lenguaje. Primero que nada, han influído en la forma en que se hacen los diccionarios. Segundo los aprendices del lenguaje han sido estudiados para mejorar el conocimiento de los maestros.

Los lingüistas creen que un análisis confiable del lenguaje ocurre mejor en ejemplos recolectados de campo; contextos naturales y con interferencia experimental mínima. Dentro del corpus lingüístico existen visiones divergentes en torno al nivel de las anotaciones. Desde John Sinclair abogando anotaciones mínimas y permitiendo a los textos «hablar por ellos mismos» a otros como el equipo de Survey of English Usage (University College, London) abogando anotaciones como un camino hacia un riguroso entendimiento linguístico.

2.3.1. Un poco de historia

El punto de inflexión en corpus linguístico moderno fué la publicación de Henry Kucera y W. Nelson Francis: Computational Analysis of Present-Day American English en 1967. Un trabajo basado en el análisis del corpus Brown, una compilación cuidadosamente seleccionada de inglés americano actual totalizando alrededor de 1 millón de palabras obtenidas de una amplia variedad de fuentes. Kucera y Francis sometieron este corpus a una gran variedad de análisis computacional desde el cual compilaron un rico y nutrido corpus combinando elementos de linguística, enseñanza de lenguaje, psicología, estadística y sociología. Una publicación clave adicional fué «Towards a description of English Usage» de Randolph Quirk (1960) en la que introdujo Survey of English Usage.

Poco después el editor de Boston Houghton-Mifflin se acercó a Kucera para suministrarle el material base de 1 millón de palabras para su nuevo diccionario American Heritage Dictionary (AHD), el primer diccionario que fué compilado utilizando corpus linguístico. El AHD dió el paso innovador de combinar elementos prescriptivos (como debe utilizarse el lenguaje) con información descriptiva (como se usa actualmente).

Otros editores siguieron el ejemplo. El editor inglés Collins creó y compiló el diccionario Cobuild utilizando el corpus Bank of English. Fué diseñado para usuarios que están aprendiendo inglés como lengua extranjera.

El corpus Brown también dió lugar a un número de corpora similarmente estructurada: el corpus LOB (1960, inglés británico), Kolhapur (inglés indio), Wellington (inglés de Nueva Zelanda), Australian Corpus of English (inglés australiano) y el Flob corpus (1990, inglés británico).

Otros corpora representan más lenguajes, variedades y modos: International Corpus of English, el British National Corpus es una colección de 100 millones de palabras provenientes de textos escritos e inglés hablado creado en los 1990s por un consorcio de editores, universidades (Oxford y Lancaster) y la British Library. Para inglés americano contemporáneo, el trabajo se ha centrado en el American National Corpus (más de 400 millones de palabras de inglés americano contemporáneo).

El primer corpus computarizado de lenguaje hablado transcripto fué construído en 1971 por el Montreal French Project, conteniendo 1 millón de palabras que inspiró a Shana Poplack a crear un corpus mucho más grande de Francés hablado.

Al lado de estos corpora de lenguaje vivo se encuentra corpora computarizado que también fué construído a partir de colecciones de textos en lenguajes antiguos. Como ejemplo tenemos la base de datos Andersen-Forbes de la biblia hebrea, desarrollada desde los años 1970, en donde cada cláusula es parseada utilizando grados que representan más de 7 niveles de sintaxis y cada segmento es etiquetado con 7 campos de información. El Quaric Arabic Corpus es un corpus anotado para el lenguaje árabe clásico del corán. Este es un proyecto reciente con múltiples capas de anotación incluyendo segmentación morfológica, etiquetado gramatical y análisis sintáctico utilizando dependencia gramatical.

2.3.2. Métodos

Los corpora linguísticos han generado una cantidad de métodos de investigación intentando trazar un camino desde los datos hacia la teoría. Wallib y Nelson (2001) introdujeron lo que ellos llamaron la perspectiva 3A: anotación, abstracción y análisis.

- **Anotación:** La anotación consiste en la aplicación de un esquema a los textos. Las anotaciones incluyen marcado estructural, etiquetado gramatical, parsing y varias representaciones más.
- **Abstracción:** La abstracción consiste en la traducción (mapeo) de términos del esquema a términos en el modelo teórico. Típicamente incluye búsqueda linguística directa y también puede incluir aprendizaje por reglas para parsers.
- **Análisis:** El análisis consiste de exploración estadística, manipulación y generalización desde los datos. También podría incluir evaluaciones estadísticas, optimización basada en reglas o métodos de descubrimiento del conocimiento. La mayoría de los corpora de hoy en día está anotado gramaticalmente y aplican algún método para aislar términos que pueden ser interesantes en las palabras circundantes.

2.4. Etiquetadores gramaticales automáticos

Como se mencionó anteriormente, el etiquetado gramatical es el proceso de asignar una etiqueta gramatical a cada palabra dentro de un texto. Generalmente las etiquetas gramaticales también son aplicadas a los signos de puntuación, por lo tanto el etiquetado requiere que los signos de puntuación sean separados

de las palabras. Este proceso se realiza previamente o como parte del etiquetado y es conocido como *tokenización*; es el proceso encargado de separar puntos, comas, paréntesis y otros caracteres de las palabras así como también desambigüar el fin de oración (por ejemplo un punto o signo de pregunta) de un signo de puntuación (como en una abreviación por ejemplo *étc.*)

La entrada para un algoritmo de etiquetación automática es una cadena de palabras y un conjunto de etiquetas. La salida es la mejor etiqueta encontrada para cada palabra. Consideremos las siguientes oraciones etiquetadas gramaticalmente:

Book/VB that/DT flight/NN ./.
Does/VBZ that/DT flight/NN serve/VB dinner/NN ?/.

Asginar una etiqueta gramatical a una palabra no es una tarea trivial incluso en estos sencillos ejemplos. Por ejemplo, la palabra *book* es ambigüa. Es decir que tiene más de un uso posible y por lo tanto más de una etiqueta gramatical posible. Puede ser un verbo (como en ***book*** *that flight* o *to* ***book*** *the suspect*) o un sustantivo (como en *hand me that* ***book*** o a ***book*** *of matches*). Análogamante *that* puede ser un determinante (como en *Does* ***that*** *flight serve dinner*) o un complementador (como en *I thought* ***that*** *your flight was earlier*).

El problema del etiquetado gramatical reside en resolver estas ambigüedades, eligiendo la etiqueta adecuada según el contexto. ¿Pero qué magnitud tiene el problema de la ambigüedad de las palabras? Podemos apreciar que la mayoría de las palabras en inglés no son ambigüas, o lo que es lo mismo, tienen una única etiqueta posible. Pero sin embargo muchas de las palabras más comunes del inglés son ambigüas, es decir que las palabras más utilizadas, las que se emplean con mayor frecuencia, pueden tener más de una etiqueta. Por ejemplo *can* puede ser un auxiliar (puede), un sustantivo (lata o contenedor de metal) o un verbo (poner algo en la lata).

Afortunadamente muchas de las palabras ambigüas son fácilmente desambigüables. Esto sucede porque las etiquetas asociadas a una palabra no suelen ocurrir con la misma frecuencia. Por ejemplo *a* puede ser un determinante o la letra *a* (quizás como parte de un acrónimo o una inicial), pero es preciso notar que el sentido de *a* es mucho más frecuente como determinante que como letra. Es decir que es mucho más frecuente encontrar *a* en oraciones como *My father bought* ***a*** *new car* o *There is* ***a*** *hair in my soup* que en oraciones como *Written by* ***A***. *Kamio* o *The letter* ***a*** *is the first letter of the alphabet.*

Existen distintos métodos computacionales para asignar una etiqueta gramatical a una palabra. La mayoría de los algoritmos de etiquetado automático pertenecen a una de dos clases: etiquetadores basados en reglas o etiquetadores estocásticos.

Los etiquetadores basados en reglas generalmente incluyen una gran cantidad de reglas de desambigüación escritas a mano que especifican, por ejemplo, que una palabra ambigüa es un sustantivo antes que un verbo si es seguida por un determinante.

Los etiquetadores estocásticos generalmente resuelven la ambigüedad de etiquetas utilizando un corpus de entrenamiento del cual "aprenden" como etiquetar. Este aprendizaje se realiza extrayendo información sobre la probabilidad de que una palabra dada tenga cierta etiqueta en cierto contexto.

Adicionalmente existe una tercera clase de etiquetadores que es una mezcla

de estos dos: etiquetadores basados en la transformación. Como los etiquetadores basados en reglas, están basados en reglas que determinan cuando una palabra ambigüa debe tener cierta etiqueta. Y como los etiquetadores estocásticos tienen un componente de aprendizaje automático; las reglas son inducidas automáticamente a partir de un corpus de entrenamiento previamente etiquetado.

2.4.1. Etiquetadores gramaticales basados en reglas

Los primeros algoritmos de asignación de etiquetas gramaticales estaban basados en un proceso de dos etapas. En la primer etapa utilizaban un diccionario para asignar a cada palabra una lista de potenciales etiquetas gramaticales. En la segunda etapa utilizaban grandes listas de reglas de desambigüación escritas a mano para reducir la lista de etiquetas hasta llegar a una para cada palabra. De esta manera eliminaban las etiquetas inconsistentes con el contexto.

Las versiones actuales de los etiquetadores gramaticales basados en reglas mantienen los principios originales teniendo en cuenta que los diccionarios y el conjunto de reglas han adquirido un tamaño considerablemente mayor: manejan alrededor de 3800 reglas y un diccionario de etiquetas del órden de las 56.000 entradas para el idioma inglés.

2.4.2. Etiquetadores gramaticales estocásticos

La inclusión de probabilidades en el proceso de etiquetación gramatical no es una idea nueva. Surge como una consecuencia natural a partir del hecho de que una palabra es empleada con un sentido gramatical mucho más frecuentemente que con otro. Como se mencionó anteriormente, *a* es mucho más frecuentemente utilizada como determinante que como letra. La inclusión de probabilidades también responde a otro factor importante: la construcción gramatical; cierta etiqueta es precedida frecuentemente por ciertas otra/s. Por ejemplo, como se mencionó anteriormente, los pronombres posesivos generalmente son sucedidos por verbos. Es decir que es más probable encontrar oraciones cuyas palabras estén etiquetadas con PP sucedida por NN que PP sucedida por otra etiqueta.

A continuación vamos a presentar 2 tipos de etiquetadores gramaticales estocásticos: etiquetadores estocásticos basados en el modelo oculto de Markov o simplemente etiquetadores HMM [2] y etiquetadores estocásticos basados en el modelo de máxima entropía.

2.4.3. Etiquetadores gramaticales basado en HMM

El uso del modelo oculto de Markov para realizar etiquetado gramatical es un caso especial de la inferencia bayesiana, un paradigma que fué conocido a partir del trabajo de Bayes (1763). La inferencia Bayesiana o clasificación Bayesiana fue aplicada exitosamente a problemas del lenguaje a partir de 1950. La clasificación bayesiana puede apreciarse como una tarea para la cual contamos con un conjunto de observaciones y el trabajo consiste en determinar a que conjunto de clases pertenece. En lo que respecta al etiquetado gramatical, se puede utilizar este mismo concepto para tratarlo como una tarea de clasificación de secuencia. En ese caso, la observación será una secuencia de palabras (digamos

[2] Por las siglas en inglés de Hiden Markov Model

una oración) para la cual el trabajo consiste en asignar una secuencia de etiquetas gramaticales. Como ejemplo tomemos la oración que aparece a continuación:

Secretariat is expected to race tomorrow

En este caso las observaciones son la secuencia de palabras (es decir la oración misma) y nuestro objetivo es asignarles las etiquetas correspondientes. Ya que una palabra puede ser ambigüa y tener más de una etiqueta posible, hay una pregunta clave que debemos hacernos: ¿Cuál es la mejor secuencia de etiquetas que corresponden a esta secuencia de palabras? La interpretación bayesiana comienza considerando todas las posibles secuencias de clases –en nuestro caso, todas las posibles secuencias de etiquetas gramaticales. El objetivo aquí es elegir la secuencia de etiquetas que es más probable dada la secuencia de observaciones de n palabras w_1^n. En otras palabras, queremos obtener, de todas las secuencias de n etiquetas t_1^n la secuencia de etiquetas tal que $P(t_1^n|w_1^n)$ sea mayor. Se utilizará la notación ˆ para decir "nuestra estimación de la secuencia de etiquetas correcta".

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} fP(t_1^n|w_1^n) \tag{1}$$

La ecuación anterior se lee así: de todas las secuencias de etiquetas de longitud n, queremos la secuencia particular t_1^n que maximiza el lado derecho.

Mientras que esta ecuación nos garantiza obtener la secuencia de etiquetas óptima, todavía no queda del todo claro como utilizarla. Es decir, para una secuencia de etiquetas dada t_1^n y una secuencia de palabras w_1^n, no sabemos como computar directamente $P(t_1^n|w_1^n)$. Aquí entra en juego la clasificación Bayesiana, ofreciendo una forma de transformar la ecuación en un conjunto de otras probabilidades más sencillas de computar. Las reglas de Bayes reemplazan la probabilidad condicional $P(x|y)$ por otras tres probabilidades:

$$P(x|y) \;=\; \frac{P(y|x)P(x)}{P(y)} \tag{2}$$

Podemos sustituir (2) en (1) para obtener (3):

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)} \tag{3}$$

Convenientemente podemos simplificar (3) eliminando el denominador $P(w_1^n)$. Esto sucede ya que estamos eligiendo una de todas las secuencias de etiquetas, computando $\frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)}$ en cada una de ellas. Pero $P(w_1^n)$ no cambia en ninguna secuencia de etiquetas, entonces estamos preguntando siempre por la misma observación w_1^n, que tiene la misma probabilidad $P(w_1^n)$. Por lo tanto podemos quitar el denominador con la garantía de que el máximo sea el mismo:

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(w_1^n|t_1^n)P(t_1^n) \tag{4}$$

En resúmen, la secuencia de etiquetas más probable $\hat{t}_1^n$ dada alguna palabra w_1^n puede ser computada tomando el producto de dos probabilidades para cada secuencia de etiquetas y eligiendo la secuencia que lo maximiza.

Desafortunadamente todavía sigue siendo muy difícil computar esta ecuación directamente. Los etiquetadores gramaticales basados en HMM realizan dos suposiciones simplificadoras. La primera es que la probabilidad de aparición de una palabra depende solo de su etiqueta gramatical, es decir que es independiente de las palabras y etiquetas que tiene alrededor. Más técnicamente:

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i) \tag{5}$$

La segunda suposición es que la probabilidad de aparición de una etiqueta gramatical depende solo de la etiqueta previa (sin tener en cuenta las etiquetas anteriores a la etiquetea previa), esto es la suposición de bigrama.

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1}) \tag{6}$$

Utilizando estas suposiciones obtenemos esta nueva ecuación, la cual es utilizada por los etiquetadores gramaticales basados en bigramas para estimar la secuencia de etiquetas gramaticales más probable.

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}}\, P(t_1^n|w_1^n) \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \tag{7}$$

La ecuación anterior contiene dos clases de probabilidades, probabilidades de transición de etiquetas y probabilidades de palabras. Tomemos un momento para ver que es lo que representan estas probabilidades.

- **Probabilidades de transición de etiquetas**: Las probabilidades de transición de etiquetas, $P(t_i|t_{i-1})$, representan la probabilidad de que ocurra una etiqueta dada la etiqueta previa. Por ejemplo, es muy probable que un determinantes preceda a un adjetivos o a un sustantivo, como *that/DD flight/NN y the/DT yellow/JJ hat/NN*. Por lo tanto esperamos que las probabilidades $P(NN|DT)$ y $P(JJ|DT)$ sean altas.

  Por otro lado, es infrecuente que los adjetivos precedan a los determinantes, entonces la probabilidad $P(DT|JJ)$ será pequeña. Podemos computar la máxima probabilidad estimada o MLE [3] de una probabilidad de transición de etiquetas $P(NN|DT)$ etiquetando y contando las etiquetas gramaticales en un corpus. Esto es: de todas las veces que vemos DT, cuántas de esas veces vemos NN después de DT. Lo expresamos más formalmente con el siguiente cociente:

  $$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_i)} \tag{8}$$

  Elijamos un corpus específico para examinar, por ejemplo el corpus Brown. Éste es un corpus de 1 millón de palabras de Inglés Americano. El corpus Brown ha sido etiquetado dos veces, la primera en los años sesenta con el conjunto de etiquetas 87-tag y de vuelta en los años noventa con el conjunto de etiquetas Treebank. En el corpus Brown etiquetado con el conjunto

[3] Por sus siglas en inglés Maximum Likelihood Estimated

de etiquetas Treebank, la etiqueta DT ocurre 116.454 veces. De esas veces, DT es seguido por NN 56.509 veces. Por lo tanto esta probabilidad de transición se calcula como sigue:

$$P(NN|DT) = \frac{C(DT, NN)}{C(DT)} = \frac{56509}{116454} = ,49 \tag{9}$$

Claramente la probabilidad de obtener un sustantivo común después de un determinante es .49 y de hecho alta como sospechábamos.

- **Probabilidades de la palabra**: Por otro lado las probabilidades de la palabra, $P(w_i|t_i)$, representan la probabilidad de que dada una etiqueta esta esté asociada con cierta palabra. Por ejemplo si tenemos la etiqueta VBZ (verbo singular de tiempo presente en tercera persona) y quisiéramos adivinar el verbo asociado a esa etiqueta, probablemente elegiríamos el verbo *is*[4], debido a que el verbo *to be* es muy común en inglés. Podemos computar $P(is|VBZ)$ de nuevo contando de cuántas veces que vemos VBZ en un corpus cuántas de esas veces VBZ está etiquetando la palabra *is*. Esto es computar el siguiente cociente:

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)} \tag{10}$$

En el corpus Brown etiquetado con Treebank, la etiqueta VBZ ocurre 21.627 veces y VBZ es la etiquetra para *is* 10.073 veces. Entonces:

$$P(is|VBZ) = \frac{C(VBZ, is)}{C(VBZ)} = \frac{10{,}073}{21{,}627} = 0{,}47 \tag{11}$$

Resumiendo, el etiquetado HMM es la tarea de elegir con la mayor probabilidad una secuencia de etiquetas para una secuencia de palabras dada. HMM incluye la suposición de ciertos hechos para simplificar las ecuaciones originales mejorando así la eficiencia de los cómputos.

2.5. Corpora de entrenamiento y corpora de verificación

Los etiquetadores gramaticales que se basan en modelos estocásticos poseen un proceso de entrenamiento sobre un corpus etiquetado previamente en el cual se generan las probabilidades que se utilizan para tomar decisiones frente a palabras ambigüas.

Dicho corpus de entrenamiento necesita ser cuidadosamente considerado. Si el corpus de entrenamiento es muy específico al dominio, es decir que el corpus de entrenamiento de alguna manera es similar al corpus que se desea etiquetar, las probailidades van a ser muy ajustadas y no tendrá un buen rendimiento en oraciones de diferentes dominios. Pero si el corpus de entrenamiento es muy general, estas probabilidades no van a llegar a hacer el trabajo suficiente de reflejar el dominio.

Supongamos que estamos intentando etiquetar una oración particular. Si nuestra oración es parte del corpus de entrenamiento, las probabilidades de las etiquetas para esa oración van a ser extraordinariamente precisas y vamos a

[4] *is* es el presente en tercera persona del verbo *to be*

sobreestimar la precisión de nuestro etiquetador. Se desprende como conclusión que el corpus de entrenamiento no debe ser parcial incluyendo esa oración. Por lo tanto al trabajar con etiquetadores basados en modelos estocásticos, dado un corpus de datos relevante, es una tarea habitual dividir los datos en un corpus de entrenamiento y un corpus de verificación.

Una vez realizada esta división se entrena el etiquetador con el corpus de entrenamiento, se ejecuta el proceso de etiquetación y luego se comparan los resultados con el corpus de verificación.

En general existen dos métodos para entrenar y verificar un etiquetador gramatical. En el primer método, se divide el corpus disponible en tres partes: un corpus de entrenamiento, un corpus de verificación y un corpus de test de desarrollo [5]. Se entrena el etiquetador con el corpus de entrenamiento. Entonces se utiliza el corpus de test de desarrollo para eventualmente afinar o ajustar algunos parámetros y en general decidir cual es el mejor modelo. Una vez que se elige el supuesto mejor modelo, se corre contra el corpus de verificación para ver su rendimiento.

En el segundo método de entrenamiento y verificación, se elige aleatoreamente una división de corpus de entrenamiento y verificación para nuestros datos. Se entrena el etiquetador y luego se calcula el error en el corpus de verificación. A continuación se repite con un corpus de entrenamiento y de verificacion diferente seleccionado aleatoreamente. La repetición de este proceso, llamado validación cruzada, generalmente es realizada 10 veces. Luego se promedian esas 10 corridas para obtener un promedio en la proporción del error.

Al comparar modelos es importante utilizar verificaciones estadísticas para determinar si la diferencia entre los modelos es significantiva.

2.6. Evaluación de etiquetadores gramaticales

Los etiquetadores gramaticales generalmente son evaluados comparando su precisión contra un corpus de verificación [6] etiquetado por humanos. Definimos precisión como el porcentaje de todas las etiquetas en el corpus de verificación donde el etiquetador y el Gold Stantard concuerdan. Los algoritmos actuales de etiquetado gramatical tienen una precisión del 96 %-97 % para conjuntos de etiquetas simples como el Penn Treebank. Estas precisiones son para palabras y puntuaciones, la precisión para palabras solas es menor.

Naturalmente uno tiende a preguntarse qué tan bueno es un 97 %. El rendimiento de un proceso de etiquetado puede ser comparado contra un límite inferior y un límite superior. Una manera de establecer un límite superior es ver que tan bien realizan la tarea los humanos.

Marcus, por ejemplo, encontró que los etiquetadores humanos concuerdan en el 96 %-97 % de las etiquetas en el corpus Brown etiquetado con etiquetas Penn Treebank. Esto sugiere que el Gold Standard debe tener un 3 %-4 % de margen de error, y por lo tanto no tiene sentido obtener una precisión del 100 %. Ratnaparkhi mostró que en las palabras donde su etiquetador ha tenido problemas de ambigüedad de etiquetación fueron exactamente las mismas en donde los humanos han etiquetado inconsistentemente el corpus de entrenamiento. Dos experimientos realizados por Voutilainen encontraron que cuando a los huma-

[5] También llamado *devtest*

[6] También llamado *Gold Standard*

nos se les permitió discutir etiquetas, alcanzaron un consenso en el 100 % de las etiquetas.

Por otro lado el límite inferior sugerido por Gale es elegir la etiqueta más probable aplicando el modelo de unigrama para cada palabra ambigüa. La etiqueta más probable para cada palabra puede ser computada desde un corpus etiquetado a mano (que puede ser el mismo que el corpus de entrenamiento para el etiquetador que está siendo evaluado).

2.7. Análisis de error

Para mejorar el rendimiento de un etiquetador gramatical necesitamos entender donde está funcionando mal. Por eso el análisis de error tiene un papel preponderante. Esta tarea se realiza construyendo una matriz de confusión o tabla de contingencia. Una matriz de confusión es una matriz de nxn donde la celda (x, y) contiene el número de veces que una palabra con correcta etiqueta x fué etiquetada por el modelo como y. Por ejemplo, la siguiente tabla muestra una porción de la matriz de confusión para los experimentos de etiquetado con HMM de Franz.

Cuadro 2: *Ejemplo de matriz de confusión*

	IN	**JJ**	**NN**	**NNP**	**RB**	**VBD**	
IN	-	.2			.7		
JJ	.2	-	3.3	2.1	1.7	.2	2.7
NN		8.7	-				.2
NNP	.2	3.3	4.1	-	.2		
RB	2.2	2.0	.5		-		
VBD		.3	.5			-	4.4
VBN		2.8				2.6	-

Las etiquetas de la fila indican las etiquetas correctas, las etiquetas de las columnas indican las etiquetas asignadas por el etiquetador, y cada celda indica el porcentaje del error de etiquetado general. Por lo tanto 4.4 % del total de errores fueron causados por fallida etiquetacion de VBD como VBN. La matriz anterior y el análisis de error relacionado en Franz, Kupiec y Ratnaparkhi sugieren que algunos de los mayores problemas que encaran los etiquetadores actuales son:

1. NN contra NNP contra JJ: Estas etiquetas son difíciles de distinguir. Es especialmente importante distinguir entre sustantivos propios para extracción de la información y traducción automática.

2. RP contra RB contra IN: Todas estas etiquetas pueden aparecer inmediatamente después del verbo.

3. VBD contra VBN contra JJ: Distinguir estas etiquetas es importante para el *parsing* parcial (los participios son utilizados para encontrar pasivos), y para etiquetar correctamente los bordes de las frases nominales.

El análisis de error es una parte crucial de cualquier aplicación lingüística computacional. Puede ayudar a encontrar *bugs*, encontrar problemas en los datos de entrenamiento y lo más importante, ayuda en el desarrollo de conocimiento y/o algoritmos para utilizar en la solución de problemas.

2.8. Palabras desconocidas

Todos los algoritmos de etiquetado gramatical presentados anteriormente requieren un diccionario que liste las posibles etiquetas de cada palabra para que posteriormente el proceso de etiquetado se encargue de identificar la etiqueta correcta. Pero claro, hay un problema: ningún diccionario es capaz de contener todas las palabras. Los sustantivos propios y los acrónimos son creados muy frecuentemente, de hecho ingresan al lenguaje nuevos sustantivos comunes y verbos en una proporción sorprendente. Por lo tanto, para construir un etiquetador completo no podemos utilizar siempre un diccionario para obtener $P(w_i|t_i)$. Necesitamos algún método para adivinar la etiqueta de una palabra desconocida.

El algoritmo más básico para manejar palabras desconocidas es suponer que cada palabra desconocida es ambigüa entre todas las posibles etiquetas, con igual probabilidad. Entonces el etiquetador debe confiar únicamente en etiquetas contextuales para sugerir la etiqueta adecuada. Un algoritmo ligeramente más complejo está basado en la idea de que la distribución de probabilidad de las etiquetas sobre las palabras desconocidas es muy similar a la distribución de las etiquetas sobre palabras que ocurren solo una vez en un corpus de entrenamiento, una idea sugerida por Baayen y Sproat (1996) y Dermatas y Kokkinakis (1995). Estas palabras que ocurren solo una vez son conocidas como *hapax legomena*.

Por ejemplo, las palabras desconocidas y *hapax legomena* son similares en el hecho de que son más probables de ser sustantivos, seguidas por verbos, pero infrecuentemente suelen ser determinantes o intersecciones. Entonces la probabilidad $P(w_i|t_i)$ para una palabra desconocida es determinada por el promedio de la distribución sobre todos los conjuntos de palabras de una sola ocurrencia en el corpus de entrenamiento. En resúmen, la idea es utilizar "cosas que hemos visto una vez" como un estimador para "cosas que nunca hemos visto".

De todas maneras, la mayoría de los algoritmos para palabras desconocidas hace uso de una fuente de información mucho más poderosa: la morfología de las palabras. Para el inglés, por ejemplo, palabras terminadas en *s* tienden a ser sustantivos plurales (NNS), palabras terminadas en *ed* tienden a ser pasado participio (VBN), palabras terminadas en *able* tienden a ser adjetivos (JJ), y así. Incluso si nunca vimos una palabra, podemos utilizar hechos sobre su forma morfológica para adivinar su etiqueta. Además la información ortográfica puede ser de mucha ayuda. Por ejemplo, palabras que comienzan con letras mayúsculas generalmente son sustantivos propios (NP). La presencia de un guión es también una característica útil; las palabras con guión en la versión Brown del Treebank son más probables de ser adjetivos (JJ).

¿Cómo son combinadas y utilizadas estas características en los etiquetadores gramaticales? Un método es entrenar por separado estimadores de probabilidad para cada característica, asumiendo independencia, y multiplicando las probabilidades. Weischedel (1993) construyó un modelo así, basado en cuatro clases específicas. Utilizaron 3 terminaciones inflecciones (*ed*, *s*, *ing*), 32 terminacio-

nes derivacionales (como *ion*, *al*, *ive* y *ly*), 4 valores de mayúscula dependiendo si una palabra es inicio de oración (+/- mayúscula, +/- inicio) y donde una palabra fué guionada. Para cada característica, entrenaron estimadores de máxima verosimilitud de la probabilidad de la característica dada una etiqueta desde un corpus de entrenamiento etiquetado. Entonces combinaron las características para estimar la probabilidad de una palabra desconocida asumiendo independencia y multiplicando:

$$P(w_i|t_i) = p(\text{palabra desconocida}|t_i)p(\text{mayúscula}|t_i)p(\text{final/guión}|t_i) \quad (12)$$

Otro acercamiento basado en HMM, proveniente del trabajo realizado por Samuelsson (1993) y Brants (2000), generaliza el uso de morfología en una manera basada en datos. En este acercamiento, en vez de preseleccionar ciertos sufijos a mano, son consideradas todas las secuencias finales de letras de todas las palabras. Consideran sufijos menores a diez letras, computando para cada sufijo de longitud i la probabilidad de la etiqueta t_i:

$$P(t_i|l_{n-i+1}, \ldots, l_n) \quad (13)$$

Estas probabilidades son suavizadas utilizando sucesivamente menores y menores sufijos. Esta información de sufijos se mantiene por separado para palabras en mayúscula y minúscula.

En general, la mayoría de los modelos de palabras desconocidas intentan capturar el hecho de que las palabras desconocidas son improbable de ser clases cerradas de palabras. Brants modela este hecho computando solamente las probabilidades de sufijos desde el corpus de entrenamiento para palabras cuya frecuencia en el corpus de enternamiento es ≤ 10.

2.9. Etiquetador Gramatical TnT

TnT(Trigrams' n' Tags) es un etiquetador gramatical estocástico basado en HMM. Según Brants este etiquetador tiene un rendimiento mejor o igual a otros etiquetadores actuales de diferentes bases teóricas, incluyendo etiquetadores basados en máxima entropía.

2.9.1. Modelo teórico

TnT utiliza modelos de Markov de segundo órden para la etiquetación gramatical. Técnicamente calcula, dada una secuencia de T palabras $w_1, \ldots, w_T$

$$argmax_{t_1,\ldots,t_T} \left[\prod_{i=1}^{T} P(t_i|t_{i-1}, t_{i-2}) P(w_i|t_i) \right] P(t_{T+1}|t_T) \quad (14)$$

para hallar las etiquetas $t_1, \ldots, t_T$. Las etiquetas adicionales t_{-1}, t_0 y t_T son delimitadores del principio y el final de la secuencia. Estas etiquetas adicionales mejoran levemente los resultados del etiquetado marcando una particularidad de TnT con respecto a otros etiquetadores. Las probabilidades son estimadas desde un corpus etiquetado previamente (el ya mencionado corpus de entrenamiento). Para ello TnT utiliza probabilidades de máxima verosimilitud $\hat{P}$ obtenidas a partir de la frecuencia relativa y luego aplica una técnica de suavizado

$$\text{Unigramas:} \hat{P}(t_3) = \frac{f(t_3)}{N} \quad (15)$$

$$\text{Bigramas:}\hat{P}(t_3|t_2) = \frac{f(t_2,t_3)}{f(t_2)} \tag{16}$$

$$\text{Trigramas:}\hat{P}(t_3|t_1,t_2) = \frac{f(t_1,t_2,t_3)}{f(t_1,t_2)} \tag{17}$$

$$\text{Léxico:}\hat{P}(w_3|t_3) = \frac{f(w_3,t_3)}{f(t_3)} \tag{18}$$

donde t_1, t_2 y t_3 pertenecen al conjunto de etiquetas y w_3 pertenece al lexicon. N es el número de *tokens* del corpus de entrenamiento. La probabilidad de máxima verosimilitud se calcula como cero si el denominador o el nominador son cero.

2.9.2. Suavizado

TnT aplica una técnica de suavizado sobre las frecuencias contextuales. Esto tiene lugar debido al problema de los datos esparsos en las probabilidades de los trigramas. Es decir, no hay suficientes instancias de cada trigrama para calcular confiablemente su probabilidad asociada. Incluso estableciendo a cero la probabilidad de un trigrama que no aparece en el corpus genera el efecto indeseado de convertir la probabilidad de una secuencia completa en cero. TnT utiliza interpolación lineal de unigramas, bigramas y trigramas para realizar este proceso de suavizado. Es decir que se estima la probabilidad de un trigrama como sigue

$$P(t_3|t_1,t_2) = \lambda_1\hat{P}(t_3) + \lambda_2\hat{P}(t_3|t_2) + \lambda_3\hat{P}(t_3|t_1,t_2) \tag{19}$$

donde $\hat{P}$ son los estimadoes de máxima verosimilitud presentados anteriormente y λ_1, λ_2 y λ_3 son los pesos asociados a estos estimadores, tales que $\lambda_1 + \lambda_2 + \lambda_3 = 1$. TnT utiliza interpolación lineal con independencia de contexto. Es decir que λ_1, λ_2 y λ_3 tienen el mismo valor para todos los trigramas, o lo que es lo mismo, λ_1, λ_2 y λ_3 son independientes del trigrama que se está calculando. Los valores λ_1, λ_2 y λ_3 son estimados por interpolación de borrado. La idea es que se dará mayor peso a la información de unigrama, bigrama o trigrama más abundante. A continuación se presenta el algoritmo utilizado para realizar esta tarea

Algoritmo 1 Cálculo de λ_1, λ_2 y $\lambda_3 = 0$

Establecer $\lambda_1 = \lambda_2 = \lambda_3 = 0$
  por cada trigrama t_1, t_2, t_3 con $f(t_1, t_2, t_3) > 0$
    según el máximo de los tres valores siguientes:
      caso $\frac{f(t_1,t_2,t_3)-1}{f(t_1,t_2)-1}$: incrementar λ_1 en $f(t_1, t_2, t_3)$
      caso $\frac{f(t_2,t_3)-1}{f(t_2)-1}$: incrementar λ_2 en $f(t_1, t_2, t_3)$
      caso $\frac{f(t_3)-1}{N-1}$: incrementar λ_3 en $f(t_1, t_2, t_3)$
    fin
  fin
normalizar λ_1, λ_2 y λ_3

2.9.3. Manejo de palabras desconocidas

TnT, al igual que muchos otros etiquetadores gramaticales, maneja las palabras desconocidas mediantes análisis de sufijos. Los sufijos son fuertes predictores del tipo de palabra. Por ejemplo las palabras terminadas en *able* en el Wall Street Journal parte del Penn Treebank son adjetivos (JJ) en el 98 % de los casos (ej.:*fashionable*, *variable*) y sustantivos (NN) en el 2 % restante.

La distribución de probabilidades para un sufijo particular es generada a partir de todas las palabras en el corpus de entrenamiento que comparten el mismo sufijo (de alguna longitud máxima predefinida). El término sufijo se entiende en este contexto como la secuencia final de letras de una palabra, que no coincide necesariamente con el significado lingüístico de sufijo.

La fórmula utilizada para calcular la probabilidad de que una etiqueta pertenezca a cierto sufijo es $P(t|l_{n-m+1}, \ldots, l_n)$, es decir, la probabilidad de una etiqueta t dadas las últimas letras l_i de una palabra de n letras. TnT aplica una técnica de suavizado utilizando sufijos cada vez más pequeños aplicando un peso θ_i a cada uno:

$$P(t|l_{n-m+1}, \ldots, l_n) = \frac{\hat{P}(t|l_{n-i+1}, \ldots, l_n) + \theta_i P(t|l_{n-i}, \ldots, l_n)}{1 + \theta_i} \tag{20}$$

para $i = m, \ldots, 0$, utilizando el estimador de máxima verosimilitud $\hat{P}$ para las frecuencias en el lexicon, los pesos θ_i y el caso base

$$P(t) = \hat{P}(t) \tag{21}$$

El estimador de máxima verosimilitud para un sufijo de longitud i es

$$\hat{P}(t|l_{n-i+1}, \ldots, l_n) = \frac{f(t, l_{n-i+1}, \ldots, l_n)}{f(l_{n-i+1}, \ldots, l_n)} \tag{22}$$

TnT utiliza desvío estándard del estimador de máxima verosimilitud para calcular los pesos θ_i.
Decisiones de diseño:

1. La primer decisión de diseño que afronta TnT es encontrar un buen valor para n, la longitud máxima de sufijo utilizada. TnT elige tomar la longitud del mayor sufijo encontrado en el corpus de entrenamiento, con la restricción de que sea menor o igual a 10.

2. Se utiliza independencia de contexto para calcular θ_i, la misma idea que se utilizó para calcular λ_i.

3. Se utilizan estimadores distintos para mayúsculas y minúsculas. Es decir, se mantienen dos árboles de sufijos distintos, uno para mayúsculas y otro para minúsculas.

4. La otra decisión relevante es: ¿Qué palabras del lexicon deben ser utilizadas para el manejo de sufijos? Basándose en el hecho de que las palabras desconocidas son más probablemente infrecuentes, TnT utiliza sufijos de palabras infrecuentes. Por lo tanto, restringe el procedimiento de cálculo de probabilidades de sufijos a palabras con una frecuencia menor o igual a 10.

Adicionalmente, TnT discrimina la información sobre mayúsculas y minúsculas. Esto es debido a que las probabilidades de las etiquetas de palabras con mayúsuculas son distintas a las de las palabras con minúsculas. Para llevar esto a cabo se utilizan flags en las probabilidades contextuales. En vez de

$$P(t_3|t_1, t_2) \tag{23}$$

se utiliza

$$P(t_3, c_3|t_1, c_1, t_2, c_2) \tag{24}$$

donde c_1, c_2 y c_3 son 1 si la palabra contiene mayúsculas y 0 en otro caso. Esto es equivalente a doblar el conjunto de etiquetas y utilizar etiquetas diferentes según si la palabra aparece en mayúscula o no.

3. Desarrollo

3.1. Diccionario COBUILD

Como se menciona anteriormente, para suplir la falta de corpus de entrenamiento sin caer en la tediosa y costosa tarea de anotar un nuevo corpus, se introduce una fuente de información existente y manualmente anotada. Estamos hablando de un diccionario, que no es ni más ni menos que un conjunto de palabras, donde cada palabra representa una entrada que posee una explicación de su significado, algunas características como su propnunciación y clase gramatical y uno o más ejemplos que muestran su uso. Entonces si pegamos o concatenamos todas esas entradas podemos ver al diccionario como un grupo de palabras con información gramatical para aquellas palabras asociadas a una entrada. Es decir, estas entradas del diccionario:

```
siren
sirens
s*a*!i*%er%e0n
A woman is described as a siren when she is attractive and dangerous to men.
One of the women, another of those sirens, haughtily regarded us as we talked.
countable noun
noun

sirloin
sirloins
s*$e*:l!o!in
A sirloin is a piece of beef which is cut from the lower part of a cows back.
...a sirloin of Scotch beef.
mass noun
noun

sissy
sissies
s*!isi1
A boy is described as a sissy, especially by other boys, if he does not like sport and is
Youre a lot of cry-babies and sissies... ...Mummys little sissy boy.
```

`countable noun: also vocative noun`

Pueden concatenarse y verse como:

A
woman
is
described
as
a
siren NN
when
she
is
attractive
and
dangerous
to
men
.
One
of
the
women
,
another
of
those
sirens NNS
,
haughtily
regarded
us
as
we
talked
.
A
sirloin NN
is
a
piece
of
beef
which
is
cut
from
the
lower
part
of
a
cows
back
.
...
a
sirloin NN
of
Scotch
beef
.
Long
fibres
are
picked
carefully
from
the
sisal NN
leaves
.
A
boy
is
described
as
a
sissy NN
,
especially
by
other
boys
,
if
he
does
not
like
sport
and
is

afraid	of
to	cry-babies
do	and
things	sissies NNS
that	...
are	...
slightly	Mummys
dangerous	little
.	sissy NN
Youre	boy
a	.
lot	

Esta última información conforma un corpus parcialmente anotado, es decir, un conjunto de oraciones donde alguna/s de las palabras que comprenden cada oración posee/n una etiqueta gramatical. Este corpus parcialmente anotado se utilizará como base para construir un nuevo corpus completamente anotado que servirá como una nueva fuente de información para entrenar etiquetadores gramaticales.

Claramente el primer paso para llevar a cabo esta tarea es elegir un diccionario y extraer la información mencionada anteriormente. El diccionario elegido fué Cobuild. A continuación se detallan las características que lo hicieron distintivo frente a otros diccionarios.

Cobuild es un diccionario basado en la información del corpus Bank of English y el corpus Collins. Su siglas significan: Collins Birmingham University International Language Database. El corpus Collins es una base de datos con alrededor de 2.5 billones de palabras en Inglés. Contiene material escrito de websites, diarios, revistas y libros publicados en todo el mundo, y material hablado de radio, TV y conversaciones diarias. A su vez el Bank of English forma parte del corpus Collins. Contiene 650 millones de palabras cuidadosamente seleccionadas para dar un reflejo preciso y balanceado del Inglés que se usa día a día. Gracias a la extensa amplitud del corpus se puede apreciar una gran cantidad de ejemplos de como la gente utiliza realmente el lenguaje. Se puede apreciar el empleo de las palabras, su significado, que palabras ocurren juntas y que tan a menudo. Para decidir que palabras incluir al diccionario Cobuild se ha utilizado información sobre la frecuencia de ocurrencia de las mismas. Por ejemplo, alrededor del 90 % del inglés hablado y escrito está constituído por 3.500 palabras aproximadamente.

The Bank of English contiene un amplio rango de tipos diferentes de lenguaje escrito y hablado proveniente de cientos de fuentes diferentes. Aunque la mayoría de las fuentes son británicas, aproximadamente el 25 % de la información proviene de fuentes de inglés americano y alrededor del 5 % de otras variedades nativas del inglés como Australia y Singapur. Los textos escritos provienen de diarios, revistas, libros de ficción y de no ficción, folletos, informes y cartas. Dos tercios del corpus están confeccionados a partir del lenguaje de los medios: diarios, revistas, radio y televisión. Esta es una categoría significativa en vista de que millones de personas escuchan y leen el lenguaje presente en los medios. También fueron incluídas publicaciones internacionales, nacionales y locales para capturar un rango general de temas importantes y estilos. Hay otros cientos de libros y revistas de especializadas que abordan temas desde aeróbicos

a zoología. Cabe destacar que no fueron incluídos en el corpus libros de texto técnicos, científicos, manuales, étc. El lenguaje hablado informal es representado por grabaciones de conversaciones diarias casuales, reuniones, entrevistas y discusiones. Alrededor de 15 millones de palabras de The Bank of English son transcripciones de lenguaje hablado de esta clase. Luego son seleccionadas para incluir un amplio espectro de temas y situaciones de habla.

El propósito de recolectar toda esta valiosa información en computadoras fué permitirles a los lingüistas (escritores de diccionarios) el acceso a la mayor cantidad de información posible sobre cada una de las palabras que deben definir. Desde luego, los lingüistas son elegidos por su habilidad con el lenguaje, pero ni siquiera el lingüista más hábil puede deducir todos los hechos relevantes sobre las palabras de un lenguaje utilizando solo su intuición. El corpus y el software que se utiliza para analizarlo ayudó al equipo de Cobuild a ordenar la información y ganar valiosa percepción sobre la manera en que se utilizan las palabras: sus significados, sus patrones gramaticales típicos y las maneras en que están relacionadas con otras palabras.

Muchas palabras tienen más de una clase de palabra gramatical asociada y a menudo es de mucha ayuda para los lingüistas mirar solo a una clase de palabra por vez. Para ayudarlos a hacer esto, se ha utilizado un software que muestra las clases de palabras en cada linea del corpus. De esta manera los lingüistas pueden mirar la información completa de la clase de palabra o pueden preguntar solo por verbos, sustantivos, étc. Este tipo de software les permite a los lingüistas tomar decisiones sobre los diferentes sentidos de las palabras, el lenguaje de las definiciones, la selección de ejemplos y la información gramatical dada. El corpus permite realizar esta tarea con confianza y exactitud. Y cuanto más grande es el corpus mayor es la confianza y la exactitud.

El diccionario Cobuild fué concebido teniendo especial atención en los ejemplos expuestos. Como se mencionó anteriormente, el proceso de agregado de palabras al diccionario es muy cuidadoso: cuando un editor quiere agregar una nueva palabra al diccionario, busca en el corpus cada ejemplo que contenga esa palabra. La palabra aparece en una larga lista de oraciones y el editor decide cuál de todos los ejemplos expresa mejor el sentido que está buscando en esa palabra. Todos los ejemplos del diccionario Cobuild muestran patrones gramaticales típicos, vocabulario típico y contextos típicos para cada palabra. En consecuencia, Cobuild presenta una cantidad exhaustiva del vocabulario inglés derivado de observaciones directas del lenguaje.

3.1.1. Método de construcción

En 1987 se publicó el diccionario Cobuild basado en un corpus de 20 millones de palabras. A continuación se construyó un nuevo corpus, el Bank of English con alrededor de 650 millones de palabras. La nueva edición del diccionario Cobuild se basa en este nuevo corpus. La construcción de Cobuild fué un proceso en donde se decidió que palabras y frases presentes en el corpus incluir. Luego se examinó el lenguaje palabra por palabra y frase por frase con el objetivo de dar clara cuenta de cada significado y uso. Entonces para cada entrada se escribió la definición, se eligieron ejemplos típicos, y se agregó información sobre la pronunciación, la gramática, semántica, pragmatismos y frecuencia.

3.1.2. Evidencia

Un diccionario debe comenzar por la evidencia, los hechos. Los hablantes de un lenguaje conocen mucho sobre éste porque cada día leen y hablan sin esfuerzo durante horas. Sin embargo no son capaces de explicar exactamente que es lo que hacen. La mayoría de las personas no son conscientes de la habilidad que poseen para utilizar un lenguaje; no pueden examinarlo en detalle, simplemente lo utilizan para comunicarse. Aquellos que aprenden a observar el lenguaje cuidadosamente pueden expresar y organizar algunos de los hechos sobre éste basados en la experiencia. De todas maneras hay muchos hechos sobre el lenguaje que no pueden ser descubiertos simplemente pensando y reflexionando sobre él, incluso leyendo y escuchando muy atentamente. Es por eso que Cobuild empleó las computadoras para identificar estos hechos.

3.1.3. Un corpus

El resultado fué que Cobuild estableció un nuevo tipo de evidencia, una colección de textos en inglés llamado corpus ubicado en una computadora de manera que pueda consultarse instantáneamente. Los creadores de Cobuild sabían que necesitaban millones de palabras de inglés, hablado y escrito, americano y británico, formal e informal, sobre hechos y sobre ficción, étc. Esta evidencia reunida durante varios años, permitió encontrar las palabras y expresiones más utilizadas. Cuando una palabra tiene varios significados existe la capacidad de ver cuales son los significados importantes, y que frases se deben incluir. Tomaron como filosofía y fueron conscientes de que todos los detalles de un uso natural de una palabra son escenciales y no pueden ser falsificados. Se dieron cuenta de que debían utilizar ejemplos reales siguiendo la tradición de los grandes lingüistas, en lugar de crearlos.

3.1.4. The Bank of English

Hace varios años que se ha hecho mucho más fácil reunir grandes cantidades de lenguaje hablado y escrito. Los editores de libros, revistas y diarios tomaron consciencia de la gran cantidad de lenguaje que pasaba a través de sus manos y de las muchas buenas razones para conservarlo en formatos electrónicos. De repente apareció un negocio para el lenguaje electrónico entre la gente que quería encontrar o verificar oraciones, particularmente en las noticias, revistas y lenguaje legal. Gradualmente millones de palabras comenzaron a estar disponibles para los estudiosos del lenguaje. Hoy en día el problema no es encontrar el lenguaje sino manejarlo y realizar selecciones sensibles y balanceadas para las tareas analíticas. Diseñando el corpus The Bank of English se balancearon un número de factores(inglés hablado y escrito, americano y británico y otras características: hablantes de comunidades nativas, libros y revistas y más clasificaciones dentro de ésas)

Dentro del componente hablado, el tipo de lenguaje más dificil de recolectar fué como siempre la conversación informal grabada en la vida diaria de la gente común, sin pensar de que su lenguaje está siendo preservado en un corpus. Cada conversación tiene que ser grabada y transcripta por expertos para luego ser ingresada en una computdora. Esta clase de lenguaje improvisado es de un interés particular para los constructores de diccionarios. El Bank of English

cuenta con un total de 15 millones de palabras de este tipo de grabaciones de lenguaje hablado.

3.1.5. La lista de palabras principales

Es mucho más fácil decidir qué palabras y frases incluir y cuales omitir, cuando se tienen cifras exactas sobre una cantidad tan grande de lenguaje. Las computadoras pueden verificar instantáneamente la actividad del lenguaje de miles de hablantes y escritores. Un diccionario (incluso un gran diccionario) es capaz de presentar solo los hechos más importantes del lenguaje y los compiladores necesitan buena evidencia para sus selecciones. Cobuild se especializa en presentar las palabras y frases que son frecuentes en el uso diario. Lejos de ser un registro histórico del lenguaje es más bien una muestra del lenguaje contamporáneo.

3.1.6. Frecuencia

Cobuild brinda información sobre la frecuencia de las palabras principales. Se establecieron 5 bandas de frecuencias. Comenzando con las palabras muy comunes (las de mayor frecuencia), oscila entre un vocabulario básico a uno intermedio hasta cubrir el vocabulario. Las palabras principales sin marca de frecuencia son las menos comunes, sin embargo vale la pena incluirlas en el diccionario. El punto es que el idioma inglés utiliza un número bastante pequeño de palabras para la mayoría de los propósitos pero también tiene disponible un rico y amplio vocabulario. Por ejemplo *be* y *because* pertenecen naturalmente a la banda de mayor frecuencia, por el otro lado, palabras como *barracuda*, *basalt* y *basrelief* no son tan frecuentes. Estas últimas son claramente utilizadas en ocasiones particulares. Cabe aclarar que incluso las palabras infrecuentes incluídas en el diccionario fueron seleccionadas por su utilidad relativa entre miles de palabras posibles.

Entonces Cobuild cuenta con un sistema de frecuencia que marca las palabras principales: una marca significa que la palabra tiene una alta frecuencia y por lo tanto es una palabra común dentro del lenguaje inglés. Dos o más marcas significan que la palabra es parte esencial del vocabulario, cuantas más marcas posee, menos frecuente es.

3.1.7. Ejemplos

Todos los ejemplos fueron seleccionados del corpus The Bank of English. Como se dijo anteriormente, los ejemplos son seleccionados cuidadosamente para mostrar los patrones que aparecen frecuentemente junto a una palabra o frase. El compilador tiene docenas, centenas o miles de ejemplos disponibles y rapidamente escoge los *colocados* [7] y las estructuras típicas en donde la palabra o frase ocurre más a menudo.

Esto significa que los ejemplos cumplen varias funciones. Desde luego ayudan a mostrar el significado de la palabra exhibiendo su uso. Las investigaciones incluso sugieren que un gran número de usuarios comienza con los ejemplos antes que con el significado. Las definiciones de Cobuild son bastante claras por sí mismas y los ejemplos muestran el fraseo característico alrededor de la

[7] Palabras particulares ubicadas cerca de la palabra principal

palabra. Como los ejemplos son piezas de texto genuinas y han sido elegidas cuidadosamente en base al uso de la palabra, pueden ser de confianza para exhibir la palabra en un contexto natural.

3.1.8. Información gramatical

Casi cada sentido de cada entrada en el diccionario Cobuild tiene junto a esta una clasificación gramatical, usualmente una clase de palabra y a menudo también una nota estructural. Esta es la información sobre la que se sustenta este trabajo, ya que en base a ella se construirá el nuevo corpus de entrenamiento.

3.1.9. Pragmatismo

Muchos usos de una palabra necesitan más de una frase para explicar apropiadamente su significado. La gente utiliza el lenguaje para realizar muchas cosas: hacer invitaciones, expresar sus sentimientos, enfatizar que es lo que está diciendo, étc. El corpus nos brinda evidencia para tales usos que son difíciles de tomar desde cualquier otra fuente.

El estudio y descripción de las formas en que la gente utiliza el lenguaje para realizar cosas es llamado pragmatismo. Este aspecto del lenguaje es muy importante y fácil de omitir. Esto sucede cuando el lenguaje está dando significado adicional. Cobuild posee mucha información sobre pragmatismo y la expone mediante un símbolo especial en cada entrada. Por ejemplo *and things like that* es definido como una expresión utilizada para ampliar el rango de una lista.

3.1.10. Definición del estilo

La característica más distintiva de Cobuild en su primera versión fué el uso de frases completas en las definiciones. El significado de una palabra fué establecido de la forma en que una persona ordinaria podría explicárselo a otra.

Generalmente los diccionarios ofrecen definiciones breves y tradicionales, mientras que Cobuild expone definiciones realmente amplias y ricas. Si se observan detenidamente las definiciones particulares se puede apreciar que cada palabra es elegida para ilustrar ciertos aspectos del significado. Y en la medida en que es posible, las palabras utilizadas en una definición son más frecuentes que la palabra que está siendo definida.

Las definiciones cortas no pueden decir demasiado. Por ejemplo, el primer sentido de verbo de *mean* podría ser definido como solo *signify*, que es cierto, pero no es todo lo que se puede decir. Cobuild expone esto: *If you want to know...* es decir que ese sentido surge cuando alguien está en la búsqueda de información. La palabra *if* indica que esta es una opción, pero una perfectamente normal, y *you* nos dice que no es una característica de ningún grupo particular de gente (compararlo con *if a policeman arrests you...*). Entonces la definición dice lo que alguien puede querer saber sobre el significado de una *palabra, código, señal o gesto*, indicando que esas son las típicas clases de temas que serán encontradas con este sentido de *mean*. Solo después de toda esta información viene el equivalente de *signify*: *lo que se refiere a o a que mensaje transmite*. Entonces hay 12 palabras antes de la palabra principal en este sentido, pero cada una de ellas transmite información vital que sería muy difícil de incluir en una definición corta.

3.2. Extracción de la información

El diccionario COBUILD guarda su información en un archivo de texto plano con un formato particular. El primer desafío de este trabajo fué comprender y extraer la información almacenada en ese archivo. A continuación se muestra un pequeño fragmento del mismo para ejemplificar

```
DICTIONARY_ENTRY
ace
aces
*e*!is
If you are or come within an ace of something, you very nearly do or experience it.
He came within an ace of being run over.
phrase: verb inflects
phrase

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
```

Cada entrada arriba presentada tiene la característica de poseer una cantidad variable de campos y no es posible identificarlos exactamente. Sin embargo, contienen algunos rasgos comunes: 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
If you are or come within an ace of something, you very nearly do or experience it.
⟶ definición
He came within an ace of being run over. ⟶ ejemplo
phrase: verb inflects ⟶ etiqueta
```

`phrase` ⟶ *etiqueta*

Estas entradas, que conforman el diccionario COBUILD y que constituyen la fuente de información principal sobre la cual se basa este trabajo, fueron cuidadosamente procesadas y refinadas intentando mantener toda la información disponible. El primer desafío de esta etapa consistió en recuperar las entradas con toda la información gramatical disponible; explícita e implícita. Una primer tarea fué reconocer y registrar información relacionada a las formas flexionadas de la palabra (plurales, pasados, etc.), es decir, obtener información gramatical implícita.

3.2.1. Reconocimiento de formas flexionadas

En muchas entradas del diccionario COBUILD ocurre la palabra, uno o más ejemplos en donde ésta aparece con cierto sentido (indicado por medio de etiquetas gramaticales) pero dentro de los ejemplos hay apariciones de formas flexionadas. Tomemos la siguiente entrada:

`DICTIONARY_ENTRY`
`bite` ⟶ *palabra*
`bites, biting, bit, bitten` ⟶ *formas flexionadas*
`b*a*!it` ⟶ *pronunciación*
`If an object or surface bites, it grips another object or surface rather than slipping on`
⟶ *definición*
`Let the clutch in slowly until it begins to bite.` ⟶ *ejemplo*
`verb` ⟶ *etiqueta para la definición*
`verb` ⟶ *etiqueta para el ejemplo*

Aquí arriba se puede observar una entrada del diccionario para la palabra *bite*, que contiene la definición y un ejemplo de esta palabra con sus respectivas etiquetas:

(1) *If an object or surface bites, it grips another object or surface rather than slipping on it or against it.*

(2) *Let the clutch in slowly until it begins to bite.*

En (2) aparece la palabra *bite* en su forma regular con la etiqueta *verb* mientras que en (1) aparece la forma flexionada *bites* con la etiqueta *verb*. En este caso (1) está ofreciendo más información gramatical que la expuesta por medio de la etiqueta. Reconociendo la forma flexionada (*bites*) podemos adicionarle información extra a la etiqueta *verb*; en vez de guardar la etiqueta de Tree Bank correspondiente a *verb* (VB), en este caso guardaríamos la etiqueta VBZ (verbo de tiempo presente en tercera persona singular) que contiene más información gramatical que VB.

Las entradas de COBUILD exponen las formas derivadas de la palabra que pueden contener los ejemplos. En el ejemplo presentado anteriormente la palabra es *bite* y las formas derivadas de *bite* que muestra la entrada son *bites*, *biting*, *bit* y *bitten*. Con esta información y la etiqueta que fué anotada en COBUILD (*verb*) se puede inferir y generar etiquetas de Tree Bank con información adicio-

nal. Como ya se mencionó anteriormente, en este caso la forma *bites* (derivada de la palabra *bite*) que aparece en la definición posee la etiqueta *verb*. La tarea aquí será reconocer que *bites* es un verbo de tiempo presente en tercera persona singular a partir de que *bites* está etiquetada como verbo y de que la palabra de la cual deriva es *bite*. Es decir, inferir el tipo de la forma derivada a partir de la palabra y la etiqueta asignada por COBUILD.

Con el objetivo de identificar las formas derivadas de una palabra se desarrollaron reglas y métodos para su reconocimiento, buscando preservar y aprovechar toda la información que ofrece COBUILD. Entonces, a partir de esta información: la palabra, la forma en que ocurre y la etiqueta asignada se aplican las siguientes reglas para reconocer información adicional a la etiqueta gramatical.

Algoritmo 2 Reconocimiento de formas derivadas

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**

Aplicando algoritmos de extracción y el algoritmo de reconocimiento de formas derivadas explicado anteriormente se obtiene un nuevo corpus parcialemente anotado a partir del diccionario Cobuild. A continuación este corpus será procesado y utilizado como corpus de entrenamiento.

3.3. Traducción de etiquetas

Para cada una de sus definiciones, el diccionario COBUILD expone información gramatical expresada mediante etiquetas. Estas etiquetas gramaticales poseen un formato propio. Por ejemplo en la siguiente entrada de COBUILD para la palabra *canary*

```
DICTIONARY_ENTRY
canary
canaries
A canary is a small yellow bird which sings beautifully.
People sometimes keep canaries in cages as pets.
countable noun
```

`noun`

Se expone la definición (1) y un ejemplo (2), ambos con información gramatical sobre la palabra:

(1) *A canary is a small yellow bird which sings beautifully.*
(2) *People sometimes keep canaries in cages as pets.*

Se puede apreciar la etiqueta *noun* asignada por COUBUILD para canary.

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 poder medir posteriormente los resultados. Es por eso que se tomó la decisión de traducir estas etiquetas propias de COBUILD en etiquetas de Tree Bank, conjunto con el cual está anotado el gold standard.

A continuación se presenta la tabla de traducción empleada:

Cuadro 3: *Tabla de traducción de etiquetas*

Etiqueta COBUILD	**Etiqueta PenTreeBank**
coordinating conjunction	CC
number	CD
determiner	DT
determiner + countable noun in singular	DT
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

Cuadro 3: *Tabla de traducción de etiquetas*

Etiqueta COBUILD	**Etiqueta PenTreeBank**
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
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 $noungrouporreflexive$	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

Cuadro 3: *Tabla de traducción de etiquetas*

Etiqueta COBUILD	**Etiqueta PenTreeBank**
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
phrase + noun group	
convention	
combining form	
prefix	
phrasal verb	
other	
phrase	
suffix	
wh	
phrase after noun	
phrase + reporting clause	

3.4. Nuevo Corpus generado

A partir del corpus parcialmente anotado obtenido en el proceso de extracción, se completarán las anotaciones automáticamente con un etiquetador gramatical manteniendo las etiquetas gramaticales obtenidas a partir de la información procedente del diccionario COBUILD. Es decir, una vez finalizado el proceso de extracción de información desde el diccionario, se obtiene un corpus nuevo con las etiquetas gramaticales correspondientes a las palabras definiadas en el diccionario. A continuación se exhibe un fragmento del mismo:

A
canary NN
is
a
small
yellow
bird
which

sings
beautifully
.
People
sometimes
keep
canaries NNS
in
cages
as
pets
.

Este es el resultado de extracción y reconocimiento de formas flexionadas correspondiente a la entrada de COBUILD:

```
DICTIONARY_ENTRY
canary
canaries
A canary is a small yellow bird which sings beautifully.
People sometimes keep canaries in cages as pets.
countable noun
noun
```

Se puede apreciar que se ha reconocido *canaries* como el plural de *canary* (etiqueta NNS) y que se han reconocido y extraído los ejemplos de estas palabras asignando las etiquetas gramaticales traducidas a partir de las etiquetas del diccionario correspondientes a *canary* (countable noun/NN) y *canaries* (noun/NNS).

El próximo paso será el de completar las anotaciones gramaticales para todas las palabras restantes. Este proceso se realiza anotando el corpus plano (sin las etiquetas obtenidas de COBUILD) con el etiquetador gramatical automático TnT. Luego se une este corpus anotado por TnT con el corpus anotado parcialmente procedente de Cobuild, preservando todas las etiquetas del diccionario. El resultado que se muestra a continuación es un nuevo corpus obtenido a partir de Cobuild, con las anotaciones que este provee y completado con anotaciones obtenidas mediante etiquetación automática utilizando TnT.

A	DT
canary	NN
is	VBZ
a	DT
small	JJ
yellow	JJ
bird	NN
which	WDT
sings	VBZ
beautifully	RB
.	.
People	NNS
sometimes	RB
keep	VB
canaries	NNS
in	IN
cages	NNS
as	IN
pets	NNS
.	.

3.5. Experimentación

3.5.1. 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
cat NN
is
a
small
furry
animal
with
a
tail
,
whiskers
,
and
sharp
claws
that
kills
smaller
animals
such
as
mice
and
birds
.
Cats NNS
are
often
kept
as
pets
.
She

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

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) [8] y luego se realizará la comparación.
La matriz de confusión[9] generada a partir de dicha comparación es la siguiente:

Cuadro 4: *Matriz de confusión para etiquetas extraídas de COBUILD vs WSJ*

	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 %

[8] Wall Street Journal es un corpus anotado, parte del Penn Treebank
[9] Las matrices de confusión presentadas de aquí en adelante contienen las primeras 15 etiquetas de mayor error

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.

3.5.2. 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:

Cuadro 5: *Matriz de confusión para WSJ etiquetado con TnT (entrenado con WSJ)*

	JJ	**NN**	**NNP**	**VBN**	**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
VBN	.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 6: *Matriz de confusión para WSJ etiquetado con TnT (entrenado con WSJ + NFI)*

	JJ	**NN**	**NNP**	**IN**	**RB**	**RP**	**VBN**	**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	-	-	-	-

Cuadro 6: *Matriz de confusión para WSJ etiquetado con TnT (entrenado con WSJ + NFI)*

	JJ	**NN**	**NNP**	**IN**	**RB**	**RP**	**VBN**	**VBD**	**NNPS**	**VBG**
RP	.0008	.0002	.0000	.0425	.0308	-	-	-	-	-
VBN	.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 7: *Matriz de confusión para la 2 mitad de WSJ etiquetado con TnT (entrenado con la primer mitad de WSJ)*

	JJ	**NN**	**NNP**	**VBN**	**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
VBN	.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 8: *Matriz de confusión para la 2 mitad de WSJ etiquetado con TnT (entrenado con la primer mitad de WSJ + NFI)*

	JJ	**NN**	**NNP**	**VBD**	**VBN**	**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	-	-	-
VBN	.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 9: *Matriz de confusión para la 1 mitad de WSJ etiquetado con TnT (entrenado con la 2 mitad de WSJ)*

	JJ	**NN**	**VBN**	**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
VBN	.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
NNPS	.0001	-	-	-	.0354	-	-	-	-	-
RP	.0007	.0002	-	-	.0001	.0213	.0289	-	-	-
VB	.0035	.0110	.0014	.0009	.0007	.0009	.0002	-	-	-

Porcentaje de aciertos: 96,20 %

Cuadro 10: *Matriz de confusión para la 1 mitad de WSJ etiquetado con TnT (entrenado con la 2 mitad de WSJ + NFI)*

	JJ	**NN**	**NNP**	**VBD**	**VBN**	**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	-	-	-	-	-
VBN	.0204	.0013	.0005	.0373	-	-	.0001	-	-	-
IN	.0022	.0005	.0005	-	-	-	.0376	.0080	-	.0003
RB	.0150	.0017	.0010	-	-	.0259	-	.0018	-	.0000

Cuadro 10: *Matriz de confusión para la 1 mitad de WSJ etiquetado con TnT (entrenado con la 2 mitad de WSJ + NFI)*

	JJ	**NN**	**NNP**	**VBD**	**VBN**	**IN**	**RB**	**RP**	**NNPS**	**VBG**
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 11: *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 %
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.

3.5.3. Tercer experimento

3.6. Conclusiones