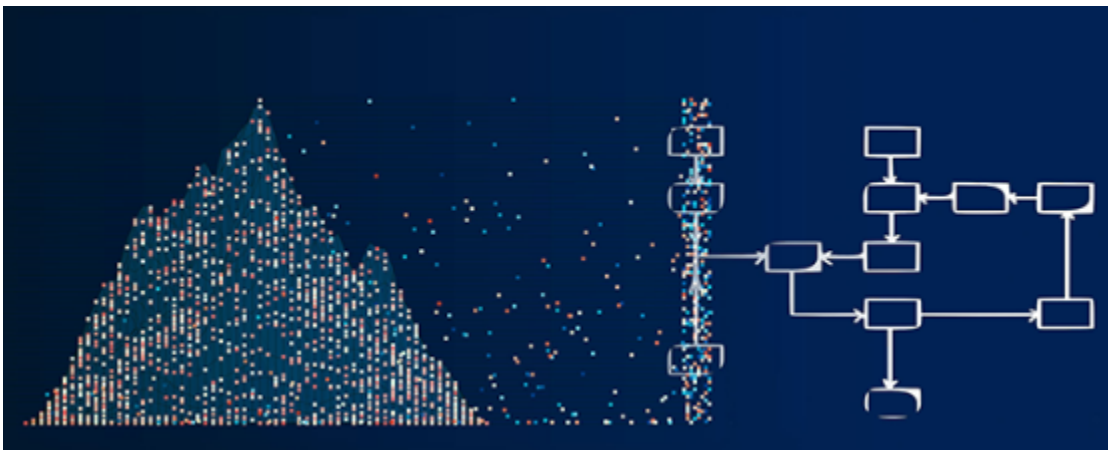


Comparativa de etiquetadores estadísticos



Asignatura: Descubrimiento de información en textos.
Profesores: M. Lourdes Araujo Serna, Raquel Martinez, Unanue
Alumno: Macià Salvà Salvà

Indice

Indice	1
Texto elegido	2
Etiquetadores	2
Stanford POS tagger	2
Instalación y ejecución	3
Ejecución con el texto elegido.	5
TreeTagger	6
Instalación y ejecución	7
Ejecución con el texto elegido.	7

Texto elegido

Se ha seleccionado este texto en inglés porque es sencillo y facilita el análisis sintáctico. Además, se eligió este texto porque el inglés no es mi lengua materna, lo que hace importante trabajar con un contenido más accesible para asegurar una mejor comprensión y análisis.

"One day, Anna wakes up early in the morning. She looks out the window and sees the sun shining. Birds are singing in the trees. Anna decides to go for a walk. She puts on her shoes and leaves the house. The street is quiet, and the air is fresh. Anna walks to the park nearby. She sees children playing with a ball. There is a small dog running around. Anna smiles and sits on a bench. She takes an apple from her bag and eats it. A man rides a bicycle past her. The wind blows softly, moving the leaves. Anna feels happy and calm. She decides to stay a bit longer and watch the clouds.

The clouds move slowly across the blue sky. A squirrel runs up a tree near her. Anna hears the sound of water from a small fountain. She stands up and walks to the fountain. She touches the cool water and feels refreshed.

Nearby, an old couple is feeding birds. The birds flap their wings and chirp loudly. Anna waves to the couple and they smile back. She then walks along a path filled with flowers. Bees buzz around, collecting nectar. Anna stops to smell a red rose. Its scent is sweet and pleasant."

Etiquetadores

Stanford POS tagger

Part-Of-Speech Tagger ([POS Tagger](#)) es un software que analiza un texto en un idioma determinado y asigna categorías gramaticales a cada palabra, como sustantivo, verbo, adjetivo, entre otras, usando etiquetas gramaticales más específicas como '*sustantivo plural*'. Esta implementación, está hecha a partir de modelos log-lineales.

El sistema requiere Java 8 o superior y memoria entre 60 y 200 MB para ejecutar un modelo entrenado; entrenar un modelo puede requerir al menos 1 GB de memoria. Se incluyen modelos entrenados para inglés, chino, árabe, francés, alemán y español. El etiquetador puede ser reentrenado para cualquier idioma si se cuenta con textos anotados para ese idioma.

Los modelos de inglés usan el conjunto de etiquetas de Penn Treebank, mientras que los modelos de francés, alemán y español emplean el conjunto UD (v2).

Instalación y ejecución

Para instalar y ejecutar este software es bastante sencillo, simplemente clickeamos en este [link](#) y nos bajamos el archivo pertinente.

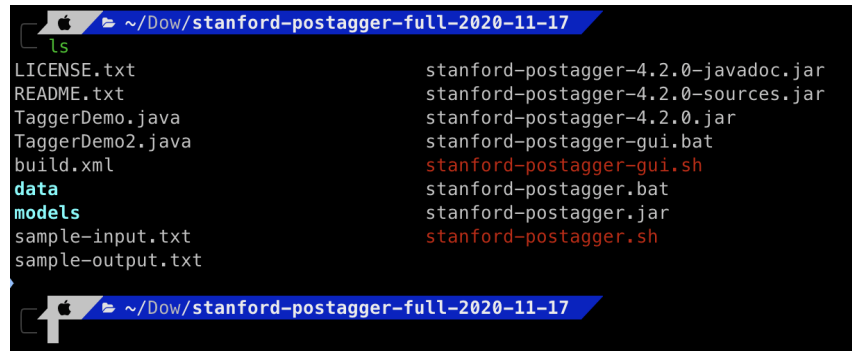
Download

Download Stanford Tagger version 4.2.0 [75 MB]

The full download is a 75 MB zipped file including models for English, Arabic, Chinese, French, Spanish, and German. This software provides a GUI demo, a command-line interface, and an API. Simple scripts are included to invoke the README.txt.

Imagen 1. Archivo de descarga del software.

Una vez descargado el archivo, procedemos a descomprimirlo, ya que viene en formato zip, y nos dirigiremos a la terminal en la ubicación del proyecto. Allí encontraremos diversos archivos y ejecutables.



```
ls
LICENSE.txt          stanford-postagger-4.2.0-javadoc.jar
README.txt           stanford-postagger-4.2.0-sources.jar
TaggerDemo.java      stanford-postagger-4.2.0.jar
TaggerDemo2.java     stanford-postagger-gui.bat
build.xml             stanford-postagger-gui.sh
data                  stanford-postagger.bat
models                stanford-postagger.jar
sample-input.txt     stanford-postagger.sh
sample-output.txt
```

Imagen 2. Ficheros del proyecto descargado.

Para ejecutar el programa de manera sencilla, seleccionaremos el ejecutable con interfaz gráfica, ya que esto facilitará su uso. Para ello, debemos ejecutar un comando específico, lo que abrirá una interfaz donde, de manera predeterminada, se cargará un modelo en inglés.

```
./stanford-postagger-gui.sh
```

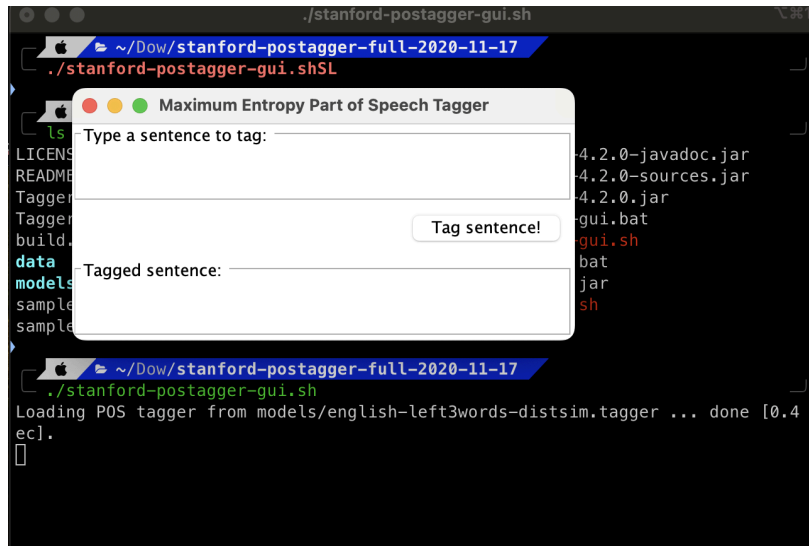


Imagen 3. Ejecución del programa con GUI.

Ahora, simplemente podemos añadir un texto y hacer clic en el botón "Tag Sentence". Esto nos permitirá ver el resultado de la ejecución de manera inmediata.

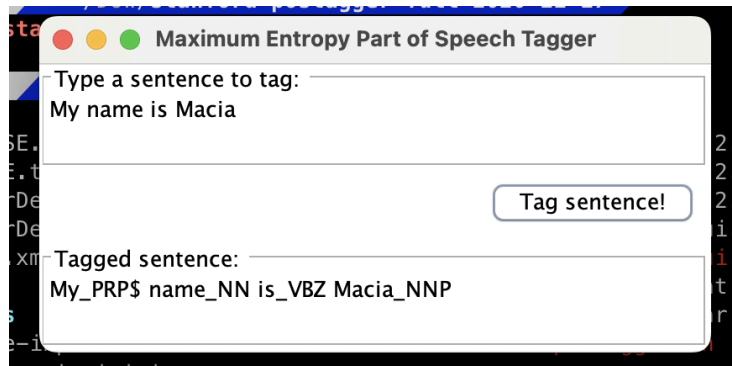


Imagen 4. Ejecución del programa con un ejemplo.

En esta ejecución encontramos que:

- **PRP\$**: Es una etiqueta de pronombre posesivo. En este caso, "My" es un pronombre posesivo, lo que significa que indica posesión. "PRP\$" significa "pronoun, possessive".
- **NN**: Es una etiqueta que indica un **sustantivo singular (singular noun)**. "name" está etiquetado como "NN", lo que significa que es un sustantivo singular.
- **VBZ**: Es una etiqueta que indica un **verbo en tercera persona singular en presente (verb, third person singular present)**. En este caso, "is" es un verbo en tercera persona

singular en presente.

- **NNP**: Es una etiqueta que indica un **sustantivo propio singular** (proper noun, singular). "Macia" está etiquetado como "NNP", lo que significa que es un sustantivo propio y singular.

Ejecución con el texto elegido.

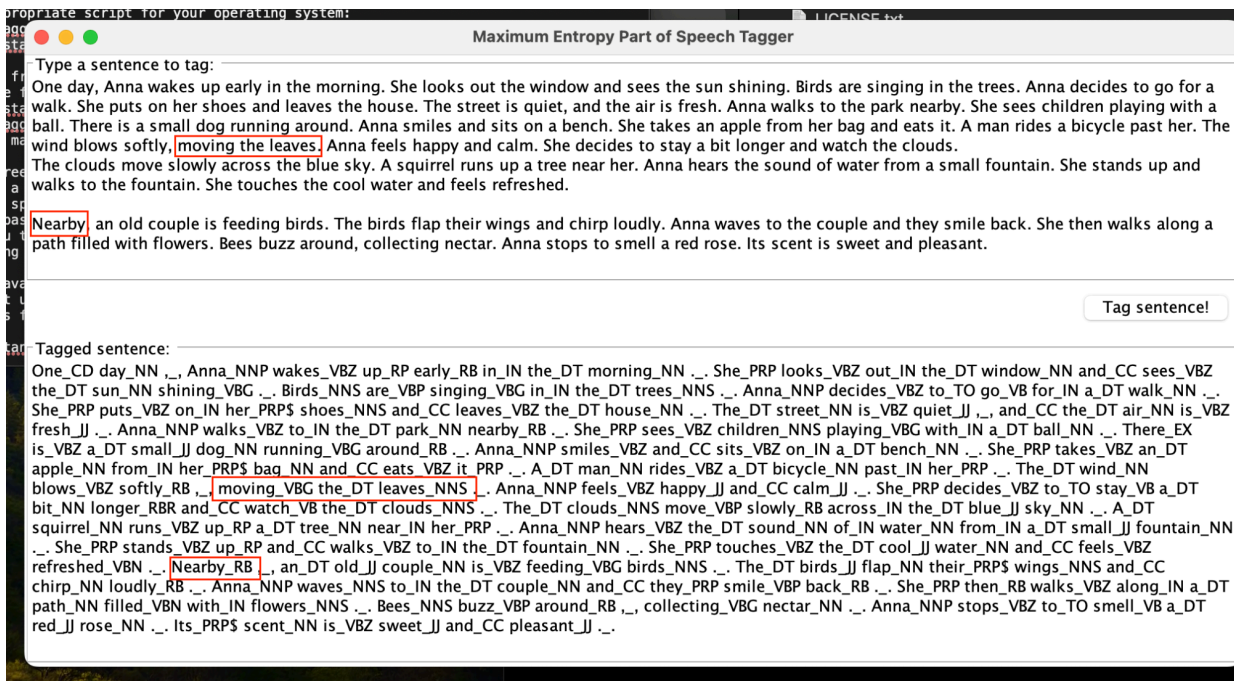


Imagen 5. Ejecución del texto elegido.

El etiquetado de partes del discurso en el texto es, en su mayoría, correcto, aunque hay algunos casos que podrían generar confusión. Un ejemplo de esto es la palabra **"Nearby"**, que está correctamente etiquetada como un adverbio (**RB**), indicando lugar. Sin embargo, si se encuentra en una estructura más compleja o en frases con múltiples modificadores, podría ser interpretado de forma diferente, lo que podría generar ambigüedad en el análisis gramatical.

Otro caso donde podría haber confusión es con el sustantivo **"leaves"** en la frase "moving the leaves". Aunque está correctamente etiquetado como sustantivo plural (**NNS**), la palabra "leaves" aparece junto al verbo en gerundio **"moving"**. Dependiendo del contexto, "leaves" podría funcionar como el objeto directo de un verbo o como parte de una estructura más compleja, lo que podría llevar a confusión sobre su función gramatical en la oración.

A pesar de estos posibles puntos de ambigüedad, el etiquetado de las partes del discurso es generalmente adecuado y refleja correctamente las funciones gramaticales de las palabras en la mayoría de los casos.

TreeTagger

El [TreeTagger](#) es una herramienta para anotar textos con información sobre partes del discurso y lemas. Fue desarrollada por Helmut Schmid en el proyecto TC del Instituto de Lingüística Computacional de la Universidad de Stuttgart. Este etiquetador ha sido utilizado exitosamente para etiquetar textos en varios idiomas, como alemán, inglés, francés, italiano, español, ruso, chino, entre otros. Además, es adaptable a otros idiomas si se dispone de un léxico y un corpus de entrenamiento etiquetado manualmente.

Ejemplo de salida del TreeTagger:

Palabra	Pos	Lema
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Tagger también puede usarse como **chunker** para idiomas como inglés, alemán, francés y español.

Instalación y ejecución

Para instalar TreeTagger de forma gratuita, puedes seguir los pasos detallados en el siguiente [enlace](#). Como alternativa, se buscó directamente en Google si la herramienta estaba disponible en línea, y efectivamente, se encontró una versión online.

En esta [página](#), puedes subir un archivo de texto o pegar el texto directamente en el campo correspondiente. Al hacer clic en ‘process’, se generará y descargará un archivo .csv con la solución al problema.

Ejecución con el texto elegido.

central.uclouvain.be/treetagger/

Online TreeTagger

Annotate your texts with part-of-speech and lemma information using [TreeTagger](#).

Type a text [Upload a file](#)

Text to process*

One day, Anna wakes up early in the morning. She looks out the window and sees the sun shining. Birds are singing in the trees. Anna decides to go for a walk. She puts on her shoes and leaves the house. The street is quiet, and the air is fresh. Anna walks to the park nearby. She sees children playing with a ball. There is a small dog running around. Anna smiles and sits on a bench. She takes an apple from her bag and eats it. A man rides a bicycle past her. The wind blows softly, moving the leaves. Anna feels happy and calm. She decides to stay a bit longer and watch the clouds. The clouds move slowly across the blue sky. A squirrel runs up a tree near her. Anna hears the sound of water from a small fountain. She stands up and walks to the fountain. She touches the cool water and feels refreshed.

Nearby, an old couple is feeding birds. The birds flap their wings and chirp loudly. Anna waves to the couple and they smile back. She then walks along a path filled with flowers. Bees buzz around, collecting nectar. Anna stops to smell a red rose. Its scent is sweet and pleasant.

Language of your text*

English

Process

Imagen 6. Se inserta el texto elegido.

treetagger_output

One	CD	one
day	NN	day
,	,	,
Anna	NP	Anna
wakes	VVZ	wake
up	RP	up
early	RB	early
in	IN	in
the	DT	the
morning	NN	morning
.	SENT	.
She	PP	she
looks	VVZ	look
out	RP	out
the	DT	the

Imagen 7. CSV descargado del texto elegido.

El término "waves" en la frase "Anna waves to the couple and they smile back." está etiquetado incorrectamente como "NNS" (sustantivo en plural), cuando en realidad debería ser un verbo en tercera persona "VVZ".

Anna	NP	Anna
waves	NNS	wave
to	TO	to
the	DT	the
couple	NN	couple
and	CC	and
they	PP	they
smile	VVP	smile
back	RB	back
.	SENT	.

Imagen 8. Error en el etiquetado de la palabra waves.

La palabra "buzz" en la frase "Bees buzz..." está etiquetada incorrectamente como "VVP" (verbo en plural), cuando debería ser "VV" o "VVZ", ya que describe una acción realizada por las abejas.

Bees	NNS	bee
buzz	VVP	buzz
around	RP	around
,	,	,
collecting	VVG	collect
nectar	NN	nectar
.	SENT	.

Imagen 9. Error en el etiquetado de la palabra buzz.

El término "red" en la frase "Anna stops to smell a red rose." está etiquetado incorrectamente como un sustantivo "NN", cuando debería ser "JJ" (adjetivo), ya que describe la característica del sustantivo "rose".

Anna	NP	Anna
stops	VVZ	stop
to	TO	to
smell	VV	smell
a	DT	a
red	NN	red
rose	VVD	rise
.	SENT	.

Imagen 10. Error en el etiquetado de la palabra red.

La palabra "rose" en la frase "red rose" está etiquetada incorrectamente como "VVD" (pasado del verbo "rise"), cuando debería ser "NN" (sustantivo), ya que se refiere a una flor.

Anna	NP	Anna
stops	VVZ	stop
to	TO	to
smell	VV	smell
a	DT	a
red	NN	red
rose	VVD	rise
.	SENT	.

Imagen 11. Error en el etiquetado de la palabra rose.