Descubrimiento de Información en Textos

Tarea del Tema 4: Comparativa de etiquetadores estadísticos

Gabriel Vázquez Torres

10 de Febrero de 2018

Resumen

En la siguiente página web: http://www-nlp.stanford.edu/links/statnlp.html
en la sección "Part of Speech Taggers" puedes encontrar numerosos etiquetadores estadísticos.Muchos de ellos se basan en modelos distintos (HMMs, Support Vector Machine, etc.), utilizan distintos corpus de entrenamiento, sirven para distintos idiomas, etc. En esta tarea debes comparar el comportamiento de al menos dos de ellos. Estúdialos, descríbelos (busca en la distribución y en la web detalles del modelo), y utilizalos para realizar el etiquetado de un pequeño texto (al menos 15 líneas), el mismo para ambos. Para ello asegurate que los etiquetados elegidos sirven para el mismo idioma. Debes elegir un texto en el que aparezcan palabras con más de una etiqueta léxica posible. Después compara los resultados: etiquetas utilizadas por cada etiquetador y precisión del etiquetado. Para analizar la corrección puedes utilizar un texto de un corpus del que conozcas el etiquetado correcto. En otro caso tendrás que realizar el etiquetado correcto manualmente.

Documentación a entregar:

- Descripción de los etiquetadores seleccionados
- Texto de prueba utilizado
- Resultado del etiquetado con cada etiquetador seleccionado
- Observaciones sobre la comparativa de los resultados

1. Introducción

En esta práctica vamos a analizar dos etiquetadores léxicos. El principal hecho de elegir ambos etiquetadores (tengo esta posibilidad por mi actual trabajo como investigador de Machine Learning en la Universidad de Sevilla) es la recomendación por parte de doctores y profesores universitarios sobre la temática ya que, según comentan, ellos trabajan o han trabajado con ellas y son bastantes utilizadas en el ámbito de los etiquetadores léxicos. Me recomendaron alguna más pero no están en página web ofrecida por enunciado. Los etiquetadores léxicos son:

1. TreeTagger
2. SVMTool

2. Etiquetadores léxicos

2.1. TreeTagger

Este proyecto fue llevado a cabo por la Universidad de Stuttgart, en concreto por el Instituto de la Lingüística Romance y el Instituto de Ciencias de la Computación departamento de inteligencia artifical.Recibió financiación al 100 % por el Ministerio de Ciencia e Investigación del Estado federado de Baden-Württemberg (MWF, Stuttgart), en 1993-1994 y 1995-1996. Entre 1993-1994 el proyecto recogió todo el material de texto para el alemán, francés e italiano, desarrollando así una representación de los textos y las marcas, junto con un lenguaje de consulta y un sistema de acceso para la exploración de corpus lingüísticos de los textos. La separación entre Los textos y análisis de resultados, tiene lugar por razones de flexibilidad y extensibilidad del sistema. Esto es

posible gracias a un enfoque particular para el almacenamiento y la representación.Actualmente algunos de los componentes de la herramienta están todavía en fase de desarrollo, un idioma específico que va desde el análisis morfosintáctico de análisis parciales, y de información mutua, la puntuación T-, la extracción de coubicación y la agrupación de etiquetado basados en HMM y etiquetado de ngrama. Hoy en día se siguen realizando investigaciones sobre modelos estadísticos para los sintagmas nominales y las colocaciones verbo-objeto.

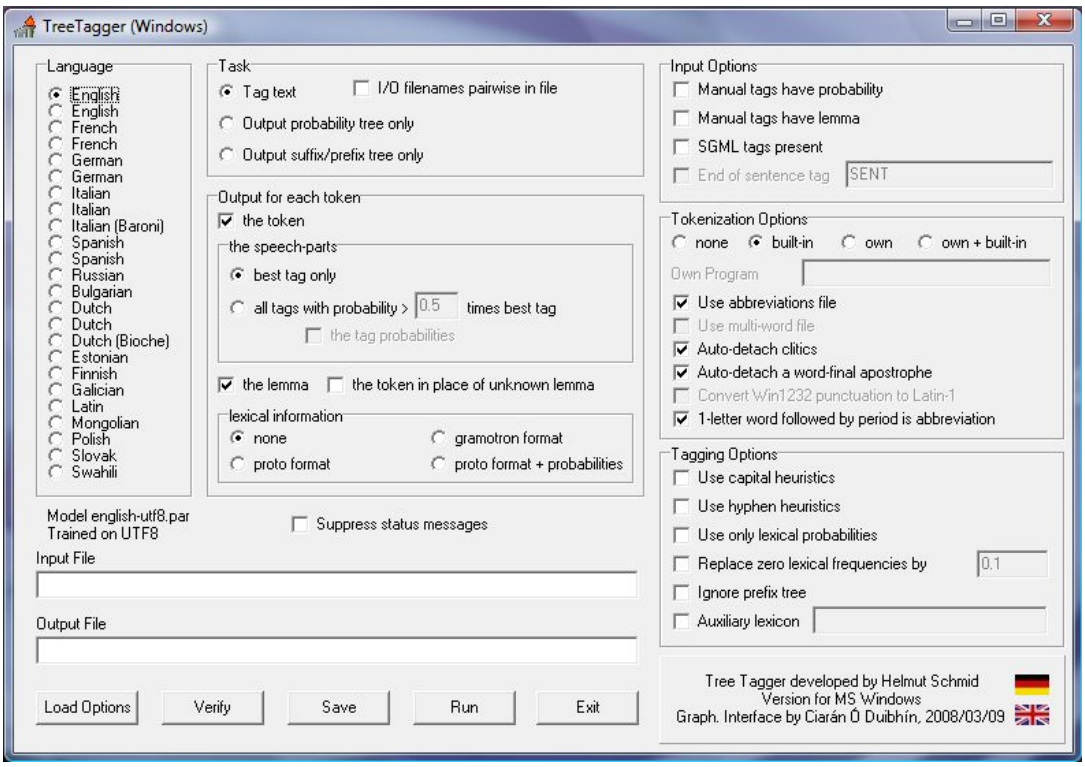

Figura 1: Menú principal de TreeTagger.

2.2. SVMTool

Esta herramienta está compuesta por tres componentes principales, el aprendedor (SVMTlearner), el etiquetador (SVMTagger) y el evaluador (SVMTeval). Antes de llevar a cabo el etiquetado, los modelos de SVM aprenden de distintos corpus usando el componente de aprendizaje utilizando diferentes estrategias.Así pues usando el SVMTagger, se selecciona la mejor estrategia para la propuesta de etiquetado que vamos a probar. Finalmente, dado un corpus anotado de forma correcta, realizado con el componente SVMTool, es evaluado por el SVMTEval.

2.2.1. SVMTlearn

El entrenamiento de los clasificadores SVM se realiza mediante un conjunto de ejemplos dado. El SVMTlearn tiene un fichero de configuración, en el que se pueden cambiar distintos parámetros que se enumeran a continuación:

- Sliding window: el tamaño de la ventana deslizante se puede ajustar. Se puede cambiar el tamaño de esta ventana, que por defecto es 5.
- Feature set: la ventana deslizante recogerá tres tipos de características distintas: características de palabras, de POS (Part of Speech) y sufijos y ortografía.
- SVM model compression: módulo que comprime los modelos de SVM para mejorar su rapidez.
- C parameter tunning: permite personalizar una serie de parámetros a la hora de realizar las pruebas.
- Dictionary repairing: permite reparar el diccionario.

- Ambigous classes: en ocasiones se encuentran palabras con ambigüedades que mediante este parámetro se pueden subsanar.
- Open classes: estas clases son para las palabras que son desconocidas.
- Backup lexicon: contiene palabras que no están normalmente en un corpus.

2.2.2. SVMTagger

Dado un corpus y una ruta para un modelo de aprendizaje SVM aprendido, se crea un etiquetado POS de una secuencia de palabras. El etiquetado está basado en una ventana deslizante que da una visión del contexto que es considerado.Este componente también tiene una serie de opciones como por ejemplo:

- Tagging scheme: se pueden utilizar dos esquemas de etiquetado distintos (Greedy y sentence-level)
- Tagging direction: la dirección del etiquetado, de izquierda a derecha, o de derecha a izquierda, o una combinación de ambos.
- One pass / two passes: Otro camino para conseguir un etiquetado en dos pasos.
- SVM Model Compression: se ignoran vectores grandes que ralentizan el etiquetado.
- Backup lexicon : de nuevo, contiene palabras que no están en los corpus.

2.2.3. SVMTeval

El SVMTeval, evalúa la ejecución y los resultados del mismo. Es un componente muy útil para personalizar los parámetros del sistema, como por ejemplo el parámetro C. En un idioma pueden existir palabras que tengan la misma forma pero que signifiquen cosas distintas y así se pueden producir ambigüedades. Con lo cual, una misma frase, puede ser evaluada de maneras distintas.

2.2.4. Conclusión de SVMTool

Como comentaremos más adelante esta herramienta destaca por su simplicidad, flexibilidad, robustez, protabilidad y exactitud. Fue realizada en su momento en Perl, y se están creando adaptaciones en C++ en la actualidad.

2.3. Texto de prueba utilizado

Se ha decidido utilizar el siguiente texto de prueba en inglés y en español cumpliendo la normativa de la práctica:

Texto en español:

Los últimos días, Chávez ha recordado en varias ocasiones su amistad con Gadafi y ha dicho que sería de cobardes culpar de las muertes en Libia al Gobierno de ese país sin conocer lo que está pasando.

Texto en inglés:

The rebel leadership said the international community had yet to inform them of any initiative from the Venezuelan president, who reportedly contacted the embattled Libyan leader earlier this week in a bid to enter the fortnight-long violent standoff.

2.4. Resultados del etiquetado

A continuación se van a mostrar los resultados de cada etiquetador.

2.4.1. TreeTragger

En español

```
root@gvazquez:/home/gvazquez/master# echo "Los últimos días, Chávez ha
recordado en varias ocasiones su amistad con Gadafi y ha dicho que sería
de "cobardes" culpar de las muertes en Libia al Gobierno de ese país sin
conocer lo que está pasando." | cmd/tree-tagger-spanish
reading parameters ...
tagging ...
Los ART el
últimos ADJ último
días NC día
, CM ,
Chávez NP Chávez
ha VHfin haber
recordado VLadj recordar
en PREP en
varias QU varios
ocasiones NC ocasión
su PPO suyo
amistad NC amistad
con PREP con
Gadafi NP <unknown>
y CC y
ha VHfin haber
dicho QU dicho
que CQUE que
sería VSfin ser
de PREP de
cobardes ADJ cobarde
culpar VLinf culpar
de PREP de
las ART el
muertes NC muerte
en PREP en
Libia NP Libia
al PAL al
Gobierno NP <unknown>
de PREP de
ese DM ese
país NC país
sin PREP sin
conocer VLinf conocer
lo ART el
que CQUE que
está VEfin estar
pasando VLadj pasar
. FS .
finished.
```

En inglés

```
root@gvazquez:/home/gvazquez/master# echo "The rebel leadership said the
international community had yet to inform them of any initiative from the
Venezuelan president, who reportedly contacted the embattled Libyan leader
earlier this week in a bid to enter the fortnight-long violent standoff."
| cmd/tree-tagger-english
reading parameters ...
tagging ...
finished.
The DT the
rebel NN rebel
leadership NN leadership
said VBD say
the DT the
international JJ international
community NN community
had VBD have
yet RB yet
to TO to
inform VB inform
them PP them
of IN of
any DT any
initiative NN initiative
from IN from
the DT the
Venezuelan JJ Venezuelan
president NN president
, , ,
who WP who
reportedly RB reportedly
contacted VBD contact
the DT the
embattled JJ embattled
Libyan JJ Libyan
leader NN leader
earlier RBR earlier
this DT this
week NN week
in IN in
a DT a
bid NN bid
to TO to
enter VB enter
the DT the
fortnight-long JJ <unknown>
violent JJ violent
standoff NN standoff
. SENT .
```

2.4.2. SVMTool

En español

```
Los DA últimos AO días NC , Fc Chávez NP ha VAI recordado VMP en SP varias DI
ocasiones NC su DP amistad NC con SP Gadafi NP y CC ha VAI dicho VMP que CS
sería VSI de SP " Fe cobardes NC " Fe culpar VMN de SP las DA muertes NC en SP
Libia NP al SPGobierno NP de SP ese DD país NC sin SP conocer VMN lo DA que PR
está VMI pasando VMG . Fp
```

En inglés

```
The DT rebel JJ leadership NN said VBD the DT international JJ community NN
had VBD yet RB to TO inform VB them PRP of IN any DT initiative NN from IN
the DT Venezuelan JJ president NN , , who WP reportedly RB contacted VBD the
DT embattled JJ Libyan JJ leaderNN earlier RBR this DT week NN in IN a DT bid
NN to TO enter VB the DT fortnight-long JJ violent JJ standoff NN . .
```

2.5. Observaciones sobre la comparativa de los resultados

Destacaremos que TreeTagger y SVMTool trabajan con el sistema de etiquetados Penn Treebank. En el etiquetado de textos en español, la herramienta SVMTool, muestra las etiquetas de Penn Treebank pero traducidas al español, mientras que en la otra herramienta no sucede esto.

Es importante comentar que las dos herramientas son bastante buenas para realizar estadísticas de etiquetados, habiendo sido escritas en lenguajes distintos cada una, ya que, TreeTagger está programada en C++ y SVMTool en Perl.

Si hacemos refrencia a la instalación, TreeTagger me pareció bastante sencillo de instalar y lo que más me gustó es la rapidez de inserción de textos a tratar, que con un simple comando en la terminal, muestra todo de forma rápida y eficaz. Por otra lado podemos hacer hincapié en que la instalación de SVMTool fue más complicada, no obstante, tienen un simulador web que no requiere instalación en el ordenador y esto es una ventaja ya que no requiere instalarlo y puedes usarlo en cualquier parte pero también como inconveniente comentar que la herramienta no funciona muy fluida en internet.

Para finalizar comentaré que intenté instalar más aplicaciones, pero daban errores durante la instalación o lograba instalarlas y tenían una documentación bastante muy desorganizada e incompleta.

En conclusión, estas herramientas son, en mi opinión, de las mejores herramientas para etiquetado léxico ofrecidas por la web en el enunciado (teniendo en cuenta las consulta con los doctores y profesores universitarios).