Descubrimiento de Información en Textos
Tarea del Tema 4: Comparativa de etiquetadores estadísticos

Jose Alberto Benítez Andrades

71454586A

Descubrimiento de Información en Textos
Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web
UNED
09/03/2011

0.Enunciado

Tarea del tema 4: Comparativa de etiquetadores estadísticos

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 utilízalos para realizar el etiquetado de un pequeño texto, el mismo para ambos. Para ello asegúrate 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 utilizas 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.Descripción de los etiquetadores seleccionados

Después de intentar completar la instalación y puesta en marcha de varios de los etiquetadores propuestos en la web dada en el enunciado, conseguí hacer funcionar correctamente tres. Los etiquetadores estadísticos seleccionados han sido los tres siguientes:

I. TreeTagger
http://www.ims.uni-stuttgart.de/projekte/tc/
II. QtTag
http://phrasys.net/uob/om/software
III. SVMTool
http://www.lsi.upc.edu/~nlp/SVMTool/lrec2004-gm.pdf

A continuación voy a describir los etiquetadores enumerados anteriormente.

I. TreeTagger

En primer lugar, cabe destacar que fue creado por el personal de la **Universidad de Stuttgart,** por el Instituto de la Lingüística Romance y el Instituto de Ciencias de la Computación departamento de inteligencia artifical).

Fue completamente financiado al 100% por el Ministerio de Ciencia e Investigación del Estado federado de Baden-Württemberg (MWF, Stuttgart), en 1993/1994 y 1995/1996.

En 1993/1994 el proyecto recogió todo el material de texto necesario para el alemán, francés e italiano, y se desarrolló 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. Los textos y análisis de resultados se mantienen separados unos de otros, por razones de flexibilidad y extensibilidad del sistema. Esto es posible gracias a un enfoque particular para el almacenamiento y la representación. Algunos de los componentes de la herramienta actualmente se encuentran en fase de desarrollo, un idioma específico y general, van 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 n-grama. Actualmente se están realizando investigaciones sobre modelos estadísticos para los sintagmas nominales, las colocaciones verbo-objeto, etc.

Instalación

El primer paso para la instalación del mismo es seleccionar el paquete de instalación correspondiente al sistema operativo que nosotros tengamos:

- PC-Linux
- Sparc-Solaris
- Mac OS-X (PowerPC)
- Mac OS-X (Intel-CPU)

En el caso que estudio en este trabajo, yo lo he instalado en un servidor local casero con una distribución de Linux, concretamente **Ubuntu 9.10 Karmic Koala Server.**

En mi caso particular, para instalar este etiquetador, realicé los siguientes pasos, después de

1.Descargar los **scripts etiquetados** en el mismo directorio.

3.Descargar el script de instalación **install-tagger.sh .**

4.Descargar los ficheros de parámetros para el sistema en el que se haya instalado la aplicación (PC, Sparc-Solaris, Mac-Power-PC, Mac-Intel).

Para finalizar, abrimos un terminal y ejecutamos el fichero de instalación:

sh install-tagger.sh

Finalmente para probar la aplicación, podemos escribir lo siguiente en la línea de comandos:

echo 'Hello world!' | cmd/tree-tagger-english

Y veremos una salida como la que muestro en la imagen

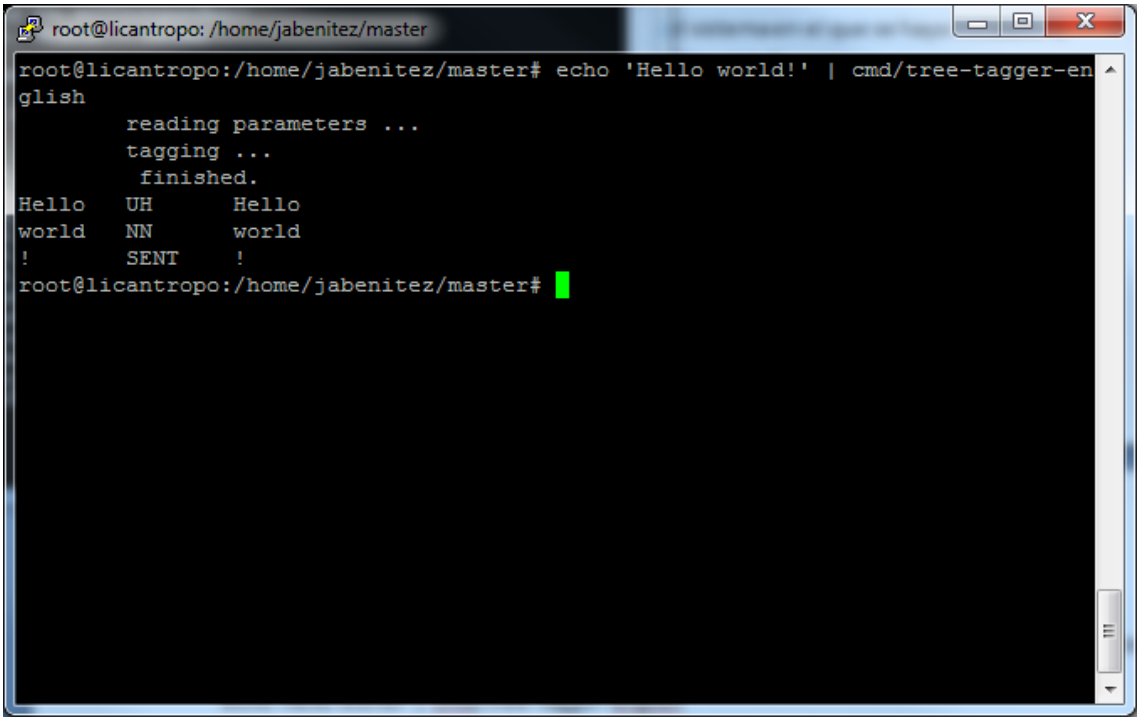

Foto 1: Ventana de putty en windows 7 conectado a servidor local.

II. QTag

QTag es un etiquetador multiplataforma libre. Está implementado en lenguaje **Java** y ha sido probado en Mac OS X, Linux y Windows. Trabaja, en principio, con cualquier idioma del que de dispone de recursos, pero para el modo "shrink-wrap" sólo trabaja con ejemplos en Inglés.

La creación de ficheros de recursos de distintos idiomas, es algo compleja, pero el autor del programa explica que no es muy común realizar ficheros de este tipo, con los lenguajes que trae por defecto suele ser suficiente. Para añadir recursos nuevos, corpus nuevos, hay que contactar con él mediante un correo electrónico.

Su creador es **Oliver Mason**, el cual se encuentra en la Universidad de Birmingham realizando distintos estudios sobre este tipo de temas.

Instalación y puesta en marcha

Este programa tiene una instalación bastante simple, sólo debemos descargar el siguiente fichero:

- Fichero QTag (http://phrasys.net/uob/downloads/qtag.jar)

Y una vez descargado, si estamos bajo sistemas unix, debemos ejecutar en una terminal:

java -jar qtag.jar

Y si estamos en un sistema Windows, teniendo bien instalada la máquina virtual de JAVA con sus correspondientes alias introducidos en el sistema, nos basta con hacer doble click sobre el programa.

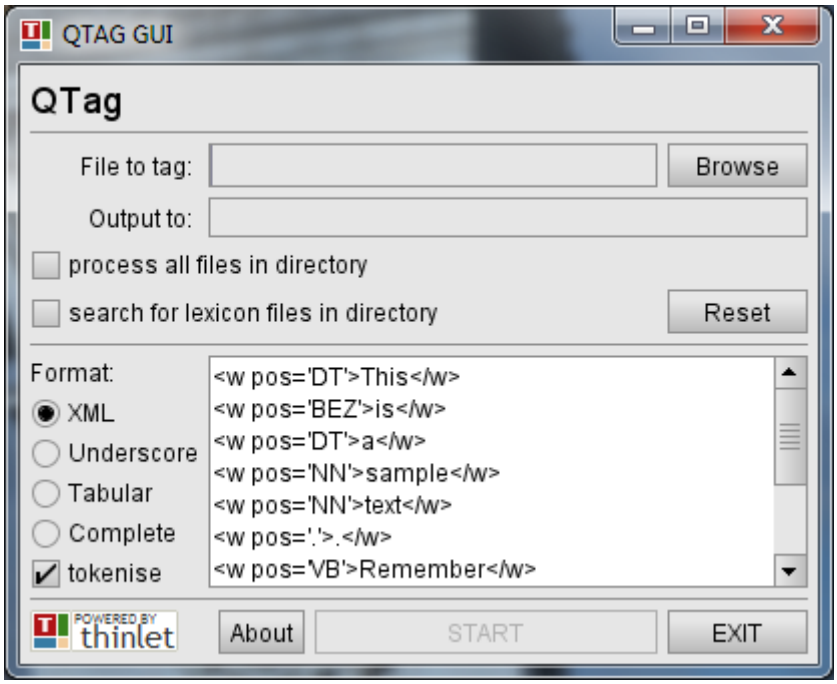

Foto 2: QTag ejecutado bajo Windows 7 Ultimate 64bits

En este caso, instalé la versión para Windows, en mi Windows 7 Ultimate 64 bits.

Su funcionamiento es sencillo, se elige el fichero que queremos analizar pulsando sobre el botón "Browse" que se encuentra justo después del campo File to Tag.

Seleccionamos después el fichero de salida, y elegimos el tipo de salida que queremos obtener (XML, con tabulaciones, completo, texto plano).

III. SVMTool

Esta herramienta está compuesta por tres componentes principales, el *aprendedor (SVMTlearner),* el *etiquetador (SVMTagger)* y el *evaluador (SVMTeval)*.

Antes de realizar el etiquetado, los modelos de SVM aprenden de distintos corpus usando el componente de aprendizaje. Se les enseñan diferentes estrategias a los distintos modelos. Entonces, en el tiempo de etiquetado, 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.

SVMTlearn

Se entrenan a unos clasificadores SVM mediante un conjunto de ejemplos dado. El SVMTlearn tiene un fichero de configuración, en el que se pueden cambiar distintos parámetros que enumeraré 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.
- Feature filtering:
- 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.

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.

SVMTeval

Dada una salida de un etiquetado predicho por la SVMTool y su correspondiente standard, 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.

Por otra parte, 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.

Conclusión de SVMTool

Es una herramienta que 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.Texto de prueba utilizado

Después de observar las distintas posibilidades que proporcionaban los etiquetadores seleccionados para la realización de este trabajo, decidí utilizar dos textos de prueba, uno de ellos en **español** y el otro en **inglés**.

El texto de prueba utilizado en **español** es el siguiente:

> 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.

El texto de prueba utilizado en **inglés** es el siguiente:

> 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.

3.Resultado del etiquetado con cada etiquetador seleccionado.

I. TreeTragger

Texto en Español

```
root@licantropo:/home/jabenitez/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.
```

Texto en Inglés

```
root@licantropo:/home/jabenitez/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    .
```

II. QTag

Texto en Español

```
<w pos='NP'>Los</w>
<w pos='NP'>últimos</w>
<w pos='???'>días</w>
<w pos=','>,</w>
<w pos='NP'>Chávez</w>
<w pos='UH'>ha</w>
<w pos='FW'>recordado</w>
<w pos='FW'>en</w>
<w pos='NNS'>varias</w>
<w pos='NP'>ocasiones</w>
<w pos='???'>su</w>
<w pos='NP'>amistad</w>
<w pos='NN'>con</w>
<w pos='???'>Gadafi</w>
<w pos='NN'>y</w>
<w pos='UH'>ha</w>
<w pos='NN'>dicho</w>
<w pos='FW'>que</w>
<w pos='???'>sería</w>
<w pos='FW'>de</w>
<w pos=''''>"</w>
<w pos='NP'>cobardes</w>
<w pos=''''>"</w>
<w pos='JJ'>culpar</w>
<w pos='NP'>de</w>
<w pos='NP'>las</w>
<w pos='NNS'>muertes</w>
<w pos='FW'>en</w>
<w pos='NP'>Libia</w>
<w pos='NP'>al</w>
<w pos='???'>Gobierno</w>
<w pos='FW'>de</w>
<w pos='NN'>ese</w>
<w pos='???'>país</w>
<w pos='NN'>sin</w>
<w pos='NN'>conocer</w>
<w pos='UH'>lo</w>
<w pos='FW'>que</w>
<w pos='???'>está</w>
<w pos='FW'>pasando</w>
<w pos='.'>.</w>
```

Texto en Inglés

<w pos='DT'>The</w>
<w pos='NN'>rebel</w>
<w pos='NN'>leadership</w>
<w pos='VBD'>said</w>
<w pos='DT'>the</w>
<w pos='JJ'>international</w>
<w pos='NN'>community</w>
<w pos='HVD'>had</w>
<w pos='RB'>yet</w>
<w pos='TO'>to</w>
<w pos='VB'>inform</w>
<w pos='PP'>them</w>
<w pos='IN'>of</w>
<w pos='DT'>any</w>
<w pos='NN'>initiative</w>
<w pos='IN'>from</w>
<w pos='DT'>the</w>
<w pos='NP'>Venezuelan</w>
<w pos='NN'>president</w>
<w pos=',',>,</w>
<w pos='WP'>who</w>
<w pos='RB'>reportedly</w>
<w pos='VBN'>contacted</w>
<w pos='DT'>the</w>
<w pos='JJ'>embattled</w>
<w pos='JJ'>Libyan</w>
<w pos='NN'>leader</w>
<w pos='RBR'>earlier</w>
<w pos='DT'>this</w>
<w pos='NN'>week</w>
<w pos='IN'>in</w>
<w pos='DT'>a</w>
<w pos='NN'>bid</w>
<w pos='IN'>to</w>
<w pos='VB'>enter</w>
<w pos='DT'>the</w>
<w pos='JJ'>fortnight-long</w>
<w pos='JJ'>violent</w>
<w pos='NN'>standoff</w>
<w pos='.'>.</w>

III. SVMTool

Texto 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 *SP*Gobierno *NP* de *SP* ese *DD* país *NC* sin *SP* conocer *VMN* lo *DA* que *PR* está *VMI* pasando *VMG* . *Fp*

Texto 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* leader*NN* 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* . .

4.Observaciones sobre la comparativa de los resultados.

Comparativa en los resultados

En primer lugar, destacar que TreeTagger y SVMTool trabajan con el sistema de etiquetados Penn Treebank y QTag trabaja con el sistema PENN Treebank y unas modificaciones que realizó su tutor en el corpus, como por ejemplo, agregarle etiquetas de este estilo:

- HAD es HVD porque es una forma del verbo Have

Y aunque en el texto de prueba que elegí, no aparezca, sucede lo mismo con diferentes formas del verbo To BE.

Y en el etiquetado de textos en español, la herramienta SVMTool, muestra las etiquetas de Penn Treebank, pero traducidas al español, mientras que en las otras 2 herramientas no sucede esto.

Comparativa de las herramientas en general

Destacar, que las tres son herramientas bastante buenas para realizar estadísticas de etiquetados, habiendo sido escritas en lenguajes distintos cada una, ya que, TreeTagger está hecha en C++, QTag en Java y SVMTool en Perl.

En cuanto a la instalación, TreeTagger me pareció bastante sencillo de instalar y lo que más me gusta es la rapidez de inserción de textos a tratar, que con un simple comando en la terminal, te muestra todo de forma rápida y eficaz. Quizá para alguien que no esté acostumbrado a trabajar con comandos en Linux, lo vea algo aparatoso (aunque no creo que alguien sin conocimientos de linux, necesite instalar este tipo de herramientas.

La instalación de QTag, es inexistente, tan sólo es descargar el fichero de su página web y ejecutar el .jar con la máquina virtual de JAVA. Para un usuario de a pie, esta es una herramienta muy sencilla para comenzar a utilizarla.

La instalación de SVMTool fue algo más engorrosa, no obstante, tienen un simulador web que no requiere instalación en el ordenador de ningún tipo y esto es por una parte una ventaja y por otra un inconveniente: ventaja porque no requiere instalarlo y puedes usarlo en cualquier parte e inconveniente porque la herramienta va algo lenta en internet.

En rapidez, QTag y TreeTagger están bastante igualadas y en cuanto a resultados, QTag puede ser más útil, porque te permite a golpe de click cambiar las configuraciones del fichero de salida, pudiendo crear ficheros XML, de texto plano, o con otros formatos.

En multiplataforma, gana claramente QTag, ya que al funcionar con la máquina virtual de JAVA, puedes utilizarlo en cualquier SO indistintamente, TreeTagger carece de versión para Windows.

He de decir, que intenté instalar más aplicaciones, pero me daban algunas bastantes fallos en la instalación o lograba instalarlas pero tenían una documentación bastante más desorganizada e incompleta. En general, estas tres herramientas desde mi punto de vista, son de las tres mejores herramientas para etiquetado léxico.