LINGÜÍSTICA COMPUTACIONAL

Antonio Frías Delgado

Área de Lógica y Filosofía de la Ciencia Departamento de Historia, Geografía y Filosofía Universidad de Cádiz



POS TAGGING

Materiales básicos de trabajo

POS TAGGING

Materiales básicos de trabajo

■ Jurafsky-Martin, capítulo 5: Part-of-Speech Tagging

POS TAGGING

Materiales básicos de trabajo

- Jurafsky-Martin, capítulo 5: Part-of-Speech Tagging
- FreeLing, NLTK, Corpus etiquetados



ELEMENTOS PREVIOS

■ Un conjunto de etiquetas (categorías gramaticales)

ELEMENTOS PREVIOS

- Un conjunto de etiquetas (categorías gramaticales)
- No existe un conjunto único (Penn Treebank, Brown, FreeLing, etc.)

ELEMENTOS PREVIOS

- Un conjunto de etiquetas (categorías gramaticales)
- No existe un conjunto único (Penn Treebank, Brown, FreeLing, etc.)
- Un Lexicon con las palabras y las distintas etiquetas



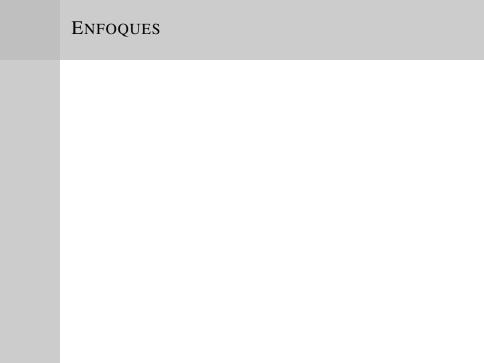
■ Oración: vino blanco

- Oración: vino blanco
- Lexicon:

- Oración: vino blanco
- Lexicon:
 - vino: V, N

- Oración: vino blanco
- Lexicon:
 - vino: V, N
 - blanco: N, ADJ

- Oración: vino blanco
- Lexicon:
 - vino: V, N
 - blanco: N, ADJ
- vino (V? N?) blanco (N? ADJ?)



■ Basado en reglas

- Basado en reglas
- Basado en métodos estadísticos

- Basado en reglas
- Basado en métodos estadísticos
- Basados en transformaciones

- Basado en reglas
- Basado en métodos estadísticos
- Basados en transformaciones
- Memory-based
 Memory-based learning is a form of supervised
 learning based on similarity-based reasoning.
 The part of speech tag of a word in a particular context is extrapolated from the most similar cases held in memory. (Daelemans)



Basados en reglas

■ Partimos de un diccionario con posibles etiquetas

- Partimos de un diccionario con posibles etiquetas
- Asignamos todas las etiquetas a todas las palabras

- Partimos de un diccionario con posibles etiquetas
- Asignamos todas las etiquetas a todas las palabras
- Escribimos reglas a mano para remover selectivamente etiquetas

- Partimos de un diccionario con posibles etiquetas
- Asignamos todas las etiquetas a todas las palabras
- Escribimos reglas a mano para remover selectivamente etiquetas
- Finalizamos cuando cada palabra tiene una etiqueta



she: PRP

promised: VBN,VBD

to: TO

back: VB, JJ, RB, NN

the: DT

bill: NN, VB



Basados en reglas

Aplicar reglas de eliminación

Aplicar reglas de eliminación Eliminar VBN (participio) si VBD (pasado) es una opción cuando sigue a PRP (pronombre personal)



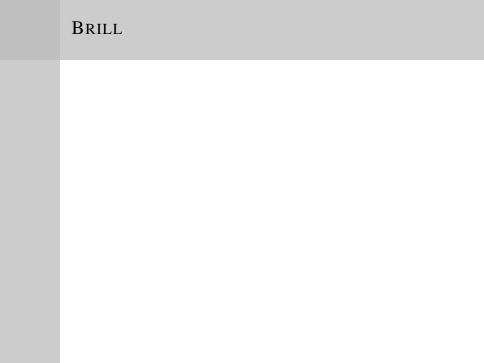
El proyecto más conocido es EngCG. Usa cerca de 4000 reglas.

El proyecto más conocido es EngCG. Usa cerca de 4000 reglas.

ENGCG, the Constraint Grammar Parser of English, performs morphosyntactic analysis (tagging) of running English text. The parser employs a morphological ("part-of-speech") disambiguator that makes 93-97% of all running-text words in Written Standard English unambiguous while 99.7% of all words retain the correct analysis.

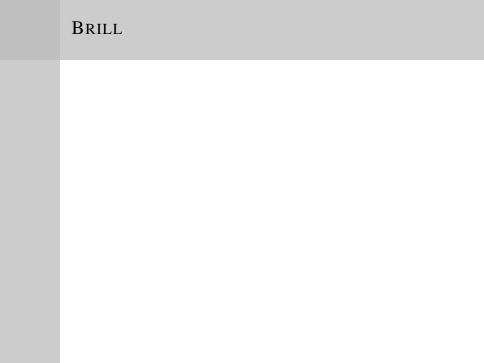


```
Adverbial-that rule
Given input: 'that'
if
      (+1 A/ADV/QUANT); if next word is adj/adv/quantifier
      (+2 SENT-LIM); following which is a sentence
           boundary
      (NOT -1 SVOC/A); and the previous word is not a
            verb like 'consider' which allows adjective
           complements
then eliminate non-ADV tags
else eliminate ADV tag
```



■ Etiquetación mediante reglas

- Etiquetación mediante reglas
- Reglas generadas automáticamente a partir de un corpus de entrenamiento

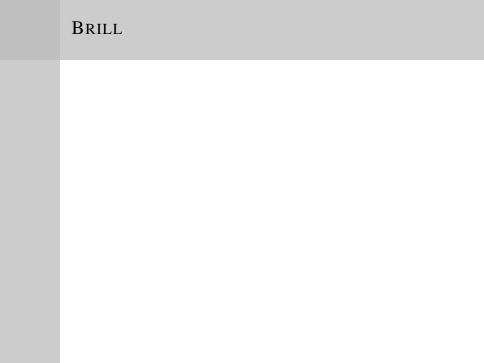


$B_{RILL} \\$

Etiqueta inicialmente cada palabra con la etiqueta más probable

- Etiqueta inicialmente cada palabra con la etiqueta más probable
- 2 Etiqueta palabras desconocidas

- Etiqueta inicialmente cada palabra con la etiqueta más probable
 - Etiqueta palabras desconocidas
- 3 Aplica en orden una serie de reglas contextuales inferidas a partir del corpus de entrenamiento



Algoritmo basado en transformaciones y dirigido por el error

■ Plantillas

- Plantillas
- Asignación inicial: etiqueta más probable

- Plantillas
- Asignación inicial: etiqueta más probable
- Comparar con el corpus de entrenamiento identificando errores

- Plantillas
- Asignación inicial: etiqueta más probable
- Comparar con el corpus de entrenamiento identificando errores
- Aplicar cada regla de transformación seleccionando la que tenga el máximo de errores eliminados - errores introducidos

- Plantillas
- Asignación inicial: etiqueta más probable
- Comparar con el corpus de entrenamiento identificando errores
- Aplicar cada regla de transformación seleccionando la que tenga el máximo de errores eliminados - errores introducidos
- Volver al corpus de entrenamiento

- Plantillas
- Asignación inicial: etiqueta más probable
- Comparar con el corpus de entrenamiento identificando errores
- Aplicar cada regla de transformación seleccionando la que tenga el máximo de errores eliminados - errores introducidos
- Volver al corpus de entrenamiento
- Parar al llegar a cierto umbral