

# LINGÜÍSTICA COMPUTACIONAL

Antonio Frías Delgado

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

# POS TAGGING

# 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

# 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

# EL PROBLEMA

# EL PROBLEMA

- Oración: vino blanco

# EL PROBLEMA

- Oración: vino blanco
- Lexicon:

# EL PROBLEMA

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

# EL PROBLEMA

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

# EL PROBLEMA

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

# ENFOQUES



# ENFOQUES

- Basado en reglas

# ENFOQUES

- Basado en reglas
- Basado en métodos estadísticos

# ENFOQUES

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

# ENFOQUES

- 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

# BASADOS EN REGLAS

- Partimos de un diccionario con posibles etiquetas

# BASADOS EN REGLAS

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

# BASADOS EN REGLAS

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



# BASADOS EN REGLAS

- 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

# BASADOS EN REGLAS

# BASADOS EN REGLAS

she: PRP

promised: VBN,VBD

to: TO

back: VB, JJ, RB, NN

the: DT

bill: NN, VB

# BASADOS EN REGLAS

# BASADOS EN REGLAS

Aplicar reglas de eliminación

# BASADOS EN REGLAS

Aplicar reglas de eliminación

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

# BASADOS EN REGLAS

# BASADOS EN REGLAS

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



## BASADOS EN 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.

# BASADOS EN REGLAS

# BASADOS EN REGLAS

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

BRILL

- Etiquetación mediante reglas

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

BRILL

- 1 Etiqueta inicialmente cada palabra con la etiqueta más probable



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

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

BRILL

Algoritmo basado en transformaciones y dirigido por el error

- Plantillas

Algoritmo basado en transformaciones y dirigido por el error

- Plantillas
- Asignación inicial: etiqueta más probable

Algoritmo basado en transformaciones y dirigido por el error

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

Algoritmo basado en transformaciones y dirigido por el error

- 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

Algoritmo basado en transformaciones y dirigido por el error

- 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



Algoritmo basado en transformaciones y dirigido por el error

- 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