Introducción al Procesamiento de Lenguaje Natural

Dr. Gabriel Infante-Lopez
Sección Informatica - FaMAF
Universidad Nacional de Cordoba
gabriel@famaf.unc.edu.ar
http://www.cs.famaf.unc.edu.ar/~gabriel

Material de Lectura

- Manning, C., Schütze, H. *Foundations of Statistical Natural Language Processing*. The MIT press. 1999. Obligatorio
- Jurafsky, D., Martin, J. Speech and Language Processing. Prentice-Hall. 2000. Referencia
- Wall, L. et al. Programming Perl. 3rd Ed. O'Reilly. Referencia
- Proceedings de conferencias: ACL, EACL, COLING

Requirimientos del Curso pesos para la nota final

- Proyectos Individuales (4) 40%
- Examen parcial mitad cuatrimestre 30%
- Examen parcial final cuatrimestre 30%
- Los examenes deben ser aprobados para que promediar los resultados.

Proyectos y Organización

- Proyectos
  - Entropia, y modelos de lenguaje
  - Clases de Palabra
  - Clasificación
  - Sintaxis
- Organización
  - algunos ejercicios de papel y lapiz, mucha programación
  - deadlines estrictos. 10 días de tardanza máximo. Ultimo proyecto solo 2 días.
  - Plagio NO pertido

Descripción del Curso

- Intro., Probabilidad, y teoría de la información
  - Conceptos basicos: definiciones, formulas, ejemplo, etc.
- Modelos de Lenguaje
  - n-grams, estimacion de parametros
  - smoothing (EM algorithm)
- Un poco de Lingüistica
  - fonologia, morfologia, sintaxis, semantica, discurso
- Palabras y Léxicon
  - clases de palabra, información mutua

Descripción del Curso

- Hidden Markov Models
  - background, algoritmos, estimación param.
- Tagging: Metodos, Algoritmos, Evaluación
  - Tagsets, morfologia, lematización
  - HMM tagging, Transformation y feature based
- Gramaticas, y Analisis Sintactico: Datos, Algs.
  - Gramaticas y automatas, parsing determ.
- Aplicaciones.

PLN: Principales problemas

- varias palabras, varios fenomenos ⇒ alta complejidad. Ejemplo: Lexicón Finlandes ~ 2 x 10^7
- irregularidades (excepción, excepción de la excepción):
  - cantar -> canté, hablar -> hablé, estar -> estuve.

PLN: Principales problemas

- Ambiguedad:
  - *bajo*: Adjetivo, nombre, preposición
  - *Prohibido girar a la izquierda de 12 a 14. Excepto taxis y remises.* ¿Los taxis y remises no pueden girar núnca?
  - *Los chorros de aguas cordobesas*. ¿Cuántas interpretaciones posibles?
  - *Gracias por no comer, beber y escuchar música sin auriculares.* ¿Podemos comer con auriculares?

¿Reglas categoricas o Estadistica?

- Preferencias:
  - casos claros: pistas en el contexto: *bajo la mesa* (bajo es una preposición)
  - casos menos claros. *El viento le voló el sombrero, el hombre lo quiso agarrar*. A quien? ¿al sombrero o al viento?
  - semanticos: *comer pizza con cuchillo*, *comer pizzas con aceitunas*.

Soluciónes

- Si sabemos, mejor no adivinar
  - Morfologia, inflexiones, exepciones
  - lexicons, lista de palabras
  - nombres inambigüos
  - frases fijas, colocaciones: ej. Aguas Cordobésa.
  - Reglas sintacticas? S -> Np Vp
- Usar estadisticas solo para preferencias

NLP Estadístico

- Supongamos que a cada oración W, se le asigna una probabilidad P(WIX) en un contexto X.
- Para cada contexto X, ordenar todas las oraciones de acuerdor a P(WIX).