Métodos estadísticos en PLN

ITALICA

Universidad de Sevilla

Víctor J. Díaz Madrigal

Introducción

- Enfoques racionalista
- Problemática
- Enfoque empírico
- Corpus
- Estudios preliminares
- Leyes de Zipf
- Colocaciones y concordancias

Enfoque racionalista

- Modelo: El conocimiento lingüístico es innato y puede ser caracterizado mediante principios y procedimientos.
- Objetivo: Distinción entre construcciones correctas e incorrectas (gramática)
- Orígenes:
  - Teoría generativa (Chomsky, 65)
- Está más orientado hacia el estudio de la capacidad (competence) que al de la habilidad (performance)

Problemática

- El concepto de corrección
  - Frases correctas sin sentido:

    “Colorless green ideas slept furiously”
  - Frases poco convencionales
  - Frases incorrectas pero inteligibles
- Cambios de uso en palabras (Fenómenos no categoriales)
- Ambigüedad
  - Léxica: Palabras con múltiples categorías (POS)
  - Sintáctica:

    “John saw a man in the park with a telescope”
  - Semántica:
- Cobertura frente a precisión

Enfoque empirista

- Modelo: Existe cierta capacidad cognitiva innata basada en criterios generales de asociación y generalización.
- Objetivo: Definición de un modelo del lenguaje
- Orígenes:
  - “You shall know a word by the company it keeps” (Firth, 57)
  - Estructuralistas americanos (Harris, 51)
- Técnicas:
  - Métodos estadísticos.
  - Reconocimiento de patrones.
  - Aprendizaje automático.
  - Inductivas: Corpus (texto) o Corpora (colección de textos)
- Está más orientado hacia el estudio de la habilidad (performance) que al de la capacidad (competence)

Corpus

- Brown Corpus (No es gratuitio)
  - Corpus balanceado (documentos procedentes de diversas fuentes) para el inglés americano
- LOB (Lancaster-Oslo-Bergen)
  - Réplica británica de Brown
- Susanne (Gratuito)
  - 130.000 palabras
  - Incluye etiquetas de estructura sintáctica
- Penn Treebank (No es gratuito)
  - Incluye etiquetas de estructura sintáctica
- Canadian Hansards (Bilingüe)
- Wordnet

Estudios preliminares

- ¿Cuál es el tamaño mínimo de un corpus?
  - ¿1 Mb de palabras ?
- ¿Cuáles son las palabras más frecuentes?
  - Abundan las palabras funcionales (function words): preposiciones, determinantes y complementos
  - Los resultados reflejan la naturaleza del texto
- ¿Cuántas palabras hay en el texto?
  - Problemas: Distinción entre palabra-ocurrencia y palabra-tipo
  - Ratio entre ocurrencia y tipo: Frecuencia media con que es usado cada tipo (No es un índice de la complejidad de un texto)
  - Ejemplo: Tom Sayer Ocurrencias 71.370 Tipos 8.018 Ratio 8.9

Leyes de Zipf

- La frecuencia f de una palabra es proporcional a su rango r

$$f \propto 1/r$$

  - hay muy pocas palabras con mucha frecuencia
  - algunas tienen una frecuencia media
  - la mayoría tienen muy poca frecuencia (Hapax Legomena).
- Ley de Malderbrot $f = P (r + p)^{-B}$
- El número de significados m de una palabra es proporcional a su frecuencia $m \propto sqrt(f)$
- La mayoría de las veces una “content word” aparece cerca de otra ocurrencia de la misma palabra $f \propto I^{-p}$
- Conclusión: El problema de seguir un enfoque basado en la frecuencia es que la mayoría de las palabras aparecen poco.

Colocaciones (Collocations) y Concordancia

- Colocaciones (Collocations) son palabras que aparecen muy cercanas y cuya semántica no es composicional.

  disk drive, make up, políticamente incorrecto
- Problemas de discontinuidad y número de palabras
- Métodos basados en calcular los N-gramas más frecuentes
- Mejoras: tienen en cuenta las frecuencias individuales o se filtran patrones funcionales P+D (of the, in the). Los patrones más frecuentes son A+N, N+N
- Concordancia (Concordances) : Búsqueda de patrones de ocurrencia (syntactic frames)
- Las herramientas KWIC (Key Word in Context) enumeran tabularmente una palabra junto con los contextos izquierdos y derechos en los que aparece en un texto

Ventajas e inconvenientes

- Ventajas de los métodos estadísticos
  - Robustos frente a errores
  - Genéricos
  - Adaptativos
  - Automatizables
- Desventajas:
  - Incompletas e inconsistentes
  - Recursos lingüísticos:
    - Documentos (Corpora)
    - Diccionarios
    - Tesaurus
    - Herramientas auxiliares