# Índice

# PLN & IR

- **Def. Procesamiento del Lenguaje Natural (PLN):** tratamiento computacional del lenguaje humano
  - Objetivo: computadora comprenda el lenguaje humano
- **Objetivo IR:** dada una **colección de documentos** y una **necesidad de información** del usuario —expresada como una **consulta (query)**—, devolver un conjunto de documentos relevantes para dicha necesidad de información (i.e., cuyo contenido satisface dicha necesidad)
- **IR como tarea de NLP:** "comprender" el contenido de los documentos

# *Keyword Retrieval Hypothesis*

- **Def.:** representación de documentos/consultas como **conjunto** de *términos índice* (a.k.a. *términos de indexación* o *palabras clave*)
- **Ppo. de composicionalidad de Frege**: "la semántica de un objeto puede obtenerse a partir de la semántica de sus componentes"
  - Si una palabra aparece en un texto, dicho texto trata dicho tema
- **Hipótesis de recuperación por palabras clave** (*keyword retrieval hypothesis*): "si una consulta y un documento comparten términos índice, es que el documento debe tratar el tema de la consulta"
  - **Problema: es insuficiente**, el lenguaje no es un mero repositorio de palabras
    - Comunicar conceptos, entidades, y relaciones de múltiples maneras
    - Las palabras se combinan en unidades lingüísticas de mayor complejidad, cuyo significado no siempre viene dado por el significado de sus palabras componente

# Variación Lingüística

- **Principal problema de la Minería de Textos** (*Text Mining*): variación lingüística
  - El mismo concepto puede expresarse de diferentes maneras (y viceversa)
  - Impide establecer correspondencias
  - Introduce ruido
- Diferentes niveles de variación:
  - Morfológica: modificaciones **flexivas** y **derivativas** (dañan *recall*)

    `cantas` $\sim$ `cantó` `cantar` $\sim$ `cantante`
    - Dependiente de la **complejidad morfológica** del lenguaje
  - Semántica: **polisemia** (dañan *precision*)

    `banda (de música)` $\neq$ `banda (franja)`
  - Léxica: **sinonimia** (dañan *recall*)

    `rápido` $=$ `veloz`
  - Sintáctica: modificaciones de la **estructura sintáctica** (dañan ambas)

    `Juan atacó a Pepe` $\neq$ `Pepe atacó a Juan`

    `Juan atacó a Pepe` $=$ `Pepe fue atacado por Juan`
  - Pueden darse simultáneamente: p.ej. morfo-sintáctica:

    `cambio climático` $=$ `cambio del clima`

# Tratamiento de la Variación Lingüística

- **Solución**: técnicas de NLP
- Dos enfoques:
  - **Normalización:** reducir las diferentes variantes de un término a una *forma canónica* común
    - Ej. *stemming*
  - **Expansión:** añadir a la consulta variantes de sus términos originales
    - Ej. añadir sinónimos

# Tratamiento de la Variación Morfológica: *Stemming*

- **Def.: reducción de una palabra a su stem o raíz supuesta** eliminando su terminación según una lista de sufijos y reglas de transformación (i.e. normalización)
  - *Stem* contiene semántica básica

$$\left.\begin{array}{l}\texttt{reloj}\\ \texttt{relojes}\\ \texttt{relojero}\end{array}\right\} \rightarrow \texttt{reloj-}$$

- **Objetivo:**
  - Principal: permitir correspondencias entre variantes (incrementar *recall*)
  - Secundario: reducir recursos almacenamiento (reducir vocabulario)
- *Stemmer* de Porter
  - Demo: `http://maya.cs.depaul.edu/~classes/ds575/porter.html`
  - `Snowball` (descargables): `http://snowball.tartarus.org`
- Nivel de normalización
  - *Superficial*: sólo morfología flexiva simplificada; p.ej., sólo plurales
  - *Profundo*: flexiva y derivativa (agresivo); p.ej., Porter

# Tratamiento de la Variación Morfológica: *Stemming* (cont.)

- Ventajas
  - Simplicidad
- Desventajas:
  - Rendimiento **dependiente de la morfología del idioma**: problemas con lenguas de morfología compleja y muchas irregularidades. Ej. español:
    - Adjetivos/nombres: +20 grupos variación género +10 grupos número
    - Verbos: 3 regulares, ±40 irregulares; 118 formas flexivas cada grupo
  - Pérdida de información de cara a procesamiento futuro: produce formas no lingüísticas

$$\texttt{recognized} \rightarrow \texttt{recogn-}$$

  - *Over-stemming*: palabras no relacionadas dan igual *stem*

$$\left.\begin{array}{l}\texttt{general}\\ \texttt{generous}\end{array}\right\} \rightarrow \texttt{gener-}$$

  - *Under-stemming*: palabras sí relacionadas dan *stems* diferentes

$$\begin{array}{ll}\texttt{recognize} & \rightarrow \texttt{recogn-}\\ \texttt{recognition} & \rightarrow \texttt{recognit-}\end{array}$$

# Tratamiento de la Variación Morfológica (cont.): Flexión

- **Expansión flexiva** (expansión): expandir la consulta con formas flexionadas:

$$\texttt{gato} \longrightarrow \begin{cases} \texttt{gata} \\ \texttt{gatos} \\ \texttt{gatas} \end{cases}$$

- **Lematización** (normalización): sustituir palabra por su **lema**

$$\text{Ej. } \texttt{gatas} \rightsquigarrow \texttt{gato}$$

  - Permite abordar la variación **flexiva**
  - Mejora resultados con idiomas de morfología compleja. Ej. lenguas romances
  - Reduce la pérdida de información: siempre obtenemos palabras
  - Google integra dicha capacidad

# Tratamiento de la Variación Morfológica (cont.): Derivación

- **Análisis morfológico:**
  - Permite abordar la variación **derivativa**
  - Necesario con las lenguas de morfología más compleja. Ej. árabe
  - Aplicaciones:
    - **Stemming lingüístico** (normalización): obtener la raíz [lingüística] mediante el análisis
    - **Clustering derivativo** (normalización): un conjunto de palabras relacionadas derivativamente son reducidas a un mismo término base común

$$\left.\begin{array}{l}\texttt{sabotear}\\ \texttt{saboteador}\\ \texttt{sabotaje}\end{array}\right\} \longrightarrow \texttt{sabotaje}$$

    - **Expansión derivativa** (expansión): expandir la consulta con palabras derivadas

$$\texttt{sabotaje} \longrightarrow \left\{\begin{array}{l}\texttt{sabotear}\\ \texttt{saboteador}\end{array}\right.$$

# Tratamiento de la Variación Léxica y Semántica

- Muy conectadas entre sí
- Requeriría aplicar técnicas de *desambiguación del sentido* (*Word Sense Disambiguation (WSD)*)
  - Necesaria alta efectividad: ∼90% (¿60%?)
- Se suele emplear **WordNet/EuroWordNet**
- Aproximaciones:
  - **Clustering semántico** (normalización): un conjunto de palabras relacionadas semánticamente son reducidas a un mismo término base común
    - Ej. indexación por sentidos (*synsets*) en lugar de palabras
  - *Fuzzy matching* en base a **distancias conceptuales**. Ej.:

$$sim(x,y) \rightarrow \begin{cases} 1 & x = y \\ 0.9 & x \in SYN(y) \\ 0.7^n & x \in HYPON_n(y)^\dagger \\ 0.5^n & x \in HYPER_n(y) \\ 0 & \text{resto} \end{cases}$$

$^\dagger$ $x \in HYPON_n(y)$ si $x$ es un hipónimo de $y$ con $n$ niveles de diferencia

# Tratamiento de la Variación Léxica y Semántica (cont.)

- Aproximaciones (cont.):
  - **Expansión semántica** (expansión): expandir la consulta con términos semánticamente relacionados. Ej. sinónimos:

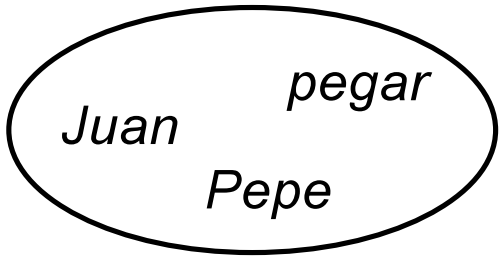
$$\texttt{bonito} \longrightarrow \begin{cases} \texttt{hermoso} \\ \texttt{bello} \end{cases}$$

    - En ocasiones la expansión es **ponderada**
    - **Poco efectivo** salvo con consultas cortas o incompletas
    - Integrado en Google:
      — Implícitamente:
      $\texttt{mantequilla de cacahuete} \leftrightarrow \texttt{crema de cacahuete}$
      — Explícitamente: operador $\sim$

$$\texttt{ape} \longrightarrow \begin{cases} \texttt{monkey} \\ \texttt{gorilla} \\ \texttt{chimpanzee} \end{cases}$$

# Tratamiento de la Variación Sintáctica

- **Problema:** *bag-of-terms* insuficiente:




$$\begin{cases} \textit{Juan pegó a Pepe ?} \\ \textit{Pepe pegó a Juan ?} \end{cases}$$

- **2 aproximaciones:**
  1. Representaciones complejas en base a estructuras sintácticas: árboles y grafos
     - Coste muy alto: inadecuado para uso práctico
  2. **Frases como términos índice:**
     - **Hipótesis:** frases denotan conceptos/entidades más significativos que las palabras
     - Términos más precisos y descriptivos
     - Uso combinado con palabras

# Tratamiento de la Variación Sintáctica (cont.): Identificación y Extracción

- Técnicas estadísticas
  - Secuencias de palabras coocurren frecuentemente
  - Análisis estadístico (frecuencias, coocurrencias, etc.)
  - No base lingüística (a veces resultados extraños)
  - Mayor simplicidad
- Sintácticas
  - Secuencias de palabras satisfacen relaciones sintácticas
  - Análisis sintáctico (complejidad diversa)
  - Sí base lingüística (teóricamente superiores)
  - Mayor complejidad
- Aproximar sintaxis mediante distancias
  - Palabras cercanas se suponen relacionadas sintácticamente

# Tratamiento de la Variación Sintáctica (cont.): Representación y Correspondencias

- Como conjuntos de palabras
- Almacenar árbol de análisis
  - Técnicas de comparación de árboles: gran complejidad
- Almacenar sólo las relaciones sintácticas interesantes
  - Sustantivo–modificador
  - Sujeto–verbo
  - Verbo–Objeto
  - ...

# Tokenización

- En cualquier aplicación de *Text Mining* es necesario **dividir el texto en** unidades lingüísticamente significativas (i.e. **palabras**) antes de procesarlo
- Generalmente obviado: los sistemas de IR emplean **tokenizadores muy sencillos** similares a los de compiladores de lenguajes de programación
- **Problemas**:
  - **Concepto lingüístico** de palabra no coincide con concepto ortográfico. Ej.:

    locuciones: `sin embargo`
    compuestos: `lebensversicherungsgesellschaftsangesteller`
    *(empleado de cía. de seguros)*

  - **Ambigüedad** en la tokenización. Ej.:

`sin embargo` $\longrightarrow$ { "`No tenía ganas, sin embargo lo hizo`" / "`Las relaciones prosiguieron sin embargo económico alguno`" }

`ténselo` $\longrightarrow$ { `ten+se+lo` (tener) / `tense+lo` (tensar) }

  - **Muy dependendiente** de los fenómenos lingüísticos de cada idioma
- **Solución: tokenizadores de base lingüística**

# Segmentación de Compuestos

- **Compuestos**: palabras concatenación de palabras ("base")
  - En ocasiones existen interfijos que las conectan.
    Ej. `spokesman`
  - Comunes en alemán, holandés, finés, sueco...

Ej. alemán: `lebensversicherungsgesellschaftsangesteller` *(empleado de cía. de seguros)*

- **Algoritmos de segmentación** de compuestos:
  - **Basados en reglas**: reglas de descomposición creadas manualmente
    - Requieren un profundo conocimiento del lenguaje
  - **Basados en diccionarios**:
    - Lexicones con palabras válidas del lenguaje como referencia
    - Intentan trocear en palabras contenidas en esos diccionarios
    - Problema: los lexicones no son exhaustivos (variantes, desconocidas, etc.)

# Segmentación de Compuestos (cont.)

- **Algoritmos de segmentación** de compuestos (cont.):
  - **Basados en corpus**: como los anteriores empleando un corpus de texto como lexicón
    - Reducen el problema de la falta de exhaustividad
    - Frecuencias de aparición en el corpus para calcular la mejor segmentación
    - Dependencia del corpus
  - **Estadísticos**: cálculo de la segmentación más probable en base a ocurrencias y co-ocurrencias en un corpus de entrenamiento segmentado a mano
    - Dependencia del corpus
  - **Híbridos**: combinan basados en diccionarios/corpus y estadísticos
  - **n-Gramas de caracteres**: segmentación en secuencias de $n$ caracteres

$$\texttt{potato} \xrightarrow{n=3} \{\ \texttt{-pot-},\ \texttt{-ota-},\ \texttt{-tat-},\ \texttt{-ato-}\ \}$$

    - Simplicidad
    - Eficiencia
    - Robustez
    - Independencia del idioma

# Un Ejemplo Extremo: el Chino

(Y otras lenguas asiáticas) Muy problemático:

- No existen separadores entre palabras: una oración es una secuencia continua de caracteres/símbolos
  - Las aproximaciones clásicas de IR no sirven
- Los caracteres (símbolos) chinos son mucho más significativos que los occidentales
  - La mayoría de ellos son palabras de por sí
- Vocabulario es extremadamente rico: el mismo concepto se puede expresar de múltiples formas (que suelen compartir símbolos). Ej. (Nie & Ren, 1999)

> For example, we can find the common character 建 in all the following words which mean "construction": 建设 (construct), 建筑 (construct), 建立 (construct, establish), 建成 (have constructed), 建造 (construct), 大建 (construct abundantly).

# Un Ejemplo Extremo: el Chino (cont.)

- La ambigüedad en la segmentación es enorme:
  - Chino-parlantes nativos concuerdan menos del 70%
  - Ej. (Nie & Ren, 1999)

现在本所有研究生活动 (there is currently an activity for graduate students in our institute) contains the following legitimate words:

| | |
|---|---|
| 现 (now), | 现 在 (now), |
| 在 (at), | |
| 本 (originally), | 本 所 (our institute), |
| 所 (institute), | 所 有 (all, belong to), |
| 有 (have), | |
| 研 究 (research), | 研 究 生 (graduate students), |
| 生 (give birth), | 生 活 (life), |
| 活 (live), | 活 动 (activity), |
| 动 (move). | |

There are 30 possible combinations of legitimate words which cover the sentence. Only the following one is correct:

现在 本所 有 研究生 活动
(now / our institute / have / graduate student / activity)

# Referencias


- [Arampatzis et al., 2000] Arampatzis, A., van der Weide, Th. P., van Bommel, P. & Koster, C.H.A. (2000). Linguistically-motivated Information Retrieval. In vol. 69 of *Encyclopedia of Library and Information Science*, pp. 201–222. Marcel Dekker, Inc.
- [Baeza-Yates & Ribeiro-Neto, 1999] Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval* Addison Wesley and ACM Press.
- [Foo & Li, 2004] Foo, S. & Li, H. (2004). Chinese word segmentation and its effect on information retrieval. *Information Processing & Management*, 40(1):161-190.
- [Hollink et al., 2004] Hollink, V., Kamps, J., Monz, C. & De Rijke. (2004). Monolingual Document Retrieval for European Languages. *Information Retrieval*, 7:33–52.
- [Koster, 2004] Koster, C.H.A. (2004). Head/Modifier Frames for Information Retrieval. In vol. 2945 of *Lecture Notes in Computer Science*, pp. 420–432. Springer-Verlag.
- [Lazarinis et al., 2009] Lazarinis, F., Vilares, J., Tait, J. & Efthimiadis, E.N. (2009). Current research issues in non-English Web searching. *Information Retrieval*, 12:230-250.

# Referencias (cont.)


- [Nie & Ren, 1999] Nie, J.-Y., Ren, F. (1999). Chinese information retrieval: using characters or words?. *Information Processing & Management*, 35(4):443-462.
- [Palmer, 2000] Palmer, D.D. (2000). *Tokenisation and Sentence Segmentation*. Chapter 2 of *Handbook of Natural Language Processing*. Dale, R. Moisl, H., Somers, H. (eds.). Marcel Dekker, Inc.
- [Vilares, 2005] Vilares, J. (2005). *Aplicaciones del Lenguaje Natural a la Recuperación de Información en Español*, PhD. Thesis, Universidade da Coruña.
- [Vilares et al., 2008] Vilares, J., Alonso, M.A. & Vilares, M. (2008). Extraction of Complex Index Terms in Non-English IR: A Shallow Parsing Based Approach. *Information Processing & Management*, 44(4):1517–1537.