

# RESUMEN AUTOMÁTICO

Maria Fuentes Fort

Dep. Informàtica i Matemàtica Aplicada

Universitat de Girona

maria.fuentes@udg.es

<http://ima.udg.es/~mfuentes>

# Contenido de la presentación

- Introducción
  - Resumen mono-documento
  - Resumen multi-documento
  - Resumen no textual
- Nuestra aportación
  - Sistema de marcadores discursivos - UB
  - Sistema de cadenas léxicas - UdG
  - La integración de ambos sistemas
- Evaluación de Sistemas de Resumen Automático

# Definición

*“A summary is a reductive transformation of a source text into a summary text by extraction or generation”*

Sparck-Jones, 2001

- Nota: Parte del material presentado aquí está tomado del tutorial de Horacio Rodríguez del doctorado en IA de LSI en la UPC

# Aplicaciones del resumen automático

- reseñaciones biográficas
- resúmenes de historiales médicos
- resúmenes de correo electrónico
- de páginas Web
- de noticias
- extracción de titulares (headlines)
- apoyo a los sistemas de recuperación de información
- resúmenes de reuniones

# Aproximaciones al resumen (elementos básicos)

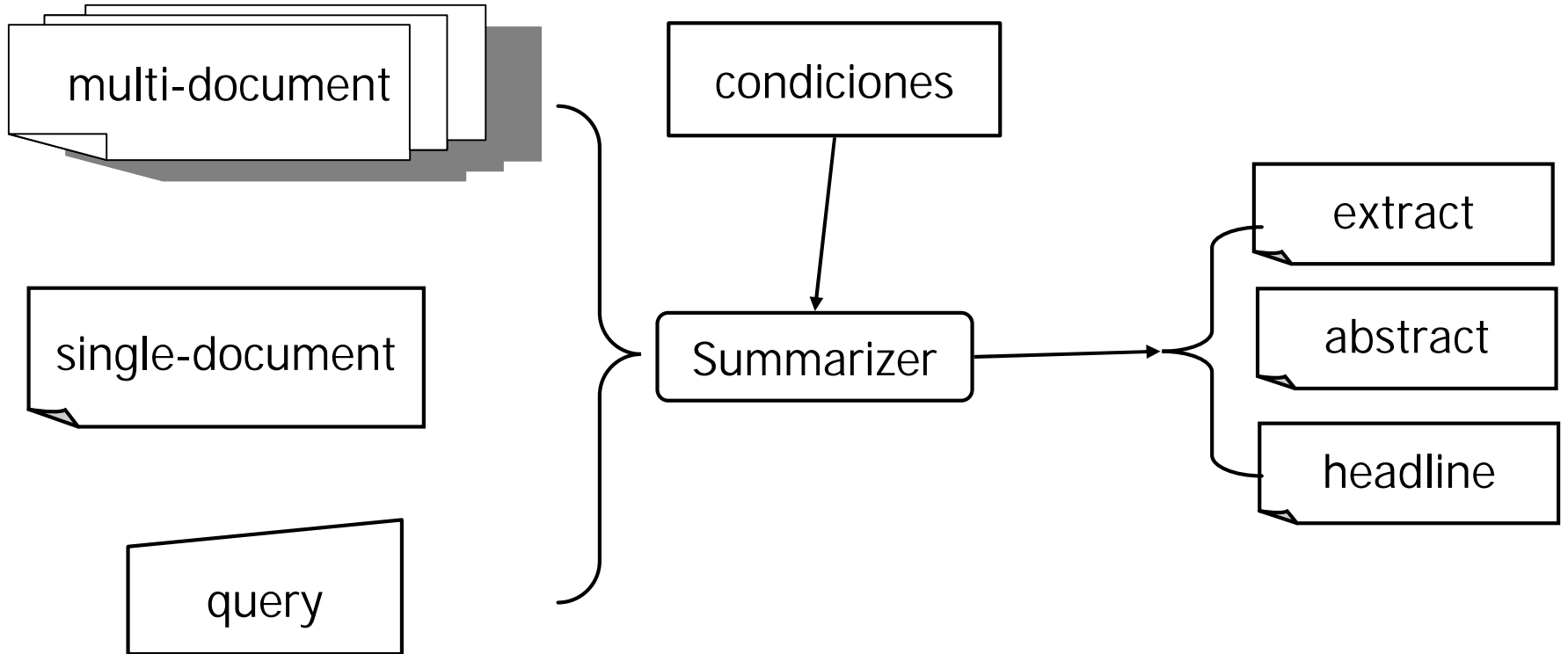
- Top-Down

- query-driven
- necesidades del usuario
  - determinada información
- sistema
  - criterio de interés específico para guiar la búsqueda (especificación de la búsqueda)
  - ejemplos:
    - plantillas con descriptores con características semánticas
    - lista de términos importantes
- Similar a IE  
(Extracción de Información)

- Bottom-up

- text-driven
- necesidades del usuario
  - todo lo importante
- sistema
  - métrica sobre la importancia genérica (estrategia)
  - ejemplos
    - grado de conectividad en un grafo semántico
    - frecuencia de (co-) aparición de términos
- Similar a IR  
(Recuperación de Información)

# Esquema básico



# Tareas

- localizar los fragmentos más relevantes (según las necesidades del usuario)
  - segmentos
  - oraciones
  - párrafos
  - pasajes
- ordenación de estos fragmentos por relevancia
- producción del resumen

# Sistemas mono-documento

## Aproximaciones al resumen

- Posición en el texto
- Términos o frases indicativos
- Frecuencia de palabras
- Estructura discursiva
  - Cohesión
  - Coherencia
- Extracción de información
- Combinación de métodos

# Posición en el texto

- Lead method
  - Lo importante aparece al principio (o al final)
    - $\Rightarrow$  Resumen = n primeras oraciones/párrafos
- Optimum position policy (OPP)
  - Lo importante aparece en posiciones dependientes del género
    - $\Rightarrow$  Aprendizaje automático de las posiciones más prometedoras (a nivel párrafo y oración)
    - Lin,Hovy,1997
    - corpus aprendizaje: 13.000 artículos de prensa (ZIFF corpus)
    - con un factor de compresión del 10% se cubre el 91% de los salient words (index terms)
- Palabras en el título o en los hyperlinks que apuntan a una página o en los índices que describen los textos, o ...

# Términos o frases indicativos (cue phrases)

- “bonus phrases”
  - concluyendo ..., en resumen ..., principalmente ...
- “stigma phrases”
  - difícilmente ..., imposible, ..., no, ...
- Técnicas de ML para localizar estas frases automáticamente
- Bonificación o penalización de las oraciones que contengan estos indicadores

# Sistemas basados en la estructura del discurso

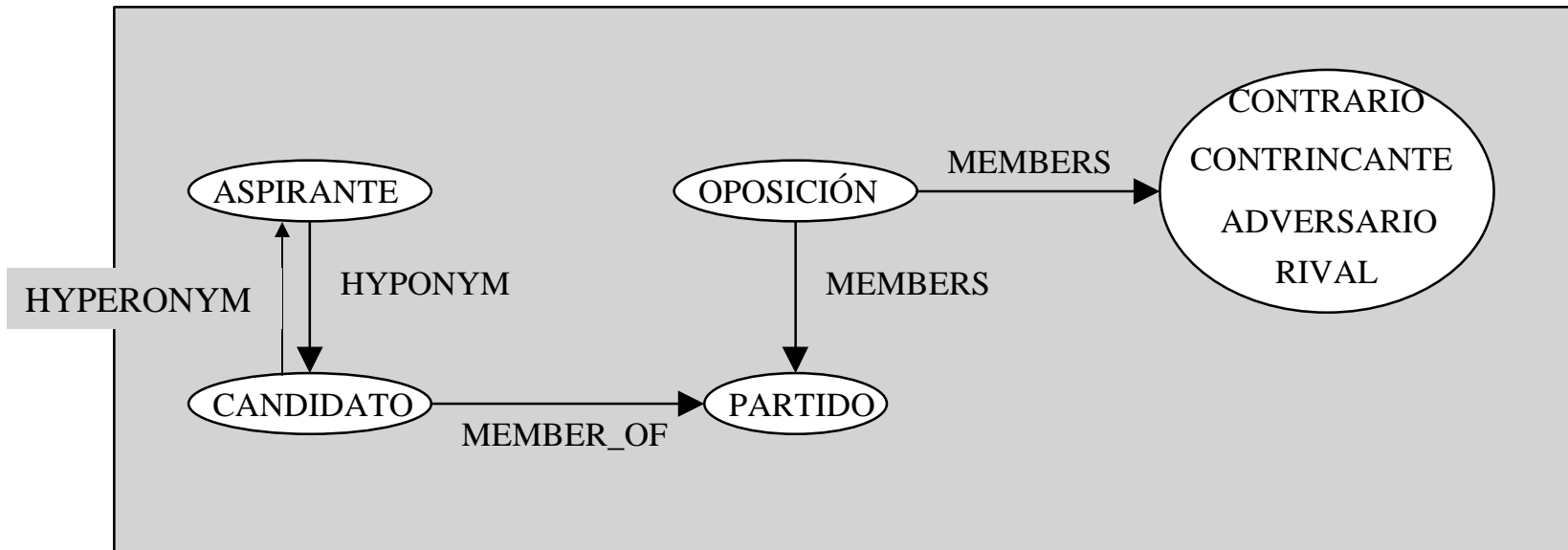
## Relaciones discursivas

- Cohesión
  - relación entre los elementos del texto
  - conectividad no estructural
  - repetición, referencia, conexión léxica
- Coherencia
  - relación entre los segmentos del texto
  - elementos del discurso conectados a través de la estructura semántica
  - elaboración, explicación

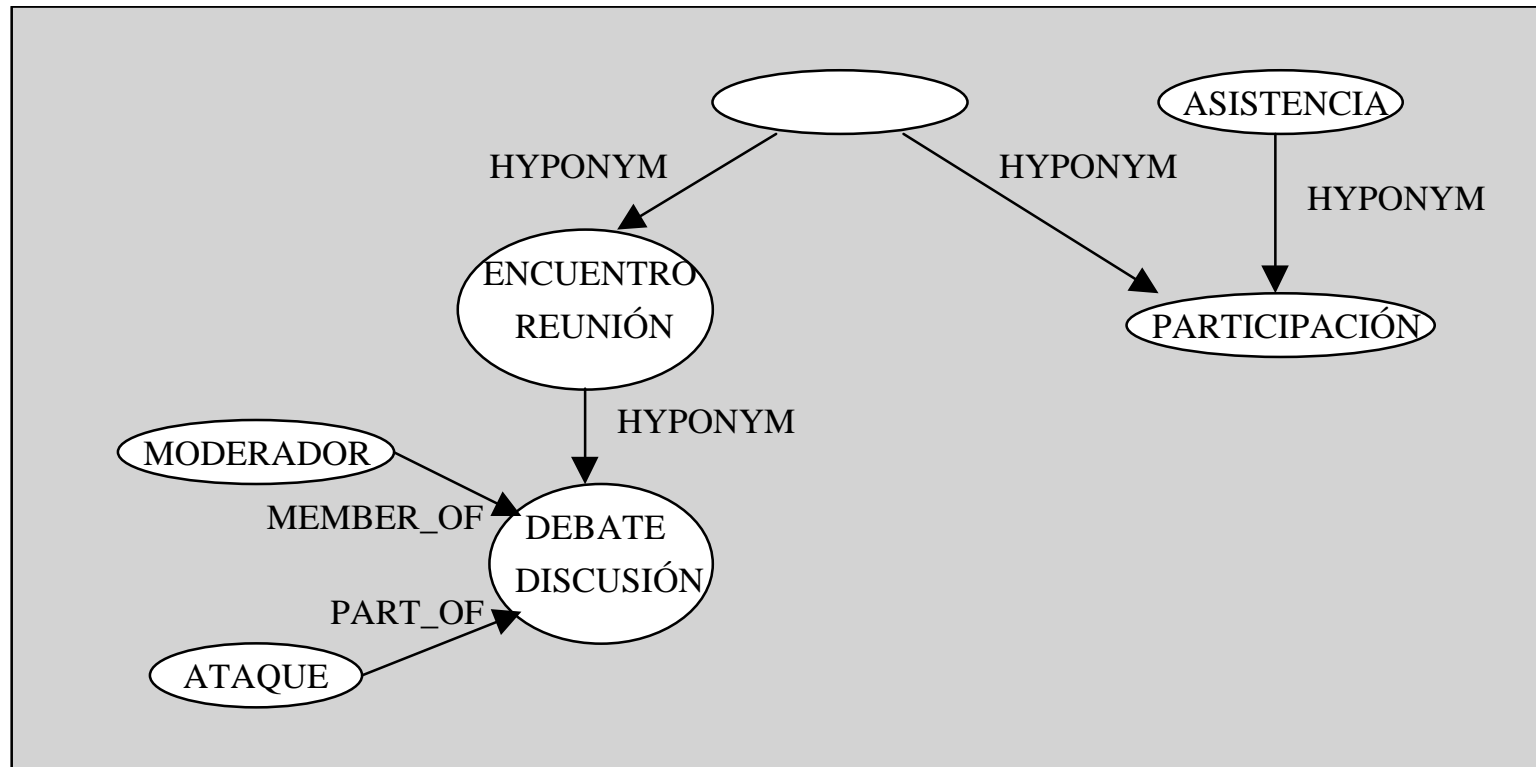
# Sistemas basados en la cohesión

- Relaciones entre los items de un texto (Barzilay, 1997)
  - referencia
  - elipsis
  - conjunción
  - cohesión léxica
    - a través de la selección de las palabras
      - reiteración
      - colocación
    - ⇒ Lexical chains
      - identity chains
        - » cohesión pronominal, repetición, equivalencia
      - similarity chains

# EWN Relaciones - Ejemplo 1



# EWN Relaciones - Ejemplo 2



El conservador **Vicente Fox**, **candidato**<sub>1</sub> del Partido Acción Nacional (PAN) de México cedió hoy ante **sus** rivales, el oficialista Francisco Labastida y el centrozquierdista Cuauhtémoc Cárdenas, en posponer para el próximo viernes el **debate** que estaba previsto para esta noche.

En un encuentro público en la casa de campaña de Cárdenas y frente a los representantes de los medios, los tres **candidatos**<sub>1</sub> discutieron durante unas dos horas sus propuestas sobre el **debate**.

El **candidato**<sub>1</sub> del PAN insistió reiteradamente en celebrar esta misma noche esta **discusión**, mientras que el **candidato** del Partido Revolucionario Institucional (PRI), Cárdenas y Labastida calificaron de "superficial", frívola", "caprichosa", "terquedad" y ca "ligereza" la insistencia de **Fox** en celebrar esta misma noche el **debate**, sin garantizar

la no. En este minidebate, los **candidatos**<sub>1</sub> evitaron ataques personales y se centraron en los puntos de procedimiento, el formato de la reunión, el tipo de moderadores y su papel, el

Con la asistencia de más de un centenar de reporteros de diversos medios, los

**candidatos**<sub>1</sub>. Al concluir la reunión, en un discurso previamente escrito, **Fox** acusó a **sus** contrarios contrincantes de ponerse de acuerdo para boicotear el **debate**, y reiteró su disposición a

Cárdenas reiteró que no existían condiciones técnicas para celebrar el **debate**<sub>2</sub> y "felicizó" a **Fox** por leer un discurso previamente elaborado dirigido a señalar la negativa de los **candidatos** del PRD y del PRI.

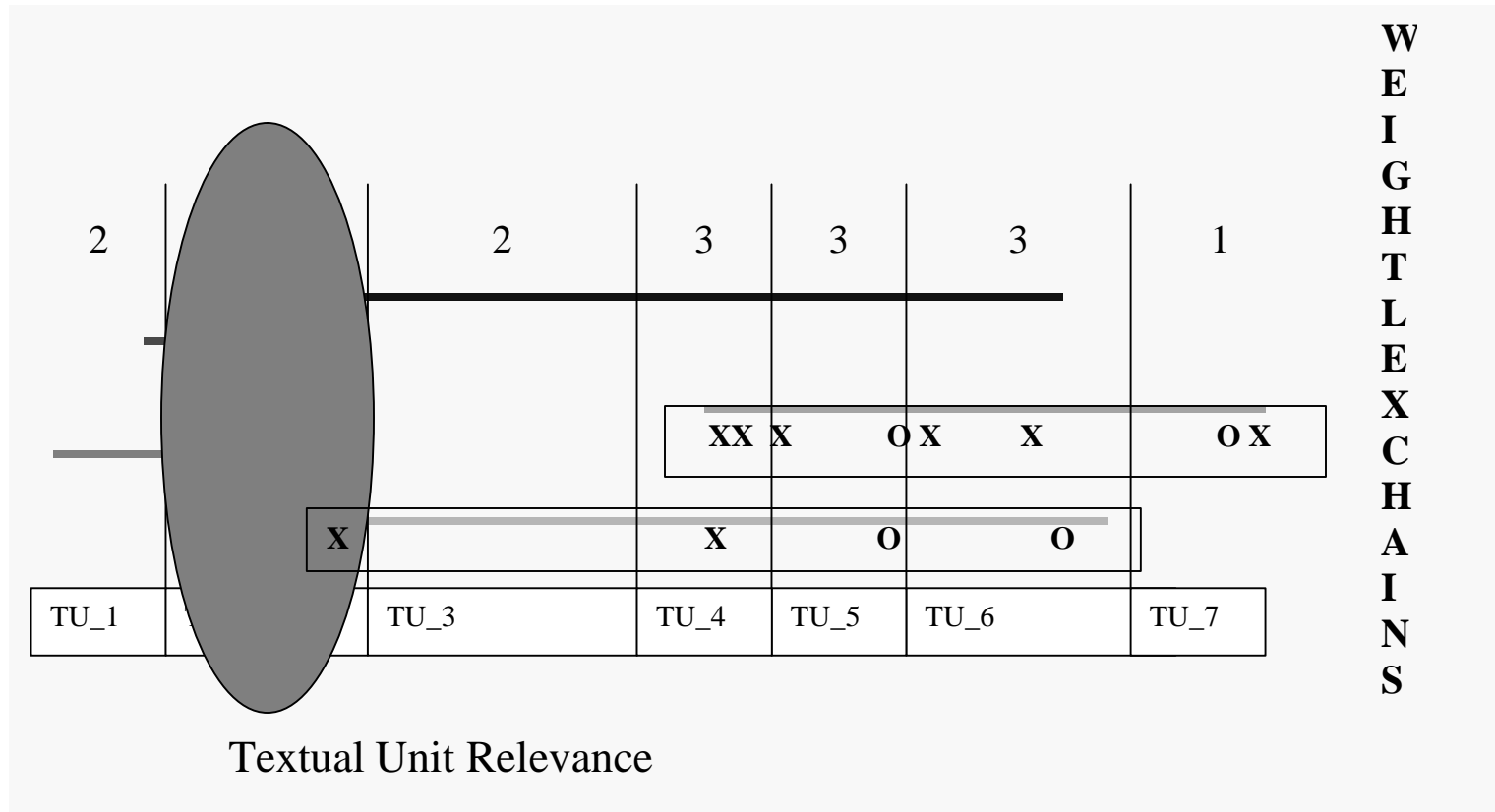
En esta **discusión** participó el presidente de la Cámara de la Industria de Radio y

Televisión. El **debate**<sub>2</sub> previsto para hoy se canceló el domingo después de que **Fox** -quien según la mayoría de los sondeos supera en unos cinco puntos a Labastida- propuso a última hora la participación de tres personas que hicieran preguntas a los **candidatos**<sub>1</sub> sobre temas específicos, pero la iniciativa la rechazaron **sus** adversarios.

El 25 de abril pasado, los seis aspirantes a la presidencia participaron en el primer **debate**<sub>2</sub> y acordaron que para el segundo sólo acudirían los tres con más posibilidades de ganar las elecciones.

La mayor parte de los especialistas han destacado que las elecciones del 2 de julio serán las más reñidas y han señalado la posibilidad de un triunfo de la oposición en la historia de México.

# Cadenas Léxicas de un Texto



# Sistemas basados en la coherencia

D. Marcu, 1997, 1999

- aproximación basada en la coherencia interna del texto.
  - Imprescindible (?) si se desea una alta calidad del resumen
- RSR (Rhetorical Structure Theory)
  - $\Rightarrow$  Representación de la estructura retórica del texto
  - Uso de esta representación para determinar las unidades más relevantes e incluirlas en el resumen

# Proceso de segmentación

Stallone se lo ha tomado con paciencia **y** ya se ha resignado, **aunque** intenta hacer mínimas variaciones.

# Proceso de segmentación

Stallone se lo ha tomado con paciencia **y** ya se ha resignado,  
**aunque** intenta hacer mínimas variaciones.

**CONCESIÓN**

Stallone se lo ha tomado con  
paciencia y ya se ha resignado,

**aunque** intenta hacer  
mínimas variaciones.

**COORDINACIÓN**

Stallone se lo ha  
tomado con paciencia

y ya se ha resignado,

# Interpretación

- Se ha localizado el contenido relevante
- Interpretación
  - a nivel conceptual  $\Rightarrow$  conocimiento semántico
  - a nivel estructural  $\Rightarrow$  conocimiento sintáctico
- generalización conceptual
  - uso de WN
  - uso de gazzeters de Nombres Propios
  - unificación de referentes

# Generación

- nivel 1
  - extracción de segmentos relevantes del texto original sin elaboración
- nivel 2
  - ensamblar diferentes porciones
  - reutilización y regeneración de lenguaje
- nivel 3
  - sistema de generación de LN
    - planificador lingüístico
      - contenido, longitud, tema, orden, realización léxica
    - generador superficial

# Resumen Multidocumento 1

## Objetivos

- Contenido de una colección de documentos

### Briefing

- concise summary of the factual matter of a set of news articles on the same or related events (SUMMONS, Radev, 1999)
- Actualización de información ya conocida
- localización de las secciones de una serie de documentos relevantes para las necesidades de información del usuario

# Resumen Multidocumento 2

## Diferencias entre resumen mono/multi-documento

- Factor de compresión más bajo
- Medidas anti-redundancia
- dimensión temporal
- mayor reto de la correferencia
- aplicación a la búsqueda de información
  - interfaz de usuario

# Resumen Multidocumento 3

## Requisitos

- Clustering de documentos y pasajes
- cobertura
- anti-redundancia
- cohesión del resumen
- calidad
  - legible
  - relevante
  - contexto
- inconsistencias de las fuentes
- actualizaciones

# Resumen multimedia 1

Zechner, 2001

- El tratamiento de documentos no textuales plantea problemas adicionales.
  - Si la entrada procede de un reconocedor de voz (ASR), dotado o no de valores de confianza en las entidades reconocidas
    - se producen disfluencias (speech disfluencies)
  - segmentación en oraciones es mucho más difícil al no existir signos de puntuación
  - La segmentación temática juega un papel importante en esta situación.
    - Un caso especialmente delicado es el de los diálogos multiparticipante (multi-party dialogs) donde las relaciones entre intervenciones (por ejemplo, los pares pregunta/respuesta) deben identificarse.

# Resumen no textual

- Las principales áreas de aplicación en este campo son:
  - diálogos orientados a tareas en dominios restringidos
  - noticias habladas en dominio abierto
  - resumen de diálogos en dominio abierto

# Necesidad de la evaluación

- **RA Esencialmente es una disciplina práctica**

**Karl Popper: Una teoría es científica si es falsable, o sea, empíricamente refutable.**

**=> Búsqueda de contra-ejemplos que puedan mostrar falsedades en nuestras teorías**

➤ **detección de errores**

# Dificultad de la evaluación

- NO EXISTE UN ÚNICO RESUMEN válido para un texto
- Lenguaje natural producido por una máquina
- Personas juzgando incrementa el coste
- Resumir conlleva compresión (reducción de tamaño). Es necesario evaluar resúmenes de distintos tamaños
- Legibilidad (puede no tener relación con la calidad del resumen)
- En relación con las expectativas de la tarea
  - Respuesta a una pregunta concreta
  - Nueva información respecto a los documentos previos
  - Con respecto a un tema concreto (Evento, Persona...)
  - ...

# Métodos de evaluación 1

- **Intrínsecos**
  - Calidad (por personas - No siempre acuerdo)
    - Legibilidad, comprensión, acrónimos, anáforas, integridad de la estructura, gramaticalidad, estilo impersonal, ...
  - Informatividad
    - Información preserva respecto al texto original (varias compresiones)
    - Información contiene respeto a un **resumen ideal**

# Métodos de evaluación 2

- Extrínsecos

Evaluar el uso del resumen en otra tarea

- Encontrar documentos relevantes en una colección
- Decisión tomada leyendo el resumen o el texto original
- Sistemas de Q&A (responden a preguntas)
- Sistemas de recuperación de información
- Contenido páginas web (buscadores)

# Corpus de evaluació

Interfaz para indicar lo relevante que es una oración en el texto

VALORS DE LES DIFERENTS ORACIONS - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: [http://ima.udg.es/~mfuentes/cgi-bin/punts\\_fr.cgi](http://ima.udg.es/~mfuentes/cgi-bin/punts_fr.cgi) What's Related

## Sistema de confecció d'un corpus d'avaluació de resums pel projecte HERMES

Puntuacions de les oracions de la notícia 000.1, per l'avaluador mfuentes

Oració 1:

México, 23 may (EFE).- El conservador Vicente Fox, candidato del Partido Acción Nacional (PAN) de México, cedió hoy ante sus rivales, el oficialista Francisco Labastida y el centroizquierdista Cuauhtémoc Cárdenas, en posponer para el próximo viernes el debate que estaba previsto para esta noche.

Oració 2:

En un encuentro público en la casa de campaña de Cárdenas y frente a los representantes de los medios, los tres candidatos discutieron durante unas dos horas sus propuestas sobre el debate.

Oració 3:

El candidato del PAN insistió reiteradamente en celebrar esta misma noche esta discusión, mientras que el candidato del Partido Revolucionario Institucional (PRI), Francisco Labastida, y Cuauhtémoc Cardenas, del Partido de la Revolución Democrática (PRD), pidieron posponerlo para el viernes a fin de garantizar las condiciones técnicas.

Oració 4:

Cárdenas y Labastida calificaron de "superficial", frívola", "caprichosa", "terquedad" y "ligereza", la insistencia de Fox en celebrar esta misma noche el debate, sin garantizar la neutralidad ni la capacidad de difusión a todos los medios.

# Resumen por evaluadores humanos

## Sistema de confecció d'un corpus d'avaluació de resums pel projecte HERMES

Noticia	Avaluador	Paraules Clau	Puntuacions										Comentaris					
[000.1]	horacio	<ul style="list-style-type: none"><li>- Vicente Fox</li><li>- Francisco Labastida</li><li>- Cuauhtémoc Cárdenas</li><li>- debate</li><li>- posponer</li></ul>	2	0	0	0	0	0	0	0	0	0	1	1	1			
	jordir	<ul style="list-style-type: none"><li>- elecciones</li><li>- México</li><li>- Fox</li><li>- PRI</li></ul>	2	0	1	0	0	0	1	1	0	0	0	0	2			
	josuka	<ul style="list-style-type: none"><li>- debate</li><li>- México</li><li>- elecciones</li></ul>	2	0	2	1	0	0	0	1	0	1	0	2				

# Ejemplo resumen (compresión 10%)

## Sistema de confecció d'un corpus d'avaluació de resums pel projecte HERMES

TIPUS RESUM: 10%

[000.1] Avaluadors (max. puntuacio possible) 5 - Paraules: 459 - Paraules del resum (10%): 45.9

fr pts n\_pars ( frases ordenades per puntuació )

1 5 44 México, 23 may (EFE).- El conservador Vicente Fox, candidato del Partido Acción Nacional (PAN) de México, cedió hoy ante sus rivales, el oficialista Francisco Labastida y el centroizquierdista Cuauhtémoc Cárdenas, en posponer para el próximo viernes el debate que estaba previsto para esta noche.

# Ejemplo resumen (Best Summary)

## Sistema de confecció d'un corpus d'avaluació de resums pel projecte HERMES

TIPUS RESUM: BEST

[000.1] Avaluadors (max. puntuacio possible) 5 - Paraules: 459 - Paraules del resum (10%): 45.9

**fr pts n\_pars ( frases ordenades per puntuació )**

- |    |     |    |  |
|----|-----|----|--|
| 1  | 5   | 44 | México, 23 may (EFE).- El conservador Vicente Fox, candidato del Partido Acción Nacional (PAN) de México, cedió hoy ante sus rivales, el oficialista Francisco Labastida y el centroizquierdista Cuauhtémoc Cárdenas, en posponer para el próximo viernes el debate que estaba previsto para esta noche. |
| 12 | 3.5 | 35 | La mayor parte de los especialistas han destacado que las elecciones del 2 de julio serán las más reñidas y han señalado la posibilidad de un triunfo de la oposición en la historia de México.  |

# Comparación automática

- Recall (Cobertura) : *oraciones del resumen ideal aparecen en el resumen automático*

Precisión : *oraciones que aparecen en el resumen automático y no en el ideal*

- Sentence Rank
- Utility-based measures
- Content-Based (extractats & abstracts)
  - Solapamiento del vocabulario
  - Sinonimia, lematización.
  - Ignoran información sintáctica:  
El gato come pescado < > El pescado come gato

# Resultados

	PRECISION	RECALL	SIMPLE COSINE
Baseline			
Lead	0.95	0.85	0.90
SweSum	0.90	0.81	0.87
Heurística 2			
LexChains	0.70	0.72	0.78
LexChains + PNChains	0.73	0.74	0.81
LexChains + PNChains + coRef Chains	0.70	0.71	0.78
LexChains + PNChains + coRef Chains + 1 <sup>st</sup> UT	0.82	0.82	0.86
Heurística 1			
LexChains	0.82	0.81	0.85
LexChains + PNChains	0.85	0.85	0.88
LexChains + PNChains + coRef Chains	0.83	0.83	0.87
LexChains + PNChains + coRef Chains + 1 <sup>st</sup> UT	0.88	0.88	0.90