Inteligencia en Redes de Comunicaciones

Tema 9

Procesamiento del Lenguaje Natural

Julio Villena Román, Raquel M. Crespo García, José Jesús García Rueda
{jvillena, rcrespo, rueda}@it.uc3m.es

Universidad Carlos III de Madrid

Objetivo

- Desarrollar sistemas informáticos capaces de comprender el lenguaje verbal humano (oral y escrito) y de utilizarlo como medio de comunicación con el usuario

Objetivo realista:

- Desarrollar sistemas informáticos capaces de trabajar con el lenguaje verbal humano (oral y/o escrito) aportando utilidad al usuario

Índice

- Lenguaje natural
- Ingeniería lingüística
- Niveles de análisis
- Aplicaciones

Lenguaje natural y artificial

- Lenguaje natural:
  - Lenguaje verbal que utilizamos los seres humanos para comunicarnos unos con otros
  - Lengua: realización concreta del lenguaje natural

- Lenguaje artificial:
  - Lenguaje creado y especificado detalladamente para ser utilizado en entorno concreto

Ingeniería Lingüística

- El lenguaje natural es de interés en múltiples disciplinas:
  - Lingüística, Filología, Psicología, Antropología, Ingeniería…
- Generalmente los estudios se centran en lenguas concretas
- La ingeniería lingüística se centra en el tratamiento computacional del lenguaje natural y cómo aplicarlo para dar solución a problemas de ingeniería
- Otros nombres:
  - procesamiento del lenguaje natural (PLN, en inglés NLP)
  - lingüística computacional (en inglés, CL)

Breve historia: Los orígenes

- Final de la década de 1940 y década de 1950

  (…antes del nacimiento del término "Inteligencia Artificial" en 1956)

- Dos campos de interés:
  - Traducción automática
  - Recuperación de información
- Muchas limitaciones:
  - Modelos morfológicos y sintácticos poco evolucionados
  - Poco interés en comprensión de significado

Década de 1960

- Cambio de enfoque:
  - Procesamiento de frases y comprensión
  - Interfaz amigable
- Varios desarrollos:
  - Acceso a base de datos (BASEBALL, DEACON, …)
  - Resolución de problemas (STUDENT, CARPS, …)
  - ELIZA
- Desarrollo de formalismos

Década de 1970

- Nuevos desarrollos basados en formalismos
- Aumentan los campos de aplicación:
  - Comprensión de lenguaje (SHRDLU, LUNAR, …)
  - Interfaz a bases de datos (REQUEST, LADDER, …)
  - Enseñanza asistida por ordenador (SCHOLAR, …)
  - Automatización de tareas de oficina (SCHED, …)
  - Programación automática
  - Procesamiento de texto científico

Situación actual

- Las aplicaciones llegan a los usuarios finales
- Busca soluciones a problemas parciales:
  - Revisión lingüística de textos
  - Recuperación de información
  - Extracción de información, resúmenes y clasificación
  - Reconocimiento y síntesis de voz
  - Traducción automática
  - Generación automática de texto…
- Como en toda la ingeniería, el mayor interés está en los aspectos prácticos: utilidad y compromiso exactitud–eficiencia
- Objetivo actual: simplificar la comunicación hombre-máquina (*Human-Computer Interaction, HCI*), desarrollando sistemas de realidad virtual/aumentada

HCI: ELIZA

- Chatbot creado por Joseph Weizenbaum en 1966 que representa el papel de un psicoanalista e imita sus respuestas típicas en una entrevista inicial
- Este contexto es una de las pocas situaciones humanas donde un humano puede responder con una pregunta sin dar ningún tipo de información al interlocutor

  http://www-ai.ijs.si/eliza/eliza.html

  http://www.manifestation.com/neurotoys/eliza.php3
- El concepto impulsó el diseño de nuevos interfaces:
  - *Dungeon* (primer "role playing game") (1974)
  - *Adventure* (primer juego de ficción interactiva) (1975)

    http://en.wikipedia.org/wiki/Interactive_fiction

Diálogo real con ELIZA

young woman: Men are all alike.
eliza: In what way?
young woman: They are always bugging us about something specific or other.
eliza: Can you think of a specific example?
young woman: Well, my boyfriend made me come here.
eliza: Your boyfriend made you come here?
young woman: He says I'm depressed much of the time.
eliza: I am sorry to hear you are depressed.
young woman: It's true. I'm unhappy.
eliza: Do you think coming here will help you not to be unhappy?
young woman: I need some help, that much seems certain.
eliza: What would it mean to you if you got some help?
young woman: Perhaps I could learn to get along with my mother.
eliza: Tell me more about your family.
young woman: My mother takes care of me.
eliza: Who else in your family takes care of you?
young woman: My father.
eliza: Your father.
young woman: You are like my father in some ways.
eliza: What resemblance do you see?
young woman: You are not very aggressive but I think you don't want me to notice that.
eliza: What makes you think I am not very aggressive?
young woman: You don't argue with me.
eliza: Why do you think I don't argue with you?
young woman: You are afraid of me.
eliza: Does it please you to believe I am afraid of you?
young woman: My father is afraid of everybody.
eliza: What else comes to mind when you think of your father?
young woman: Bullies.
eliza: Does that have anything to do with the fact that your boyfriend made you come here?

Diálogo ficticio con ELIZA

young woman: Green cows **are** all **alike**.
eliza: In what way?
young woman: Seven **are always** five plus two.
eliza: Can you think of a specific example?
young woman: Well, my teacher **made me** study a lot.
eliza: Your teacher made you study a lot?
young woman: He says **I'm depressed and bla and bla and more bla bla.**
eliza: I am sorry to hear you are depressed and bla and bla and more bla bla.
young woman: No se me ocurren más tonterías que escribir. **I'm unhappy**.
eliza: Do you think coming here will help you not to be unhappy?
…
…

Realiza análisis superficial y transformaciones básicas

HCI: ALICE

- Desarrollado por Dr. Richard Wallace (1999-) y ganador durante años del premio Loebner (último en 2004)
- Chatbot bastante más evolucionado, que se basa en el paradigma estímulo-respuesta escritos mediante patrones en el lenguaje AIML

```
<category>
  <pattern>WHAT ARE YOU</pattern>
  <template>
    <think><set name="topic">Me</set></think>
    I am the latest result in artificial intelligence.
  </template>
</category>
```

```
<category>
  <pattern>KNOCK KNOCK</pattern>
  <template>Who is there?</template>
</category>
<category>
  <pattern>*</pattern>
  <that>WHO IS THERE</that>
  <template><person/> who?</template>
</category>
```

Tiene memoria y "estado mental", con lo que puede recordar sesiones anteriores

- Puede entrenarse de forma relativamente sencilla para diferentes ámbitos de conocimiento

http://alice.pandorabots.com/

HCI en la actualidad

- Jabberwacky y Joan: desarrollados por Rollo Carpenter y ganadores del premio Loebner (2005 y 2006)
- Guarda todo lo que se va diciendo y proporciona la respuesta más apropiada que existe en su base de datos usando búsqueda de patrones conceptuales ➔ APRENDE
- En cierta forma modela la forma en que los humanos aprendemos el idioma, los hechos y las reglas

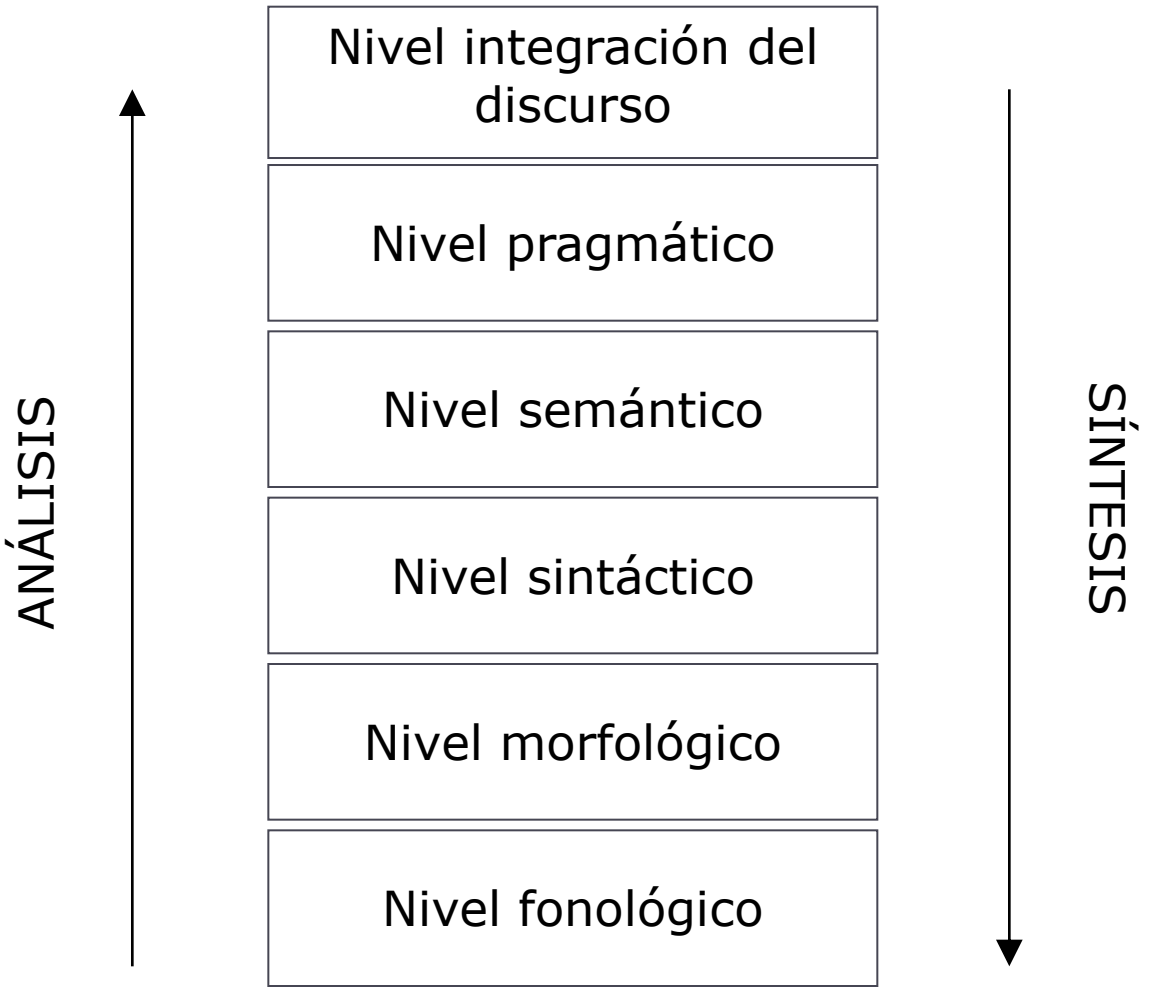
  http://www.jabberwacky.com

- Más sobre el premio Loebner:

  http://www.loebner.net/Prizef/loebner-prize.html

  http://loebner.net/Prizef/2007_Contest/Rules.html

Arquitectura de niveles



Nivel fonológico

- Conversión Voz ⇔ Texto
- Requisitos:
  - Conocimiento de fonemas
  - Algoritmo de reconocimiento
- Es muy importante el tratamiento de la ambigüedad:

  /baca/, /ora/

  - Requiere conocimiento de niveles superiores (al menos, morfológico y sintáctico)
  - Confusión del significado, pérdida de información

Nivel morfológico

- Palabra ⇔ Análisis morfológico *(POS: part-of-speech)*
  - Lema
  - Categoría gramatical
  - Atributos propios de categoría
- Requisitos:
  - Conocimiento de los formantes
    - raíz (*cas-*, *com-*) + desinencias (*-a*, *-s*, *-o*, *-ía*, *super-*)
  - Gramática de palabra
- Ambigüedad…

  casa, *sobre*, *bajo*

Recursos morfológicos: Base léxica

- Una lista de palabras no suele valer (en general)
- Base léxica: almacén de información fundamentalmente morfológica, aprovechando las regularidades de la lengua y escrita para linguistas
- Para español:
  - Modelos de flexión nominal y verbal
    - Palabras formadas por uno o dos formantes

      sobre perr-o perr-os com-emos
    - Cada formante aporta parte de información
  - Derivación de adverbios en -mente (adj→fem→-mente)
  - Generación automática de alomorfos
- No sobregenerar ni sobreaceptar

Nivel sintáctico

- Análisis morfológico ⇔ Análisis sintáctico
  - Estructura en árbol de agrupaciones de palabras y relaciones
- Requisitos:
  - Información morfológica de palabras (léxico)
  - Gramática de frase
- Una gramática general es difícil (por no decir imposible)
- Complejidad del léxico vs. complejidad de la gramática (directamente proporcional)
- Ambigüedad…

  Se comió el helado con cuchara

  Se comió el helado con vainilla

Nivel semántico

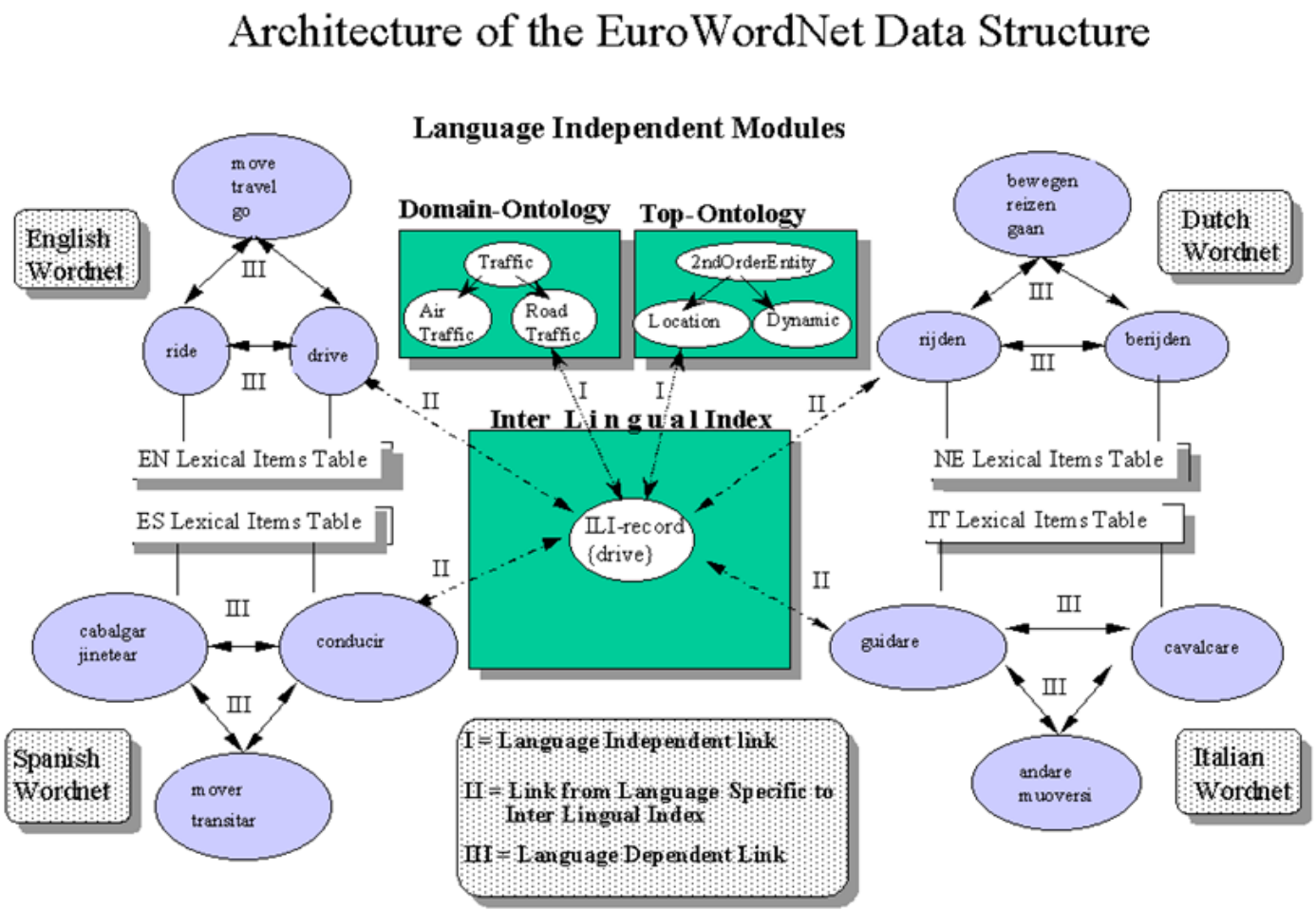
- Análisis sintáctico ⇔ Semántica de frase
  - Significado literal de la frase
- Requisitos:
  - Modelo del mundo
  - Reglas semánticas
- Como es muy complicado, es totalmente dependiente de la aplicación concreta (dominio restringido)
- Ambigüedad…

  Pasé delante del banco

Recursos semánticos: WordNet

- Red semántica de conceptos o *synsets* (agrupaciones de sinónimos) http://wordnet.princeton.edu
- Guarda diferente información:
  - Nombres
    - Hiperónimos: Y es un hiperónimo de X si todo X es un tipo de Y
    - Hipónimos: Y es un hipónimo de X si cada Y es un tipo de X
    - Términos coordinados: Y es un término coordinado de X si X y Y comparten un hiperónimo
    - Holónimos: Y es un holónimo de X si X es parte de Y
    - Merónimos: Y es un merónimo de X si Y es parte de X
  - Verbos
    - Hiperónimos: Y es un hiperónimo de X si la actividad X es un tipo de Y (viajar→moverse)
    - Tropónimos: Y es un tropónimo de X si la actividad Y está haciendo X de alguna manera (susurrar→hablar)
    - Vinculación: Y está vinculado a X si al hacer X también se está haciendo Y (dormir→roncar)
    - Términos coordinados: verbos que comparten un hiperónimo común
  - Adjetivos
    - Nombres relacionados
    - Participios verbales
  - Adverbios
    - Adjetivos origen
- Pensado para uso por personas:
  - Significado de palabras en forma textual
  - Demasiada información

EuroWordnet



Niveles pragmático y de integración del discurso

- Nivel pragmático:

  Significado literal de frase ⇔ Significado real de frase

  ¿Puedes pasarme la sal?

- Nivel de integración del discurso:

  Significado de frase aislada ⇔ Significado en contexto

  Me dijo que se lo daría

- Ambigüedad intrínseca

Dificultades

- **Ambigüedad**
  - En la mayoría de casos, para resolver la ambigüedad en un nivel se requiere de los análisis de niveles superiores
- Modelos lingüísticos insuficientes
- Sintaxis implica gramática dependiente de contexto
- Tratamiento de semántica
- Niveles superiores a semántica aún más complejos
- Abordable sólo parcialmente con arquitectura de niveles
- Aplicaciones muy variadas ➔ Solución general difícil
- Diferencias entre lenguas
- Inserción de conocimiento manual

Caso del español (o castellano)

- Problemas (nivel morfológico):
  - Altamente flexivo: Múltiples procesos (flexión, derivación, composición)
  - No existen modelos morfológicos generales (muchas excepciones)
  - Número de palabras inmenso (decenas de millones)
  - 1,6-1,9 análisis por palabra (media)
- Problemas (nivel sintáctico):
  - Carencia de estructura fija como en otros idiomas (ambigüedad)

Pero…

- Para resolver grandes problemas deben resolverse antes subproblemas pequeños
- Es posible desarrollar sistemas realmente útiles
- El tiempo corre a nuestro favor
  - Ordenadores más potentes
  - Formalismos más desarrollados
  - Más experiencias y desarrollos

Aplicación: Análisis y síntesis de voz

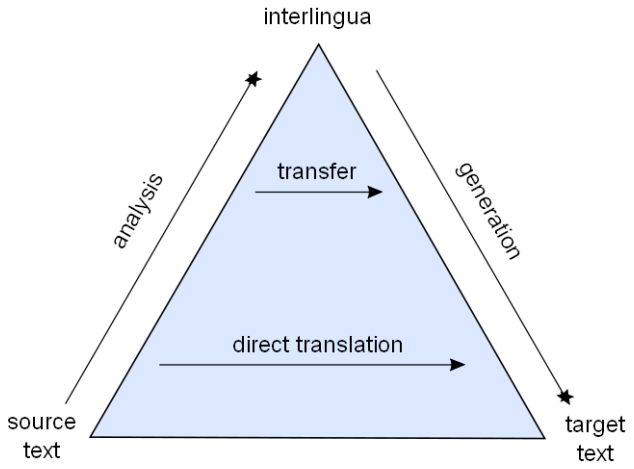
- Primeros productos realmente útiles desde 1997
- Técnicas de procesado de señal de audio + clasificadores + vocabulario (+ gramáticas del lenguaje)
- Actualmente múltiples *motores*:
  - IBM
  - Scansoft/Nuance
  - Microsoft (Speech API)
    - Integrado en Windows Vista
  - Loquendo
  - Desarrollos gratuitos…
- Interfaz para muchos entornos e idiomas

http://cepstral.com/demos/
http://www.loquendo.com/en/demo-center/interactive-tts-demo/

Aplicación: Traducción automática

- Desde los primeros tiempos del PLN
- Caso de éxito: TAUM-METEO (U. Montreal, 1975) para la traducción de partes meteorológicos inglés→francés

  http://en.wikipedia.org/wiki/METEO_System
- Vocabularios y memorias de traducción + gramáticas de transformación de estructuras
- Corpus alineados
- Muchos sistemas:
  - SYSTRAN
  - Altavista Babelfish
  - Google Translate
  - Apertium, OpenTrad (libres)



Aplicación: Revisión lingüística

- Gramática de errores
- Lenguaje declarativo compilado
- Análisis en niveles:
  - Nivel I: estructuras independientes
    *me se ha olvidado, *¡cuanto tiempo sin verte!
  - Nivel II: errores intrasintagmáticos
    *los coches rojo
  - Nivel III: errores intersintagmáticos
    *los niños juega, *la película es divertido
- Compromiso precisión – rendimiento:
  - Sólo considera análisis más probable
  - Reglas particulares / generales
- Mucho interés hoy en día para el aprendizaje de idiomas

Ejemplo de regla

```
/*
 - está mucho loco
 + está muy loco
 + hay mucho loco por aquí
*/
REGLA("MuchoPorMuy")
   FORMA_I_EXISTENCIAL(POS(N), "mucho") Y
   (ANALISIS_EXISTENCIAL(POS(N+1), Eti_AdjetivoOParticipio) O
    FORMA_I_EXISTENCIAL(POS(N+1), "bien|mal")) Y
   !ANALISIS_EXISTENCIAL(POS(N+1), Eti_AdjComp"|"Eti_AdjSup) Y
   GN(POS(N), POS(N+1)) Y
   LEMA_EX_VERBO_PRINC(POS(N-1), "estar|ser")
ENTONCES
   SUG_PALABRA(POS(N), "muy ");
   SUG_PALABRA(POS(N+1),LETRAS(POS(N+1)));
   DAR_ERROR(Error_Gramatical, POS(N), POS(N+1),
                        "Posible secuencia incorrecta de palabras"));
FIN
```

Aplicación: Recuperación de información

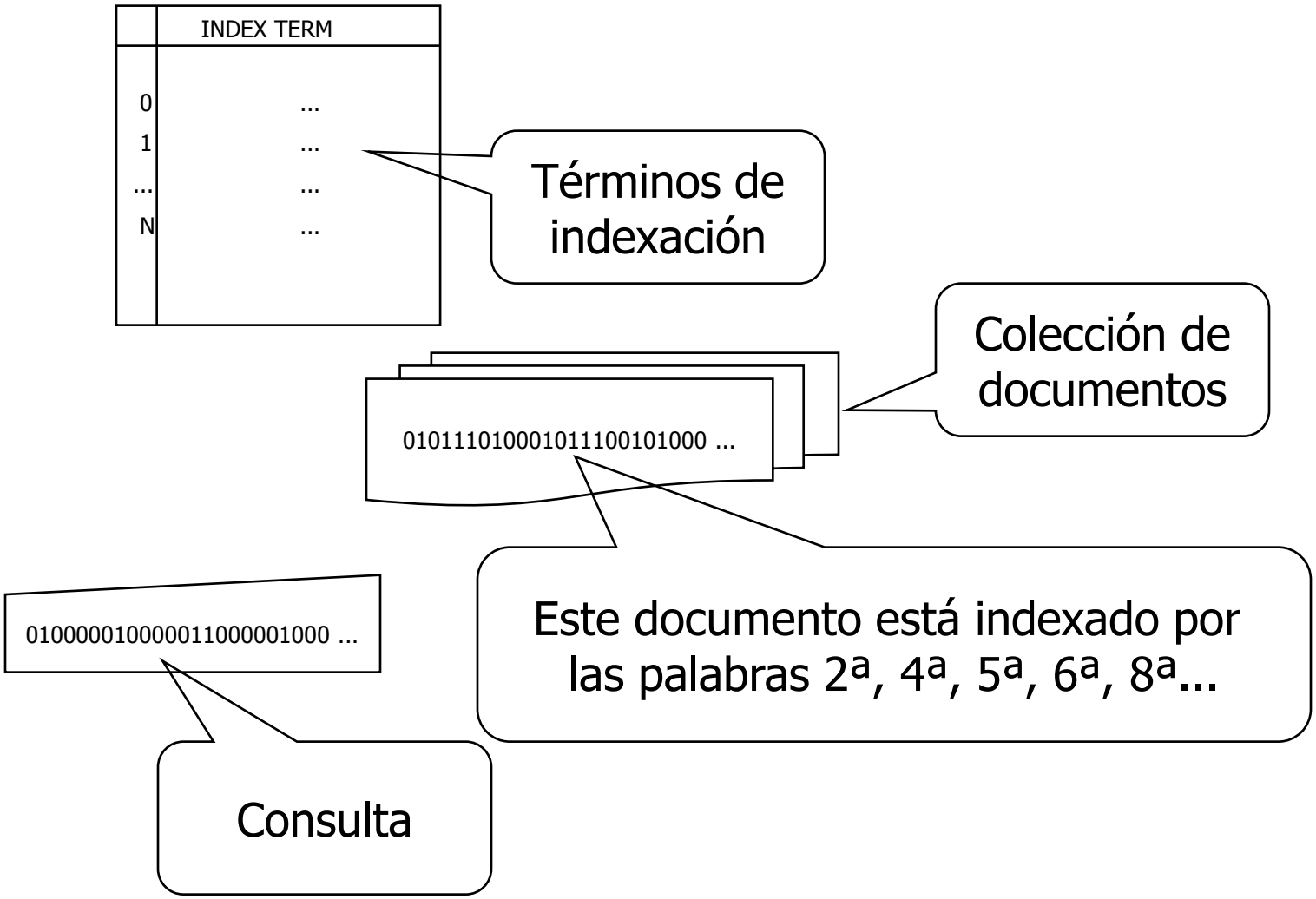
- Los sistemas de RI son aquellos que
  - Basándose en distintas técnicas y modelos,
  - Permiten buscar de forma rápida y eficiente
  - En grandes colecciones de objetos que contienen información
  - Aquellos resultados más relevantes para la consulta de usuario.
- Los objetos pueden, en general, contener información en una gran variedad de formatos, incluyendo texto escrito, ficheros de audio, fotografías y otras imágenes, vídeo, etc.

Proceso de RI

- Indexación:
  - Extraer los atributos de cada uno de los objetos
    - Texto: frecuencia de palabra
    - Imágenes: extracción de características de la señal
    - Audio (hablado): conversión a texto + frecuencia de palabra
    - Audio (música): ¿partitura?
  - Almacenarlos en una base de datos de acceso rápido
- Búsqueda:
  - Comparar la consulta del usuario con todos los objetos indexados, obteniendo una medida de parecido (relevancia)
  - Presentar los resultados ordenando por relevancia decreciente
- Modelos:
  - Modelo booleano
  - Modelo probabilístico
  - Modelo de espacio de vectores
  - Latent Semantic Indexing

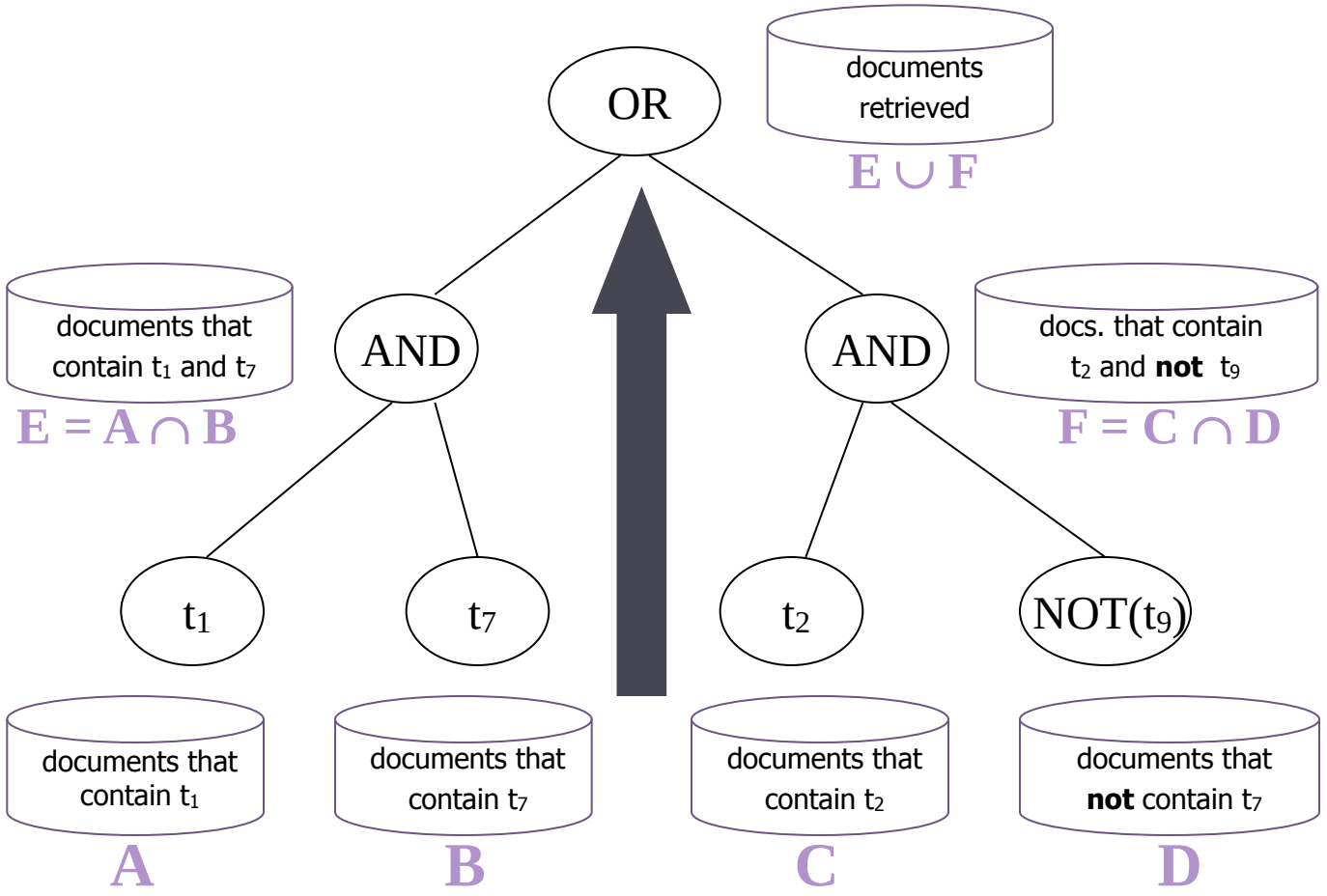
http://en.wikipedia.org/wiki/Information_retrieval

Modelo Booleano

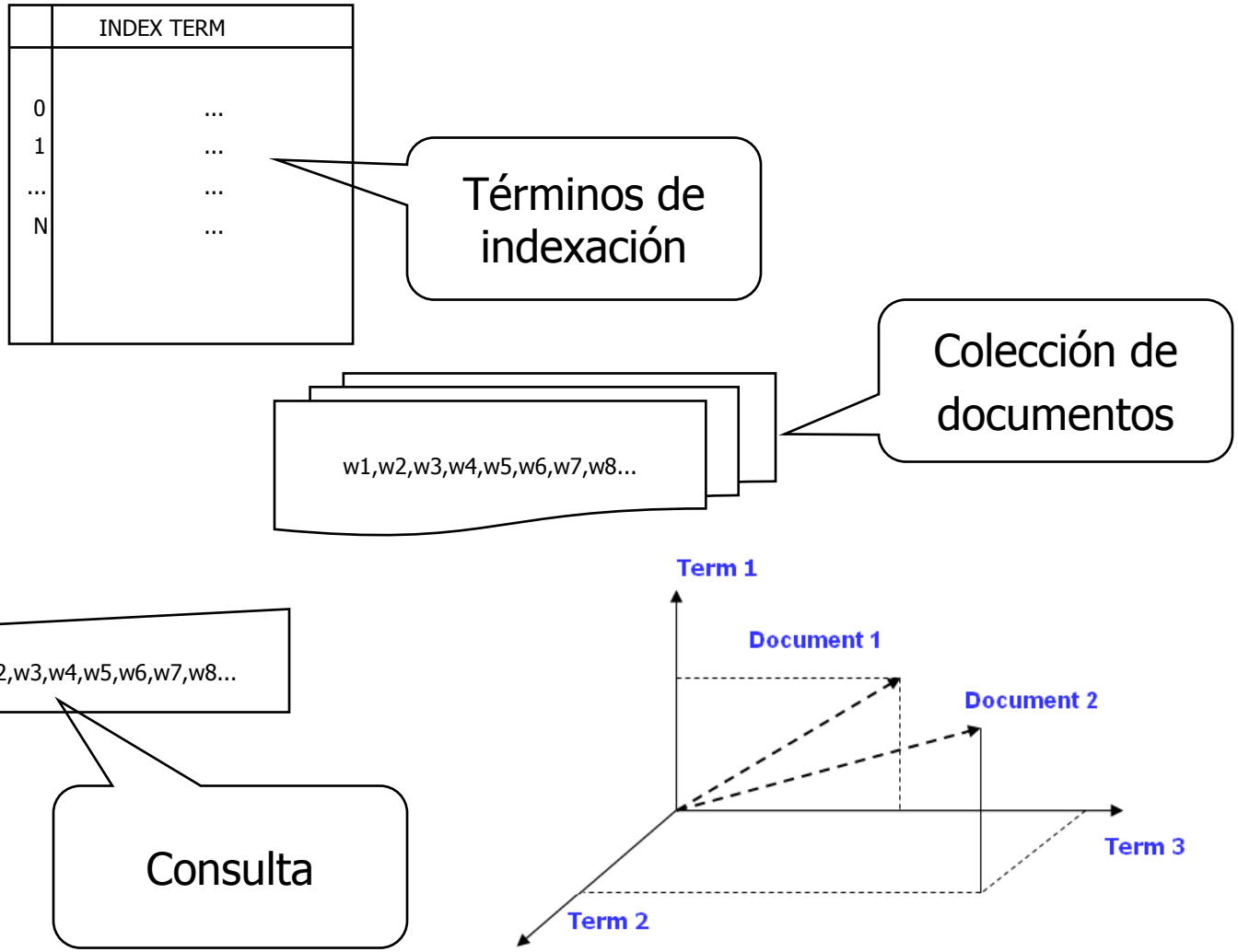


Ejemplo

(t1 AND t7) OR (t2 AND NOT(t9))



Modelo de Espacio de Vectores



Modelo de Espacio de Vectores (2)

- Modelo algebraico clásico de RI [Salton, 1968]
- Representa los objetos (documentos y consulta) mediante un vector de términos en un espacio multidimensional:

$$\mathbf{v}_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T$$

- El peso de cada término se calcula con el modelo TF·IDF:

$$w_i = tf_i * log\left(\frac{D}{df_i}\right)$$

- Para calcular la relevancia de cada documento se utiliza habitualmente la fórmula del coseno:

$$\mathrm{Sim}(Q, D_i) = \frac{\sum_i w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \sqrt{\sum_i w_{i,j}^2}}$$

Ejemplo

Query, Q: "gold silver truck"
D_1: "Shipment of gold damaged in a fire"
D_2: "Delivery of silver arrived in a silver truck"
D_3: "Shipment of gold arrived in a truck"
$D = 3$; $IDF = \log(D/df_i)$

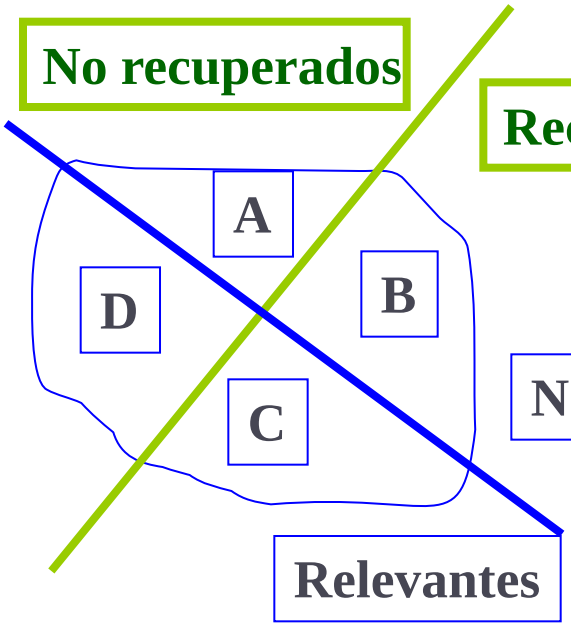
	Counts, tf_i							Weights, $w_i = tf_i * IDF_i$			
Terms	Q	D_1	D_2	D_3	df_i	D/df_i	IDF_i	Q	D_1	D_2	D_3
a	0	1	1	1	3	3/3 = 1	0	0	0	0	0
arrived	0	0	1	1	2	3/2 = 1.5	0.1761	0	0	0.1761	0.1761
damaged	0	1	0	0	1	3/1 = 3	0.4771	0	0.4771	0	0
delivery	0	0	1	0	1	3/1 = 3	0.4771	0	0	0.4771	0
fire	0	1	0	0	1	3/1 = 3	0.4771	0	0.4771	0	0
gold	1	1	0	1	2	3/2 = 1.5	0.1761	0.1761	0.1761	0	0.1761
in	0	1	1	1	3	3/3 = 1	0	0	0	0	0
of	0	1	1	1	3	3/3 = 1	0	0	0	0	0
silver	1	0	2	0	1	3/1 = 3	0.4771	0.4771	0	0.9542	0
shipment	0	1	0	1	2	3/2 = 1.5	0.1761	0	0.1761	0	0.1761
truck	1	0	1	1	2	3/2 = 1.5	0.1761	0.1761	0	0.1761	0.1761

$$\text{Sim}\,\theta_{D_1} = \frac{0.0310}{0.5382 * 0.7192} = 0.0801$$

$$\text{Sim}\,\theta_{D_2} = \frac{0.4862}{0.5382 * 1.0955} = 0.8246$$

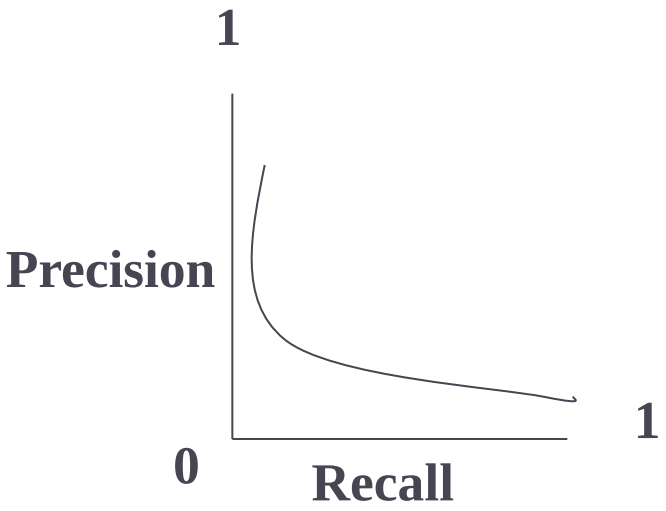
$$\text{Sim}\,\theta_{D_3} = \frac{0.0620}{0.5382 * 0.3522} = 0.3271$$

Evaluación de los sistemas de RI



$$\text{Recall} = \frac{C}{D + C}$$

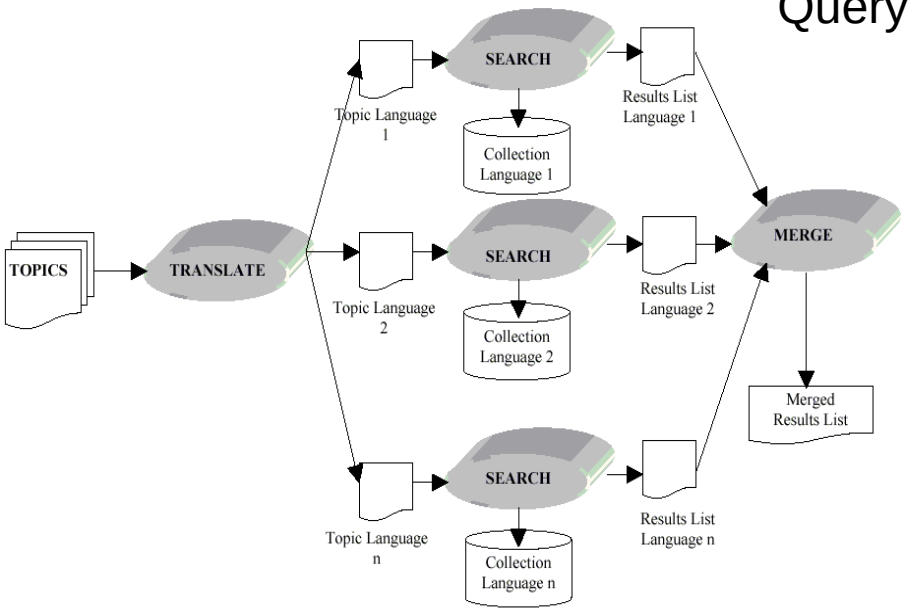
$$\text{Precision} = \frac{C}{B + C}$$



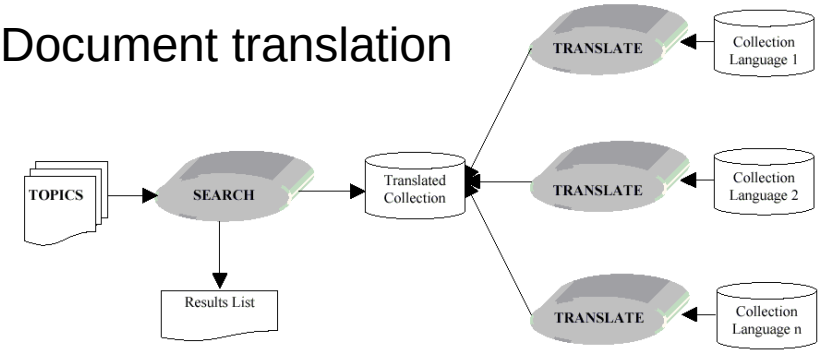
Cross-lingual IR – CLIR

Enfoques para CLIR

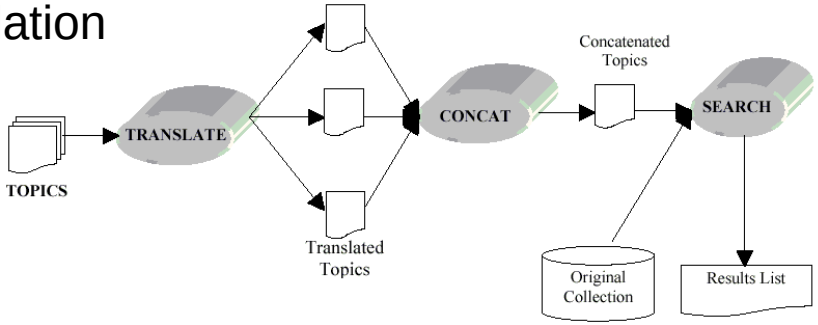
Query translation



Document translation



Mixed translation



RI multimedia (imágenes/vídeo/audio)

Show me images of an infected wound.
Zeige mir Bilder einer infizierten Wunde.
Montre-moi des images d'une plaie infectée.

Show me images of findings with Alzheimer's Disease.
Zeige mir Bilder von Fällen mit einer Alzheimer Diagnose.
Montre-moi des images d'observations avec la maladie d'Alzeimer.

Aplicación: Extracción de información

- Evolución de la recuperación de información
- El sistema no sólo presenta la lista de objetos que contienen la información, sino que la extrae de ellos
- Information Extraction

  http://en.wikipedia.org/wiki/Information_extraction
- Question answering

  http://en.wikipedia.org/wiki/Question_answering

  http://www.answers.com/bb/
- Generación de resúmenes
- Esteganografía (Cifrado)

  http://en.wikipedia.org/wiki/Steganography

  http://www.spammimic.com/

Aplicación: Clasificación de información

- Clasificación de texto en categorías
  - Clasificación de noticias
  - Filtros antispam
  - Sistemas de diagnóstico automático
- Extracción del vector de características del texto + segmentación o clasificación