Grau d'Intel.ligència Artificial

Processament del Llenguatge Humà

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Índice

Índice

Definición

- TLH es la tecnología enfocada al estudio del lenguaje humano desde el punto de vista computacional.
- TLH comprende métodos computacionales, recursos y modelos específicamente diseñados para el tratamiento de todo tipo de texto:
  - lista de palabras
  - pregunta en lenguaje natural
  - documento en formato electrónico (p.e., texto plano, página web, sms, tweet, transcripción oral)
  - **corpus**: colección de documentos en formato electrónico

Definición

- TLH es una área multidisciplinar:
  - **Procesamiento del Lenguaje Natural (PLN)**
  - Lingüística Computacional
  - Inteligencia Artificial
  - Procesamiento del Habla
  - Ciencia Cognitiva , Psicología
  - Lógica, Matemáticas

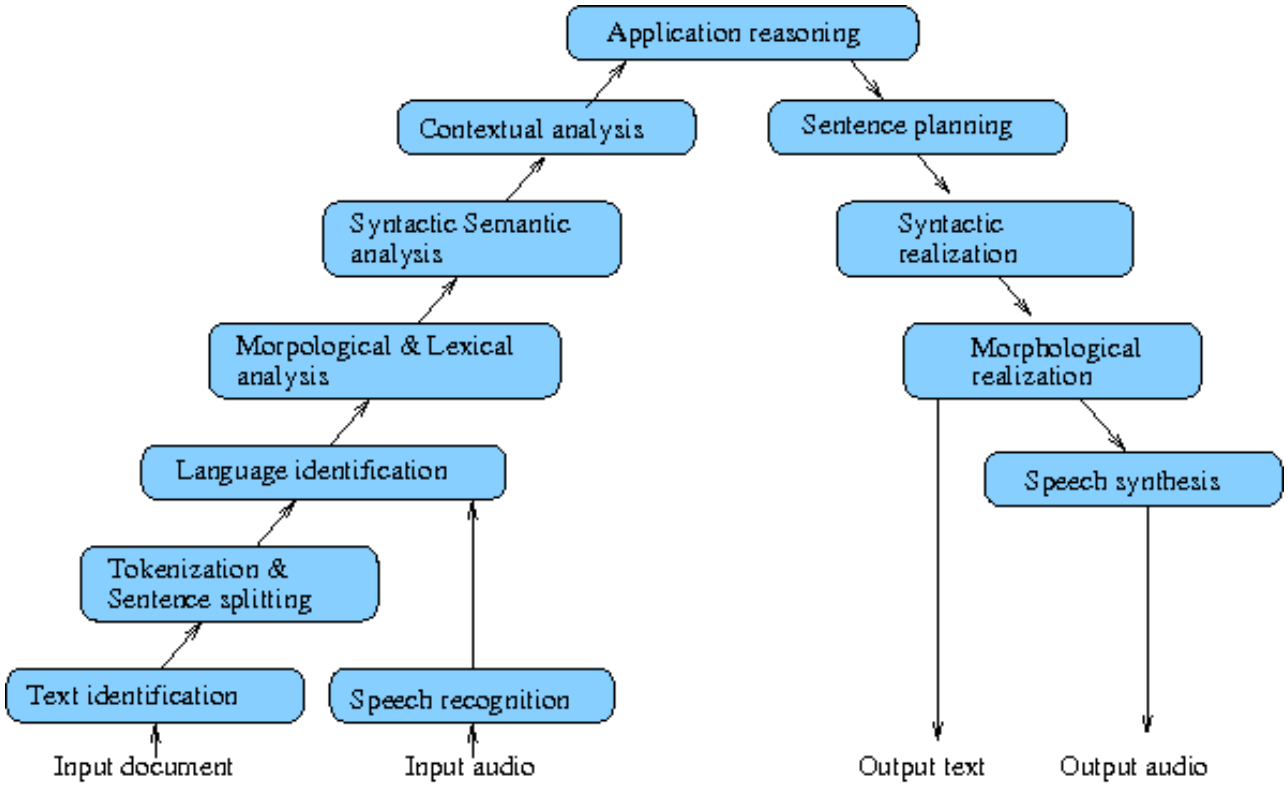
Índice

Definiciones

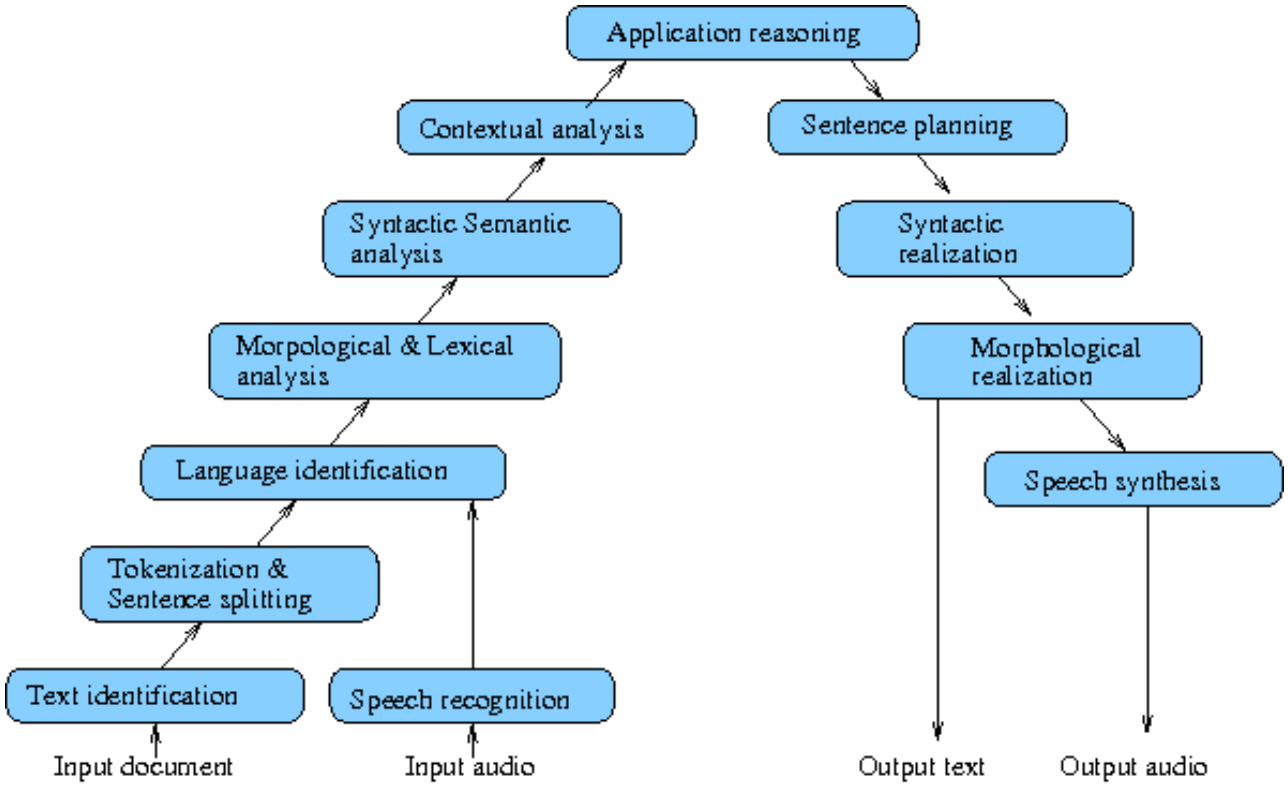
La estrategia general sigue las subáreas estandares de la lingüística

- Fonética: sonidos del habla humana.
  P.e., *infrequent* $\rightarrow$ `/ɪn'frikwənt/`
- Morfología: formación estructural de las palabras.
  P.e., *in-frequent-ly*.
- Sintaxis: relaciones estructurales entre palabras de una misma oración.
  P.e., *un determinante es seguido de un nombre común*.
- Semántica: significados de las palabras aisladas y en el contexto de una oración.
  P.e., *the president of USA is Donald Trump* $\rightarrow$ `president(USA, Donald_Trump)`
- Pragmática: significado efectivo del texto.
  P.e., **He** *is very well known in* **his country** [sarcasm]

Arquitectura general

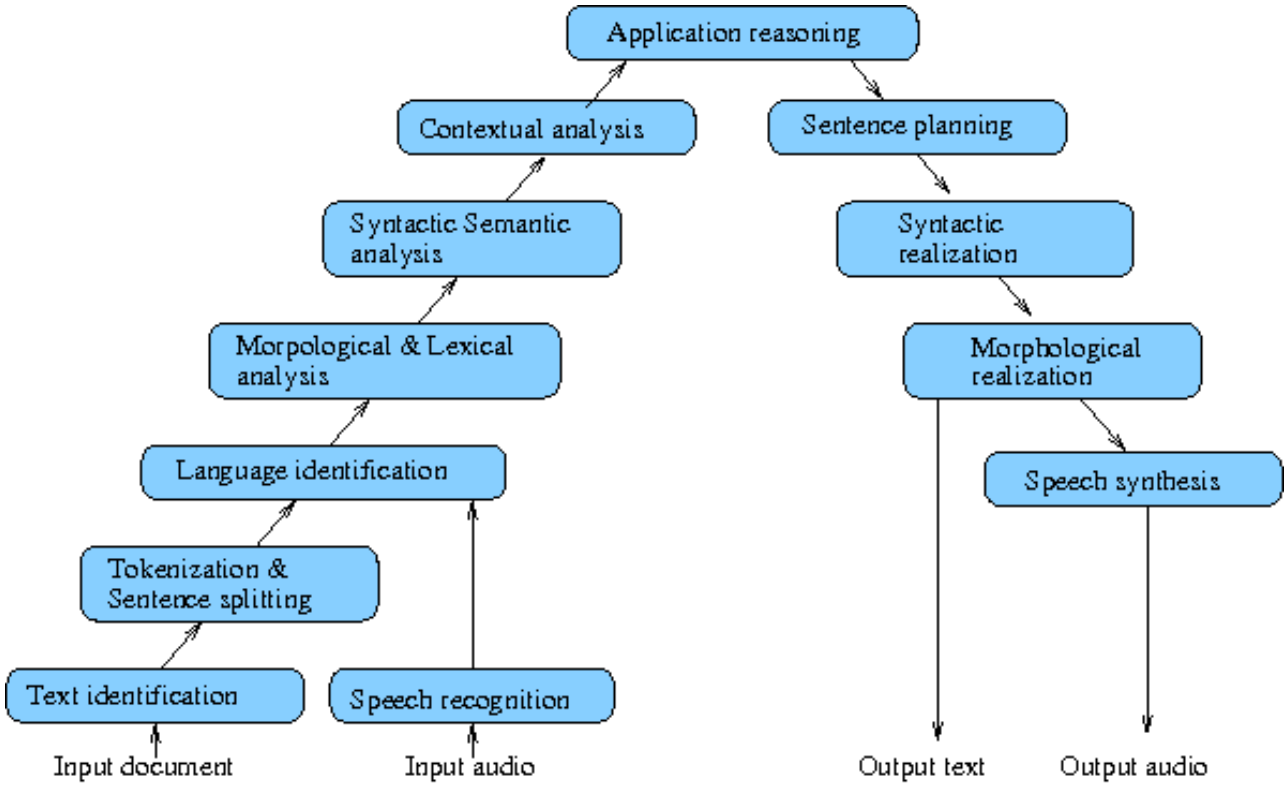


Arquitectura general



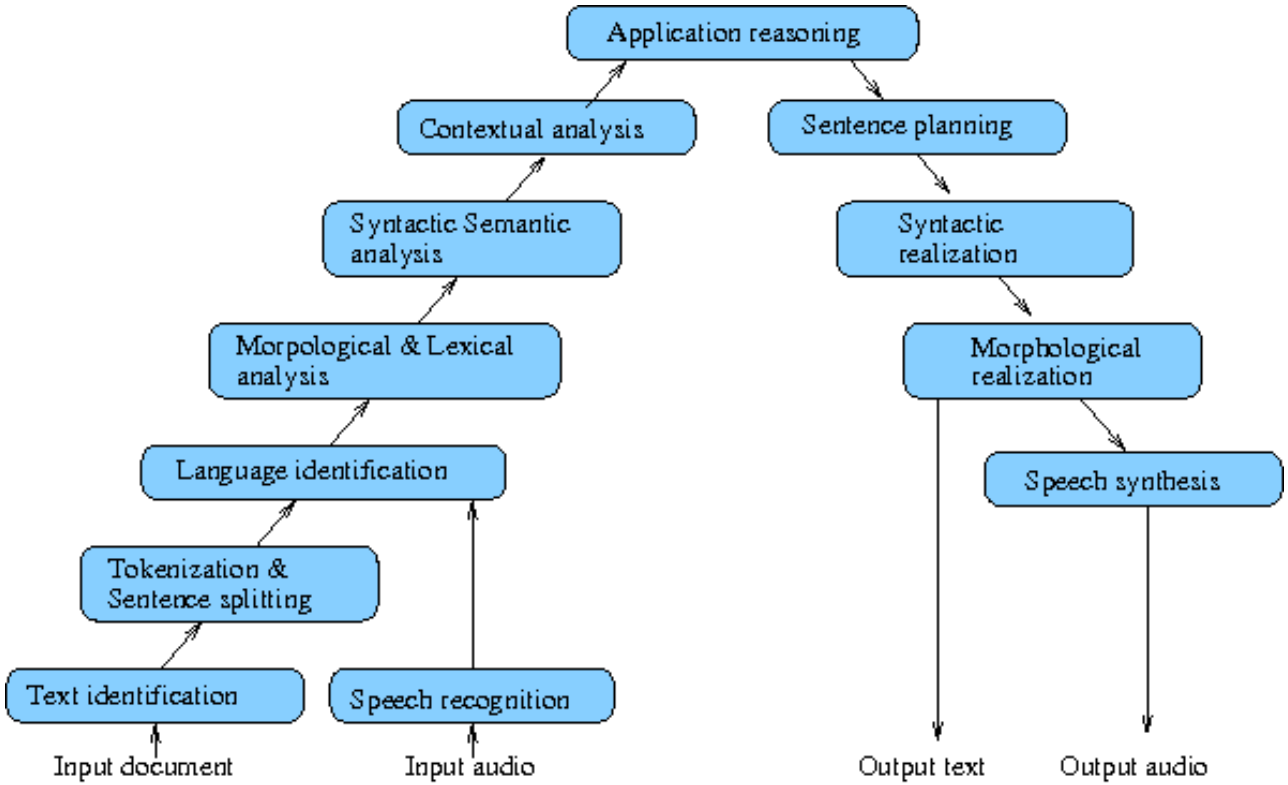
- Ramas: Interpretación de LN y Generación de LN.

Arquitectura general



- Ramas: Interpretación de LN y Generación de LN.
- Aproximaciones: basadas en el conocimiento *vs.* basadas en corpus.

Arquitectura general



- Ramas: Interpretación de LN y Generación de LN.
- Aproximaciones: basadas en el conocimiento *vs.* basadas en corpus.
- Métodos superficiales (*pattern matching*) *vs* métodos complejos (análisis semántico, inferencia lógica)

Índice

Problemática

- Conocimiento del mundo
  - Representar el conocimiento del mundo es necesario para interpretar el LN (AI-completitiud)
    p.e, Yago - hechos, OpenCyc - sentido común
- Multilingualidad
  - Diferentes idiomas requieren diferentes modelos y recursos
  - Uso de palabras de otros idiomas
    `Estoy a full!` (texto non estandar en castellano)
- Evaluación
  - Correctitud/Idoneidad de una traducción/resumen
- Variabilidad
  - diferentes oraciones con un mismo significado
    `Where can I get a map?`
    `I need a map`
    `need map` (inglés no estandar)
- Ambigüedad
  - Una oración con varios significados
    `Esther said about Alice: ''I made her duck''`

Ambigüedad

E.g., `Esther said about Alice: ''I made her duck''`

- `Cociné pato para ella`
- `Cociné su pato`
- `Creé su pato (escultura)`
- `Hice que ella se inclinase`
- `La convertí en pato (mágia)`

Palabra	Ambigüedad	Alternativas
make	semántica	cook o create
her	sintáctica	pronombre posesivo o dativo
	pragmática	Esther o Alice
duck	sint-sem	nombre o verbo

Índice

Ejemplos de aplicaciones

- Agrupación de Documentos
- Clasificación de Documentos (p.e.. anti-spamming, email routing, sentiment polarity, language identification)
- Recuperación de Información (RI)
- Corrección de Texto
- Detección de plagio
- Extracción de Información (EI)
- Generación de Resumen
- *Question Answering* (QA)
- Traducción Automática (TA)
- Sistemas de Diálogo

...

Recuperación de Información (RI)

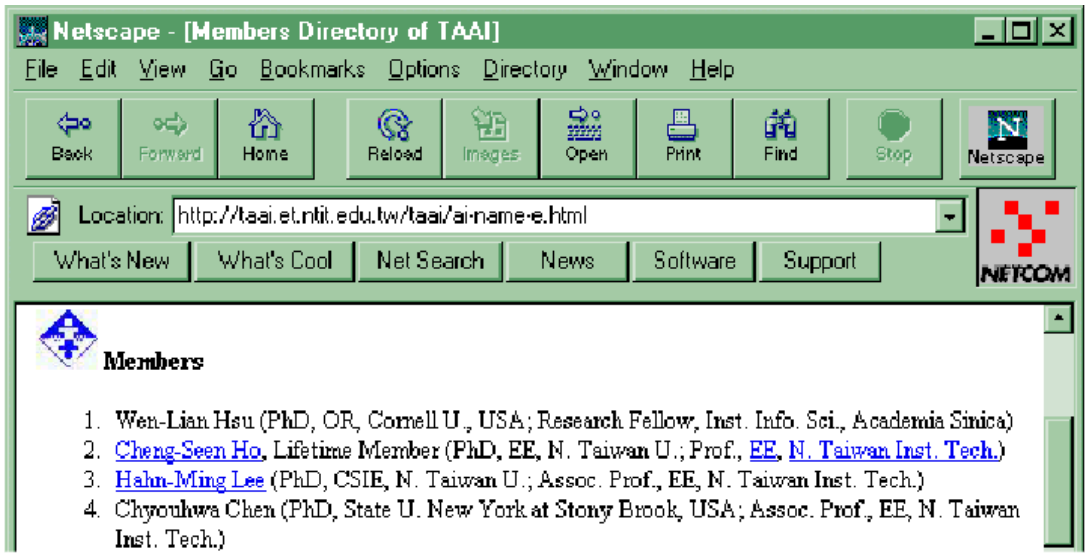
- P.e.: Buscadores (Google, Yahoo, ...)
- dado un corpus, $D = \{D_i\}$, y una *query* (lista de palabras), Q , devolver $\hat{D} \subset D$ que mejor casa con Q.
- $sim(v(Q), v(D_i))$, donde $v(X)$ representa X en un espacio vectorial
- ¿Qué espacio vectorial funciona mejor?
  - palabras? $Q =$"window", $D_i =$"...he closed the windows..."
  - lemas? $Q =$"window", $D_i =$"...he closed Windows..."
  - compuestos? $Q =$"Energie", $D_i =$"...Sonnenenergie..."
  - ...
  - Técnicas complejas de PLN no parecen ser productivas

Extracción de Información (EI)

- P.e.: ampliar el contenido de BDs o BCs. Indiciar colecciones de documentos. Análisis del Sentimiento.
- Extraer la **información relevante** contenida en texto (entidades, propiedades, relaciones and eventos).
- Subtareas principales:
  - Reconocimiento y clasificación de entidades nombradas (NERC)
  - Extracción de propiedades (Slot Filling)
  - Extracción de relaciones
  - Extracción de eventos
- Dependiendo de la tarea específica, se requiere un PLN más complejo, así como técnicas de aprendizaje automático (AA)

Extracción de Información (EI)

- Ejemplo 1: Nombre, Grado, Colegio y Aficliación de páginas WEB.



Name	Degree	Affiliation	School
Wen-Lan Hsu	PhD, OR, Cornell U., USA	Research Fellow	Inst. Info. Sci. Academia Sinica
Chen-Seen Hu	PhD, EE, N. Taiwan U.	Prof.	EE, N. Taiwan Inst. Tech
Hahn-Ming Lee	PhD, CSIE, N. Taiwan U.	Prof.	EE,N. Taiwan Inst. Tech
...			

Extracción de Información (EI)

- Ejemplo 2: incidentes a partir de texto plano (tipo de incidente perpetrador, objetivo, fecha, lugar, efectos, instrumento).

```
At 5pm on  Thursday , a  white Fiat van  veered off the road and into a
crowd outside the Plaça de Catalunya metro station in Barcelona.   The  van
continued down  Las Ramblas  for more than 500 metres while  crashing  into
 pedestrians .   13 people have been killed .   100 people were injured  and
 15 are in serious condition .  Las Ramblas attacker  Younes Abouyaaqoub
 was  killed in Subirats.
```

Extracción de Información (EI)

- Ejemplo 2: incidentes a partir de texto plano (tipo de incidente perpetrador, objetivo, fecha, lugar, efectos, instrumento).

At 5pm on Thursday, a white Fiat van veered off the road and into a crowd outside the Plaça de Catalunya metro station in Barcelona. The van continued down Las Ramblas for more than 500 metres while crashing into pedestrians. 13 people have been killed. 100 people were injured and 15 are in serious condition. Las Ramblas attacker Younes Abouyaaqoub was killed in Subirats.

tipo de incidente = crash
lugar = Las Ramblas (Barcelona)
fecha = 17/8/2017
perpetrador = Younes Abouyaaqoub
objetivo = pedestrians
instrumento = white Fiat van
efectos = 13 people killed, 100 people injured, 15 people in serious condition

Resumen Automático

- P.e., Generar biografías, minutas de reuniones, *abstracts* o extractos de documentos escritos.
- Métodos abstractivos:
  - Generan nuevo texto a partir de la información relevante del texto de entrada.
  - Requieren la interpretación y la generación del lenguaje
- Métodos extractivos:
  - Seleccionan las oraciones mas relevantes del texto de entrada.
  - El conjunto de oraciones seleccionado debe maximizar la relevancia y coherencia del resultado y minimizar la redundancia.
- ¿Cómo se computan la *relevancia* y la *redundancia*?
- La semántica, la prágmática y las técnicas de AA ayudan

Question Answering (QA)

- P.e.: Preguntas realizadas a coches o salas inteligentes.
- Dado un corpus $D = \{D_i\}$, y una pregunta, Q, producir la respuesta exacta a Q a partir de D.
  - QA factoide: las respuestas son hechos concretos extraidos de D

    P.e..: `Who was the president of the USA in 1987?`
  - QA no factoides: la respuesta es una definición, una explicación, un resumen biográfico

    P.e..: `Tell me what has been said so far in the meeting`
- Principales subtareas:
  - Indiciación del corpus
  - Procesamiento de la pregunta (tipo de pregunta, foco de la pregunta)
  - Producción de la respuesta
- La semántica, la prágmática y las técnicas de AA ayudan

Traducción Automática (TA)

- P.e.: Traducción de texto escrito, ayuda a la comunicación humano-humano, traducción online

- TA clásica (TA basada en reglas o TA estadística):
  - Parte la oración origen o en palabras o en grupos sintácticos
  - Mapea palabra con palabra o grupo con grupo usando poco contexto para tomar decisiones
- TA neuronal:
  - Mapea oración origen con oración destino
  - Usa el contexto amplio de las palabras y grupos sintácticos a cada paso

- Resultados aceptables pero no comparables con la traducción humana

Traducción Automática (TA)

Ejemplos de errores: (con Google Translate - traducción neuronal)

- Trabajan oración a oración: falta de contexto

```
ES: Ana no aprobó el examen.  Su amigo sí.
EN: Ana did not pass the exam.  Your friend yes.
ok:  Ana did not pass the exam.  Her friend did.
```

Traducción Automática (TA)

Ejemplos de errores: (con Google Translate - traducción neuronal)

- Trabajan oración a oración: falta de contexto

  ES: Ana no aprobó el examen. Su amigo sí.
  EN: Ana did not pass the exam. **Your** friend **yes**.
  ok: Ana did not pass the exam. Her friend did.

- Falta de conocimiento del mundo: Named entities

  ES: Disfrutar es el mejor nuevo restaurante de Europa
  EN: **Enjoy** is the best new restaurant in Europe
  ok: Disfrutar is the best new restaurant in Europe

Traducción Automática (TA)

Ejemplos de errores: (con Google Translate - traducción neuronal)

- Trabajan oración a oración: falta de contexto

  ES: Ana no aprobó el examen. Su amigo sí.
  EN: Ana did not pass the exam. **Your** friend **yes**.
  ok: Ana did not pass the exam. Her friend did.

- Falta de conocimiento del mundo: Named entities

  ES: Disfrutar es el mejor nuevo restaurante de Europa
  EN: **Enjoy** is the best new restaurant in Europe
  ok: Disfrutar is the best new restaurant in Europe

- Dominios específicos: terminología

  ES: El níscalo se cría bajo pinos
  EN: **The níscalo** grows under pines
  ok: Red pine mushroom grows under pines

  ES: Los níscalos se crían bajo pinos
  EN: **The chanterelles are raised** under pines
  ok: Red pine mushrooms grow under pines

Sistemas de Diálogo

- E.g.: chatbots, asistencia médica automática, QA dirigida por diálogo en coches o salas inteligentes.
- Ayuda a los usuarios a conseguir objetivos específicos via interaccion en lenguale natural
- Principales subtareas:
  - Interpretar la intervención del usuario
  - Determinar la siguiente acción del sistema considerando la intervención del usuario (responder una pregunta, preguntar por más información, ...)
  - Generar la intervención del sistema
- Requiere interpretación y generación del lenguaje. Métodos de AA ayudan en cada subtarea

Índice

Contenido

Podeis encontrar el contenido del curso y su planificación dentro del Racó. ▸ LINK

Evaluación de la asignatura

- Examen final: todo el contenido. Se realiza en periodo de examenes
- Sesiones de laboratorio: grupos de 2 estudiantes.
  - Se realizan ejercicios prácticos en cada sesión
  - Desarrollo de 4 prácticas
- Nota final = 50% Examen + 50% Prácticas