Introducción

Departamento de Ingeniería de Sistemas Telemáticos
http://moodle.dit.upm.es

Vision

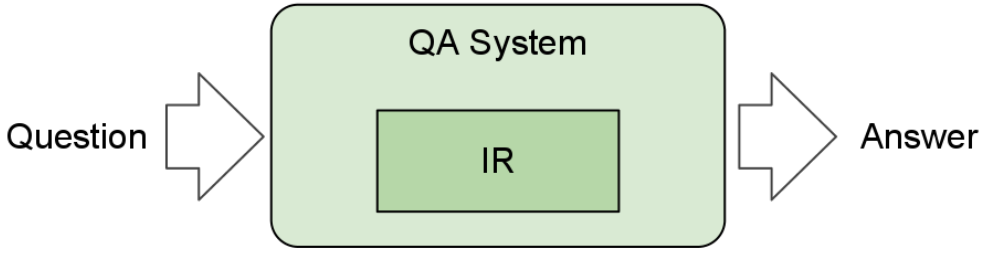
"When I think about computers in the future I imagina myself talking to them"

Dr. David Ferrucci

Question Answering

- IR: encontrar documentos relevantes a una determinada consulta (query)
  - **I**nformation **R**etrieval
  - Query: generalmente una combinación de palabras clave – keywords
- QA: encontrar la respuesta a una pregunta
  - Expresada en lenguaje natural
  - La respuesta será una frase corta

Question Answering



Ejemplos

Q: Who shot President Abraham Lincoln?
A: John Wilkes Booth

Q: How many lives were lost in the Pan Am crash in Lockerbie?
A: 270

Q: How long does it take to travel from London to Paris through the Channel?
A: three hours 45 minutes

Q: Which Atlantic hurricane had the highest recorded wind speed?
A: Gilbert (200 mph)

Q: Which country has the largest part of the rainforest?
A: Brazil (60%)

Características

- Independiente del tiempo
  - Respuesta en tiempo real
  - Capaz de dar a respuesta a preguntas sobre hechos recientes o del pasado
- Exactitud
  - Dar respuesta que debe ser idónea
- Usabilidad
  - Respuesta comprensible por el usuario
- Completitud

Tipos de preguntas

- Preguntas de hechos
  - Requieren una respuesta específica: fechas, nombres, cantidades, etc.
  - Ej: *¿Cuál es la capital de España?*
- Preguntas de definición:
  - Buscan la explicación de un hecho o concepto
  - Ej: *¿Cómo se produce un Eclipse?*

Tipos de preguntas

- Preguntas sobre listas:
  - Una lista de factores que puede estar estructurada a modo de resumen
  - Ej: *¿Cuáles han sido los factores que han propiciado la actual crisis mundial?*
- Preguntas contextuales:
  - No son un tipo en sí. Se refieren al contexto
  - Ej: *¿Quiénes han sido los responsables de ello?*

Tipos de preguntas

- Preguntas especulativas
  - Alta complejidad
  - Requieren de técnicas deductivas
  - Ej: *¿Qué podría pasar si cayese un rayo sobre un coche?*

Etapas de un sistema QA

Análisis de la pregunta

- Pretende clasificar cada pregunta en categorías
  - Sobre personas, cosas, lugares, fechas, etc
  - Sobre propiedades o atributos
  - Sobre una clasificación o taxonomía

Análisis de la pregunta

- El tipo de pregunta determina qué procesado/análisis/consulta se realiza
  - Qué respuesta se espera obtener
  - Palabras clave
- Ejemplo
  - *¿Quién es el rey de España?*
    - Respuesta esperada: **nombre**
    - Palabras clave: **rey, España**

Patrones de pregunta

- Extraen información de la pregunta
  - Palabras clave

```
(who|what|which) be <target>
(who|what|which) be <target>
   (in|of|on) <context>
(who|what) be <context>'s <target>
(how|what) do you call <target>
```


Patrones de pregunta

- Pueden responder al mismo patrón preguntas de muy distinta naturaleza

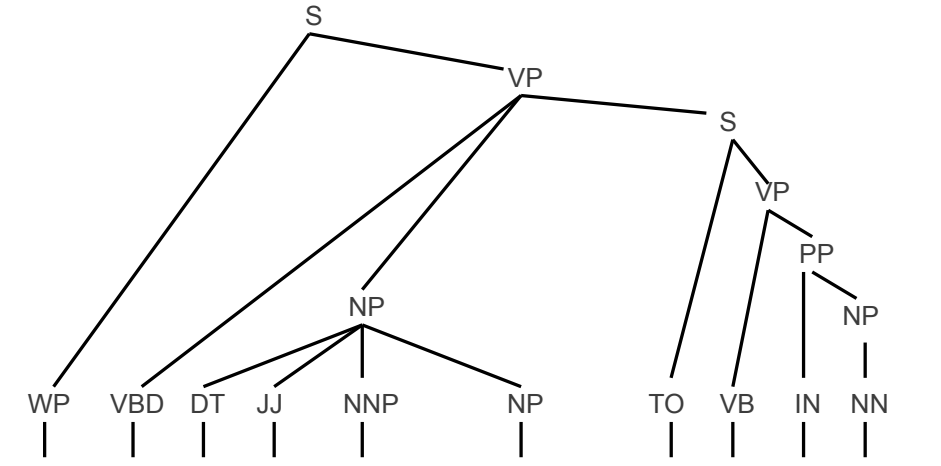
```
(who|what|which) be <target>
```

  - "Who is the king of Spain?"
  - "Who is the most famous football player?"
  - "Who was the first Russian astronaut to walk in space?"
  - "Who is your favourite singer?"

Procesado lenguaje (NLP)

- Las herramientas de procesado de lenguaje natural son muy útiles para
  - OpenNLP, Treetagger, Freeling, Stanford NLP
  - http://nlp.lsi.upc.edu/freeling/demo/demo.php
- Permiten crear patrones de mayor complejidad
  - En función de análisis sintáctico
  - NLP RegExp

Procesado lenguaje (NLP)



Herramientas NLP

Funcionalidad	OpenNLP	Stanford NLP	Freeling	TreeTagger
Tokenización	✓	✓	✓	✗
Retokenización	✓	✗	✗	✗
Detección de frases	✓	✗	✓	✗
Etiquetado PoS	✓	✓	✓	✓
Chunking	✓	✓	✓	✓
Lematización	✗	✗	✓	✓
Análisis sintáctico	✓	✓	✓	✗
NER	✓	✓	✓	✗
Correferencia	✓	✓	✓	✗
Lenguaje	Java	Java	C++	Perl
Wrapper para Java	—	—	✓	✓

Herramientas NLP

Funcionalidad	OpenNLP	Stanford NLP	Freeling	TreeTagger
Tokenización	✗	✗	✓	✗
Retokenización	✗	✗	✗	✗
Detección de frases	✗	✗	✓	✗
Etiquetado PoS	✗	✗	✓	✓
Chunking	✗	✗	✓	✗
Lematización	✗	✗	✓	✓
Análisis sintáctico	✗	✗	✓	✗
NER	✓	✗	✓	✗
Correferencia	✗	✗	✓	✗
Lenguaje	Java	Java	C++	Perl
Wrapper para Java	—	—	✓	✓

Expansión de la pregunta

- Variantes morfológicas
  - Inventor → Inventó
- Variantes léxicas
  - Matón → Asesino
  - Lejos → distancia
  - Sobrepeso → Gordo
- Variantes semánticas
  - Preferir ↔ gustar

Tipo de respuesta

- Se obtiene a partir del tipo de pregunta
  - Taxonomía de tipos de respuesta
  - Sirve de guía al sistema IR saber qué debe buscar en los textos seleccionado
- No es tan sencillo como parece
  - *"Who-questions"* pueden tener una organización como respuesta
  - *"Which-questions"* puede tener una persona como respuesta

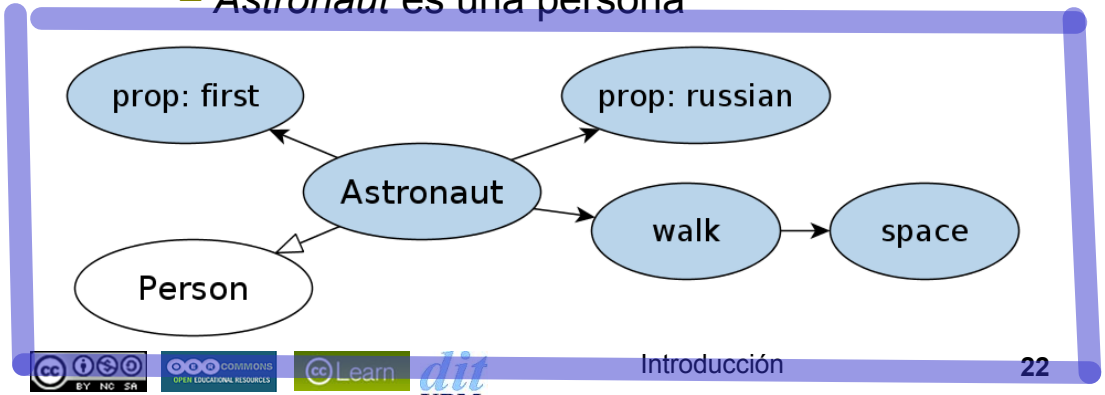
Tipo de respuesta

`Who be <target>`

- "Who is the king of Spain?"
- "Who is the most famous football player?"
- "Who was the first Russian astronaut to walk in space?"
- "Who is your favourite singer?"
- "Who fought against Rome in the Punic Wars?"

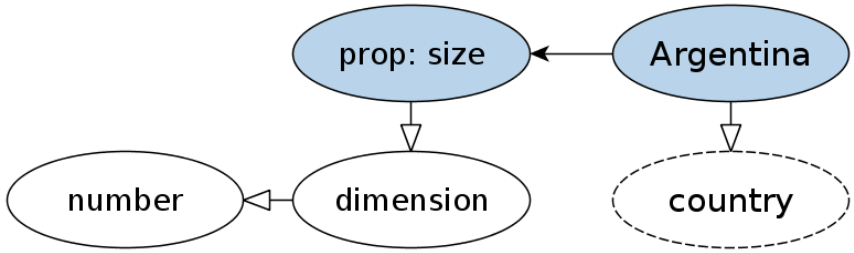
Tipo de respuesta

- *"Who was the first russian Astronaut to walk in space?"*
  - El target es *Astronaut*
  - *Astronaut* es una persona

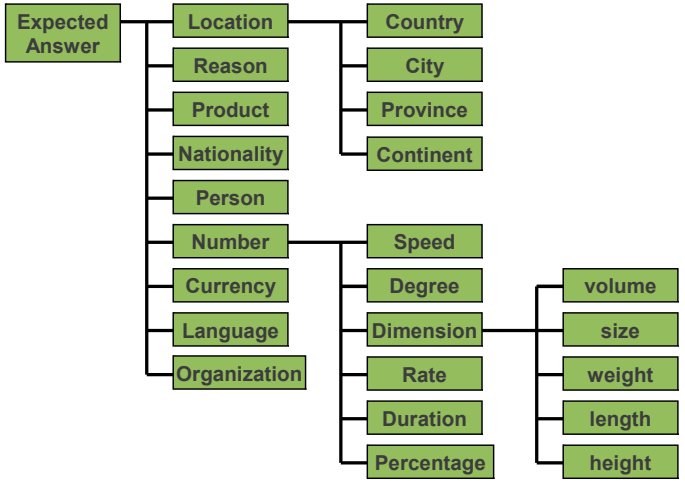


Tipo de respuesta

- *"What size is Argentina?"*
  - El target es *Argentina*
  - Pregunta por la propiedad **size** del target



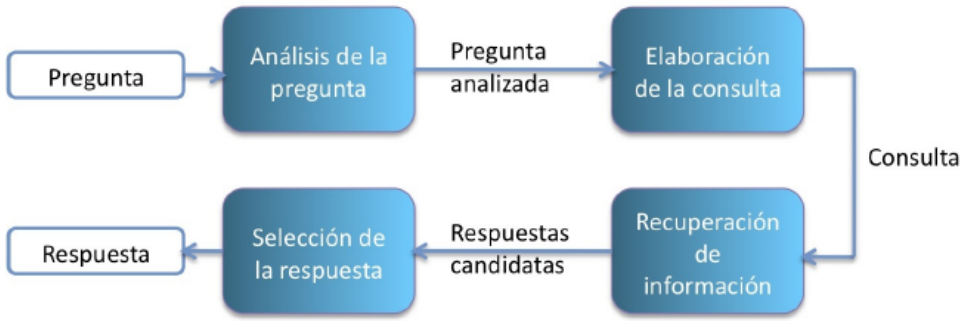
Taxonomía de respuestas



Taxonomía de respuestas

- La mayoría de los programas QA trabajan sobre una taxonomía inmensa (escrita a mano)
  - Estrategias para escalar el sistema
    - Aumentar el número de patrones de pregunta
    - ... a la vez que los tipos de respuestas
  - Si no está en la taxonomía de respuestas el programa puede desestimar las respuestas correctas
- Reglas de aprendizaje para completar dicha taxonomía

Etapas de un sistema QA



Selección de documentos

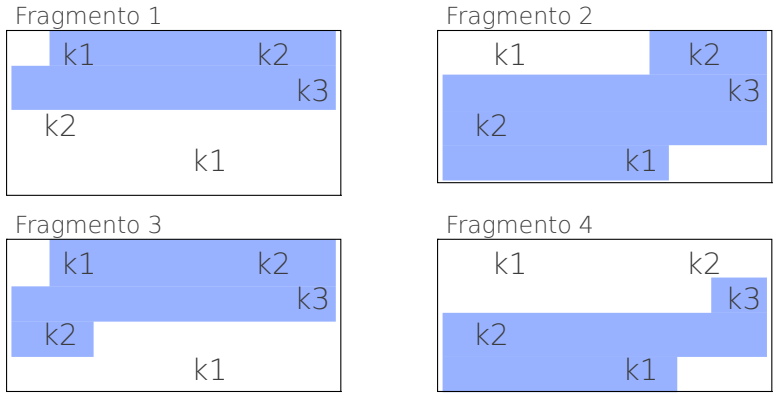
- Consiste en seleccionar aquellos documentos que son relevantes a la consulta realizada
  - Utilizan palabras clave
    - n-grammas
  - También el tipo de respuesta esperado
  - Muchas veces no son más que un filtro de aquellos documentos que **no** son relevantes
- Arroja un lista de documentos relevantes

Extracción de fragmentos

- A partir de la lista de documentos relevantes
  - Selecciona fragmentos en los que aparecen la palabras clave
  - Tamaño e inicio del fragmento dinámico
  - De un documento se pueden sacar multitud de fragmentos

Extracción de fragmentos

- k1, k2, k3, k4 son palabras clave
- El tamaño de la ventana es dinámico (pero con límites)



Extracción de la respuesta

- En función del tipo de respuesta esperada
  - Estos pasajes pueden ser la propia respuesta y deben tener completitud sintáctica
  - O deben contener la respuesta concreta
    - Para generar a partir de ella la respuesta en lenguaje natural

Patrones de respuesta

- Para generar la respuesta en lenguaje natural

```
(a|an|the) <target> is (a|an|the)
    <property>

<property> is the <target>

On <property>, <target>

The <property> is the official <target>

<target> <property>:
```


Extracción de la respuesta

- Criterios para clasificar respuestas
  - Similitud
  - Popularidad
  - Relación de patrones
  - Validación de la respuesta

Extracción de la respuesta

- Similitud
  - Where is the Louvre Museum located?
  - The Louvre Museum is located in Paris

```
Q: Where be <target>
A: g(<target>) <answer_type>
```


Extracción de la respuesta

Name the first private citizen to fly in space.

- Answer type: **Person**
- Text passage

  "... among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in "Raiders of the Lost Ark", plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith..."

Extracción de la respuesta

Name the first private citizen to fly in space.

- Answer type: **Person**
- Text passage

  "... among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in "Raiders of the Lost Ark", plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith..."

- Best candidate: **Christa McAuliffe**

Validación de la respuesta

- Reglas para garantizar que el dato dado como respuesta es:
  - Coherencia
  - Magnitud de unidades
  - Sistemas deductivos más complejos
    - Fechas
    - Lógica

Validación de la respuesta

- Validación por parte del usuario
  - Constituye una realimentación del usuario hacia el sistema
  - Porcentaje de acierto de patrones
  - Probabilidades

Conclusión

- Investigar en QA es un desafío
- Engloba varias técnicas
  - IR
  - NLP
  - AI
  - Manejo de grandes conjuntos de datos
  - Machine Learning