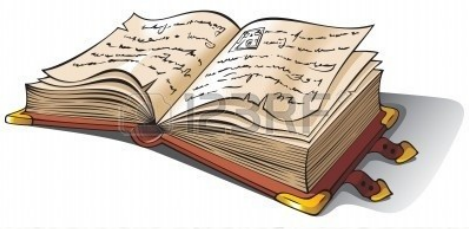
Extracción de información y clasificación de documentos médicos



Lourdes Araujo (lurdes@lsi.uned.es)
Grupo de investigación en Procesamiento del Lenguaje Natural y Recuperación de Información
http://nlp.uned.es/web-nlp/

Motivación

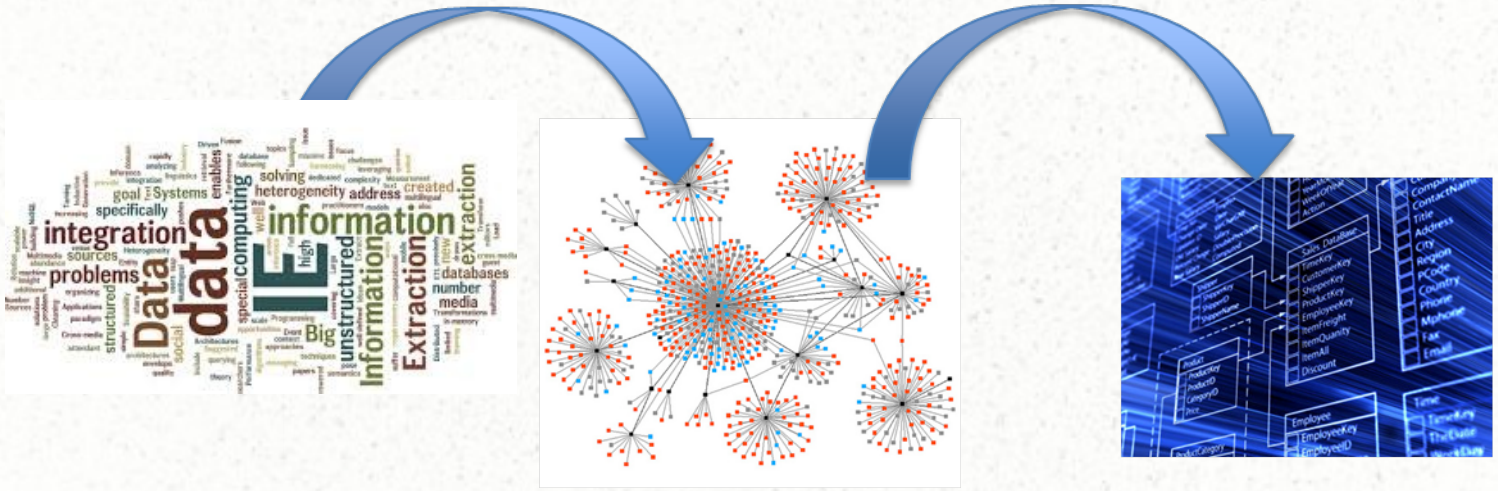
Comunicación humana: LENGUAJE NATURAL

Motivación

Suponemos que el lenguaje se recoge en textos que podemos procesar:

Motivación

- Necesitamos información estructurada:



Información no estructurada

Procesamiento

Información estructurada

Motivación

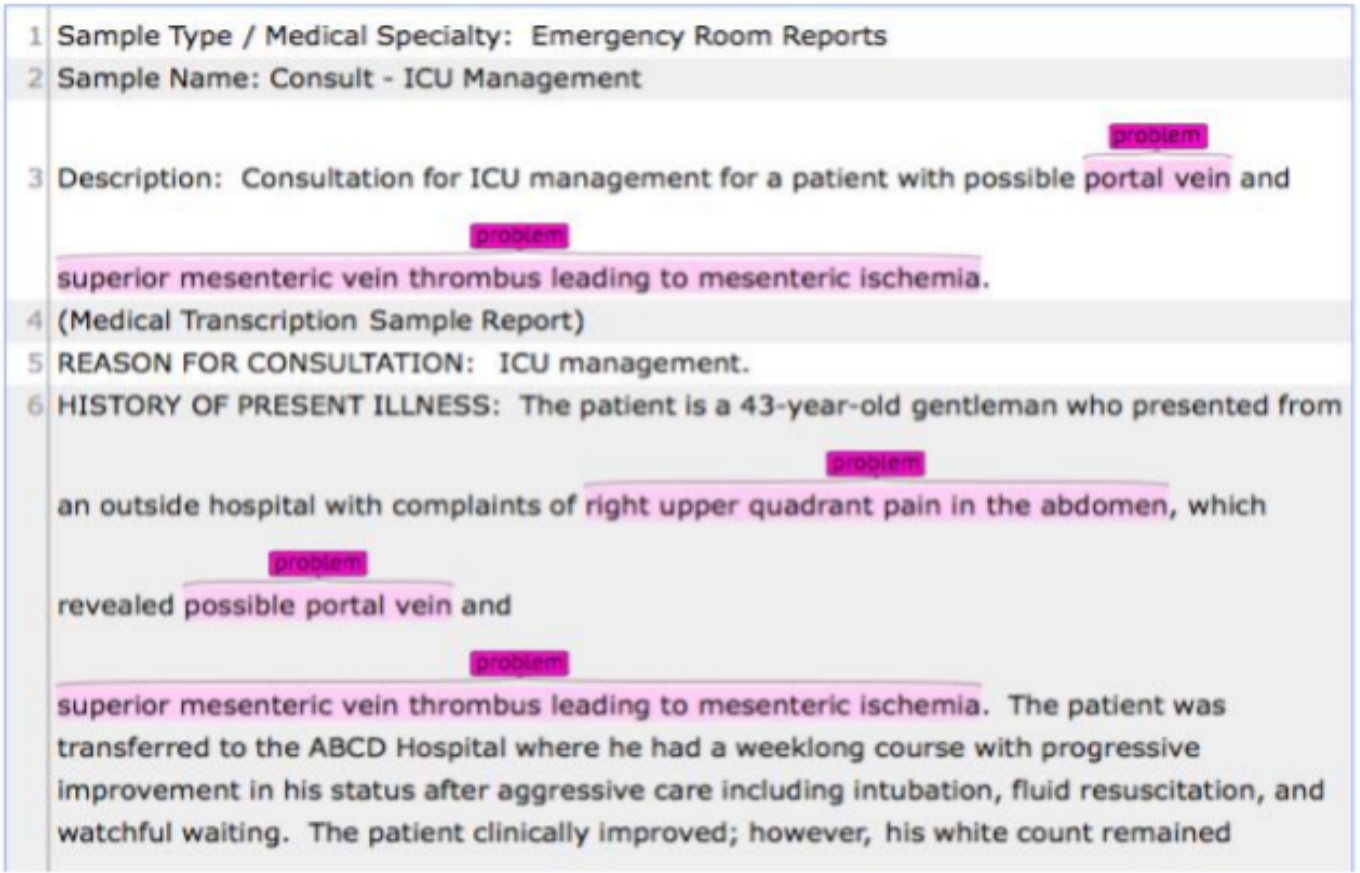
- Extracción de información:

Paciente varón de 86 años con antecedentes de TBC pulmonar, refiere F no cuantificada desde hace dos semanas, así como tos esporádica productiva y perdida de peso.

Documento: *Informe de urgencias*
- Sexo: *hombre*
- Edad: *86*
- Antecedentes:
  - *Tuberculosis pulmonar*
- Síntomas:
  - *Fiebre no cuantificada*
  - *Tos esporádica productiva*
  - *Perdida de peso*
- Tiempo: *dos semanas*

Motivación

Anotación de conceptos



Motivación

- Anotación/clasificación

Manual

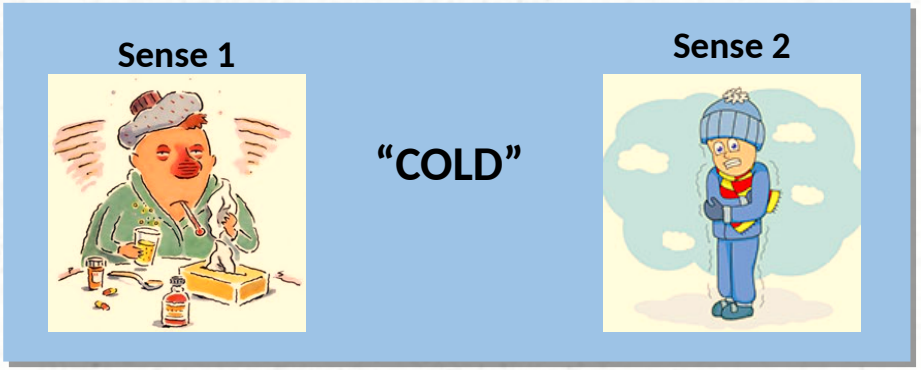
Automática

Motivación

- Facilitar el uso de los ordenadores con una forma de acceso más sencilla y natural
- Sacar partido de información contenida de forma NO explícita en la enorme cantidad de documentos médicos electrónicos:
  - Hacer predicciones
  - Entender mejor el funcionamiento del problema
  - Extraer nuevo conocimiento

Dificultades

- Ambigüedad léxica



- Acrónimos y abreviaturas (corresponden a muchas posibles formas extendidas):
  - AAC: Ácido aminocapróico, Actividad anticomplementaria, Alopecia areata circunscrita, Angiopatía amiloidea cerebral, Anticuerpos anticardiolipídicos,etc.
- Erratas en la escritura

Dificultades

- Negación y su ámbito:

  "*the syndrome appears to be related to diminished speech and language capacity, rather than the specific social deficits central to autism.*"

- Especulación:

  "*Recessive mutations of the SLC26A4 (PDS) gene on chromosome 7q31 can cause sensorineural hearing loss with goiter (Pendred syndrome).*"

- Conceptos expresados en lenguaje libre

Aplicaciones

Que pueden hacer las técnicas de procesamiento del lenguaje natural?

- Identificar conceptos: Enfermedades, Medicamentos, Síntomas, Procedimientos, etc.
- Extraer relaciones entre conceptos: Cura, produce, coaparece, efectos adversos a medicamentos, discapacidades asociadas a enfermedades raras, etc.
- Detección de negación y de relaciones negadas.

Aplicaciones

- Inducción de conocimiento nuevo: Descubrir relaciones que no aparecen explícitamente en los documentos
  - No están recogidas
  - No se conocen
- Identificación de reglas de asociación entre conjuntos de enfermedades

Aplicaciones

- Anonimización de historia clínica
- Identificación y desambiguación de acrónimos en historia clínica.
- Recomendación de códigos CIE-10 en historia clínica (informes de alta hospitalaria, partes de defunción, sospechas diagnósticas)

Aplicaciones

- Acceso a la información en foros y redes sociales de salud:
  - Minería de textos y opiniones: análisis de sentimientos, etc.
  - Generación automática de resúmenes: opiniones positivas de un tratamiento, etc.
  - Monitorización y fiabilidad de la información.

Tipos de documentos considerados

- Informes médicos:
  - Lenguaje específico
  - Uso masivo de siglas y abreviaturas especificas.
  - Erratas de escritura frecuentes
- Artículos científicos:
  - Lenguaje más formal
- Redes sociales:
  - Lenguaje poco cuidado, textos cortos, etc.
- Distintos idiomas: español, inglés, francés, ...

Técnicas

- Análisis de textos: segmentación de palabras, normalización de textos, análisis léxicos y sintácticos, etc.
- Aprendizaje automático y redes neuronales profundas (keras, LSTM, Convolución)
- Técnicas no supervisadas: grafos, métodos estadísticos

Organización de campañas de evaluación

- Ibereval: Evaluation of Human Language Technologies for Iberian Languages
  - DIANN: Disability annotation on documents from the biomedical domain
- IberLEF: Iberian Languages Evaluation Forum

Contacto:

Lourdes Araujo
lurdes@lsi.uned.es
Tf.: 913987318
Dpto. lenguajes y Sistemas Informáticos. ETSI. Informática UNED

Gracias !