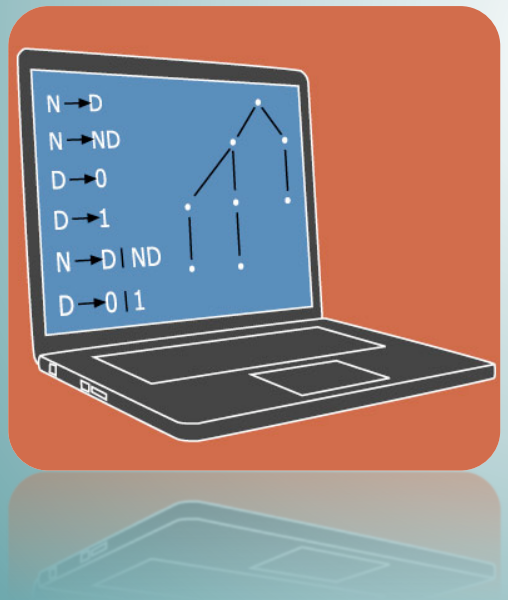
Procesadores de Lenguaje

Procesamiento de Lenguaje Natural



Cristina Tirnauca

DPTO. DE MATEMÁTICAS,

ESTADÍSTICA Y COMPUTACIÓN

¿Por qué es una tarea dificil?

Versión en inglés

- The thieves stole the paintings.
  They were subsequently sold.
- The thieves stole the paintings.
  They were subsequently caught.
- The thieves stole the paintings.
  They were subsequently found.

Traducción a español

- Los ladrones robaron las pinturas.
  Se vendieron posteriormente.
- Los ladrones robaron las pinturas.
  Ellos fueron capturados posteriormente.
- Los ladrones robaron las pinturas.
  Fueron encontrados posteriormente.

Un poco de historia

- El PLN, parte de la IA (1956!!!)
  traducción automática:
  - siglo XVII - diccionarios mecánicos, lengua universal
  - Weaver, 1949 - "Translation" memorándum
  - visión simplista: diferencias entre las lenguas residen en su vocabulario y en el orden de las palabras.
  - Chomsky, 1957, gramáticas generativas
  - el informe de ALPAC (Automatic Language Processing Advisory Committee of the National Academy of Science - National Research Council) en 1966
- análisis sintáctico de frases en el lenguaje natural: la profundidad de la ambigüedad en los lenguajes naturales
  Time flies like an arrow vs. Fruit flies like a banana.
- prototipos de sistemas: ELIZA (1964-1966), SHRDLU (1968-1970), PARRY (1972)
- enfoque estadístico - éxito en muchos problemas genéricos de la lingüística computacional (desambiguación del significado de la palabra, part-of-speech tagging,...)

Aplicaciones en el procesamiento de voz

Tareas de trabajo:

- Reconocimiento de voz = Reconocimiento del habla = Comprensión del lenguaje (Speech Recognition)
- Síntesis de voz (Speech Synthesis)
- Reconocimiento del hablante
- Mejora de la señal de voz
- Codificación de voz para compresión de datos y transmisión de la voz.
- Análisis de voz con propósitos médicos

Aplicaciones en el procesamiento de texto

Tareas de trabajo:

- Traducción automática (Machine Translation)
- Resumen automático (Automatic Summarization)
- Generación de texto (Natural Language Generation)
- Comprensión de texto (Natural Language Understanding)
- Respuesta a preguntas = búsqueda de respuestas (Question Answering)
- Recuperación de la información (Information Retrieval)
- Extracción de la información (Information Extraction)

Componentes

- Análisis fonológico: fonemas
  Su función se basa en la relación de las palabras con el sonido asociado a su pronunciación
- Análisis morfológico: morfemas
  Su función consiste en detectar la relación que se establece entre las unidades mínimas que forman una palabra, como puede ser el reconocimiento de sufijos o prefijos. Este nivel de análisis mantiene una estrecha relación con el léxico. Normalmente el léxico sólo contiene la raíz de las palabras con formas regulares, siendo el analizador morfológico el que se encarga de determinar si el género, número o flexión que componen el resto de la palabra son adecuados.
- Análisis léxico (POS: part-of-speech tagging): nombre, adjetivo, artículo, pronombre, verbo...
  El léxico es el conjunto de información sobre cada palabra que el sistema utiliza para el procesamiento. Las palabras que forman parte del diccionario están representadas por una entrada léxica, y en caso de que ésta tenga más de un significado o diferentes categorías gramaticales, tendrá asignada diferentes entradas.
  En el léxico se incluye la información morfológica, la categoría gramatical, irregularidades sintácticas y representación del significado. Asimismo, a este nivel podemos reemplazar aquellas palabras que sólo tienen un significado con su representación semantica.
- Análisis sintáctico (gramática, analizador): sujeto, predicato, complemento...
  Tiene como función etiquetar cada uno de los componentes sintácticos que aparecen en la oración y analizar cómo las palabras se combinan para formar construcciones gramaticalmente correctas. El resultado de este proceso consiste en generar la estructura correspondiente a las categorías sintácticas formadas por cada una de las palabras que aparecen en la oración.
- Análisis semántico
  Hay que distinguir entre significado independiente o dependiente del contexto. El significado independiente del contexto (tratado por la semántica), hace referencia al significado que las palabras tienen por sí mismas, ignorando la influencia del contexto o las intenciones del hablante. El significado dependiente del contexto (estudiado por la pragmática) se refiere al significado de las palabras asociado a las circunstancias.
- Análisis del discurso: anáforas, catáforas

Dificultades en el PLN

- Ambigüedad (a nivel léxico, referencial: anáforas y catáforas, estructural: de agrupamiento o funcional, pragmático,...)
  Han puesto un banco nuevo en la plaza.
  Los ladrones robaron los cuadros. Los encontraron posteriormente.
  A esto me refiero: a que te has portado mal.
  María guardó las revistas que Paco dejó bajo la cama.
  Compraré solo este regalo.
  Salió de la cárcel con tanta honra, que le acompañaron doscientos cardenales sino que a ninguno llamaban eminencia.
- Desambiguación: word-category disambiguation (POS), word-sense disambiguation (WSD)
- Detección de separación entre las palabras o frases (text segmentation: word segmentation, sentence segmentation)

  En la lengua hablada no se suelen hacer pausas entre palabra y palabra. El lugar en el que se debe separar las palabras a menudo depende de cuál es la posibilidad que mantenga un sentido lógico tanto gramatical como contextual. En la lengua escrita, idiomas como el chino mandarín tampoco tienen separaciones entre las palabras.
- Recepción imperfecta de datos

  Acentos extranjeros, regionalismos o dificultades en la producción del habla, errores de mecanografiado o

Gramáticas para el NLP, una retrospectiva

- El enfoque estructuralista (Ferdinand de Saussure, 1920-1950) en Europa y los constituentes (Leonard Bloomfield) en EEUU. El lenguaje natural = estructura de elementos mutualmente vinculados entre sí
- Chomsky: generative grammars (CFGs,...)
- Transformational grammars (Chomsky)
- Valencias
- Constraints (limitaciones): generalized phrase structure grammars (GPSG)
- Head-driven phrase structure grammars (HDPSG)
- Unification
- Meaning-text theory (MTT) - dependency grammars

El NLP en la actualidad

- "Trade-off" entre gramáticas con propriedades bonitas y gramáticas que son facil de analizar desde el punto de vista computacional
- Los analizadores modernos son, al menos parcialmente, estadísticos: se basan en un corpus de datos de entrenamiento previamente anotados
  Colorless green ideas sleep furiously. / Furiously sleep ideas green colorless.
- La mayoría de los algoritmos de análisis estadísticos se basan en una forma (modificada) de chart parsing
  - Ejemplos: Earley parser, Cocke-Younger-Kasami (CYK) parser
  - adecuados para gramáticas ambiguas
  - programación dinámica
- Herramientas para la traducción automática: tree transducers, synchronous grammars, tree bimorphisms