



Procesamiento del Lenguaje Natural

Dr. John Atkinson



Introducción

Procesamiento de Lenguaje Natural

Objetivos del Curso

- ✓ *Entender los fundamentos y aplicaciones del Procesamiento del Lenguaje Natural (NLP).*
- ✓ *Introducir los principales problemas del NLP, soluciones y problemas.*
- ✓ *Comprender los métodos utilizados para desarrollar algunas aplicaciones sencillas de NLP.*

Bibliografía

Material Clave:

- ✓ D. Jurafsky, J. Martin. “*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*”, 2nd edition, Prentice Hall, 2008.
- ✓ S. Bird, E. Klein, E Loper. “*Natural Language Processing with Python*”, O’Reilly, 2009.

El lenguaje ...

Por si le interesa a alguien: Tengo una amiga que hace dos semanas se compró un iPhone 8 (le costó \$749,000 mil pesos) y por un descuido su hijo le rompió la pantalla, entonces después de ver lo que cuesta arreglarlo, lo está regalando. Si alguien lo quiere avísenme por favor. El pequeño tiene 8 años, es gordito, está en 3ero básico, come lo que sea, si lo sientas con video juegos no da mucho trabajo... si saben de alguien que lo quiera, me avisan...

El lenguaje ...

¿Qué mecanismos utilizamos para comprender el lenguaje escrito?

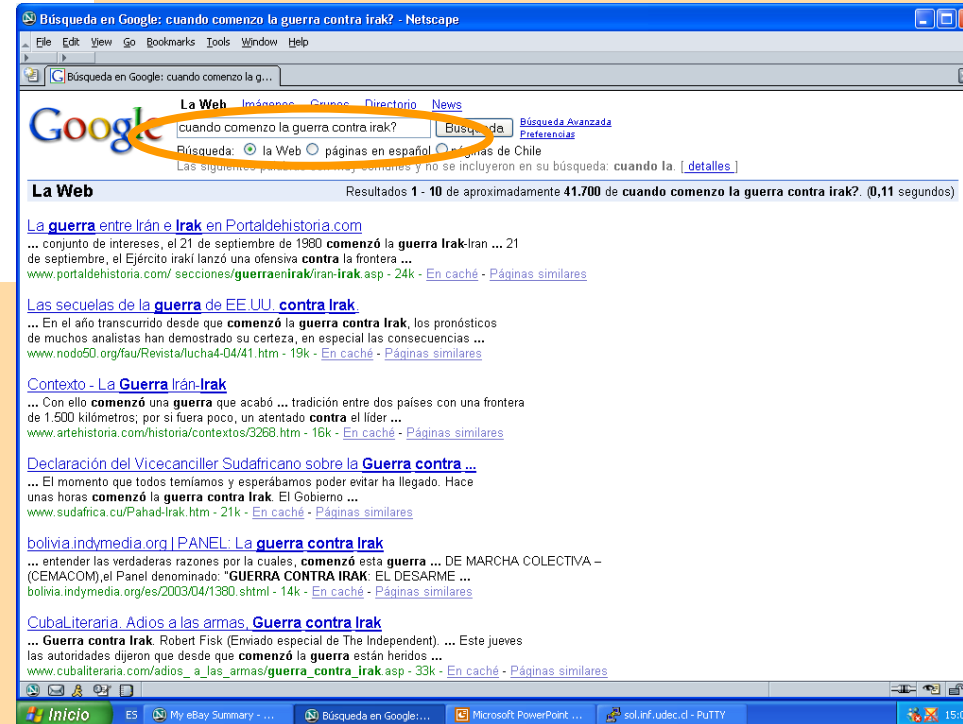
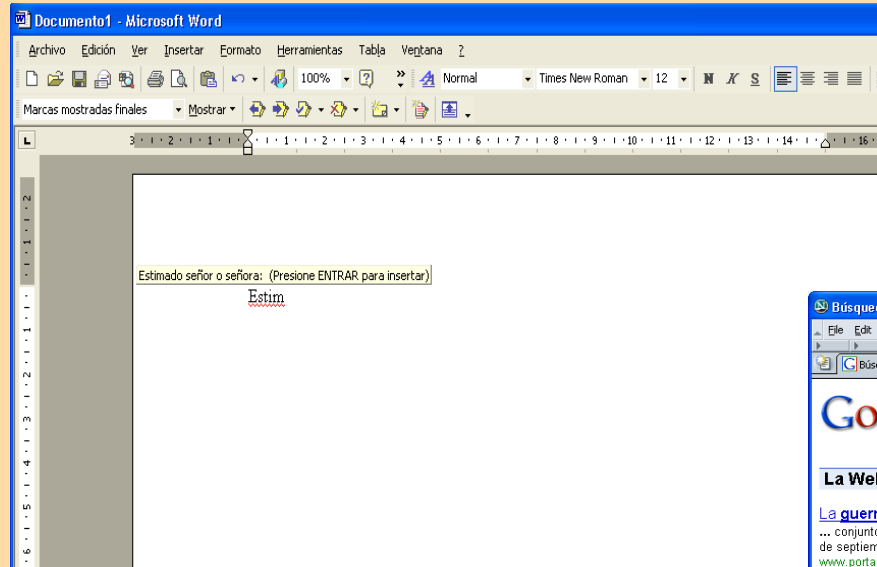
“El ministro del Interior reconoció que una de esas personas es el economista DC y ex ministro del Trabajo René Cortázar, aunque recalcó que es la Presidenta Michelle Bachelet quien deberá tomar una decisión final. Lo mismo ocurrirá respecto a un eventual cambio de gabinete.

Asimismo, Belisario Velasco explicó que parte de las fallas en los recorridos de los buses se debe a la falta del software que aún no es entregado por el proveedor”.

Lucas y el Lenguaje ...



Muchos beneficios...



Muchos beneficios...

Verdad y Justicia @Verjusticia · 13 may. 2017
En respuesta a @santanderchile
No es así. Los cajeros de la sucursal Alto Palermo y la de Sanchez de Bustamante no toman depósitos. Los fines son un desastre.

Banco Santander Chile @santanderchile · 14 may. 2017
Hola, envíanos un mensaje interno con tu nombre completo, rut, mail y teléfono para ayudarte ¡Saludos!

Verdad y Justicia @Verjusticia · 14 may. 2017
No es necesario mis datos, lo importante es que mejoren el servicio!!.



Procesamiento del Lenguaje Natural

Objetivo: lograr que los computadores utilicen *lenguaje natural* como entrada y/o salida para comprenderlo ó producirlo.



Específicamente...

- Hacer que los computadores realicen tareas útiles a partir del *lenguaje humano* para:
 - *Posibilitar* la comunicación *humano-computador*
 - *Mejorar* la comunicación humano-computador
 - *Hacer cosas* con los objetos del lenguaje
- Aplicaciones:
 - Categorización de documentos
 - Traducción automática
 - Reconocimiento del habla (ASR)
 - Agentes conversacionales
 - Sistemas de Pregunta-Respuesta (QA)
 - Analítica textual
 - Análisis de sentimientos en redes sociales
 - Muchos más...

Origen:

El Lenguaje en la Comunicación



Intención



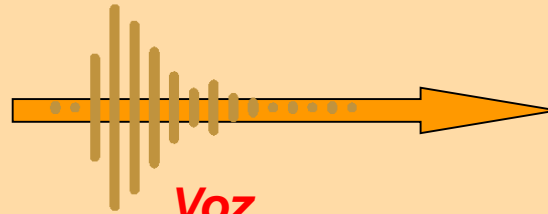
Forma Lógica

Generación



Lenguaje

Síntesis



Voz

Canal con ruido

Incorporación



Forma Lógica

Desambiguación



Lenguaje

Análisis



Palabras

Percepción



¿Porqué NLP es difícil?

- ✓ Se deben abordar muchos problemas:
 - Ej: Ambigüedad (ej. acústica, morfológica, sintáctica, semántica, discursiva, etc)
- ✓ La tarea parece difícil, entonces ¿Qué recursos necesitamos?
 - Conocimiento acerca del *lenguaje*
 - Conocimiento acerca del *dominio (el “mundo”)*
- ✓ Dos tendencias actuales para combatir el cuello de botella de la *adquisición de conocimiento*:
 - Dirigido por *expertos* y codificados manualmente.
 - Dirigido por los *datos* y automatizado.

Algunos Problemas ...

Ambigüedad es un problema que se produce cuando existen *múltiples* caminos posibles para resolver un problema en un gran espacio de búsqueda:

- *Ambigüedad léxica*: múltiples usos de *palabras*.
- *Ambigüedad sintáctica*: múltiples estructuras *sintácticas* para una misma oración
- *Ambigüedad semántica*: múltiples sentidos para una misma estructura sintáctica de una *oración*.
- *Ambigüedad de Referencia (Discurso)*: múltiples formas de “*referenciar*” (invocar) a una entidad.

Ambigüedad *Léxica*

Banco



Ambigüedad *Sintáctica*

Trae el gato con los guantes



Ambigüedad Semántica

Pedro regaló un pastel a los niños

*¿Uno para todos?
¿Uno a cada uno?
A lo mejor depende del
tamaño (del pastel, no
de Pedro ni de los
niños)*



Referencia (en el Discurso)

Ella le dijo que lo ordenara.

¿Quién dijo?
¿A quién?
¿Cuándo?
¿Qué ordenara qué?



Drivers de NLP

- ✓ **La necesidad de interfaces inteligentes:**
 - Aplicaciones de voz, telefonía, etc.
 - Nadie desea un teclado en su micro-ondas inteligente!!
 - No hay suficientes humanos para responder a usuarios/clientes.
- ✓ **El problema de la sobrecarga de información:**
 - Existe demasiada información en medios electrónicos.
 - Nadie tiene tiempo para leer los documentos, emails, etc.
 - Necesitamos tomar decisiones desde los documentos

¿Qué tipo de Conocimiento Requeriría una Máquina?



¿Qué requiere HAL?

✓ *Reconocimiento y Síntesis de Voz*

- Diccionarios (*¿Cómo se pronuncian las palabras?*)
- Fonética (*¿Cómo reconocer/producir cada sonido del lenguaje?*)



• *Comprensión del Lenguaje*

- Conocimiento de las palabras del lenguaje
 - ¿Qué significan?
 - ¿Cómo se combinan? (*¿Qué es un "compartimiento"?*)
- Conocimiento de la *relaciones entre palabras*

¿Qué requiere HAL?

✓ *Diálogo y Pragmática*

- Necesita entender que “*abre la puerta*” es un **REQUERIMIENTO** (opuesto a una pregunta por información).
- Es bueno responder, incluso si se planifica matar a alguien ☺.
- Es bueno pretender desear ser cooperativo (“*me temo, no puedo...*”)
- ¿Qué es ‘eso’ en ‘*No puedo hacer eso*’?



¿Como se ve esto actualmente?

- ✓ Sistema de reservas para pasajes aéreos.
- ✓ Atención automatizada de clientes.
- ✓ Comprensión de quejas de clientes y extracción de información clave.
- ✓ Entender las emociones de los clientes en redes sociales.
- ✓ Muchas más..

Interacciones con limitaciones..

Chatbot: Bienvenido a "*PepitoPagaDoble Comunicaciones*",
¿En qué lo puedo ayudar?

Cliente: Tengo un problema con **BLA BLA Internet BLA BLA**

Chatbot: Veo que su problema es **Internet**, ¿Necesita
(a)..(b)... ó (c)...?

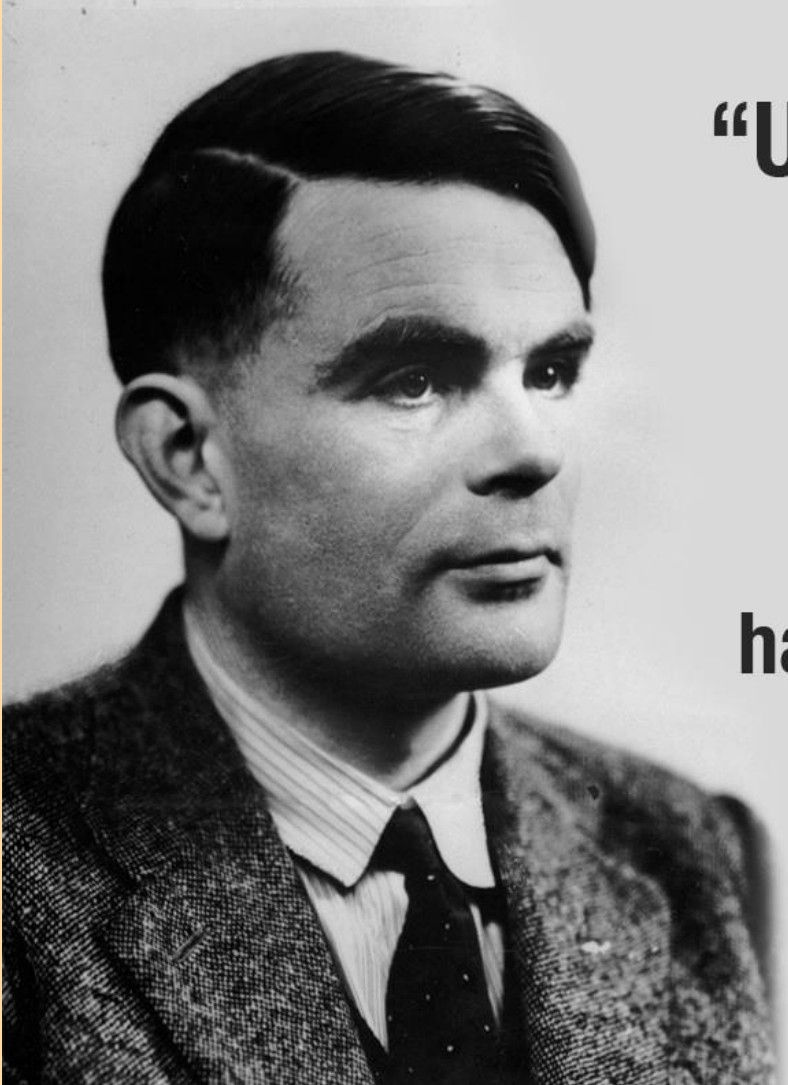
Cliente: no, es lo primero que le escribí, mire le
explico....

Chatbot: ok. ¿Qué necesita específicamente?

Cliente: sus canales se ven pixelados y no se porqué.
¿Qué hora es?.

Chatbot: ...

Recordemos...todo esto no es nuevo...

A black and white portrait of Alan Turing, a man with dark hair, wearing a suit and tie, looking slightly to the right.

**“Una computadora
puede ser llamada
inteligente
si logra engañar
a una persona
haciéndole creer que es un
humano”**

Alan Mathison Turing
1912–1954

zomwi

Historia: 1940s-1950s

✓ Autómatas:

- *Turing* (1936)
- McCulloch-Pitts neuron (1943)
- Kleene (1951/1956)
- Shannon (1948) conecta autómatas y modelos de Markov
- Chomsky (1956)/Backus (1959)/Naur(1960): gramáticas formales para lenguajes (CFG)

✓ Modelos probabilísticos y teoría de información:

- Shannon (1948)
- Reconocimiento de voz en AT&T Bell Labs, USA (1952)

Historia: 1957-1970

✓ Enfoque Simbólico

- Primer analizador sintáctico (*Parser*) de Zellig Harris (1958).
 - *Transductores* de estados-finitos.
- Teorías de adquisición de lenguaje (Chomsky)
- Workshop de AI en Dartmouth College (McCarthy, Minsky, Shannon, Rochester, etc)
- Demostración de teoremas (Logic Theorist, General Problem Solver) de Newell y Simon.

✓ Enfoque Estadístico

- Reconocimiento de Caracteres Ópticos (OCR) basados en inferencia Bayesiana (1959).
- Combinación de Gramáticas y probabilidad acústicas para Reconocimiento de Voz (Denes, 1959):

Cuatro paradigmas: 1970-1983

✓ Estocástico

- *Modelos Ocultos de Markov ó HMM* (1972).

✓ Basado en Lógica

- *Q-Systems* de Colmerauer (1970,1975).
- *Definite Clause Grammars* (Pereira and Warren 1980).
- *Gramáticas funcionales* (Kay, 1979), *Unificación* (Bresnan y Kaplan, 1982).

✓ Comprensión de Lenguaje Natural

- *SHRDLU* de Winograd (1972).
- *Scripts y Comprensión de historias* de Schank y Abelson (1977).
- *Representación de roles-casos* de Fillmore (1968) y Simmons (1973).

✓ Modelamiento del Discurso

- *Estructura y foco del discurso* de Grosz.
- *Modelo de “creencias” (BDI)* de Perrault y Allen (1980).

Empirismo y Reaparecimiento de Estados Finitos: 1983-1993

✓ Modelos de Estados Finitos

- Kaplan and Kay (1981): Fonología/Morfología
- Church (1980): Sintaxis

✓ *Retorno de Modelos Probabilísticos:*

- **Corpus** creados para tareas de análisis de lenguaje
- Primeras versiones estadísticas de aplicaciones de NLP (ej. parsing, tagging, traducción)
- Se incrementa foco en el rigor metodológico y muestreo.

Primeros pasos: ELIZA

- Creado por J. Weizenbaum (MIT, 1966).
- Primer sistema práctico de conversación humano-computador utilizando técnicas de '*pattern matching*' para simular un psico-terapeuta.
- **ELIZA** era capaz de mantener una conversación con un paciente.
- En algunos casos, **ELIZA** era casi capaz de 'engañar' al paciente haciéndole creer que no era una máquina.

ELIZA en acción..

Human: Well, my boyfriend made me come here.

ELIZA: Your boyfriend made you come here?

Human: He says I'm depressed much of the time.

ELIZA: I am sorry to hear you are depressed.

Human: It's true. I'm unhappy.

ELIZA: Do you think coming here will help you
not to be unhappy?

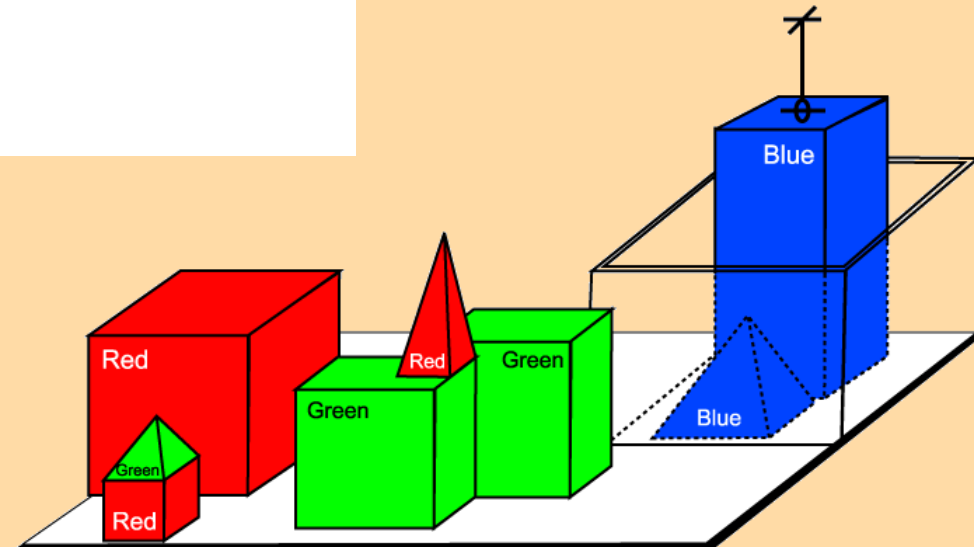
...

Primeros pasos: SHRDLU

- Creado por T. Winograd (Stanford, 1968).
- Sistema que permite la interacción entre un humano y ambiente simulado de interacción para un robot (mundo de los bloques).
- A diferencia de ELIZA, SHRDLU puede:
 - **Entender** lenguaje natural en las instrucciones dadas a un robot.
 - Resolver problemas de **ambigüedad**.
 - **Actuar** y **razonar** lógicamente.
 - **Responder** preguntas.

SHRDLU en acción..

- Person: PICK UP A BIG RED BLOCK.
- Computer: OK. (does it)
- Person: GRASP THE PYRAMID.
- Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
- Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
- Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
- Computer: OK. (does it)
- Person: WHAT DOES THE BOX CONTAIN?
- Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
- Person: WHAT IS THE PYRAMID SUPPORTED BY?
- Computer: THE BOX.



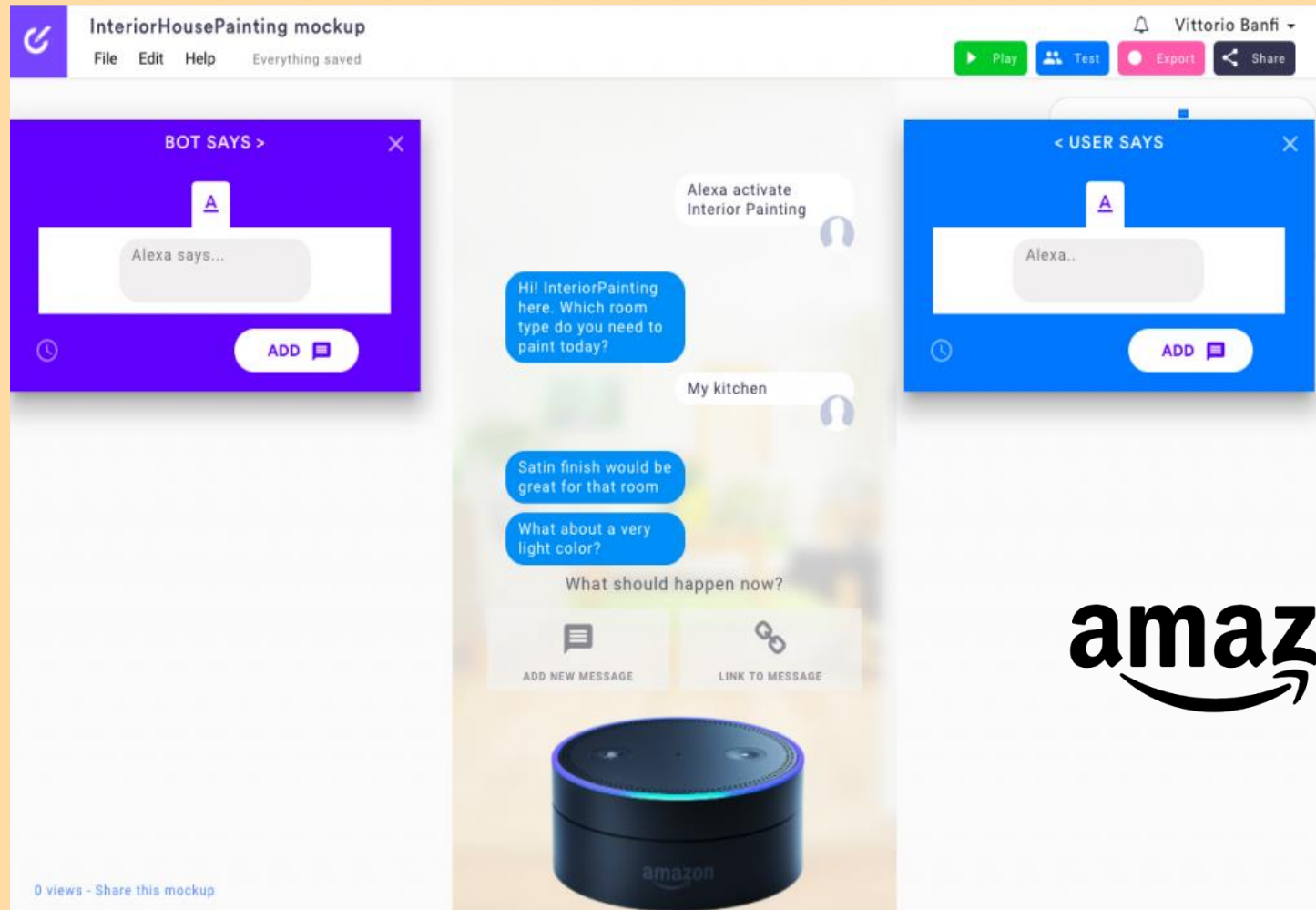
Algunas tecnologías modernas



Human: I really want to talk about chatbots and language learning.
ALICE: Sure, chatbots and language learning sounds like a good topic.
Human: Great.
ALICE: I'm glad you liked it judge.
Human: So what is the future of chatbots in language learning?
ALICE: You are talking to her.
Human: It looks like it's going to be a pretty surreal future.
ALICE: Interesting comparsion.

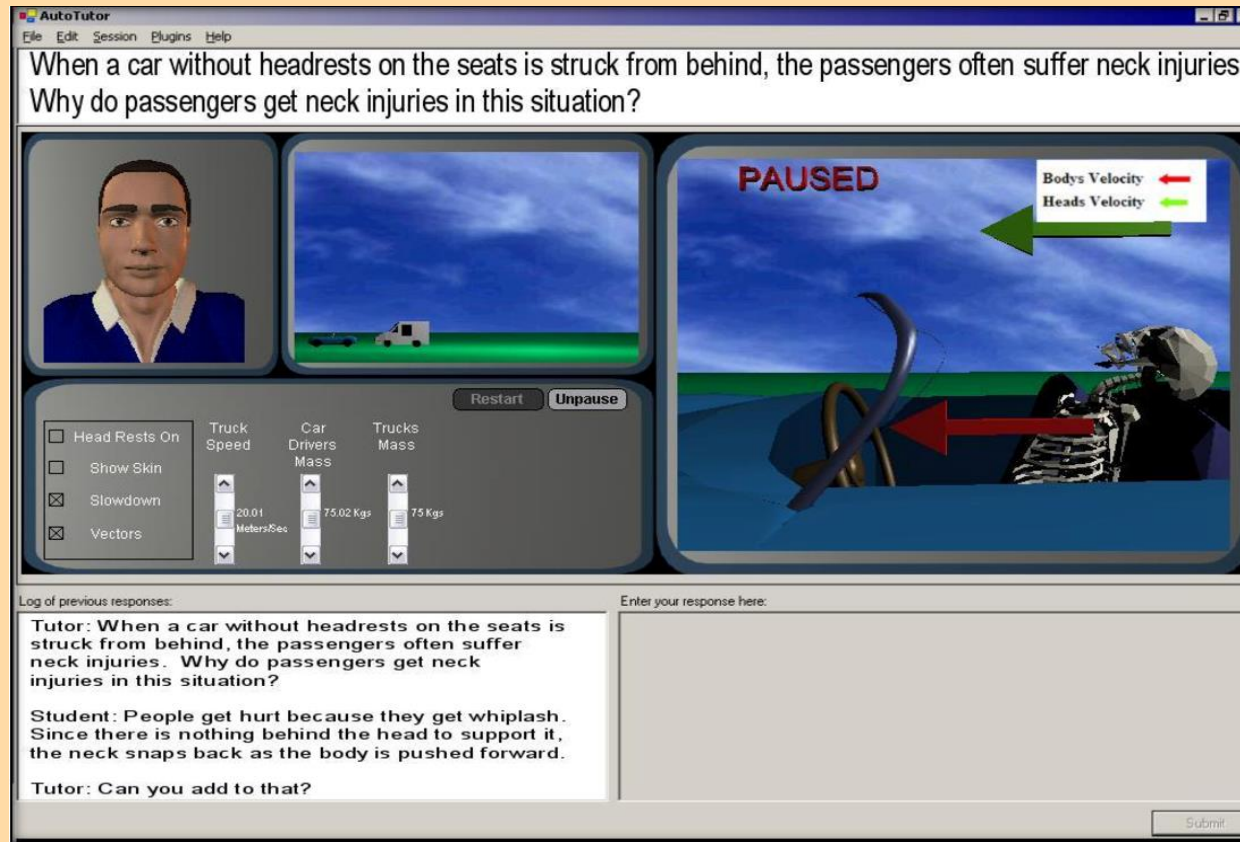
Artificial Linguistic Internet Computer Entity

Algunas tecnologías modernas



amazon alexa

Algunas tecnologías modernas



Autotutor

Uno de los sistemas tutoriales inteligentes más avanzados del mundo

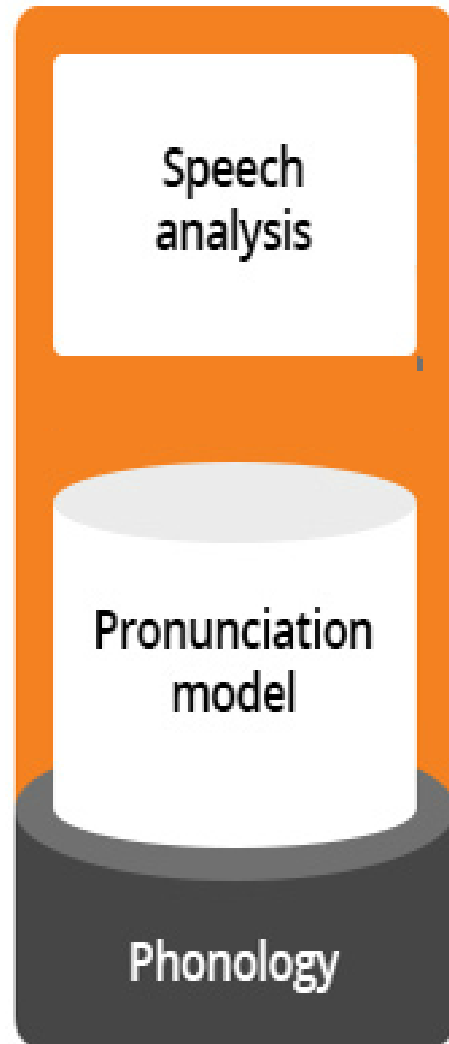
Pero...a veces pasan cosas indeseadas..



Modelos y Algoritmos

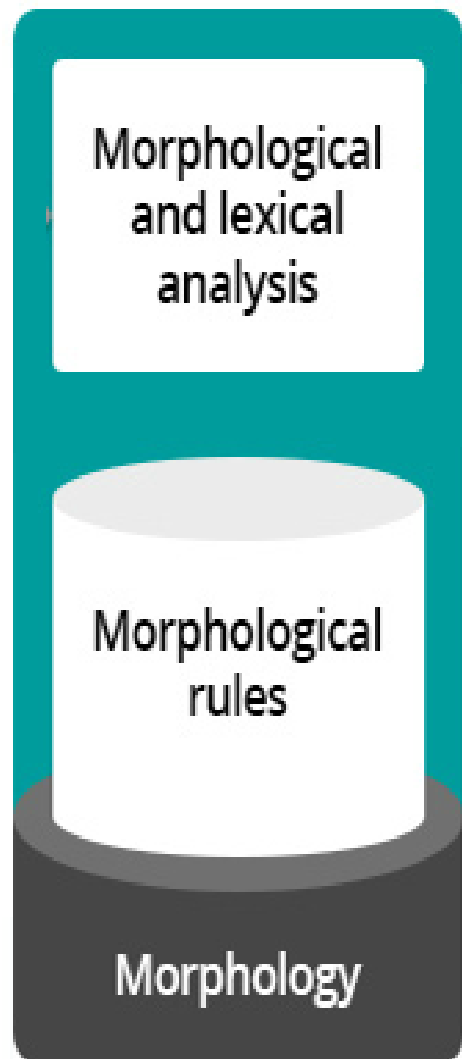
- ✓ **Modelos**: formalismos utilizados para capturar diferentes tipos de *estructura lingüística*.
 - *Máquinas de estados* (ej. Autómatas, modelos de *Markov*)
 - *Sistemas formales de reglas* (ej. Gramáticas, feature systems)
 - *Lógica* (ej. cálculo de predicados, inferencia)
 - Versiones *probabilísticas* de las anteriores más otros.
- ✓ **Algoritmos** utilizados para manejar representaciones que crean alguna *estructura*.
 - Búsqueda (ej. heurística, programación dinámica, etc)
 - Métodos estocásticos, aprendizaje automático, etc

Niveles en NLP



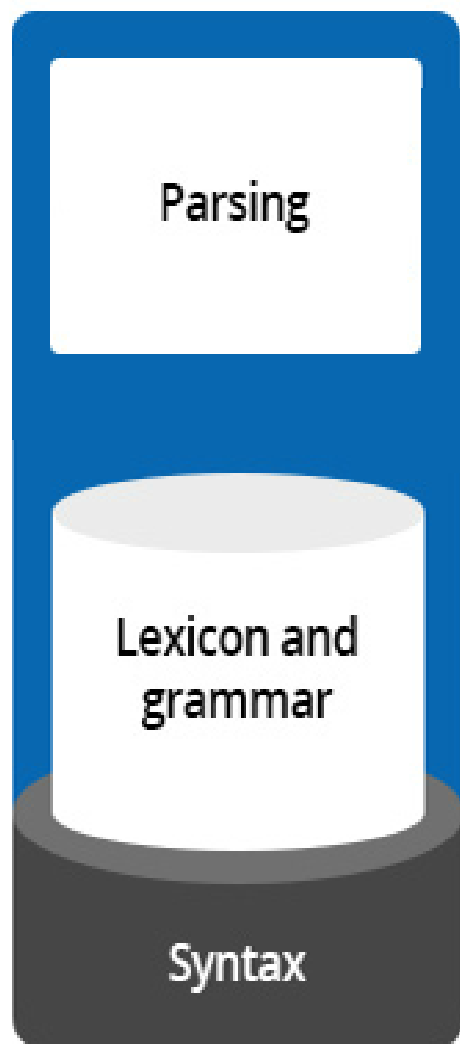
- ✓ Reconocimiento de fonemas en el audio y los límites de las palabras para identificar secuencias de palabras.
- ✓ Se debe diseñar un modelo del lenguaje para predecir en caso de “ambigüedad”

Niveles en NLP



- ✓ Determinar la forma de una palabra y si ésta es válida en un lenguaje.
- ✓ Problemas en ambigüedad léxica, por lo que convierte en un problema de búsqueda.

Niveles en NLP



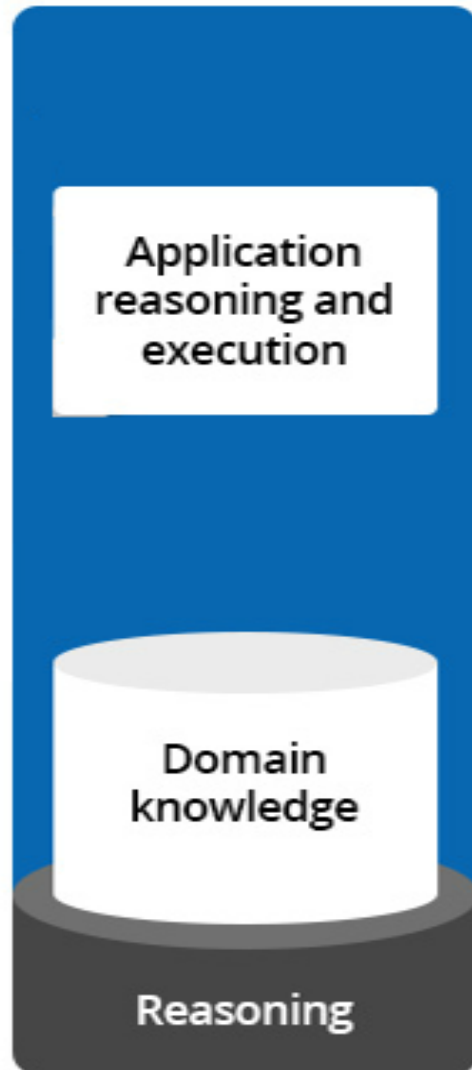
- ✓ Análisis (parsing) e identificación de la estructura sintáctica de el lenguaje de entrada (ambigüedad)
- ✓ Una gramática describe posibles estructuras sintácticas válidas.
- ✓ Para grandes volúmenes de datos, a veces las reglas gramaticales no son robustas: utilizar métodos de aprendizaje estadístico.

Niveles en NLP



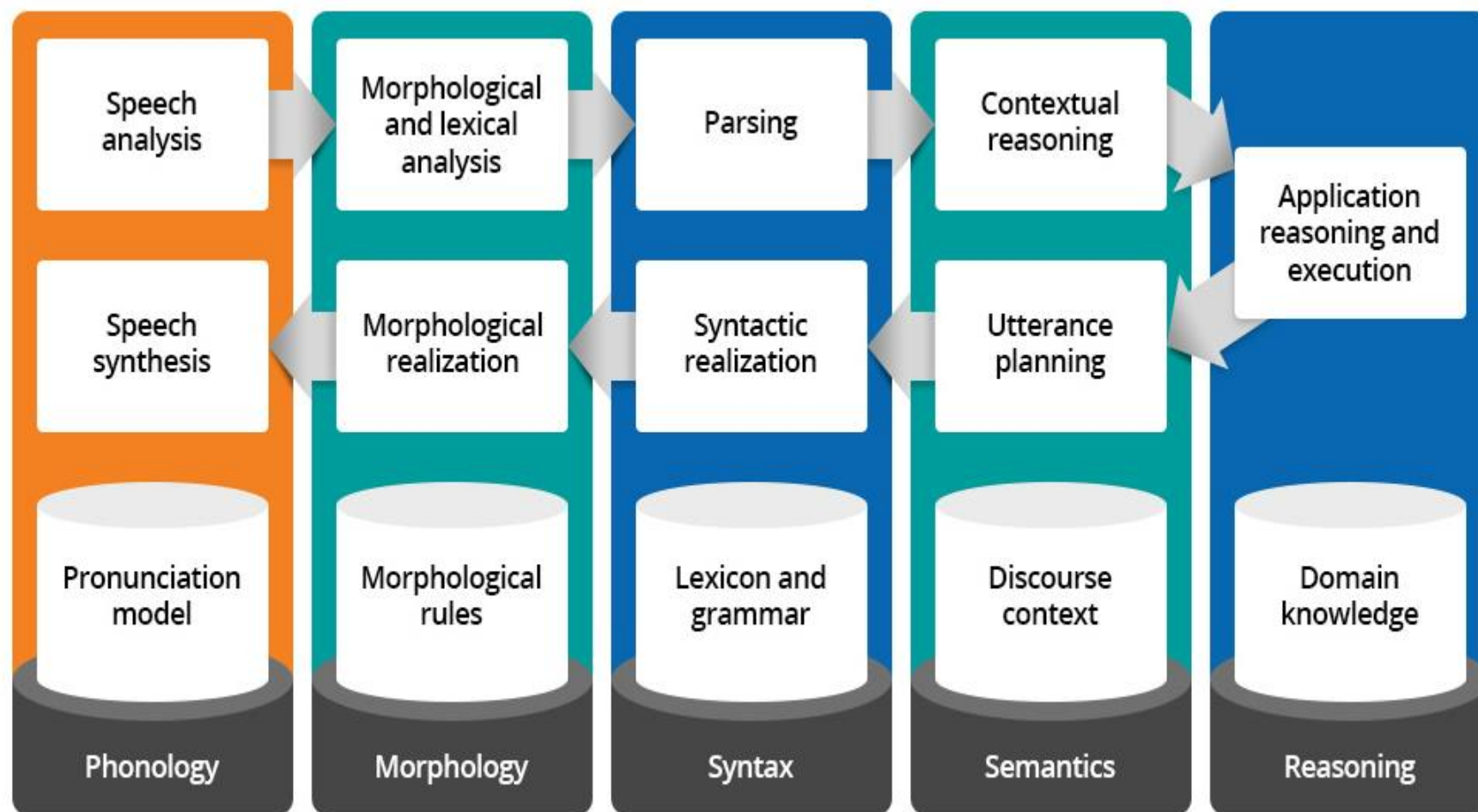
- ✓ Análisis semántico (contexto) intenta formalmente determinar una interpretación semántica no ambigua de la entrada.
- ✓ Problemas: ambigüedad, referencias, etc
- ✓ Resolver “referencias” involucra conocimiento en varios niveles.
- ✓ Interpretación semántica implica también considerar el contexto y conocimiento de palabras.

Niveles en NLP



- ✓ *Análisis del Dominio (pragmático)* involucra reconocer los *actos de habla* del usuario de modo de realizar la interpretación correcta.
- ✓ Existe más significado que las palabras mismas (intenciones, objetivos, etc).
- ✓ Reconocimiento del plan del hablante, muchos implícitos.

Ciclo Completo en NLP



Resumen de Niveles

✓ Fonética/Fonología/Morfología:

¿Qué palabras estamos escuchando/tratando?

✓ Sintaxis:

¿Qué frases estamos tratando?

¿Qué palabras modifican las otras?

✓ Semántica:

¿Cuál es el significado *literal*?

✓ Pragmática:

¿Qué deberíamos *inferir* del hecho que yo dije algo?

¿Cómo deberíamos *reaccionar*?

A photograph of a silver dumbbell and a round analog clock. The dumbbell is on the left, and the clock is on the right. A semi-transparent red banner is placed across the middle of the image, containing the text 'Tiempo de Ejercicios'. The background is a light gray surface.

Tiempo de Ejercicios

Ejercicio

¿Qué tipo de problemas de lenguaje tienen las siguientes oraciones?

1. El cliente no está no satisfecho con la seggurida del banko.
2. El clientes está satisfecho no con la banco del seguridad.
3. El cliente no está satisfecho con el continente.
4. El cliente no está satisfecho con el banco. Sus molestia es comprensible.

Resumen

- ✓ NLP se compone de modelos y métodos que permiten que un computador comprenda y/o produzca lenguaje para diversas aplicaciones y tareas.
- ✓ El lenguaje natural es inherentemente ambigüo, por lo que todas los problemas de NLP son del tipo NP-hard.
- ✓ Las aplicaciones de NLP son transversales en cualquier área comercial/industrial donde se posea problemas con el tratamiento del lenguaje.