Ciencia de Datos y Lenguaje Natural

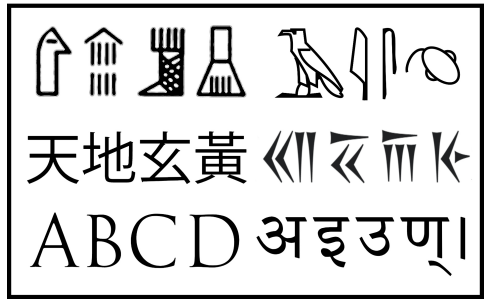
Clase 1.1

Grupo PLN - INCO

Universidad de la República

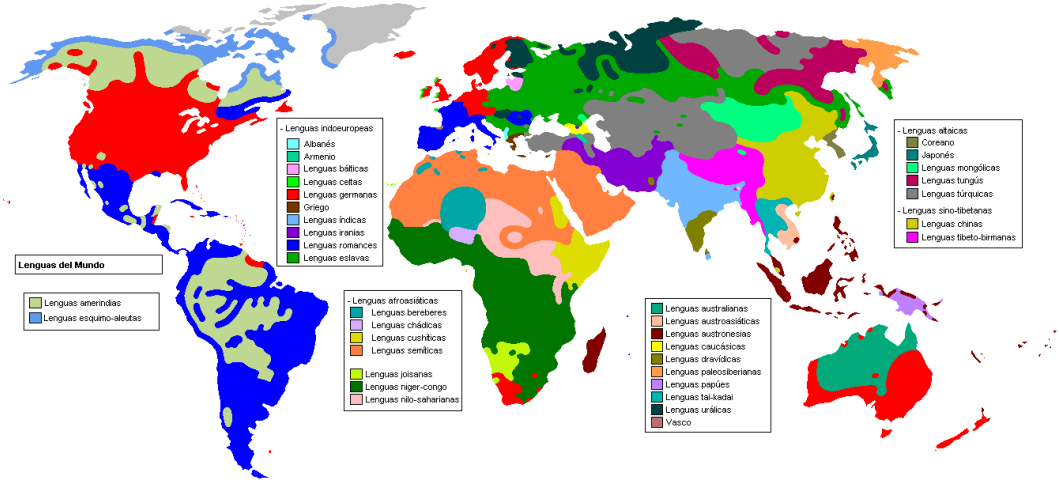
8 de agosto de 2022

- Por Lenguaje Natural entendemos el lenguaje de los humanos.
- Son en realidad diversas lenguas, según www.ethnologue.com hay más de 7.000 lenguas vivas.
- No nos entendemos hablando con hablantes de otras lenguas, ni podemos leer muchas veces siquiera textos escritos en otros alfabetos.



Sistemas de escritura, wikipedia

Mapa de lenguas del mundo



Ciencia de Datos y Lenguaje Natural

- Nos referimos al Procesamiento Automático del Lenguaje Natural, o sea, a programas que comprendan y/o generen lenguaje "en situación".
- Al hablar de Ciencia de Datos y Lenguaje Natural nos referimos a métodos empíricos, basados en datos, para realizar las tareas asociadas al procesamiento del lenguaje.

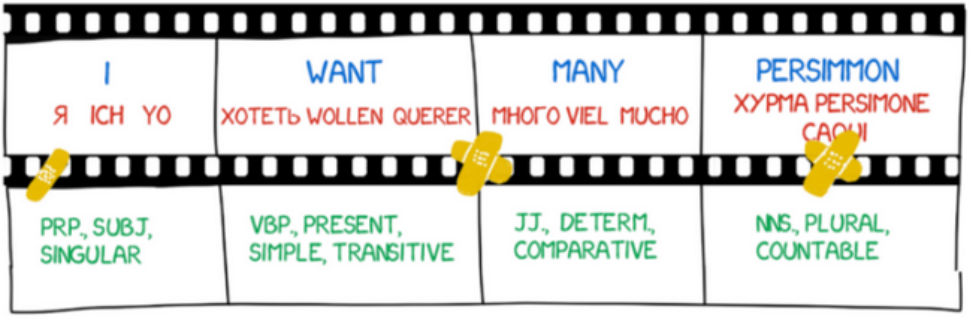
Historia del PLN

- Traducción automática
- Eliza (y los chatbots)
- Ubicuidad del procesamiento de lenguaje

La traducción automática

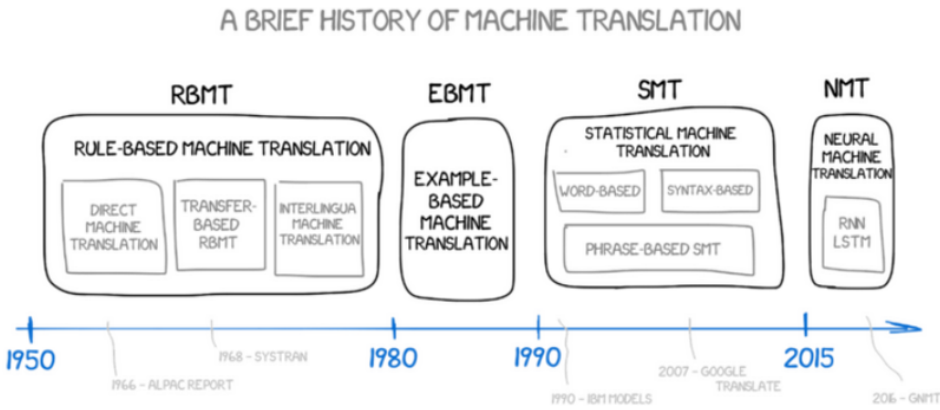
- Uno de los primeros grandes proyectos de la informática
- Antecedente en la URSS en el 1936
- Muy relevante en la guerra fria, traducción del ruso al ingles
- los primeros traductores eran poco más que un sistema de búsqueda en un diccionario informatizado + alguna regla morfológica (género y número de sustantivos, conjugación de verbos)

La traducción automática, URSS 1933, Peter Troyanskii



A history of machine translation from the Cold War to deep learning, Ilya Pestov
https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/

La traducción automática, etapas hasta hoy



Traducción automática, sistemas basados en reglas

RBMT, Rule Based Machine Translation

- Son sistemas de traducción automática basados en información lingüística sobre los idiomas de origen y destino, básicamente recuperados de diccionarios y gramáticas.
- Reglas manuales, directas o pasando por una *interlingua*.
- Técnica que sigue siendo útil, muchas veces como 1er abordaje.
- En el curso se va a experimentar con expresiones regulares, un caso de esta técnica.

Traducción basada en ejemplos

EBMT, Example Based Machine Translation

- Japón 1984
- Sistema basado en corpus paralelos
- Esa idea se mantiene hasta el día de hoy
- Idea de que los datos (corpus) son la fuente de conocimiento.

Traducción estadística

SBMT, Statistical Based Machine Translation

- IBM 1994
- Sistema basado en modelos estadísticos generado a partir de corpus paralelos.
- Se utilizan modelos de ambas lenguas y un modelo de transferencia
- Paradigma dominante hasta la segunda década del siglo XXI

Traducción por Redes Neuronales Profundas

NMT Neural Machine Translation

- Sigue la línea de los SMBT...
- ... con varias innovaciones:
  - representaciones vectoriales de las palabras
  - modelo secuencial único, predice cada palabra a partir de toda la entrada
  - el formalismo es una red neuronal profunda
- Estado del arte al momento actual

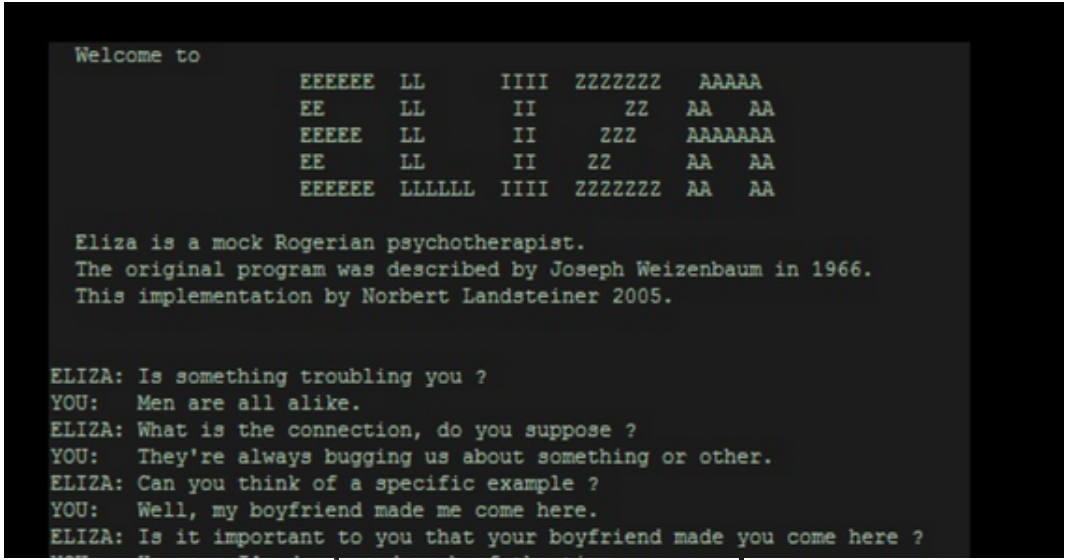
Sistemas que conversan, Respuestas a preguntas

- Asistentes personales, orientados a tareas, guiados por objetivo Gus, Siri, otros
- Respuestas a preguntas Abiertas a la web : Siri, Alexa, Wolfram, Google Colecciones cerradas : Squad
- pasando el test de Turing

Psico-terapeuta en la computadora

- Jospeh Weizanbaum,1964 1966, MIT
- Programa de demostración, imitando un psicólogo.
- Consecuencias insospechadas, la gente se sentía bien y comprendida hablando con Eliza. Primer ejemplo de la concepción antropomórfica de una computadora
- 1964 !! no había monitor !! comunicaciones por una teletipo

Psico-terapeuta?



Otros sistemas de diálogo

- Asistentes personales, orientados a tareas, guiados por objetivo Gus, Siri, otros
- Respuestas a preguntas Abiertas a la web : Siri, Alexa, Wolfram, Google Colecciones cerradas : Squad
- pasando el test de Turing

Aplicaciones PLN en la vida corriente

- Búsqueda de información, respuestas a preguntas
- Corrección ortográfica y gramatical
- Autocompletado al escribir
- Análisis de sentimientos
- Sistemas de recomendación