

Test de Georgetown

a) Háblame del Test de Georgetown: indica sus características, el impacto que produjo en la sociedad y su objetivo. ¿Qué sistemas hay actualmente para la traducción automática?

Fue una demostración realizada sobre traducción automática el 7 de enero de 1954. Desarrollado conjuntamente por la Universidad de Georgetown e IBM, el experimento consistió en la traducción de más de sesenta frases del ruso al inglés.

- Características del test:
 1. Traducción ruso-inglés: Se utilizó un sistema computacional para traducir 60 frases del ruso al inglés.
 2. Lenguaje controlado: Las frases seleccionadas eran específicas y no representaban la complejidad del lenguaje cotidiano, lo que facilitó la tarea.
 3. Reglas básicas: El sistema empleaba reglas gramaticales y un vocabulario limitado, manejado por un algoritmo computacional.
 4. Primera demostración pública: Se presentó como una prueba de concepto más que un sistema funcional de traducción.

Impacto en la sociedad

Fue un test publicitado por los periodistas y percibido como un gran éxito, el experimento animó a los gobiernos a invertir en lingüística computacional. Los autores declararon que en tres o cuatro años, la traducción automática podría ser un problema resuelto. Sin embargo, el progreso real ha sido mucho más lento.

Objetivo del Test

Demostrar que las máquinas podían realizar traducciones automáticas y generar interés en desarrollar sistemas más avanzados para superar barreras lingüísticas.

Sistemas actuales de traducción automática

- Basados en reglas (RBMT - Rule-Based Machine Translation)
- Basados en estadísticas (SMT - Statistical Machine Translation)
- Traducción automática neuronal (NMT - Neural Machine Translation)
- Sistemas híbridos
- Modelos generativos avanzados:

b) Indica porqué entre el año 1954 y los 80 hubo un parón en el desarrollo del procesamiento del lenguaje natural.

El parón del procesamiento del lenguaje natural (PLN) entre 1954 y los años 80, tuvo varias razones clave:

- Expectativas no cumplidas respecto al test Georgetown
- Las limitaciones tecnológicas de la época
- Informe ALPAC (1966) : Un informe del Automatic Language Processing Advisory Committee (ALPAC), encargado por el gobierno de EE.UU., concluyó que la traducción automática no era rentable ni alcanzaba una calidad adecuada para competir con los traductores humanos.
- Dificultades lingüísticas: Los sistemas basados en reglas no podían manejar múltiples significados de palabras o estructuras gramaticales complejas.
- Cambio de enfoque en IA : Los investigadores de inteligencia artificial comenzaron a enfocarse en dominios considerados más alcanzables, como la resolución de problemas matemáticos dejando de lado el procesamiento del lenguaje natural.

c) ¿Cuál es la teoría de Noam Chomsky? Explícala.

La teoría de Noam Chomsky, conocida como gramática generativa, propone que los humanos tienen una capacidad innata para el lenguaje.

Chomsky argumenta que la capacidad para aprender un lenguaje no se adquiere sólo por imitación, sino que es una facultad cognitiva inherente, lo que diferencia a los humanos de otros animales.

d) CORPUS.

¿Qué es un corpus dentro de la IA?

Un Corpus en el ámbito de la inteligencia artificial se refiere a una colección estructurada de textos escritos o grabaciones orales que se utilizan para entrenar y evaluar modelos de aprendizaje automático, especialmente en el campo del procesamiento del lenguaje natural.

¿Qué significa que está etiquetado?

Un corpus etiquetado incluye anotaciones adicionales que describen características específicas del contenido de dicho corpus, como por ejemplo:

- Etiquetas gramaticales: Palabras marcadas con su categoría gramatical (sustantivo, verbo, adjetivo, etc.).
- Entidades nombradas: Identificación de nombres propios, lugares, fechas, etc.
- Sentimiento: Anotaciones que indican si el texto tiene connotación positiva, negativa o neutral.
- Relaciones semánticas: Cómo se conectan las palabras o frases dentro del texto.

Indica sus usos dentro de la IA.

- Entrenamiento de modelos de procesamiento del lenguaje:
- Evaluación: Ayudan a medir el rendimiento de modelos en tareas como clasificación de texto, generación de lenguaje y traducción.
- Análisis lingüístico: Permiten estudiar patrones del lenguaje para mejorar los sistemas de procesamiento del lenguaje.

Pon ejemplos de lingüísticos, masivos y específicos.

1. Lingüísticos:

Penn Treebank: Un corpus etiquetado con información sintáctica, usado para tareas de análisis gramatical.

2. Masivos:

Common Crawl: Contiene datos de internet a gran escala, utilizado para modelos como GPT o BERT.

3. Específicos:

PubMed Corpus: Contiene textos biomédicos para aplicaciones en salud y farmacología.

¿Qué significa un corpus curado?

Un corpus curado es una colección de datos cuidadosamente seleccionada, limpiada y anotada para garantizar que sea precisa, relevante y de alta calidad para su propósito.

