Test de Georgetown

a) Test de Georgetown: características, impacto y objetivo

El Test de Georgetown fue una demostración de traducción automática desarrollada por investigadores de la Universidad de Georgetown y la IBM. Se realizó en 1954 con el objetivo de traducir 60 oraciones del ruso al inglés usando una computadora IBM 701.

1. **Limitaciones deliberadas:** El sistema estaba diseñado para trabajar con un vocabulario reducido (alrededor de 250 palabras) y reglas gramaticales simplificadas.
2. **Propósito experimental:** Fue más una prueba conceptual que un sistema práctico.
3. **Reglas basadas en gramática:** Usó un enfoque basado en reglas lingüísticas, típicas de la época.

El éxito percibido de esta demostración generó un gran entusiasmo en la comunidad científica y en la sociedad en general.

- Se creyó que la traducción automática estaba a punto de resolver problemas de comunicación global, lo que atrajo financiamiento significativo para investigaciones en esta área.
- Inspiró la creación de numerosos proyectos gubernamentales y académicos en traducción automática, especialmente en EE. UU. durante la Guerra Fría, para manejar el volumen de documentos en ruso.

Mostrar que las máquinas podían realizar traducción automática entre lenguajes naturales y sentar las bases para una era de colaboración entre lingüística y computación fue su principal objetivo

Encontramos los siguientes sistemas actuales de traducción automática:

1. **Modelos estadísticos:** Por ejemplo, Google Translate en sus inicios utilizaba modelos basados en estadísticas para traducir textos mediante patrones.
2. **Traducción automática neuronal:** Sistemas modernos como DeepL, Google Translate y ChatGPT emplean redes neuronales profundas para generar traducciones más precisas.
3. **Modelos híbridos:** Algunos sistemas combinan enfoques basados en reglas, estadísticos y redes neuronales para mejorar la calidad.

b) Parón en el desarrollo del PLN entre 1954 y los 80

Este parón, conocido como el "Invierno de la Traducción Automática," fue causado por varios factores:

1. **Expectativas no realistas:** Tras el Test de Georgetown, se creyó que la traducción automática completa sería posible en pocos años, lo cual resultó ser demasiado optimista.
2. **Informe ALPAC (1966):** El Comité Asesor de Procesamiento Automático de Lenguaje de los Estados Unidos publicó un informe crítico que concluyó que los resultados de la traducción automática eran mediocres en comparación con el costo. Esto resultó en una reducción drástica del financiamiento.
3. **Limitaciones tecnológicas:** Las computadoras de la época no tenían suficiente poder de cómputo ni almacenamiento para manejar la complejidad del lenguaje humano.
4. **Enfoques insuficientes:** El énfasis excesivo en reglas lingüísticas estrictas no era capaz de capturar la ambigüedad y la diversidad del lenguaje humano.

c) Teoría de Noam Chomsky

Chomsky introdujo el concepto de **gramática generativa** en su obra "Syntactic Structures." Su teoría se centra en describir las reglas abstractas que generan todas las oraciones gramaticales en un idioma. Elementos a destacar son:

1. **Competencia vs. desempeño:**
    - La **competencia** se refiere al conocimiento implícito de un hablante sobre su idioma.
    - El **desempeño** se refiere a cómo se utiliza este conocimiento en situaciones reales, incluidas pausas y errores.
2. **Estructura jerárquica:** Las oraciones tienen una estructura profunda (abstracta) y una estructura superficial (cómo se pronuncian o escriben).
3. **Transformaciones:** Las reglas transformacionales explican cómo se derivan las estructuras superficiales de las estructuras profundas.

La teoría de Chomsky revolucionó la lingüística moderna y sentó las bases para enfoques computacionales más avanzados del lenguaje. Fue fundamental para el desarrollo de teorías formales en PLN.

d) Corpus en IA

Un corpus es un conjunto grande y estructurado de textos (o datos lingüísticos) que se utiliza para analizar, entrenar o evaluar sistemas de procesamiento de lenguaje natural.

Un corpus etiquetado contiene anotaciones adicionales, como marcas gramaticales, categorías semánticas o etiquetas sintácticas. Por ejemplo, las palabras en un texto pueden estar etiquetadas como sustantivos, verbos, etc., o las frases pueden incluir análisis de dependencia. Los usos del corpus en la IA son el entrenamiento de modelos de PLN, como sistemas de traducción, análisis de sentimientos y modelos de lenguaje. La evaluación de algoritmos para garantizar precisión y eficacia. Y por último investigación lingüística para entender patrones en lenguas naturales.

Ejemplos de corpus:

1. **Lingüísticos:**
    - Corpus Brown: Un corpus etiquetado con categorías gramaticales.
    - Penn Treebank: Contiene anotaciones sintácticas.
2. **Masivos:**
    - Common Crawl: Una colección masiva de datos web utilizada para entrenar modelos como GPT.
    - Wikipedia Dumps: Conjunto de textos de Wikipedia usados para modelos de lenguaje.
3. **Específicos:**
    - MEDLINE: Corpus de artículos médicos.
    - Corpora legales como EULEX para procesamiento jurídico.

Un corpus curado es aquel que ha sido cuidadosamente recopilado, revisado y anotado por expertos para garantizar su calidad y precisión. Esto es crucial para obtener modelos más precisos y confiables.