Traducción automática

TECNOLOGÍAS DE LA TRADUCCIÓN

¿Qué son los sistemas de TA?

- Sistemas informáticos que llevan a cabo traducciones de una lengua a otra con o sin intervención humana

Ideas preconcebidas sobre la TA

- *La traducción automática es inútil porque produce traducciones de muy baja calidad*.
  - La TA se utiliza productivamente en muchas empresas y organismos (El Periódico, Microsoft, Ford, VW, SAP, Unión Europea, etc.).
  - No siempre es necesario tener traducciones perfectas (*information gisting*).
  - La calidad generalmente es menor pero tiene valores añadidos como el tiempo y el coste.

Ideas preconcebidas sobre la TA

- *La traducción automática amenaza el trabajo de los/as traductores/as.*
  - Gran parte del volumen de textos traducibles no podría ser asumido por traductores profesionales por razones económicas y de tiempo.
  - La TA libera al traductor de la parte más repetitiva y rutinaria de su trabajo.
  - Integración de los sistemas de traducción automática en el entorno de la traducción profesional (funciones de pre-edición y post-edición). Nuevas salidas profesionales.

Automatización de la traducción

Hutchins & Somers, 1992

Tipos de texto y TA

- Textos escritos en lenguaje formal y estereotipado, relativamente predecible:
  - Disposiciones legales y administrativas
  - Textos jurídicos (normativas, contratos, etc.)
  - Manuales técnicos
  - Boletines informativos (partes meteorológicos, bolsa, teletexto, anuncios por palabras, ofertas de empleo, etc.)
  - Resúmenes de publicaciones científicas, informes técnicos y textos expositivos.

Tipos de textos y TA

- La TA es inadecuada para procesar:
  - Textos creativos: narrativa, poesía, teatro, ensayo.
  - Textos expresivos: lenguaje coloquial, lenguaje humorístico, chistes, juegos de palabras, etc.

Necesidades de la TA

- Factores socio-políticos:
  - Crecimiento exponencial de la información
  - Multilingüismo
- Factores comerciales:
  - Incremento del comercio mundial
  - Manuales y documentación técnica (alrededor del 30% del coste total del producto)
  - Actualización de páginas web
- Factores científicos:
  - Reto científico de varias disciplinas: lingüística computacional, inteligencia artificial, ingeniería del conocimiento.

Métodos de TA

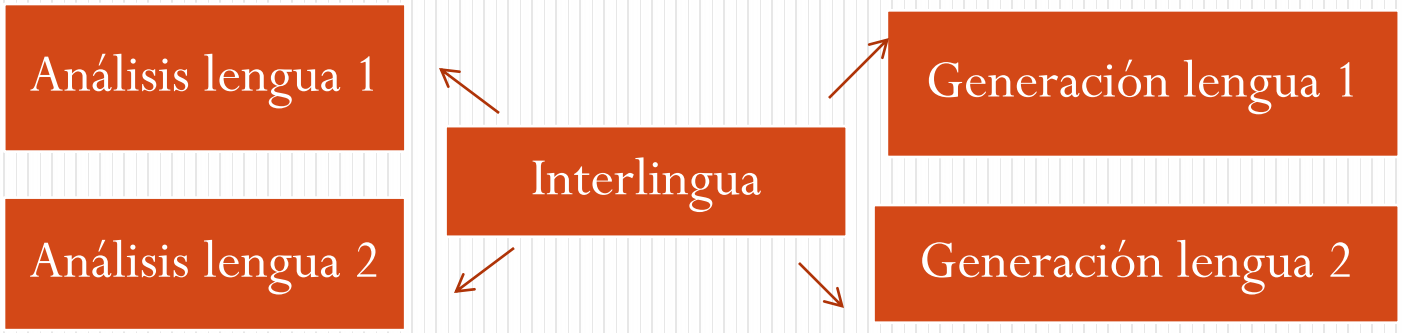
- Métodos basados en reglas
  - Método directo
  - Métodos indirectos (con representación intermedia)
    - Interlingua
    - Transferencia
  - Métodos basados en conocimiento
- Métodos sin reglas lingüísticas
  - Métodos estadísticos
  - Métodos basados en ejemplos

Método directo

- Método más primitivo, utilizado en los años 50.
-También llamado "traducción palabra por palabra".
-No hay análisis sintáctico ni semántico.
-Tres fases: análisis morfológico, consulta diccionario, reordenación.

Ejemplo: SALT (Generalitat Valenciana)

Método interlingua (indirecto)



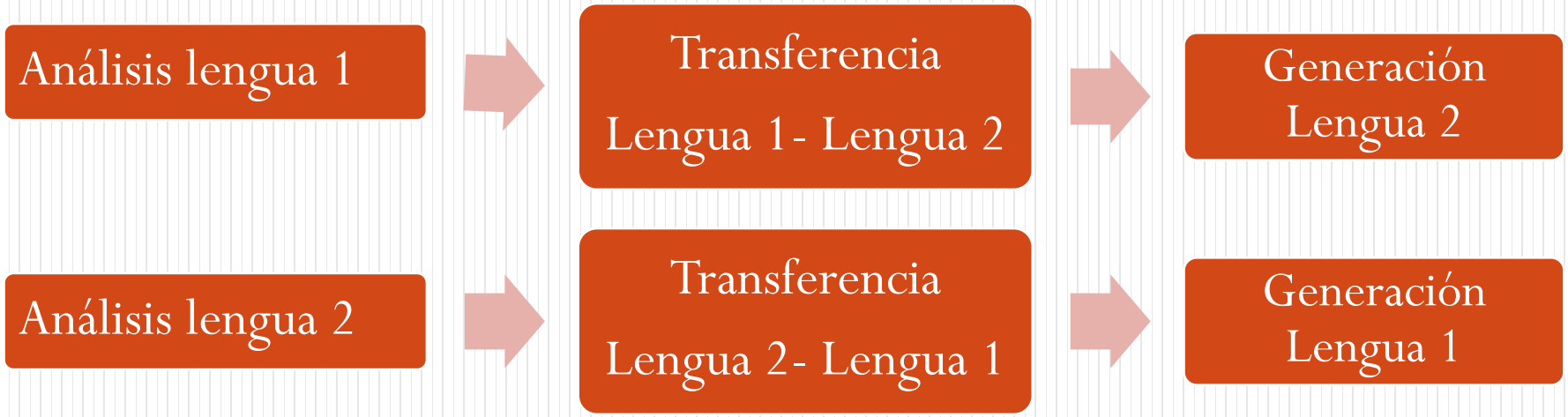
-Única representación intermedia común a todas las lenguas (representación conceptual).

- Basada en la representación del conocimiento que propuso la Inteligencia Artificial en los años 80.

-Ventaja: desarrollo de sistemas multilingües.

- Dificultad: establecer esta interlingua universal que sirva para cualquier lengua.

Método de transferencia



-Representación intermedia dependiente del par de lenguas.

-Ventaja: representación intermedia menos abstracta que la interlingua.

-El módulo de transferencia contiene el diccionario y el conjunto de reglas sintácticas y morfológicas para la transferencia.

-Ex: adj subst1 subst2 subst3 → subst3 adj prep de subst2 prep de subst1

-Dificultad: diseñar un modelo de representación para cada para de lenguas.

-Ejemplos: SYSTRAN, ProMT

Método basado en conocimiento

- Premisa: para conseguir una traducción de calidad es necesario comprender el significado del texto origen
- Estos sistemas deben poseer los siguientes componentes: gramática, lexicón (diccionario), ontología (conjunto de conceptos de un dominio relacionados entre sí), reglas sintácticas para el análisis y reglas sintácticas para la generación.
- La ontología ayuda a establecer las restricciones sobre conocimientos del mundo, a clasificar personas, lugares, roles, etc.
- Problemas: sistema muy complejo de crear y gestionar incluso para dominios muy restringidos.

Método basado en ejemplos

- Basados en corpus paralelos
- No hay análisis lingüístico profundo sino sustitución de frases
- Ventaja: Exige menos inversión tecnológica
- Se fundamenta en la reutilización de traducciones humanas
- Problema: es necesaria gran cantidad de corpus paralelos anotados lingüísticamente

Método estadístico

- Iniciado por IBM a finales de los 80
- Objetivo: Utilizar corpus paralelos para extraer la información estadística necesaria con el objeto de "entrenar" el sistema de TA.
- A partir de este análisis, realizan cálculos de probabilidades:
  - Probabilidad de que las palabras de una oración en lengua meta se ordenen de una determinada manera.
  - Probabilidad de que una palabra se traduzca de una determinada manera en otra lengua.
  - Probabilidad de que la traducción de una palabra se traduzca mediante una expresión de varias palabras en la lengua meta.
  - Etc.
- Ejemplos: Google, Asia Online

Problemas de la TA

- Ambigüedad léxica
- Ambigüedad estructural
- Diferencias conceptuales en el vocabulario
- Construcciones elípticas y resolución de anáforas
- Construcciones incorrectas gramaticalmente

Ambigüedad léxica

- Una palabra puede tener varios sentidos y, por lo tanto, varias traducciones diferentes:
  - Liverpool was eliminated in the first **round**.
  - The cowboy started to **round** up the cattle.
  - I went to buy a **round** table.
  - We are going to a cruise **round** the world.
  - The tree measured six feet **round**.

Ambigüedad estructural

- Una frase puede analizarse sintácticamente de más de una forma.
  - Flying planes can be dangerous
    - It can be dangerous to fly planes.

    Puede ser peligroso pilotar aviones
    - Planes which are flying can be dangerous.

    Los aviones que están volando pueden ser peligrosos

Diferencias conceptuales en el vocabulario

- Comer

  1) sujeto humano -> essen
  2) en los demás casos -> fressen

- Biblioteca

  1) Institución académica-> Bibliothek
  2) Bib. Pública -> Bücherei

Resolución de anáforas

- Un pronombre puede tener varios antecedentes potenciales, lo que en ocasiones da lugar a distintos significados y posiblemente varias traducciones.
  - The soldiers shot at the women and some of them fell.

    (Los soldados dispararon a las mujeres y algunas de ellas/algunos de ellos cayeron)
  - The soldiers shot at the women and some of them missed.

    (Los soldados dispararon a las mujeres y algunas de ellas/algunos de ellos fallaron)

Resolución de anáforas

- Mientras el sentido común o conocimiento del mundo permite que el humano decida la interpretación correcta, el ordenador requiere reglas específicas que le indiquen qué interpretación elegir y por tanto cómo traducir 'some of them'
- The soldiers shot at the women and some of them fell.

  Los soldados dispararon a las mujeres y algunas de ellas cayeron
- The soldiers shot at the women and some of them missed.

  Los soldados dispararon a las mujeres y algunos de ellos fallaron

Los programas de TA en la práctica

- Traducción totalmente automática de alta calidad (no realista en la actualidad)
- Traducción automática con participación humana
  - Pre-edición: preparación del texto antes de ser traducido automáticamente
  - Post-edición: corrección del texto después de la traducción automática
  - Traducción interactiva: interacción humana durante el proceso de traducción automática.

TA para sublenguajes

- El programa está diseñado para tratar el vocabulario y las construcciones típicas de un determinado campo de conocimiento y un tipo específico de documento.
  - Ej.: Météo (partes meterológicos canadienses, inglés > francés)

TA de baja calidad

- Uso del producto final sin revisar:
  - El resultado es de baja calidad, pero proporciona información suficiente para extraer al menos cierta noción del contenido del texto.
  - Las personas con limitaciones de tiempo y dinero prefieren una traducción en borrador que ninguna.
  - Permiten decidir si se desea una traducción de calidad del texto, según su interés.

Lenguajes controlados

- Objetivo:
  - Aumentar la legibilidad de la documentación técnica
  - Aumentar la traducibilidad
  - Facilitar el uso de tecnologías de la traducción
- Cómo:
  - Imposición de estilo directo y claro en la redacción
  - Reducción de ambigüedades sintácticas
  - Reducción de ambigüedades léxicas
- Resultado:
  - Consistencia en el estilo
  - Reusabilidad de los textos
  - Ahorro en los procesos de redacción y traducción/localización

Reglas de pre-edición

- Pre1) Mantener las oraciones cortas y simples
- Pre2) Evitar las coordinación múltiple de oraciones
- Pre3) Insertar determinantes donde sea posible
- Pre4) Insertar *that*, *which*, *in order to* en las oraciones subordinadas cuando sea posible
- Pre5) Evitar los pronombres con referencia anafórica (*it*, *they*, *them*)
- Pre6) Evitar las construcciones elípticas
- Pre7a) Reescribir *when*, *while, before, after* seguido de –ing
- Pre7b) Reescribir *when*, *where, if* seguido de participio pasado
- Pre8) Evitar los verbos frasales
- Pre9) Evitar los adjetivos, participios pasados (-ed) y participios de presente (–ing) en posición post-nominal
- Pre10a) Repetir el nombre cuando vaya modificado por dos o más adjetivos coordinados
- Pre10b) Repetir el adjetivo cuando vaya modificando a dos o más nombres
- Pre11) Repetir las preposiciones en la coordinación de sintagmas preposicionales
- Pre12) Rescribir los compuestos nominales de más de 3 nombres

Post-edición

- Corrección realizada por humanos del texto resultante de la traducción automática
- ¿Cuánta **corrección** hay que hacer?
  - Depende del nivel de aceptabilidad
  - Depende del esfuerzo humano necesario
- Diferentes niveles:
  - Sin post-edición
  - Con post-edición rápida
  - Con post-edición completa

Principios generales de post-edición

- Principios generales de la Comisión Europea:
  - **Do retain as much of the raw translation as possible**
  - **Neither delete nor rewrite too much**
  - **Remember that many of the words are there, but in the wrong order**
  - **Don't worry if the style is repetitive or pedestrian**
  - **Correct only non-sensical or plainly wrong words**

Post-edición mínima basada en J2450

- *Post1*) Corrija todo término incorrecto en el texto, sea técnico o no técnico.
- *Post2*) Corrija cualquier error sintáctico:
  - categoría gramatical equivocada
  - estructura incorrecta de sintagma
  - orden de palabras equivocado
- *Post3*) Añada texto que no ha sido traducido
- *Post4*) Corrija cualquier error de desinencia morfológica o de concordancia morfológica
- *Post5*) Corrija cualquier error ortográfico
- *Post6*) Corrija cualquier error de puntuación
- *Post7*) Corrija cualquier otro error que no haya sido corregido con las reglas anteriores

Tipos de errores de la TA

- **Error léxico**
  - La traducción de un término no se corresponde con el equivalente adecuado en un campo de especialidad.
  - El término se ha traducido de distintas formas en el mismo texto.
- **Error sintáctico**
  - Palabra con categoría gramatical incorrecta.
  - Orden de las palabras incorrecto.
  - Estructura sintáctica incorrecta.
  - Otro
- **Error morfológico**
  - Una palabra tiene morfología incorrecta: género, número, tiempo verbal, etc.
  - Problema de concordancia entre dos o más palabras.
- **Ortografía**
  - El texto no respeta las normas de ortografía.
- **Puntuación**
  - El texto no respeta las normas de puntuación.
- **Omisión**
  - Alguna palabra o parte del texto no ha sido traducida.