Diseño de sistemas de TA

- A la hora de diseñar un sistema se decide principalmente su aplicación
  - Esto generalmente determina que tipo de sistema a implementar
- Los sistemas pueden clasificarse por distintos aspectos (no mutuam. excluyentes)
  1. Pares de idiomas y direcciones tratados
  2. Intervención humana en el proceso
  3. Tipo de tarea a realizar
  4. Estrategias de traducción implementada

1. Idiomas y direcciones tratados

- Idiomas: sistemas bilingüe vs. Multilingüe
  - Sistemas que tratan 1 par de lenguas o muchos
- Direcciones: de qué LO a qué LM traduce
  - LO = lengua origen, LM = lengua meta (o SL/TL)
  - Sistema bilingüe puede ser uni-direccional o bi-direccional
  - Sistemas multilingue puede ir de un idioma a muchos o en cualquier direccion entre muchos idiomas o tratar varios pares pero no todas las direcciones
- Ejemplos
  - EUROTRA de/hacia 9 idiomas → 9 * 8 = 72 pares
  - Primeras versiones de Systran eran una colección de sistemas bilingues uni-direccionales

Sistemas multilingue

- Los “verdaderos” sistemas multilingüe mantienen el tratamiento de cada idioma en módulos independientes
  - Ej. tratan EN como LO siempre igual aunque LM sea ZH, ES o QU
- Cuando es preferible un sistema puramente multilingüe?
  - Discusión Ariane o KANT vs. Météo

2. Intervención humana (1/2)

- El proceso de pasar de LO a LM puede realizarse sin intervención humana (off-line) o con asistencia humana (interactivos)
- Sistemas interactivos involucran al usuario durante el proceso de traducción
  - Sobre ciertas decisiones de traducción, ej para resolver ambigüedades
  - Si el texto a traducir contiene sintaxis o vocabulario que no reconoce
  - Propone traducciones a medida que el usuario tipea (predicción de palabras y/o frases)
- Tienen requerimientos estrictos (rapidez para tiempo real, usuarios expertos en LMs)
- Eliminan el costo de pre/post edición pero podrían no reducir el tiempo de producción (ver biblio 4)

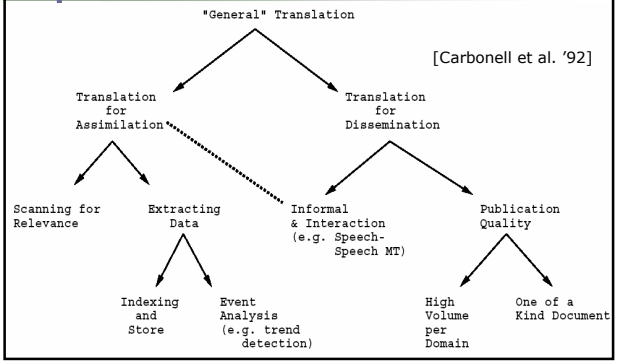
2. Intervención humana (2/2)

- Proceso off-line puede requerir intervención humana
- Antes de la traducción = pre-edición del texto
  - Antes de TA eliminar ambiguedades, construcciones complejas, vocabulario desconocido, etc → asociado a lenguajes controlados
  - Requiere alto conocimiento del sistema a usar
- Después de la traducción = post-edición Un humano arregla la TA → cuesta menos que traducir de cero y es mas rápido
  - Requiere competencia casi nativo en LM

3. Tareas a realizar (1/2)

- Algunas tareas requieren cierto nivel de calidad de la TA para realizarse con éxito
- En general hay 2 tareas principales
  - Asimilación: sacar la idea del texto o encontrar palabras clave para indexar o evaluar pertinencia de un documento a una búsqueda
  - Diseminación: apunta a publicar el texto sin editar o usarlo en etapa siguiente de un pipeline
- Mejorar la calidad implica mejorar el sistema? → ver “Good applications for crummy MT”

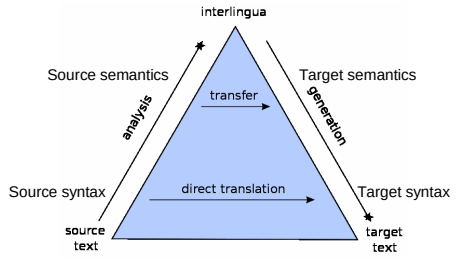
3. Tareas a realizar (2/2)



4. Estrategias de traducción

- Clasificar los sistemas por el tipo de paradigma que implementan es lo mas común
- Hay muchos y siguen apareciendo
  - Basada en reglas
    - Directa
    - Transferencia
    - Interlingua
  - Basada en ejemplos
  - Estadística
  - Híbridos

Pirámide de Vaquois (1968)



Traducción directa

- También llamados "primera generación" traducen de origen a destino mediante diccionarios y gramáticas
- Proceso
  - Traducción palabra a palabra
  - Reglas de reordenación local
- Puede suceder que
  - Frases sin sentido
  - Podría ser útil en lenguaje controlado

Modelo simplista

1. Identificar y separar oraciones
2. Separar en tokens (tokenisation)
3. Manejo de mayúsculas (suelen sacarse o usarse para identificar entidades)
4. Búsqueda en diccionario y sustitución (puede incluir resolución de ambigüedades)
5. Copiar palabras desconocidas, dígitos, puntuación, etc.

Modelo mejorado (Tucker 1987)

1. Búsqueda en diccionario origen y análisis morfológico
2. Identificación de homógrafos
3. Identificación de sustantivos compuestos
4. Identificación de frases verbales y sustantivos
5. Procesamiento de expresiones idiomáticas

Modelo mejorado (Tucker 1987)

6. Procesamiento de preposiciones
7. Identificación de sujeto – predicado
8. Identificación de ambigüedades sintácticas
9. Síntesis y procesamiento morfológico de texto destino
10. Reordenamiento de palabras o frases en texto destino

Analisis del modelo

- Problemas
  - La base es el análisis morfológico y un diccionario bilingüe → reordenamientos de larga distancia, estructura sintáctica
- Ventajas
  - podría ser útil para traducir entre idiomas muy cercanos
  - Es fácil de implementar

Ejemplos de un sistema RU →EN [Hutchins&Somers]

My trebuem mira.
We require world
'We want peace.'

On dopisal stranitsu i otložil ručku v storonu.
It wrote a page and put off a knob to the side.
'He finished writing the page and laid his pen aside.'

Včera my tselyi čas katalis' na lodke,
Yesterday we the entire hour rolled themselves on a boat.
'Yesterday we went out boating for a whole hour.'

Modelos indirectos

- También llamados “segunda generación”
- Usan representaciones intermedias para pasar de origen a destino
  - La idea es capturar la semántica del texto
- Dos variantes principales
  - Por transferencia
  - Interlingua

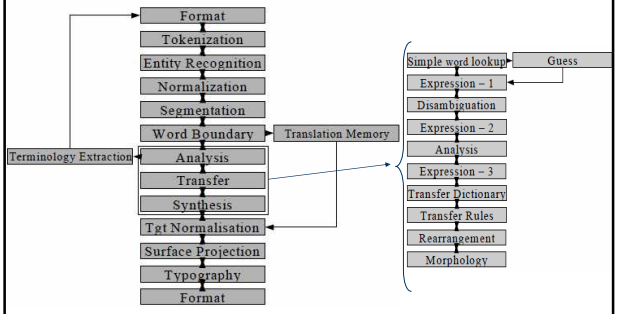
Por transferencia

- Transferencia a nivel sintáctico
  - Transferir = pasar de LO a LM = traducir
- Se analizan ambos lenguajes con módulos dependientes del lenguaje
  - Ej. un modulo para EN→ES, otro para EN→FR, etc
- Se relacionan ambas estructuras con gramáticas bilingües
  - Reglas que indican como transformar LO en LM

Por transferencia

- Proceso varía en cada sistema pero incluye
  - Análisis morfológico (POS tags)
  - Análisis sintáctico (parsing)
  - Transferencia léxica (diccionarios)
  - Transferencia estructural (gramáticas bilingues)
  - Síntesis (generación morfológica, reordenamiento, etc)

Ejemplo: Systran



- http://wwwv5.systransoft.com/IDC/26459.html

	A $100 wrist-watch.
	<html><hr>A $100 wrist-watch.</html>
first xml representation	<?xml version="1.0"?> <document original format="html"><tag><hr></tag><par id="1">A <ts face="bold">$100</ts> wrist-watch.</par> </document>
tokenized text after entity recognition	[...] <par id="p1" xml:lang="en"> <token type="word" capit="first" id="t1">A</token> <ts face="bold"><entity type="monval" id="t2">$100</entity></ts> <token type="word" norm="wrist-watch" id="t3">wristwatch</token> <token type="punct" id="t4">.</token> </par> [...]
segmented, translated text	[...] <par id="1"> <tu group id="s1"> <tu xml:lang="en"> <token type="word" id="t1" capit="first">A</token> <ts face="bold"><entity type="monval" id="t2">$100</entity></ts> <token type="word" norm="wrist-watch" id="t3">wristwatch</token> <token type="punct" id="t4">.</token> </tu> <transtu xml:lang="fr"> <token id="t1" synt="pos=det">Une</token> <token type="word" type="noun" id="t3">montre-bracelet</token> <token type="det" id="t2-0">de</token> <ts face="bold"><entity type="monval" id="t2">$100</entity></ts> <token type="punct" id="t4">.</token> </transtu> </tu group> </par> [...]
output text	<html><hr>Une montre-bracelet de $100.</html> Une montre-bracelet de $100.

Análisis del modelo

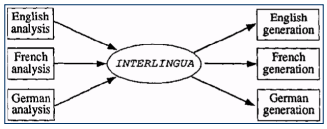
- Ventajas
  - Análisis más profundo que permite solucionar orden de palabras o constituyentes
- Desventajas
  - Es costoso si hay muchos pares de lenguas
    - Específicamente *n* analizadores LO, *n* generadores LM y *n * (n - 1)* módulos de transferencia
  - Reglas generalmente escritas a mano → se necesita un lingüista para esta tarea

Modelo interlingua

- Se analiza el texto LO y se crea un representación abstracta del mismo
  - Interlingua o idioma pivote = lenguaje artificial neutro para cualquier idioma
- Interlingua contiene toda la información necesaria para generar texto en LM sin referirse a LO (proyección)
- Cada modulo de análisis puede ser independiente de módulos de generación y de otros módulos de análisis

Modelo interlingua

- Interlingua permite traducir de un idioma a si mismo → “back-translation”
  - Usado durante el desarrollo para testear análisis y generación
- Agregar módulos de aumenta exponencialmente el nro de pares tratados



Ejemplo de interlingua

- Representación para “Does coffee give you headaches?”
- Quasi Logical Form vs. interlingua alternativa simplificada

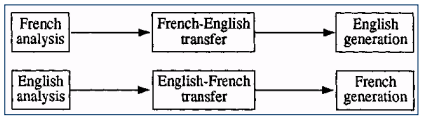
```
[dcl,
 form(verb(present,no,no,no,yes),E,
      [[give1,
        E,
        term(q(bare,sing),X,
             [coffee1, X])
        term(ref(pro,you,sing),Y,
             [person,Y]),
        term(q(bare,plur),Z,
             [headache1, Z])]])]
```

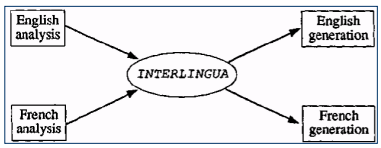
```
[[utterance_type,ynq],
 [action,give],
 [cause,coffee],
 [pronoun,you],
 [symptom,headache],
 [tense,present], [voice,active]]
```

Análisis del modelo

- Ventajas
  - Eficiente para el desarrollo de sistemas multilingües
  - Se podrían capturar diferencias sintácticas
    - The bottle floated into the cave →La botella entró a la cueva flotando (→ the bottle entered in the cave floating)
- Desventajas
  - Creación interlingua es muy difícil
  - Interlingua sigue siendo una representación en texto

Transferencia vs Interlingua





- Cap 4 Hutchins & Somers

La próxima clase

- Será sobre TA basada en ejemplos (example-based), estadística y traducción del habla
- Lectura pre-clase, con (*) alcanzaría
  - "Example-Based Machine Translation: A New Paradigm" [Kit et al 2000]
  - (*) "Towards a definition of example-based machine translation" [Hutchins 2005]
  - (* pags 1 a 4) "A Statistical MT Tutorial Workbook" [Knight 1999]
  - "The mathematics of SMT: parameter estimation" [Brown et al 1993]

Bibliografía

1) An introduction to machine translation- Hutchins&Somers
2) Machine translation: An introductory guide – Arnold et al
3) Carbonell, J. G. and Tomita, M., ''New Approaches to Machine Translation,''
4) User-Friendly Text Prediction for Translators – Foster et al 2002
5) "Current strategies in machine translation research and development", Tucker 1987
6) A Framework Of A Mechanical Translation Between Japanese And English By Analogy Principle – Nagao 1984
7) The Mathematics of Statistical Machine Translation: Parameter Estimation [Brown et al 1993]
8) A Statistical MT Tutorial Workbook, Kevin Knight, 1999
9) The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System, Bach et al 2007
10) Performance Evaluation of Speech Translation Systems – Weiss et al 2008
11) Recursos para TA http://www.computing.dcu.ie/~mforcada/fosmt.html