Traducción basada en ejemplos

- Objetivo: obtener una nueva traducción en base a traducciones previas
  - Textos alineadas y almacenadas en memorias de traducción o DBs
  - Alinear textos = crear pares (LO, LM) que correspondan a la misma frase
  - A diferencia de sistemas basados en reglas EBMT es basado en corpus
- Originalmente llamada traducción por analogía [Nagao 1984] ahora example-based (EBMT)
  - Intenta superar los inconvenientes de sistemas basados en reglas para idiomas que difieren mucho (EN ← → JP)
  - Inspirado por metodologías de aprendizaje de idiomas
    - "We learn a language not by deep linguistic analysis and rule application, but by rote learning and analogy-based generalization."

Traducción basada en ejemplos

- Proceso
  1. Analizar frase a traducir (parser)
  2. Partir la entrada en fragmentos
  3. Encontrar los equivalentes entre los ejemplos
     1. Si el matching es exacto = traducción directa
     2. Si no lo es ~ trad. indirecta con análisis y generación
  4. Selección y extracción de frases equivalentes en LM
  5. Adaptación y/o combinación de ejemplos LM
  6. Generación de frase completa (output)

Ejemplo de [Sato & Nagao 1990]

Input

He buys a book on international politics

Matches

He buys a notebook.

Kare wa nōto o kau.

I read a book on international politics.

Watashi wa kokusai seiji nitsuite kakareta hon o yomu.

Result

Kare wa kokusai seiji nitsuite kakareta hon o kau.

Otro ejemplo

- Corpus

  Ich wohne in Dublin ⇔ I live in Dublin

  Ich kaufe viele Sachen in Frankreich ⇔I buy many things in France

  Ich gehe gern spazieren mit meinem Ehemann ⇔ I like to go for a walk with my husband
- Traducción

  Ich wohne in Frankreich mit meinem Ehemann ⇔ I live in France with my husband

Análisis del modelo

- Ventajas
  - No necesita reglas explicitas (menos costoso desarrollar)
  - Económico en dominios acotados (traducciones repetitivas o muy regulares)
  - Podría ser sencillo mejorar la calidad agregando mas ejemplos
- Desventajas
  - Necesita un corpus de ejemplos
  - Medir la similitud de fragmentos es complicado
    - Se pueden usar tesauros para medir la similitud semántica, medidas de distancia entre palabras (levenshtein), anotar el corpus con info lingüística (pos tags, relaciones, etc), generalización del corpus
  - Determinar el tamaño de los fragmentos
    - Si son muy largos hay poca probabilidad de match exacto

Traducción estadística

- Modelo basado en corpus pero a diferencia de EBMT las traducciones se generan a partir de probabilidades calculadas sobre un corpus paralelo
- Modelo relativamente nuevo, propuesto en los 90 por grupo en IBM [Brown 1993]
- La idea es ver cuál es la probabilidad de traducir una palabra *e* como la palabra *f*
  - Como Brown et al trabajaban en Fr→En el modelo considera *e* = *English* y *f* = *French*
  - Mas adelante vamos a ver que el modelo considera también otras opciones (frases, árboles, …)

Traducción estadística

- Si traducían Fr→En por qué hablan de la probabilidad de traducir e como f ??
- Modelo basado en un “canal ruidoso”
  - En el proceso de traducción el canal ruidoso “deforma” la frase en Ingles y la convierte en una frase en Francés
- El problema es encontrar la frase en Ingles que generó tal transformación
  - Analogía con la criptografía → idea expresada por Weaver ~1950

Traducción estadística

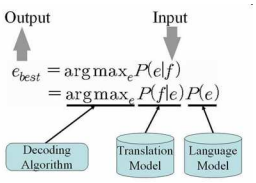
- [Knight 99] Pensemos *f* como la escena de un crimen: queremos descubrir cómo sucedió todo, es decir
  - Quién puede haber tomado la decisión (motivo, personalidad, etc)
  - Cómo es que lo hizo (armas, transportes, etc)
  - Estas cosas podrían contradecirse
    - Podría haber alguien con buen motivo pero sin medios
    - Podría haber alguien con medios pero sin motivos
- Esta idea se modela con ciertas probabilidades → razonamiento bayesiano

Traducción estadística

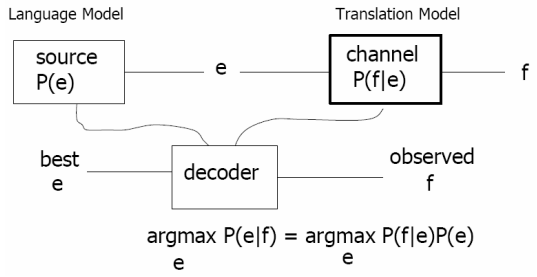
- Probabilidades básicas [Knight 99]
  - P(e): prob. a priori, chance de que ocurra e
    - ej. si e = “I like snakes” → P(e) = prob de que alguien diga *e* en algún momento
  - P(f | e): prob condicional, chance de que pase *f* si pasa *e*.
    - Ej. Con *e* anterior y *f* = “maison bleue” → P(f | e) es la chance de que un traductor genere *f* habiendo visto *e*
  - P(e,f): prob conjunta, chances de pasen e y f a la vez
    - Si e y f son independientes P(e,f) = P(e) * P(f)
    - Si e y f no son independientes P(e,f) = P(e) * P(f | e).
  - Teorema de Bayes P(e|f) = P(e) * P(f | e) / P(f)

Traducción estadística

- Dado *f* la tarea del sistema de TA es encontrar la *e* original
- Ecuación fundamental de la TAE



Traducción estadística



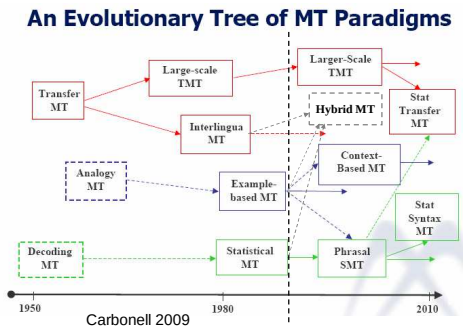
Traducción estadística

- Proceso
  1. Pre-procesar textos (tokens, mayúsculas, etc)
  2. Alineamiento de textos
  3. Entrenar un modelo del LM (P(e))
  4. Entrenar modelo de traducción (P(f|e))
  5. Mejora de parámetros (tuning)
  6. Decodificar (búsqueda de ê)

Análisis del modelo

- Ventajas
  - No necesitamos mucho conocimiento lingüístico
  - Muchas herramientas libres para desarrollo/mejora de sistemas
  - Generaliza mejor que otros modelos (mas robusto para textos OOC)
- Desventajas
  - Necesita grandes cantidades de texto para entrenar modelos
  - No funciona muy bien para idiomas ricos estructuralmente
  - No apropiado para lenguas con escasos recursos disponibles

Que vimos hasta ahora?



Carbonell 2009

Traducción del habla

- La base es la TA en cualquiera de sus sabores (RBMT, EBMT, SMT, etc)
- Pero presenta otros desafíos pre/post-traducción
  - Pre: Reconocimiento del habla (silencio vs habla, segmentación de palabras), factores del entorno como ruidos, tiempo de respuesta si es real-time, procesamiento de frases poco gramaticales….
  - Post: síntesis del habla (text-to-speech), manejo de errores de traducción y/o palabras desconocidas, pronunciación, efectos de personalidad/humor sobre la voz, etc
- Estos problemas están relacionadas a otras ares de investigación → solo veremos algunos ejemplos

Ejemplo 1: OddCast

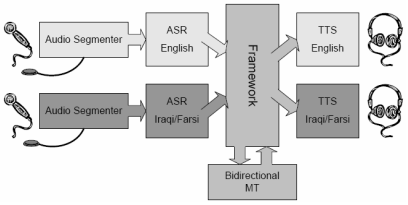
- Empresa OddCast dedicada al marketing web → tecnologías audiovisuales multimodales
  - Son los que hicieron la campaña de Quilmes "hacé confesar a un amigo" http://www.oddcast.com/clients/quilmes/
  - No especifican mucho sobre las tecnologías usadas
- Ofrecen demos de sus productos para:
  - Notar que usan avatars sincronizados con la pronunciación y algunos efectos
  - Traducción de texto a voz
    - ~25 idiomas, varias voces (con/sin dialectos o variaciones)
    - http://www.oddcast.com/home/demos/tts/tts_tran_example.php
  - Texto a voz (TTS)
    - http://www.oddcast.com/demos/tts/tts_example.php?clients
  - Texto a canción (TTSo)
    - http://host-d.oddcast.com/php/application_UI/doorId=373/clientId=1/

Ejemplo 2: Iraqcomm

- Sistema de traducción del habla bidireccional (En→Farsi/arabe iraquí coloquial) desarrollado durante el proyecto TransTac
  - TransTac = Spoken Language Communication and Translation System for Tactical Use → financiado por DARPA
- Objetivo es proveer capacidades de traducción movil para personal en operaciones de campo (militares, de auxilio, de reconstrucción, etc)
- Sistema adaptado para traducir conversaciones espontáneas en temas de seguridad, servicios médicos y municipales,…

Ejemplo 2: Iraqcomm

- Arquitectura del sistema
  - Para Iraqui usan sólo TA estadística
  - Para Farsi usan TA estadística para frases complicadas y una especie de diccionario de frases para las usadas con mas frecuencia



Ejemplo 2: Iraqcomm

- El modulo de segmentación de voz detecta automáticamente silencios y comandos de voz, también puede activarse manualmente
- Comandos disponibles:

Voice Commands	Functions
transtac instructions	play pre-recorded instructions to the foreign language speaker
transtac say again	repeat the last translation
transtac say recognition	say the ASR result of the last utterance
transtac say translation	say the back-translation of the last utterance
transtac automatic mode	switch to automatic mode (hands-free mode)
transtac manual mode	switch to manual mode (push-to-talk)
transtac stand by	turn translation off
transtac listen	turn translation on
transtac status	report system ready and mode information

Ejemplo 2: Iraqcomm

Originalmente este era el setup

Y evolucionó a este otro, concepto "Eyes-free and Hands-free"

Ejemplo 2: Iraqcomm

- Además de las pruebas de laboratorio se realizaron pruebas de campo
- Escenarios con/sin ruido, interacción "dentro/fuera de dominio"
  - Situaciones más reales como checkpoints, uso de vehículos, etc
- Resultados [Weiss 2008]
  - "Detailed results of the evaluations cannot be reported due to restrictions on releasing the data. "
  - "some anonymous results are presented in Sanders et al. (2008) and Condon et al. (2008) "
  - Se valora presentar la metodología y los problemas/soluciones encontrados

Ejemplo 2: Iraqcomm

- Principales desafíos de este tipo de sistemas
  - Usabilidad: personal que puede ser entrenado vs usuarios ocasionales
    - ej pacientes, gente local en alguna situación de emergencia
  - Aspectos culturales: aceptación de usuarios de las tecnologías del sistema
    - Ej todos los usuarios aceptarían usar auriculares?
  - Aspectos técnicos: procesamiento en tiempo real, portabilidad, acceso a recursos
    - ej energía para baterías, reemplazo de partes, etc
  - Recolección de datos para testing: debe ser lo más cercano al uso del sistema
    - Incluyendo nativos de los idiomas/dialectos tratados, condiciones ambientales similares, etc

Ejemplo 3: Mastor

- MASTOR = Multilingual Automatic Speech-to-Speech Translator
  - Desarrollado en IBM parcialmente financiado por DARPA
- Sistema speech-to-speech bidireccional En→ ZH Mandarin en dominios restringidos (viajes, emergencias médicas, diagnósticos, seguridad, …)
  - http://domino.watson.ibm.com/comm/research.nsf/pages/r.uit.innovation.html