PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Curso de traducción automática de lenguas naturales

César Antonio Aguilar

Facultad de Lenguas y Letras

20/11/2015

Cesar.Aguilar72@gmail.com

Síntesis de la clase pasada (1)

En la clase anterior, abordamos una primera aproximación a los conceptos, métodos y algoritmos que están detrás de **Google Translate**, la herramienta más popular que existe hoy en día para hacer traducción automática.

Síntesis de la clase pasada (2)

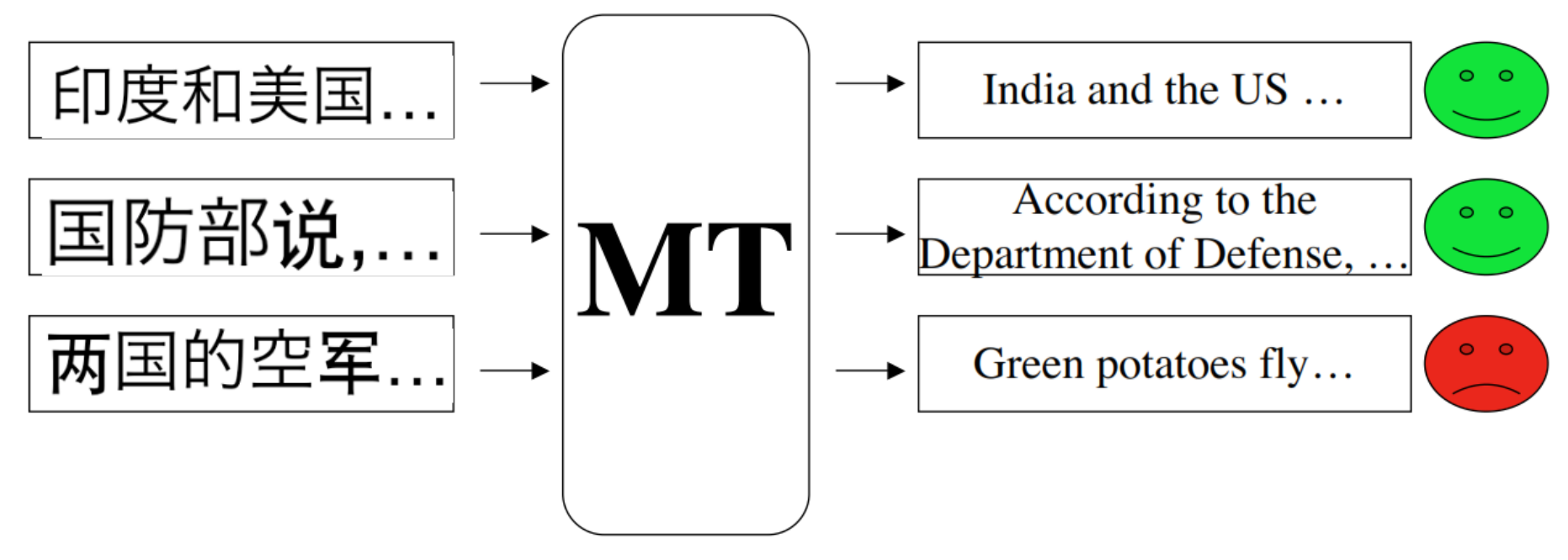
El principal responsable del diseño de Google Translate es **Franz Josef Och**, quien consideró los siguientes criterios en su diseño:

1. Es necesario priorizar los enfoques basados en enfoques estadísticos, dada la gran complejidad que implica hacer una descripción lingüística de cada lengua natural representada en la WEB.
2. Este enfoque estadístico debe aprovechar al máximo los datos brindados por la WEB, de modo que estos sirvan como un **baseline** que ayude a entrenar un sistema de traducción.

Franz Josef Och
(1971)

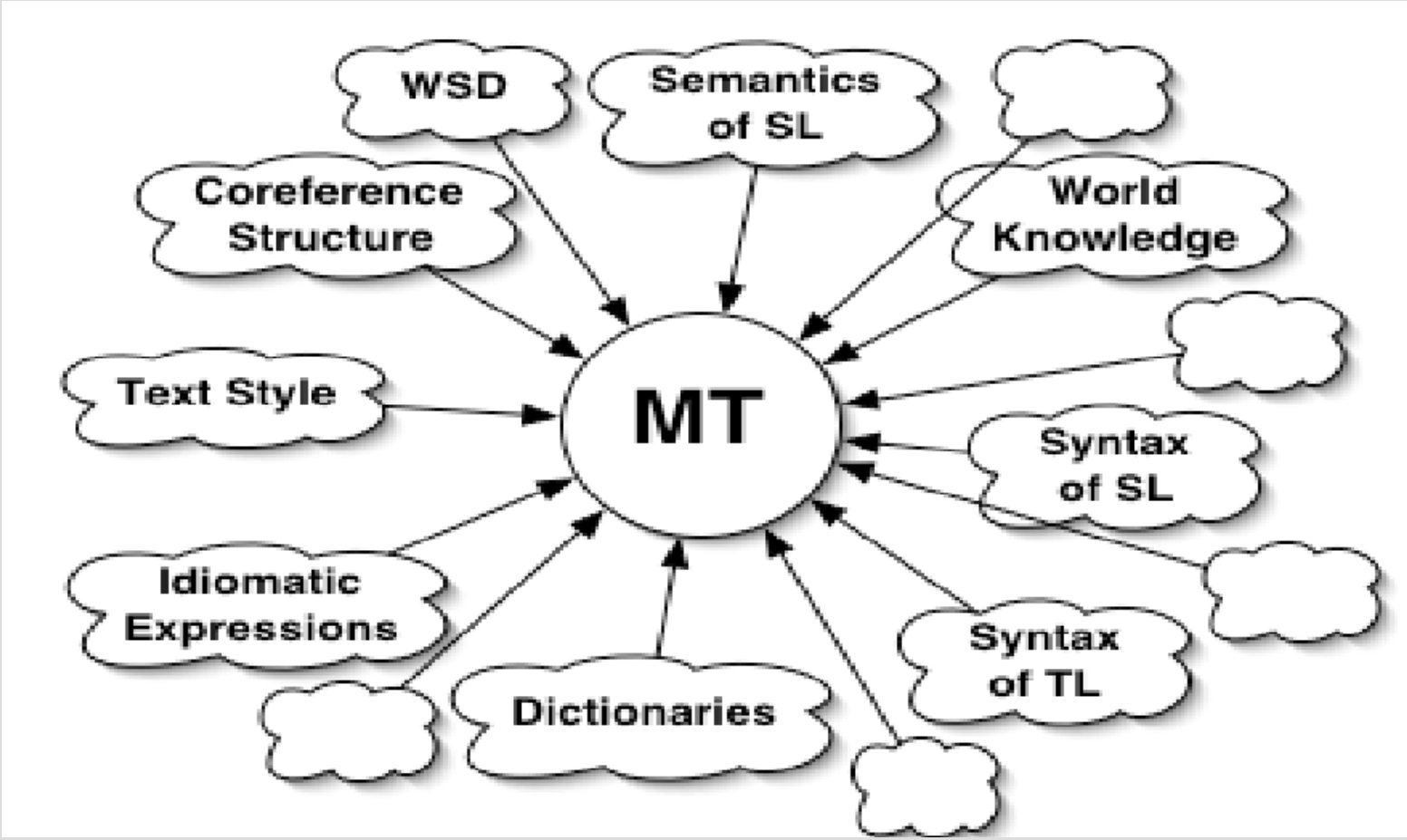
Aspectos generales (1)

Como veíamos en la clase pasada, Google Translate intenta resolver un problema específico en traducción automática: ¿cómo lograr la mejor versión de una lengua a otra, usando exclusivamente recursos computacionales?:



Aspectos generales (2)

Resolver esto, no es fácil, pues implica aumentar muchas veces la complejidad en los sistemas para lograr una traducción aceptable:



Aspectos generales (3)

Para resolver este problema, Och y sus colegas parten de los siguientes principios:

Machine translation is decision problem

- Given source language sentence: decide for the best translation

In machine translation we have to make decisions under uncertainty

- Fundamental uncertainties (e.g. ambiguity)
- Simplified model assumptions (e.g. just trigram dependencies)

Goal is to make good, ideally optimal decisions

- General theory: statistics, statistical decision theory

For machine translation: statistical MT

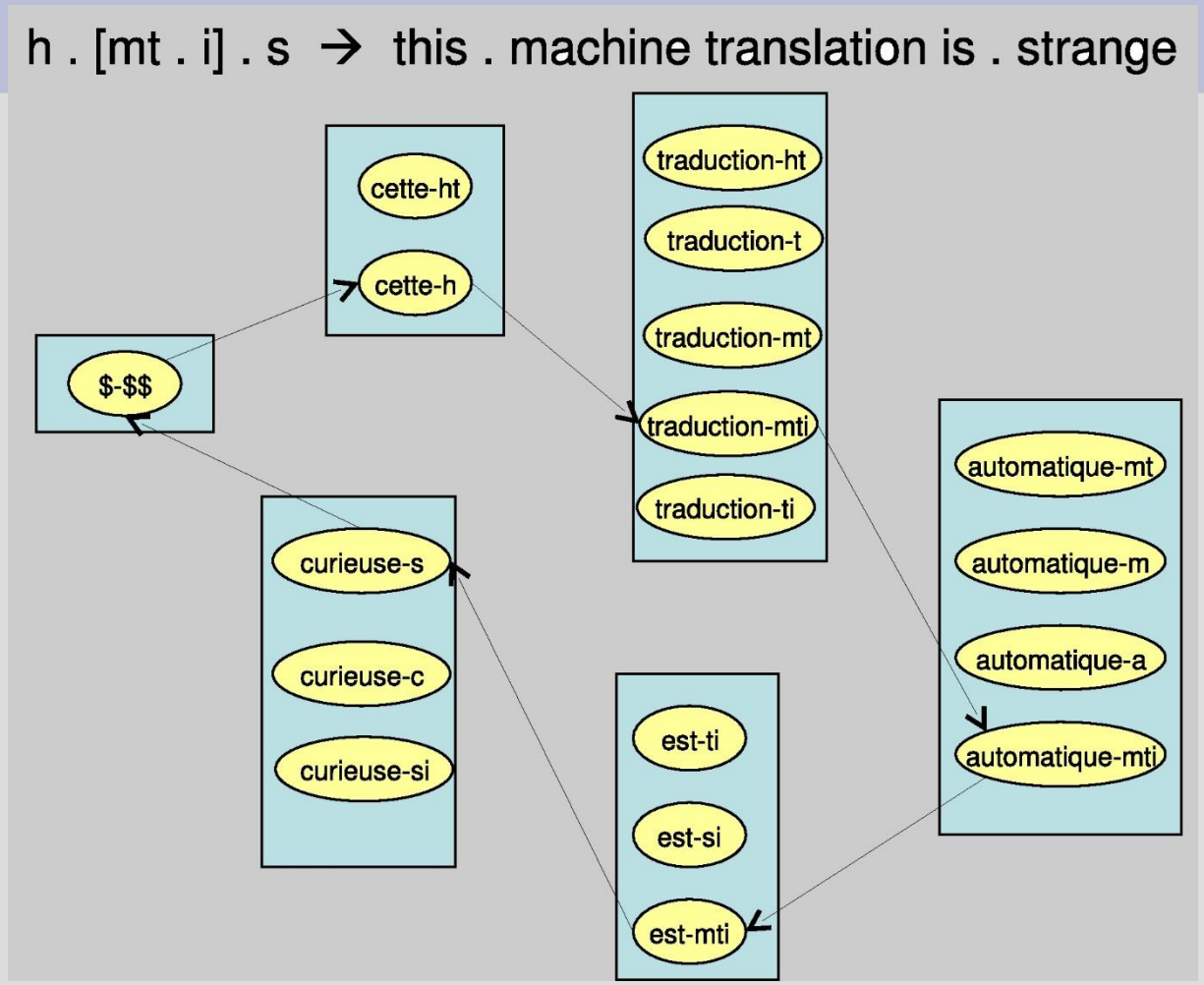
= trying to make good, ideally optimal decisions in MT

Hermann Ney: "statistical approach is expressed by equation

MT = Linguistic Modeling + Statistical Decision Theory"

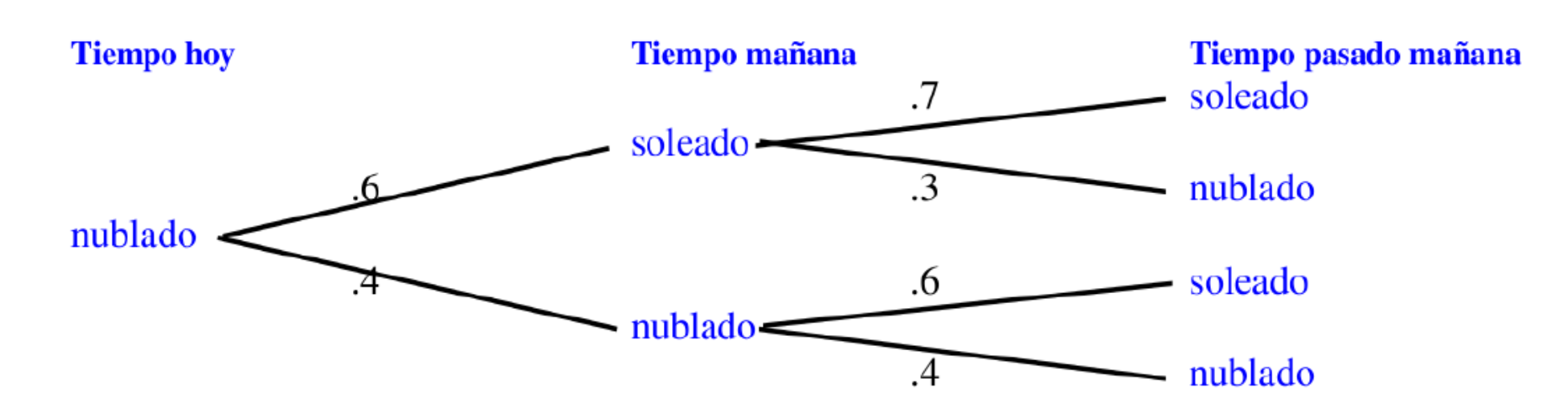
Aspectos generales (4)

En pocas palabras: si vemos a la traducción como un problema de elección, lo que debemos de centrarnos en desarrollar un algoritmo (o varios, si es necesario) que nos permitan calcular probabilísticamente cuál es la mejor versión posible a escoger, dentro de un conjunto de posibilidades.



Aspectos generales (5)

Para poder entender esto mejor, pensemos en el siguiente ejemplo: supongamos que diseñamos un sistema capaz de predecir el tiempo, considerando los datos que tiene a la mano. Claro está, estas predicciones son aproximativas, esto es: únicamente puede dar un pronóstico si cuenta con información suficiente que le indique una tendencia en el clima:



Implementado una traducción *naïve* (1)

¿Cómo relacionamos lo anterior con lo que hemos visto hasta ahora en el curso? Primero, Och y sus colegas no dejan de lado el modelo estadístico de la traducción, pero plantean las siguientes preguntas:

Assumption: we have the true probability distribution Pr(**e**|**f**)

Question: What is the 'best translation' **e** for a specific Foreign **f** sentence?

- What means 'best translation'?
- Some errors are more 'costly' than others

L(**e**,**e'**,**f**): costs (loss) to produce **e'** given that **e** is the correct translation of Foreign sentence **f**?

Reformulated question: What is the translation where our expected loss is minimal?

Implementado una traducción *naïve* (2)

Considerando lo anterior, podemos establecer dos conjuntos: el primero contiene todas las mejores traducciones posibles de generar, y el segundo comprende al menos una (o varias) que son las mínimamente aceptables. Esto lo calculamos con las siguientes fórmulas:

(General) Minimum loss decision rule:

$$\hat{\mathbf{e}}(\mathbf{f}) = \text{argmin}_{\mathbf{e}} \sum_{\mathbf{e}'} L(\mathbf{e}, \mathbf{e}', \mathbf{f}) \cdot Pr(\mathbf{e}|\mathbf{f})$$

Assume: L(**e**,**e'**,**f**)=0 for identical **e** and **e**' and L(**e**,**e'**,**f**)=1 otherwise

- So-called 0-1 loss function
- MAP (maximum a posteriori) decision rule

$$\hat{\mathbf{e}}(\mathbf{f}) = \text{argmax}_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f})$$

Implementado una traducción *naïve* (3)

Una planteadas estas fórmulas, pasamos a resolver tres tareas concretas: (i) hacer un modelo de traducción, (ii) fijar un proceso de entrenamiento para nuestro modelo, y (iii) iniciar una búsqueda de todos aquellos candidatos que se ajusten a nuestro modelo:

Modeling:

- Coming up with a model p(e|f) for Pr(e|f)
- Describing the relevant (linguistic) dependencies

Training (or Parameter Estimation):

- Assigning specific values to model parameters given data
- E.g. EM-algorithm, relative frequency estimate, maximum likelihood

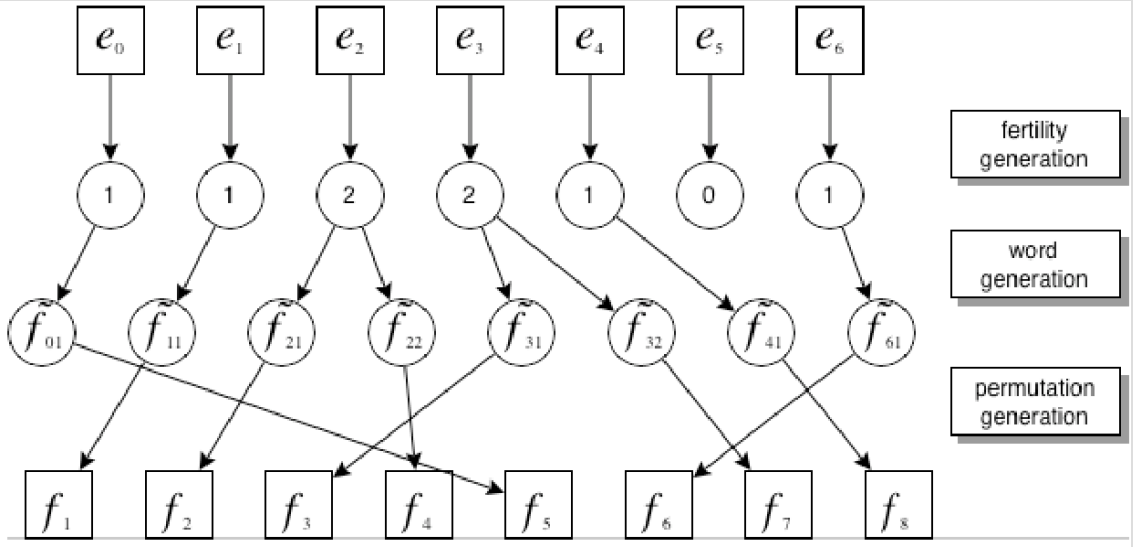
Search (or Decoding):

- Actual translation process
- Find 'best' translation e for input sentence f
- Very large search space: typically heuristic search algorithms

Buscando el menos peor de todos (1)

¿Qué es lo que se busca hacer con esto? Algo muy simple: de todas las traducciones posibles que puedan hacerse de una palabra, una frase o una oración en una lengua extranjera, sabemos que siempre habrá (hipotéticamente) una que sea impecable y perfecta.

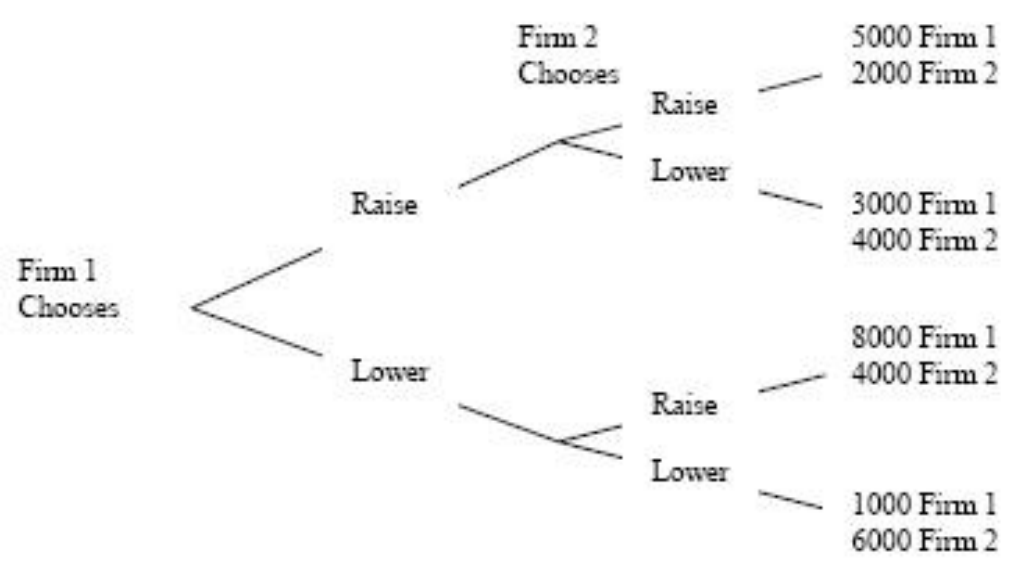
Sin embargo, considerando que demos con una versión no tan buena de lo que queremos traducir (esto es lo más seguro que ocurra), lo mínimo que esperamos es que ésta **no tenga demasiados errores**, de tal suerte que entre la traducción perfecta y la traducción real haya una diferencia mínima.



Buscando el menos peor de todos (2)

Y aquí es en donde las cosas se ponen interesantes: de entrada, sabemos que jamás podremos dar con la mejor traducción posible, de modo que lo único que nos queda es *apostar* por la versión que, si bien no es perfecta, **por lo menos es la que nos provoca menos pérdidas** (en concreto, de fluidez y fidelidad).

Al final, desde el punto de vista de Och y sus colegas, la traducción automática es como jugar al *Monopoly*: si en una jugada te toca perder, digamos que aceptas el castigo *a condición de no perder demasiadas fichas*.



Técnicas de evaluación (1)

¿Cómo se evalúan los resultados de un sistema de traducción automática según Google? La apuesta es muy alta: si una máquina es capaz de traducir de forma muy cercana a la de un humano, es buena…; si no es así, entonces lo que tenemos claro es que la máquina traduce de manera muy diferente a como lo hacemos nosotros. En otras palabras:

Ideal criterion: user satisfaction

Problems:

- Expensive, Slow, Inconsistent, Subjective
- Problematic to use in system development

Goal: automatic + objective evaluation of machine translation quality

- Idea: Compute similarity of MT output with good human translations (=reference translations)
- Hope:
  - If MT output is good: similar to good human translations
  - If MT output is bad: very different from human translations

Técnicas de evaluación (2)

Veamos el siguiente ejemplo: suponiendo que tenemos una oración en japonés que dice: *Japón continúa siendo el mayor socio comercial*, ¿cuáles son las mejores opciones en inglés, contrastando traductores humanos y automáticos?:

Human reference translations:

- Reference 1: japan remained the biggest trading partner
- Reference 2: japan is still the largest trade partner
- Reference 3: japan still remains the number one trade partner

MT output:

- MT 1: **japan** will continue to be **partner** big
- MT 2: **japan is still the biggest trading partner**
- MT 3: ukraine won 2:1 against poland

Técnicas de evaluación (3)

Una manera de ayudar a reconocer automáticamente los mejores candidatos de traducción que se asimilen a los resultados generados por los humanos es usar modelos de *n-gramas*, de acuerdo con el trabajo que reportan varios investigadores de IBM Research, en una ponencia presentada en la Conferencia Anual de la Asociación de Lingüística Computacional (ACL), del año 2002.

En este trabajo presentaron el algoritmo **BLUE** (*Bilingual Evaluation Understudy*), el cual explica cómo se pueden ocupar *n-gramas* para evaluar la calidad de las traducciones automáticas.

IBM Research

¿Cómo trabaja *BLEU*? (1)

Como dijimos, **BLEU** parte de la idea de aprovechar modelos de *n-gramas* para detectar parecidos entre traducciones, tanto humanas como automáticas, de modo que al compararlas establece grados de similaridad entre las mismas. Lo interesante aquí es que **BLEU** es capaz de usar más de un modelo de referencia.

Veamos el siguiente ejemplo: ¿cuál es la versión al inglés con menos errores para la oración en español *el gato está sobre la mesa*, descontando la opción que sirve como referente?

Candidata 1	A	cat	is	on	the	mat
Candidata 2	The	cat	exists	in	the	board
Referencia	The	cat	is	on	the	table

¿Cómo trabaja *BLEU*? (2)

Ahora, recordemos: no somos exigentes, simplemente queremos determinar cuál es la opción que menos errores tiene respecto a nuestro referente. Así, si segmentamos estas oraciones en secuencias de *unigramas,* como aparece en la lámina anterior, ¿cuál es la más cercana al modelo de referencia? Veamos:

En este ejemplo la precisión en 1-gramas sería de $\frac{4}{6}$ para la candidata 1 y de $\frac{3}{6}$ para la 2.

Dicho resultado lo obtenemos aplicando la siguiente fórmula:

$$P = \frac{\text{ngramas comunes}}{\text{ngramas candidata}}$$

¿Cómo trabaja *BLEU*? (3)

¿Cuál es el peor resultado que podemos esperar? Hay por lo menos dos opciones: en una, logramos generar la misma secuencia de n-gramas de nuestro modelo de referencia, pero a la hora de contrastar solo logramos una aproximación mínima, p.e.:

Ejemplo de mala traducción con una alta precisión

Candidata	the	the	the	the	the	the
Referencia	The	cat	is	on	the	table

En breve, de acuerdo con este ejemplo logramos acertar perfectamente con dos unigramas concretos: el artículo que precede al nombre *cat*, y el que precede a *table*. Empero, si bien somos muy precisos, apenas logramos una similitud de 2/6, lo cual es bajo respecto a nuestras opciones previas.

¿Cómo trabaja *BLEU*? (4)

En contraste, la peor de nuestras opciones es:

Ejemplo de mala traducción con alta precisión

Candidata	the	the				
Referencia	The	cat	is	on	the	table

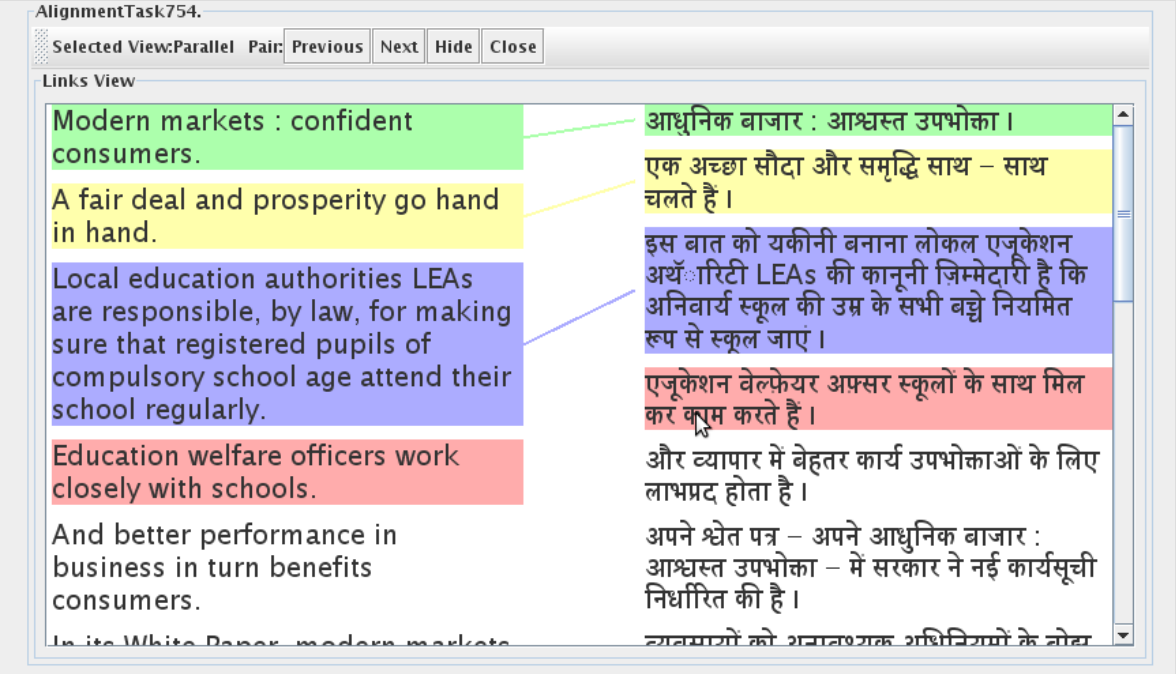
De entrada, tenemos un escenario similar al caso anterior: logramos ubicar dos apariciones del artículo *the* con relación a la oración de referencia…, el problema es que ya no pudimos proponer más unigramas ligados.

Así, aunque tenemos una buena precisión, ésta es de apenas 2/2 (dos aciertos que coinciden con dos unigramas del modelo de referencia). BLEU descartará a este candidato por su brevedad, ya que en comparación con las opciones anteriores, ésta es que comete más errores.

Grandes conjuntos de datos generan mejores resultados (1)

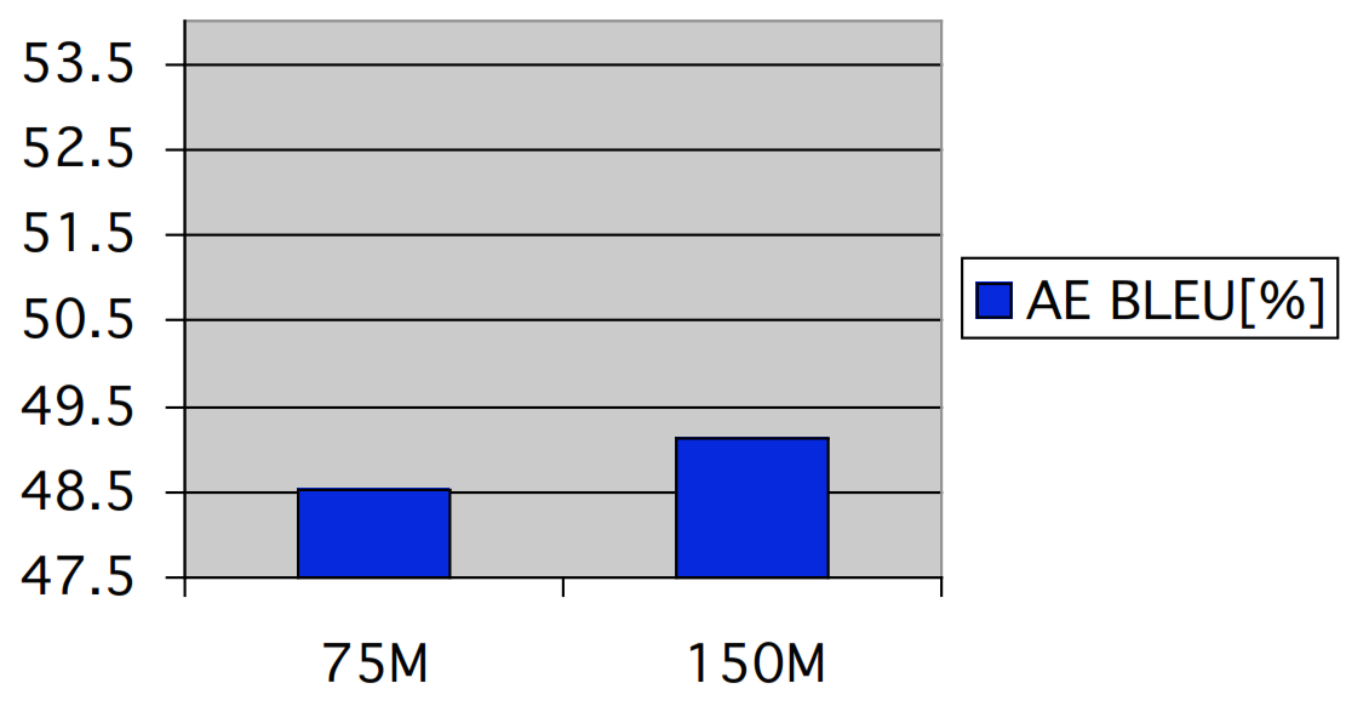
La siguiente etapa es ahora buscar datos que nos permitan entrenar nuestro embrión de Google Translate.

Para iniciar esta etapa, nos conviene trabajar con **corpus paralelos alineados**, de tal suerte que vayamos haciendo inferencias sobre cuál es nuestra mejor opción de traducción teniendo dos modelos de referencia a contrastar.



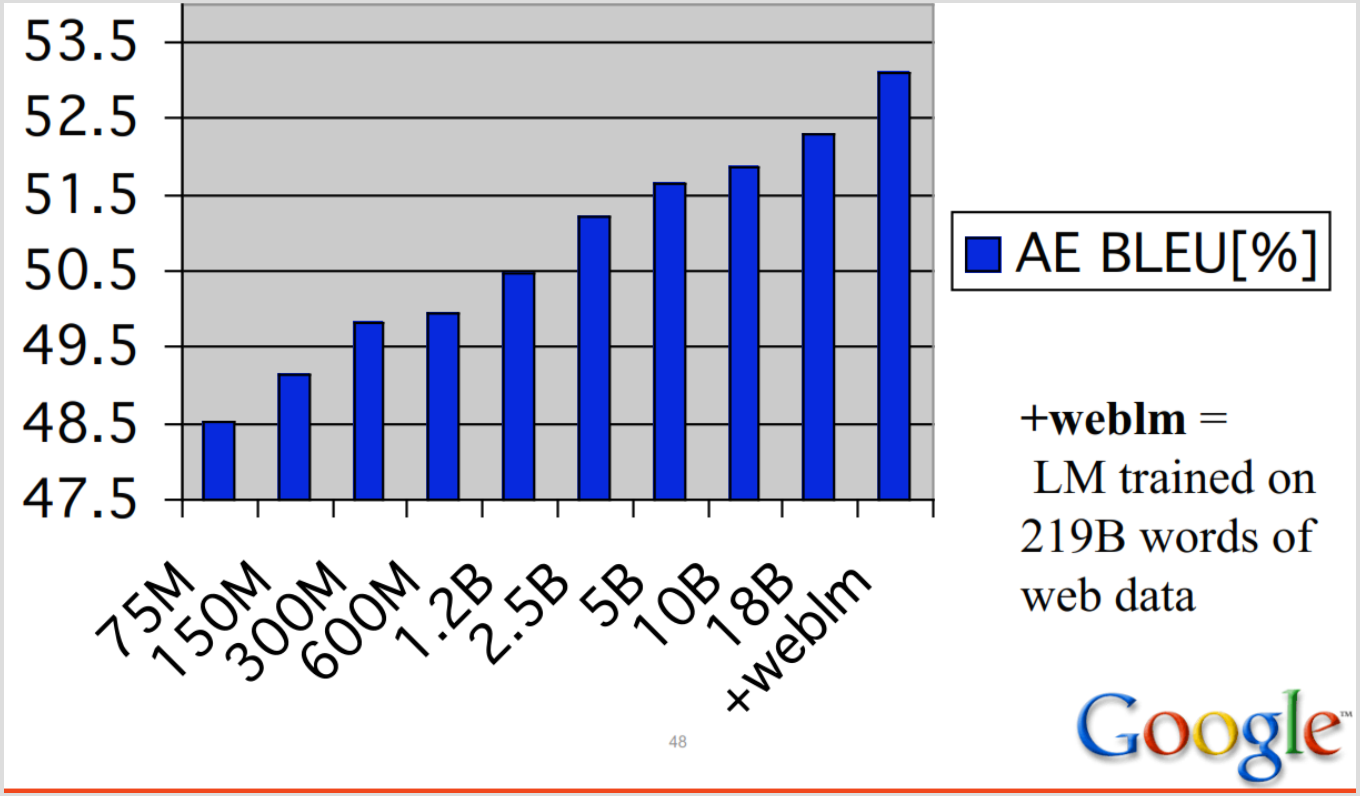
Grandes conjuntos de datos generan mejores resultados (2)

Y es aquí en donde Google lleva una enorme ventaja. Imaginemos lo siguiente: pongamos que queremos desarrollar un sistema de traducción entre inglés y árabe, y tomamos como modelos dos corpus monolingües que van de los 75 a los 150 millones de palabras. Son grandes, claro…, pero nuestro nivel de precisión apenas pasa del 48.5%. La siguiente tabla lo explica mejor:



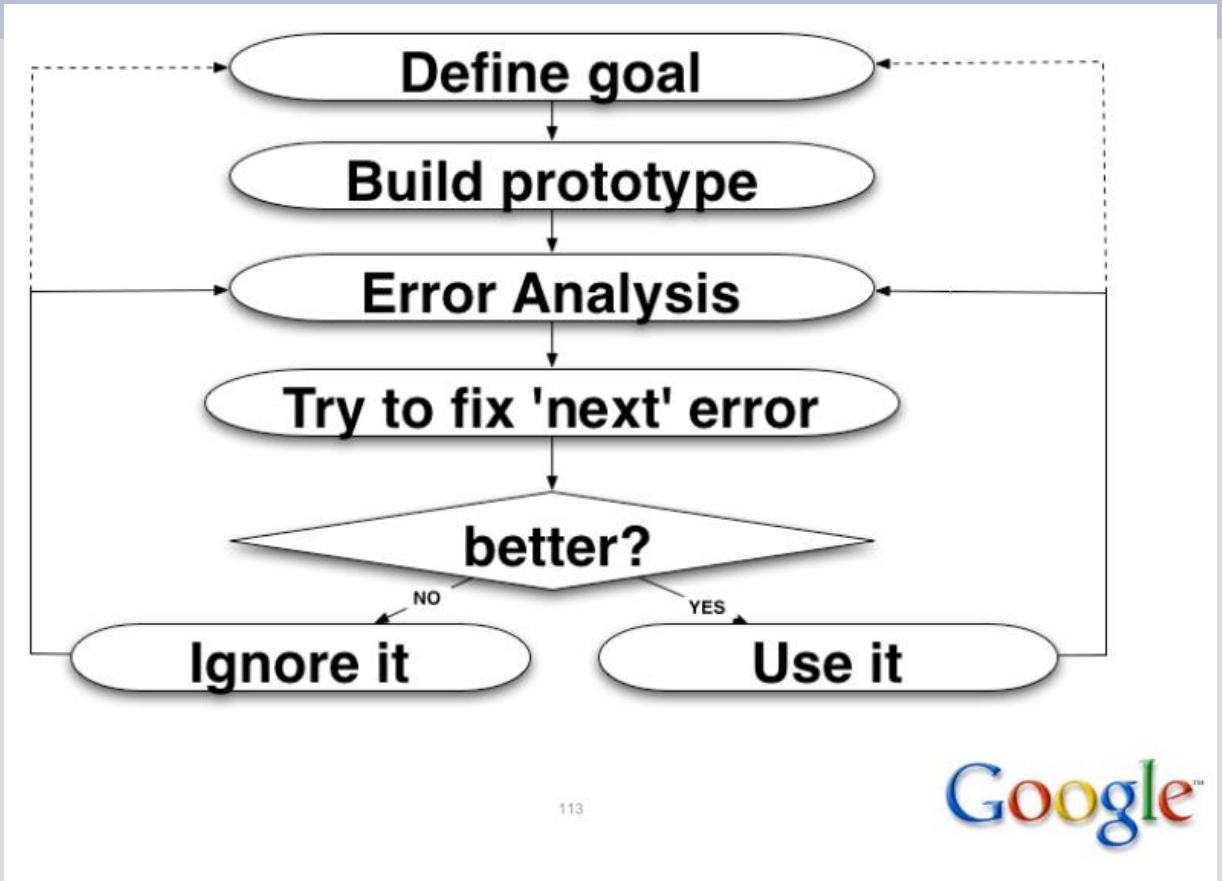
Grandes conjuntos de datos generan mejores resultados (3)

¿Cómo mejorar lo anterior? Parece que Google lo tiene claro: aumentar la cantidad de datos para entrenar al sistema, de tal suerte que lo que conviene es tomar **todos los textos de la WEB posibles** en inglés y árabe como **dos corpus monolingües**. Usando **BLEU**, la mejoría respecto al grado de precisión es más que notoria:



La arquitectura de Google Translate

En conclusión, Google ha logrado integrar diversas propuestas de traducción automática en una arquitectura muy simple, que aprovecha la enorme cantidad de información que le proporciona su buscador para aumentar sus opciones de dar con la traducción idónea, conforme a las necesidades de sus usuarios:



Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/traduccioacuten-automaacutetica.html