Método de Traducción Automática por Traducción Estadística

Francisco Guzmán
CSI - ITESM

18 de abril de 2008

2° CONGRESO NACIONAL DE SISTEMAS Y COMPUTACION
Instituto Tecnológico de Orizaba
Orizaba, Veracruz

CSI CENTRO DE SISTEMAS INTELIGENTES

Outline

1. Introducción
2. Principios de SMT
3. Creando un traductor estadístico
4. Conclusiones SMT
5. Paráfrasis de Traducción
6. Conclusiones

Motivación...

- 4000-5000 diferentes lenguajes en el mundo

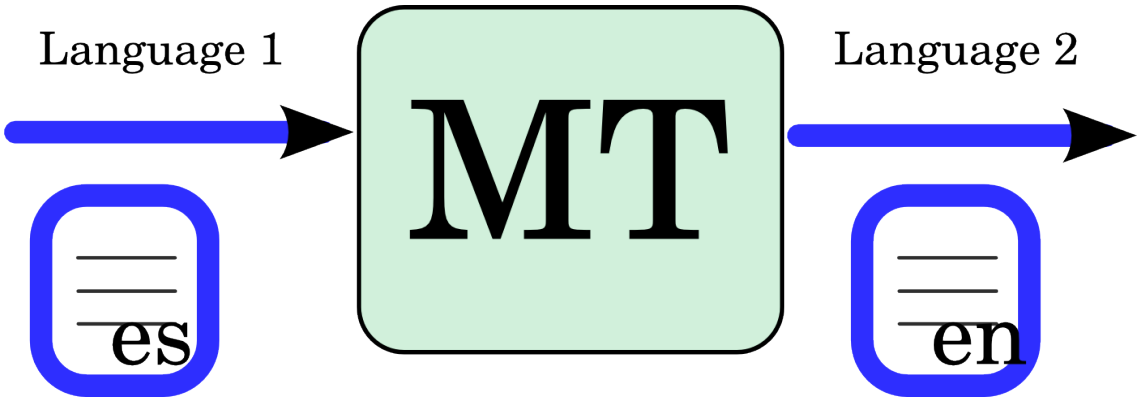
Motivación...

- 4000-5000 diferentes lenguajes en el mundo
- Información en el internet en incremento exponencial

Motivación...

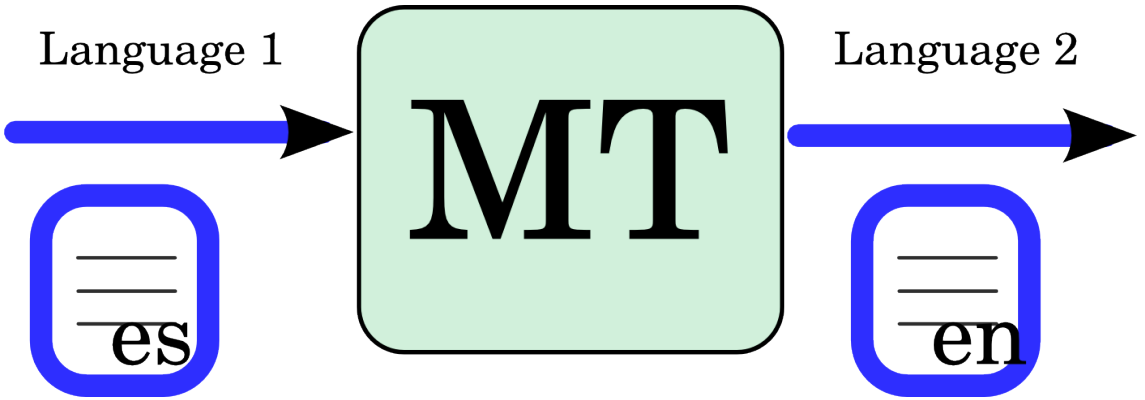
- 4000-5000 diferentes lenguajes en el mundo
- Información en el internet en incremento exponencial
- El acceso a la información está limitado por la barrera del lenguaje.

...para la Traducción Automática...



- El niño juega en el parque.
- Ella comprará un libro de política internacional.
- El agua está clara pero fría.

...para la Traducción Automática...



- El niño juega en el parque.
- Ella comprará un libro de política internacional.
- El agua está clara pero fría.

- The boy plays in the park.
- She will buy a book on international politics.
- The water is clear but cold.

...para la Traducción Automática...

- Es un problema difícil.
- En sus inicios: Traducciones Basadas en Reglas
- Los 90's fueron una etapa fructífera
- Surgimiento de nuevos paradigmas, **empíricos**, con colecciones de textos (corpus) paralelos.
  - Traducciones Basadas en Ejemplos
  - Traducciones por Métodos Estadísticos

...para la Traducción Automática...

- Es un problema difícil.
- En sus inicios: Traducciones Basadas en Reglas
- Los 90's fueron una etapa fructífera
- Surgimiento de nuevos paradigmas, **empíricos**, con colecciones de textos (corpus) paralelos.
  - Traducciones Basadas en Ejemplos
  - Traducciones por Métodos Estadísticos

...Estadística (SMT)

- Por Brown y sus colegas en los 90's.
- Modela el proceso de traducción en terminos de probabilidades
- En su forma tradicional no hace uso de información lingüística.
- No está limitado por los pares de lenguas.

Modelar el lenguaje

e: enunciado en lenguaje objetivo (ej. Inglés)
f: enunciado en lenguaje origen (ej. Español)
$Pr(e|f)$= distribución de probabilidad de que **e** sea la traducción de **f**.

- Problema: no se conoce la verdadera $Pr(e|f)$

Modelar el lenguaje

e: enunciado en lenguaje objetivo (ej. Inglés)
f: enunciado en lenguaje origen (ej. Español)
$Pr(e|f)$= distribución de probabilidad de que **e** sea la traducción de **f**.

- Problema: no se conoce la verdadera $Pr(e|f)$
- Aproximación: se estima un modelo $p(e|f)$

Modelar el lenguaje

e: enunciado en lenguaje objetivo (ej. Inglés)
f: enunciado en lenguaje origen (ej. Español)
$Pr(e|f)$= distribución de probabilidad de que **e** sea la traducción de **f**.

- Problema: no se conoce la verdadera $Pr(e|f)$
- Aproximación: se estima un modelo $p(e|f)$
- Objetivo: obtener una frase $\hat{\mathbf{e}}$ que maximice tal probabilidad:

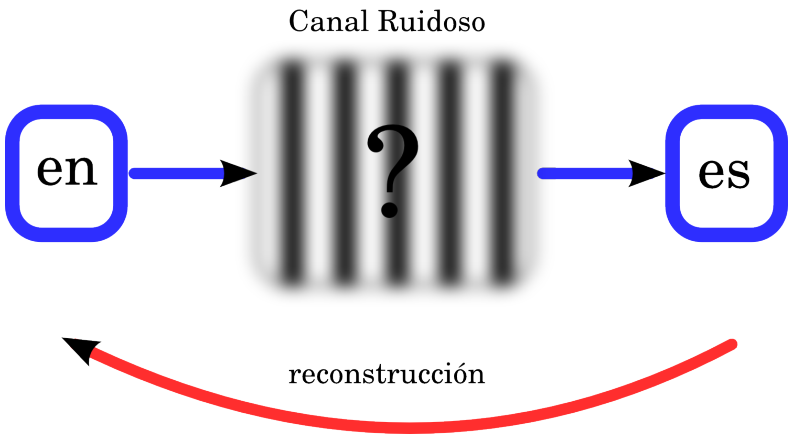
Modelar el lenguaje

e: enunciado en lenguaje objetivo (ej. Inglés)
f: enunciado en lenguaje origen (ej. Español)
$Pr(e|f)$= distribución de probabilidad de que **e** sea la traducción de **f**.

- Problema: no se conoce la verdadera $Pr(e|f)$
- Aproximación: se estima un modelo $p(e|f)$
- Objetivo: obtener una frase **ê** que maximice tal probabilidad:

$$\hat{e}(f) = arg \max_{e} Pr(e|f)$$

¿Canal ruidoso?



- Tenemos que suponer "la historia a la inversa"
- Pensamos que el texto se originó en Inglés
- Algún proceso "ruidoso" lo transformó en español
- E intentaremos "reconstruirlo" al inglés con la evidencia que tenemos.

¿Canal ruidoso?

$$
\begin{aligned}
\hat{e}(f) &= \arg\max_{e} Pr(e|f) \\
&= \arg\max_{e} \frac{Pr(f|e)Pr(e)}{Pr(f)} \\
&= \arg\max_{e} Pr(f|e)Pr(e) \\
&\approx \arg\max_{e} p(f|e)p(e)
\end{aligned}
$$

- Tenemos que suponer "la historia a la inversa"
- Pensamos que el texto se originó en Inglés
- Algún proceso "ruidoso" lo transformó en español
- E intentaremos "reconstruirlo" al inglés con la evidencia que tenemos.

Modelos

$$\hat{e} = arg\ \max_{e} p(f|e)p(e)$$

modelo de lenguaje $p(e)$

- Nos da una probabilidad de qué tan "común" es cierta construcción en el lenguaje objetivo.
- Esto nos asegura cierta validez gramatical.
- p(Voy Orizaba conferencia a unas mañana) => muy poco probable
- p(Mañana voy a Orizaba a una conferencia) => mucho mejor

Modelos

$$\hat{e} = arg\ \max_{e} p(f|e)p(e)$$

modelo de traducción $p(f|e)$

- Es la probabilidad de que una frase se traduzca en otra
- Es tan complejo como estimar $p(e|f)$
- Se basa en alineaciones
- p(la casa es azul | the martians are going to attack!!)=> poco probable
- p(la casa es azul | the house is blue)=> mejor

Modelos

$$\hat{e} = arg\ \max_{e} p(f|e)p(e)$$

modelo de traducción $p(f|e)$

- Es la probabilidad de que una frase se traduzca en otra
- Es tan complejo como estimar $p(e|f)$
- Se basa en alineaciones
- p(la casa es azul | the martians are going to attack!!)=> poco probable
- p(la casa es azul | the house is blue)=> mejor

Alineaciones

- Primeros modelos basados en palabras

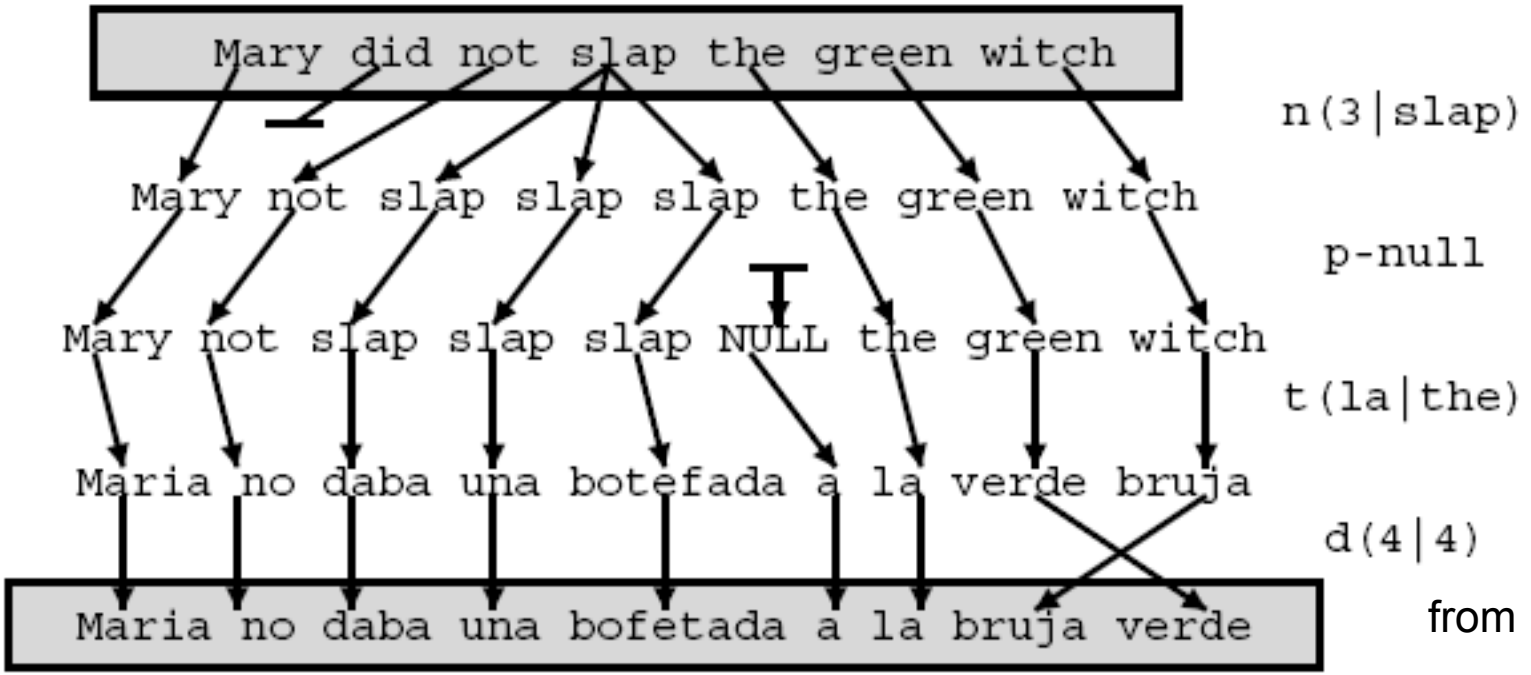
Alineaciones

- Primeros modelos basados en palabras
- Modelos más recientes basados en “frases”

Alineaciones

- Primeros modelos basados en palabras
- Modelos más recientes basados en “frases”
- Phrase-Based Statistical Machine Translation

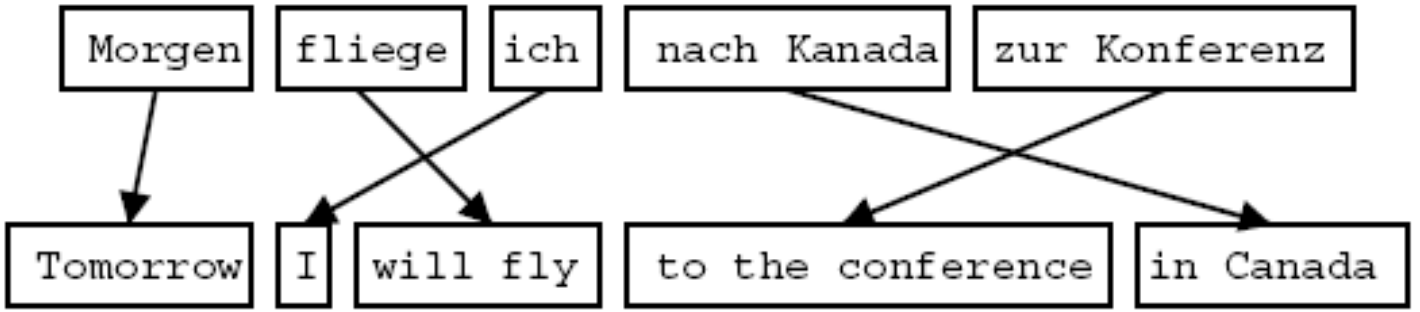
Modelos de alineación palabra-palabra



from Koehn et al.

El proceso de traducción se descompone en varios pasos
Modelos originales de Brown (1993) (IBM-1, IBM-2 ... IBM-5)

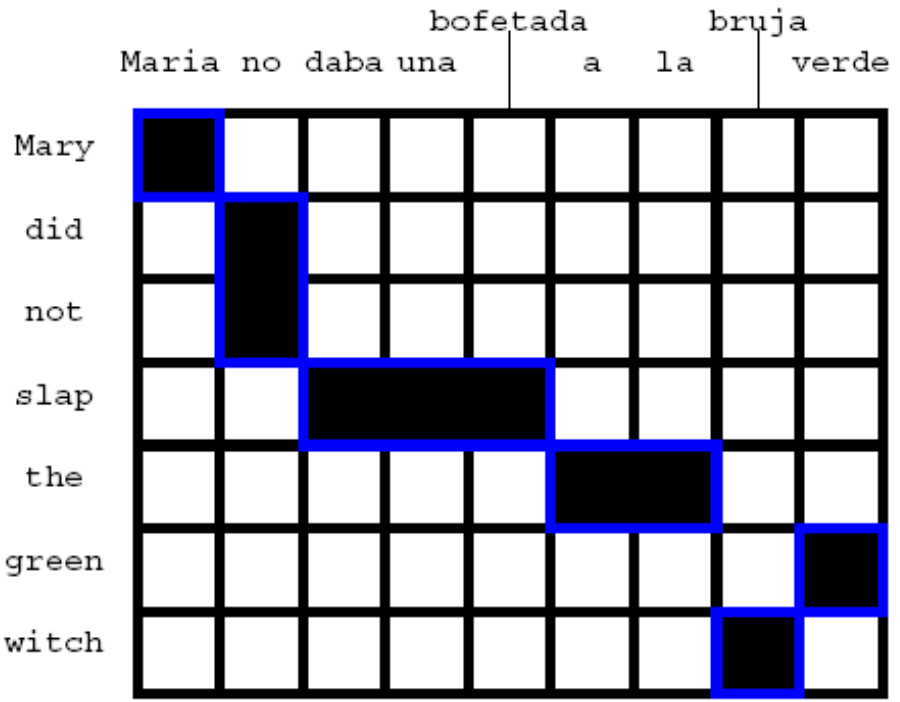
Un paso adelante... frases



from Koehn et al.

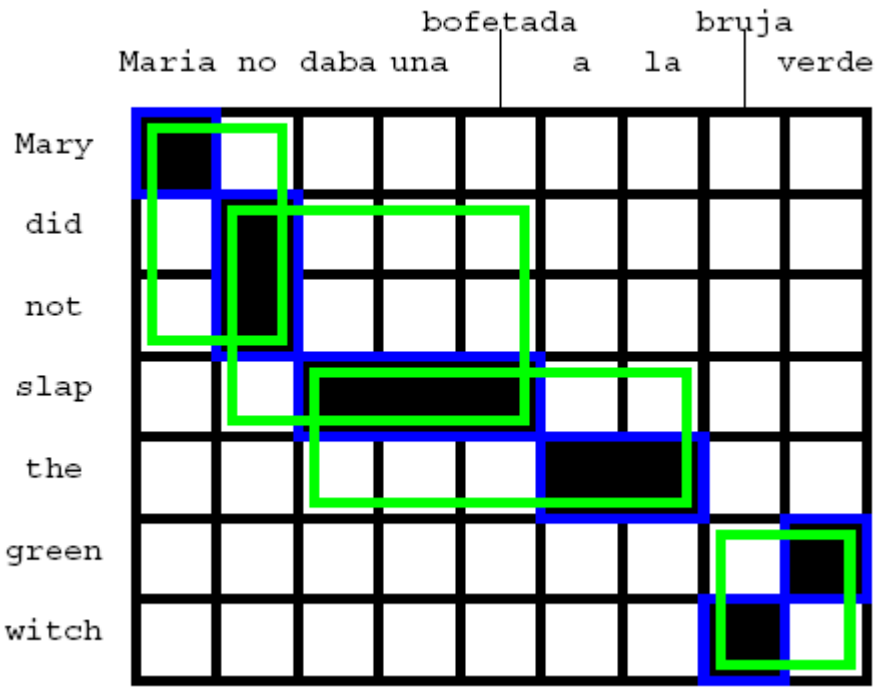
- La entrada a traducir se segmenta en secuencias de palabras que se conocen como "frases".
- Esta denominación no tiene una motivación lingüística.
- Las frases se reordenan.
- Se usan heurísticas para determinar las alineaciones, basadas en las alineaciones palabra-palabra.
- Las probabilidades obtenidas se guardan en "tablas de frases".

Ejemplo



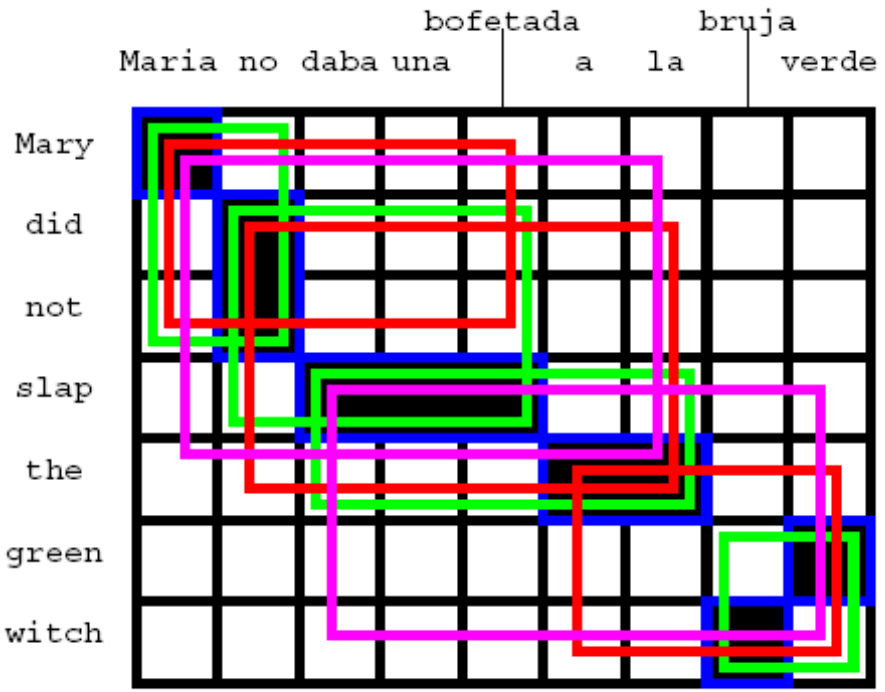
from Koehn et al.

Ejemplo



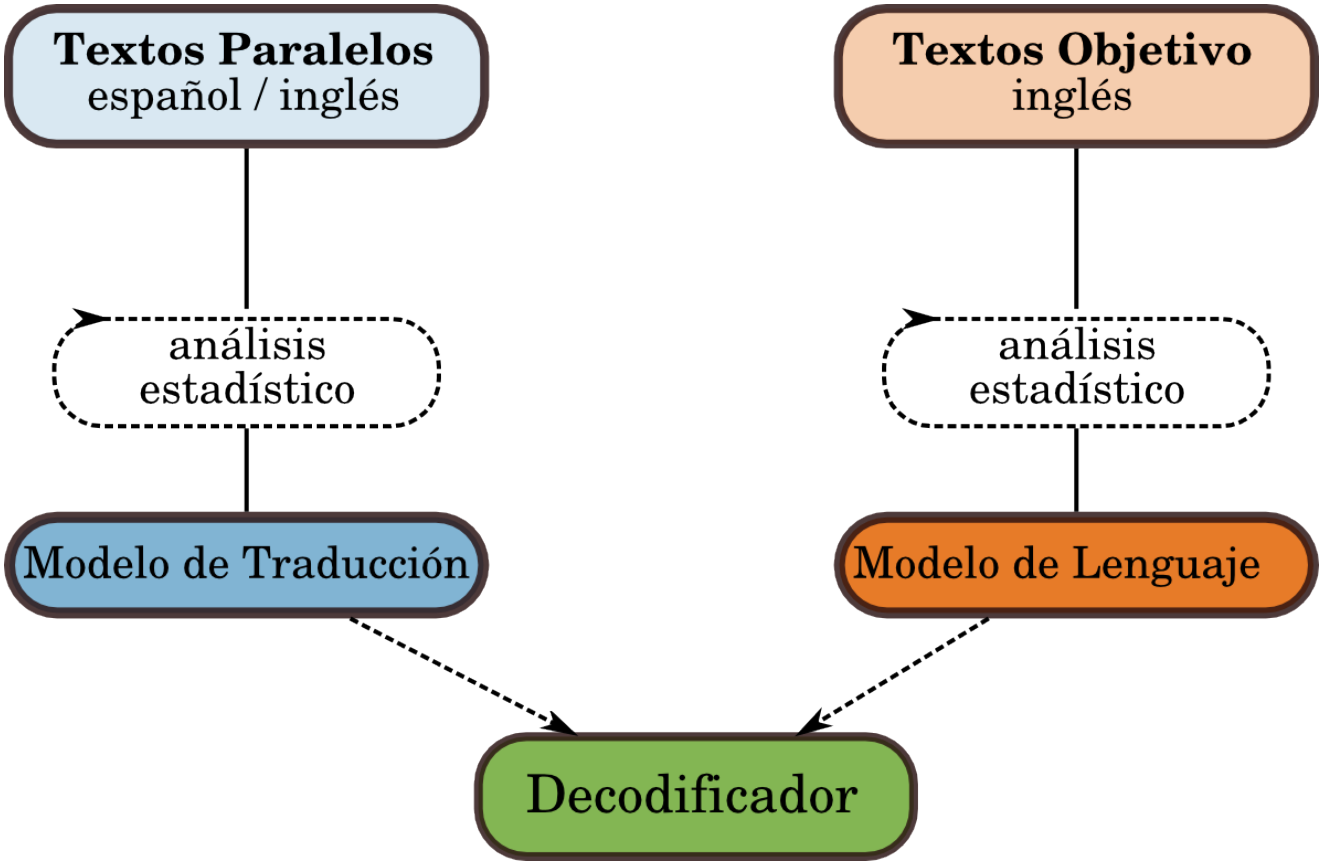
from Koehn et al.

Ejemplo



from Koehn et al.

Proceso de Traducción



Decodificación

- Existen muchas posibles formas de segmentar la frase de entrada.
- Existen muchas posibles traducciones.
- Se van creando "hipótesis" de traducción
- Se trata de un problema de búsqueda entre la "hipótesis" que maximice el resultado.
- Explosión exponencial del espacio de búsqueda.
- Se ha comprobado que es NP-Difícil.
- Se han creado diferentes heurísticas para atacar este problema

¿Cómo crear un traductor?

Obtención de datos

- Se necesitan GRANDES cantidades de texto para entrenar...

¿Cómo crear un traductor?

Obtención de datos

Preprocesamiento

- Se limpian los corpus para eliminar datos “no deseados”.

¿Cómo crear un traductor?

Obtención de datos

Preprocesamiento

Entrenamiento

- Se entrena el modelo de lenguaje usando SRILM.
- Se entrena el modelo de traducción usando GIZA++.
- Se obtienen tablas de frases.

¿Cómo crear un traductor?

- Obtención de datos
- Preprocesamiento
- Entrenamiento
- Decodificacion

- Se hace entrenamiento de parámetros (MERT) del decodificador.
- Se usan las tablas de frases para decodificar.

¿Cómo crear un traductor?

- Obtención de datos
- Preprocesamiento
- Entrenamiento
- Decodificacion
- Evaluación

- Se utilizan métricas automáticas (BLEU)

Disponibilidad

Textos Paralelos:

- Europarl: 30 millones de palabras en 11 lenguajes http://www.statmt.org/europarl/
- Acquis Communitaire: 8-50 millones de palabras en 20 lenguajes de la UE.
- Muchos más...

Textos Monolingües (para modelado de lenguaje)

- 2800 millones de palabras en Inglés del LDC
- Cientos de millones en la web

Ejemplos:

Europarl

1 adoption of the minutes of the previous sitting	aprobación del acta de la sesión anterior
2 the minutes of yesterday's sitting have been distributed .	el acta de la sesión anterior ha sido distribuida .
3 are there any comments ?	¿ hay alguna objeción ?

Evaluación

- ¿Por qué se hace una evaluación automática?
- Evaluación manual es muy lenta (y costosa)
- Evaluación automática por excelencia: BLEU
- Se trata de comparar una traducción hecha por un sistema, contra una traducción de referencia...
- Se ha comprobado que BLEU tiene alta correlación con las evaluaciones de los jueces humanos.

La batalla no está ganada

- La SMT ha probado ser MUY efectiva
- Permite traducir de A $\rightarrow$ B indiscriminadamente

La batalla no está ganada

- La SMT ha probado ser MUY efectiva
- Permite traducir de A $\rightarrow$ B indiscriminadamente

pero...

La batalla no está ganada

- La SMT ha probado ser MUY efectiva
- Permite traducir de A $\rightarrow$ B indiscriminadamente

pero...

- Está limitada por el **entrenamiento** y la disponibilidad de recursos
- Esto es un problema cuando tratamos de traducir entre lenguajes "poco densos"

¿Qué tal si...

¿Qué tal si...

Pudiéramos aprender de un lenguaje, mirando a otros lenguajes?

¿Qué tal si...

Pudiéramos aprender de un lenguaje, mirando a otros lenguajes?

- Se extraen paráfrasis mediante la traducción de A→B
- Se usan esas paráfrasis para PB-SMT (Callison-Burch) como una variable más

¿Qué tal si...

Pudiéramos aprender de un lenguaje, mirando a otros lenguajes?

- Se extraen paráfrasis mediante la traducción de A→B
- Se usan esas paráfrasis para PB-SMT (Callison-Burch) como una variable más

Pudiéramos aprender cómo traducir de A a B, mirando a C

¿Qué tal si...

Pudiéramos aprender de un lenguaje, mirando a otros lenguajes?

- Se extraen paráfrasis mediante la traducción de A→B
- Se usan esas paráfrasis para PB-SMT (Callison-Burch) como una variable más

Pudiéramos aprender cómo traducir de A a B, mirando a C

- Traducciones indirectas (Eng→Esp→Cat) (Gispert)
- Traducción de una parte del corpus.

Paráfrasis de traducción

- **Paráfrasis** son diferentes frases con la misma carga semántica. (mismo significado)

Paráfrasis de traducción

- **Paráfrasis** son diferentes frases con la misma carga semántica. (mismo significado)

The child wants a candy

Paráfrasis de traducción

- **Paráfrasis** son diferentes frases con la misma carga semántica. (mismo significado)

The child wants a candy

⇕

The infant desires a sweet

Paráfrasis de traducción

- **Paráfrasis** son diferentes frases con la misma carga semántica. (mismo significado)

The child wants a candy

⇕

The infant desires a sweet

⇕

El niño quiere un dulce

Paráfrasis de traducción

- **Paráfrasis** son diferentes frases con la misma carga semántica. (mismo significado)
- **Paráfrasis de Traducción** diferentes pares de frases con la misma carga semántica

Paráfrasis de traducción

- **Paráfrasis** son diferentes frases con la misma carga semántica. (mismo significado)
- **Paráfrasis de Traducción** diferentes pares de frases con la misma carga semántica

el niño quiere un dulce $\leftrightarrow$ *the child wants a candy*

Paráfrasis de traducción

- **Paráfrasis** son diferentes frases con la misma carga semántica. (mismo significado)
- **Paráfrasis de Traducción** diferentes pares de frases con la misma carga semántica

el niño quiere un dulce $\leftrightarrow$ *the child wants a candy*

$\Updownarrow$

l'enfant veut un bombon $\leftrightarrow$*the child desires a sweet*

... en acción ...

spanish- french

El niño quiere un dulce
L'enfant veut un bombon

+

french-english

L'enfant veut un bombon
The child wants a candy

⇓

spanish-english

El niño quiere un dulce
The child wants a candy

- Se usa un lenguaje intermediario para obtener las PTs.
- Se combinan dos pares de traducción para obtener uno nuevo.
- Las probabilidades se calculan de la siguiente manera:

$$p(es|en) = \sum_{fr} p(es|fr)p(fr|en)$$

... en acción ...

spanish- french

El niño quiere un dulce
L'enfant veut un bombon

+

french-english

L'enfant veut un bombon
The child wants a candy

⇓

spanish-english

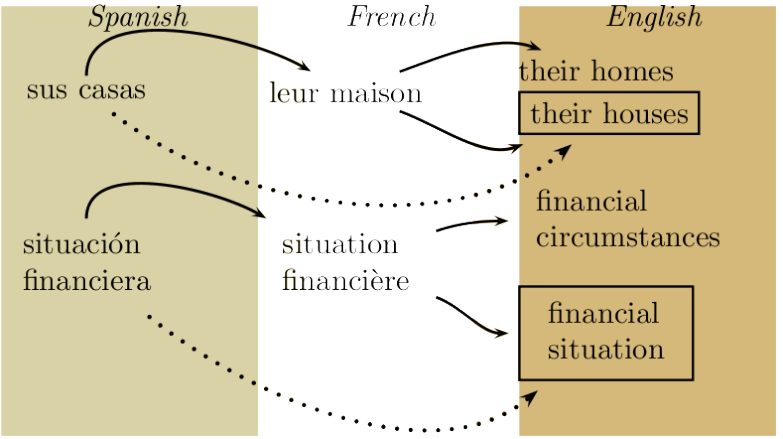
El niño quiere un dulce
The child wants a candy

- Se usa un lenguaje intermediario para obtener las PTs.
- Se combinan dos pares de traducción para obtener uno nuevo.
- Las probabilidades se calculan de la siguiente manera:

$$p(es|en) = \sum_{fr} p(es|fr)p(fr|en)$$

... para enriquecer las tablas de frases

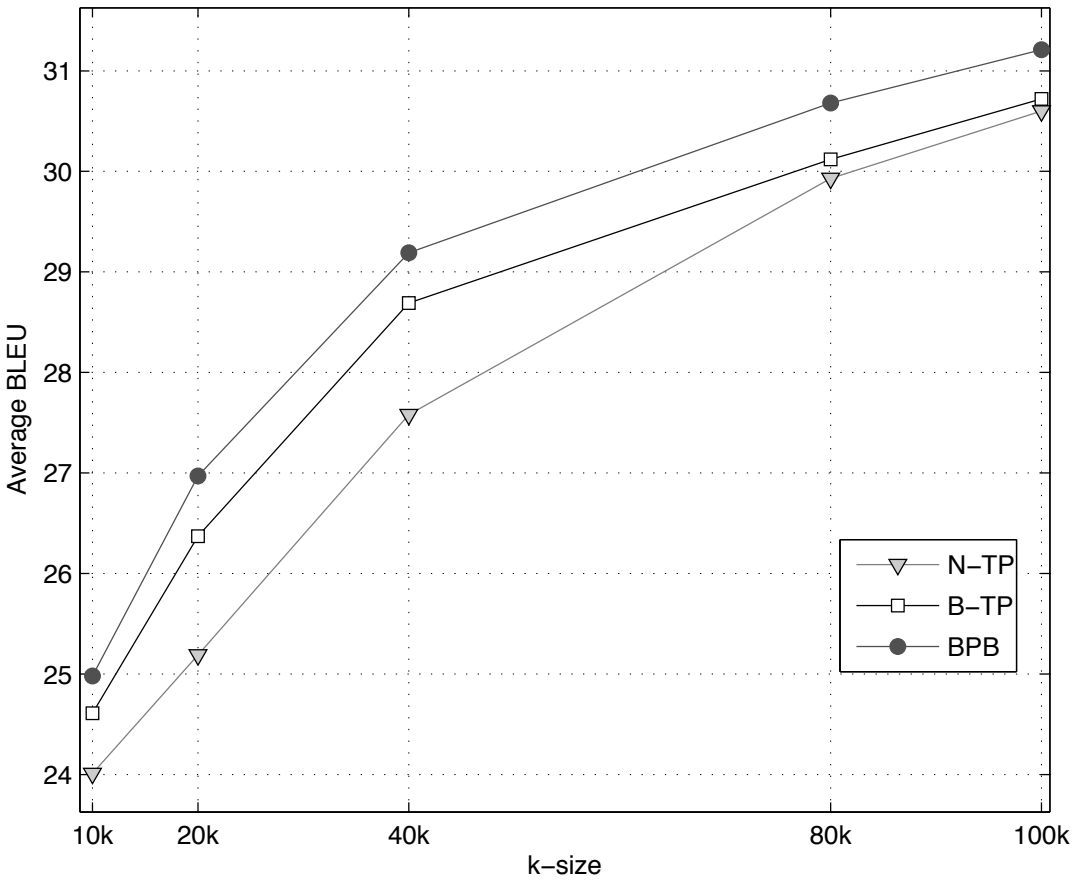
- Se obtienen por medio de un lenguaje intermediario
- Nos ayudan a obtener interpretaciones más flexibles



Question

¿Qué tanto se puede mejorar la traducción usando las PTs?

Resultados de Comparaciones



Conclusiones

- Mientras se incrementa el tamaño de los corpus de entrenamiento, las TPs tienen menor impacto
- En casos donde hay poco entrenamiento, la mejora de la calidad es SIGNIFICATIVA
- Las TPs por sí solas, no dan buenos resultados.

¿Preguntas?

Centro de Sistemas Inteligentes

Si están interesados en la Traducción Automática o la Inteligencia Artificial en general:

- Maestrías y Doctorados en ITESM Campus Monterrey:
- http://sistemas-inteligentes.mty.itesm.mx

CSI CENTRO DE SISTEMAS INTELIGENTES