Introducción al Procesamiento de Lenguaje Natural

Grupo de PLN – InCo

Traducción Automática

Agenda

- Introducción
- Traducción Automática Basada en Reglas
- Corpus Paralelos
- Traducción Automática Estadística
- Traducción Automática Neuronal
- Evaluación

Introducción

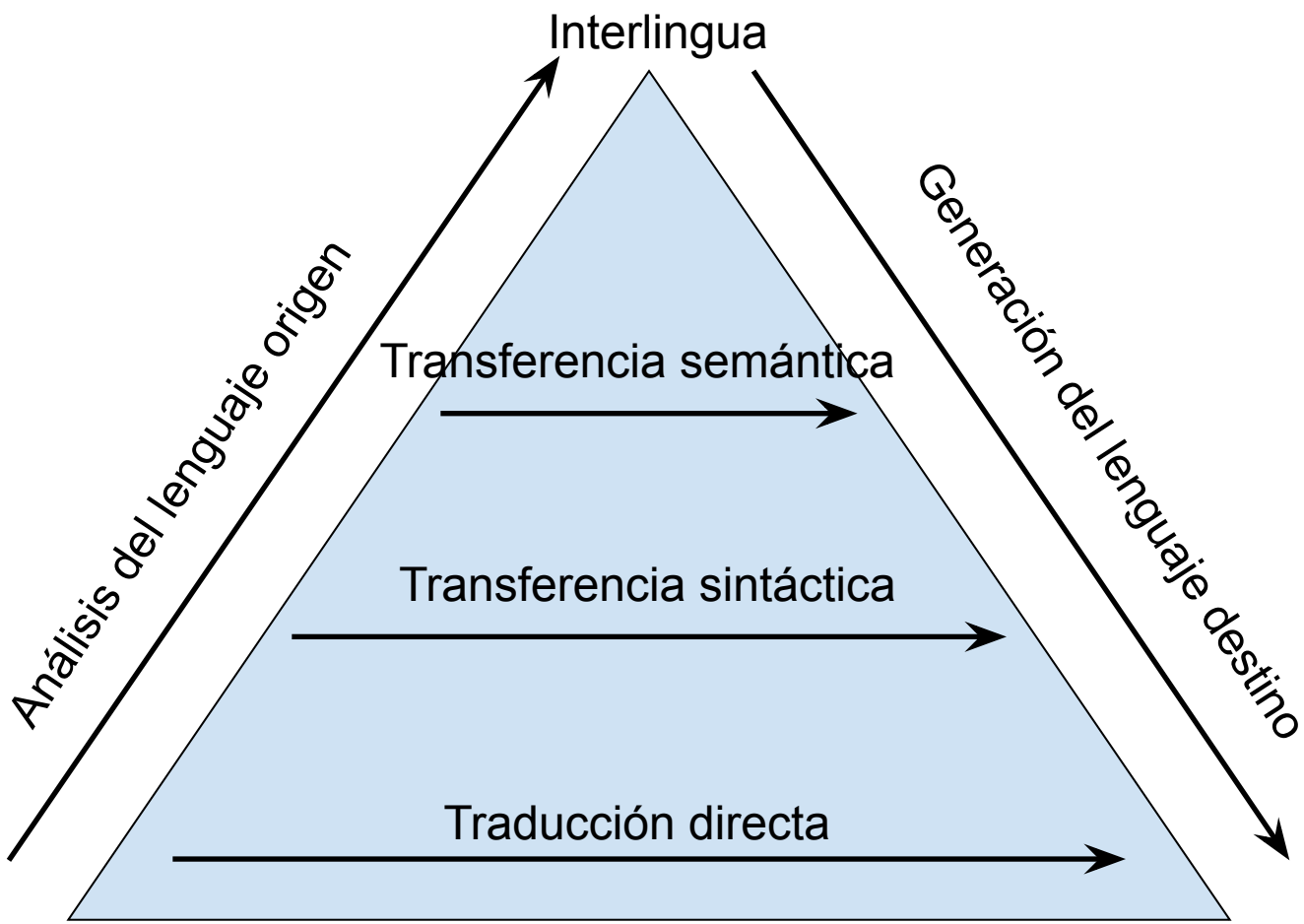
Uno de los primeros problemas de PLN

¿Por qué es difícil?

- Tipologías lingüísticas: SVO vs SOV
- Divergencia léxica: pata vs pierna vs leg
- Diferencias morfológicas: aglutinante vs fusional
- Densidad referencial: sujetos omitidos

Traducción Automática Basada en Reglas

Triángulo de Vauquois

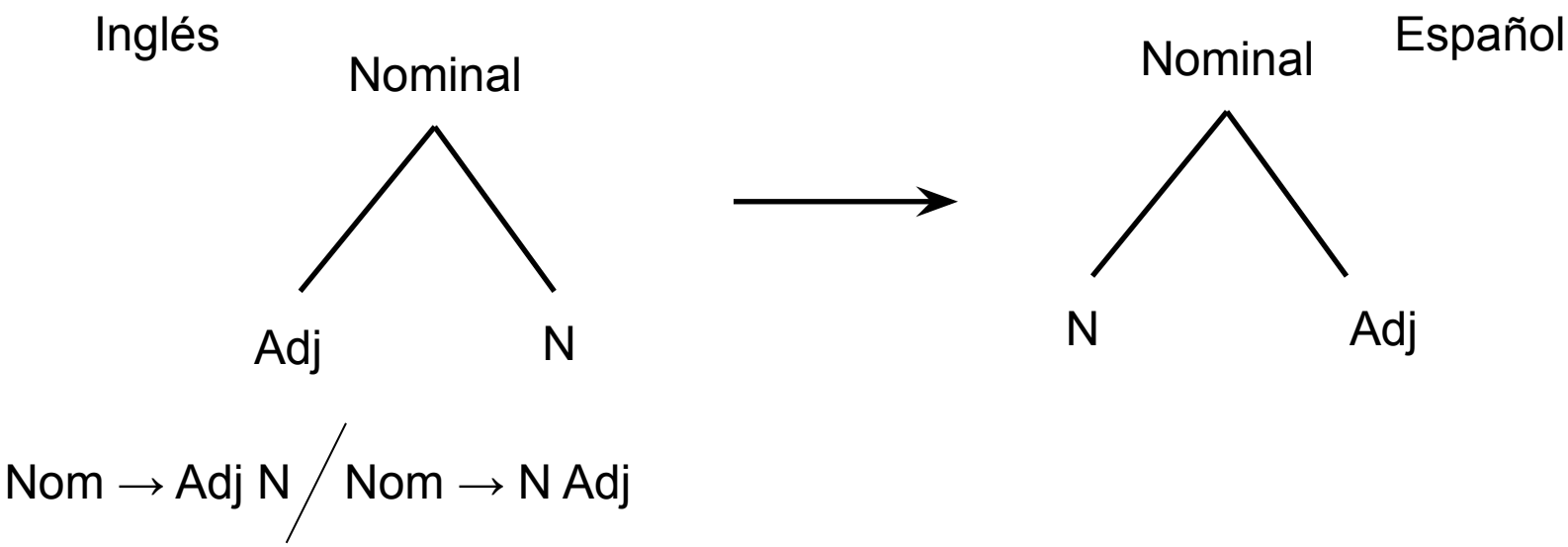


Transferencia Sintáctica

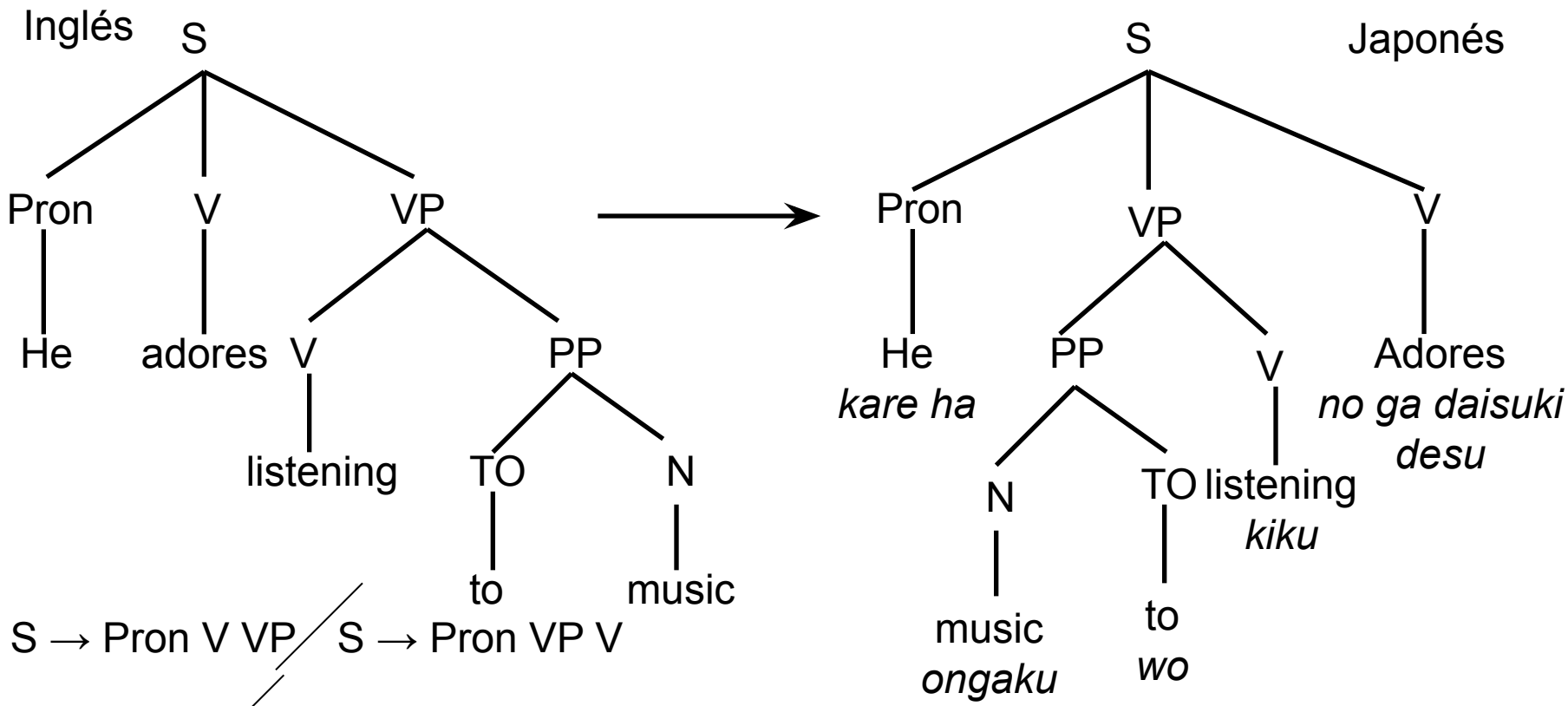
Parsing del lenguaje origen

Generación en en lenguaje destino

Reglas de transferencia entre árboles y subárboles



Transferencia Sintáctica



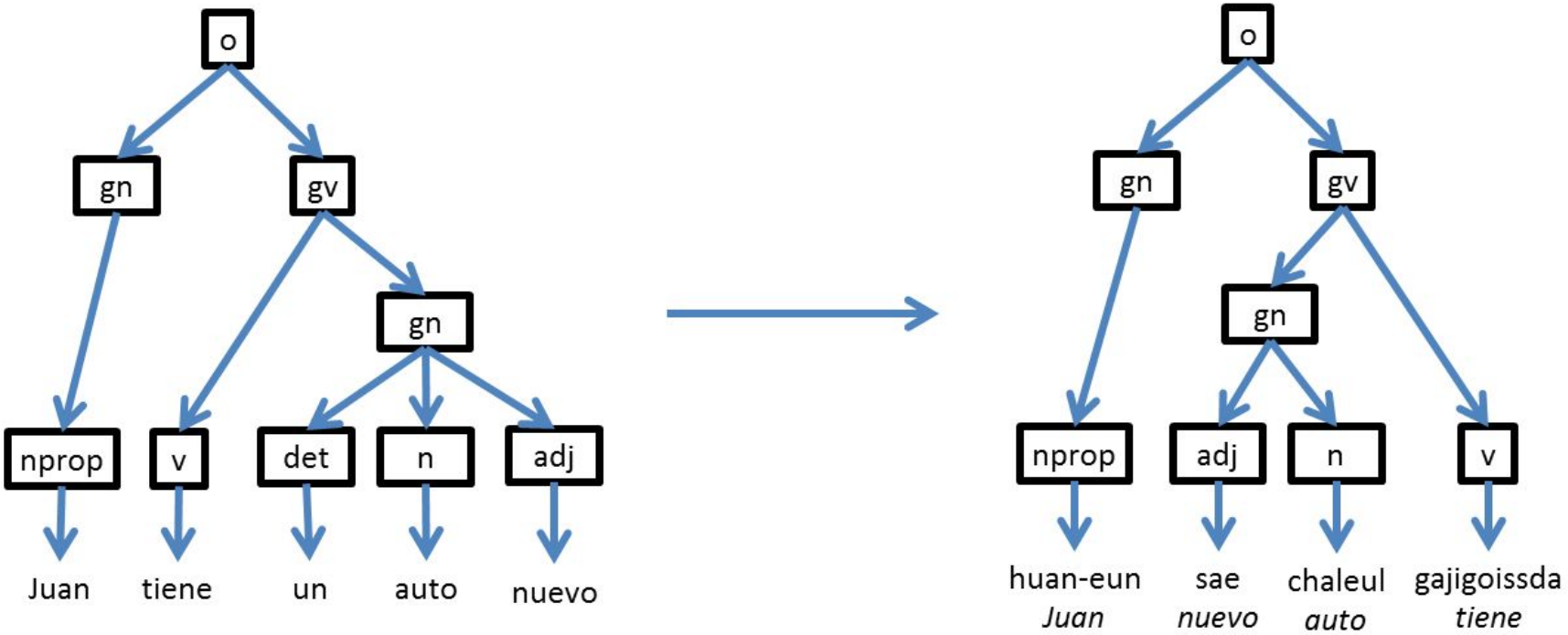
S → Pron V VP / S → Pron VP V

VP → V PP / VP → PP V

PP → TO N / PP → N TO

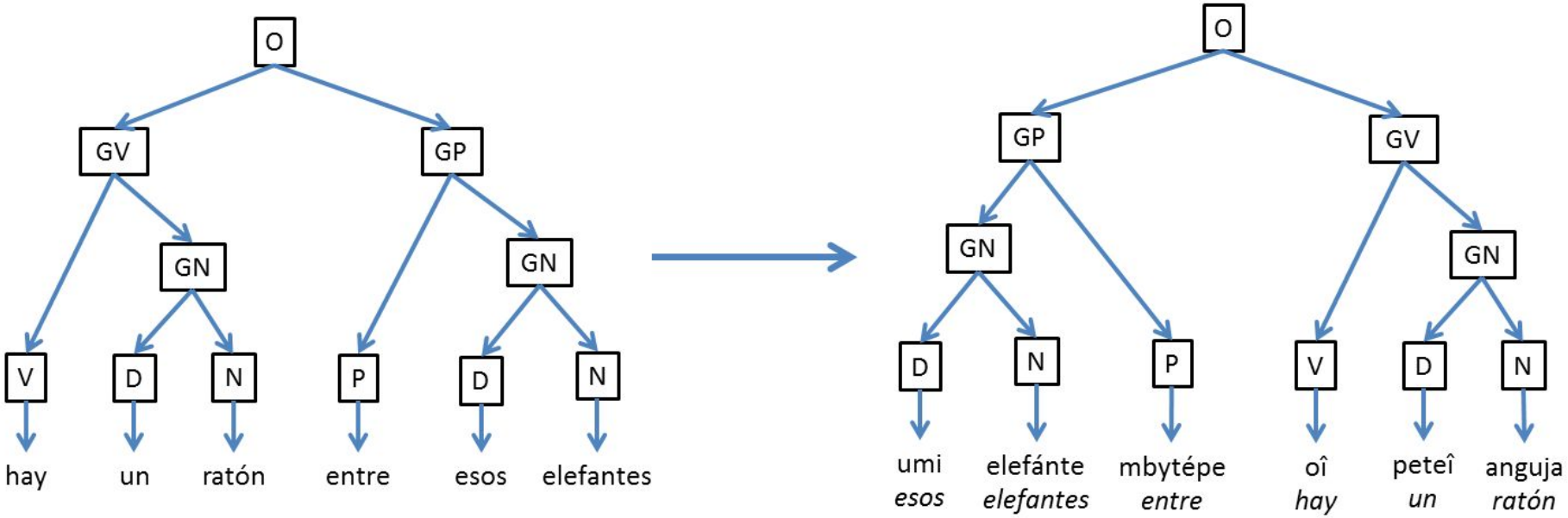
Ejercicio

(Febrero 2015) Suponga que se quiere construir un sistema de traducción automática del español al coreano basado en transferencia sintáctica. Infiera las reglas de transferencia que utilizaría dicho sistema, basándose en la siguiente traducción de ejemplo:



Ejercicio

(Febrero 2018) Suponga que se quiere construir un sistema de traducción automática del español al guaraní basado en transferencia sintáctica. Infiera las reglas de transferencia que utilizaría dicho sistema, basándose en la siguiente traducción de ejemplo:

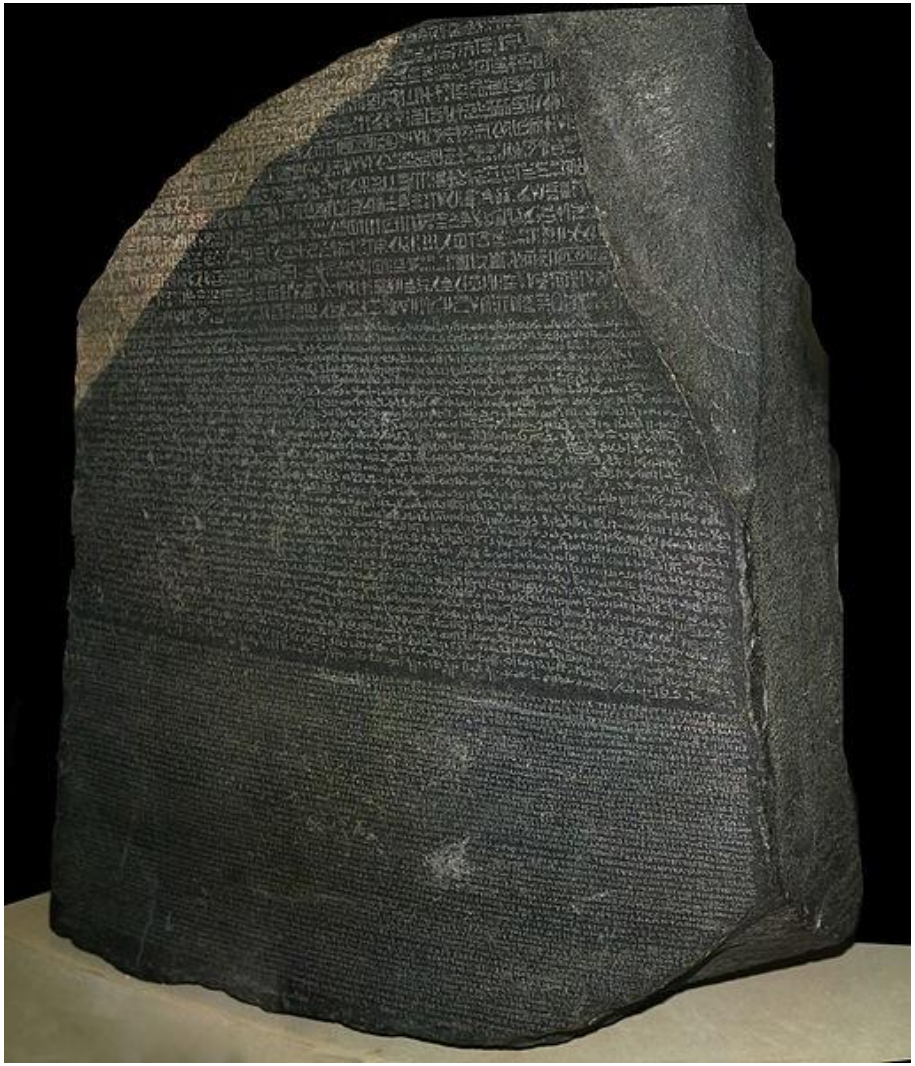


Corpus Paralelos

Corpus Paralelos

Un ejemplo famoso de corpus paralelo:

La Piedra de Rosetta.



Corpus Paralelos

- Conjuntos de pares de textos

  Un texto en el lenguaje origen y otro en el lenguaje destino

- Hay corpus paralelos para un puñado de pares de lenguajes
  - Árabe-Inglés
  - Chino-Inglés
  - La mayoría de los lenguajes europeos (más usados)
    - Europarl: 11 lenguajes, c.44M de palabras por lenguaje
  - United Nations Parallel Corpus
    - 10M palabras - árabe, chino, español, francés, inglés, ruso

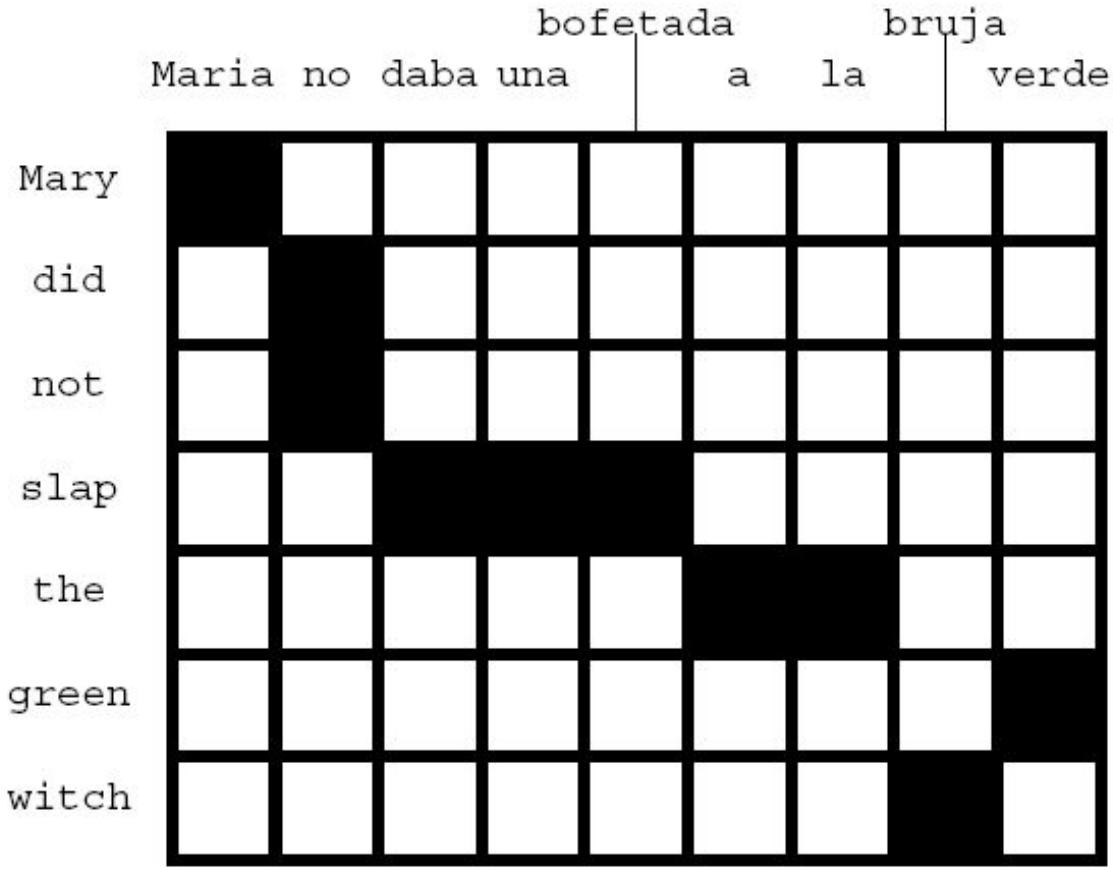
- ...pero para la mayoría de los pares de lenguajes, no existen corpus

Corpus Paralelos

Diferentes tipos de *alineación*:

- Alineados a nivel de documento
- Alineados a nivel de oración
  - Algoritmo de Gale y Church
  - Distancia coseno con embeddings multilingües
- Alineados a nivel de palabra
  - Este es el ideal, pero en general no existen

Alineación


bofetada
bruja
Maria no daba una a la verde
Mary
did
not
slap
the
green
witch

Traducción Automática Estadística

MT Estadística

Quiero traducir una oración del idioma F al idioma E

$$\hat{e} = \arg\max_e P(f \mid e) P(e)$$

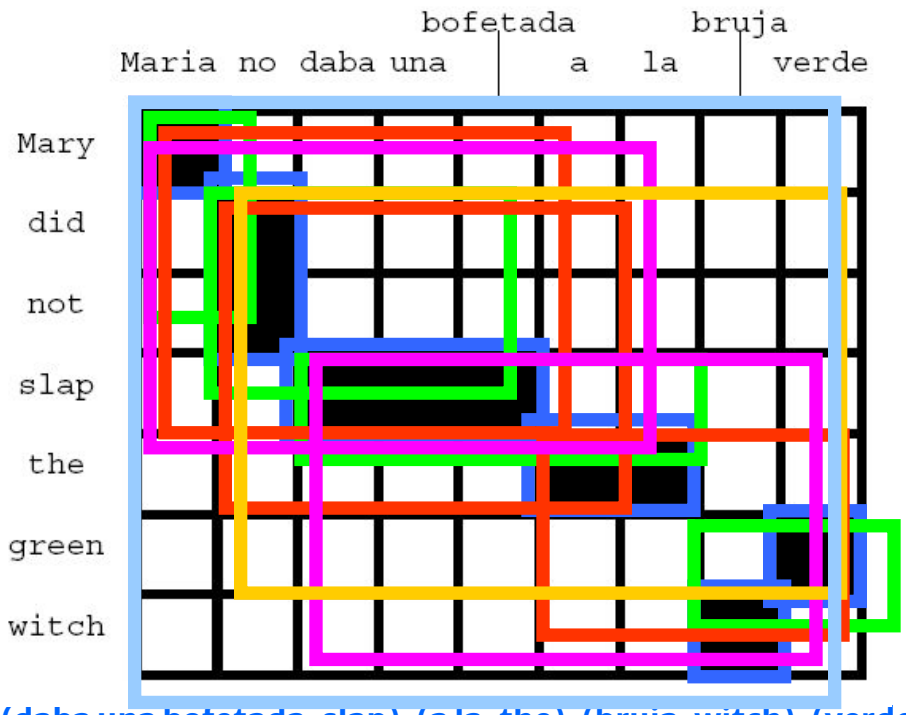
Para resolverla necesitamos:

- Un **modelo de lenguaje** P(*e*)
- Un **modelo de traducción** P(*f*|*e*)
- Un **decodificador**: proceso para buscar, de todas las oraciones *e* del idioma destino, cuáles son las más probables dado *f*

MT Estadística

- Modelos de lenguaje P(e)
  - n-gramas
- Modelos de traducción P($f|e$)
  - Uso de corpus paralelos (bilingües)
  - Basados en palabras
  - Basados en frases
- Decodificador:
  - Beam search

Modelo basado en frases



(María, Mary), (no, did not), (daba una bofetada, slap), (a la, the), (bruja, witch), (verde, green)

(María no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

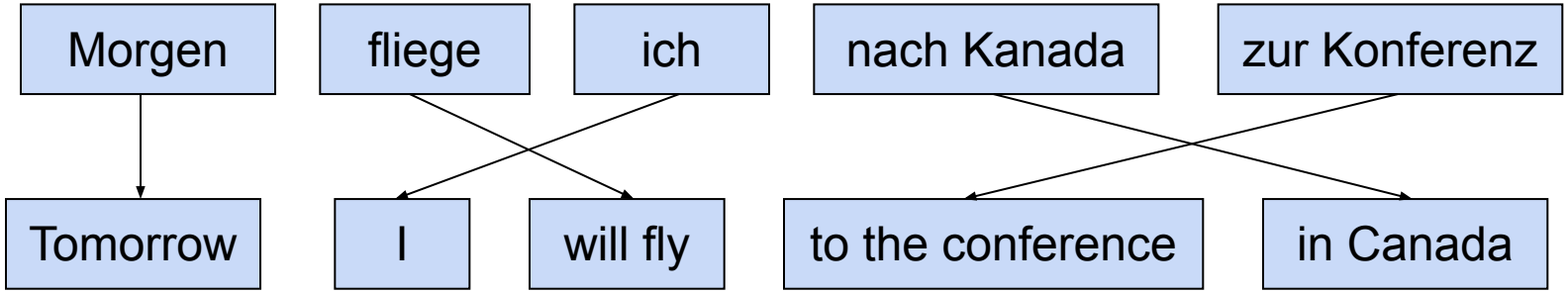
(María no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

(María no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch)

(no daba una bofetada a la bruja verde, did not slap the green witch),

(María no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Modelo basado en frases



La entrada en idioma origen se segmenta en frases

Cualquier secuencia de palabras, no es necesario que tenga significado lingüístico

Cada frase se traduce al idioma destino $\Phi(f,e)$

Se reordenan las frases $\Omega(f,e)$

$P(f|e) = \Phi(f,e)\ \Omega(f,e)$

Traducción Automática Neuronal

Traducción Automática Neuronal

La traducción automática es un problema de correspondencia entre pares de secuencias

- Respuestas a preguntas
- Resúmenes automáticos
- Otros...

Redes tipo seq2seq (secuencia a secuencia)

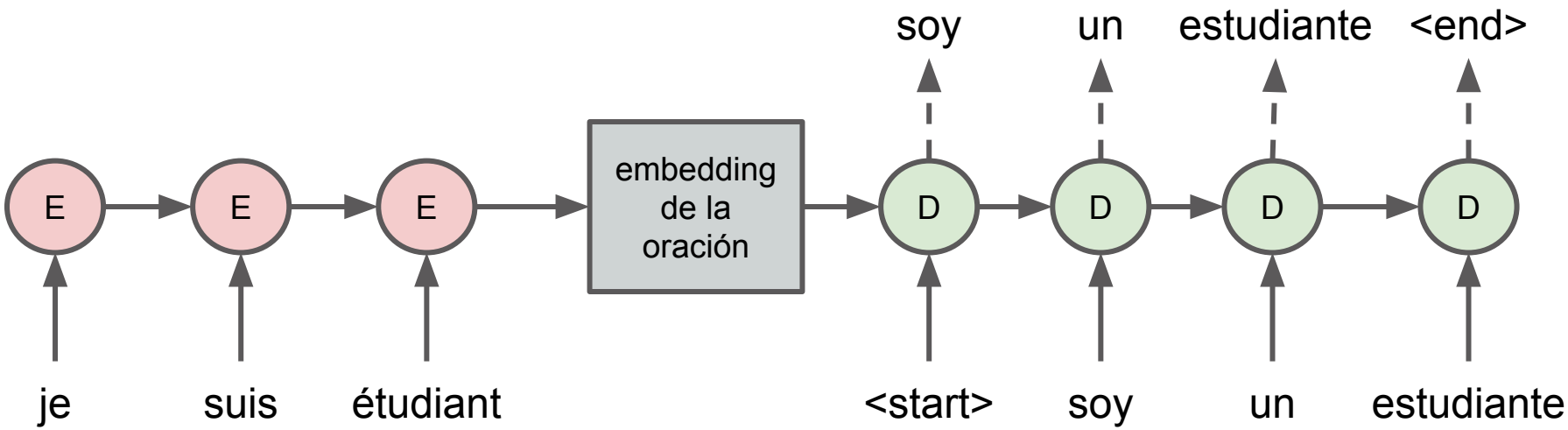
Traducción Automática Neuronal

Modelo codificador-decodificador (encoder-decoder)

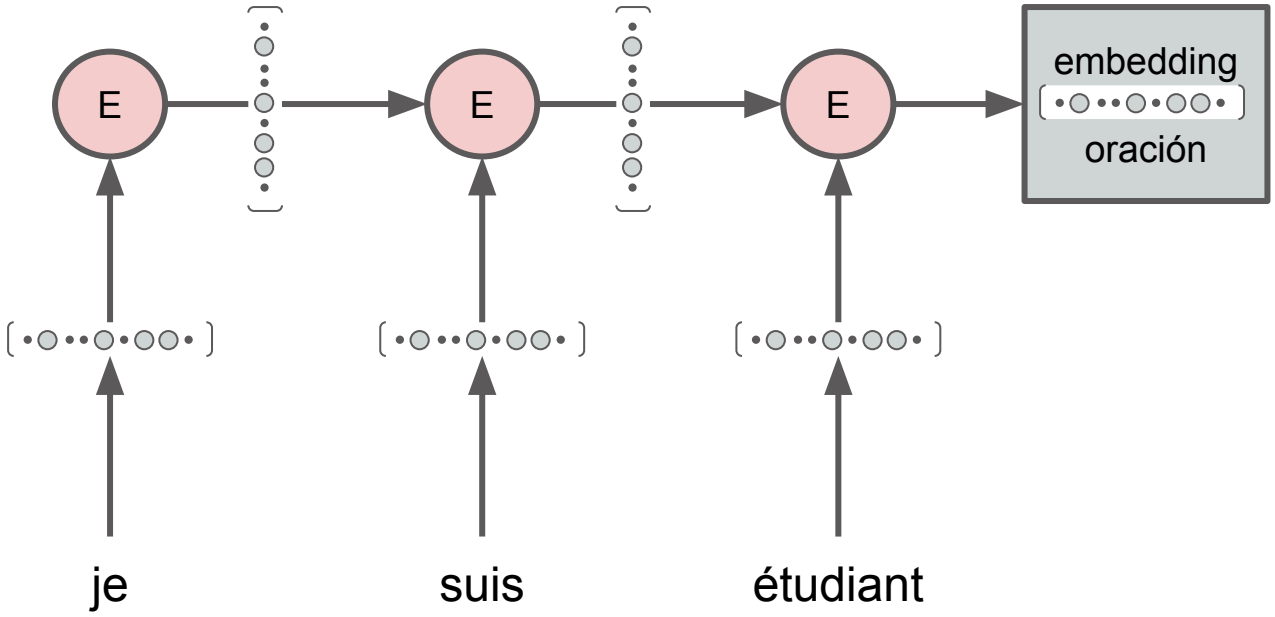
Una red codifica la oración en el idioma origen

Una red genera la decodificación en el idioma destino

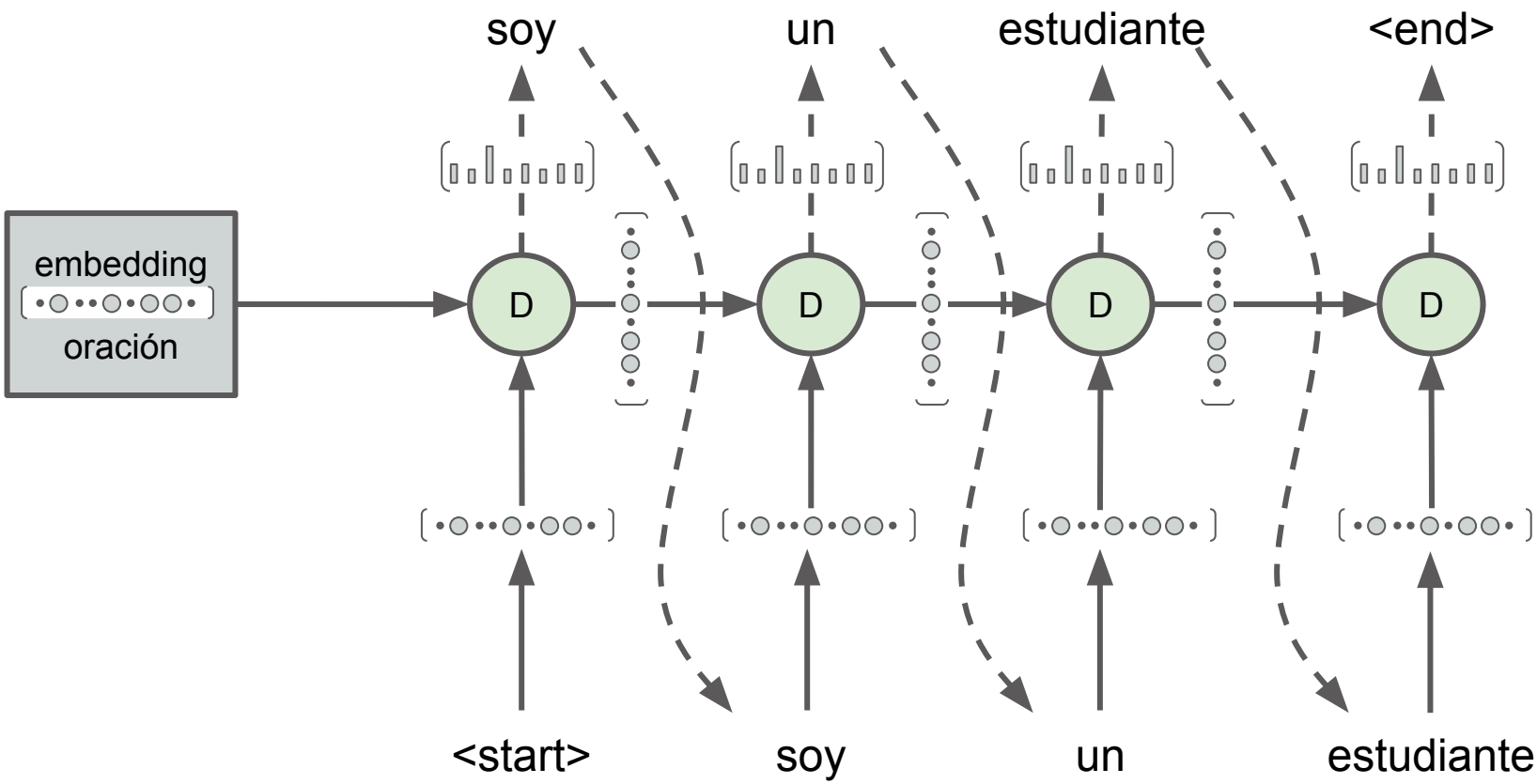
Modelo Encoder-Decoder



Encoder



Decoder


soy
un
estudiante
<end>
embedding
oración
D
D
D
D
<start>
soy
un
estudiante

Entrenamiento

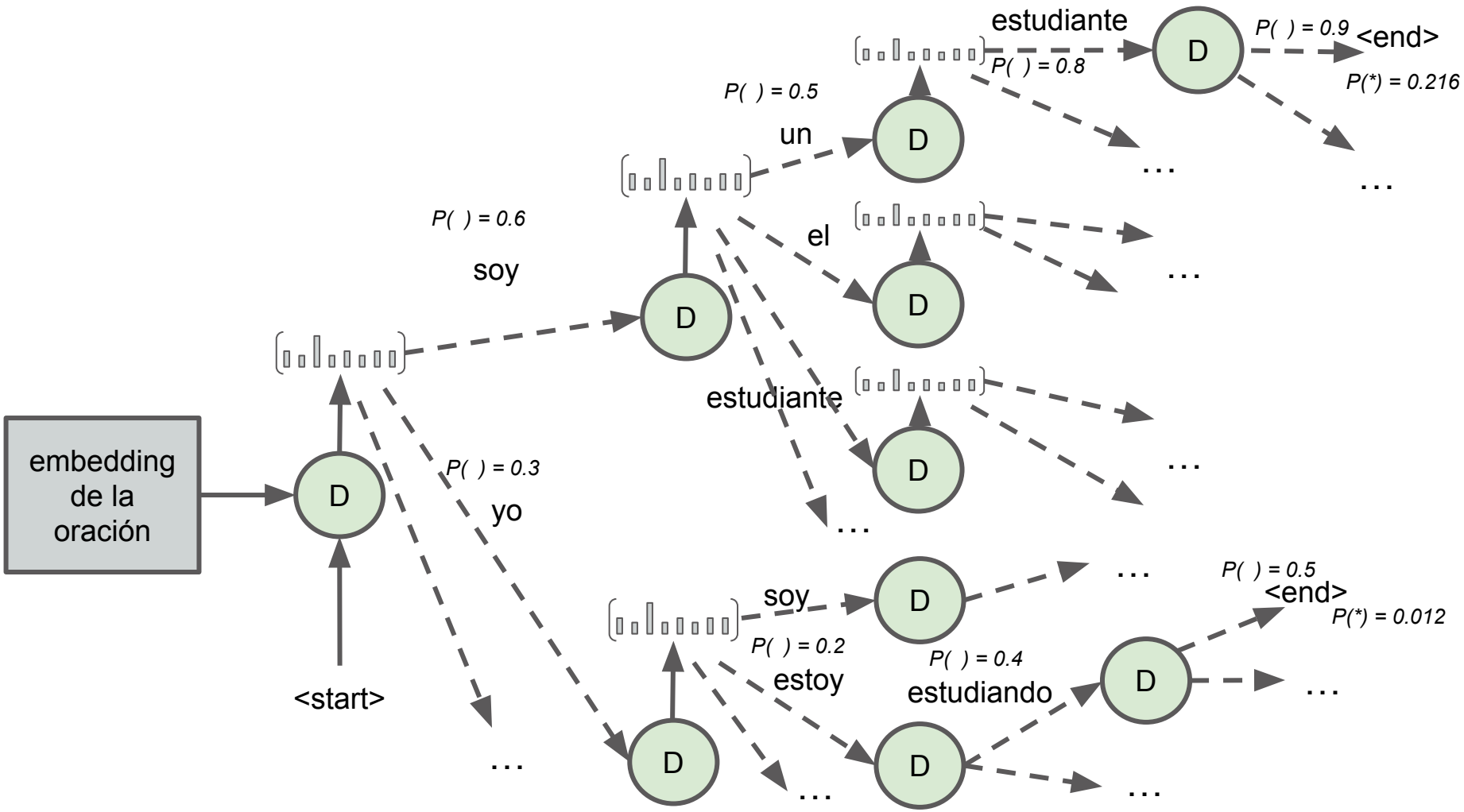
End-to-end: usamos pares de oraciones alineados como ejemplos de entrada y salida esperadas

Cada entrada del decoder usa la palabra correcta esperada, y no la salida del paso anterior

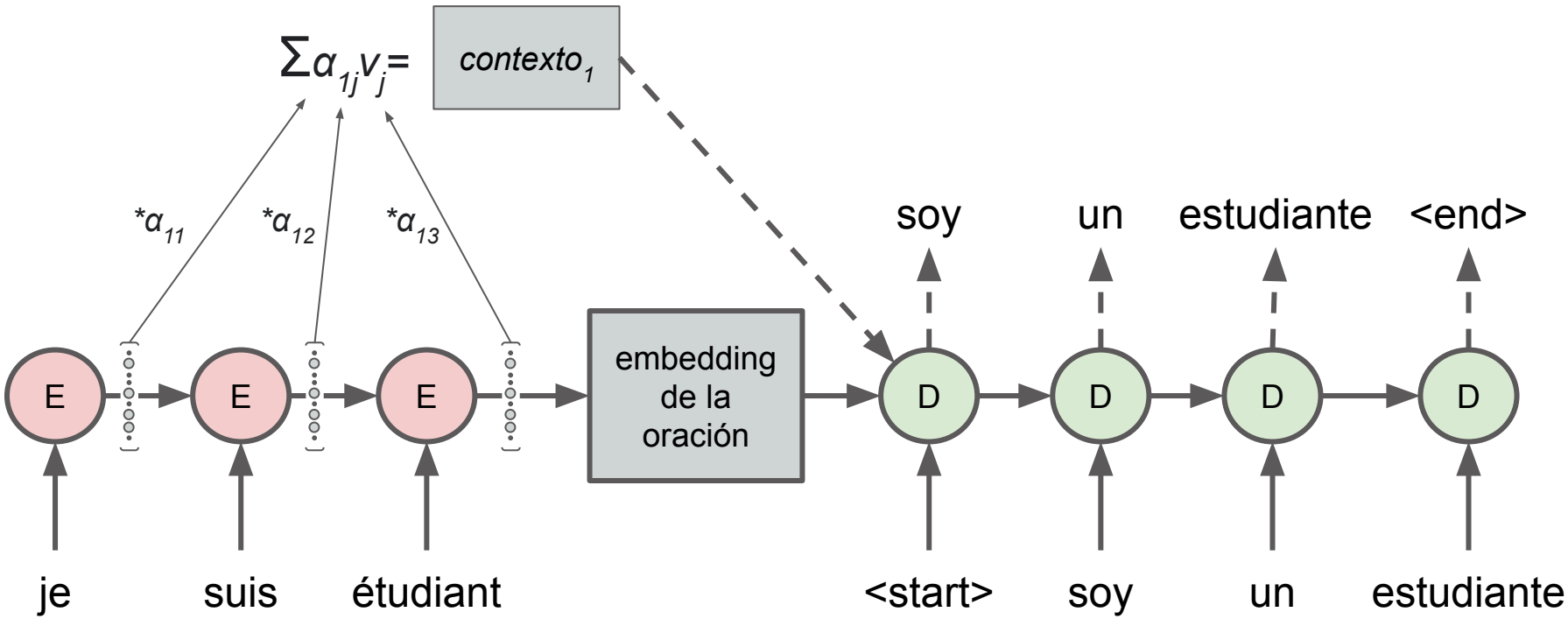
Tokenización:

- Se define un vocabulario fijo para cada idioma
- O se puede usar un vocabulario compartido entre idiomas
- Esto permite usar embeddings compartidos para inicialización

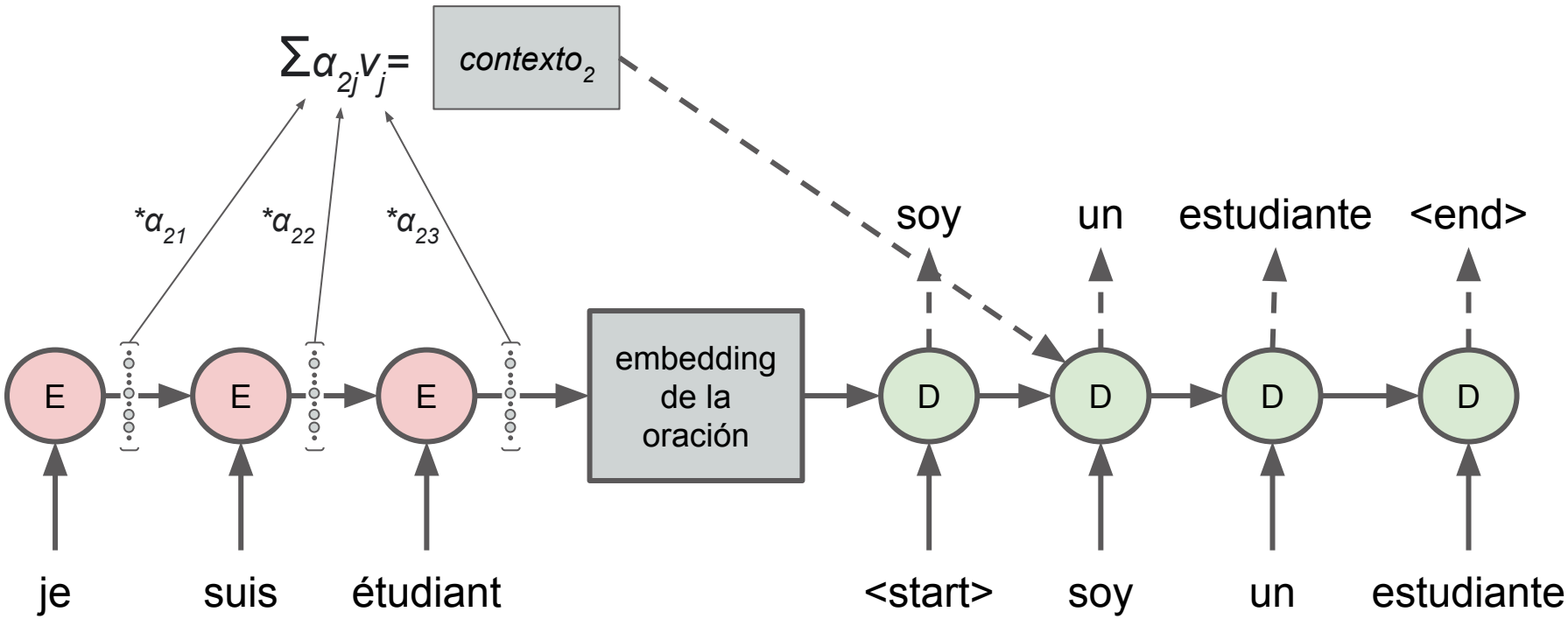
Beam Search


embedding de la oración
D
<start>
P( ) = 0.3
yo
P( ) = 0.6
soy
P( ) = 0.5
un
el
estudiante
P( ) = 0.8
estudiante
P( ) = 0.9
<end>
P(*) = 0.216
soy
P( ) = 0.2
estoy
P( ) = 0.4
estudiando
P( ) = 0.5
<end>
P(*) = 0.012
…

Modelo Atencional



Modelo Atencional



Modelo Atencional

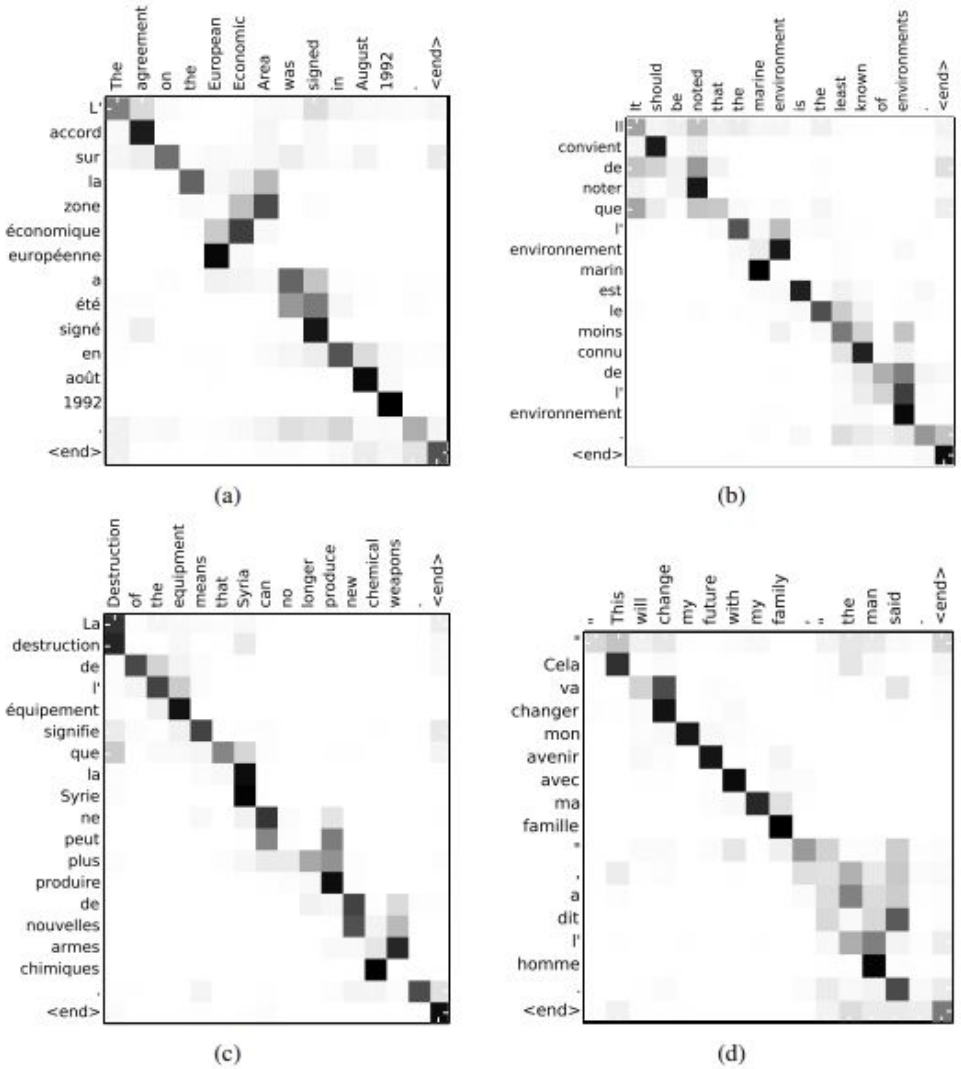
En cada paso del decoder creo un vector de contexto

Promedio ponderado de los embeddings del encoder

Ponderaciones α_{ij} : matriz de largo origen * largo destino

Las ponderaciones se aprenden durante el entrenamiento

Modelo Atencional



Evaluación

Evaluación

Medida BLEU

$$BLEU = BP \exp\left(\sum_{1}^{N} w_n \log p_n \right)$$

- Compara un conjunto de traducciones candidatas con un conjunto de traducciones de referencia
- Cuenta n-gramas presentes en los candidatos que también estén en las referencias (n = 1,2,3,4)
- Incluye una penalización por brevedad (BP) para que las traducciones demasiado cortas tengan menos puntos

Evaluación

Definición de BP:

$$BP = \begin{cases} 1 & (C' \geq R') \\ \exp\left(1 - \frac{R'}{C'}\right) & (C' < R') \end{cases}$$

Donde R' es el largo total (en palabras) de todas las referencias (documento referencia) y C' es la el largo total de las traducciones candidatas (documento candidato)

Ejercicio

Calcular BLEU para los siguientes candidatos a traducción inglés-español:

Original	Referencia	Candidato 1	Candidato 2
Yesterday it was raining very heavily	Ayer estaba lloviendo muy fuerte	Llovía muy fuerte ayer	Ayer estaba lloviendo pesadamente
I wasn't carrying an umbrella	No llevaba paraguas	Yo no llevaba un paraguas	Él llevaba paraguas
Then I took shelter under a roof	Entonces me refugié bajo un techo	Entonces me refugié debajo de un techo	Entonces refugié bajo techo

Considere n-gramas hasta n=3 y pesos w_i equiprobables

Evaluación

BLEU generalmente se correlaciona con la evaluación subjetiva humana

Captura nociones de adecuación y fluidez

- Unigramas vs n-gramas mayores

Difícil de interpretar

- Valor entre 0 y 1
- Estado del arte inglés-francés 0.48
- Una buena traducción que no esté en el conjunto de referencia será penalizada