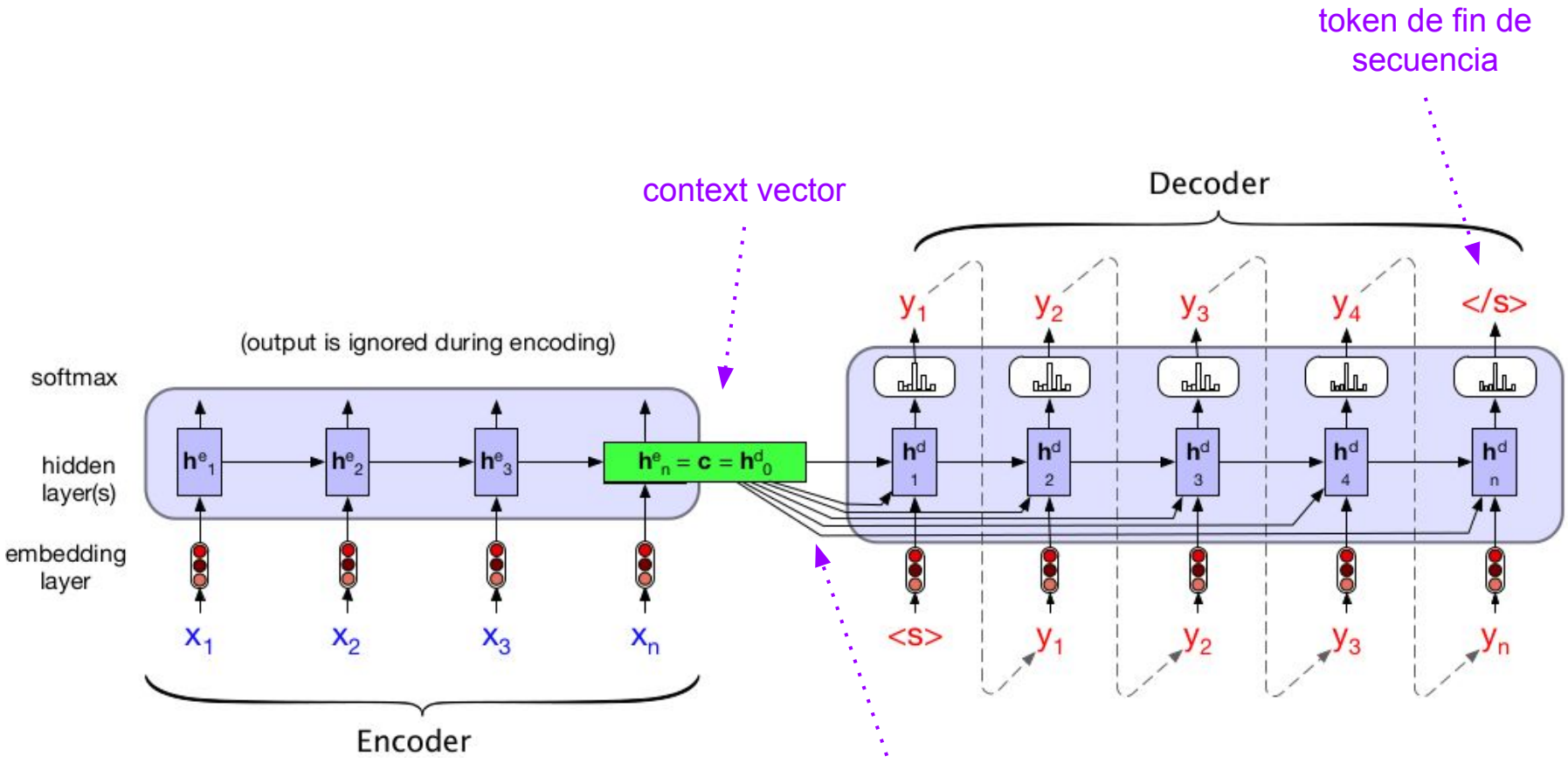
Redes Neuronales para Lenguaje Natural

2024

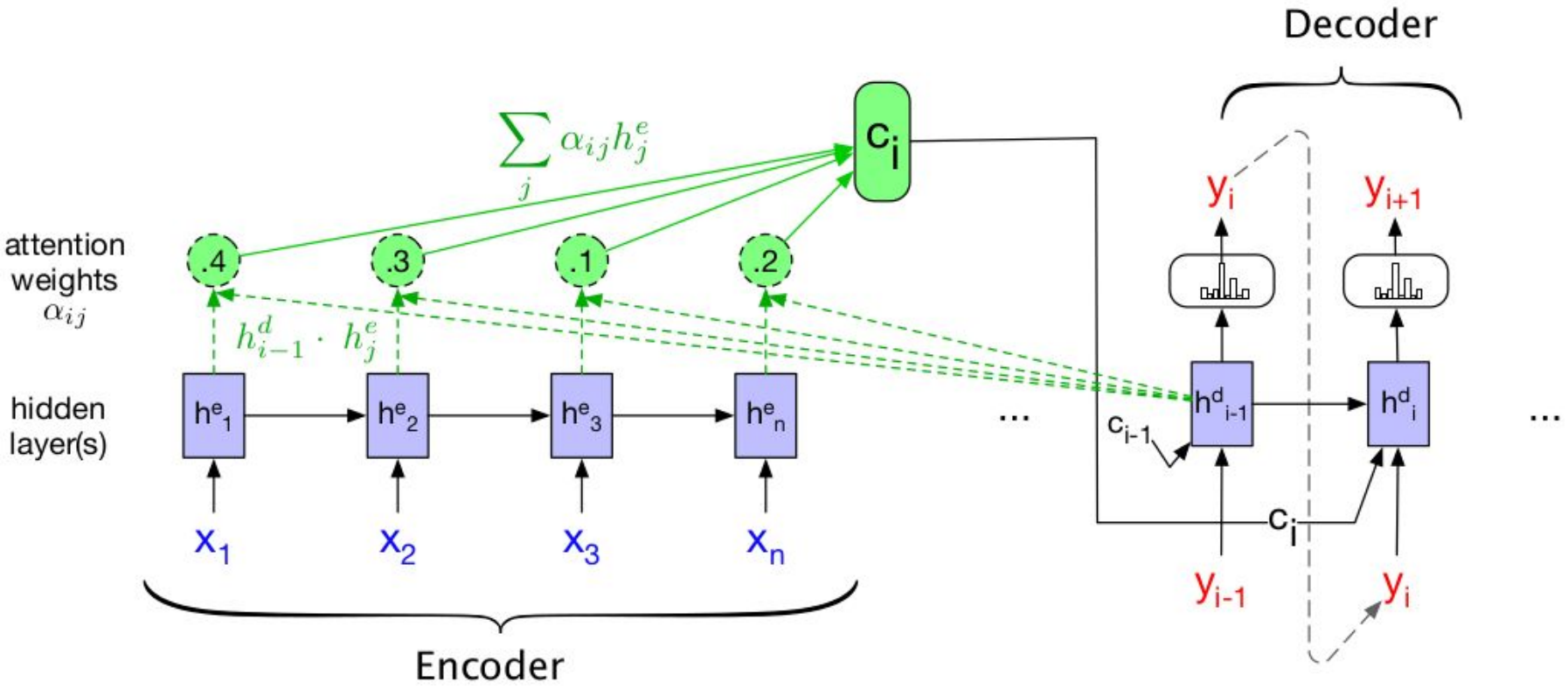
Grupo de Procesamiento de Lenguaje Natural
Instituto de Computación

Encoder-Decoder



debido a que la influencia de c sobre la decodificación de la secuencia puede ir disminuyendo, se suele incluir en cada paso

Mecanismo Atencional




Traducción Automática

Traducción Automática

Machine Translation (MT)

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

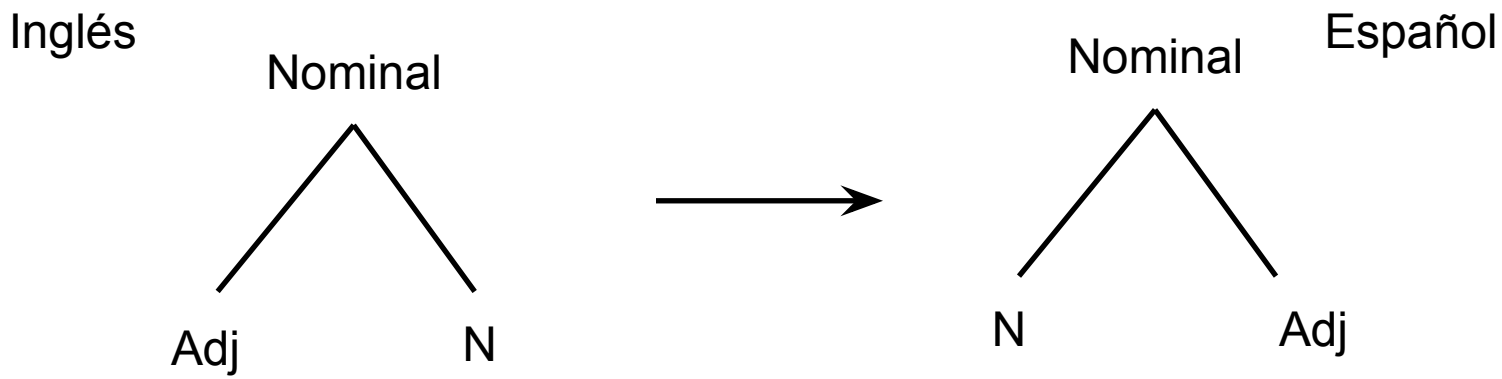
Ha ido evolucionando junto con el PLN (y el aprendizaje automático)

- Métodos basados en reglas (1950s - 1990s)
- Métodos estadísticos (2000s - 2010s)
- Métodos neuronales (2010s hasta hoy)

Métodos Basados en Reglas

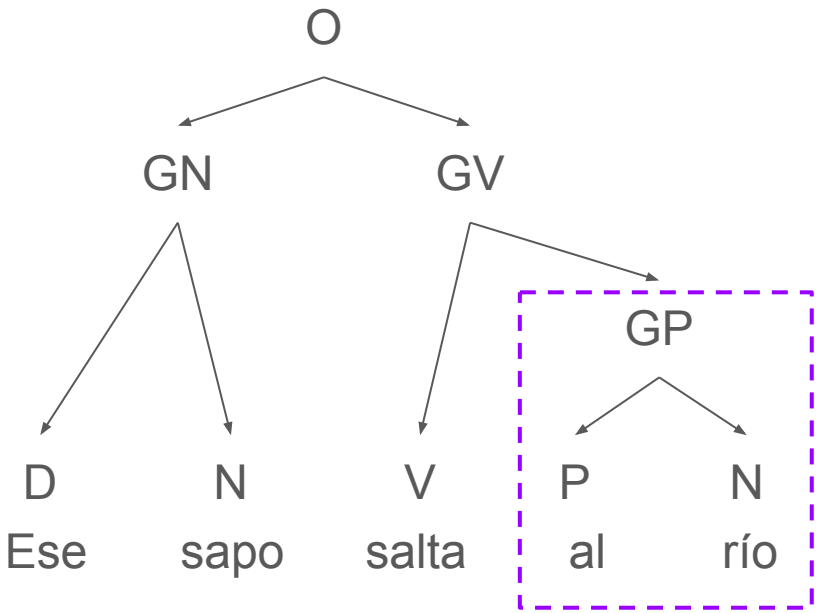
Transferencia Sintáctica

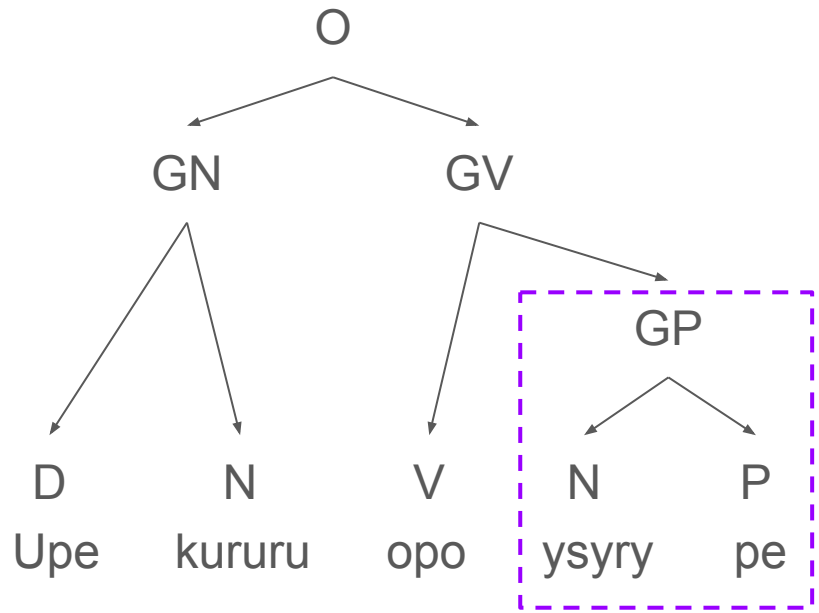
- Parsing del lenguaje origen
- Generación en en lenguaje destino
- Reglas de transferencia entre árboles y subárboles



$\text{Nom} \rightarrow \text{Adj N} \;/\; \text{Nom} \rightarrow \text{N Adj}$

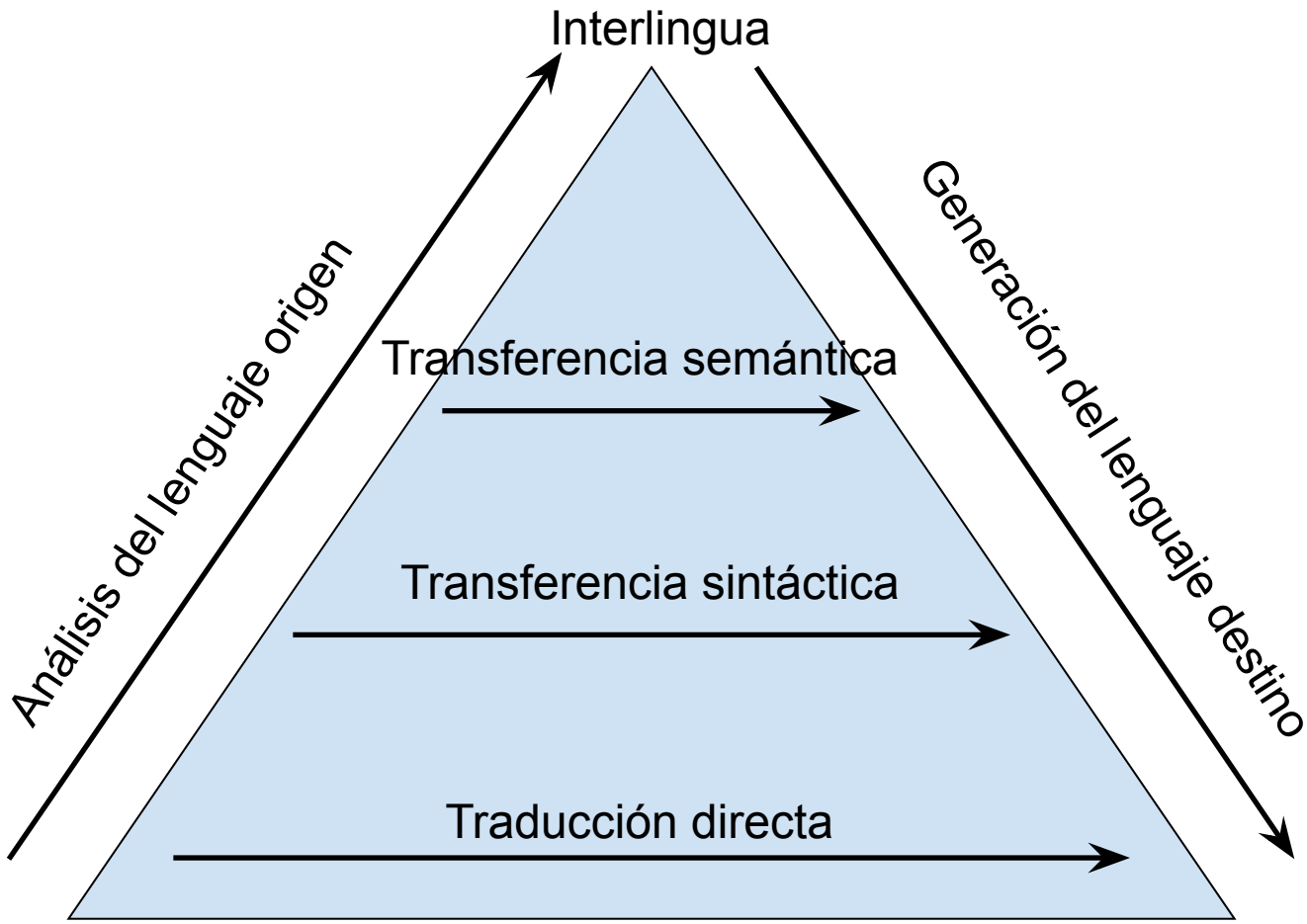
Métodos Basados en Reglas





GP → P N / GP → N P

Métodos Basados en Reglas



A survey of formal grammars and algorithms for recognition and transformation in machine translation
Vauquois, 1968

Métodos Neuronales

Vemos el problema de traducción como un problema de aprendizaje automático

Cosas que necesitamos definir

- ¿Cómo son los datos en MT?
  - Corpus paralelos
- ¿Cómo representamos las palabras?
  - Tokenización, embeddings
- ¿Qué modelos se utilizan?
  - Modelos seq2seq: encoder-decoder con RNNs, LSTMs, mecanismo atencional, transformers…
- ¿Cómo medimos la performance del sistema?
  - Diferentes métricas: BLEU, chrF…

Corpus Paralelos

- Conjuntos de pares de textos

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

  A diferencia de los corpus monolingües, no es tan común que existan “naturalmente”

- Dónde aparecen? Países multilingües y entidades multinacionales
  - Francés-Inglés, Chino-Inglés
  - Europarl: 11 idiomas más usados de la UE, 44M de palabras por idioma
  - United Nations Parallel Corpus: 10M palabras en árabe, chino, español, francés, inglés, ruso

- Hay alrededor de 7000 idiomas en el mundo

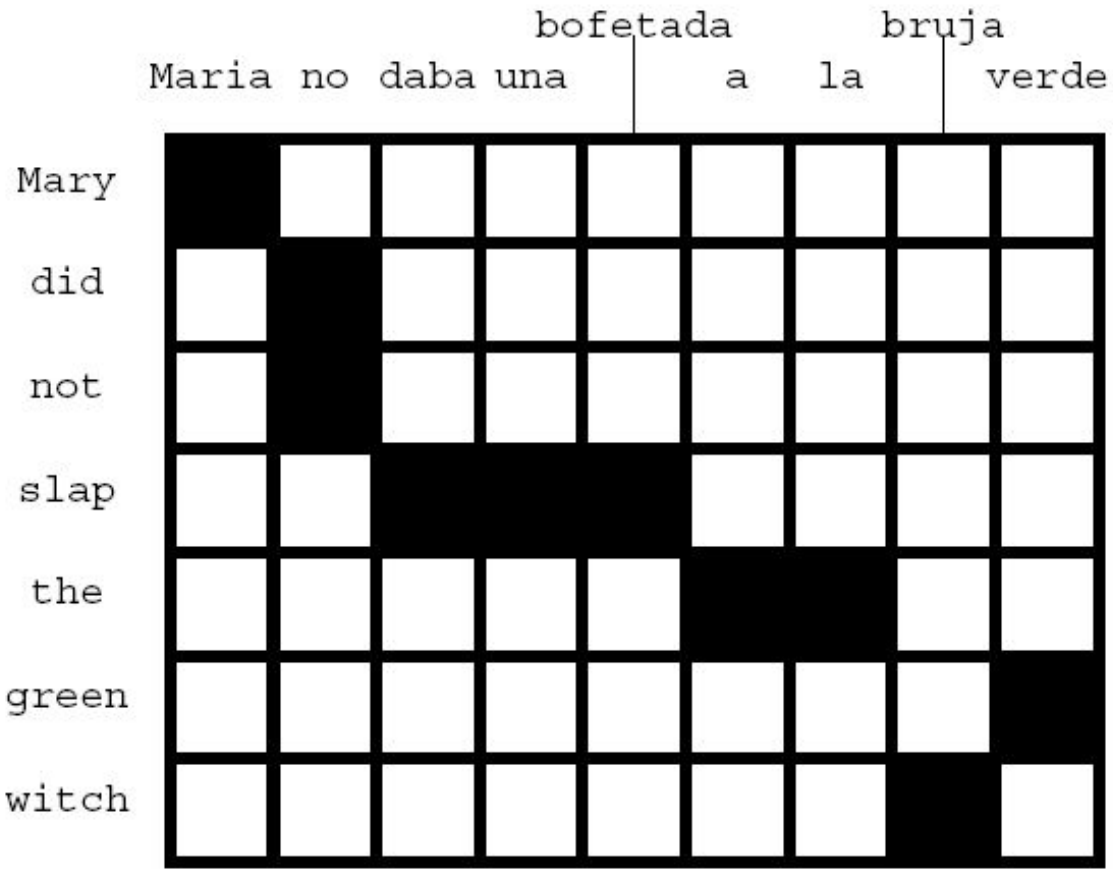
  La mayoría no tienen ningún tipo de corpus, mucho menos paralelo!

Corpus Paralelos

Diferentes tipos de alineación:

- Alineados a nivel de documento
- Alineados a nivel de oración
  Programación dinámica (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 Neuronal

Traducción Automática Neuronal

Modelo codificador-decodificador (encoder-decoder)

Una red codifica la oración en el idioma origen

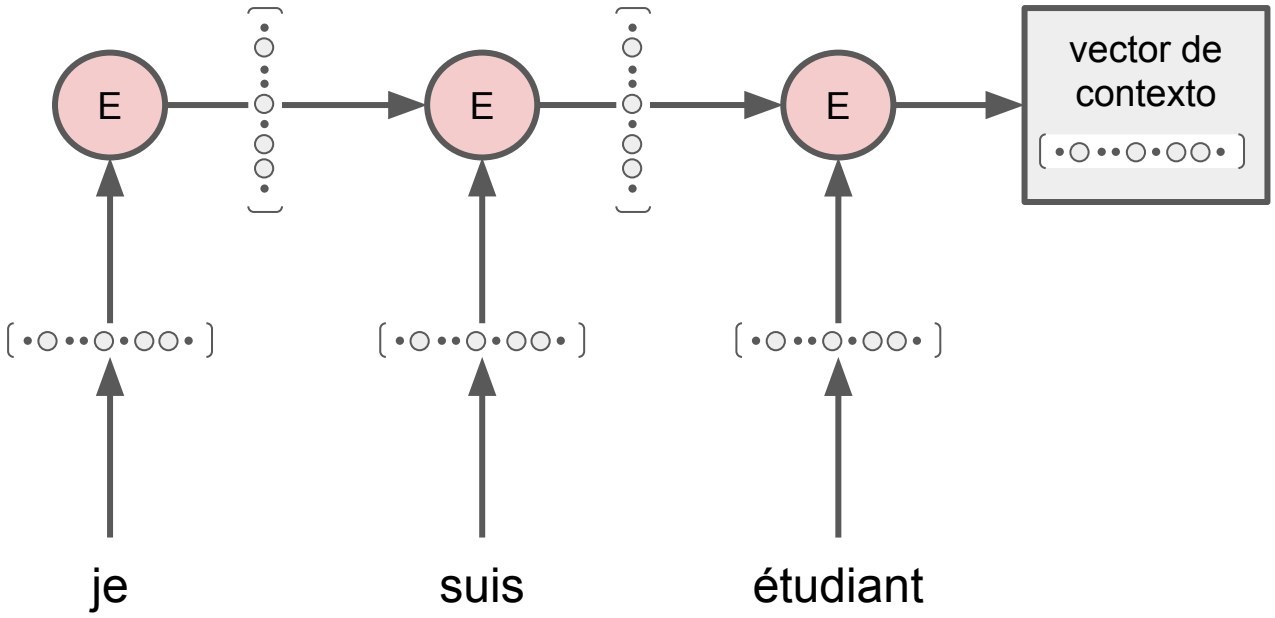
Una red genera la decodificación en el idioma destino

Implementaciones habituales:

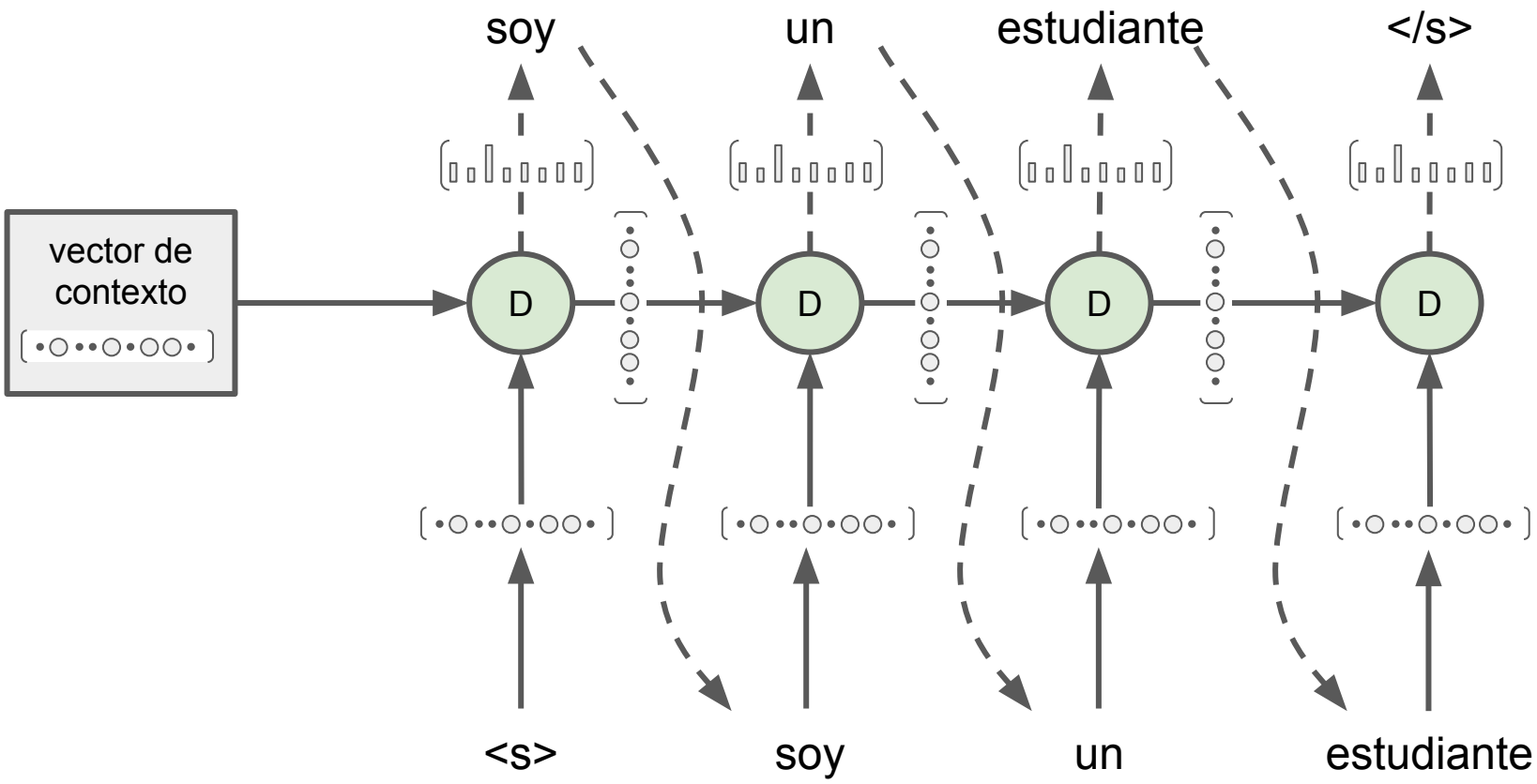
- RNNs (LSTMs) con mecanismo atencional
- Transformers

Entrenado end-to-end con pares alineados

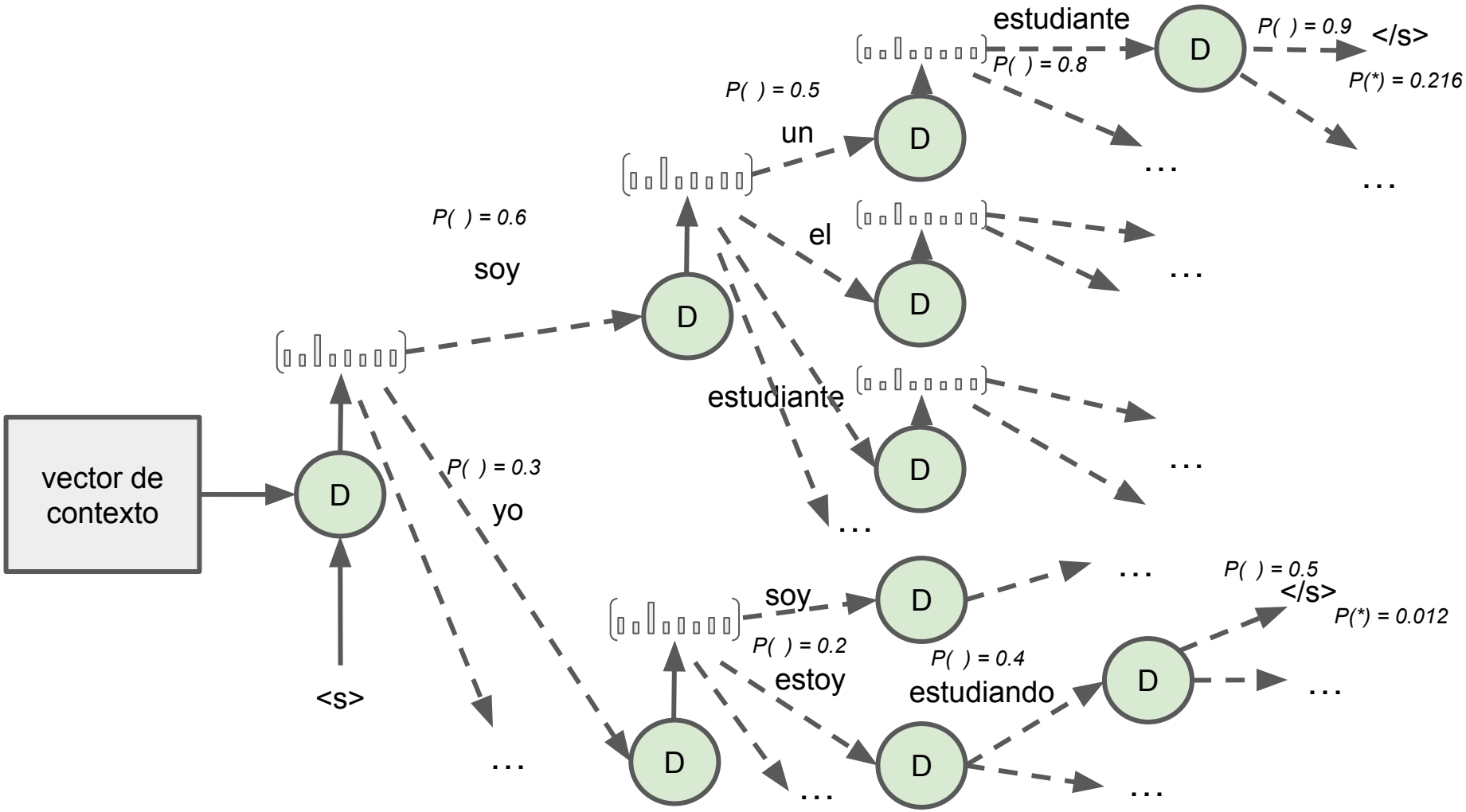
Encoder



Decoder


soy
un
estudiante
</s>
vector de contexto
D
D
D
D
<s>
soy
un
estudiante

Beam Search


vector de contexto
D
<s>
P( ) = 0.6
soy
P( ) = 0.3
yo
un
el
estudiante
P( ) = 0.5
P( ) = 0.8
estudiante
P( ) = 0.9
</s>
P(*) = 0.216
soy
estoy
P( ) = 0.2
estudiando
P( ) = 0.4
P( ) = 0.5
</s>
P(*) = 0.012

Beam Search

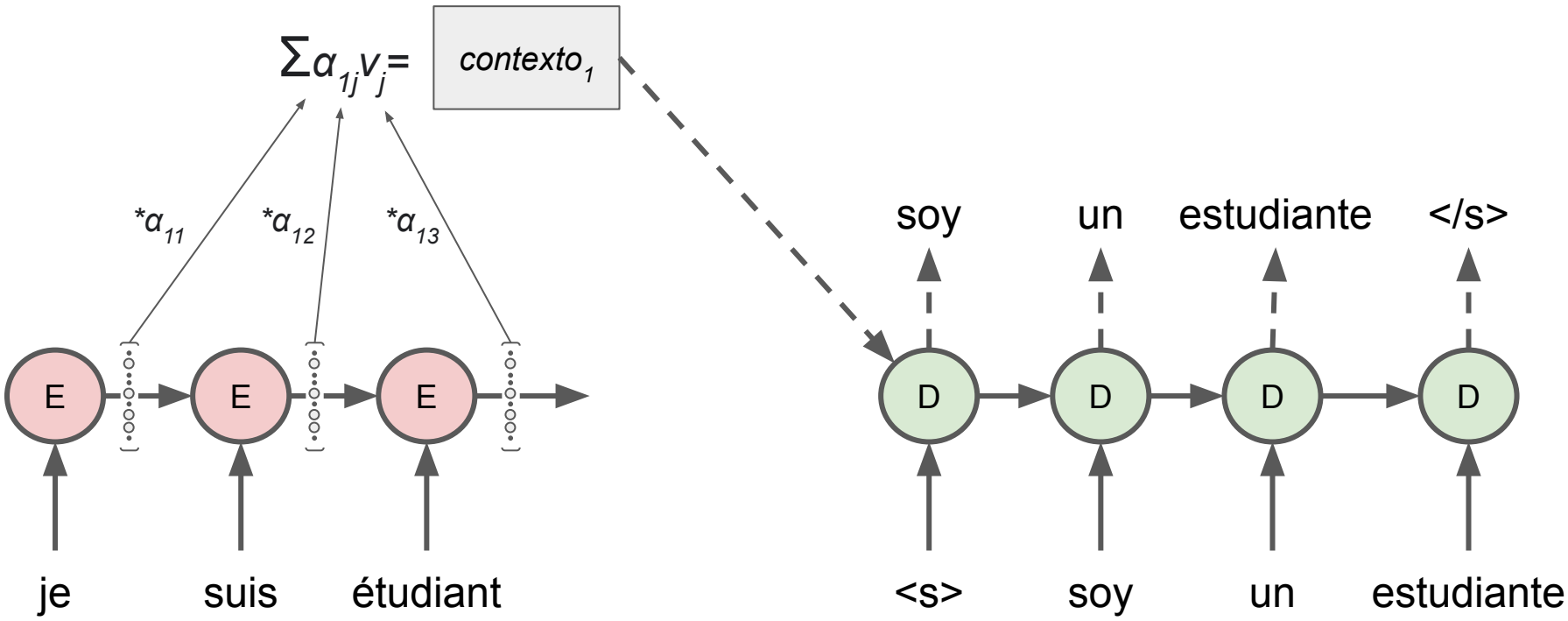
- Seleccionamos un tamaño del beam *n*
- En el primer paso del decoder, elegimos las *n* palabras más probables
- Expandimos a partir de esas *n*, y nos quedamos con los *n* pares más probables
- Iteramos el proceso hasta llegar a generar las </s>, siempre nos quedamos con las *n* más probables

Beam Search

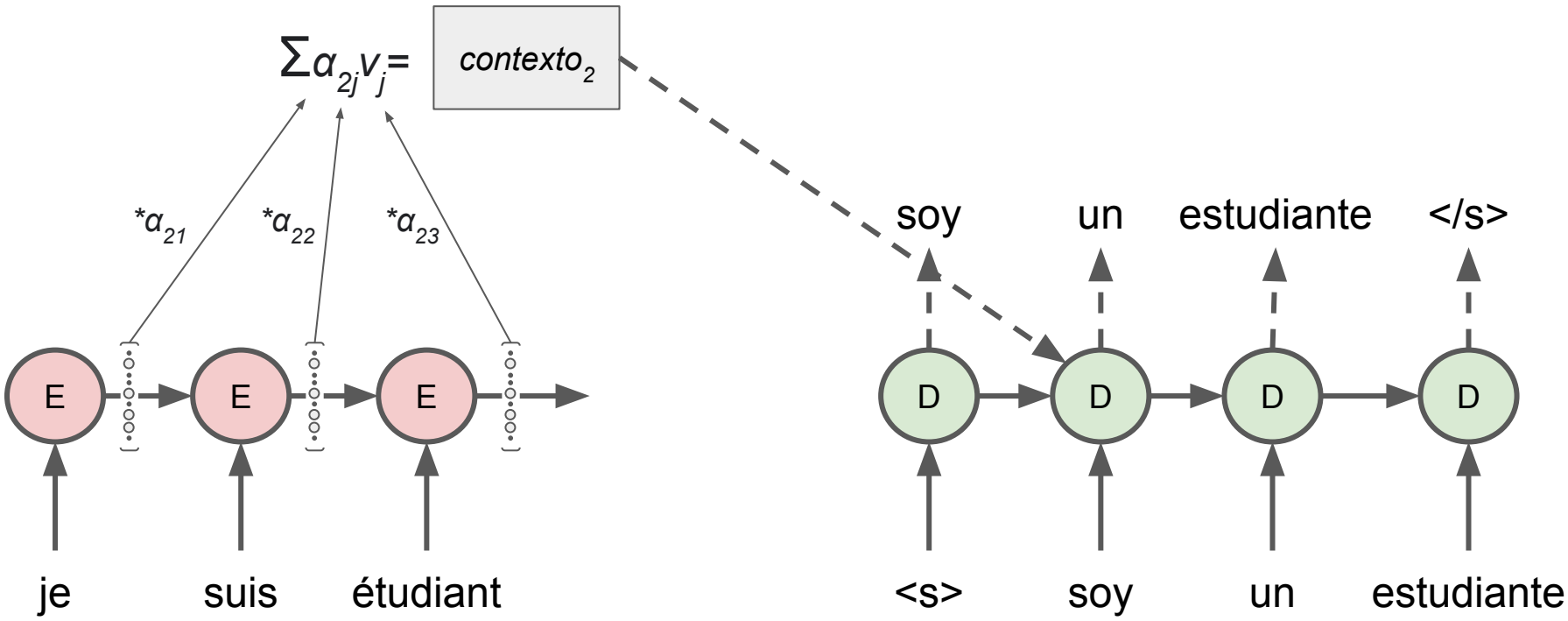
No es un proceso perfecto!

- No asegura que la decodificación más probable esté en el beam
- Pero sí asegura que la nueva solución encontrada va a ser al menos tan buena como la greedy
- Aún puede ser lento, se le pueden hacer mejoras de performance como podas tempranas

Modelo Atencional



Modelo Atencional



Modelo Atencional

En cada paso del decoder creo un vector de contexto

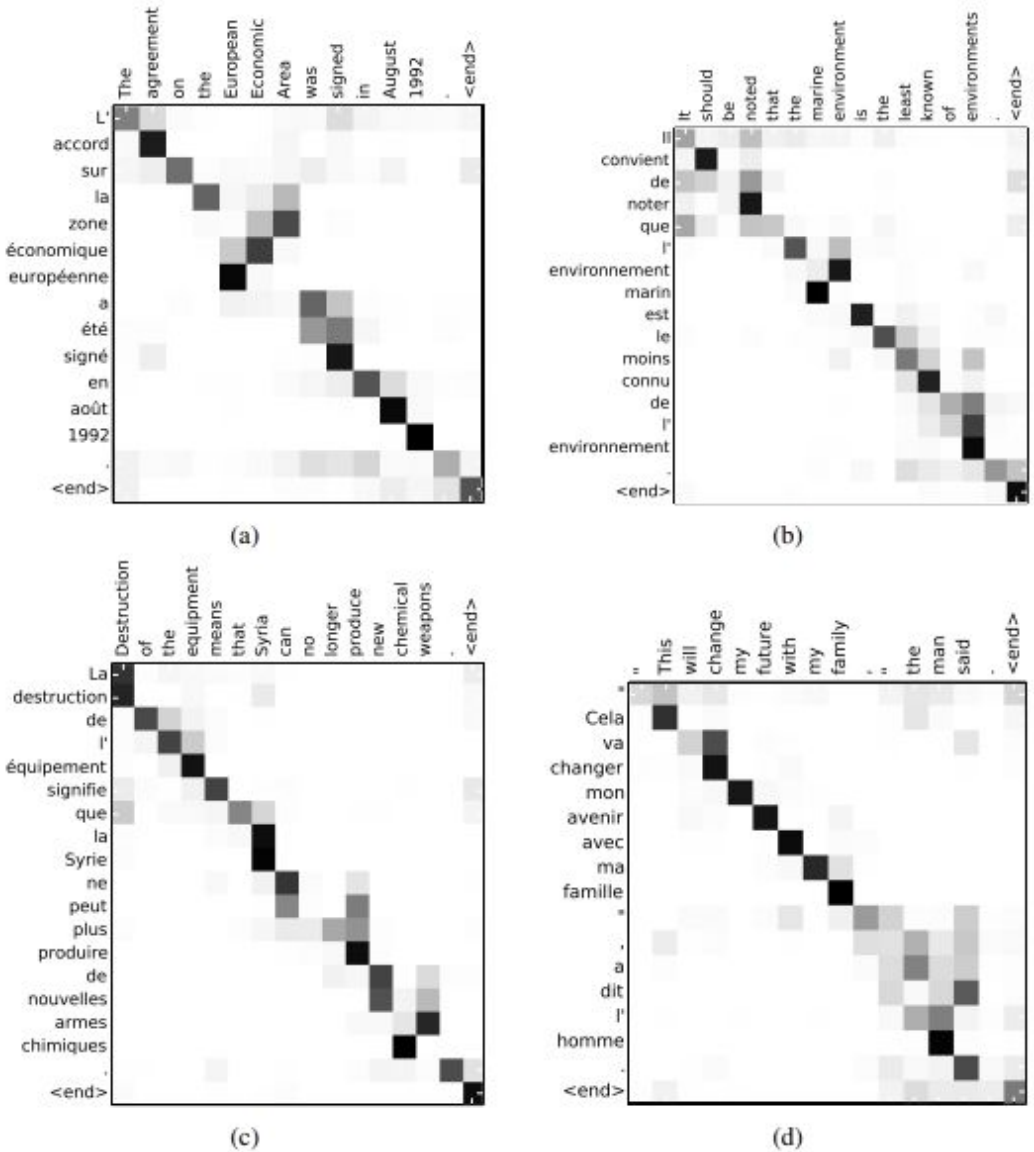
Promedio ponderado de los embeddings del encoder

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

Vimos diferentes maneras de aprender esas ponderaciones

- Atención aditiva
- Atención multiplicativa

Modelo Atencional




Tokenización

Representación de las palabras

¿Cómo tratamos las palabras de los idiomas?

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

¿Pero cómo son esos tokens?

Tokenización

Es un tópico general de PLN. Formas clásicas de tokenización incluyen:

- Por espacios
- Expresiones regulares
- Morfemas

En MT y métodos modernos se suelen usar métodos estadísticos que obtienen información sub-palabra:

- BPE, ULM, wordpiece

Algoritmo BPE

Codificación de a pares de bytes (Byte-Pair Encoding)

El corpus a procesar es una colección de tokens con separación simple por espacios

El vocabulario inicial está compuesta de todas las letras (bytes) encontradas en el corpus, más un símbolo especial de fin de palabra “_”

En cada paso, uniremos un par de tokens del vocabulario hasta alcanzar la cantidad *k* tokens en total (parámetro del sistema)

Algoritmo BPE

Corpus

```
5  l o w _
2  l o w e s t _
6  n e w e r _
3  w i d e r _
2  n e w _
```

9 veces

Vocabulario

```
_, d, e, i, l, n, o, r,
s, t, w
```

Algoritmo BPE

Corpus

```
5  l o w _
2  l o w e s t _
6  n e w er _
3  w i d er _
2  n e w _
```

9 veces

Vocabulario

```
_, d, e, i, l, n, o, r,
s, t, w, er
```

Algoritmo BPE

Corpus

```
5  l o w _
2  l o w e s t _
6  n e w er_
3  w i d er_
2  n e w _
```

8 veces

Vocabulario

```
_, d, e, i, l, n, o, r,
s, t, w, er, er_
```

Algoritmo BPE

Corpus

```
5  l o w _
2  l o w e s t _
6  ne w er_
3  w i d er_
2  ne w _
```

8 veces

Vocabulario

```
_, d, e, i, l, n, o, r,
s, t, w, er, er_, ne
```

Algoritmo BPE

Corpus

7 veces

```
5  l o w _
2  l o w e s t _
6  new er_
3  w i d er_
2  new _
```

Vocabulario

```
_, d, e, i, l, n, o, r,
s, t, w, er, er_, ne,
new
```

Algoritmo BPE

Corpus

```
5  lo w _

2  lo w e s t _

6  new er_

3  w i d er_

2  new _
```

Vocabulario

```
_, d, e, i, l, n, o, r,
s, t, w, er, er_, ne,
new, lo
```

Algoritmo BPE

Sigue uniendo de a dos palabras:

`(lo, w) → _, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low`

`(new, er_) → _, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_`

`(low, _) → _, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_, low_`

Hasta que hayamos hecho *k* uniones

Notar que palabras enteras quedan en el vocabulario (las frecuentes)

Algoritmo BPE

Al momento de codificar un texto nuevo, recorremos la lista haciendo las uniones en el orden que las fuimos encontrando

`_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_, low_`

`l o w e r _ → l o w er _ → l o w er_ → lo w er_ → low er_`

“lower” no estaba en el vocabulario, y la podemos representar con dos tokens: (low, er_)


Evaluación

Evaluación

Es difícil porque puede haber más de una salida correcta!

La forma ideal para evaluarlo es con anotadores humanos

Dada la oración en el idioma origen, evaluar la traducción candidata

- Adecuación (1 al 5): qué tanto se preserva la semántica
- Fluidez (1 al 5): qué tan bien suena en el idioma destino

Problemas:

- Muy caro!
- No reutilizable

Evaluación

Aunque no son perfectas, las métricas automáticas son lo más habitual

- WER
- BLEU
- METEOR
- chrF

Todas se basan en tener una o más traducciones de referencia

Se olvidan de la oración origen: solo se compara las traducciones candidatas con las de referencia

Métrica BLEU

Métrica 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 (de tokens) 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

BLEU: a Method for Automatic Evaluation of Machine Translation
Papineni et al., 2002

Métrica chrF

A diferencia de BLEU, que cuenta n-gramas de tokens, chrF utiliza n-gramas de caracteres

$$\mathrm{chrF}\beta = (1+\beta)\frac{\mathrm{chrP}\cdot\mathrm{chrR}}{\beta^2\cdot\mathrm{chrP}+\mathrm{chrR}}$$

- chrP: cantidad de *n-gramas* de caracteres de la hipótesis que están en la referencia
- chrP: cantidad de *n-gramas* de caracteres de la referencia que están en la hipótesis
- En general n≤4 o 6, y β=3

chrF: character n-gram F-score for automatic MT evaluation
Popović, 2015

Evaluación

Tanto BLEU como chrF toman valores entre 0 y 1 como BLEU

Las dos penalizan traducciones que no estén entre las referencias

En general se correlacionan con la evaluación subjetiva humana, pero el número resultante en sí es difícil de interpretar

BLEU es más estricta que chrF porque cuenta palabras exactas, mientras que chrF igual da algunos puntos si se tradujo una sub-palabra

chrF es mejor para hacer comparaciones con idiomas morfológicamente ricos, ya que puede haber muchas flexiones de una misma palabra

Ejemplo:
Traducción Guaraní-Español

Idioma Guaraní

- Lengua indígena de América del Sur
- Hablada por entre 6 y 10 millones de personas

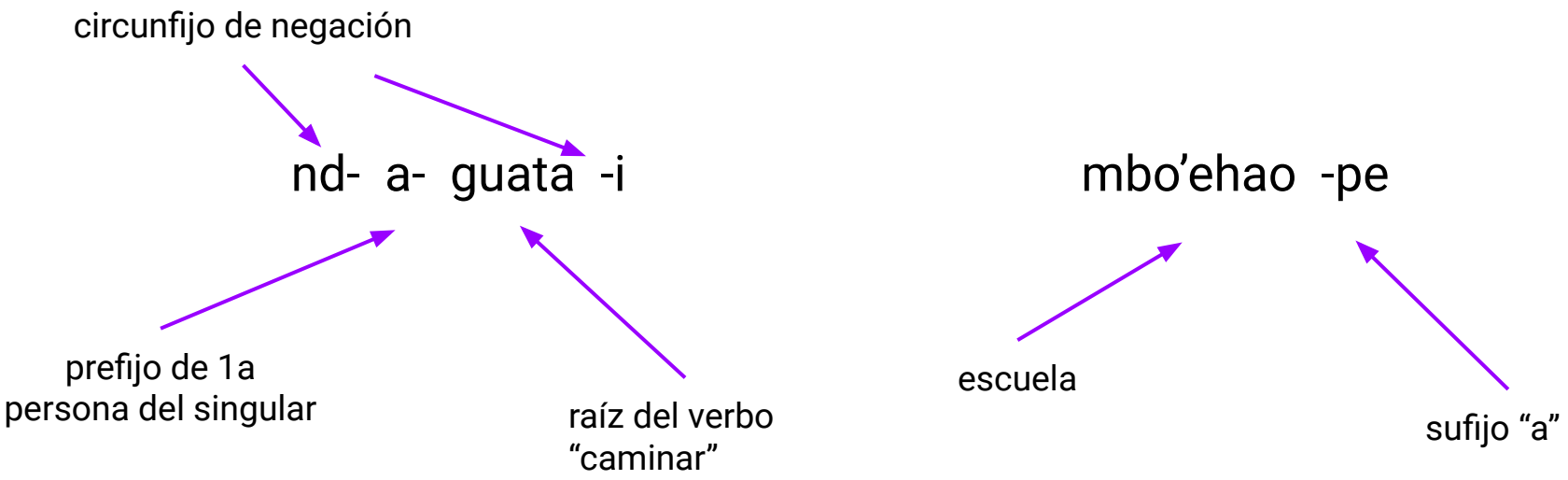
  Paraguay, Argentina, Bolivia, Brasil

- Contacto con español y otras lenguas europeas por alrededor de 500 años
- Hablada por toda la sociedad, no solo población indígena

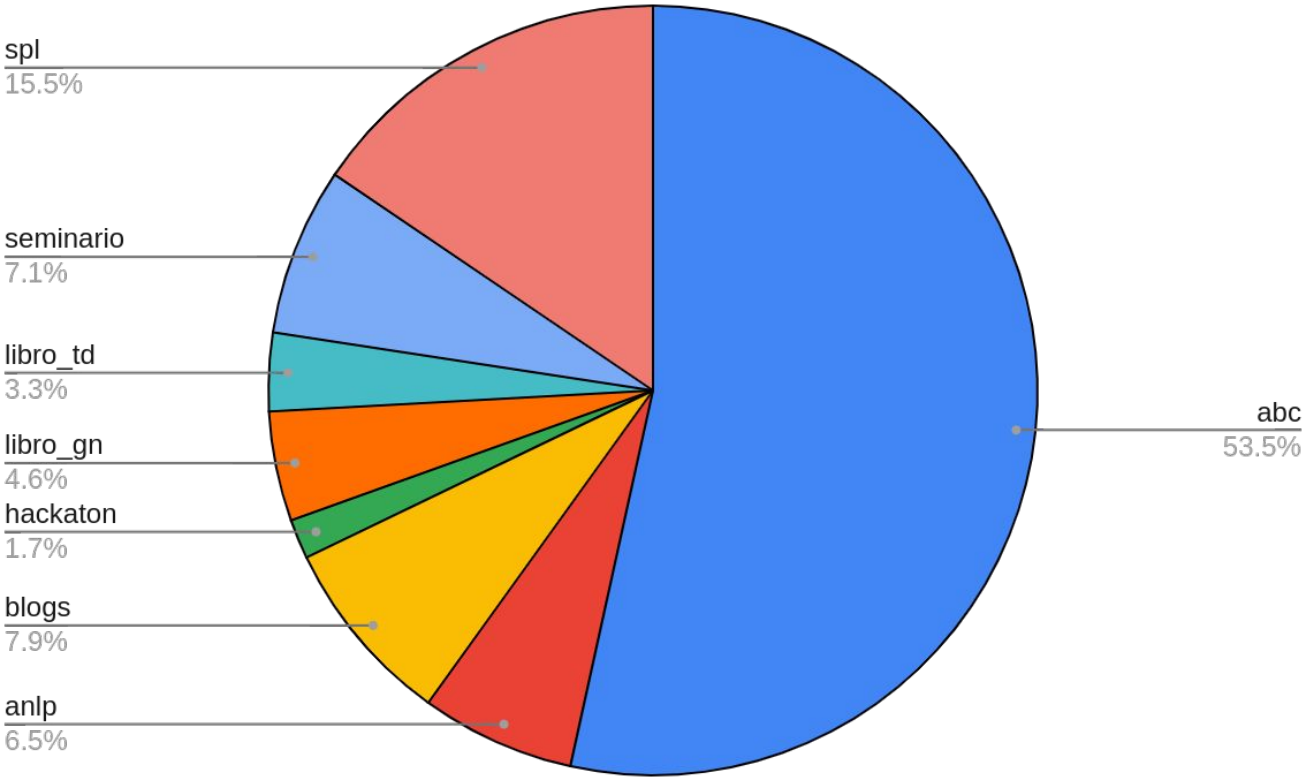
Sintaxis del Guaraní

ndaguatái mbo'ehaópe

No camino a la escuela



Corpus Jojajovai

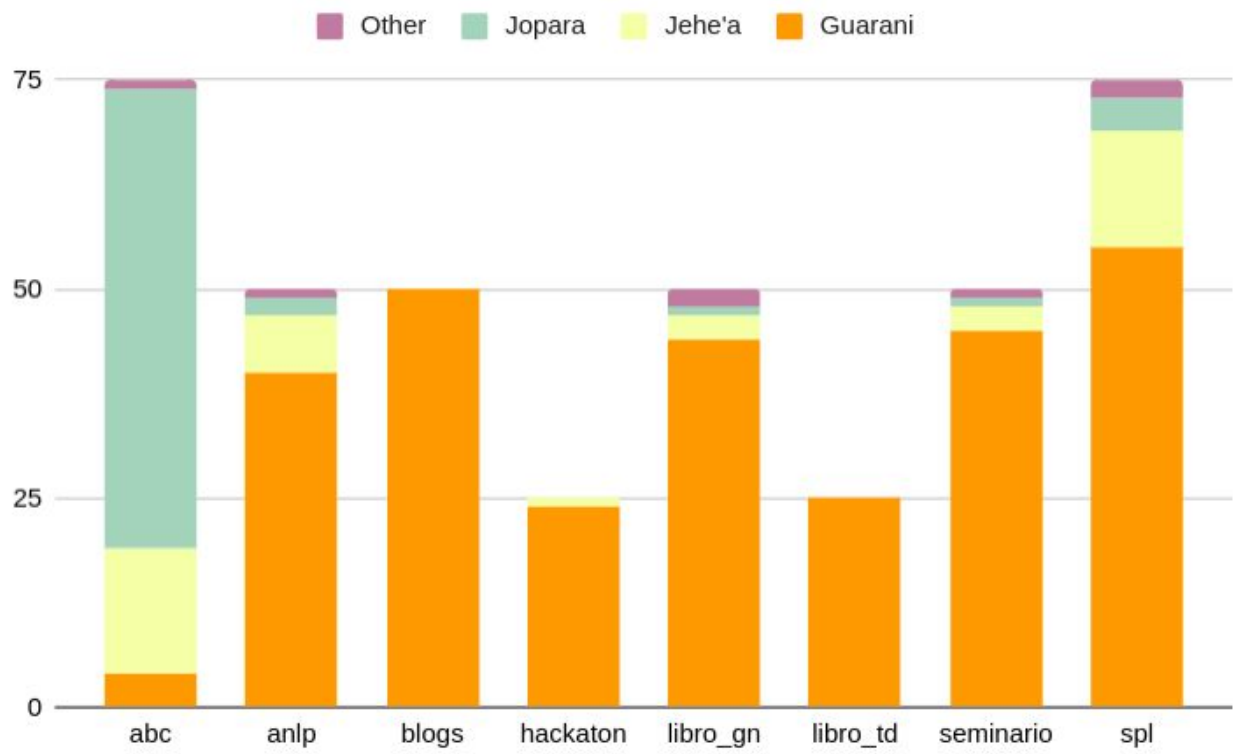


Alrededor de 30 mil pares en total

https://github.com/pln-fing-udelar/jojajovai

Jojajovai: A Parallel Guarani-Spanish Corpus for MT Benchmarking
Chiruzzo, Góngora, Alvarez, Giménez-Lugo, Agüero-Torales, Rodrıguez. 2022

Variedades



El code-switching entre guaraní y español es muy frecuente

Variedades

- *Embohasamína ko marandu umi rehayhuvévape…*

  Por favor, pasa este mensaje a las personas que estimas…

- *Afara orenunsiáta ko'êrõ*

  Afara renuncia mañana

- *Ojuhúma 52 allanamiento Argentina gotyo ha 21 detenido, 200.000 munición ha 2.500 fusil ojokóva.*

  En Argentina ya han realizado unos 52 allanamientos, 21 detenidos, 200.000 municiones con 2.500 fusiles secuestrados.

Alineación

Itaugua omokyre'ÿ "omopotî" Congreso

En Itauguá promueven "limpiar" el Congreso

Omopotîvo hikuái tetãme vicio política, ko'ã itaugüeño he'íva ombotovévo pokarême umi elemento omopotîva.

Con el propósito de limpiar al país de los vicios de la política, los itaugüeños expresan su repudio a los corruptos con elementos de limpieza.

Ko'ã 50 tapicha oñembyaty parroquía Virgen del Rosario plazoleta-pe ko distrito onemanifestávo político pokarême.

Unas 50 personas se encuentran en la plazoleta de la parroquía Virgen del Rosario de este distrito manifestando su repudio hacia los políticos corruptos.

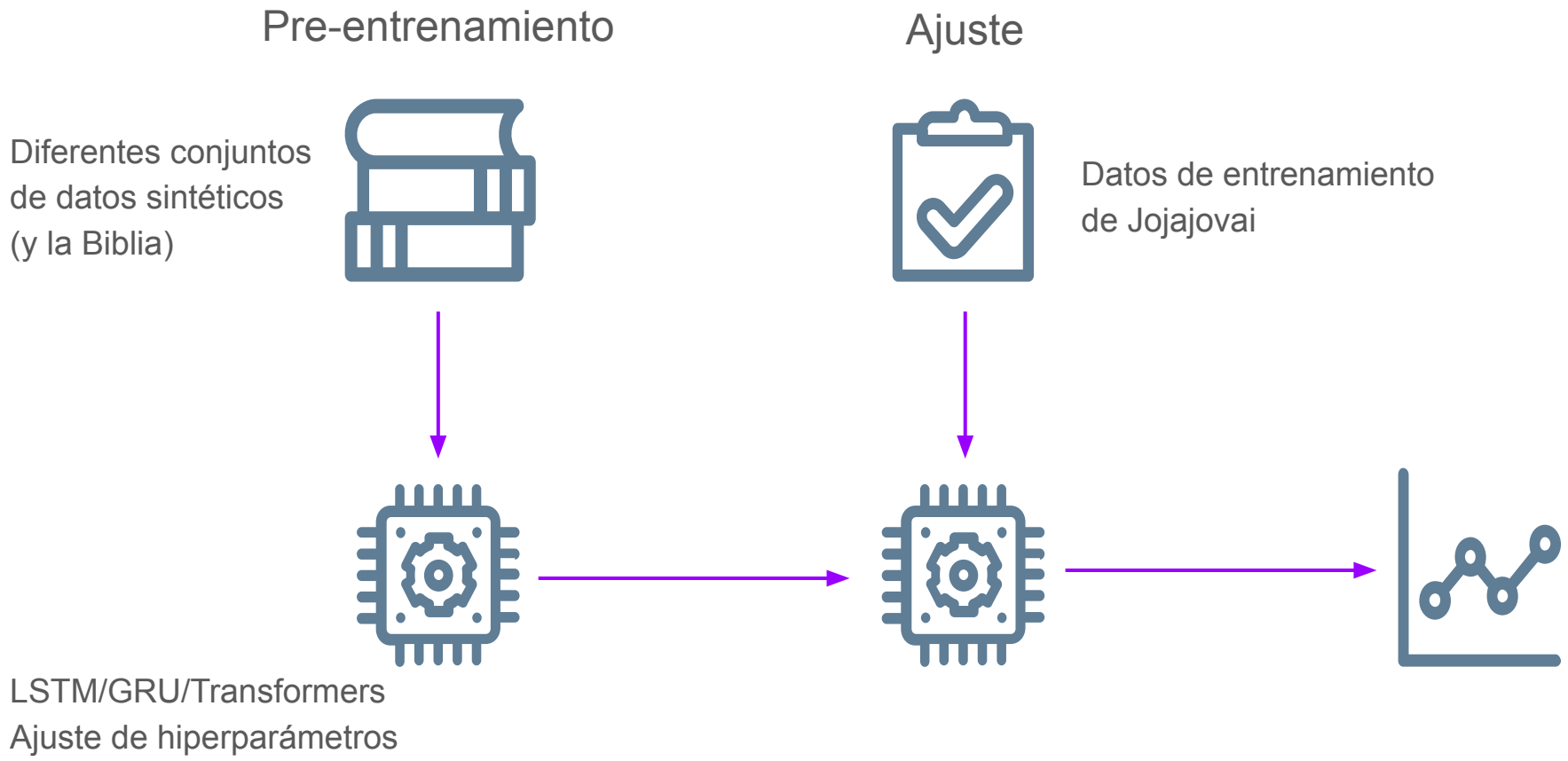
Durante el encuentro "limpiaron" un muñeco de un diputado acusado de corrupción.

Itaugüeño oipotáva ohechauka ipotîha itáva ha ikatúha paraguayo ikatu omopotî parlamento ha upévare hi'agui, orekóva yvyra orepasa haguã, trapo de piso, typycha ha lavandina oguahëva plaza parroquial rovái.

Los itaugüeños quieren demostrar que los paraguayos pueden limpiar el parlamento y por eso se acercaron, con palos de repasar, trapos de piso, escobas y lavandina hasta la plaza parroquial.

La manifestación fue acompañada por aplausos y vítores por parte de los vecinos.

Entrenamiento



Grammar-based Data Augmentation for Low-Resource Languages: The Case of Guarani-Spanish Neural Machine Translation
Lucas, Baladón, Pardiñas, Agüero-Torales, Góngora, Chiruzzo. 2024

Gramáticas de Rasgos

Spanish	Guarani
Part-of-speech: V (verb)	
Type: M (main), A (auxiliary), S (semi-auxiliary)	
Mood: I (indicative), S (subjunctive), M (imperative), P (participle), G (gerund), N (infinitive)	
Tense: P (present), I (past imperfective), F (future), S (past perfect), C (conditional)	
Person: 1 (first), 2 (second), 3 (third)	
Number: S (singular), P (plural)	
Gender: M (male), F (female), C (common)	Inclusiveness (only for first-person plural): I (inclusive), E (exclusive)
	Pronoun position (only for third-person plural): B (before the verb), A (after the verb), 0 (not relevant)
Transitivity: I (intransitive), T (transitive), D (ditransitive)	

Verbo

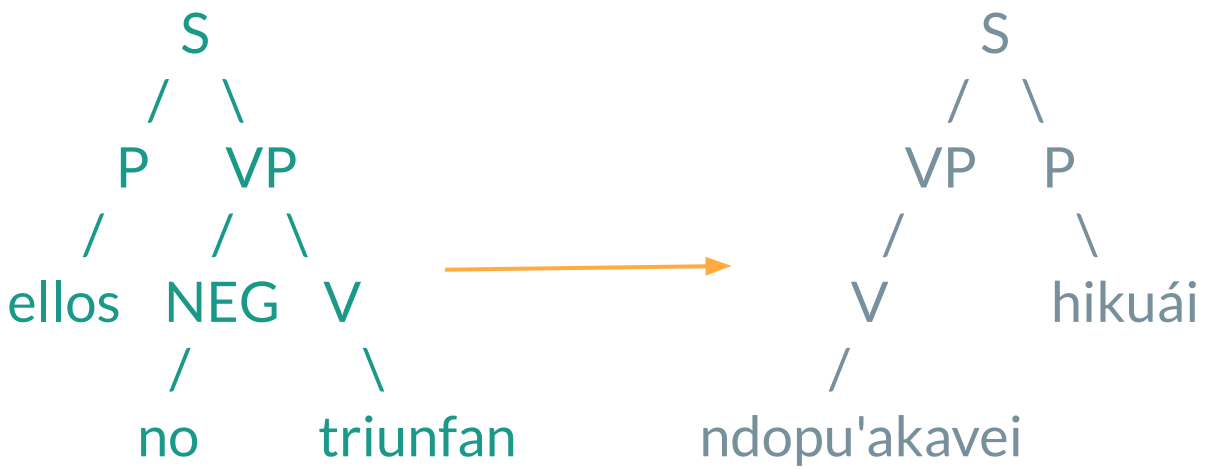
miro / amaña
mira / omaña
miré / amañakuri
miró / omañakuri
…

Spanish	Guarani
Part-of-speech: N (noun)	
Type: C (common), P (proper)	
Gender: M (male), F (female), C (common)	Gender: 0 (there is no gender for nouns)
Number: S (singular), P (plural), N (invariable)	
	Nasalization: N (nasal), O (oral)

Sustantivo

perro / jagua
perros / jaguakuéra
río / ysyry
amistad / ñoirũ
piedra / ita
piedras / itakuéra
…

Transferencia Sintáctica



Reglas de transferencia entre idiomas

```
"VP -> NEG V": [
        "VP[AGR='?a', POS='?p'] -> V[AGR='?a', NEG='1', POS='?p']"
    ]
```

Corpus Sintético

Generar oraciones aleatorias en español

Transferir a guaraní

Creamos tantos datos como queramos!

Set paralelo con más de 1M de palabras en guaraní

Pero…

Una piedra mintió / *Peteĩ ita oñe'ẽreikuri*

Él no apretará nuestras sopas / *Ha'e nomombeita ore jukysykuéra*

Corpus AnCora

Conjunto de noticias en español con 14,000 oraciones, 500,000 palabras

Detectar y transferir porciones traducibles con nuestra gramática

Josep y Angel Ortiz desbordan la ilusión en su mirada

Josep ha Angel Ortiz ochovi ilusión ima'eme

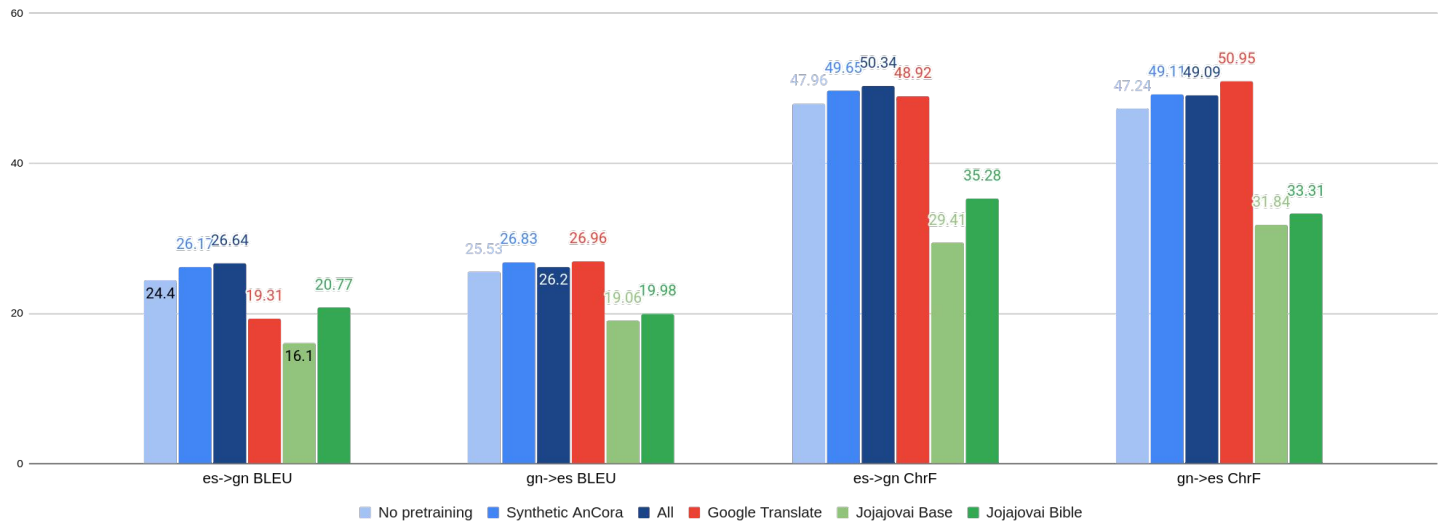
Oraciones más realistas

Genera un code-switching artificial!

Experimentos de Traducción Automática

Dir	Model	ChrF	BLEU
es→gn	jojajovai base	29.41	16.10
	jojajovai + bible	35.28	20.77
	Google	**48.92**	**19.31**
	no-pre	47.96	24.40
	pre-ancora	49.65	26.17
	pre-all	**50.34**	**26.64**

Dir	Model	ChrF	BLEU
gn→es	jojajovai base	31.84	19.06
	jojajovai + bible	33.31	19.98
	Google	**50.95**	**26.96**
	no-pre	47.24	25.53
	pre-ancora	**49.11**	**26.83**
	pre-all	49.09	26.20



Experimentos de Traducción Automática

Dir	Metric	Model	abc	anlp	blogs	hackathon	libro_gn	libro_td	seminario	spl
es→gn	ChrF	s2s - All	**58.76**	24.58	32.30	34.69	**30.16**	**39.38**	28.88	48.50
		s2s - AnCora	58.34	23.59	31.55	31.65	28.93	37.00	29.71	46.99
		Google Translate	56.61	**37.05**	**39.38**	**41.71**	28.82	28.15	**35.94**	**49.49**
		Jojajovai Base	37.44	14.10	21.35	20.02	16.98	24.10	19.83	37.49
		Jojajovai Bible	46.14	18.67	25.45	23.39	19.15	28.25	22.32	39.63
	BLEU	s2s - All	**31.45**	3.01	16.10	5.47	7.72	**10.49**	7.78	29.58
		s2s - AnCora	31.16	2.66	15.34	3.67	**10.86**	8.63	8.76	28.38
		Google Translate	23.56	**6.01**	**16.27**	**5.75**	8.30	3.09	**9.00**	**30.01**
		Jojajovai Base	18.24	0.75	7.73	3.09	3.44	5.15	3.02	20.73
		Jojajovai Bible	24.48	1.76	11.26	3.06	7.46	3.38	5.15	23.51
gn→es	ChrF	s2s - All	56.17	21.48	34.54	31.09	28.56	36.02	30.15	**48.61**
		s2s - AnCora	56.31	21.17	33.37	30.34	30.36	**37.69**	30.64	48.58
		Google Translate	**56.73**	**42.04**	**45.25**	**46.32**	**31.88**	29.62	**36.73**	44.49
		Jojajovai Base	40.25	14.77	24.71	19.35	17.15	24.02	23.15	41.68
		Jojajovai Bible	42.03	17.19	25.40	23.58	19.08	26.45	23.05	41.24
	BLEU	s2s - All	30.06	4.33	18.44	14.69	9.70	15.69	9.95	30.59
		s2s - AnCora	**30.83**	4.04	18.13	10.86	10.50	**18.41**	10.10	**31.21**
		Google Translate	30.81	**19.80**	**24.45**	**18.44**	**11.29**	9.02	**13.16**	23.58
		Jojajovai Base	20.84	1.55	11.89	6.45	5.40	10.25	6.37	25.93
		Jojajovai Bible	22.14	2.52	12.50	6.48	7.80	8.56	6.80	25.83

Bibliografía

- Jurafsky & Martin, 3rd Ed. (draft) - Capítulos 9, 13 y 2
- Clase de Traducción Automática de IntroPLN
- Papers…