

La Traducción Automática Estadística (TAE) ¿Pueden las máquinas traducir?

Adrià de Gispert

Machine Intelligence Lab, Dept. of Engineering, University of Cambridge

9 de noviembre del 2007
Seminario de la Unidad de Idiomas

Índice

- 1 **Introducción**
- 2 **Fundamentos TAE**
- 3 **Arquitectura básica**
- 4 **Reordenamiento**
- 5 **Conclusiones**

Ejercicio 1: TA basada en reglas

le otn aba noc azr al lañ

Ejercicio 1: TA basada en reglas

le

otn

aba

noc

azr

al

lañ

DT

NC

VP

PP

NC

DT

NC

Ejercicio 1: TA basada en reglas

le otn aba noc azr al lañ
 DT NC VP PP NC DT NC

ORIGEN [ANÁLISIS]	TRANSFERENCIA	DESTINO [SÍNTESIS]
DT NC → SN	SUJ SV → NP VP	NP → DT N
PP NC → SP		VP → V
VP PP SN → SV	DICCIONARIO	VP → V NP
SN → SUJ	otn → wind	VP → V NP ADV
SUJ SV → O	lañ → sign	
	azr → strength	
	aba/rat → to shake	

Ejercicio 1: TA basada en reglas

le otn aba noc azr al lañ
 DT NC VP PP NC DT NC

ORIGEN [ANÁLISIS]	TRANSFERENCIA	DESTINO [SÍNTESIS]
DT NC → SN	SUJ SV → NP VP	NP → DT N
PP NC → SP		VP → V
VP PP SN → SV	DICCIONARIO	VP → V NP
SN → SUJ	otn → wind	VP → V NP ADV
SUJ SV → O	lañ → sign	
	azr → strength	
	aba/rat → to shake	

The wind shook the sign violently ???

Ejercicio 2: TA estadística

ore euq al alr aes ed ort sér

Ejercicio 2: TA estadística

ore euq al alr aes ed ort sér

ORIGEN → DESTINO p	ORIGEN → DESTINO p
ore → I hope	alr → chats
ore → I wait	alr → he chats
ore → I wait for	aes → whether
ore euq → I hope that	aes → is
euq → which	aes → will be
al → the	aes ed ort sér → interests you
al alr → the talk	ed ort sér → of your interest
al alr → a mistake	

Ejercicio 2: TA estadística

ore euq al alr aes ed ort sér

ORIGEN → DESTINO p	ORIGEN → DESTINO p
ore → I hope	alr → chats
ore → I wait	alr → he chats
ore → I wait for	aes → whether
ore euq → I hope that	aes → is
euq → which	aes → will be
al → the	aes ed ort sér → interests you
al alr → the talk	ed ort sér → of your interest
al alr → a mistake	

I hope that the talk is of your interest ???

I hope that a mistake is of your interest ???

Ejercicio 2: TA estadística

ore euq al alr aes ed ort sér

ORIGEN → DESTINO	p	ORIGEN → DESTINO	p
ore → I hope		alr → chats	
ore → I wait		alr → he chats	
ore → I wait for		aes → whether	0.33
ore euq → I hope that		aes → is	0.33
euq → which		aes → will be	0.33
al → the		aes ed ort sér → interests you	
al alr → the talk	0.9	ed ort sér → of your interest	
al alr → a mistake	0.1		

Ejercicio 2: TA estadística

ore euq al alr aes ed ort sér

ORIGEN → DESTINO	p	ORIGEN → DESTINO	p
ore → I hope		alr → chats	
ore → I wait		alr → he chats	
ore → I wait for		aes → whether	0.33
ore euq → I hope that		aes → is	0.33
euq → which		aes → will be	0.33
al → the		aes ed ort sér → interests you	
al alr → the talk	0.9	ed ort sér → of your interest	
al alr → a mistake	0.1		

I hope that the talk ...

Ejercicio 2: TA estadística

ore euq al alr aes ed ort sér

ORIGEN → DESTINO	p	ORIGEN → DESTINO	p
ore → I hope		alr → chats	
ore → I wait		alr → he chats	
ore → I wait for		aes → whether	0.33
ore euq → I hope that		aes → is	0.33
euq → which		aes → will be	0.33
al → the		aes ed ort sér → interests you	
al alr → the talk	0.9	ed ort sér → of your interest	
al alr → a mistake	0.1		

I hope that the talk ...

is of your interest / whether of your interest ???

El modelo de lenguaje: definición

- Dada la siguiente secuencia de N palabras t : $w_1w_2w_3\dots w_N$
- El modelo de lenguaje se define como:

$$p_{TM}(t) = \prod_{n=1}^N p(w_n | w_1, \dots, w_{n-1})$$

- ▷ Cada palabra w_n depende de todas las anteriores

$$p(A|B) = \frac{C(A,B)}{C(B)}$$

El modelo de lenguaje: definición

- Dada la siguiente secuencia de N palabras t : $w_1w_2w_3\dots w_N$
- El modelo de lenguaje se define como:

$$p_{TM}(t) = \prod_{n=1}^N p(w_n | w_1, \dots, w_{n-1})$$

- ▷ Cada palabra w_n depende de todas las anteriores
- **Estimación del modelo es inviable** \Rightarrow Simplificación:

$$p_{TM}(t) = \prod_{n=1}^N p(w_n | w_{n-N+1}, \dots, w_{n-1})$$

- ▷ Modelo de Ngramas: sólo se consideran las $N-1$ palabras precedentes

$$p(A|B) = \frac{C(A,B)}{C(B)}$$

El modelo de lenguaje: estimación

- Para $N=3$, tenemos un modelo de trigramas

$$p(w_n | w_{n-1}, w_{n-2}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})}$$

El modelo de lenguaje: estimación

- Para $N=3$, tenemos un modelo de trigramas

$$p(w_n | w_{n-1}, w_{n-2}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})}$$

- Ejemplo: the talk ...

$$p(is | the, talk) = \frac{C(the, talk, is)}{C(the, talk, -)} = \frac{36}{135} = 0.267$$

$$p(whether | the, talk) = \frac{C(the, talk, whether)}{C(the, talk, -)} = \frac{1}{135} = 0.007$$

El modelo de lenguaje: estimación

- Para $N=3$, tenemos un modelo de trigramas

$$p(w_n | w_{n-1}, w_{n-2}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})}$$

- Ejemplo: the talk ...

$$p(is | the, talk) = \frac{C(the, talk, is)}{C(the, talk, -)} = \frac{36}{135} = 0.267$$

$$p(whether | the, talk) = \frac{C(the, talk, whether)}{C(the, talk, -)} = \frac{1}{135} = 0.007$$

- Se requieren estrategias de suavizado (smoothing)

$$p(will | the, talk) = \frac{C(the, talk, will)}{C(the, talk, -)} = \frac{0}{135} = 0.000 \quad !!!$$

El modelo de lenguaje: estimación

- Para $N=3$, tenemos un modelo de trigramas

$$p(w_n | w_{n-1}, w_{n-2}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})}$$

- Ejemplo: the talk ...

$$p(is | the, talk) = \frac{C(the, talk, is)}{C(the, talk, -)} = \frac{36}{135} = 0.267$$

$$p(whether | the, talk) = \frac{C(the, talk, whether)}{C(the, talk, -)} = \frac{1}{135} = 0.007$$

- Se requieren estrategias de suavizado (smoothing)

$$p(will | the, talk) = \frac{C(the, talk, will)}{C(the, talk, -)} = \frac{0}{135} = 0.000 \text{ !!!}$$

- Muchas aplicaciones: reconocimiento y síntesis de voz, traducción automática, análisis morfosintáctico, resumen automático, etc...

Breve historia de la TA

- años 50-60: sistemas de búsqueda palabra a palabra en diccionarios
- años 70-90: sistemas basados en conocimiento:
 - ▷ reglas de transferencia
 - ▷ interlingua
 - × desarrollo costoso, dominios restringidos, poca robustez
- años 90: sistemas estadísticos palabra a palabra
 - ▷ modelos de IBM (Brown et al., 1990)
 - ✓ desarrollo rápido, resultados equivalentes, mayor robustez y traducción del habla
- años 2000: sistemas estadísticos **basados en frases** (*phrase-based*)

Principio básico

“Cualquier secuencia de palabras de un idioma destino es una traducción posible de una frase dada en el idioma origen”

- Cada candidata tiene una **probabilidad** asociada
- Esta probabilidad se define **de acuerdo a un modelo estocástico** que describe el proceso de traducción
- Los parámetros del modelo se **estiman automáticamente a partir de textos paralelos** grandes

Traducción \equiv **encontrar** la candidata de mayor probabilidad

$$\hat{t} = \arg \max_t P(t|s)$$

Puntos importantes

- Unidad de traducción + conjunto de modelos estocásticos
 - ▷ palabra, frase, tupla, árbol, ...
 - ▷ **frase**: secuencia de palabras consecutivas

Puntos importantes

- Unidad de traducción + conjunto de modelos estocásticos
 - ▷ palabra, frase, tupla, árbol, ...
 - ▷ **frase**: secuencia de palabras consecutivas
- Textos paralelos (corpora de aprendizaje)
 - ▷ tamaño (*cuanto mayor, mejor*)
 - ▷ pares de idiomas
 - ▷ dominio (escalabilidad entre tareas)
 - ▷ *calidad* (datos con *ruido* o errores)

Puntos importantes

- Unidad de traducción + conjunto de modelos estocásticos
 - ▷ palabra, frase, tupla, árbol, ...
 - ▷ **frase**: secuencia de palabras consecutivas
- Textos paralelos (corpora de aprendizaje)
 - ▷ tamaño (*cuanto mayor, mejor*)
 - ▷ pares de idiomas
 - ▷ dominio (escalabilidad entre tareas)
 - ▷ *calidad* (datos con *ruido* o errores)
- Medidas de evaluación
 - ▷ manuales [**costosas**]: adecuación, fluencia, **HTER**,...
 - ▷ automáticas (comparación con referencias) [**pesimistas**]: **WER**, **BLEU**, **NIST**, **TER**,...

TAE basada en frases (phrase-based)

Enfoque de Máxima Entropía con entrenamiento discriminativo (combinación log-lineal):

$$\hat{t}_1^I = \arg \max_{t_1^I} \left\{ \sum_{m=1}^M \lambda_m h_m(s_1^J, t_1^I) \right\}$$

- **Modelo de traducción** basado en frases
 - ▷ fuente→destino **y también** destino→fuente
- **Modelo de lenguaje** del idioma destino
 - ▷ se compensa con bonificación de palabra
- **Modelos léxicos** (traducción palabra a palabra)
 - ▷ fuente→destino **y también** destino→fuente
- otros

- 1 **Introducción**
- 2 **Fundamentos TAE**
- 3 Arquitectura básica**
 - Alineado por palabras
 - Extracción de unidades
 - Modelos de la combinación
 - Decodificación
 - Resultados y discusión
- 4 **Reordenamiento**
- 5 **Conclusiones**

Alineado por palabras (1)

A partir de un corpus de frases paralelas (traducción mutua), queremos encontrar:

- relaciones translacionales entre palabras o conjuntos de palabras
- de forma completamente automática

Alineado por palabras (1)

A partir de un corpus de frases paralelas (traducción mutua), queremos encontrar:

- relaciones translacionales entre palabras o conjuntos de palabras
- de forma completamente automática
- Ejemplo:

AA	BB	HH	KK	NN		NN	
suerte	siempre	se	necesita			necesita	
<hr/>							
AA	BB	GG				HH	GG
mucha	suerte					mucha	siempre

Alineado por palabras (1)

A partir de un corpus de frases paralelas (traducción mutua), queremos encontrar:

- relaciones translacionales entre palabras o conjuntos de palabras
- de forma completamente automática
- Ejemplo:

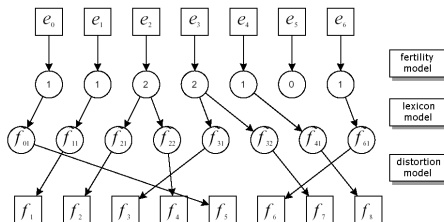


Alineado por palabras (2)

- Modelos de traducción + alineado de IBM (Brown et al., 1990)

$$Pr(\tau, \pi | e) = Pr(\phi | e) \cdot Pr(\tau | \phi, e) \cdot Pr(\pi | \tau, \phi, e)$$

- ▷ fertilidad: ¿cuántas palabras generará?
- ▷ traducción: ¿qué traducciones tiene?
- ▷ distorsión: ¿en qué posición se colocarán?



Alineado por palabras (2)

- × No existen algoritmos para estimar todos estos parámetros de forma exacta
- ✓ El **algoritmo EM** (Expectation - Maximization) ofrece una aproximación
 - ▷ Resultado: tablas de probabilidad (fertilidad, traducción, distorsión)
- ✓ El **algoritmo de Viterbi** permite encontrar el **alineado por palabras** más probable dadas las tablas y un par de frases (tools: GIZA++, MTK)

Alineado por palabras (2)

- ✗ No existen algoritmos para estimar todos estos parámetros de forma exacta
- ✓ El **algoritmo EM** (Expectation - Maximization) ofrece una aproximación
 - ▷ Resultado: tablas de probabilidad (fertilidad, traducción, distorsión)
- ✓ El **algoritmo de Viterbi** permite encontrar el **alineado por palabras** más probable dadas las tablas y un par de frases (tools: GIZA++, MTK)

miembro	+	.
Estado	+
otro	+	.
en	.	.	.	+	.	.	.
obras	.	+	+
de
aparejador	.	.	+
un	+
	a	building	contractor	.in	another	Member	State

Extracción de frases

- secuencia de hasta N palabras origen
- alineada con una secuencia de palabras destino de forma excluyente

miembro	+	.
Estado	+
otro	+	.	.
en	.	.	.	+	.	.	.
obras	.	+	+
de
aparejador	.	.	+
un	+
a		building	contractor
			in	another	Member	State	

Extracción de frases

- secuencia de hasta N palabras origen
- alineada con una secuencia de palabras destino de forma excluyente

miembro	+	.
Estado	+
otro	+	.	.
en	.	.	.	+	.	.	.
obras	.	+	+
de
aparejador	.	.	+
un	+

a	building	contractor	.in	another	Member	State
---	----------	------------	-----	---------	--------	-------

Phrases (N=4):

un # a
 un apdr. de obras # a bldg. contr.
 un apdr. de obras en # a bldg. contr. in
 apdr. de obras # bldg. contr.
 apdr. de obras en # bldg. contr. in
 apdr. de obras en otro # bldg. contr. in an.
 en # in
 en otro # in another
 en otro Est. miem. # in an. Mem. State
 otro Est. miem. # an. Mem. State
 Estado miembro # Member State

...

Modelo de traducción de frases

- Estimación de máxima verosimilitud (frecuencia relativa):

$$h_{MT}(s, t) = p(u_k | v_k) = \frac{C(u_k, v_k)}{\sum_x C(u_x, v_k)}$$

$$h_{MT2}(s, t) = p(v_k | u_k) = \frac{C(v_k, u_k)}{\sum_x C(v_x, u_k)}$$

- ▷ sin suavizado
- ▷ ambas direcciones son útiles

DESTINO → ORIGEN	C	$p(u v)$	ORIGEN → DESTINO	C	$p(v u)$
is → aes	48	0.96	aes → whether	25	0.25
is → euq	2	0.04	aes → is	70	0.70
			aes → will be	5	0.05

Modelo de lenguaje

- Modelo de Ngramas del idioma destino:

$$h_{ML}(s, t) = h_{ML}(t) = \log \prod_{k=1}^K p(w_k | w_{k-N+1}, \dots, w_{k-1})$$

- ✓ favorece las hipótesis con mayor fluencia (naturalidad)
- ✗ tiene preferencia por las hipótesis más cortas
- Modelo de bonificación de palabra: $h_{BP}(s, t) = h_{BP}(t) = K$
- ✓ compensa dicha preferencia fomentando la producción de palabras

Modelos léxicos

- modelo léxico origen→destino:

$$h_{LEX}(s, t) = \log \frac{1}{(I + 1)^J} \prod_{j=1}^J \sum_{i=0}^I p_{IBM}(t_j^n | s_i^n)$$

- modelo léxico destino→origen (análogo)

- ⇒ a partir de las tablas de traducción palabra a palabra (del alineado)
- ⇒ es una información complementaria al modelo de traducción

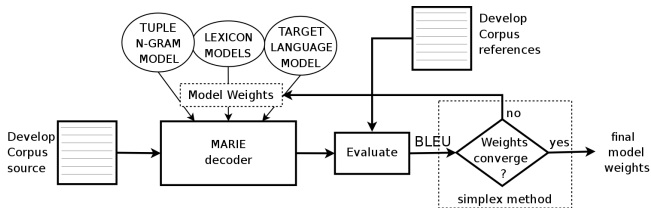
Ejemplo de decodificación

- Contribución de cada modelo (sólo 4) para 2 hipótesis de traducción:

PHRASE	MT	LEX	ML	BP	PHRASE	MT	LEX	ML	BP
always # siempre	4.22	0.06	1.85	-0.56	always # siempre	4.22	0.06	1.85	-0.56
has # tiene	2.13	0.88	1.38	-0.56	has # tiene	2.13	0.88	1.38	-0.56
a # una	1.27	0.51	0.98	-0.56	a # un	1.21	0.48	0.19	-0.56
difficult # difícil	2.49	0.14	2.42	-0.56	difficult # difícil	2.49	0.14	1.69	-0.56
role_to_play # papel	5.39	2.06	2.62	-0.56	role_to_play # papel	5.39	2.06	2.58	-0.56
COST/model	15.50	3.65	9.25	-2.8	COST/model	15.44	3.62	7.69	-2.8
FINAL COST	25.60				FINAL COST	23.95			

El decodificador

- El decodificador debe buscar la traducción más probable dados:
 - ▷ los 6 modelos anteriores
 - ▷ sus respectivos pesos
- algoritmos de **búsqueda en haz** (beam-search) basados en programación dinámica
- la **eficiencia** se regula mediante umbrales de poda en histograma y frecuencia
- existen varios decodificadores gratuitos (open-source):
Pharaoh, MARIE, MOSES,...



Resultados. Campañas de evaluación

- 1a evaluación del proyecto europeo TC-STAR (otoño 2004)
- Datos castellano↔inglés del Parlamento Europeo:
 - ▷ entrenam.: ~1.2M frases, ~35M palabras por idioma
 - ▷ texto editado y también salida de un reconocedor
 - ▷ des/test: 1k frases, 2 referencias manuales

	Spanish→English			English→Spanish		
	site	BLEU	NIST	site	BLEU	NIST
ASR (1-best)	RWTH	41.5	9.12	RWTH	38.7	8.73
	IBM	39.7	8.81	IBM	34.3	8.13
	UPC	37.7	8.56	UPC	33.8	8.00
	ITC-irst	34.7	7.97	UKA	33.0	7.94
	UKA	32.3	7.85	UPV	19.1	5.46
	UPV	16.0	4.35			
Text	UPC	53.3	10.55	UPC	46.2	9.65
	IBM	53.1	10.38	IBM	45.2	9.44
	ITC-irst	47.5	9.60	RWTH'	38.9	8.72
	RWTH'	46.1	9.68	UKA	37.6	8.46
	UKA	40.5	8.96	UPV	34.1	7.51
	UPV	32.7	6.80			

Análisis de errores

¿Qué tipo de errores se cometen?

Análisis de errores

¿Qué tipo de errores se cometen?

- selección léxica **errónea** (~24%):

2	This	3	is not	3	acceptable	3	either	1	.
	Esto		no es		acceptable		ni		.

Análisis de errores

¿Qué tipo de errores se cometen?

- selección léxica **errónea** (~24%):

2	This	3	is not	3	acceptable	3	either	1	.
	Esto		no es		aceptable		ni		.

- generación de forma verbal **incorrecta** (~28%):

2	What	3	is	2	basically	1	happening
	¿ Qué		es		básicamente		sucediendo

Análisis de errores

¿Qué tipo de errores se cometen?

- selección léxica **errónea** (~24%):

2	This	3	is not	3	acceptable	3	either	1	.
	Esto		no es		aceptable		ni		.

- generación de forma verbal **incorrecta** (~28%):

2	What	3	is	2	basically	1	happening
	¿ Qué		es		básicamente		sucediendo

- orden de las palabras **incorrecto** (~16%):

3	toda	3	la	3	sociedad	1	neerlandesa	2	,
	whole		of		society		Dutch		,

Análisis de errores

¿Qué tipo de errores se cometen?

- selección léxica **errónea** (~24%):

2	This	3	is not	3	acceptable	3	either	1	.
	Esto		no es		acceptable		ni		.

- generación de forma verbal **incorrecta** (~28%):

2	What	3	is	2	basically	1	happening
	¿ Qué		es		básicamente		sucediendo

- orden de las palabras **incorrecto** (~16%):

3	toda	3	la	3	sociedad	1	neerlandesa	2	,
	whole		of		society		Dutch		,

- omisiones** (~23%), falta de **concordancia** (~8%), etc.

3	Cuba	2	no	2	debe	2	cambiar	3	.
	Cuba		NULL		must		change		.

Discusión

- sistema básico: extrae mucha información de forma automática
- se obtienen resultados en función del par de idiomas implicados
 - ✓ catalán↔castellano (www.n-ii.org)
 - × chino↔inglés

Discusión

- sistema básico: extrae mucha información de forma automática
- se obtienen resultados en función del par de idiomas implicados
 - ✓ catalán↔castellano (www.n-ii.org)
 - ✗ chino↔inglés
- para el caso castellano↔inglés:
 - ▷ gran dificultad en la **generación de la forma verbal**
 - ▷ omisiones y errores de morfología básica (concordancias)
- ¿Puede ayudar la información morfosintáctica?

Discusión

- sistema básico: extrae mucha información de forma automática
- se obtienen resultados en función del par de idiomas implicados
 - ✓ catalán↔castellano (www.n-ii.org)
 - ✗ chino↔inglés
- para el caso castellano↔inglés:
 - ▷ gran dificultad en la **generación de la forma verbal**
 - ▷ omisiones y errores de morfología básica (concordancias)
- ¿Puede ayudar la información morfosintáctica?
- ¿qué hacer para pares de idiomas **no monótonos** ?
 - ▷ estrategias de reordenamiento de frases

- 1 Introducción
- 2 Fundamentos TAE
- 3 Arquitectura básica
- 4 Reordenamiento**
 - Búsqueda reordenada
 - Patrones de reordenamiento morfosintácticos
 - Comparativa y discusión
- 5 Conclusiones

Búsqueda reordenada

El decodificador puede incluir reordenamiento:

- se permiten cubrir las posiciones origen en cualquier orden
- el orden del destino no cambia (modelo lenguaje)
- **coste CPU elevado**, dos parámetros limitantes:
 - límite de distorsión (m): mayor distancia permitida de salto, en número de palabras
 - límite de reordemiento (j): mayor número de saltos permitidos en una frase
- Modelo de distorsión (coste de cada salto):

$$h_{DIST} = \sum_{k=1}^K d_k$$

donde d_k es la distancia entre la primera palabra origen de la k -ésima frase, y la última palabra origen de la frase anterior, más 1.

Resultados

IWSLT'06	Chi→Eng (dev123)			Chi→Eng (official)		
	BLEU	NIST	WER	BLEU	NIST	WER
monótono	0.3972	8.584	47.02	–	–	–
reordenamiento	0.4626	8.846	40.85	0.1863	5.571	68.04

IWSLT'06	Ara→Eng (dev123)			Ara→Eng (official)		
	BLEU	NIST	WER	BLEU	NIST	WER
monótono	0.5242	10.179	33.95	–	–	–
reordenamiento	0.5511	10.445	31.67	0.2323	6.238	62.71

EUROPAL	Eng→Spa			Spa→Eng		
	BLEU	NIST	WER	BLEU	NIST	WER
monótono	0.4775	9.73	41.89	0.5532	10.70	34.28
reordenamiento	0.4780	9.81	41.94	0.5447	10.63	34.95

tamaño entren.: Chi: 40k, Ara: 20k

parámetros: $m=5$, $j=3$

dev123: dev'04 + test'04 + test'05 (16 refs)

Discusión

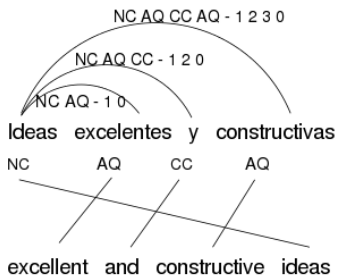
- Es un enfoque de fuerza bruta
 - ▷ espera que el modelo de lenguaje elija el orden adecuado
- Incorpora el reordenamiento con resultados positivos para tareas de chino/árabe → inglés
- Sin embargo, incluso con limitaciones, los **costes computacionales son elevados**
- **No mejora** en castellano↔inglés

Discusión

- Es un enfoque de fuerza bruta
 - ▷ espera que el modelo de lenguaje elija el orden adecuado
- Incorpora el reordenamiento con resultados positivos para tareas de chino/árabe → inglés
- Sin embargo, incluso con limitaciones, los **costes computacionales son elevados**
- **No mejora** en castellano↔inglés
- Enfoque alternativo: (Crego and Mariño, 2006)
 - ▷ Permitir el reordenamiento **sólo en ciertas situaciones** cuando la frase de test tenga las mismas propiedades que las frases reordenadas del entrenamiento
 - ▷ Extender el grafo de búsqueda monótono con caminos reordenados en función de las **secuencias de etiquetas morfológicas**

Extracción de patrones de reordenamiento

- A partir del alineado por palabras y las secuencias de etiquetas POS
- Buscamos alineados cruzados en el entrenamiento
- Filtramos patrones poco frecuentes:
 - min ocurrencias (1000), min reo/mon ratio (0.2)



Patrones castellano→inglés

Sólo se extraen 17 patrones:

Patrón	Num.oc.	Ejemplo
NC AQ 1 0	877,580	preguntas serias
NC RG 1 0	54,968	actividades aparentemente
AQ AQ 1 0	46,509	medioambientales europeas
RN VM 1 0	45,777	no promuevan
NC AQ AQ 2 1 0	35,661	decisiones políticas delicadas
NC RG AQ 1 2 0	32,887	ideas muy sencillas
NC AQ CC AQ 1 2 3 0	27,119	programa ambicioso y realista
RG VA 1 0	9,824	ahora habíamos
AQ RG 1 0	8,701	suficiente todavía
RG VS 1 0	5,043	supuestamente somos
VM PP 1 0	4,769	estar ustedes
NC CC NC AQ 3 0 1 2	3,355	mezquitas y centros islámicos
AQ RG AQ 1 2 0	2,777	europea más sólida
NC RG AQ CC 1 2 3 0	2,226	ideas muy sencillas y
NC AQ RG AQ 2 3 1 0	1,971	control fronterizo más estricto
NC RG RG 1 2 0	1,473	texto mucho más
NC RG AQ CC AQ 1 2 3 4 0	1,406	ideas muy sencillas y elementales

Patrones de inglés→castellano

Se extraen 29 patrones:

Patrón	Num.oc.	Ejemplo
JJ NN 1 0	784,572	Italian parliamentarians
NN NN 1 0	472,809	food scandals
MD RB 1 0	55,226	will actively
JJ JJ 1 0	40,825	liberal European
JJ NN NN 2 1 0	31,395	Belgian Supreme Court
CC JJ NN 2 0 1	30,287	and pro-European forces
JJ JJ NN 2 1 0	29,834	American occupying forces
RB JJ NN 2 0 1	29,379	absolutely rigid control
JJ CC JJ NN 3 0 1 2	27,795	political and symbolic issues
NN PO 1 0	19,216	Barroso 's
NN PO NN 2 0 1	16,493	children 's questions
PO NN 1 0	13,875	's problems
NN JJ 1 0	13,359	EU military
CC NN NN 2 0 1	12,642	and Mrs Zimmer
NN CC NN NN 3 0 1 2	10,559	Lambert and Mrs Zimmer
NN JJ NN 2 1 0	6,351	EU military operation
NN NN PO 2 0 1	3,860	President Bush 's
NN PO JJ 2 0 1	3,576	Bush 's foreign
NN NN PO NN 3 0 1 2	2,684	European Union 's appreciation
JJ CC NN NN 3 0 1 2	2,656	political and policy complexion
NN PO JJ NN 3 2 0 1	2,013	Union 's targeted sanctions
...

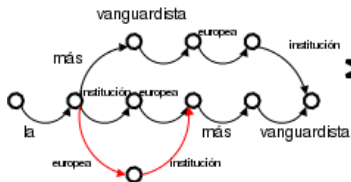
Extensión del grafo de búsqueda

- Cuando se encuentra un patrón en el test, se añade un camino reordenado

DT NC AQ RG AQ
 la institución europea más vanguardista

NC AQ -> AQ NC
 NC AQ RG AQ -> RG AQ NC AQ

la # the
 institución europea # european institution
 institución # institution
 europea # european
 más # most
 vanguardista # avant-garde



the | most | avant-garde | european | institution

the | european institution | most | avant-garde

Resultados (cont.)

IWSLT'06	Chi→Eng (dev123)			Chi→Eng (official)		
	BLEU	NIST	WER	BLEU	NIST	WER
monótono	0.3972	8.584	47.02	–	–	–
reordenamiento	0.4626	8.846	40.85	0.1863	5.571	68.04
patrones reo.	0.4463	8.993	43.55	0.1834	5.740	69.76

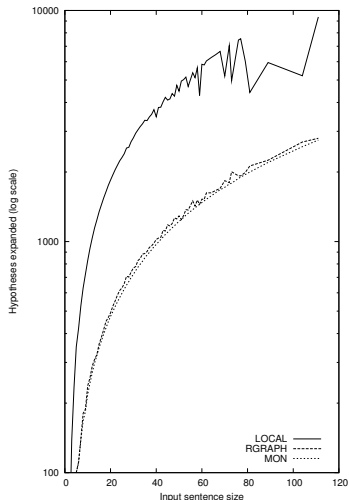
IWSLT'06	Ara→Eng (dev123)			Ara→Eng (official)		
	BLEU	NIST	WER	BLEU	NIST	WER
monótono	0.5242	10.179	33.95	–	–	–
reordenamiento	0.5511	10.445	31.67	0.2323	6.238	62.71
patrones reo.	0.5471	10.412	31.79	0.2267	6.135	63.10

EUROPARL	Eng→Spa			Spa→Eng		
	BLEU	NIST	WER	BLEU	NIST	WER
monótono	0.4775	9.73	41.89	0.5532	10.70	34.28
reordenamiento	0.4780	9.81	41.94	0.5447	10.63	34.95
patrones reo.	0.5006	10.00	39.73	0.5611	10.76	33.59

parámetros: $m=5$, $j=3$

dev123: dev'04 + test'04 + test'05 (16 refs)

Comparativa de eficiencia



- ▷ el grafo de búsqueda extendido con patrones de reordenamiento posee un número de nodos similar al caso monótono
- ▷ requiere muchos menos recursos que la búsqueda reordenada de fuerza bruta

Discusión

- mejora significativamente en castellano↔inglés
- eficiencia similar al caso monótono
- no ayuda para chino/árabe→inglés
 - ▷ se requiere reordenamiento de larga distancia
 - ▷ las secuencias de etiquetas Part-Of-Speech no generalizan
 - ▷ se necesitan patrones que tengan en cuenta *sintaxis*

NIST'06	Chi→Eng (test'05)			Chi→Eng (official)	
	BLEU	NIST	WER	BLEU	NIST
reordenamiento	0.2097	7.500	73.69	0.2071	7.217
patrones reo.	0.2120	7.601	72.85	0.2075	7.231

Comentarios finales

- la TAE permite aprender de forma automática el proceso de traducción
 - ▷ a partir de corpora paralelos (tamaño, calidad, etc.)
 - ▷ utilizando técnicas de estimación generales
 - ▷ requiere algoritmos eficientes (coste computacional)
- cada par de idiomas requiere de enfoques ligeramente diferentes
- se cometen errores de orden y morfología básicos
- actualmente, se trabaja en enfoques híbridos
 - ▷ incorporación de información morfológica
 - ▷ aprendizaje automático de estructuras sintácticas
- integración viable con el reconocimiento del habla

¿Preguntas y/o comentarios?

Adrià de Gispert
ad465@eng.cam.ac.uk

TIME FLIES LIKE AN ARROW