A Reordering Framework for SMT

Josep M. Crego
TALP Research Center - UPC

Outline

- Motivation
- Reordering Patterns
- Reordered Input Graph
- Baseline
- Experiments
- Conclusions
- Further Work

Motivation I

- In general, spa-eng does not contain important disparities in word order...

suspect products → productos sospechosos

American occupying forces → fuerzas de ocupación americanas

Parliament 's good name → buen nombre del Parlamento

Parlamento Europeo → European Parliament

situación claramente insatisfactoria → clearly unsatisfactory situation

temas importantes y complejos → important and complex issues

...

Motivation II

- Reordering is not needed:

  | Parlamento Europeo | ⇒ European Parliament

- Reordering is needed: SPARSENESS

  | situación | claramente | insatisfactoria | ⇒ situation clearly unsatisfactory

  it should be: clearly unsatisfactory situation

  | temas | importantes | y | complejos | ⇒ issues important and complex

  it should be: important and complex issues

Motivation III

- When reordering is applied HUGE computational expenses:

  | temas | y | importantes | complejos | ⇒ issues and important complex

  | temas | importantes | complejos | y | ⇒ issues important complex and

  | temas | complejos | y | importantes | ⇒ issues complex and important

  | importantes | y | complejos | temas | ⇒ important and complex issues

  | importantes | complejos | temas | y | ⇒ important complex issues and

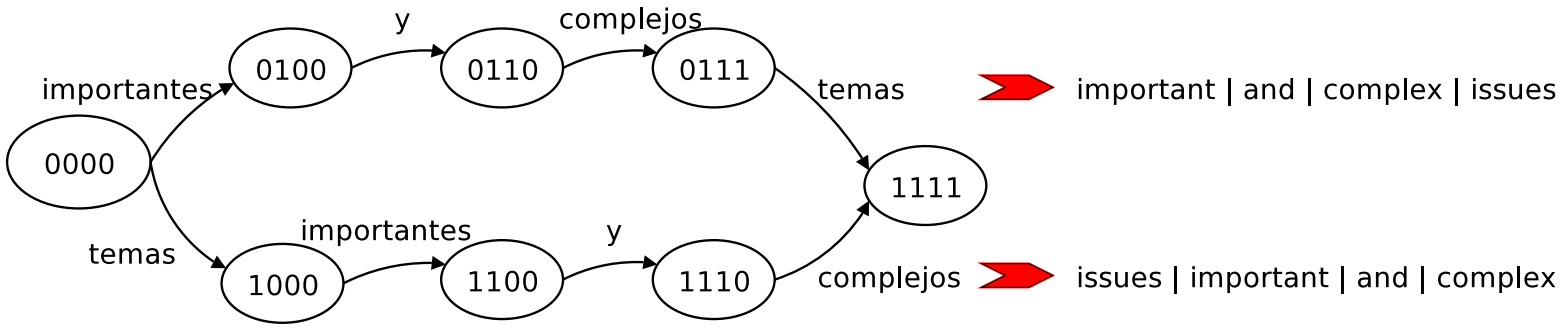
  ...

- Using 'local' constraints (ITG, IBM, maxJumps, ...)
  still NP-complete and BAD model

Motivation IV

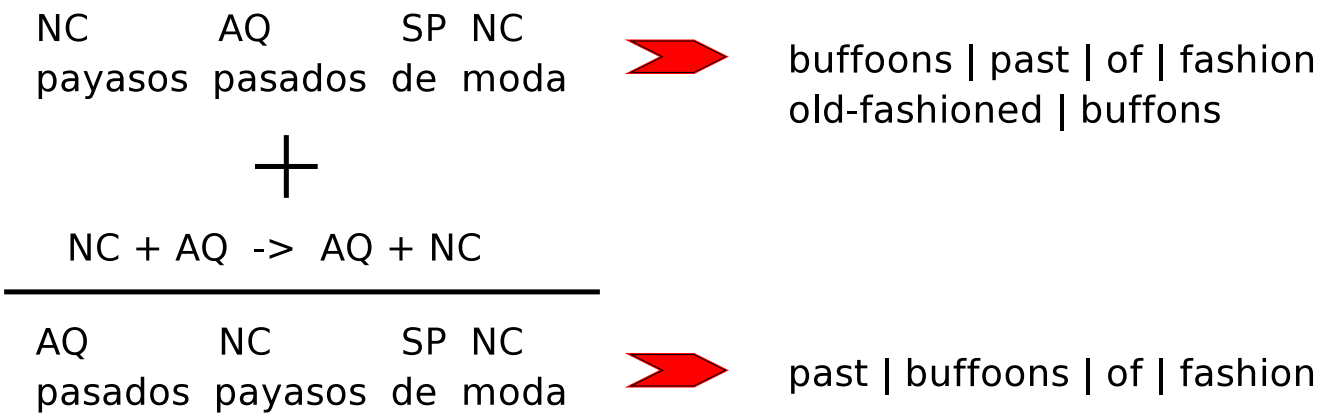
How to alleviate the computational and modeling problem of reordering ?

- Pre-computing reorderings using linguistic information (linguistically-based)
- Taking the final decision using the whole models in the search (fully-informed decision)



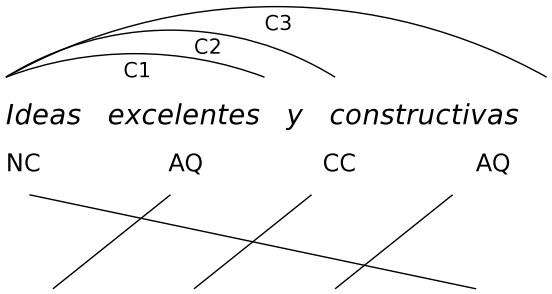
Motivation V

- Hard decisions of reorderings in preprocess introduce unrecoverable errors



Reordering Patterns

- Using word-to-word alignments (union) and source POS tags (first 2 characters), Freeling (es) and TNT (en).
- Crossings produced in word-to-word alignments (train)



C1: NC AQ - 1 0
C2: NC AQ CC - 1 2 0
C3: NC AQ CC AQ - 1 2 3 0

Reordering Patterns

- Pruning out patterns
  - max pattern diffsize: 3 words
  - max pattern size: 5 tokens
  - min pattern occurrences: 1000
  - min average: 0.2
- Spa-to-Eng: 17
- Eng-to-Spa: 29

Reordering Patterns II

Pattern	Insts.	Example
NC RG AQ CC AQ 1 2 3 4 0	1,406	ideas muy sencillas y elementales
NC AQ CC AQ 1 2 3 0	27,119	programa ambicioso y realista
NC AQ RG AQ 2 3 1 0	1,971	control fronterizo más estricto
NC CC NC AQ 3 0 1 2	3,355	mezquitas y centros islámicos
NC RG AQ CC 1 2 3 0	2,226	ideas muy sencillas y
AQ RG AQ 1 2 0	2,777	europea más sólida
NC AQ AQ 2 1 0	35,661	decisiones políticas delicadas
NC RG AQ 1 2 0	32,887	ideas muy sencillas
NC RG RG 1 2 0	1,473	texto mucho más
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
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

Reordering Patterns III

Pattern	Insts.	Example
JJ CC JJ NN 3 0 1 2	27,795	political and symbolic issues
NN CC NN NN 3 0 1 2	10,559	Lambert and Mrs Zimmer
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
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
NN PO NN 2 0 1	16,493	children 's questions
CC NN NN 2 0 1	12,642	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
JJ NN 1 0	784,572	Italian parliamentarians
NN NN 1 0	472,809	monster Berlusconi
MD RB 1 0	55,226	will actively
JJ JJ 1 0	40,825	liberal European
NN PO 1 0	19,216	Barroso 's
PO NN 1 0	13,875	's problems
NN JJ 1 0	13,359	EU military

Reordered Input Graph

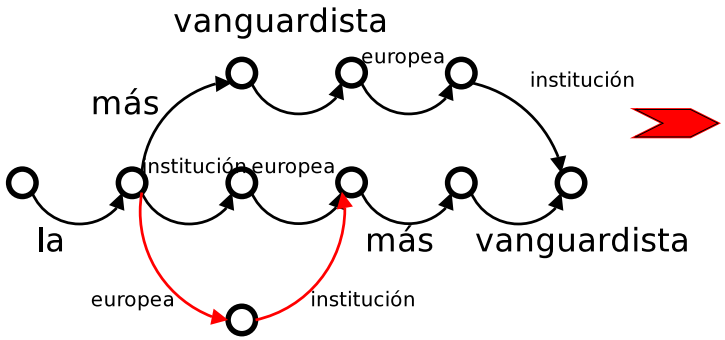
- Used MARIE decoder (Version 1.3), with the ability to read input graphs which encode the allowed reorderings
- Whenever a reordering pattern is found in the test file, the (reordered) path is added.
- Reordering paths are not included in the input graph when the translation unit (tuple) exists

Reordered Input Graph II

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

Baseline (Ngram-based SMT)

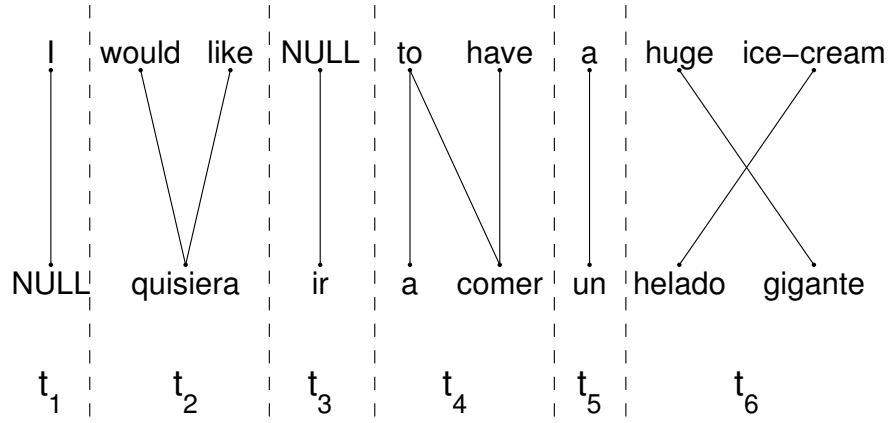
- Maximization of a log-linear combination of feature functions:

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

  - a target language model
  - a word bonus model
  - a source-to-target lexicon model
  - a target-to-source lexicon model
  - a tagged target language model (using POS tags)
  - a translation language model

Baseline II

- Given a word alignment (union), monotonic segmentation of each bilingual sentence.
- Language model of a particular language of bilingual units (tuples).



Experiments

- Euparl corpus (Spanish-English), using 2005 dev set and 2004 test set.

Euparl ES-EN	sent	words	voc	POSvoc
Train set				
English	1.28 M	34.9 M	106 k	44
Spanish		36.6 M	153 k	328
Dev set				
English	735	18,764	3,193	41
Spanish	430	15,332	3,217	181
Test set				
English	1,094	26,917	3,958	42
Spanish	840	22,774	4,081	196

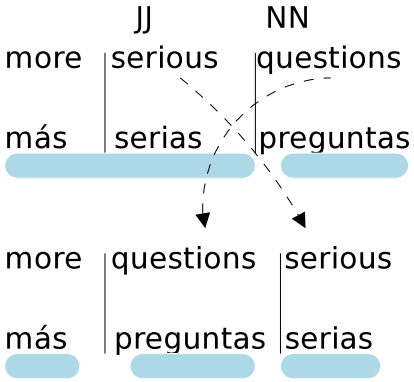
Experiments II

- Consistent improvements in BLEU (dev and test)
- The rest of measures remain very similar

Conf	bleu'	bleu	nist	mwer	per
Spanish-to-English					
base	0,529	0,552	10,69	34,40	25,32
rgraph	0,533	0,556	10,70	34,23	25,50
English-to-Spanish					
base	0,481	0,480	9,84	41,18	31,11
rgraph	0,490	0,485	9,81	41,15	31,87

Experiments III

- Problem Context in tuples (bad modeling)



- Solution Reordering model as Ngram LM of reordered source tags (POS).
  - Reorder source POS tags in train when crossing detected
  - Learn a tagged source Ngram LM

Experiments IV

- All measures correlated (PER does not account for reorderings)
- Consistent improvements in dev and test

Conf	bleu'	bleu	nist	mwer	per
Spanish-to-English					
base	0,529	0,552	10,69	34,40	25,32
rgraph	0,533	0,556	10,70	34,23	25,50
pos	0,539	0,564	10,75	33,75	25,41
English-to-Spanish					
base	0,481	0,480	9,84	41,18	31,11
rgraph	0,490	0,485	9,81	41,15	31,87
pos	0,491	0,489	9,91	40,29	31,27

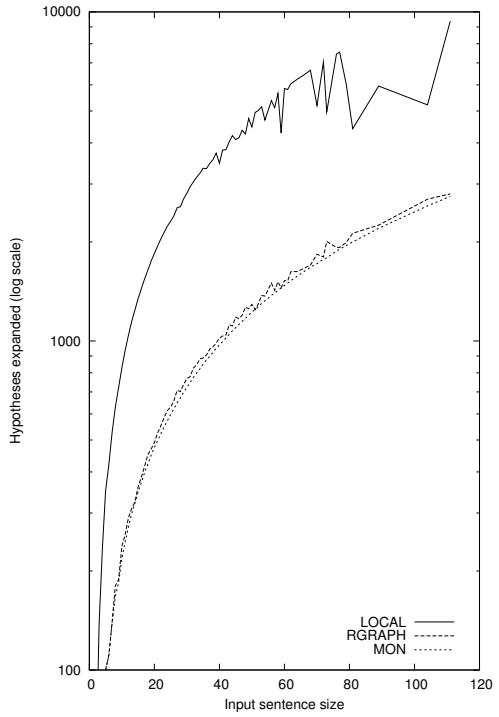
Experiments V

- Human evaluation of the Spanish-to-English test task.

Pattern	Test	Swap	Error	Example
NC RG AQ CC AQ 1 2 3 4 0	1	1	0	ideas muy sencillas y elementales
NC AQ CC AQ 1 2 3 0	23	17	2	programa ambicioso y realista
NC AQ RG AQ 2 3 1 0	4	1	0	control fronterizo más estricto
NC CC NC AQ 3 0 1 2	12	6	3	mezquitas y centros islámicos
NC RG AQ CC 1 2 3 0	2	0	0	ideas muy sencillas y
AQ RG AQ 1 2 0	7	2	1	europea más sólida
NC AQ AQ 2 1 0	24	18	3	decisiones políticas delicadas
NC RG AQ 1 2 0	35	26	1	ideas muy sencillas
NC RG RG 1 2 0	3	3	2	texto mucho más
NC AQ 1 0	142	110	16	preguntas serias
NC RG 1 0	47	7	7	actividades aparentemente
AQ AQ 1 0	40	4	2	medioambientales europeas
RN VM 1 0	2	1	1	no promuevan
RG VA 1 0	2	1	0	ahora habíamos
AQ RG 1 0	21	4	*2*	suficiente todavía
RG VS 1 0	1	1	0	supuestamente somos
VM PP 1 0	13	12	2	estar ustedes
Total (17)	379	214	42	

Experiments VI

- Hypotheses expanded in the search

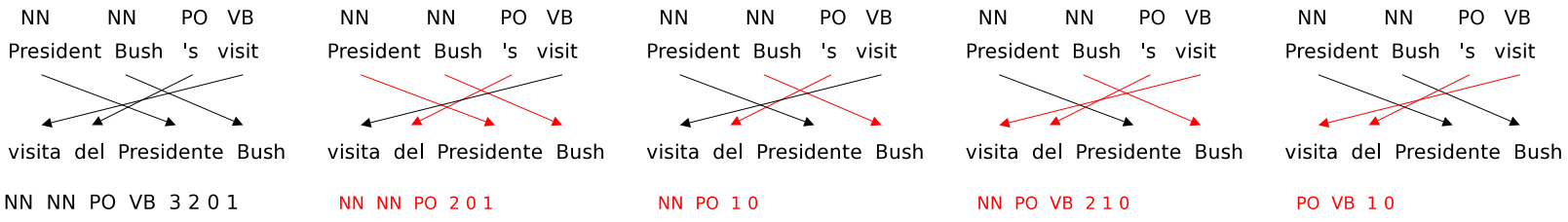


Conclusions

- Translation accuracy improved when allowing for reordering
- Despite of using 'bad' patterns, the decoder (models) shows the ability of discarding bad reorderings
- Clear efficiency improvement (pseudo-monotone decoding)

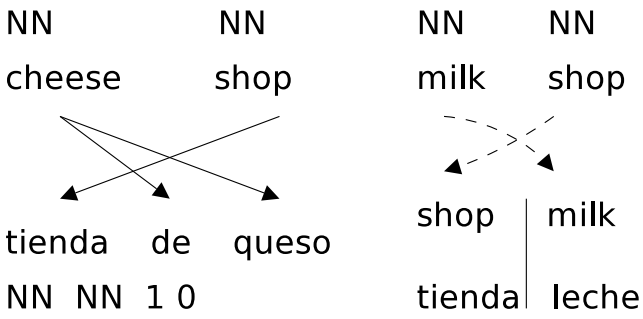
Further Work

- Use the approach in different tasks:
  - zh-en, ar-en, ...
- Better pattern extraction:
  - No inner patterns



Further Work

- Better pattern extraction:
  - Lexicalized patterns
  - New pattern structures:



NN NN 1 NULL 0 ???