Short Range Re-ordering using POS Tags

Hieu Hoang

Contents

Introduction

Motivation

- Re-ordering is a Challenge in Machine Translation
- Two separate problems
  - Long-range re-ordering
  - Short-range re-ordering
- Address short-range re-ordering

Examples of Long Range Re-ordering

Re-Order across Clauses

Source: Wir haben daher nicht für diesen Bericht gestimmt.

Ref: We did not, consequently, vote in favour of this report.

Moses: thus , we have not voted for this report .

Source: Er ist bisher existent, wird aber möglicherweise durch bürokratische Intransparenz und den offensichtlichen, aber nicht öffentlichen Widerstand einiger Regierungen gefährdet.

Ref: It exists so far, but will possibly be jeopardised by a bureaucratic lack of transparency and the obvious, but not public resistance of certain governments.

Moses: he is so far existent , but possibly by bureaucratic intransparenz and the obvious , but not public opposition of some governments jeopardized .

Short Range Re-ordering

Not a Solved Problem for Phrase-Based Model

union européenne → European Union
nom adj

minorités insignifiantes → insignificant minorities
nom adj

Moses: minorities insignificant

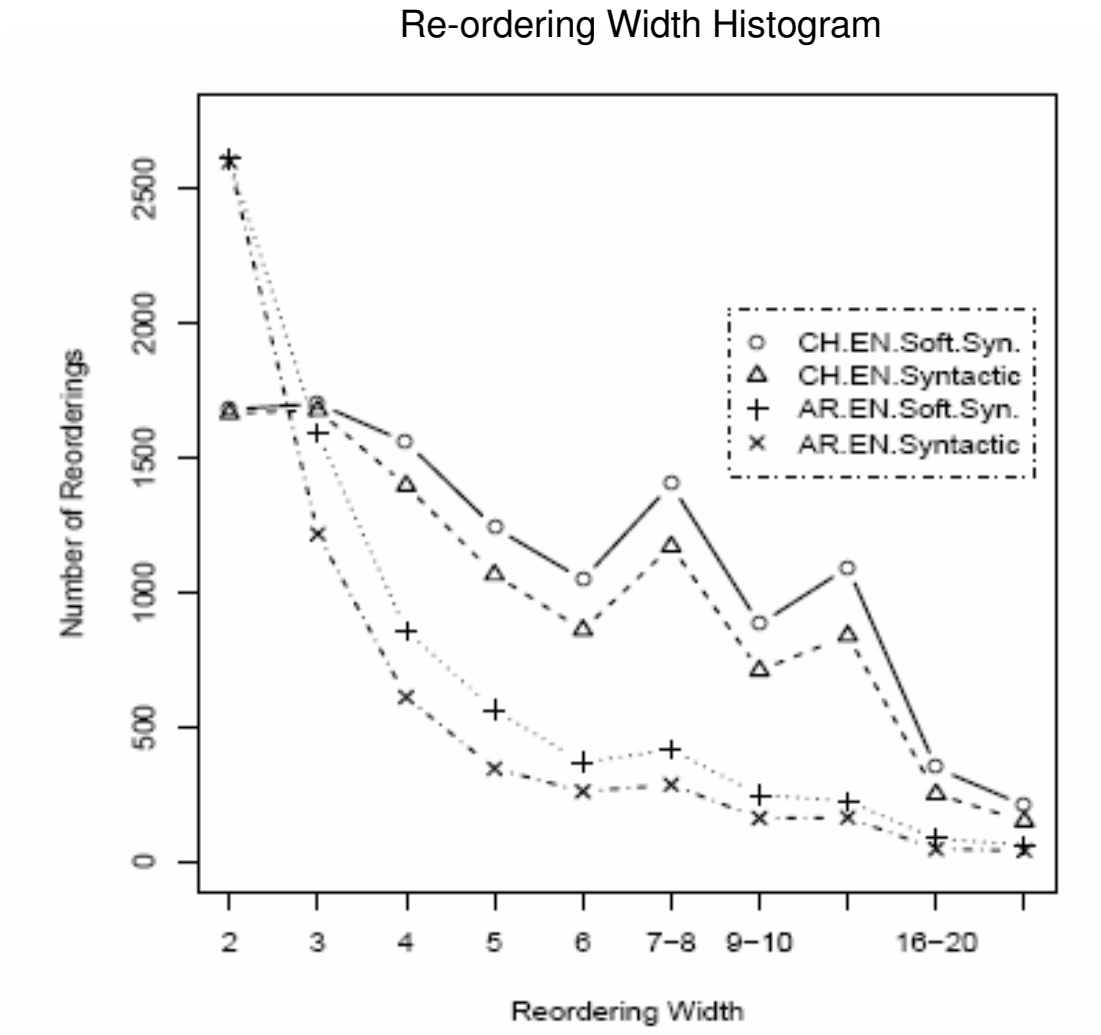
difficultés économiques et sociales → economic and social problems
nom adj kon adj

Moses: difficulties economic and social

justifications militaires importantes → important military justifications
nom adj adj

Moses: justifications military important

Short-range Re-ordering is Important



Manually aligned corpus

3380 Chinese-English
4337 Arabic English

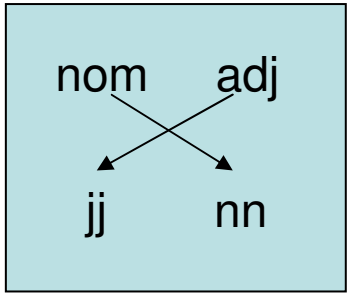
Corpus: Subset of Gale corpus

(Birch et al, unpublished, 2008)

Capturing POS Re-ordering from alignment

Example:

minorités insignifiantes → *insignificant minorities*
nom adj → jj nn



Translations for 'nom adj'

Source	Target	Num of exampl	%
nom adj	jj nn	7362	32%
nom adj	jj nns	3547	15%
nom adj	nn	1027	4%
nom adj	nnp nnp	867	4%
nom adj	jj	747	3%
nom adj	nn nn	733	3%
nom adj	nns	546	2%

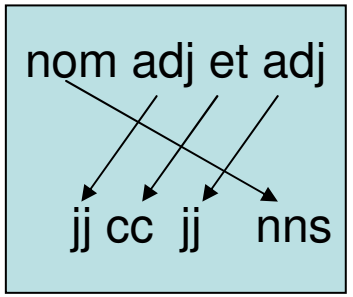
GIZA++ alignments for 'nom adj' → 'jj nn'

Alignment	Num of examples	%
1-0 0-1	6704	29%
1-0 0-1 1-1	168	1%
0-0 1-0 0-1	156	1%
0-0 1-0 1-1	75	0%

Capturing POS Re-ordering from alignment

Example: difficultés économiques et sociales → economic and social difficulties

nom adj kon adj → jj cc jj nns



Translations for ‘nom adj kon adj’

Source	Target	Num of examples	%
nom adj kon adj	jj cc jj nns	137	21%
nom adj kon adj	jj cc jj nn	134	21%
nom adj kon adj	jj cc jj	31	5%
nom adj kon adj	jj , jj nn	15	2%
nom adj kon adj	dt jj cc jj nn	8	1%
nom adj kon adj	jj cc jj nns ,	7	1%

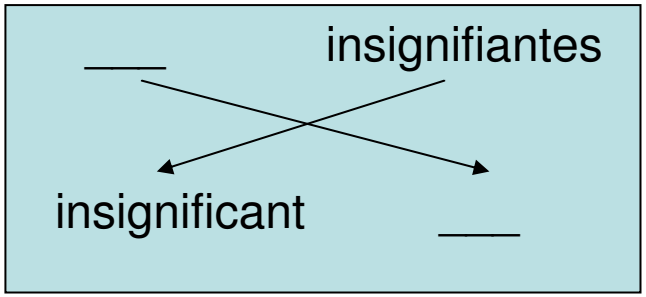
GIZA++ alignments for ‘nom adj kon adj’ → ‘jj cc jj nns’

Alignment	Num of examples	%
1-0 2-1 3-2 0-3	103	16%
3-0 2-1 1-2 0-3	18	3%
1-0 2-1 3-2 3-3	6	1%
1-0 2-1 3-2	3	0%
0-0 1-0 2-1 3-2	2	0%
0-0 1-0 2-1 3-2 3-3	1	0%
1-0 1-1 2-1 3-2 0-3	1	0%
1-0 2-1 1-2 3-2 0-3	1	0%
2-1 3-2 0-3	1	0%
3-0 2-1 1-2 3-2 0-3	1	0%

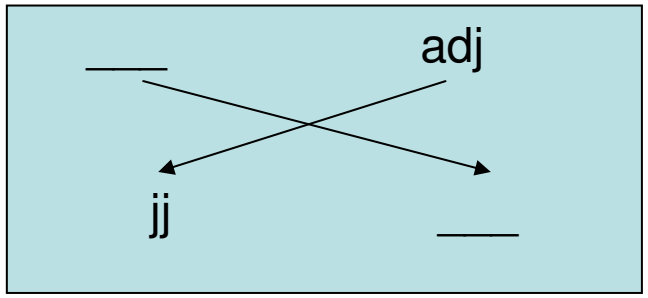
Current Solutions

Lexicalized Reordering

– Create probabilistic rules to swap phrases
  - eg. translates ‘insignifiantes’ before translating the preceding source phrase



– Can also apply to POS tags
  - eg. translates ‘adj’ before translating the preceding source phrase



Language Models

- Can be used for re-ordering
- Disadvantages
  - No access to source sentence
  - Insufficient coverage may still be a problem

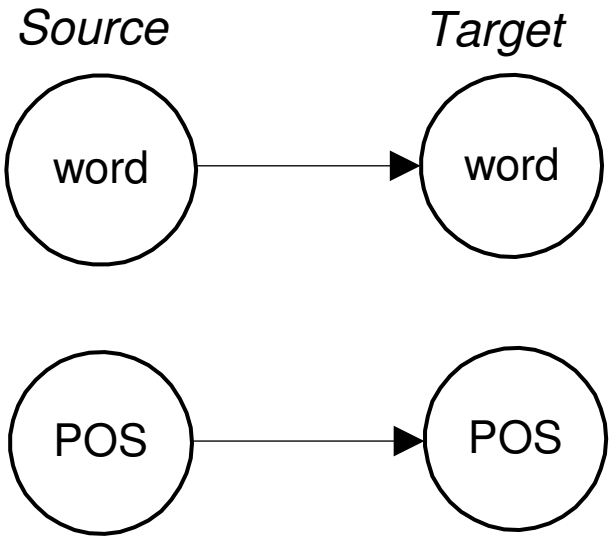
		Log prob		Log prob
Correct:	union european	-10.59	european union	-6.02
	nn jj (English)	-6.18	jj nn (English)	-5.37
Coverage Problem:	minorities insignificant	-12.56	insignificant minorities	-12.56
No knowledge of source:	man bites dog	-18.05	dog bites man	-17.63

(homme mord chien)

Proposed Solution

Factored Models

- Represent word as vector of factors
  - User-defined: surface form, POS tag, lemma…
- Decompose translation into multiple steps



Factored Models

Example of factored translation

Source phrase:

union européenne	
nom	adj

2 step translation process:

Step 1. Translate POS tags

Phrase Table 1

nom	adj	→	*jj*	*nn*

Step 2. Translate surface words

Phrase Table 2

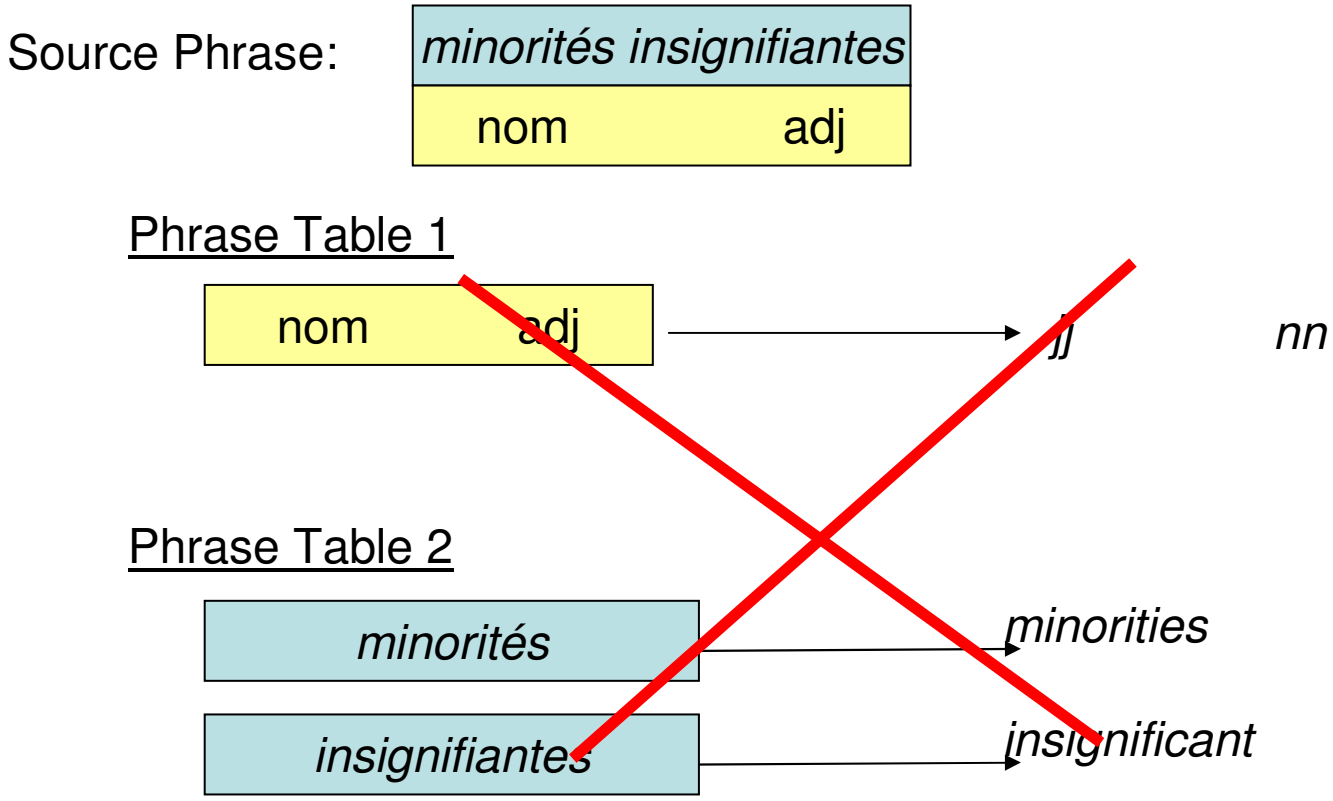
union européenne	→	*european union*

Factored Models

Problem: Coverage of both phrase tables must be the same

In example below:

Phrase table 1 translate source with 1 translation
Phrase table 2 can only translate source with 2 translations

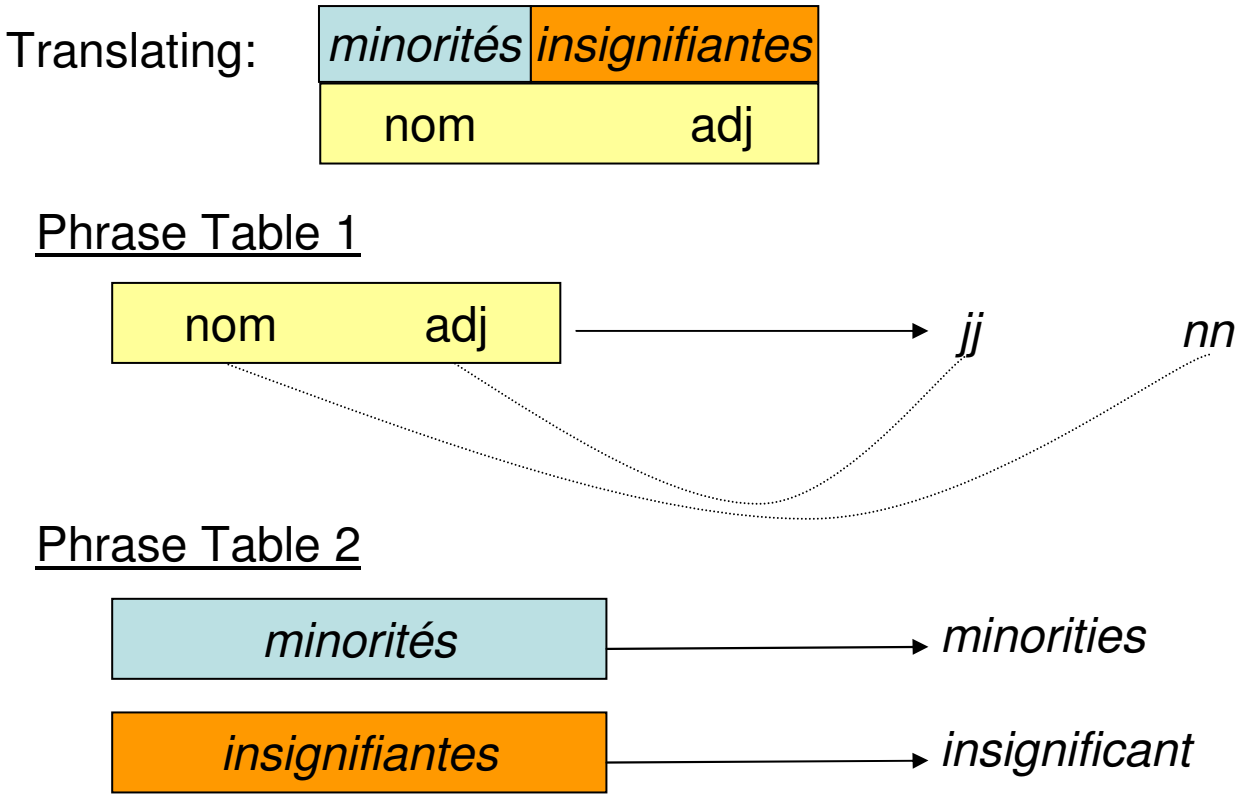


Proposed Solution

Idea: Relaxing constraint to enable multiple surface phrase per POS phrase

Implementation:

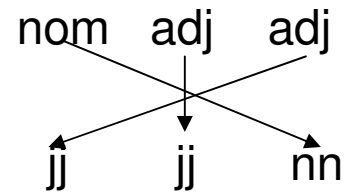
1. Retain GIZA++ alignment in phrase tables
2. Alignment for surface and POS must 'match'

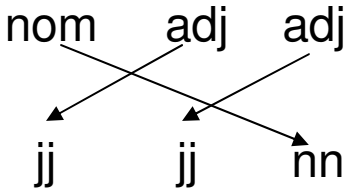


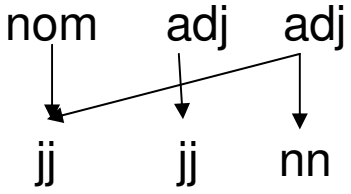
Creating alignment info

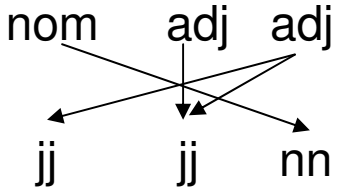
Example: nom adj adj → jj jj nn

Alignment	Num of examples	%age
2-0 1-1 0-2	337	21%
1-0 2-1 0-2	8	0%
0-0 2-0 1-1 2-2	7	0%
2-0 1-1 2-1 0-2	6	0%

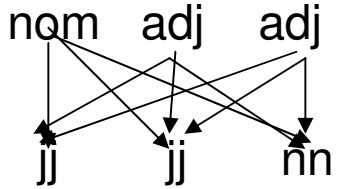




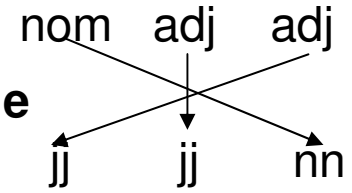




Union



Most probable



Experimental Results

Experimental Setup

News Commentary Corpus

Training:	De	En	Fr	En
Sentences	60k		43k	
Run. Word	1.3M	1.2M	1.0M	0.9M
Voc.	111K	70K	64K	58K

1. Tuned weights (MERT) using Europarl development set. (2000 sentences)

2. Evaluation on test set
   - In-domain: News Commentary (1064 sentences)
   - Out-of-domain: Europarl (2000 sentences)

3. Taggers
   - Brill Tagger (English), Treetagger (French), LoPar Tagger (German)

4. Trigram LM for surface words and POS tags

Results

German to English	Out of domain (BLEU)	In domain (BLEU)
Baseline (non-factored)	14.55	18.23
+ Factors	15.02	18.84
+ POS Tag Templates (union)	15.27	18.83
Comparison with other re-ordering strategy:		
Baseline + Lex. re-ordering	15.30	19.27

Little short range re-ordering

unbedeutenden Minderheiten → insignificant minorities

wirtschaftlichen und sozialen Schwierigkeiten → economic and social problems

große militärische Begründungen → important military justifications

Results

French to English	Out of domain (BLEU)	In domain (BLEU)
Baseline (non-factored)	19.59	23.12
+ Factors	19.77	22.99
+ POS Tag Templates		
union of alignments	20.51	24.32
most probable alignment	20.61	24.09
Comparison with other re-ordering strategy:		
Baseline + Lex. re-ordering (non-factored)	20.24	23.96

More short range re-ordering

minorités insignifiantes → minorities insignificant

difficultés économiques et sociales → economic and social problems

justifications militaires importantes → military important justifications

Conclusion

- POS Tag translation can improve short-range re-ordering
- Performance is dependent on language pair

Further Work

- Dealing with NULL alignments
- Longer range re-ordering
- Combine POS tag template with lexicalized re-ordering

Thank you !