



Bilingual segmentation for phrasetable pruning in Statistical Machine Translation

Germán Sanchis-Trilles Daniel Ortiz-Martínez Jesús González-Rubio
Jorge González Francisco Casacuberta

{gsanchis,dortiz,jegonzalez,jgonzalez,fcn}@dsic.upv.es

Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática • Universidad Politécnica de Valencia

Motivation

- Typical SMT systems require inferring huge tables of phrase pairs
- Large phrasetables lead to an elevated computational cost
- Bottleneck for the widespread application of SMT in portable devices
- Remove phrase pairs that have no influence on final translation
- Develop phrasetable pruning technique that:
 - is straightforward
 - is independent on the extraction algorithm
 - does not affect translation quality

Introduction

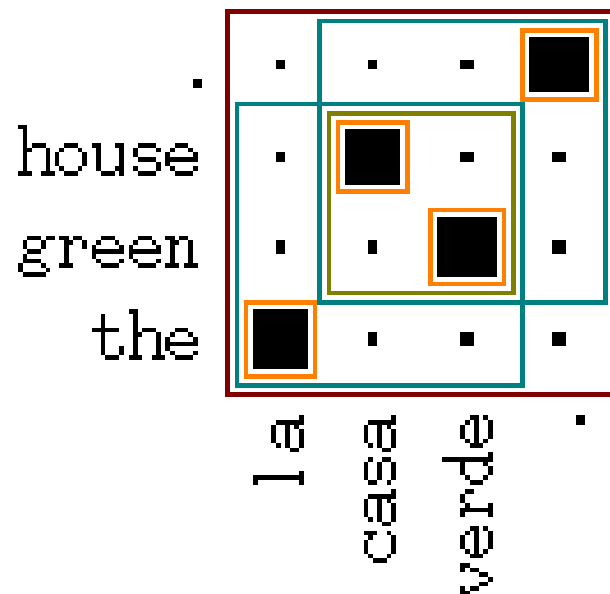
- Fundamental equation of SMT:

$$\hat{\mathbf{e}} = \operatorname{argmax}_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f})$$
$$\approx \operatorname{argmax}_{\mathbf{e}} \sum_{m=1}^M \lambda_m h_m(\mathbf{f}, \mathbf{e})$$

- Current SMT systems strongly based on phrases (i.e. word sequences)
- Work performed in PB models and PBSFSTs
- Phrase-extraction obtains multiple overlapping segmentations per sentence pair
→ reduce redundancy

Bilingual segmentation

- Selecting a single bilingual segmentation per sentence is a difficult problem
- In SMT, bilingual segmentation can be derived from phrase-based alignment



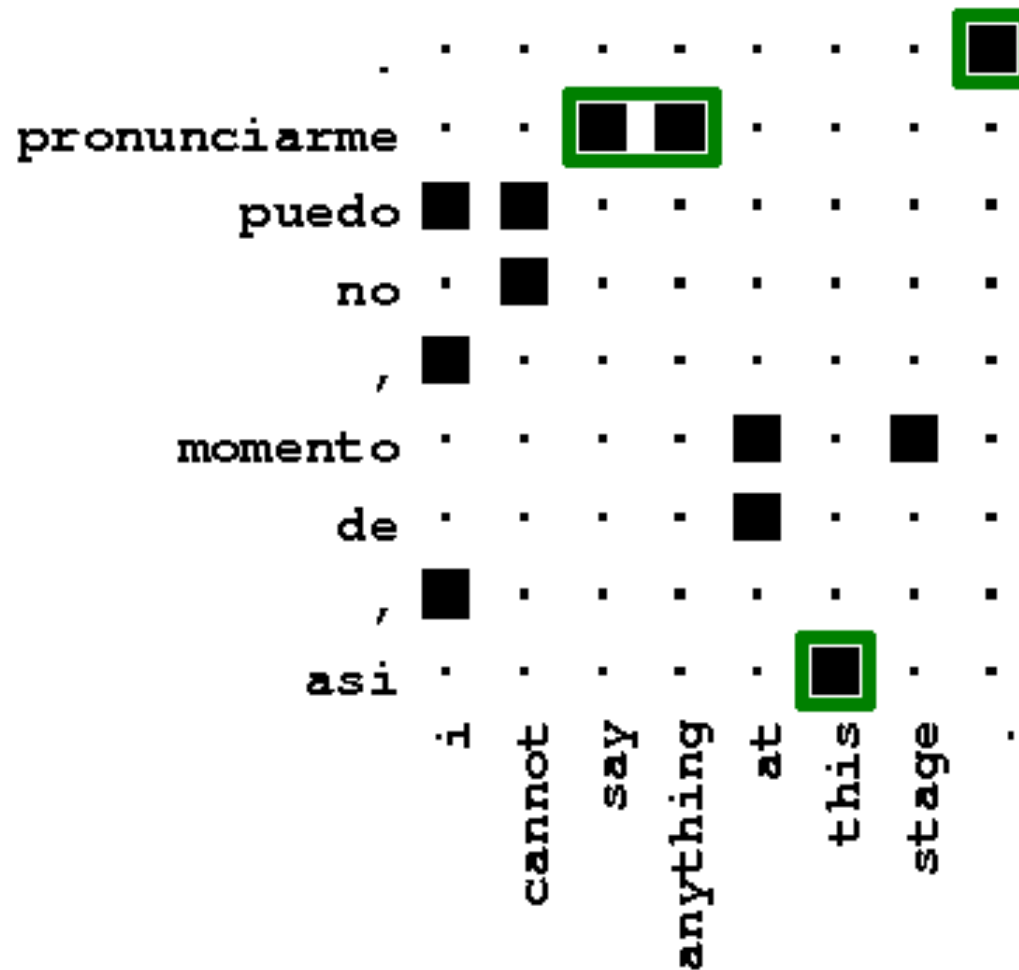
* Words are aligned into phrases, building supersets

* The best phrase-alignment can be defined as

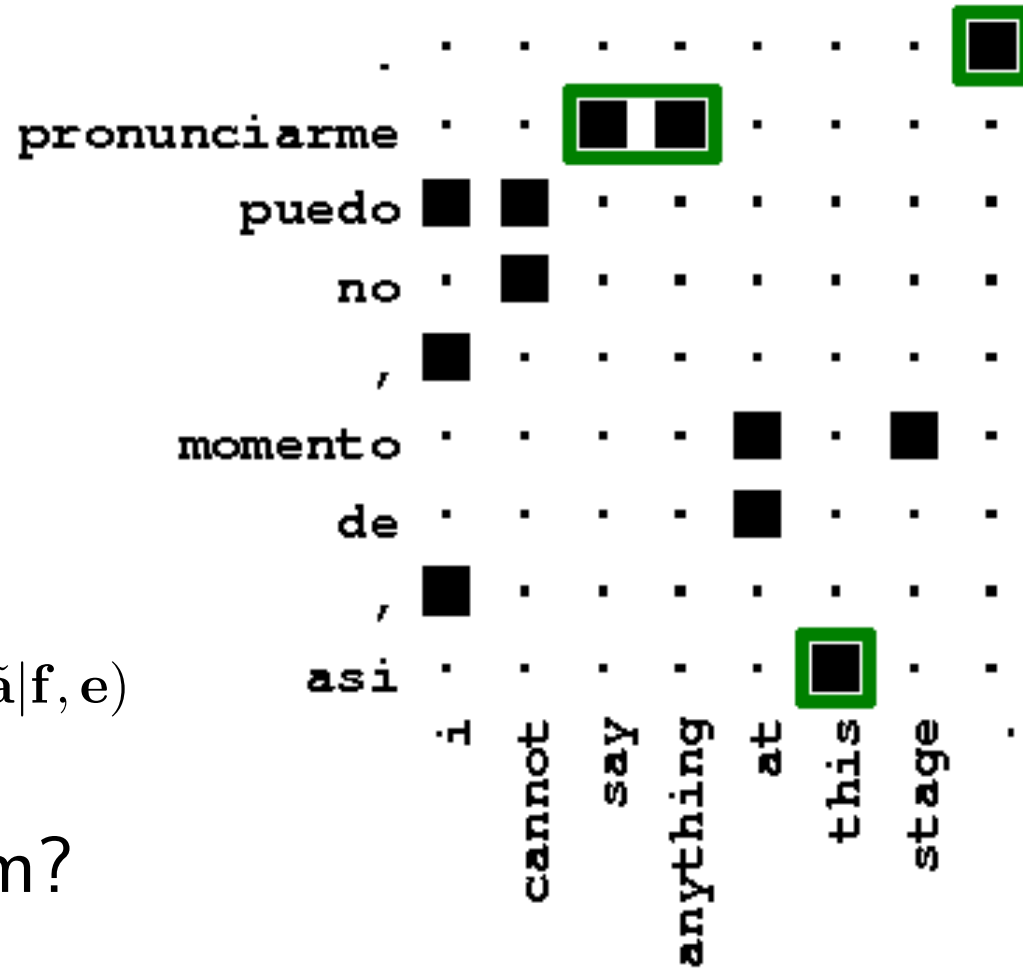
$$\tilde{A}_V(\mathbf{f}, \mathbf{e}) = \operatorname{argmax}_{\tilde{\mathbf{a}}} p(\tilde{\mathbf{a}}|\mathbf{f}, \mathbf{e}) \quad (1)$$

Search problem?

Bilingual segmentation: coverage problem



Bilingual segmentation: coverage problem



$$\tilde{A}_V(\mathbf{f}, \mathbf{e}) = \operatorname{argmax}_{\tilde{\mathbf{a}}} p(\tilde{\mathbf{a}} | \mathbf{f}, \mathbf{e})$$

Search problem?

Outline

- Motivation
- Introduction
- Bilingual segmentation
 - *True bilingual segmentation*
 - *Source-driven bilingual segmentation*
- Experiments
 - Phrase-based models
 - Phrase-based SFSTs
- Conclusions

True bilingual segmentation

- If output sentence is fixed, coverage problems imply that smoothing is needed
- Use a log-linear model to control different aspects of the segmentation

$$\tilde{A}_V(\mathbf{f}, \mathbf{e}) = \operatorname{argmax}_{\tilde{\mathbf{a}}} p(\tilde{\mathbf{a}}|\mathbf{f}, \mathbf{e}) = \operatorname{argmax}_{\tilde{\mathbf{a}}} p(\tilde{\mathbf{a}}, \mathbf{e}|\mathbf{f})$$

- Not a decoding problem, since maximisation takes place only over alignments
- However, underlying log-linear model not the same as in decoding time
 - Once optimal segmentations are available, a new phrasetable can be built
 - New phrases are introduced as a side-effect of smoothing

Source-driven bilingual segmentation

- True bilingual segmentation solves the coverage problem and fixes the output sentence
- In translation time, such restriction may introduce an inappropriate bias
 - Score function is modified due to smoothing
 - New phrase pairs are introduced
- Heuristic algorithm has proved to provide appropriate bilingual phrases
- Relax the output sentence restriction:

$$\tilde{A}_V(\mathbf{f}, \mathbf{e}) \approx \operatorname{argmax}_{\tilde{\mathbf{a}}, \mathbf{e}} p(\tilde{\mathbf{a}}|\mathbf{f}, \mathbf{e}) = \operatorname{argmax}_{\tilde{\mathbf{a}}, \mathbf{e}} p(\tilde{\mathbf{a}}, \mathbf{e}|\mathbf{f})$$

⇒ SMT search problem

- Output sentence is allowed to be different from reference
- Only segments in the current phrasetable are used
- Segmentation induced by input sentence

Experimental setup

- Experiments conducted by means of Thot & GREAT toolkits toolkit
- Similar experiments with Moses led to the same conclusions
- Translation quality measured with BLEU, TER and speedup ($S_p = T_b/T_r$)
- Experiments conducted on Europarl
- 95% level confidence intervals were about 0.65 points in every case

Experimental setup

	Subset features	De	En	Es	En
Training	Sentences	751k		731k	
	Run. words	15.3M	16.1M	15.7M	15.2M
	Mean length	20.3	21.4	21.5	20.8
	Vocabulary	195k	66k	103k	64k
Development	Sentences	2000		2000	
	Run. words	55k	59k	61k	59k
	Mean length	27.6	29.3	30.3	29.3
	OoV words	432	125	208	127
Test	Sentences	3064		3064	
	Run. words	82k	85k	92k	85k
	Mean length	26.9	27.8	29.9	27.8
	OoV words	1020	488	470	502

Results: Phrase-based models

Pair	Baseline		Source-driven			True		
	BLEU	w/s	BLEU	w/s	S_p	BLEU	w/s	S_p
Es–En	28.2	93	27.5	1500	16	23.8	380	4
En–Es	27.6	76	27.2	700	9	24.7	250	3
De–En	21.6	100	21.1	1500	15	17.5	280	3
En–De	15.2	46	15.1	400	9	14.7	170	4

- For source-driven segmentation:
 - BLEU (not significantly) lower, TER unaltered
 - Number of parameters reduced by two orders of magnitude ($\pm 2\%$ of original)
 - Translation speed increased by a factor of 9–16
- For true segmentation:
 - Translation quality drops significantly

Results: Phrase-based SFSTs

Pair	Source-driven			PB
	BLEU	w/s	S_p	
Es–En	25.8	92k	986	(28.2)
En–Es	25.3	28k	374	(27.6)
De–En	18.8	41k	412	(21.6)
En–De	13.0	14k	309	(15.2)

- PBSFSTs require monotonic bilingual segmentation (no "*baseline*")
- Baseline PB models produce better translation quality (although with more models)
- Speed increased by almost two orders of magnitude (more)

Conclusions

- Technique for reducing size of phrasetables
- Select most probable phrase pairs in a Viterbi fashion
- Source-driven segmentation leads to important improvements in decoding speed
 - Subset of original phrasetable
- True bilingual segmentation provides worse translation results:
 - Smoothing techniques are introduced
 - New phrase pairs introduced (10%–50%)
 - Important role in estimation of new model parameters
- Further work needed to understand true bilingual segmentation

Questions? Comments? Suggestions?