Phrase-based alignment combining corpus cooccurrences and linguistic knowledge

Adrià de Gispert
José B. Mariño
Josep Maria Crego

Outline

- Introduction
- Proposed phrase alignment strategy
- Experimental results
- Discussion
- Further research

Outline

- **Introduction**
  - Motivation
  - Word and phrases association measures
- Proposed phrase alignment strategy
- Experimental results
- Discussion
- Further research

Motivation

- Word alignment is crucial to train SMT systems
- GIZA++ alignments are state-of-the-art, but...
  - Symmetrization strategies are non-linguistic
  - Model complexity to introduce additional knowledge
- Cooccurrence-based algorithms perform well too, but...
  - Their output must be a many-to-many alignment

Goal: *phrase alignment following linguistic criteria*

Word & phrase cooccurrence measures

- ϕ^2 **score**, t-score, Dice, ...
- Can be computed between words but also phrases
- **Phrase cooccurrence** measures give complementary and stronger evidence

	please	
por	22.4	**0.9**
favor	1.2	

	maybe	
a	23.1	**8.0**
lo	18.2	
mejor	12.2	

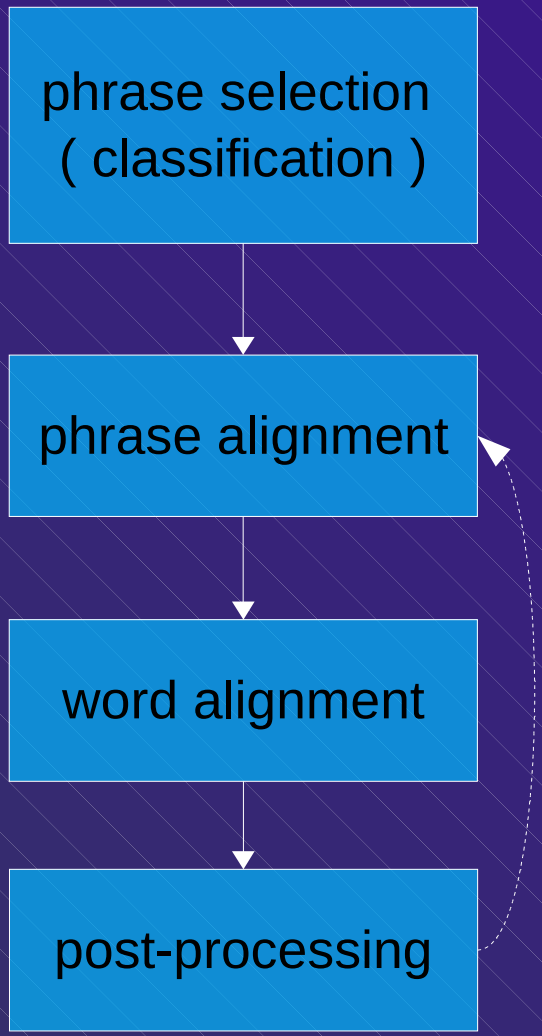
- Not efficient to compute for all possible phrase pairs
- A selection of candidate phrases is needed

Outline

- Introduction
- **Proposed phrase alignment strategy**
  - Candidate phrase selection and classification
  - Phrase-to-phrase alignment
  - Word alignment algorithm
- Experimental results
- Discussion
- Further research

Phrase alignment strategy

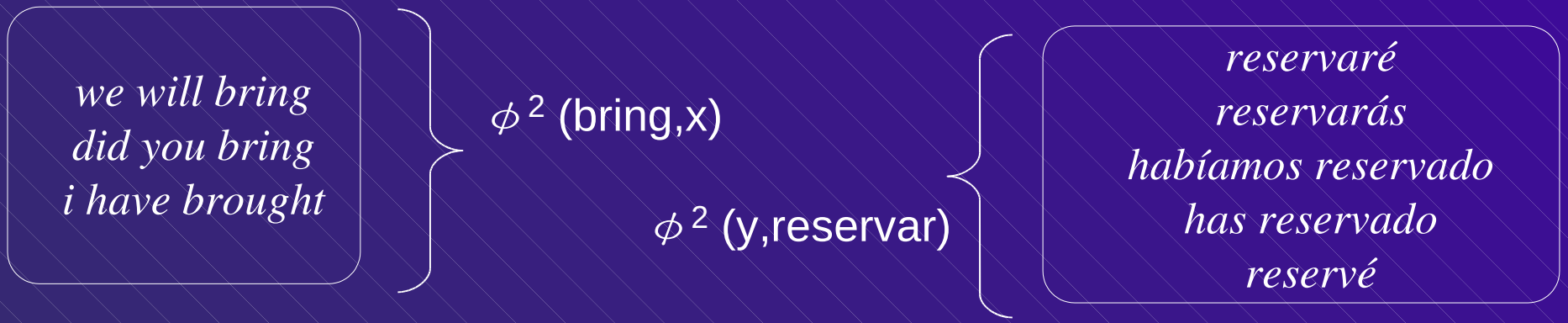
- Four stages:



  - Linguistically-guided selection of candidate phrases
  - Verb groups and idiomatic expressions
  - Add knowledge limiting cooc. counts table size
  - ϕ^2 -based competitive linking until threshold
  - Very-high precision required
  - one-to-one word alignment with unaligned tokens
  - final global decisions on word alignment

Candidate selection: Verbs

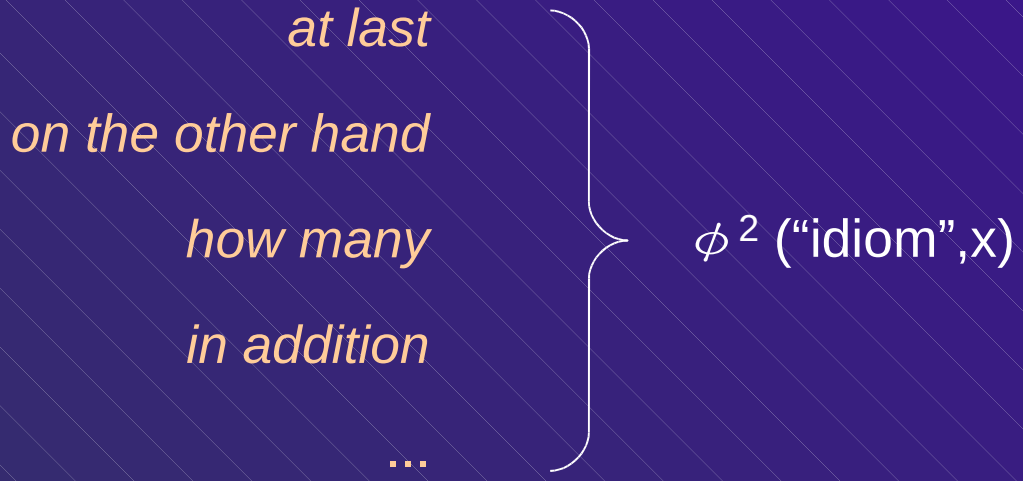
- **Rule-based** detection
  - Using word, POS and base form
  - Classification according to head verb base form
  - Check base forms against lists to avoid tagging errors



  - Single-word verbs substituted by base form
  - Reduction in cooc. table size
  - Limit: Base form ambiguity not tackled

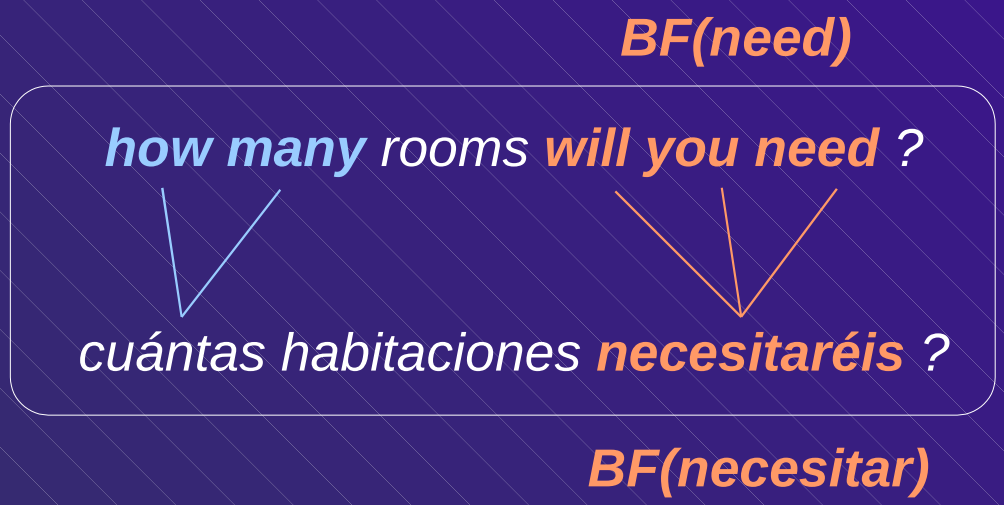
Candidate selection: Idioms

- Lists of **frequently-used idioms**
  - Spanish: 1496 idioms
  - English: 49 idioms
- No further classification
  - Compute coocs. against all other language tokens
  - Slight increase in cooc. table size



Phrase-to-phrase alignment

- Competitive linking strategy until threshold is met
- Verb groups and idioms treated separately
- Example



ϕ^2 (“how many”,cuántas) = **2.5**

ϕ^2 (“how many”,habitaciones) = 23.0

ϕ^2 (“how many”,”BF(necesitar)”) = 33.4

ϕ^2 (“BF(need)”,cuántas) = 31.05

ϕ^2 (“BF(need)”,habitaciones) = 19

ϕ^2 (“BF(need)”,”BF(necesitar)”) = **0.9**

Word alignment algorithm

(Cherry and Lin, 2003)

- One-to-one alignment
- Iterative best-first search
- Heuristic based on link probabilities
  - Initial alignment generated using ϕ^2 scores
  - Estimate link probabilities
  - Realignment using new estimates

- Syntax-guided cohesion constrain included

Outline

- Introduction
- Proposed phrase alignment strategy
- **Experimental results**
  - Data used
  - Partial results: phrase alignment
  - Complete AER results
- Discussion
- Further research

Data used

- Verbmobil Spa-Eng corpus *30 K sentences*

	words	vocab	singlet.	Lmax	Lavg
English	230 K	3.2 K	39 %	66	7.6
Spanish	220 K	5.0 K	43 %	66	7.3

  - Preprocessing
    - Normalization of contracted forms *we've=we have / del=de el*
    - Tagging and base form *Eng:TnT + wnmorph / Spa:maco+ relax*
    - Date and time expressions
    - No punctuation
- Evaluation scheme with AER
  - Dev. + test sets: 100 + 400 sentences
  - Manual alignment (80% Sure, 20% Poss) *stress on Recall*

Partial results: phrase alignment

- Results before word alignment

	Recall	Precision
Verbs $\phi^2 < 8$	8.07	99.02
Verbs $\phi^2 < 10$	9.00	99.12
Verbs $\phi^2 < 15$	9.68	98.69
Idioms $\phi^2 < 5$	2.01	98.48
Idioms $\phi^2 < 10$	3.06	99.00
Idioms $\phi^2 < 15$	3.50	97.41

- Straightforward approach, but ...
  - About 10% Recall at nearly no Precision cost
  - Complementary links between Verbs and Idioms
  - Complexity reduction for word alignment algorithm

Complete AER results

	Recall	Precision	AER
giza++ eng2spa	76.99	93.15	15.51
giza++ spa2eng	78.75	94.19	13.94
giza++ union	**84.47**	90.85	**12.30**
giza++ intersection	71.27	**97.58**	17.52

- union: precision loss, but very high recall

Complete AER results

	Recall	Precision	AER
giza++ eng2spa	76.99	93.15	15.51
giza++ spa2eng	78.75	94.19	13.94
giza++ union	**84.47**	90.85	**12.30**
giza++ intersection	71.27	**97.58**	17.52
one-to-one word aligner	72.56	96.69	16.96

- union: precision loss, but very high recall
- intersection vs. one-to-one aligner

Complete AER results

	Recall	Precision	AER
giza++ eng2spa	76.99	93.15	15.51
giza++ spa2eng	78.75	94.19	13.94
giza++ union	**84.47**	90.85	**12.30**
giza++ intersection	71.27	**97.58**	17.52
one-to-one word aligner	72.56	96.69	16.96
phrase aligner $\phi^2 < 10$	76.31	**97.48**	**13.36**
phrase aligner $\phi^2 < 15$	76.88	**97.35**	**13.20**

- intersection vs. one-to-one aligner
- union: precision loss, but very high recall
- proposed: high-precision, much higher recall
- phrase alignment is accurate and helps word alignment algorithm to perform better

Outline

- Introduction
- Proposed phrase alignment strategy
- Experimental results
- **Discussion**
- Further research

Discussion

- Promising results
  - competitive results still making small use of ling. knowledge
  - open to new knowledge sources
- Evaluation in translation task
- Evaluation with other corpora

Outline

- Introduction
- Proposed phrase alignment strategy
- Experimental results
- Discussion
- **Further research**

Further research

- Postprocessing techniques
- Extension of phrase detection rules
  - 'Gapped' structures

	Recall	Precision	AER
phrase aligner $\phi^2 < 15$	76.88	**97.35**	**13.20**
+ Gapped verbs	77.67	**97.55**	**12.85**

- Ambiguity in classifying detected phrases
  - numbers, times, different head verbs,...
- Training data reduction

Thanks for attention

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona