technology
from seed

Building a golden collection of parallel Multi-Language Word Alignment

João Graça, Joana Paulo Pardal,

Luísa Coheur and Diamantino Caseiro

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

L^{2}F - Spoken Language Systems Laboratory

Resume

- Needed resources for Machine Translation
  - Manual Word Alignments for Portuguese
  - Comparable Word Alignments between different languages

Resume

- Needed resources for Machine Translation
  - Manual Word Alignments for Portuguese
  - Comparable Word Alignments between different languages
- Created and made available
  - 100 sentences manually aligned between 6 language pairs
  - Detailed guidelines for Multi-language Word Alignment
- http://www.l2f.inesc-id.pt/resources/translation

Outline

- Manual Word Alignment
- Corpus
- Alignment Process
- Evaluation
- Future Work

Manual Word Alignment

	0	1	2	3	4	5	6	7	8	
0	□	·	·	·	·	·	·	·	·	i
1	■	·	·	·	·	·	·	·	·	did
2	■	·	·	·	·	·	·	·	·	receive
3	·	·	·	■	·	·	·	·	·	the
4	·	·	·	·	■	·	·	·	·	request
5	·	·	·	·	·	·	·	□	·	you
6	·	·	·	·	·	·	·	■	·	sent
7	·	·	·	·	·	·	■	·	·	me
8	·	·	·	·	·	·	·	·	■	.
	recebi	de	facto	o	pedido	que	me	dirigiu	.	

Corpus

- European Parliament proceedings (Koehn 2006)
- First 100 sentence from common test set (2000-10 2000-12)

Sentences	100			
	English	Portuguese	French	Spanish
Words	1072	1131	1227	1106
Types	466	513	474	472
Avg. Size	10.72	11.31	12.27	11.06

Alignment Process

- Team
  - Four annotators (h_1, h_2, h_3, h_4)
  - Proficient speakers of annotated languages

Alignment Process

- Team
  - Four annotators (h_1, h_2, h_3, h_4)
  - Proficient speakers of annotated languages
- Starting point
  - Guidelines for Spanish-English (EPPS) (Mariño 2005)

Annotation Procedure – Part I

	1-20	20-40	40-60	60-80	80-100
EN-PT	h_1				
EN-ES	h_1				
EN-FR	h_2				
PT-ES	h_1				
PT-FR	h_2				
ES-FR	h_2				

- New alignment guidelines created
- Annotated all languages a sentence at a time
- Leave guidelines as unambiguous as possible

Annotation Procedure – Part II

	1-20	20-40	40-60	60-80	80-100
EN-PT	h_1	$h_1\&h_3$			
EN-ES	h_1	$h_1\&h_4$			
EN-FR	h_2	$h_2\&h_3$			
PT-ES	h_1	$h_1\&h_4$			
PT-FR	h_2	$h_2\&h_4$			
ES-FR	h_2	$h_2\&h_4$			

- First contact of new annotators with new guidelines
- First evaluation of the guidelines
- Disagreements were discussed and guidelines updated

Annotation Procedure – Part III

	1-20	20-40	40-60	60-80	80-100
EN-PT	h_1	$h_1 \& h_3$	$h_1 \& h_3$		
EN-ES	h_1	$h_1 \& h_4$	$h_1 \& h_4$		
EN-FR	h_2	$h_2 \& h_3$	$h_2 \& h_3$		
PT-ES	h_1	$h_1 \& h_4$	$h_1 \& h_4$		
PT-FR	h_2	$h_2 \& h_4$	$h_2 \& h_4$		
ES-FR	h_2	$h_2 \& h_4$	$h_2 \& h_4$		

- Final evaluation of the guidelines
- Guidelines refinement

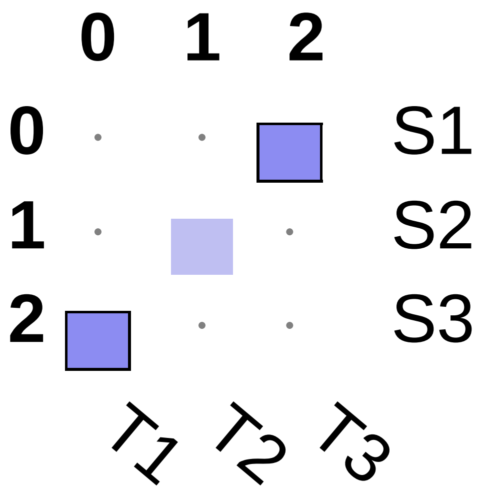
Annotation Procedure – Part IV

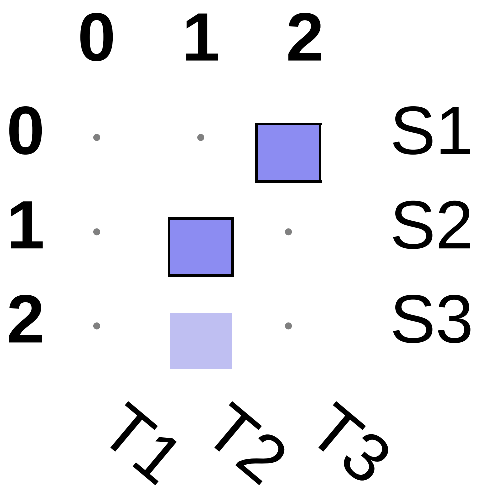
	1-20	20-40	40-60	60-80	80-100
EN-PT	h_1	$h_1 \& h_3$	$h_1 \& h_3$	h_1	h_3
EN-ES	h_1	$h_1 \& h_4$	$h_1 \& h_4$	h_1	h_4
EN-FR	h_2	$h_2 \& h_3$	$h_2 \& h_3$	h_2	h_3
PT-ES	h_1	$h_1 \& h_4$	$h_1 \& h_4$	h_1	h_4
PT-FR	h_2	$h_2 \& h_4$	$h_2 \& h_4$	h_2	h_4
ES-FR	h_2	$h_2 \& h_4$	$h_2 \& h_4$	h_2	h_4

- All sentences were reviewed one at a time for all languages
- Current version of the guidelines

Evaluation

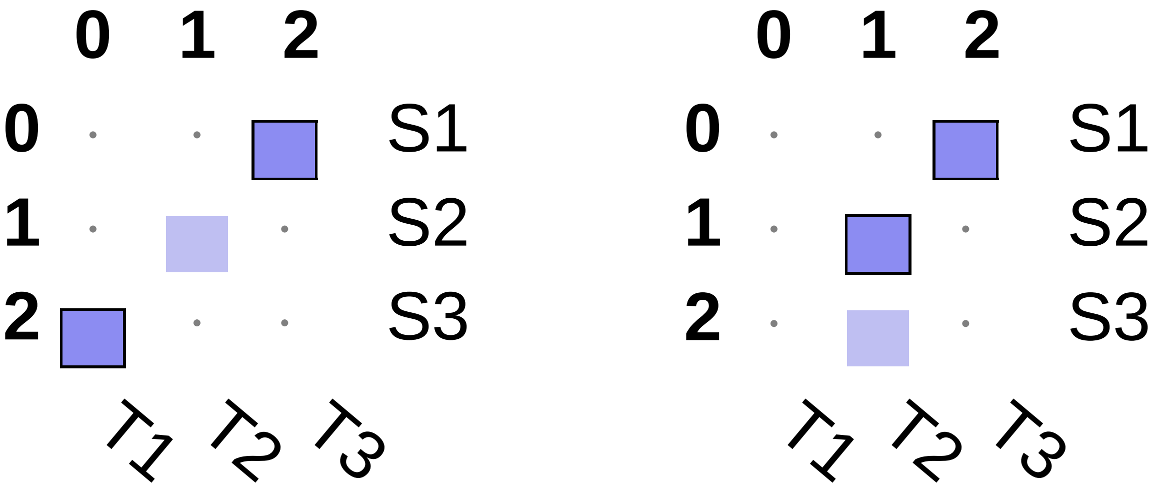
- Structures Output





Evaluation

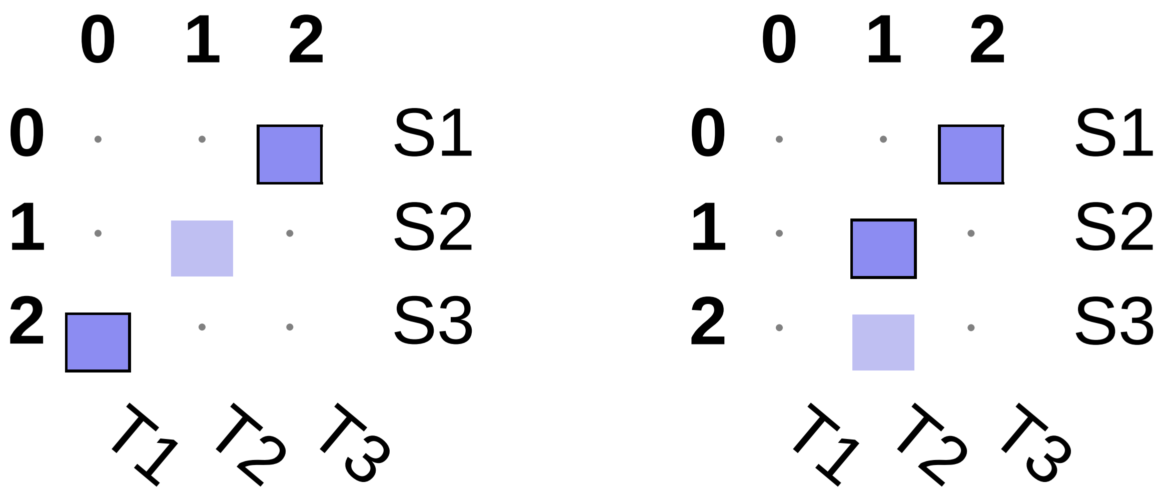
- Structures Output
- How to count misaligned points



- $\frac{|I_{1-2}|}{|A1|+|A2|}$ – 50 %

Evaluation

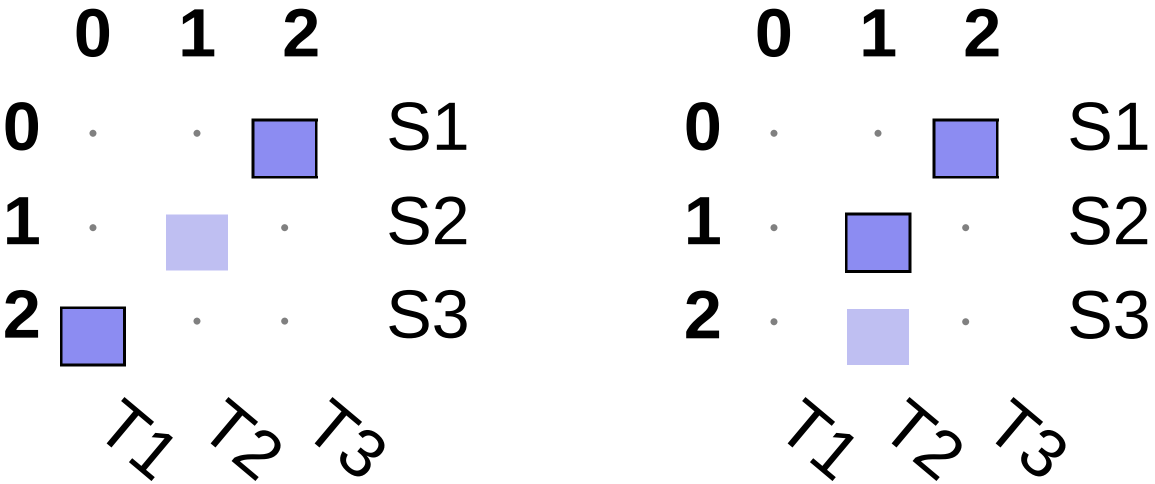
- Structures Output
- How to count misaligned points



- $\frac{2*|I_{1-2}|}{|A1|+|A2|}$ – 66 %
- Melamed (1998), Kruijff-Korbayová (2006)

Evaluation

- Structures Output
- How to count misaligned points
- Sure and Possible Alignment



	0	1	2	
0	·	·	🟦	S1
1	·	▨	·	S2
2	🟦	·	·	S3
	T1	T2	T3	

	0	1	2	
0	·	·	🟦	S1
1	·	🟦	·	S2
2	·	▨	·	S3
	T1	T2	T3	

Inter Annotator Agreement

- Distinguish between
  Sure and Possible alignment mistakes
  - *Strong Agreement* - $\frac{2*(|I_{s-s}|+|I_{p-p}|)}{|A_1|+|A_2|}$
  - *Weak Agreement* - $\frac{2*(|I_{p-s}|+|I_{s-p}|)}{|A_1|+|A_2|}$
  - *Weak Disagreement* - $\frac{2*(|I_{p-0}|+|I_{0-p}|)}{|A_1|+|A_2|}$
  - *Strong Disagreement* - $\frac{2*(|I_{s-0}|+|I_{0-s}|)}{|A_1|+|A_2|}$

First Evaluation: 20–40

	$\lvert A_1 \rvert$	$\lvert A_2 \rvert$	SA	WA	WD	SD
EN-PT (h_1,h_4)	269	196	67.0	11.2	15.3	6.5
EN-ES (h_1,h_3)	271	314	73.2	13.3	7.9	5.6
EN-FR (h_2,h_4)	332	256	61.9	19.0	9.3	9.7
PT-ES (h_1,h_3)	260	259	78.6	8.9	6.7	5.8
PT-FR (h_2,h_4)	331	260	73.8	10.2	9.1	6.9
ES-FR (h_2,h_3)	324	349	75.8	11.0	3.3	10.0
Average			71.7	12.3	8.6	7.4
Undifferentiated Average			84.0		16.0	

- Errors due to different interpretation of the guidelines

Second Evaluation: 40–60

	$\|A_1\|$	$\|A_2\|$	SA	WA	WD	SD
EN-PT (h_1,h_4)	287	316	82.9	6.6	9.5	1.0
EN-ES (h_1,h_3)	277	272	80.9	5.8	10.6	2.7
EN-FR (h_2,h_4)	262	281	80.8	10.0	5.0	4.2
PT-ES (h_1,h_3)	298	307	86.9	6.3	4.8	2.0
PT-FR (h_2,h_4)	273	268	86.5	7.0	4.1	2.4
ES-FR (h_2,h_3)	290	305	87.1	9.4	0.4	3.2
Average			84.2	7.5	5.7	2.6
Undifferentiated Average			91.6		8.4	

Future Work

- Increase each alignment set
  - External contributions
- Add data from new domains
- Add other types of annotations

technology
from seed

INSTITUTO
SUPERIOR
TÉCNICO

inesc id lisboa

L^2F - Spoken Language Systems Laboratory