Discriminative Weighted Alignment Matrices for Statistical Machine Translation

EAMT 2011 - Leuven

Nadi Tomeh Alexandre Allauzen François Yvon

LIMSI-CNRS / Université Paris-Sud 11

May 31, 2011

Outline

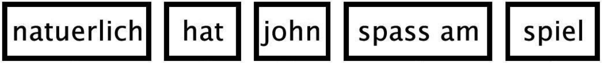
Phrase-based Translation Model

To translate the German sentence:

natuerlich hat john spass am spiel

Phrase-based Translation Model

To translate the German sentence: - phrase segmentation



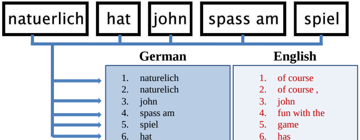
Phrase-based Translation Model

To translate the German sentence: - phrase segmentation - look up their translations in a *phrase-table*

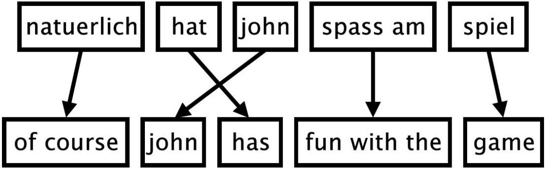
	German	English
1.	naturelich	of course
2.	naturelich	of course ,
3.	john	john
4.	spass am	fun with the
5.	spiel	game
6.	hat	has
7.	michael geht davon aus	michael assumes
8.	michael geht davon aus ,	michael assumes
9.	michael geht davon aus , dass	michael assumes that
10.	geht davon aus	assumes
11.	geht davon aus ,	assumes
12.	geht davon aus , dass	assumes that
13.	geht davon aus , dass er	assumes that he

Phrase-based Translation Model

To translate the German sentence: - phrase segmentation - look up their translations in a *phrase-table* - reorder English phrases



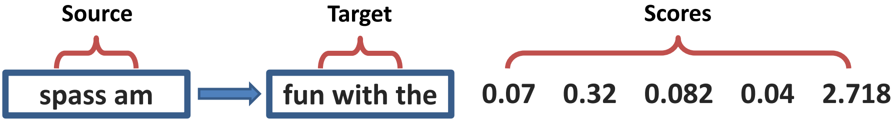
	German	English
1.	naturelich	of course
2.	naturelich	of course ,
3.	john	john
4.	spass am	fun with the
5.	spiel	game
6.	hat	has
7.	michael geht davon aus	michael assumes
8.	michael geht davon aus ,	michael assumes
9.	michael geht davon aus , dass	michael assumes that
10.	geht davon aus	assumes
11.	geht davon aus ,	assumes
12.	geht davon aus , dass	assumes that
13.	geht davon aus , dass er	assumes that he



Phrase Table Construction

Source		Target	Scores				
spass am	→	fun with the	0.07	0.32	0.082	0.04	2.718

Phrase Table Construction



1. **Extraction:** identify the set of all phrase-pairs
2. **Scoring:** translation probabilities and lexical scores

Phrase Table Construction

Source		Target	Scores				
spass am	→	fun with the	0.07	0.32	0.082	0.04	2.718

1. **Extraction:** identify the set of all phrase-pairs
2. **Scoring:** translation probabilities and lexical scores

Models

- Joint phrase alignment model (Marcu and Wang, 2002)

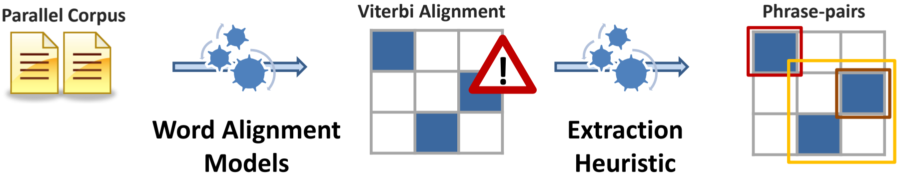
Phrase Table Construction

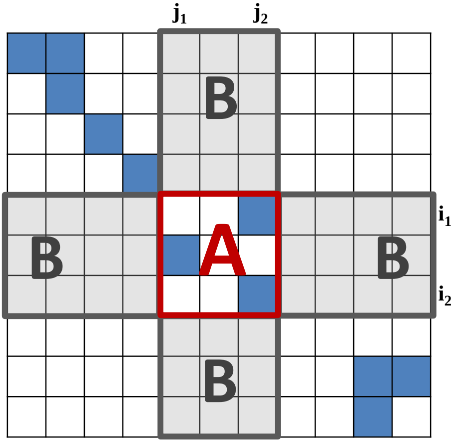
Source		Target	Scores				
spass am	→	fun with the	0.07	0.32	0.082	0.04	2.718

1. **Extraction:** identify the set of all phrase-pairs
2. **Scoring:** translation probabilities and lexical scores

Models

- Joint phrase alignment model (Marcu and Wang, 2002)
- Word alignment based pipeline (Zens et al., 2002)





Viterbi Alignment

$$f_E(p) = \alpha(p) \times \beta(p)$$

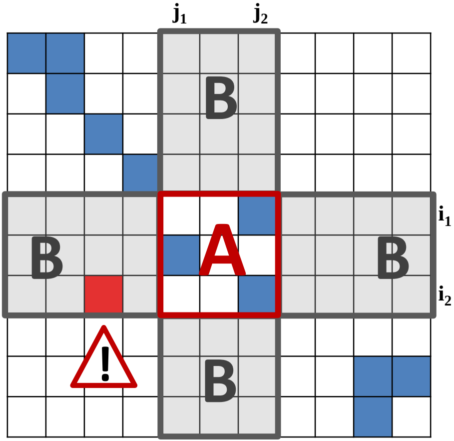
Viterbi-based:

$$\alpha \begin{cases} 1 & \text{if } \mathbf{A} \text{ contains a link} \\ 0 & \text{otherwise} \end{cases}$$

$$\beta \begin{cases} 0 & \text{if } \mathbf{B} \text{ contains a link} \\ 1 & \text{otherwise} \end{cases}$$

Accepted phrase:

$$f_C(p) = f_E(p) = 1$$



Viterbi Alignment

$$f_E(p) = \alpha(p) \times \beta(p)$$

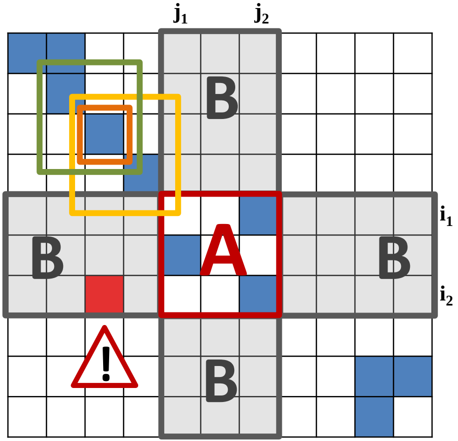
Viterbi-based:

$$\alpha \begin{cases} 1 & \text{if } \mathbf{A} \text{ contains a link} \\ 0 & \text{otherwise} \end{cases}$$

$$\beta \begin{cases} 0 & \text{if } \mathbf{B} \text{ contains a link} \\ 1 & \text{otherwise} \end{cases}$$

Accepted phrase:

$$f_C(p) = f_E(p) = 1$$



Viterbi Alignment

$$f_E(p) = \alpha(p) \times \beta(p)$$

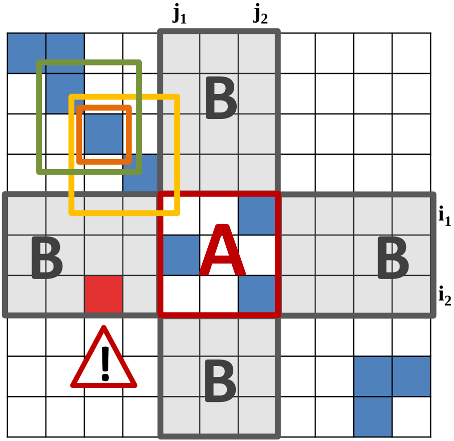
Viterbi-based:

$$\alpha \begin{cases} 1 & \text{if } \mathbf{A} \text{ contains a link} \\ 0 & \text{otherwise} \end{cases}$$

$$\beta \begin{cases} 0 & \text{if } \mathbf{B} \text{ contains a link} \\ 1 & \text{otherwise} \end{cases}$$

Accepted phrase:

$$f_C(p) = f_E(p) = 1$$



Viterbi Alignment

$$f_E(p) = \alpha(p) \times \beta(p)$$

Viterbi-based:

$$\alpha \begin{cases} 1 & \text{if } \mathbf{A} \text{ contains a link} \\ 0 & \text{otherwise} \end{cases}$$

$$\beta \begin{cases} 0 & \text{if } \mathbf{B} \text{ contains a link} \\ 1 & \text{otherwise} \end{cases}$$

Accepted phrase:

$$f_C(p) = f_E(p) = 1$$

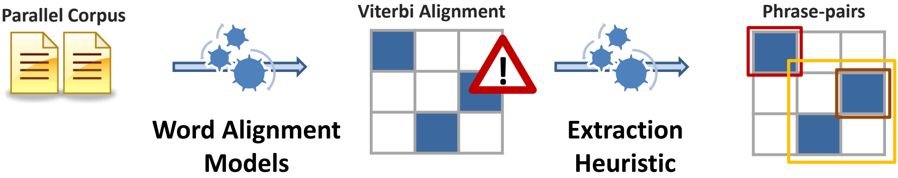
Phrase Table Construction

Source		Target	Scores				
spass am	→	fun with the	0.07	0.32	0.082	0.04	2.718

1. **Extraction:** identify the set of all phrase-pairs
2. **Scoring:** translation probabilities and lexical scores

Models

- Joint phrase alignment model (Marcu and Wang, 2002)
- Word alignment based pipeline (Zens et al., 2002)



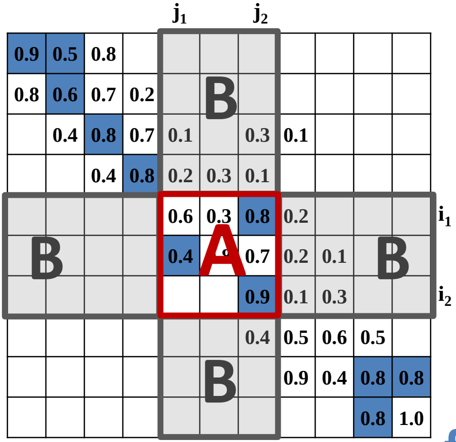
Weighted Alignment Matrices (Liu et al., 2009)

WAM-based extraction

				j_1		j_2					
0.9	0.5	0.8									
0.8	0.6	0.7	0.2								
	0.4	0.8	0.7	0.1		0.3	0.1				
		0.4	0.8	0.2	0.3	0.1					
				0.6	0.3	0.8	0.2				i_1
				0.4	0.8	0.7	0.2	0.1			
						0.9	0.1	0.3			i_2
						0.4	0.5	0.6	0.5		
							0.9	0.4	0.8	0.8	
									0.8	1.0	

Weighted Matrix

$$p(a_{i,j} \mid e, f)$$



Weighted Matrix

$$f_E(p) = \alpha(p) \times \beta(p)$$

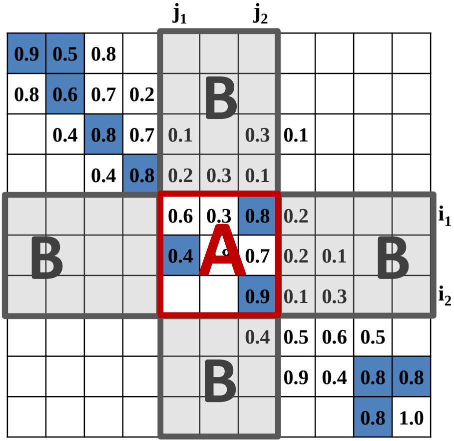
WAM-based:

$$\alpha(j1, j2, i1, i2) = 1 - \prod_{(j,i) \in in(j1,j2,i1,i2)} \bar{p}_m(j,i)$$

$$\beta(j1, j2, i1, i2) = \prod_{(j,i) \in out(j1,j2,i1,i2)} \bar{p}_m(j,i)$$

Accepted phrase:

$$f_C(p) = f_E(p) \text{ in } [0,1]$$



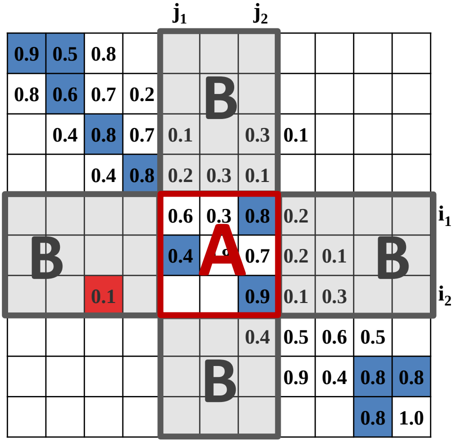
Weighted Matrix

$$f_E(p) = \alpha(p) \times \beta(p)$$

WAM-based:

$$f_E = 0.999 \times 0.069 = 0.068$$

$$f_C = 0.068$$



Weighted Matrix

$$f_E\,(p) = \alpha\,(p) \times \beta\,(p)$$

WAM-based:

$$f_E = 0.999 \times 0.069 = 0.068$$

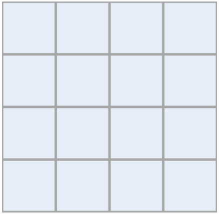
$$f_C = 0.068$$

$$f_E = 0.999 \times 0.062 = 0.062$$

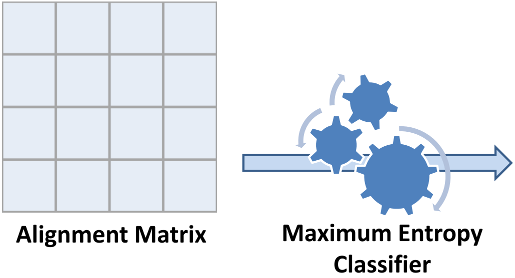
$$f_C = 0.062$$

Estimation of the Weighted Matrix

The MaxEnt Framework



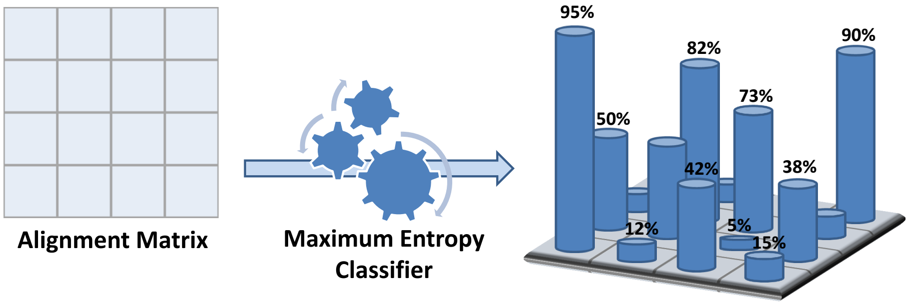
Alignment Matrix

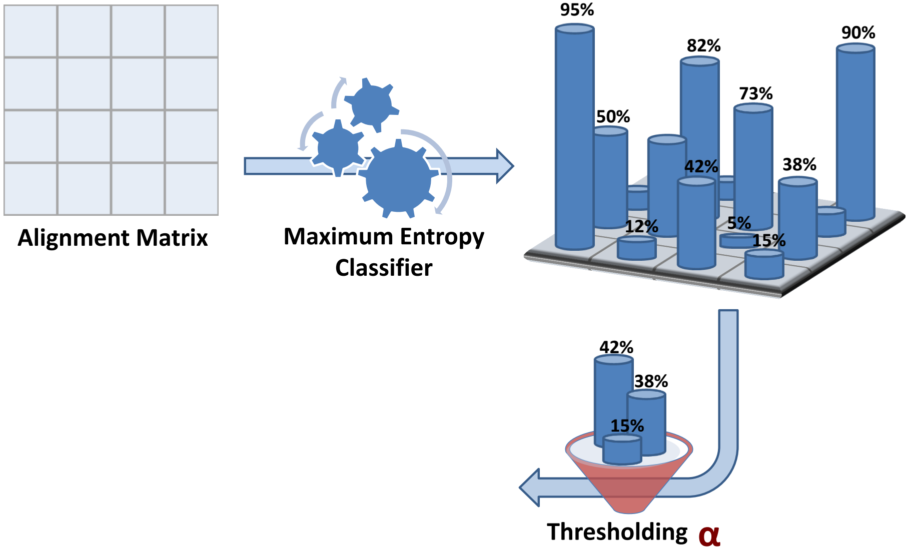


Alignment Matrix
Maximum Entropy Classifier

Estimation of the Weighted Matrix

The MaxEnt Framework

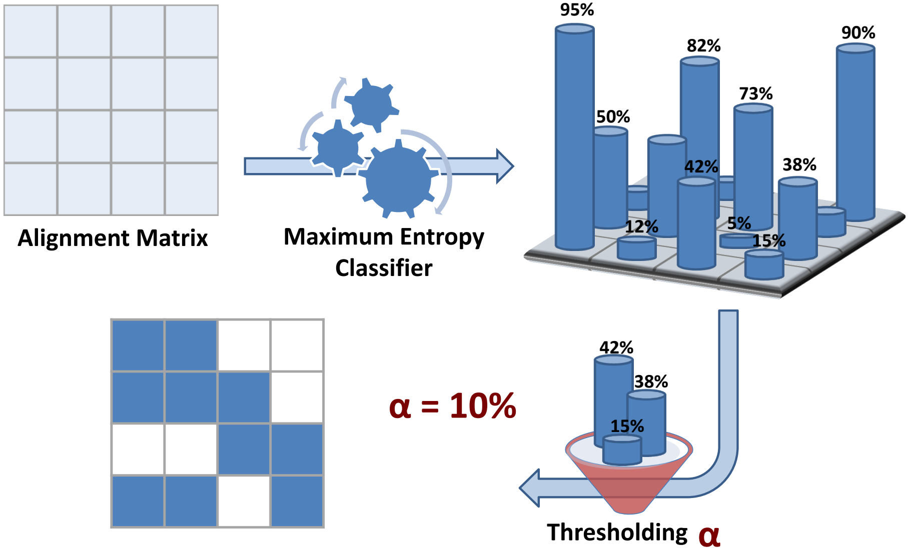


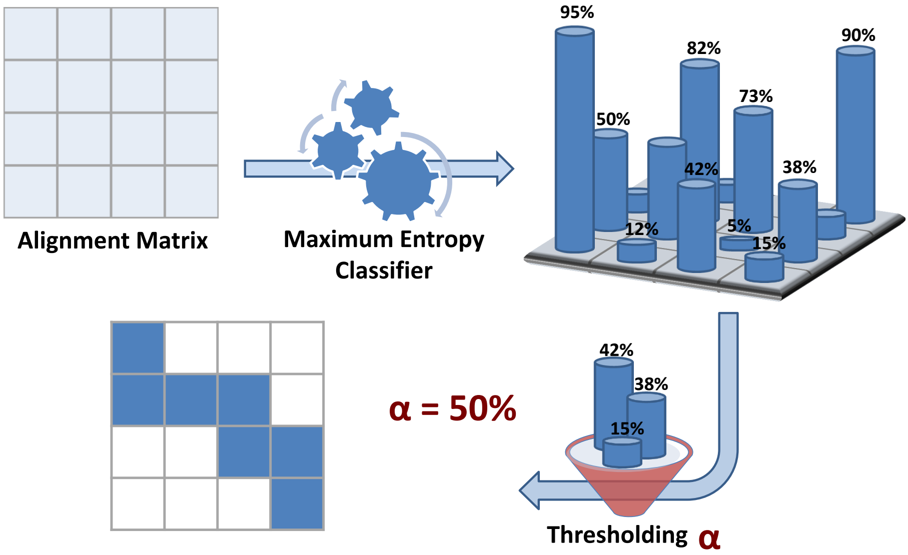


Alignment Matrix
Maximum Entropy Classifier
95%
50%
82%
73%
90%
42%
38%
12%
5%
15%
42%
38%
15%
Thresholding α

Estimation of the Weighted Matrix

The MaxEnt Framework

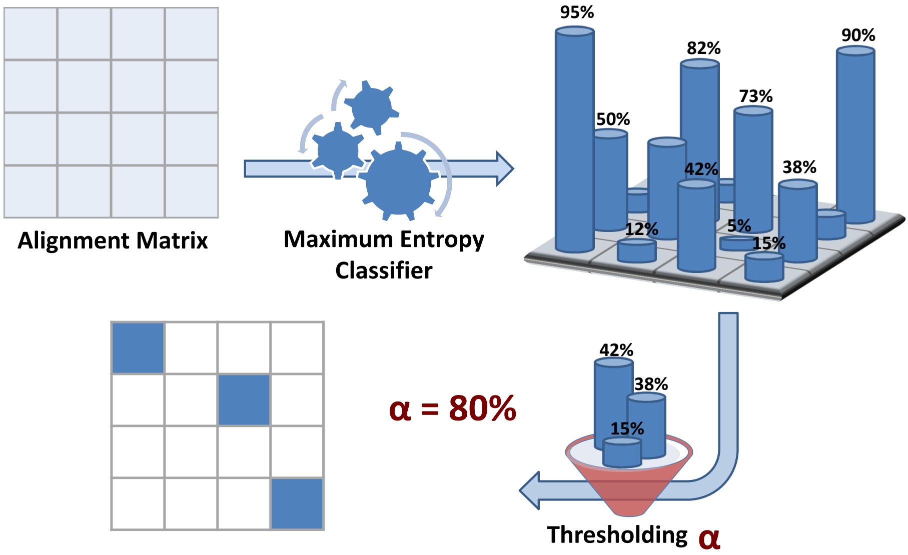




Alignment Matrix
Maximum Entropy Classifier
95%
50%
82%
73%
90%
42%
38%
12%
5%
15%
α = 50%
42%
38%
15%
Thresholding α

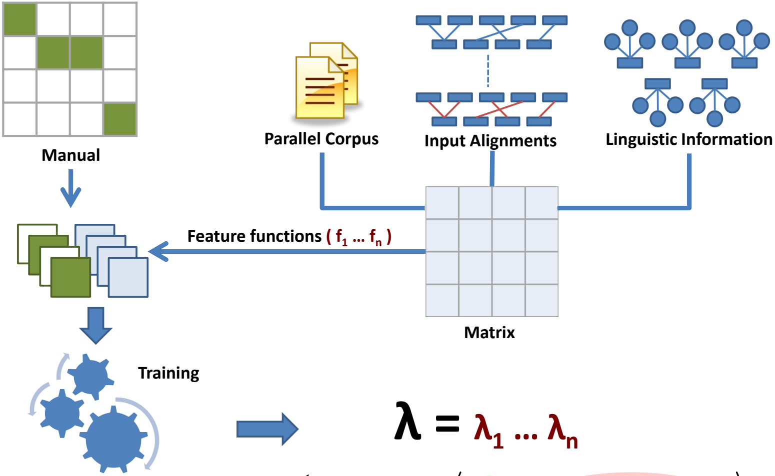
Estimation of the Weighted Matrix

The MaxEnt Framework

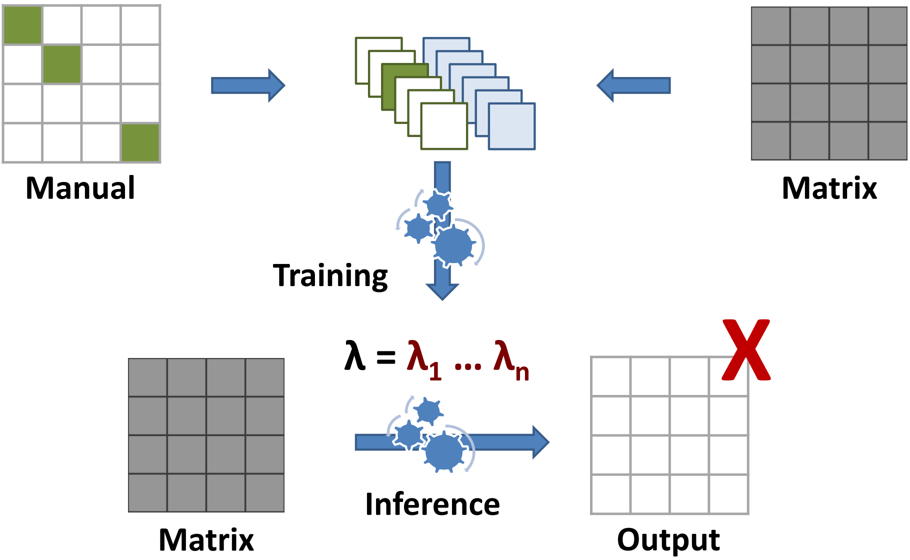


Estimation of the Weighted Matrix

Training the MaxEnt Model



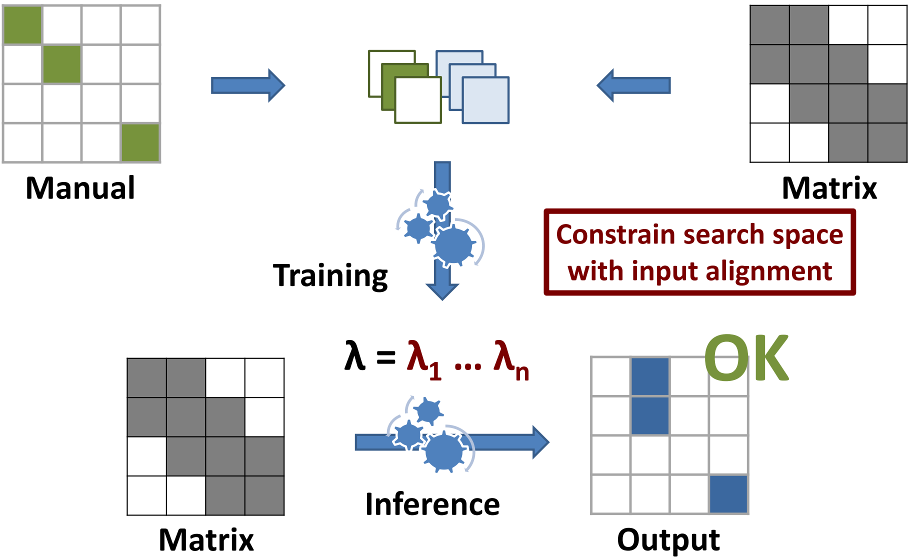
$$p(y_{ij} | \mathbf{e}, \mathbf{f}; \lambda) = \frac{1}{Z(\mathbf{e}, \mathbf{f}, i, j)} \cdot \exp \left\langle \lambda, \phi(\mathbf{e}, \mathbf{f}, i, j) \right\rangle$$



Manual
Matrix
Training
λ = λ1 … λn
Matrix
Inference
Output
X

Estimation of the Weighted Matrix

The Problem of Imbalanced Data



Goals

- Compare Viterbi-based with WAM-based extraction
- Compare different approaches to populate the matrix

Data

- Discriminative alignment: IBMAC (Ittycheriah et al., 2006)
- Training: two NIST MT'09 sub corpora - 30K and 130K
- Development: NIST MT'06 test data (4 refs)
- Evaluation: NIST MT'08 test data (4 refs)

Tools

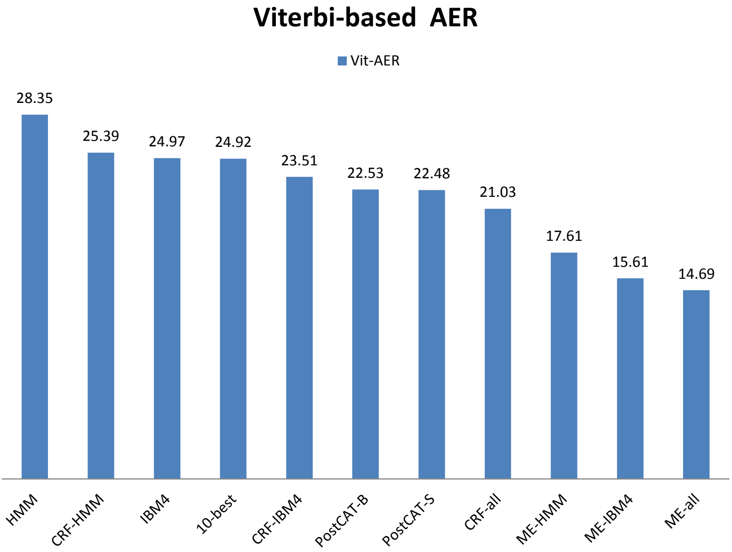
- Moses - MERT - SRILM

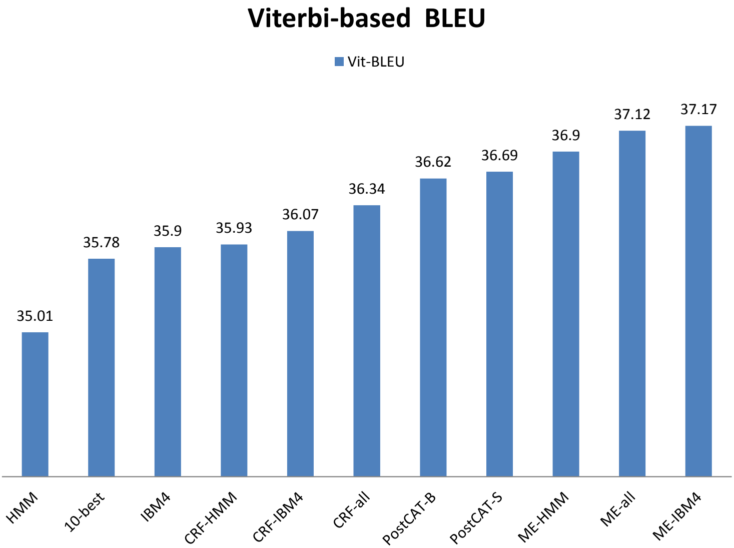
Generative

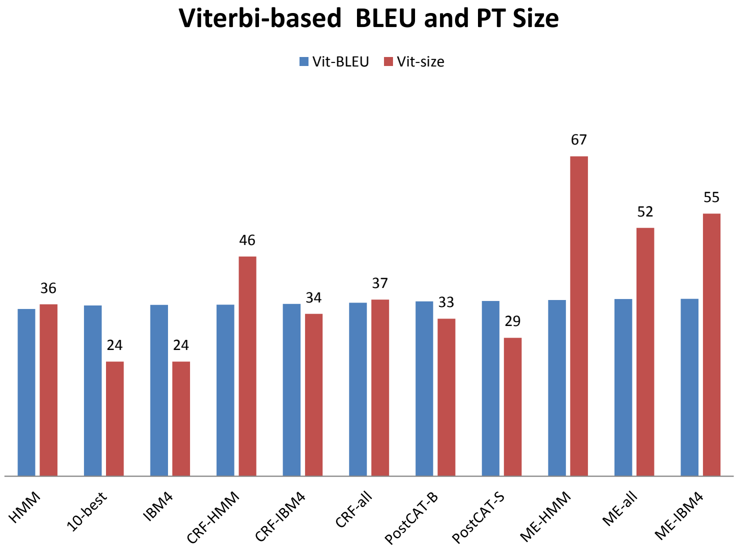
- **MGIZA++** IBM models - Features
- **10-best WAM** average link occurrences over MGIZA++ N-best alignments (Liu et al., 2009)
- **PostCAT** constrain the posteriors of latent variables in the EM algorithm for HMM (Graça et al., 2007)

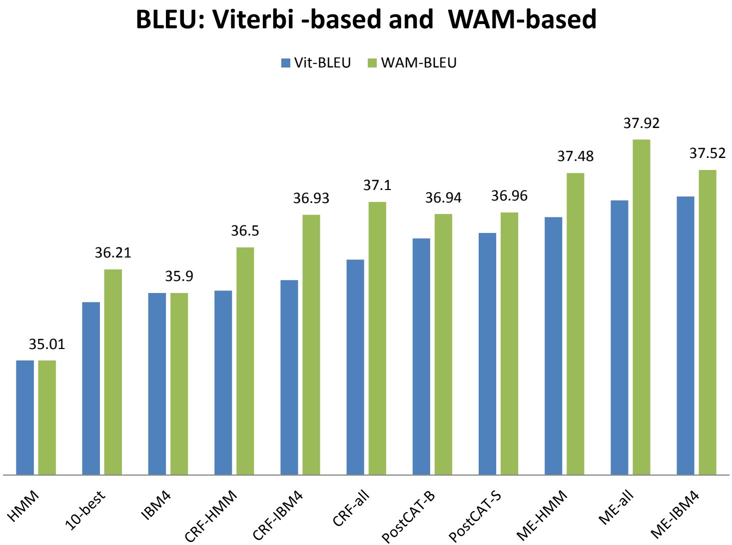
Discriminative

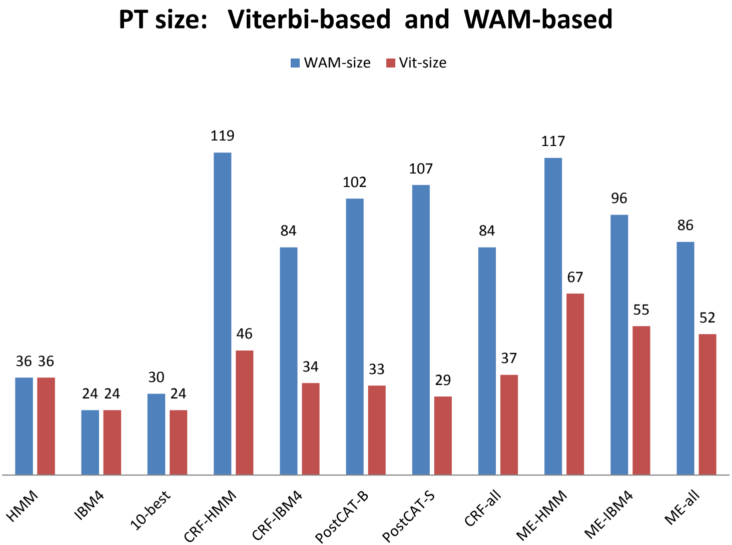
- **CRF** real-valued features - optimize L and AER (Niehues and Vogel, 2008)
- **MaxEnt** Maximum entropy based alignment system (Tomeh et al., 2010)


Viterbi-based AER
Vit-AER
28.35
25.39
24.97
24.92
23.51
22.53
22.48
21.03
17.61
15.61
14.69
HMM
CRF-HMM
IBM4
10-best
CRF-IBM4
PostCAT-B
PostCAT-S
CRF-all
ME-HMM
ME-IBM4
ME-all


Viterbi-based BLEU
Vit-BLEU
35.01
35.78
35.9
35.93
36.07
36.34
36.62
36.69
36.9
37.12
37.17
HMM
10-best
IBM4
CRF-HMM
CRF-IBM4
CRF-all
PostCAT-B
PostCAT-S
ME-HMM
ME-all
ME-IBM4


Viterbi-based BLEU and PT Size
Vit-BLEU
Vit-size
36
24
24
46
34
37
33
29
67
52
55
HMM
10-best
IBM4
CRF-HMM
CRF-IBM4
CRF-all
PostCAT-B
PostCAT-S
ME-HMM
ME-all
ME-IBM4


BLEU: Viterbi -based and WAM-based
Vit-BLEU
WAM-BLEU
35.01
36.21
35.9
36.5
36.93
37.1
36.94
36.96
37.48
37.92
37.52
HMM
10-best
IBM4
CRF-HMM
CRF-IBM4
CRF-all
PostCAT-B
PostCAT-S
ME-HMM
ME-all
ME-IBM4


PT size: Viterbi-based and WAM-based
WAM-size
Vit-size
36 36
24 24
30 24
119 46
84 34
102 33
107 29
84 37
117 67
96 55
86 52
HMM
IBM4
10-best
CRF-HMM
CRF-IBM4
PostCAT-B
PostCAT-S
CRF-all
ME-HMM
ME-IBM4
ME-all

Conclusion

- A generic algorithm to construct the translation model
- Two instantiations: Viterbi-based and WAM-based
- Generative and discriminative alignment models
- WAM-based outperforms Viterbi-based
- MaxEnt-based estimation performs best (Viterbi and WAM)

Thank you !

Questions ?