

Statistical NLP

Spring 2007



Lecture 11: Phrase Alignment

Dan Klein – UC Berkeley

Examples: Translation and Fertility

the

f	$t(f e)$	ϕ	$n(\phi e)$
le	0.497	1	0.746
la	0.207	0	0.254
les	0.155		
l'	0.086		
ce	0.018		
cette	0.011		

not

f	$t(f e)$	ϕ	$n(\phi e)$
ne	0.497	2	0.735
pas	0.442	0	0.154
non	0.029	1	0.107
rien	0.011		

farmers

f	$t(f e)$	ϕ	$n(\phi e)$
agriculteurs	0.442	2	0.731
les	0.418	1	0.228
cultivateurs	0.046	0	0.039
producteurs	0.021		

Example: Idioms

nodding

f	$t(f e)$	ϕ	$n(\phi e)$
signe	0.164	4	0.342
la	0.123	3	0.293
tête	0.097	2	0.167
oui	0.086	1	0.163
fait	0.073	0	0.023
que	0.073		
hoche	0.054		
hocher	0.048		
faire	0.030		
me	0.024		
approuve	0.019		
qui	0.019		
un	0.012		
faites	0.011		

Example: Morphology

should

f	$t(f e)$	ϕ	$n(\phi e)$
devrait	0.330	1	0.649
devraient	0.123	0	0.336
devrions	0.109	2	0.014
faudrait	0.073		
faut	0.058		
doit	0.058		
aurait	0.041		
doivent	0.024		
devons	0.017		
devrais	0.013		

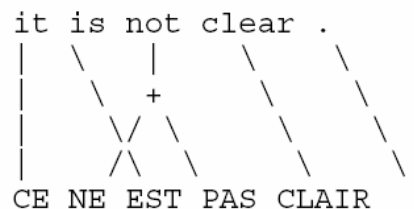
Some Results

- [Och and Ney 03]

Model	Training scheme	0.5K	8K	128K	1.47M
Dice		50.9	43.4	39.6	38.9
Dice+C		46.3	37.6	35.0	34.0
Model 1	1^5	40.6	33.6	28.6	25.9
Model 2	$1^5 2^5$	46.7	29.3	22.0	19.5
HMM	$1^5 H^5$	26.3	23.3	15.0	10.8
Model 3	$1^5 2^5 3^3$	43.6	27.5	20.5	18.0
	$1^5 H^5 3^3$	27.5	22.5	16.6	13.2
Model 4	$1^5 2^5 3^3 4^3$	41.7	25.1	17.3	14.1
	$1^5 H^5 3^3 4^3$	26.1	20.2	13.1	9.4
	$1^5 H^5 4^3$	26.3	21.8	13.3	9.3
Model 5	$1^5 H^5 4^3 5^3$	26.5	21.5	13.7	9.6
	$1^5 H^5 3^3 4^3 5^3$	26.5	20.4	13.4	9.4
Model 6	$1^5 H^5 4^3 6^3$	26.0	21.6	12.8	8.8
	$1^5 H^5 3^3 4^3 6^3$	25.9	20.3	12.5	8.7

Decoding

- In these word-to-word models
 - Finding best alignments is easy
 - Finding translations is hard (why?)



Bag “Generation” (Decoding)

soon me your as give possible please as response
the some let me disadvantages mention now of
missions research our has in two organization

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

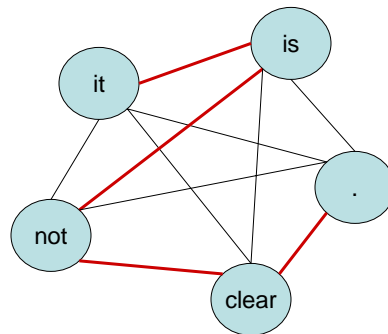
Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

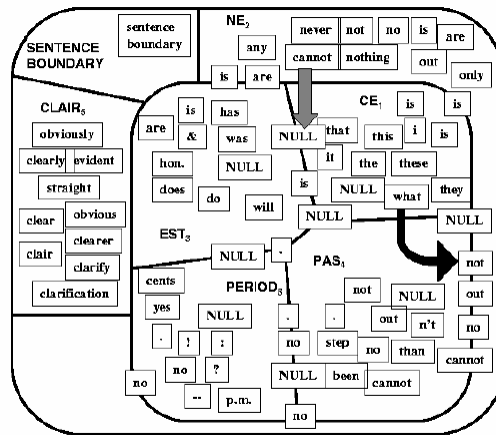
In our organization research has two missions.
⇒ In our missions research organization has two.

Bag Generation as a TSP

- Imagine bag generation with a bigram LM
 - Words are nodes
 - Edge weights are $P(w|w')$
 - Valid sentences are Hamiltonian paths
- Not the best news for word-based MT!



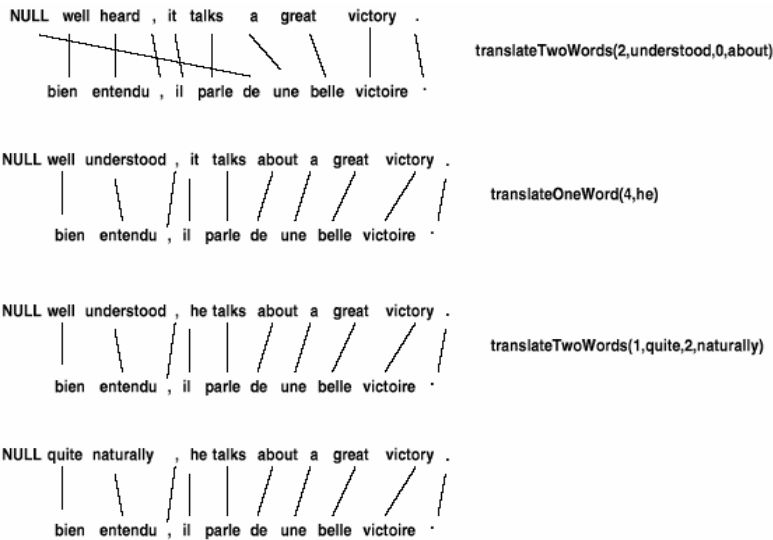
IBM Decoding as a TSP



Decoding, Anyway

- Simplest possible decoder:
 - Enumerate sentences, score each with TM and LM
- Greedy decoding:
 - Assign each French word it's most likely English translation
 - Operators:
 - Change a translation
 - Insert a word into the English (zero-fertile French)
 - Remove a word from the English (null-generated French)
 - Swap two adjacent English words
 - Do hill-climbing (or annealing)

Greedy Decoding

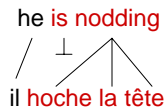


Stack Decoding

- **Stack decoding:**
 - Beam search
 - Usually A* estimates for completion cost
 - One stack per candidate sentence length
- **Other methods:**
 - Dynamic programming decoders possible if we make assumptions about the set of allowable permutations

sent length	decoder type	time (sec/sent)	search errors	translation errors (semantic and/or syntactic)	NE	PME	DSE	FSE	HSE	CE
6	IP	47.50	0	57	44	57	0	0	0	0
6	stack	0.79	5	58	43	53	1	0	0	4
6	greedy	0.07	18	60	38	45	5	2	1	10
8	IP	499.00	0	76	27	74	0	0	0	0
8	stack	5.67	20	75	24	57	1	2	2	15
8	greedy	2.66	43	75	20	38	4	5	1	33

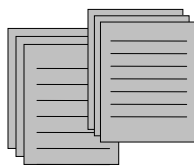
Phrases in IBM Models



nodding

f	$t(f e)$	ϕ	$n(\phi e)$
signe	0.164	4	0.342
la	0.123	3	0.293
tête	0.097	2	0.167
oui	0.086	1	0.163
fait	0.073	0	0.023
que	0.073		
hoche	0.054		
hocher	0.048		
faire	0.030		
me	0.024		
approuve	0.019		
qui	0.019		
un	0.012		
faites	0.011		

Overview: Extracting Phrases

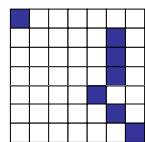


Sentence-aligned corpus

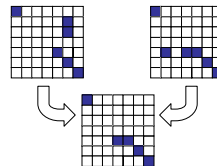
```

cat ||| chat ||| 0.9
the cat ||| le chat ||| 0.8
dog ||| chien ||| 0.8
house ||| maison ||| 0.6
my house ||| ma maison ||| 0.9
language ||| langue ||| 0.9
...
    
```

Phrase table
(translation model)

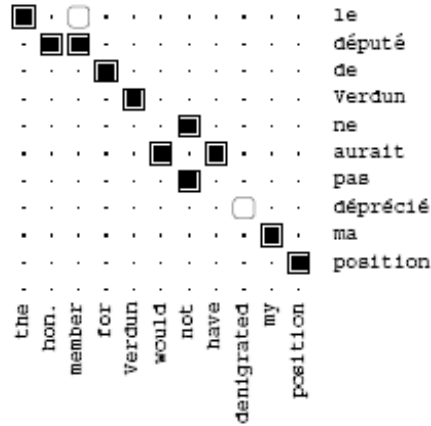


Directional word alignments

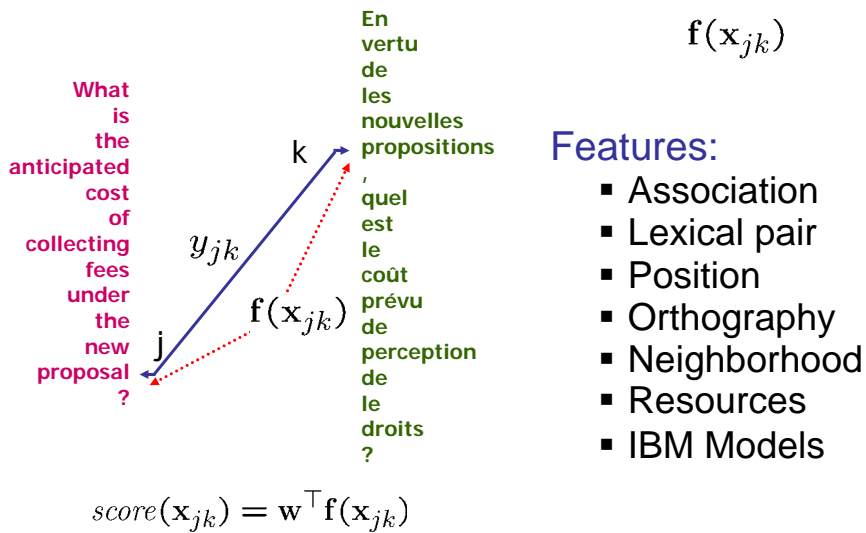


Intersected and grown word alignments

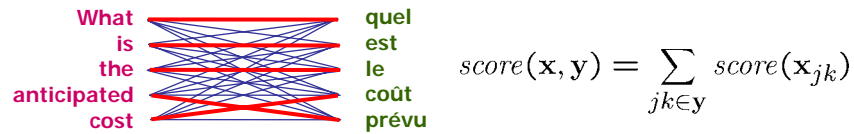
Complex Configurations



Feature-Based Alignment



Finding Viterbi Alignments



- Complete bipartite graph
- Maximum score matching with node degree ≤ 1

$$\mathbf{y} = \arg \max_{\mathbf{y}' \in \mathcal{Y}} score(\mathbf{x}, \mathbf{y}') = \arg \max_{\mathbf{y}' \in \mathcal{Y}} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}')$$

⇒ Weighted bipartite matching problem

[Lacoste-Julien, Taskar, Jordan, and Klein, 05]

Learning \mathbf{w}

- Supervised training data

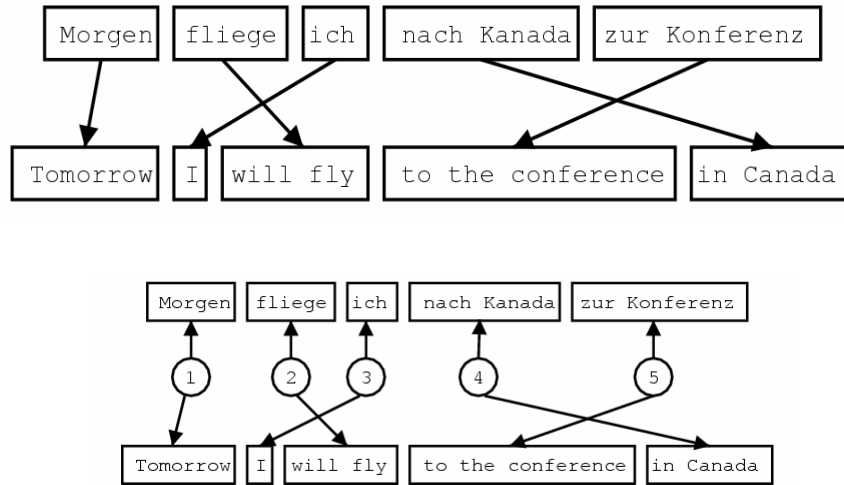
$(\mathbf{x}^i, \mathbf{y}^i)$



- Training methods
 - Maximum likelihood/entropy
 - Perceptron
 - Maximum margin

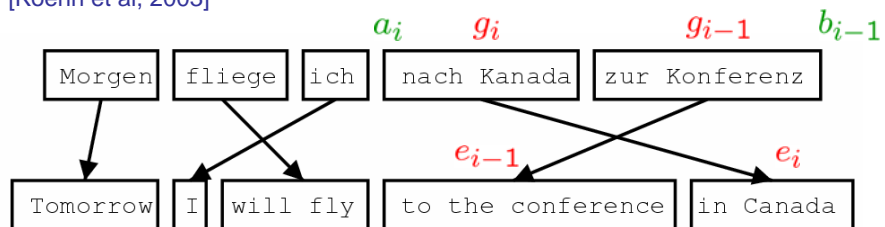
[Lacoste-Julien, Taskar, Jordan, and Klein, 05]

Phrase-Based Systems



Pharaoh's Model

[Koehn et al, 2003]



$$P(e|g) = P(\{\bar{g}_i\}|g) \prod_i \phi(\bar{e}_i|\bar{g}_i) d(a_i - b_{i-1})$$

Segmentation

Translation

Distortion

Pharaoh's Model

$$P(f|e) = P(\{\bar{e}_i\}|e) \prod_i \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

$\frac{1}{K}$
 $\frac{\text{count}(\bar{f}_i, \bar{e}_i)}{\text{count}(\bar{e}_i)}$
 $\alpha^{|a_i - b_{i-1}|}$

Where do we get these counts?

Phrase-Based Decoding

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

the	7 people	including	by some	and	the russian	the	the astronauts	,
it	7 people included		by france	and the	the russian		international astronautical	of rapporteur .
this	7 out	including the	from	the french	and the russian	the fifth		members .
these	7 among	including from		the french and	of the russian	of	space	members .
that	7 persons	including from the		of france	and to	russian	of the	members .
	7 include		from the	of france and	russian		astronauts	the
	7 numbers include		from france	and russian		of astronauts who		."
	7 populations include		those from france	and russian		astronauts .		
	7 deportees included		come from	france	and russia	in	astronautical	personnel ;
	7 philtrum	including those from		france and	ruusia	a space		member
		including representatives from		france and the	ruusia		astronaut	
		include	came from	france and russia		by cosmonauts		
		include representatives from	french	and russia		cosmonauts		
		include	came from france	and russia 's		cosmonauts .		
		includes	coming from	french and	russia 's	cosmonaut		
			french and	russian	russia 's	's	astronavigation	member .
			french and	and russia		astronauts		
				and russia 's				special rapporteur
				, and	ruusia			rapporteur
				, and russia				rapporteur .
				, and russia				
				or	russia 's			

Decoder design is important: [Koehn et al. 03]

Phrase Scoring

$$\phi_{\text{new}}(e_j | f_i) = \frac{c(f_i, e_j)}{c(f_i)}$$

	les	chats	aiment	le	poisson	frais	.
cats							
like							
fresh							
fish							
.							

} } }

- Learning weights has been tried, several times:
 - [Marcu and Wong, 02]
 - [DeNero et al, 06]
 - ... and others

- Seems not to work, for a variety of only partially understood reasons

- Main issue: big

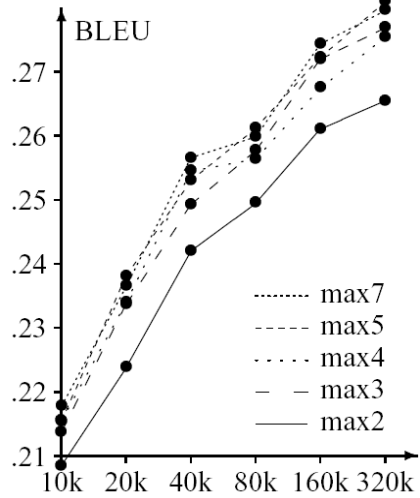
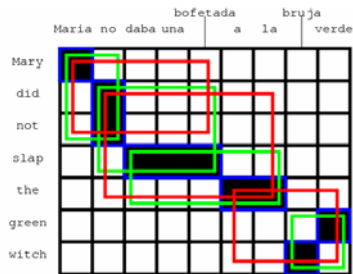
Extracting Phrases

	Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary									
did									
not									
slap									
the									
green									
witch									

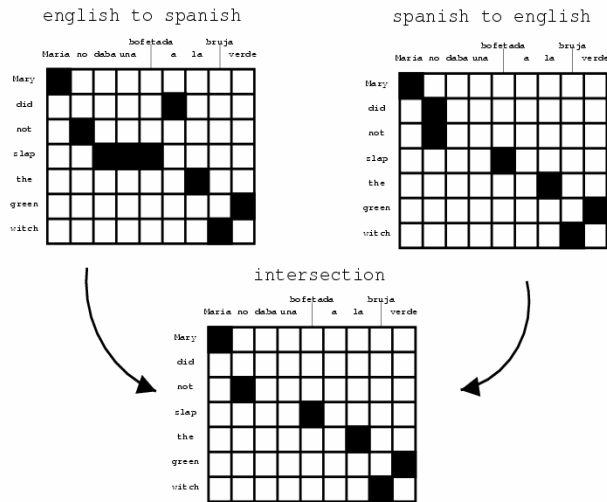
} } }

Phrase Size

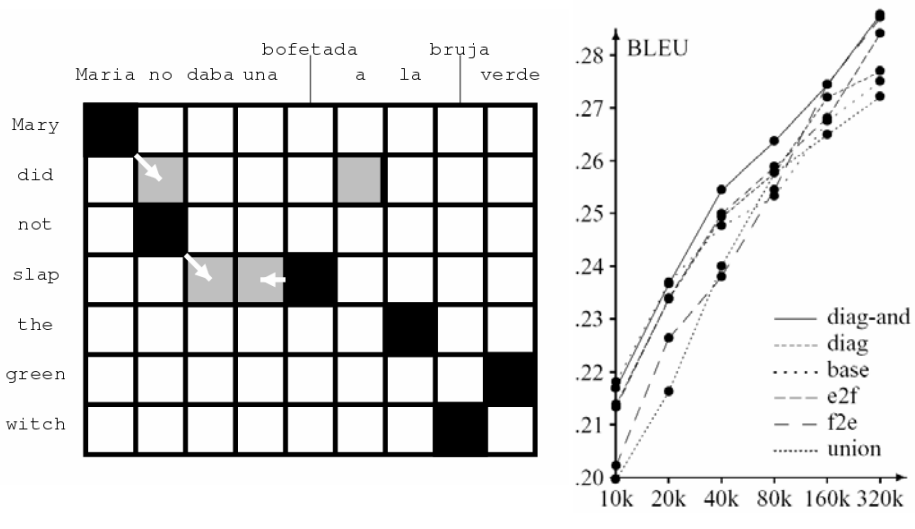
- Phrases do help
 - But they don't need to be long
 - Why should this be?



Bidirectional Alignment

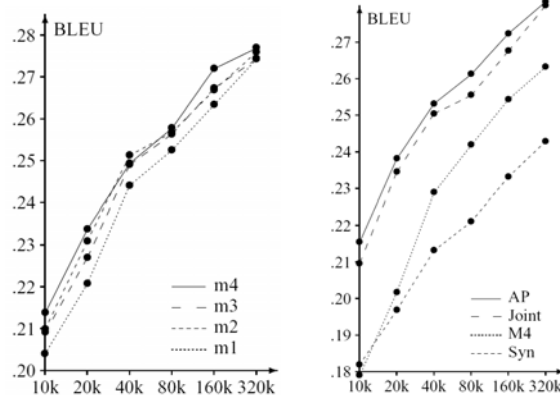


Alignment Heuristics



Sources of Alignments

Method	Training corpus size					
	10k	20k	40k	80k	160k	320k
AP	84k	176k	370k	736k	1536k	3152k
Joint	125k	220k	400k	707k	1254k	2214k
Syn	19k	24k	67k	105k	217k	373k



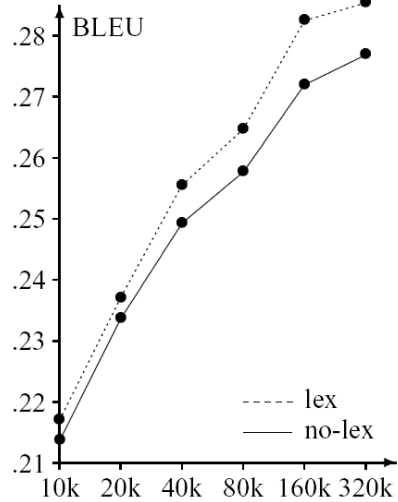
Lexical Weighting

$$\phi(\bar{f}_i|\bar{e}_i) = \frac{\text{count}(\bar{f}_i, \bar{e}_i)}{\text{count}(\bar{e}_i)} p_w(\bar{f}_i|\bar{e}_i)$$

```

      f1 f2 f3
NULL -- -- ##
e1  ## -- --
e2  -- ## --
e3  -- ## --
    
```

$$\begin{aligned}
 p_w(\bar{f}|\bar{e}, a) &= p_w(f_1 f_2 f_3 | e_1 e_2 e_3, a) \\
 &= w(f_1|e_1) \\
 &\quad \times \frac{1}{2}(w(f_2|e_2) + w(f_2|e_3)) \\
 &\quad \times w(f_3|\text{NULL})
 \end{aligned}$$



The Pharaoh Decoder

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a	slap	by		green	witch
	no		slap		to the			
	did not give				to			
					the			
				slap		the	witch	

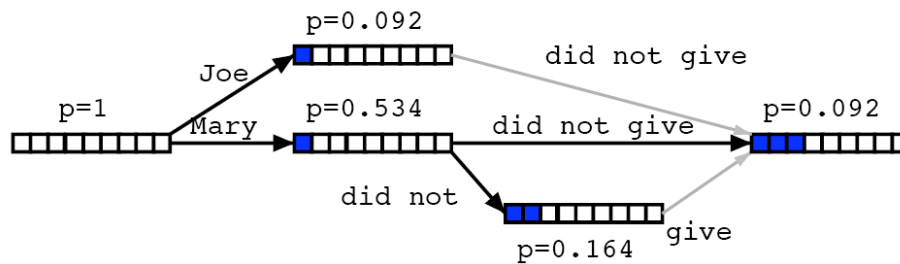
Maria	no	dio una bofetada	a la	bruja	verde
-------	----	------------------	------	-------	-------

Mary	did not	slap	the	green	witch
------	---------	------	-----	-------	-------

- Probabilities at each step include LM and TM

Hypothesis Lattices

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		by		green witch	
	no		slap		to the			
	did not give				to			
					the			
				slap			the witch	



Pruning

Maria no dio una bofetada a la bruja verde

e: Mary did not
f: **-----
p: 0.154

**better
partial
translation**

e: the
f: -----*--
p: 0.354

**covers
easier part
--> lower cost**

- Problem: easy partial analyses are cheaper
 - Solution 1: use beams per foreign subset
 - Solution 2: estimate forward costs (A*-like)

WSD?

- Remember when we discussed WSD?
 - Word-based MT systems rarely have a WSD step
 - Why not?

What's Next?

- Modeling syntax
 - PCFGs and phrase structure
 - Syntactic parsing
 - Grammar induction
 - Syntactic language and translation models
- Speech systems
 - Acoustics
 - Applications