CA446

Statistical Machine Translation

Week 8: Decoding

Lecturer: Qun Liu

Lab Tutor: Xiaofeng Wu, Iacer Calixto

2nd Semester, 2014-2015 Academic Year

http://computing.dcu.ie/~qliu/CA446

Content

- **Phrase-based Translation Model**
- Distance-based Reordering
- Log-linear Model
- Decoding
- Make Decoding Manageable
- Exercises

Recap of Noisy Channel Model

$$p(e|f) = p(f|e) * p(e)$$

- $p(f|e)$ is the translation model
- $p(e)$ is the language model

Decomposition of the translation model

$$p(f|e) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1)$$

Decomposition of the translation model

$$p(f|e) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1)$$

- f is segmented into I phrases

Decomposition of the translation model

$$p(f|e) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1)$$

- f is segmented into I phrases
- ϕ is the phrase table translation probability

Decomposition of the translation model

$$p(f|e) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1)$$

- f is segmented into I phrases
- ϕ is the phrase table translation probability
- d is the distance-based reordering function.

Decomposition of the translation model

$$p(f|e) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1)$$

- f is segmented into I phrases
- ϕ is the phrase table translation probability
- d is the distance-based reordering function.
- $start_i$ is the position of the first word of f_i

Decomposition of the translation model

$$p(f|e) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1)$$

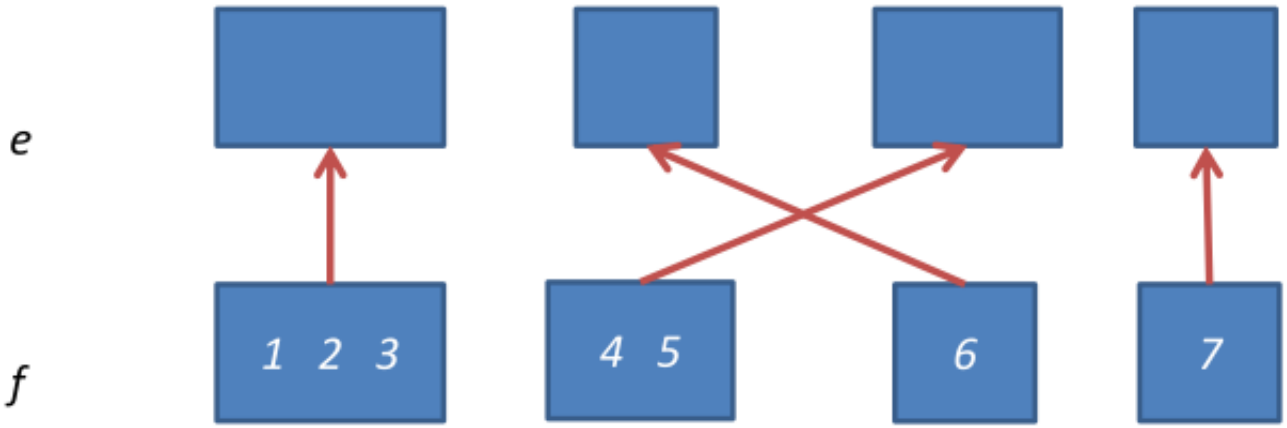
- f is segmented into I phrases
- ϕ is the phrase table translation probability
- d is the distance-based reordering function.
- $start_i$ is the position of the first word of f_i
- end_i is the position of the last word in f_i

Content

- Phrase-based Translation Model
- **Distance-based Reordering**
- Log-linear Model
- Decoding
- Make Decoding Manageable
- Exercises

Distance based Reordering

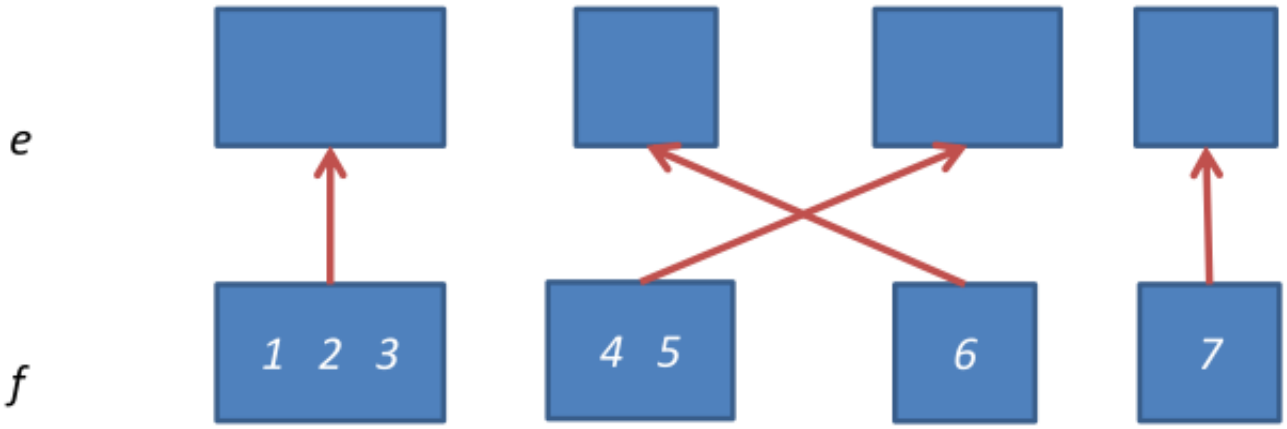
What are the distances associated with each of the English phrases?



i	f_i	$start_i - end_i - 1$	Distance
1	1-3	?	?
2	6	?	?
3	4-5	?	?
4	7	?	?

Distance based Reordering

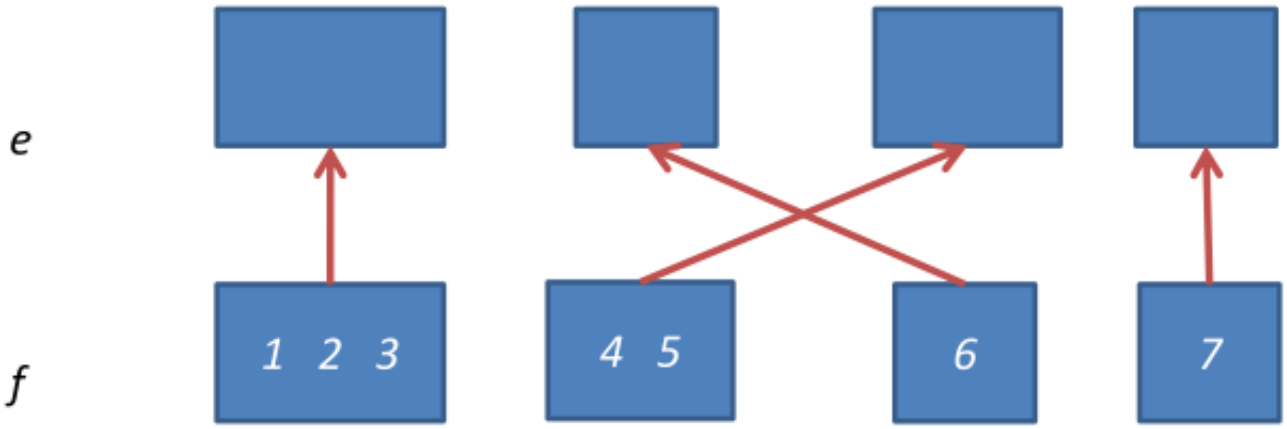
What are the distances associated with each of the English phrases?



i	f_i	$start_i - end_i - 1$	Distance
1	1-3	1-0-1	0
2	6	?	?
3	4-5	?	?
4	7	?	?

Distance based Reordering

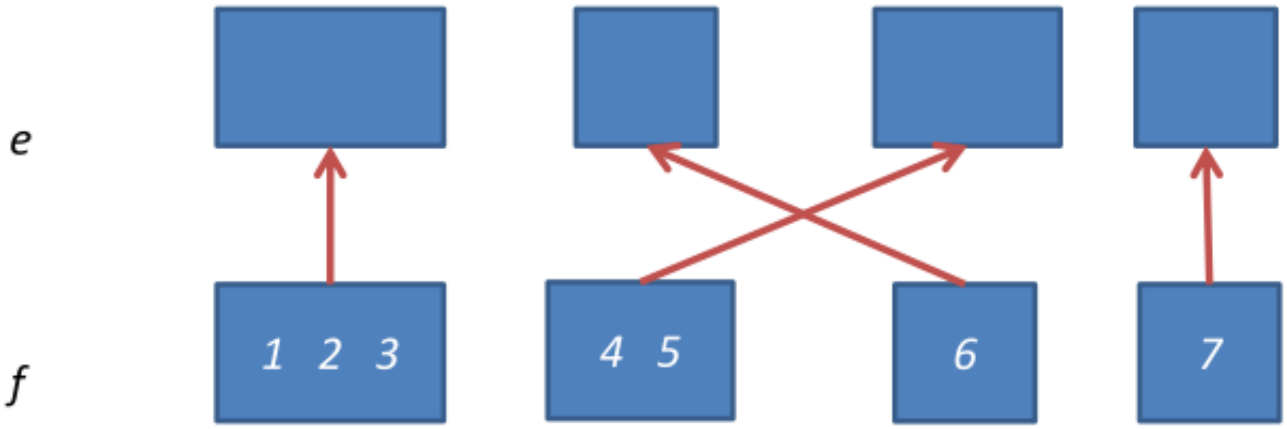
What are the distances associated with each of the English phrases?



i	f_i	$start_i - end_i - 1$	Distance
1	1-3	1-0-1	0
2	6	6-3-1	2
3	4-5	?	?
4	7	?	?

Distance based Reordering

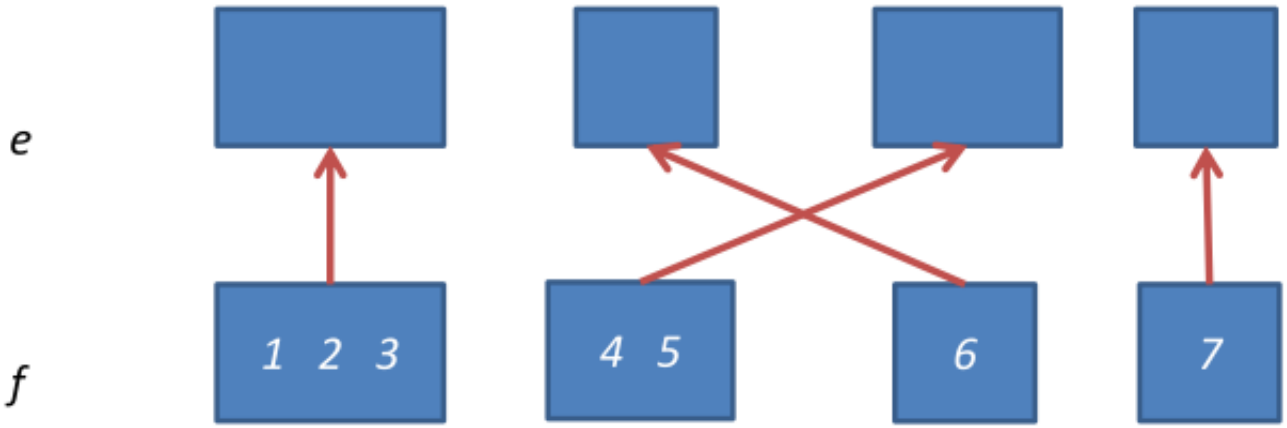
What are the distances associated with each of the English phrases?



i	f_i	$start_i - end_i - 1$	Distance
1	1-3	1-0-1	0
2	6	6-3-1	2
3	4-5	4-6-1	-3
4	7	?	?

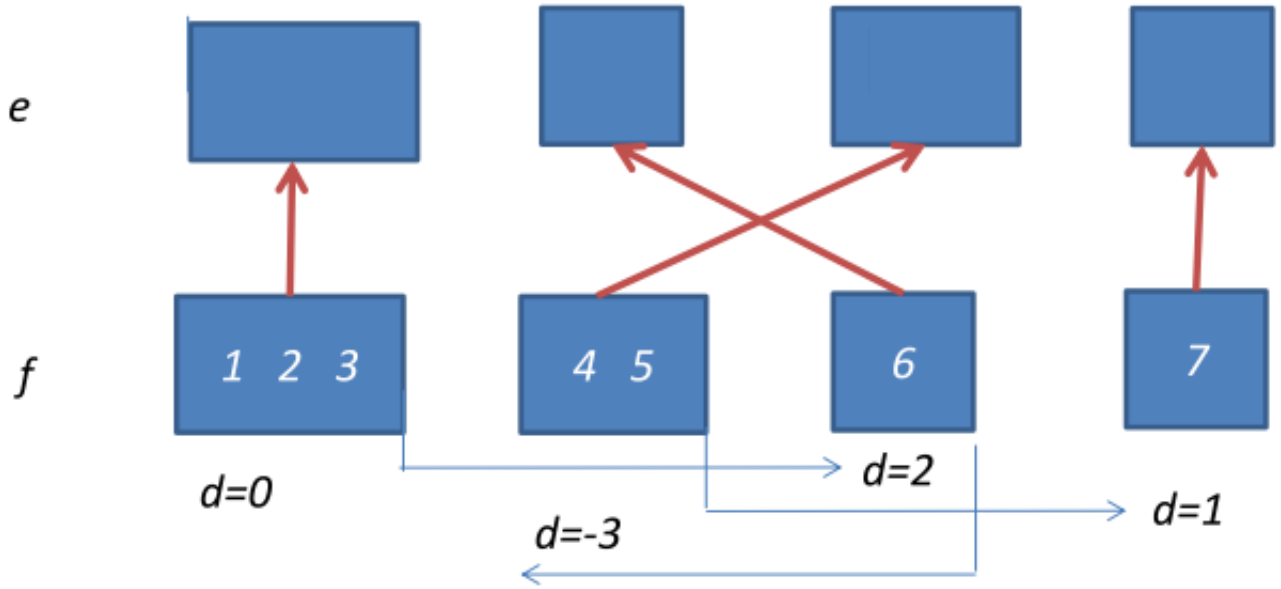
Distance based Reordering

What are the distances associated with each of the English phrases?



i	f_i	$start_i - end_i - 1$	Distance
1	1-3	1-0-1	0
2	6	6-3-1	2
3	4-5	4-6-1	-3
4	7	7-5-1	1

Distance based Reordering



The distance-based reordering function d is defined so that it penalises reordering over long distances.

Content

- Phrase-based Translation Model
- Distance-based Reordering
- **Log-linear Model**
- Decoding
- Make Decoding Manageable
- Exercises

Noisy Channel Model Revisited

$$p(e|f) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1) \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \dots e_{i-1})$$

3 submodels: ϕ, d, p_{LM}

Noisy Channel Model Revisited

$$p(e|f) = \prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1) \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \dots e_{i-1})$$

We may want to emphasise one submodel over another.

Noisy Channel Model Revisited

$$\prod_{i=1}^{I} \phi(f_i|e_i)\, d(start_i - end_i - 1) \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \ldots e_{i-1})$$

Add weights to yield:

$$\prod_{i=1}^{I} \phi(f_i|e_i)^{\lambda_\phi}\, d(start_i - end_i - 1)^{\lambda_d} p_{LM}(e_i|e_1 \ldots e_{i-1})^{\lambda_{p_{LM}}}$$

Log-linear Model

What we end up with is a log-linear model:

$$p(x) = \exp(\sum_{i=1}^{n} \lambda_i h_i(x))$$

where:

- $n = 3$
- $h_1(x) = \log \phi$
- $h_2(x) = \log d$
- $h_3(x) = \log p_{LM}$

Log-linear Model

What is the advantage of reformulating the translation formula in this way?

Log-linear Model

What is the advantage of reformulating the translation formula in this way?

It makes it easier to add in more information sources:

Log-linear Model

What is the advantage of reformulating the translation formula in this way?

It makes it easier to add in more information sources:

- Multiple translation models
- Multiple language models
- Linguistic information
- Lexical probabilities as well as phrase probabilities

Content

- Phrase-based Translation Model
- Distance-based Reordering
- Log-linear Model
- **Decoding**
- Make Decoding Manageable
- Exercises

Noisy Channel Model Revisited

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e)p(f|e)$$

Decoder

Translation Model

Language Model

What is Decoding?

Process of searching for the best translation among all possible translations:

$$e_{best} = argmax_{e} p(e|f)$$

Decoding Process

Maria no dio una bofetada a la bruja verde

Build translation left to right

- Select phrase to be translated

Decoding Process

Build translation left to right

- Select phrase to be translated
- Find phrase translation

Decoding Process

Build translation left to right

- Select phrase to be translated
- Find phrase translation
- Add phrase to end of partial translation

Decoding Process

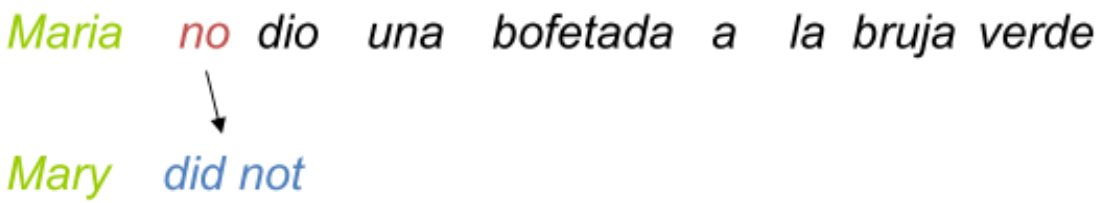
Maria no dio una bofetada a la bruja verde

Mary

Build translation left to right

- Select phrase to be translated
- Find phrase translation
- Add phrase to end of partial translation
- Mark words as translated

Decoding Process



One to many translation

Decoding Process

Many to one translation

Decoding Process

Many to one translation

Decoding Process

Reordering

Decoding Process

Translation finished!

Translation Options

- Many different ways to segment words into phrases
- Many different ways to translate each phrase

Decoding is a Complex Process!

Phrase-Based Translation

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	including the	**from**	the french	and the russian		the fifth		.	
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			
	7 include		from the	of france and				**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		russia	a space		**member**	
		including representatives from		france and the		**russia**		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			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

Thanks to Kevin Knight

Content

- Phrase-based Translation Model
- Distance-based Reordering
- Log-linear Model
- Decoding
- **Make Decoding Manageable**
- Exercises

Making decoding manageable

The decoding problem is NP-complete which means that exhaustively examining all possible translations, scoring them and picking the best is computationally too expensive for an input sentence of even modest length (Koehn, 2010, p.155).

Making decoding manageable

Two strategies:

1. Hypothesis Recombination
2. Pruning the search space (heuristic)

Hypothesis Recombination

- A translation *hypothesis* is a partial translation.

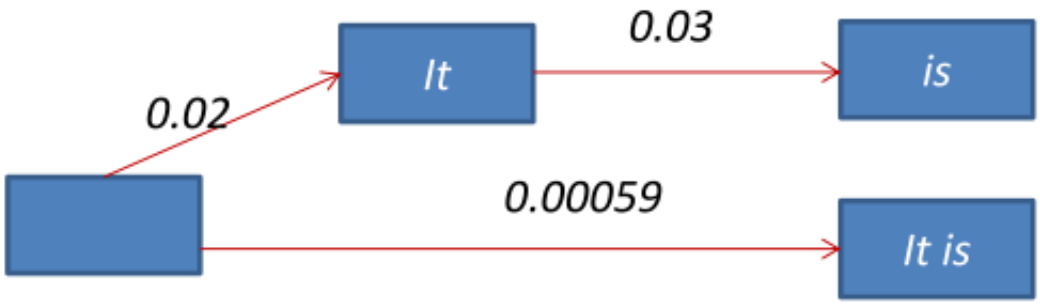
Hypothesis Recombination

- A translation *hypothesis* is a partial translation.
- We can arrive at the same partial translation in more than one way.

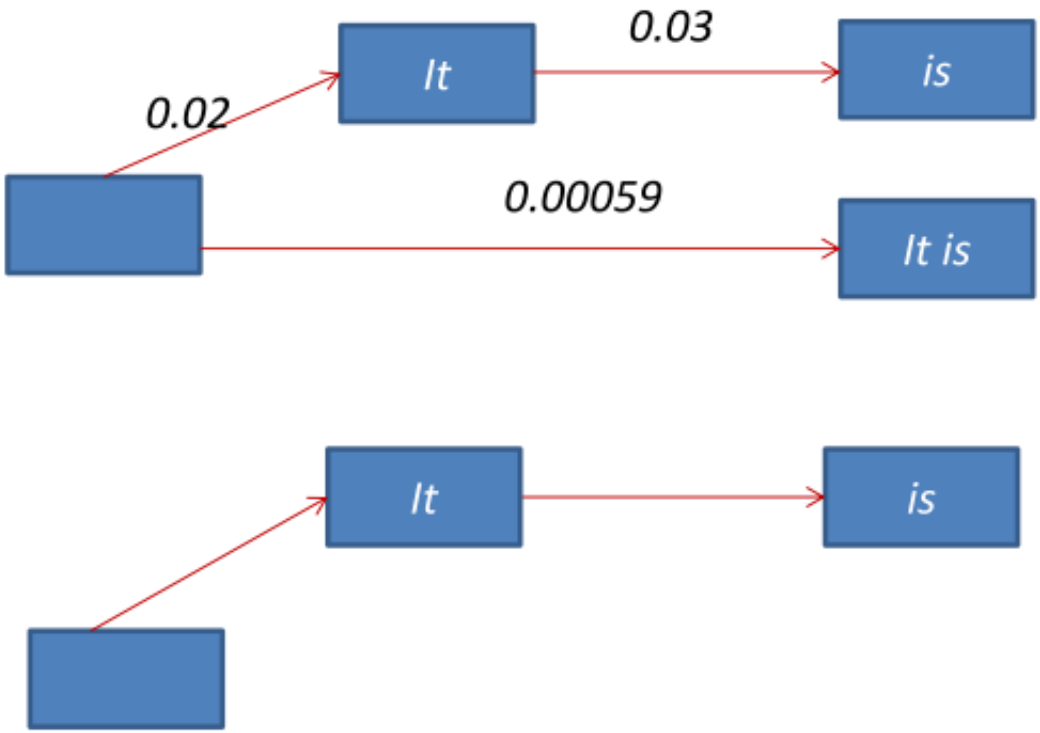
Hypothesis Recombination

- A translation *hypothesis* is a partial translation.
- We can arrive at the same partial translation in more than one way.
- Hypothesis recombination takes advantage of this by storing only the most likely path associated with a particular hypothesis.

Hypothesis Recombination



Hypothesis Recombination



Pruning the search space

- Pruning is the process of deleting unlikely hypotheses to reduce the search space.
- One way to do this is to store hypotheses in stacks based on the number of words translated.
- Unlikely hypotheses can be pruned from each stack.

Pruning the search space

Two types of pruning strategies

1. Histogram pruning: keep a maximum of m hypotheses in a stack

Pruning the search space

Two types of pruning strategies

1. Histogram pruning: keep a maximum of m hypotheses in a stack
2. Threshold or beam pruning: keep only those hypotheses that are within a threshold α of the best hypothesis. Any hypothesis that is α times worse than the best hypothesis is pruned.

Histogram Pruning

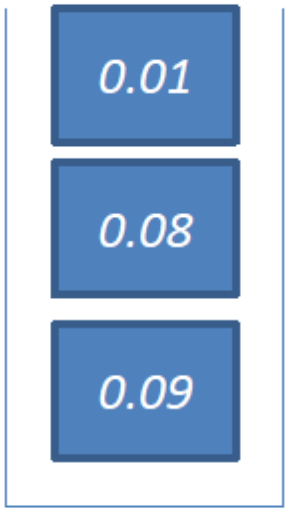
Keep a maximum of *m* hypotheses

Threshold Pruning

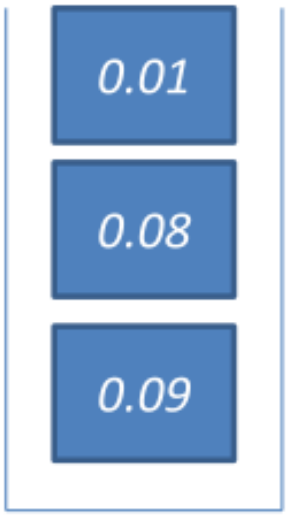
Keep only those hypotheses that are within a threshold α of the best hypothesis.

Histogram Pruning

How many hypotheses will be pruned if m = 1?

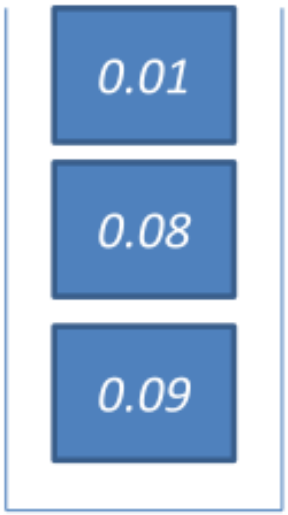


Threshold Pruning



How many hypotheses will be pruned if we prune all hypotheses that are at least 0.5 times worse than the best hypothesis?

Threshold Pruning



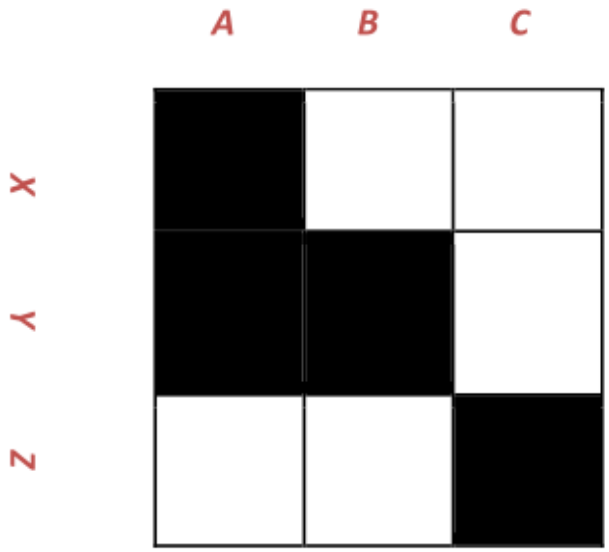
Threshold pruning is more flexible than histogram pruning since it takes into account the difference between the scores of the best and worst hypothesis.

Content

- Phrase-based Translation Model
- Distance-based Reordering
- Log-linear Model
- Decoding
- Make Decoding Manageable
- **Exercises**

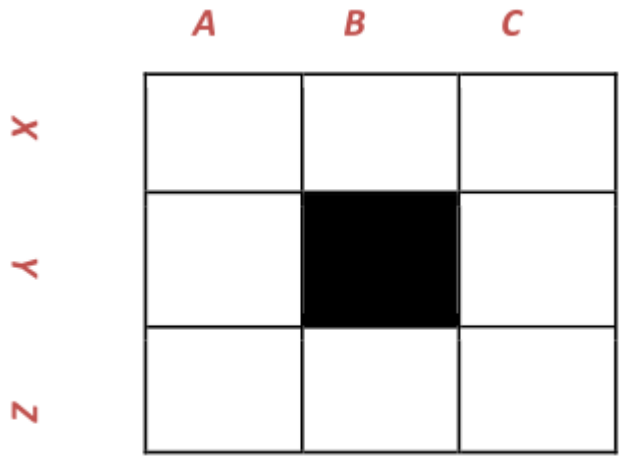
Exercises

List all phrase pairs that are consistent with the following word alignment:



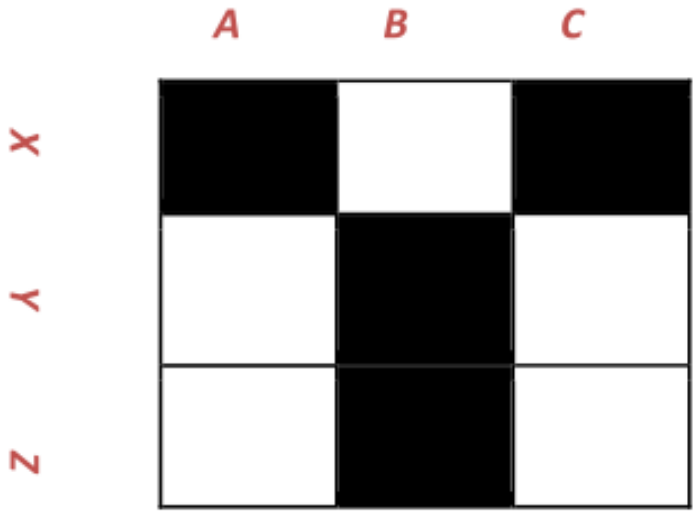
Exercises

List all phrase pairs that are consistent with the following word alignment:



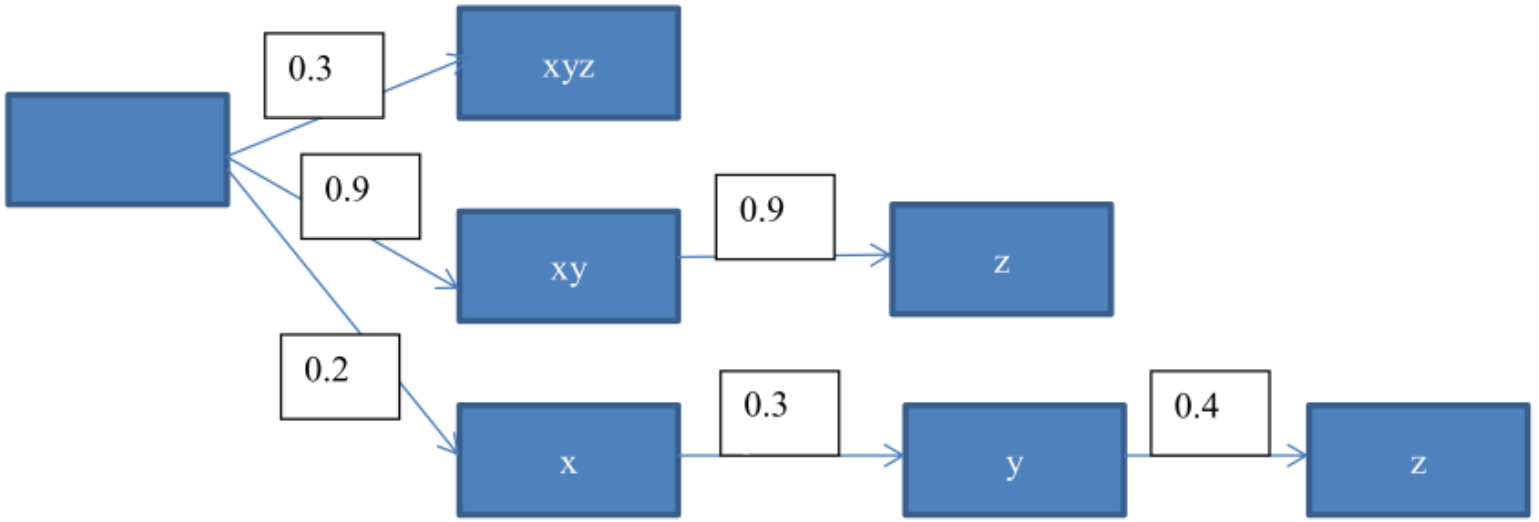
Exercises

List all phrase pairs that are consistent with the following word alignment:



Exercises

Which hypothesis will remain after hypothesis recombination?



Discussion

Acknowledgement

Parts of the content of this lecture are taken from previous lectures and presentations given by Jennifer Foster, Declan Groves, Yvette Graham, Kevin Knight, Josef van Genabith, Andy Way.