

# Phrase-based SMT

Miguel Rios

Universiteit van Amsterdam

April 27, 2018

# Content

① Introduction

② Model

③ Prediction

# Recap

We looked into Alignment a directional word-based model.

- Parametrisation: Categorical.
- Estimation techniques: EM vs VB.

# Recap

We looked into Alignment a directional word-based model.

- Parametrisation: Categorical.
- Estimation techniques: EM vs VB.

We have not look into generation:

- No model of length
- No model of segmentation
- Bad model for translation

# Translation

Model:

$$P(E|F) = \frac{P(E)P(F|E)}{P(F)}$$

Prediction:

$$\hat{E} = \arg \max_E P(E)P(F = f|E)$$

Estimation:

- $P(E)$   $n$ -gram LM.
- $P(F|E)$  TM.

# Word-based SMT

[Brown et al., 1993]

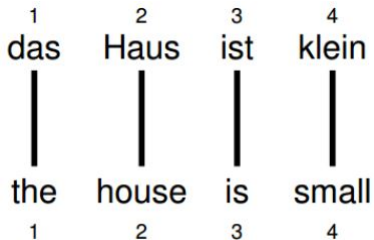


Figure: Koehn [2010]

# Limitations of word-based approach

## Linguistically

- Can not translate many-to-one or many-to-many
- Compositionality of translation  
multi-word / idiomatic expressions.

## Computationally during prediction

- $n!$  permutations in decoding.

# Phrase-based model

Change of units: phrase.

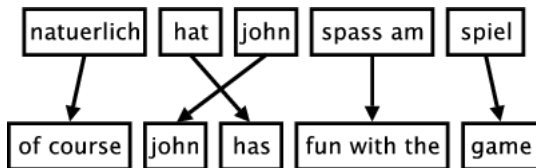


Figure: Koehn [2010]



# Phrase-based model

## Phrase pairs as translation units

- Capture non-compositional translations.
- Exploit (local) reordering patterns.

# Illustration

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

# Illustration

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

$J'_1$   $ai_2$   $les_3$   $yeux_4$   $noirs_5$

input

# Illustration

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

$J'_1$   $ai_2$   $les_3$   $yeux_4$   $noirs_5$   
 $[J'_1 \ ai_2]$   $[les_3 \ yeux_4]$   $[noirs_5]$

input  
segmentation

# Illustration

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

$J'_1$   $ai_2$   $les_3$   $yeux_4$   $noirs_5$

$[J'_1 \ ai_2]$   $[les_3 \ yeux_4]$   $[noirs_5]$

$[J'_1 \ ai_2]_1$   $[noirs_5]_3$   $[les_3 \ yeux_4]_2$

input

segmentation

ordering

# Illustration

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

$J'_1$   $ai_2$   $les_3$   $yeux_4$   $noirs_5$

$[J'_1 \ ai_2]$   $[les_3 \ yeux_4]$   $[noirs_5]$

$[J'_1 \ ai_2]_1$   $[noirs_5]_3$   $[les_3 \ yeux_4]_2$

$[I \ have]_1$   $[black]_3$   $[eyes]_2$

input

segmentation

ordering

translation

# Illustration

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

J'<sub>1</sub> ai<sub>2</sub> les<sub>3</sub> yeux<sub>4</sub> noirs<sub>5</sub>

[J'<sub>1</sub> ai<sub>2</sub>] [les<sub>3</sub> yeux<sub>4</sub>] [noirs<sub>5</sub>]

[J'<sub>1</sub> ai<sub>2</sub>]<sub>1</sub> [noirs<sub>5</sub>]<sub>3</sub> [les<sub>3</sub> yeux<sub>4</sub>]<sub>2</sub>

[I have]<sub>1</sub> [black]<sub>3</sub> [eyes]<sub>2</sub>

input

segmentation

ordering

translation

**Derivation**

# Modelling Derivations

$$P(e, d|f) = \frac{\exp(S_{\theta}(e, d, f))}{\sum_{e'} \sum_{d'} \exp(S_{\theta}(e', d', f))}$$



# Modelling Derivations

$$P(e, d|f) = \frac{\exp(S_{\theta}(e, d, f))}{\sum_{e'} \sum_{d'} \exp(S_{\theta}(e', d', f))}$$

Challenging normalisation.

Large space of derivations:

- Number of segments.
- Number of permutations.
- Number of translations.

# Discriminative classifier

- Give up on marginalisation of  $d$
- Give up on probabilistic modelling
- How?

# Discriminative classifier

- Give up on marginalisation of  $d$
- Give up on probabilistic modelling
- How?
- If we look at the prediction:

$$\begin{aligned}
 \hat{e}, \hat{d} &= \arg \max_{e, d|f} \log P(e, d|f) \\
 &= \arg \max_{e, d|f} S_{\theta}(e, d, f) - \underbrace{\log \sum_{e'} \sum_{d'} \exp(S_{\theta}(e', d', f))}_{\text{constant for any}(e, d|f)} \\
 &= \arg \max_{e, d|f} S_{\theta}(e, d, f)
 \end{aligned}$$

Trained discriminatively (e.g. structured perceptron).

# Linear model

The score function  $S_\theta$  is defined as a linear model.

$$S_\theta(e, d, f) = \theta^T H(e, d, f)$$

where  $\theta$  are parameters

$h$  are feature functions.

# Linear model

The score function  $S_\theta$  is defined as a linear model.

$$S_\theta(e, d, f) = \theta^T H(e, d, f)$$

where  $\theta$  are parameters

$h$  are feature functions.

Linear model decomposes over phrases.

$$S_\theta(e, d, f) = \theta^T \sum_i^n \underbrace{h_i(d_i|e, f)}_{\text{local feature function}}$$

Model featurises steps in the derivation independently.

# PBSMT Model

- Feature functions  $n = 3$
- Translation feature function:

$$h_1 = \log P(\bar{f}|\bar{e})$$

- Language Model feature function:

$$h_2 = \log P(e|e_{\text{past}})$$

- Distortion feature function:

$$h_3 = \log d(\text{start}_k - \text{end}_{k-1} - 1)$$

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair



# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair

# Phrase pairs from word alignments

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an “observed” phrase pair
- we extract all of them once, irrespective of derivation

# Phrase Table

- Goal: Learn phrase translation table from parallel corpus.



# Phrase Table

- Goal: Learn phrase translation table from parallel corpus.
- Three stages:
- Word alignment given IBM.
- Extraction of phrase pairs.
- Phrase scoring.

# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

Let  $A$  be an alignment matrix

# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

Let  $A$  be an alignment matrix

$(\bar{f}, \bar{e})$  consistent with  $A$  if, and only if:

# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

Let  $A$  be an alignment matrix

$(\bar{f}, \bar{e})$  consistent with  $A$  if, and only if:

- Words in  $\bar{f}$ , if aligned, align only with words in  $\bar{e}$

# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

Let  $A$  be an alignment matrix

$(\bar{f}, \bar{e})$  consistent with  $A$  if, and only if:

- Words in  $\bar{f}$ , if aligned, align only with words in  $\bar{e}$

C

•		
	•	•

C

•		
	•	•

I

•		
	•	•

# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

Let  $A$  be an alignment matrix

$(\bar{f}, \bar{e})$  consistent with  $A$  if, and only if:

- Words in  $\bar{f}$ , if aligned, align only with words in  $\bar{e}$

C

•		
	•	•

C

•		
	•	•

I

•		
	•	•

- Words in  $\bar{e}$ , if aligned, align only with words in  $\bar{f}$

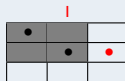
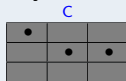
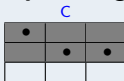
# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

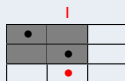
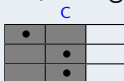
Let  $A$  be an alignment matrix

$(\bar{f}, \bar{e})$  consistent with  $A$  if, and only if:

- Words in  $\bar{f}$ , if aligned, align only with words in  $\bar{e}$



- Words in  $\bar{e}$ , if aligned, align only with words in  $\bar{f}$



# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

Let  $A$  be an alignment matrix

$(\bar{f}, \bar{e})$  consistent with  $A$  if, and only if:

- Words in  $\bar{f}$ , if aligned, align only with words in  $\bar{e}$

C

•		
	•	•

C

•		
	•	•

I

•		
	•	•

- Words in  $\bar{e}$ , if aligned, align only with words in  $\bar{f}$

C

•		
	•	
	•	

C

•		
	•	
	•	

I

•		
	•	
	•	

- $(\bar{f}, \bar{e})$  must contain at least one alignment point



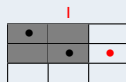
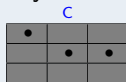
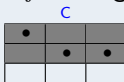
# Phrase extraction

Let  $(\bar{f}, \bar{e})$  be a phrase pair

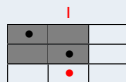
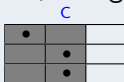
Let  $A$  be an alignment matrix

$(\bar{f}, \bar{e})$  consistent with  $A$  if, and only if:

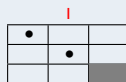
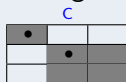
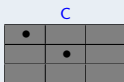
- Words in  $\bar{f}$ , if aligned, align only with words in  $\bar{e}$



- Words in  $\bar{e}$ , if aligned, align only with words in  $\bar{f}$



- $(\bar{f}, \bar{e})$  must contain at least one alignment point



# Feature Translation Model

Features

$$\log P(\bar{f}, \bar{e})$$

and

$$\log P(\bar{e}, \bar{f})$$

Number of times a (consistent) phrase pair is “observed”

$$c(\bar{f}, \bar{e})$$

Relative frequency counting

$$\varphi(\bar{f}|\bar{e}) = \frac{c(\bar{f}, \bar{e})}{\sum_{\bar{f}'} c(\bar{f}', \bar{e})}$$

# Feature Distortion

Feature

$$h_3 = \log d(\text{start}_k - \text{end}_{k-1} - 1)$$

Example

		I	have	black	eyes
1	J'	1			
2	ai				
3	les				3
4	yeux				
5	noirs			2	

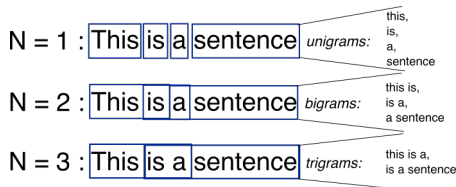
- $\bar{f}_1 = \text{J' ai}$
- $\bar{e}_1 = \text{I have}$
- $\text{start}_1 = 1$
- $\text{end}_1 = 2$
- $\bar{f}_2 = \text{noirs}$
- $\bar{e}_2 = \text{black}$
- $\text{start}_2 = 5$
- $\text{end}_2 = 5$
- $\bar{f}_3 = \text{les yeux}$
- $\bar{e}_3 = \text{eyes}$
- $\text{start}_3 = 3$
- $\text{end}_3 = 4$

# Feature Language Model

Feature n-gram language model

$$\log P(e|e_{\text{past}})$$

Estimated independently on monolingual data.



<http://recognize-speech.com/images/Antonio/Unigram.png>

# Decoding

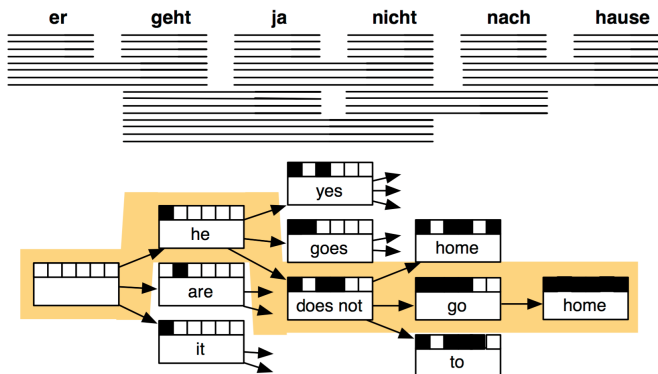


Figure: Koehn [2010]

# Translation Options

- Europarl phrase table: 2727 matching phrase pairs for a sentence.
- Search problem with beam search:
  - ① From phrase translation table for all input phrases.
  - ② Initial hypothesis: no input words covered, no output produced.
  - ③ Pick any translation option, create new hypothesis.
  - ④ Expand hypotheses from created partial hypothesis.
  - ⑤ Backtrack from highest scoring complete hypothesis.

Questions?

# References I

- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, June 1993. ISSN 0891-2017. URL <http://dl.acm.org/citation.cfm?id=972470.972474>.
- Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. ISBN 0521874157, 9780521874151.
- Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075117. URL <http://www.aclweb.org/anthology/P03-1021>.