Phrase-based SMT

Miguel Rios

Universiteit van Amsterdam

April 20, 2017

Content

Recap

We looked into Alignment a directional word-based model.

- Different parametrisations: Categorical vs Logistic.
- Estimation techniques: EM vs VB.

Recap

We looked into Alignment a directional word-based model.

- Different parametrisations: Categorical vs Logistic.
- 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]

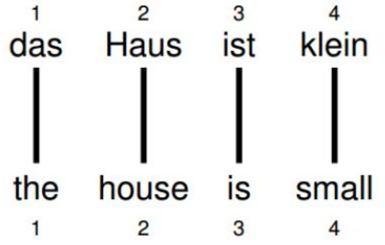


Figure: Koehn [2010]

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.

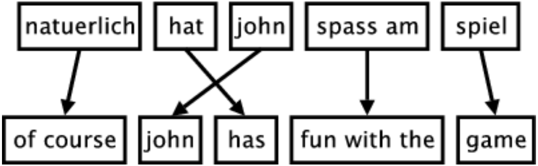


Figure: Koehn [2010]

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$_5$ input

[J'$_1$ ai$_2$] [les$_3$ yeux$_4$] [noirs$_5$] 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$_5$ — input

[J'$_1$ ai$_2$] [les$_3$ yeux$_4$] [noirs$_5$] — segmentation

[J'$_1$ ai$_2$]$_1$ [noirs$_5$]$_3$ [les$_3$ yeux$_4$]$_2$ — 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$_5$ — input

[J’$_1$ ai$_2$] [les$_3$ yeux$_4$] [noirs$_5$] — segmentation

[J’$_1$ ai$_2$]$_1$ [noirs$_5$]$_3$ [les$_3$ yeux$_4$]$_2$ — ordering

[I have]$_1$ [black]$_3$ [eyes]$_2$ — translation

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

[J'$_1$ ai$_2$] [les$_3$ yeux$_4$] [noirs$_5$] — segmentation

[J'$_1$ ai$_2$]$_1$ [noirs$_5$]$_3$ [les$_3$ yeux$_4$]$_2$ — ordering

[I have]$_1$ [black]$_3$ [eyes]$_2$ — 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 functionS_θ is defined as a linear model.

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

where θ are parameters
h are feature functions.

Linear model

The score functionS_θ is defined as a linear model.

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

where θ 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

$$
\begin{aligned}
P(F|E) &= \sum_{S} \sum_{A} P(S, A, F|E) \\
&= \sum_{S} \sum_{A} P(S|E) \times P(A|S, E) \times P(F|A, S, E)
\end{aligned}
$$

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

$$h_1 = \log P(\hat{f}, \hat{e})$$

- Language Model feature function:

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

- Distortion feature function:

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

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:

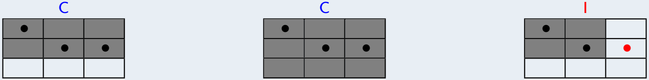
- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$

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 $\bar{e}$

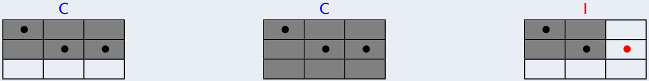


Phrase extraction

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 $\bar{f}$

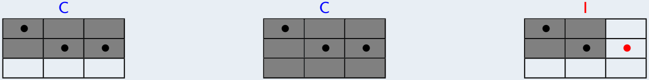
Phrase extraction

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

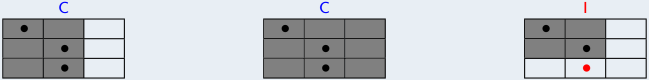
Let A be an alignment matrix

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

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



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



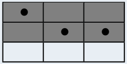
Phrase extraction

Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

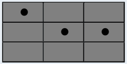
$(\bar{f}, \bar{e})$ consistent with A if, and only if:

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

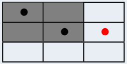
C



C

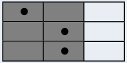


I

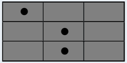


- Words in $\bar{e}$, if aligned, align only with words in $\bar{f}$

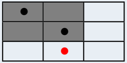
C



C



I



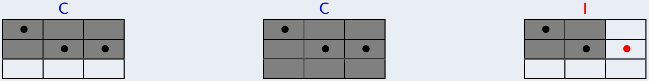
- $(\bar{f}, \bar{e})$ must contain at least one alignment point

Phrase extraction

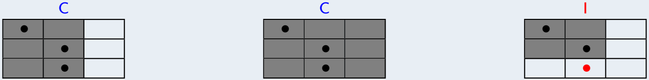
Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

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

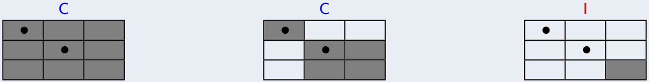
- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$



- 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

Feature

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

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

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

Example

<table>
<tr><th></th><th></th><th>I</th><th>have</th><th>black</th><th>eyes</th></tr>
<tr><td>1</td><td>J'</td><td colspan="2" rowspan="2">1</td><td></td><td></td></tr>
<tr><td>2</td><td>ai</td><td></td><td></td></tr>
<tr><td>3</td><td>les</td><td></td><td></td><td></td><td rowspan="2">3</td></tr>
<tr><td>4</td><td>yeux</td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>noirs</td><td></td><td></td><td>2</td><td></td></tr>
</table>

- $\bar{f}_1 =$ J' ai
- $\bar{e}_1 =$ I have
- $\mathsf{start}_1 = 1$
- $\mathsf{end}_1 = 2$

- $\bar{f}_2 =$ noirs
- $\bar{e}_2 =$ black
- $\mathsf{start}_2 = 5$
- $\mathsf{end}_2 = 5$

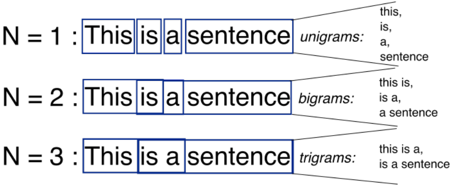
- $\bar{f}_3 =$ les yeux
- $\bar{e}_3 =$ eyes
- $\mathsf{start}_3 = 3$
- $\mathsf{end}_3 = 4$

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

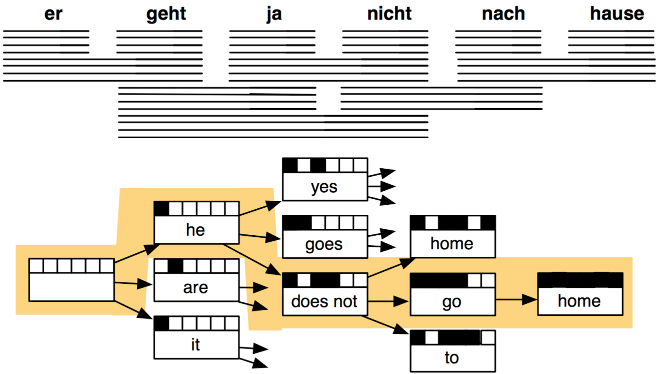


Figure: Koehn [2010]

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`.