CSE 517
Natural Language Processing
Winter 2017

Machine Translation

Yejin Choi

Slides from Dan Klein, Luke Zettlemoyer, Dan Jurafsky, Ray Mooney
When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’

- Warren Weaver (1955:18, quoting a letter he wrote in 1947)
Brief History of NLP

- Mid 1950’s – mid 1960’s: Birth of NLP and Linguistics
  - At first, people thought MT would be easy! Researchers predicted that “machine translation” can be solved in 3 years or so.

- Mid 1960’s – Mid 1970’s: A Dark Era
  - People started believing that machine translation is impossible.

- 1970’s and early 1980’s – Slow Revival of NLP
  - Small toy problems, linguistic heavy, weak empirical evaluation

- Late 1980’s and 1990’s – Statistical Revolution!
  - By this time, the computing power increased substantially.
  - Data-driven, statistical approaches with simple representation.

> “Whenever I fire a linguist, our MT performance improves.” (Jelinek, 1988)

- 2000’s – Statistics Powered by Linguistic Insights
  - More complex statistical models & richer linguistic representations.
Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.
Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:

- Yo lo haré mañana
  - I will do it tomorrow
- Hasta pronto
  - See you soon
- Hasta pronto
  - See you around

Machine translation system:

- Yo lo haré pronto
  - I will do it soon
  - I will do it around
  - See you tomorrow
Levels of Transfer

"Vauquois Triangle"

- Levels: interlingua, semantics, syntax, phrases, words
- Source: source language
- Target: target language

Example: English (E) to Spanish

| English (E)      | P(E | lo haré) |
|------------------|-------------|
| will do it       | 0.8         |
| will do so       | 0.2         |

| English (E)      | P(E | mañana) |
|------------------|------------|
| tomorrow         | 0.7        |
| morning          | 0.3        |
General Approaches

- **Rule-based approaches**
  - Expert system-like rewrite systems
  - Interlingua methods (analyze and generate)
  - Lexicons come from humans
  - Can be very fast, and can accumulate a lot of knowledge over time (e.g. Systran)

- **Statistical approaches**
  - Word-to-word translation
  - Phrase-based translation
  - Syntax-based translation (tree-to-tree, tree-to-string)
  - Trained on parallel corpora
  - Usually noisy-channel (at least in spirit)
Translation is hard!

zi    zhu     zhong   duan
自助 终 端

self help terminal device

help oneself terminating machine

(ATM, “self-service terminal”)

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
or even...

Examples from Liang Huang
Human Evaluation

Madame la présidente, votre présidence de cette institution a été marquante.
Mrs Fontaine, your presidency of this institution has been outstanding.
Madam President, president of this house has been discoveries.
Madam President, your presidency of this institution has been impressive.

Je vais maintenant m'exprimer brièvement en irlandais.
I shall now speak briefly in Irish.
I will now speak briefly in Ireland.
I will now speak briefly in Irish.

Nous trouvons en vous un président tel que nous le souhaitions.
We think that you are the type of president that we want.
We are in you a president as the wanted.
We are in you a president as we the wanted.

Evaluation Questions:
• Are translations fluent/grammatical?
• Are they adequate (you understand the meaning)?
MT: Automatic Evaluation

- **Human evaluations**: subject measures, fluency/adequacy

- **Automatic measures**: n-gram match to references
  - NIST measure: n-gram recall (worked poorly)
  - BLEU: n-gram precision (no one really likes it, but everyone uses it)

- **BLEU**:
  - P1 = unigram precision
  - P2, P3, P4 = bi-, tri-, 4-gram precision
  - Weighted geometric mean of P1-4
  - Brevity penalty (why?)
  - Somewhat hard to game…

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:
The American airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail, which sends out; The threat will be able after public place and so on the airport to start the biochemistry attack, [?] highly alerts after the maintenance.
Automatic Metrics Work (?)
MT System Components – Noisy Channel Model

\[ \text{argmax } P(e|f) = \text{argmax } P(f|e)P(e) \]

Diagram:
- **Language Model**: source \( P(e) \)
- **Translation Model**: channel \( P(f|e) \)
- **Decoder**
- **Best Output**: \( P(e|f) = P(f|e)P(e) \)
Part I – Word Alignment Models
What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?
Word Alignment

1. Align words with a probabilistic model
2. Infer presence of larger structures from this alignment
3. Translate with the larger structures

Example:

Yo lo haré mañana
I will do it tomorrow
Unsupervised Word Alignment

- Input: a *bitext*, pairs of translated sentences

- Output: *alignments*: pairs of translated words
  - When words have unique sources, can represent as a (forward) alignment function $a$ from French to English positions

```
ous acceptons votre opinion .
we accept your view .
```
We describe a series of five statistical models of the translation process and give algorithms for estimating the parameters of these models given a set of pairs of sentences that are translations of one another.  We define a concept of word-by-word alignment between such pairs of sentences.  For any given pair of such sentences each of our models assigns a probability to each of the possible word-by-word alignments.  We give an algorithm for seeking the most probable of these alignments.  Although the algorithm is suboptimal, the alignment thus obtained accounts well for the word-by-word relationships in the pair of sentences.  We have a great deal of data in French and English from the proceedings of the Canadian Parliament.  Accordingly, we have restricted our work to these two languages; but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages.  We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus.
IBM Model 1 (Brown 93)

- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer
- 3667 citations.
IBM Model 1 (Brown 93)

- Model parameters: \( t(f|e) := p('e' \text{ is translated into } 'f'|e) \)
- A (hidden) alignment vector \((a_1, ..., a_m)\) where \( a_i = j \) means \(i\)'th target word is translated from \(j\)'th source word.
- Include a “null” word on the source side
- This alignment vector defines 1-to-many mappings. (why?)

\[ p(f_1 \ldots f_m, a_1 \ldots a_m | e_1 \ldots e_l, m) = \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i}) \]

Uniform alignment model!!
IBM Model 1: Learning

- If given data with alignment \( \{(e_1 \ldots e_i, a_1 \ldots a_m, f_1 \ldots f_m)_{k=1 \ldots n}\} \)

\[
\hat{t}_{ML}(f | e) = \frac{c(e, f)}{c(e)} \quad \text{where} \quad \delta(k, i, j) = 1 \text{ if } a_i^{(k)} = j, \ 0 \text{ otherwise}
\]

\[
c(e, f) = \sum_k \sum_{i \text{ s.t. } e_i = e} \sum_{j \text{ s.t. } f_j = f} \delta(k, i, j)
\]

- In practice, no such data available at large scale.
- Thus, learn the translation model parameters while keeping alignment as latent variables, using EM,
  - Repeatedly re-compute the expected counts:

\[
\delta(k, i, j) = \frac{t(f_i^{(k)} | e_j^{(k)})}{\sum_{j'} t(f_i^{(k)} | e_{j'}^{(k)})}
\]

- Basic idea: compute expected source for each word, update co-occurrence statistics, repeat
Sample EM Trace for Alignment (IBM Model 1 with no NULL Generation)

<table>
<thead>
<tr>
<th>Training Corpus</th>
<th>green house</th>
<th>casa verde</th>
<th>the house</th>
<th>la casa</th>
</tr>
</thead>
<tbody>
<tr>
<td>verde</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td>casa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>la</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assume uniform initial probabilities

<table>
<thead>
<tr>
<th>Translation Probabilities</th>
<th>green house</th>
<th>casa verde</th>
<th>the house</th>
<th>la casa</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
</tr>
</tbody>
</table>

Compute Alignment Probabilities

\[
P(A, F | E) = 1/3 \times 1/3 = \frac{1}{9} \quad 1/3 \times 1/3 = \frac{1}{9} \quad 1/3 \times 1/3 = \frac{1}{9} \quad 1/3 \times 1/3 = \frac{1}{9}
\]

Normalize to get

\[
P(A | F, E) = \frac{1/9}{2/9} = \frac{1}{2} \quad \frac{1/9}{2/9} = \frac{1}{2} \quad \frac{1/9}{2/9} = \frac{1}{2} \quad \frac{1/9}{2/9} = \frac{1}{2}
\]
### Example cont.

<table>
<thead>
<tr>
<th>green house</th>
<th>verde</th>
<th>casa</th>
<th>la</th>
</tr>
</thead>
<tbody>
<tr>
<td>casa verde</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>the house</td>
<td>1/2</td>
<td>1/2 + 1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>the</td>
<td>0</td>
<td>1/2</td>
<td>1/2</td>
</tr>
</tbody>
</table>

**Compute weighted translation counts**

<table>
<thead>
<tr>
<th>green house</th>
<th>verde</th>
<th>casa</th>
<th>la</th>
</tr>
</thead>
<tbody>
<tr>
<td>casa verde</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>the house</td>
<td>1/4</td>
<td>1/2</td>
<td>1/4</td>
</tr>
<tr>
<td>the</td>
<td>0</td>
<td>1/2</td>
<td>1/2</td>
</tr>
</tbody>
</table>

**Normalize rows to sum to one to estimate P(f | e)**
### Example cont.

<table>
<thead>
<tr>
<th>Translation Probabilities</th>
<th>verde</th>
<th>casa</th>
<th>la</th>
</tr>
</thead>
<tbody>
<tr>
<td>green house</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>the</td>
<td>1/4</td>
<td>1/2</td>
<td>1/4</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1/2</td>
<td>1/2</td>
</tr>
</tbody>
</table>

Recompute Alignment Probabilities

\[
P(A, F | E) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8} \quad \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \quad \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \quad \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}
\]

Normalize to get

\[
P(A | F, E) = \frac{1}{8} = \frac{1}{3} \quad \frac{1}{4} = \frac{2}{3} \quad \frac{1}{4} = \frac{2}{3} \quad \frac{1}{8} = \frac{1}{3}
\]

Continue EM iterations until translation parameters converge
IBM Model 1 - EM intuition

Step 1

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

Step 2

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

Step 3

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

Step N

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

Example from Philipp Koehn
IBM Model 1: Inference

- Model parameters: $t(f|e) := p('e' \text{ is translated into } 'f'|e)$
- A (hidden) alignment vector $(a_1, \ldots, a_m)$ where $a_i = j$ means ‘$i$’th target word is translated from ‘$j$’th source word.

Uniform alignment model!

Inference: Find the best alignment a given $(f,e)$ pairs. Is this hard?
Evaluating Alignments

- How do we measure quality of a word-to-word model?
  - Method 1: use in an end-to-end translation system
    - Hard to measure translation quality
    - Option: human judges
    - Option: reference translations (NIST, BLEU)
    - Option: combinations (HTER)
    - Actually, no one uses word-to-word models alone as TMs
  - Method 2: measure quality of the alignments produced
    - Easy to measure
    - Hard to know what the gold alignments should be
    - Often does not correlate well with translation quality (like perplexity in LMs)
Alignment Error Rate

- **Alignment Error Rate**

  - = Sure
  - = Possible
  - = Predicted

- **A := predicted alignments**
- **S := sure alignments**
- **P := possible alignments**
  (including sure alignments)

\[
AER(A, S, P) = \left( 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \right)
= \left( 1 - \frac{3 + 3}{3 + 4} \right) = \frac{1}{7}
\]
Problems with Model 1

- There’s a reason they designed models 2-5!
- Problems: alignments jump around, align everything to rare words
- Experimental setup:
  - Training data: 1.1M sentences of French-English text, Canadian Hansards
  - Evaluation metric: alignment error Rate (AER)
  - Evaluation data: 447 hand-aligned sentences
Intersected Model 1

- **Post-intersection:** standard practice to train models in each direction then intersect their predictions [Och and Ney, 03]

- Second model is basically a filter on the first
  - Precision jumps, recall drops
  - End up not guessing hard alignments

<table>
<thead>
<tr>
<th>Model</th>
<th>P/R</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 E→F</td>
<td>82/58</td>
<td>30.6</td>
</tr>
<tr>
<td>Model 1 F→E</td>
<td>85/58</td>
<td>28.7</td>
</tr>
<tr>
<td>Model 1 AND</td>
<td>96/46</td>
<td>34.8</td>
</tr>
</tbody>
</table>
Joint Training?

- “Alignment by agreement” (Liang et al, 2006)
  - Similar high precision to post-intersection
  - But recall is much higher
  - More confident about positing non-null alignments

<table>
<thead>
<tr>
<th>Model</th>
<th>P/R</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 E→F</td>
<td>82/58</td>
<td>30.6</td>
</tr>
<tr>
<td>Model 1 F→E</td>
<td>85/58</td>
<td>28.7</td>
</tr>
<tr>
<td>Model 1 AND</td>
<td>96/46</td>
<td>34.8</td>
</tr>
<tr>
<td>Model 1 INT</td>
<td>93/69</td>
<td>19.5</td>
</tr>
</tbody>
</table>
Independent Training

### Figure 1: An example of the Viterbi output of a pair of independently trained HMMs (top) and a pair of jointly trained HMMs (bottom), both trained on 1.1 million sentences. Rounded boxes denote possible alignments, square boxes are sure alignments, and solid boxes are model predictions. For each model, the overall Precision/Recall/AER on the development set is given. See Section 4 for details.

In this example, COJO is a rare word that becomes a garbage collector (Moore, 2004) for the models in both directions. Intersection eliminates the spurious alignments, but at the expense of recall. Intersection after training produces alignments that both models agree on. The garbage-collecting rare word is no longer a problem. Not only are the individual E→F and F→E jointly-trained models better than their independently-trained counterparts, the jointly-trained intersected model also provides a significant overall gain over the independently-trained intersected model. We maintain both high precision and recall.

Before we introduce the objective function for joint training, we will write the two directional models in a symmetric way so that they share the same...
Joint Training

- Joint training
  - Independent training
  - we deemed it inadvisable to attend the meeting and so informed cojo.

**Figure 1:** An example of the Viterbi output of a pair of independently trained HMMs (top) and a pair of jointly trained HMMs (bottom), both trained on 1.1 million sentences. Rounded boxes denote possible alignments, square boxes are sure alignments, and solid boxes are model predictions. For each model, the overall Precision/Recall/AER on the development set is given. See Section 4 for details.

In this example, COJO is a rare word that becomes a garbage collector (Moore, 2004) for the models in both directions. Intersection eliminates the spurious alignments, but at the expense of recall. Intersection after training produces alignments that both models agree on. The garbage-collecting rare word is no longer a problem. Not only are the individual E→F and F→E jointly-trained models better than their independently-trained counterparts, the jointly-trained intersected model also provides a significant overall gain over the independently-trained intersected model. We maintain both high precision and recall.

Before we introduce the objective function for joint training, we will write the two directional models in a symmetric way so that they share the same...
Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes
Local Order Change

Japan is at the junction of four tectonic plates

Le Japon est au confluent de quatre plaques tectoniques
IBM Model 2 (Brown 93)

- Alignments: a hidden vector called an *alignment* specifies which English source is responsible for each French target word.

\[
p(f_1 \ldots f_m, a_1 \ldots a_m | e_1 \ldots e_l, m) = \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i})
\]

- Same decomposition as Model 1, but we will use a multi-nominal distribution for q!
IBM Model 2: Learning

- Given data \( \{(e_1,...,e_l,a_1,...,a_m,f_1,...,f_m)_{k=1..n}\} \) where
  \[
  t_{ML}(f|e) = \frac{c(e,f)}{c(e)} \quad q_{ML}(j|i,l,m) = \frac{c(j|i,l,m)}{c(i,l,m)} \quad \delta(k,i,j) = 1 \text{ if } a^{(k)}_i = j, \ 0 \text{ otherwise}
  \]

- Better approach: re-estimated generative models with EM,
  - Repeatedly compute counts, using redefined deltas:
    \[
    \delta(k,i,j) = \frac{q(j|i,l_k,m_k)t(f^{(k)}_i|e^{(k)}_j)}{\sum_{j'} q(j'|i,l_k,m_k)t(f^{(k)}_i|e^{(k)}_{j'})}
    \]

- Basic idea: compute expected source for each word, update co-occurrence statistics, repeat

- Q: What about inference? Is it hard?
Example

les embranchements
que ils songeaient à fermer
the branches they intend to close
On Tuesday Nov. 4, earthquakes rocked Japan once again.

Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.
The HMM Model

E: Thank you, I shall do so gladly.

A: 1 → 3 → 7 → 6 → 8 → 8 → 8 → 8 → 9

F: Gracias, lo haré de muy buen grado.

Model Parameters

Emissions: \( P( F_1 = \text{Gracias} | E_{A_1} = \text{Thank} ) \)

Transitions: \( P( A_2 = 3 | A_1 = 1 ) \)
The HMM Model

- Model 2 can learn complex alignments
- We want local monotonicity:
  - Most jumps are small
- HMM model (Vogel 96)

\[ P(f, a|e) = \prod_j P(a_j|a_{j-1})P(f_j|e_i) \]

\[ P(a_j - a_{j-1}) \]

- Re-estimate using the forward-backward algorithm
- Handling nulls requires some care

What are we still missing?

| f           | t(f | e) |
|-------------|--------|
| nationale   | 0.469  |
| national    | 0.418  |
| nationaux   | 0.054  |
| nationales  | 0.029  |
HMM Examples
AER for HMMs

<table>
<thead>
<tr>
<th>Model</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 INT</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM E→F</td>
<td>11.4</td>
</tr>
<tr>
<td>HMM F→E</td>
<td>10.8</td>
</tr>
<tr>
<td>HMM AND</td>
<td>7.1</td>
</tr>
<tr>
<td>HMM INT</td>
<td>4.7</td>
</tr>
<tr>
<td>GIZA M4 AND</td>
<td>6.9</td>
</tr>
</tbody>
</table>
IBM Models 3/4/5

Mary did not slap the green witch

Mary not slap slap slap the green witch

Mary not slap slap slap NULL the green witch

Mary no daba una botefada a la verde bruja

Mary no daba una botefada a la bruja verde

[from Al-Onaizan and Knight, 1998]
# Overview of Alignment Models

Table 1
Overview of the alignment models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Alignment model</th>
<th>Fertility model</th>
<th>E-step</th>
<th>Deficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>uniform</td>
<td>no</td>
<td>exact</td>
<td>no</td>
</tr>
<tr>
<td>Model 2</td>
<td>zero-order</td>
<td>no</td>
<td>exact</td>
<td>no</td>
</tr>
<tr>
<td>HMM</td>
<td>first-order</td>
<td>no</td>
<td>exact</td>
<td>no</td>
</tr>
<tr>
<td>Model 3</td>
<td>zero-order</td>
<td>yes</td>
<td>approximative</td>
<td>yes</td>
</tr>
<tr>
<td>Model 4</td>
<td>first-order</td>
<td>yes</td>
<td>approximative</td>
<td>yes</td>
</tr>
<tr>
<td>Model 5</td>
<td>first-order</td>
<td>yes</td>
<td>approximative</td>
<td>no</td>
</tr>
<tr>
<td>Model 6</td>
<td>first-order</td>
<td>yes</td>
<td>approximative</td>
<td>yes</td>
</tr>
</tbody>
</table>
Some Results

- [Och and Ney 03]

<table>
<thead>
<tr>
<th>Model</th>
<th>Training scheme</th>
<th>0.5K</th>
<th>8K</th>
<th>128K</th>
<th>1.47M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td></td>
<td>50.9</td>
<td>43.4</td>
<td>39.6</td>
<td>38.9</td>
</tr>
<tr>
<td>Dice+C</td>
<td></td>
<td>46.3</td>
<td>37.6</td>
<td>35.0</td>
<td>34.0</td>
</tr>
<tr>
<td>Model 1</td>
<td>$1^5$</td>
<td>40.6</td>
<td>33.6</td>
<td>28.6</td>
<td>25.9</td>
</tr>
<tr>
<td>Model 2</td>
<td>$1^52^5$</td>
<td>46.7</td>
<td>29.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM</td>
<td>$1^5H^5$</td>
<td>26.3</td>
<td>23.3</td>
<td>15.0</td>
<td>10.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>$1^52^53^3$</td>
<td>43.6</td>
<td>27.5</td>
<td>20.5</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^3$</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 4</td>
<td>$1^52^53^43^3$</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^43^3$</td>
<td>26.1</td>
<td>20.2</td>
<td>13.1</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>$1^5H^54^3$</td>
<td>26.3</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Model 5</td>
<td>$1^5H^43^5^3$</td>
<td>26.5</td>
<td>21.5</td>
<td>13.7</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^43^5^3$</td>
<td>26.5</td>
<td>20.4</td>
<td>13.4</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 6</td>
<td>$1^5H^4^36^3$</td>
<td>26.0</td>
<td>21.6</td>
<td>12.8</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^4^36^3$</td>
<td>25.9</td>
<td>20.3</td>
<td>12.5</td>
<td>8.7</td>
</tr>
</tbody>
</table>
Part II - Phrase Translation Model
Phrase-Based Systems

Morgen  fliege  ich  nach Kanada  zur Konferenz

Tomorrow  I will fly  to the conference  in Canada

Sentence-aligned corpus

Word alignments

Phrase table (translation model)
Phrase Translation Tables

- Defines the space of possible translations
  - each entry has an associated “probability”
- One learned example, for “den Vorschlag” from Europarl data

| English              | $\phi(\bar{e}|f)$ | English              | $\phi(\bar{e}|f)$ |
|----------------------|-------------------|----------------------|-------------------|
| the proposal         | 0.6227            | the suggestions      | 0.0114            |
| ’s proposal          | 0.1068            | the proposed         | 0.0114            |
| a proposal           | 0.0341            | the motion           | 0.0091            |
| the idea             | 0.0250            | the idea of          | 0.0091            |
| this proposal        | 0.0227            | the proposal ,       | 0.0068            |
| proposal             | 0.0205            | its proposal         | 0.0068            |
| of the proposal      | 0.0159            | it                   | 0.0068            |
| the proposals        | 0.0159            | ...                  | ...               |

- This table is noisy, has errors, and the entries do not necessarily match our linguistic intuitions about consistency.….
Extracting Phrases

- We will use word alignments to find phrases

- Question: what is the best set of phrases?
Extracting Phrases

- **Phrase alignment must**
  - Contain at least one alignment edge
  - Contain all alignments for phrase pair

- Extract all such phrase pairs!
Phrase Pair Extraction Example

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

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

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

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

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

Score by relative frequency:
Phrase pair scoring: assign probabilities to phrase translations
Phrase pair extraction: collect all phrase pairs from the data
Phrase Size

- Phrases do help
  - But they don’t need to be long
  - Why should this be?
Why not Learn Phrases w/ EM?

EM Training of the Phrase Model

• We presented a heuristic set-up to build phrase translation table (word alignment, phrase extraction, phrase scoring)

• Alternative: align phrase pairs directly with EM algorithm
  – initialization: uniform model, all $\phi(\bar{e}, \bar{f})$ are the same
  – expectation step:
    * estimate likelihood of all possible phrase alignments for all sentence pairs
  – maximization step:
    * collect counts for phrase pairs $(\bar{e}, \bar{f})$, weighted by alignment probability
    * update phrase translation probabilities $p(\bar{e}, \bar{f})$

• However: method easily overfits
  (learns very large phrase pairs, spanning entire sentences)
Phrase Scoring

\[ g(f, e) = \log \frac{c(e, f)}{c(e)} \]

\[ g(\text{les chats, cats}) = \log \frac{c(\text{cats, les chats})}{c(\text{cats})} \]

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

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

- Main issue: big chunks get all the weight, obvious priors don’t help
  - Though, [DeNero et al 08]
Part III - Decoding
### Phrase-Based Translation

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>the 7 people including by some and the russian the astronauts ,</td>
<td>7人中包括来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>it 7 people included by france and the russian international astronomical of rapporteur .</td>
<td>7个来自法国和俄罗斯的国际天文学会的报告员。</td>
</tr>
<tr>
<td>this 7 out including the from the french and the russian the fifth .</td>
<td>这个来自法国和俄罗斯的第五位。</td>
</tr>
<tr>
<td>these 7 among including from the french and of the russian of space members .</td>
<td>这些来自法国和俄罗斯的太空成员。</td>
</tr>
<tr>
<td>that 7 persons including from the of france and to russian of the aerospace members .</td>
<td>那些来自法国的和太空的成员。</td>
</tr>
<tr>
<td>7 include from the of france and russian astronauts . the numbers include from france and russian of astronauts who .</td>
<td>7个来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>7 populations include those from france and russian astronauts .</td>
<td>7个来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>7 deportees included came from france and russian in astronomical personnel ;</td>
<td>7个来自法国和俄罗斯的成员。</td>
</tr>
<tr>
<td>7 philtrum including those from france and russia a space member</td>
<td>涉及来自法国和俄罗斯的成员。</td>
</tr>
<tr>
<td>include came from france and russia by cosmonauts .</td>
<td>包括来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>include representatives from french and russia cosmonauts</td>
<td>包括法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>include came from france and russia ’s cosmonauts .</td>
<td>包括来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>includes coming from french and russian ’s cosmonaut</td>
<td>包括来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>french and russia ’s astronomavigation member .</td>
<td>法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>french and russia astronauts</td>
<td>法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>and russia ’s special rapporteur</td>
<td>法国和俄罗斯的特别报告员。</td>
</tr>
<tr>
<td>, and russia rapporteur</td>
<td>和俄罗斯的报告员。</td>
</tr>
<tr>
<td>, and russia rapporteur .</td>
<td>和俄罗斯的报告员。</td>
</tr>
<tr>
<td>or russia ’s</td>
<td>或俄罗斯的。</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Phrase-Based Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>这七人中包括来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>7 people including from France and the Russian astronaut.</td>
</tr>
</tbody>
</table>

<p>| Scoring: Try to use phrase pairs that have been frequently observed.  |
| Try to output a sentence with frequent English word sequences. |</p>
<table>
<thead>
<tr>
<th>Phrase-Based Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>这7人中包括来自法国和俄罗斯的宇航员。</td>
</tr>
<tr>
<td>the 7 people including by some and the russian the the astronauts ,</td>
</tr>
<tr>
<td>it 7 people included by france and the the russian international astronomical of rapporteur .</td>
</tr>
<tr>
<td>these 7 among including from the french and the russian the fifth .</td>
</tr>
<tr>
<td>these 7 persons including from the of the russian of space members .</td>
</tr>
<tr>
<td>that 7 include from the of france and the russian . the</td>
</tr>
<tr>
<td>7 numbers include from france and russian of astronauts who .</td>
</tr>
<tr>
<td>7 populations include those from france and russian of astronauts .</td>
</tr>
<tr>
<td>7 deportees included come from france and russian of astronauts who .</td>
</tr>
<tr>
<td>7 philtrum including those from france and russian of a space member</td>
</tr>
<tr>
<td>including representatives from france and the russian of astronaut</td>
</tr>
<tr>
<td>include came from france and russian by cosmonauts</td>
</tr>
<tr>
<td>include representatives from french and russian cosmonauts</td>
</tr>
<tr>
<td>include came from french and russian ′ s cosmonaut</td>
</tr>
<tr>
<td>includes coming from french and russian ′ s cosmonaut</td>
</tr>
<tr>
<td>french and russian ′ s astronavigation member .</td>
</tr>
<tr>
<td>french and russian astronauts</td>
</tr>
<tr>
<td>and russian ′ s special rapporteur</td>
</tr>
<tr>
<td>, and russian rapporteur</td>
</tr>
<tr>
<td>, and russian rapporteur .</td>
</tr>
<tr>
<td>or russian ′ s</td>
</tr>
</tbody>
</table>

Table 1: 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.
### Scoring

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

- Basic approach, sum up phrase translation scores and a language model
  - Define $y = p_1 p_2 \ldots p_L$ to be a translation with phrase pairs $p_i$
  - Define $e(y)$ be the output English sentence in $y$
  - Let $h()$ be the log probability under a tri-gram language model
  - Let $g()$ be a phrase pair score (from last slide)
  - Then, the full translation score is:
    \[
    f(y) = h(e(y)) + \sum_{k=1}^{L} g(p_k)
    \]

- Goal, compute the best translation
  \[
  \hat{y}(x) = \arg \max_{y \in \mathcal{Y}(x)} f(y)
  \]
The Pharaoh Decoder

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
</tr>
<tr>
<td>did not</td>
<td></td>
<td>a slap</td>
<td>by</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td></td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>slap</td>
<td>to</td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Scores at each step include LM and TM
The Pharaoh Decoder

- Space of possible translations
  - Phrase table constrains possible translations
  - Output sentence is built left to right
    - but source phrases can match any part of sentence
  - Each source word can only be translated once
  - Each source word must be translated

```
Morgen  fliege  ich  nach Kanada  zur Konferenz
```

```
Tomorrow  I  will fly  to the conference  in Canada
```
Scoring:

- In practice, much like for alignment models, also include a distortion penalty
  - Define \( y = p_1 p_2 \ldots p_L \) to be a translation with phrase pairs \( p_i \)
  - Let \( s(p_i) \) be the start position of the foreign phrase
  - Let \( t(p_i) \) be the end position of the foreign phrase
  - Define \( \eta \) to be the distortion score (usually negative!)
  - Then, we can define a score with distortion penalty:

\[
f(y) = h(e(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \eta \times |t(p_k) + 1 - s(p_{k+1})|
\]

- Goal, compute the best translation

\[
y^*(x) = \arg \max_{y \in \mathcal{Y}(x)} f(y)
\]
Hypothesis Expansion

- ... until all foreign words *covered*
  - find *best hypothesis* that covers all foreign words
  - *backtrack* to read off translation
Hypothesis Explosion!

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mary</td>
<td>did not give</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>a slap to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>the witch green</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>by</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>slap to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>the green witch</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>did not give</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ e: \text{slap} \]
\[ f: \text{******} \]
\[ p: .004283 \]
\[ e: \text{the green witch} \]
\[ f: \text{*******} \]
\[ p: .000271 \]

**Q:** How much time to find the best translation?
- Exponentially many translations, in length of source sentence
- NP-hard, just like for word translation models
- So, we will use approximate search techniques!
Hypothesis Lattices

Can recombine if:
• Last two English words match
• Foreign word coverage vectors match
Decoder Pseudocode

Initialization: Set beam $Q=\{q_0\}$ where $q_0$ is initial state with no words translated

For $i=0$ ... $n-1$ [where $n$ in input sentence length]
  • For each state $q \in \text{beam}(Q)$ and phrase $p \in \text{ph}(q)$
    1. $q'=\text{next}(q,p)$ [compute the new state]
    2. $\text{Add}(Q,q',q,p)$ [add the new state to the beam]

Notes:
  • $\text{ph}(q)$: set of phrases that can be added to partial translation in state $q$
  • $\text{next}(q,p)$: updates the translation in $q$ and records which words have been translated from input
  • $\text{Add}(Q,q',q,p)$: updates beam, $q'$ is added to $Q$ if it is in the top-$n$ overall highest scoring partial translations
Decoder Pseudocode

Initialization: Set beam $Q=\{q_0\}$ where $q_0$ is initial state with no words translated

For $i=0 \ldots n-1$ [where $n$ in input sentence length]

• For each state $q \in \text{beam}(Q)$ and phrase $p \in \text{ph}(q)$
  1. $q' = \text{next}(q,p)$ [compute the new state]
  2. $\text{Add}(Q,q',q,p)$ [add the new state to the beam]

Possible State Representations:

• Full: $q = (e, b, \alpha)$, e.g. (“Joe did not give,” 11000000, 0.092)
  • $e$ is the partial English sentence
  • $b$ is a bit vector recorded which source words are translated
  • $\alpha$ is score of translation so far
Decoder Pseudocode

Initialization: Set beam $Q=\{q_0\}$ where $q_0$ is initial state with no words translated

For $i=0 \ldots n-1$ [where $n$ in input sentence length]

• For each state $q \in$ beam($Q$) and phrase $p \in$ ph($q$)
  1. $q'=$next($q$,p) [compute the new state]
  2. Add($Q$,q',q,p) [add the new state to the beam]

Possible State Representations:
• Full: $q = (e, b, \alpha)$, e.g. (“Joe did not give,” 11000000, 0.092)
• Compact: $q = (e_1, e_2, b, r, \alpha)$,
  • e.g. (“not,” “give,” 11000000, 4, 0.092)
  • $e_1$ and $e_2$ are the last two words of partial translation
  • $r$ is the length of the partial translation
• Compact representation is more efficient, but requires back pointers to get the final translation