CSEP 517
Natural Language Processing
Autumn 2013

Machine Translation: Word Alignment

Luke Zettlemoyer

Most slides from Dan Klein
**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

Model of translation

I will do it around

See you tomorrow
Levels of Transfer

\[
P(\text{English} \mid \text{lo haré}) = \begin{array}{c|c}
\text{will do it} & 0.8 \\
\text{will do so} & 0.2 
\end{array}
\]

\[
P(\text{English} \mid \text{mañana}) = \begin{array}{c|c}
\text{tomorrow} & 0.7 \\
\text{morning} & 0.3 
\end{array}
\]
World-Level MT: Examples

- **la politique de la haine .** (Foreign Original)
  - politics of hate . (Reference Translation)
  - the policy of the hatred . (IBM4+N-grams+Stack)

- **nous avons signé le protocole .** (Foreign Original)
  - we did sign the memorandum of agreement . (Reference Translation)
  - we have signed the protocol . (IBM4+N-grams+Stack)

- **où était le plan solide ?** (Foreign Original)
  - but where was the solid plan ? (Reference Translation)
  - where was the economic base ? (IBM4+N-grams+Stack)
Le président américain Barack Obama doit annoncer lundi de nouvelles mesures en faveur des constructeurs automobile. General motors et Chrysler avaient déjà bénéficié fin 2008 d'un prêt d'urgence cumulé de 17,4 milliards de dollars, et ont soumis en février au Trésor un plan de restructuration basé sur un total de 22 milliards de dollars d'aides publiques supplémentaires.

Interrogé sur la chaine CBS dimanche, le président a toutefois clairement précisé que le gouvernement ne preterait pas d'argent sans de fortes contreparties. "Il faudra faire des sacrifices à tous les niveaux", a-t-il prévenu. "Tout le monde devra se réunir autour de la table et se mettre d'accord sur une restructuration en profondeur".

General Motors et Chrysler sont engagés dans des négociations avec le principal syndicat de l'automobile. Les constructeurs souhaitent diminuer leurs cotisations aux caisses de retraites, et accorder en échange des actions aux syndicats. Ils souhaiteraient également négocier des baisses des salaires.

U.S. President Barack Obama to announce Monday new measures to help automakers. General Motors and Chrysler had already received late in 2008 a cumulative emergency loan of 17.4 billion dollars, and submitted to the Treasury in February in a restructuring plan based on a total of 22 billion dollars in additional aid.

Interviewed on CBS Sunday, the president has clearly stated that the government does not lend money without strong counterparts. "We must make sacrifices at all levels," he warned. "Everyone should gather around the table and agree on a profound restructuring."

General Motors and Chrysler are engaged in negotiations with the major union of the car Manufacturers wishing to reduce their contributions to pension funds, and give in exchange for the shares to trade unions. They would also negotiate lower wages.
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)
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)?
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:
- $P_1 =$ unigram precision
- $P_2, P_3, P_4 =$ bi-, tri-, 4-gram precision
- Weighted geometric mean of $P_1$-$4$
- 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 office all receive 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 (?)
argmax \ P(e|f) = argmax \ P(f|e)P(e)
Today

- The components of a simple MT system
  - You already know about the LM
  - Word-alignment based TMs
    - IBM models 1 and 2, HMM model
  - A simple decoder

- Next few classes
  - More complex word-level and phrase-level TMs
  - Tree-to-tree and tree-to-string TMs
  - More sophisticated decoders
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
Unsupervised Word Alignment

- Input: a bitext, pairs of translated sentences
  - nous acceptons votre opinion.
  - we accept your view.

- Output: alignments: pairs of translated words
  - When words have unique sources, can represent as a (forward) alignment function a from French to English positions
And₁ the₂ program₃ has₄ been₅ implemented₆

Le₁ programme₂ a₃ été₄ mis₅ en₆ application₇
Many-to-Many Alignments

The\textsubscript{1} poor\textsubscript{2} don’t\textsubscript{3} have\textsubscript{4} any\textsubscript{5} money\textsubscript{6}

Les\textsubscript{1} pauvres\textsubscript{2} sont\textsubscript{3} demunis\textsubscript{4}
The IBM Translation Models

The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

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)

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

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

Uniform alignment model!
IBM Model 1: Learning

- Given data \( \{(e_1, a_1, \ldots, e_m, a_m, f_1, \ldots, f_m, k) | k = 1 \ldots n \} \)

\[
t_{ML}(f|e) = \frac{c(e, f)}{c(e)}
\]

where

\[
\delta(k, i, j) = 1 \text{ if } a_{i}^{(k)} = j, \ 0 \text{ otherwise}
\]

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

- Better approach: re-estimated generative models with EM,
  - Repeatedly compute counts, using redefined deltas:

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

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

- Q: What about inference? Is it hard?
IBM Model 1: Example

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
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**
  - $\square = \text{Sure}$
  - $\square = \text{Possible}$
  - $\blacksquare = \text{Predicted}$

$$AER(A, S, P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$

$$= \left(1 - \frac{3 + 3}{3 + 4}\right) = \frac{1}{7}$$
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?

- **Overall:**
  - 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>
Monotonic Translation

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_i,a_1...a_m,f_1...f_m)_k|k=1..n\} where
  \[ t_{ML}(f|e) = \frac{c(e, f)}{c(e)} \]
  \[ q_{ML}(j|i, l, m) = \frac{c(j|i, l, m)}{c(i, l, m)} \]
  \[ \delta(k, i, j) = 1 \text{ if } a_i^{(k)} = j, \text{ 0 otherwise} \]
  \[ c(e, f) = \sum_k \sum_{i \text{ s.t. } e_i=e} \sum_j \delta(k, i, j) \]

- **Better approach:** re-estimated generative models with EM,
  - Repeatedly compute counts, using redefined deltas:
    \[ \delta(k, i, j) = \frac{q(j|i, l, m_k)t(f_i^{(k)}|e_j^{(k)})}{\sum_{j'} q(j'|i, l, m_k)t(f_i^{(k)}|e_{j'}^{(k)})} \]

- **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.
Phrase Movement

Diagram showing phrase movement with words like "days", "both", "on", "eight", "at", "it", "make", "can", "we", "if", "think", "I", "well", "ja", "ich", "denke", "wenn", "wir", "das", "hinkriegen", "an", "beiden", "Tagen", "acht", "Uhr".
The HMM Model

E: Thank you, I shall do so gladly.

A: 

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)

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

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

| f               | t(f|e) |
|-----------------|-------|
| nationale       | 0.469 |
| national        | 0.418 |
| nationaux       | 0.054 |
| nationales      | 0.029 |

What are we still missing?
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

n(3|slap)

P(NULL)

t(la|the)

d(j|i)

Mary no daba una botefada a la bruja verde

[from Al-Onaizan and Knight, 1998]
Examples: Translation and Fertility

<table>
<thead>
<tr>
<th>f</th>
<th>( t(f \mid e) )</th>
<th>( \phi )</th>
<th>( n(\phi \mid e) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>0.497</td>
<td>1</td>
<td>0.746</td>
</tr>
<tr>
<td>la</td>
<td>0.207</td>
<td>0</td>
<td>0.254</td>
</tr>
<tr>
<td>les</td>
<td>0.155</td>
<td></td>
<td></td>
</tr>
<tr>
<td>l'</td>
<td>0.086</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ce</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cette</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>f</th>
<th>( t(f \mid e) )</th>
<th>( \phi )</th>
<th>( n(\phi \mid e) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ne</td>
<td>0.497</td>
<td>2</td>
<td>0.735</td>
</tr>
<tr>
<td>pas</td>
<td>0.442</td>
<td>0</td>
<td>0.154</td>
</tr>
<tr>
<td>non</td>
<td>0.029</td>
<td>1</td>
<td>0.107</td>
</tr>
<tr>
<td>rien</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>f</th>
<th>( t(f \mid e) )</th>
<th>( \phi )</th>
<th>( n(\phi \mid e) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>agriculteurs</td>
<td>0.442</td>
<td>2</td>
<td>0.731</td>
</tr>
<tr>
<td>les</td>
<td>0.418</td>
<td>1</td>
<td>0.228</td>
</tr>
<tr>
<td>cultivateurs</td>
<td>0.046</td>
<td>0</td>
<td>0.039</td>
</tr>
<tr>
<td>producteurs</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Example: Idioms

### Nodding

| Word  | $t(f | e)$ | $\phi$ | $n(\phi | e)$ |
|-------|-----------|--------|--------------|
| 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     |        |              |

He is nodding il hoche la tête
Example: Morphology

\textit{should}

<table>
<thead>
<tr>
<th>$f$</th>
<th>$t(f \mid e)$</th>
<th>$\phi$</th>
<th>$n(\phi \mid e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>devrait</td>
<td>0.330</td>
<td>1</td>
<td>0.649</td>
</tr>
<tr>
<td>devraient</td>
<td>0.123</td>
<td>0</td>
<td>0.336</td>
</tr>
<tr>
<td>devrions</td>
<td>0.109</td>
<td>2</td>
<td>0.014</td>
</tr>
<tr>
<td>faudrait</td>
<td>0.073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faut</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doit</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aurait</td>
<td>0.041</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doivent</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>devons</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>devrais</td>
<td>0.013</td>
<td></td>
<td></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>$1^52^53^34^3$</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^34^3$</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>$1^5H^54^35^3$</td>
<td>26.5</td>
<td>21.5</td>
<td>13.7</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^34^35^3$</td>
<td>26.5</td>
<td>20.4</td>
<td>13.4</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 6</td>
<td>$1^5H^54^36^3$</td>
<td>26.0</td>
<td>21.6</td>
<td>12.8</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^34^36^3$</td>
<td>25.9</td>
<td>20.3</td>
<td>12.5</td>
<td>8.7</td>
</tr>
</tbody>
</table>
Decoding

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

```
  it is not clear.
  CE  NE  EST  PAS  CLAIR.
```
Bag “Generation” (Decoding)

**Exact reconstruction**

Please give me your response as soon as possible.

⇒ Please give me your response as soon as possible.

**Reconstruction preserving meaning**

Now let me mention some of the disadvantages.

⇒ Let me mention some of the disadvantages now.

**Garbage reconstruction**

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 its 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 your favorite search technique)
Greedy Decoding

NULL well heard, it talks a great victory.

Null entendu, il parle de une belle victoire.

translateTwoWords(2, understood, 0, about)

NULL well understood, it talks about a great victory.

Null entendu, il parle de une belle victoire.

translateOneWord(4, he)

NULL well understood, he talks about a great victory.

Null entendu, il parle de une belle victoire.

translateTwoWords(1, quite, 2, naturally)

NULL quite naturally, he talks about a great victory.

Null entendu, il parle de une belle victoire.
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

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