Lecture 6:
Word Alignments

Slav Petrov – Google

With thanks to Chris Dyer

Machine Translation

- Translate text from one language to another
- Recombines fragments of example translations

Challenges:
- What fragments? [Learning to translate]
- How to make efficient? [Fast translation search]
- Fluency vs fidelity

Machine Translation (French)

WikiLeaks promises "new revelations" every week until the US election

It is an annual anniversary that celebrated, Tuesday, October 4, Julian Assange, founder of WikiLeaks. There decade, he recorded a domain name that would become famous: WikiLeaks org. One hundred and twenty months later, the organization ginned during a press conference in Berlin, a long list of publications (US diplomatic cables, emails Hillary Clinton, etc.) for welcoming for making public "ten million documents" that would "remain inaccessible to the general public" without its action.

The site’s support, including the United States, expected that Julian Assange to take the opportunity to publish further revelations about Hillary Clinton - the organization and its relatives have referred repeatedly to the fact that holds other documents incriminating the Democratic candidate after the publication this summer of the emails Democratic party, which showed that she had been favored by the party bodies in his fight against his rival for the nomination, Bernie Sanders.

Machine Translation (Japanese)

Mr. Trump, "I vividly with tax law" tax coverage of escape

Of the US presidential election Republican obligation Trump said of candidates (70) in a speech in Colorado three days, is that I vividly came with the tax law, "me to the tax legally possible as long as the small amount He said that there is ", admitted that it has been" tax saving. "US newspaper obtained a portion of the tax records of Mr. Trump. New York Times. 13, to report huge losses in the 1980s, in response to the fact that reported that as a result there is a possibility to have escaped the tax over 18 years seen the remarks."
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
- No novel sentence

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)

Levels of Transfer

MT: 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 themselves the South Arabian Osama bin Laden and threatening a biological chemical attack against public places such as the airport.

Machine translation:
The American (7) international airport is closed after the office received one e-mails left on the sand Arab not business (7) writer on electronic mail which sends out the threats to stop the public place and so on the airport. To start the biological attack, (7) highly alert after the maintenance.
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
- Future classes
  - More complex word-level and phrase-level TMs
  - Tree-to-tree and tree-to-string TMs
  - More sophisticated decoders

Word Alignment

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

A sci-fi example (Knight, 1997)

Your assignment: translate this Centauri sentence into Arcturan

**FAROK CRRROK HIHOK YOROK CLOK KANTOK OK-YURP**

<table>
<thead>
<tr>
<th>Farok</th>
<th>Crrrok</th>
<th>Hihok</th>
<th>Yorok</th>
<th>Clok</th>
<th>Kantok</th>
<th>Ok-yurp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1c. OK-VOON OK-VOON SPROK</td>
<td>7c. LALOK OK-VOON OK-VOON SPROK IZOK ENEMOK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a. AT-VOON BICHAT DAT</td>
<td>7a. WAT JJAT BICHAT WAT DAT VAT ENENAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2c. OK-DROBIL OK-VOON ANOK PLOK SPROK</td>
<td>8c. LALOK BROK ANOK PLOK NOK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2a. AT-DROBIL AT-VOON PIPPAT KRAAT DAT</td>
<td>8a. IAT LAT PIPPAT KRAAT NNAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c. EROK SPROK IZOK HIHOK CHIROK</td>
<td>9c. WIWOK NOK IZOK KANTOK OK-YURP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3a. TOTAT DAT ARRAT VAT HILAT</td>
<td>9a. TOTAT NNAT QUAT OLOT AT-YURP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4c. OK-VOON ANOK DROK BROK JOK</td>
<td>10c. LALOK MOK NOK YOROK CHIROK CLOK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4a. AT-VOON KRAAT PIPPAT IAT LAT</td>
<td>10a. WAT NNAT CAT MAT BAT HILAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5c. WIWOK FAROK IZOK JIK</td>
<td>11c. LALOK NOK ERABOK HIHOK YOROK ZANZANOK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5a. TOTAT JJAT QUAT CAT</td>
<td>11a. WAT NNAT ARRAAT MAT ZANZANAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6c. LALOK SPROK IZOK JOK JOK</td>
<td>12c. LALOK KAROK NOK IZOK HIHOK MOK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6a. WAT DAT KRAAT QUAT CAT</td>
<td>12a. WAT NNAT FORAT ARRAAT VAT CAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### A sci-fi example (Knight, 1997)

Your assignment: translate this Centauri sentence into Arcturan

Farok crrrok hirok yorok clok kantok ok-yurp

Jjat arrat mat bat oloat at-yurp

Are these Arcturan words in Arcturan order?

### Clients do not sell pharmaceuticals in Europe

<table>
<thead>
<tr>
<th>1e. Garcia and associates.</th>
<th>7e. the clients and the associates are enemies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1s. Garcia y asociados.</td>
<td>7s. los clientes y los asociados son enemigos.</td>
</tr>
<tr>
<td>2e. Carlos Garcia has three associates.</td>
<td>8e. the company has three groups.</td>
</tr>
<tr>
<td>2s. Carlos Garcia tiene tres asociados.</td>
<td>8s. la empresa tiene tres grupos.</td>
</tr>
<tr>
<td>3e. his associates are not strong.</td>
<td>9e. its groups are in Europe.</td>
</tr>
<tr>
<td>3s. sus asociados no son fuertes.</td>
<td>9s. sus grupos estan en Europa.</td>
</tr>
<tr>
<td>4e. Garcia has a company also.</td>
<td>10e. the modern groups sell strong pharmaceuticals.</td>
</tr>
<tr>
<td>4s. Garcia tambien tiene una empresa.</td>
<td>10s. los grupos modernos venden medicinas fuertes.</td>
</tr>
<tr>
<td>5e. its clients are angry.</td>
<td>11e. the groups do not sell zenzanine.</td>
</tr>
<tr>
<td>5s. sus clientes estan enfadados.</td>
<td>11s. los grupos no venden zanzanina.</td>
</tr>
<tr>
<td>6e. the associates are also angry.</td>
<td>12e. the small groups are not modern.</td>
</tr>
<tr>
<td>6s. los asociados tambien estan enfadados.</td>
<td>12s. los grupos pequenos no son modernos.</td>
</tr>
</tbody>
</table>

### Monotonic Translation

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>das</td>
<td>Haus</td>
<td>ist</td>
<td>klein</td>
</tr>
</tbody>
</table>

the house is small

### Word Deletion

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>das</td>
<td>Haus</td>
<td>ist</td>
<td>klein</td>
</tr>
</tbody>
</table>

house is small
Word Insertion

1-to-Many Alignments

Many-to-1 Alignments

Many-to-Many Alignments
Evaluating TMs

- 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

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

  \[ = \left( 1 - \frac{3 + 3}{3 + 4} \right) = \frac{1}{7} \]

IBM Model 1 (Brown 93)

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

\[
P(f, a | e) = \prod_j P(a_j = i) P(f_j | e_i) = \prod_j \frac{1}{I + 1} P(f_j | e_i)
\]

\[
P(f | e) = \sum_a P(f, a | e)
\]

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
IBM Model 2

- Alignments tend to the diagonal (broadly at least)

\[ P(f, a | e) = \prod_j P(a_j = i|j, I, J) P(f_j | e_i) \]
\[ P(\text{dist} = i - j) = \frac{1}{Z} e^{-\alpha (i - j)^2} \]

- Other schemes for biasing alignments towards the diagonal:
  - Relative vs absolute alignment
  - Asymmetric distances
  - Learning a full multinomial over distances

EM for Model 1

- Start with uniform translation probabilities

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

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

- After a few iterations the alignments start becoming apparent

EM for Models 1/2

- Model 1 Parameters:
  - Translation probabilities (1+2)
  - Distortion parameters (2 only)
- Start with \( P(f_j | e_i) \) uniform, including \( P(f_j | \text{null}) \)
- For each sentence:
  - For each French position \( j \)
    - Calculate posterior over English positions

\[ P(a_j = i | f, e) = \frac{P(a_j = i | j, I, J) P(f_j | e_i)}{\sum_i P(a_j = i | j, I, J) P(f_j | e_i)} \]

- (or just use best single alignment)
- Increment count of word \( f \) with word \( e \) by these amounts
- Also re-estimate distortion probabilities for model 2
- Iterate until convergence

IBM Model 1 Pseudocode

initialize \( t(e | f) \) uniformly

do until convergence
  set count(e | f) to 0 for all e, f
  set total(f) to 0 for all f
  for all sentence pairs (e_s, f_s)
    set total_s(e) = 0 for all e
    for all words e in e_s
      for all words f in f_s
        total_s(e) += t(e | f)
      end
    end
    for all words e in e_s
      for all words f in f_s
        count(e | f) += t(e | f) / total_s(e)
        total(f) += t(e | f) / total_s(e)
      end
    end
  end
  for all f
    for all e
      t(e | f) = count(e | f) / total(f)
  end
end

The final implementation should include the NULL word as position 0 of \( f_s \).
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>
<tr>
<td>Model 1 INT</td>
<td>93/69</td>
<td>19.5</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>

Example

Phrase Movement

On Tuesday Nov. 4, earthquakes rocked Japan once again

Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.
Phrase Movement

**IBM Models 1/2**

E:

Thank you, I shall do so gladly.

A:

Gracias, lo haré de muy buen grado.

Model Parameters

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

Transitions: $P(A_2 = 3)$

---

The HMM Model

E:

Thank you, I shall do so gladly.

A:

Gracias, lo haré de muy buen grado.

F:

Model Parameters

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

Transitions: $P(A_2 = 3)$

---

The HMM Model

- Model 2 preferred global monotonicity
- We want local monotonicity:
  - Most jumps are small
- HMM model (Vogel 96)
  
  \[ P(f, a \mid e) = \prod_j P(a_j \mid a_{j-1})P(f_j \mid e_i) \]

- Re-estimate using the forward-backward algorithm
- Handling nulls requires some care
- What are we still missing?
Examples

IBM Models 3/4/5

AER for HMMs

Examples: Translation and Fertility

Examples

41

42

43

44

[from Al-Onaizan and Knight, 1998]
Example: Idioms

| f | t(f | e) | φ | n(φ | e) |
|---|-------|---|------|
| signer | 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

| f | t(f | e) | φ | n(φ | e) |
|---|-------|---|------|
| devrait | 0.330 | 1 | 0.649 |
| devraient | 0.123 | 0 | 0.336 |
| devrons | 0.109 | 2 | 0.014 |
| faudrait | 0.073 |
| faut | 0.058 |
| doit | 0.058 |
| aurait | 0.041 |
| doivent | 0.024 |
| devons | 0.017 |
| devrais | 0.013 |

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+C</td>
<td>46.3</td>
<td>37.6</td>
<td>35.0</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>$1^1$</td>
<td>40.6</td>
<td>33.6</td>
<td>28.6</td>
<td>25.9</td>
</tr>
<tr>
<td>Model 2</td>
<td>$1^2$</td>
<td>46.7</td>
<td>29.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM</td>
<td>$1^3$</td>
<td>26.3</td>
<td>23.3</td>
<td>15.0</td>
<td>10.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>$1^2$</td>
<td>43.6</td>
<td>27.5</td>
<td>20.5</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>$3^2$</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 4</td>
<td>$1^2$</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>$3^2$</td>
<td>26.1</td>
<td>20.2</td>
<td>13.1</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>$4^3$</td>
<td>26.3</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Model 5</td>
<td>$1^2$</td>
<td>26.5</td>
<td>21.5</td>
<td>13.7</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>$3^3$</td>
<td>26.5</td>
<td>20.4</td>
<td>13.4</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 6</td>
<td>$1^3$</td>
<td>26.0</td>
<td>21.6</td>
<td>12.8</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>$2^3$</td>
<td>25.9</td>
<td>20.3</td>
<td>12.5</td>
<td>8.7</td>
</tr>
</tbody>
</table>