Lecture 7:
EM (cont.) + Word Alignment

Ankur Parikh – Google

With thanks to Slav Petrov, Chris Dyer, and Michael Collins
Recap:

\[ P(s, w) = \prod_{i} P(s_i | s_{i-1}) P(w_i | s_i) \]

transition matrix

emission matrix
Unsupervised learning

Maximize marginal likelihood

\[ \ell(\theta) = \sum_{m=1}^{M} \log \sum_{s} P(w^{(m)}, s | \theta) \]

non-concave, has local optima
Expectation Maximization

Maximize lower bound on log-likelihood

\[ \ell(\theta) = \sum_{m=1}^{M} \log \sum_{s} P(w^{(m)}, s|\theta) \]

\[ = \sum_{m=1}^{M} \log \sum_{s} Q^{(m)}(s) \frac{P(w^{(m)}, s|\theta)}{Q^{(m)}(s)} \]

\[ \geq \sum_{m=1}^{M} \sum_{s} Q^{(m)}(s) \log \frac{P(w^{(m)}, s|\theta)}{Q^{(m)}(s)} \]
Expectation Maximization

E-step: Set Q to posterior distribution

\[ Q^{(m)}(s) = P(s|w^{(m)}) \]

M-step: Maximize weighted likelihood

\[ \arg\max_{\theta} \sum_{m=1}^{M} \sum_{s} Q^{(m),t}(s) \log P(w^{(m)}, s|\theta) \]
Expectation Maximization

\[
\text{argmax}_{\theta} \sum_{m=1}^{M} \sum_{s} Q^{(m),t}(s) \log P(w^{(m)}, s|\theta)
\]

we can decompose this efficiently using the structure of the HMM
Expectation Maximization

\[
\sum_{m=1}^{M} \sum_{s} Q^{(m), t}(s) \log P(w^{(m)}, s | \theta) \\
= \sum_{m=1}^{M} \sum_{s} Q^{(m), t}(s) \log \prod_{i} (P(s_i | s_{i-1}, \theta) P(w_i | s_i, \theta)) \\
= \sum_{m=1}^{M} \sum_{s} Q^{(m), t}(s) \sum_{i} \log P(s_i | s_{i-1}, \theta) + \sum_{m=1}^{M} \sum_{s} Q^{(m), t}(s) \sum_{i} \log P(w_i | s_i, \theta) \\
= \sum_{m=1}^{M} \sum_{s_i, s_{i-1}} Q^{(m), t}(s_i, s_{i-1}) \sum_{i} \log P(s_i | s_{i-1}, \theta) + \sum_{m=1}^{M} \sum_{s_i} Q^{(m), t}(s_i) \sum_{i} \log P(w_i | s_i, \theta)
\]

\[Q^{(m), t}(s_i, s_{i-1}) = P(s_i, s_{i-1} | w^{(m)}, \theta^{(t)})\]
Let us just take one term:

\[
\sum_{m=1}^{M} \sum_{s_i, s_{i-1}} Q^{(m), t}(s_i, s_{i-1}) \log P(s_i | s_{i-1}, \theta)
\]

\[
= \sum_{m=1}^{M} \sum_{s_i, s_{i-1}} Q^{(m), t}(s_i, s_{i-1}) \log \theta_{s_i, s_{i-1}}
\]

\[
\theta^{(t+1)}_{s_i, s_{i-1}} = \arg\max_{\theta} \sum_{m=1}^{M} \sum_{s_i, s_{i-1}} Q^{(m), t}(s_i, s_{i-1}) \log \theta_{s_i, s_{i-1}}
\]

\[
\theta^{(t+1)}_{s_i, s_{i-1}} = \frac{\sum_{m=1}^{M} Q^{(m), t}(s_i, s_{i-1})}{\sum_{m=1}^{M} \sum_{s_i} Q^{(m), t}(s_i, s_{i-1})}
\]
Comparison

Unsupervised (Hidden states are latent)

\[
\theta^{(t+1)}_{s_i, s_{i-1}} = \frac{\sum_{m=1}^{M} Q(m,t(s_i, s_{i-1}))}{\sum_{m=1}^{M} \sum_{s_i} Q(m,t(s_i, s_{i-1}))}
\]

Supervised

\[
\theta^{*}_{j,k} = \frac{\sum_{m=1}^{M} \mathbb{I}[s^{(m)} = j, s_{i-1}^{(m)} = k]}{\sum_{m=1}^{M} \sum_{j} \mathbb{I}[s^{(m)} = j, s_{i-1}^{(m)} = k]}
\]

EM essentially replaces hard counts with soft counts
Efficiently Computing Q

Efficiently computing $Q^{(m),t}(s_i, s_{i-1}) = P(s_i, s_{i-1}|w^{(m)}, \theta^{(t)})$

Can be computed with dynamic program.

But we have to do this $O(n)$ times (one for each transition/emission)

Can we do better?
Yes!! Compute everything with one dynamic program! Called “forward-backward” or message-passing.

Essentially run the dynamic program twice:
(1) once from $s_1$ to $s_n$
(2) once from $s_n$ to $s_1$

Get all you need by combining the results from both.
Let us say we want to compute the following probability:

\[ P(s_2, s_3|w_1, w_2, w_3, w_4) \propto P(s_2, s_3, w_1, w_2, w_3, w_4) \]
Forward Backward - Example

\[ P(s_2, s_3, w_1, w_2, w_3, w_4) = P(s_3, s_2, w_1, w_2) P(w_3, w_4 | s_3) \]

compute forward pass

compute with backward pass
Forward Pass

\[ \delta(s_2) = \sum_{s_1} P(s_2|s_1)P(w_1|s_1) \psi(s_1, s_2) \]

\[ \delta(s_3) = \sum_{s_2} P(s_3|s_2)P(w_2|s_2)\delta(s_2) \psi(s_2, s_3) \]
Summing over "soft" counts instead of hard counts.

\[ \beta(s_4) = P(w_4|s_4) \]

\[ \beta(s_3) = \sum_{s_4} P(s_4|s_3)P(w_4|s_4) \]

\[ \beta(s_2) = \sum_{s_3} P(s_3|s_2)P(w_3|s_2) \]
Putting it together

\[ P(s_2, s_3, w_1, w_2, w_3, w_4) = P(s_3, s_2, w_1, w_2)P(w_3, w_4|s_3) = \psi(s_2, s_3)\beta(s_3) \]

Similarly,

\[ P(s_1, s_2, w_1, w_2, w_3, w_4) = \psi(s_1, s_2)\beta(s_2) \]

\[ P(s_3, s_4, w_1, w_2, w_3, w_4) = \psi(s_3, s_4)\beta(s_4) \]
Run the dynamic program twice (storing intermediate quantities along the way:

(1) once from $s_1$ to $s_n$
(2) once from $s_n$ to $s_1$

Get all you need by combining the results from both.
Word Alignments

- Primary motivation: Machine Translation

- But the general concept is one of the most fundamental ideas in NLP.
  - paraphrasing
  - entailment
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
WikiLeaks promises "new revelations" every week until the US election

It is an informal 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.) 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.
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 is, to report huge losses in the 1990s, for 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
-Mostly sentence
-Model of translation
-I will do it around
-See you tomorrow
Levels of Transfer

The diagram illustrates the concept of Levels of Transfer in translation, showing the relationship between source and target languages at different levels: words, phrases, syntax, semantics, and interlingua. The example sentence "Yo lo haré mañana" (I will do it tomorrow) is analyzed at the syntax level, with its English translation "I will do it tomorrow" shown in the target language. The probability of this translation is 0.8.

| 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)
  - Neural Machine Translation
MT System Components

Language Model

source
P(e)

best

observed

Translation Model

channel
P(f|e)

decoder

e

argmax P(e|f) = argmax P(f|e)P(e)

e

today

Lecture 1
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

Yo, lo hare mañana
I will do it tomorrow
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 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 Crrrok Hihok Yorok Clok Kantok Ok-Yurp</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1C. Ok-Voon Ororok Sprok.</strong></td>
</tr>
<tr>
<td><strong>1A. At-Voon Bichat Dat.</strong></td>
</tr>
<tr>
<td><strong>2C. Ok-Drubel Ok-Voon Anok Plok Sprok.</strong></td>
</tr>
<tr>
<td><strong>2A. At-Drubel At-Voon Pippat Rrat Dat.</strong></td>
</tr>
<tr>
<td><strong>3C. Erok Sprok Izok Hihok Ghirok.</strong></td>
</tr>
<tr>
<td><strong>3A. Totat Dat Arrat Wat Hilat.</strong></td>
</tr>
<tr>
<td><strong>4C. Ok-Voon Anok Drok Brok Jok.</strong></td>
</tr>
<tr>
<td><strong>4A. At-Voon Krat Pippat Sat Lat.</strong></td>
</tr>
<tr>
<td><strong>5C. Wiwok Farok Izok Stok.</strong></td>
</tr>
<tr>
<td><strong>5A. Totat Jjat Quat Cat.</strong></td>
</tr>
<tr>
<td><strong>6C. Lalok Sprok Izok Jok Stok.</strong></td>
</tr>
<tr>
<td><strong>6A. Wat Dat Krat Quat Cat.</strong></td>
</tr>
<tr>
<td>Column 1</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>1C. OK-VOON OROROK SPROK .</td>
</tr>
<tr>
<td>1A. AT-VOON BICHAT DAT .</td>
</tr>
<tr>
<td>2C. OK-DRUDEL OK-VOON ANOK PLOK SPROK .</td>
</tr>
<tr>
<td>2A. AT-DRUDEL AT-VOON PIPPAT RRAT DAT .</td>
</tr>
<tr>
<td>3C. EROK SPROK IZOK HIHOK GHIROK .</td>
</tr>
<tr>
<td>3A. TOTAT DAT ARRAKT VAT HILAT .</td>
</tr>
<tr>
<td>4C. OK-VOON ANOK DROK BROK JOK .</td>
</tr>
<tr>
<td>4A. AT-VOON KRAT PIPPAT SAT LAT .</td>
</tr>
<tr>
<td>5C. WIWOK FAROK IZOK STOK .</td>
</tr>
<tr>
<td>5A. TOTAT JJAT QUAT CAT .</td>
</tr>
<tr>
<td>6C. LALOK SPROK IZOK JOK STOK .</td>
</tr>
<tr>
<td>6A. WAT DAT KRAT QUAT CAT .</td>
</tr>
</tbody>
</table>
Your assignment: translate this Centauri sentence into Arcturan

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

das Haus ist klein

the house is small
Word Deletion

das Haus ist klein

house is small
Word Insertion

0 1 2 3 4
NULL das Haus ist klein

the house is just small

1 2 3 4 5
1-to-Many Alignments

das | Haus | ist | klitzeklein

the | house | is | very | small
Many-to-1 Alignments

1. das → the
2. Haus → house
3. brach → collapsed
4. zusammen
Many-to-Many Alignments

The_1 poor_2 don’t_3 have_4 any_5 money_6

Les_1 pauvres_2 sont_3 demunis_4
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 translation models
  - 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

- Hypothesized Alignment (A)
- Sure (S)
- Possible (P)

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

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

\[ \text{Precision}(A, P) = \frac{|P \cap A|}{|A|} \]

\[ \text{Recall}(A, S) = \frac{|S \cap A|}{|S|} \]

\[ AER(A, S, P) = \left( 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \right) \]
IBM Models [Brown 93] (following Mike Collin’s notes)

English sentence:
\[ e = e_1 e_2 \ldots e_l \]
\[ e_0 := \text{NULL} \]

French sentence:
\[ f = f_1 f_2 \ldots f_m \]

Introduce (latent) alignment variables
\[ a = a_1 a_2 \ldots a_m \]
\[ a_i \in \{0, 1, \ldots, l\} \]  
    english word aligned to french word i
IBM Models (Brown 93)

Note that each French word is aligned to exactly one English word. The alignment is many-to-one: more than one French word can be aligned to a single English word (e.g., "mis", "en", and "application" are all aligned to "implemented"). Some English words may be aligned to zero French words: for example, the word "And" is not aligned to any French word in this example.

Note also that the model is asymmetric, in that there is no constraint that each English word is aligned to exactly one French word: each English word can be aligned to any number (zero or more) French words. We will return to this point later.

As another example alignment, we could have:

\[ a_1 = 2, a_2 = 3, a_3 = 4, a_4 = 5, a_5 = 6, a_6 = 6, a_7 = 6 \]

specifying the following alignment:

- Le \( \Rightarrow \) the
- Programme \( \Rightarrow \) program
- a \( \Rightarrow \) has
- ete \( \Rightarrow \) been
- mis \( \Rightarrow \) implemented
- en \( \Rightarrow \) implemented
- application \( \Rightarrow \) implemented
IBM Models

Want to model the following conditional probability:

$$P(f_1, \ldots, f_m, a_1, \ldots, a_m | e_1, \ldots, e_l, m)$$

But need to make some independence assumptions!
IBM Model 2

We describe Model 2 first, since Model 1 follows as a special case.

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

conditional probability of alignment

conditional probability of generating french word given its aligned english word
What assumptions did we make?

\[ P(f_1, \ldots, f_m, a_1, \ldots, a_m | e_1, \ldots, e_l, m) \]

\[ = \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i}) \]

(1) Each target word only depends on its aligned source word

(2) Each alignment variable only depends on the relevant positions and lengths of the sentences
IBM Model 1

- IBM Model 1 is a special case of Model 2

\[ P(f_1, \ldots, f_m, a_1, \ldots, a_m | e_1, \ldots, e_l, m) = \prod_{i=1}^{m} \frac{1}{\ell + 1} t(f_i | e_{a_i}) \]

uniform probability
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 Variations

- Generic form (full multinomial) has too many free parameters
  \[
  P(f_1, \ldots, f_m, a_1, \ldots, a_m | e_1, \ldots, e_l, m) = \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i})
  \]

- Restrict parameterization to favor diagonal alignments [Dyer et al. 2013]
IBM Model 2 Variations [Dyer et al. 2013]

- Restrict parameterization to favor diagonal alignments

\[
h(j, i, l, m) = -\left| \frac{i}{m} - \frac{j}{l} \right|
\]

\[
q(a_i = j | i, l, m) = \begin{cases} 
p_0 & j = 0 \text{ (NULL)} \\
(1 - p_0) \times \frac{\exp(\lambda h(j, i, l, m))}{Z_\lambda(i, l, m)} & 0 < j \leq l
\end{cases}
\]
Estimating Model 2

- First, assume the alignment variables are provided in the training data.

- Define (hard alignments):

\[ \delta^{(k)}(i, j) = \mathbb{I}[a_i^{(k)} = j] \]

- Estimate parameters from counting (assume K = # data points):

\[
t(f|e) = \frac{\sum_{k=1}^{K} \sum_{i,j} \delta^{(k)}(i, j) \mathbb{I}[f_i^{(k)} = f, e_j^{(k)} = e]}{\sum_{f'} \sum_{k=1}^{K} \sum_{i,j} \delta^{(k)}(i, j) \mathbb{I}[f_i^{(k)} = f', e_j^{(k)} = e]}
\]

\[
q(j|i, l, m) = \frac{\sum_{k=1}^{K} \delta^{(k)}(i, j) \mathbb{I}[m^{(k)} = m, l^{(k)} = l]}{\sum_{j'} \sum_{k=1}^{K} \delta^{(k)}(i, j') \mathbb{I}[m^{(k)} = m, l^{(k)} = l]}
\]
Estimating Model 2

- In the unsupervised case, we use Expectation Maximization. Alternate between following steps:

- E-step (compute soft alignments):
  \[
  \delta^{(k)}(i, j) = \frac{q(j \mid i, l^{(k)}, m^{(k)}) t(f_i^{(k)} \mid e_j^{(k)})}{\sum_{j'} q(j' \mid i, l^{(k)}, m^{(k)}) t(f_i^{(k)} \mid e_{j'}^{(k)})}
  \]

- M-step (assume K = # data points),
  \[
  t(f \mid e) = \frac{\sum_{k=1}^{K} \sum_{i,j} \delta^{(k)}(i, j) \mathbb{I}[f_i^{(k)} = f, e_j^{(k)} = e]}{\sum_{f'} \sum_{k=1}^{K} \sum_{i,j} \delta^{(k)}(i, j) \mathbb{I}[f_i^{(k)} = f', e_j^{(k)} = e]}
  \]

  \[
  q(j \mid i, l, m) = \frac{\sum_{k=1}^{K} \delta^{(k)}(i, j) \mathbb{I}[m^{(k)} = m, l^{(k)} = l]}{\sum_{j'} \sum_{k=1}^{K} \delta^{(k)}(i, j') \mathbb{I}[m^{(k)} = m, l^{(k)} = l]}
  \]
EM for IBM Models

- 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
- Initialize Model 2 from Model 1
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>
Example

Les embranchements que ils songeaient à fermer.

The branches they intend to close.
Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.

On Tuesday Nov. 4, earthquakes rocked Japan once again.
Phrase Movement
Thank you, I shall do so gladly.

Gracias, lo haré de muy buen grado.

Model Parameters

Emissions: \( P( F_1 = \text{Gracias} | E_{A1} = \text{Thank} ) \)

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

E:
Thank you, I shall do so gladly.

A:

F:
Gracias, lo haré de muy buen grado.

Model Parameters

Emissions: \( P( F_1 = \text{Gracias} | E_{A1} = \text{Thank} ) \)
Transitions: \( P( A_2 = 3 | A_1 = 1 ) \)
The HMM Model

- Model 2 preferred global monotonicity
- We want local monotonicity:
  - Most jumps are small
- HMM model (Vogel 96)

\[
P(f, a \mid e) = \prod_i P(a_i \mid a_{i-1}) P(f_i \mid e_{a_i})
\]

\[
P(a_i \mid a_{i-1}) \propto s(a_i - a_{i-1})
\]

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

[from Al-Onaizan and Knight, 1998]
Some Results (AER) [Och and Ney 03]

- Typical strategy is to use simpler models to initialize more complex ones

<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>$1^5H^45^3$</td>
<td>26.5</td>
<td>21.5</td>
<td>13.7</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>$1^5H^34^35^3$</td>
<td>26.5</td>
<td>20.4</td>
<td>13.4</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 6</td>
<td>$1^5H^46^3$</td>
<td>26.0</td>
<td>21.6</td>
<td>12.8</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>$1^5H^34^36^3$</td>
<td>25.9</td>
<td>20.3</td>
<td>12.5</td>
<td>8.7</td>
</tr>
</tbody>
</table>
Neural Alignment - Attention

- Just because IBM models aren’t the state of the art for machine translation anymore doesn’t mean alignment is useless.

- The neural counterpart for is called “attention” originally proposed by
  - Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio 2014 (>2K citations!!)
  - Proposed in the context of neural machine translation.
  - We will discuss their formulation later in this course

- For now, just to show the potential for alignment, we will discuss a paper we wrote last year.
Natural Language Inference

- Determine entailment/contradiction relationships between a premise and a hypothesis.

<table>
<thead>
<tr>
<th>Premise</th>
<th>Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 1</td>
<td>Bob is awake.</td>
</tr>
<tr>
<td>Hypothesis 2</td>
<td>It is sunny outside.</td>
</tr>
<tr>
<td>Hypothesis 3</td>
<td>Bob has a big house.</td>
</tr>
</tbody>
</table>
Motivation for this Work

Often times, alignment is sufficient, don’t need real sentence representations.

**Premise**
Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.

**Hypothesis 1**
Bob is awake.

**Premise**
Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.

**Hypothesis 2**
It is sunny outside.
Motivation for this Work

• Alignment plays key role in many NLP tasks:

  • Machine translation [Koehn, 2009]

  • Sentence Similarity [Haghighi et al., 2005; Koehn, 2009; Das and Smith, 2009, Chang et al., 2010; Fader et al., 2013]

  • Natural Language Inference [Marsi and Krahmer, 2005; McCartney et al., 2006; Hickl and Bensley, 2007; McCartney et al., 2008]

  • Semantic Parsing [Andreas et al., 2013]

• Attention is the neural counterpart to alignment [Bahdanau et al., 2014]
Decomposable Attention

1. Attend

\[
F(\text{alice, flute, music})
\]

2. Compare

\[
G(\text{park, outside}) = G(\text{alice, someone}) = \ldots
\]

3. Aggregate

\[
\hat{y} = H(\text{flute+ solo, music})
\]
Step 1: Attend

\[ e_{ij} = F^*(a_i, b_j) \]

In practice,

\[ e_{ij} = F(a_i)^\top F(b_j) \]

\[ \alpha_j = \sum_{i=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{kj})} a_i \]

\[ \beta_i = \sum_{j=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} b_j \]

sub-phrase in sentence 1 aligned to \( b_j \)

sub-phrase in sentence 2 aligned to \( a_i \)
Attend 2: Compare

Separately compare aligned subphrases:

\[ \mathbf{v}_{1,i} := G([a_i, \beta_i]) \quad \forall i \in [1, \ldots, n] \]

\[ \mathbf{v}_{2,j} := G([b_j, \alpha_j]) \quad \forall j \in [1, \ldots, n] \]

\( G \) is a feed forward network
Step 3: Aggregate

- Combine results and classify.

\[
\hat{y} = H([v_1, v_2])
\]

\[
v_1 = \sum_{i=1}^{n} v_{1,i}
\]

\[
v_2 = \sum_{j=1}^{n} v_{2,j}
\]

In practice, H is a feed forward neural network + normalization
Decomposable Attention

1. Attend

\[ F(\text{flute}, \text{music}) \]

2. Compare

\[ G(\text{park outside}, \text{flute+}) = G(\text{alice someone}, \text{flute+}) = G(\text{flute+}, \text{music}) \]

3. Aggregate

\[ \hat{y} = H(\text{park outside} + \text{alice someone} + \text{flute+} + \text{music}) \]
Empirical Results

Dataset: Stanford Natural Language Inference Corpus (SNLI, Bowman et al. 2015)

http://nlp.stanford.edu/projects/snli/

<table>
<thead>
<tr>
<th>Text</th>
<th>Judgments</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man inspects the uniform of a figure in some East Asian country.</td>
<td>contradiction</td>
<td>The man is sleeping</td>
</tr>
<tr>
<td></td>
<td>C C C C</td>
<td></td>
</tr>
<tr>
<td>An older and younger man smiling.</td>
<td>neutral</td>
<td>Two men are smiling and laughing at the cats playing on the floor.</td>
</tr>
<tr>
<td></td>
<td>N N E N N</td>
<td></td>
</tr>
<tr>
<td>A black race car starts up in front of a crowd of people.</td>
<td>contradiction</td>
<td>A man is driving down a lonely road.</td>
</tr>
<tr>
<td></td>
<td>C C C C</td>
<td></td>
</tr>
<tr>
<td>A soccer game with multiple males playing.</td>
<td>entailment</td>
<td>Some men are playing a sport.</td>
</tr>
<tr>
<td></td>
<td>E E E E E</td>
<td></td>
</tr>
<tr>
<td>A smiling costumed woman is holding an umbrella.</td>
<td>neutral</td>
<td>A happy woman in a fairy costume holds an umbrella.</td>
</tr>
<tr>
<td></td>
<td>N N E C N</td>
<td></td>
</tr>
</tbody>
</table>

549,367 sentence pairs for training
9,842 pairs for development
9,824 pairs for testing
# Empirical Results

[Parikh et al. 2016]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>78.2</td>
<td>80.6</td>
<td>81.4</td>
<td>82.1</td>
<td>83.2</td>
<td>83.5</td>
<td>86.1</td>
<td>86.3</td>
<td>86.3</td>
</tr>
<tr>
<td>Lexicalized Classifiers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM RNN Encoders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretrained GRU Encoders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree-Based CNN Encoders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPINN-PI Encoders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM with Attention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mLSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTMN w/ Attention Fusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This Work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This Work with Self Attention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 3M: 3 million parameters
- 15M: 15 million parameters
- 3.5M: 3.5 million parameters
- 3.7M: 3.7 million parameters
- 1.9M: 1.9 million parameters
- 3.4M: 3.4 million parameters
- 382K: 382,000 parameters
- 252K: 252,000 parameters
- 582K: 582,000 parameters
Conclusions

- Alignment is a central part of NLP.

- Unsupervised word alignment the foundation of phrased based machine translation

- The “neural” counterpart to alignment “attention” (Bahdanau et al. 2014) is a powerful tool today in deep learning for NLP.