**Administrivia**

Project 2 due one week from today!

P1 test set results: top 3

Su Wang: 84.03 F1 (86.10 P / 82.05 R)

Larger window size and Wikipedia gazetteer

Prateek Shrishail Kolhar: 82.32 F1 (82.61 P / 82.07 R)

Conjunctions of words, POS, and shapes in neighborhood

Very fast vectorized implementation (15s per epoch)

Yasumasa Onoe: 78.55 F1 (78.27 P / 78.83 R)

Used transition probabilities from HMM, character 5-grams and other feature tuning.

---

**Machine Translation: Examples**

**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 fedeli persone a palazzo di Giustizia e che ha poi avuto l'agente di dogana, s'è consegnato alla polizia, dopo aver cercato di togliere l'arresto di una donna in un complesso d'appartamenti alla periferia della città. Per tutt'oggi, il centro della città, sede della Contessa e dei Giudici 1990, 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 has given the customs agent has retreated. S' has been 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 day, the center of the city, center of the Contessa and of the Judges 1990, heart of a populous metropolitan area, remained paralyzed.
Levels of Transfer

Word-Level MT: Examples

la politique de la haine.
the policy of the hatred.

nous avons signé le protocole.
we have signed the protocol.

où était le plan solide?
where was the economic base?

Phrasal MT: Examples

metrics
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)
  - Lots more: TER, HTER, METEOR, …
- 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...
- Magnitude only meaningful on same language, corpus, number of references, probably only within system types...

Automatic Metrics Work (?)

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
Phrase-Based System Overview

Sentence-aligned corpus  Word alignments  Phrase table (translation model)

Tomorrow \rightarrow \text{I will fly to the conference in Canada}

Phrase table $P(f|e)$

Unlabeled English data

Noisy channel model: combine scores from translation model + language model to translate foreign to English

Many slides and examples from Philipp Koehn or John DeNero

Word Alignment

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

"Translate faithfully but make fluent English"
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?

Unsupervised Word Alignment

- Input: a bitext: pairs of translated sentences
- Output: alignments: pairs of translated words
  - Not always one-to-one!

1-to-Many Alignments

Evaluating Models

- How do we measure quality of a word-to-word model?
  - Method 1: use in an end-to-end translation system
    - Slow development cycle
    - Misleading if your MT system was “tuned” for certain aspects of bad alignments
  - 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
Alignment Error Rate

- Alignment Error Rate

- = Sure
- = Possible
■ = Predicted

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

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

---

IBM Model 1

Alignments: a hidden vector called an alignment specifies which English source (or a special null token) is responsible for each French target word.

\[ a = a_1 \ldots a_J \]

\[ P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i) \]

\[ = \prod_j \frac{1}{I + 1} P(f_j|e_i) \]

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

---

IBM Model 1 (Brown 93)

Model Parameters

- \( P(A_1 = 1) = 1/10 \), nothing to learn
- \( P( F_1 = \text{Gracias} | A_1 = 1) = P(\text{Gracias} | \text{Thank}) < \) learn these translation probs

---

IBM Model 1

E: Thank you , I shall do so gladly .

A: Gracias , lo haré de muy buen grado .
### EM for Model 1

- **Model 1 Parameters:**
  - Translation probabilities \( P(f_j|e_i) \)
- Start with \( P(f_j|e_i) \) uniform, including \( P(f_j|\text{null}) \)
- For each sentence, for each foreign position \( j \):
  - Calculate posterior over English positions
    \[
    P(a_j = i|f, e) = \frac{P(f_j|e_i)}{\sum_{i'} P(f_j|e_{i'})}
    \]
  - Increment count of word \( f \), with word \( e \), by these amounts
- Do for whole corpus, re-estimate \( P(f|e) \) with M-step

### 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

### HMM Model: Local Monotonicity

<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>
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

The HMM Model

- 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

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>

Language Modeling

Phrase-Based System Overview

\[ P(e|f) \propto P(f|e)P(e) \]

Noisy channel model: combine scores from translation model + language model to translate foreign to English

N-gram Language Modeling

- Could give several lectures on this!
- Estimate \( P(w_n | w_{n-k}, w_{n-k+1}, \ldots, w_{n-1}) \)
- Generative model: read off counts and normalize
  - \( P(\text{fox} | \text{the quick brown}) = 0.9, \text{etc.} \)
- Very complex distributions, need to smooth
  - Interpolate with lower-order models
  - Lots of complex techniques

Many slides and examples from Philipp Koehn or John DeNero
Phrase-Based MT

Phrase-Based System Overview

Unlabeled English data

“Translate faithfully but make fluent English”

Phrase table $P(f|e)$

Language model $P(e)$

Noisy channel model: combine scores from translation model + language model to translate foreign to English

Phrase-Based Translation Overview

Input: lo harto-rápidamente.

Translations: I’ll do it quickly. translates phrase by phrase, and considers reorderings.

Objective: $\arg \max_e [P(f|e) \cdot P(e)]$

$\arg \max_e \left[ \prod_{(e,f)} P(f|e) \cdot \prod_{i=1}^{w} P(e_i|e_{i-1}, e_{i-2}) \right]$

Many slides and examples from Philipp Koehn or John DeNero
Phrase-Based Decoding

Monotonic Word Translation

Decoder design is important: [Koehn et al. 03]

Beam Decoding

- For real MT models, this kind of dynamic program is a disaster (why?)
- Standard solution is beam search: for each position, keep track of only the best k hypotheses

```
for (fPosition in 1…|f|) 
  for (eContext in bestEContexts[fPosition]) 
    for (eOption in translations[fPosition]) 
      score = scores[fPosition-1][eContext] * LM(eContext+eOption) * TM(eOption, fWord[fPosition]) 
      bestEContexts.maybeAdd(eContext[2]+eOption, score)
```

Example from David Chiang
If monotonic, almost an HMM; technically a semi-HMM

- If distortion... now what?

```
for (fPosition in 1...|f|)
  for (lastPosition < fPosition)
    for (eContext in eContexts)
      for (eOption in translations[fPosition])
        ... combine hypothesis for (lastPosition ending in eContext) with eOption
```

Pruning: Beams + Forward Costs

- Problem: easy partial analyses are cheaper
  - Solution 1: use beams per foreign subset
  - Solution 2: estimate forward costs (A*-like)
Hypothesis Lattices

Syntactic Models

Translating with Tree Transducers

Input | Output
--- | ---
lo haré de muy buen grado . | Grammar

Grammar

ADV \rightarrow \langle de muy buen grado ; gladly \rangle
Translating with Tree Transducers

**Input**

<table>
<thead>
<tr>
<th>ADV</th>
<th>de muy buen grado</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo haré</td>
<td></td>
</tr>
</tbody>
</table>

**Output**

<table>
<thead>
<tr>
<th>ADV</th>
<th>gladly</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo haré</td>
<td>de muy buen grado</td>
</tr>
</tbody>
</table>

**Grammar**

```plaintext
ADV → ( de muy buen grado ; gladly )
```

S → ( lo haré ADV ; I will do it ADV )

S → ( de muy buen grado ; gladly )
Translating with Tree Transducers

**Input**

| ADV | 
|---|---|
| lo haré | de muy buen grado |

**Output**

| ADV | 
|---|---|
| gladly | 

**Grammar**

\[
S \rightarrow \langle \text{lo haré} \text{ ADV } . \; ; \text{ I will do it } \text{ ADV } . \rangle \\
ADV \rightarrow \langle \text{de muy buen grado} ; \text{ gladly} \rangle \\
\]

---

Translating with Tree Transducers

**Input**

| VP | 
|---|---|
| lo haré | de muy buen grado |

**Output**

| VP | 
|---|---|
| will do it | gladly |

**Grammar**

\[
VP \rightarrow \langle \text{lo haré} \text{ ADV } ; \text{ will do it } \text{ ADV } \rangle \\
S \rightarrow \langle \text{lo haré} \text{ ADV } . \; ; \text{ I will do it } \text{ ADV } . \rangle \\
ADV \rightarrow \langle \text{de muy buen grado} ; \text{ gladly} \rangle \\
\]
Syntactic Translation

- Lots of complexity: large phrase tables, errors introduced by parsers, parses don’t agree, inference is harder, ...

- Good for some languages (Japanese->English), but generally more trouble than it’s worth

- Easier method: syntactic “pre-reordering”

MT: Takeaways

- Word alignments: unsupervised process for finding word-level correspondences. Turn these into phrase level correspondences -> phrase table

- Language model: estimate n-gram model on a very large corpus

- Translation process: use beam search to find the best translation \( \arg\max_e P(f|e)P(e) \)