Objectives

- Understand the complexity of decoding in translation
- Discuss possible decoders
- Discuss metrics for end-to-end MT

Decoding

- Goal: Find a good translation
- Use word alignment models and language models
- Brown et al. (1993) left out the decoding algorithm!
- They left that for an IBM patent (1995)

Start with Bag "Generation"

1. as as give me please possible response soon your

2. disadvantages let me mention now of some the

Bag “Generation” (Decoding)

Evact 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.

TSP Reduced to Bag Generation

- Imagine bag generation with a bigram LM:
  - Vertices are words
  - Edges are transitions
  - Edge weights are $P(w' | w)$
  - Valid sentences are Hamiltonian paths
  - Looking for Hamiltonian path with least cost (e.g., negative log probability)
- Not the best news for word-based MT!
Why “Decoding”?  

**Decoding**  
Inferring something useful from a signal. Often by means of optimization.

---

TSP Reduced to Translation  

**Source**: French  
**Target**: English  

- Visit all French cities in any legal order.  
- Produce one English word for each city OR Produce a word on the border for bordered cities.  
- Apply your LM and TM

---

Solution  

**Source**: French  
**Target**: English  

- CE NE EST PAS CLAIR

---

Ingredients  

- What must a decoder do?  
  - Optimize: find the best translation, according to your models  
  - Fluency: Choose target word order  
  - Fidelity: For each source word(s), choose target word  
- If decoding is state-space search, what does the state space consist of?  
  - Partial (e.g., left) prefixes and whole sentences  
  - Together with (partial) alignments  
- How should we search through that space?  
  - Incremental state construction  
  - Prefer high scoring states

---

Simple Decoders  

- **Simplest possible decoder:**  
  - Enumerate sentences and alignments, score each with TM and LM  
- **Monotone decoder:**  
  - Greedy monotone: assign each French word it’s most likely English translation  
    - Given $f$, let $\hat{\theta} = \text{argmax}_{\theta} P(f|\theta)$  
    - Notice: only uses TM!  
  - Better monotone: Viterbi  
    - Uses LM and TM

---

Stack Decoding (IBM)  

- More commonly called: A*  
- One “stack” (priority queue) per  
  - candidate sentence length, or  
  - subset of input words  
- Employ heuristic estimates for completion cost  
  - (Which admissible heuristics have people tried?)  
- Beam Search  
  - Limit size of priority queues
Other Algorithms

- Dynamic programming: possible if we make assumptions about the set of allowable permutations
- Integer Linear Programming (IP)
  - Complete
  - Optimal

Possible Greedy Search Decoder

- Start: with monotone decoding
- Apply operators:
  - Change a word translation
  - Change two word translations
  - Insert a word into the English (zero-fertile French word)
  - Remove a word from the English (null-generated French word)
  - Swap two adjacent English words
- Pick highest scoring state, according to LM and TM.
- Do hill-climbing (move to higher scoring states)
- Stop when no improvement is to be had
- Could do annealing or genetic algorithm or something even smarter

Greedy Decoding Example

Given French source sentence

Think of this as the start state for the Greedy Decoder

Greedy Decoding Example

It has many possible child states, determined by possible operators
Greedy Decoding Example

One of those child states has the best score according to LM and TM.

Greedy Decoding Example

Here’s the best child state and the operator that produced this state. The numbers denote English positions.

Greedy Decoding Example

Here’s the best child state in the third generation.

Greedy Decoding Example

Experimental Results

- From [Germann et al., 2001]

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Type</th>
<th>Length error</th>
<th>Length energy</th>
<th>Error error</th>
<th>Error energy</th>
<th>Error/energy</th>
<th>Word</th>
<th>Word energy</th>
<th>Word error</th>
<th>Word/energy</th>
<th>Pron</th>
<th>Pron energy</th>
<th>Pron error</th>
<th>Pron/energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>32</td>
<td>8.79</td>
<td>5</td>
<td>53</td>
<td>44</td>
<td>35</td>
<td>2</td>
<td>6</td>
<td>58.05</td>
<td>58.05</td>
<td>58.05</td>
<td>5</td>
<td>58.05</td>
<td>58.05</td>
<td>58.05</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
<td>3.01</td>
<td>6</td>
<td>55</td>
<td>53</td>
<td>39</td>
<td>2</td>
<td>6</td>
<td>61.77</td>
<td>61.77</td>
<td>61.77</td>
<td>5</td>
<td>61.77</td>
<td>61.77</td>
<td>61.77</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>1.67</td>
<td>13</td>
<td>54</td>
<td>55</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>56.97</td>
<td>56.97</td>
<td>56.97</td>
<td>5</td>
<td>56.97</td>
<td>56.97</td>
<td>56.97</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>1.68</td>
<td>13</td>
<td>54</td>
<td>55</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>56.97</td>
<td>56.97</td>
<td>56.97</td>
<td>5</td>
<td>56.97</td>
<td>56.97</td>
<td>56.97</td>
</tr>
</tbody>
</table>

Suppose a decoder outputs $\tilde{c}$ while the optimal decoding turns out to be $c$. Then we consider six possible outcomes:

- no error (NE): $\tilde{c} = c$, and $\tilde{c}$ is a perfect translation.
- pure model error (PME): $\tilde{c} = c$, but $\tilde{c}$ is not a perfect translation.
- directly search error (DSE): $\tilde{c} \neq c$, and $\tilde{c}$ is a perfect translation, while $c$ is not.
- compounded error (CE): $\tilde{c} \neq c$, and $\tilde{c}$ is a perfect translation, while $c$ is not.
**Major Focus of Ongoing Research**

- Better models
- Keeping phrases together
- Using syntactic constraints to keep words syntactically close and coherent

**Evaluating Translation**

- How do we measure translation quality?
- It’s difficult!
  - Option: ask human judges
  - Option: use reference translations
    - NIST
    - BLEU
    - others

**Bleu**

- Mean n-gram precision
- Multiple reference sentences
  
  \[ \text{Ref}_1: \text{the ball was kicked by the boy.} \\
  \text{Ref}_2: \text{the boy kicked the ball} \\
  \text{Hyp}: \text{the boy was kicked by the ball} \\
  \text{Sec}: f_1, f_2, f_3, f_4, f_5 \]

- Used everywhere now
- WANTED: Better metrics

**Next**

- Phrase-based Translation
  - Last lecture!