Machine Translation 1

CS 287

Quiz: CRF

If we have a conditionat

Answer

Today's Lecture

- ► History of Translation
- Statistical Machine Translation
- Simplified Translation Models
- Search for Translation

Next Class: Neural Machine Translation

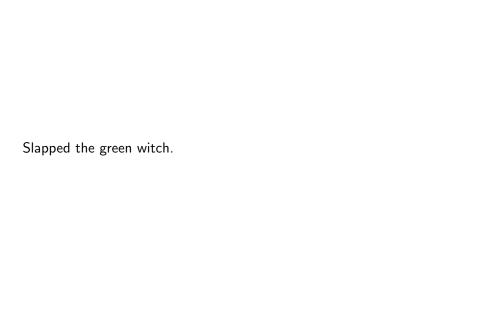
Contents

True Translation

Datasets for Machine Translation

Hansard's Corpus

Warren Weaver's View of Translation



Shannon Noisy Channel Model

Encoder-Decoder

- ► Idea
- ► Encoder
- Decoder

Contents

True Translation

Thought Experiment: One-to-One Ordered MT

What if the two languages just involved word to word translation?

- $\mathbf{x} = [w_1^s \ w_2^s \ w_3^s \dots w_n^s]$
- $\mathbf{y} = [w_1^t \ w_2^t \ w_3^t \dots w_n^t]$

Noisy-Channel Model

$$p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$$

Translation

- 1. Language Model (p(y))
- 2. Translation Model $(p(\mathbf{x}|\mathbf{y}))$

Example: Hidden Markov Model

y 1	y 2	y 3	•••	y n
x_1	\mathbf{x}_2	x ₃		\mathbf{x}_n

Example: Hidden Markov Model

$$p(\mathbf{y}_i|\mathbf{y}_{i-1})$$

Language Model

$$p(\mathbf{y}) = \prod_{i=1} p(w_i^t | w_{i-n+1}^t, \dots, w_{i-1}^t)$$

Language Model

$$p(\mathbf{y}) = \prod_{i=1} p(w_i^t | w_{i-n+1}^t, \dots, w_{i-1}^t)$$

How might you estimate this?

$$p(\mathbf{y})$$

- ► Language model. Standard forms of Markov model estimation
- Could use n-gram model or NNLM

Translation Model

$$ho(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{n}
ho(\mathbf{x}_{i}|\mathbf{y}_{i})$$

$$ho(\mathbf{x}_{i}|\mathbf{y}_{i})$$

Assume we have many examples of language. Why estimate separate LM and TM?

Conditional Random Field

$$\mathop{\arg\max}_{w^t_{1:n}} f(\mathbf{x}, w^t_{1:n}) =$$

Contents

True Translation

Thought Experiment 2: Out-of-Order

Assume 1-to-1 still but allow any order.

Latent Variable

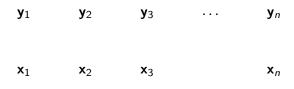
 \boldsymbol{a} maps each source word to a target word

$$\rho(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{a}} \rho(\mathbf{y}) \rho(\mathbf{x}, \mathbf{a}|\mathbf{y})$$

For efficiency, max-over-alignment,

$$\underset{\mathbf{a}, w_{1:n}^t}{\operatorname{arg max}} f(\mathbf{x}, w_{1:n}^t, \mathbf{a})$$

Example: Possible Alignment



Decoding Quiz

▶ In monotonic case $O(|\mathcal{C}|^2)$, what is complexity?

Answer ()

- Finding optimal translation is NP-Hard!
- Reduction from TSP:
 - 1. Each city becomes a source word with a single translation word.
 - 2. Distance between cities is a bigram LM score $p(w_i^t|w_{i-1}^t)$ between words.
 - 3. A tour is a complete translation (each word used = each city visited)

How do you find answer?

$$f(\mathbf{x}, w_{1:n}^t, \mathbf{a}) = \sum_{i=1}^n \log \hat{y}(w_{i-1})_{w_i, a_i}$$

with constraint that a_i uses each word once.



Bit-Set Beam Search

 $[\mathsf{Describe}\ \mathsf{on}\ \mathsf{board}]$

$p(w^t w^s)$	

How do you estimate this score?

Compute Alignments



More Statistical Machine Translation

- Handling Length Issues
- Producing and Symmetrizing Alignments
- ► Tuning Systems and MERT
- Rare and Unseen Words
- Syntactic Translation