Machine Translation 1

CS 287

Review: Conditional Random Field (Lafferty et al, 2001)

- Model consists of unnormalized weights

$$\log \hat{\mathbf{y}}(c_{i-1})_{c_i} = feat(\mathbf{x}, c_{i-1})\mathbf{W} + \mathbf{b}$$

- Out of log space,

$$\hat{\mathbf{y}}(c_{i-1})_{c_i} = \exp(feat(\mathbf{x}, c_{i-1})\mathbf{W} + \mathbf{b})$$

- Score of the sequence, (same as last few classes)

$$f(\mathbf{x}, c_{1:n}) = \sum_{i=1}^{n} \log \hat{y}(c_{i-1})_{c_i}$$

- Objective is based on global NLL of this sequence distribution

$$\mathbf{z}_{c_{1:n}} = f(\mathbf{x}, c_{1:n})$$

Review: Computing the Softmax

Want to compute:

$$p(\mathbf{y} = \delta(c_{1:n})|\mathbf{x}) = \frac{\prod_{i=1}^{n} \hat{\mathbf{y}}(c_{i-1})_{c_i}}{\sum_{c'_{1:n}} \prod_{i=1}^{n} \hat{\mathbf{y}}(c'_{i-1})_{c'_i}}$$

- $\prod_{i=1}^{n} \hat{\mathbf{y}}(c_{i-1})_{c_i}$; easy to compute
- $\sum_{c'_{1:n}} \prod_{i=1}^{n} \hat{\mathbf{y}}(c'_{i-1})_{c'_i}$; can use forward algorithm.

j

Softmax goes from $O(|\mathcal{C}|^n)$ to $O(|\mathcal{C}|^2)$.

Review: Final Gradients

$$
\begin{aligned}
\frac{\partial L}{\partial \log \hat{y}_i(c'_{i-1})_{c'_i}} &= \sum_{d_{1:n}} \frac{\partial z_{d_{1:n}}}{\partial \log \hat{y}_i(c'_{i-2})_{c'_i}} \frac{\partial L}{\partial z_{d_{1:n}}} \\
&= \sum_{c'_{1:i-2}, c'_{i+1:n}} \frac{\partial L}{\partial z_{c'_{1:n}}} \\
&= p(\mathbf{y}_{i-1} = c'_{i-1}, \mathbf{y}_i = c'_i | \mathbf{x}) - \mathbf{1}(c'_{i-1} = c_{i-1} \wedge c'_i = c_i)
\end{aligned}
$$

- First term, marginals of the CRF.
- Second term, indicator of whether edge is in gold.

Quiz: CRF

Note: Nothing in our definition of CRFs relied on $\mathbf{y}_i$ to align with $\mathbf{x}_i$ (conditioned on full sequence).

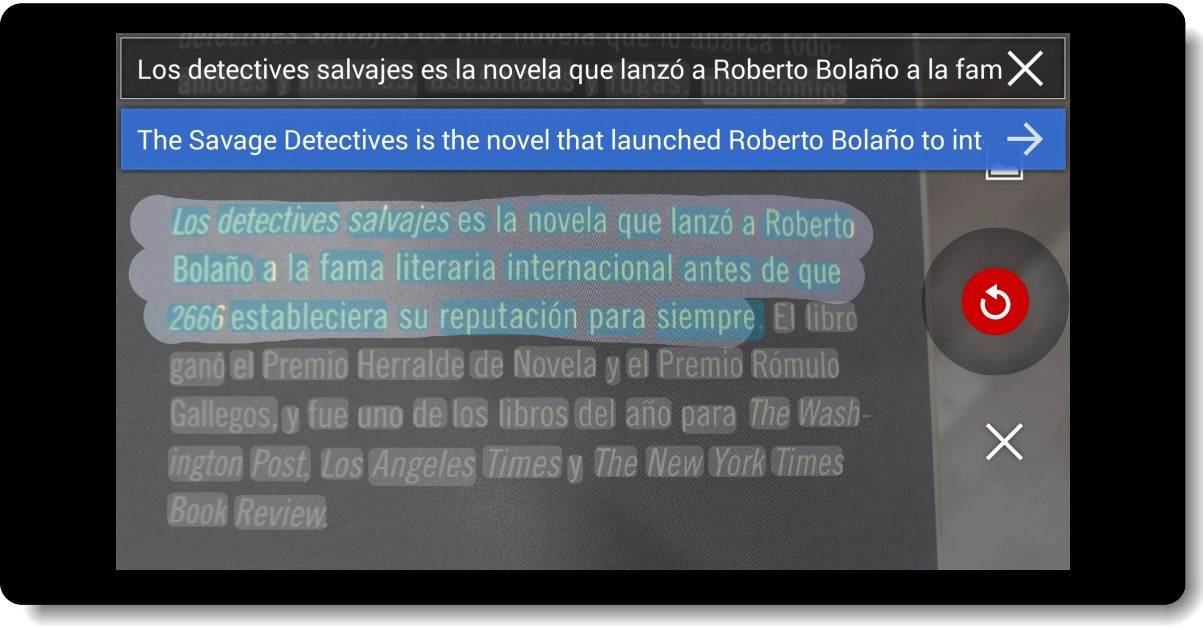
For this quiz, imagine we have an input sequence $\mathbf{x}$, and we want to find the optimal output sequence $\mathbf{y}$ but we do not fix $n < N$. For instance finding the best word segmentation of an unsegmented input $\mathbf{x}$.

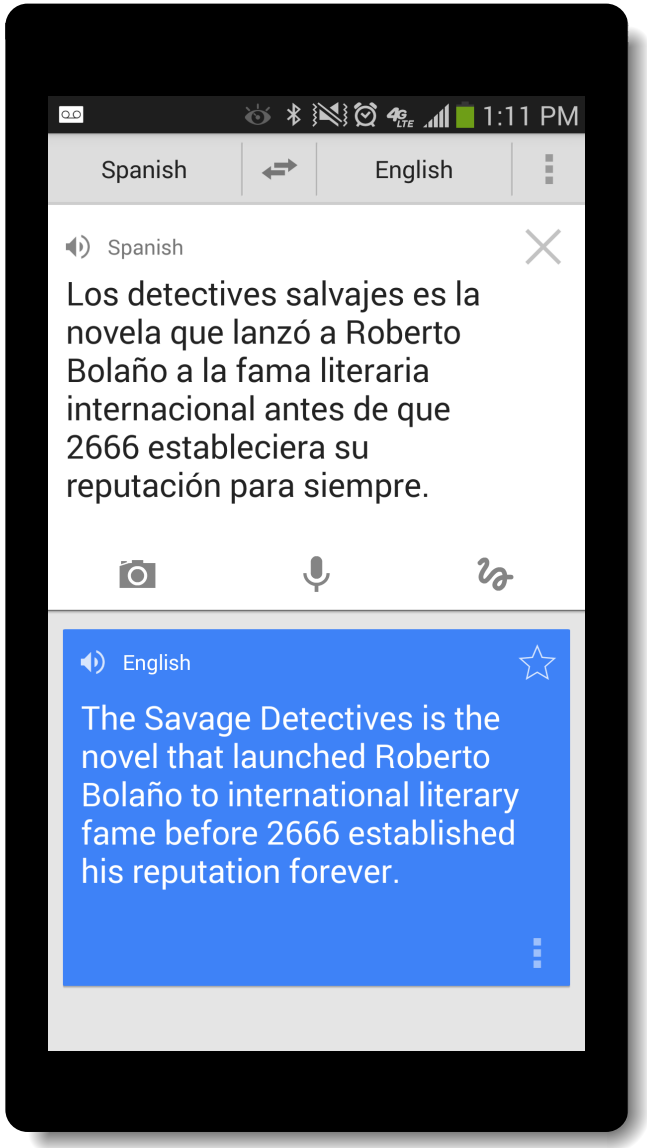
- How would you find

$$\arg\max_{n, c_{1:n}} f(\mathbf{x}, c_{1,n})?$$

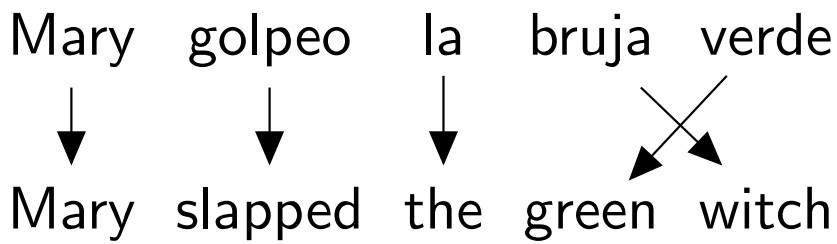
- How would you train

$$f(\mathbf{x}, c_{1,n}; \theta)?$$


Los detectives salvajes es la novela que lanzó a Roberto Bolaño a la fam
The Savage Detectives is the novel that launched Roberto Bolaño to int
Los detectives salvajes es la novela que lanzó a Roberto Bolaño a la fama literaria internacional antes de que 2666 estableciera su reputación para siempre. El libro ganó el Premio Herralde de Novela y el Premio Rómulo Gallegos, y fue uno de los libros del año para The Washington Post, Los Angeles Times y The New York Times Book Review.


1:11 PM
Spanish
English
Spanish
Los detectives salvajes es la novela que lanzó a Roberto Bolaño a la fama literaria internacional antes de que 2666 estableciera su reputación para siempre.
English
The Savage Detectives is the novel that launched Roberto Bolaño to international literary fame before 2666 established his reputation forever.

Machine Translation



Today's Lecture

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

Next Class: Neural Machine Translation

Contents

Early Ideas of Translation

> *. . . one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.*

Letter from Warren Weaver to Norbert Weiner, 1947

Shannon's Noisy Channel (Shannon, 1948)

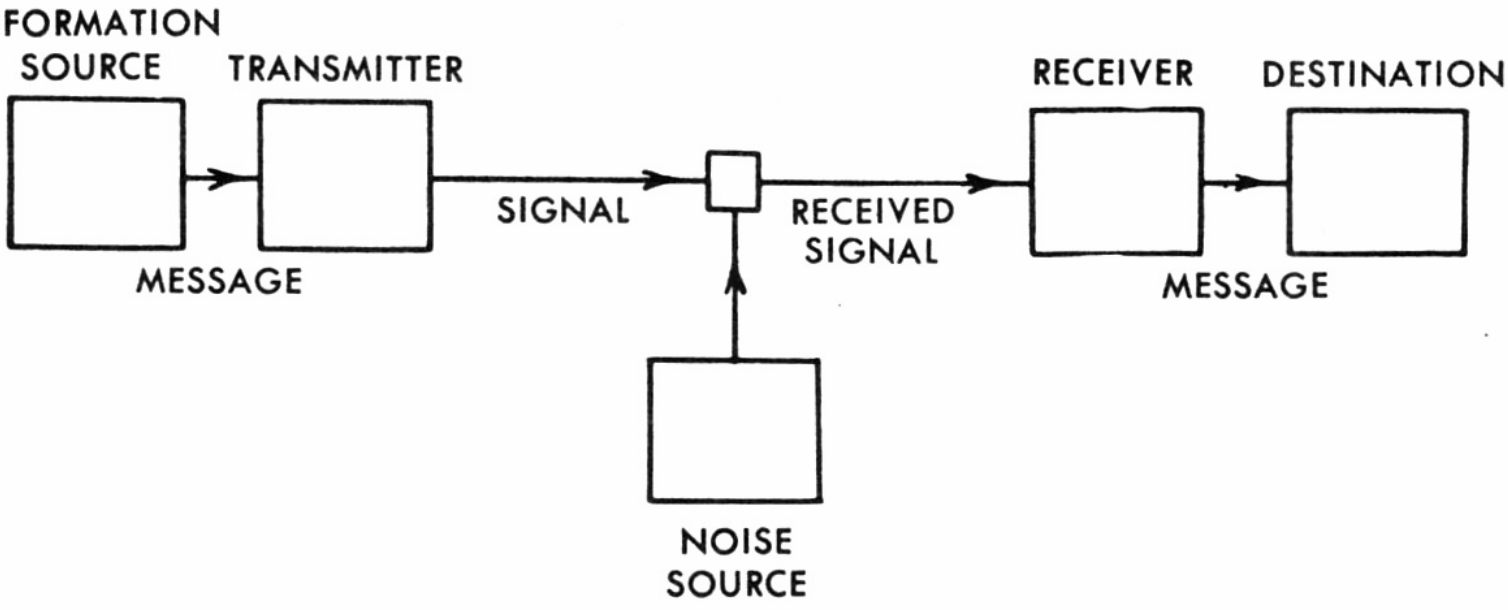


Fig. 1. — Schematic diagram of a general communication system.

Noisy Channel

- Method provides a basis for thinking about translation.
- However how do you actually learn what the encoder/decoder are?
- We will focus on learning from data.

The 35th Parliament having been dissolved by proclamation on Sunday, April 27, 1997, and writs having been issued and returned, a new Parliament was summoned to meet for the dispatch of business on Monday, September 22, 1997, and did accordingly meet on that day. Monday, September 22, 1997 This being the day on which Parliament was convoked by proclamation of His Excellency the Governor General of Canada for the dispatch of business, and the members of the House being assembled: Robert Marleau, Esquire, Clerk of the House of Commons, read to the House a letter from the Administrative Secretary to the Governor General informing him that the Right Honourable Antonio Lamer, in his capacity as Deputy Governor General, would proceed to the Senate chamber to open the first session of the 36th Parliament of Canada on Monday, September 22 at Ottawa. A message was delivered by the Gentleman Usher of the Black Rod as follows: Members of the House of Commons:

La trente-cinquime lgislature ayant t proroge et les Chambres dissoutes par proclamation le dimanche 27 avril 1997, puis les brefs ayant t mis et rapports, les nouvelles Chambres ont t convoques pour l'expdition des affaires le lundi 22 septembre 1997 et, en consquence, se sont runies le jour dit. Le lundi 22 septembre 1997. Le Parlement ayant t convoqu pour aujourd'hui, par proclamation de Son Excellence le Gouverneur gnral du Canada pour l'expdition des affaires, et les dputs tant runis: M. Robert Marleau, greffier de la Chambre, donne lecture d'une lettre du directeur administratif du Gouverneur gnral annonant que le trs honorable Antonio Lamer, titre de supplant du Gouverneur gnral, se rendra la salle du Snat le lundi 22 septembre 1997, Ottawa, pour ouvrir la premire session de la trente-sixime lgislature. Le gentilhomme huissier de la verge noire apporte le message suivant: Membres de la Chambre des communes:

Hansard's Corpus

Statistical Machine Translation

The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

We describe a series of five statistical models of the translation process and give algorithms for estimating the parameters of these models given a set of pairs of sentences that are translations of one another. We define a concept of word-by-word alignment between such pairs of sentences. For any given pair of such sentences each of our models assigns a probability to each of the possible word-by-word alignments. We give an algorithm for seeking the most probable of these alignments. Although the algorithm is suboptimal, the alignment thus obtained accounts well for the word-by-word relationships in the pair of sentences. We have a great deal of data in French and English from the proceedings of the Canadian Parliament. Accordingly, we have restricted our work to these two languages; but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages. We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus.

Modern Statistical Translation

Translation systems are trained with a vast amount of data,

- Training uses 2.5 billion parallel documents.
- Language model trained with 500 billion English words.

Google Translate has used a statistical system since 2006

- 80 languages
- 200 million users a month
- Over 10 billion words translated a day

Evaluation

How do you evaluate machine translation output?

- Model produces one output, compared to several references.
- Want a corpus-wide metric (short sentences count less)

BLEU (Papineni et al, 2002)

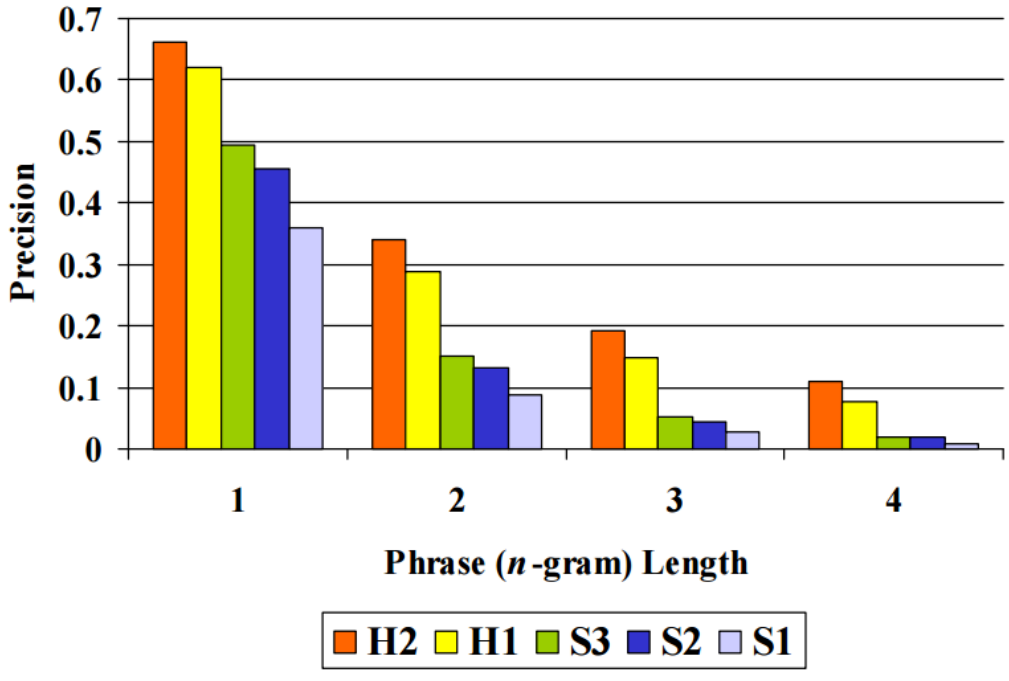
Main metric: BLEU (bilingual evaluation understudy)

- Calculate the *precision* of unigrams, bigrams, trigrams, 4-grams
- Take the geometric mean of corpus precision scores
- Use length penalty to ensure appropriately long translations

$$\log BLEU = \min(0, 1 - \frac{\text{ref len}}{\text{cand len}}) + \text{ mean of log precisions}$$

BLEU

Figure 2: Machine and Human Translations



Contents

Noisy-Channel Model

Notation: Source words and target words

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

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

Translation is reversing noisy channel-process,

- $p(\mathbf{y})$ - prob generating target sentence
- $p(\mathbf{x}|\mathbf{y})$ - prob of converting to source language

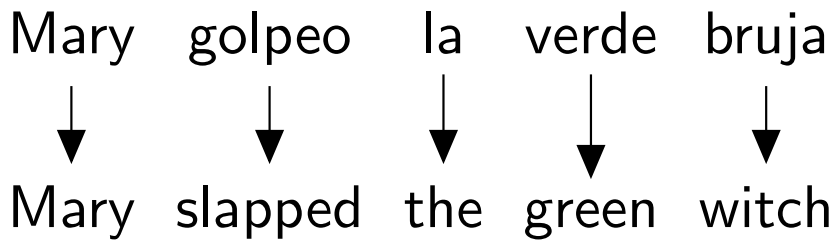
Translation

How do we model these two distributions?:

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

One-to-One In-Order Translation

Thought Experiment 1: What if the two languages just involved word to word translation?



Notation: Source words and target words

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

Simple One-to-One Model

1. Language Model; words depend on previous word

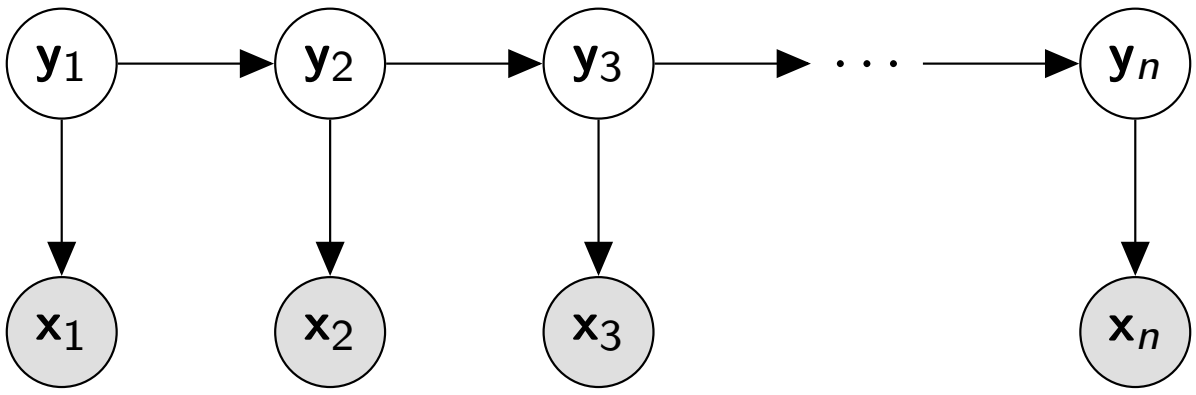
$$p(\mathbf{y}) = \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{y}_{i-1})$$

2. Translation Model; source word depends on current position

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

What model is this?

Answer: Hidden Markov Model



How might you estimate this?

- Language model. Standard forms of Markov model estimation (Could use n-gram model or NNLM)
- Translation Model

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

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

Assume we have many examples of language.

Why estimate separate LM and TM?

Conditional Random Field

- Could also utilize CRF model.
- Finding the optimal translation (as in quiz)

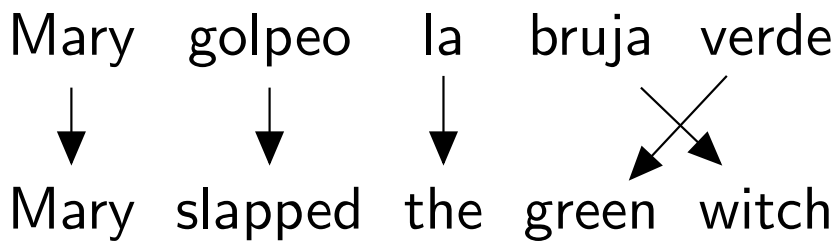
$$\arg\max_{w_{1:n}^t} f(\mathbf{x}, w_{1:n}^t)$$

- What would be benefits? Downsides?

Contents

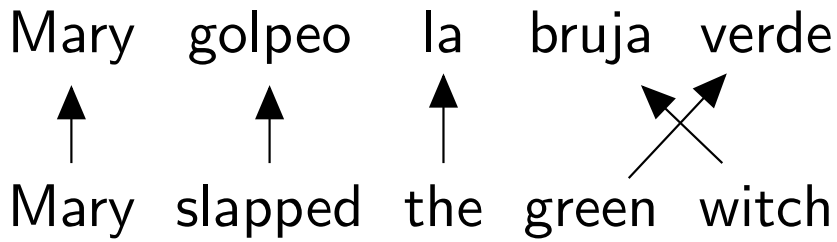
Out-of-Order One-to-One Translation

Thought Experiment 2: Assume 1-to-1 still but allow any order.



Alignment

- **a**; alignment mapping each target word to a source word
- Assuming one-to-one



	Mary	golpeo	la	bruja	verde
Mary	X				
slapped		X			
the			X		
green					X
witch				X	

Alignment Model

- Probability of alignment order,

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

- Models typically look at movement and past alignment choices,

$$p(\mathbf{a} = c_{1:n}|\mathbf{y}) = \prod_{i=1}^{n} p(a_i = c_i|a_{i-1} = c_{i-1}, i)$$

- But with constraint that all words used exactly once.
- (Vastly Simplified version, many different approaches)

Using Alignments

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

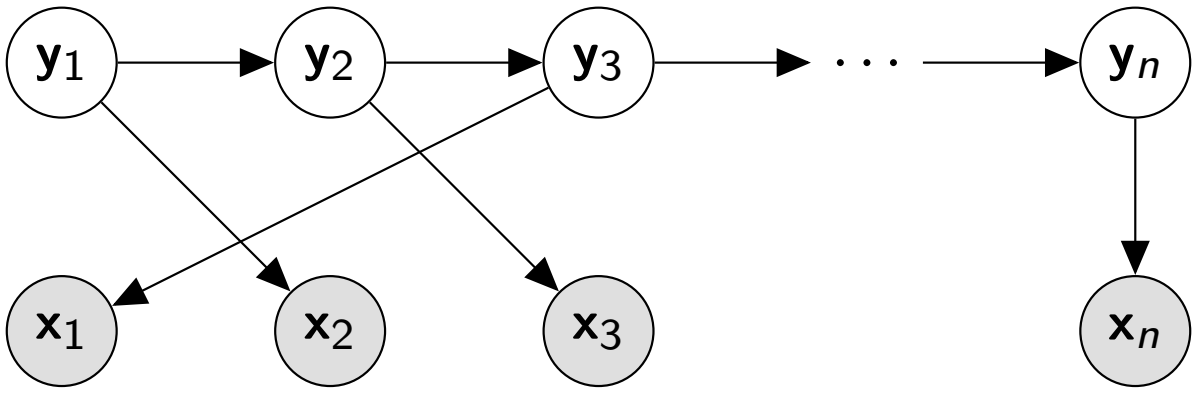
With alignment,

$$p(\mathbf{x}|\mathbf{a},\mathbf{y}) = \prod_{i=1}^{n} p(\mathbf{x}_{a_i}|\mathbf{y}_i)$$

Sum-over-alignment approximated with a max-over-alignment,

$$\arg\max_{j,w_{1:n}^t} \prod_{i=1}^{n} p(\mathbf{x}_{a_i}|\mathbf{y}_i = w_i^t)p(\mathbf{y}_i = w_i^t|\mathbf{y}_{i-1} = w_{i-1}^t)p(a_i = c_i|a_{i-1} = c_{i-1}, i)$$

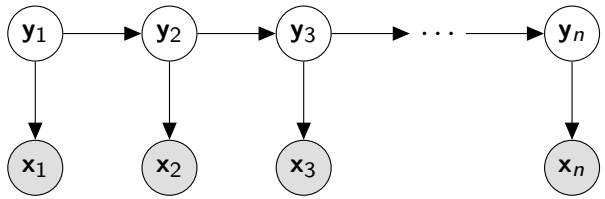
Example: Possible Alignment



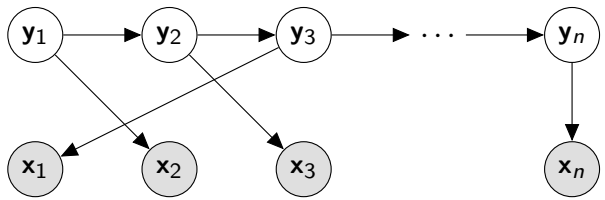
Decoding Quiz

We have seen two translation models, one with a fixed order and one where we had alignment as a latent variable.

- What is the complexity in the fixed-order case?



- What is the complexity when we max-over-alignments?



Answer

- In order time is $O(|\mathcal{W}|^2)$. (But exact still intractable).
- 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?

$$\prod_{i=1}^{n} p(\mathbf{x}_{a_i}|\mathbf{y}_i = w_i^t)p(\mathbf{y}_i = w_i^t|\mathbf{y}_{i-1} = w_{i-1}^t)p(a_i = c_i|a_{i-1} = c_{i-1}, i)$$

With constraint that c_i uses each word once.

-

Bit-Set Beam Search

[Describe on board]

Contents

Training the Translation Model

$$p(w^t|w^s)$$

How do you estimate this score?

MOSES

[Show Intro]

More Statistical Machine Translation

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