Phrase-Based Translation

Michael Collins, Columbia University

Roadmap for the Next Few Lectures

- Last time: IBM Models 1 and 2
- Today: *phrase-based* models

Overview

- Learning phrases from alignments
- A phrase-based model
- Decoding in phrase-based models

Phrase-Based Models

- First stage in training a phrase-based model is extraction of a *phrase-based (PB) lexicon*
- A PB lexicon pairs strings in one language with strings in another language, e.g.,

nach Kanada	$\leftrightarrow$	in Canada
zur Konferenz	$\leftrightarrow$	to the conference
Morgen	$\leftrightarrow$	tomorrow
fliege	$\leftrightarrow$	will fly
...		

An Example (from tutorial by Koehn and Knight)

- A training example (Spanish/English sentence pair):

  Spanish: Maria no daba una bofetada a la bruja verde

  English: Mary did not slap the green witch

- Some (not all) phrase pairs extracted from this example:

  (Maria $\leftrightarrow$ Mary), (bruja $\leftrightarrow$ witch), (verde $\leftrightarrow$ green),
  (no $\leftrightarrow$ did not), (no daba una bofetada $\leftrightarrow$ did not slap),
  (daba una bofetada a la $\leftrightarrow$ slap the)

- We'll see how to do this using *alignments* from the IBM models (e.g., from IBM model 2)

Recap: IBM Model 2

- IBM model 2 defines a distribution $p(a, f|e, m)$ where f is foreign (French) sentence, e is an English sentence, a is an *alignment*, m is the length of the foreign sentence
- A useful by-product: once we've trained the model, for any (f, e) pair, we can calculate

$$a^* = \arg\max_a p(a|f, e, m) = \arg\max_a p(a, f|e, m)$$

under the model. a^* is the **most likely alignment**

English: Mary did not slap the green witch

Spanish: Maria no daba una bofetada a la bruja verde

Representation as Alignment Matrix

	Maria	no	daba	una	bof'	a	la	bruja	verde
Mary	●								
did						●			
not		●							
slap			●	●	●				
the							●		
green									●
witch								●	

(Note: “bof”' = “bofetada”)

In IBM model 2, each foreign (Spanish) word is aligned to exactly one English word. The matrix shows these alignments.

Finding Alignment Matrices

- Step 1: train IBM model 2 for $p(f \mid e)$, and come up with most likely alignment for each (e, f) pair
- Step 2: train IBM model 2 for $p(e \mid f)$ and come up with most likely alignment for each (e, f) pair
- We now have two alignments:
  take intersection of the two alignments as a starting point

Alignment from $p(f \mid e)$ model:

	Maria	no	daba	una	bof'	a	la	bruja	verde
Mary	●								
did						●			
not		●							
slap			●	●	●				
the							●		
green									●
witch								●	

Alignment from $p(e \mid f)$ model:

	Maria	no	daba	una	bof'	a	la	bruja	verde
Mary	●								
did		●							
not		●							
slap					●				
the							●		
green									●
witch								●	

Intersection of the two alignments:

	Maria	no	daba	una	bof'	a	la	bruja	verde
Mary	●								
did									
not		●							
slap					●				
the							●		
green									●
witch								●	

The intersection of the two alignments has been found to be a very reliable starting point

Heuristics for Growing Alignments

- Only explore alignment in **union** of $p(f \mid e)$ and $p(e \mid f)$ alignments
- Add one alignment point at a time
- Only add alignment points which align a word that currently has no alignment
- At first, restrict ourselves to alignment points that are "neighbors" (adjacent or diagonal) of current alignment points
- Later, consider other alignment points

The final alignment, created by taking the intersection of the two alignments, then adding new points using the growing heuristics:

	Maria	no	daba	una	bof'	a	la	bruja	verde
Mary	●								
did		●							
not		●							
slap			●	●	●				
the						●	●		
green									●
witch								●	

Note that the alignment is no longer many-to-one: potentially multiple Spanish words can be aligned to a single English word, and vice versa.

Extracting Phrase Pairs from the Alignment Matrix

	Maria	no	daba	una	bof'	a	la	bruja	verde
Mary	●								
did		●							
not		●							
slap			●	●	●				
the						●	●		
green									●
witch								●	

- A phrase-pair consists of a sequence of English words, e, paired with a sequence of foreign words, f
- A phrase-pair (e, f) is *consistent* if: 1) there is at least one word in e aligned to a word in f; 2) there are no words in f aligned to words outside e; 3) there are no words in e aligned to words outside f
  e.g., (Mary did not, Maria no) is consistent. (Mary did, Maria no) is *not* consistent
- We extract all consistent phrase pairs from the training example.

Probabilities for Phrase Pairs

- For any phrase pair (f, e) extracted from the training data, we can calculate

$$t(f|e) = \frac{Count(f, e)}{Count(e)}$$

e.g.,

$$t(\text{daba una bofetada} \mid \text{slap}) = \frac{Count(\text{daba una bofetada}, \text{slap})}{Count(\text{slap})}$$

An Example Phrase Translation Table

An example from Koehn, EACL 2006 tutorial. (Note that we have $t(e|f)$ not $t(f|e)$ in this example.)

- Phrase Translations for *den Vorschlag*

English	t(e|f)	**English**	t(e|f)
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Overview

- Learning phrases from alignments
- A phrase-based model
- Decoding in phrase-based models

Phrase-Based Systems: A Sketch

$\boxed{\text{Today}}$

$\boxed{\text{Heute}}$ werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

$$
\begin{aligned}
\text{Score} \quad = \quad & \underbrace{\log q(\text{Today} \mid *, *)}_{\text{Language model}} \\
+ \quad & \underbrace{\log t(\text{Heute} \mid \text{Today})}_{\text{Phrase model}} \\
+ \quad & \underbrace{\eta \times 0}_{\text{Distortion model}}
\end{aligned}
$$

Phrase-Based Systems: A Sketch

Today $\boxed{\text{we shall be}}$

Heute $\boxed{\text{werden wir}}$ uber die Wiedereroffnung
des Mont-Blanc-Tunnels diskutieren

$$\text{Score} = \underbrace{\log q(\text{we}|*,\ \text{Today}) + \log q(\text{shall}|\text{Today, we}) + \log q(\text{be}|\text{we, shall})}_{\text{Language model}}$$

$$+ \underbrace{\log t(\text{werden wir} \mid \text{we shall be})}_{\text{Phrase model}}$$

$$+ \underbrace{\eta \times 0}_{\text{Distortion model}}$$

Phrase-Based Systems: A Sketch

$\boxed{\text{Today we shall be debating}}$

Today we shall be $\boxed{\text{debating}}$

Heute werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels $\boxed{\text{diskutieren}}$

$$\text{Score} = \underbrace{\log q(\text{debating}|\text{shall, be})}_{\text{Language model}}$$

$$+ \underbrace{\log t(\text{diskutieren} \mid \text{debating})}_{\text{Phrase model}}$$

$$+ \underbrace{\eta \times 6}_{\text{Distortion model}}$$

Phrase-Based Systems: A Sketch

Today we shall be debating $\boxed{\text{the reopening}}$

Heute werden wir $\boxed{\text{uber die Wiedereroffnung}}$ des Mont-Blanc-Tunnels diskutieren

Phrase-Based Systems: A Sketch

Today we shall be debating the reopening
$\boxed{\text{of the Mont Blanc tunnel}}$

Heute werden wir uber die Wiedereroffnung
$\boxed{\text{des Mont-Blanc-Tunnels}}$ diskutieren