COMS 4705, Fall 2011

Machine Translation Part III

Roadmap for the Next Few Lectures

- Lecture 1 (last time): IBM Models 1 and 2
- Lecture 2 (today): *phrase-based* models
- Lecture 3: Syntax in statistical machine translation

Overview

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

(Thanks to Philipp Koehn for giving me the slides from his EACL 2006 tutorial)

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

  $$P(\mathbf{a}, \mathbf{f}|\mathbf{e})$$

  where $\mathbf{f}$ is foreign (French) sentence, $\mathbf{e}$ is an English sentence, $\mathbf{a}$ is an *alignment*

- A useful by-product: once we've trained the model, for any $(\mathbf{f}, \mathbf{e})$ pair, we can calculate

  $$\mathbf{a}^* = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{f}, \mathbf{e}) = \arg\max_{\mathbf{a}} P(\mathbf{a}, \mathbf{f}|\mathbf{e})$$

  under the model. $\mathbf{a}^*$ is the **most likely alignment**

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 $P(\mathbf{f} \mid \mathbf{e})$, and come up with most likely alignment for each $(\mathbf{e}, \mathbf{f})$ pair

- Step 2: train IBM model 4 for $P(\mathbf{e} \mid \mathbf{f})$ and come up with most likely alignment for each $(\mathbf{e}, \mathbf{f})$ pair

- We now have two alignments:
  take intersection of the two alignments as a starting point

Alignment from $P(\mathbf{f} \mid \mathbf{e})$ model:

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

Alignment from $P(\mathbf{e} \mid \mathbf{f})$ 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 there are no words in f aligned to words outside e, and 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: "no" is aligned to "not", which is not in the string "Mary did"
- We extract all consistent phrase pairs from the training example. See Koehn, EACL 2006 tutorial, **pages 103-108** for illustration.

Probabilities for Phrase Pairs

- For any phrase pair (f, e) extracted from the training data, we can calculate
$$P(f|e) = \frac{Count(f, e)}{Count(e)}$$
e.g.,
$$P(\text{daba una bofetada} \mid \text{slap}) = \frac{Count(\text{daba una bofetada}, \text{slap})}{Count(\text{slap})}$$

An Example Phrase Translation Table

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

- Phrase Translations for *den Vorschlag*

English	P(e|f)	**English**	P(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

Translate using a greedy, left-to-right decoding method

Today

Heute werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

$$\text{Score} = \underbrace{\log P(\text{Today} \mid \text{START})}_{\text{Language model}}$$

$$+ \underbrace{\log P(\text{Heute} \mid \text{Today})}_{\text{Phrase model}}$$

$$+ \underbrace{\log P(\text{1-1} \mid \text{1-1})}_{\text{Distortion model}}$$

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

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

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

$$\text{Score} = \underbrace{\log P(\text{we shall be} \mid \text{today})}_{\text{Language model}}$$

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

$$+ \underbrace{\log P(\text{2-3} \mid \text{2-4})}_{\text{Distortion model}}$$

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

Today we shall be debating

Heute werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

Today we shall be debating the reopening

Heute werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

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

Phrase-Based Systems: Formal Definitions

(following notation in Jurafsky and Martin, chapter 25)

- We'd like to translate a French string $\mathbf{f}$
- E is a sequence of l English phrases, $e_1, e_2, \ldots, e_l$.
  For example,

  $e_1 =$ Mary, $e_2 =$ did not, $e_3 =$ slap, $e_4 =$ the, $e_5 =$ green witch

  E defines a possible translation, in this case $e_1 e_2 \ldots e_5$ = *Mary did not slap the green witch.*
- F is a sequence of l foreign phrases, $f_1, f_2, \ldots, f_l$.
  For example,

  $f_1 =$ Maria, $f_2 =$ no, $f_3 =$ dio una bofetada, $f_4 =$ a la, $f_5 =$ bruja verde
- a_i for $i = 1 \ldots l$ is the position of the first word of f_i in $\mathbf{f}$. b_i for $i = 1 \ldots l$ is the position of the last word of f_i in $\mathbf{f}$.

Phrase-Based Systems: Formal Definitions

- We then have

$$Cost(E, F) = P(E) \prod_{i=1}^{l} P(f_i|e_i) d(a_i - b_{i-1})$$

- $P(E)$ is the language model score for the string defined by E
- $P(f_i|e_i)$ is the phrase-table probability for the i'th phrase pair
- $d(a_i - b_{i-1})$ is some probability/penalty for the distance between the i'th phrase and the $(i-1)$'th phrase. Usually, we define

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$$

  for some $\alpha < 1$.
- Note that this is *not* a coherent probability model

An Example

Position	**1**	**2**	**3**	**4**	**5**
English	Mary	did not	slap	the	green witch
Spanish	Maria	no	dio una bofetada	a la	bruja verde

In this case,

$$
\begin{aligned}
Cost(E,F) \quad = \quad & P_L(\text{Mary did not slap the green witch}) \times \\
& P(\text{Maria}|\text{Mary}) \times d(1) \times P(\text{no}|\text{did not}) \times d(1) \times \\
& P(\text{dio una bofetada}|\text{slap}) \times d(1) \times P(\text{a la}|\text{the}) \times d(1) \times \\
& P(\text{bruja verde}|\text{green witch}) \times d(1)
\end{aligned}
$$

P_L is the score from a language model

Another Example

Position	**1**	**2**	**3**	**4**	**5**	**6**
English	Mary	did not	slap	the	green	witch
Spanish	Maria	no	dio una bofetada	a la	verde	bruje

The original Spanish string was *Maria no dio una bofetada a la bruje verde*, so notice that the last two phrase pairs involve reordering

In this case,

$$
\begin{aligned}
Cost(E, F) \quad = \quad & P_L(\text{Mary did not slap the green witch}) \times \\
& P(\text{Maria}|\text{Mary}) \times d(1) \times P(\text{no}|\text{did not}) \times d(1) \times \\
& P(\text{dio una bofetada}|\text{slap}) \times d(1) \times P(\text{a la}|\text{the}) \times d(1) \times \\
& P(\text{verde}|\text{green}) \times d(2) \times P(\text{bruja}|\text{witch}) \times d(1)
\end{aligned}
$$

Overview

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

The Decoding Problem

- For a given foreign string $\mathbf{f}$, the decoding problem is to find

$$\arg\max_{(E,F)} Cost(E, F)$$

  where the $\arg\max$ is over all (E, F) pairs that are consistent with $\mathbf{f}$

- See Koehn tutorial, EACL 2006, slides 29–57
- See Jurafsky and Martin, Chapter 25, Figure 25.30
- See Jurafsky and Martin, Chapter 25, section 25.8