CSCI 5832
Natural Language Processing

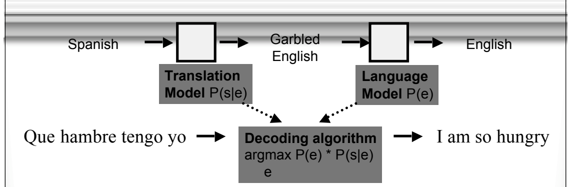
Jim Martin

Lecture 26

Today 4/29

More MT

Bayes Rule/Noisy Channel



Given a source sentence s, the decoder should consider many possible translations … and return the target string e that maximizes

P(e | s)

By Bayes Rule, we can also write this as:

P(e) x P(s | e) / P(s)

and maximize that instead. P(s) never changes while we compare different e's, so we can equivalently maximize this:

P(e) x P(s | e)

Three Sub-Problems of Statistical MT

- Language model
  - Given an English string e, assigns P(e) by formula
  - good English string -> high P(e)
  - random word sequence -> low P(e)
- Translation model
  - Given a pair of strings <f,e>, assigns P(f | e) by formula
  - <f,e> look like translations -> high P(f | e)
  - <f,e> don't look like translations -> low P(f | e)
- Decoding algorithm
  - Given a language model, a translation model, and a new sentence f … find translation e maximizing P(e) * P(f | e)

Parts List

- We need probabilities for
  - n (x|y) The probability that word y will yield x outputs in the translation… (fertility)
  - p The probability of a null insertion
  - t The actual word translation probability table
  - d(j|i) the probability that a word at position i will make an appearance at position j in the translation

Parts List

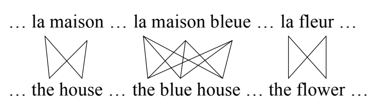
- Every one of these can be learned from a sentence aligned corpus…
  - Ie. A corpus where sentences are paired but nothing else is specified
- And the EM algorithm

Word Alignment

Inherent hidden structure revealed by EM training!

EM: Worked out example

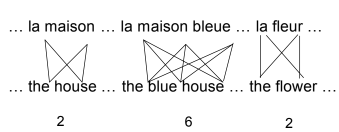
- Focus only on the word translation probs



- How many alignments are there for each of these sentence pairs?

EM: Worked out Knight example

- Focus only on the word translation probs



- How many alignments are there?

EM: Step 1

- Make up some numbers for the parameters of interest. In this case, just the word translation probabilities.

(la|the) (m|the) (b|the) (f|the)	(la|house) (m|house) (b|house)	(la|blue) (m|blue) (b|blue)	(la|flower) (f|flower)

5/2/08 10

Reminder

- P(la | the) is P(la aligned with the)/P(the)
- Which is Count(la aligned with the)/Count(the) in a word-aligned corpus.
- Which we don't have.

5/2/08 11

EM: Step 1

- Make up some numbers for the parameters of interest. In this case, just the translation probs.

(la|the) 1/4 (m|the) 1/4 (b|the) 1/4 (f|the) 1/4	(la|house) 1/3 (m|house) 1/3 (b|house) 1/3	(la|blue) 1/3 (m|blue) 1/3 (b|blue) 1/3	(la|flower) 1/2 (f|flower) 1/2

5/2/08 12

EM: Step 2

- Make some simple assumptions and produce some normalized alignment probabilities

la maison / the house	la maison / the house	la maison bleue / the blue house
¼ * 1/3 = 1/12	¼ * 1/3 = 1/12	¼ * 1/3 * 1/3 = 1/36

5/2/08 13

EM: Step 2 (normalize)

- Make some simple assumptions and produce some normalized alignment probabilities

la maison / the house	la maison / the house	la maison bleue / the blue house
¼ * 1/3 = 1/12	¼ * 1/3 = 1/12	¼ * 1/3 * 1/3 = 1/36
(1/12) / (2/12) = 1/2	(1/12) / (2/12) = 1/2	1/6 For each

5/2/08 14

EM: Step 3

- Now that we have the probability of each alignment we can go back and count the evidence in the alignments for each translation pair and prorate them based on the alignments they come from.

5/2/08 15

EM: Step 3

- Now that we have the probability of each alignment we can go back and count the evidence in the alignments for each translation pair and prorate them based on the alignments they come from. Huh?
- Let's just look at (la | the).
  - What evidence do we have?

EM: Step 3

- Evidence for (la|the)

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

1 2 1

EM: Step 3

- Evidence for (la|the)

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

1 2 1

1* ½ 2* 1/6 1* ½

EM: Step 3

- Evidence for (la|the)

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

1 2 1

1* ½ 2* 1/6 1* ½

8/6 = Discounted count for (la|the)

5/2/08 19

EM: Step 3

- Do that for the other (?|the) and normalize

(la| the) = 8/6	(la| the) = 8/6 / 18/6	(la| the) = .44
(m| the) = 5/6	(m| the) = 5/6 / 18/6	(m| the) = .27
(b| the) = 2/6	(b| the) = 2/6 /18/6	(b| the) = .11
(f | the) = 3/6	(f | the) = 3/6 / 18/6	(f | the) = .16

5/2/08 20

EM: Step 4

- Do that for all the words in the table
- Recompute the alignment probabilities using the new word translation probabilities
- Recompute the fractional counts
  - Normalize to get new word translation probs
- Iterate until done
- When you're done you can use the numbers to get the most probable alignment. From which you can read off all the parameters you need.

5/2/08 21

EM/Alignment

- Ok. I know this seems weird
- We need some parameters
  - which we don't have
- We can get them from a word-aligned corpus
  - which we don't have
- So we make up some parameters to get the alignment and then use that alignment to get the right numbers.

5/2/08 22

Parts List

- Given a sentence alignment we can induce a word alignment
- Given that word alignment we can get the p, t, d and n parameters we need for the model.
- Ie. We can argmax P(e|f) by max over P(f|e)*P(e)… and we can do that by iterating over some large space of possibilities.

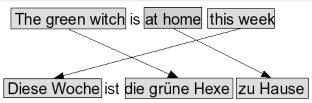
5/2/08 23

Break

- I will try to cover a subset of the material in Chapter 23 next week.

5/2/08 24

Intuition of phrase-based translation (Koehn et al. 2003)



- Generative story has three steps
  1) Group words into phrases
  2) Translate each phrase
  3) Move the phrases around

5/2/08 25

Generative story again

1) Group English source words into phrases e_1, e_2, …, e_n
2) Translate each English phrase e_i into a Spanish phrase f_j.
   - The probability of doing this is $\phi(f_j|e_i)$
3) Then (optionally) reorder each Spanish phrase
   - We do this with a **distortion** probability
   - A measure of distance between positions of a corresponding phrase in the 2 lgs.
   - "What is the probability that a phrase in position X in the English sentences moves to position Y in the Spanish sentence?"

5/2/08 26

Distortion probability

- The distortion probability is parameterized by
  - a_i-b_{i-1}
  - Where a_i is the start position of the foreign (Spanish) phrase generated by the *i*th English phrase e_i.
  - And b_{i-1} is the end position of the foreign (Spanish) phrase generated by the I-1th English phrase e_{i-1}.
- We'll call the distortion probability $d(a_i$-$b_{i-1})$.
- And we'll have a really stupid model:
  - $d(a_i-b_{i-1}) = \alpha^{|ai-bi-1|}$
  - Where α is some small constant.

5/2/08 27

Final translation model for phrase-based MT

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

$$P(F|E) = P(\text{Maria, Mary}) \times d(1) \times P(\text{no}|\text{did not}) \times d(1) \times P(\text{dió una bofetada}|\text{slap}) \times d(1) \times P(\text{a la}|\text{the}) \times d(1) \times P(\text{bruja verde}|\text{green witch}) \times d(1)$$

5/2/08 28

Phrase-based MT

- Language model P(E)
- Translation model P(F|E)
  - Model
  - How to train the model
- Decoder: finding the sentence E that is most probable

5/2/08 29

Training P(F|E)

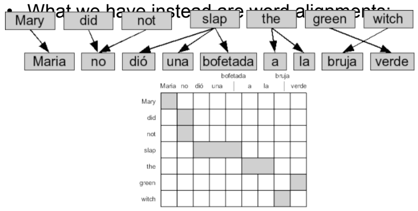
- What we mainly need to train is $\phi(f_j|e_i)$
- Suppose we had a large bilingual training corpus
  - A **bitext**
  - In which each English sentence is paired with a Spanish sentence
- And suppose we knew exactly which phrase in Spanish was the translation of which phrase in the English
- We call this **ent**
- If we had th ount-and-divide:

$$\phi(\bar{f},\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f},\bar{e})}$$

5/2/08 30

But we don't have phrase alignments

- What we have instead are word alignments:



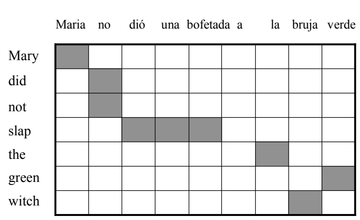
- (actually the word alignments we have are more restricted than this, as we'll see in two slides)

Getting phrase alignments

- To get phrase alignments:
  1) We first get word alignments
  2) Then we "symmetrize" the word alignments into phrase alignments

How to Learn the Phrase Translation Table?

- One method: "alignment templates" (Och et al, 1999)
- Start with word alignment, build phrases from that.

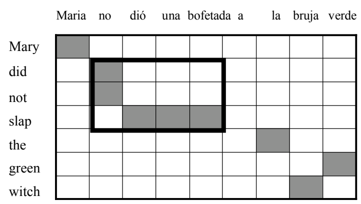


This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.

This is the best (or "Viterbi") alignment.

How to Learn the Phrase Translation Table?

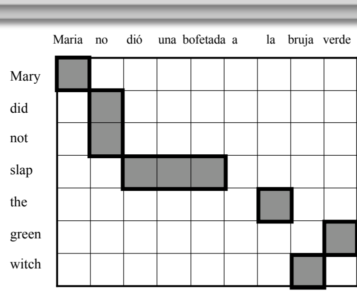
- One method: “alignment templates” (Och et al, 1999)
- Start with word alignment, build phrases from that.



This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.

This is the best (or “Viterbi”) alignment.

Word Alignment Induced Phrases

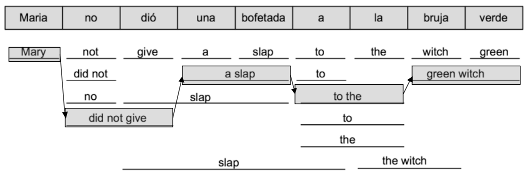


(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

Decoding

- Given the phrase alignments and translation probabilities how to decode?
- Basically stack decoding (ala A*; heuristic best first).
- Goal is to cover/account for all the foreign words with the best (highest prob) english sequence.

Decoding



5/2/08

37