From No Data to Sentence Pairs

- Really hard way: pay $$$
  - Suppose one billion words of parallel data were sufficient
  - At 20 cents/word, that’s $200 million
- Pretty hard way: Find it, and then earn it!
  - De-formatting
  - Remove strange characters
  - Character code conversion
  - Document alignment
  - Sentence alignment
- Tokenization (also called Segmentation)
- Easy way: Linguistic Data Consortium (LDC)

Centauri/Arcturan [Knight, 1997]:

- It’s Really Spanish/English
- Clients do not sell pharmaceuticals in Europe => Cliente no venden medicinas en Europa

From From No Data to Sentence Pairs:

- Ready-to-Use Online Bilingual Data
- Tokenization (or Segmentation)

Machine Translation:
Word alignment models

Christopher Manning
CS224N

[Based on slides by Kevin Knight, Dan Klein, Dan Jurafsky]
Sentence Alignment

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

A division of labor

- Use of Bayes Rule (“the noisy channel model”) allows a division of labor:
  - Job of the translation model P(E|S) is just to model how various Spanish words typically get translated into English (perhaps in a certain context)
    - P(E|S) doesn’t have to worry about language particular facts about English word order: that’s the job of P(E)
  - The job of the language model is to choose felicitous bags of words and to correctly order them for English
    - P(E) can do bag generation: putting a bag of words in order:
      - E.g., hungry I am so ...
      - I am so hungry
  - Both can be incomplete/sloppy

Statistical MT Systems

- Spanish/English Bilingual Text
- English Text
- Que hambre tengo yo
- I am so hungry

Statistical Analysis

- Statistical Analysis
- Statistical Analysis

Translation Model P(s|e)

Language Model P(e)

Decoding algorithm

argmax P(e) * P(s|e)
Word Alignment Examples: Grid

Word alignment examples: easy

Japan shaken by two new quakes
Le Japon secoué par deux noveaux séismes

Extra word appears in French: “spurious” word

Alignments: harder

“Zero fertility” word: not translated

And1 the2 program3 has4 been5 implemented6
Le1 programme2 a3 est4 été5 implementé6

One word translated as several words

Alignments: harder

The balance was the territory of the aboriginal people
Le reste appartenait aux autochtones

Several words translated as one

Alignments: hard

The1 poor3 don’t3 have4 any5 money6
Le1 pauvre2 a3 pas4 et5 avoir6 aucun7 argent8

Many to many

• A line group linking a minimal subset of words is called a ‘cept’ in the IBM work

Statistical Machine Translation

… la maison … la maison bleue … la fleur …
… the house … the blue house … the flower …

All word alignments equally likely
All P(french-word | english-word) equally likely
“la” and “the” observed to co-occur frequently, so \( P(\text{la} \mid \text{the}) \) is increased.

“house” co-occurs with both “la” and “maison”, but \( P(\text{maison} \mid \text{house}) \) can be raised without limit, to 1.0, while \( P(\text{la} \mid \text{house}) \) is limited because of “the” (pigeonhole principle)

settling down after another iteration

Hidden structure revealed by EM training!

That was IBM Model 1. For details, see later and:

- “A Statistical MT Tutorial Workbook” (Knight, 1999)
- Software: GIZA++

new French sentence

Possible English translations, to be rescored by language model

<table>
<thead>
<tr>
<th>IBM StatMT Translation Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>• IBM1 – lexical probabilities only</td>
</tr>
<tr>
<td>• IBM2 – lexicon plus absolute position</td>
</tr>
<tr>
<td>• HMM – lexicon plus relative position</td>
</tr>
<tr>
<td>• IBM3 – plus fertilities</td>
</tr>
<tr>
<td>• IBM4 – inverted relative position alignment</td>
</tr>
<tr>
<td>• IBM5 – non-deficient version of model 4</td>
</tr>
</tbody>
</table>

• All the models we discuss handle 0:1, 1:0, 1:1, 1:n alignments only

[Brown, et al. 93, Vogel, et al. 96]
IBM models 1,2,3,4,5

- Models for $P(F|E)$
- There is a set of English words and the extra English word NULL
- Each English word generates and places 0 or more French words
- Any remaining French words are deemed to have been produced by NULL

Model 1: Word alignment learning with Expectation-Maximization (EM)

- Start with $P(f|e)$ uniform, including $P(f|$null$)$
- For each sentence
  - For each French position $j$
    - Calculate posterior over English positions $P(a_j)$
      
      $$
      P(a_j = i \mid f, e) = \frac{P(f_j \mid e_i) \cdot \prod_{j=1}^{n} P(a_j = i)}{\sum_{i=1}^{m} \prod_{j=1}^{n} P(a_j = i)}
      $$
    - Increment count of word $f_j$ with word $e_i$
    - Renormalize counts to give probabilities
    - Iterate until convergence

IBM models 1,2,3,4,5

- In model 2, the placement of a word in the French depends on where it was in the English
  - Unlike Model 1, Model 2 captures the intuition that translations should usually “lie along the diagonal”.
  - The main focus of PA #2.

IBM models 1,2,3,4,5

- In model 3 we model how many French words and English word can produce, using a concept called fertility

  The proposal will not now be implemented

  Les propositions ne seront pas mises en application maintenant

Figure 32.3
Alignment example.

IBM Model 3, Brown et al., 1993

Generative approach:

- Mary did not slap the green witch
- Mary not slap slap slap the green witch
- Mary not slap slap NULL the green witch
- Maria no dío una boteada a la bruja verde

Probabilities can be learned from raw bilingual text.
IBM Model 3 (from Knight 1999)

- For each word $e_i$ in English sentence, choose a fertility $\Phi_i$. The choice of $\Phi_i$ depends only on $e_i$, not other words or $\Phi$’s.
- For each word $e_i$, generate $\Phi_i$ Spanish words. Choice of French word depends only on English word $e_i$, not English context or any Spanish words.
- Permute all the Spanish words. Each Spanish word gets assigned absolute target position slot (1,2,3, etc). Choice of Spanish word position dependent only on absolute position of English word generating it.

Model 3: P(S|E) training parameters

- What are the parameters for this model?
- Words: P(casa|house)
- Spurious words: P(a|null)
- Fertilities: n(1|house): prob that “house” will produce 1 Spanish word whenever ‘house’ appears.
- Distortions: d(5|2) prob. that English word in position 2 of English sentence generates French word in position 5 of French translation
  - Actually, distortions are d(5,2,4,6) where 4 is length of English sentence, 6 is Spanish length

Spurious words

- We could have n(3|NULL) (probability of being exactly 3 spurious words in a Spanish translation)
- But instead, of n(0|NULL), n(1|NULL) … n(25|NULL), have a single parameter $p_1$
- After assign fertilities to non-NUL English words we want to generate (say) z Spanish words.
- As we generate each of z words, we optionally toss in spurious Spanish word with probability $p_1$
- Probability of not tossing in spurious word $p_0=1-p_1$

Distortion probabilities for spurious words

- Can’t just have d(5|0,4,6), i.e. chance that NULL word will end up in position 5.
- Why? These are spurious words! Could occur anywhere!! Too hard to predict
- Instead,
  - Use normal-word distortion parameters to choose positions for normally-generated Spanish words
  - Put Null-generated words into empty slots left over
  - If three NULL-generated words, and three empty slots, then there are 3!, or six, ways for slotting them all in
  - We’ll assign a probability of 1/6 for each way

Real Model 3

- For each word $e_i$ in English sentence, choose fertility $\Phi_i$ with prob $n(\Phi_i|e_i)$
- Choose number $\Phi_0$ of spurious Spanish words to be generated from e0-NUL using $p_1$ and sum of fertilities from step 1
- Let m be sum of fertilities for all words including NULL
- For each $i=0,1,2,…,L$, $k=1,2,…,\Phi_i$:
  - Choose Spanish word $\tau_i$ with probability $t(\tau_i|e_i)$
  - For each $i=1,2,…,L$, $k=1,2,…,\Phi_i$:
    - Choose target Spanish position $\pi_k$ with prob $d(\pi_k|\tau_i,L,m)$
  - For each $k=1,2,…,\Phi_0$, choose position $\pi_k$ from $\Phi_0, \Phi_0+1$ remaining vacant positions in $1,2,…,m$ for total prob of 1/ $\Phi_0$!
  - Output Spanish sentence with words $\tau_k$ in positions $\pi_k$ ($0<\pi_k<1, 1<\pi_k<\Phi_i$)

Model 3 parameters

- n,t,p,d
- Again, if we had complete data of English strings and step-by-step rewritings into Spanish, we could:
  - Compute n(0|did) by locating every instance of “did”, and seeing how many words it translates to
  - t(maison|house) how many of all French words generated by “house” were “maison”
  - d(5)(2,4,6) out of all times some word2 was translated, how many times did it become words?
Since we don't have word-aligned data...

- We bootstrap alignments from incomplete data
- From a sentence-aligned bilingual corpus
  1) Assume some startup values for n,d, etc
  2) Use values for n,d, etc to use model 3 to work out chances of different possible alignments. Use these alignments to retrain n,d, etc
  3) Go to 2
- This is a more complicated case of the EM algorithm

---

IBM models 1,2,3,4,5

- In model 4 the placement of later French words produced by an English word depends on what happened to earlier French words generated by that same English word

---

Alignments: linguistics

On Tuesday Nov. 4, earthquakes rocked Japan once again

Des tremblements de terre ont à nouveau touché le Japon mardi 4 novembre

---

IBM models 1,2,3,4,5

- In model 5 they do non-deficient alignment. That is, you can’t put probability mass on impossible things.

---

Why all the models?

- We don't start with aligned text, so we have to get initial alignments from somewhere.
- Model 1 is words only, and is relatively easy and fast to train.
- We are working in a space with many local maxima, so output of model 1 can be a good place to start model 2. Etc.
- The sequence of models allows a better model to be found faster [the intuition is like deterministic annealing].

---

Alignments: linguistics

the green house

la maison verte

- There isn’t enough linguistics to explain this in the translation model … have to depend on the language model … that may be unrealistic … and may be harming our translation model