Machine Translation: Word alignment models

Christopher Manning
CS224N / Ling 284
[Based on slides by Kevin Knight, Dan Klein, Dan Jurafsky]
IBM StatMT Translation Models

- IBM1 – lexical probabilities only
- IBM2 – lexicon plus absolute position
- HMM – lexicon plus relative position
- IBM3 – plus fertilities
- IBM4 – inverted relative position alignment
- IBM5 – non-deficient version of model 4

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

[Brown, et.al. 93, Vogel, et.al. 96]
Word alignment examples: easy

Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes

Extra word appears in French: “spurious” word
Alignments: harder

“Zero fertility” word: not translated

And₁ the₂ program₃ has₄ been₅ implemented₆

Le₁ programme₂ a₃ été₄ mis₅ en₆ application₇

One word translated as several words
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 parameters

- \( P(F|E) = \prod_{(f,e)} P(f|e) \)
- \( P(f|e) = P(J|I) \sum_a P(f, a|e) \)
- \( P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i) = \prod_j [1/(l+1)] P(f_j|e_i) \)

And_1 the_2 program_3 has_4 been_5 implemented_6

Le_1 programme_2 a_3 été_4 mis_5 en_6 application_7
Model 1: Word alignment learning with Expectation-Maximization (EM)

- Start with $P(f_i|e^i)$ uniform, including $P(f_i|\text{NULL})$
- For each sentence
  - For each French position $j$
    - Calculate posterior over English positions $P(a_j|i)$
      
      \[
P(a_j = i \mid f,e) = \frac{P(f_j \mid e_i)}{\sum_{i'} P(f_j \mid e_{i'})}
      \]
    
    - Increment count of word $f_j$ with word $e_{a_j}$
      
      \[C(f_j|e_i) += P(a_j = i \mid f,e)\]
  
  - Renormalize counts to give probs $P(f^p \mid e^q) = \frac{C(f^p \mid e^q)}{\sum_{f^x} C(f^x \mid e^q)}$
  - 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 an 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.
Generative approach:

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 $n(3|\text{NULL})$ (probability of being exactly 3 spurious words in a Spanish translation).
- But instead, of $n(0|\text{NULL})$, $n(1|\text{NULL})$ … $n(25|\text{NULL})$, have a single parameter $p_1$.
- After assign fertilities to non-NULL 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 $e_0=NULL$ using $p1$ 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 $\tau_{ik}$ with probability $t(\tau_{ik}|e_i)$
- For each $i=1,2,...L$, $k=1,2,... \Phi_i$:
  - choose target Spanish position $\pi_{ik}$ with prob $d(\pi_{ik}|I,L,m)$
- For each $k=1,2,..., \Phi_0$ choose position $\pi_{0k}$ from $\Phi_0 - k + 1$ remaining vacant positions in $1,2,...m$ for total prob of $1/ \Phi_0$
- Output Spanish sentence with words $\tau_{ik}$ in positions $\pi_{ik}$ ($0<=l<=1,1<=k<=\Phi_i$)
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 word5?
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 $n,d,\Phi$, etc
  2) Use values for $n,d, \Phi$, etc to use model 3 to work out chances of different possible alignments. Use these alignments to retrain $n,d, \Phi$, 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
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.
Sample Translation Probabilities

| Translation Model | e     | f       | P(f | e) |
|-------------------|-------|---------|--------|
| national          | nationale | 0.47   |        |
|                   | national  | 0.42   |        |
|                   | nationaux | 0.05   |        |
|                   | nationales | 0.03   |        |
| the               | le      | 0.50   |        |
|                   | la      | 0.21   |        |
|                   | les     | 0.16   |        |
|                   | l’      | 0.09   |        |
|                   | ce      | 0.02   |        |
|                   | cette   | 0.01   |        |
| farmers           | agriculteurs | 0.44 |        |
|                   | les     | 0.42   |        |
|                   | cultivateurs | 0.05 |        |
|                   | producteurs | 0.02 |        |

[Brown et al 93]
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: impossible (in IBM)

The balance was the territory of the aboriginal people

Le reste appartenait aux autochtones

Several words translated as one
Alignments: impossible (in IBM)

- A minimal aligned subset of words is called a ‘cept’ in the IBM work; often a ‘bead’ or ‘(aligned) statistical phrase’ elsewhere.
Alignments: linguistics

the green house
\[\text{\underline{X}}\]
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
Getting Sentence Pair Data

• Really hard way: pay $$$
  – Suppose one billion words of parallel data were sufficient
  – At 5 cents/word, that’s $50 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)
Ready-to-Use Online Bilingual Data

(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

+ 1m-50m words for many language pairs, e.g., all EU languages)
Tokenization (or Segmentation)

- **English**
  - Input (some character stream):
    "There," said Bob.
  - Output (7 “tokens” or “words”):
    "There," said Bob.

- **Chinese**
  - Input (char stream):
    美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。
  - Output:
    美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。
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.
<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The old man is happy.</td>
<td>1. El viejo está feliz porque ha pescado muchos veces.</td>
</tr>
<tr>
<td>2. He has fished many times.</td>
<td>2. Su mujer habla con él.</td>
</tr>
<tr>
<td>3. His wife talks to him.</td>
<td>3. Los tiburones esperan.</td>
</tr>
<tr>
<td>4. The fish are jumping.</td>
<td></td>
</tr>
<tr>
<td>5. The sharks await.</td>
<td></td>
</tr>
</tbody>
</table>
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.

Done by similar Dynamic Programming or EM: see FSNLP ch. 13 for details
MT Evaluation
Illustrative translation results

- *la politique de la haine*.  
  *(Foreign Original)*
- *politics of hate*.  
  *(Reference Translation)*
- *the policy of the hatred*.  
  *(IBM4+N-grams+Stack)*

- *nous avons signé le protocole*.  
  *(Foreign Original)*
- *we did sign the memorandum of agreement*.  
  *(Reference Translation)*
- *we have signed the protocol*.  
  *(IBM4+N-grams+Stack)*

- *où était le plan solide?*  
  *(Foreign Original)*
- *but where was the solid plan?*  
  *(Reference Translation)*
- *where was the economic base?*  
  *(IBM4+N-grams+Stack)*

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and
MT Evaluation

• Manual (the best!?):
  – SSER (subjective sentence error rate)
  – Correct/Incorrect
  – **Adequacy and Fluency** (5 or 7 point scales)
  – Error categorization
  – **Comparative ranking of translations**

• Testing in an application that uses MT as one sub-component
  – Question answering from foreign language documents

• Automatic metric:
  – WER (word error rate) – why problematic?
  – **BLEU (Bilingual Evaluation Understudy)**
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:

The American international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail, which sends out; The threat will be able after public place and so on the airport to start the biochemistry attack, [?] highly alerts after the maintenance.

BLEU Evaluation Metric
(Papineni et al, ACL-2002)

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
  - An n-gram is an sequence of n words
  - Not allowed to match same portion of reference translation twice at a certain n-gram level (two MT words airport are only correct if two reference words airport; can’t cheat by typing out “the the the the the”)
  - Do count unigrams also in a bigram for unigram precision, etc.

- Brevity Penalty
  - Can’t just type out single word “the” (precision 1.0!)

- It was thought quite hard to “game” the system (i.e., to find a way to change machine output so that BLEU goes up, but quality doesn’t)
Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail, which sends out; The threat will be able after public place and so on the airport to start the biochemistry attack, [?] highly alerts after the maintenance.

BLEU Evaluation Metric
(Papineni et al, ACL-2002)

- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it’s precision-oriented
- BLEU4 formula
  (counts n-grams up to length 4)

\[
\exp(1.0 \times \log p_1 + 0.5 \times \log p_2 + 0.25 \times \log p_3 + 0.125 \times \log p_4 - \max(\text{words-in-reference} / \text{words-in-machine} - 1, 0))
\]

- \(p_1 = 1\)-gram precision
- \(P_2 = 2\)-gram precision
- \(P_3 = 3\)-gram precision
- \(P_4 = 4\)-gram precision

Note: only works at corpus level (zeroes kill it); there’s a smoothed variant for sentence-level
**BLEU in Action**

枪手被警方击毙。  

(Foreign Original)

<table>
<thead>
<tr>
<th>Translation</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>the gunman was shot to death by the police.</td>
<td>#1</td>
</tr>
<tr>
<td>the gunman was police kill.</td>
<td>#2</td>
</tr>
<tr>
<td>wounded police jaya of</td>
<td></td>
</tr>
<tr>
<td>the gunman was shot dead by the police.</td>
<td>#3</td>
</tr>
<tr>
<td>the gunman arrested by police kill.</td>
<td>#4</td>
</tr>
<tr>
<td>the gunmen were killed.</td>
<td>#5</td>
</tr>
<tr>
<td>the gunman was shot to death by the police.</td>
<td>#6</td>
</tr>
<tr>
<td>gunmen were killed by police ?SUB&gt;0 ?SUB&gt;0</td>
<td>#7</td>
</tr>
<tr>
<td>al by the police.</td>
<td>#8</td>
</tr>
<tr>
<td>the ringer is killed by the police.</td>
<td>#9</td>
</tr>
<tr>
<td>police killed the gunman.</td>
<td>#10</td>
</tr>
</tbody>
</table>

**4-gram match**

- **green** = 4-gram match (good!)
- **red** = word not matched (bad!)
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Reference translation 1:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an email from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places.

Machine translation:
The American [?] international airport and its [the] office all receives one calls self the sand Arab [rich] business [?] and so on electronic mail, which sends out; The threat will be able after public place and so on the airport to start the biochemistry attack, [?] highly alerts after the maintenance.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden, which threatens to launch a biochemical attack on such public places as airport. Guam authority has been on alert.

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other [rich] businessman from Saudi Arabia. They said there would be a biochemistry air raid to Guam Airport and other public places. Guam needs to be in high precaution about this matter.
Initial results showed that BLEU predicts human judgments well.

slide from G. Doddington (NIST)
Quiz question!

MT Hypothesis: *the gunman was shot dead by police.*

– Ref 1: The gunman was shot to death by the police.
– Ref 2: The cops shot the gunman dead.

• What is the:
  – Unigram precision?
  – Trigram precision?

Note: punctuation tokens are counted in calculation but not sentence boundary tokens
Automatic evaluation of MT

- People started optimizing their systems to maximize BLEU score
  - BLEU scores improved rapidly
  - The correlation between BLEU and human judgments of quality went way, way down
  - StatMT BLEU scores now approach those of human translations but their true quality remains far below human translations
- Coming up with automatic MT evaluations has become its own research field
  - There are many proposals: TER, METEOR, MaxSim, SEPIA, our own RTE-MT
  - TERpA is a representative good one that handles some word choice variation.
- MT research really requires some automatic metric to allow a rapid development and evaluation cycle.
Decoding for IBM Models

• Of all conceivable English word strings, find the one maximizing \( P(e) \times P(f \mid e) \)

• Decoding is NP hard
  – (Knight, 1999)

• Several search strategies are available
  – Usually a beam search where we keep multiple stacks for candidates covering the same number of source words

• Each potential English output is called a hypothesis.
Search for Best Translation

voulez – vous vous taire !
Search for Best Translation

voulez – vous vous taire !

you – you you quiet !
Search for Best Translation

voulez – vous vous taire !

quiet you – you you !
Search for Best Translation

voulez – vous vous taire !

you shut up !
Dynamic Programming Beam Search

Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

Dynamic Programming Beam Search

Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)