Machine Translation: Word alignment models

Christopher Manning

CS224N / Ling 284

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

IBM StatMT Translation Models

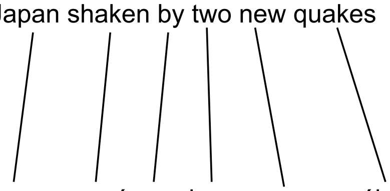
- IBM1 – lexical probabilities only
- IBM2 – lexicon plus absolute position
- HMM – lexicon plus relative position
- IBM3 – plus fertilities
- IBM4 – inverted relative position alignment
- IBM5 – non-deficient version of model 4

- 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

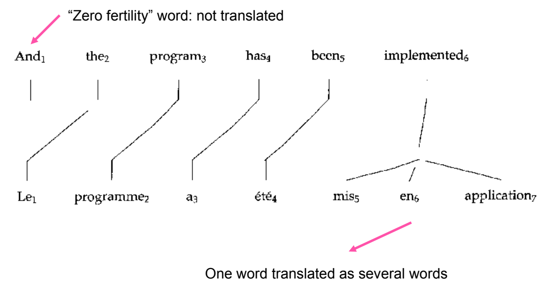
Japan shaken by two new quakes



Le Japon secoué par deux noveaux séismes

Extra word appears in French: "spurious" word

Alignments: harder

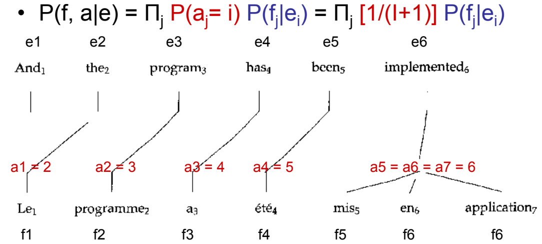


IBM models 1,2,3,4,5

- 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

- $P(F|E) = \Pi_{(f,e)} P(f|e)$
- $P(f|e) = P(J|I)\ \Sigma_a P(f, a|e)$
- $P(f, a|e) = \Pi_j P(a_j = i)\ P(f_j|e_i) = \Pi_j [1/(I+1)]\ P(f_j|e_i)$



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

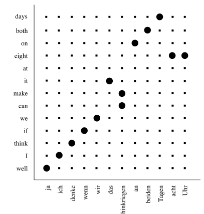
- Start with $P(f^j|e^i)$ uniform, including $P(f^j|\text{NULL})$
- For each sentence
  - For each French position j
    - Calculate posterior over English positions $P(a_j|i)$

$$P(a_j = i \mid f, e) = \frac{P(f_j \mid e_i)}{\sum_{i'} P(f_j \mid e_{i'})}$$

    - Increment count of word f_j with word e_{a_j}
      - $C(f_j|e_i) \mathrel{+}= P(a_j = i \mid f,e)$
- 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,<u>2</u>,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,<u>3</u>,4,5

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

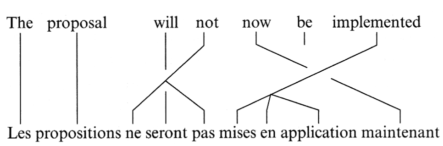
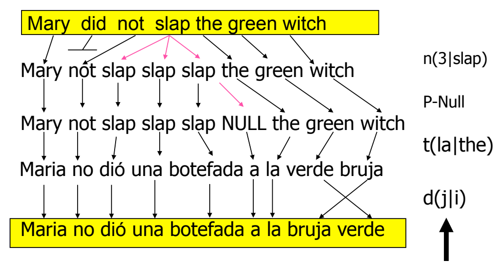


Figure 32.3
Alignment example.

IBM Model 3, Brown et al., 1993

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 Φ_i. The choice of Φ_i depends only on e_i, not other words or Φ's.
- For each word e_i, generate Φ_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 p1
- 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 p1
- Probability of not tossing in spurious word p0=1–p1

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 Φ_i with prob $n(\Phi_i| e_i)$
- Choose number Φ_0 of spurious Spanish words to be generated from e0=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,… Φ_I :
  - choose Spanish word τ_{ik} with probability $t(\tau_{ik}|e_i)$
- For each i=1,2,…L , k=1,2,… Φ_I :
  - choose target Spanish position π_{ik} with prob $d(\pi_{ik}|I,L,m)$
- For each k=1,2,…, Φ_0 choose position π_{0k} from Φ_0 -k+1 remaining vacant positions in 1,2,…m for total prob of 1/ Φ_0!
- Output Spanish sentence with words τ_{ik} in positions π_{ik} (0<=I<=1,1<=k<= Φ_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?

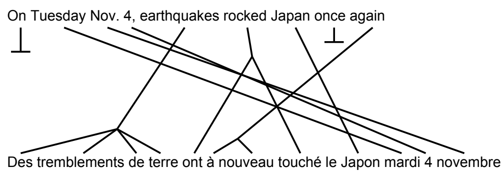
Since we don't have word-aligned data…

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

IBM models 1,2,3,<u>4</u>,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



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

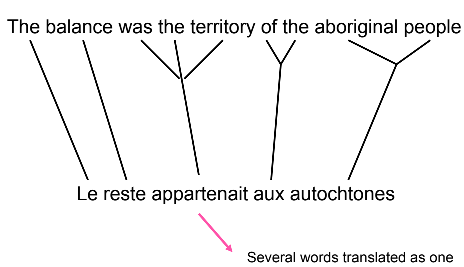
e	f	P(f | e)
national	nationale	0.47
	national	0.42
	nationaux	0.05
	nationales	0.03
the	le	0.50
	la	0.21
	les	0.16
	l'	0.09
	ce	0.02
	cette	0.01
farmers	agriculteurs	0.44
	les	0.42
	cultivateurs	0.05
	producteurs	0.02

[Brown et al 93]

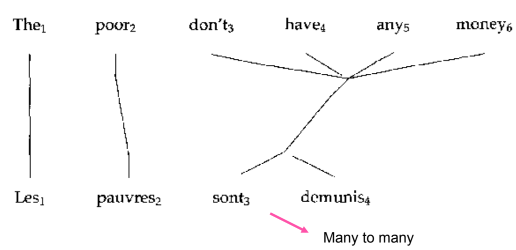
Why all the models?

- 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)



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

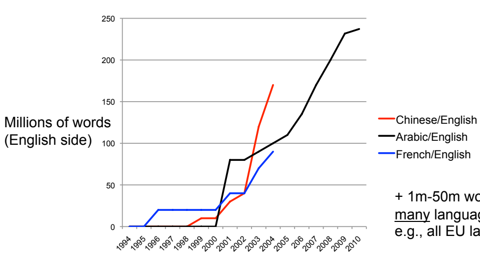
la maison verte

- 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



+ 1m-50m words for <u>many</u> language pairs, e.g., all EU languages)

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

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。

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.

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)**

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

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.

- 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)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

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 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 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Note: only works at corpus level (zeroes kill it); there's a smoothed variant for sentence-level

BLEU in Action

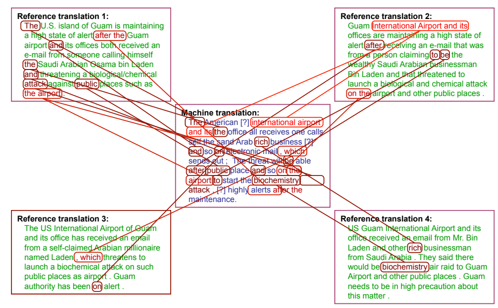
枪手被警方击毙。 **(Foreign Original)**

the gunman was shot to death by the police . **(Reference Translation)**

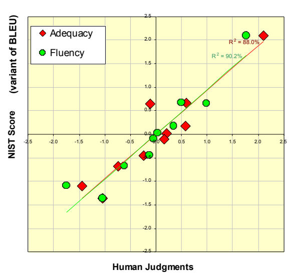
the gunman was police kill . **#1**
wounded police jaya of **#2**
the gunman was shot dead by the police . **#3**
the gunman arrested by police kill . **#4**
the gunmen were killed . **#5**
the gunman was shot to death by the police . **#6**
gunmen were killed by police ?SUB>0 ?SUB>0 **#7**
al by the police . **#8**
the ringer is killed by the police . **#9**
police killed the gunman . **#10**

green = 4-gram match (good!)
red = word not matched (bad!)

Multiple Reference Translations



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) x P(f | 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 !

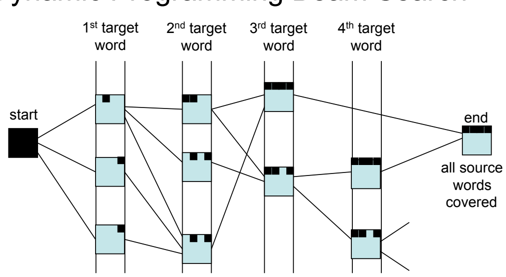
quiet you – you you !

Search for Best Translation

voulez – vous vous taire !

you shut up !

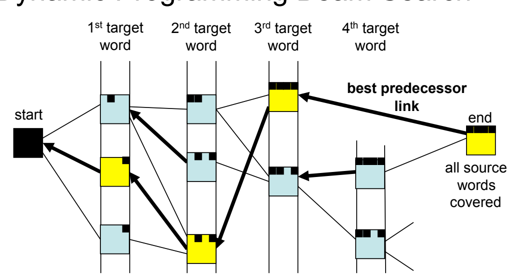
Dynamic Programming Beam Search



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)

[Jelinek, 1969; Brown et al, 1996 US Patent; (Och, Ueffing, and Ney, 2001]

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)

[Jelinek, 1969; Brown et al, 1996 US Patent; (Och, Ueffing, and Ney, 2001]