Machine Translation: Word alignment models

Christopher Manning
CS224N
[Based on slides by Kevin Knight, Dan Klein, Dan Jurafsky]

Centauri/Arcturan [Knight, 1997]: It's Really Spanish/English

Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

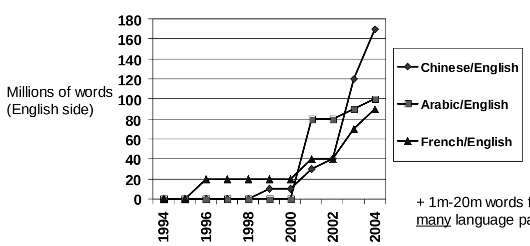
Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . zero fertility
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

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)

Ready-to-Use Online Bilingual Data



+ 1m-20m words for <u>many</u> language pairs

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

Tokenization (or Segmentation)

- English
  - Input (some byte stream):
    `"There," said Bob.`
  - Output (7 "tokens" or "words"):
    `" There , " said Bob .`
- Chinese
  - Input (byte 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.

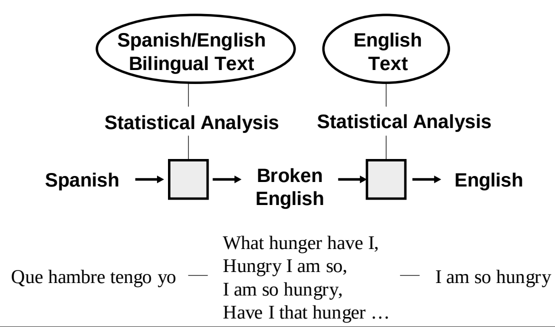
Sentence Alignment

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 Dynamic Programming: see FSNLP ch. 13 for details

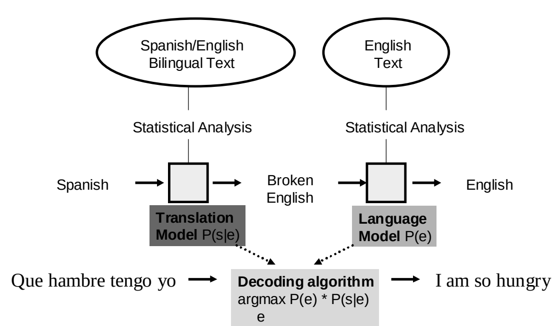
Statistical MT Systems



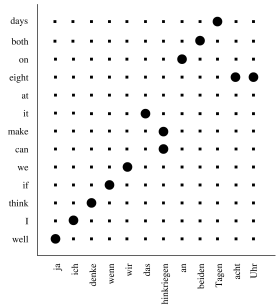
A division of labor

- 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

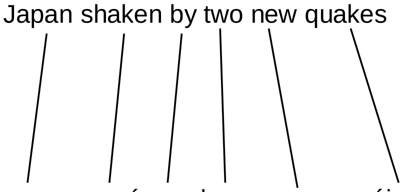


Word Alignment Examples: Grid



Word alignment examples: easy

Japan shaken by two new quakes

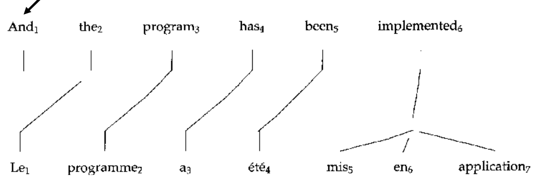


Le Japon secoué par deux noveaux séismes

Extra word appears in French: “spurious” word

Alignments: harder

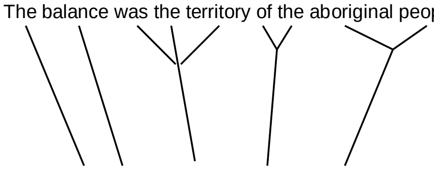
“Zero fertility” word: not translated



One word translated as several words

Alignments: harder

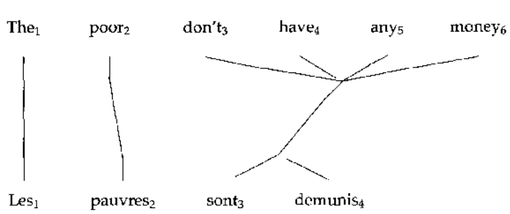
The balance was the territory of the aboriginal people



Le reste appartenait aux autochtones

Several words translated as one

Alignments: hard

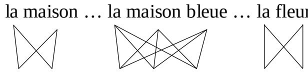


Many to many

- 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 …



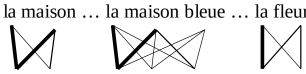
… the house … the blue house … the flower …

All word alignments equally likely

All P(french-word | english-word) equally likely

Statistical Machine Translation

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

"la" and "the" observed to co-occur frequently,
so P(la | the) is increased.

Statistical Machine Translation

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

"house" co-occurs with both "la" and "maison", but
P(maison | house) can be raised without limit, to 1.0,
while P(la | house) is limited because of "the"

(pigeonhole principle)

Statistical Machine Translation

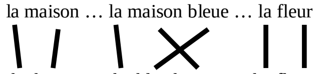
… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

settling down after another iteration

Word alignment learning with EM

… la maison … la maison bleue … la fleur …



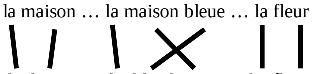
… the house … the blue house … the flower …

Hidden structure revealed by EM training!

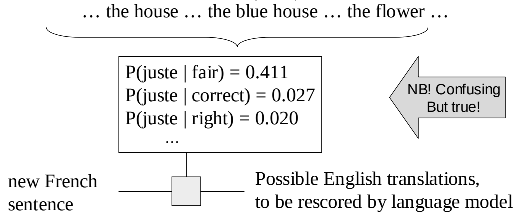
That was IBM Model 1. For details, see later and:
- "A Statistical MT Tutorial Workbook" (Knight, 1999).
- "The Mathematics of Statistical Machine Translation" (Brown et al, 1993)
- Software: GIZA++

Statistical Machine Translation

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …



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 handle 0:1, 1:0, 1:1, 1:n alignments *only*

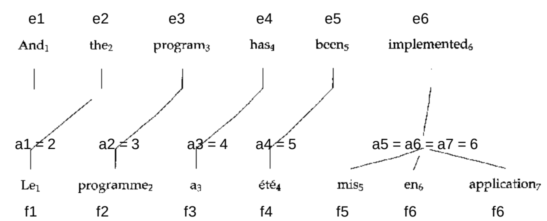
[Brown, et.al. 93, Vogel, et.al. 96]

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 parameters

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



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

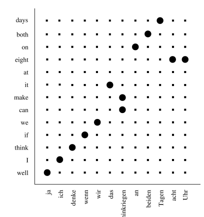
- Start with $P(f^j|e^i)$ uniform, including $P(f^j|null)$
- For each sentence
  - For each French position j
    - Calculate posterior over English positions $P(a_j|i)$

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

    - 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 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

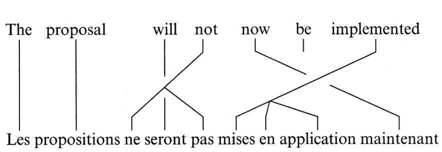
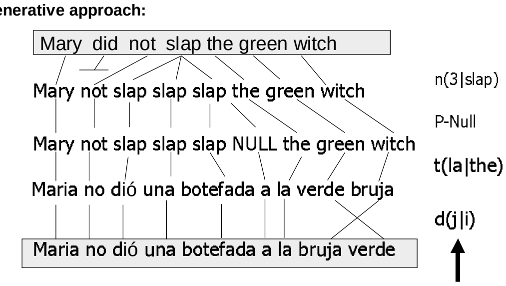


Figure 32.3
Alignment example.

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?

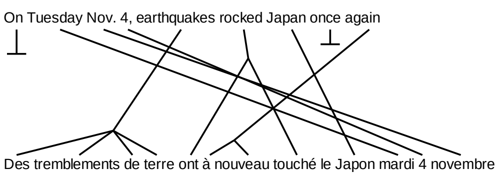
Since we don't have word-aligned data…

- 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



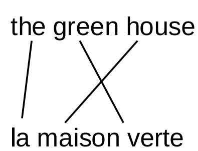
IBM models 1,2,3,4,<u>5</u>

- In model 5 they do <u>non-deficient alignment.</u> 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



- 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