CS 294-5: Statistical Natural Language Processing

Machine Translation
Lecture 10: 10/10/05

Machine Translation: Examples

Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

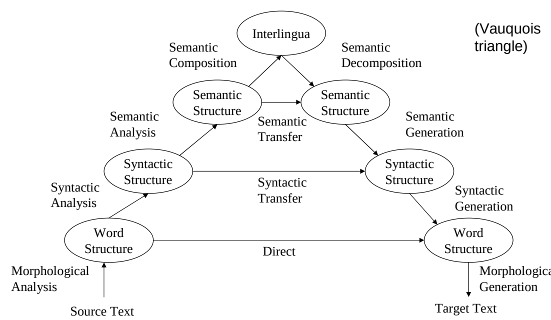
ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

History

- 1950's: Intensive research activity in MT
- 1960's: Direct word-for-word replacement
- 1966 (ALPAC): NRC Report on MT
  - Conclusion: MT no longer worthy of serious scientific investigation.
- 1966-1975: `Recovery period'
- 1975-1985: Resurgence (Europe, Japan)
- 1985-present: Gradual Resurgence (US)

http://ourworld.compuserve.com/homepages/WJHutchins/MTS-93.htm

Levels of Transfer



(Vauquois triangle)

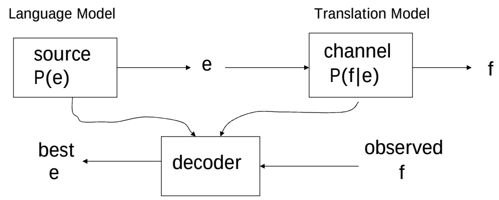
General Approaches

- Rule-based approaches
  - Expert system-like rewrite systems
  - Interlingua methods (analyze and generate)
  - Lexicons come from humans
  - Can be very fast, and can accumulate a lot of knowledge over time (e.g. Systran)
- Statistical approaches
  - Word-to-word translation
  - Phrase-based translation
  - Syntax-based translation (tree-to-tree, tree-to-string)
  - Trained on parallel corpora
  - Usually noisy-channel (at least in spirit)

The Coding View

- "One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' "
  - Warren Weaver (1955:18, quoting a letter he wrote in 1947)

MT System Components



$$\underset{e}{\operatorname{argmax}}\, P(e|f) = \underbrace{\underset{e}{\operatorname{argmax}}\, P(f|e)P(e)}$$

Finds an English translation which is both fluent and semantically faithful to the French source

Today

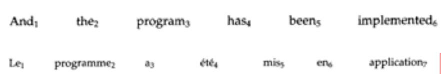
- The components of a simple MT system
  - You already know about the LM
  - Word-alignment based TMs
    - IBM models 1 and 2, HMM model
  - A simple decoder
- Next few classes
  - More complex word-level and phrase-level TMs
  - Tree-to-tree and tree-to-string TMs
  - More sophisticated decoders

A Word-Level TM?

- What might a model of P(f|e) look like?

$e = e_1 \dots e_I$

$f = f_1 \dots f_J$



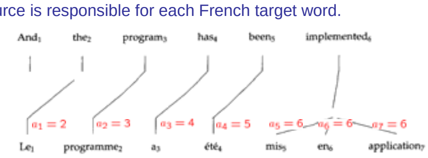
$$P(f|e) = \prod_j \underbrace{P(f_j|e_1 \dots e_I)}_{\text{How to estimate this?}}$$

What can go wrong here?

IBM Model 1 (Brown 93)

- Alignments: a hidden vector called an *alignment* specifies which English source is responsible for each French target word.

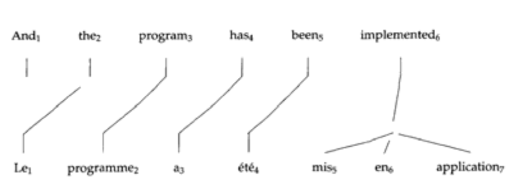
$a = a_1 \dots a_J$



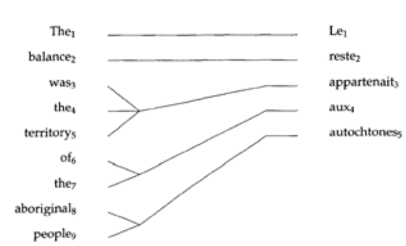
$$P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i)$$
$$= \prod_j \frac{1}{I+1} P(f_j|e_i)$$

$$P(f|e) = \sum_a P(f, a|e)$$

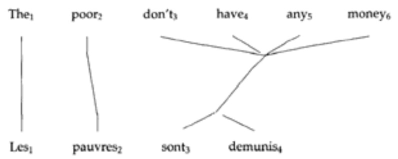
1-to-Many Alignments



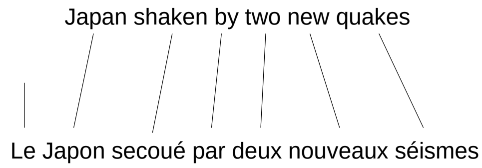
Many-to-1 Alignments



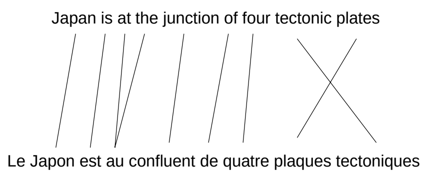
Many-to-Many Alignments



Monotonic Translation



Local Order Change



IBM Model 2

- Alignments tend to the diagonal (broadly at least)

$$P(f,a|e) = \prod_j P(a_j = i|j,I,J)P(f_j|e_i)$$

$$P(i - j\frac{I}{J})$$

$$\frac{1}{Z}e^{-\alpha(i-j\frac{I}{J})}$$

- Other schemes for biasing alignments towards the diagonal:
  - Relative alignment
  - Asymmetric distances
  - Learning a multinomial over distances

EM for Models 1/2

- Model 1 Parameters:
  Translation probabilities (1+2) $P(f_j|e_i)$
  Distortion parameters (2 only) $P(a_j = i|j,I,J)$
- 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

$$P(a_j = i|f,e) = \frac{P(a_j = i|j,I,J)P(f_j|e_i)}{\sum_{i'} P(a_j = i'|j,I,J)P(f_j|e'_i)}$$

    - (or just use best single alignment)
    - Increment count of word f_j with word e_i by these amounts
    - Also re-estimate distortion probabilities for model 2
- Iterate until convergence

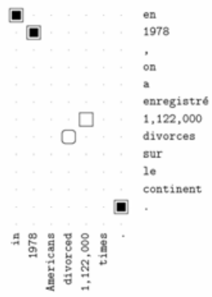
Evaluating TMs

- How do we measure TM quality?
  - Method 1: use in an end-to-end translation system
    - Hard to measure translation quality
    - Option: human judges
    - Option: reference translations (NIST, BLEU scores)
  - Method 2: measure quality of the alignments produced
    - Easy to measure
    - Hard to know what the gold alignments should be
    - May not correlate with translation quality (like perplexity in LMs)

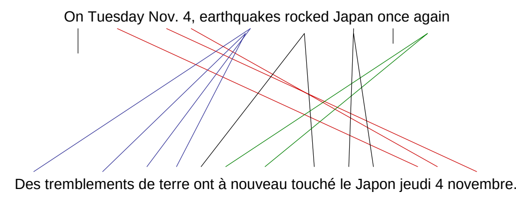
Alignment Error Rate

- Alignment Error Rate
  - □ = Sure
  - ○ = Possible
  - ■ = Predicted

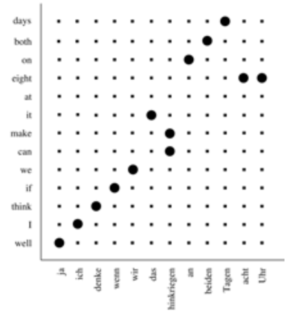
$$AER(A,S,P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$
$$= \left(1 - \frac{3+3}{3+4}\right) = \frac{1}{7}$$



Phrase Movement



Phrase Movement



The HMM Model

- Model 2 preferred global monotonicity
- We want local monotonicity:
  - Most jumps are small
- HMM model (Vogel 96)

$$P(f,a|e) = \prod_j P(a_j|a_{j-1}) P(f_j|e_i)$$
$$P(a_j - a_{j-1})$$

  - Re-estimate using the forward-backward algorithm
  - Handling nulls requires some care
- What are we still missing?

Some Results

- [Och and Ney 03]

Model	Training scheme	0.5K	8K	128K	1.47M
Dice		50.9	43.4	39.6	38.9
Dice+C		46.3	37.6	35.0	34.0
Model 1	1^5	40.6	33.6	28.6	25.9
Model 2	$1^5 2^5$	46.7	29.3	22.0	19.5
HMM	$1^5 H^5$	26.3	23.3	15.0	10.8
Model 3	$1^5 2^5 3^3$	43.6	27.5	20.5	18.0
	$1^5 H^5 3^3$	27.5	22.5	16.6	13.2
Model 4	$1^5 2^5 3^3 4^3$	41.7	25.1	17.3	14.1
	$1^5 H^5 3^3 4^3$	26.1	20.2	13.1	9.4
	$1^5 H^5 4^3$	26.3	21.8	13.3	9.3
Model 5	$1^5 H^5 4^3 5^3$	26.5	21.5	13.7	9.6
	$1^5 H^5 3^3 4^3 5^3$	26.5	20.4	13.4	9.4
Model 6	$1^5 H^5 4^3 6^3$	26.0	21.6	12.8	8.8
	$1^5 H^5 3^3 4^3 6^3$	25.9	20.3	12.5	8.7

Decoding

- In these word-to-word models
  - Finding best alignments is easy
  - Finding translations is hard (why?)

```
it is not clear .
|   \  |    \   \
|    \ +     \   \
|     \/ \    \   \
|     /\  \    \   \
CE NE EST PAS CLAIR  .
```

Bag “Generation” (Decoding)

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

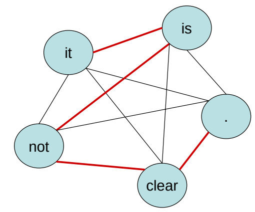
Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

In our organization research has two missions.
⇒ In our missions research organization has two.

Bag Generation is a TSP

- Imagine bag generation with a bigram LM
  - Words are nodes
  - Edge weights are P(w|w’)
  - Valid sentences are Hamiltonian paths
- Not the best news for word-based MT!



Decoding, Anyway

- Simplest possible decoder:
  - Enumerate sentences, score each with TM and LM
- Greedy decoding:
  - Assign each French word it’s most likely English translation
  - Operators:
    - Change a translation
    - Insert a word into the English (zero-fertile French)
    - Remove a word from the English (null-generated French)
    - Swap two adjacent English words
  - Do hill-climbing (or annealing)
- You should be able to build a model 1/2 translator now
- More on word alignment, decoding next class

WSD?

- Remember when we discussed WSD?
  - Word-based MT systems rarely have a WSD step
  - Why not?