

6.863J Natural Language Processing

Lecture 18: Machine translation 2

Robert C. Berwick
berwick@ai.mit.edu

The Menu Bar

- Administrivia:
 - Start w/ final projects, unless there are objections; No 'enrichment project'
- *Agenda*:
 - MT: the statistical approach
 - Formalize what we did last time
 - Divide & conquer: 4 steps
 - Noisy channel model
 - Language Model
 - Translation model
 - Scrambling & Fertility

Submenu

- The basic idea: moving from Language A to Language B
- The noisy channel model
- Use of Bayes' Rule
- Juggling words in translation – bag of words model; divide & translate
- Using n-grams – the Language Model
- The Translation Model
- Estimating parameters
- Searching for the best solution

6.863J/9.611J Lecture 18 Sp03

Translation B to A

- **13(B)iat lat pippat eneat hilat oloat at-yurp**
- Consult dictionary – 7 words can be looked up
- **iat lat pippat eneat hilat oloat at-yurp**

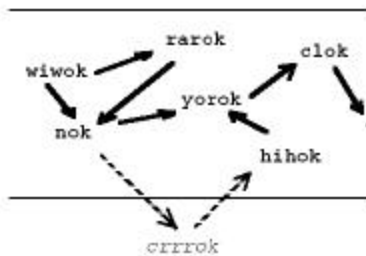
6.863J/9.611J Lecture 18 Sp03

The translation (answer sheet)

- iat lat pippat eneat hilat oloat at-yurp
- [you provide this]
- totat nnat forat arrat mat bat
- wat dat quat cat uskrat at-drubel

6.863J/9.611J Lecture 18 Sp03

Various possibilities



6.863J/9.611J Lecture 18 Sp03

The actual sentences

1. Garcia and associates.
Garcia y asociados.
2. Carlos Garcias has three associates.
Carlos Garcias tiene tres asociados.
3. His associates are not strong.
Sus asociados no son fuertes.
4. Garcia has a company also.
5. Its clients are angry.
6. The associates are also angry.
7. The clients and the associates are enemies.

6.863J/9.611J Lecture 18 Sp03

Statistical Machine Translation

- The fundamental idea of statistical MT is to let the computer learn how to do MT through studying the translation statistics from a bilingual corpus

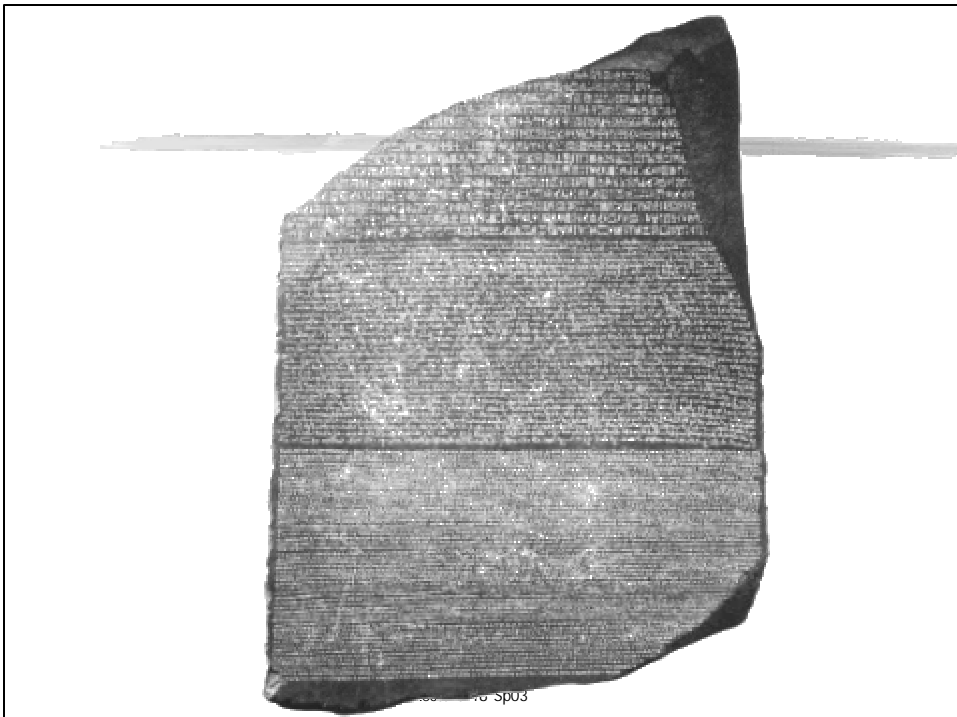
6.863J/9.611J Lecture 18 Sp03

What's the data? What are we doing?

- Pairs of sentences that are translations of one another are used
- Learn parameters for a probability model
- Source, Target pairs (S,T)

Find pr distribution over (S,T)

6.863J/9.611J Lecture 18 Sp03



Honourable Members of the Senate,
Members of the House of Commons,
Ladies and Gentlemen:
Honorables sénateurs et sénatrices,
Messdames et Messieurs les députés,
Mesdames et Messieurs,

My wife, Duna, and I were happy to welcome Her Majesty the Queen and the Duke of Edinburgh when they arrived in Canada last June and to be their hosts during their stay in the National Capital over Canada Day.
Ma femme et moi avons eu la joie d'accueillir Sa Majesté la Reine et le duc d'Édimbourg à leur arrivée au Canada en juin dernier et d'être leurs hôtes pendant leur séjour dans la région de la capitale nationale à l'occasion de la Fête du Canada.

As Governor General I have visited every province and territory, and I wish every Canadian could share that experience.

De plus, en tant que Gouverneur général, j'ai visité toutes les provinces ainsi que les territoires. C'est une expérience que je souhaite à tous les Canadiens.

6.863J/9.611J Lecture 18 Sp03

Example alignment

The proposal will not now be implemented
| | / \ / \
Les propositions ne seront pas mises en application maintenant

6.863J/9.611J Lecture 18 Sp03

Statistical Machine Translation



- Warren Weaver (4 March 1947): (letter to Weiner)

6.863J/9.611J Lecture 18 Sp03

Weaver



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'.

6.863J/9.611J Lecture 18 Sp03

Example of what Weaver had in mind?

The proposal will not now be implemented

Les propositions ne seront pas mises en application maintenant

6.863J/9.611J Lecture 18 Sp03

We have to estimate these

- Training model from parallel aligned sentences (where do we get parallel texts; how do we align?)
- How much data needed?

6.863J/9.611J Lecture 18 Sp03

So, how does English become French?

- Story 1. English gets converted to some sort of mental logic (predicate logic, or lexical-conceptual structures...), e.g., "I must not like ice-cream" into (obligatory (not (event like :obj ice-cream...))) blah blah blah

Rest of story: how this gets mapped to French

Call this story interlingua

6.863J/9.611J Lecture 18 Sp03

How does English become French?

- Story 2. English sentences gets syntactically parsed, into heads & modifiers, a binary tree say – phrases
- Then transformed into a French tree (a vine, say) – phrases swapped, english words replaced by french words.
- Call this syntactic transfer

6.863J/9.611J Lecture 18 Sp03

How does English become French?

- Story 3. Words in English sentence replaced by French words, which are scrambled
- Zany!
- Heh: this is IBM Model 3 story

6.863J/9.611J Lecture 18 Sp03

Like our alien system

- We will have two parts:
 1. A bi-lingual dictionary that will tell us what e words go w/ what f words
 2. A shake-n-bake idea of how the words might get scrambled around

We get these from cycling between alignment & word translations – re-estimation loop on which words linked with which other words

6.863J/9.611J Lecture 18 Sp03

IBM “Model 3”

- First to do this, late 80s: Brown et al, “The Mathematics of Statistical Machine Translation”, Computational Linguistics, 1990 (orig 1988 conference) – “Candide”
- We’ll follow that paper

6.863J/9.611J Lecture 18 Sp03

How to estimate?

- Formalize alignment
- Formalize dictionary in terms of $P(f|e)$
- Formalize shake-n-bake in terms of $P(e)$
- Formalize re-estimation in terms of the EM Algorithm
 - Give initial estimate (uniform), then up pr’s of some associations, lower others

6.863J/9.611J Lecture 18 Sp03

IBM toujours...

ISSUED: Apr. 23, 1996

FILED: Oct. 28, 1993

US PATENT NUMBER: 5510981

SERIAL NUMBER: 144913

INTL. CLASS (Ed. 6): G06F 17/28;

U.S. CLASS: 364-419.02; 364-419.08; 364-419.16;
381-043;

FIELD OF SEARCH: 364-419.02,419.08,419.16,200
MS File ; 381-43,51 ;

ABSTRACT: An apparatus for translating a series
of source words in a first language...

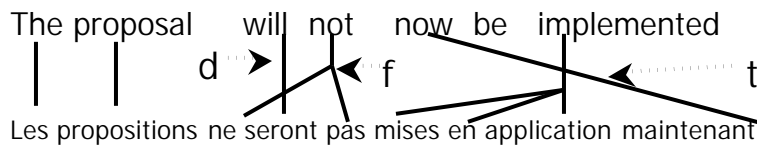
6.863J/9.611J Lecture 18 Sp03

The IBM series

- 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

6.863J/9.611J Lecture 18 Sp03

Example alignment



4 parameters for $P(f|e)$

1. Word translation, t Spurious word toss-in, p
2. Distortion (scrambling), d
3. Fertility, f

6.863J/9.611J Lecture 18 Sp03

4 Parameters

- Word Translation, $t(f_j | e_i)$
- Distortion, scrambling, $d(a_j | j) d(a_j | j m l)$
- Fertility, $\phi(n | e_i)$
- Spurious word appearance, p_i
- Q: how much space?
- Other:
- Class-based alignment 50 classes
- Nondeficient alignments (nulls)

6.863J/9.611J Lecture 18 Sp03

Bake-off – how to evaluate?

Tricky: not like speech (why?)

- Proposed measures...
 - Round-trip – ok, not always. E.g., “why in the world” → Sp → English → “why in the world” but
 - The Spanish is *porque en el mundo* (???)
- 1. Compare human & machines –
- 2. Categorize as same; equally good; different meaning; wrong; (=‘fluency’); ungrammatical (= ‘adequacy’)
- 3. Humans take test based on translated text...

6.863J/9.611J Lecture 18 Sp03

Bake-off Candide vs. Systran (Darpa) - 1995

	<u>Fluency</u>		<u>Adequacy</u>	
	1992	1993	1992	1993
Systran	47%	54%	69%	74%
Candide	51%	58%	58%	67%
Human		.83%		84%


6.863J/9.611J Lecture 18 Sp03



OK, now back to the game

6.863J/9.611J Lecture 18 Sp03

What's the data? What are we doing?



- Pairs of sentences that are translations of one another are used
- Learn parameters for a probability model
- Source, Target pairs (S,T)

Find pr distribution over (S,T)

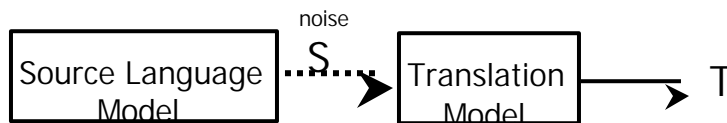
6.863J/9.611J Lecture 18 Sp03

How does English become French?

- Story 3. Words in English sentence replaced by French words, which are scrambled
- Zany!
- Heh: this is IBM Model 3 story

6.863J/9.611J Lecture 18 Sp03

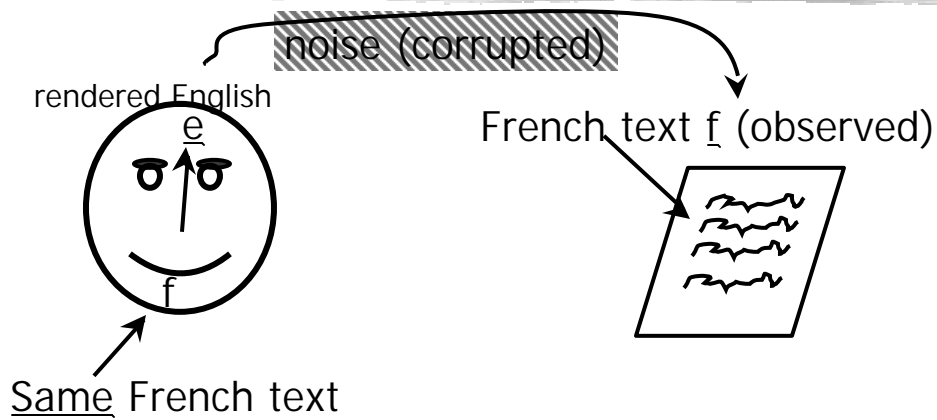
Noisy channel model to the rescue



Find pr distribution over (S,T)

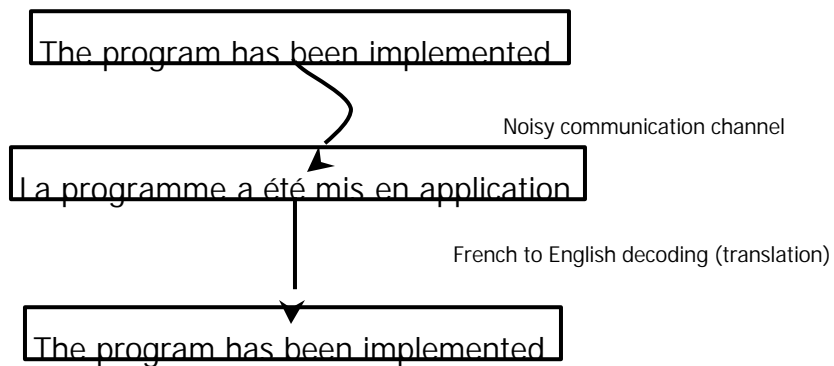
6.863J/9.611J Lecture 18 Sp03

'George Bush' model of translation (noisy channel)



6.863J/9.611J Lecture 18 Sp03

Noisy channel model



6.863J/9.611J Lecture 18 Sp03

George Bush Model of translation

- Somewhere in the noisy channel between (French) speaker's brain and mouth, the sentence E got "corrupted" to its French translation F
- Crazy?
- No stranger than the view that an English sentence gets corrupted into an acoustic signal in passing from the person's brain to his mouth

6.863J/9.611J Lecture 18 Sp03

We need to estimate $p(r)$'s

- Need to know:
 - What people say in English (source)
 - How E gets turned into French (channel)

6.863J/9.611J Lecture 18 Sp03

How do we do this?

- English sentence \underline{e} , French sentence \underline{f}
- An English sentence \underline{e} can be translated to *any* French sentence \underline{f}
- But some translations are more equal than others... (more likely)
- We use probabilities to measure this!

6.863J/9.611J Lecture 18 Sp03

OK, to begin

- $P(\underline{e})$ = pr of producing some English sentence \underline{e} (e.g., "cheese-eating surrender monkeys")
- $P(\underline{e}|\underline{f})$ = pr on encountering \underline{f} , will produce \underline{e}
- E.g., \underline{f} = "Lincoln était un bon avocat"
 \underline{e} = "cheese-eating surrender monkeys"

$P(\underline{e}|\underline{f})$ Not bloody likely!

Note: in general, \underline{e} and \underline{f} can be anything, not just words...

6.863J/9.611J Lecture 18 Sp03

In our case...

- What we see is \underline{f}
- We want to find is \underline{e} (the most likely translation \underline{e})
- In other words, compute:

$$\operatorname{argmax}_{\underline{e}} P(\underline{e}|\underline{f})$$

What's wrong with this plan???

Why can't we just figure out $P(\underline{e}|\underline{f})$?

6.863J/9.611J Lecture 18 Sp03

What's wrong with just $P(\underline{e}|\underline{f})$?

- We are extending from words:
'sol' \leftrightarrow 'sun'
'to pull the wool over someone's eyes' \leftrightarrow 'deitar
areia para os olhos de alguém'

To sentences:

cheese eating surrender monkeys

fromage mangeant des singes de reddition

- What's wrong with this plan?
- Probably won't see a sentence match more than once, probably not at all!

6.863J/9.611J Lecture 18 Sp03

So,

- If we compute $P(e|f)$ directly, we had better be good – but there's no data....
- $P(e|f)$ directly makes sense only if words in french are translations of words in english...
- A nice model for mutating bad french into bad english
- Note that it also gives no guarantee on the well-formedness of e !
- But: We can use Bayes' Rule to get good translations even if the pr estimates are crummy!

6.863J/9.611J Lecture 18 Sp03

Decoupling by Bayes' Rule

- $$P(e|f) = \frac{P(e) \times P(f|e)}{P(f)}$$
- We want to *maximize* this quantity $P(e|f)$, so we can simply maximize:
$$P(e) \times P(f|e)$$

Q: What happened to $P(f)$?

6.863J/9.611J Lecture 18 Sp03

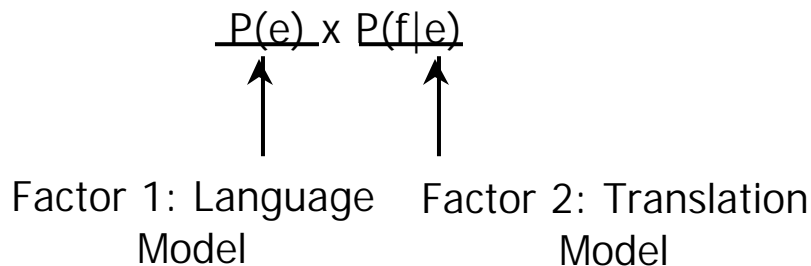
So, if this works...

- Our job has been reduced to three things:
 1. Estimate the parameters for $P(e)$
 2. Estimate the parameters for $P(f|e)$
 3. Search the product space to maximize

Let's see what each of the pr quantities mean, and what role they play

6.863J/9.611J Lecture 18 Sp03

Let's see what this means



6.863J/9.611J Lecture 18 Sp03

Factor 1: $P(e)$, language model

- $P(e)$ says that 'John ate ice-cream' has high pr, but 'ate ice-cream John' has lower pr
- Indeed, ungrammatical sentence – pr 0 (but this could be hard to figure out)
- $P(e)$ is really lowering pr of ungrammatical S's
- So really, this is like our alien language case (what part?)
- Several possible collections of words ('bags') – pick most probable sequence

6.863J/9.611J Lecture 18 Sp03

Language model $P(e)$

- So in fact, we have to choose between many grammatical sentences, e.g.,
- Which of these is better translation?
Fred viewed Sting in the television
Fred saw Sting on TV
- So, we are back to N-grams again!
- This will let us model word order

6.863J/9.611J Lecture 18 Sp03

Language model & N-grams

- In general – next word could depend on all preceding context
- But there are too many parameters to estimate, so we use just bigrams or trigrams
- To find pr for a whole sentence, multiply conditional pr's of the n-grams it contains

6.863J/9.611J Lecture 18 Sp03

Language and N-gram example

- $P(\text{I found riches in my backyard}) =$
 $P(\text{I} \mid \text{start-of-sentence}) \times$
 $P(\text{found} \mid \text{I}) \times$
 $P(\text{riches} \mid \text{found}) \times$
 $P(\text{in} \mid \text{riches}) \times$
 $P(\text{my} \mid \text{in}) \times$
 $P(\text{backyard} \mid \text{my}) \times$
 $P(\text{end-of-sentence} \mid \text{backyard})$
- Note how this will be higher pr than “my I in riches backyard found” – just as we want

6.863J/9.611J Lecture 18 Sp03

Language model $P(e)$

- So, if this does word order...
- Question: restore order for
actual the hashing is since not collision-free
usually the is less perfectly the of somewhat
capacity table
- Question: what knowledge are you using?
- Amazingly, this alone can be used to restore
scrambled English sentences (63-80%)
- Question: restore order for
loves John Mary

6.863J/9.611J Lecture 18 Sp03

A final use of $P(e)$

- Choose between alternative translations
I found the riches in my backyard
I found the riches on my backyard
In Spanish, 'in' and 'on' correspond to 'en'
We can use trigram counts to tell the
difference and select the higher pr one...

6.863J/9.611J Lecture 18 Sp03

Problemes? Problemos?

6.863J/9.611J Lecture 18 Sp03

The estimation catch

- Where do these pr numbers come from?
Which has higher pr:
‘I hate ice-cream’, or ‘I like ice-cream’?
use Google!
- What happens when $P(y | x)$ is zero? (not observed in training)
- The whole product would be zero
- Bad, because then “I like cheese eating monkeys” = same pr as “like I monkeys cheese eating”

6.863J/9.611J Lecture 18 Sp03

Estimation

- Acute issue for trigrams - 'found riches in' probably never seen
- Solution: smoothing (see textbook & next lecture - large literature on this)

6.863J/9.611J Lecture 18 Sp03

Problems...

- Won't always work – consider
Underline it
Emphasize it
- English might prefer the first, but must look at Spanish – 'subrayar' translates as both, but mostly as 'underline'; Spanish uses 'acentuar' for emphasis
- But this means we need to look at connections between 2 languages, ie, $P(f|e)$ that bridge between them, not just in English... that is the job of the Translation Model

6.863J/9.611J Lecture 18 Sp03

Language model & translation model

- Factoring knowledge out this way makes estimation easier
- Since $P(e)$ takes care of word order, the translation model, $P(f|e)$ doesn't have to worry about this – it can give crummy pr estimates, it can be sloppy, as long as it has the right words
- But as we've seen, $P(e)$ can't do all the work for this...

6.863J/9.611J Lecture 18 Sp03

Translation model $P(f|e)$

- What was it in our alien example?
- It was the bilingual dictionary
- What does it do?
- Ensure the words of \underline{e} express the ideas of \underline{f}
- So, responsibility is divided between $P(e)$ and $P(f|e)$

An example (Spanish)

6.863J/9.611J Lecture 18 Sp03

Spanish-English

- $P(e) \times P(s|e)$ to get $P(e|s)$ – assume ‘subrayar’ input...
- 1. Underline it.
 $P(\text{underline}) \times$
 $P(\text{it} \mid \text{underline}) \times$
 $P(\text{subrayar} \mid \text{underline})$
- 2. Emphasize it.
 $P(\text{emphasize}) \times$
 $P(\text{it} \mid \text{emphasize}) \times$
 $P(\text{subrayar} \mid \text{emphasize})$
- (1) is preferred because ‘underline’ is common and it is usually translated as ‘subrayar’

Language model can give crummy pr's

- As long as it has the right words
- This gives some measure of robustness
- Example – all of these could have roughly the same pr, despite being lousy translations...

Lousy translations

- $P(\text{Yo no comprendo} | \text{I don't understand})$
 - $P(\text{Comprendo yo no} | \text{Don't understand I})$
 - $P(\text{No yo comprendo} | \text{I don't understand})$
 - $P(\text{Comprendo yo no} | \text{I don't understand})$
 - $P(\text{Yo no comprendo} | \text{I understand don't})$
 - $P(\text{Yo no comprendo} | \text{Understand I don't})$
- In fact, this gives a first-cut way to estimate $P(f|e)$! Do you see how?

6.863J/9.611J Lecture 18 Sp03

Cheap and dirty $P(s|e)$

- Just product of individual translation probabilities!
- $P(\text{yo no comprendo} | \text{I don't understand}) =$
 - $P(\text{yo} | \text{I}) \times$
 - $P(\text{yo} | \text{don't}) \times$
 - $P(\text{yo} | \text{understand}) \times$
 - $P(\text{no} | \text{I}) \times$
 - $P(\text{no} | \text{don't}) \times$
 - $P(\text{no} | \text{understand}) \times$
 - $P(\text{comprendo} | \text{I}) \times$
 - $P(\text{comprendo} | \text{don't}) \times$
 - $P(\text{comprendo} | \text{understand})$


6.863J/9.611J Lecture 18 Sp03

Any problemas?

- Si...
- $P(\text{comprendo} \mid \text{understand})$ will be too low
- $P(\text{la} \mid \text{understand})$ will be too high – just because la is frequent in Spanish
- Use our method for alien languages!
- If we have previously established a link between 'the' and 'la', then we should boost 'comprendo'
- That will reduce translation of 'don't' as 'comprendo' because that will co-occur only when 'understand' is already nearby
- $P(\text{comprendo} \mid \text{understand})$ should work out close to 1, and $P(\text{la} \mid \text{the})$ say 0.4, rest going to $P(\text{el} \mid \text{the})$...

6.863J/9.611J Lecture 18 Sp03

In other words...

- Use alignments to assist with $P(e)$, $P(f|e)$
- 
- Use $P(e)$ to assist with alignments

problème de poulet et d'oeufs et
problema del pollo y del huevo y
problema dell'uovo e del pollo e
Huhn- und Eiproblem und

6.863J/9.611J Lecture 18 Sp03

Problemos

- Alignments help us get the translations
- Translations help us get the alignments...
- Where do we start???

6.863J/9.611J Lecture 18 Sp03

For the example...

- Yo no comprendo / I don't understand
- There are six possible alignments (for now...assuming no null maps, etc)
- All possible word combinations...
- This is just like our alien language case

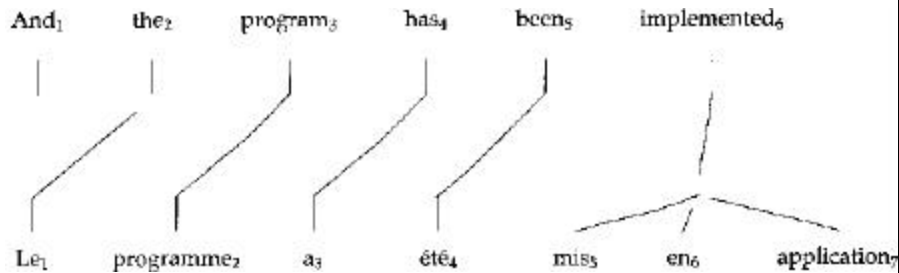
6.863J/9.611J Lecture 18 Sp03

Like this...

$$\begin{aligned} & \bullet P(\text{yo no comprendo} \mid \text{I don't understand}) = \\ & \quad P(\text{Alignment1}) \times P(\text{yo} \mid \text{I}) \times P(\text{no} \mid \text{don't}) \times \\ & \quad \quad \quad P(\text{comprendo} \mid \text{understand}) \\ & + P(\text{Alignment2}) \times P(\text{yo} \mid \text{don't}) \times P(\text{no} \mid \text{I}) \times \\ & \quad \quad \quad P(\text{comprendo} \mid \text{understand}) \\ & + P(\text{Alignment3}) \times P(\text{yo} \mid \text{understand}) \times P(\text{no} \mid \text{I}) \times \\ & \quad \quad \quad \dots \\ & + P(\text{Alignment6}) \times P(\text{yo} \mid \text{understand}) \times P(\text{no} \mid \text{don't}) \\ & \quad \quad \quad P(\text{comprendo} \mid \text{I}) \end{aligned}$$

6.863J/9.611J Lecture 18 Sp03

Example



6.863J/9.611J Lecture 18 Sp03

More problems...

- Can't assume direct word-for-word translation – some sentence pairs are different lengths
- An English word might correspond to more than one French word, or none at all
- So we model this -

6.863J/9.611J Lecture 18 Sp03

Procrustean bed

- For each word e_i in the sentence, $i = 1, 2, \dots, l$ we choose a fertility $\phi(e_i)$, equal to $0, 1, 2, \dots$
- This value is dependent solely on the English word, not other words or the sentence, or the other fertilities
- For each word e_i we generate $\phi(e_i)$ French words – not dependent on English context
- The French words are permuted ('distorted') – assigned a position slot (this is the scrambling phase)
- Call this a distortion parameter $d(i|j)$

6.863J/9.611J Lecture 18 Sp03

Summary of components

- The language model: $P(e)$
- The translation model for $P(f|e)$
 - Word translation t
 - Distortion (scrambling) d
 - Fertility ϕ
- (really evil): (for next time)
- Maximize (A^* search) through product space

6.863J/9.611J Lecture 18 Sp03

What's the input data? Aligned S's

The high turnover rate was largely due to an increase in the sales volume.

Employment and investment levels have also climbed.

Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988.

Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.

La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.

L'emploi et les investissements ont également augmenté.

La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autre les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

6.863J/9.611J Lecture 18 Sp03

What's the data?

- Hansard – Canadian Parliament since early 1800s, dual language
- 100M words, > 1M sentences
- Each on separate tape (!)
- Corresponding sentences not marked, paragraphs missing
- We want this – how to we get to it?

6.863J/9.611J Lecture 18 Sp03

Issues with alignment

- Clues include...
 - French sentences usually in same order as English sentences (but word order diff)
 - Short French sentences ↔ short English sentences, and v.v.
 - Corresponding French and English sentences often contain many of the same character sequences (why?)

6.863J/9.611J Lecture 18 Sp03

Weaver knew...

Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed.

But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers.

Thus it may be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication—the real but as yet undiscovered universal language—and—then re-emerge by whatever particular route is convenient.

6.863J/9.611J Lecture 18 Sp03

Lost in the translation



- Proust: "A la recherche du temps perdu:
"Depuis longtemps..."
- Translation: "For a long time I would go
to bed early..."
- Last word in book: depuis