CS 4650/7650
Machine Translation 1

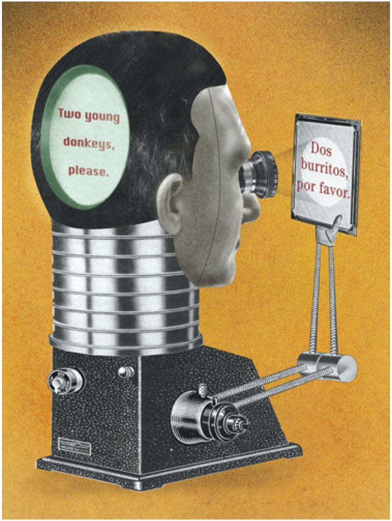
Jacob Eisenstein[1]

November 30, 2015

[1] with slides from David Chiang, Chris Dyer, and Phillip Koehn

Overview of machine translation

Why is MT hard?



Word order

- English: subject-verb-object
- Japanese: subject-object-verb
- Examples
  - English: IBM bought Lotus
    Japanese: IBM Lotus bought

Word order

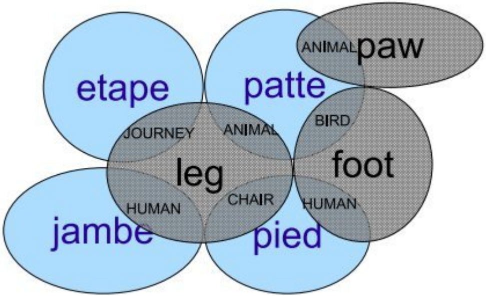
- English: subject-verb-object
- Japanese: subject-object-verb
- Examples
  - English: IBM bought Lotus
    Japanese: IBM Lotus bought
- French: subject-verb-object... except for pronouns
  - English: I will buy it
    French: Je vais l'acheter (I will **it** buy)
  - English: I bought it
    French: Je l'ai acheté (I **it** have bought)
- How many orderings are there?

Word sense ambiguity

- We've already talked about how bill translates as pico (bird) or cuenta (cost).

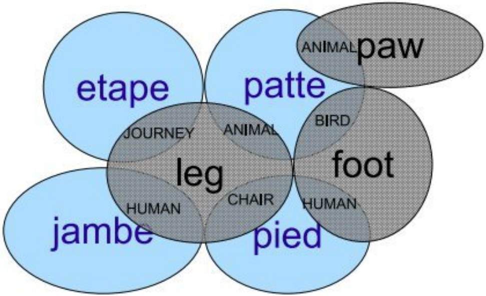
Word sense ambiguity

- We've already talked about how bill translates as pico (bird) or cuenta (cost).
- Legs and feet in English and French:



Word sense ambiguity

- We've already talked about how bill translates as pico (bird) or cuenta (cost).
- Legs and feet in English and French:



- **Lexical gaps** are when a language lacks a word for a concept.
  - My favorite English lexical gap: schadenfreude, a German word for "pleasure taken from the misfortune of others."

Pronouns

Pronoun morphology can convey different information in each language:

- English possessive pronouns take the gender of the owner: Marie rides **her** bike
- French possessive pronouns take the gender of the object: Marie monte sur **son** vélo

Pronouns

Pronoun morphology can convey different information in each language:

- English possessive pronouns take the gender of the owner:
  Marie rides **her** bike
- French possessive pronouns take the gender of the object:
  Marie monte sur **son** vélo
- Translating into English requires:
  - Anaphora resolution: son to Marie.
  - Gender determination: Marie is female.
  - Google Translate's treatment of this one is interesting.
    Compare Marie monte son vélo versus Marie vend son vélo.

Pronouns

Many languages (Spanish, Japanese, Chinese, Turkish, Arabic etc.) can drop pronouns.

- In Spanish, you can recover the pronoun from verb inflection:
  Viv**imos** en Atlanta → **We** live in Atlanta
- Again, discourse context is often crucial:
  Vive en Atlanta → **She/he/it** lives in Atlanta

Chinese example:

这块蛋糕很美味。谁烤的?

Zhè kuài dàngāo hěn měiwèi. Shéi kǎo de?

This piece cake very beautiful taste. Who bake?

"This cake is very tasty. Who baked **it**?"

Tense and case

- Spanish has two past tenses.
  - The **preterite** tense is for events with a definite time, e.g.
    I biked to work this morning
  - The **imperfect** is for events with indefinite times, e.g.
    I biked to work all last summer
  - To translate English to Spanish, we must pick the right tense.
    This seems to require some deeper semantic understanding.

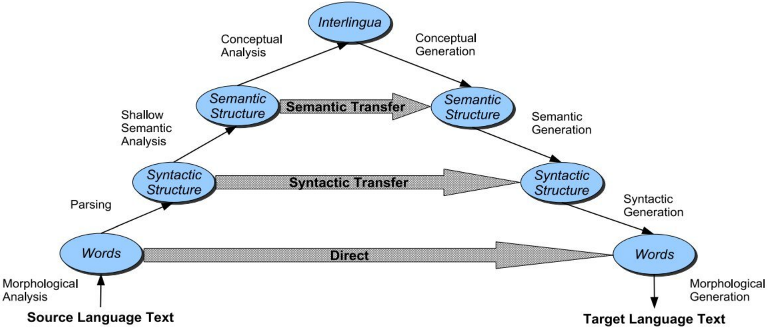
Tense and case

- Spanish has two past tenses.
  - The **preterite** tense is for events with a definite time, e.g.
    I biked to work this morning
  - The **imperfect** is for events with indefinite times, e.g.
    I biked to work all last summer
  - To translate English to Spanish, we must pick the right tense. This seems to require some deeper semantic understanding.
- Many languages have richer case morphology than English; translating to these languages requires identifying whether the word should be in the accusative, dative, etc.

Idioms

- Why in the world
- Kick the bucket
- Lend me your ears
- ...

The Vauquois Triangle



Outline

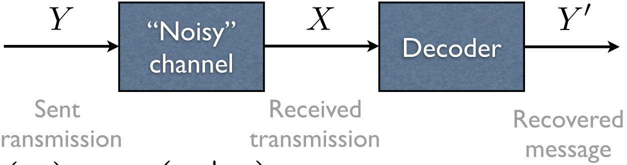
The noisy channel model

- Remember the noisy channel model?[2]
- This is a general framework for thinking about translation (and many other NLP problems).

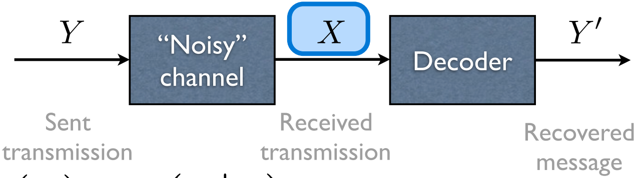
[2] next few slides from Chris Dyer

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.***'

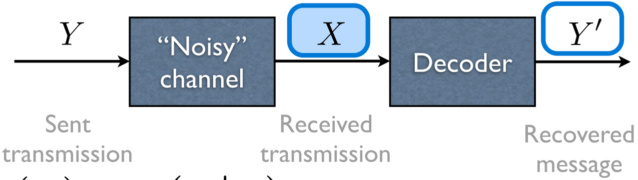
Warren Weaver to Norbert Wiener, March, 1947


Y
"Noisy" channel
X
Decoder
Y'
Sent transmission
Received transmission
Recovered message

$$p(y) \quad p(x|y)$$

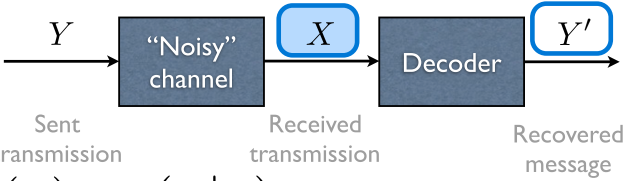

Y
"Noisy" channel
X
Decoder
Y'
Sent transmission
Received transmission
Recovered message

$$p(y) \quad p(x|y)$$



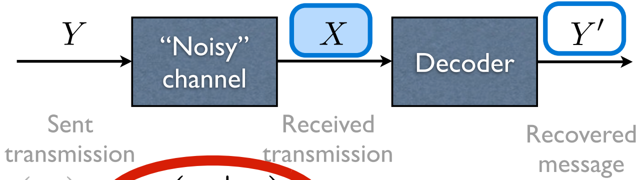
Y
"Noisy" channel
X
Decoder
Y'
Sent transmission
Received transmission
Recovered message

$$p(y) \quad p(x|y)$$



$$p(y) \qquad p(x|y)$$

$$\boxed{y'} = \arg\max_{y} p(y|x)$$



$$p(y) \quad p(x|y)$$

$$\boxed{y'} = \arg\max_y p(y|x)$$

Y → "Noisy" channel → $\boxed{X}$ → Decoder → $\boxed{Y'}$

Sent transmission

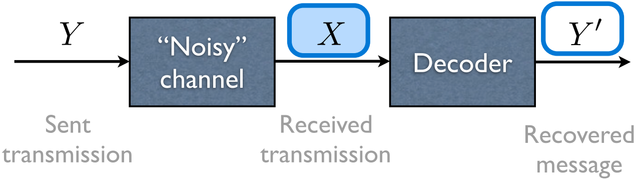
Received transmission

Recovered message

$p(y)$ $p(x|y)$

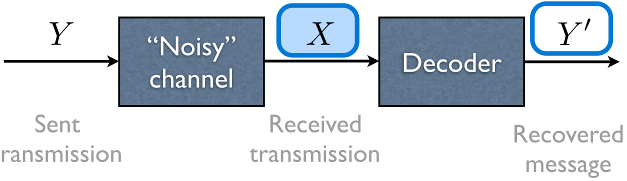
$$\boxed{y'} = \arg\max_{y} p(y|x)$$

I can help.



$$\boxed{y'} = \arg\max_y p(y|x)$$

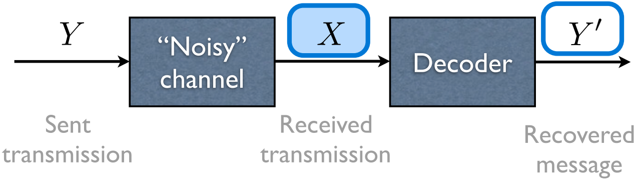
$$= \arg\max_y \frac{p(x|y)p(y)}{p(x)}$$



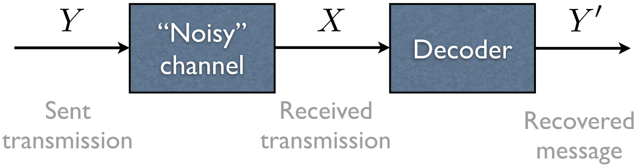
$$\boxed{y'} = \arg\max_y p(y|x)$$

$$= \arg\max_y \frac{p(x|y)p(y)}{p(x)}$$

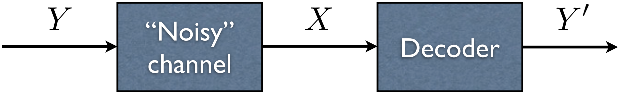
Denominator doesn't depend on y.



$$
\begin{aligned}
\boxed{y'} &= \arg\max_{y} p(y|x) \\
&= \arg\max_{y} \frac{p(x|y)p(y)}{p(x)} \\
&= \arg\max_{y} p(x|y)p(y)
\end{aligned}
$$



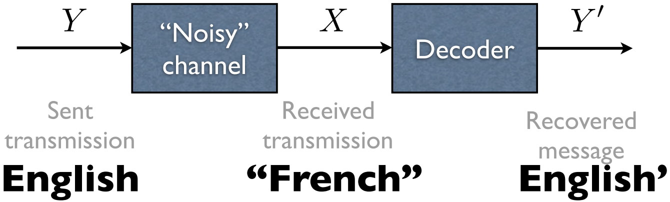
$$y' = \arg\max_y p(x|y)p(y)$$



Sent transmission	Received transmission	Recovered message
English	**"French"**	**English'**

$$\cancel{y' = \arg\max_{y} p(x|y)p(y)}$$

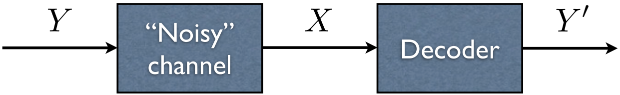
$$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$



$$\cancel{y' = \arg\max_{y} p(x|y)p(y)}$$

$$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

translation model



Sent transmission | Received transmission | Recovered message

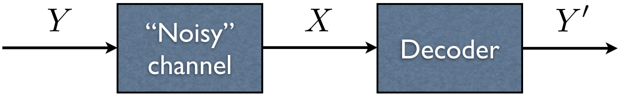
English | **"French"** | **English'**

$$\cancel{y' = \arg\max_{y} p(x|y)p(y)}$$

$$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

translation model

language model



Sent transmission

English

Received transmission

"French"

Recovered message

English'

$$\cancel{y' = \arg\max_{y} p(x|y)p(y)}$$

$$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

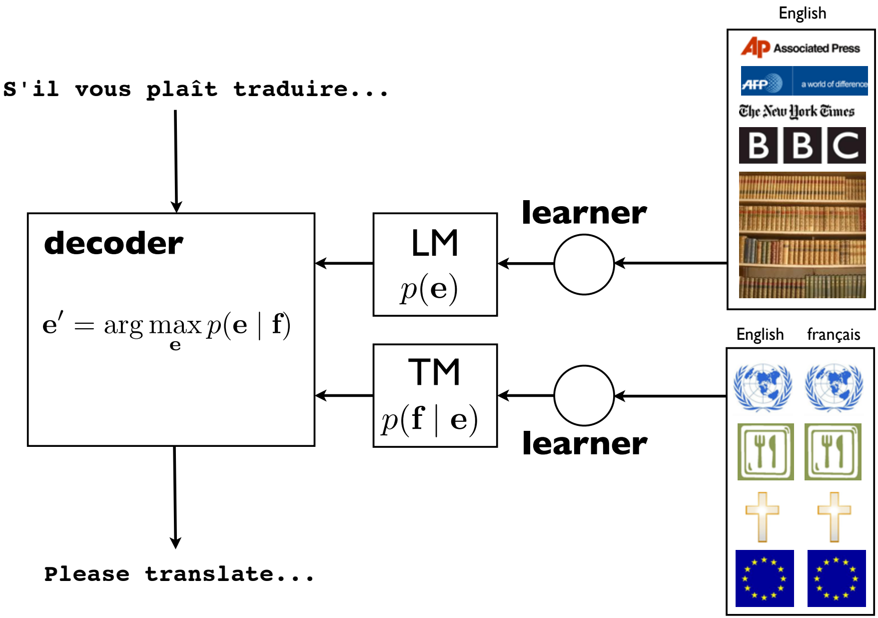
translation model

language model

Other noisy channel applications: OCR, speech recognition, spelling correction...

Division of labor

- **Translation model**
  - probability of translation *back* into the source
  - ensures **adequacy** of translation
- **Language model**
  - is a translation hypothesis "good" English?
  - ensures **fluency** of translation


S'il vous plaît traduire...
decoder
$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{e} \mid \mathbf{f})$
Please translate...
LM
$p(\mathbf{e})$
TM
$p(\mathbf{f} \mid \mathbf{e})$
learner
learner
English
Associated Press
AFP
a world of difference
The New York Times
BBC
English
français

Outline

Estimating the components

- We've already learned how to estimate language models.
  - Google uses n-grams of order 5-7.
  - Smoothing is really important;
    so is being clever about decoding.

Estimating the components

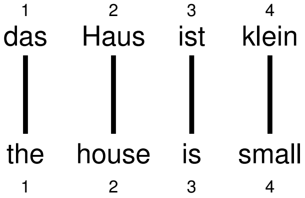
- We've already learned how to estimate language models.
  - Google uses n-grams of order 5-7.
  - Smoothing is really important;
    so is being clever about decoding.
- Estimating the translation model is more difficult.
  - $\boldsymbol{e}$ = And the program was implemented
  - $\boldsymbol{f}$ = La programmation a été mise en application

Estimating the components

- We've already learned how to estimate language models.
  - Google uses n-grams of order 5-7.
  - Smoothing is really important;
    so is being clever about decoding.
- Estimating the translation model is more difficult.
  - $\boldsymbol{e} = \text{And the program was implemented}$
  - $\boldsymbol{f} = \text{La programmation a été mise en application}$
  - $P(\boldsymbol{f}|\boldsymbol{e})$ is really hard to define.
  - Easier: something like
    $P(\text{la}|\text{the}) \times P(\text{programme}|\text{programmation}) \times \ldots$
  - But we are given aligned sentences, not aligned words.

Alignment

- In a parallel text (or when we translate), we **align** words in one language with the words in the other



- Word *positions* are numbered 1–4

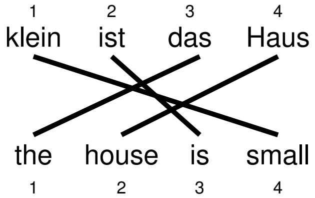
Alignment function

- Formalizing *alignment* with an **alignment function**
- Mapping an English target word at position i to a German source word at position j with a function $a : i \to j$
- Example

$$a : \{1 \to 1, 2 \to 2, 3 \to 3, 4 \to 4\}$$

Reordering

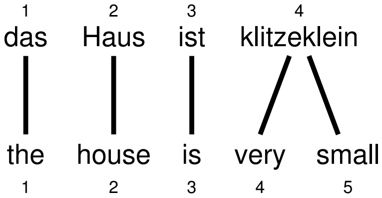
- Words may be **reordered** during translation



$$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$$

One-to-many translation

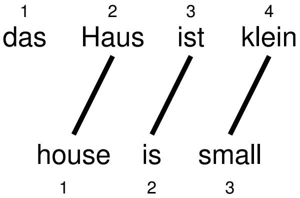
- A source word may translate into **multiple** target words



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Dropping words

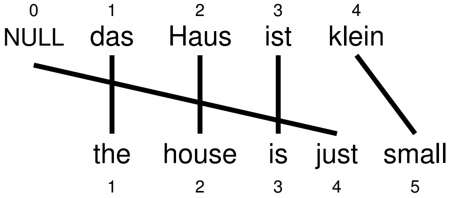
- Words may be **dropped** when translated
  - The German article *das* is dropped



$$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$$

Inserting words

- Words may be **added** during translation
  - The English *just* does not have an equivalent in German
  - We still need to map it to something: special NULL token



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$$

Translation with alignments

Let's add the alignments as a variable into the probability model:

$$
\begin{aligned}
P(f_1 \dots f_m | e_1 \dots e_\ell) &= \sum_{\boldsymbol{a}} P(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_\ell) \\
&= \dots \\
&= \prod_i \sum_{a_i} q(a_i | i, \ell, m) t(f_i | e_{a_i})
\end{aligned}
$$

- $t(f_i | e_{a_i})$ is the translation probability
- $q(a_i | i, \ell, m)$ is the alignment probability

Translation with alignments

Let's add the alignments as a variable into the probability model:

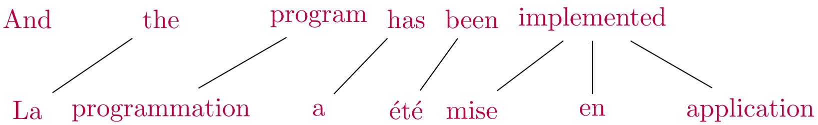
$$
\begin{aligned}
P(f_1 \ldots f_m | e_1 \ldots e_\ell) &= \sum_{\boldsymbol{a}} P(f_1 \ldots f_m, a_1 \ldots a_m | e_1 \ldots e_\ell) \\
&= \ldots \\
&= \prod_i \sum_{a_i} q(a_i | i, \ell, m) t(f_i | e_{a_i})
\end{aligned}
$$

- $t(f_i | e_{a_i})$ is the translation probability
- $q(a_i | i, \ell, m)$ is the alignment probability

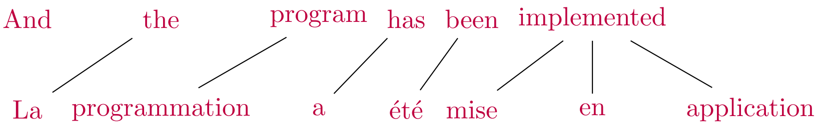
Independence assumptions:

- Word translations are independent given the alignments.
- Alignments are independent of each other.

Example


And the program has been implemented
La programmation a été mise en application

Example



$$
\begin{aligned}
P(\boldsymbol{f}, \boldsymbol{a} | \boldsymbol{e}) = & q(2, 1, 6, 7) \times t(\text{La}|\text{the}) \\
& \times q(3|2, 6, 7) \times t(\text{Programmation}|\text{program}) \\
& \times q(4|3, 6, 7) \times t(\text{a}|\text{has}) \\
& \times q(5|4, 6, 7) \times t(\text{été}|\text{been}) \\
& \times q(6|5, 6, 7) \times t(\text{mise}|\text{implemented}) \\
& \times q(6|6, 6, 7) \times t(\text{en}|\text{implemented}) \\
& \times q(6|7, 6, 7) \times t(\text{application}|\text{implemented})
\end{aligned}
$$

IBM alignment models

- IBM Model 1: $q(j|i, \ell, m) = \frac{1}{\ell}$
  (All alignments are equally likely.)

Estimation and alignment

- To translate, we need $t(f|e)$, the translation probabilities.
- If we knew the alignments, estimation would be easy.

$$
\begin{aligned}
t(\text{programme}|\text{program}) &= \frac{\text{count-of-programme-aligned-to-program}}{\text{count-of-program}} \\
&= \sum_{\langle f,e\rangle \in \text{examples}} \frac{\sum_i \delta(f_i = \text{programme}, e_{a_i} = \text{program})}{\sum_j \delta(e_j = \text{program})}
\end{aligned}
$$

Estimation and alignment

- To translate, we need $t(f|e)$, the translation probabilities.
- If we knew the alignments, estimation would be easy.

$$t(\text{programme}|\text{program}) = \frac{\text{count-of-programme-aligned-to-program}}{\text{count-of-program}}$$

$$= \sum_{\langle f,e\rangle \in \text{examples}} \frac{\sum_i \delta(f_i = \text{programme}, e_{a_i} = \text{program})}{\sum_j \delta(e_j = \text{program})}$$

- Conversely, getting the alignments would be easy if we knew the translation probabilities.

$$P(\boldsymbol{a}|\boldsymbol{e},\boldsymbol{f}) = \frac{P(\boldsymbol{f}|\boldsymbol{e},\boldsymbol{a})P(\boldsymbol{a})}{\sum_{\boldsymbol{a}'} P(\boldsymbol{f}|\boldsymbol{e},\boldsymbol{a}')P(\boldsymbol{a}')}$$

Estimation and alignment

- To translate, we need $t(f|e)$, the translation probabilities.
- If we knew the alignments, estimation would be easy.

$$t(\text{programme}|\text{program}) = \frac{\text{count-of-programme-aligned-to-program}}{\text{count-of-program}}$$
$$= \sum_{\langle f,e \rangle \in \text{examples}} \frac{\sum_i \delta(f_i = \text{programme}, e_{a_i} = \text{program})}{\sum_j \delta(e_j = \text{program})}$$

- Conversely, getting the alignments would be easy if we knew the translation probabilities.

$$P(\boldsymbol{a}|\boldsymbol{e}, \boldsymbol{f}) = \frac{P(\boldsymbol{f}|\boldsymbol{e}, \boldsymbol{a})P(\boldsymbol{a})}{\sum_{\boldsymbol{a}'} P(\boldsymbol{f}|\boldsymbol{e}, \boldsymbol{a}')P(\boldsymbol{a}')}$$

- How to solve a chicken-and-egg problem? **EM**!

Example

- Translation probabilities

	the	house
la	0.4	0.1
maison	0.1	0.6

- Alignments

	$P(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$	$P(\boldsymbol{a} \mid \boldsymbol{f}, \boldsymbol{e})$
the $\rightarrow$ la, house $\rightarrow$ la		
the $\rightarrow$ la, house $\rightarrow$ maison		
the $\rightarrow$ maison, house $\rightarrow$ la		
the $\rightarrow$ maison, house $\rightarrow$ maison		

- Counts

	the	house
la		
maison		

Example

- Translation probabilities

	the	house
la	0.4	0.1
maison	0.1	0.6

- Alignments

	$P(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$	$P(\boldsymbol{a} \mid \boldsymbol{f}, \boldsymbol{e})$
the $\rightarrow$ la, house $\rightarrow$ la	$0.4 \times 0.1 = 0.04$	
the $\rightarrow$ la, house $\rightarrow$ maison	$0.4 \times 0.6 = 0.24$	
the $\rightarrow$ maison, house $\rightarrow$ la	$0.1 \times 0.6 = 0.06$	
the $\rightarrow$ maison, house $\rightarrow$ maison	$0.1 \times 0.1 = 0.01$	

- Counts

	the	house
la		
maison		

Example

- Translation probabilities

	the	house
la	0.4	0.1
maison	0.1	0.6

- Alignments

	$P(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$	$P(\boldsymbol{a} \mid \boldsymbol{f}, \boldsymbol{e})$
the $\rightarrow$ la, house $\rightarrow$ la	$0.4 \times 0.1 = 0.04$	0.11
the $\rightarrow$ la, house $\rightarrow$ maison	$0.4 \times 0.6 = 0.24$	0.69
the $\rightarrow$ maison, house $\rightarrow$ la	$0.1 \times 0.6 = 0.06$	0.17
the $\rightarrow$ maison, house $\rightarrow$ maison	$0.1 \times 0.1 = 0.01$	0.03

- Counts

	the	house
la		
maison		

Example

- Translation probabilities

	the	house
la	0.4	0.1
maison	0.1	0.6

- Alignments

	$P(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$	$P(\boldsymbol{a} \mid \boldsymbol{f}, \boldsymbol{e})$
the $\rightarrow$ la, house $\rightarrow$ la	$0.4 \times 0.1 = 0.04$	0.11
the $\rightarrow$ la, house $\rightarrow$ maison	$0.4 \times 0.6 = 0.24$	0.69
the $\rightarrow$ maison, house $\rightarrow$ la	$0.1 \times 0.6 = 0.06$	0.17
the $\rightarrow$ maison, house $\rightarrow$ maison	$0.1 \times 0.1 = 0.01$	0.03

- Counts

	the	house
la	$0.11 + 0.69 = 0.8$	$0.11 + 0.17 = 0.28$
maison	$0.17 + 0.03 = 0.2$	$0.69 + 0.03 = 0.72$

Example

- Translation probabilities

	the	house
la	0.4	0.1
maison	0.1	0.6

- Alignments

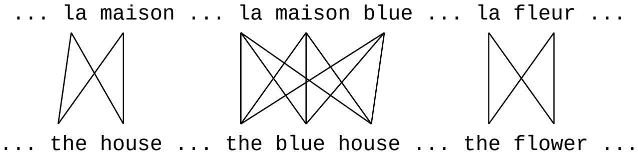
	$P(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$	$P(\boldsymbol{a} \mid \boldsymbol{f}, \boldsymbol{e})$
the $\rightarrow$ la, house $\rightarrow$ la	$0.4 \times 0.1 = 0.04$	0.11
the $\rightarrow$ la, house $\rightarrow$ maison	$0.4 \times 0.6 = 0.24$	0.69
the $\rightarrow$ maison, house $\rightarrow$ la	$0.1 \times 0.6 = 0.06$	0.17
the $\rightarrow$ maison, house $\rightarrow$ maison	$0.1 \times 0.1 = 0.01$	0.03

- Counts

	the	house
la	$0.11 + 0.69 = 0.8$	$0.11 + 0.17 = 0.28$
maison	$0.17 + 0.03 = 0.2$	$0.69 + 0.03 = 0.72$

Then we add up the counts across all the examples and update the translation probabilities

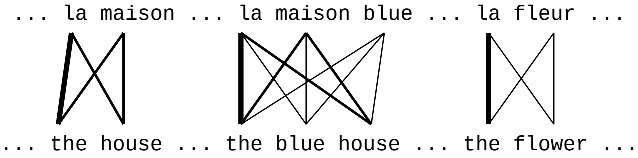
EM algorithm



- Initial step: all alignments equally likely

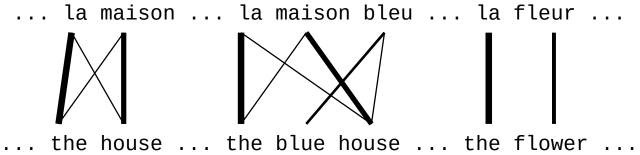
- Model learns that, e.g., *la* is often aligned with *the*

EM algorithm



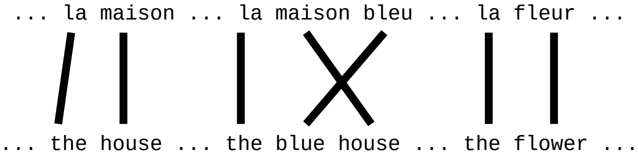
- After one iteration
- Alignments, e.g., between *la* and *the* are more likely

EM algorithm



- After another iteration
- It becomes apparent that alignments, e.g., between *fleur* and *flower* are more likely (**pigeon hole principle**)

EM algorithm



- Convergence
- Inherent hidden structure revealed by EM

EM algorithm

```
... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...
```

- Parameter estimation from the aligned corpus

IBM Model 1 and EM

- EM Algorithm consists of two steps
- **Expectation-Step**: Apply model to the data
  - parts of the model are hidden (here: alignments)
  - using the model, assign probabilities to possible values
- **Maximization-Step**: Estimate model from data
  - take assign values as fact
  - collect counts (weighted by probabilities)
  - estimate model from counts
- Iterate these steps until **convergence**

IBM alignment models

- IBM Model 1: $q(j|i, \ell, m) = \frac{1}{\ell}$
  (All alignments are equally likely.)

IBM alignment models

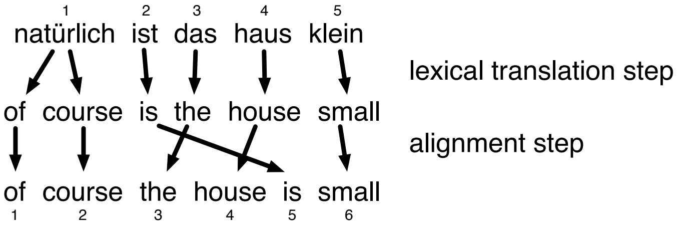
- IBM Model 1: $q(j|i,\ell,m) = \frac{1}{\ell}$
  (All alignments are equally likely.)
- IBM Model 2: $q(j|i,\ell,m) = \frac{c(j|i,\ell,m)}{c(i,\ell,m)}$
  (Alignment probability is a parameter of the model.)

IBM Model 2

Add a prior probability on alignments $q(a_i|i, m, \ell)$.

- We estimate q along with t during the M-step.
- The joint probability still decomposes across words: $P(\boldsymbol{e}, \boldsymbol{a}|\boldsymbol{f}) = \prod_i t(e_i|f_{a_i})q(a_i|i, \ell_e, \ell_f)$
- But adding this parameter makes the likelihood non-convex.
  - This means initialization affects the outcome.
  - Initializing from IBM model 1 works well in practice.

IBM Model 2


1 natürlich
2 ist
3 das
4 haus
5 klein
of course is the house small
of course the house is small
1 2 3 4 5 6
lexical translation step
alignment step

IBM alignment models

- IBM Model 1: $q(j|i, \ell, m) = \frac{1}{\ell}$
  (All alignments are equally likely.)
- IBM Model 2: $q(j|i, \ell, m) = \frac{c(j|i,\ell,m)}{c(i,\ell,m)}$
  (Alignment probability is a parameter of the model.)

IBM alignment models

- IBM Model 1: $q(j|i,\ell,m) = \frac{1}{\ell}$
  (All alignments are equally likely.)
- IBM Model 2: $q(j|i,\ell,m) = \frac{c(j|i,\ell,m)}{c(i,\ell,m)}$
  (Alignment probability is a parameter of the model.)
- IBM Model 3
  (model "fertility", the number of alignments per word)

IBM Model 3


Mary did not slap the green witch
Mary not slap slap slap the green witch
Mary not slap slap slap NULL the green witch
Maria no daba una botefada a la verde bruja
Maria no daba una bofetada a la bruja verde
n(3|slap)
p-null
t(la|the)
d(4|4)

IBM alignment models

- IBM Model 1: $q(j|i, \ell, m) = \frac{1}{\ell}$
  (All alignments are equally likely.)
- IBM Model 2: $q(j|i, \ell, m) = \frac{c(j|i,\ell,m)}{c(i,\ell,m)}$
  (Alignment probability is a parameter of the model.)
- IBM Model 3
  (model "fertility", the number of alignments per word)

IBM alignment models

- IBM Model 1: $q(j|i,\ell,m) = \frac{1}{\ell}$
  (All alignments are equally likely.)
- IBM Model 2: $q(j|i,\ell,m) = \frac{c(j|i,\ell,m)}{c(i,\ell,m)}$
  (Alignment probability is a parameter of the model.)
- IBM Model 3
  (model "fertility", the number of alignments per word)
- IBM Models 4 and 5
  - Condition on the alignment of the preceding word
  - Like an HMM: $P(a_i|a_{i-1},\ell,m)$

IBM alignment models

- IBM Model 1: $q(j|i,\ell,m) = \frac{1}{\ell}$
  (All alignments are equally likely.)
- IBM Model 2: $q(j|i,\ell,m) = \frac{c(j|i,\ell,m)}{c(i,\ell,m)}$
  (Alignment probability is a parameter of the model.)
- IBM Model 3
  (model "fertility", the number of alignments per word)
- IBM Models 4 and 5
  - Condition on the alignment of the preceding word
  - Like an HMM: $P(a_i|a_{i-1},\ell,m)$

There are lots of papers on alignment alone, but Fraser and Marcu (2007) note that improvements in alignment accuracy may not improve overall translation.

Outline

Problems with word-based translation

Word-based translation has obvious limitations:

- Multi-word alignments aren't modeled well:
  $P(\text{daba una botefada}|\text{slap}) \stackrel{?}{=} P(\text{daba}|\text{slap})P(\text{una}|\text{slap})P(\text{botefada}|\text{slap})$
- Many phrasal translations are non-compositional:
  faire (make) le (the) menage (home) $\rightarrow$ clean up
- Alignment decisions for phrasal units should be made jointly:
  la comida me gusta mucho $\rightarrow$ i like the food a lot

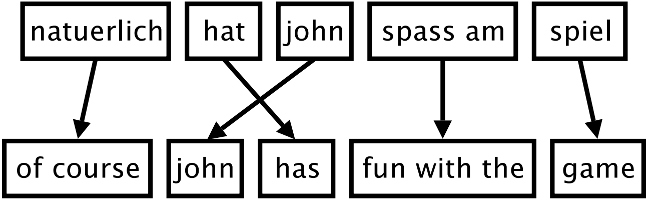
Problems with word-based translation

Word-based translation has obvious limitations:

- Multi-word alignments aren't modeled well:
  $P(\text{daba una botefada}|\text{slap}) \stackrel{?}{=}$
  $P(\text{daba}|\text{slap})P(\text{una}|\text{slap})P(\text{botefada}|\text{slap})$
- Many phrasal translations are non-compositional:
  faire (make) le (the) menage (home) $\rightarrow$ clean up
- Alignment decisions for phrasal units should be made jointly:
  la comida me gusta mucho $\rightarrow$ i like the food a lot

Two solutions: **phrases** and **syntax**

Phrase-Based Model



- Foreign input is segmented in phrases
- Each phrase is translated into English
- Phrases are reordered

Phrase Translation Table

- Main knowledge source: table with phrase translations and their probabilities

- Example: phrase translations for natuerlich

Translation	Probability $\phi(\bar{e}\|f)$
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

Linguistic Phrases?

- Model is not limited to linguistic phrases
  (noun phrases, verb phrases, prepositional phrases, ...)

- Example non-linguistic phrase pair

$$\text{spass am} \rightarrow \text{fun with the}$$

- Prior noun often helps with translation of preposition

- Experiments show that limitation to linguistic phrases hurts quality

Learning a Phrase Translation Table

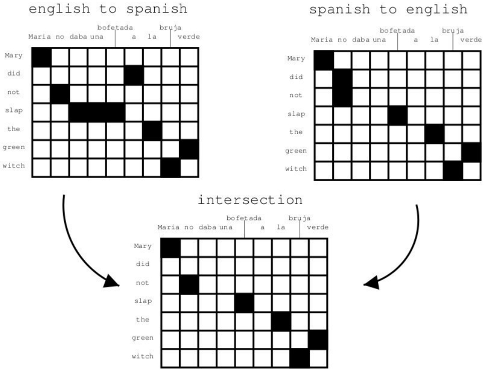
- Task: learn the model from a parallel corpus

- Three stages:
  - word alignment: using IBM models or other method
  - extraction of phrase pairs
  - scoring phrase pairs

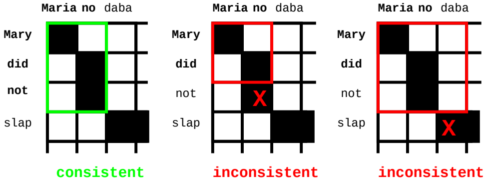
Phrase-based translation

Typically, we start with a *symmetrized* set of word alignments

- Align e to f
- Align f to e (not generally the same!)
- Take the intersection

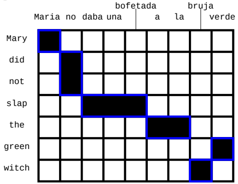


Phrase Extraction Criteria



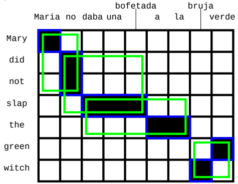
- Phrase alignment has to *contain all alignment points* for all covered words
- Phrase alignment has to *contain at least one alignment point*

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

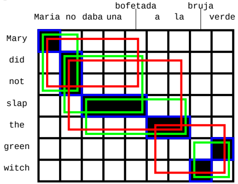
Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

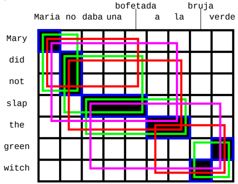
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Word alignment induced phrases



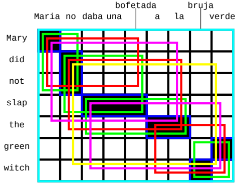
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)

Word alignment induced phrases (5)



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Scoring Phrase Translations

- Phrase pair extraction: collect all phrase pairs from the data
- Phrase pair scoring: assign probabilities to phrase translations
- Score by relative frequency:

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)}$$

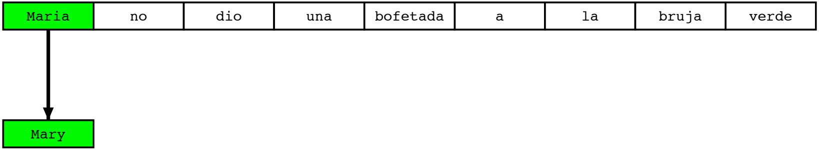
Outline

Decoding Process

Maria	no	dio	una	bofetada	a	la	bruja	verde

- Build translation left to right
  - *select foreign* words to be translated

Decoding Process



- Build translation *left to right*
  - select foreign words to be translated
  - *find English* phrase translation
  - *add English* phrase to end of partial translation

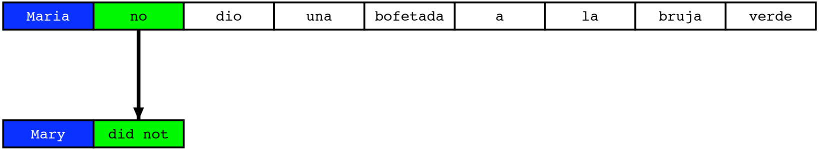
Decoding Process

Maria	no	dio	una	bofetada	a	la	bruja	verde

Mary

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
  - *mark foreign* words as translated

Decoding Process



- *One to many* translation

Decoding Process

- Many to one translation

Decoding Process

- *Many to one* translation

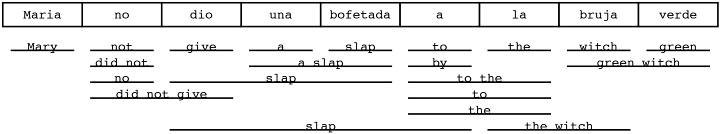
Decoding Process

- *Reordering*

Decoding Process

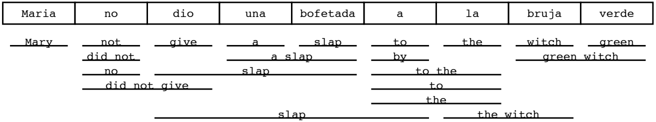
- Translation *finished*

Translation Options



- Look up *possible phrase translations*
  - many different ways to *segment* words into phrases
  - many different ways to *translate* each phrase

Hypothesis Expansion




```
e:
f: ---------
p: 1
```

- Start with **empty hypothesis**
  - e: no English words
  - f: no foreign words covered
  - p: probability 1

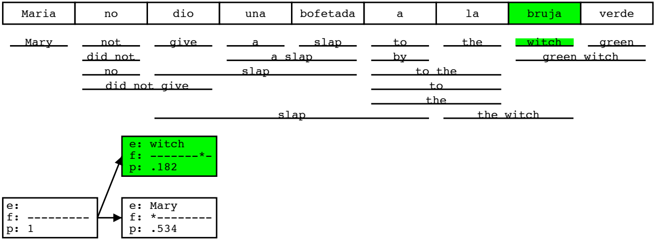
Hypothesis Expansion

- Pick *translation option*
- Create *hypothesis*
  - e: add English phrase `Mary`
  - f: first foreign word covered
  - p: probability 0.534

A Quick Word on Probabilities

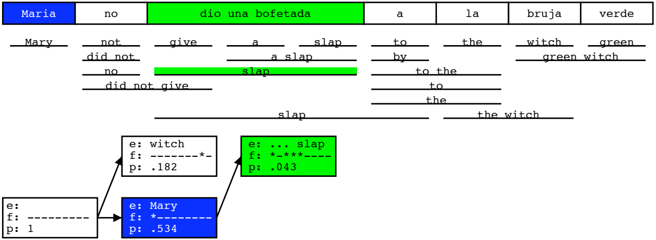
- Not going into detail here, but...
- *Translation Model*
  - phrase translation probability p(Mary|Maria)
  - reordering costs
  - phrase/word count costs
  - ...
- *Language Model*
  - uses trigrams:
  - $p(\text{Mary did not}) =$
    $p(\text{Mary}|\text{START}) \times p(\text{did}|\text{Mary,START}) \times \text{p}(\text{not}|\text{Mary did})$

Hypothesis Expansion



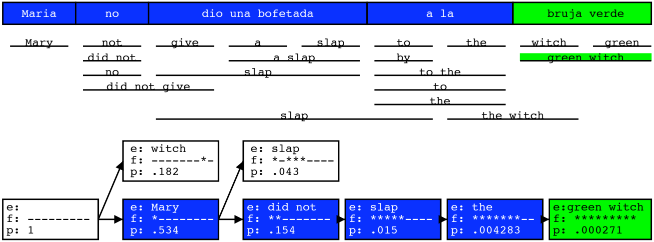
- Add another *hypothesis*

Hypothesis Expansion



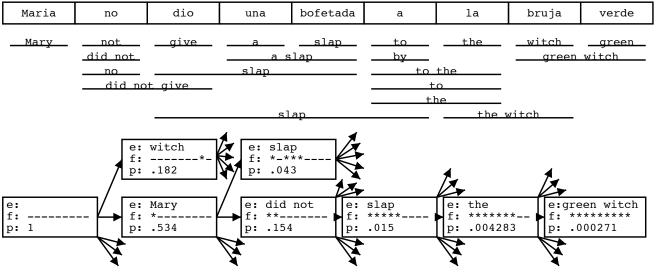
- Further *hypothesis expansion*

Hypothesis Expansion



- ... until all foreign words *covered*
  - find *best hypothesis* that covers all foreign words
  - *backtrack* to read off translation

Hypothesis Expansion



- Adding more hypothesis

⇒ *Explosion* of search space

Explosion of Search Space

- Number of hypotheses is *exponential* with respect to sentence length

⇒ Decoding is NP-complete [Knight, 1999]

⇒ Need to *reduce search space*
  - risk free: hypothesis **recombination**
  - risky: **histogram/threshold pruning**

Outline

Freedom from the rusty manacles of probability

- In the noisy channel model, decoding maximizes:

$$\log P(f|e) + \log P(e)$$

Freedom from the rusty manacles of probability

- In the noisy channel model, decoding maximizes:

$$\log P(f|e) + \log P(e)$$

- We might decide these components are not equally important:

$$\lambda_1 \log P(f|e) + \lambda_2 \log P(e)$$

Freedom from the rusty manacles of probability

- In the noisy channel model, decoding maximizes:

$$\log P(f|e) + \log P(e)$$

- We might decide these components are not equally important:

$$\lambda_1 \log P(f|e) + \lambda_2 \log P(e)$$

- But this is just a log-linear model.
  Why not add other features?

$$\lambda_1 \log P(f|e) + \lambda_2 \log P(e) + \lambda_3 \log P(e|f) + \ldots$$

Minimum error-rate training (MERT)

Our new objective:

$$\lambda_1 \log P(f|e) + \lambda_2 \log P(e) + \lambda_3 \log P(e|f) + \ldots$$

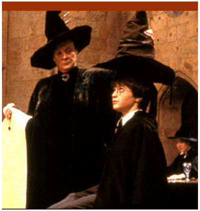
- How to set λ?
  Maximize Bleu score on dev-set.

Minimum error-rate training (MERT)

Our new objective:

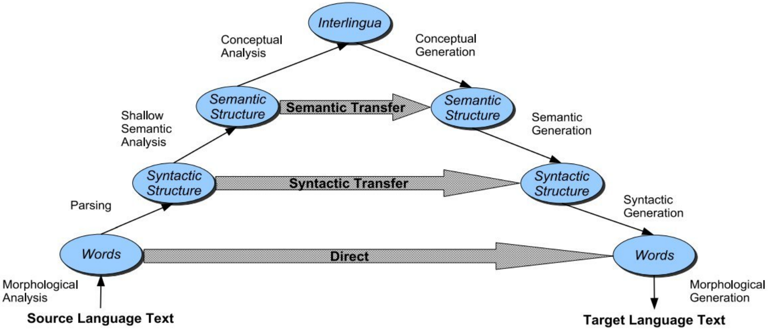
$$\lambda_1 \log P(f|e) + \lambda_2 \log P(e) + \lambda_3 \log P(e|f) + \dots$$

- How to set λ?
  Maximize Bleu score on dev-set.
  - This **will** help you get a higher Bleu score on test data. ☺
  - But how much do we really trust Bleu?
  - We can't get anything like a gradient of the Bleu score with respect to λ, so learning is difficult, especially with many features. ☹
  - (But see Galley et al EMNLP 2013 for some progress)



Outline

The Vauquois Triangle



Is translation easier at the syntactic level?

la empresa tiene enemigos fuertes en Europa .

the company has strong enemies in Europe .

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

la empresa tiene enemigos fuertes en Europa .

the company has strong enemies in Europe .

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

la empresa tiene enemigos fuertes en Europa .

the company has **strong enemies** in Europe .

Same pattern:

NN JJ → JJ NN

sus asociados no son fuertes .

sus grupos estan en Europa .

Garcia has a company also .

Garcia tambien tiene una empresa .

the **modern groups** sell **strong pharmaceuticals** .

los grupos modernos venden medicinas fuertes .

its clients are angry .

sus clientes estan enfadados .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the associates are also angry .

los asociados tambien estan enfadados .

the **small groups** are not modern .

los grupos pequenos no son modernos .

la empresa tiene enemigos fuertes en Europa .

the company has strong enemies in Europe .

Same pattern:

NN JJ → JJ NN

Finite-state models do not capture this generalization.

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

Let's do an example

Context-Free Grammar

Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu

Context-Free Grammar

S

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu

Context-Free Grammar

S

$S \rightarrow NP\ VP$

$NP \rightarrow$ watashi wa

$NP \rightarrow$ hako wo

$VP \rightarrow NP\ V$

$V \rightarrow$ akemasu

Context-Free Grammar

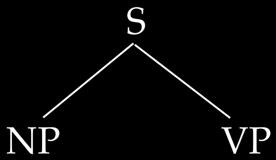
S → NP VP

NP → watashi wa

NP → hako wo

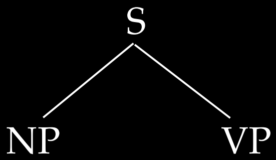
VP → NP V

V → akemasu



Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



Context-Free Grammar

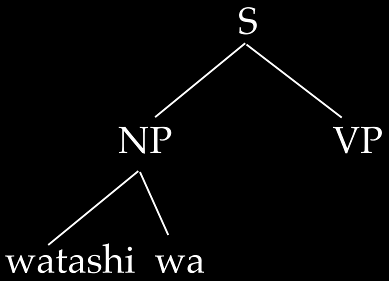
S → NP VP

NP → watashi wa

NP → hako wo

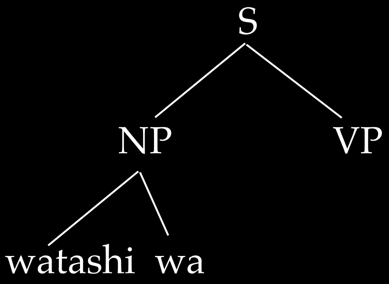
VP → NP V

V → akemasu



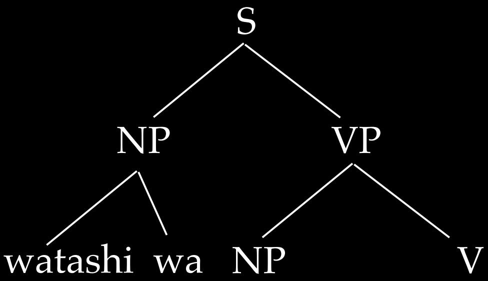
Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



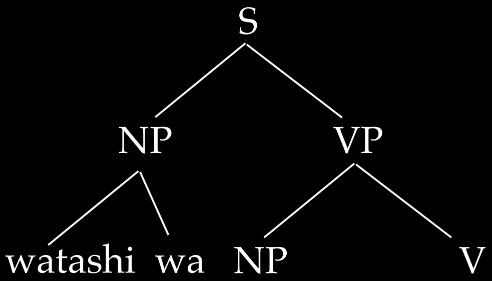
Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



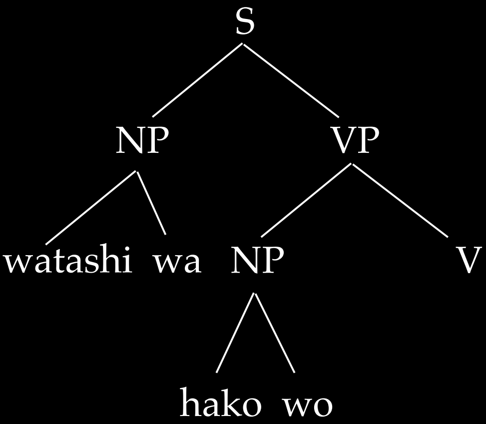
Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



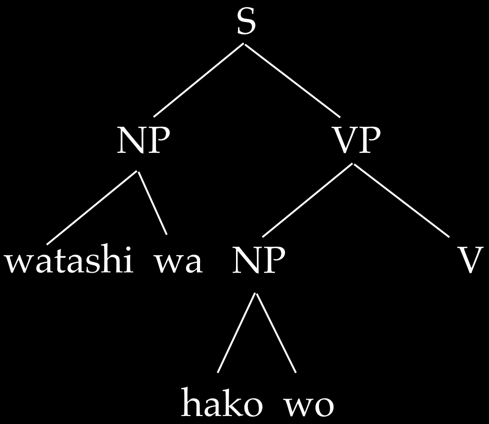
Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



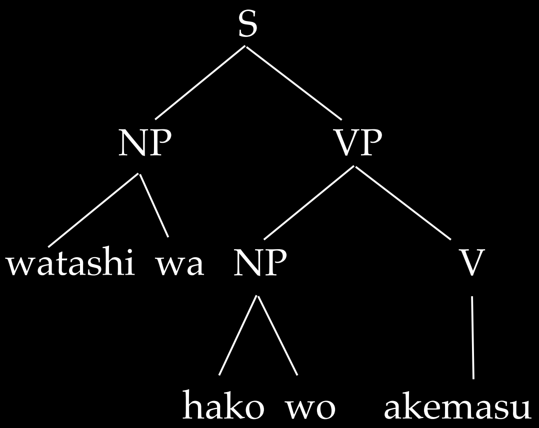
Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



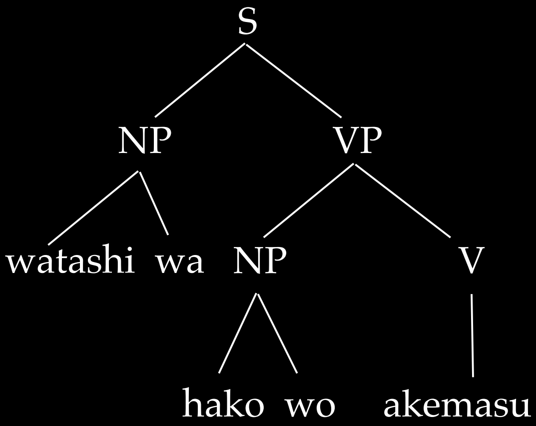
Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
$\boxed{\text{V → akemasu}}$



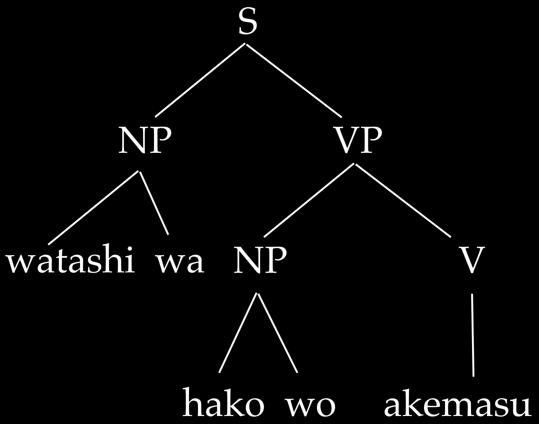
Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



Context-Free Grammar

S → NP VP
NP → watashi wa
NP → hako wo
VP → NP V
V → akemasu



watashi wa hako wo akemasu

Synchronous Context-Free Grammar

$$
\begin{aligned}
S &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{watashi wa} \\
\text{NP} &\rightarrow \text{hako wo} \\
\text{VP} &\rightarrow \text{NP V} \\
V &\rightarrow \text{akemasu}
\end{aligned}
$$

Synchronous Context-Free Grammar

S → NP VP	S → NP VP
NP → watashi wa	NP → I
NP → hako wo	NP → the box
VP → NP V	VP → V NP
V → akemasu	V → open

Synchronous Context-Free Grammar

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP}_1\ \mathrm{VP}_2 \ /\ \mathrm{NP}_1\ \mathrm{VP}_2 \\
\mathrm{NP} &\rightarrow \text{watashi wa} \ /\ \text{I} \\
\mathrm{NP} &\rightarrow \text{hako wo} \ /\ \text{the box} \\
\mathrm{VP} &\rightarrow \mathrm{NP}_1\ \mathrm{V}_2 \ /\ \mathrm{V}_1\ \mathrm{NP}_2 \\
\mathrm{V} &\rightarrow \text{akemasu} \ /\ \text{open}
\end{aligned}
$$

Synchronous Context-Free Grammar

$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

$NP \rightarrow$ watashi wa / I

$NP \rightarrow$ hako wo / the box

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar

S S

$S \rightarrow NP_1\ VP_2$ / $NP_1\ VP_2$
$NP \rightarrow$ watashi wa / I
$NP \rightarrow$ hako wo / the box
$VP \rightarrow NP_1\ V_2$ / $V_1\ NP_2$
$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar

S ······························ S

$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

$NP \rightarrow$ watashi wa / I

$NP \rightarrow$ hako wo / the box

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar

S .... S

$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

$NP \rightarrow$ watashi wa / I

$NP \rightarrow$ hako wo / the box

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar

$\boxed{S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2}$

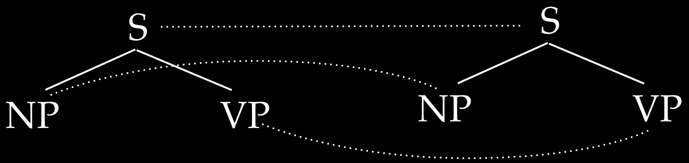
$NP \rightarrow$ watashi wa / I

$NP \rightarrow$ hako wo / the box

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar



$S \rightarrow NP_1\ VP_2 \ /\ NP_1\ VP_2$

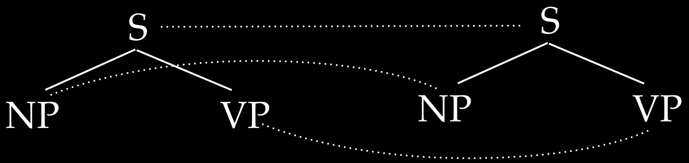
$NP \rightarrow$ watashi wa / I

$NP \rightarrow$ hako wo / the box

$VP \rightarrow NP_1\ V_2 \ /\ V_1\ NP_2$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar



$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

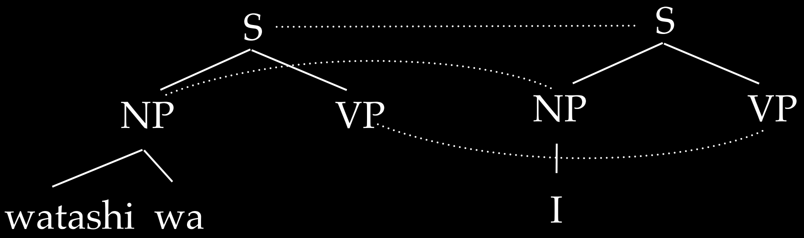
$\boxed{NP \rightarrow \text{watashi wa}\ /\ \text{I}}$

$NP \rightarrow \text{hako wo}\ /\ \text{the box}$

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow \text{akemasu}\ /\ \text{open}$

Synchronous Context-Free Grammar



$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

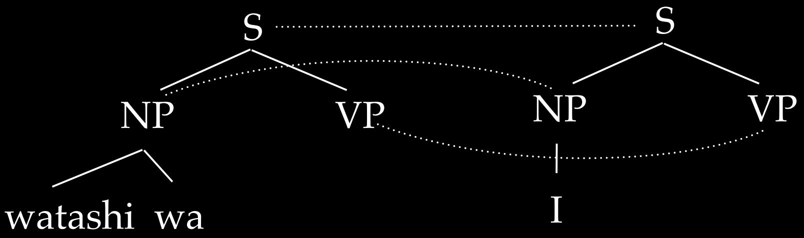
$NP \rightarrow \text{watashi wa}\ /\ \text{I}$

$NP \rightarrow \text{hako wo}\ /\ \text{the box}$

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow \text{akemasu}\ /\ \text{open}$

Synchronous Context-Free Grammar



$$S \rightarrow NP_1\ VP_2 \ /\ NP_1\ VP_2$$

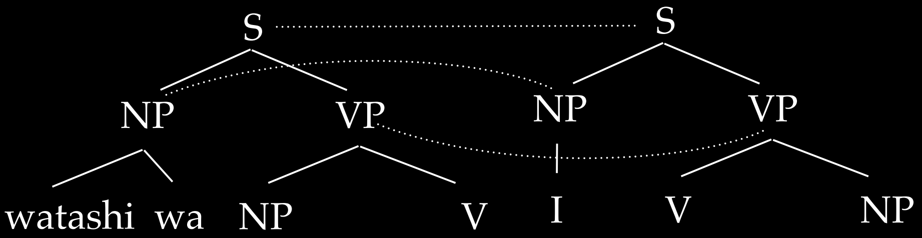
$$NP \rightarrow \text{watashi wa} \ /\ \text{I}$$

$$NP \rightarrow \text{hako wo} \ /\ \text{the box}$$

$$\boxed{VP \rightarrow NP_1\ V_2 \ /\ V_1\ NP_2}$$

$$V \rightarrow \text{akemasu} \ /\ \text{open}$$

Synchronous Context-Free Grammar



$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

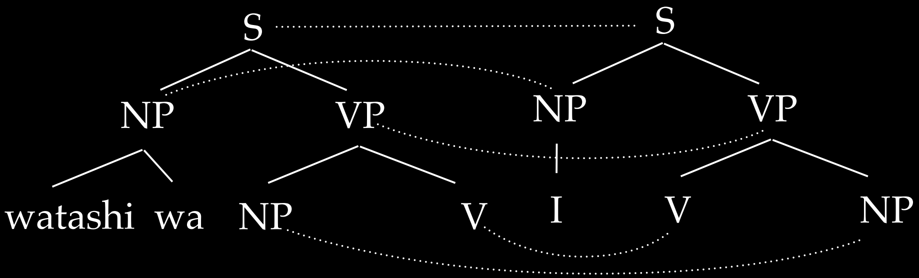
$NP \rightarrow$ watashi wa / I

$NP \rightarrow$ hako wo / the box

$\boxed{VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2}$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar



$\mathrm{S} \rightarrow \mathrm{NP}_1\ \mathrm{VP}_2\ /\ \mathrm{NP}_1\ \mathrm{VP}_2$

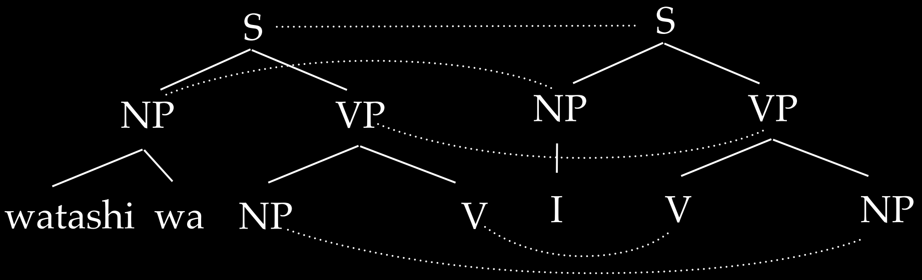
$\mathrm{NP} \rightarrow$ watashi wa / I

$\mathrm{NP} \rightarrow$ hako wo / the box

$\boxed{\mathrm{VP} \rightarrow \mathrm{NP}_1\ \mathrm{V}_2\ /\ \mathrm{V}_1\ \mathrm{NP}_2}$

$\mathrm{V} \rightarrow$ akemasu / open

Synchronous Context-Free Grammar



$$S \rightarrow NP_1\ VP_2 \ /\ NP_1\ VP_2$$

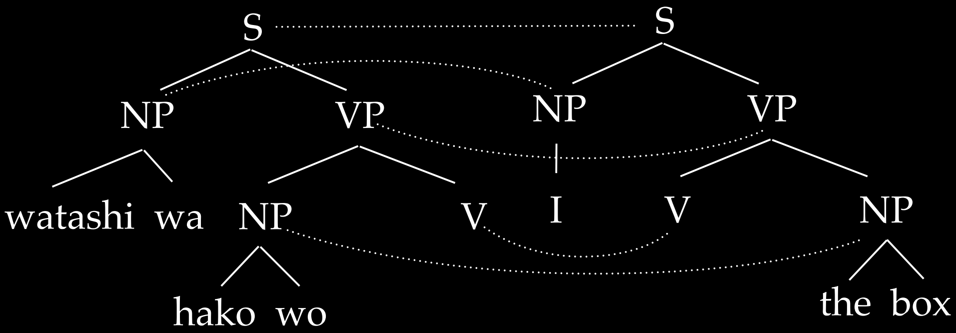
$$NP \rightarrow \text{watashi wa} \ /\ \text{I}$$

$$\boxed{NP \rightarrow \text{hako wo} \ /\ \text{the box}}$$

$$VP \rightarrow NP_1\ V_2 \ /\ V_1\ NP_2$$

$$V \rightarrow \text{akemasu} \ /\ \text{open}$$

Synchronous Context-Free Grammar



$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

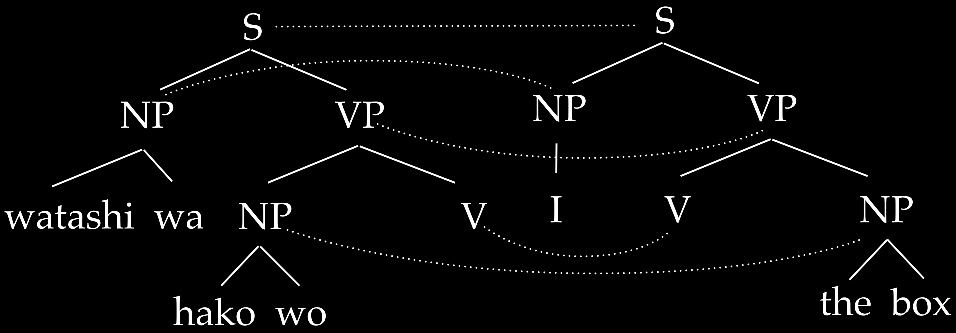
$NP \rightarrow$ watashi wa / I

$\boxed{NP \rightarrow \text{hako wo / the box}}$

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar



$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$

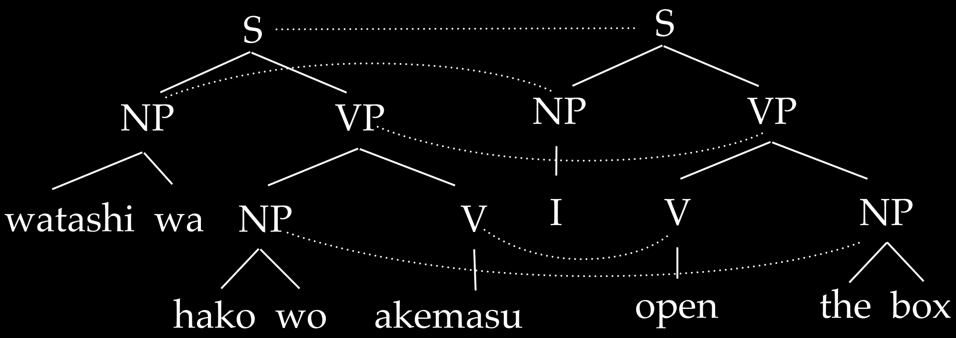
$NP \rightarrow$ watashi wa / I

$NP \rightarrow$ hako wo / the box

$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$

$V \rightarrow$ akemasu / open

Synchronous Context-Free Grammar



$\mathrm{S} \rightarrow \mathrm{NP}_1\ \mathrm{VP}_2\ /\ \mathrm{NP}_1\ \mathrm{VP}_2$

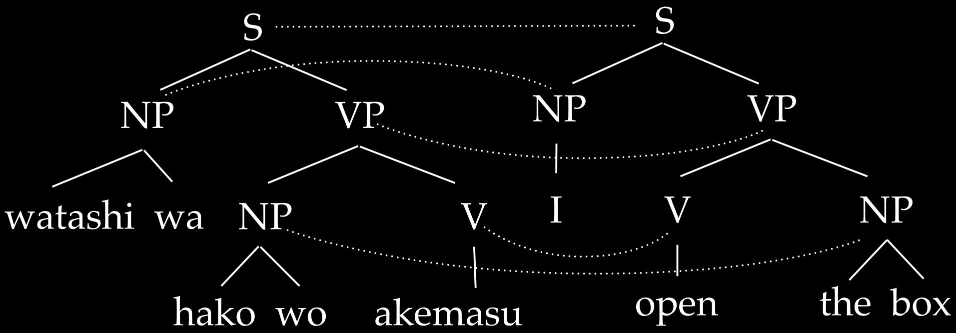
$\mathrm{NP} \rightarrow$ watashi wa / I

$\mathrm{NP} \rightarrow$ hako wo / the box

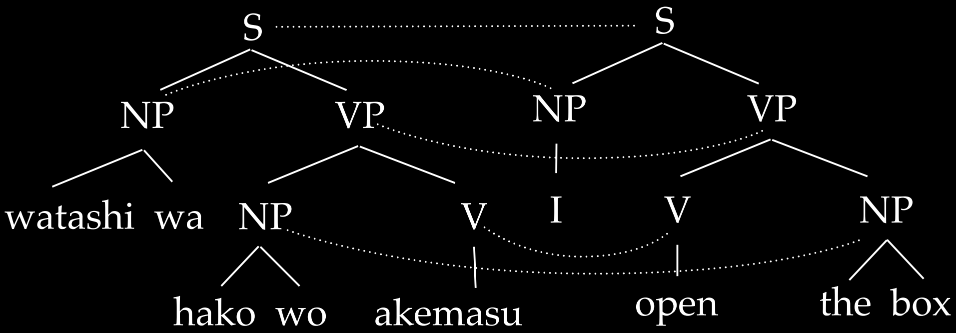
$\mathrm{VP} \rightarrow \mathrm{NP}_1\ \mathrm{V}_2\ /\ \mathrm{V}_1\ \mathrm{NP}_2$

$\boxed{\mathrm{V} \rightarrow \text{akemasu / open}}$

Synchronous Context-Free Grammar

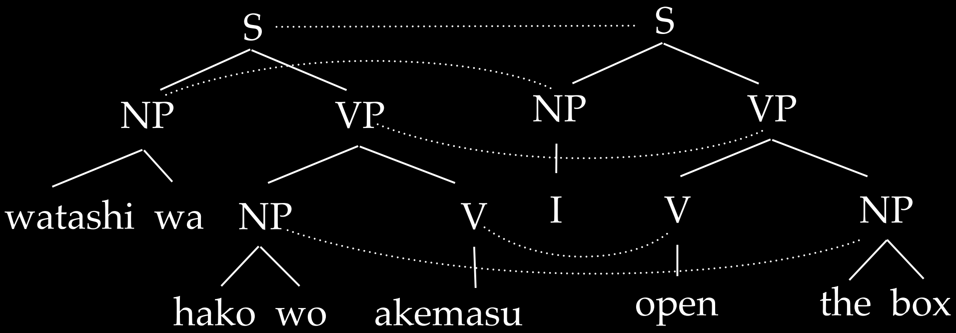


Synchronous Context-Free Grammar



watashi wa hako wo akemasu

Synchronous Context-Free Grammar

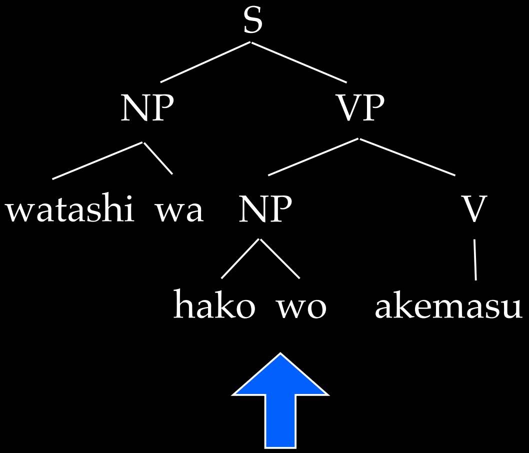


watashi wa hako wo akemasu   I open the box

Translation as Parsing

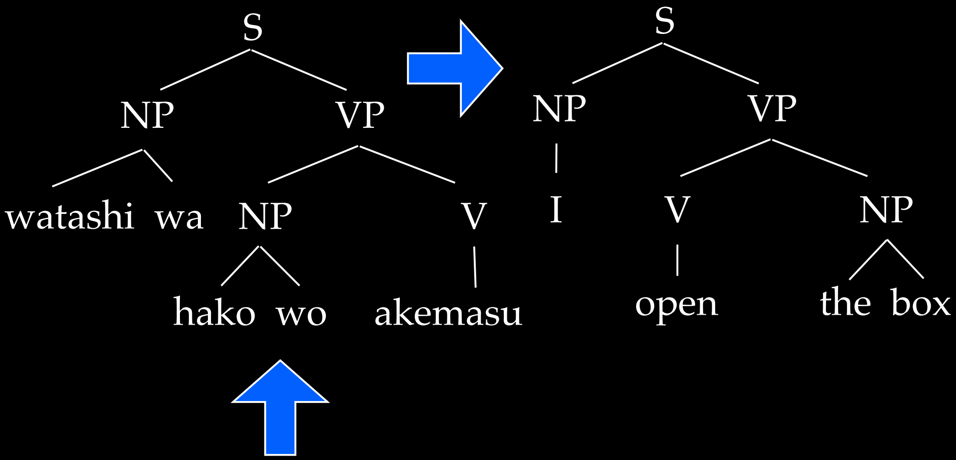
watashi wa hako wo akemasu

Translation as Parsing



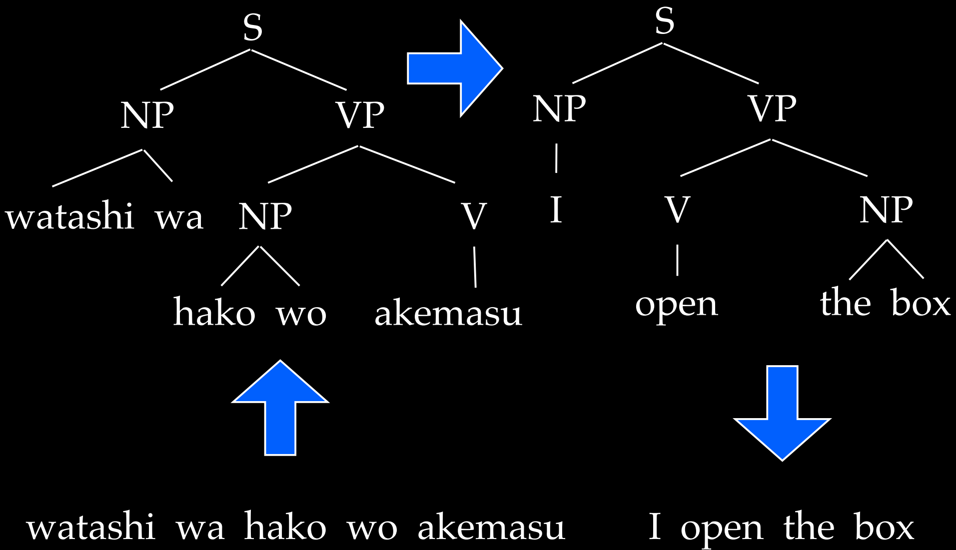
watashi wa hako wo akemasu

Translation as Parsing



watashi wa hako wo akemasu

Translation as Parsing

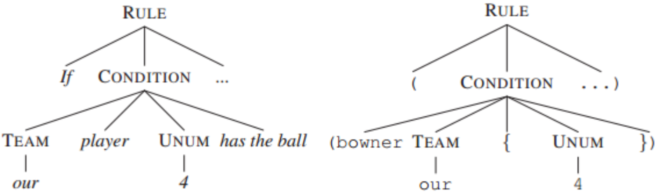


Synchronous grammars for semantic parsing

`((bowner our {4}) (do our {6} (pos (left (half our)))))`

If our player 4 has the ball, then our player 6 should stay in the left side of our half.

Figure 3.1: A meaning representation in CLANG and its English gloss



(a) English (b) CLANG

Figure 3.2: Partial parse trees for the string pair in Figure 3.1

Wong and Mooney (2007)

The big question

The Big Question

Where do the categories come from?

The Big Question

Where do the categories come from?

Answer #1: there are no categories!

The Big Question

Where do the categories come from?

Answer #1: there are no categories!

$X \rightarrow X_1\ X_2\ /\ X_1\ X_2$

$X \rightarrow X_1\ X_2\ /\ X_2\ X_1$

$X \rightarrow$ watashi wa / I

$X \rightarrow$ hako wo / the box

$X \rightarrow$ akemasu / open

The Big Question

Where do the categories come from?

Answer #1: there are no categories!

$X \rightarrow X_1\ X_2\ /\ X_1\ X_2$ ← Keep order

$X \rightarrow X_1\ X_2\ /\ X_2\ X_1$

$X \rightarrow$ watashi wa / I

$X \rightarrow$ hako wo / the box

$X \rightarrow$ akemasu / open

The Big Question

Where do the categories come from?

Answer #1: there are no categories!

$X \rightarrow X_1\ X_2\ /\ X_1\ X_2$ ⟵ Keep order

$X \rightarrow X_1\ X_2\ /\ X_2\ X_1$ Swap order

$X \rightarrow$ watashi wa / I

$X \rightarrow$ hako wo / the box

$X \rightarrow$ akemasu / open

The Big Question

Where do the categories come from?

Answer #1: there are no categories!

$X \rightarrow X_1\ X_2\ /\ X_1\ X_2$ ⟵ Keep order

$X \rightarrow X_1\ X_2\ /\ X_2\ X_1$ ⟵ Swap order

$X \rightarrow$ watashi wa / I

$X \rightarrow$ hako wo / the box

$X \rightarrow$ akemasu / open

The Big Question

Where do the categories come from?

Answer #1: there are no categories!

$X \rightarrow X_1\ X_2\ /\ X_1\ X_2$ ← Keep order

$X \rightarrow X_1\ X_2\ /\ X_2\ X_1$ ← Swap order

$X \rightarrow$ watashi wa / I

$X \rightarrow$ hako wo / the box

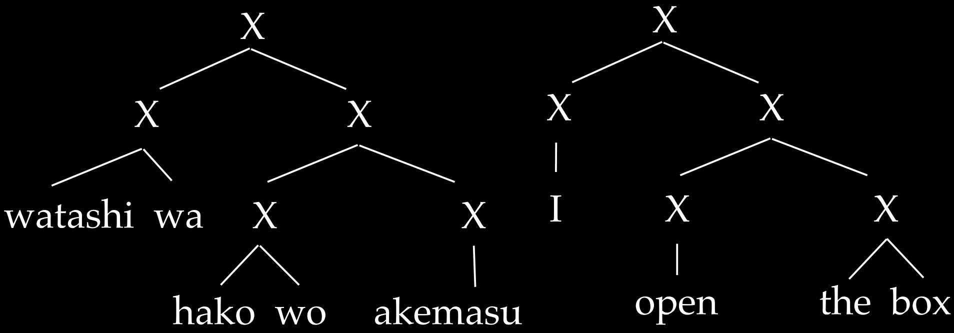
$X \rightarrow$ akemasu / open

Translate words or phrases

The Big Question

Where do the categories come from?

Answer #1: there are no categories!



Inversion Transduction Grammar

Inversion Transduction Grammar

Parsing is polynomial. We must be giving up *something* in order to acheive polynomial complexity.

Inversion Transduction Grammar

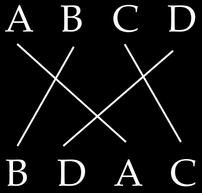
Parsing is polynomial. We must be giving up *something* in order to acheive polynomial complexity.

A B C D

B D A C

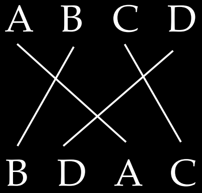
Inversion Transduction Grammar

Parsing is polynomial. We must be giving up *something* in order to acheive polynomial complexity.



Inversion Transduction Grammar

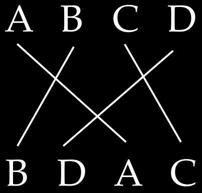
Parsing is polynomial. We must be giving up *something* in order to acheive polynomial complexity.



ITG cannot produce this kind of reordering.

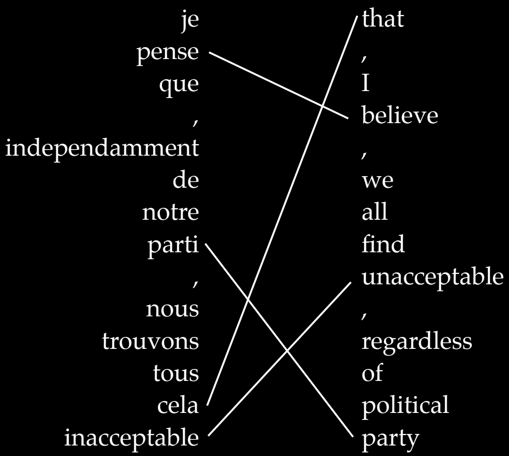
Inversion Transduction Grammar

Parsing is polynomial. We must be giving up *something* in order to acheive polynomial complexity.



ITG cannot produce this kind of reordering.
Does this matter? Do such reorderings occur in real data?

Inversion Transduction Grammar

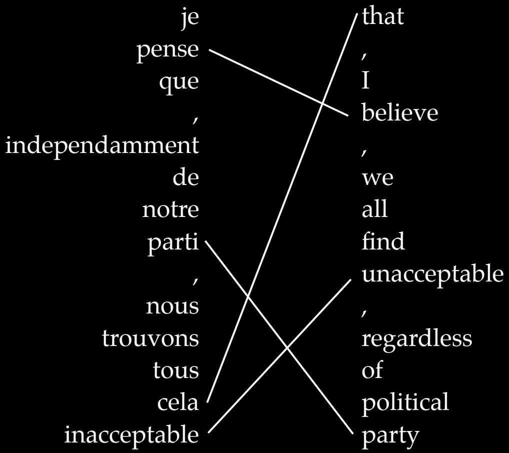


ITG cannot produce this kind of reordering.

Does this matter? Do such reorderings occur in real data?

YES!

Inversion Transduction Grammar



ITG cannot produce this kind of reordering.

Does this matter? Do such reorderings occur in real data?

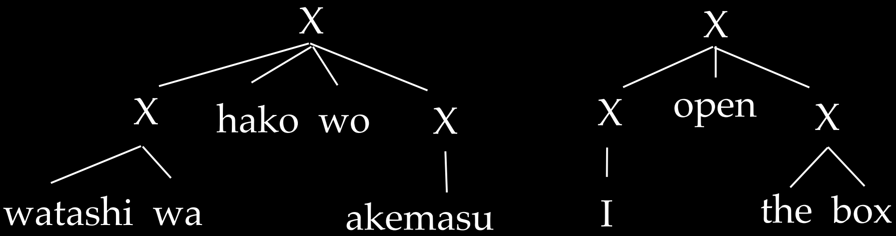
YES! (but they're very rare)

Hierarchical Phrase-Based Translation

$X \rightarrow X_1$ hako wo X_2 / X_1 open X_2

$X \rightarrow$ hako wo / the box

$X \rightarrow$ akemasu / open



The Big Question

Where do the categories come from?

The Big Question

Where do the categories come from?

Answer #2: from a parser.

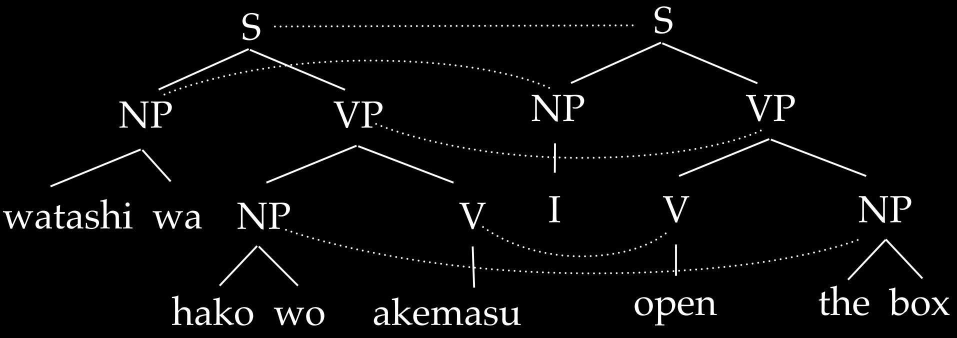
The Big Question

Where do the categories come from?

Answer #2: from a parser.

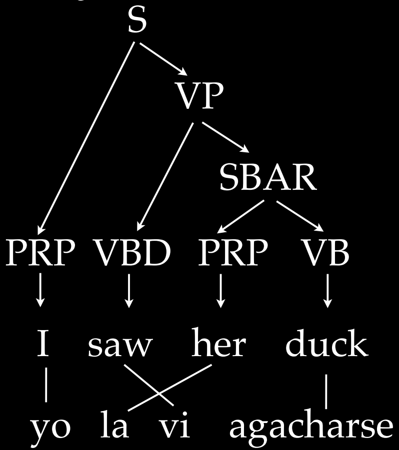
$$S \rightarrow NP_1\ VP_2\ /\ NP_1\ VP_2$$
$$NP \rightarrow \text{watashi wa}\ /\ \text{I}$$
$$NP \rightarrow \text{hako wo}\ /\ \text{the box}$$
$$VP \rightarrow NP_1\ V_2\ /\ V_1\ NP_2$$
$$V \rightarrow \text{akemasu}\ /\ \text{open}$$

Syntax-based Translation



Are reorderings in real data consistent with isomorphisms on linguistic parse trees?

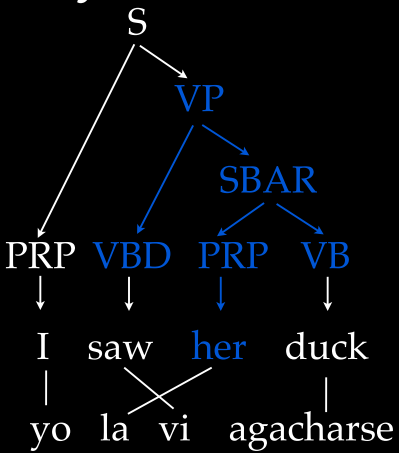
Syntax-based Translation



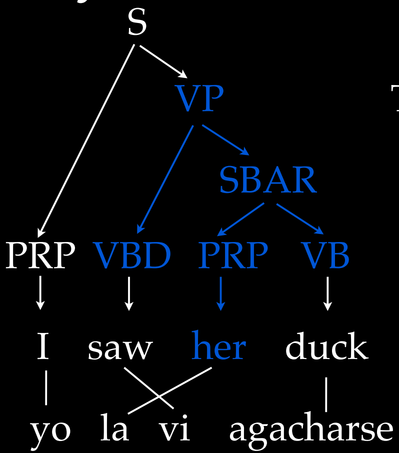
Are reorderings in real data consistent with isomorphisms on linguistic parse trees?

Of course not.

Syntax-based Translation

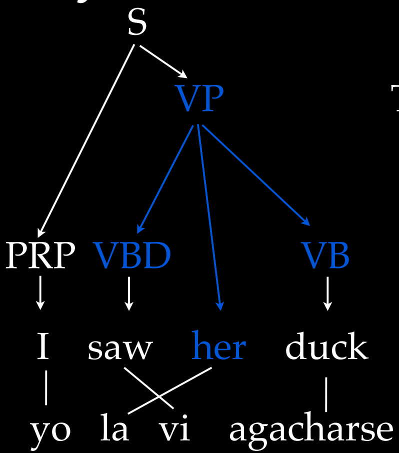


Syntax-based Translation



Tree substitution grammar

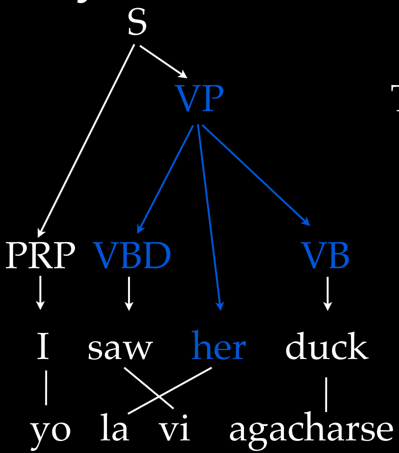
Syntax-based Translation



Tree substitution grammar

weakly equivalent SCFG

Syntax-based Translation



Tree substitution grammar

weakly equivalent SCFG

$$\mathrm{VBD} \rightarrow \text{saw} \ / \ \text{vi}$$

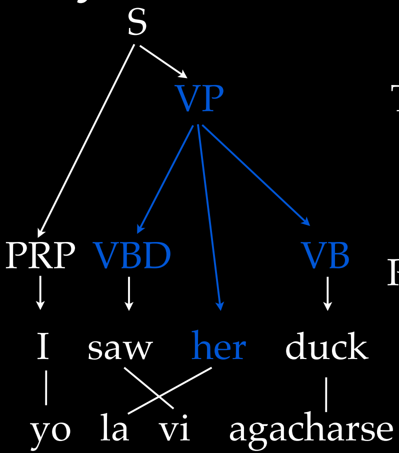
$$\mathrm{VB} \rightarrow \text{duck} \ / \ \text{agacharse}$$

$$\mathrm{S} \rightarrow \mathrm{PRP}_1\ \mathrm{VP}_2 \ / \ \mathrm{PRP}_1\ \mathrm{VP}_2$$

$$\mathrm{PRP} \rightarrow \text{I} \ / \ \text{yo}$$

$$\mathrm{VP} \rightarrow \mathrm{VBD}_1\ \text{her}\ \mathrm{VB}_2 \ / \ \text{la}\ \mathrm{VBD}_1\ \mathrm{VB}_2$$

Syntax-based Translation



Tree substitution grammar

weakly equivalent SCFG

Problem: we need a parser!

$$\text{VBD} \rightarrow \text{saw} \;/\; \text{vi}$$

$$\text{VB} \rightarrow \text{duck} \;/\; \text{agacharse}$$

$$\text{S} \rightarrow \text{PRP}_1\ \text{VP}_2 \;/\; \text{PRP}_1\ \text{VP}_2$$

$$\text{PRP} \rightarrow \text{I} \;/\; \text{yo}$$

$$\text{VP} \rightarrow \text{VBD}_1\ \text{her}\ \text{VB}_2 \;/\; \text{la}\ \text{VBD}_1\ \text{VB}_2$$

The Big Question

Where do the categories come from?

The Big Question

Where do the categories come from?

Answer #3: they are automatically induced!

The Big Question

Where do the categories come from?

Answer #3: they are automatically induced!

This is an area of active research.

www.clsp.jhu.edu/workshops/ws10/groups/msgismt/

Outline

More has been written about machine translation evaluation than about machine translation itself.

- Yorrick Wilks

Human evaluation

Have annotators read and assess translations

- **adequacy**: "$\boldsymbol{e}$ covers the same content as $\boldsymbol{f}$"
- **fluency**: "$\boldsymbol{e}$ looks like English"

Problems with human evaluation

- People hate evaluating translation, especially bad translation.
- People don't tend to agree with each other.

You try it:

- A: furious nAgA on wednesday , the tribal minimum pur of ten schools also was burnt
- B: furious nAgA on wednesday the tribal pur mini ten schools of them was also burnt

Automatic evaluation

Automatic evaluation of MT is hard:

- There are many correct ways to translate something.
- Measuring fluency is an open problem.
- Measuring adequacy is even harder (AI-complete?)

Automatic evaluation

Automatic evaluation of MT is hard:

- There are many correct ways to translate something.
- Measuring fluency is an open problem.
- Measuring adequacy is even harder (AI-complete?)

That said...

- Computers don't mind the work.
- Rapid evaluation supports iterative development.
- Evaluation can use fancier models than translation itself, since only a few hypotheses must be considered.

BLEU

- Most widespread automatic evaluation statistic by far
- 0.0 (worst) -1.0 (best)
- Computes *n*-gram overlap of a hypothesis with **one or more references**
  - Weighted average of precisions
  - "Brevity penalty" that kicks in if the hypothesis translation is too short

Computing BLEU

- n-gram overlap is computed on a per-segment basis
- a reference length is determined for each segment: what is the closest length in the set of reference?
- Statistics are aggregated over the corpus

$$\mathrm{BLEU}_4 = e^{\max\{|h|-r,0\}} \prod_{n=1}^{4} prec(n)^{1/4}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{1}{1}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{1}{1+1}$$

hypothesis: ' extension of isi in uttar pradesh '

ref 1: ' isi 's expansion in uttar pradesh '

ref 2: ' the spread of isi in uttar pradesh '

ref 3: ' isi spreading in uttar pradesh '

ref 4: the spread of isi in uttar pradesh

$$\mathrm{prec}(1) = \frac{1+1}{1+1+1}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{1+1+1}{1+1+1+1}$$

hypothesis:	‘ extension of isi in uttar pradesh ’
ref 1:	‘ isi ’s expansion in uttar pradesh ’
ref 2:	‘ the spread of isi in uttar pradesh ’
ref 3:	‘ isi spreading in uttar pradesh ’
ref 4:	the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{1 + 1 + 1 + 1}{1 + 1 + 1 + 1 + 1}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{|+|+|+|+|}{|+|+|+|+|+|}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{1+1+1+1+1}{1+1+1+1+1+1}$$

hypothesis:	‘ extension of isi in uttar pradesh ’
ref 1:	‘ isi ’s expansion in uttar pradesh ’
ref 2:	‘ the spread of isi in uttar pradesh ’
ref 3:	‘ isi spreading in uttar pradesh ’
ref 4:	the spread of isi in uttar pradesh

$$\mathrm{prec}(1) = \frac{1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\mathrm{prec}(1) = \frac{1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875$$

$$\mathrm{prec}(2) = \frac{}{1}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{|+|+|+|+|+|+|}{|+|+|+|+|+|+|+|} = 0.875$$

$$\text{prec}(2) = \frac{\qquad}{|+|}$$

hypothesis: ' extension of isi in uttar pradesh '

ref 1: ' isi 's expansion in uttar pradesh '

ref 2: ' the spread of isi in uttar pradesh '

ref 3: ' isi spreading in uttar pradesh '

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875$$

$$\text{prec}(2) = \frac{1}{1+1+1}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\mathrm{prec}(1) = \frac{1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875$$

$$\mathrm{prec}(2) = \frac{1+1}{1+1+1+1}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\mathrm{prec}(1) = \frac{|+|+|+|+|+|+|}{|+|+|+|+|+|+|+|} = 0.875$$

$$\mathrm{prec}(2) = \frac{|+|+|}{|+|+|+|+|}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\mathrm{prec}(1) = \frac{|+|+|+|+|+|+|}{|+|+|+|+|+|+|+|} = 0.875$$

$$\mathrm{prec}(2) = \frac{|+|+|+|}{|+|+|+|+|+|}$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{|+|+|+|+|+|+|}{|+|+|+|+|+|+|+|} = 0.875$$

$$\text{prec}(2) = \frac{|+|+|+|+|}{|+|+|+|+|+|+|} = 0.714$$

hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

$$\text{prec}(1) = \frac{|+|+|+|+|+|+|}{|+|+|+|+|+|+|+|} = 0.875$$

$$\text{prec}(2) = \frac{|+|+|+|+|}{|+|+|+|+|+|+|} = 0.714$$

$$\text{prec}(3) = 0.666$$

$$\text{prec}(4) = 0.6$$

Outline

What you need to do MT

- Parallel corpus
- Word alignment
- Language modeling
- Decoder

Parallel corpora

- The Linguistic Data Consortium will sell you:
  - United Nations data
  - Canadian Hansards
  - Hong Kong laws parallel text
  - Parallel newswire
  - `http://projects.ldc.upenn.edu/TIDES/`
- Or, you can download:
  EuroParl `http://www.statmt.org/europarl/`

Word alignment

- Giza++ is an open-source implementation of the IBM models
- `http://code.google.com/p/giza-pp/`

Language modeling

SRILM (Stanford Research Institute Language Model)

- Developed for speech recognition, but works for MT
- All kinds of fancy smoothing algorithms
- `http://www.speech.sri.com/projects/srilm/`

Decoder

- cdec (`http://cdec-decoder.org/`)
- Moses (`http://www.statmt.org/moses/`)
  - Contains code for the entire MT pipeline, including decoding.
  - Decent-looking documentation
  - Doing MT for a course project might be amibitious, but you might be able to experiment with one of the components and leave the rest as black boxes.

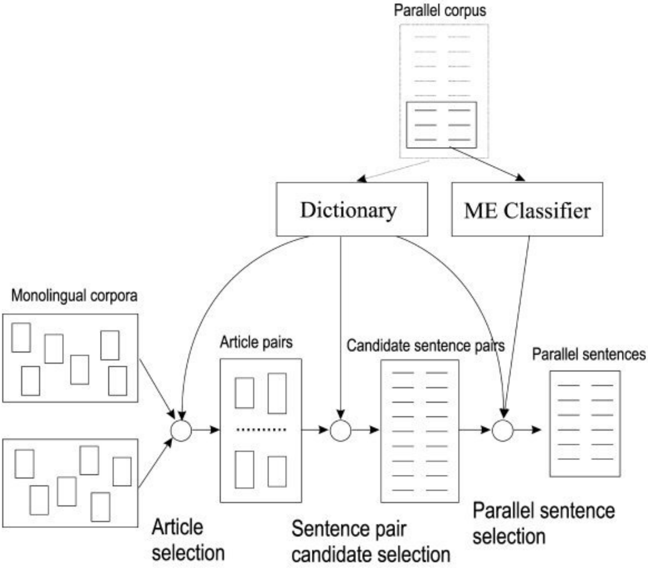
Outline

What can you do without much parallel data?

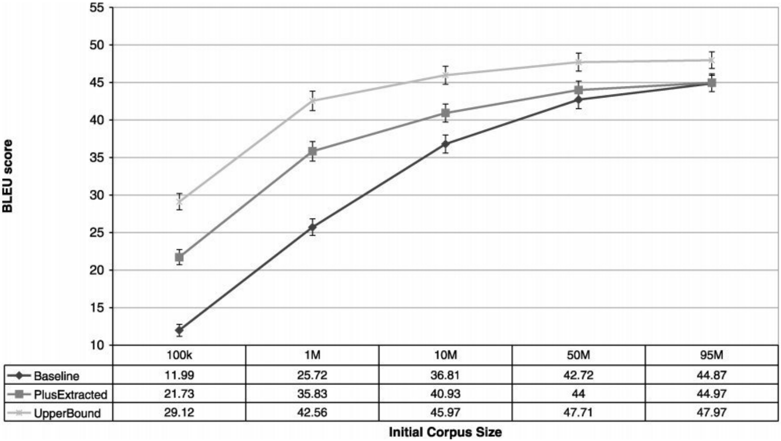
Munteneau and Marcu (2005), “Improving Machine Translation Performance by Exploiting Non-Parallel Corpora”

- suppose we only have a few aligned sentences
- but lots of possibly-aligned documents
- idea: use the parallel corpus to train a system to find more parallel documents and sentences.

What can you do without much parallel data?



What can you do without much parallel data?



	100k	1M	10M	50M	95M
Baseline	11.99	25.72	36.81	42.72	44.87
PlusExtracted	21.73	35.83	40.93	44	44.97
UpperBound	29.12	42.56	45.97	47.71	47.97

Wikipedia as a parallel corpus

Smith, Quirk, and Toutanova, "Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment"

- Thanks to Wikipedia, aligned *documents* are easy to obtain for many language pairs:

French 496K	German 488K	Polish 384K	Italian 380K	Dutch 357K	Portuguese 323K	Spanish 311K	Japanese 252K
Russian 232K	Swedish 197K	Finnish 146K	Chinese 142K	Norwegian 141K	Volapük 106K	Catalan 103K	Czech 87K

Table 1: Number of aligned bilingual articles in Wikipedia by language (paired with English).

- Similar idea to M&M: train a model (CRF) to identify aligned sentences, add these to the MT system.
- Note: you still need some sentence-aligned data to train your initial model.

Wikipedia as a parallel corpus

For "medium" sized-corpora, adding wikipedia sentences helps.

Language pair	Training data	Dev A	Test A	Wikitest
Spanish-English	Medium	32.6	30.5	33.0
	Medium+Wiki	36.7 (+4.1)	33.8 (+3.3)	39.1 (+6.1)
	Large	39.2	**37.4**	38.9
	Large+Wiki	**39.5** (+0.3)	37.3 (-0.1)	**41.1** (+2.2)
German-English	Medium	28.7	26.6	13.0
	Medium+Wiki	31.5 (+2.8)	29.6 (+3.0)	18.2 (+5.2)
	Large	**35.0**	33.7	17.1
	Large+Wiki	34.8 (-0.2)	**33.9** (+0.2)	**20.2** (+3.1)
Bulgarian-English	Medium	36.9	26.0	27.8
	Medium+Wiki	37.9 (+1.0)	27.6 (+1.6)	37.9 (+10.1)
	Large	**51.7**	**49.6**	36.0
	Large+Wiki	**51.7**(+0.0)	49.4 (-0.2)	**39.5**(+3.5)

Wikipedia as a parallel corpus

For "medium" sized-corpora, adding wikipedia sentences helps.

Language pair	Training data	Dev A	Test A	Wikitest
Spanish-English	Medium	32.6	30.5	33.0
	Medium+Wiki	36.7 (+4.1)	33.8 (+3.3)	39.1 (+6.1)
	Large	39.2	**37.4**	38.9
	Large+Wiki	**39.5** (+0.3)	37.3 (-0.1)	**41.1** (+2.2)
German-English	Medium	28.7	26.6	13.0
	Medium+Wiki	31.5 (+2.8)	29.6 (+3.0)	18.2 (+5.2)
	Large	**35.0**	33.7	17.1
	Large+Wiki	34.8 (-0.2)	**33.9** (+0.2)	**20.2** (+3.1)
Bulgarian-English	Medium	36.9	26.0	27.8
	Medium+Wiki	37.9 (+1.0)	27.6 (+1.6)	37.9 (+10.1)
	Large	**51.7**	**49.6**	36.0
	Large+Wiki	**51.7**(+0.0)	49.4 (-0.2)	**39.5**(+3.5)

If you're evaluating on a test set of Wikipedia,
it's helpful to add Wikipedia to your training set.

What can you do without ANY parallel data?

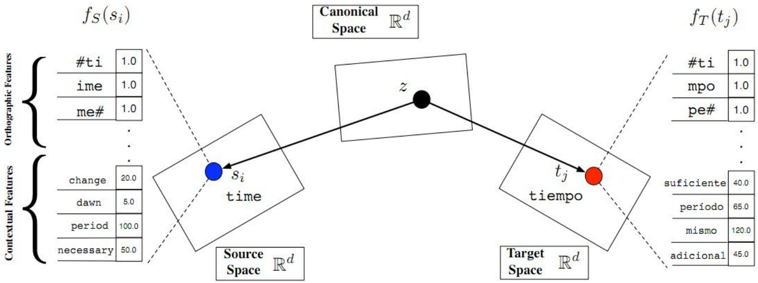
Haghighi et al, "Learning Bilingual Lexicons from Monolingual Corpora"

- Latent space models capture word similarity by factoring a matrix of local context counts, $f_i \approx Wz_i$
  - f_i is the feature vector for word i (e.g., context counts)
  - z_i is the latent representation for word i
  - W is some projection matrix

What can you do without ANY parallel data?

Haghighi et al, "Learning Bilingual Lexicons from Monolingual Corpora"

- Latent space models capture word similarity by factoring a matrix of local context counts, $f_i \approx W z_i$
  - f_i is the feature vector for word i (e.g., context counts)
  - z_i is the latent representation for word i
  - W is some projection matrix
- Key idea: apply this to two languages at once, and automatically discover a translation lexicon.
- Aligned words should have identical z_i.
  - $f_S(s_i) \approx W_S z_i$
  - $f_T(t_i) \approx W_T z_i$



$f_S(s_i)$
Canonical Space $\mathbb{R}^d$
$f_T(t_j)$
Orthographic Features
Contextual Features
#ti 1.0
ime 1.0
me# 1.0
change 20.0
dawn 5.0
period 100.0
necessary 50.0
z
s_i
time
t_j
tiempo
Source Space $\mathbb{R}^d$
Target Space $\mathbb{R}^d$
#ti 1.0
mpo 1.0
pe# 1.0
suficiente 40.0
período 65.0
mismo 120.0
adicional 45.0

Learning the lexicon

- Start with a seed set of 100 words
- Viterbi (hard) EM. Iterate:
  - Compute W_S, W_T, z_i for all i
  - Find the best alignment (bipartite mapping)

Results

- The approach works well for English and French:

Setting	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	Best-F_1
EDITDIST	58.6	62.6	61.1	—-	47.4
ORTHO	76.0	81.3	80.1	52.3	55.0
CONTEXT	**91.1**	81.3	80.2	65.3	58.0
MCCA	87.2	**89.7**	**89.0**	**89.7**	**72.0**

- Not so well for Chinese and Arabic.
- Having a good seed helps.
  (but inducing seeds from edit distance works ok)
- Having similar corpora (Wikipedia vs Gigaword) helps.

Results

(b) English-French

Rank	Source	Target	Correct
3.	xenophobia	xénophobie	Y
4.	corruption	corruption	Y
5.	subsidiarity	subsidiarité	Y
6.	programme	programme-cadre	N
8.	traceability	traçabilité	Y

(c) English-Chinese

Rank	Source	Target	Correct
1.	prices	价格	Y
2.	network	网络	Y
3.	population	人口	Y
4.	reporter	孙	N

Results

(b) Interesting Incorrect Pairs

liberal	partido
Kirkhope	Gorsel
action	reacción
Albanians	Bosnia
a.m.	horas
Netherlands	Bretaña

Recap

- Key pieces for machine translation:
  - **Alignment**: word-to-word, phrase-to-phrase, or syntactic.
  - **Estimation**: MLE for noisy-channel, MERT for component weights.
  - **Decoding**: Reordering makes beam search crucial.
  - **Evaluation**: BLEU counts n-gram overlap

Recap

- Key pieces for machine translation:
  - **Alignment**: word-to-word, phrase-to-phrase, or syntactic.
  - **Estimation**: MLE for noisy-channel, MERT for component weights.
  - **Decoding**: Reordering makes beam search crucial.
  - **Evaluation**: BLEU counts n-gram overlap
- Syntactic machine translation is mainly based on **synchronous context-free grammars**.

Recap

- Key pieces for machine translation:
  - **Alignment**: word-to-word, phrase-to-phrase, or syntactic.
  - **Estimation**: MLE for noisy-channel, MERT for component weights.
  - **Decoding**: Reordering makes beam search crucial.
  - **Evaluation**: BLEU counts n-gram overlap
- Syntactic machine translation is mainly based on **synchronous context-free grammars**.
- Translation without parallel text:
  - Find parallel documents on the web.
  - Use Wikipedia, learn to classify parallel sentences.
  - Use latent space models to build bilingual lexicons.