Machine Translation

Overview

- Challenges in machine translation
- Classical machine translation
- Statistical MT: phrase-based models

An Example

French-to-English from Google translate:

Dans une évaluation stratégique du conflit afghan remise à Robert Gates le 30 août, et révélée par le Washington Post lundi, le général McChrystal prévient que sans augmentation des moyens militaires en Afghanistan, la coalition risquait d'y subir "un échec".

$$\Downarrow$$

In a strategic assessment of the Afghanistan conflict given to Robert Gates August 30, revealed by The Washington Post Monday, General McChrystal warned that without an increase in military resources in Afghanistan, the coalition might undergo a "failure."

Lexical Ambiguity

Example 1:

book the flight $\Rightarrow$ reservar

read the book $\Rightarrow$ libro

Example 2:

the box was in the pen

the pen was on the table

Example 3:

kill a man $\Rightarrow$ matar

kill a process $\Rightarrow$ acabar

Differing Word Orders

- English word order is *subject – verb – object*
- Japanese word order is *subject – object – verb*

English: IBM bought Lotus
Japanese: *IBM Lotus bought*

English: Sources said that IBM bought Lotus yesterday
Japanese: *Sources yesterday IBM Lotus bought that said*

Syntactic Structure is not Preserved Across Translations

The bottle floated into the cave

$\Downarrow$

La botella entro a la cuerva flotando
(the bottle entered the cave floating)

Syntactic Ambiguity Causes Problems

John hit the dog with the stick

$\Downarrow$

John golpeo el perro con el palo/que tenia el palo

Pronoun Resolution

The computer outputs the data; it is fast.

$\Downarrow$

La computadora imprime los datos; es rapida

The computer outputs the data; it is stored in ascii.

$\Downarrow$

La computadora imprime los datos; estan almacendos en ascii

Differing Treatments of Tense

From Dorr et. al 1998:

Mary went to Mexico. During her stay she learned Spanish.

Went $\Rightarrow$ iba (simple past/preterit)

Mary went to Mexico. When she returned she started to speak Spanish.

Went $\Rightarrow$ fue (ongoing past/imperfect)

Overview

- Challenges in machine translation
- Classical machine translation
- Statistical MT: phrase-based models

Transfer-Based Approaches

- Three phases in translation:

  Analysis: Analyze the source language sentence; for example, build a syntactic analysis of the source language sentence.

  Transfer: Convert the source-language parse tree to a target-language parse tree.

  Generation: Convert the target-language parse tree to an output sentence.

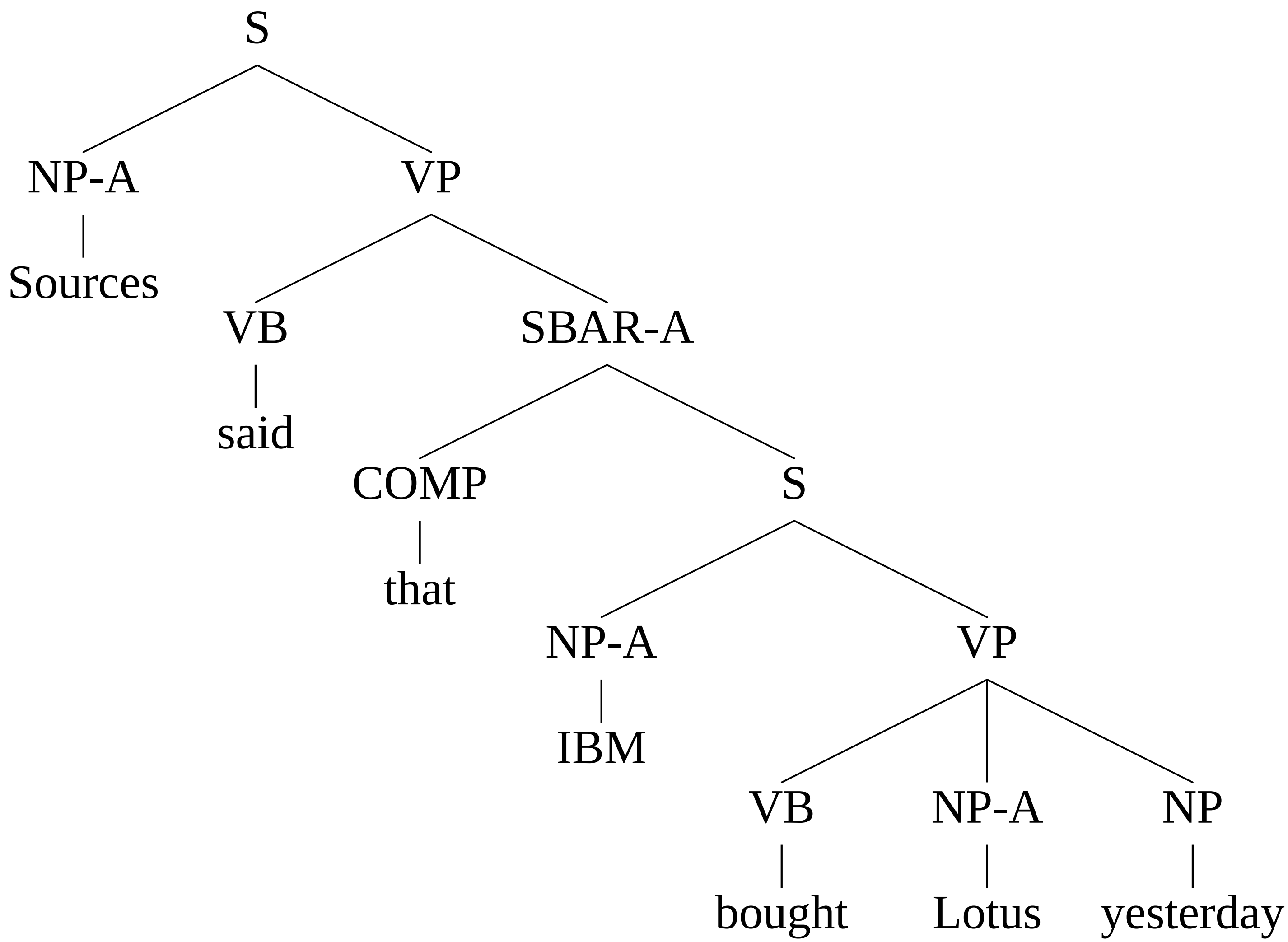
Transfer-Based Approaches

- The “parse trees” involved can vary from shallow analyses to much deeper analyses (even semantic representations).

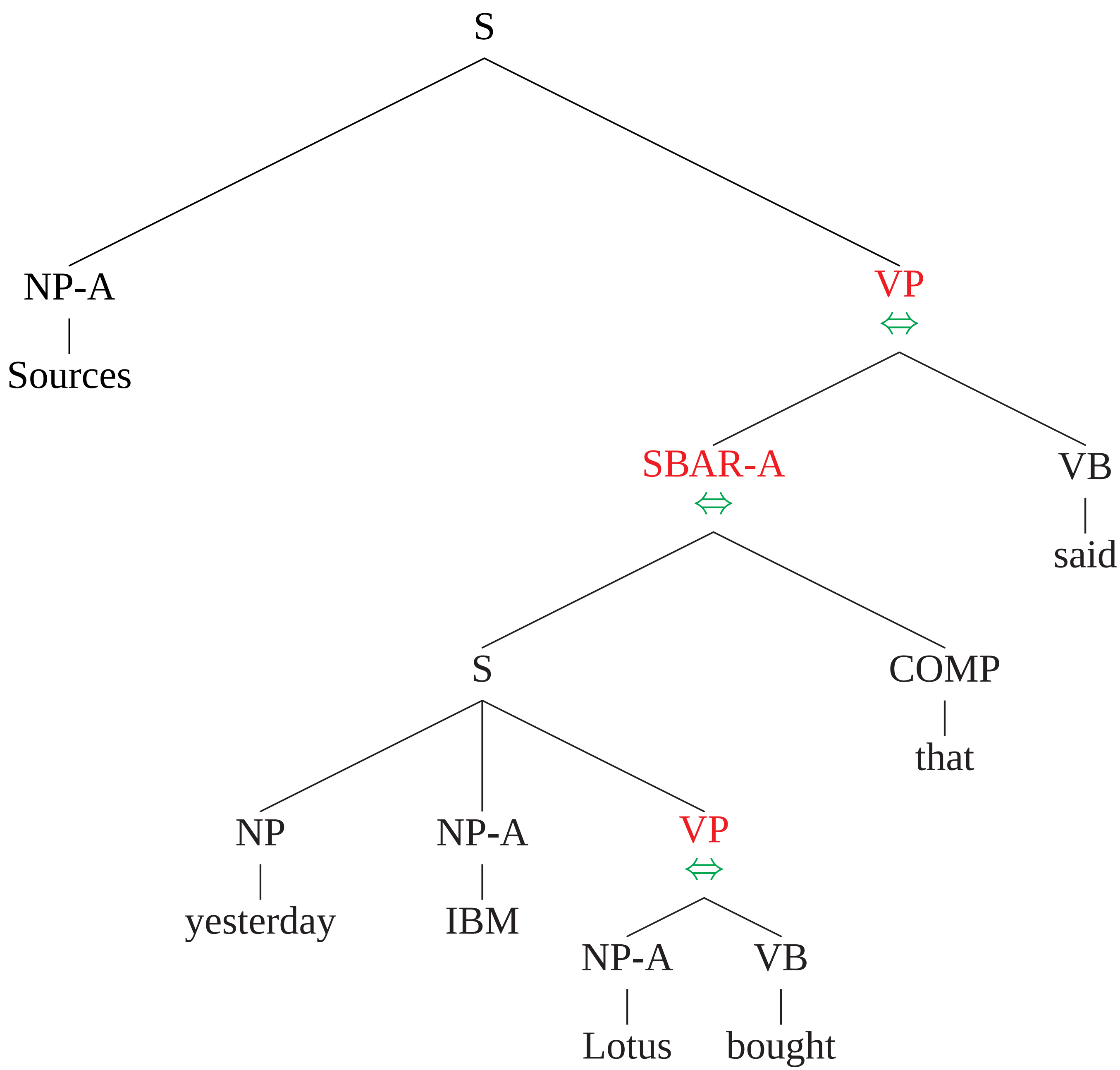
- The transfer rules might look quite similar to the rules for direct translation systems. But they can now operate on syntactic structures.

- It’s easier with these approaches to handle long-distance reorderings

- The *Systran* systems are a classic example of this approach



⇒ Japanese: *Sources yesterday IBM Lotus bought that said*


S
NP-A
Sources
VP
⇔
SBAR-A
⇔
VB
said
S
COMP
that
NP
yesterday
NP-A
IBM
VP
⇔
NP-A
Lotus
VB
bought

Interlingua-Based Translation

- Two phases in translation:

  Analysis: Analyze the source language sentence into a (language-independent) representation of its meaning.

  Generation: Convert the meaning representation into an output sentence.

Interlingua-Based Translation

One Advantage: If we want to build a translation system that translates between n languages, we need to develop n analysis and generation systems. With a transfer based system, we'd need to develop $O(n^2)$ sets of translation rules.

Disadvantage: What would a language-independent representation look like?

Interlingua-Based Translation

- How to represent different concepts in an interlingua?
- Different languages break down concepts in quite different ways:

  German has two words for *wall*: one for an internal wall, one for a wall that is outside

  Japanese has two words for *brother*: one for an elder brother, one for a younger brother

  Spanish has two words for *leg*: *pierna* for a human's leg, *pata* for an animal's leg, or the leg of a table

- An interlingua might end up simple being an intersection of these different ways of breaking down concepts, but that doesn't seem very satisfactory...

Overview

- Challenges in machine translation
- Classical machine translation
- Statistical MT: phrase-based models

Machine Translation: Background

- A long-standing problem in Artificial Intelligence
- Much work in 1970's, 1980's, on rule-based systems (e.g., Systran)
- In early 1990's, IBM researchers began work on statistical systems

  Basic idea: given a *parallel corpus* of example translations, can we induce a statistical model of a translation process between two languages?

  Goes back to ideas from Warren Weaver (1949), who had the idea of framing machine translation as a cryptographic problem.

Phrase-Based Models

- First stage in training a phrase-based model is extraction of a *phrase-based (PB) lexicon*

- A PB lexicon pairs strings in one language with strings in another language, e.g.,

nach Kanada	$\leftrightarrow$	in Canada
zur Konferenz	$\leftrightarrow$	to the conference
Morgen	$\leftrightarrow$	tomorrow
fliege	$\leftrightarrow$	will fly
...		

An Example (from tutorial by Koehn and Knight)

- A training example (Spanish/English sentence pair):

  Spanish: Maria no daba una bofetada a la bruja verde

  English: Mary did not slap the green witch

- Some (not all) phrase pairs extracted from this example:

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

- We'll see how to do this using *alignments* from the IBM models (e.g., from IBM model 2)

An Example Alignment

French:
le conseil a rendu son avis , et nous devons à présent adopter un nouvel avis sur la base de la première position .

English:
the council has stated its position , and now , on the basis of the first position , we again have to give our opinion .

Alignment:
the/le council/conseil has/à stated/rendu its/son position/avis ,/, and/et now/présent ,/NULL on/sur the/le basis/base of/de the/la first/première position/position ,/NULL we/nous again/NULL have/devons to/a give/adopter our/nouvel opinion/avis ./.

Recovered using the EM algorithm, typically using models developed at IBM in the early 1990s

Representation as an Alignment Matrix

	Maria	no	daba	una	bof’	a	la	bruja	verde
Mary	●								
did		●							
not		●							
slap			●	●	●				
the						●	●		
green									●
witch								●	

Note that the alignment is potentially many-to-many: multiple Spanish words can be aligned to a single English word, and vice versa.

Extracting Phrase Pairs from the Alignment Matrix

	Maria	no	daba	una	bof'	a	la	bruja	verde
Mary	●								
did		●							
not		●							
slap			●	●	●				
the						●	●		
green									●
witch								●	

- A phrase-pair consists of a sequence of English words, e, paired with a sequence of foreign words, f
- A phrase-pair (e, f) is *consistent* if there are no words in f aligned to words outside e, and there are no words in e aligned to words outside f
  e.g., (Mary did not, Maria no) is consistent. (Mary did, Maria no) is *not* consistent: "no" is aligned to "not", which is not in the string "Mary did"
- We extract all consistent phrase pairs from the training example. See Koehn, EACL 2006 tutorial, **pages 103-108** for illustration.

Probabilities for Phrase Pairs

- For any phrase pair (f, e) extracted from the training data, we can calculate

$$P(f|e) = \frac{Count(f, e)}{Count(e)}$$

e.g.,

$$P(\text{daba una bofetada} \mid \text{slap}) = \frac{Count(\text{daba una bofetada}, \text{slap})}{Count(\text{slap})}$$

An Example Phrase Translation Table

An example from Koehn, EACL 2006 tutorial. (Note that we have $P(e|f)$ not $P(f|e)$ in this example.)

- Phrase Translations for *den Vorschlag*

English	P(e|f)	**English**	P(e|f)
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

$\boxed{\text{Today}}$

$\boxed{\text{Heute}}$ werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

$$\text{Score} = \underbrace{\log P(\text{Today} \mid \text{START})}_{\text{Language model}}$$

$$+ \underbrace{\log P(\text{Heute} \mid \text{Today})}_{\text{Phrase model}}$$

$$+ \underbrace{\log P(1\text{-}1 \mid 1\text{-}1)}_{\text{Distortion model}}$$

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

Today $\boxed{\text{we shall be}}$

Heute $\boxed{\text{werden wir}}$ uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

$$\text{Score} = \underbrace{\log P(\text{we shall be} \mid \text{today})}_{\text{Language model}}$$

$$+ \underbrace{\log P(\text{werden wir} \mid \text{we will be})}_{\text{Phrase model}}$$

$$+ \underbrace{\log P(\text{2-3} \mid \text{2-4})}_{\text{Distortion model}}$$

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

Today we shall be $\boxed{\text{debating}}$

Heute werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels

$\boxed{\text{diskutieren}}$

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

Today we shall be debating the reopening

Heute werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

Phrase-Based Systems: A Sketch

Translate using a greedy, left-to-right decoding method

Today we shall be debating the reopening of the Mont Blanc tunnel

Heute werden wir uber die Wiedereroffnung des Mont-Blanc-Tunnels diskutieren

Phrase-Based Systems: Formal Definitions

(following notation in Jurafsky and Martin, chapter 25)

- We'd like to translate a French string $\mathbf{f}$
- E is a sequence of l English phrases, $e_1, e_2, \ldots, e_l$.
  For example,

  $e_1 =$ Mary, $e_2 =$ did not, $e_3 =$ slap, $e_4 =$ the, $e_5 =$ green witch

  E defines a possible translation, in this case $e_1 e_2 \ldots e_5$ = *Mary did not slap the green witch.*
- F is a sequence of l foreign phrases, $f_1, f_2, \ldots, f_l$.
  For example,

  $f_1 =$ Maria, $f_2 =$ no, $f_3 =$ dio una bofetada, $f_4 =$ a la, $f_5 =$ bruja verde
- a_i for $i = 1 \ldots l$ is the position of the first word of f_i in $\mathbf{f}$. b_i for $i = 1 \ldots l$ is the position of the last word of f_i in $\mathbf{f}$.

Phrase-Based Systems: Formal Definitions

- We then have

$$Cost(E, F) = P(E) \prod_{i=1}^{l} P(f_i|e_i) d(a_i - b_{i-1})$$

- $P(E)$ is the language model score for the string defined by E
- $P(f_i|e_i)$ is the phrase-table probability for the i'th phrase pair
- $d(a_i - b_{i-1})$ is some probability/penalty for the distance between the i'th phrase and the $(i-1)$'th phrase. Usually, we define

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$$

  for some $\alpha < 1$.
- Note that this is *not* a coherent probability model

An Example

Position	1	2	3	4	5
English	Mary	did not	slap	the	green witch
Spanish	Maria	no	dio una bofetada	a la	bruja verde

In this case,

$$
\begin{aligned}
Cost(E, F) \quad = \quad & P_L(\text{Mary did not slap the green witch}) \times \\
& P(\text{Maria}|\text{Mary}) \times d(1) \times P(\text{no}|\text{did not}) \times d(1) \times \\
& P(\text{dio una bofetada}|\text{slap}) \times d(1) \times P(\text{a la}|\text{the}) \times d(1) \times \\
& P(\text{bruja verde}|\text{green witch}) \times d(1)
\end{aligned}
$$

P_L is the score from a language model

Another Example

Position	**1**	**2**	**3**	**4**	**5**	**6**
English	Mary	did not	slap	the	green	witch
Spanish	Maria	no	dio una bofetada	a la	verde	bruje

The original Spanish string was *Maria no dio una bofetada a la bruje verde*, so notice that the last two phrase pairs involve reordering

In this case,

$$
\begin{aligned}
Cost(E, F) \quad = \quad & P_L(\text{Mary did not slap the green witch}) \times \\
& P(\text{Maria}|\text{Mary}) \times d(1) \times P(\text{no}|\text{did not}) \times d(1) \times \\
& P(\text{dio una bofetada}|\text{slap}) \times d(1) \times P(\text{a la}|\text{the}) \times d(1) \times \\
& P(\text{verde}|\text{green}) \times d(2) \times P(\text{bruja}|\text{witch}) \times d(1)
\end{aligned}
$$

The Decoding Problem

- For a given foreign string $\mathbf{f}$, the decoding problem is to find

$$\arg\max_{(E,F)} Cost(E, F)$$

  where the $\arg\max$ is over all (E, F) pairs that are consistent with $\mathbf{f}$

- See Koehn tutorial, EACL 2006, slides 29–57
- See Jurafsky and Martin, Chapter 25, Figure 25.30
- See Jurafsky and Martin, Chapter 25, section 25.8