

What's New in Statistical Machine Translation

Kevin Knight

USC/Information Sciences Institute
USC/Computer Science Department



U A I 2 0 0 3

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

The classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

Hard -- ambiguous source words, target fluency, re-ordering

About \$10 billion spent annually on human translation.

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Three basic uses:

1. Communication (as in speech/chat translation)
2. Dissemination (publish your stuff to the world)
3. Assimilation (be able to understand what's out there)

Despite decades of work, not a lot of progress!

Progress

slide from C. Wayne, DARPA

2002

insistent Wednesday may
recurred her trips to Libya
tomorrow for flying

Cairo 6-4 (AFP) - an official
announced today in the
Egyptian lines company for
flying Tuesday is a company "
insistent for flying " may
resumed a consideration of a
day Wednesday tomorrow her
trips to Libya of Security Council
decision trace international the
imposed ban comment .

And said the official " the
institution sent a speech to
Ministry of Foreign Affairs of
lifting on Libya air , a situation
her receiving replying are so a
trip will pull to Libya a morning
Wednesday " .

2003

Egyptair Has Tomorrow to
Resume Its Flights to Libya

Cairo 4-6 (AFP) - said an official
at the Egyptian Aviation
Company today that the
company Egyptair may resume
as of tomorrow, Wednesday its
flights to Libya after the
International Security Council
resolution to the suspension of
the embargo imposed on Libya.

" The official said that the
company had sent a letter to the
Ministry of Foreign Affairs,
information on the lifting of the
air embargo on Libya, where it
had received a response, the
first take off a trip to Libya on
Wednesday morning " .

Caveats

- Good for news text
 - Need to test on other genres
- Good for assimilation
 - Dissemination requires publication quality
- Fortunately, only scratched the surface of good ideas

Why Is It Getting Better?

“I don’t want to talk about how we might do this task in the future. I’m going to talk about what we can do right now, with today’s technology.”

-- Marti Hearst

opening a talk on text segmentation

Why Is It Getting Better?

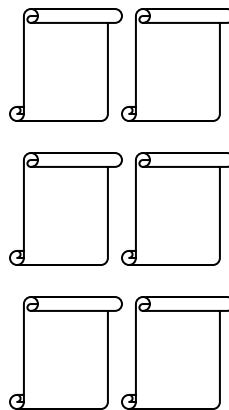
- Large-scale data resources
 - bilingual text
 - Linguistic modeling of the translation relation
 - can string x a translation of string y, or not?
 - Machine learning
 - Heuristic search
 - both in learning and runtime
 - Significant engineering of training data and software
 - 4 Gb memory & fast dual processor at \$5000
 - New automatic evaluation metric to drive day-to-day experimentation
 - Competition between sites on translation of same previously-unseen test documents
 - Collaborative sharing of complex software tools, intensive workshops
- since
2003
- 1990s
- 1990s
- 1990s
- 2003
- 2001
- 2002
- 1999

Data-Driven Machine Translation

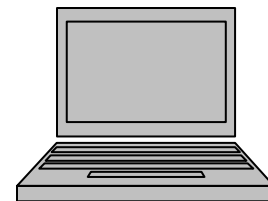
(Nagao 84,
Brown et al 88,
etc.)

Man, this is so boring.

Hmm, every time he sees
“banco”, he either types
“bank” or “bench” ... but if
he sees “banco de...”,
he always types “bank”,
never “bench”...



Translated documents



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrok hihok yorok klok kantok ok-yurp

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

It's Really Spanish/English

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

1a. Garcia and associates .

1b. Garcia y asociados .

7a. the clients and the associates are enemies .

7b. los clients y los asociados son enemigos .

2a. Carlos Garcia has three associates .

2b. Carlos Garcia tiene tres asociados .

8a. the company has three groups .

8b. la empresa tiene tres grupos .

3a. his associates are not strong .

3b. sus asociados no son fuertes .

9a. its groups are in Europe .

9b. sus grupos estan en Europa .

4a. Garcia has a company also .

4b. Garcia tambien tiene una empresa .

10a. the modern groups sell strong pharmaceuticals .

10b. los grupos modernos venden medicinas fuertes .

5a. its clients are angry .

5b. sus clientes estan enfadados .

11a. the groups do not sell zenzanine .

11b. los grupos no venden zanzanina .

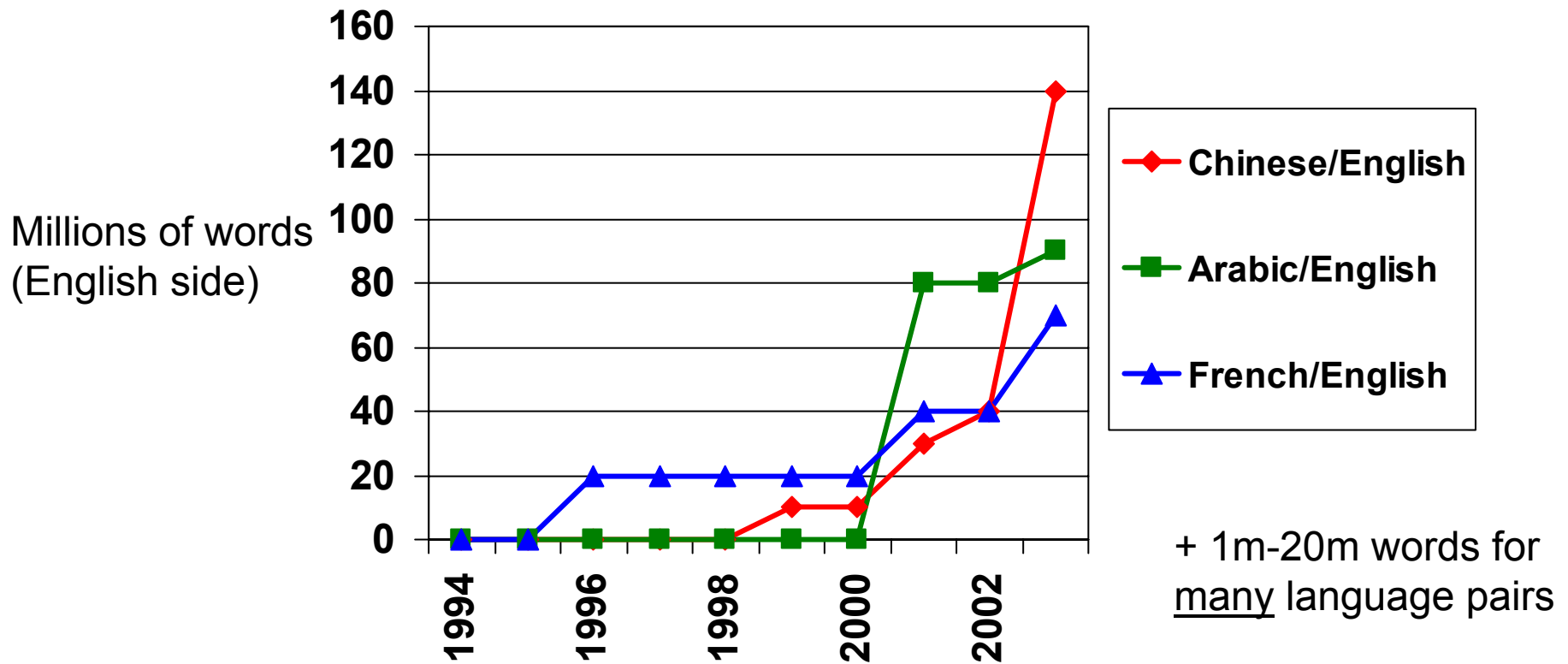
6a. the associates are also angry .

6b. los asociados tambien estan enfadados .

12a. the small groups are not modern .

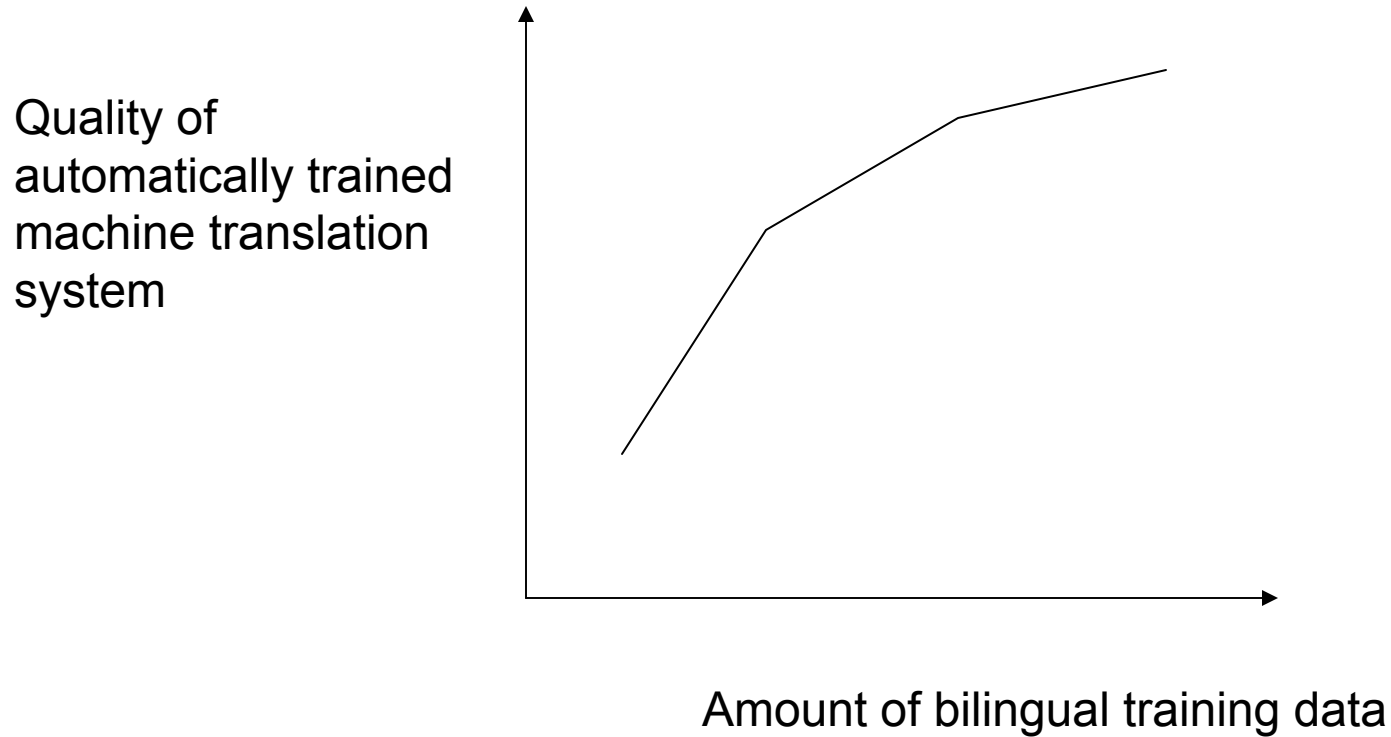
12b. los grupos pequenos no son modernos .

Ready-to-Use Bilingual Data



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

Possible to Do Learning Curves



BLEU Evaluation Metric

(Papineni et al 02)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision
 - Look for machine n-grams in the reference translation
 - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")
- Brevity penalty
 - Can't cheat by typing out just the single word "the"
- Multiple reference translations
 - To account for variation

Multiple Reference Translations

Reference translation 1:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Reference translation 2:

Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

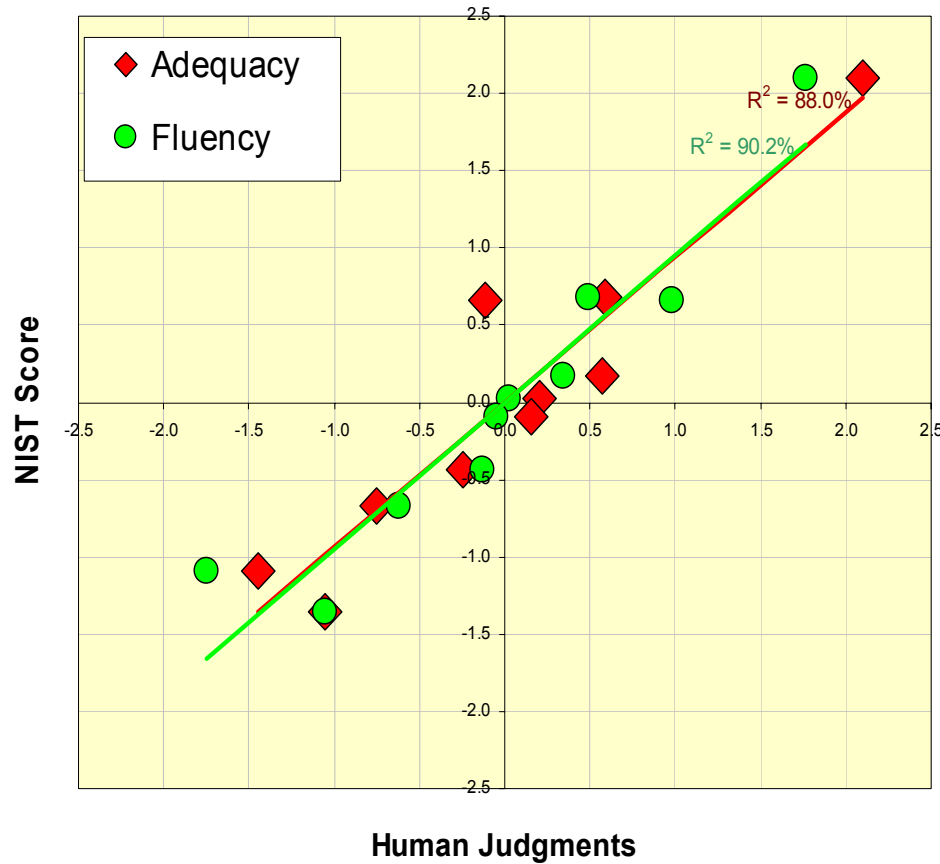
Reference translation 3:

The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:

US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

BLEU Tends to Predict Human Judgments



**Results of June 2002
DARPA evaluations of
MT quality**

**(Experiment by
George Doddington, NIST)**

BLEU in Action

枪手被警方击毙。

(Foreign Original)

the gunman was shot to death by the police .

(Reference Translation)

the gunman was police kill .

#1

wounded police jaya of

#2

the gunman was shot dead by the police .

#3

the gunman arrested by police kill .

#4

the gunmen were killed .

#5

the gunman was shot to death by the police .

#6

gunmen were killed by police ?SUB>0 ?SUB>0

#7

al by the police .

#8

the ringer is killed by the police .

#9

police killed the gunman .

#10

BLEU in Action

枪手被警方击毙。

(Foreign Original)

the gunman was shot to death by the police .

(Reference Translation)

the gunman was police kill .

#1

wounded police jaya of

#2

the gunman was shot dead by the police .

#3

the gunman arrested by police kill .

#4

the gunmen were killed .

#5

the gunman was shot to death by the police .

#6

gunmen were killed by police ?SUB>0 ?SUB>0

#7

al by the police .

#8

the ringer is killed by the police .

#9

police killed the gunman .

#10

green = 4-gram match (good!)

red = word not matched (bad!)

BLEU in Action

枪手被警方击毙。

(Foreign Original)

the gunman was shot to death by the police .

(Reference Translation)

the gunman was police kill .

#1 Machine

wounded police jaya of

#2 Machine

the gunman was shot dead by the police .

#3 Human

the gunman arrested by police kill .

#4 Machine

the gunmen were killed .

#5 Machine

the gunman was shot to death by the police .

#6 Human

gunmen were killed by police ?SUB>0 ?SUB>0

#7 Machine

al by the police .

#8 Machine

the ringer is killed by the police .

#9 Machine

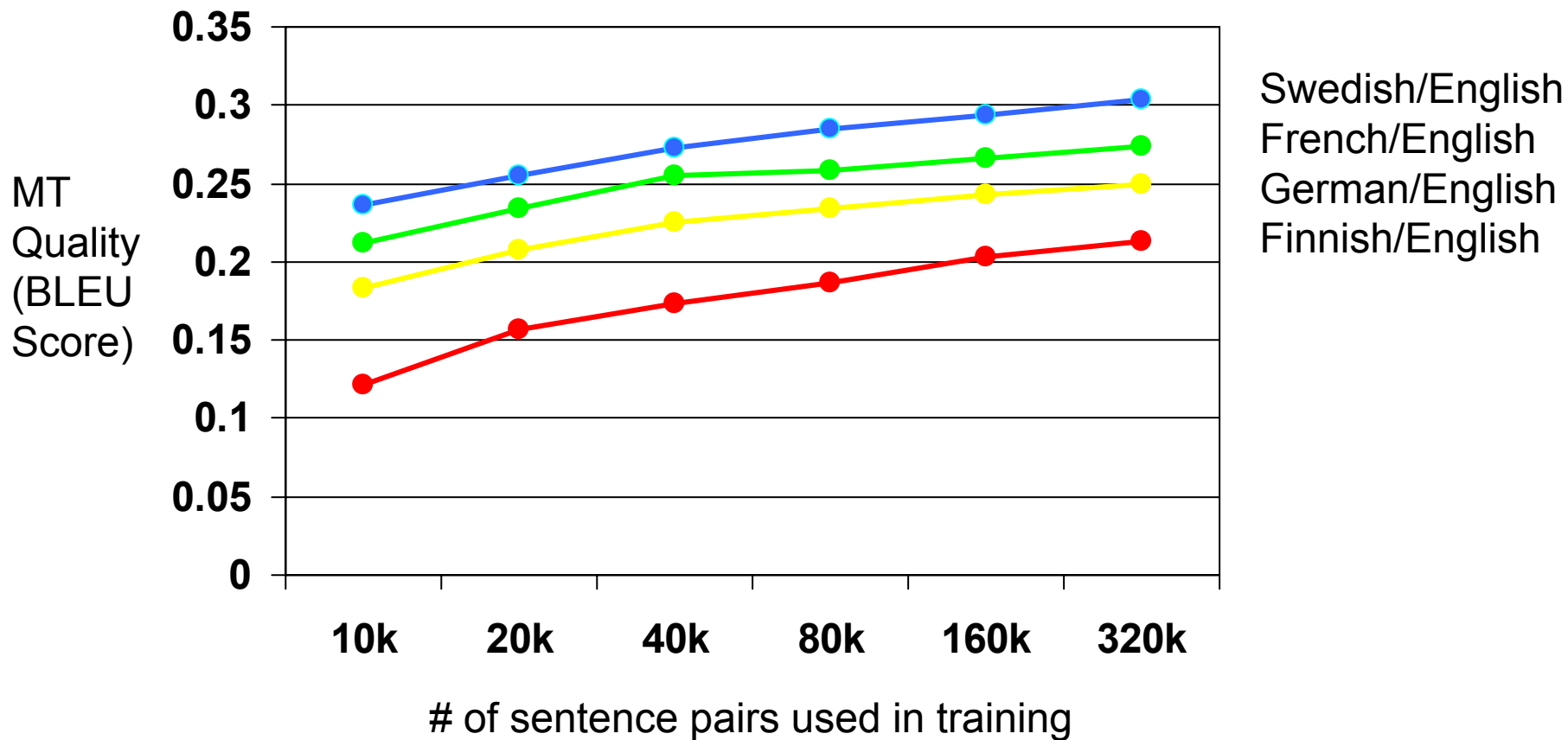
police killed the gunman .

#10 Human

green = 4-gram match (good!)

red = word not matched (bad!)

Sample Learning Curves

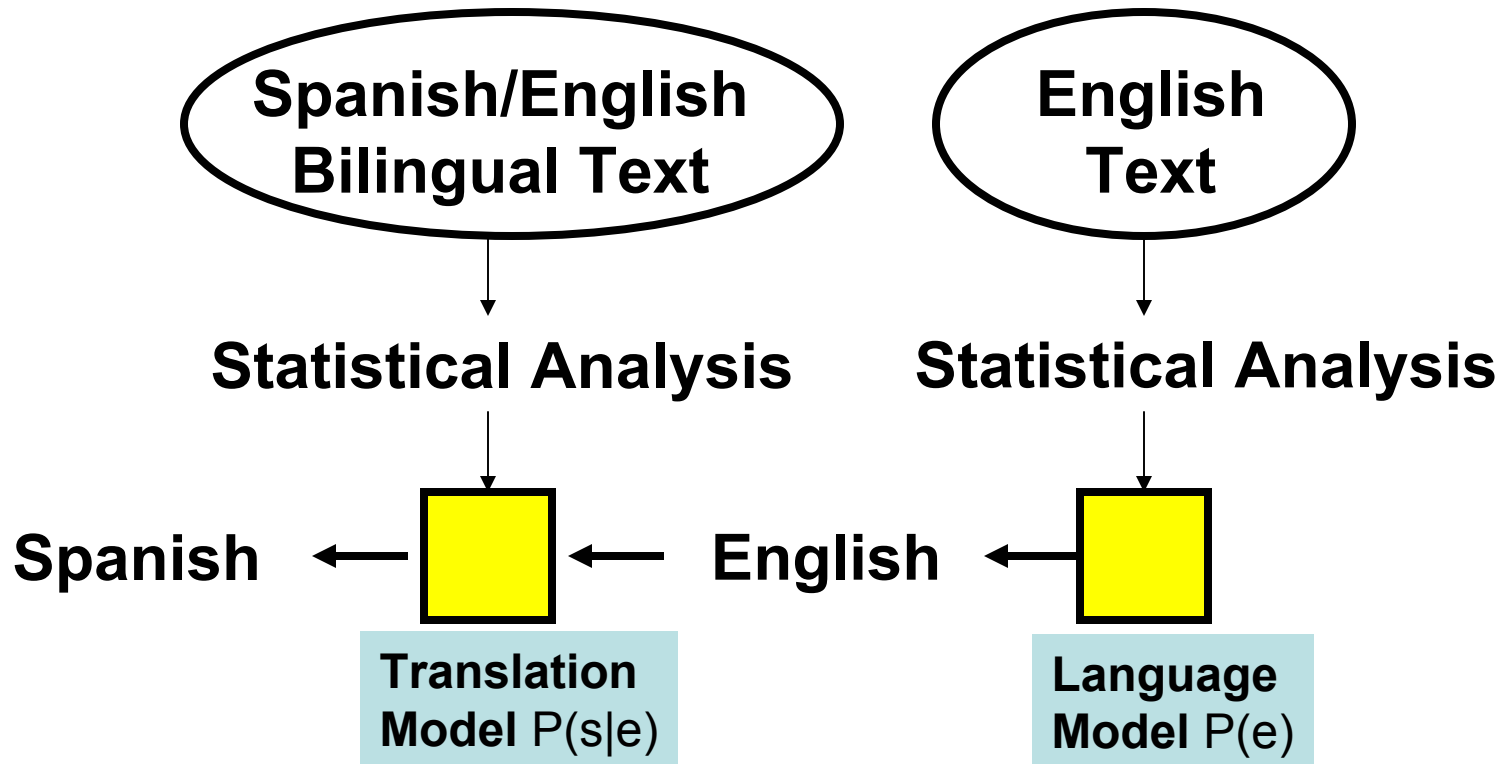


Experiments by P. Koehn on Europarl data

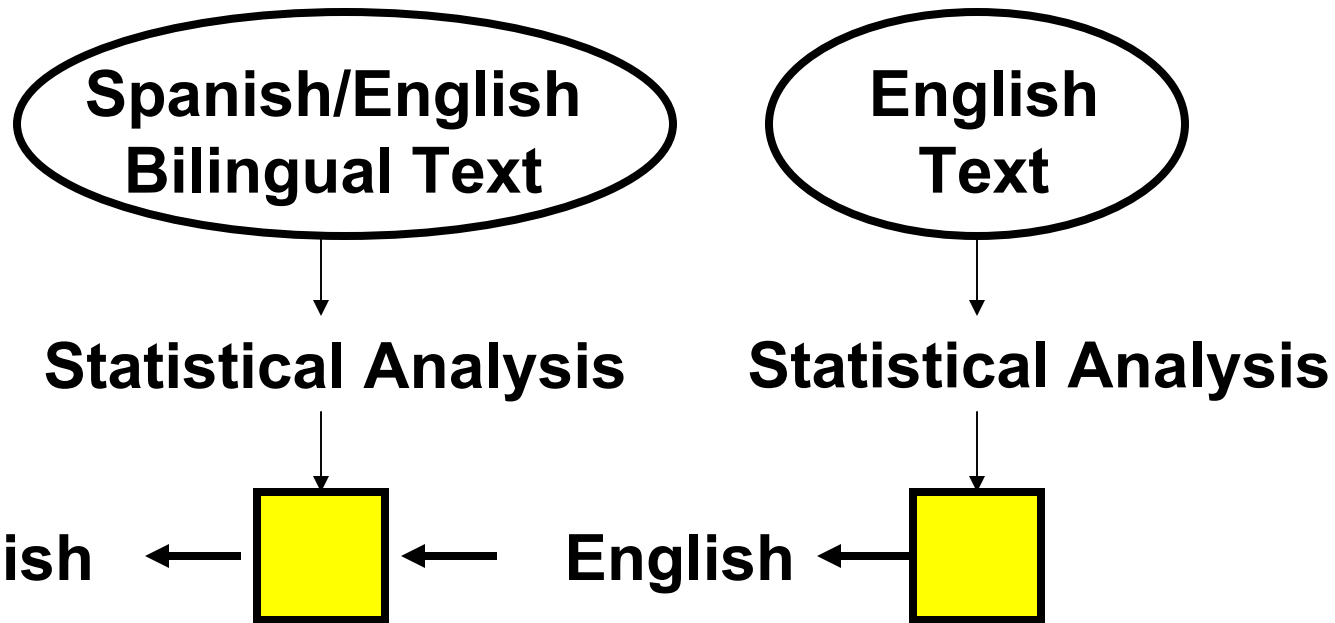
Revolution of 2001

- Permits rapid experimentation with ideas:
 - Should hyphenated words be split up?
 - Should the component X be given more weight in the scoring of translation candidates?
 - Does this form of syntactic control over word re-ordering help translation accuracy?
 - If we exploit non-parallel corpora, does the knowledge gained help translation?
 - Do learned word-formation rules work better than manually-built ones? What if the word-formation learning is tweaked like in such-and-such manner?
- Allows different systems to be compared:
 - Researchers learn from each other

Statistical MT: Noisy Channel Style



Statistical MT: Noisy Channel Style



Decoding algorithm
 $\text{argmax}_e P(e|s) =$
 $\text{argmax}_e P(e) * P(s|e)$

Que hambre tengo yo →

What hunger have I,
Hungry I am so,
I am so hungry,
Have I that hunger ...

→ I am so hungry

Noisy-Channel Statistical MT

- Language model
 - Given an English string e , assigns $P(e)$ by formula
 - good English string \rightarrow high $P(e)$
 - random word sequence \rightarrow low $P(e)$
- Translation model
 - Given a pair of strings $\langle s, e \rangle$, assigns $P(s | e)$ by formula
 - $\langle s, e \rangle$ look like translations \rightarrow high $P(s | e)$
 - $\langle s, e \rangle$ don't look like translations \rightarrow low $P(s | e)$
- Decoding algorithm
 - Given a language model, a translation model, and a new sentence s ... find translation e maximizing $P(e) * P(s | e)$

The Classic Language Model

Word N-Grams

Generative story:

$w_1 = \text{START}$

repeat until END is generated:

 produce word w_2 according to a big table $P(w_2 \mid w_1)$

$w_1 := w_2$

$P(\text{I saw water on the table}) =$

$P(\text{I} \mid \text{START}) *$

$P(\text{saw} \mid \text{I}) *$

$P(\text{water} \mid \text{saw}) *$

$P(\text{on} \mid \text{water}) *$

$P(\text{the} \mid \text{on}) *$

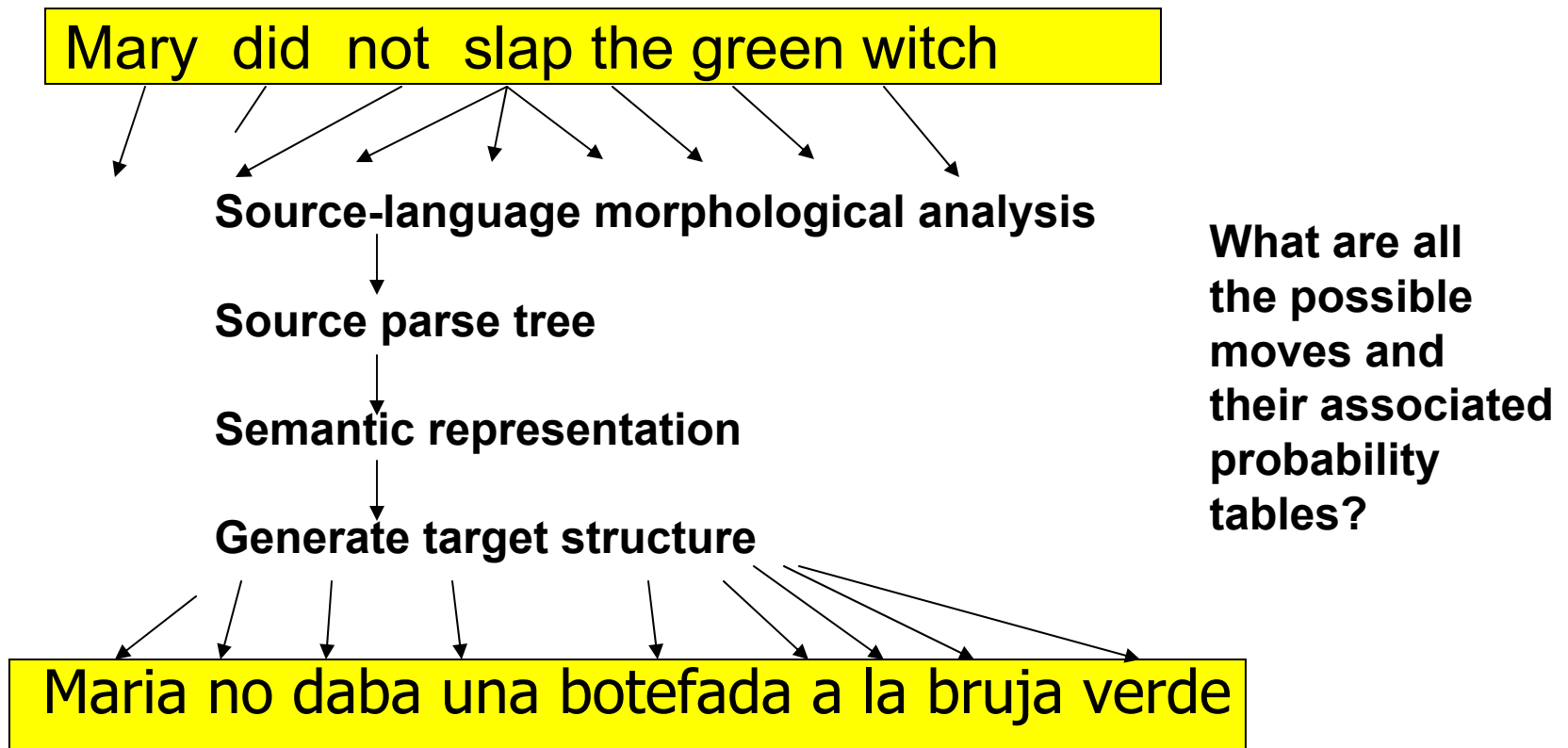
$P(\text{table} \mid \text{the}) *$

$P(\text{END} \mid \text{table})$

**Probabilities can be learned
from online English text.**

Translation Model, $P(s | e) = ?$

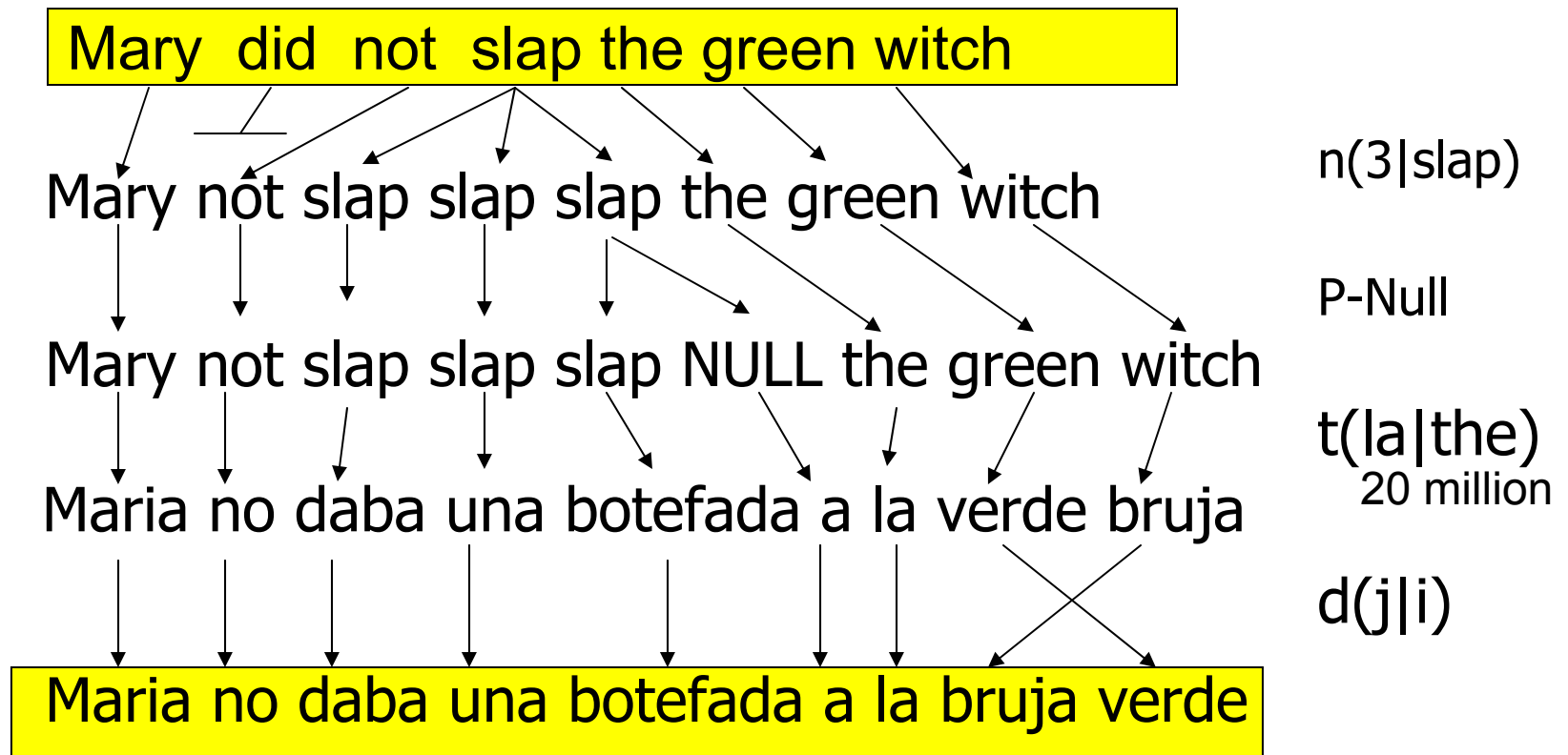
Generative story:



The Classic Translation Model

Word Substitution/Permutation (Brown et al 93)

Generative story:



The Classic Translation Model

Word Substitution/Permutation (Brown et al 93)

$P(s_1 \dots s_m \mid e_1 \dots e_n) =$ a big hairy formula with lots of products and sums, in terms of hidden variables:

$t(s_j \mid e_i)$
 $d(j \mid l, m, n)$
 p_0
 $n(\phi_i \mid e_i)$

EM can set these by maximizing

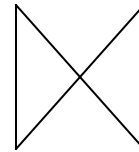
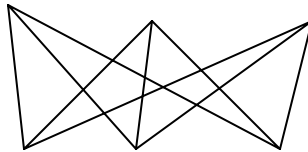
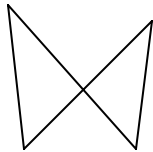
$P(\text{SPANISH CORPUS} \mid \text{ENGLISH CORPUS})$

Full EM is impossible, as there are too many ways to align a sentence pair.

→ Model bootstrapping

Statistical Machine Translation

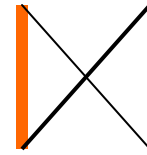
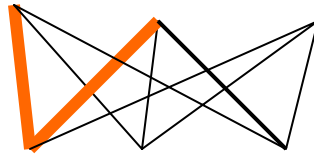
... la maison ... la maison bleue ... la fleur ...



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

Statistical Machine Translation

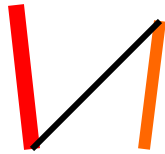
... la maison ... la maison bleue ... la fleur ...



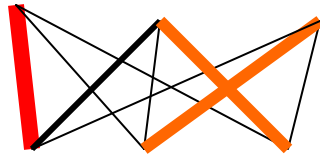
... the house ... the blue house ... the flower ...

Statistical Machine Translation

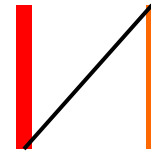
... la maison ... la maison bleue ... la fleur ...



... the house



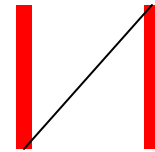
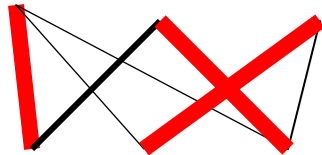
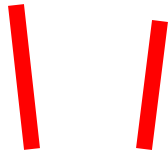
... the blue house



... the flower ...

Statistical Machine Translation


... la maison ... la maison bleue ... la fleur ...



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

Statistical Machine Translation

... la maison ... la maison bleue ... la fleur ...
... the house ... the blue house ... the flower ...



Inherent hidden structure revealed by EM training

For details, see

- Brown et al 93
- “A Statistical MT Tutorial Workbook” (Knight, 1999).
- 37 easy sections, final section promises a free beer.

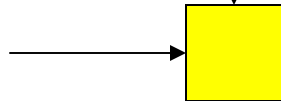
Statistical Machine Translation

... la maison ... la maison bleue ... la fleur ...

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

$P(\text{juste} \mid \text{fair}) = 0.411$
 $P(\text{juste} \mid \text{correct}) = 0.027$
 $P(\text{juste} \mid \text{right}) = 0.020$
...

new French
sentence



Possible
English translations

Statistical Machine Translation

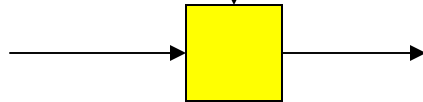
Translation model

$P(\text{juste} \mid \text{fair}) = 0.411$
 $P(\text{juste} \mid \text{correct}) = 0.027$
 $P(\text{juste} \mid \text{right}) = 0.020$
...

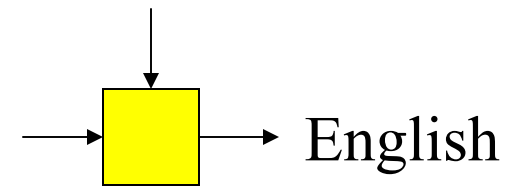
Language model

$P(\text{is not correct it}) = 0.00001$
 $P(\text{it is not correct}) = 0.001$
 $P(\text{correct not it is}) = 0.00001$

new French
sentence



Possible
English translations



English

Decoding

- Of all conceivable English word strings e , find the one maximizing $P(e) * P(s | e)$
- Decoding “Classic” models is NP-complete
 - Knight 99

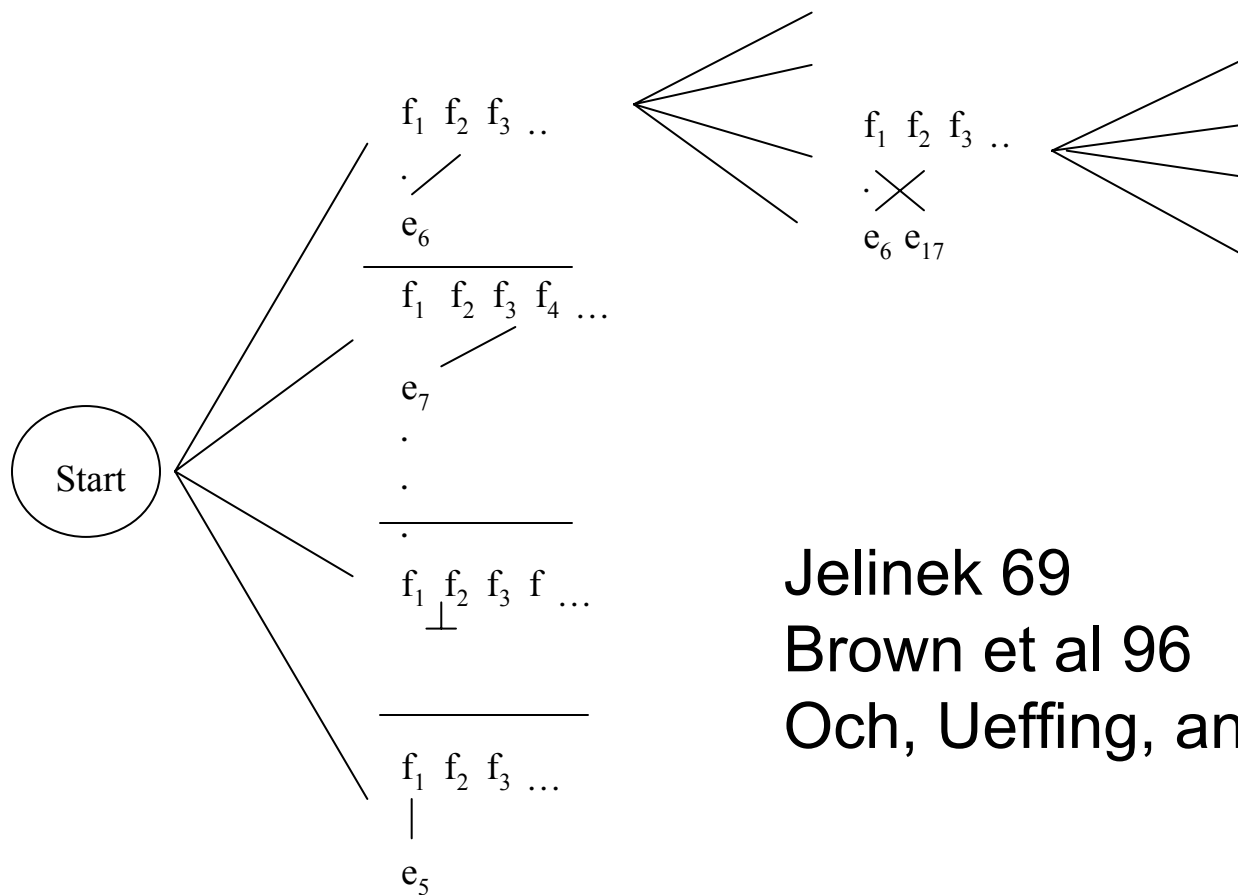
DP Beam Search (optional A*)

To decode: $f_1 f_2 f_3 f_4 f_5 \dots$

First
English word

Second
English word

Third
English word



Jelinek 69

Brown et al 96

Och, Ueffing, and Ney, 2001

Greedy decoding

(Germann et al 01)

Action

NULL well heard , it talking a beautiful victory .
 | | | | | | | | | |
 bien entendu , il parle de une belle victoire .

translateTwoWords(5,talks,7,great)

NULL well heard , it talks a great victory .
 | | | | | | | | | |
 bien entendu , il parle de une belle victoire .

translateTwoWords(2,understood,0,about)

NULL well understood , it talks about a great victory .
 | | | | | | | | | |
 bien entendu , il parle de une belle victoire .

translateOneWord(4,he)

NULL well understood , he talks about a great victory .
 | | | | | | | | | |
 bien entendu , il parle de une belle victoire .

translateTwoWords(1,quite,2,naturally)

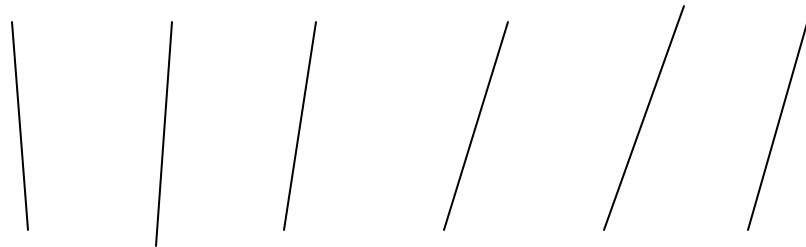
NULL quite naturally , he talks about a great victory .
 | | | | | | | | | |
 bien entendu , il parle de une belle victoire .

IBM Models 1-5

voulez – vous vous taire !

IBM Models 1-5

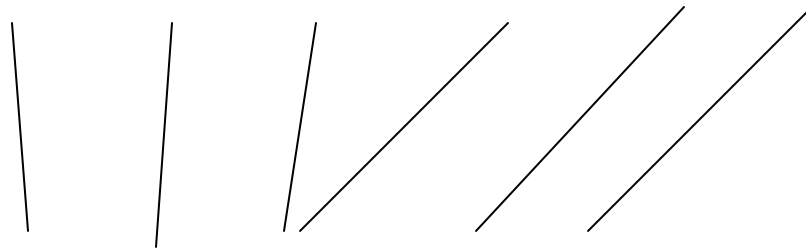
voulez – vous vous taire !



you – you you quiet !

IBM Models 1-5

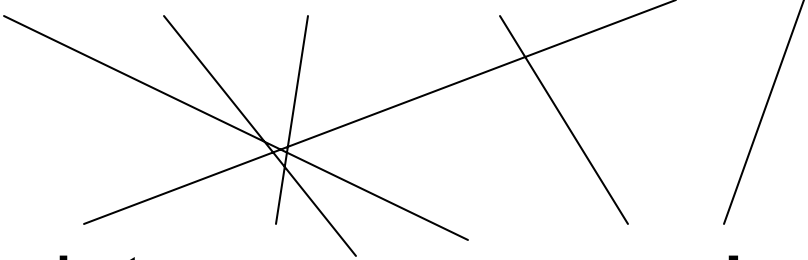
voulez – vous vous taire !



you – you quiet !

IBM Models 1-5

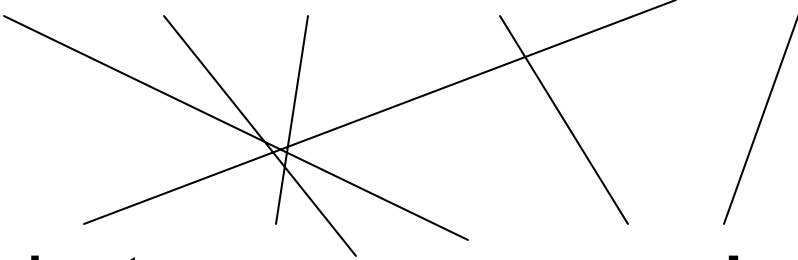
voulez – vous vous taire !



quiet you – you you !

IBM Models 1-5

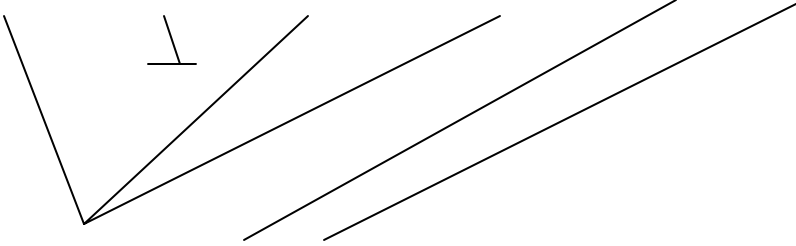
voulez – vous vous taire !



shut you – you you !

IBM Models 1-5

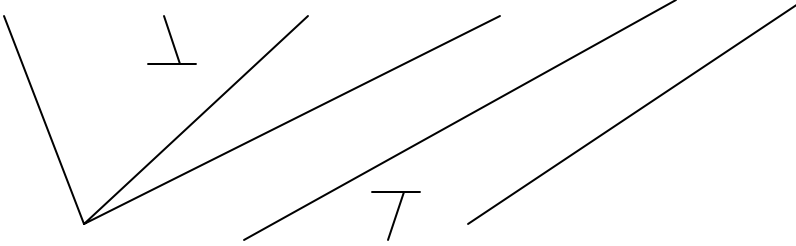
voulez – vous vous taire !



you shut !

IBM Models 1-5

voulez – vous vous taire !



you shut up !

The Classic Results

- *la politique de la haine .* (Foreign Original)
- politics of hate . (Reference Translation)
- the policy of the hatred . (IBM4+N-grams+Stack)

- *nous avons signé le protocole .* (Foreign Original)
- we did sign the memorandum of agreement . (Reference Translation)
- we have signed the protocol . (IBM4+N-grams+Stack)

- *où était le plan solide ?* (Foreign Original)
- but where was the solid plan ? (Reference Translation)
- where was the economic base ? (IBM4+N-grams+Stack)

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and

Next Idea:

From Word Translation to Phrase Translation

Word models have $P(f | e)$

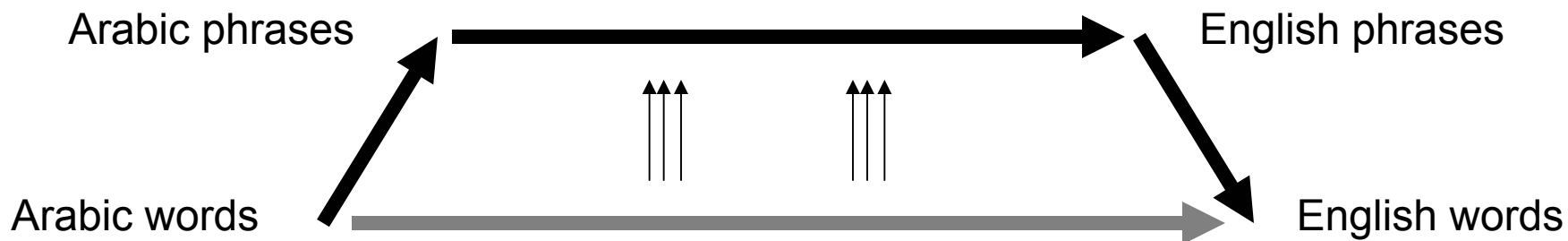
Intuition prefers $P(f f f | e e)$

By phrase we mean any
sequence of words, e.g.:

“...real estate...”

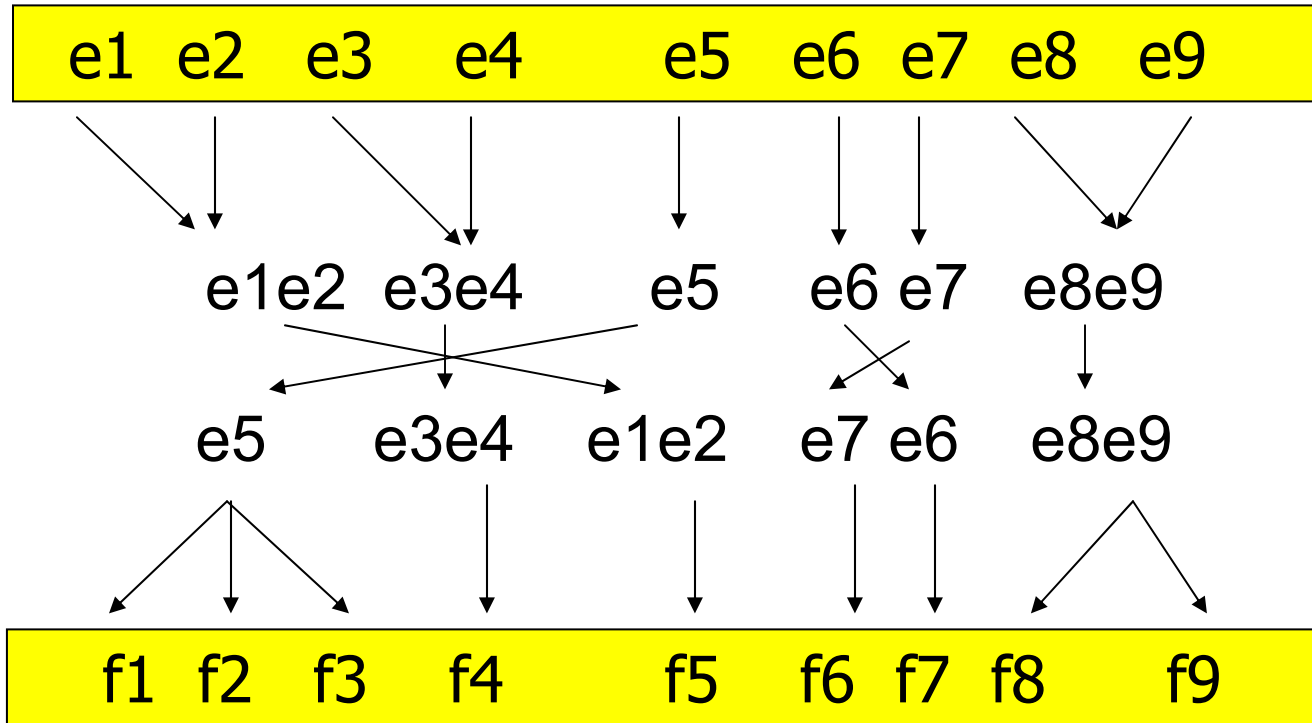
“...note that...”

“...of the same...”



Alignment Templates Model

(Och, Tillmann, & Ney 99 simplified)



Segmentation to Source Phrases

Re-Ordering of Phrases

Translation to Target Phrases
t(f8f9 | e8e9)

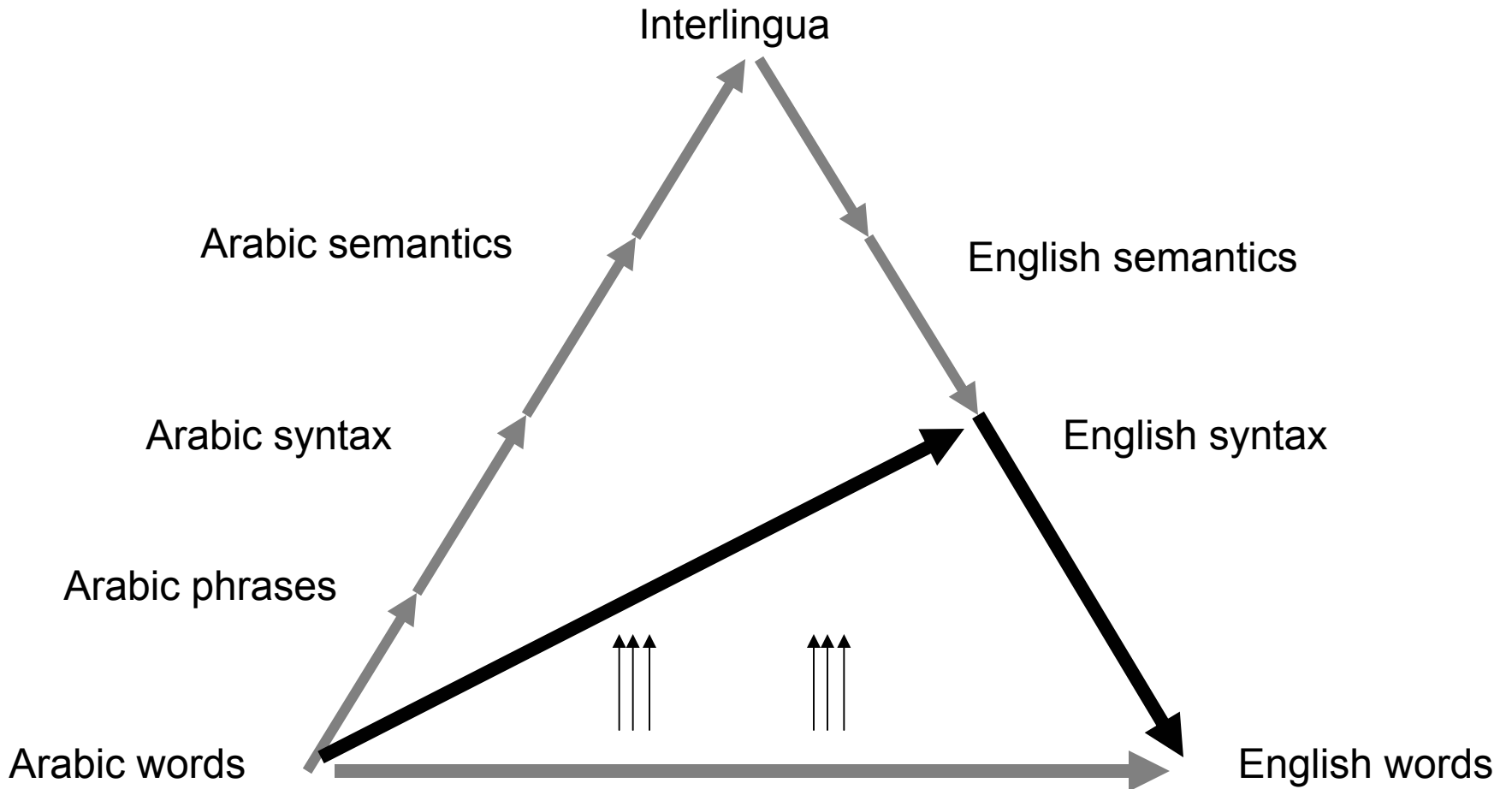
Many Recent Phrase-Based Models

- USC/ISI EMNLP 02
 - CMU ACL 03
 - ATR ACL 03
 - IBM ACL 03
-
- Significantly harder to deal with than words
 - Many more parameters (highest = 1 billion!)
 - More severe estimation problems

Empirical Results

- Phrase models are a huge win!
- Why did it take so long?
 - 1990s IBM had anticipated “multi-word cepts”
 - Complained about “vast featureless desert” in choosing which word sequences might be candidates
 - Only now is the computing hardware there
 - \$5000 for a 4Gb fast dual processor
 - Still lack memory for easily deployable systems!
 - Intuition was lacking

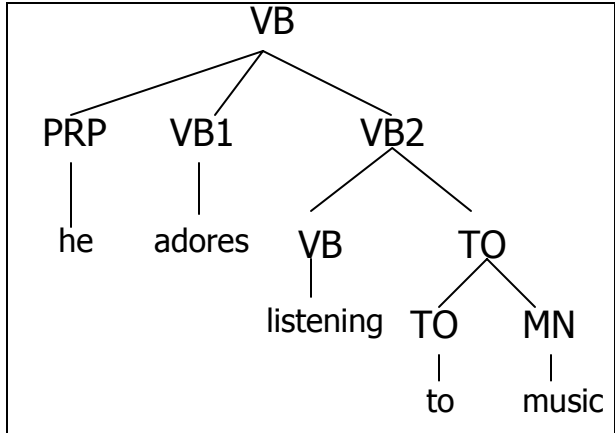
Another Idea: Syntax-Sensitive Translation



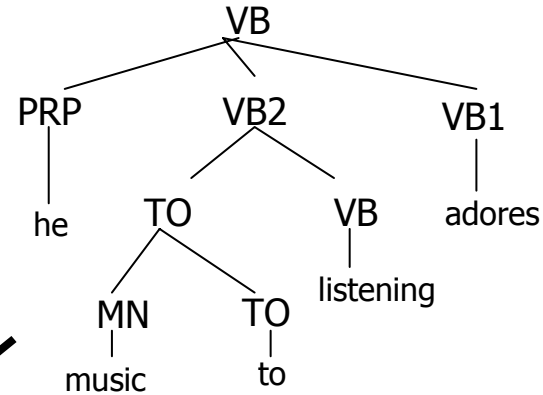
Syntax-Based Model

(Yamada and Knight 01, 02)

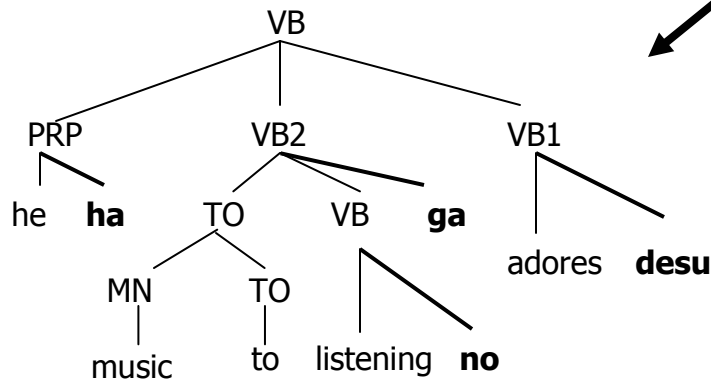
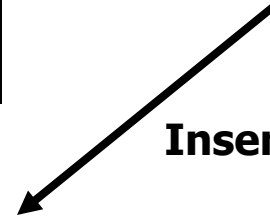
Parse Tree(E)



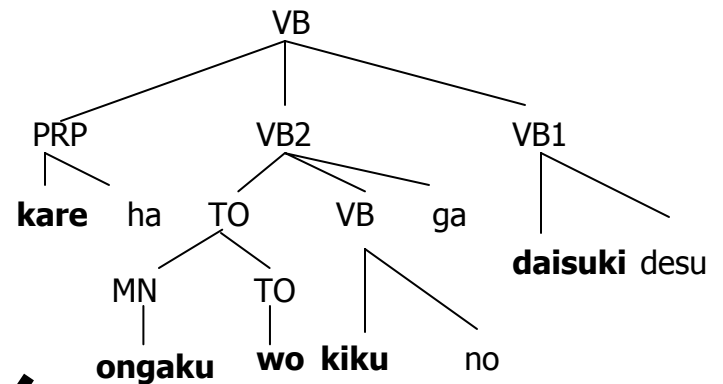
Reorder



Insert



Translate



Take Leaves



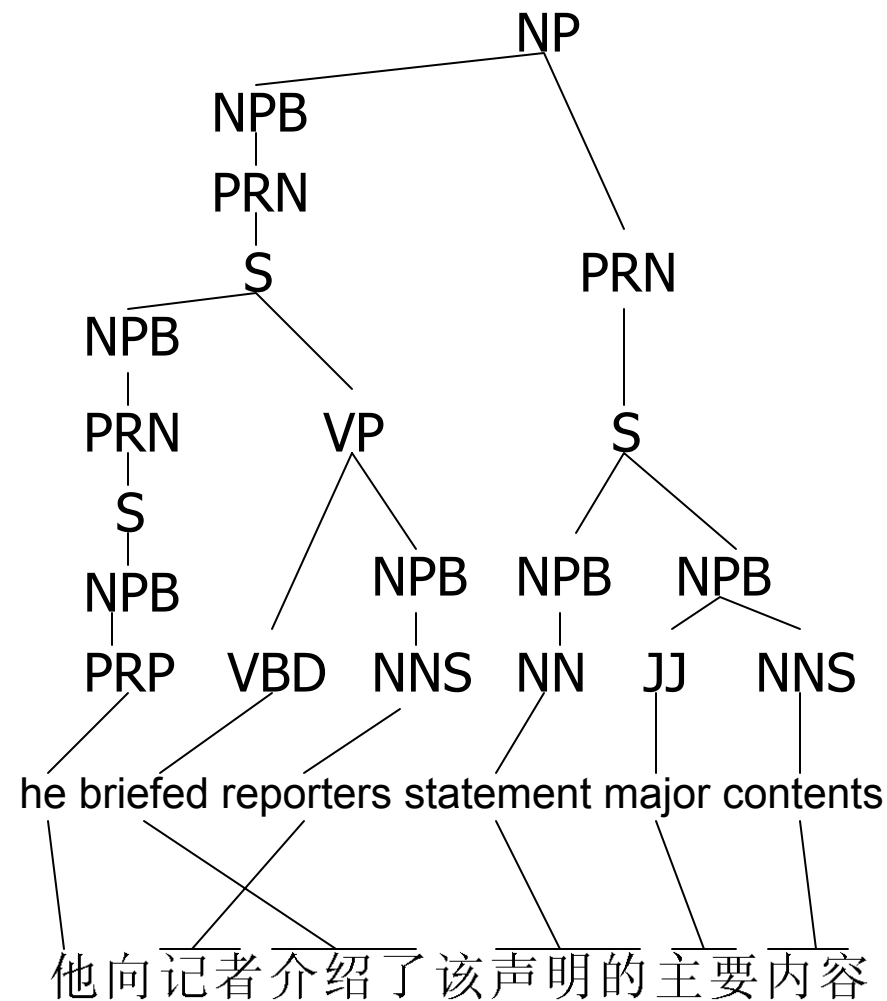
Sentence(J)

Kare ha ongaku wo kiku no ga daisuki desu

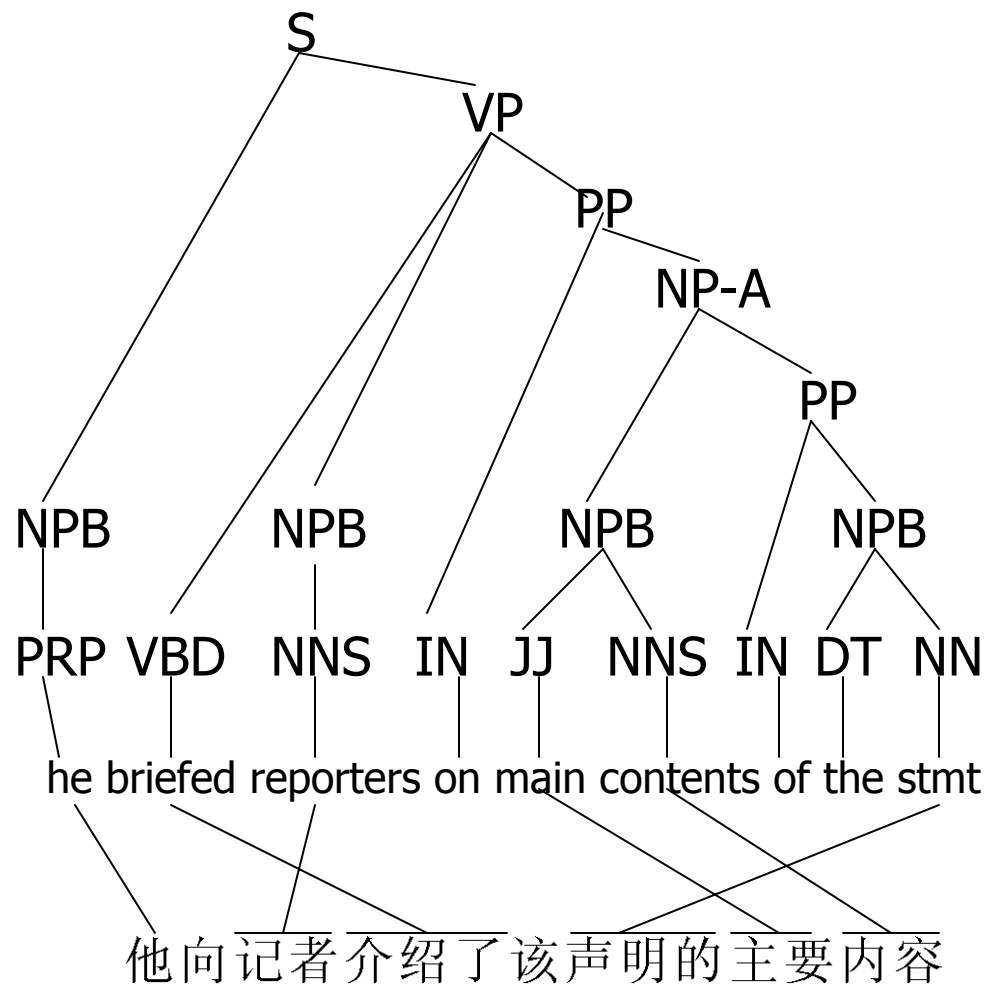
Parameter Table: Reorder

Original Order	Reordering	P(reorder original)
PRP VB1 VB2	PRP VB1 VB2	0.074
	PRP VB2 VB1	0.723
	VB1 PRP VB2	0.061
	VB1 VB2 PRP	0.037
	VB2 PRP VB1	0.083
	VB2 VB1 PRP	0.021
VB TO	VB TO	0.107
	TO VB	0.893
TO NN	TO NN	0.251
	NN TO	0.749

Decoded Tree



Decoding with Trigrams



Decoding with Charniak 01 LM

Charniak, Knight, Yamada 03

Translation model	Language model	# of perfect translations/317
IBM Model 4	Trigram	23
Syntax-Based	Trigram	31
Syntax-Based	Charniak 01	45
		77 (oracle)

But: No improvement in semantic accuracy (or BLEU score)!

Many factors left out.

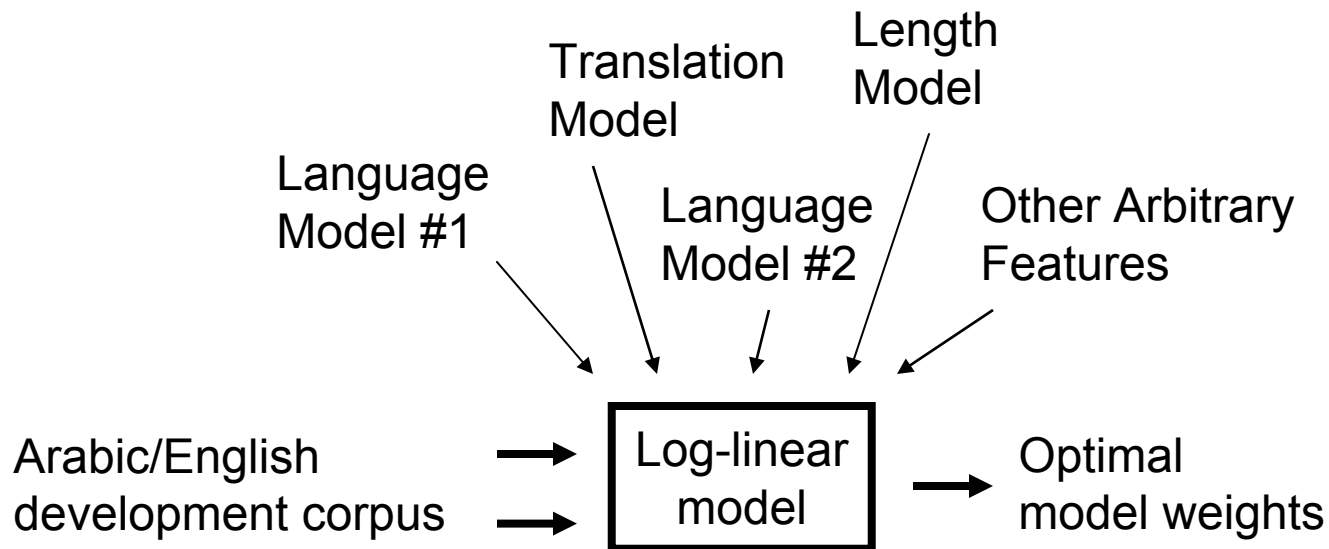
Line seems promising:

-- current work by Mark Hopkins (UCLA) and Michel Galley

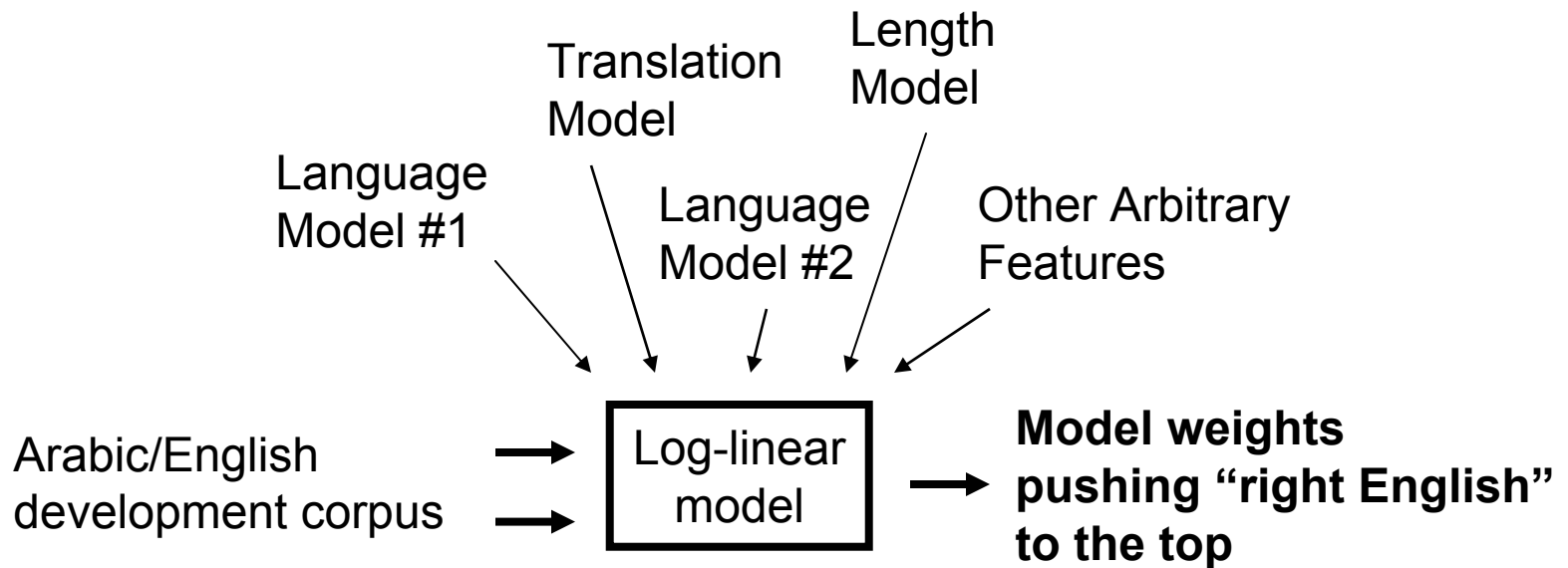
Recent Alternatives to Generative Models and Maximum Likelihood

Log-linear Models (Och 01)

Speech people often do $\operatorname{argmax}_{w \dots w} P(w \dots w)^{10} * P(a \dots a | w \dots w)$

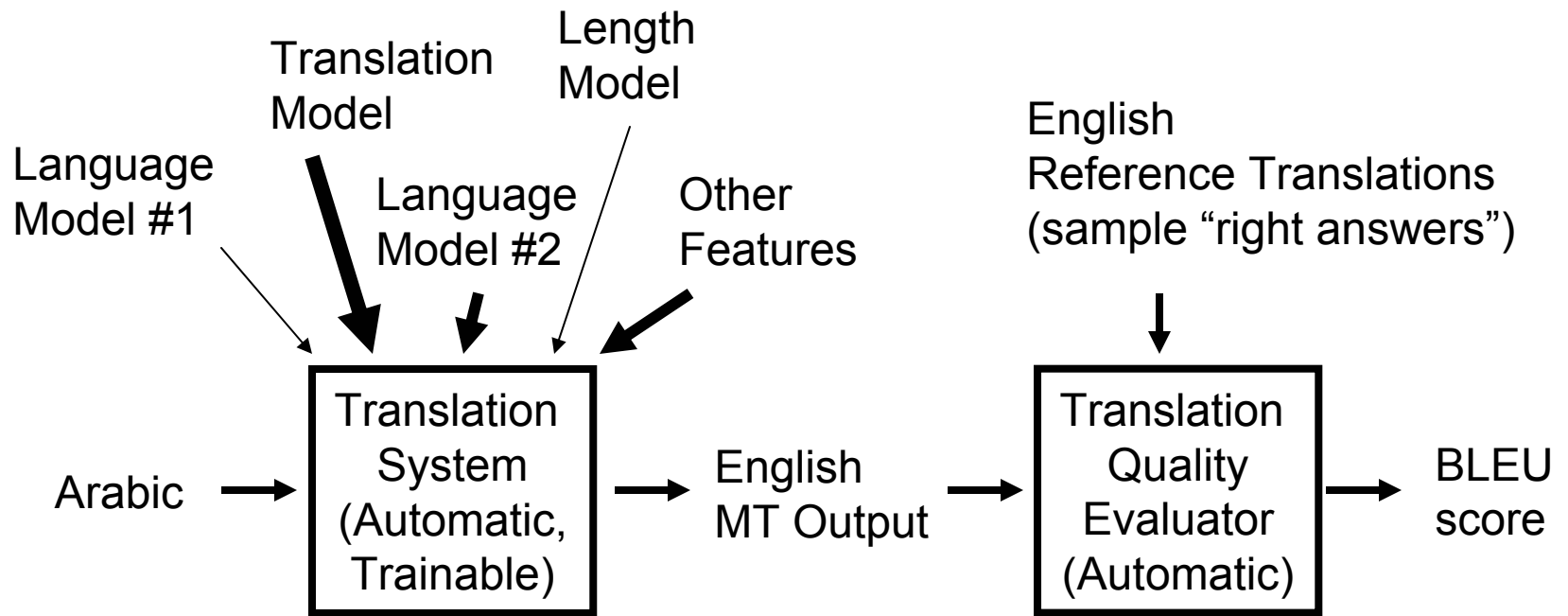


Discriminative Training (Och 02)



Not easy to do, since "right English" isn't in the n-best list.

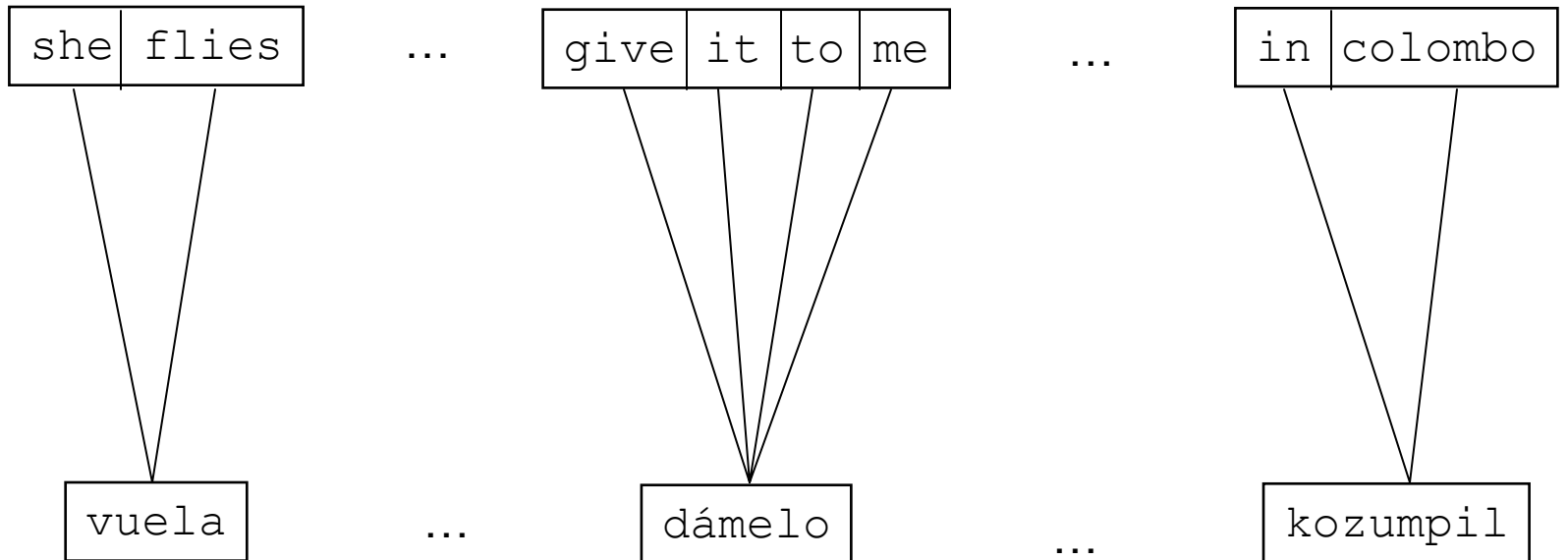
Error-Based Training (Och 03)



Learning Algorithm for Directly Reducing Translation Error
(Not easy to do because search space is very bumpy).

Some Currently Active Areas

Character-Based Models



Character-Based Models, or Universal Morphology

s	h	e	f	l	i	e	s
---	---	---	---	---	---	---	---

g	i	v	e	i	t	t	o	m	e
---	---	---	---	---	---	---	---	---	---

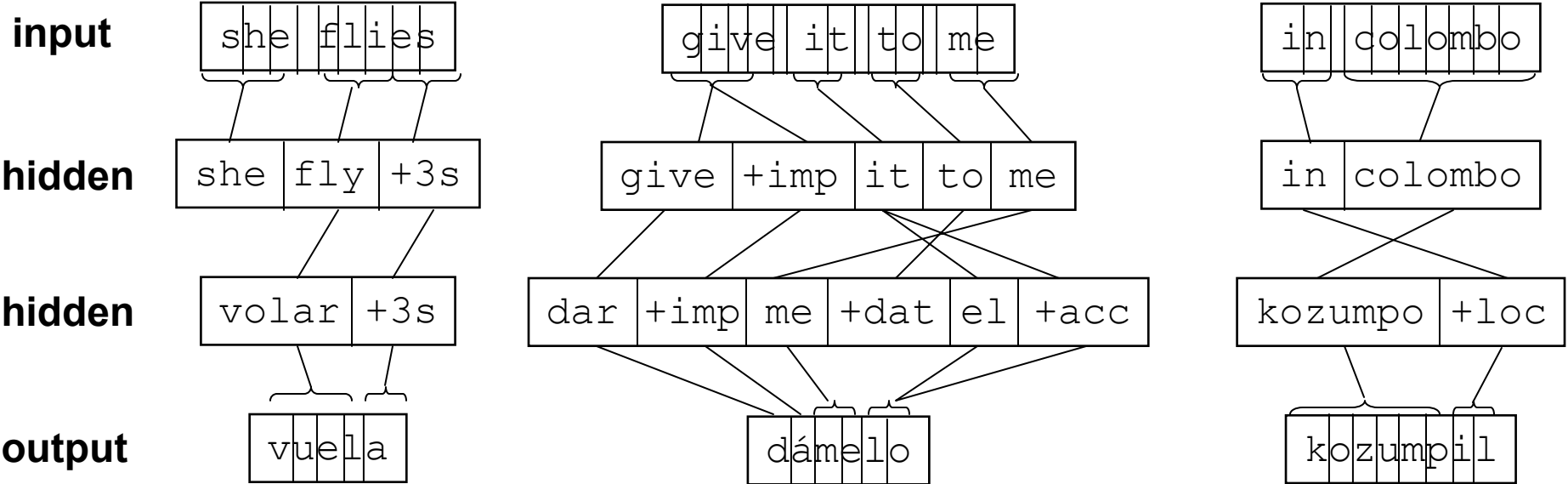
i	n	c	o	l	o	m	b	o
---	---	---	---	---	---	---	---	---

v	u	e	l	a
---	---	---	---	---

d	a	m	e	l	o
---	---	---	---	---	---

k	o	z	u	m	p	i	l
---	---	---	---	---	---	---	---

Character-Based Models, or Universal Morphology



Tree-Based Models

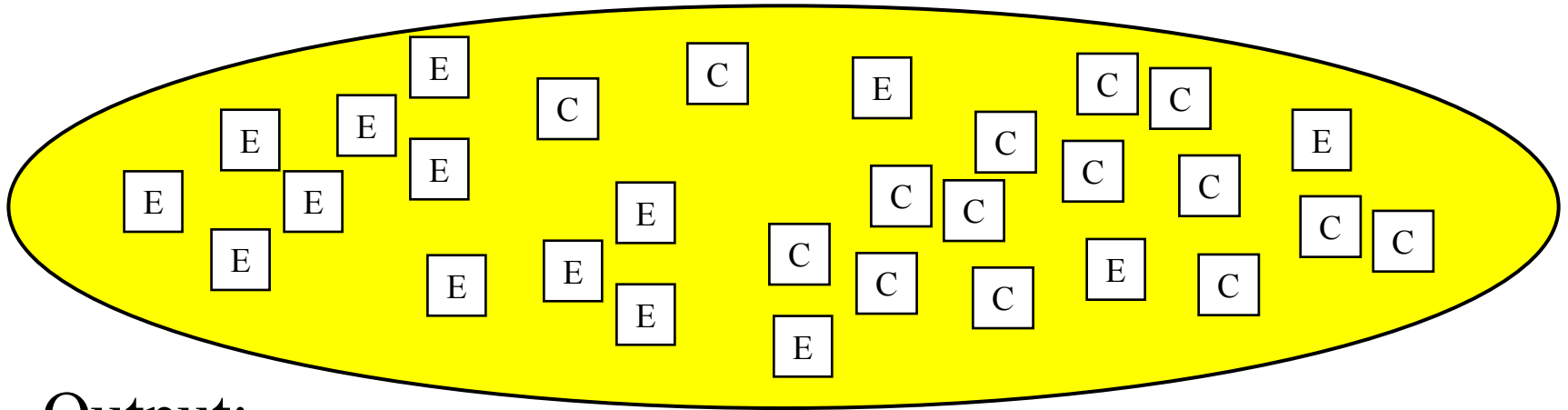
- How does an English tree become a foreign-language string?
- How can a syntax-based TM produce more good candidates and fewer bad ones?
- What happens to passive sentences when they translate to Chinese? Arabic?
- What happens to NP-of-NP?
- How much does the translation of a verb depend on the translation of its object? Is this best captured in LM or TM?
- Can long sentences be efficiently decoded?

Web as Bilingual Corpus

- Only a tiny bit of work on this important problem
 - e.g., Resnik & Smith, 2001
- Train on the web, to deploy on the web
 - Potentially very heterogeneous source
 - Much wider vocabulary/phrase coverage

Document Alignment

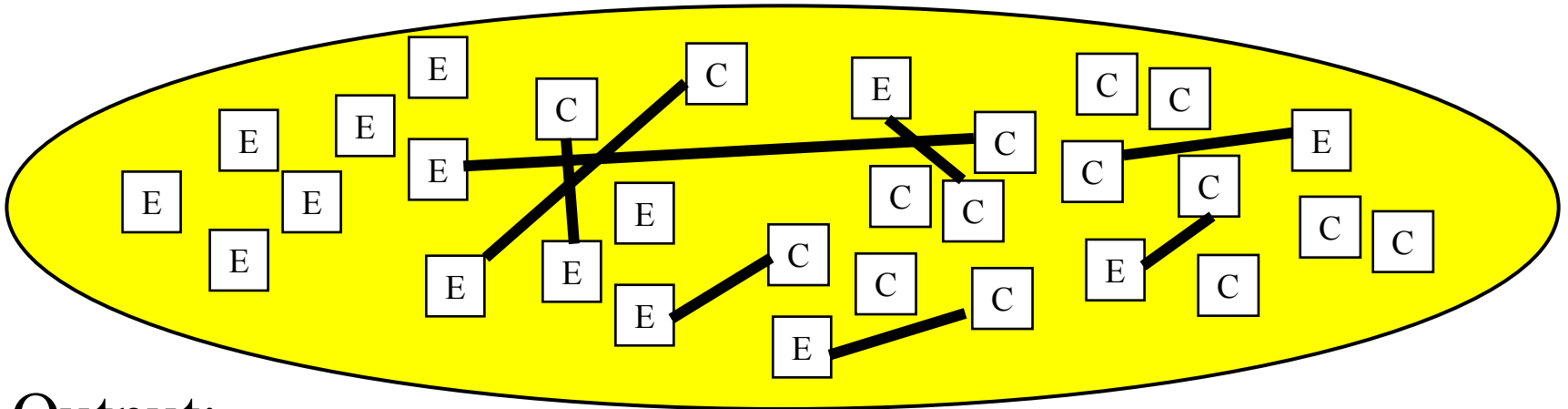
- Input:
 - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.



- Output:
 - List of pairs of files that are actually translations.

Document Alignment

- Input:
 - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.



- Output:
 - List of pairs of files that are actually translations.

All Sorts of Expertise Needed

- Linguistics:
 - building models of a human process
- Machine Learning:
 - appropriate model templates and algorithms
 - unsupervised and supervised problems
 - generative models, feature-based models
 - EM, maximum entropy, discriminative training, error-based training
 - Model bootstrapping
- Computer Science and AI:
 - heuristic search
 - efficient data structures
- Resources:
 - acquiring data from the Internet and other sources

Thanks!