Natural Language Models and Interfaces

lecture 9

Ivan Titov

Institute for Logic, Language and Computation

UNIVERSITEIT VAN AMSTERDAM

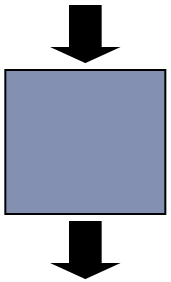
Today

- Machine translation: outlook
  - motivating the task
  - word-based models
  - Integrating phrases and syntax

Machine Translation

[In this part, many slides are from Kevin Knight]

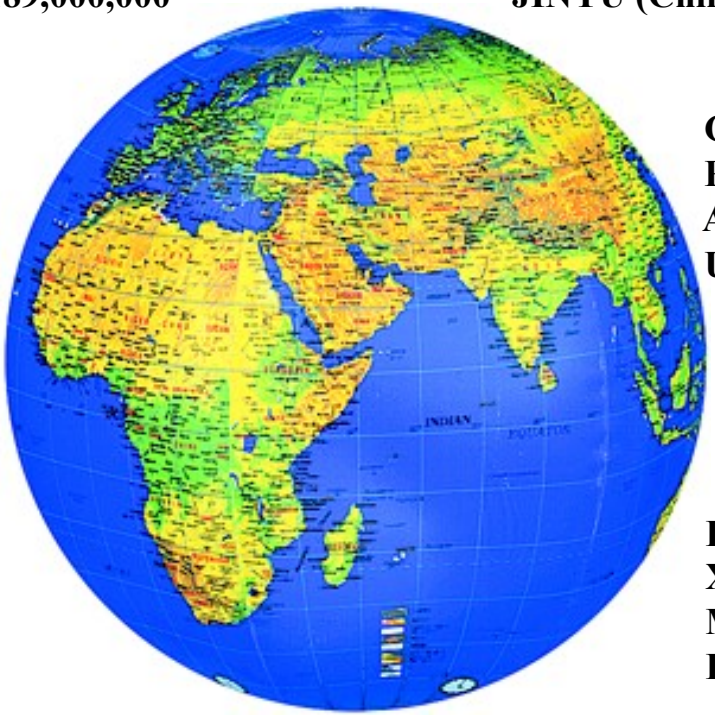
美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



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.

Thousands of Languages Are Spoken

MANDARIN	885,000,000
SPANISH	332,000,000
ENGLISH	322,000,000
BENGALI	189,000,000
HINDI	182,000,000
PORTUGUESE	170,000,000
RUSSIAN	170,000,000
JAPANESE	125,000,000
GERMAN	98,000,000
WU (China)	77,175,000
JAVANESE	75,500,800
KOREAN	75,000,000
FRENCH	72,000,000
VIETNAMESE	67,662,000
TELUGU	66,350,000
YUE (China)	66,000,000
MARATHI	64,783,000
TAMIL	63,075,000
TURKISH	59,000,000
URDU	58,000,000
MIN NAN (China)	49,000,000
JINYU (China)	45,000,000
GUJARATI	44,000,000
POLISH	44,000,000
ARABIC	42,500,000
UKRAINIAN	41,000,000
ITALIAN	37,000,000
XIANG (China)	36,015,000
MALAYALAM	34,022,000
HAKKA (China)	34,000,000
KANNADA	33,663,000
ORIYA	31,000,000
PANJABI	30,000,000
SUNDA	27,000,000



Warren Weaver (1947)

ingcmpnqsnwf cv fpn owoktvcv

hu ihgzsnwfv rqcffnw cw owgcnwf

kowazoanv ...

Warren Weaver (1947)

e e e e

ingcmpnqsnwf cv fpn owoktvcv

e e e

hu ihgzsnwfv rqcffnw cw owgcnwf

e

kowazoanv ...

Warren Weaver (1947)

e e e the

ingcmpnqsnwf cv fpn owoktvcv

e e e

hu ihgzsnwfv rqcffnw cw owgcnwf

e

kowazoanv ...

Warren Weaver (1947)

e he e the

ingcmpnqsnwf cv fpn owoktvcv

e e e t

hu ihgzsnwfv rqcffnw cw owgcnwf

e

kowazoanv ...

Warren Weaver (1947)

e he e of the
ingcmpnqsnwf cv fpn owoktvcv
e e e t
hu ihgzsnwfv rqcffnw cw owgcnwf
e
kowazoanv ...

Warren Weaver (1947)

e he e of the fof
ingcmpnqsnwf cv fpn owoktvcv
e f o e o oe t
hu ihgzsnwfv rqcffnw cw owgcnwf
ef
kowazoanv ...

Warren Weaver (1947)

e he e ~~of~~ the

ingcmpnqsnwf cv fpn owoktvcv

e e e t

hu ihgzsnwfv rqcffnw cw owgcnwf

e

kowazoanv ...

Warren Weaver (1947)

e he e is the sis

ingcmpnqsnwf cv fpn owoktvcv

e s i e i ie t

hu ihgzsnwfv rqcffnw cw owgcnwf

es

kowazoanv ...

Warren Weaver (1947)

decipherment is the analysis
ingcmpnqsnwf cv fpn owoktvcv
of documents written in ancient
hu ihgzsnwfv rqcffnw cw owgcnwf
languages ...
kowazoanv ...

“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, March 1947

“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, March 1947

“... as to the problem of mechanical translation, I frankly am afraid that the [semantic] boundaries of words in different languages are too vague ... to make any quasi-mechanical translation scheme very hopeful.”

- Norbert Wiener, April 1947

Spanish/English corpus

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 .

Spanish/English corpus

Translate: Clients do not sell pharmaceuticals in Europe.

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 .

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

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

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

process of elimination

Centauri/Arcturan [Knight 97]

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

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

Centauri/Arcturan [Knight 97]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

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

"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."

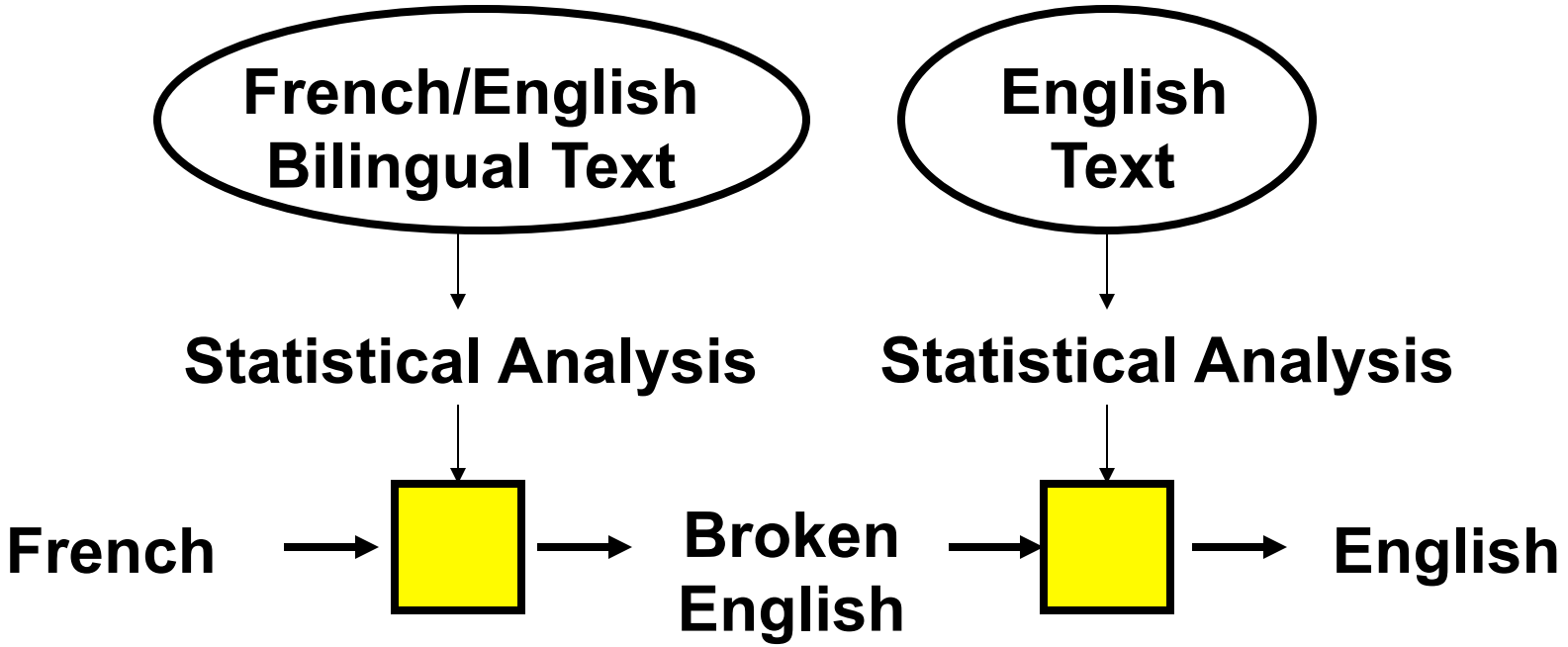
- Warren Weaver, March 1947

The required statistical tables have millions of entries…?

Too much for the computers of Weaver's day.

→ Not enough RAM!

IBM Candide Project
[Brown et al 93]



J’ai si faim → What hunger have I, Hungry I am so, I am so hungry, Have me that hunger … → I am so hungry

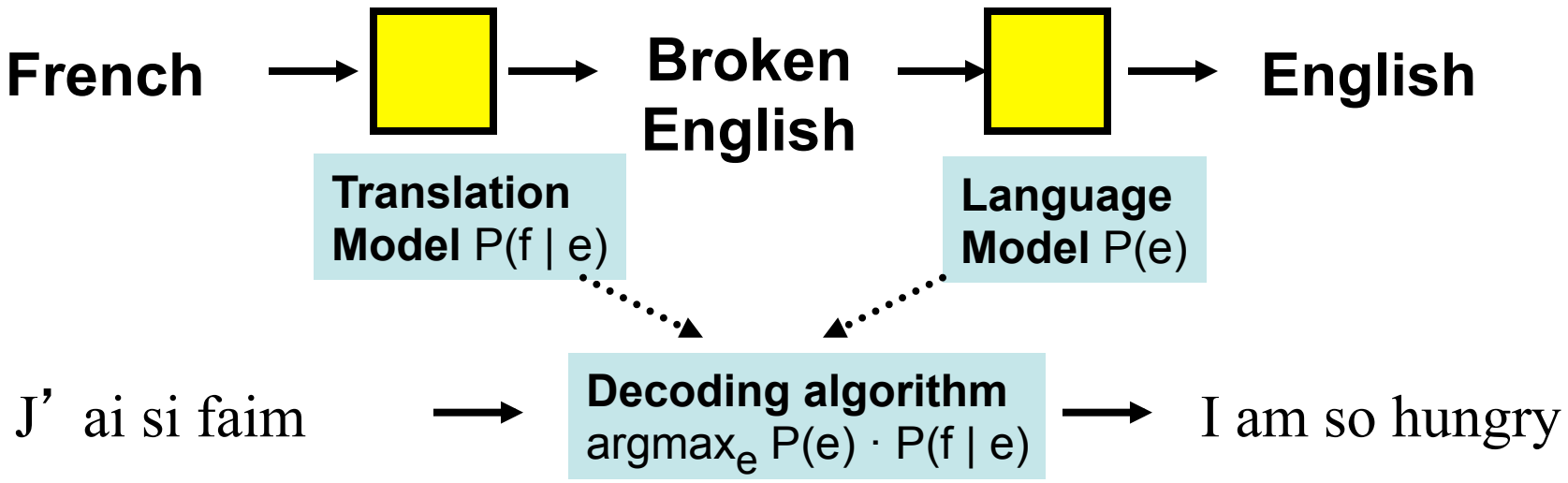
Mathematical Formulation

Given source sentence f:

$$\text{argmax}_e \quad P(e \mid f) =$$

$$\text{argmax}_e \quad P(f \mid e) \cdot P(e) / P(f) = \qquad \textit{by Bayes Rule}$$

$$\text{argmax}_e \quad P(f \mid e) \cdot P(e) \qquad \textit{P(f) same for all e}$$



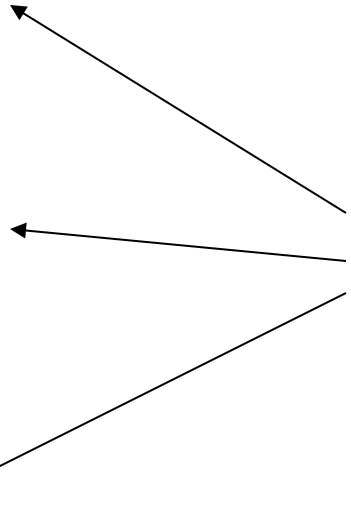
Language Modeling

Goal of a language model for MT:

He is on the soccer field
He is in the soccer field

Is table the on cup the
The cup is on the table

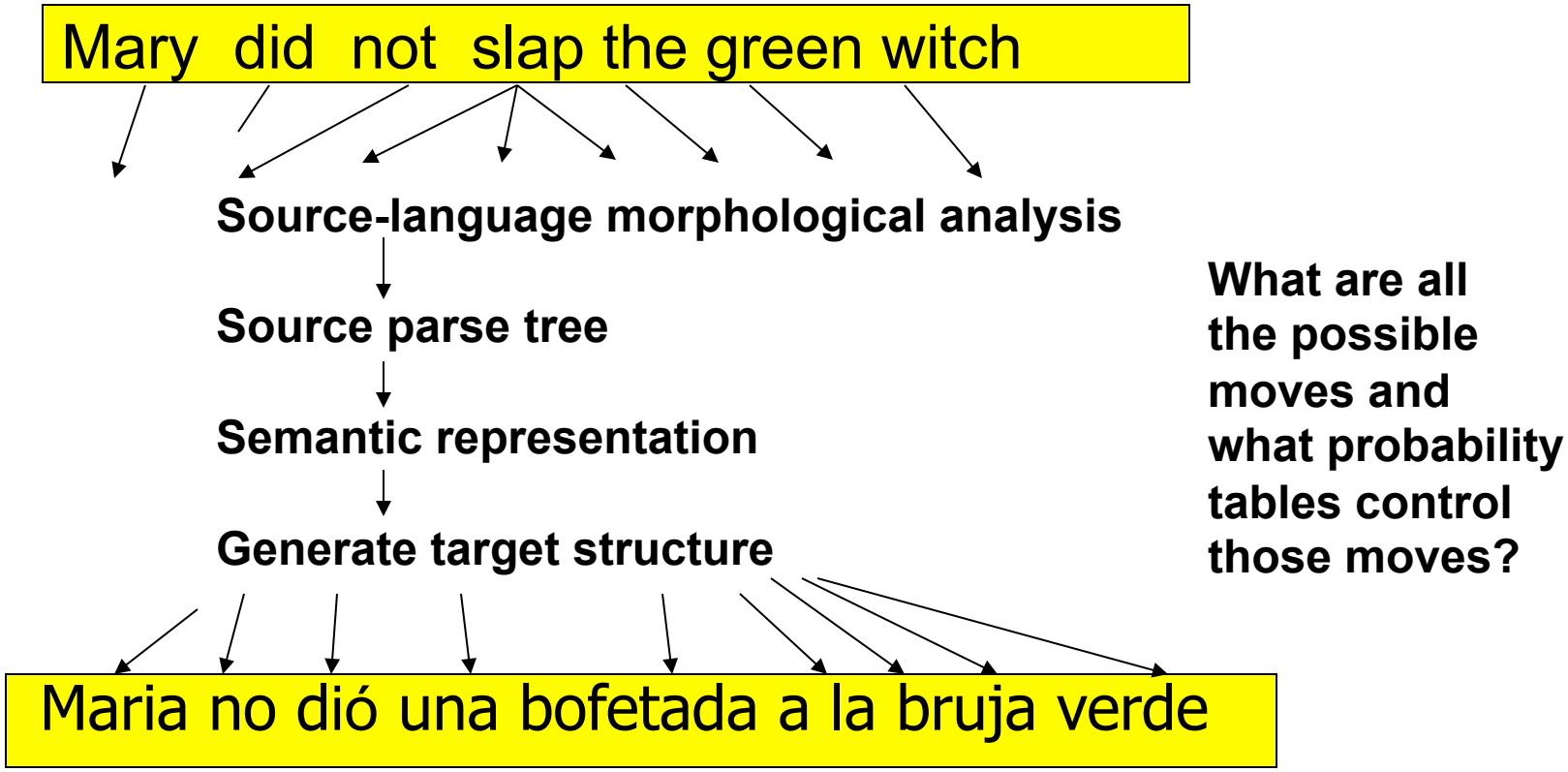
American shrine
American company



Need to make these decisions, because translation model may not have a lot of context information!

Translation Model?

Process model of translation:

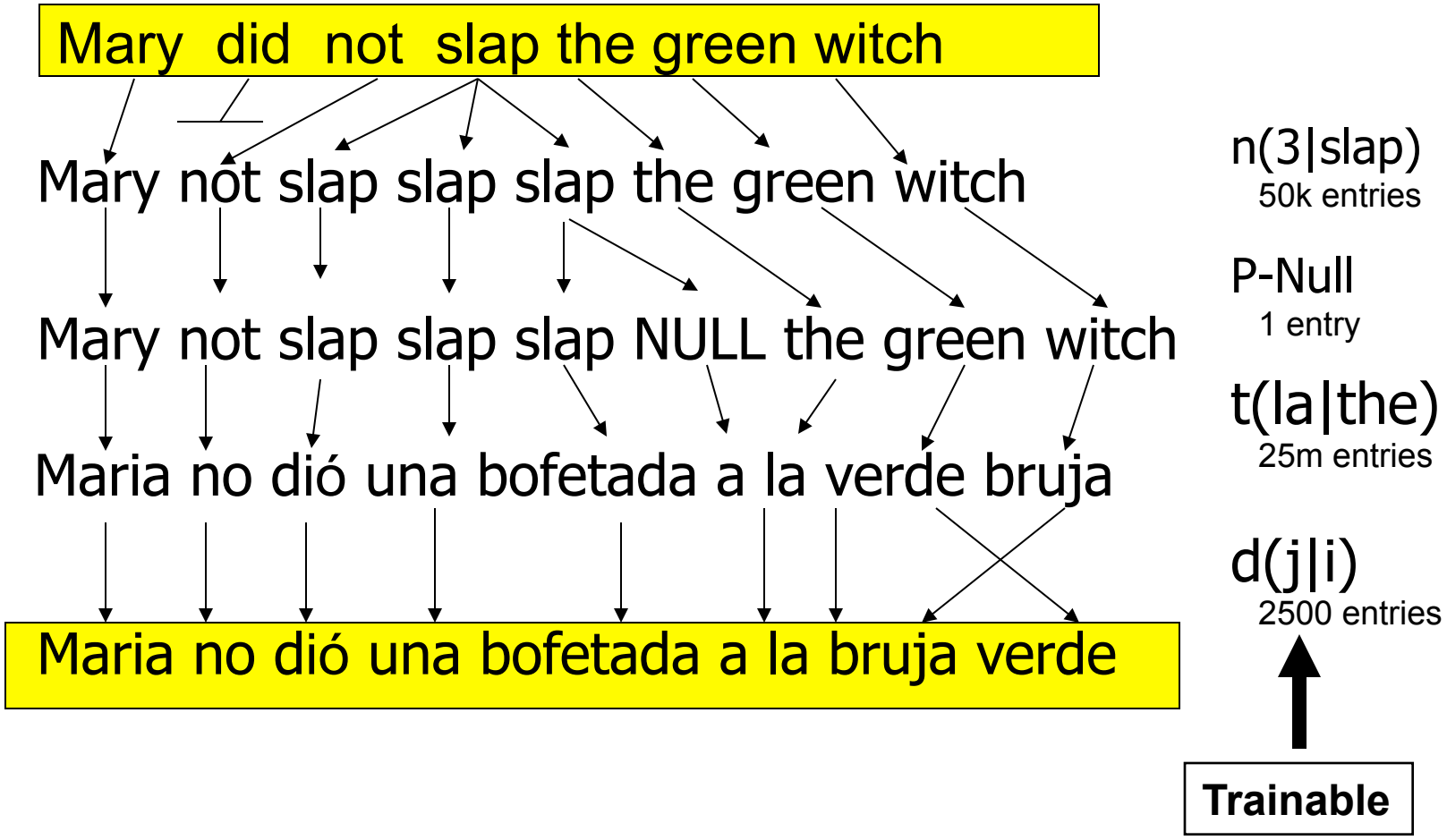


What are all the possible moves and what probability tables control those moves?

The Classic Translation Model

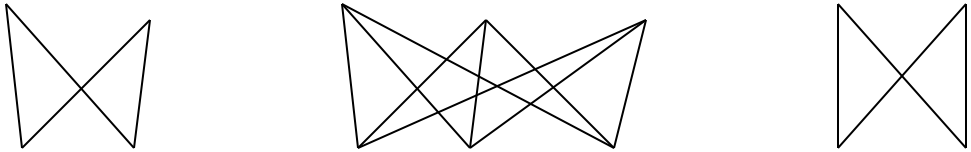
Word Substitution/Permutation [Brown et al., 1993]

Process model of translation:



Unsupervised EM Training

… la maison … la maison bleue … la fleur …

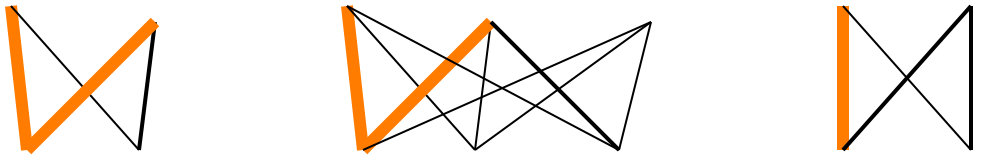


… the house … the blue house … the flower …

All P(french-word | english-word) equally likely

Unsupervised EM Training

… la maison … la maison bleue … la fleur …

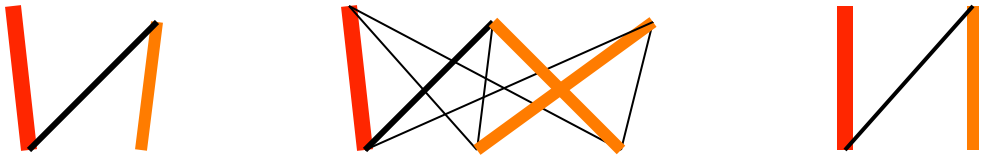


… the house … the blue house … the flower …

“la” and “the” observed to co-occur frequently,
so P(la | the) is increased.

Unsupervised EM Training

… la maison … la maison bleue … la fleur …



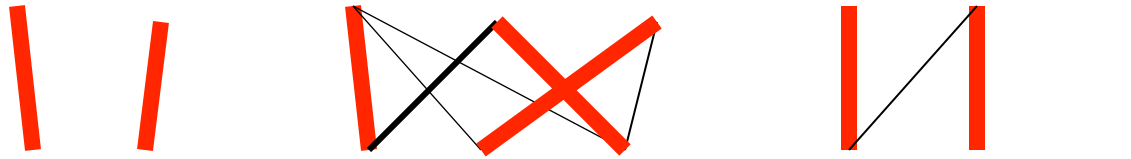
… the house … the blue house … the flower …

“maison” co-occurs with both “the” and “house”, but P(maison | house) can be raised without limit, to 1.0, while P(maison | the) is limited because of “la”

(pigeonhole principle)

Unsupervised EM Training

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

settling down after another iteration

Unsupervised EM Training

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

Inherent hidden structure revealed by EM training!

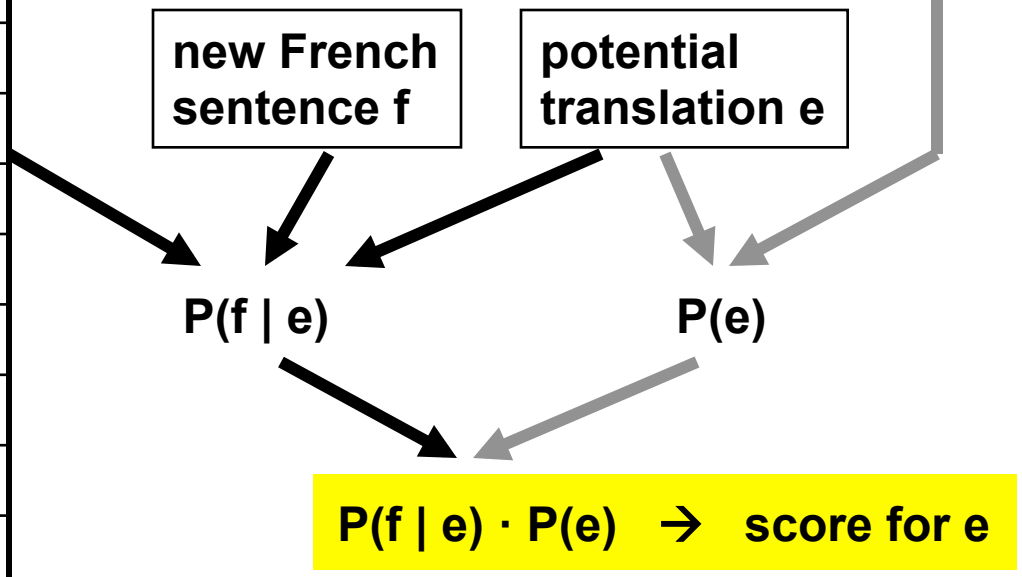
- "A Statistical MT Tutorial Workbook" (Knight, 1999).
- "The Mathematics of Statistical Machine Translation" (Brown et al, 1993)
- Software: GIZA++

Language Model

w1	w2	P(w2 | w1)
of	the	0.13
	a	0.09
	another	0.01
	some	0.01
hong	kong	0.98
	said	0.01
	stated	0.01

Translation Model

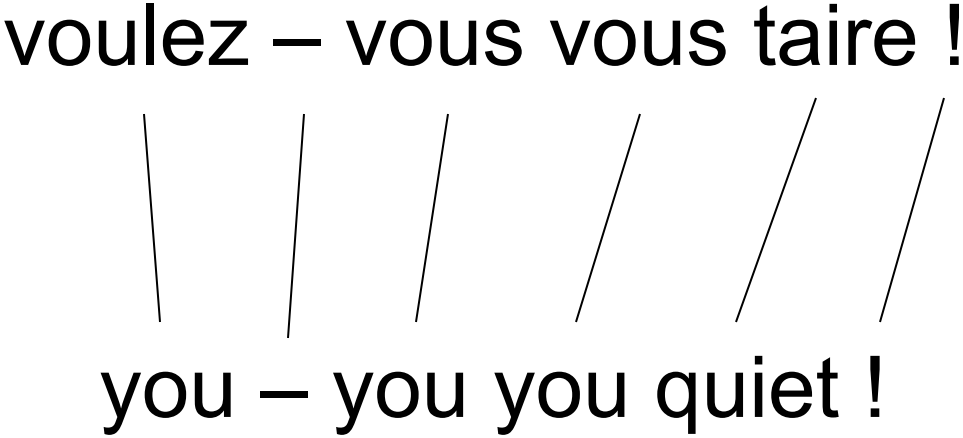
e	f	P(f | e)
national	nationale	0.47
	national	0.42
	nationaux	0.05
	nationales	0.03
the	le	0.50
	la	0.21
	les	0.16
	l’	0.09
	ce	0.02
	cette	0.01
farmers	agriculteurs	0.44
	les	0.42
	cultivateurs	0.05
	producteurs	0.02



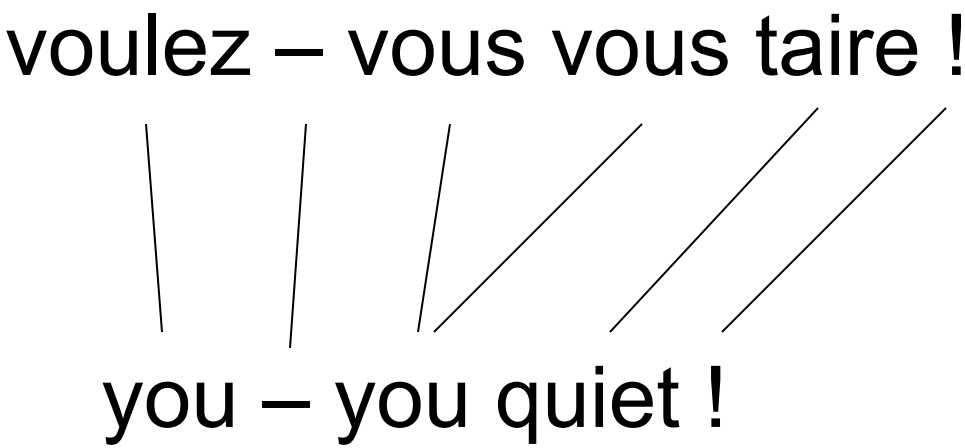
Search for Best Translation

voulez – vous vous taire !

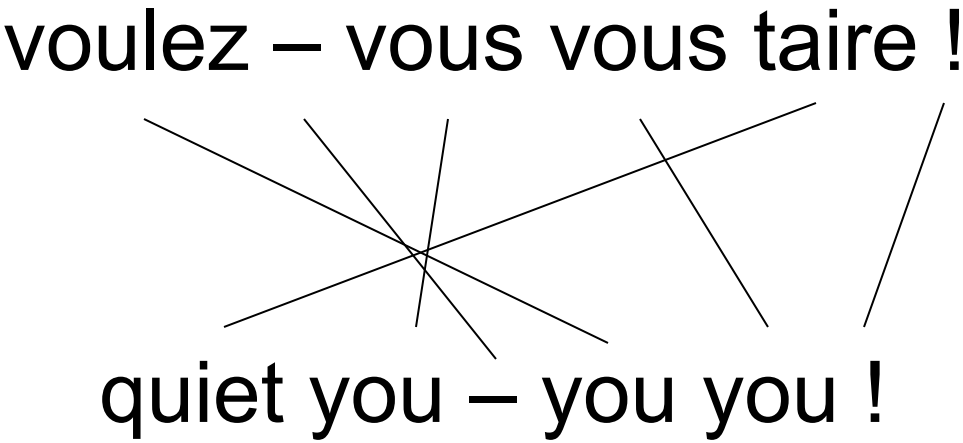
Search for Best Translation



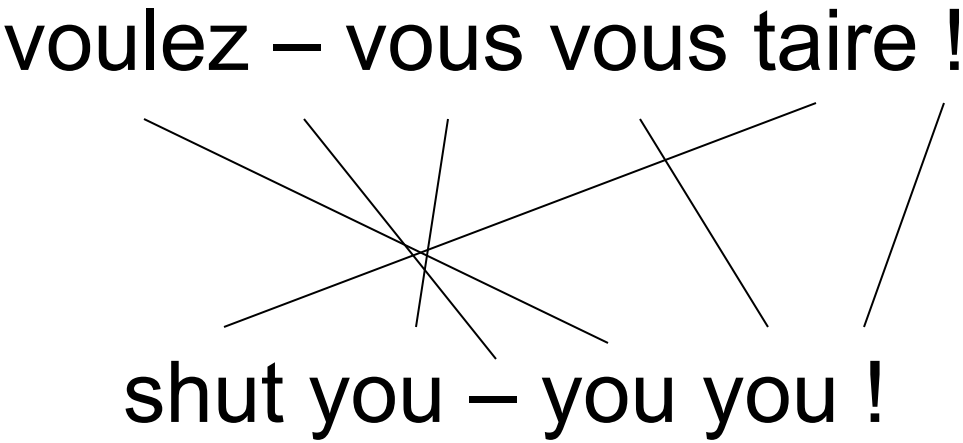
Search for Best Translation



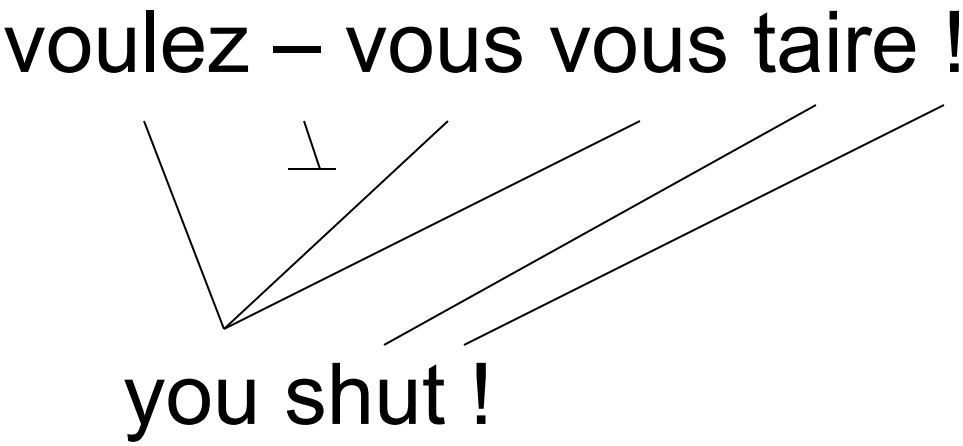
Search for Best Translation



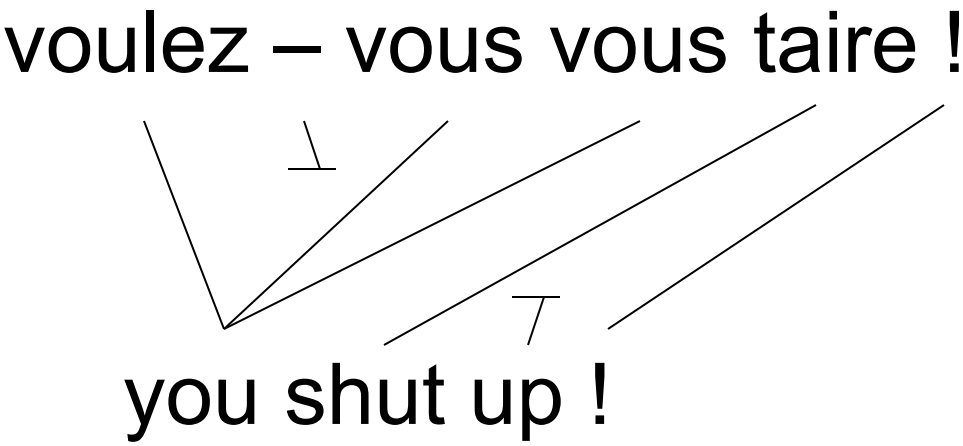
Search for Best Translation



Search for Best Translation



Search for Best Translation



Classic Decoding Algorithm

Given f, find the English string e that maximizes $P(e) \cdot P(f \mid e)$

NP-Complete [Knight 99].

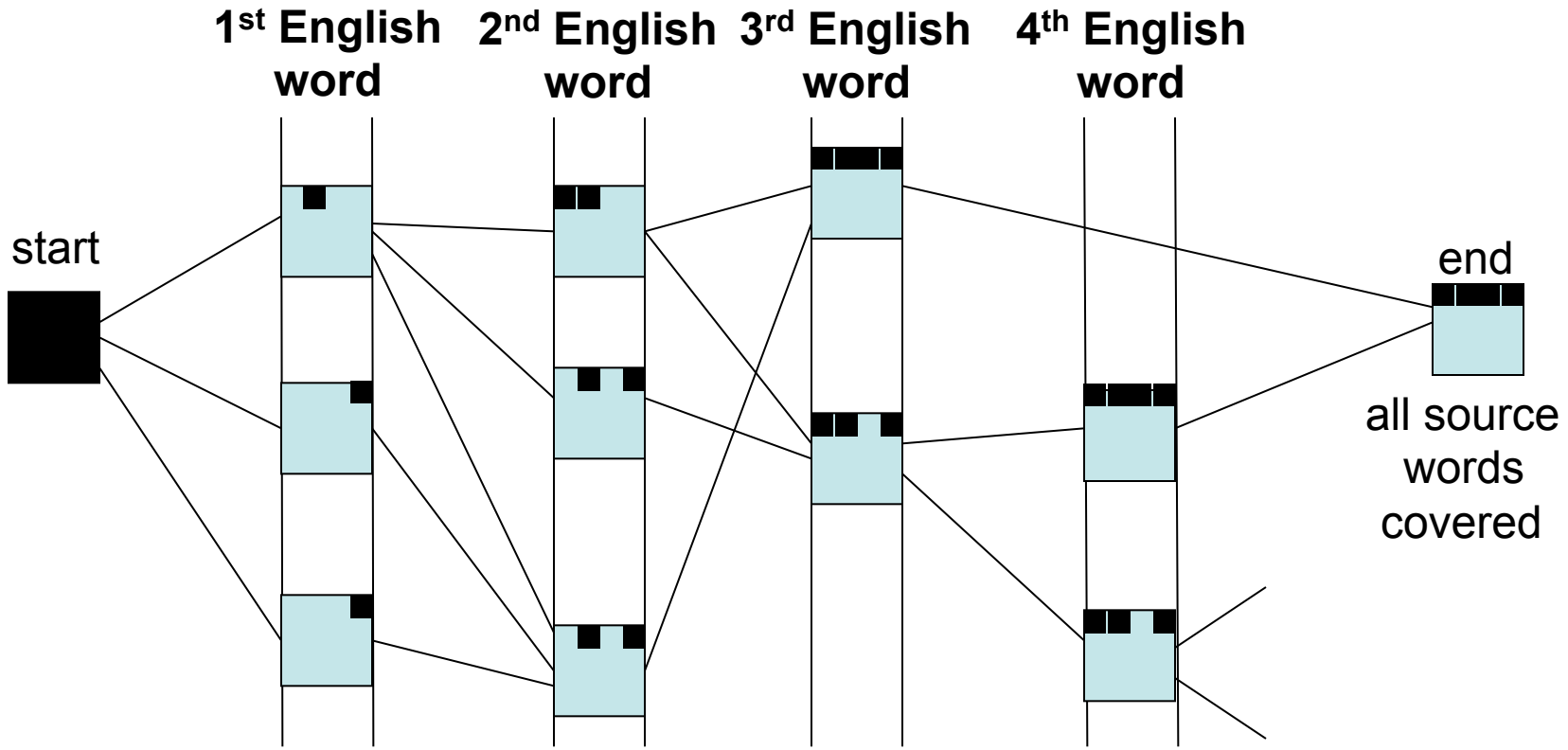
Brown et al 93:

> "In this paper, we focus on the translation modeling problem. We hope to deal with the [decoding] problem in a later paper."

Beam search can be used instead

Beam Search Decoding

[Brown et al US Patent #5,477,451]

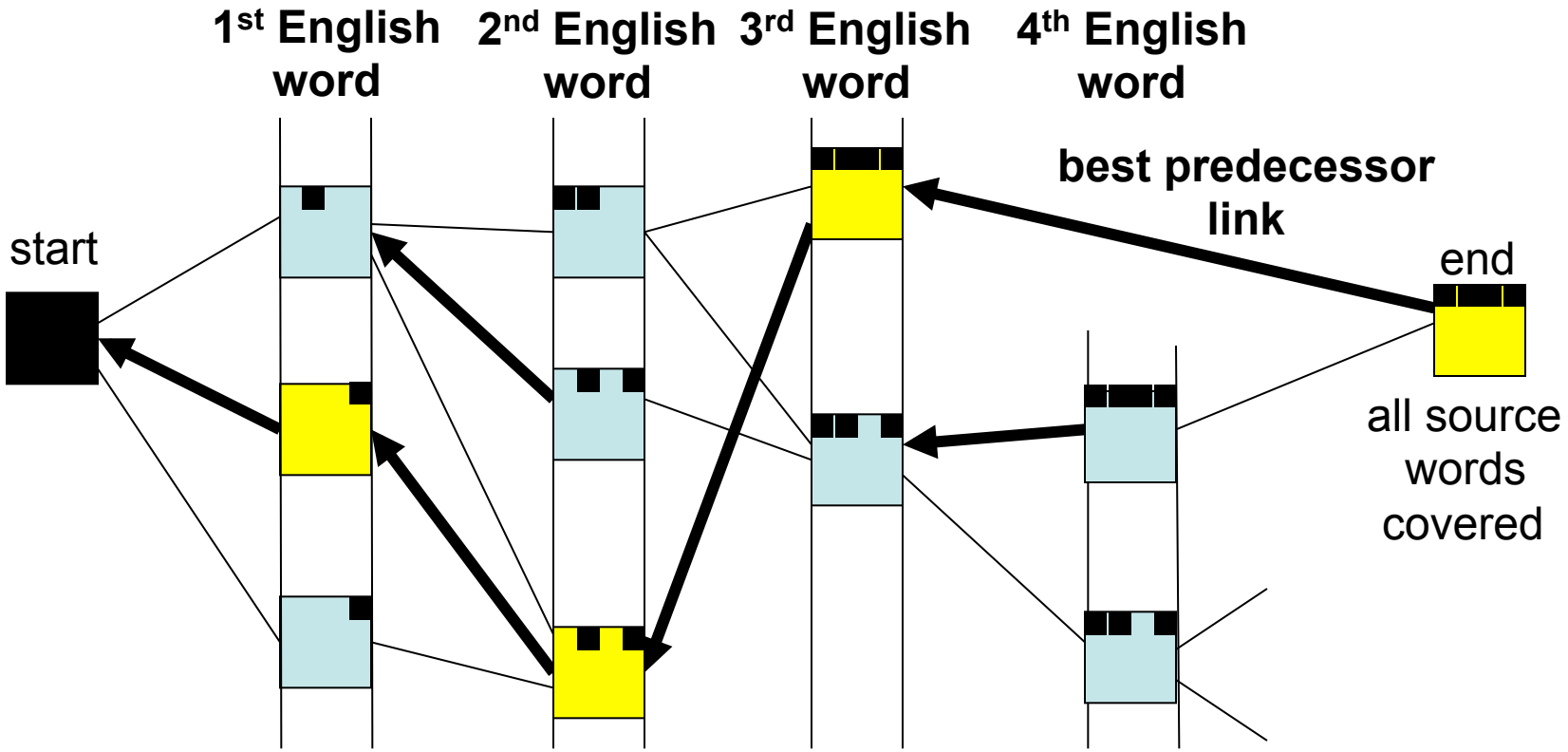


Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek 69;
Och, Ueffing, and Ney, 01]

Beam Search Decoding

[Brown et al US Patent #5,477,451]



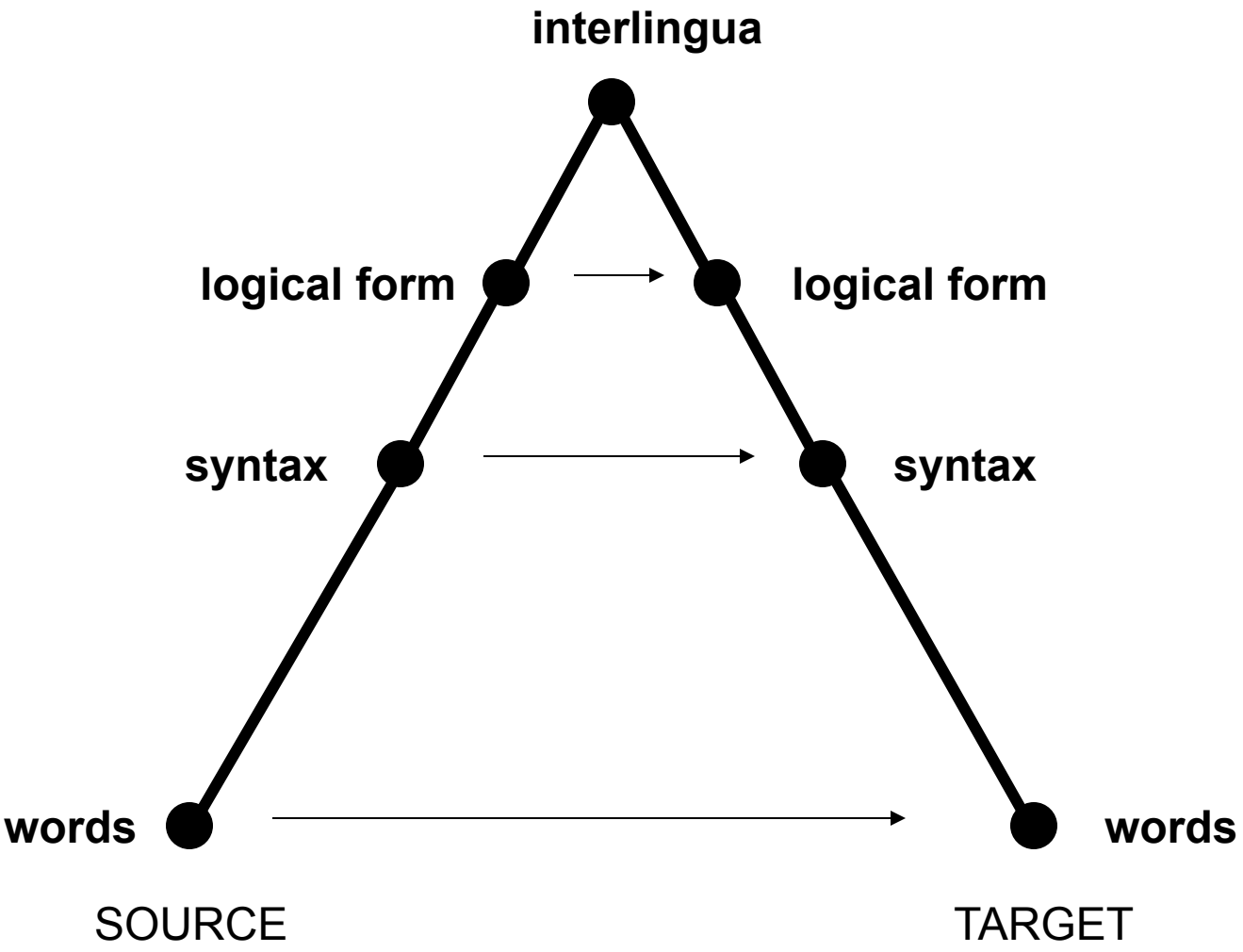
Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek 69;
Och, Ueffing, and Ney, 01]

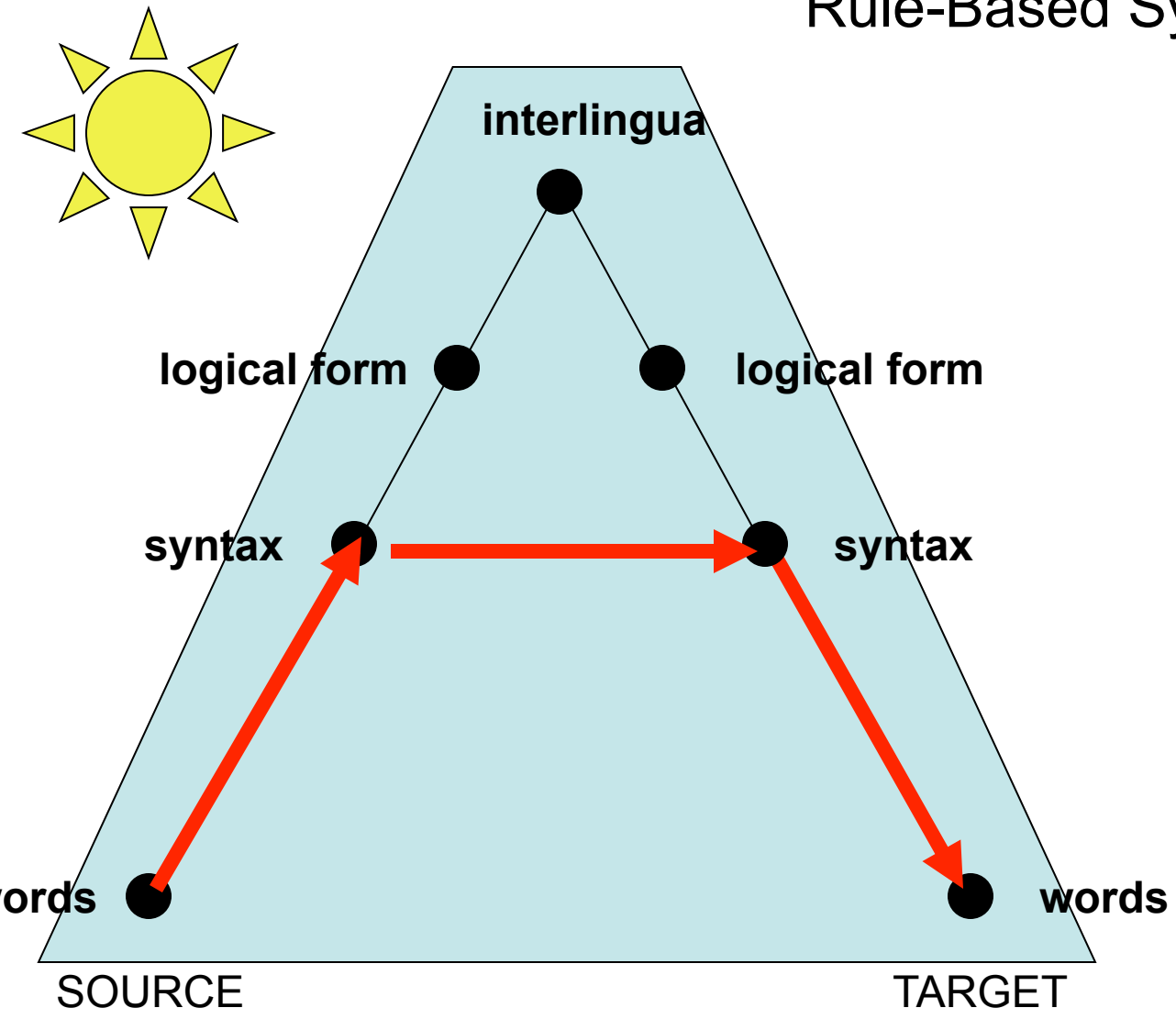
Flaws of Word-Based MT

- Can't translate multiple English words to one French word
- Can't translate phrases
  - "real estate", "note that", "interest in"
- Isn't sensitive to syntax
  - Adjectives/nouns should swap order
  - Verb comes at the beginning in Arabic
- Doesn't understand the meaning (?)

The MT Triangle

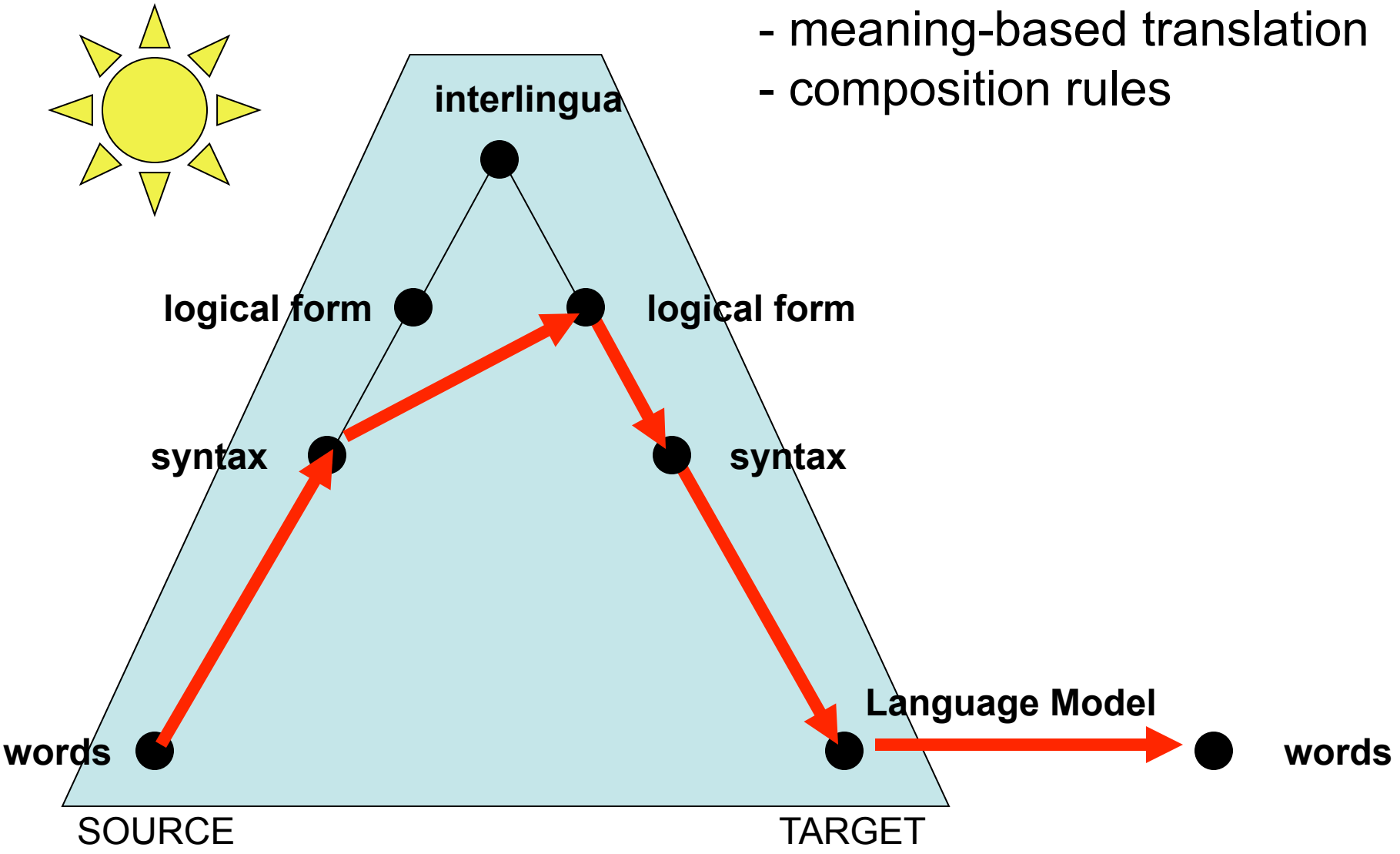


Commercial
Rule-Based Systems



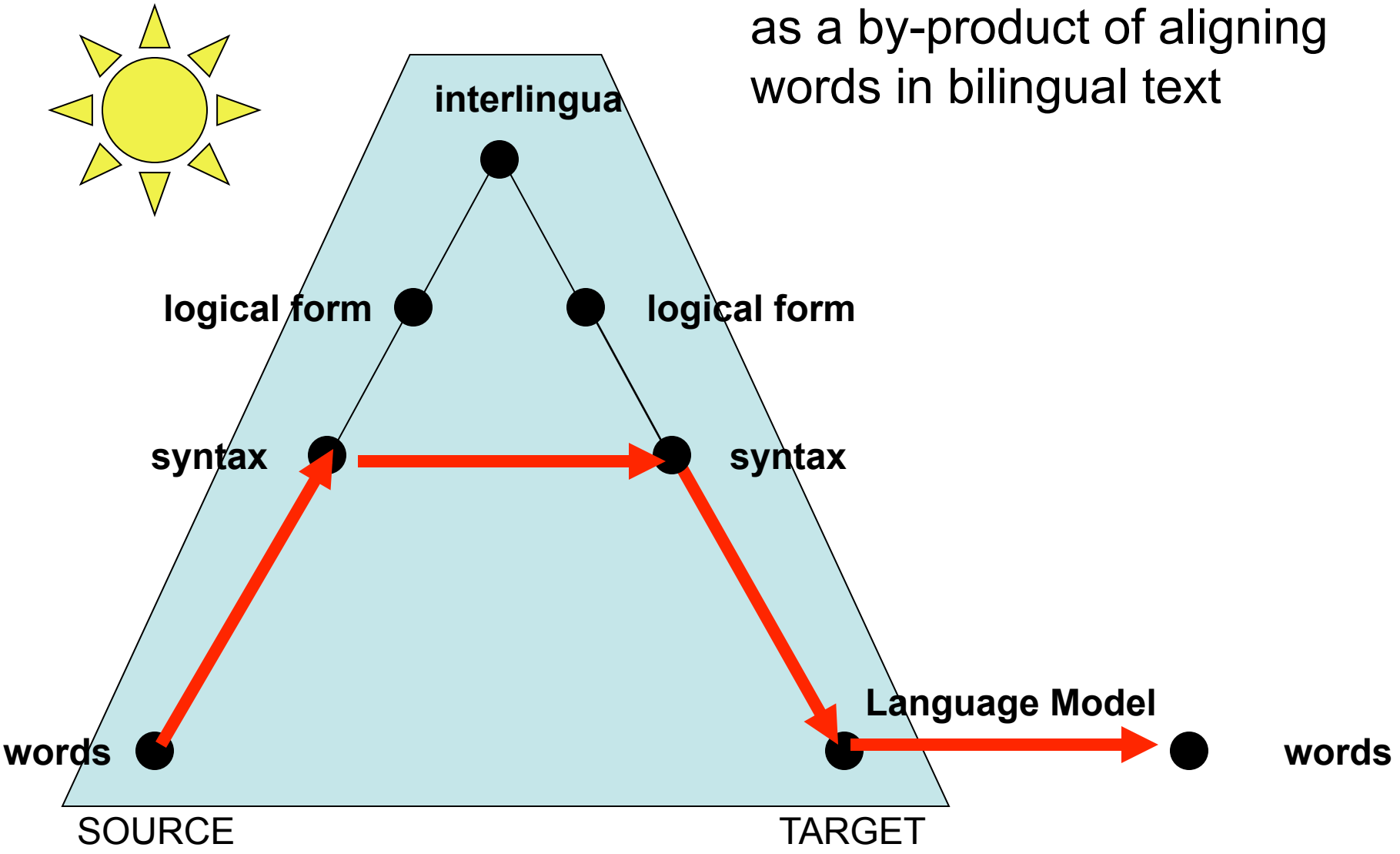
Knight et al 95

- meaning-based translation
- composition rules



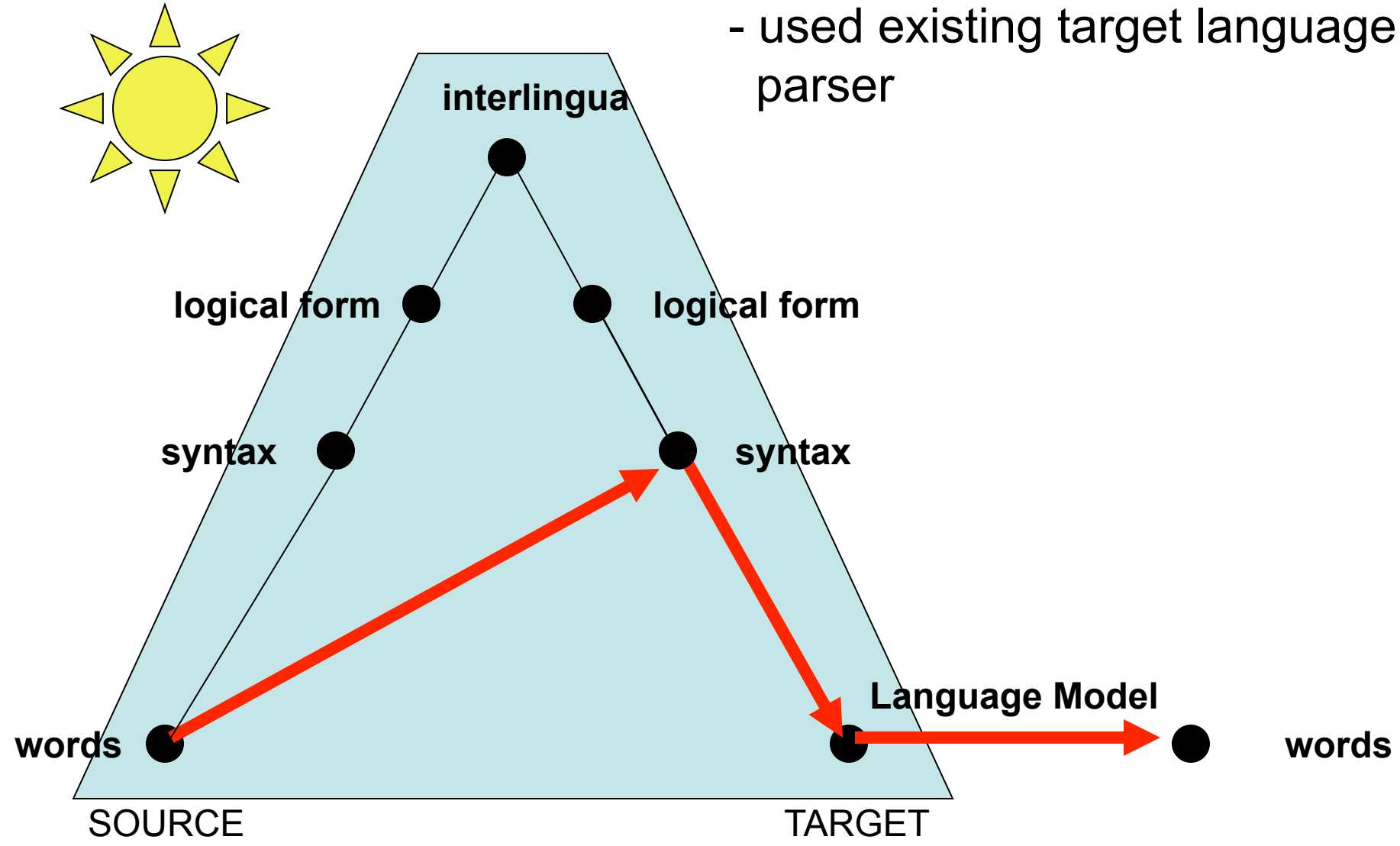
Wu 97, Alshawi 98

- inducing syntactic structure as a by-product of aligning words in bilingual text

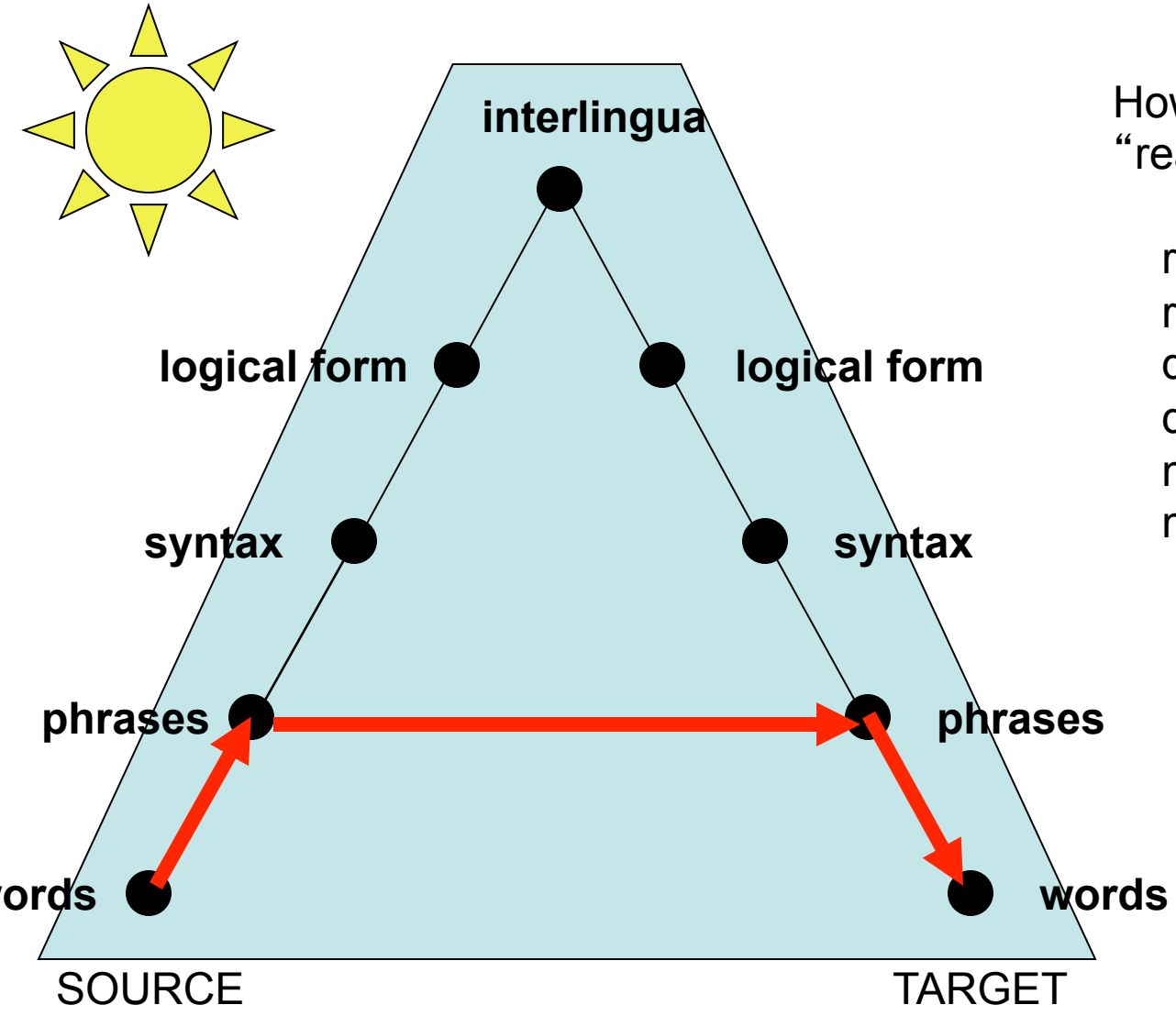


Yamada/Knight (01,02)

- tree/string model
- used existing target language parser



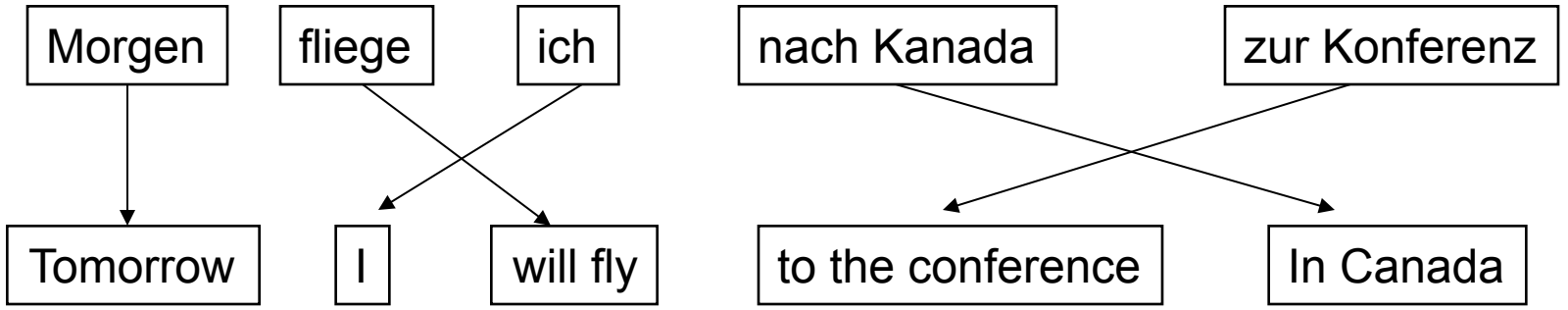
Phrases



How do you translate
“real estate” into French?

real estate
real number
dance number
dance card
memory card
memory stick
…

Phrase-Based Statistical MT



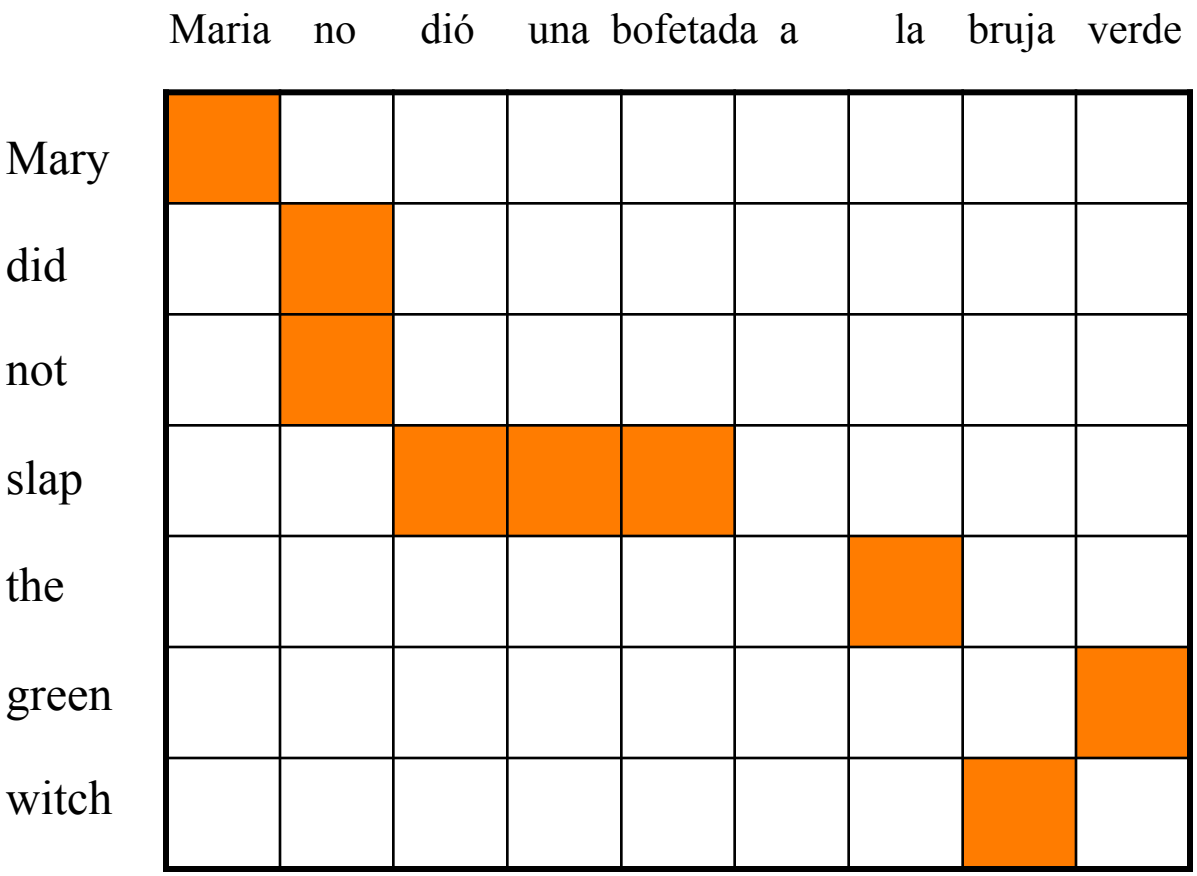
- Foreign input segmented into phrases
  - "phrase" just means "word sequence"
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered

See [Koehn et al, 2003] for an overview.

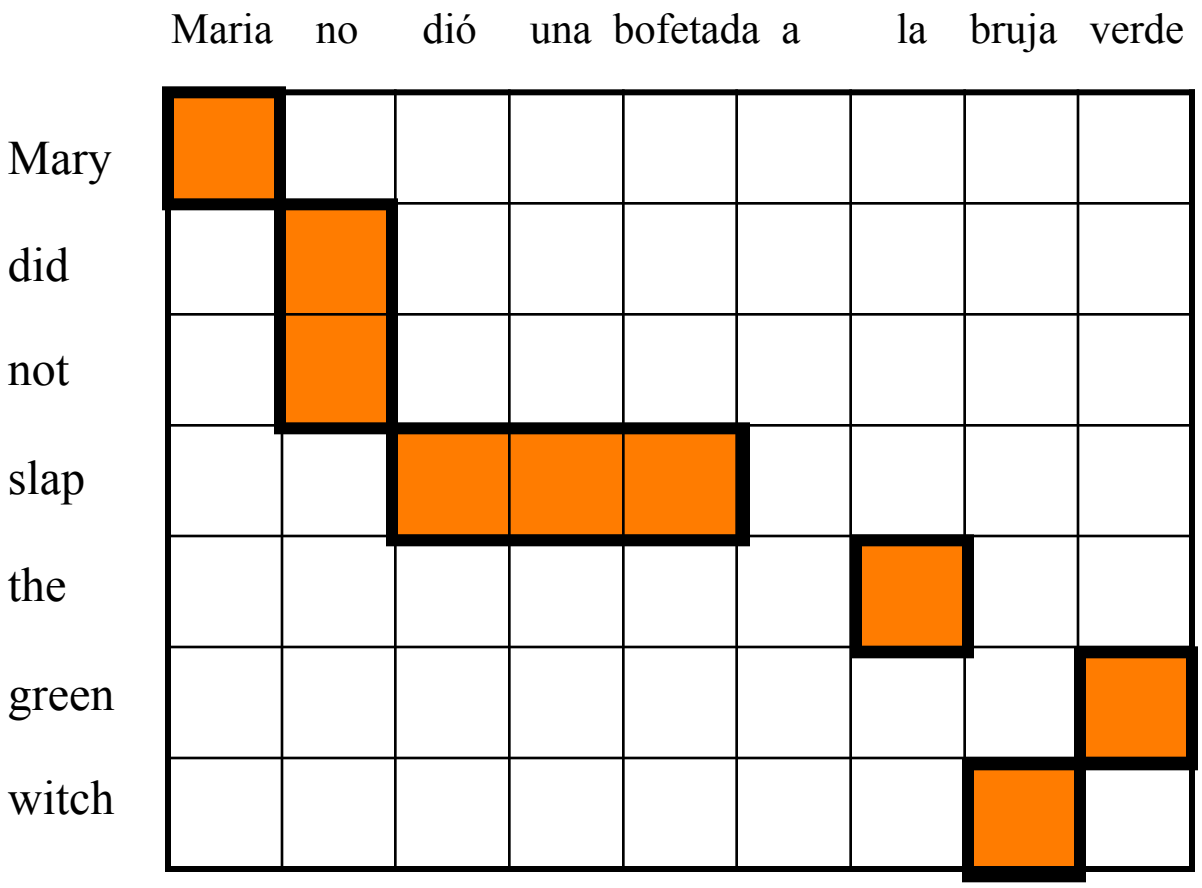
How to Learn the Phrase Translation Table?

- One method: “alignment templates” [Och et al 99]
- Start with word alignment
- Collect all phrase pairs that are *consistent with the word alignment*

Word Alignment Induced Phrases

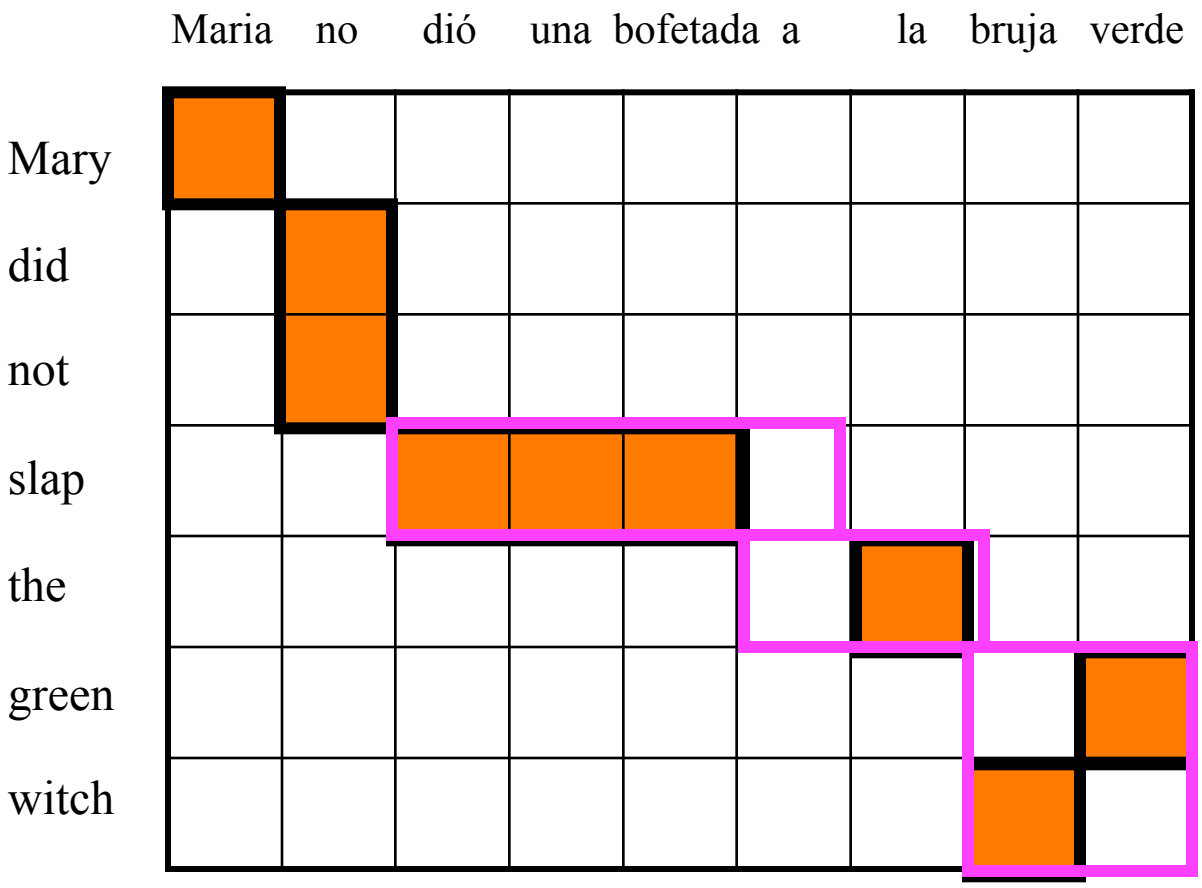


Word Alignment Induced Phrases



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

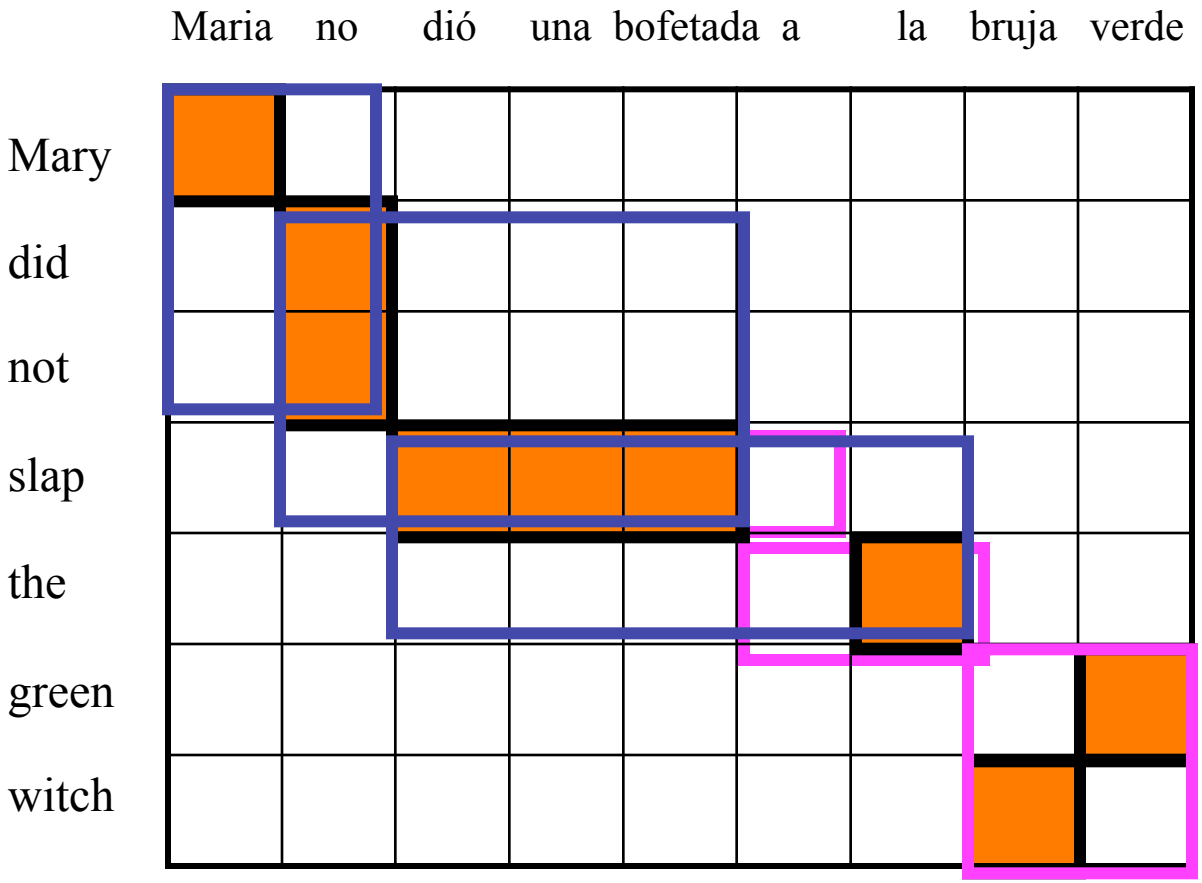
Word Alignment Induced Phrases



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

(a la, the) (dió una bofetada a, slap) (bruja verde, green witch)

Word Alignment Induced Phrases

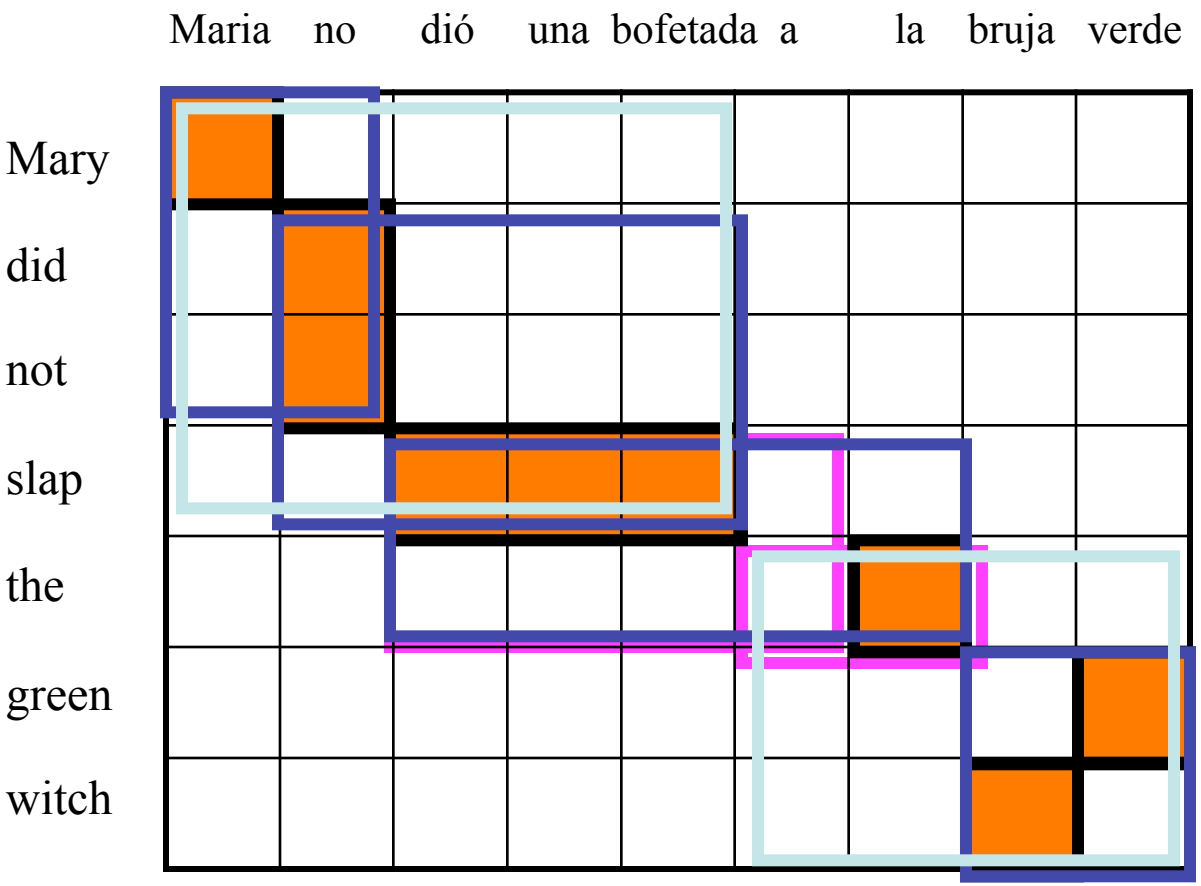


(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(a la, the) (dió una bofetada a, slap) (bruja verde, green witch)

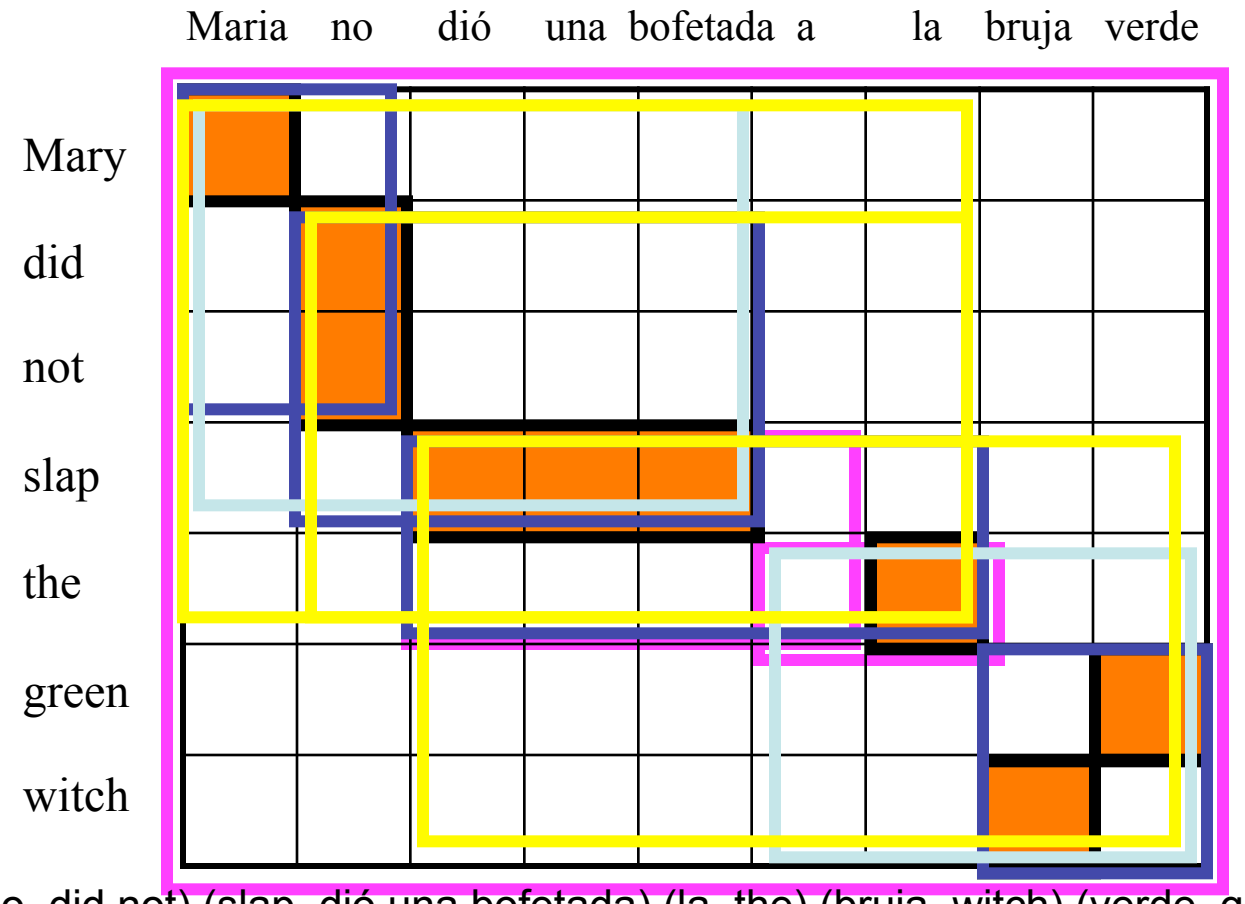
(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)

Word Alignment Induced Phrases



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

Word Alignment Induced Phrases



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

Phrase Pair Probabilities

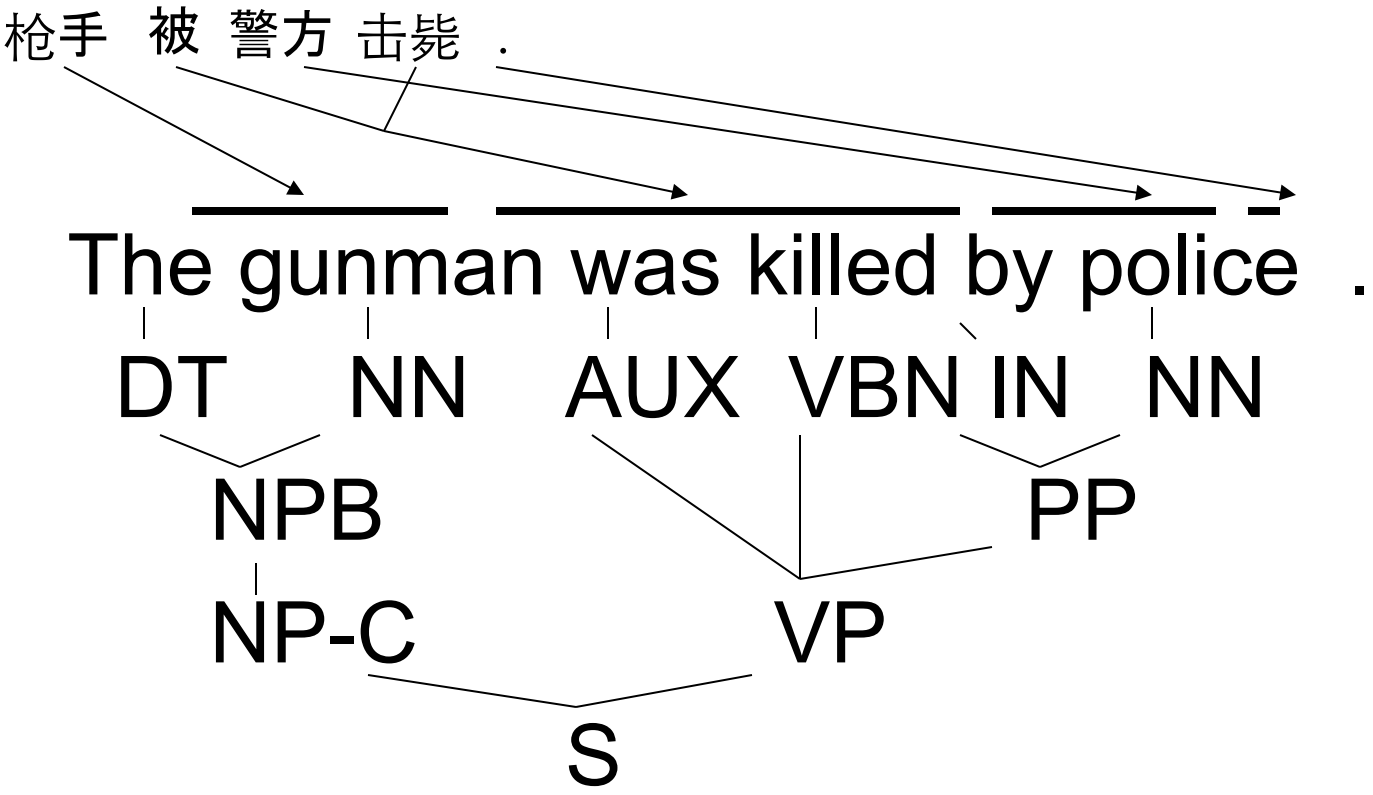
- A certain phrase pair (f-f-f, e-e-e) may appear many times across the bilingual corpus.
- No EM training
- Just relative frequency:

$$P(\text{f-f-f} \mid \text{e-e-e}) = \frac{\text{count(f-f-f, e-e-e)}}{\text{count(e-e-e)}}$$

Phrase-Based MT

- It was the best way to do Statistical MT until very recently
- Now syntax starts play the role

Tree Output


枪手 被 警方 击毙 .
The gunman was killed by police .
DT NN AUX VBN IN NN
NPB PP
NP-C VP
S

Synchronous CFGs [Chiang, 2005]

- Developed in the 60's for programming-language compilation [Aho1969]
- In NLP synchronous CFGs have been used for
  - Machine translation
  - Semantic interpretation

Synchronous CFGs [Chiang, 2005]

- Like CFGs, but production have two right hand sides
  - *Source* side
  - *Target* side
  - Related through linked non-terminal symbols
    - E.g. VP → <V[1] NP[2],NP[2] V[1]>
    - One-to-one correspondence
    - Productions applied in parallel to both sides to linked non-terminals

Synchronous CFGs [Chiang, 2005]

$$\langle S_{\boxed{1}}, S_{\boxed{1}} \rangle \Rightarrow \langle S_{\boxed{2}} X_{\boxed{3}}, S_{\boxed{2}} X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle S_{\boxed{4}} X_{\boxed{5}} X_{\boxed{3}}, S_{\boxed{4}} X_{\boxed{5}} X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle X_{\boxed{6}} X_{\boxed{5}} X_{\boxed{3}}, X_{\boxed{6}} X_{\boxed{5}} X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \text{Aozhou } X_{\boxed{5}} X_{\boxed{3}}, \text{Australia } X_{\boxed{5}} X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \text{Aozhou shi } X_{\boxed{3}}, \text{Australia is } X_{\boxed{3}} \rangle$$

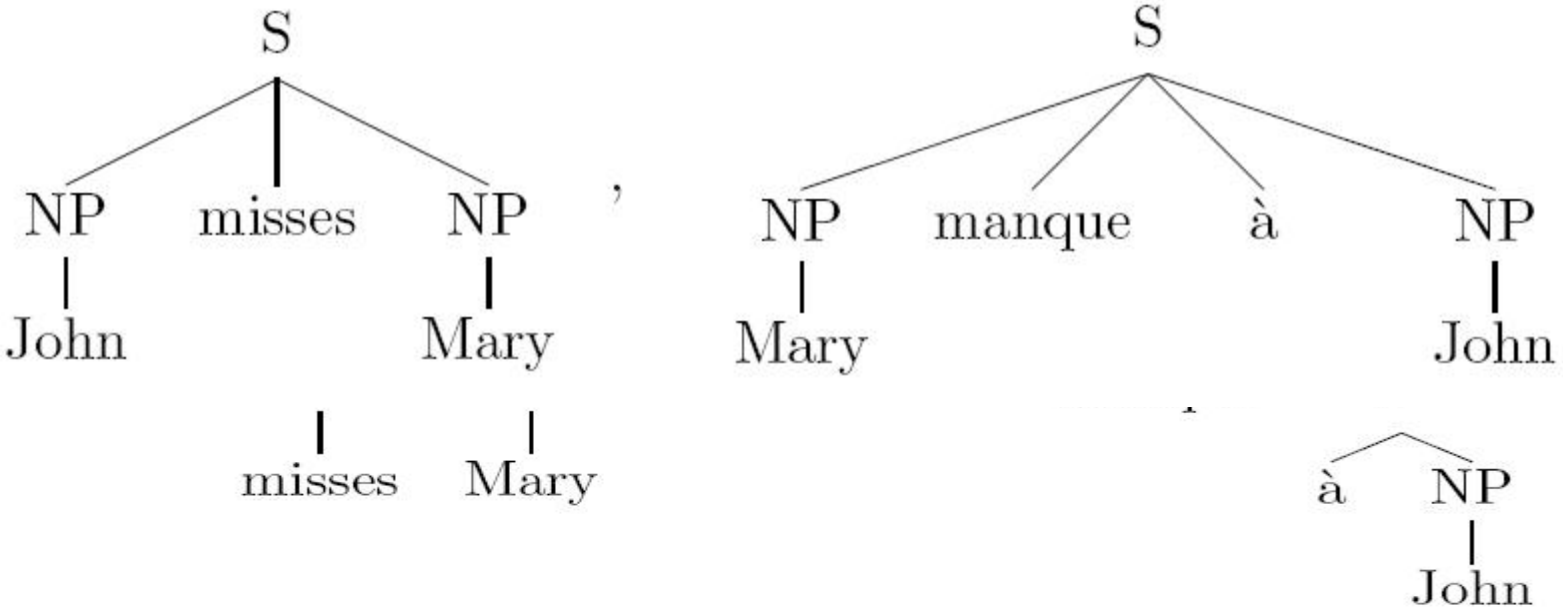
$$\Rightarrow \langle \text{Aozhou shi } X_{\boxed{7}} \text{ zhiyi, Australia is one of } X_{\boxed{7}} \rangle$$

$$\Rightarrow \langle \text{Aozhou shi } X_{\boxed{8}} \text{ de } X_{\boxed{9}} \text{ zhiyi, Australia is one of the } X_{\boxed{9}} \text{ that } X_{\boxed{8}} \rangle$$

$$\Rightarrow \langle \text{Aozhou shi yu } X_{\boxed{1}} \text{ you } X_{\boxed{2}} \text{ de } X_{\boxed{9}} \text{ zhiyi, Australia is one of the } X_{\boxed{9}} \text{ that have } X_{\boxed{2}} \text{ with } X_{\boxed{1}} \rangle$$

Synchronous CFGs [Chiang, 2005]

- Limitations
  - No Chomsky normal form
    - Has implications for complexity of decoder
  - Sister-reordering only



Summary MT

- An important application
- There has been an important progress
- Interdisciplinary work
  - Natural language processing
  - Machine learning
  - Linguistics
  - Automata theory
- More classes in Masters of AI