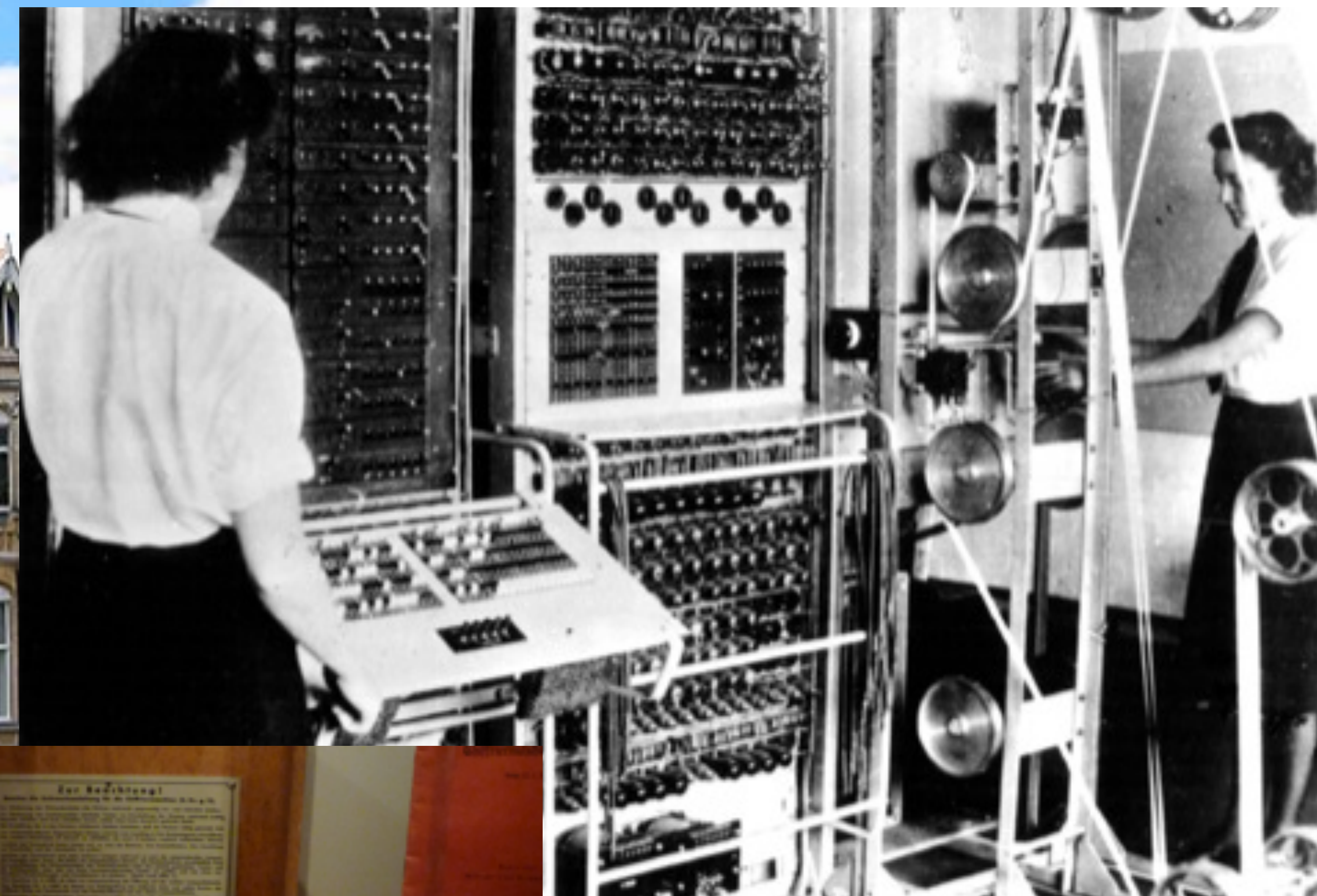
Statistical Machine Translation

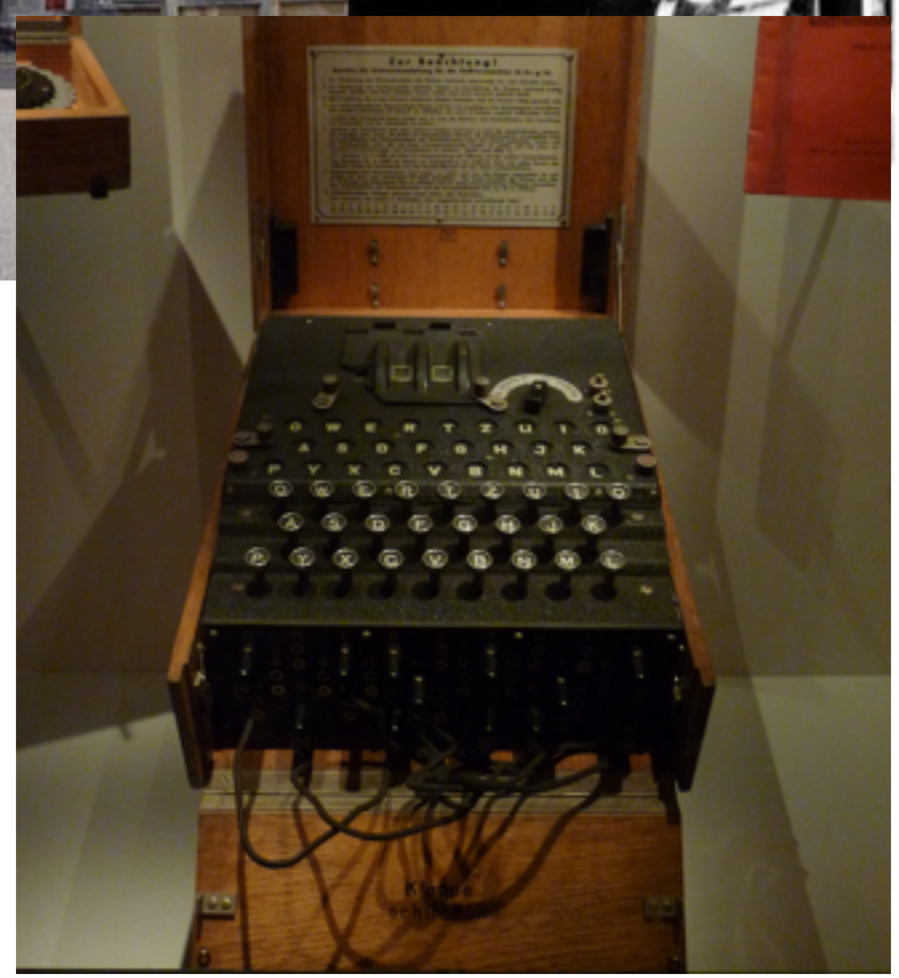
Chris Dyer

2011 MT Marathon - Trento - FBK

Outline

- What is statistical machine translation?
- A quick survey:
  - Language modeling
  - Phrase-based translation and decoding
  - Word alignment
  - Translation evaluation

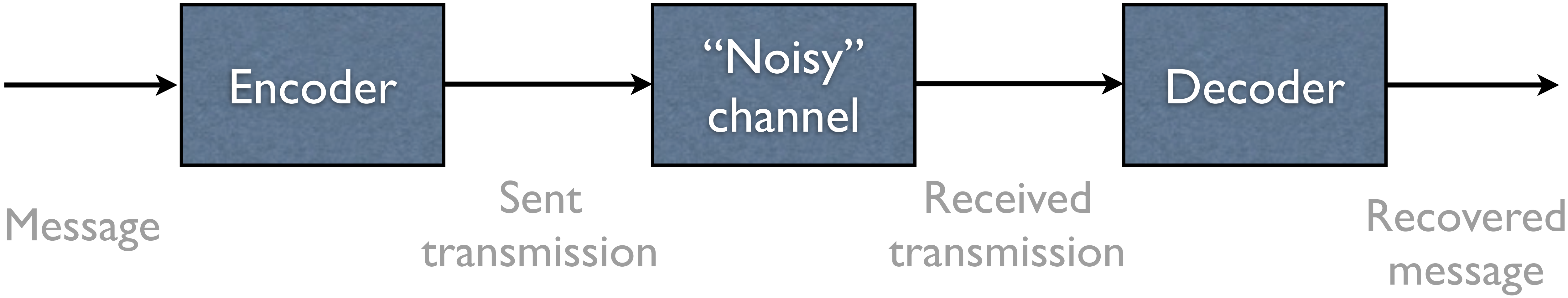




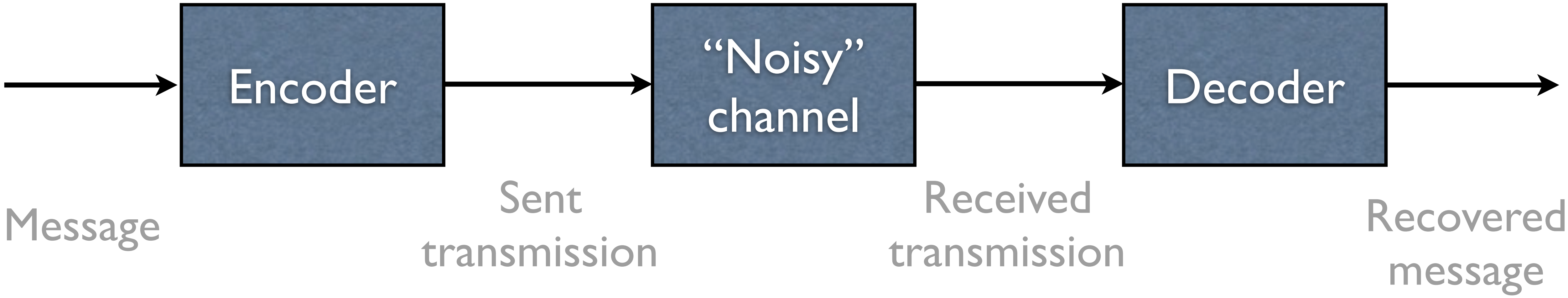
One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: '***This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode***.'

Warren Weaver to Norbert Wiener, March, 1947

How do we model coding problems like this?



Claude Shannon. "A Mathematical Theory of Communication" 1948.



Shannon's theory tells us:

1) the limits of compression
2) why your download is so slow
3) how to recognize speech
4) **how to translate**

Claude Shannon. "A Mathematical Theory of Communication" 1948.

Probability and language

- Event spaces are the output spaces of various processes that **generate language**
- Language is a discrete combinatoric system
  - (Nice math: sums instead of integrals!)
- Probability is robust: even with inaccurate models, we can do well
- The "art" of probability modeling is coming up with models that are easy to work with ***and*** close to realty

Example

Imagine a many-sided die,
only instead of numbers,
there are **English words**.

By rolling this die, we can "generate" a single word.

Example

Imagine a many-sided die,
only instead of numbers,
there are **English words**.

By rolling this die, we can "generate" a single word.

What if we want to generate **sentences**?

Example

Imagine a many-sided die,
only instead of numbers,
there are **English words**.

By rolling this die, we can "generate" a single word.

What if we want to generate **sentences**?

After each word, flip a coin to decide
whether to **stop**, or **roll again**.

Some notation

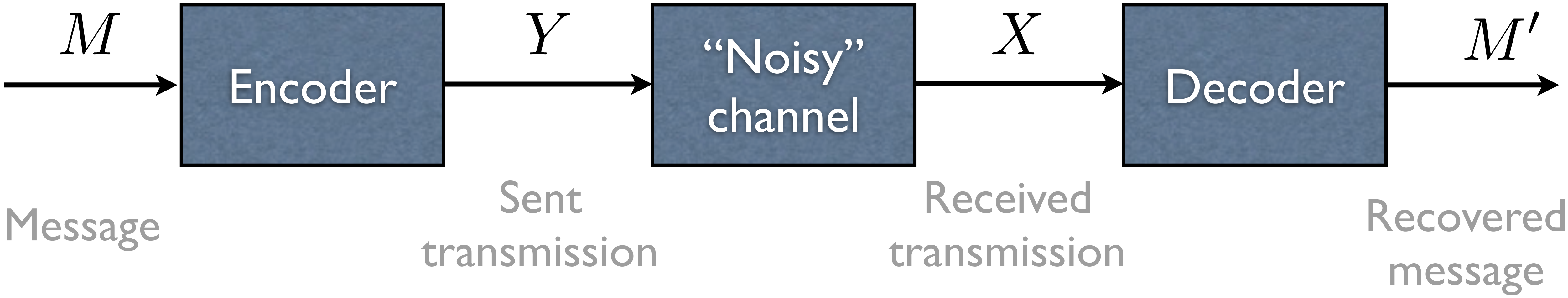
$$P(A, B) = \text{the probability that both } A \text{ and } B \text{ occur}$$

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \quad \left(= \frac{P(A \cap B)}{P(B)} \right)$$

$$P(A, B) = P(B, A)$$

$$P(A \mid B) \neq P(B \mid A)$$

Bivariate models are be useful for relating words/sentences/documents in two languages.



Claude Shannon. "A Mathematical Theory of Communication" 1948.

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

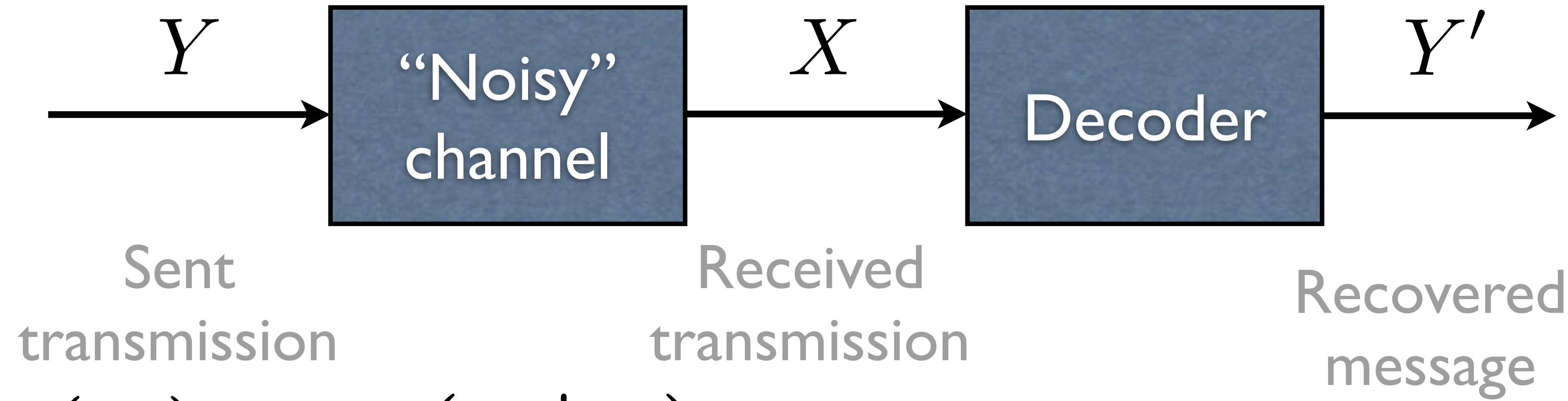
Claude Shannon. "A Mathematical Theory of Communication" 1948.

$$\boxed{p(y) \quad p(x|y)} \quad p(x)$$

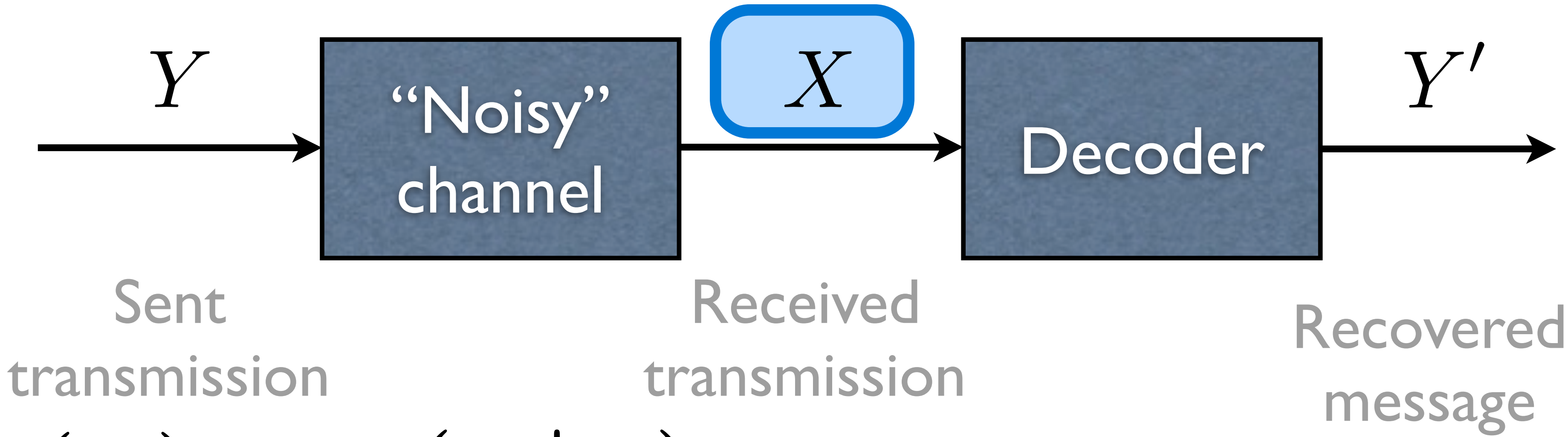
Claude Shannon. "A Mathematical Theory of Communication" 1948.

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

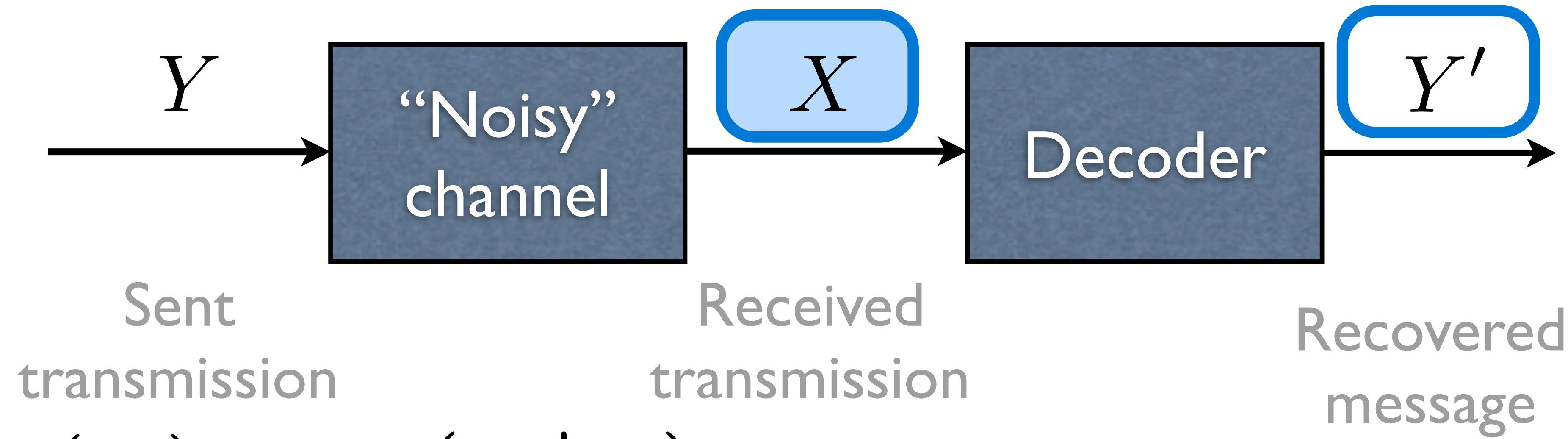
Claude Shannon. “A Mathematical Theory of Communication” 1948.


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

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

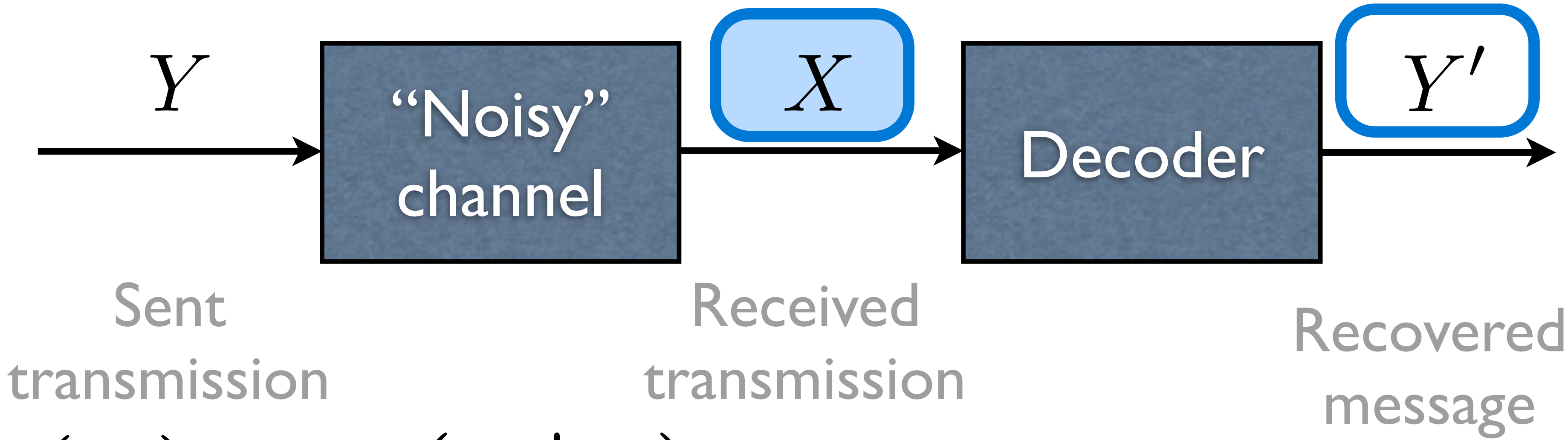

Y
"Noisy" channel
X
Decoder
Y'
Sent transmission
Received transmission
Recovered message

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



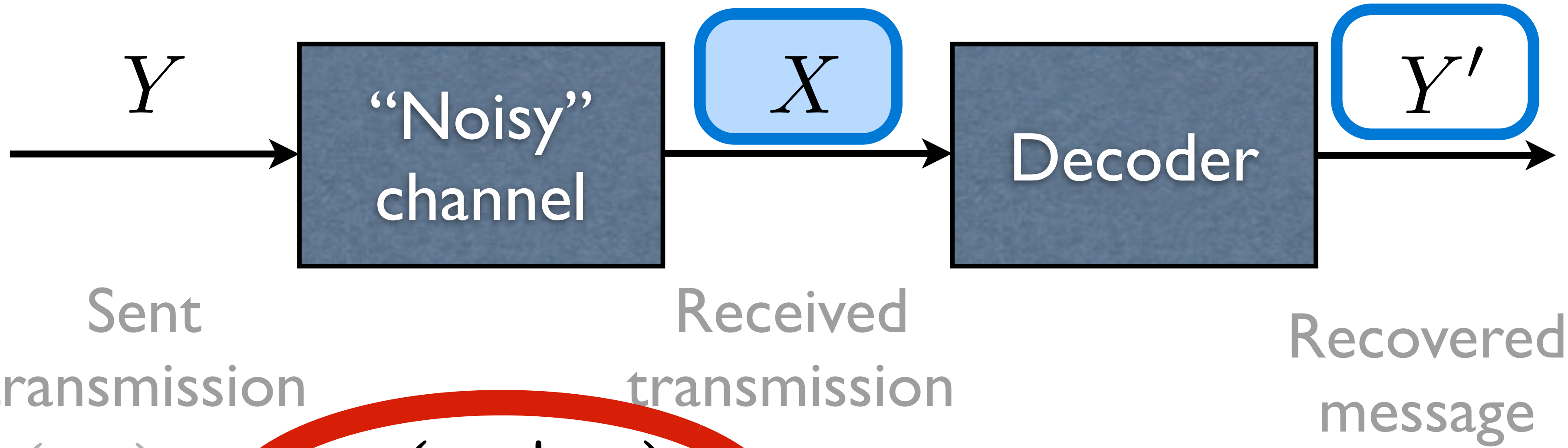
Y
"Noisy" channel
X
Decoder
Y'
Sent transmission
Received transmission
Recovered message

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



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

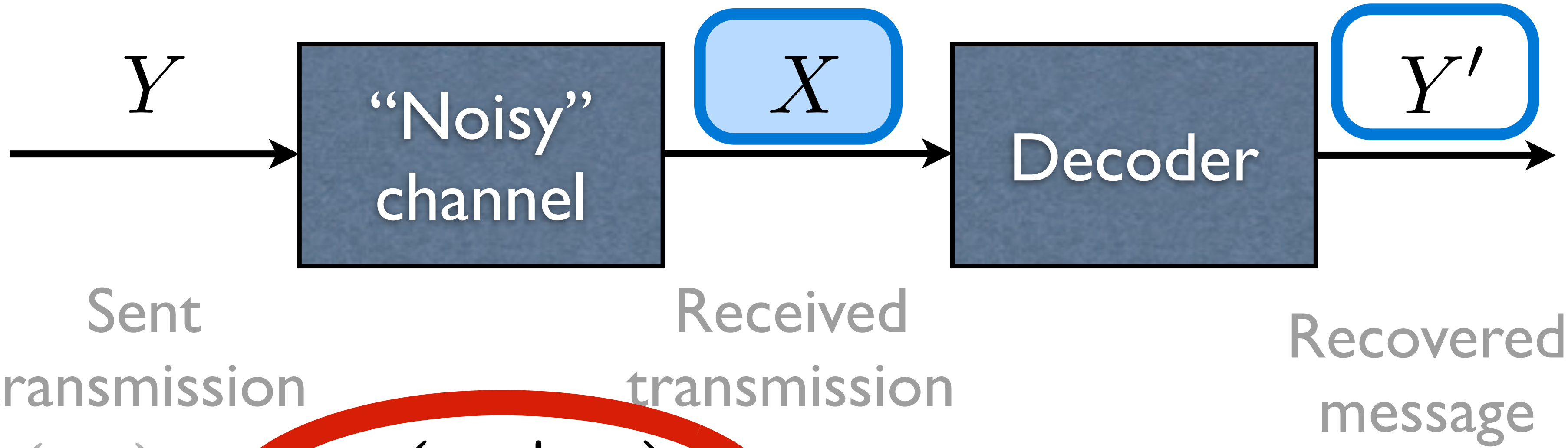
$$\boxed{y'} = \arg\max_{y} p(y|x)$$



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

$$\propto$$

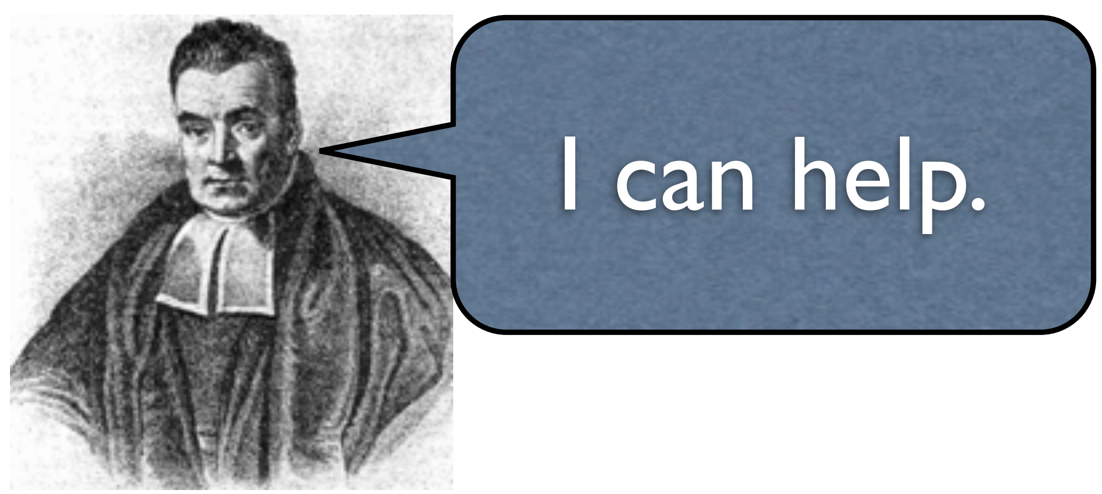
$$\boxed{y'} = \arg\max_y p(y|x)$$

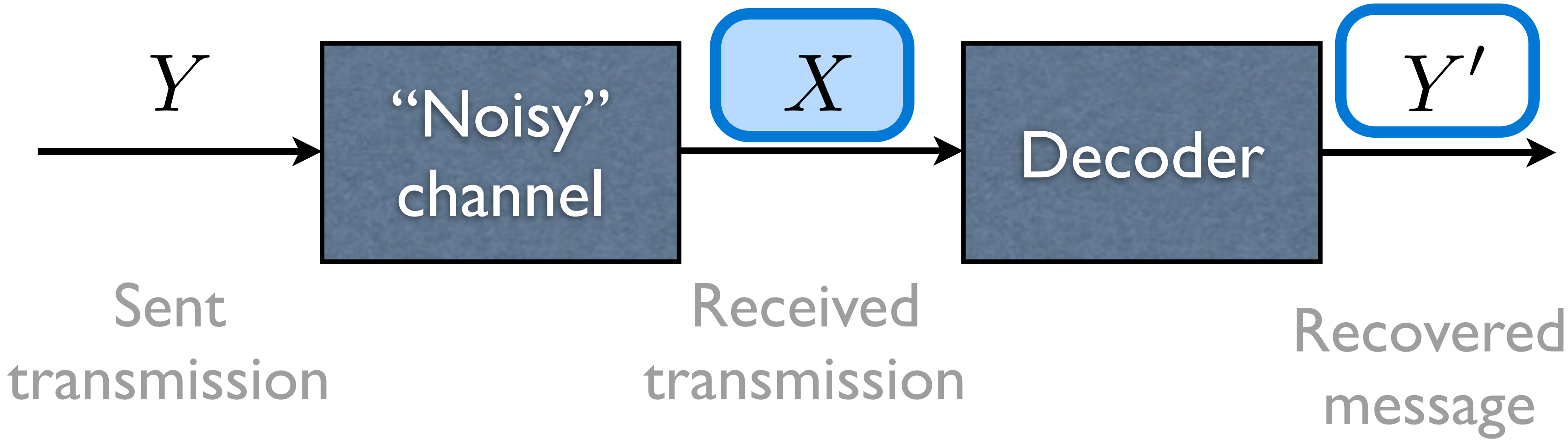


$$p(y) \quad p(x|y)$$

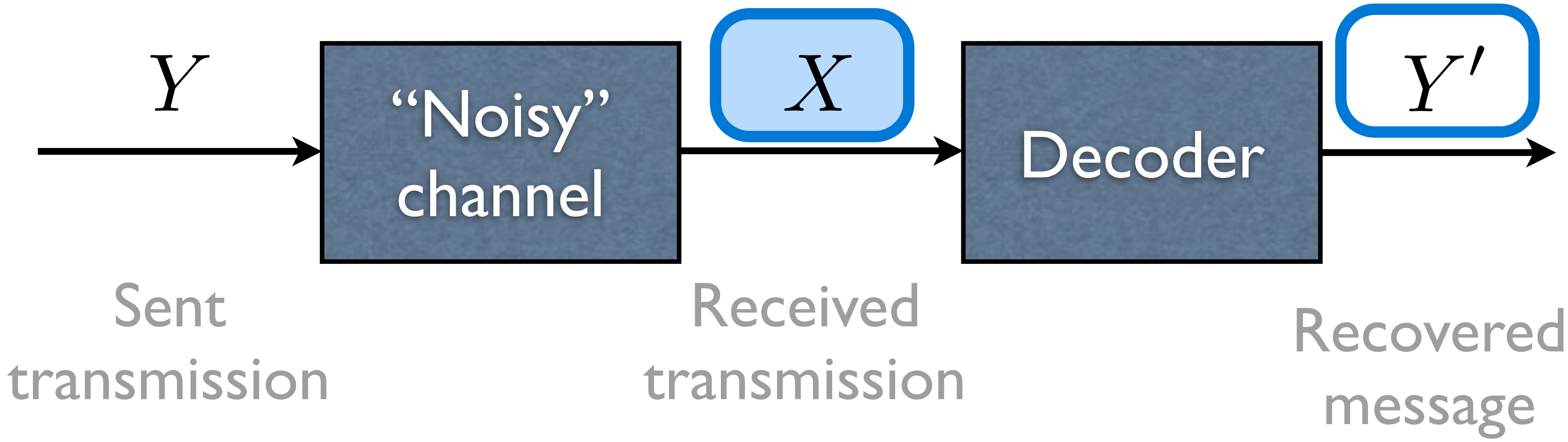
$$\not\propto$$

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





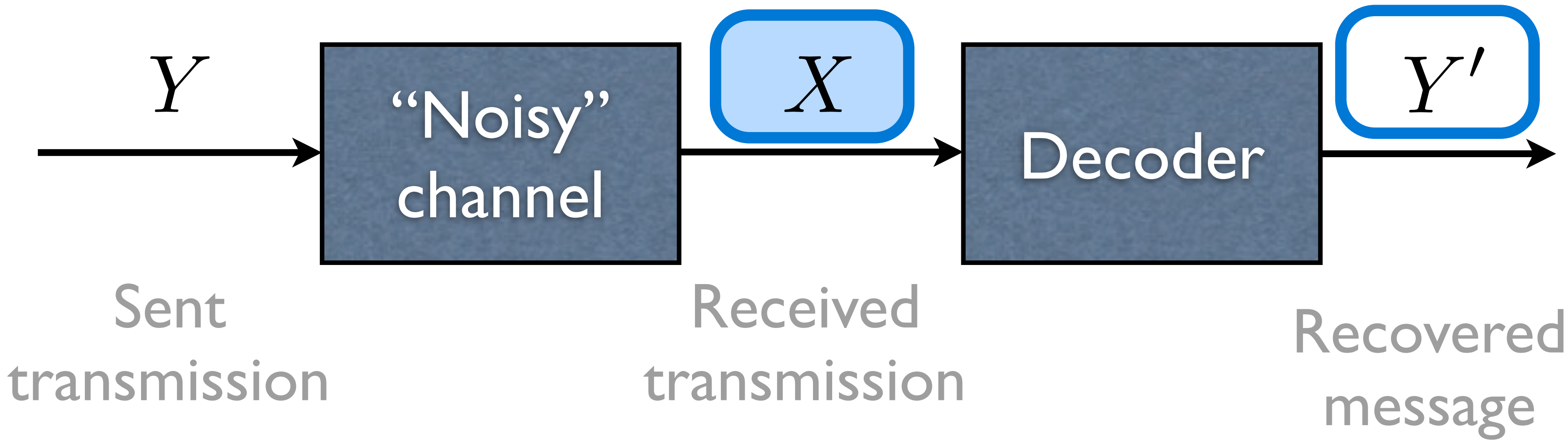
$$\begin{aligned}\boxed{y'} &= \arg\max_{y} p(y|x) \\ &= \arg\max_{y} \frac{p(x|y)p(y)}{p(x)}\end{aligned}$$



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

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

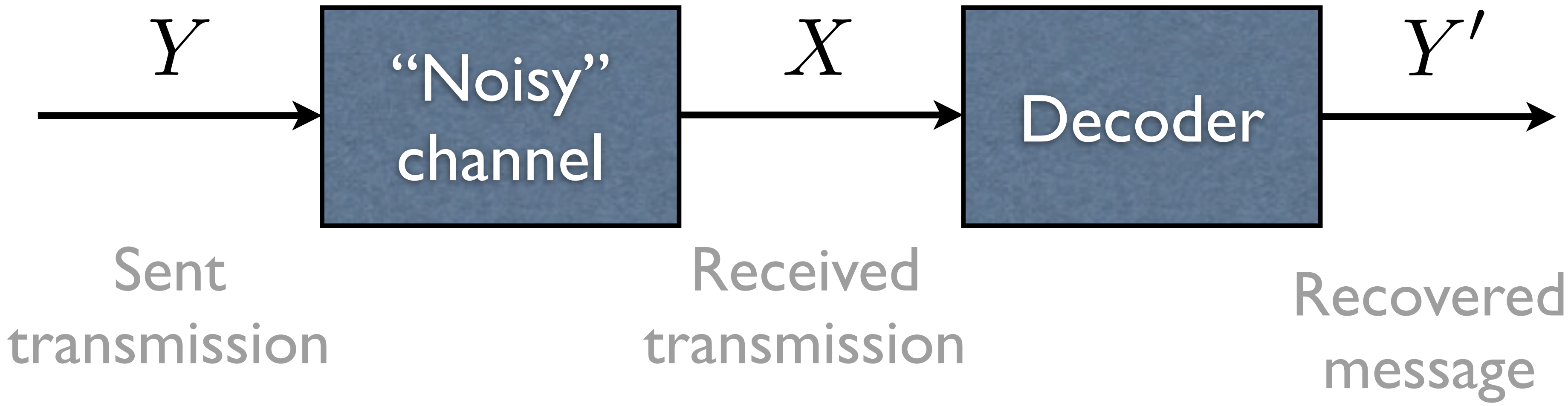
Denominator doesn't depend on y.



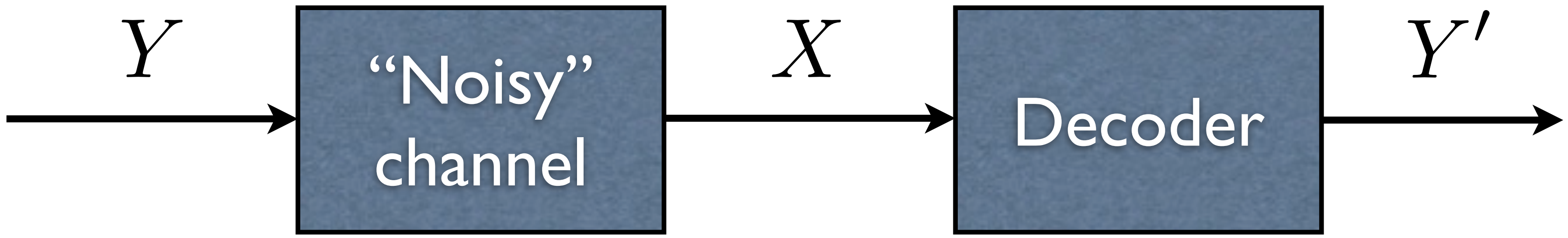
$$\boxed{y'} = \arg\max_y p(y|x)$$

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

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



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



Sent transmission

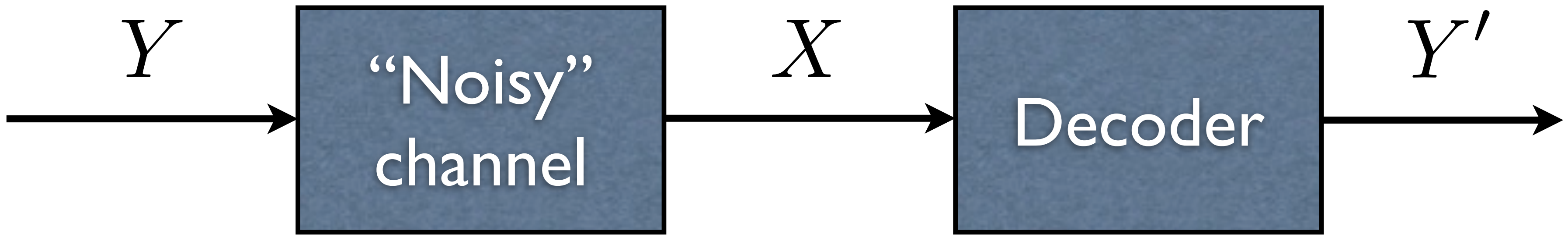
Received transmission

Recovered message

English **“French”** **English’**

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

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

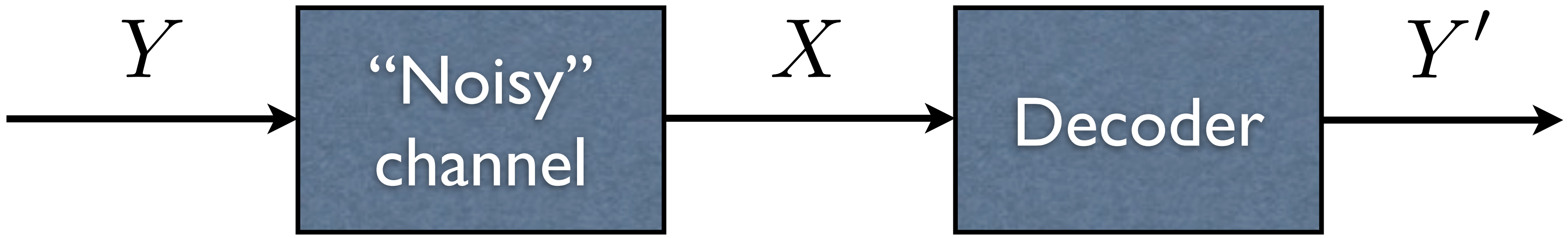


Sent transmission	Received transmission	Recovered message
English	**"French"**	**English'**

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

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

translation model



Sent transmission

English

Received transmission

"French"

Recovered message

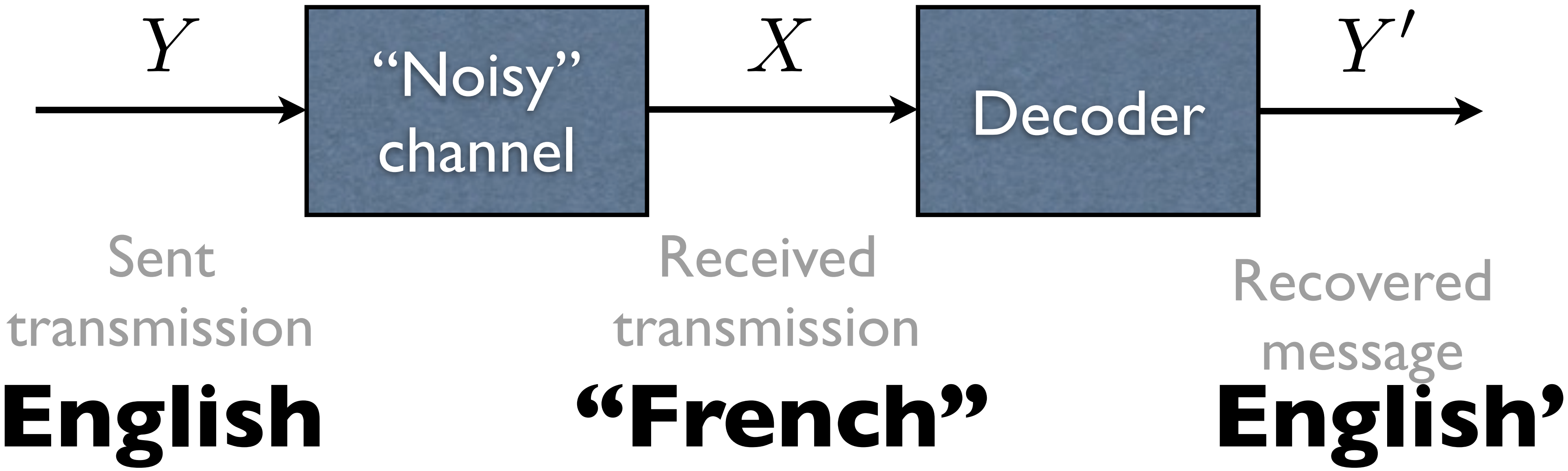
English'

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

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

translation model

language model



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

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

translation model   language model

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

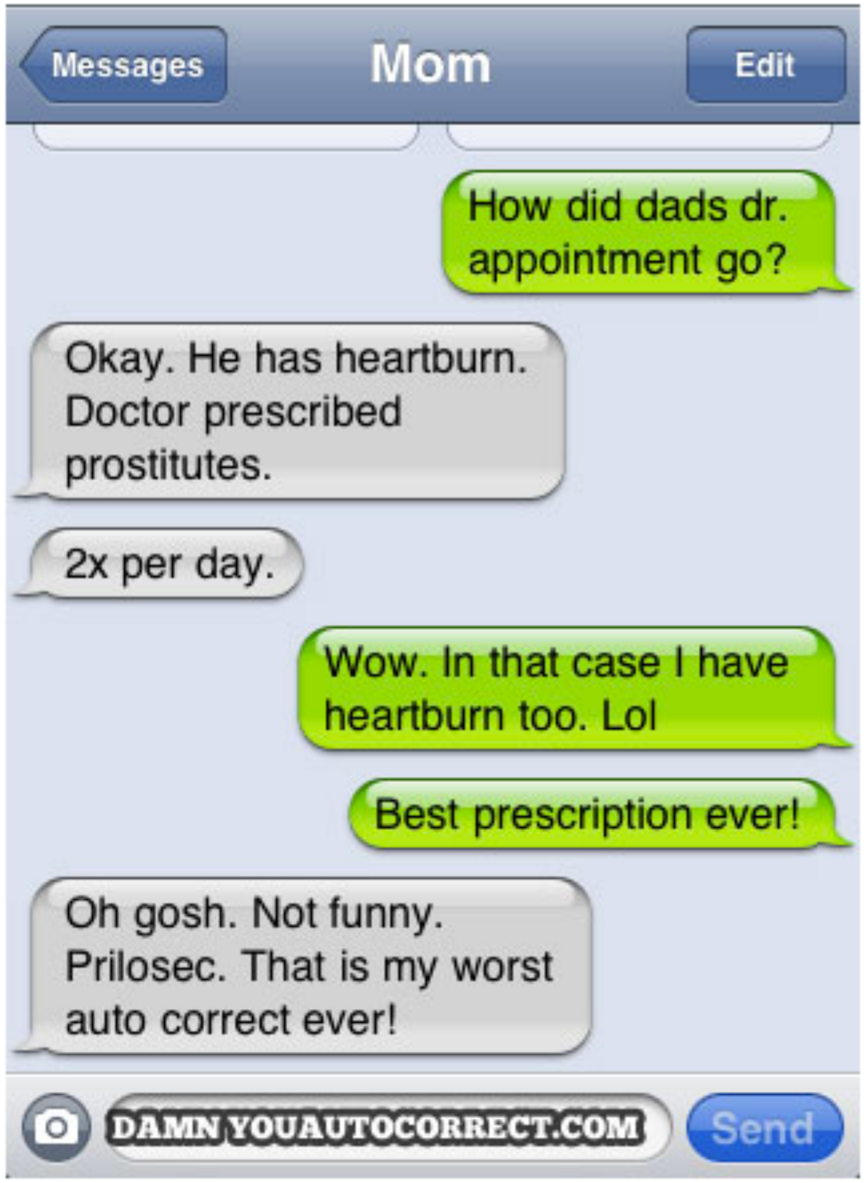
Division of labor

- **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

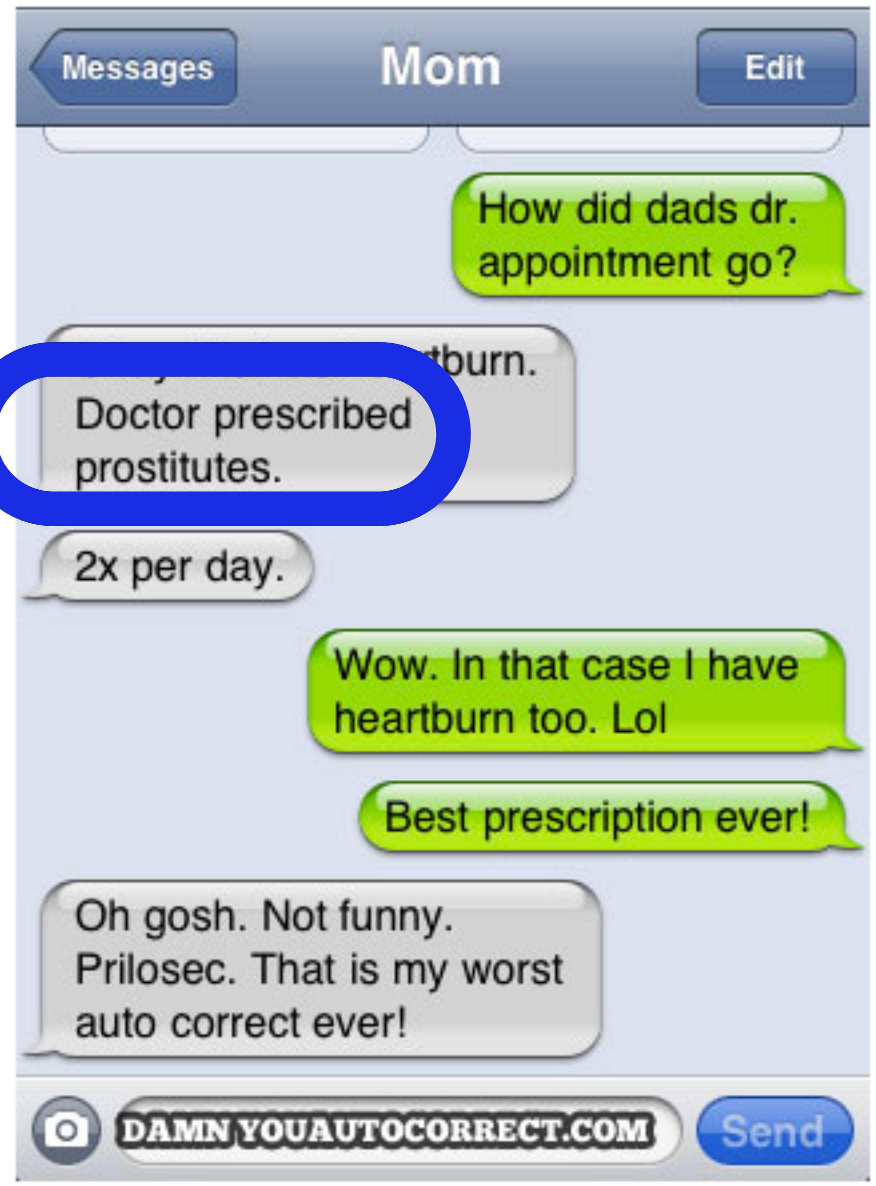
Language Model

- p(e) is typically modeled with n-grams
  (Lecture later this week)
- State-of-the-art in MT is 5-grams
- Why does context matter?

DAMN YOU, AUTO CORRECT!



DAMN YOU, AUTO CORRECT!



Translation Model

$$p(\mathbf{f}|\mathbf{e})$$

Q: What translates into what?

A: Look in a dictionary?

A: Everything into everything?

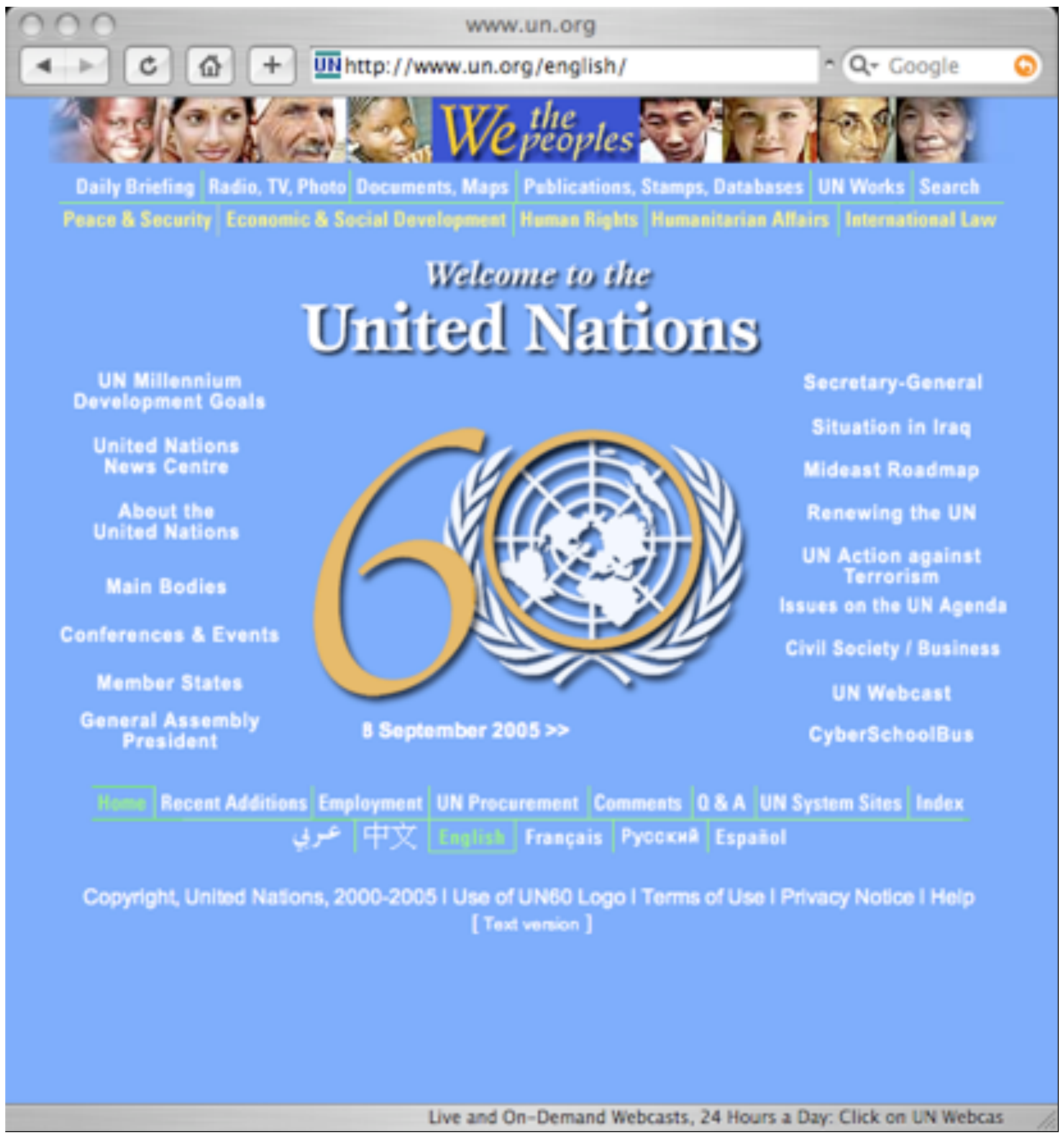
Q: And with what probability?

A: Ask people what they think?

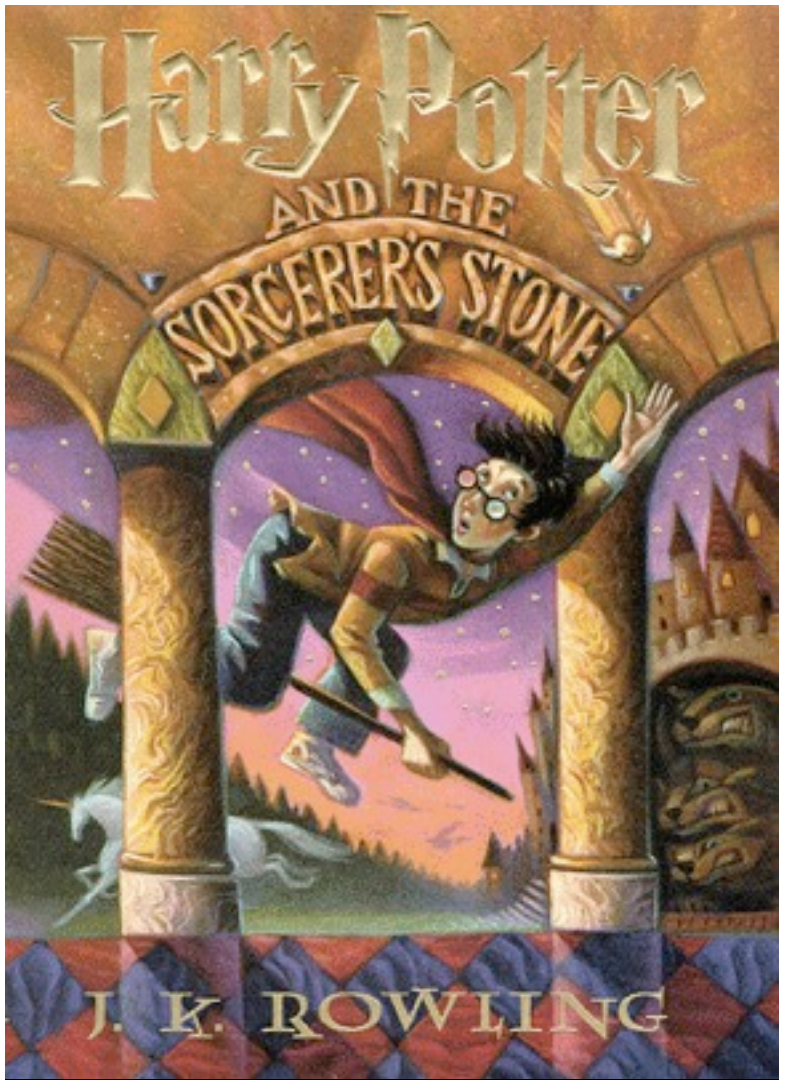
A: Let's figure it out from data!

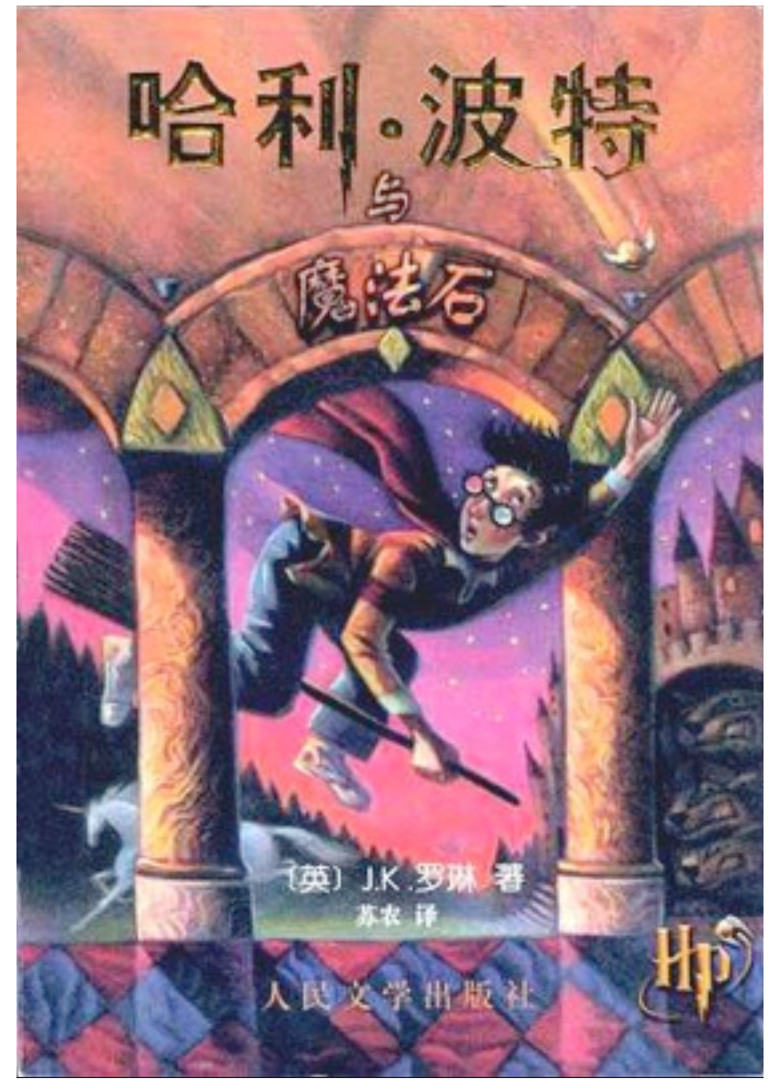
CLASSIC SOUPS

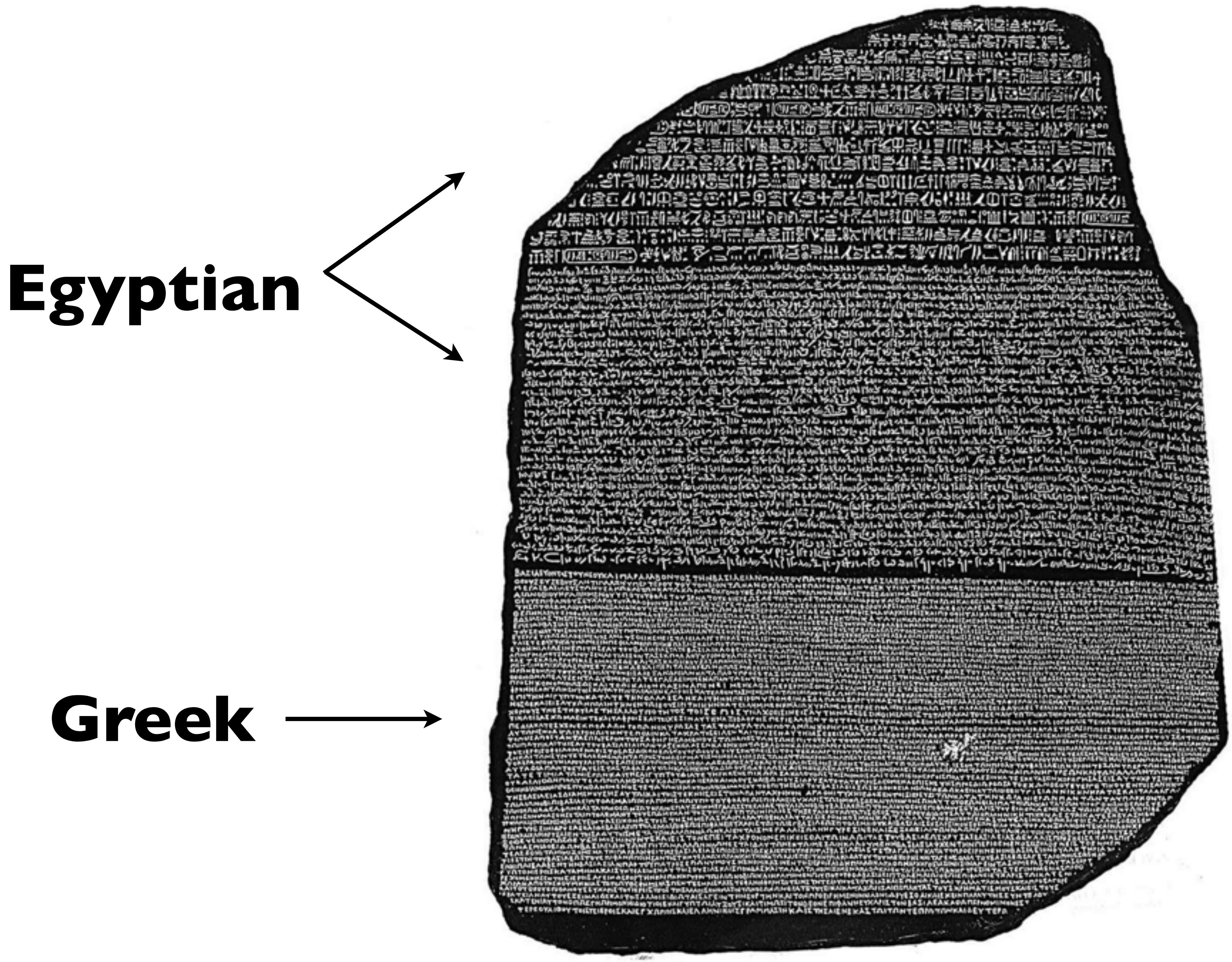
			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
蛋花湯	64.	Egg Drop Soup	1.10	2.10
雲蛋湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50


www.un.org
http://www.un.org/english/
Daily Briefing | Radio, TV, Photo | Documents, Maps | Publications, Stamps, Databases | UN Works | Search
Peace & Security | Economic & Social Development | Human Rights | Humanitarian Affairs | International Law
Welcome to the
United Nations
UN Millennium Development Goals
United Nations News Centre
About the United Nations
Main Bodies
Conferences & Events
Member States
General Assembly President
8 September 2005 >>
Secretary-General
Situation in Iraq
Mideast Roadmap
Renewing the UN
UN Action against Terrorism
Issues on the UN Agenda
Civil Society / Business
UN Webcast
CyberSchoolBus
Home | Recent Additions | Employment | UN Procurement | Comments | Q & A | UN System Sites | Index
عربي | 中文 | English | Français | Русский | Español
Copyright, United Nations, 2000-2005 | Use of UN60 Logo | Terms of Use | Privacy Notice | Help
[Text version]
Live and On-Demand Webcasts, 24 Hours a Day: Click on UN Webcas

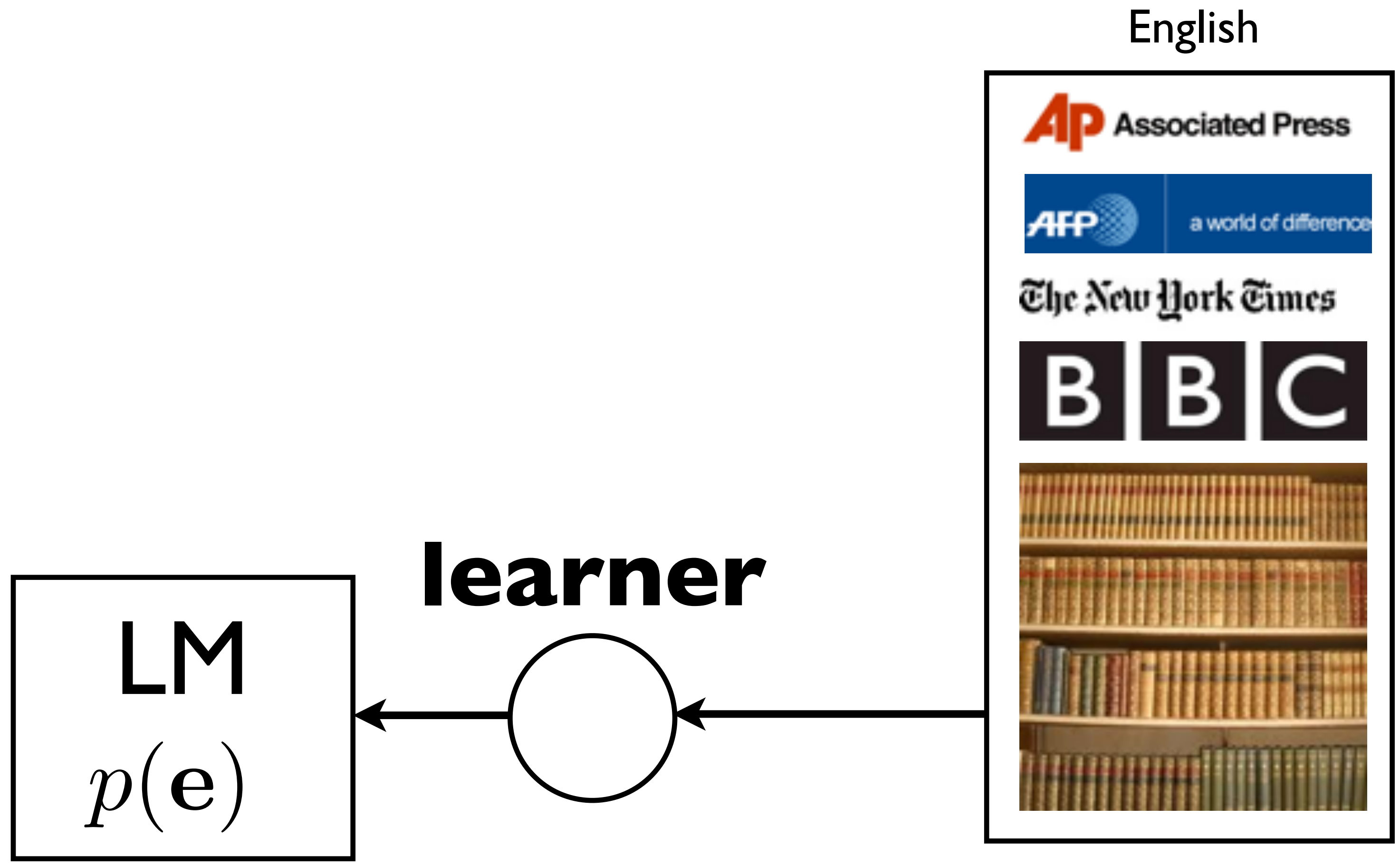

联合国主页
http://www.un.org/chinese/
我们人民
每日简报 | 多媒体 | 文件与地图 | 出版物 邮票 数据库 | 服务全球 | 网址搜索
和平与安全 | 经济与社会发展 | 人权 | 人道主义事务 | 国际法
欢迎来到
联合国
联合国千年发展目标
联合国新闻
联合国概况
联合国主要机关
会议与活动
联合国会员国
联合国大会主席
联大第60届会议一般性辩论
联合国秘书长
伊拉克局势
中东路线图
更新联合国
反恐怖主义
联合国日常议题
民间团体/商业
联合国网络直播
空中校车
新增内容 | 工作机会 | 联合国采购 | 建议 | 问题与解答 | 其他网址 | 网址索引
عربي | 中文 | English | Français | Русский | Español
联合国2000-2005年版权|联合国60周年徽标使用准则|使用条件|隐私通告|帮助
[纯文字版]

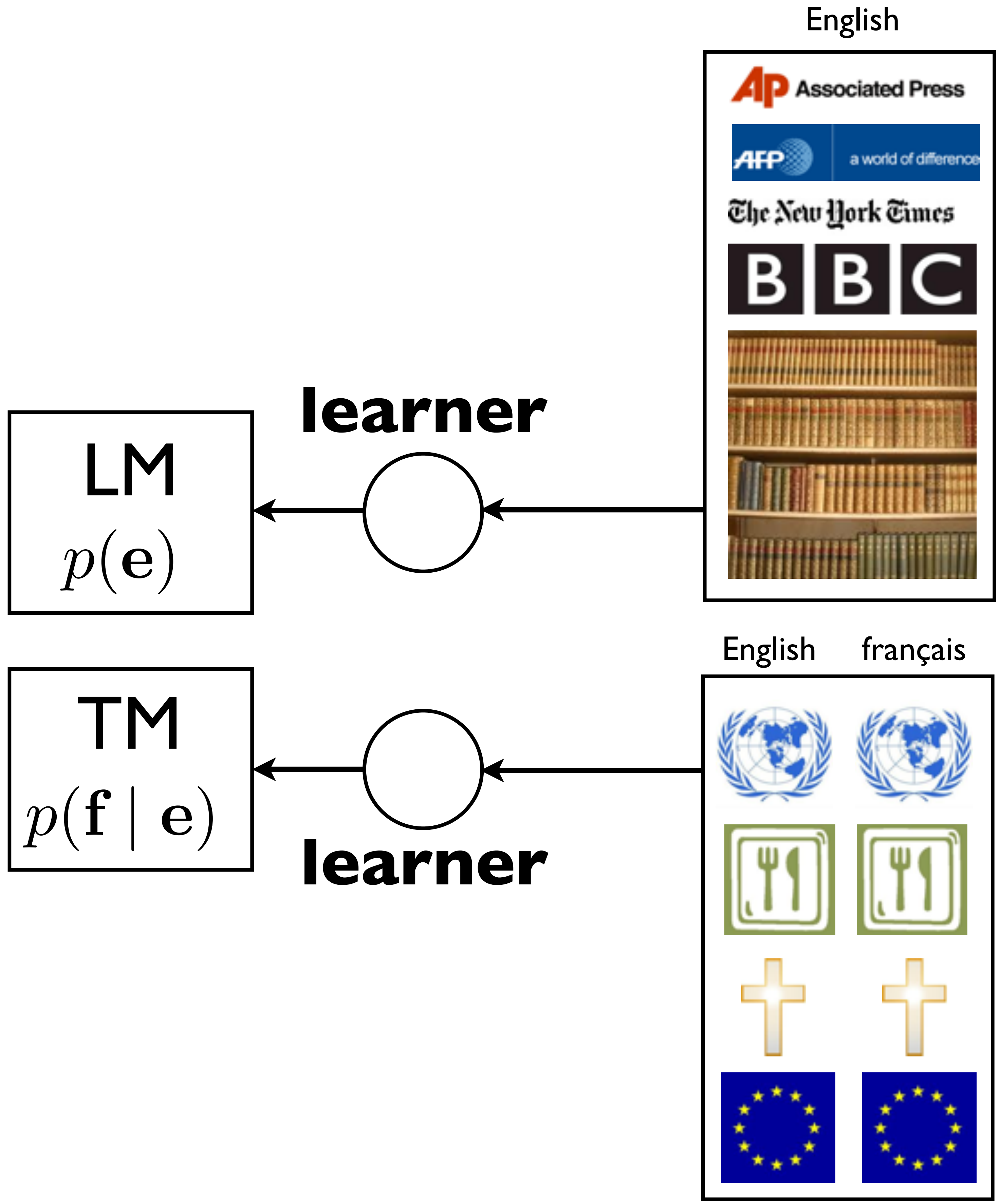

Harry Potter
AND THE
SORCERER'S STONE
J. K. ROWLING

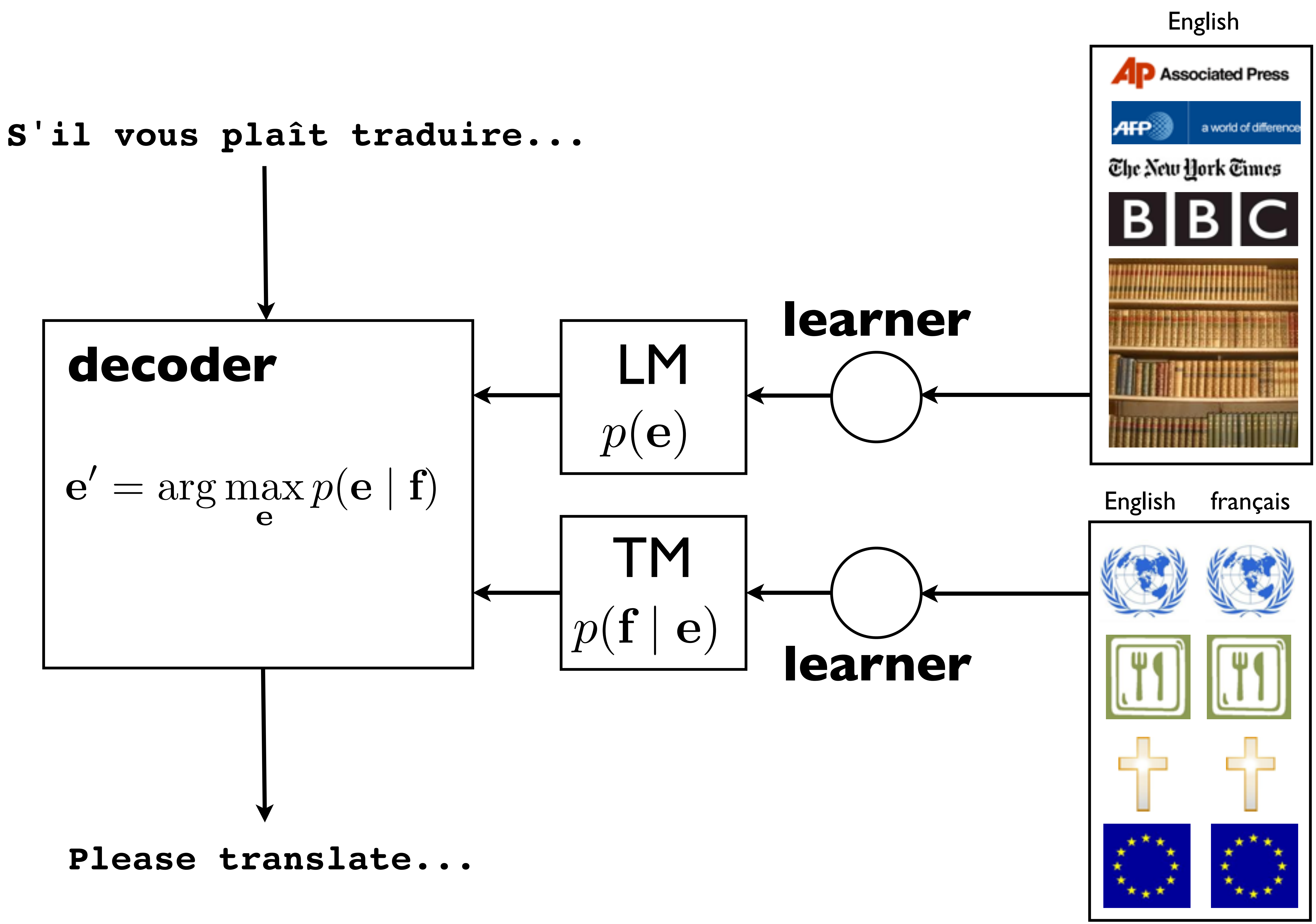

哈利·波特
与
魔法石
〔英〕J.K.罗琳 著
苏农 译
人民文学出版社


Egyptian
Greek

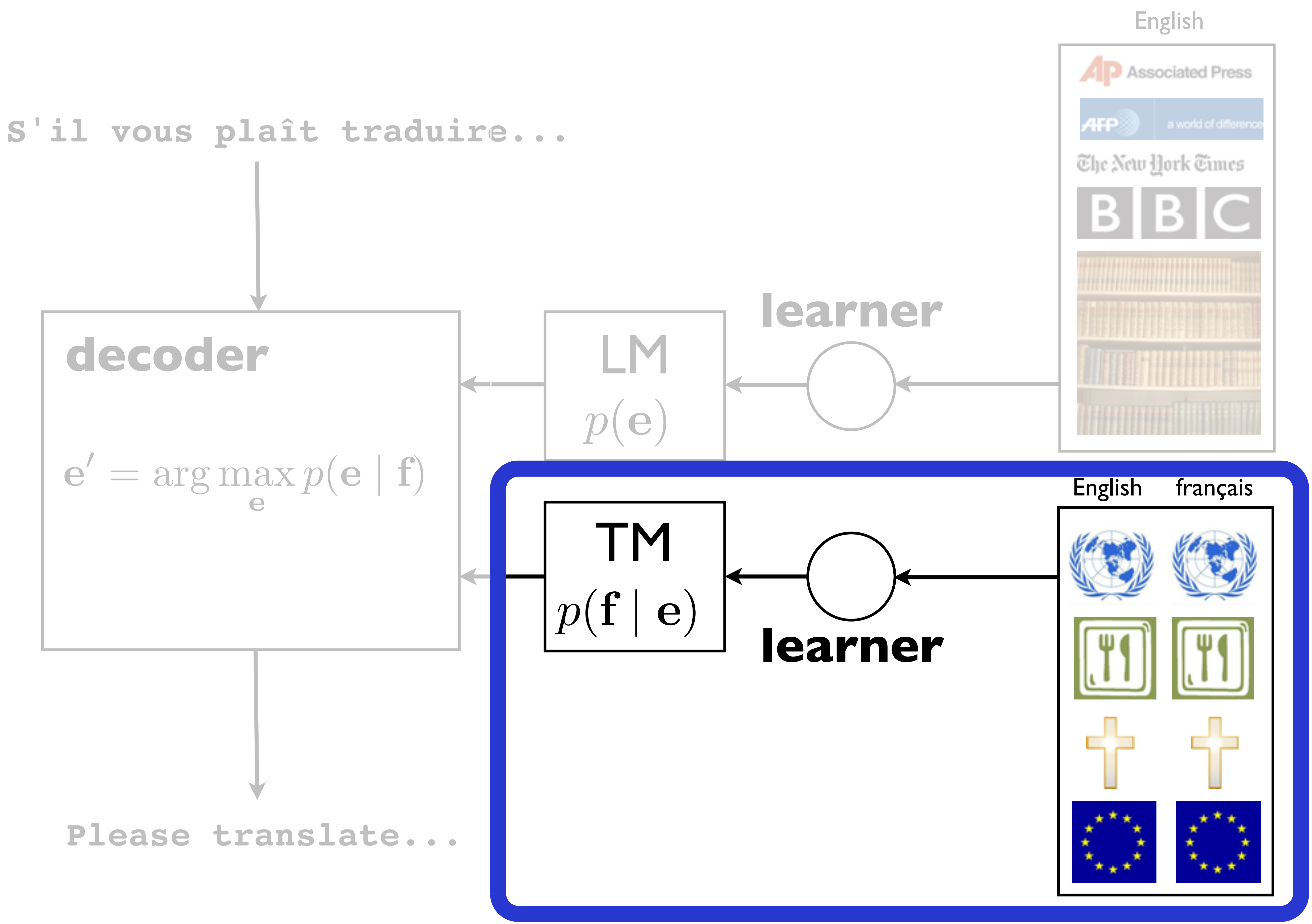
Putting the pieces together


English
Associated Press
AFP
a world of difference
The New York Times
BBC
learner
LM
$p(\mathbf{e})$


English
AP Associated Press
AFP a world of difference
The New York Times
BBC
learner
LM
$p(\mathbf{e})$
TM
$p(\mathbf{f} \mid \mathbf{e})$
learner
English
français



S'il vous plaît traduire...
decoder
$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{e} \mid \mathbf{f})$
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
Please translate...


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

Phrase-based translation

Koehn

Och

Marcu

Translation model

- $p(\mathbf{f}|\mathbf{e})$ - the probability of a foreign sentence **given** an English translation
- $\mathbf{f}$ = *Je voudrais un peu de frommage.*

Translation model

- $p(\mathbf{f}|\mathbf{e})$ - the probability of a foreign sentence **given** an English translation
- $\mathbf{f}$ = *Je voudrais un peu de frommage.*
- $\mathbf{e}_1$ = *I would like some cheese.*
- $\mathbf{e}_2$ = *I would like a little of cheese.*
- $\mathbf{e}_3$ = *There is no train to Barcelona.*

Translation model

- $p(\mathbf{f}|\mathbf{e})$ - the probability of a foreign sentence **given** an English translation
- $\mathbf{f}$ = *Je voudrais un peu de frommage.*
- $\mathbf{e}_1$ = *I would like some cheese.* **~ 0.9**
- $\mathbf{e}_2$ = *I would like a little of cheese.* **~ 1.0**
- $\mathbf{e}_3$ = *There is no train to Barcelona.* **>> 0.00001**

Translation model

- How do we parameterize p(f|e)?

$$p(\mathbf{f} \mid \mathbf{e}) = \frac{count(\mathbf{f}, \mathbf{e})}{count(\mathbf{e})} \quad ?$$

- This will be really sparse! Plus, we know that translations of "similar" sentences are themselves "similar".

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

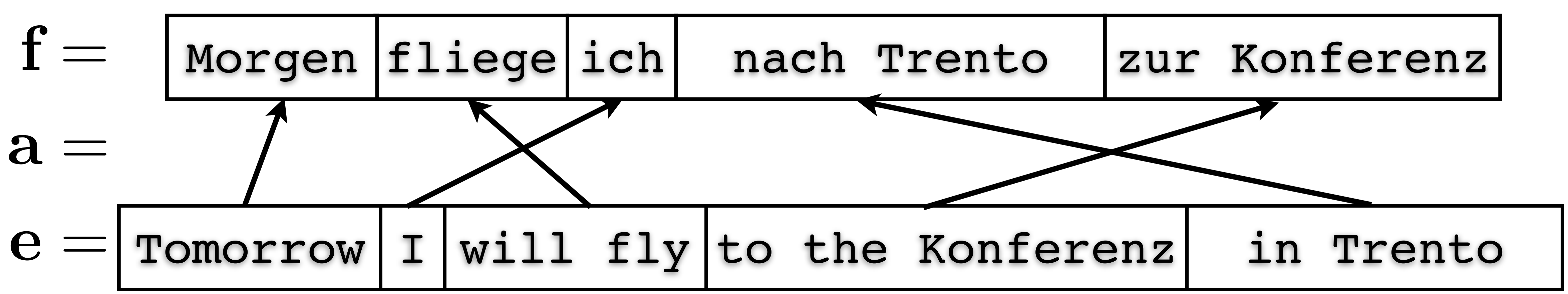
$\mathbf{f} =$ `Morgen fliege ich` `nach Trento` `zur Konferenz`

$\mathbf{e} =$ `Tomorrow I will fly to the Konferenz` `in Trento`

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

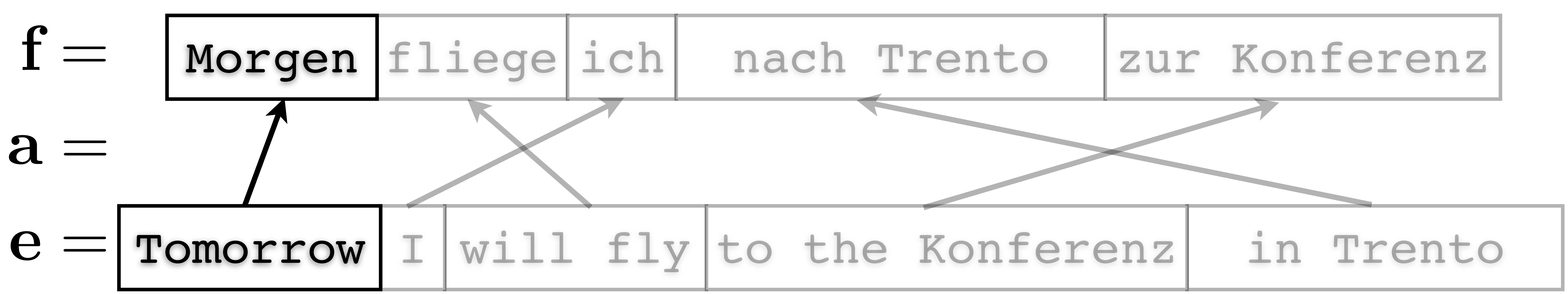
$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$



Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

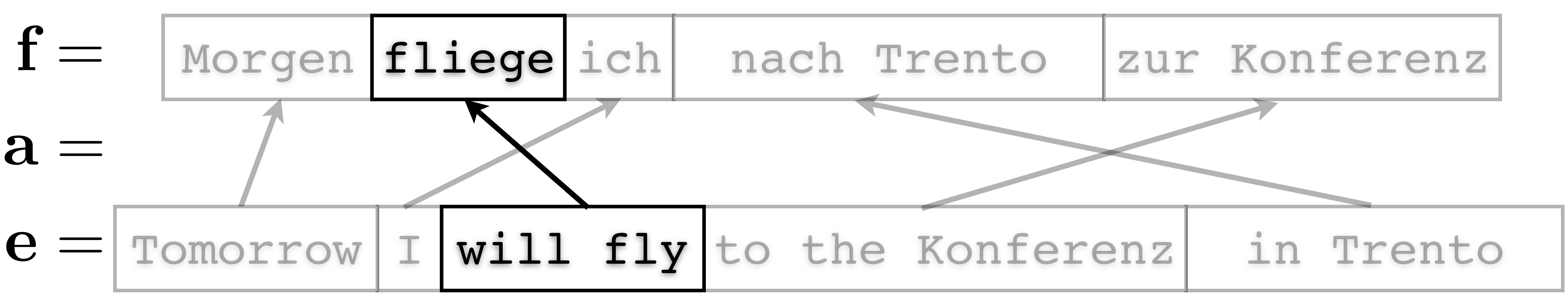


p(Morgen|Tomorrow)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

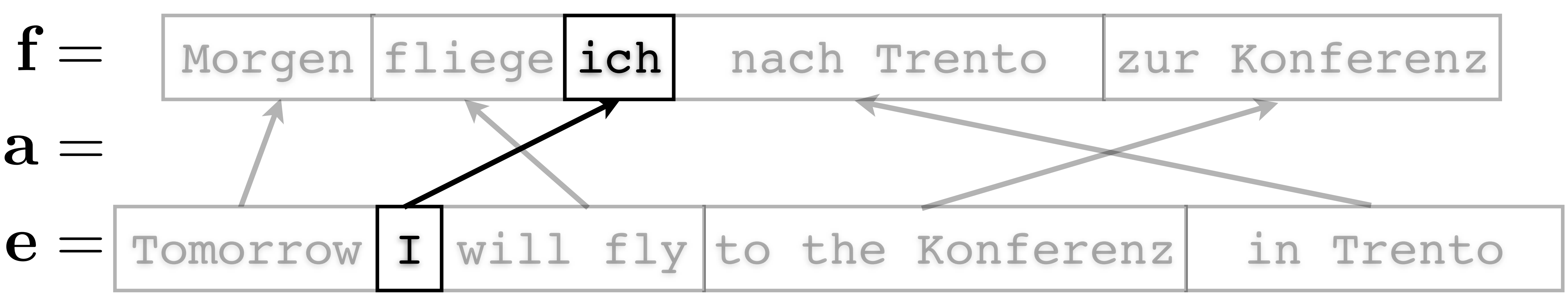


p(Morgen|Tomorrow) x *p*(fliege|will fly)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

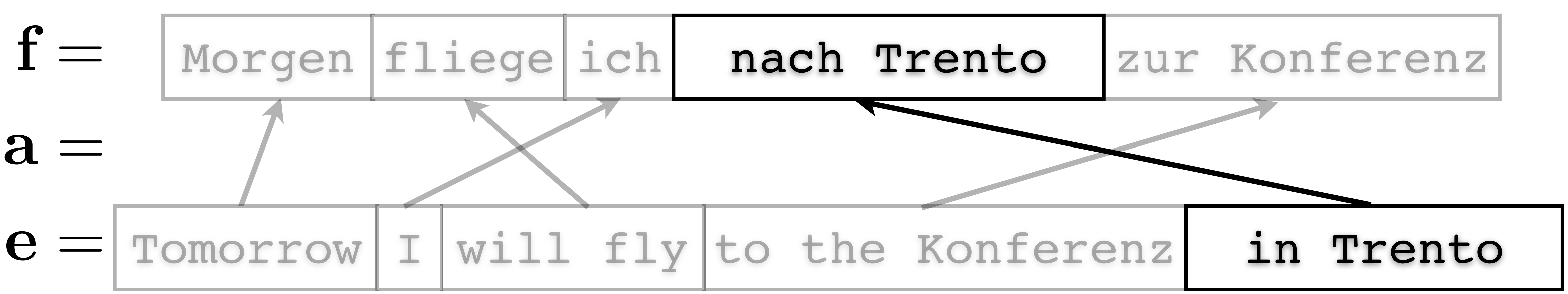


p(Morgen|Tomorrow) x *p*(fliege|will fly) x *p*(ich|I)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$



p(Morgen|Tomorrow) x *p*(fliege|will fly) x *p*(ich|I) x ...

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

Marginalize* to get p(f|e):

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a} \in \mathcal{A}} p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

*Searching in a model with this marginalization turns out to be NP-hard; it's often approximated with a max operator.

Phrases

- Contiguous strings of words (discontiguous variants after the break and in the lab)
- Phrases are not necessarily syntactic constituents
- Extracted from parallel corpora annotated with **word alignments**

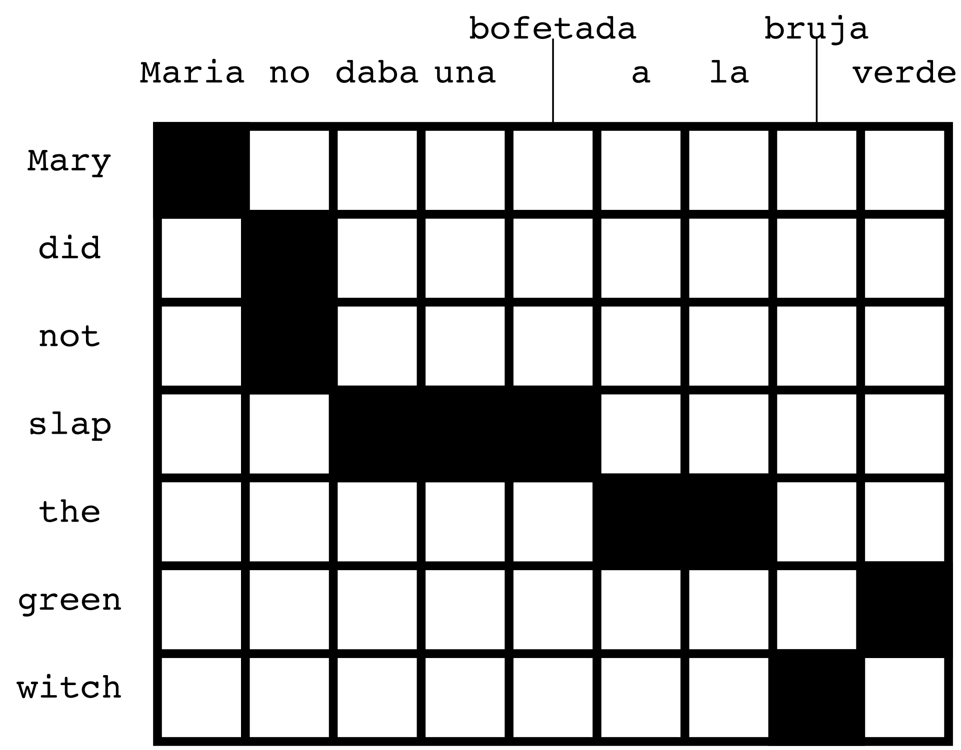
Phrase Tables

$\bar{\mathbf{f}}$	$\bar{\mathbf{e}}$	$p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$
das Thema	the issue	0.41
	the point	0.72
	the subject	0.47
	the thema	0.99
es gibt	there is	0.96
	there are	0.72
morgen	tomorrow	0.9
fliege ich	will I fly	0.63
	will fly	0.17
	I will fly	0.13

*In practice many other features can be included in phrase tables which are used in a different, but related, parameterization.

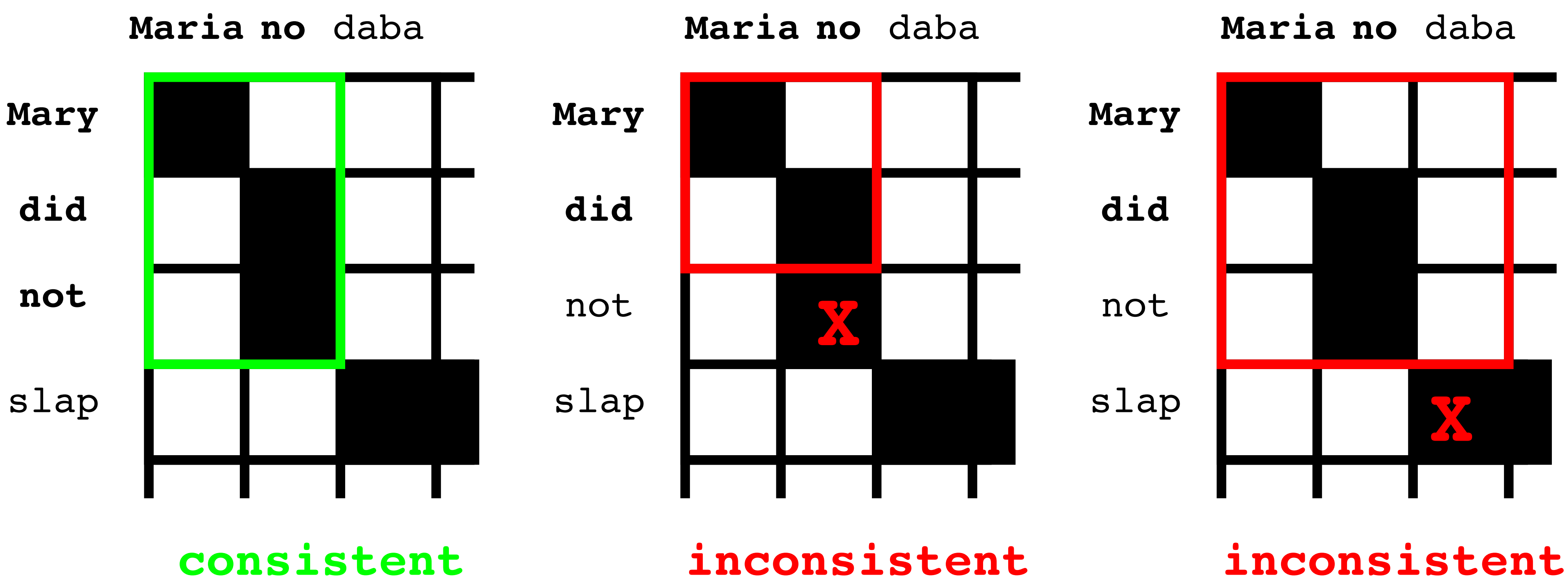
Learning the Phrase Table

- Start with the word alignment:



- Collect all phrase pairs that are consistent with the word alignment

Consistent Phrases



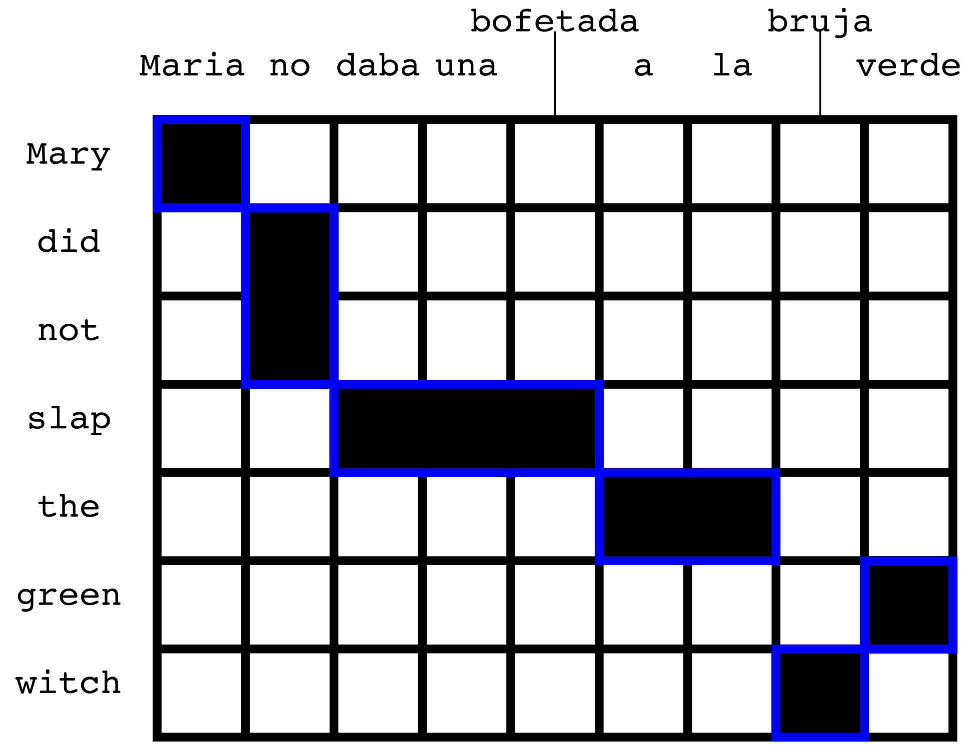
- Consistent with the word alignment :=

  phrase alignment has to contain all alignment points for all covered words

$$(\bar{e}, \bar{f}) \in BP \Leftrightarrow \quad \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f}$$

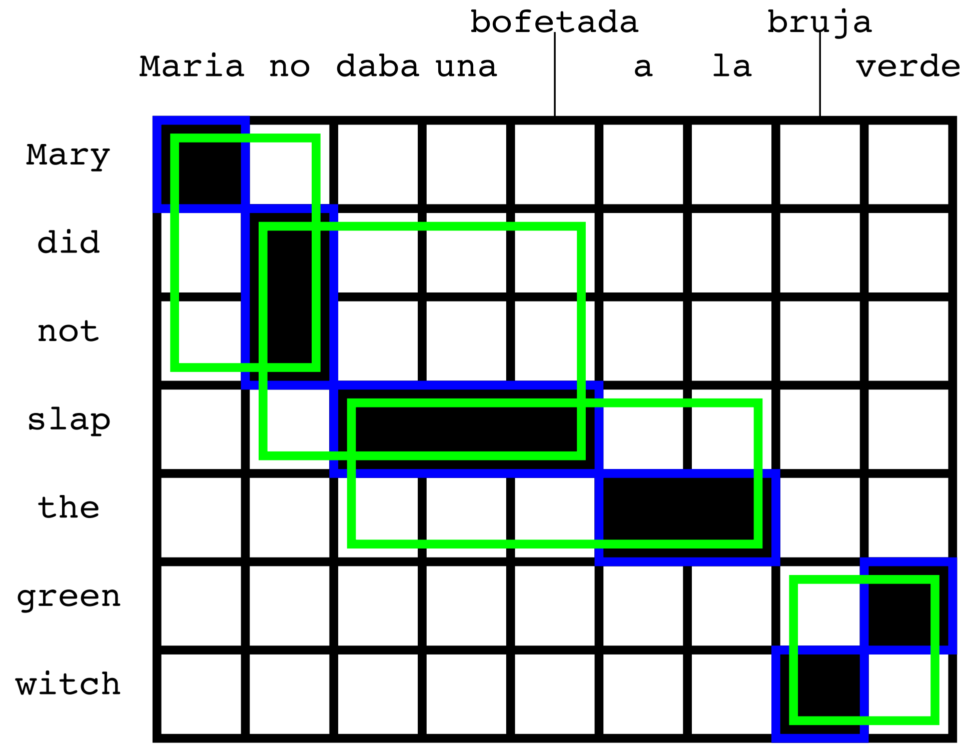
$$\text{AND} \quad \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e}$$

Consistent Phrases



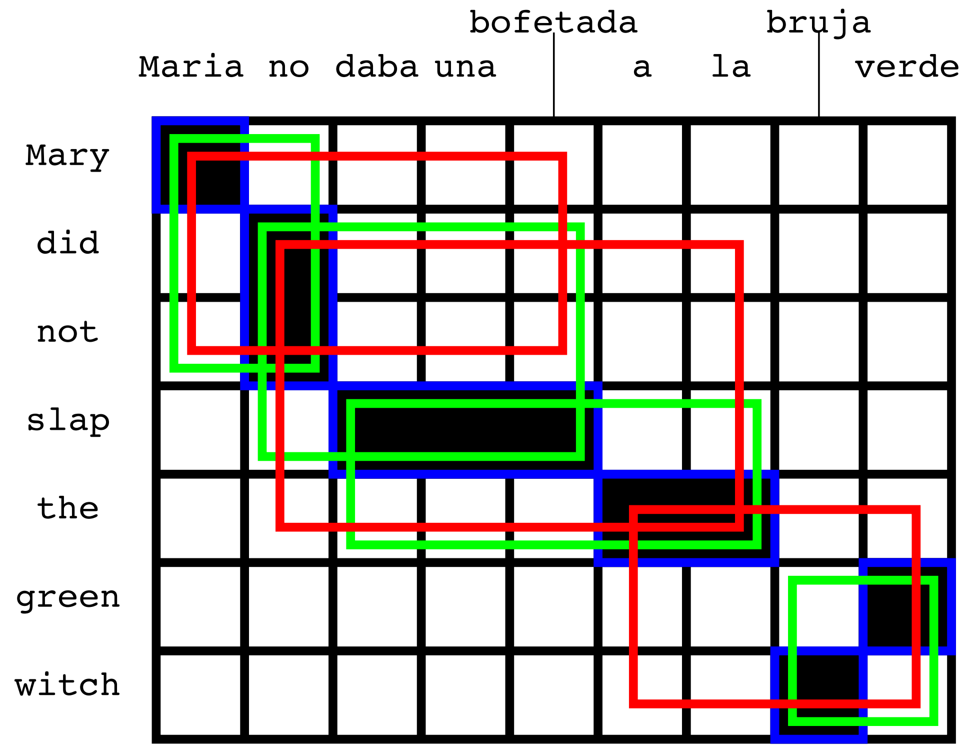
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

Consistent Phrases



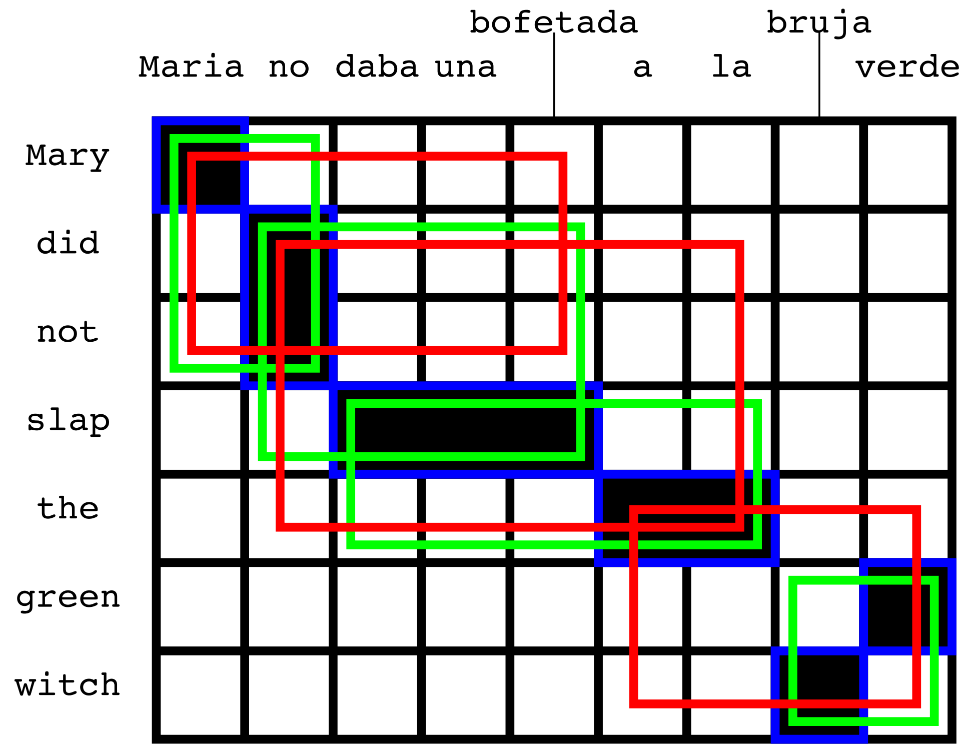
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Consistent Phrases



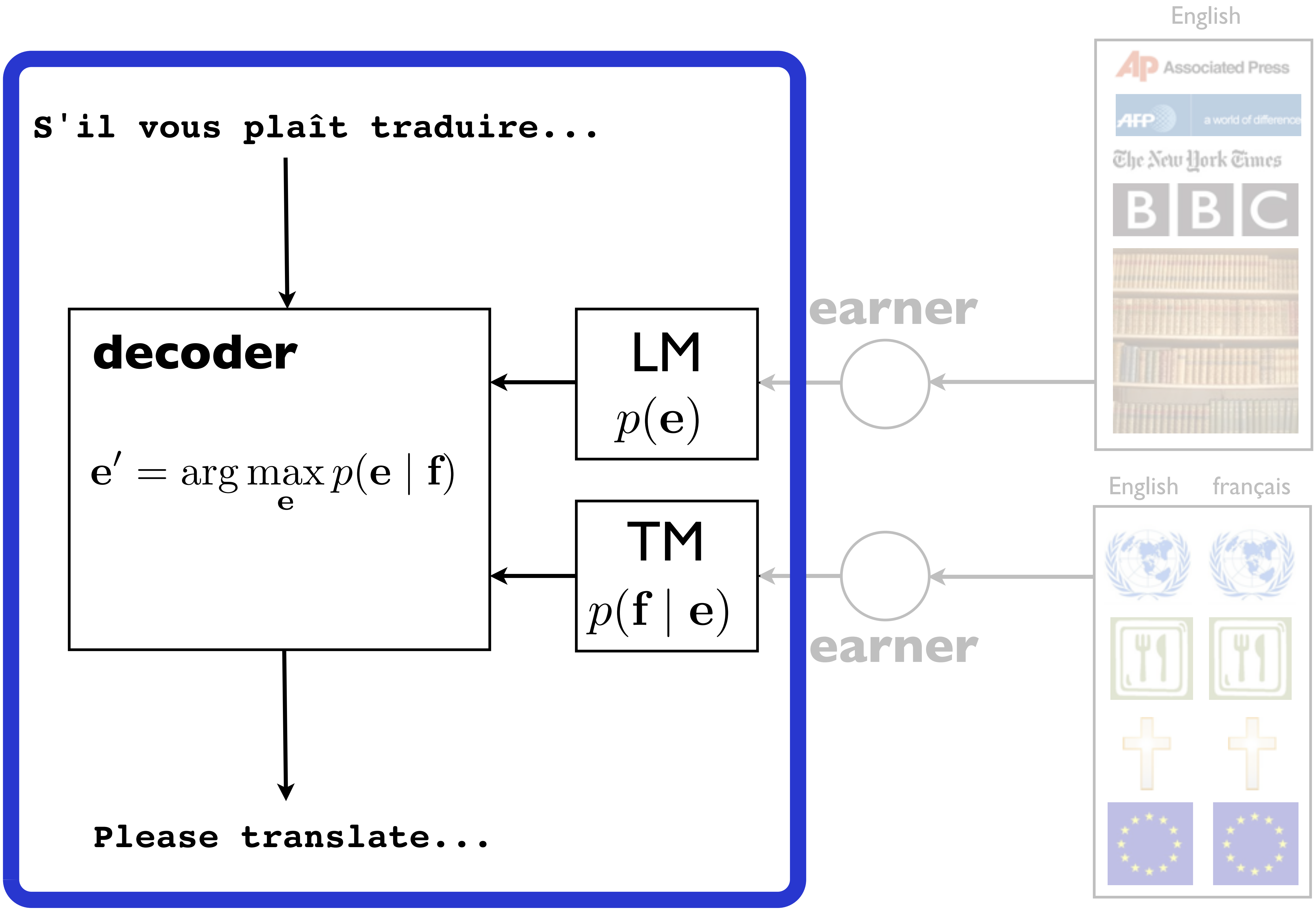
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap), (no daba una bofetada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

Consistent 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)

All that's left is to count and normalize!



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

Decoding

- **Decoding** is the process of searching for the best translation under a translation model (and language model)
- Two problems
  - Find the right words (~ easy)
  - Get them in the right order (~ hard)

Naive phrase-based decoding

- Partial hypothesis keeps track of
  - which source words have been translated (*coverage vector*)
  - *n*-1 most recent words of English (for LM!)
  - a back pointer list to the previous hypothesis + (e,f) phrase pair used
  - the (partial) translation probability
- Extend a partial hypothesis by translating something untranslated
- Start state: no translated words, E=<s>, bp=nil
- Goal state: all translated words

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		by		hag	bawdy
	no	slap			to the		green witch	
	did not give				the			
						the witch		

Adapted from Koehn (2006)

<table>
<tr><th>Maria</th><th>no</th><th>dio</th><th>una</th><th>bofetada</th><th>a</th><th>la</th><th>bruja</th><th>verde</th></tr>
<tr><td>Mary</td><td>not</td><td>give</td><td>a</td><td>slap</td><td>to</td><td>the</td><td>witch</td><td>green</td></tr>
<tr><td></td><td>did not</td><td></td><td colspan="2">a slap</td><td>by</td><td></td><td>hag</td><td>bawdy</td></tr>
<tr><td></td><td>no</td><td colspan="3">slap</td><td colspan="2">to the</td><td colspan="2">green witch</td></tr>
<tr><td></td><td colspan="2">did not give</td><td></td><td></td><td colspan="2">the</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="2">the witch</td><td></td></tr>
</table>

Adapted from Koehn (2006)

<table>
<tr><th>Maria</th><th>no</th><th>dio</th><th>una</th><th>bofetada</th><th>a</th><th>la</th><th>bruja</th><th>verde</th></tr>
<tr><td>Mary</td><td>not</td><td>give</td><td>a</td><td>slap</td><td>to</td><td>the</td><td>witch</td><td>green</td></tr>
<tr><td></td><td>did not</td><td></td><td colspan="2">a slap</td><td>by</td><td></td><td>hag</td><td>bawdy</td></tr>
<tr><td></td><td>no</td><td colspan="3">slap</td><td colspan="2">to the</td><td colspan="2">green witch</td></tr>
<tr><td></td><td colspan="2">did not give</td><td></td><td></td><td colspan="2">the</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="2">the witch</td><td></td></tr>
</table>

Adapted from Koehn (2006)

Reordering

- Language express words in different orders
  - **bruja verde** *vs.* **green witch**
- Phrase pairs can "memorize" some of these
- But we must generalize and consider **reorderings**
- **Problem**
  - If you search all reorderings (and translate each source word exactly one time), you can encode *arbitrary traveling salesperson problems* (TSPs) in your translation model!
- **Solution**
  - Don't search everything. But what? Find out later this week.

Word Alignment

How do we get alignments?

	pervez	musharrafs	langer	abschied
pervez				?
musharraf				
’s		?		
long			?	
goodbye	?			

Word alignment with EM

- EM lets you estimate parameters of probability distributions when not all random variables in the model are observable
- For alignment:
  - parallel data is observed
  - alignment is unobserved

Lexical Translation

- Every target word aligns to a single source word (or "null")
- Given this alignment, translations are conditionally independent of each other
- The pros:
  - Parameters are just the probabilities that words in the source translate into words in the target
  - If two words don't co-occur in the training data their probability is zero
- The cons:
  - 1-1 and 1-many alignments possible, but not many-1 or many-many
  - Do we really believe this independence assumption is reasonable?

Learning Model 1

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \approx p(\mathbf{a}) \prod_{j=1}^{|\mathbf{f}|} t(f_j \mid e_{a_j})$$

$$p(\mathbf{a}) = \mathrm{Uniform}(|\mathbf{e}|)^{|\mathbf{f}|}$$

The only parameters are the lexical translation probabilities:

$$t(F \mid E)$$

If we **had** word alignments, we could count and normalize

$$t(f|e) = \frac{count(\mathrm{e,f})}{count(\mathrm{e})}$$

Number of times e is aligned to f

Number of times e is occurs

EM to the rescue!

- Start with some random translation probabilities
- Repeat until convergence
  - Infer the best alignments under the current model [E step]

$$a_j^* = \arg\max_{i=0}^{|\mathbf{e}|} t(f_j|e_i)$$

  - Assume these inferred alignments are correct
  - Count and normalize! [M step]
- Details tomorrow!

Translation Evaluation

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

- Yorrick Wilks

The gold standard?

- Human evaluation
  - Have annotators read and assess translations
    - (Fluency, adequacy)
  - Have annotators read translations and do something

Is the cake delicious?

Human evaluation

- Problems
  - Humans don't like to evaluate translation, especially bad translation
  - Humans don't tend to agree with each other

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

- Evaluating translation automatically is hard
  - Many correct ways to say something
  - If we could measure if a sentence was grammatical, we would have solved the language modeling problem!
  - Same goes for if a translation is correct.

Questions?