CS447: Natural Language Processing

http://courses.engr.illinois.edu/cs447

Lecture 16: Machine Translation

Julia Hockenmaier

juliahmr@illinois.edu

3324 Siebel Center

An exercise: Centauri and Arcutran

Kevin Knight. AI Magazine Volume 18 Number 4 (1997)

1a. ok-voon ororok sprok.
1b. at-voon bichat dat.

2a. ok-drubel ok-voon anok plok sprok.
2b. at-drubel at-voon pippat rrat dat.

3a. erok sprok izok hihok ghirok.
3b. totat dat arrat vat hilat.

4a. ok-voon anok drok brok jok.
4b. at-voon krat pippat sat lat.

5a. wiwok farok izok stok.
5b. totat jjat quat cat.

6a. lalok sprok izok jok stok.
6b. wat dat krat quat cat.

7a. lalok farok ororok lalok sprok izok enemok
7b. wat jjat bichat wat dat vat eneat.

8a. lalok brok anok plok nok.
8b. iat lat pippat rrat nnat.

9a. wiwok nok izok kantok ok-yurp.
9b. totat nnat quat oloat at-yurp.

10a. lalok mok nok yorok ghirok clok.
10b. wat nnat gat mat bat hilat.

11a. lalok nok crrrok hihok yorok zanzanok.
11b. wat nnat arrat mat zanzanat.

12a. lalok rarok nok izok hihok mok.
12b. wat nnat forat arrat vat gat.

1a. ok-**voon** ororok **sprok**.
1b. at-**voon** bichat **dat**.

2a. ok-drubel ok-**voon** anok plok **sprok**.
2b. at-drubel at-**voon** pippat rrat **dat**.

3a. **erok sprok izok hihok** ghirok.
3b. **totat dat arrat vat** hilat.

4a. ok-**voon** anok drok brok **jok**.
4b. at-**voon** **krat** pippat sat lat.

5a. **wiwok** farok **izok** stok.
5b. **totat** jjat **quat** cat.

6a. **lalok sprok izok jok** stok.
6b. **wat dat krat quat** cat.

7a. **lalok** farok ororok **lalok sprok izok** enemok
7b. **wat** jjat bichat **wat dat vat** eneat.

8a. **lalok** brok anok plok **nok**.
8b. **iat** lat pippat rrat **nnat**.

9a. **wiwok nok izok** kantok ok-yurp.
9b. **totat nnat quat** oloat at-yurp.

10a. **lalok** mok **nok** yorok ghirok clok.
10b. **wat nnat** gat mat bat hilat.

11a. **lalok nok crrrok hihok** yorok zanzanok.
11b. **wat nnat arrat** mat zanzanat.

12a. **lalok** rarok **nok izok hihok** mok.
12b. **wat nnat** forat **arrat vat** gat.

1a. ok-**voon** ororok **sprok**.
1b. at-**voon** bichat **dat**.

2a. ok-drubel ok-**voon** anok plok **sprok**.
2b. at-drubel at-**voon** pippat rrat **dat**.

3a. **erok sprok izok hihok** ghirok.
3b. **totat dat arrat vat** hilat.

4a. ok-**voon** anok drok brok **jok**.
4b. at-**voon krat** pippat sat lat.

5a. **wiwok** farok **izok** stok.
5b. **totat** jjat **quat** cat.

6a. **lalok sprok izok jok** stok.
6b. **wat dat krat quat** cat.

7a. **lalok** farok ororok **lalok sprok izok** enemok
7b. **wat** jjat bichat **wat dat vat** eneat.

8a. **lalok** brok anok plok **nok**.
8b. **iat** lat pippat rrat **nnat**.

9a. **wiwok nok izok** kantok ok-yurp.
9b. **totat nnat quat** oloat at-yurp.

10a. **lalok** mok **nok** yorok ghirok clok.
10b. **wat nnat** gat mat bat hilat.

11a. **lalok nok crrrok hihok** yorok zanzanok.
11b. **wat nnat arrat** mat zanzanat.

12a. **lalok** rarok **nok izok hihok** mok.
12b. **wat nnat** forat **arrat vat** gat.

1a. **Garcia** and **associates**.
1b. **Garcia** y **asociados**.

2a. Carlos **Garcia** has three **associates**.
2b. Carlos **Garcia** tiene tres **asociados**.

3a. **his associates are not** strong.
3b. **sus asociados no son** fuertes.

4a. **Garcia** has a company **also**.
4b. **Garcia tambien** tiene una empresa.

5a. **its** clients **are** angry.
5b. **sus** clientes **están** enfadados.

6a. **the associates are also** angry.
6b. **los asociados tambien están** enfadados.

7a. **the** clients and **the associates are** enemies.
7b. **los** clientes y **los asociados son** enemigos.

8a. **the** company has three **groups**.
8b. **la** empresa tiene tres **grupos**.

9a. **its groups are** in Europe.
9b. **sus grupos están** en Europa.

10a. **the** modern **groups** sell strong pharmaceuticals
10b. **los grupos** modernos venden medicinas fuertes

11a. **the groups do not** sell zanzanine.
11b. **los grupos no** venden zanzanina.

12a. **the** small **groups are not** modern.
12b. **los grupos** pequeños **no son** modernos.

Machine translation approaches

Today's key concepts

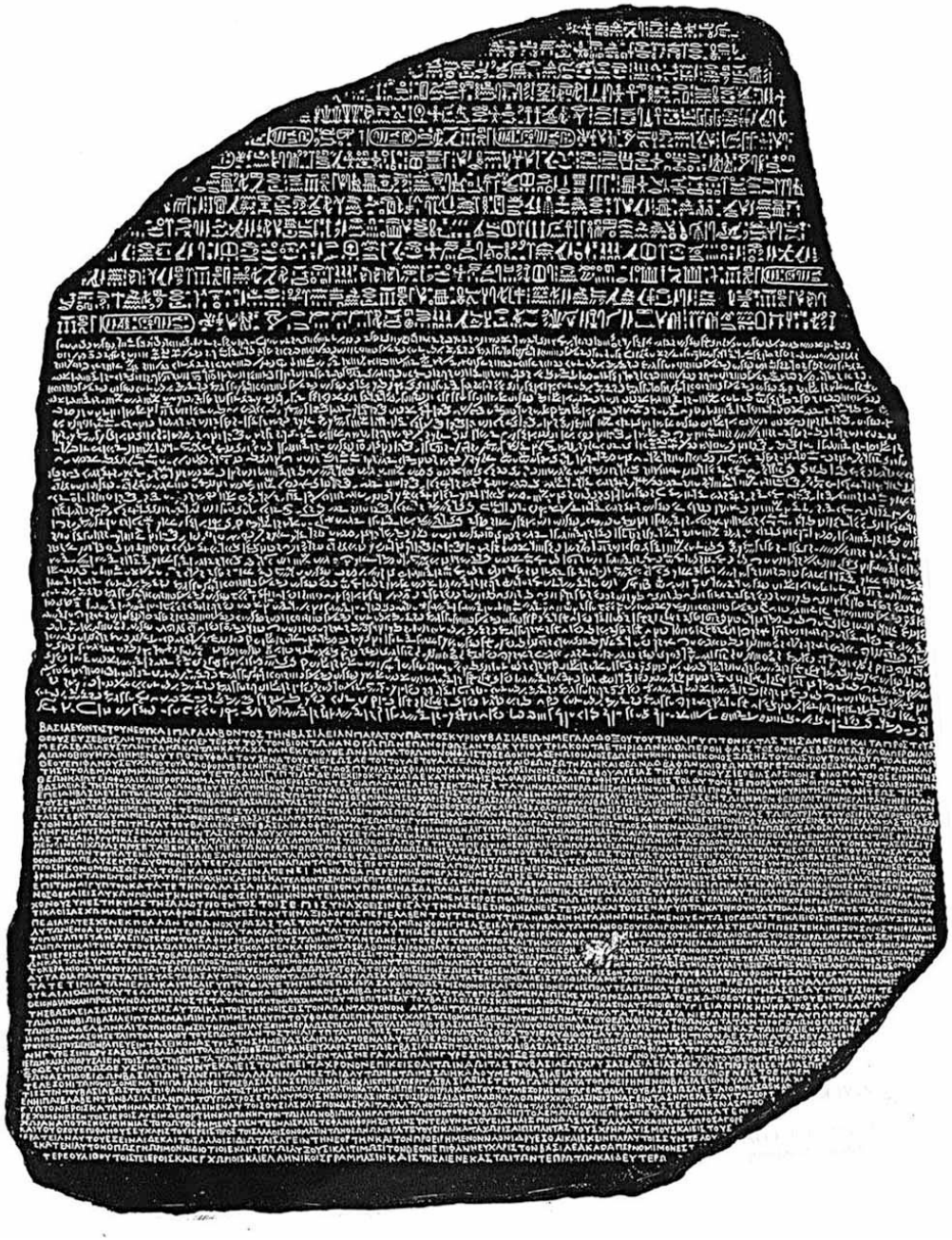
Why is machine translation hard?

- Linguistic divergences: morphology, syntax, semantics

Different approaches to machine translation:

- Vauquois triangle
- Statistical MT
- Neural MT

Evaluation: BLEU score



The Rosetta Stone

The same text in three languages:

- Hieroglyphic Egyptian (used by priests)
- Demotic Egyptian (used for daily purposes)
- Classical Greek (used by the administration)

Instrumental in our understanding of ancient Egyptian

This is an instance of **parallel text**

The Greek inscription allowed scholars to decipher the hieroglyphs

Machine Translation History

WW II: Code-breaking efforts at Bletchley Park, England (Alan Turing)

1948: Shannon/Weaver: Information theory

1949: Weaver's memorandum defines the machine translation task

1954: IBM/Georgetown demo: 60 sentences Russian-English

1960: Bar-Hillel: MT to difficult

1966: ALPAC report: human translation is far cheaper and better:
kills MT for a long time

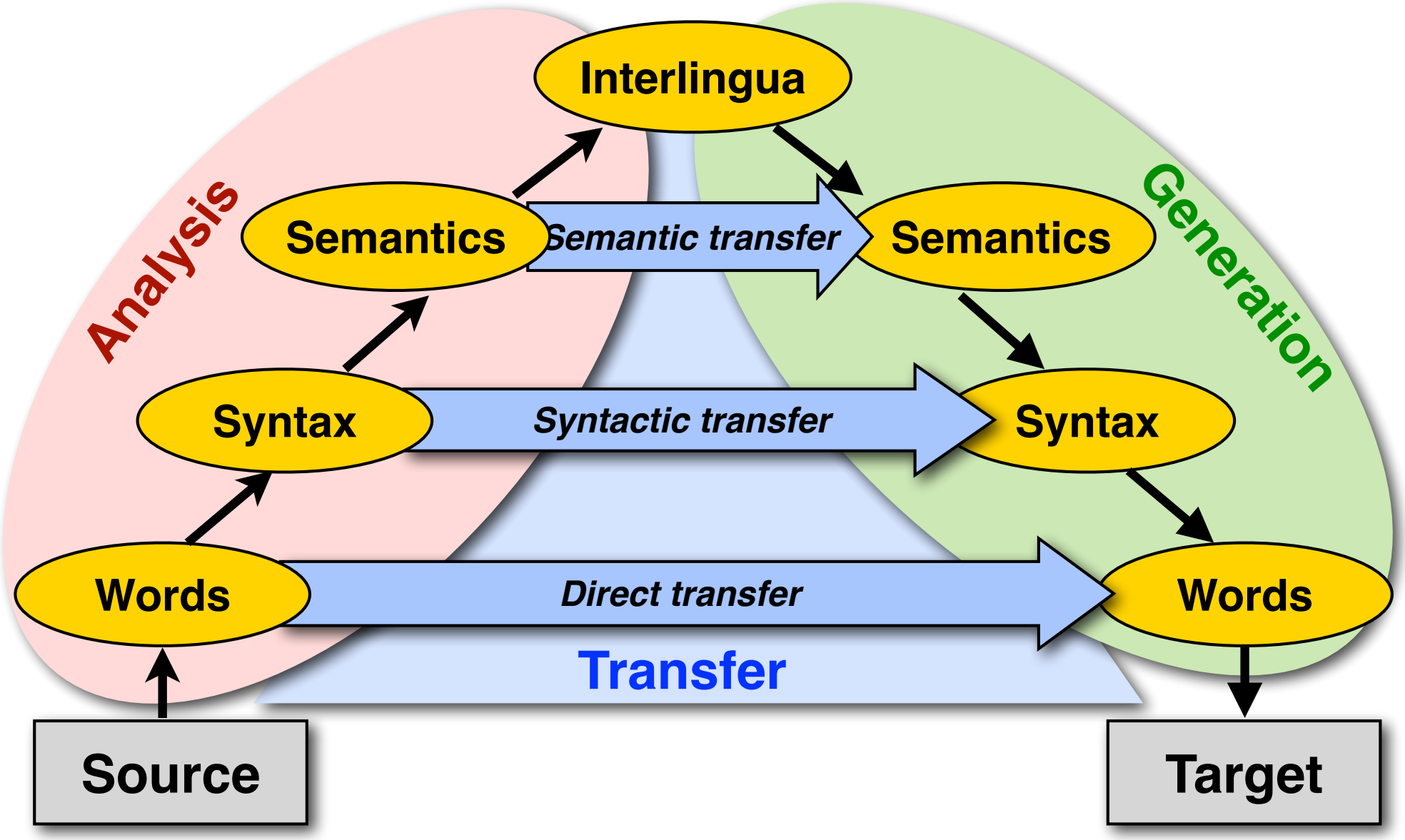
1980s/90s: Transfer/interlingua-based approaches (**Vauquois** triangle)

1990: **IBM's** CANDIDE system (first modern **statistical MT** system)

2000s: Huge interest and progress in wide-coverage statistical MT:
phrase-based MT, syntax-based MT, open-source tools

since mid/late 2010's: Neural machine translation
(**seq2seq models with attention**)

The Vauquois triangle



Machine Translation in 2012

新华新闻 地方联播 > 正文

家庭生活日记账：计价精确到“8分8厘3”

2012年11月09日 20:57:04
来源： 北京晚报

【字号：大 中 小】【打印】 【纠错】

“婴儿纸尿裤，384元；手机贴膜，10元；报纸，1元……”十年来的每一天，在北京这座城市，都有5000个收入不同、构成各异的家庭在细心填写着统一格式的生活账本——《城镇居民家庭生活情况日记账》。

这一行行的“针头线脑”真实记录着他们的生活轨迹，也勾勒着“城镇居民可支配收入”、“城镇住户调查收入情况”等事关国计民生的统计数据。

而长年和这些原生态数据打交道的基层调查员们，总能亲身感知到生活在这座城市中的人们赚钱、花钱的那些事……

我们每天的生活如何变成数字，数字又如何影响我们的生活？

北京有5000个记账户家庭，他们的记录将成为政府了解居民收入、生活、物价等多方面信息的渠道，为制订社会发展计划和进行科学决策提供重要依据，包括最低生活保障线、最低工资标准等等。这些家庭由统计调查单位遵循随机抽样的原则选取，记账家庭三年整体轮换一次。

他们填报的数据，经过系统的整理、汇总和分析后，每个月都会形成《城镇居民人均可支配收入》等多份数据报告，公众可以进入北京统计信息网（www.bjstats.gov.cn）查询。

Google Translate
translate.google.com

Xinhua News Local network >

Family life journals: pricing is accurate to "8%"

November 9, 2012 20:57:04
Source: Beijing Evening News

Views: big in small 】【 Print [Error correction] 】

Baby diapers, 384 yuan; mobile phone film, 10; newspapers, ... "decades every day in the city of Beijing has 5000 income, constitute a diverse family carefully fill unified format life books - "family life of urban residents journal.

This the the trekking of "Zhentouxiannao" a true record of the trajectory of their lives, and also outlines the disposable income of urban residents, urban household survey income "statistics related to the national economy and people's livelihood.

Many years dealing with and original ecological data, primary investigator have always personally perception to the people living in the city to make money, spend money on those things ...

How our daily lives become digital, digital, how to affect our lives?

Beijing 5000 Hutchison account family, their records will be the government understand residents' income, life, price and other information channels, and provide an important basis for the formulation of social development plans and scientific decision-making, including the minimum living security line, minimum wage and so on. These families to follow the principle of random sampling survey units selected family bookkeeping three overall rotation once.

Data they reported, after finishing, summary and analysis of the system, every month the formation of urban residents per capita disposable income "data reporting, the public can enter the Beijing Statistical Information Net (www.bjstats.gov.cn) query.

Machine Translation in 2018

习近平在上海考察

2018-11-07 19:41:21　来源：新华网

习近平在上海考察时强调

坚定改革开放再出发信心和决心

加快提升城市能级和核心竞争力

新华社上海11月7日电　中共中央总书记、国家主席、中央军委主席习近平近日在上海考察时强调，坚持以新时代中国特色社会主义思想为指导，坚决贯彻落实党中央决策部署，坚定改革开放再出发的信心和决心，坚持稳中求进工作总基调，全面贯彻新发展理念，坚持以供给侧结构性改革为主线，加快建设现代化经济体系，打好三大攻坚战，加快提升城市能级和核心竞争力，更好为全国改革发展大局服务。

Xi Jinping inspected in Shanghai

Xi Jinping stressed during his visit to Shanghai

Strengthening reform and opening up and starting to build confidence and determination

Accelerate the improvement of urban energy level and core competitiveness

Xinhua News Agency, Shanghai, November 7th, Xi Jinping, general secretary of the CPC Central Committee, president of the State Council and chairman of the Central Military Commission, stressed during his recent visit to Shanghai that he should adhere to the guidance of socialism with Chinese characteristics in the new era, resolutely implement the decision-making and deployment of the Party Central Committee, and strengthen reform and opening up. Confidence and

Machine translation in 2019

(http://www.xinhuanet.com/2019-10/16/c_1125113117.htm)

10月16日，国家主席习近平在北京人民大会堂会见新西兰前总理约翰·基。 新华社记者 庞兴雷 摄

习近平指出，当前国际形势正在经历深刻复杂变化。新形势下，中国对外合作的意愿不是减弱了，而是更加强了。中国坚持和平发展，中国开放的大门必将越开越大。欢迎世界各国包括各国企业抓住中国发展机遇，更好实现互利共赢。习近平表示，约翰·基先生担任总理期间，为推动中新关系发展作出积极贡献，希望你继续为增进两国人民友好合作添砖加瓦。

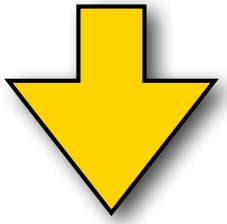
On October 16, President Xi Jinping met with former New Zealand Prime Minister John Key at the Great Hall of the People in Beijing. Xinhua News Agency reporter Pang Xinglei photo

Xi Jinping pointed out that the current international situation is undergoing profound and complex changes. Under the new situation, China's willingness to cooperate with foreign countries has not weakened, but has been strengthened. China adheres to peaceful development, and the door to China's opening is bound to grow. We welcome all countries in the world, including national enterprises, to seize the opportunities of China's development and better achieve mutual benefit and win-win results. Xi Jinping said that during his tenure as Prime Minister, Mr. John Kee made positive contributions to promoting the development of China-Singapore relations. I hope that you will continue to contribute to the friendship and cooperation between the two peoples.

Machine translation in 2019

"Noch immer ist Notre-Dame gefährdet"

Am Morgen des 16. April schauten die Pariser schweigend und übernächtigt auf rußgeschwärzte Steine, auf eine Kathedrale, die kein Dach mehr hatte. Der markante Spitzturm des Architekten Eugène Viollet-Le-Duc fehlte. Krachend war er am Abend zuvor um kurz vor 20 Uhr unter den entsetzten Schreien der Umstehenden in die Tiefe gestürzt.



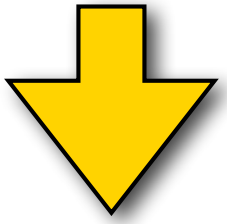
"Still is Notre-Dame at risk"

On the morning of April 16, the Parisians looked in silence and blackened on soot-blackened stones, on a cathedral, which had no roof. The striking pinnacle of the architect Eugène Viollet-Le-Duc was missing. He had crashed the night before at just before 20 clock under the horrified screams of those around in the depths.

Machine translation in 2023

"Noch immer ist Notre-Dame gefährdet"

Am Morgen des 16. April schauten die Pariser schweigend und übernächtigt auf rußgeschwärzte Steine, auf eine Kathedrale, die kein Dach mehr hatte. Der markante Spitzturm des Architekten Eugène Viollet-Le-Duc fehlte. Krachend war er am Abend zuvor um kurz vor 20 Uhr unter den entsetzten Schreien der Umstehenden in die Tiefe gestürzt.



"Notre-Dame is still at risk"

On the morning of April 16, the Parisians looked on, silent and weary, at the soot-blackened stones, at a cathedral that no longer had a roof. The distinctive pointed tower designed by the architect Eugène Viollet-Le-Duc was missing. He fell with a crash just before 8 p.m. the night before, to the horrified screams of those around him.


Why is MT difficult?

Examples of **lexical** divergences

Lexical specificity

German ***Kürbis*** = English ***pumpkin*** or ***(winter) squash***

English ***brother*** = Chinese ***gege*** (older) or ***didi*** (younger)

Morphological divergences

English: ***new*** *book(s),* ***new*** *story/stories*

French: un **nouveau** *livre (sg.m), une* **nouvelle** *histoire (sg.f),*
des **nouveaux** *livres (pl.m), des* **nouvelles** *histoires (pl.f)*

- How much **inflection** does a language have?
  (cf. Chinese vs.Finnish)
- How many **morphemes** does each word have?
- How easily can the morphemes be **separated**?

Examples of **lexical** divergences

The different senses of **homonymous words** generally have different translations:

English-German: (river) bank - Ufer
(financial) bank - Bank

The different senses of **polysemous words** may also have different translations:

I **know that** he bought the book: Je **sais qu**'il a acheté le livre.

I **know** Peter: Je **connais** Peter.

I **know** math: Je **m'y connais en** maths**.**

Examples of **syntactic** divergences

Word order: fixed or free?

If fixed, which one? [SVO (Sbj-Verb-Obj), SOV, VSO,…]

Head-marking vs. dependent-marking

Dependent-marking (English) *the man's house*

Head-marking (Hungarian) *the man house-**his***

Pro-drop languages can omit pronouns:

Italian (with inflection): *I eat = mangi**o**; he eats = mangi**a***

Chinese (without inflection): *I/he eat: chīfàn*

Negation

English *I didn't drink any coffee*

German: *Ich habe keinen Kaffee getrunken* *("I have no coffee drank")*

Examples of **semantic** divergences

Aspect:

- English has a progressive aspect:
  'Peter swims' vs. 'Peter is swimming'
- German can only express this with an adverb:
  'Peter schwimmt' vs. 'Peter schwimmt gerade' ('swims currently')

Motion events have two properties:

- manner of motion (*swimming*)
- direction of motion (*across* the lake)

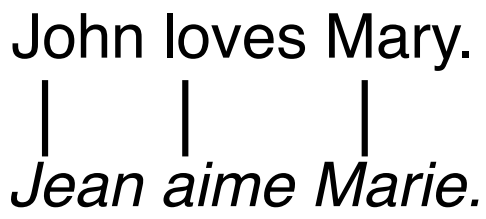
Languages express either the manner with a verb
and the direction with a 'satellite' or vice versa (L. Talmy):

English (satellite-framed): *He* [*swam*]$_{\text{MANNER}}$ **[*across*]$_{\text{DIR}}$** *the lake*

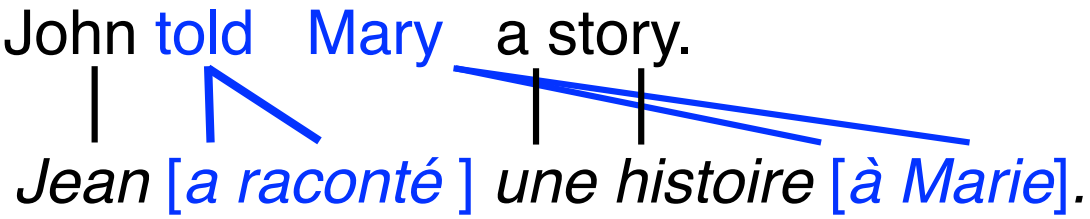
French (verb-framed): *Il a* **[*traversé*]$_{\text{DIR}}$** *le lac* **[*à la nage*]**$_{\text{MANNER}}$

Word-to-Word Correspondences

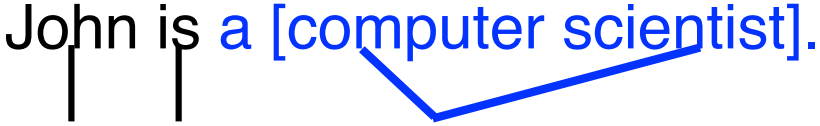
One to-one:



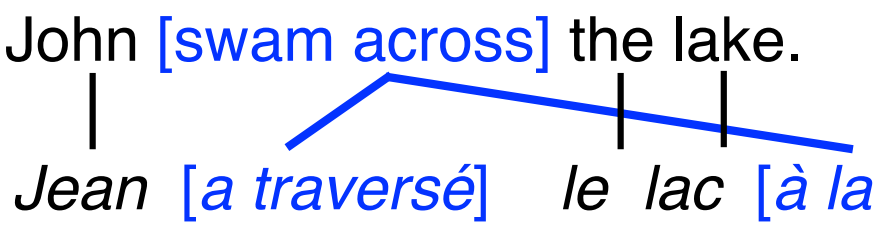
One-to-many:
(and reordering)

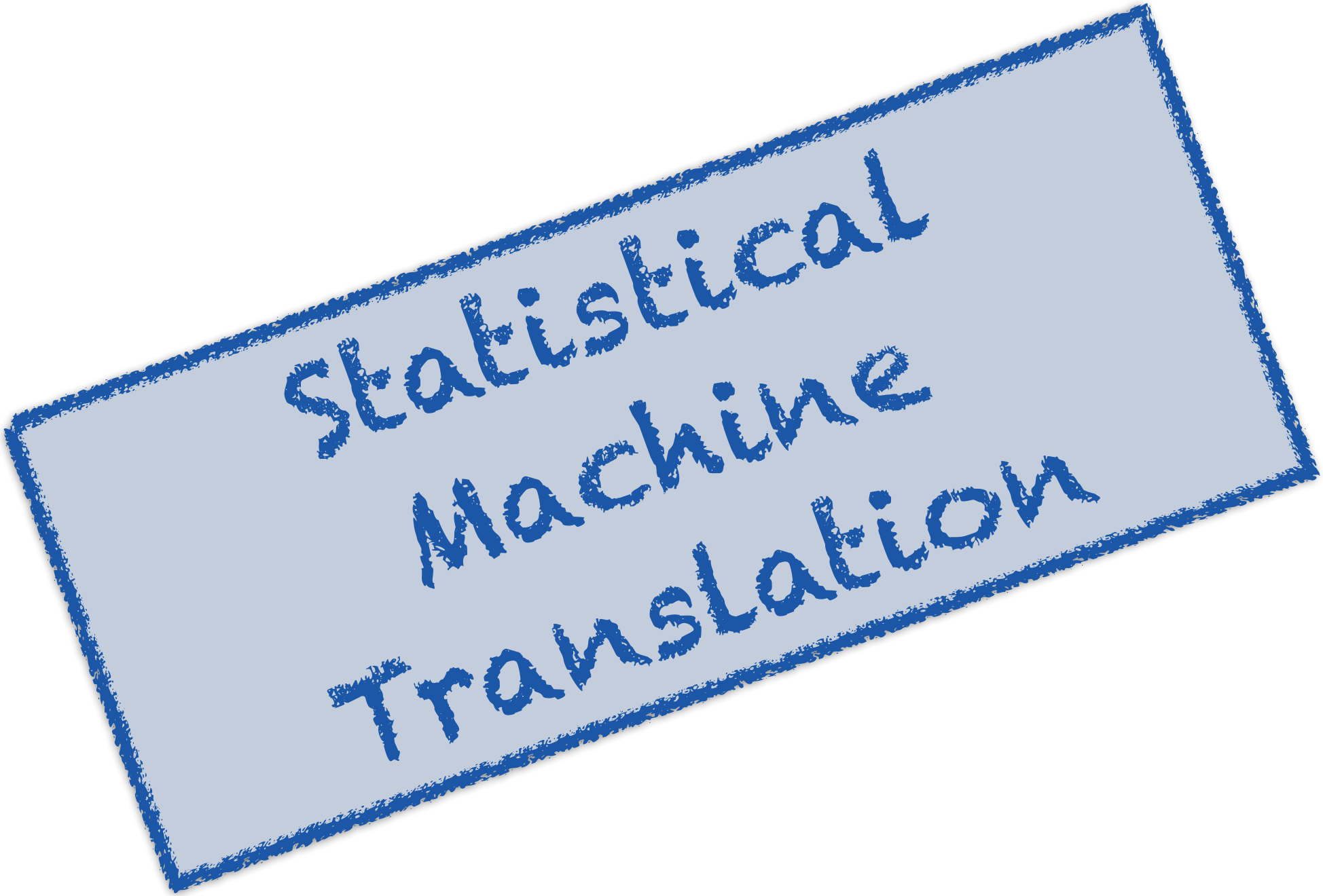


Many-to-one:
(and elision)



Many-to-many:




Statistical
Machine
Translation

Statistical Machine Translation

Given input in the source language, *S*,…

e.g. a Chinese sentence…

主席：各位議員，早晨。

… return the best translation in the target language, *T

e.g in English:

President: Good morning, Honourable Members.

We can formalize this as $T^* = \text{argmax}_T P(T \mid S)$

The noisy channel model

(This is really just an application of **Bayes' rule**):

$$T^* = \text{argmax}_T \; P(T \mid S)$$

$$= \text{argmax}_T \; \underbrace{P(S \mid T)}_{\text{Translation Model}} \; \underbrace{P(T)}_{\text{Language Model}}$$

The **translation model** $P(S \mid T)$ is intended to capture the **faithfulness** of the translation. [this is the noisy channel]

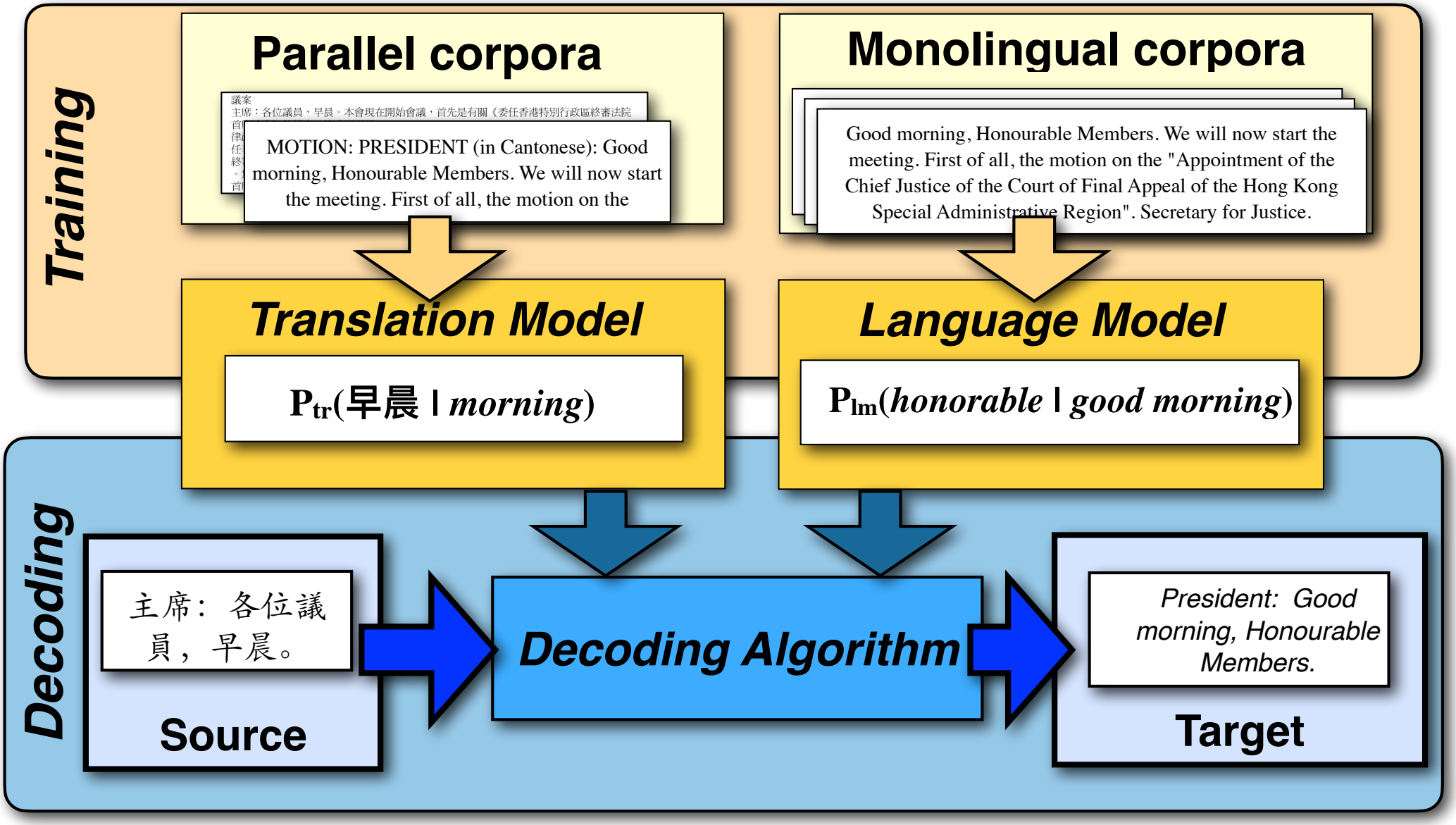
Since we only need $P(S \mid T)$ to score S, and don't need it to generate a grammatical S, it can be a relatively simple model.

$P(S \mid T)$ needs to be trained on a **parallel corpus**

The **language model** $P(T)$ is intended to capture the **fluency** of the translation.

$P(T)$ can be trained on a (very large) **monolingual corpus**

Statistical MT: Training and Decoding



IBM models

First statistical MT models, based on noisy channel:

Translate from (French/foreign) source $\mathbf{f}$ to (English) target $\mathbf{e}$
via a **translation model** $P(\mathbf{f} \mid \mathbf{e})$ and a **language model** $P(\mathbf{e})$

The translation model goes **from target $\mathbf{e}$ to source $\mathbf{f}$**
via **word alignments $\mathbf{a}$**: $P(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$

Original purpose: Word-based translation models
Later: Were used to obtain word alignments,
which are then used to obtain phrase alignments
for phrase-based translation models

Sequence of 5 translation models

Model 1 is too simple to be used by itself,
but can be trained very easily on parallel data.

n-gram language models for MT

With training on data from the web and clever parallel processing (MapReduce/Bloom filters), *n* can be quite large

- Google (2007) uses 5-grams to 7-grams,
- This results in huge models, but the effect on translation quality levels off quickly:

Size of models

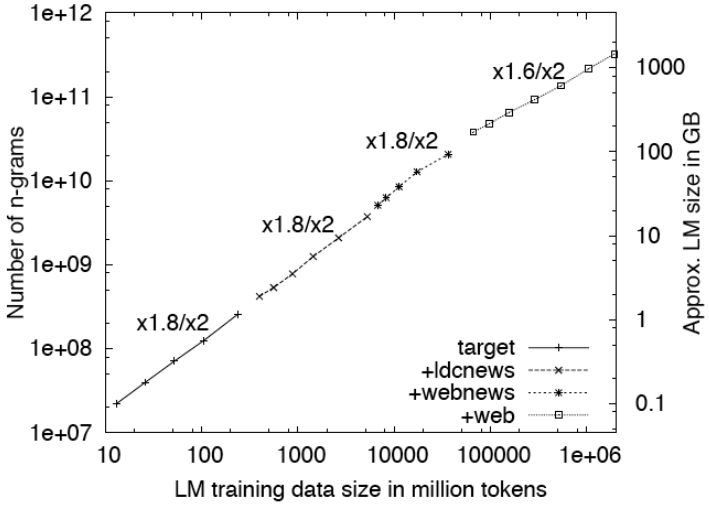


Figure 3: Number of n-grams (sum of unigrams to 5-grams) for varying amounts of training data.

Effect on translation quality

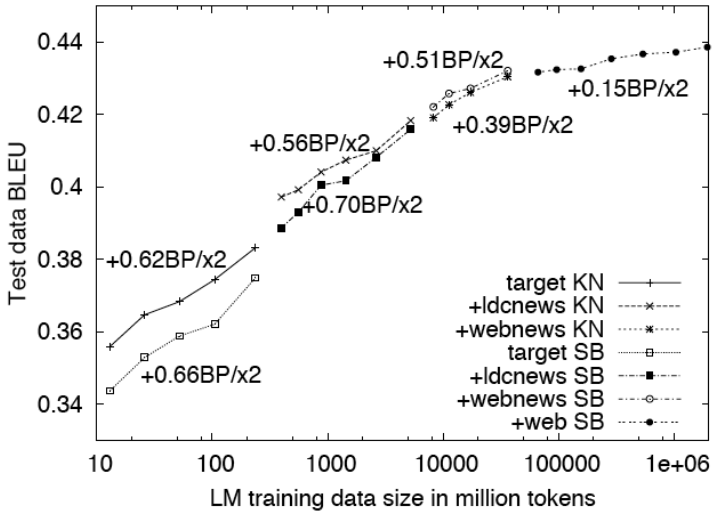


Figure 5: BLEU scores for varying amounts of data using Kneser-Ney (KN) and Stupid Backoff (SB).

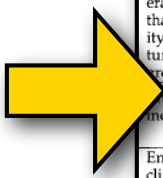
Getting translation probabilities

A **parallel corpus** consists of the same text in two (or more) languages.

Examples: Parliamentary debates: Canadian Hansards; Hong Kong Hansards, Europarl; Movie subtitles (OpenSubtitles)

In order to train translation models, we need to **align the sentences** (Church & Gale '93)

English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment. La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. L'emploi et les investissements ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.



English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

We can learn **word** and **phrase alignments** from these aligned sentences

IBM models

First statistical MT models, based on noisy channel:

Translate from (French/foreign) source $\mathbf{f}$ to (English) target $\mathbf{e}$
via a **translation model** $P(\mathbf{f} \mid \mathbf{e})$ and a **language model** $P(\mathbf{e})$

The translation model goes **from target $\mathbf{e}$ to source $\mathbf{f}$**
via **word alignments $\mathbf{a}$**: $P(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$

Original purpose: Word-based translation models
Later: Were used to obtain word alignments,
which are then used to obtain phrase alignments
for phrase-based translation models

Sequence of 5 translation models

Model 1 is too simple to be used by itself,
but can be trained very easily on parallel data.

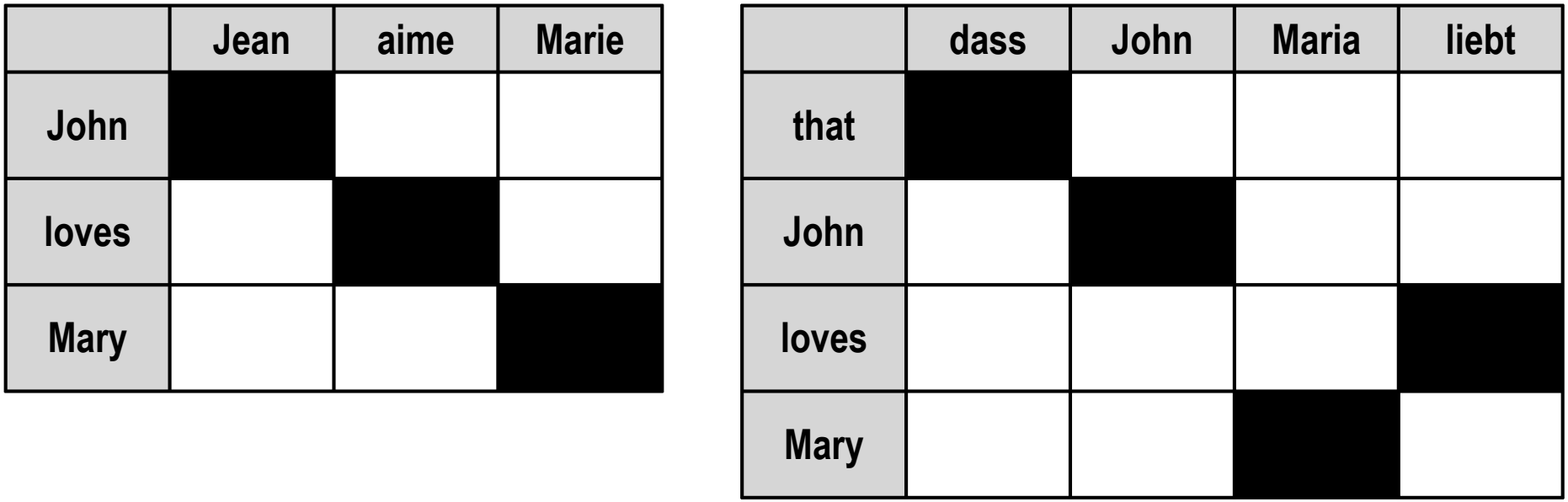
Word alignment: target (E) to source (F)

John loves Mary.

Jean aime Marie.

… that John loves Mary.

… dass John Maria liebt.



	Jean	aime	Marie
John	■		
loves		■	
Mary			■

	dass	John	Maria	liebt
that	■			
John		■		
loves				■
Mary			■	

Representing word alignments

		1	2	3	4	5	6	7	8
		Marie	a	traversé	le	lac	à	la	nage
0	NULL						■	■	
1	Mary	■							
2	swam								■
3	across		■	■					
4	the				■				
5	lake					■			

Position	1	2	3	4	5	6	7	8
Source	Marie	a	traversé	le	lac	à	la	nage
Alignment	1	3	3	4	5	0	0	2

Every source word **f[i]** is aligned to **one** target word **e[j]** (incl. NULL).
We represent alignments as a vector **a** (of the same length as the source) with **a[i] = j**

IBM models

The IBM models are **word-based** MT systems based on the **"noisy channel"** (Bayes rule)

- They model the probability of the input sentence S [source] given the output translation *T* [target], $P(S \mid T)$
- Because *S* is known, the translation model $P(S \mid T)$ does not need to be very sophisticated
- The language model $P(T)$ is needed to create fluent output

Word alignments were extended to phrase alignments, leading to **phrase-based MT** approaches that used logistic regression ("Maximum Entropy") models of $P(T \mid S)$ (still with language-model based features)

Word alignments also inspired **attention in seq2seq models**

Phrase-based MT models

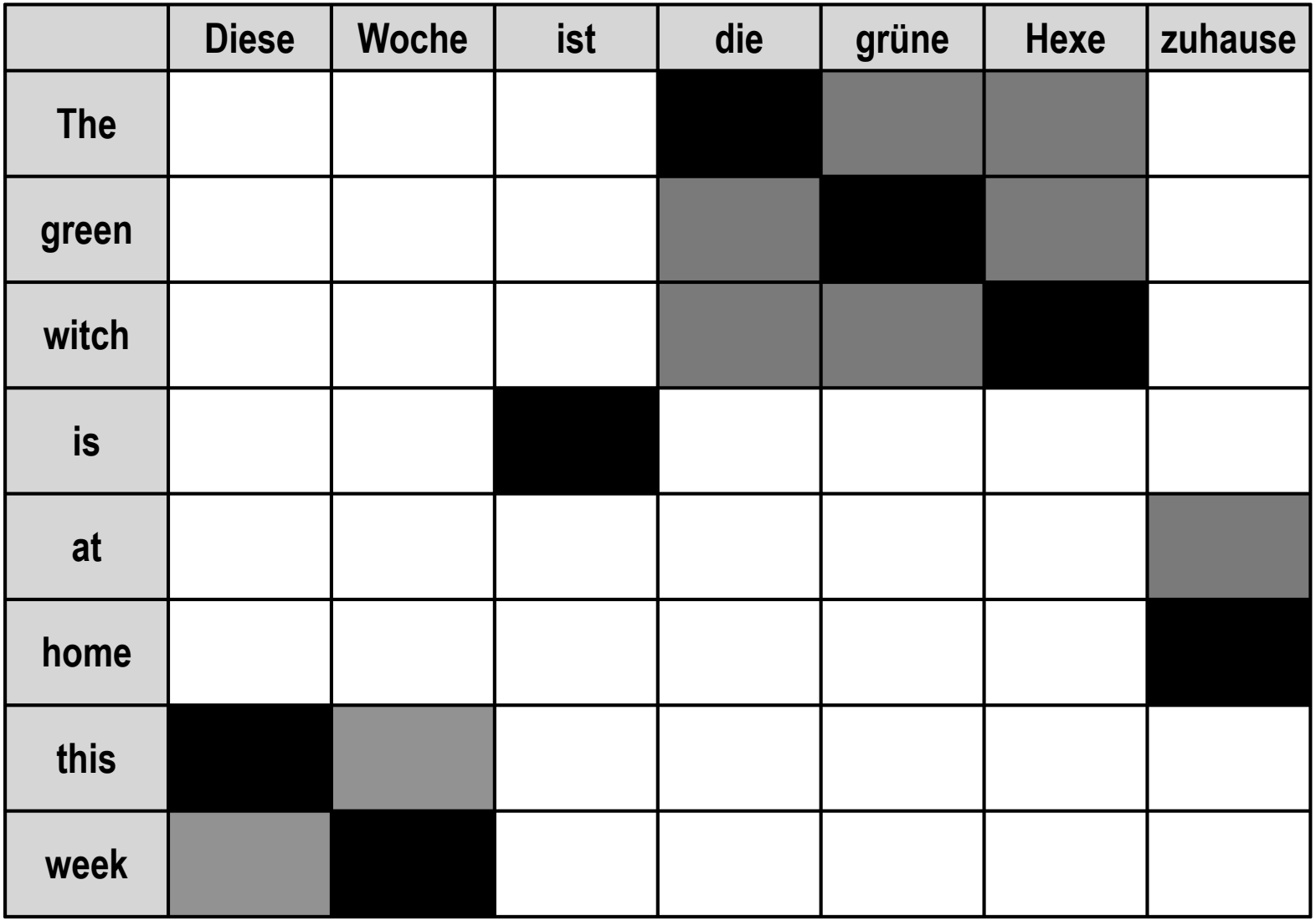
IBM models were replaced by **phrase-based MT approaches** that used logistic regression ("**Maximum Entropy**") models of $P(T \mid S)$ (still with language-model based features)

This required **phrase alignments** that can be used to obtain **phrase tables**

From word alignment…

	Diese	Woche	ist	die	grüne	Hexe	zuhause
The				■			
green					■		
witch						■	
is			■				
at							
home							■
this	■						
week		■					

… to phrase alignment



Phrase-based translation models

$P(sp_i \mid tp_i)$

Phrase translation probabilities of source phrases given target phrases can be obtained from a **phrase table:**

TP	SP	count
green witch	*grüne Hexe*	…
at home	*zuhause*	10534
at home	*daheim*	9890
is	*ist*	598012
this week	*diese Woche*	….

This requires **phrase alignment** on a parallel corpus.

Word alignments today

Machine translation inspired the creation of seq2seq architectures:

The encoder reads in source language input,
The decoder returns target language output (translation)

The concept of word alignments inspired **attention** in **seq2seq models**


MT
evaluation

How do we evaluate machine translation output?

What do we need to evaluate?

- Correctness of the translation
- Fluency of the translation, appropriateness, …

We need appropriate evaluation metrics

Automatic evaluation:

Inexpensive, can be done on a large scale,
but may not capture what we want to evaluate.

Human evaluation:

Expensive, and not easily reproducible or comparable across evaluations (different judges, different questions, …)

Automatic evaluation: BLEU

Evaluate candidate translations against several reference translations.

C1: It is a guide to action which ensures that the military always obeys the commands of the party.
C2: It is to insure the troops forever hearing the activity guidebook that party direct

R1: It is a guide to action that ensures that the military will forever heed Party commands.
R2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
R3: It is the practical guide for the army always to heed the directions of the party.

The **BLEU score** is based on **N-gram precision**:

How many n-grams in the candidate translation occur also in one of the reference translation?

BLEU details

For $n \in \{1,\ldots,4\}$, compute the (modified) **precision of all *n*-grams:**

$$Prec_n = \frac{\sum_{c \in C} \sum_{n\text{-gram} \in c} \text{MaxFreq}_{\text{ref}}(n\text{-gram})}{\sum_{c \in C} \sum_{\text{-gram} \in c} \text{Freq}_{\text{c}}(n\text{-gram})}$$

$\text{MaxFreq}_{\text{ref}}$(*'the party'*) = max. count of *'the party'* in **one** reference translation.

Freq_{c}(*'the party'*) = count of *'the party'* in candidate translation c.

Penalize short candidate translations by a **brevity penalty** BP

c = length (number of words) of the whole candidate translation corpus

r = Pick for each candidate the reference translation that is closest in length; sum up these lengths.

Brevity penalty $\text{BP} = \exp(1-\text{c/r})$ for $\text{c} \leq \text{r}$; $\text{BP} = 1$ for $\text{c} > \text{r}$

(BP ranges from e for c=0 to 1 for c=r)

BLEU score

The BLEU score is the geometric mean of the modified n-gram precision (for n=1..4), weighted by a brevity penalty BP:

$$\mathbf{BLEU} = BP \times \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log Prec_n\right)$$

Geometric mean for $a_1, \ldots, a_N > 0$ = N-th root of $\prod_{n=1}^{N} a_n$

$$\sqrt[N]{\prod_{n=1}^{N} a_n} = \left(\prod_{n=1}^{N} a_n\right)^{\frac{1}{N}} = \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log a_n\right)$$

BLEU details

Compute the (modified) precision of all *n*-grams (for $n = 1 \ldots 4$)

Sum over the translations c of any sentence in the test corpus C…

…sum over all n-grams occurring in c..

… the **maximum frequency** of that n-gram in any ***one*** of c's **reference** translations.

For n = 1..4:

$$Prec_n = \frac{\sum_{c \in C} \sum_{n\text{-gram} \in c} \text{MaxFreq}_{\text{ref}}(n\text{-gram})}{\sum_{c \in C} \sum_{\text{-gram} \in c} \text{Freq}_{\text{c}}(n\text{-gram})}$$

Sum over the translations c of any sentence in the test corpus C…

…sum over all n-grams occurring in c..

… the **frequency** of that n-gram in c.

Penalize short candidate translations by a brevity penalty BP

BP = exp(1–c/r) for $c \leq r$; BP = 1 for $c > r$

(BP ranges from 1 for c=r to e for c=0)

c = Total length (number of words) of the whole candidate translation corpus

r = Total length of all reference translations closest in length to candidates

Human evaluation

We want to know…

whether the translation is **"good" English**, and…

… whether it is an **accurate translation** of the original.

— Ask human raters to judge the **fluency** and the **adequacy** of the translation (e.g. on a scale of 1 to 5)

— Correlated with fluency is accuracy on **cloze task**:

Give rater the sentence with one word replaced by blank.
Ask rater to guess the missing word in the blank.

— Similar to adequacy is **informativeness**

Can you use the translation to perform some task
(e.g. answer multiple-choice questions about the text)