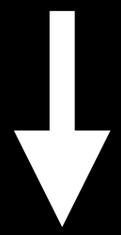
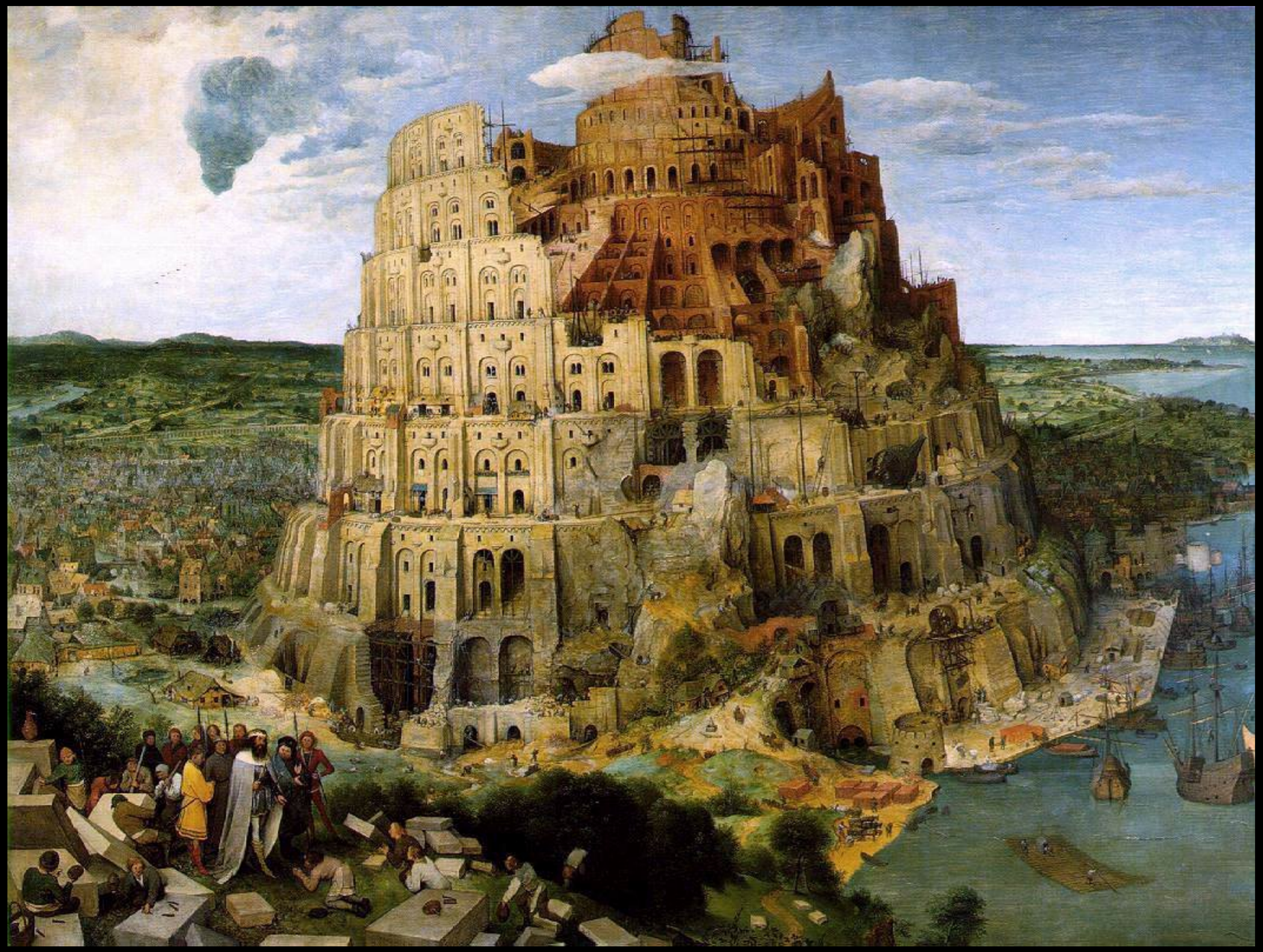
Machine Translation

ペンシルベニア大大学で講演をしています。

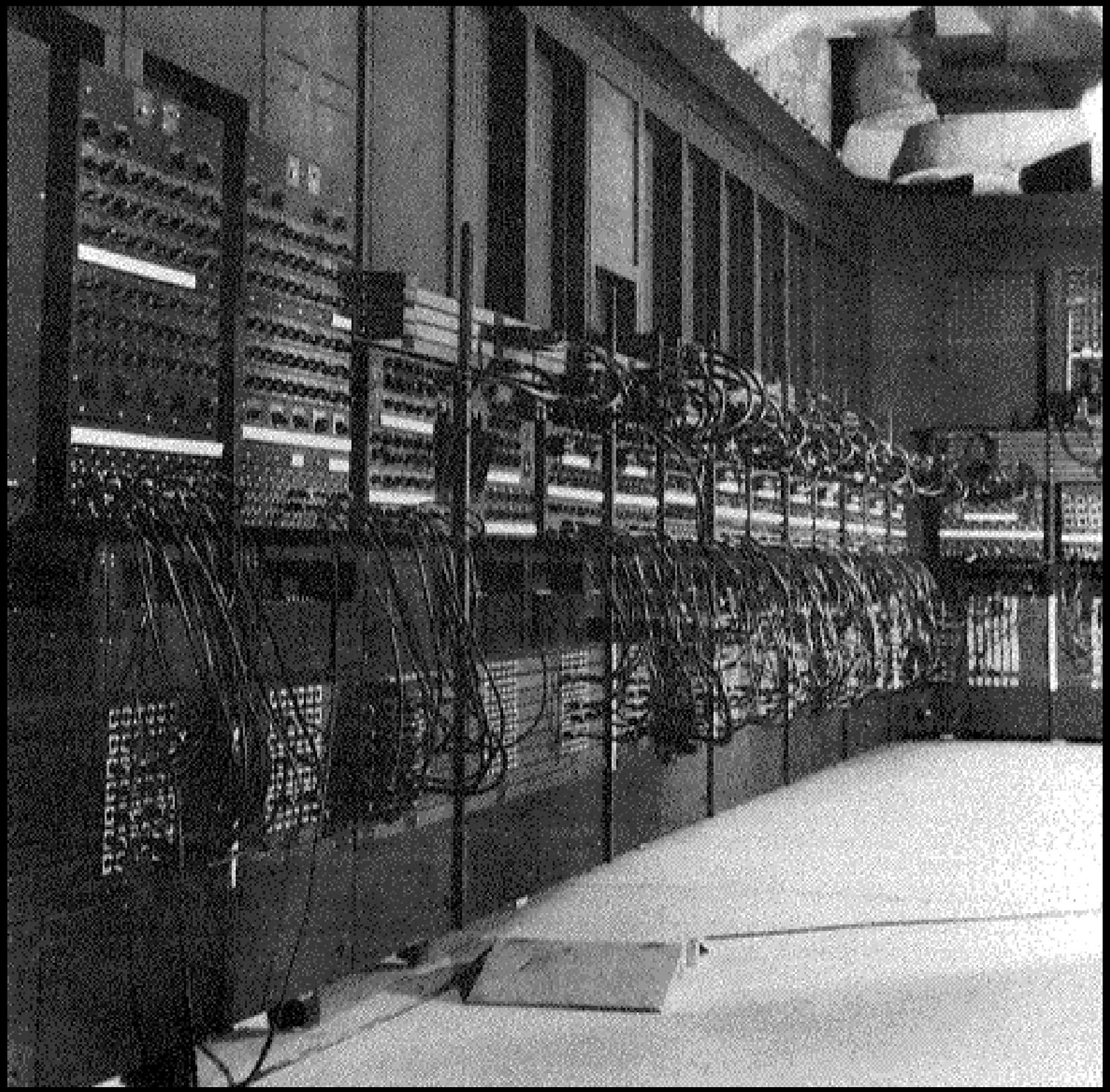


I'm giving a talk at University of Pennsylvania



The Tower of Babel

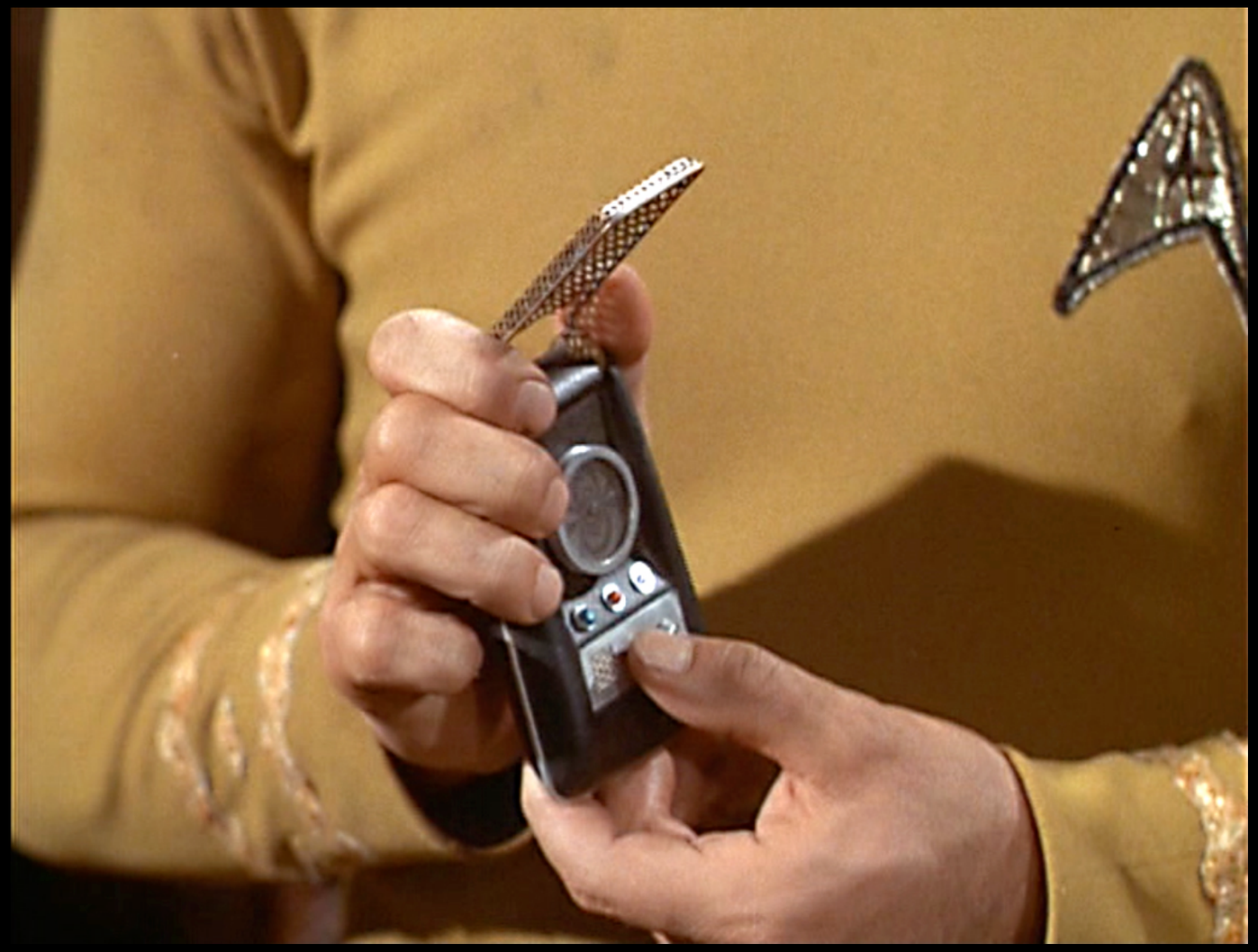
Pieter Brueghel the Elder (1563)



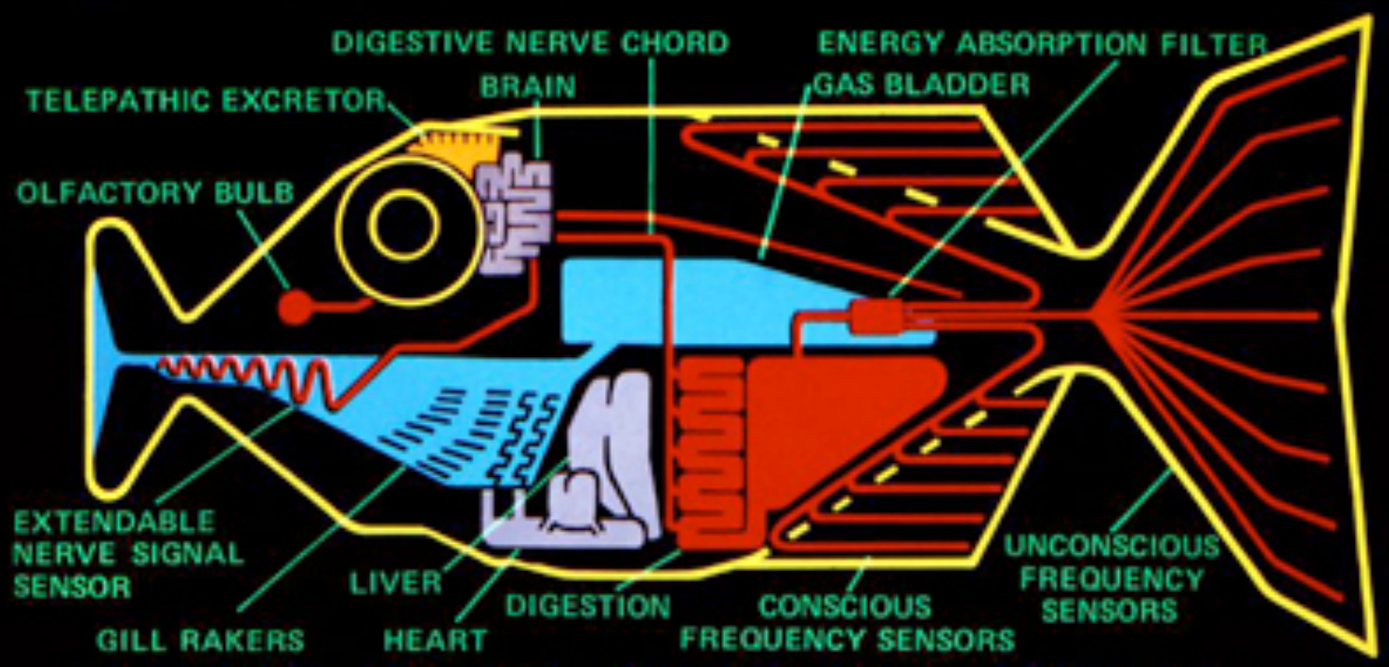
ENIAC (1946)

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 (1949)



Star Trek



Hitchhiker's Guide to the Galaxy

Statistical Machine Translation Live

4/28/2006
Franz Och

Because we want to provide everyone with access to all the world's information, including information written in every language, one of the exciting projects at Google Research is machine translation... Now you can see the results for yourself. We recently launched an online version of our system for Arabic-English and English-Arabic. Try it out!

Google AI Blog

The latest news from Google AI

A Neural Network for Machine Translation, at Production Scale

Tuesday, September 27, 2016

Posted by Quoc V. Le & Mike Schuster, Research Scientists, Google Brain Team

Ten years ago, we announced the launch of Google Translate, together with the use of Phrase-Based Machine Translation as the key algorithm behind this service. Since then, rapid advances in machine intelligence have improved our speech recognition and image recognition capabilities, but improving machine translation remains a challenging goal.

Today we announce the Google Neural Machine Translation system (GNMT), which utilizes state-of-the-art training techniques to achieve the largest improvements to date for machine translation quality. Our full research results are described in a new technical report we are releasing today: "*Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*" [1].

A few years ago we started using Recurrent Neural Networks (RNNs) to directly learn the mapping between an input sequence (e.g. a sentence in one language) to an output sequence (that same sentence in another language) [2]. Whereas Phrase-Based Machine Translation (PBMT) breaks an input sentence into words and phrases to be translated largely independently, Neural Machine Translation (NMT) considers the entire input sentence as a unit for translation.The advantage of

Google Translate

Text | Documents

DETECT LANGUAGE | SPANISH | TURKISH | ENGLISH

ENGLISH | INDONESIAN | BOSNIAN

Search languages

Detect language	Czech	Hebrew	Latin	Portuguese	Tajik
Afrikaans	Danish	Hindi	Latvian	Punjabi	Tamil
Albanian	Dutch	Hmong	Lithuanian	Romanian	Telugu
Amharic	English	Hungarian	Luxembourgish	Russian	Thai
Arabic	Esperanto	Icelandic	Macedonian	Samoan	Turkish
Armenian	Estonian	Igbo	Malagasy	Scots Gaelic	Ukrainian
Azerbaijani	Filipino	Indonesian	Malay	Serbian	Urdu
Basque	Finnish	Irish	Malayalam	Sesotho	Uzbek
Belarusian	French	Italian	Maltese	Shona	Vietnam
Bengali	Frisian	Japanese	Maori	Sindhi	Welsh
Bosnian	Galician	Javanese	Marathi	Sinhala	Xhosa
Bulgarian	Georgian	Kannada	Mongolian	Slovak	Yiddish
Catalan	German	Kazakh	Myanmar (Burmese)	Slovenian	Yoruba
Cebuano	Greek	Khmer	Nepali	Somali	Zulu
Chichewa	Gujarati	Korean	Norwegian	✓ Spanish	
Chinese	Haitian Creole	Kurdish (Kurmanji)	Pashto	Sundanese	
Corsican	Hausa	Kyrgyz	Persian	Swahili	
Croatian	Hawaiian	Lao	Polish	Swedish	


www.un.org
http://www.un.org/english/
Google
National Bureau
http://www.stats.gov.cn/engl
NBS
中文简体
www.stats.gov.cn
Agency Information
Laws & Regulations
Survey Programs
Statistical Standards
Statistical Indicators
New Release
Int. Cooperation
NBS/IMF Meeting
Int. Conference On Ag Census
International Symposium on Statistical Cooperation
18th Roundtable On Business Survey Frames
News and Coming
Memorial Ceremony for Late Deputy Commissioner Zhu Xiangdong Held in Beijing(09.16)
The Urban Investment in Fixed Assets Continued Increasing in August(09.16)
German Delegation Visited the National Bureau of Statistics of China(09.15)
The Value-added of Industry up by 16 Percent in August(09.15)
The Total Retail Sale of Consumer Goods Increased in August(09.14)
The Consumer Price Index (CPI) Increased in August(09.13)
The producers' Price Index (PPI) For Manufactured Goods Kept Advancing in August(09.12)
Global Manager of ICP of World Bank Visited Beijing(09.08)
What's New
Monthly Data Updated(09.15)
Statistical Data: Women and Men in China----Facts and Figures 2004(09.08)
Monthly Data Updated(09.07)
Monthly Data Updated(08.29)
Monthly Data Updated(08.23)
Statistical Data
Monthly
Yearly
Census
Others
Related links
Chinese Version
Others
Search
Aug Oct Dec Feb Apr Jun Aug
Live and On-Demand Webcasts, 24 Hours a Day: Click on UN Webcas


联合国主页
http://www.un.org/chinese/
Google
中华人民共和国家统计局
http://www.stats.gov.cn/
中华人民共和国国家
National Bureau of Statistics
中文繁体
ENGLISH
105年9月18日 星期日
统计公报
统计数据
进度数据 年度数据
专题数据 部门数据
统计分析
进度信息 分析报告
专题分析 统计图表
地方信息 企业景气
统计出版
统计法规
统计管理
部门统计调查项目
涉外调查管理
数据直报
统计链接
国内统计网站 政府网站
国外统计网站 国际网站
最新统计信息
2005年全国早稻总产量比上年减产43万吨（09.16）
8月份"国房景气指数"为101.86 同比下降3.10点（09.16）
1-8月湖南城镇居民人均可支配收入同比增长10.4%（09.16）
1-8月甘肃固定资产投资增长17.64% 增幅回落3.41%（09.16）
株洲：商品房预售制度对房地产市场的影响浅析（09.16）
经济全球化对江西国民经济发展产生六大影响（09.16）
统计数据：8月份工业产品产量 各地区产品销售率（09.15）
统计数据：8月份工业增加值 各地区工业增加值（09.15）
1-8月份全国城镇固定资产投资同比增长27.4%（09.15）
加快云南人口城市化进程需解决四大关键问题（09.15）
丹江口：遏止教育乱收费 "一费制"深入人心（09.15）
1-8月浙江限额以上固定资产投资同比增长16.4%（09.15）
8月份广西消费品零售额与去年同期相比增长13.6%（09.15）
8月份我国工业实现增加值5968亿 同比增长16%（09.14）
调查显示：广东省企业流动资金短缺问题日益突出（09.14）
实施品牌战略 推动吉林省经济快速可持续发展（09.14）
无锡：城乡居民收入剪刀差十年扩大0.46倍（09.14）
8月份甘肃工业品价格呈现四特点 波动频率有所加快（09.14）
最新统计动态
河南省县级政府领导干部统计知识培训班举行（09.16）
白银市副市长火统元勉励统计工作者当好"参谋"（09.16）
吉林磐石市统计机构建设加强 成立实体统计工作站（09.16）
《新疆五十年》出版发行 自治区领导为该书题词（09.16）
国家统计局局长李德水撰文谈心得：沧桑巨变话地球（09.15）
湖南省副省长肖捷强调统计要实事更要求是（09.15）
山东省统计局实行行政许可办理"窗口"式办公（09.15）
重要公告
讣 告
关于申报2005年度全国统计科研计划项目的通知
印发《关于统计上对公有和非公有控股经济的分类办法》的通知
统计机构
统计动态
专项统计工作 业务资料下载
统计标准
区划代码 行业分类 企业划型
统计制度
统计知识
10月 12月 2月 4月 6月 8月
联合国网络直播

In-class exercise

Word aligner

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

Word aligner

Garcia and associates .
Garcia y asociados .

Carlos Garcia has three associates .
Carlos Garcia tiene tres asociados .

his associates are not strong .
sus asociados no son fuertes .

Garcia has a company also .
Garcia tambien tiene una empresa .

its clients are angry .
sus clientes estan enfadados .

the associates are also angry .
los asociados tambien estan enfadados .

the clients and the associates are enemies .
los clientes y los asociados son enemigos .

the company has three groups .
la empresa tiene tres grupos .

its groups are in Europe .
sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .
los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .
los grupos no venden zanzanina .

the small groups are not modern .
los grupos pequenos no son modernos .

a : una
also : tambien
and : his
angry : enfadados
are : estan
are : son
associates : asociados
clients : clientes
Carlos : Carlos
company : empresa
do : ???
enemies : enemigos
Garcia : Garcia
groups : groupos
has : tiene
his : sus
its : sus
modern : godernos
not : no
pharmaceuticals : medicinas
sell : venden
small : pequeños
strong : fuertes
the : los
three : tres
zanzanine : zanzanina

Lexical Translation

- How do we translate a word? Look it up in the dictionary

  Haus : house, home, shell, household

- Multiple translations
  - Different word senses, different registers, different inflections (?)
  - *house, home* are common
  - *shell* is specialized (the Haus of a snail is a shell)

How common is each?

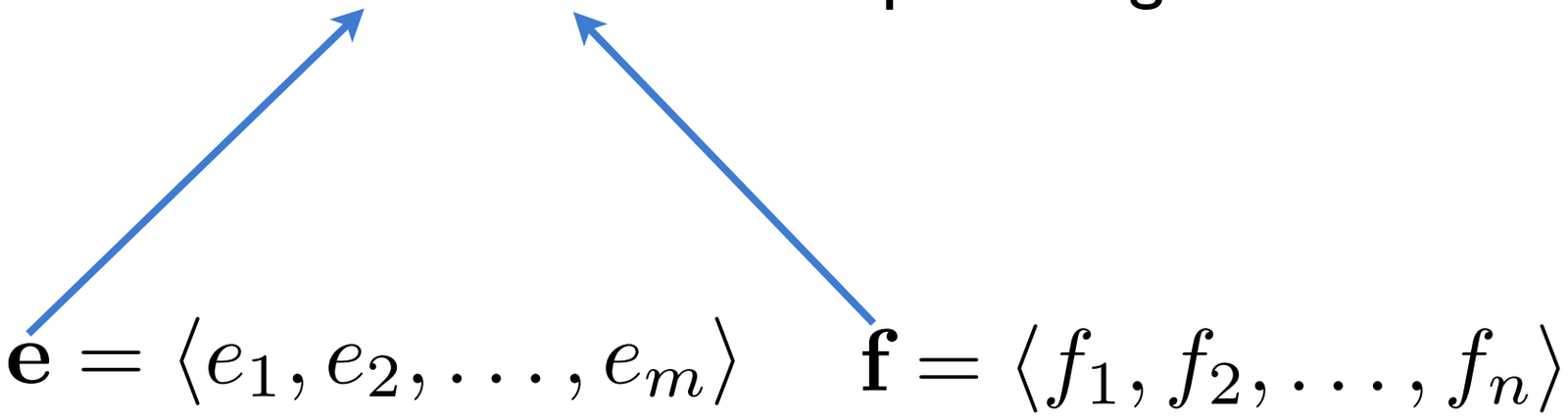
Translation	Count
house	5000
home	2000
shell	100
household	80

MLE

$$\hat{p}_{\mathrm{MLE}}(e \mid \texttt{Haus}) = \begin{cases} 0.696 & \text{if } e = \texttt{house} \\ 0.279 & \text{if } e = \texttt{home} \\ 0.014 & \text{if } e = \texttt{shell} \\ 0.011 & \text{if } e = \texttt{household} \\ 0 & \text{otherwise} \end{cases}$$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences



Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where **e** and **f** are complete English and Foreign sentences
- Lexical translation makes the following ***assumptions***:
  - Each word e_i in **e** is generated from exactly one word in **f**
  - Thus, we have an *alignment* a_i that indicates which word e_i "came from", specifically it came from f_{a_i} .
  - Given the alignments **a**, translation decisions are conditionally independent of each other and depend *only* on the aligned source word f_{a_i} .

Lexical Translation

- Putting our assumptions together, we have:

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in [0,n]^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

Alignment × Translation | Alignment

Lexical Translation

$$p(e_i \mid f_{a_i})$$

$p(\texttt{house} \mid \texttt{Haus})$ $p(\texttt{shell} \mid \texttt{Haus})$

Remember bigram models...

Lexical Translation

- Putting our assumptions together, we have:

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in [0,n]^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

Alignment × Translation | Alignment

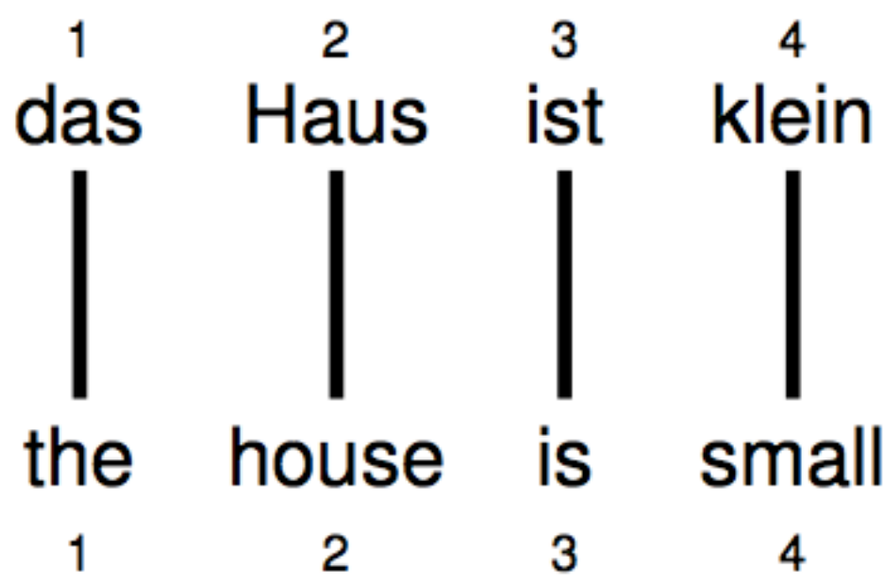
Alignment

$$p(\mathbf{a} \mid \mathbf{f}, m)$$

Most of the action for the early days of statistical MT was here. Words weren't the problem, word *order* was hard.

Alignment

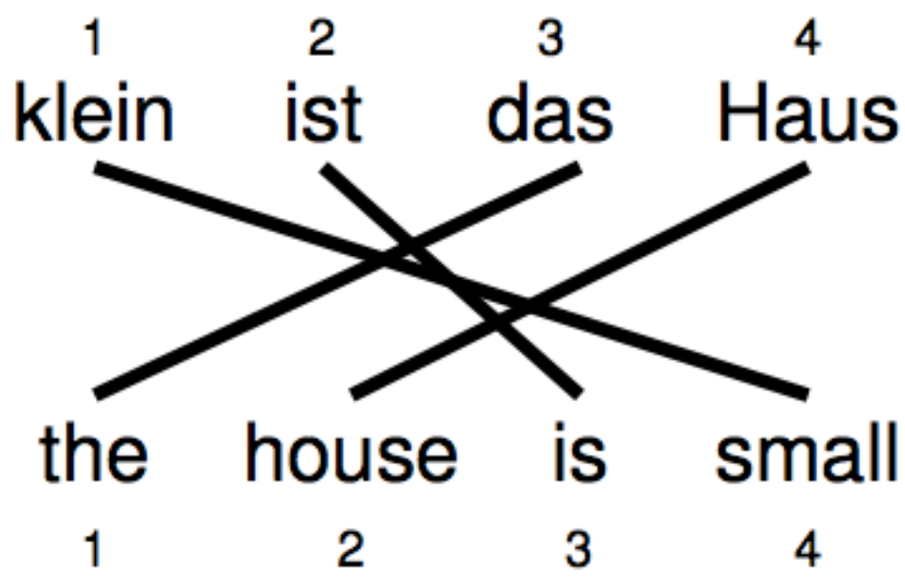
- Alignments can be visualized in by drawing links between two sentences, and they are represented as vectors of positions:



$$\mathbf{a} = (1, 2, 3, 4)^\top$$

Reordering

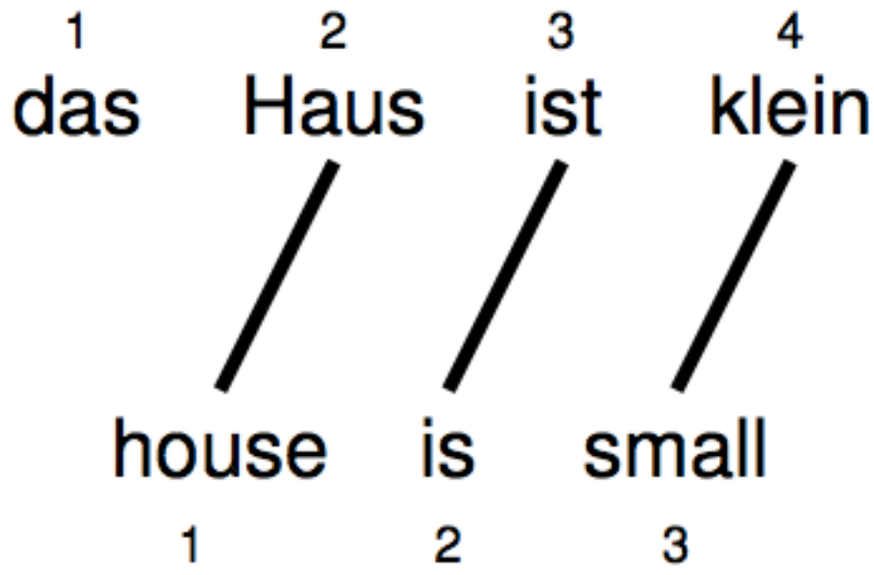
- Words may be reordered during translation.



$$\mathbf{a} = (3, 4, 2, 1)^\top$$

Word Dropping

- A source word may not be translated at all

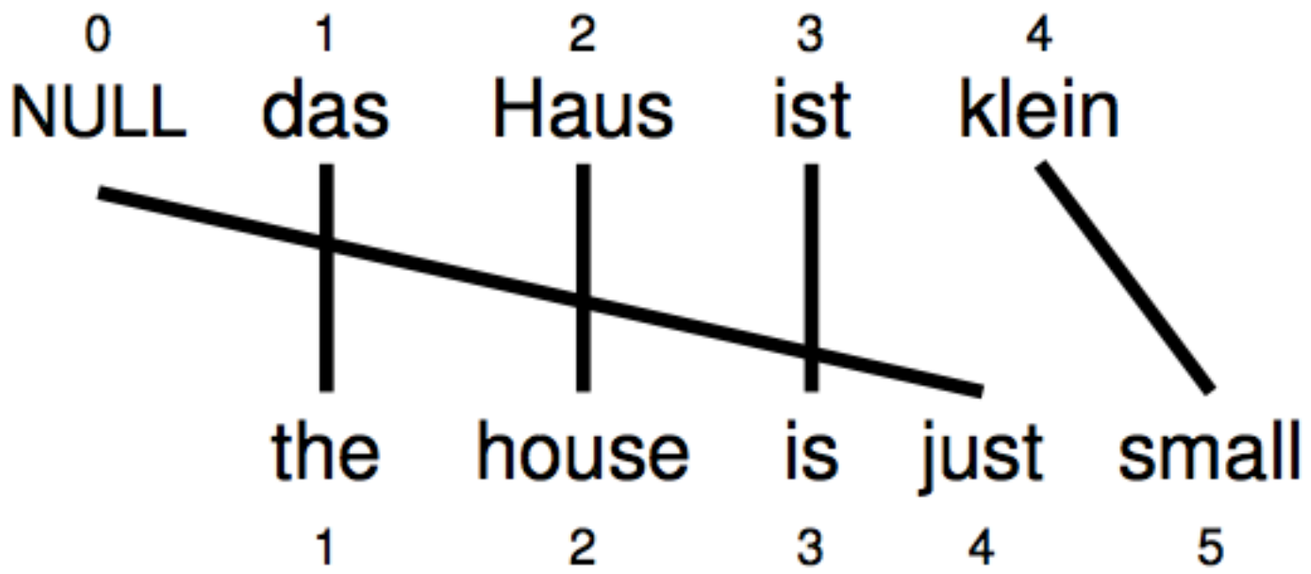


$$\mathbf{a} = (2, 3, 4)^\top$$

Word Insertion

- Words may be inserted during translation
  English just does not have an equivalent

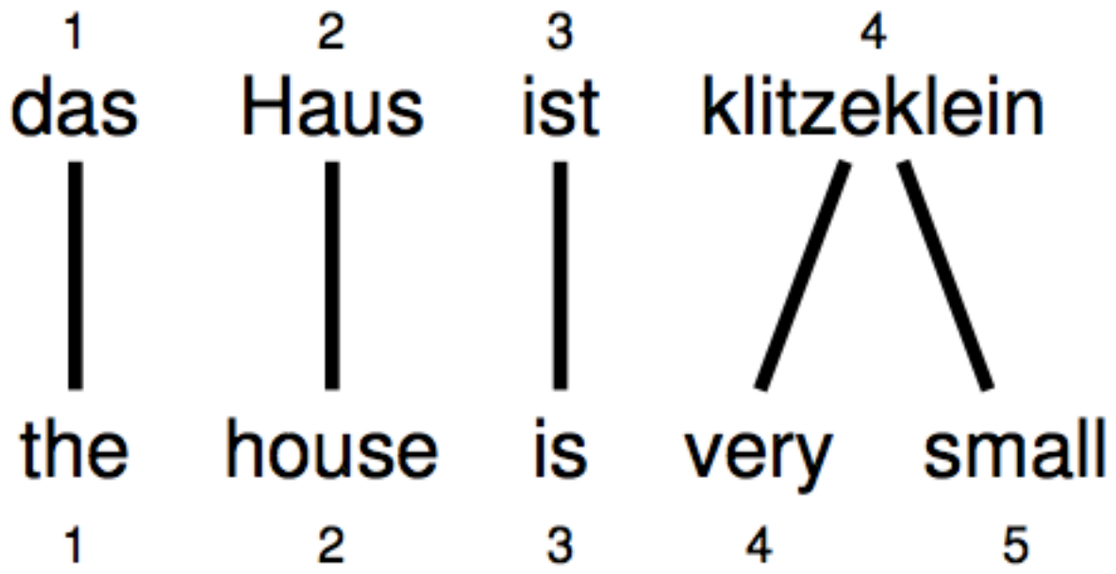
  But it must be explained - we typically assume every source sentence contains a NULL token



$$\mathbf{a} = (1, 2, 3, 0, 4)^\top$$

One-to-many Translation

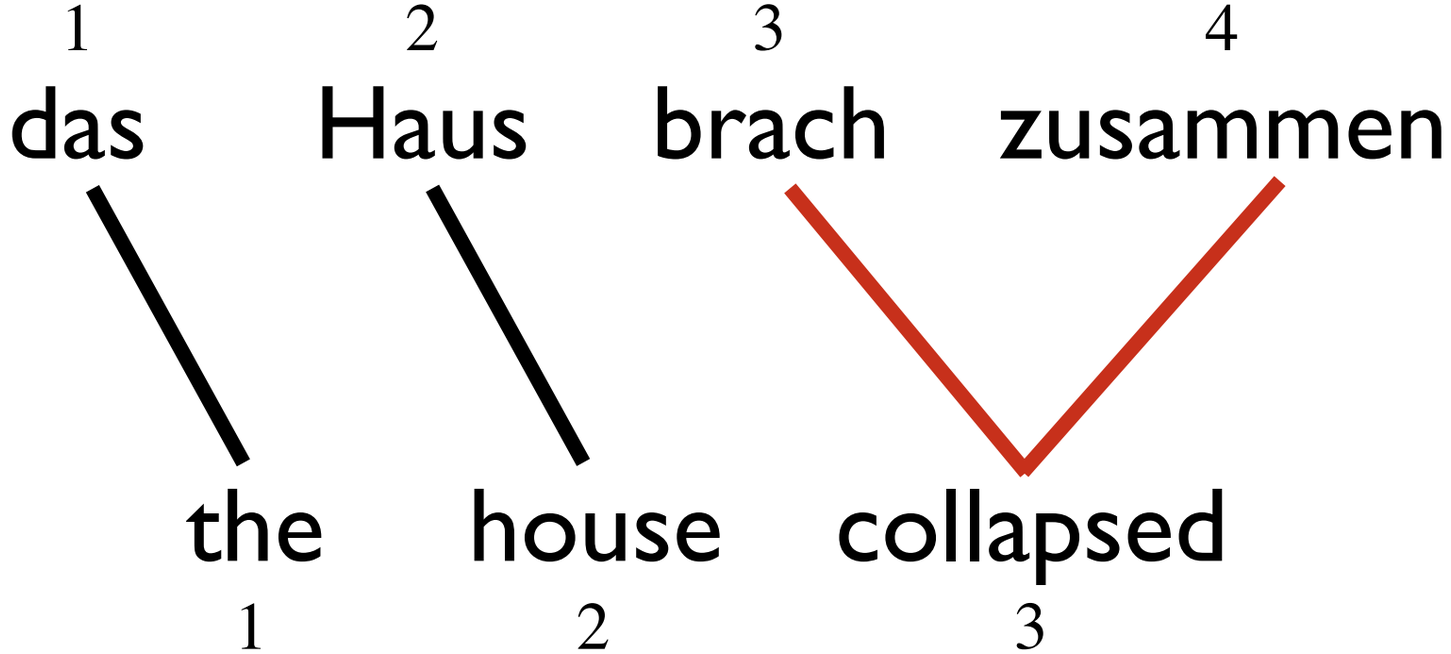
- A source word may translate into **more than one** target word



$$\mathbf{a} = (1, 2, 3, 4, 4)^\top$$

Many-to-one Translation

- **More than one source word** may **not** translate as a unit in lexical translation



$\mathbf{a} = ???$ $\mathbf{a} = (1, 2, (3, 4)^\top)^\top$ **?**

IBM Model 1

- Simplest possible lexical translation model
- Additional assumptions
  - The m alignment decisions are independent
  - The alignment distribution for each $a_i$is uniform over all source words and NULL

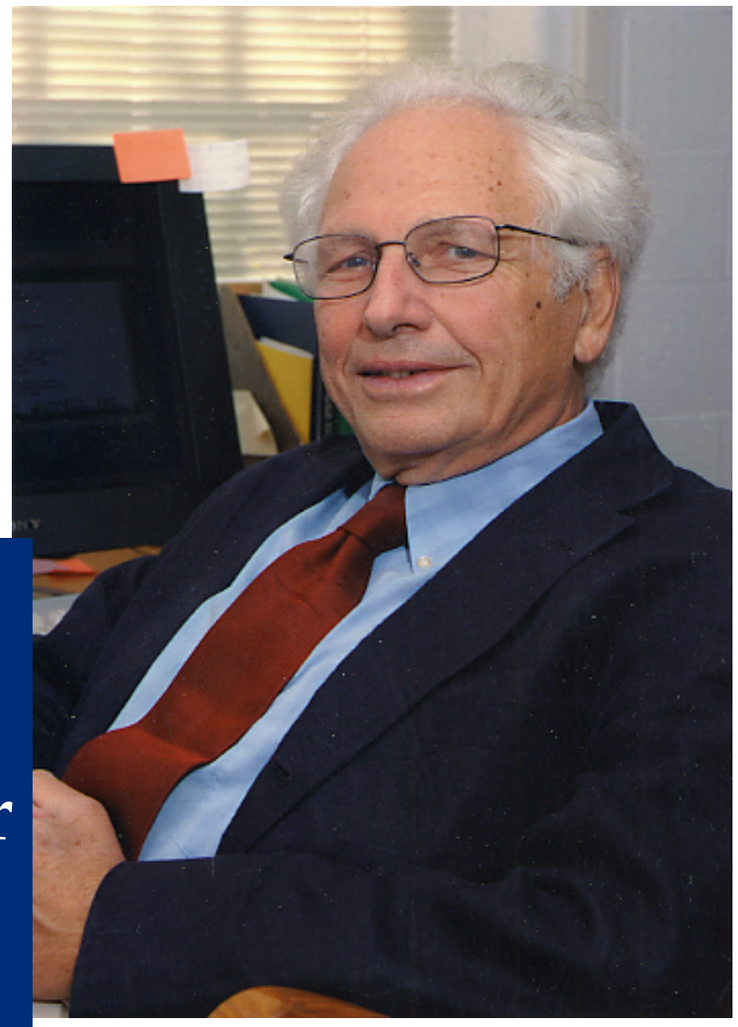
$$\text{for each } i \in [1, 2, \ldots, m]$$
$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$
$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

Historical Note

IBM Models

Renaissance

"The validity of a statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950 (cf. Hutchins, MT – Past, Present, Future, Ellis Horwood, 1986, p. 30ff and references therein). The crude force of computers is not science. The paper is simply beyond the scope of COLING."



Fred Jelinek
(1932-2010)

The Center For Language and Speech Processing
at the Johns Hopkins University

1. Introduction

The growing availability of bilingual, machine-readable texts has stimulated interest

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m}$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n}$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n}\, p(e_i \mid f_{a_i})$$

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i, a_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given the alignments then all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

$$p(a, b, c, d) = p(a)p(b)p(c)p(d)$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

$$= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

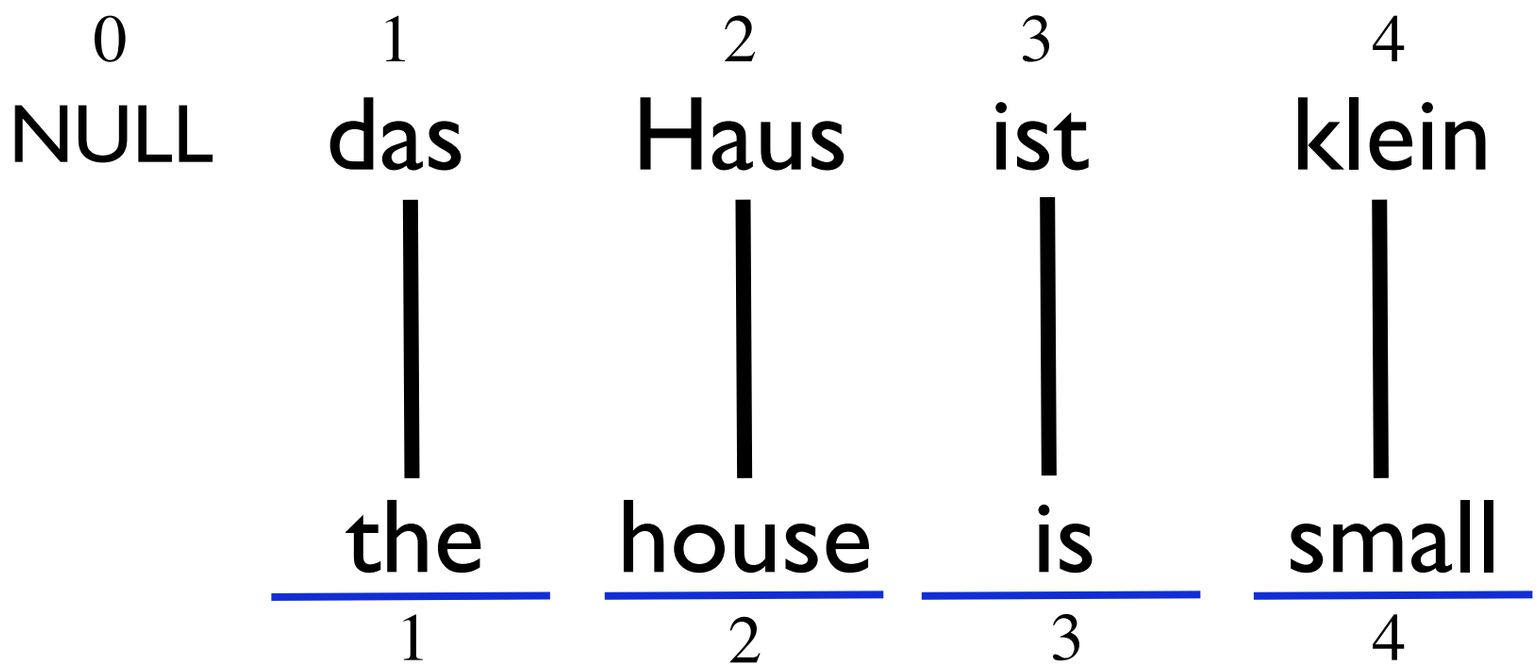
$$\begin{aligned} p(\mathbf{e} \mid \mathbf{f}, m) &= \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m) \\ &= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i}) \\ &= \frac{1}{(1+n)^m} \prod_{i=1}^{m} \sum_{a_i=0}^{n} p(e_i \mid f_{a_i}) \end{aligned}$$

Example

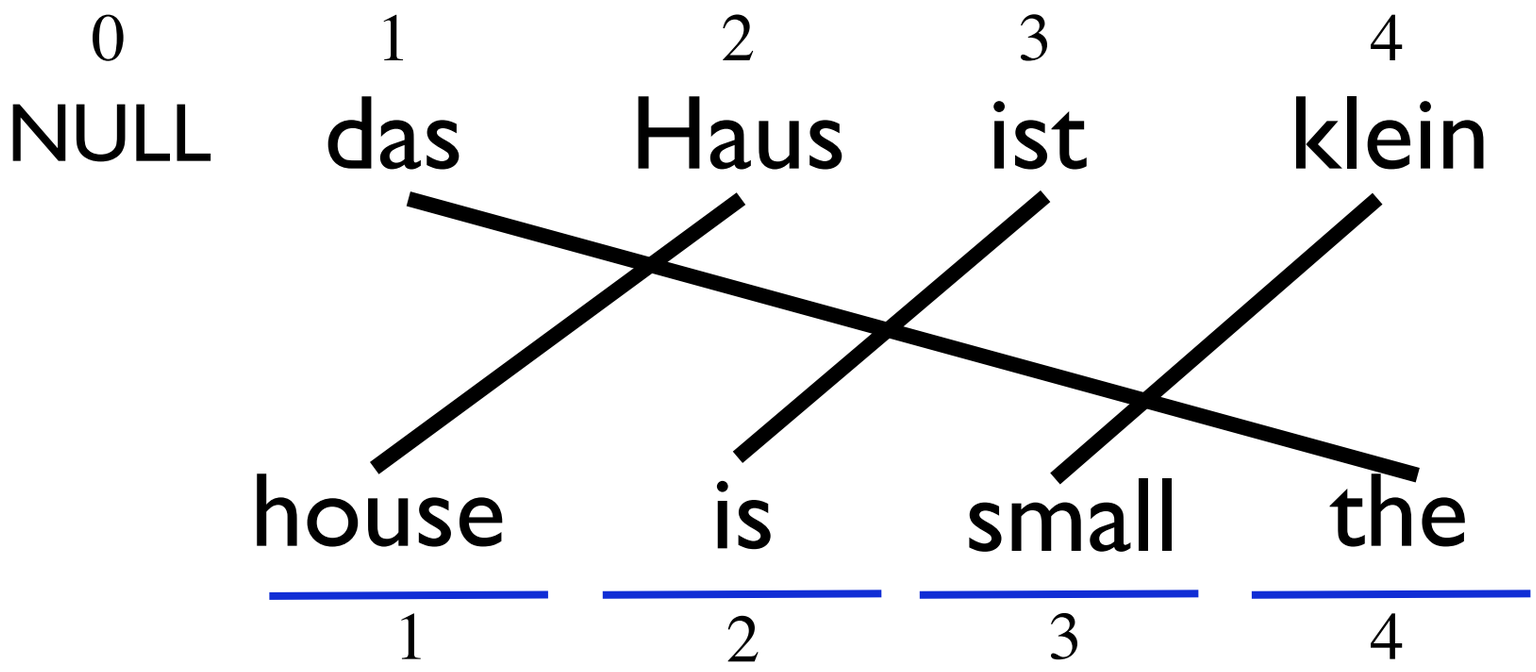
0	1	2	3	4
NULL	das	Haus	ist	klein

Start with a foreign sentence and a target length.

Example



Example



Finding the Viterbi Alignment

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in [0,1,\ldots,n]^m} p(\mathbf{a} \mid \mathbf{e}, \mathbf{f})$$

Historical Note #2

The **Viterbi algorithm** is a dynamic programming algorithm for finding the most likely sequence of hidden states – called the **Viterbi path** – that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models.

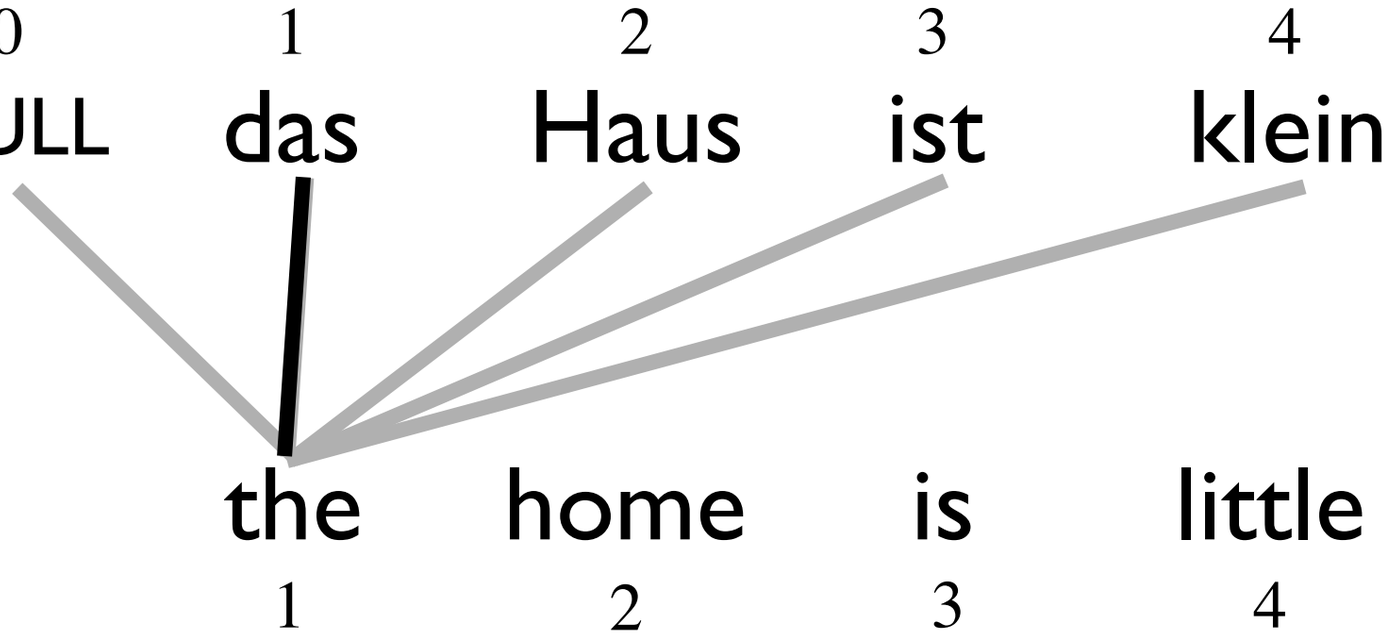
Andrew Viterbi

Professor at USC

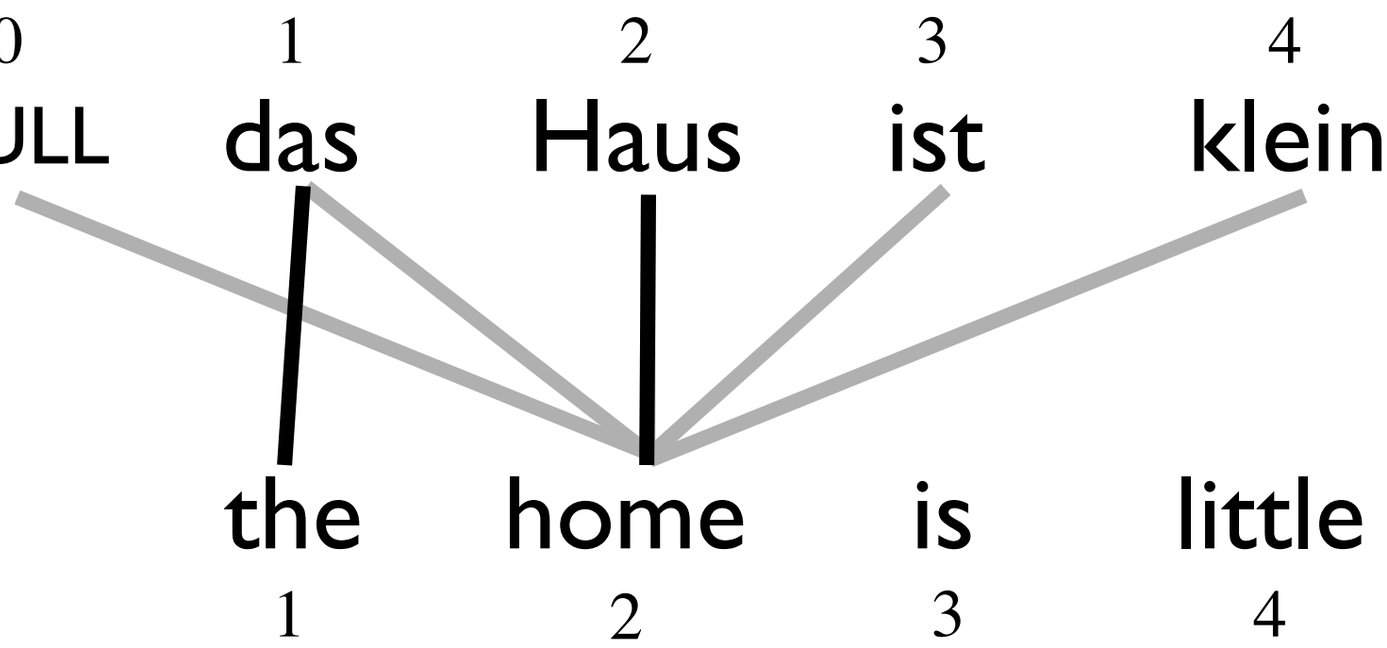
co-founder of Qualcomm

classmates with Fred Jelinek

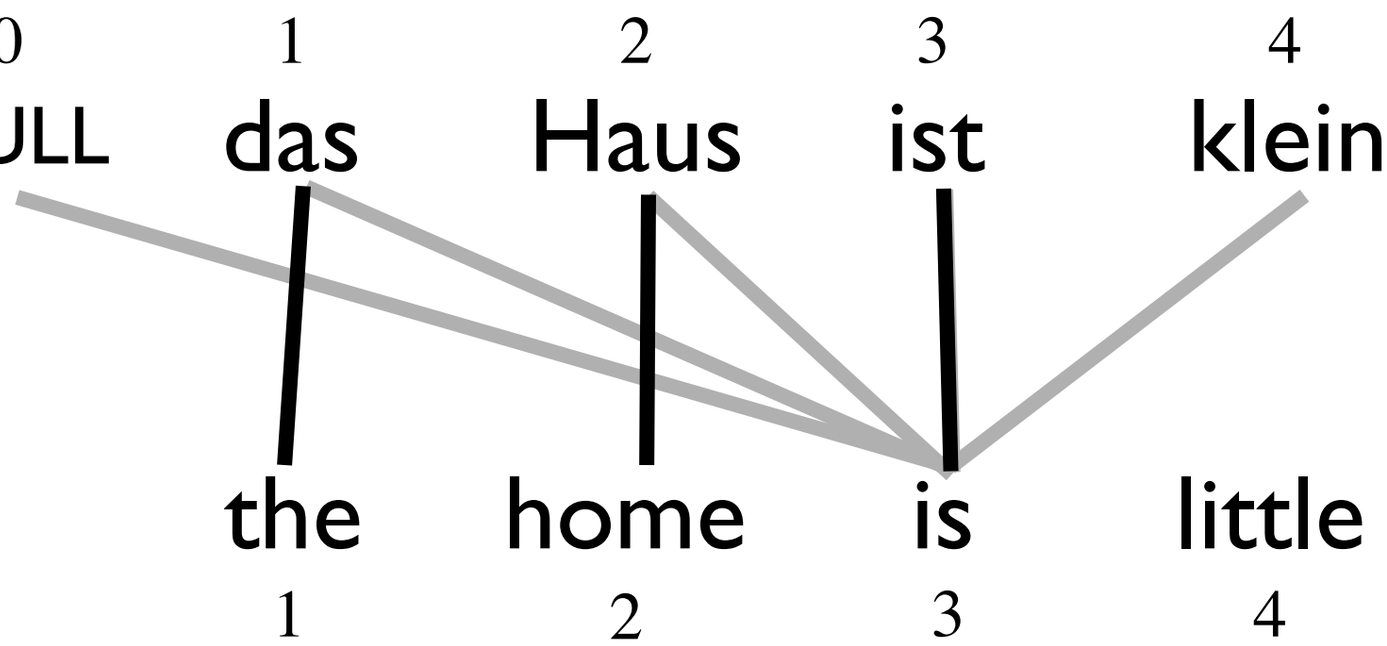
Finding the Viterbi Alignment



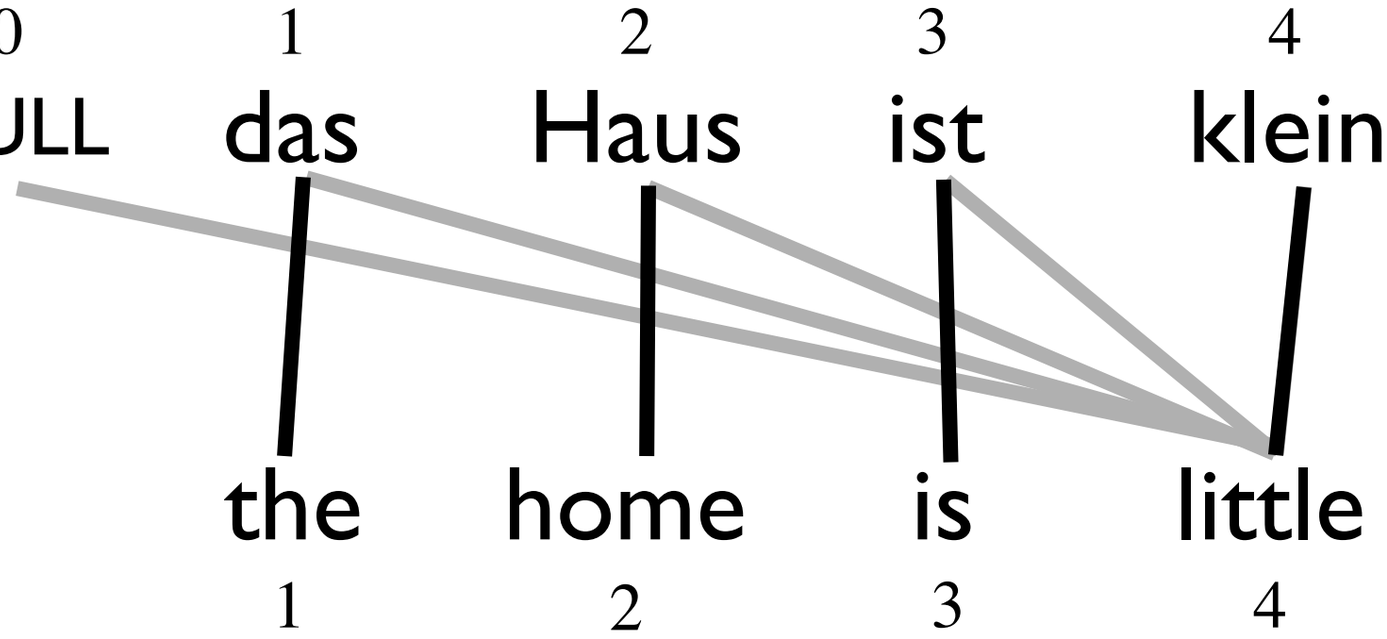
Finding the Viterbi Alignment



Finding the Viterbi Alignment

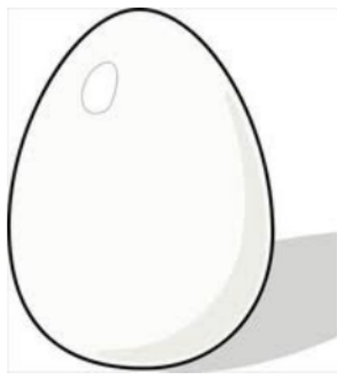


Finding the Viterbi Alignment



Learning Lexical Translation Models

- How do we learn the parameters $p(e \mid f)$
- "Chicken and egg" problem
  - If we had the alignments, we could estimate the parameters (MLE)
  - If we had parameters, we could find the most likely alignments



EM Algorithm

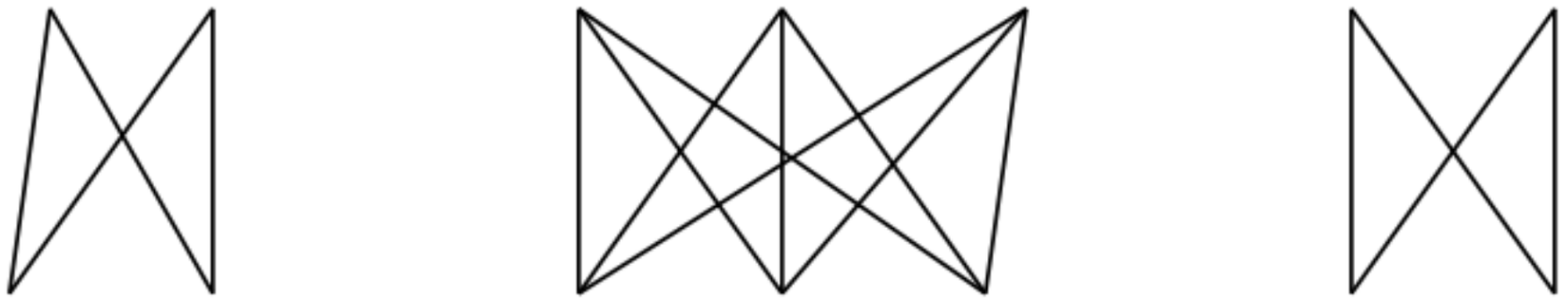
- pick some random (or uniform) parameters
- Repeat until you get bored (~ 5 iterations for lexical translation models)
  - using your current parameters, compute "expected" alignments for every target word token in the training data

$$p(a_i \mid \mathbf{e}, \mathbf{f})$$

  - keep track of the expected number of times f translates into e throughout the whole corpus
  - keep track of the expected number of times that f is used as the source of any translation
  - use these expected counts as if they were "real" counts in the standard MLE equation

EM for Model 1

... la maison ... la maison blue ... la fleur ...

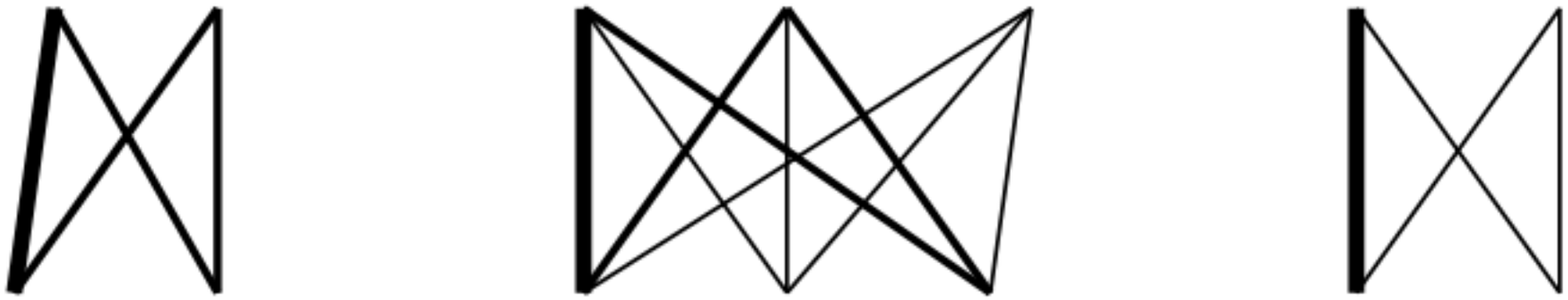


... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely
- Model learns that, e.g., la is often aligned with the

EM for Model 1

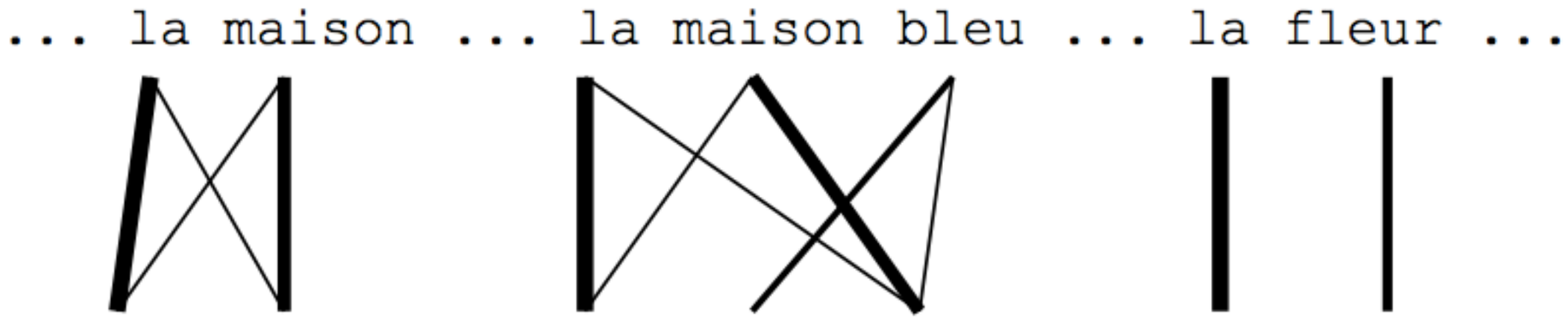
... la maison ... la maison blue ... la fleur ...



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

- After one iteration
- Alignments, e.g., between la and the are more likely

EM for Model 1



- After another iteration
- It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)

EM for Model 1

- Convergence
- Inherent hidden structure revealed by EM

EM for Model 1

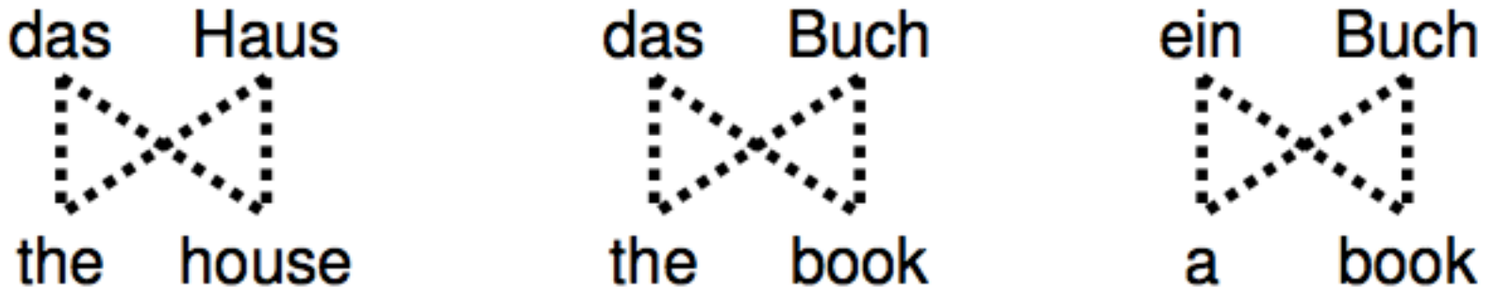
... la maison ... la maison bleu ... la fleur ...

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

p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...

- Parameter estimation from the aligned corpus

Convergence



e	f	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

Evaluation

- Since we have a probabilistic model, we can evaluate **perplexity**.

	Iter 1	Iter 2	Iter 3	Iter 4	...	Iter ∞
-log likelihood	-	7.66	7.21	6.84	...	-6
perplexity	-	2.42	2.3	2.21	...	2