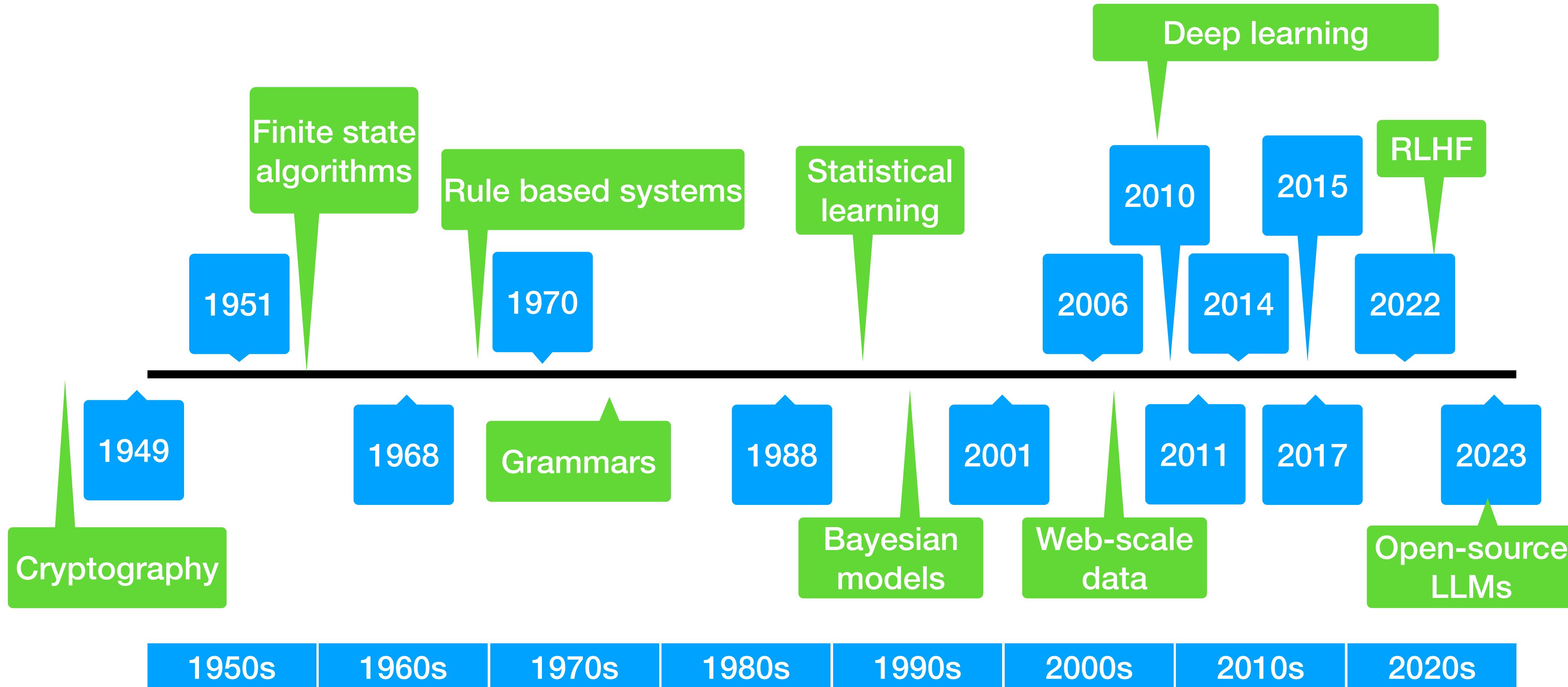




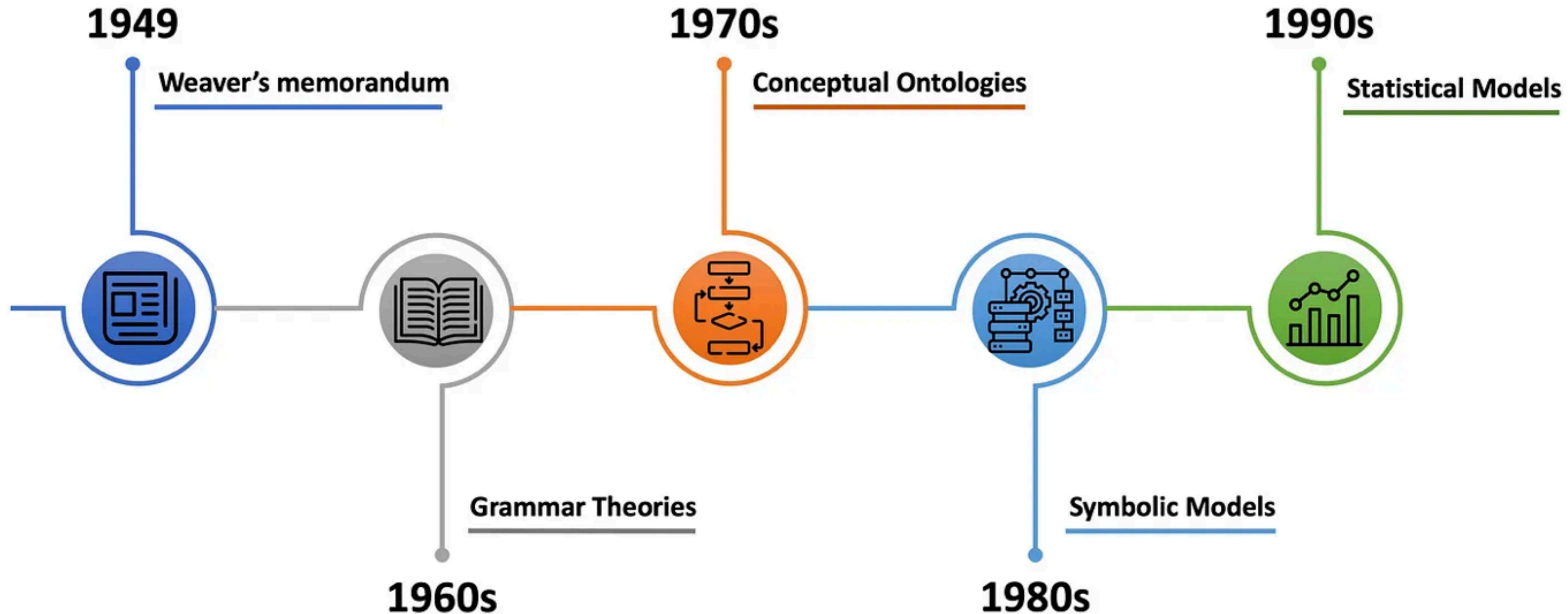
# Brief History of NLP

Natural Language Processing

Fall 2024



# NLP before deep learning



The big stages of NLP before the deep learning era.

**1948**

# Alan Turing



Instead we propose to try and see what can be done with a ‘brain’ which is more or less without a body providing, at most, organs of sight, speech, and hearing. We are then faced with the problem of finding suitable branches of thought for the machine to exercise its powers in. The following fields appear to me to have advantages:

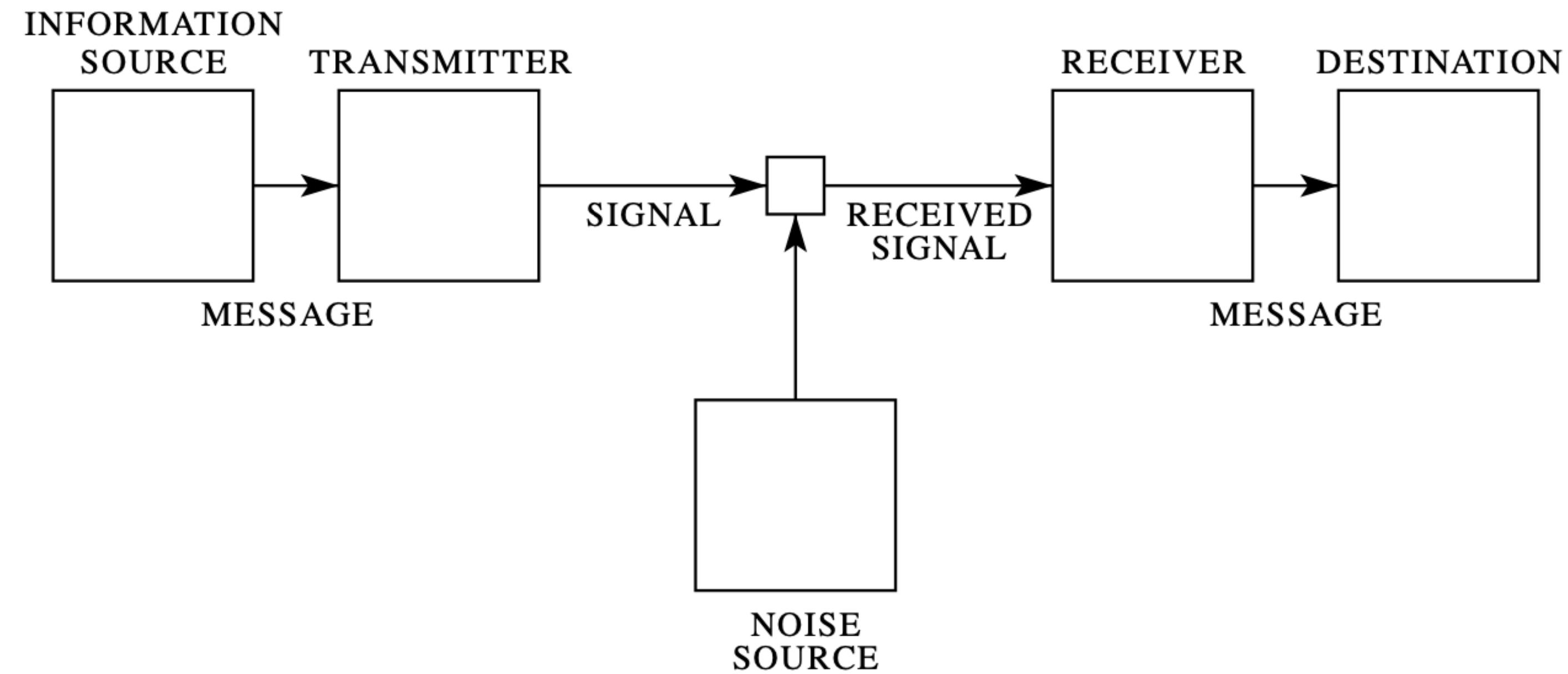
- (i) Various games, e.g., chess, noughts and crosses, bridge, poker
- (ii) The learning of languages
- (iii) Translation of languages
- (iv) Cryptography
- (v) Mathematics.

The field of cryptography will perhaps be the most rewarding.

# A Mathematical Theory of Communication

July 1948

By C. E. SHANNON



Noisy Channel Model

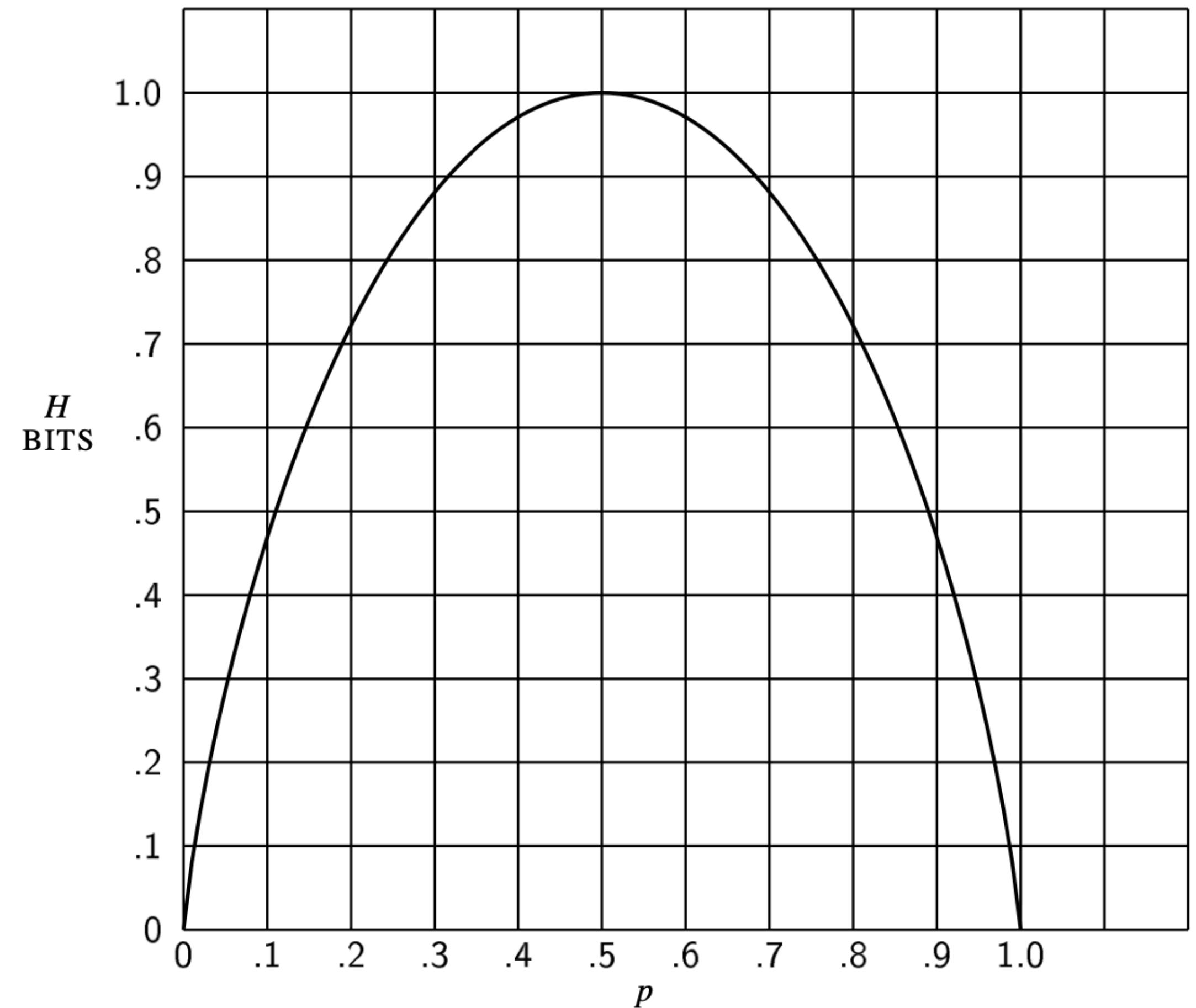


Fig. 7 — Entropy in the case of two possibilities with probabilities  $p$  and  $(1 - p)$ .

### 3. THE SERIES OF APPROXIMATIONS TO ENGLISH

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol “alphabet,” the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNESBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

<sup>5</sup>Letter, digram and trigram frequencies are given in *Secret and Urgent* by Fletcher Pratt, Blue Ribbon Books, 1939. Word frequencies are tabulated in *Relative Frequency of English Speech Sounds*, G. Dewey, Harvard University Press, 1923.

We can think of a discrete source as generating the message, symbol by symbol. It will choose successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question. A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities, is known as a stochastic process. We may consider a discrete source, therefore, to be represented by a stochastic process. Conversely, any stochastic process which produces a discrete sequence of symbols chosen from a finite set may be considered a discrete source. This will include such cases as: Natural written languages such as English, German, Chinese . . .

**1949**

Warren Weaver  
Carlsbad, New Mexico  
July 15, 1949



*Warren Weaver*

During the war a distinguished mathematician whom we will call P, an ex-German who had spent some time at the University of Istanbul and had learned Turkish there, told W.W. the following story.

Let's call him Peter

A mathematical colleague, let's call him Max, asks Peter to provide him a cipher.

Peter thinks, 'Max doesn't know that I can speak Turkish, so I'll encipher some Turkish text.' Peter reduces a sentence in Turkish into a column of five digit numbers.

Max comes back the next day and says he failed. All he could produce was some gibberish text "bu ne anlama geliyor"

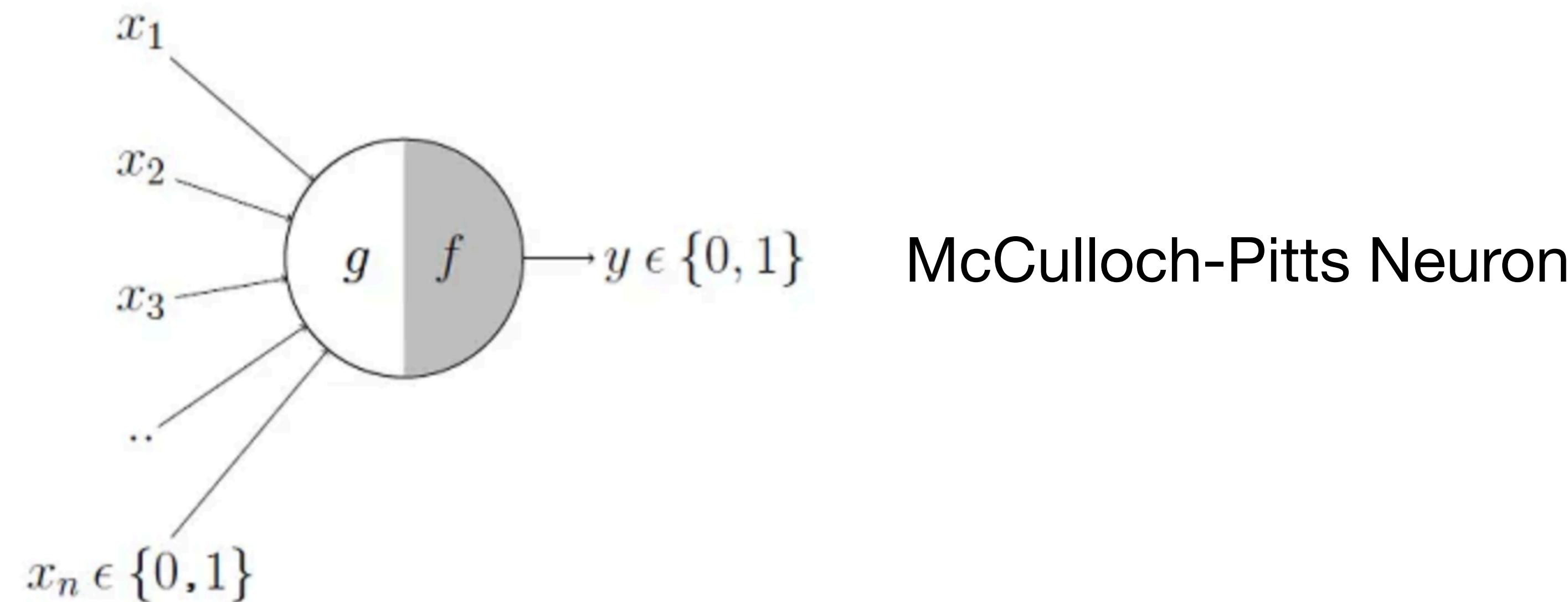
Peter was amazed that Max had produced (with some minor errors) the original message in Turkish!

"Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - 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.



Every single NLP system that produces output uses what we now call a **decoder**

A more general basis for hoping that a computer could be designed which would cope with a useful part of the problem of translation is to be found in a theorem which was proved in 1943 by McCulloch and Pitts.



This is where it all began..

**1966**

# An Application of PTAH to the Voynich Manuscript (U)

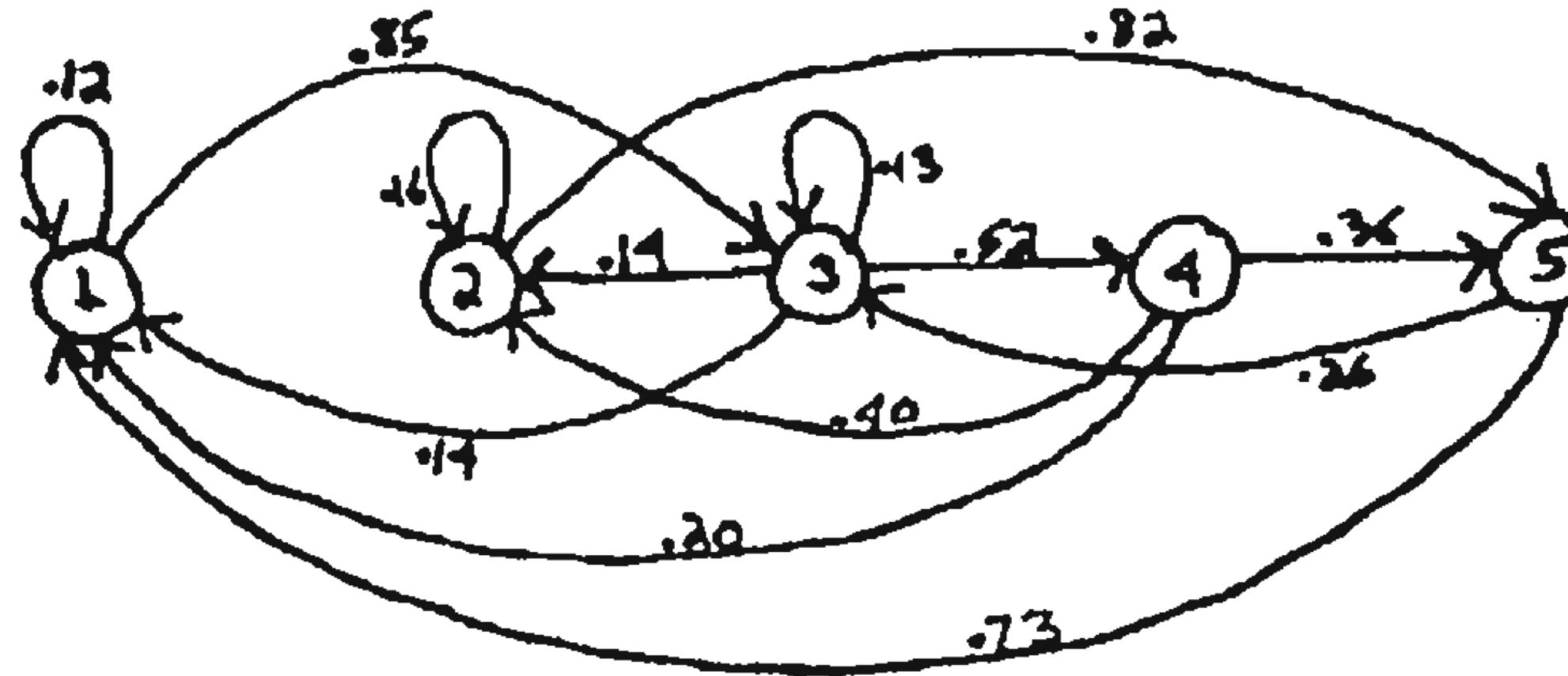
BY MARY E. D'IMPERIO

~~Top Secret Umbra~~

(U) This article is the second in a series of studies applying some modern statistical techniques to the problems posed by the Voynich manuscript. This study attempts to discover and demonstrate regularities of patterning in the Voynich text subjectively noted by many earlier students of the manuscript. Three separate PTAH studies are described, attacking the Voynich text at three levels: single symbols, whole "words," and a carefully chosen set of substrings within "words." These analyses are applied to samples of text from the "Biological B" section of the manuscript, in Currier's transcription. A brief general characterization of PTAH is provided, with an explanation of how it is used in the present application.

[https://ia601301.us.archive.org/view\\_archive.php?archive=/19/items/NSA-FOIA-Vault/TechJournals.7z&file=Tech Journals/Application\\_of\\_PTAH.pdf](https://ia601301.us.archive.org/view_archive.php?archive=/19/items/NSA-FOIA-Vault/TechJournals.7z&file=Tech Journals/Application_of_PTAH.pdf)

# Hidden Markov Models



**UNCLASSIFIED**

**State Transition Diagram (U)**

State	Associated English Letters
1	t b c j m k p v z w q
2	s y e d g
3	a o h i u
4	n r f l x
5	word space

[https://ia601301.us.archive.org/  
view\\_archive.php?archive=19/  
items/NSA-FOIA-Vault/Tech  
Journals.7z&file=Tech Journals/  
Application\\_of\\_PTAH.pdf](https://ia601301.us.archive.org/view_archive.php?archive=19/items/NSA-FOIA-Vault/Tech Journals.7z&file=Tech Journals/Application_of_PTAH.pdf)

## REFERENCES

- [1] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 37, pp. 1554–1563, 1966.
- [2] L. E. Baum and J. A. Egon, "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology," *Bull. Amer. Meteorol. Soc.*, vol. 73, pp. 360–363, 1967.
- [3] L. E. Baum and G. R. Sell, "Growth functions for transformations on manifolds," *Pac. J. Math.*, vol. 27, no. 2, pp. 211–227, 1968.
- [4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164–171, 1970.
- [5] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.
- [6] J. K. Baker, "The dragon system—An overview," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-23, no. 1, pp. 24–29, Feb. 1975.
- [7] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM J. Res. Develop.*, vol. 13, pp. 675–685, 1969.
- [8] L. R. Bahl and F. Jelinek, "Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition," *IEEE Trans. Informat. Theory*, vol. IT-21, pp. 404–411, 1975.
- [9] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Informat. Theory*, vol. IT-21, pp. 250–256, 1975.
- [10] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–536, Apr. 1976.

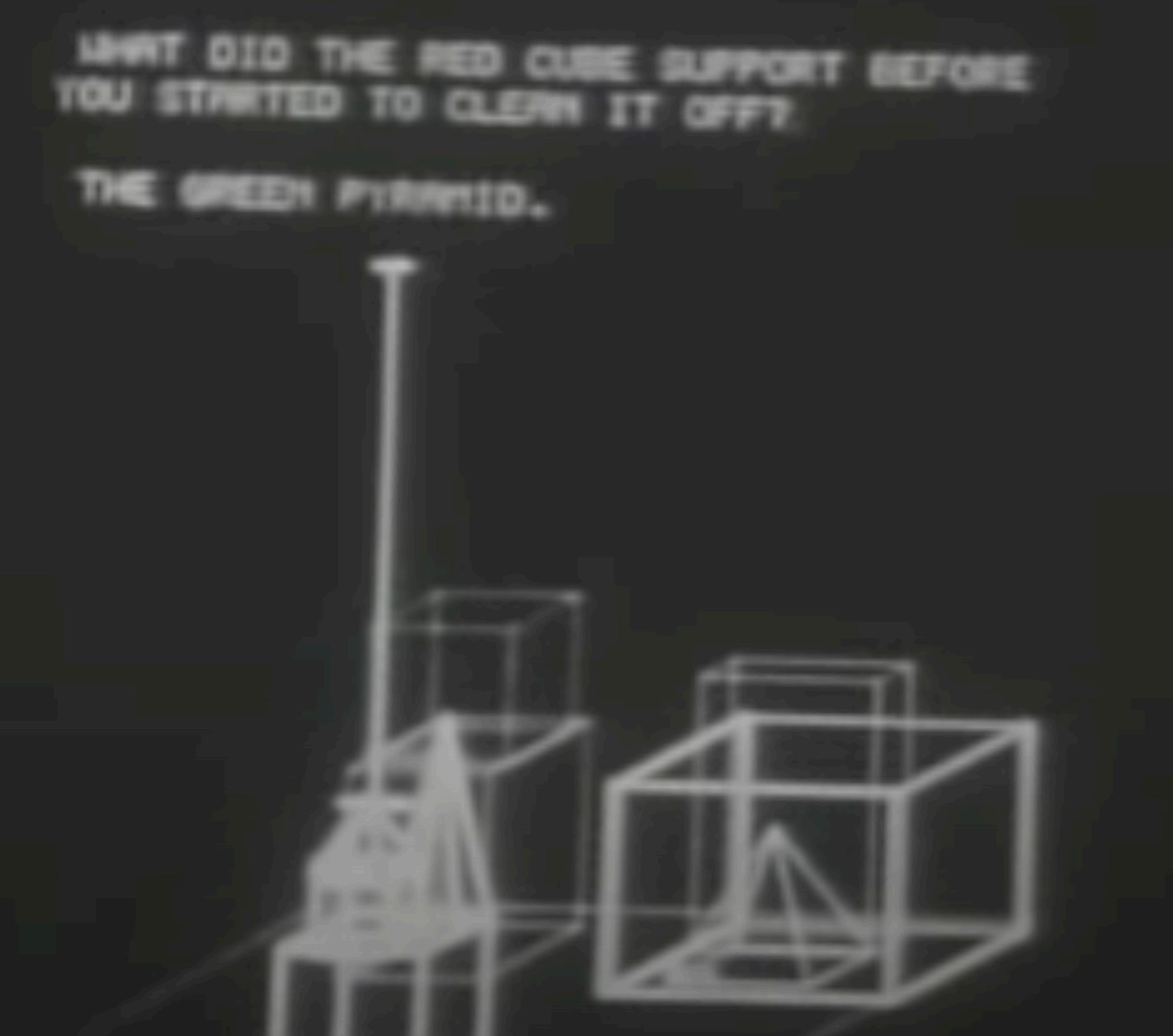
Lawrence R. Rabiner (February 1989). "[A tutorial on Hidden Markov Models and selected applications in speech recognition](#)". Proceedings of the IEEE. 77 (2): 257–286.

**1968**

# SHRDLU

Terry Winograd

Rule-based NLP  
(grounding)



<https://www.youtube.com/watch?v=QAJz4YKUwqw>

Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.

Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Computer: OK.

Person: What does the box contain?

Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.

Person: What is the pyramid supported by?

Computer: THE BOX.

Person: How many blocks are not in the box?

Computer: FOUR OF THEM.

Person: Is at least one of them narrower than the one which i told you to pick up?

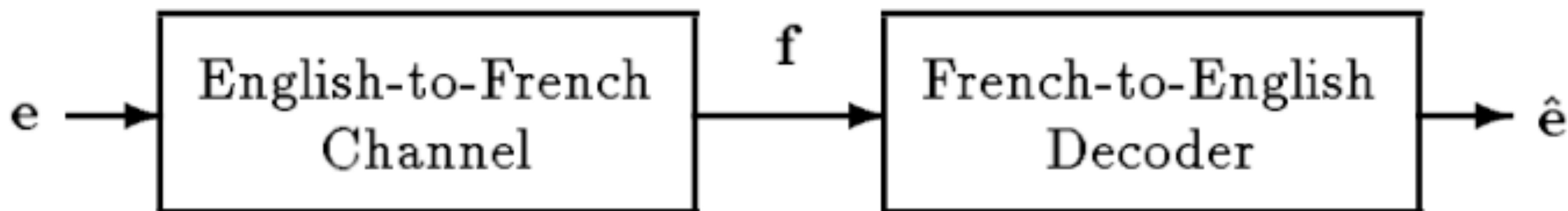
Computer: YES, THE RED CUBE.

**1988**

# The Candide System for Machine Translation

*Adam L. Berger, Peter F. Brown\*, Stephen A. Della Pietra, Vincent J. Della Pietra,  
John R. Gillett, John D. Lafferty, Robert L. Mercer\*, Harry Printz, Luboš Ureš*

IBM Thomas J. Watson Research Center  
P.O. Box 704  
Yorktown Heights, NY 10598



$$\hat{e} = \operatorname{argmax}_e \Pr(e | f) = \operatorname{argmax}_e \Pr(f | e) \Pr(e)$$

Bayes Rule

**2001**

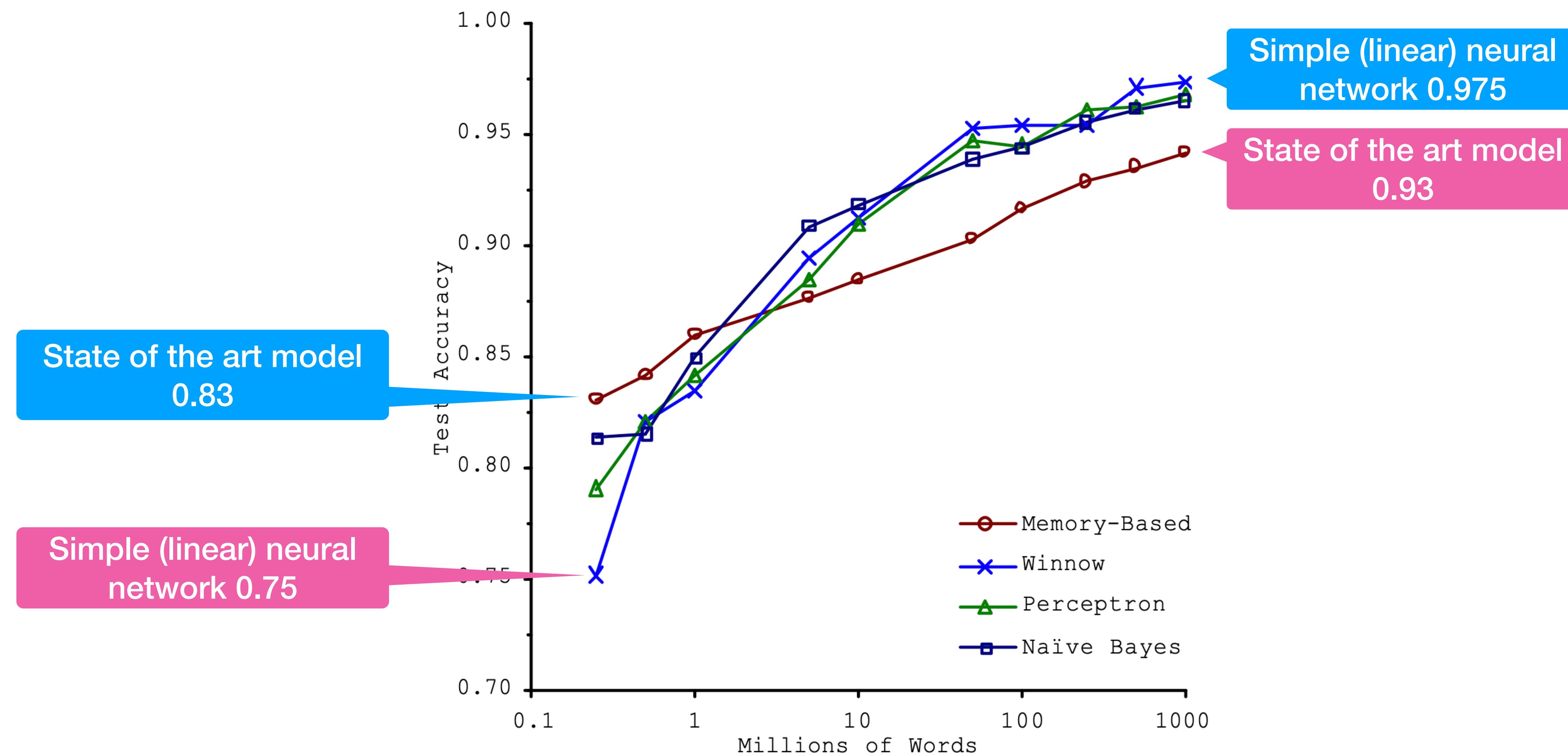
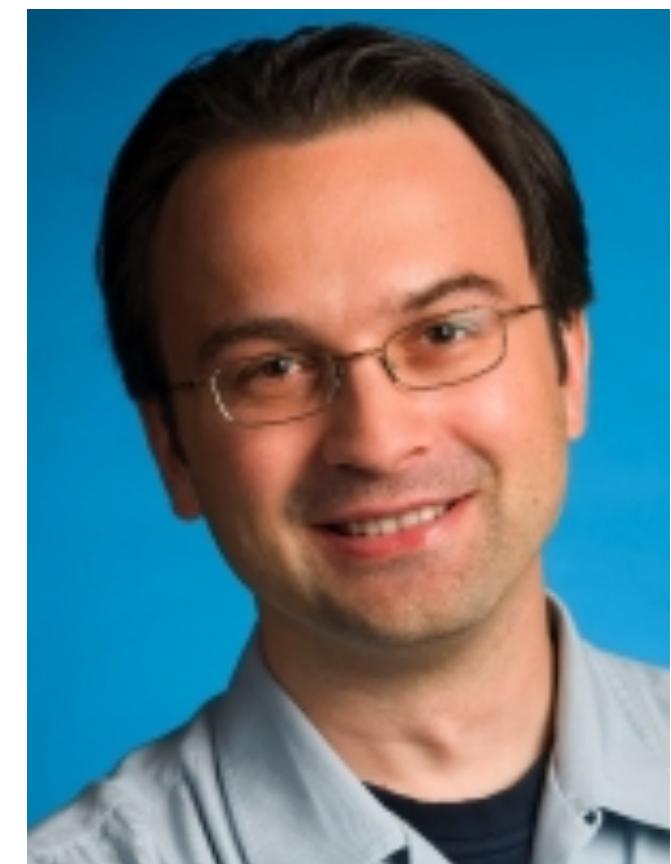


Figure 1. Learning Curves for Confusion Set Disambiguation

Scaling to very very large corpora for natural language disambiguation.  
Banko and Brill, 2001

**2006**



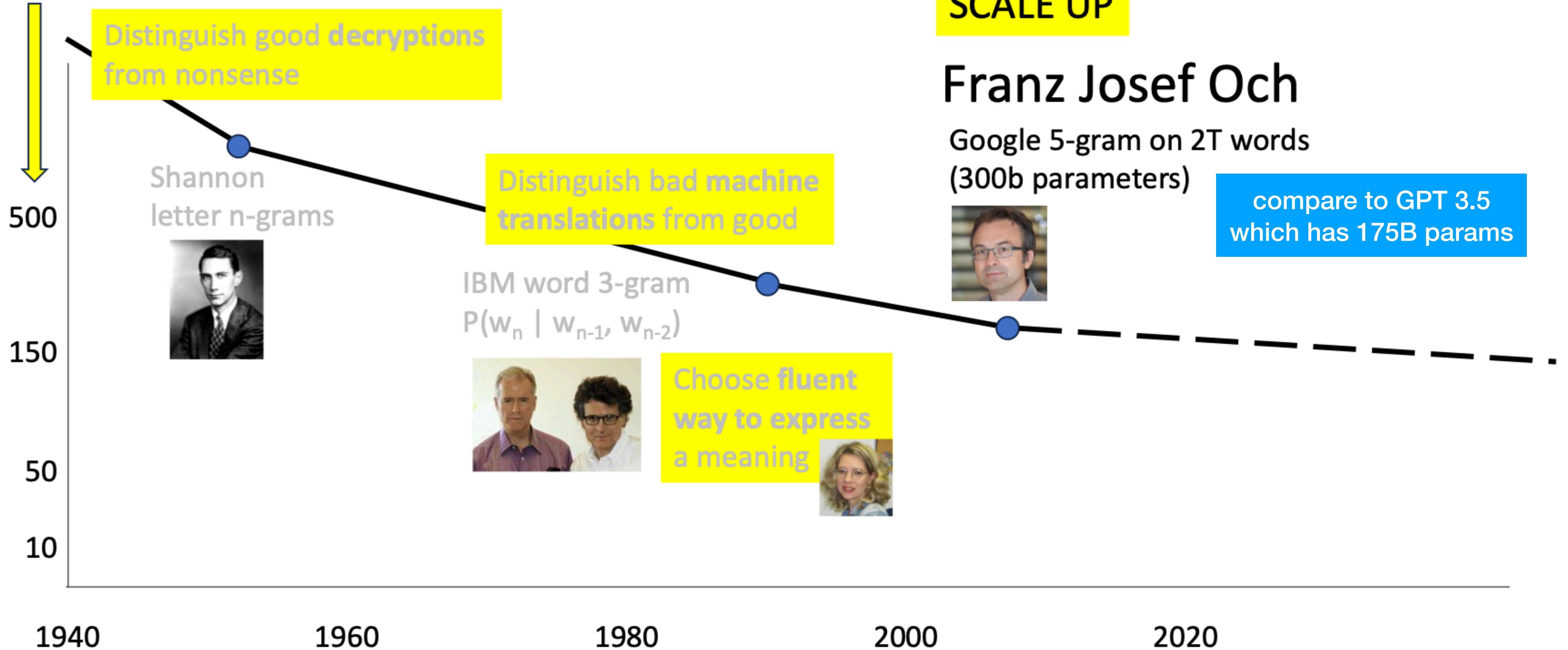
Franz Och, Apr 28  
2006

Most state-of-the-art commercial machine translation systems in use today have been developed using a rules-based approach ...

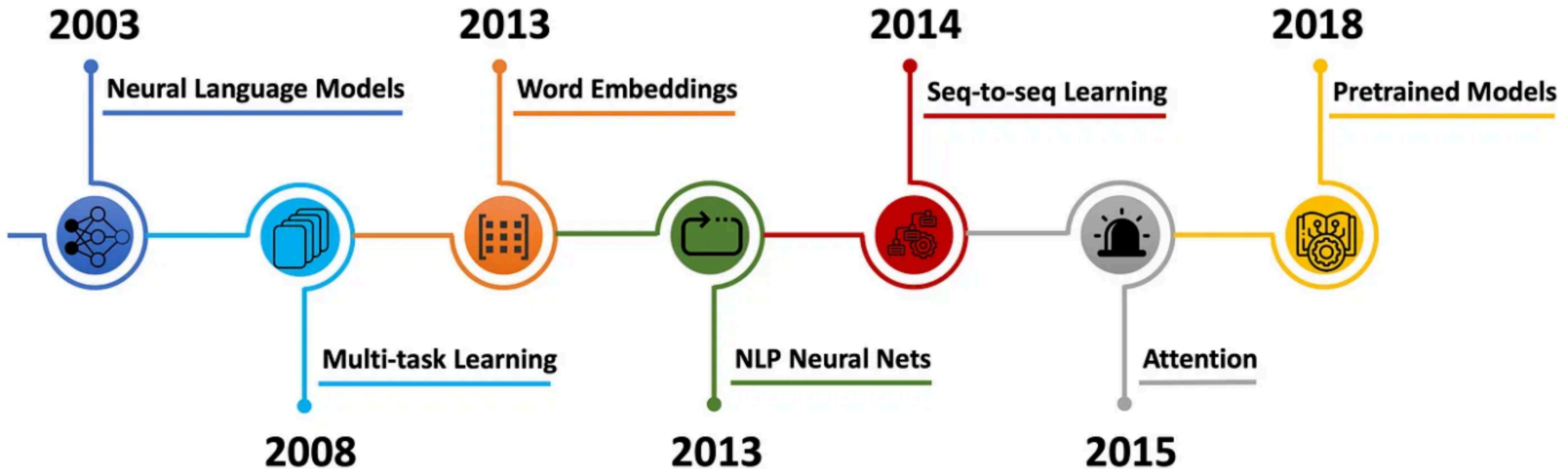
Several research systems, including ours, take a different approach: **we feed the computer with billions of words of text**, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages.

**We then apply statistical learning techniques to build a translation model.**

Perplexity (ability to predict the next word)



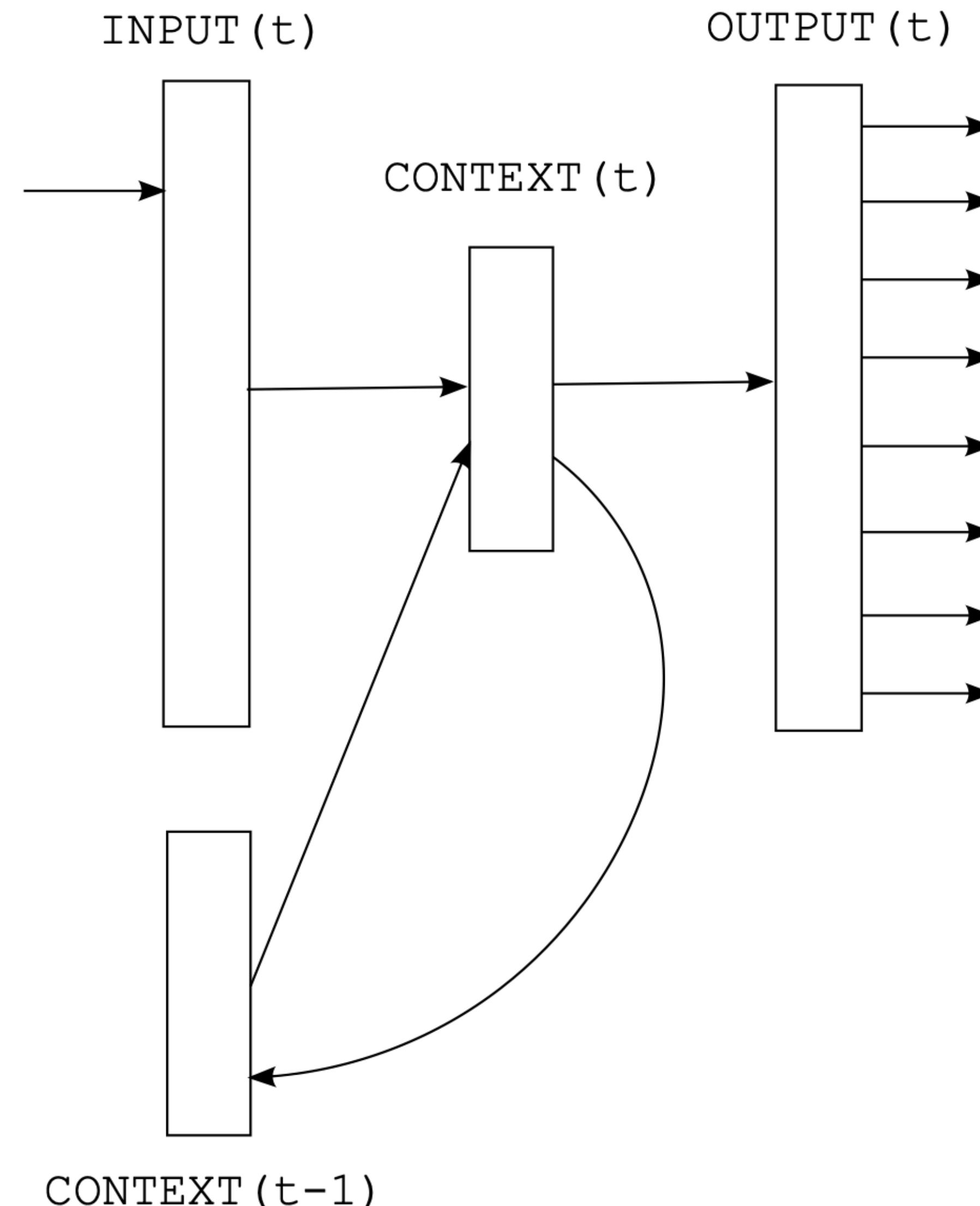
# NLP in the deep learning era



The big stages of NLP in the deep learning era.

**2010**

# Mikolov, Recurrent Networks for Language Models (RNN-LM)



**2011**

# IBM Watson plays Jeopardy!



and wins ...

**2011**

# Natural Language Processing (Almost) from Scratch

**Ronan Collobert\***

RONAN@COLLOBERT.COM

**Jason Weston<sup>†</sup>**

JWESTON@GOOGLE.COM

**Léon Bottou<sup>‡</sup>**

LEON@BOTTOU.ORG

**Michael Karlen**

MICHAEL.KARLEN@GMAIL.COM

**Koray Kavukcuoglu<sup>§</sup>**

KORAY@CS.NYU.EDU

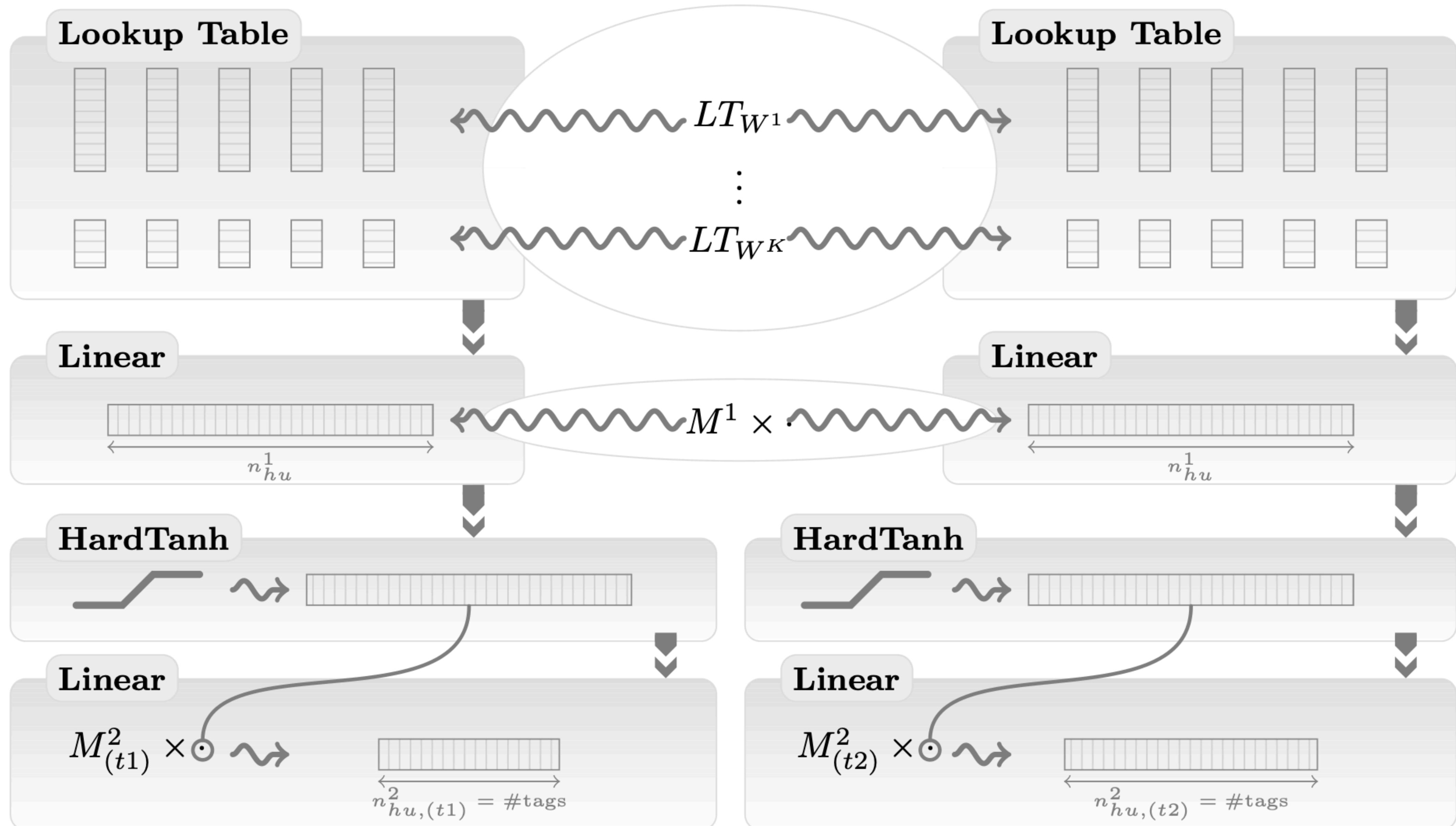
**Pavel Kuksa<sup>¶</sup>**

PKUKSA@CS.RUTGERS.EDU

*NEC Laboratories America*

*4 Independence Way*

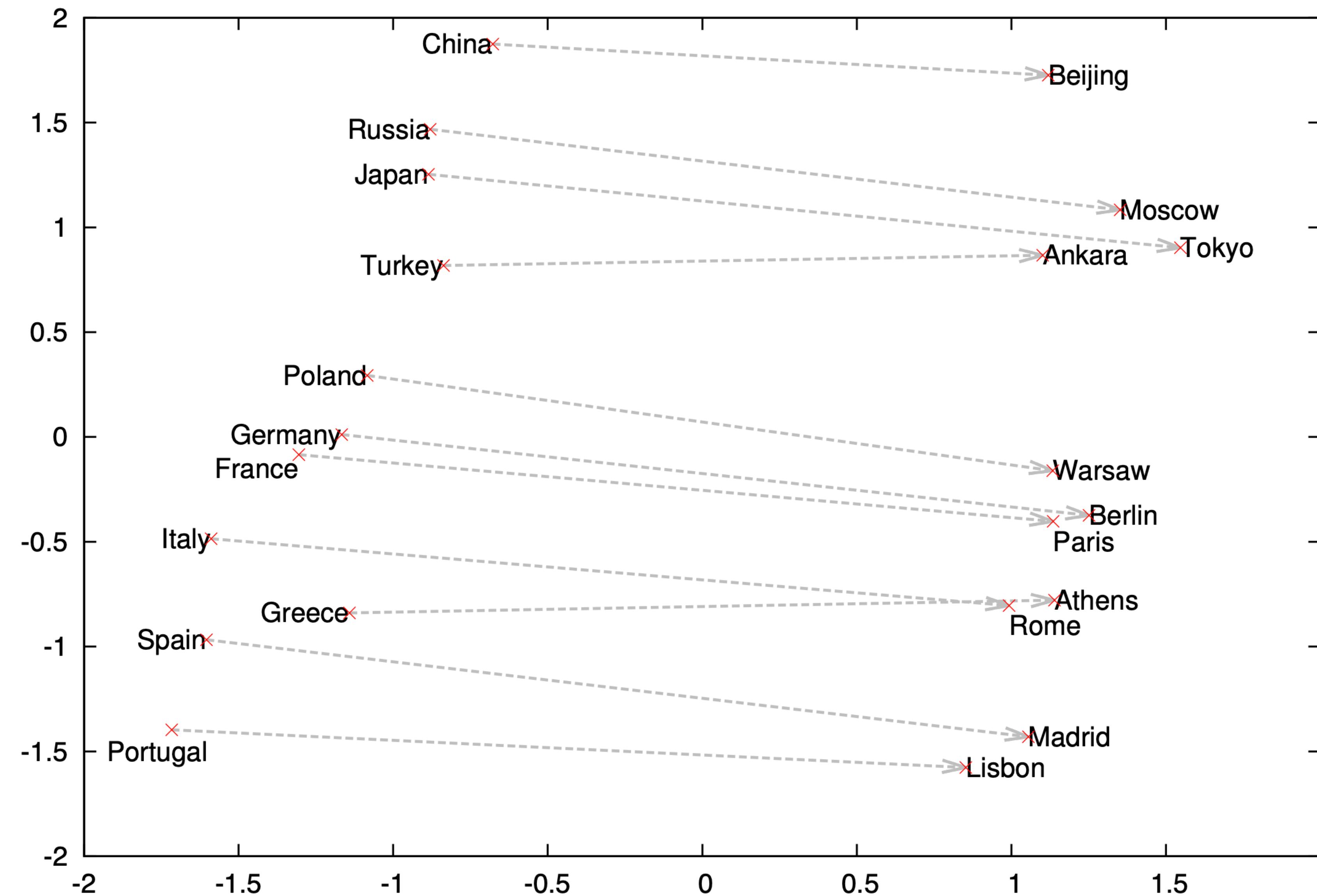
*Princeton, NJ 08540*



**2012**

Mikolov,  
Continuous  
Bag of Words  
(word2vec)

## Country and Capital Vectors Projected by PCA



**2014**

“You can’t cram the meaning of a whole  
sentence into a single vector!”

<https://yoavartzi.com/sp14/slides/mooney.sp14.pdf>

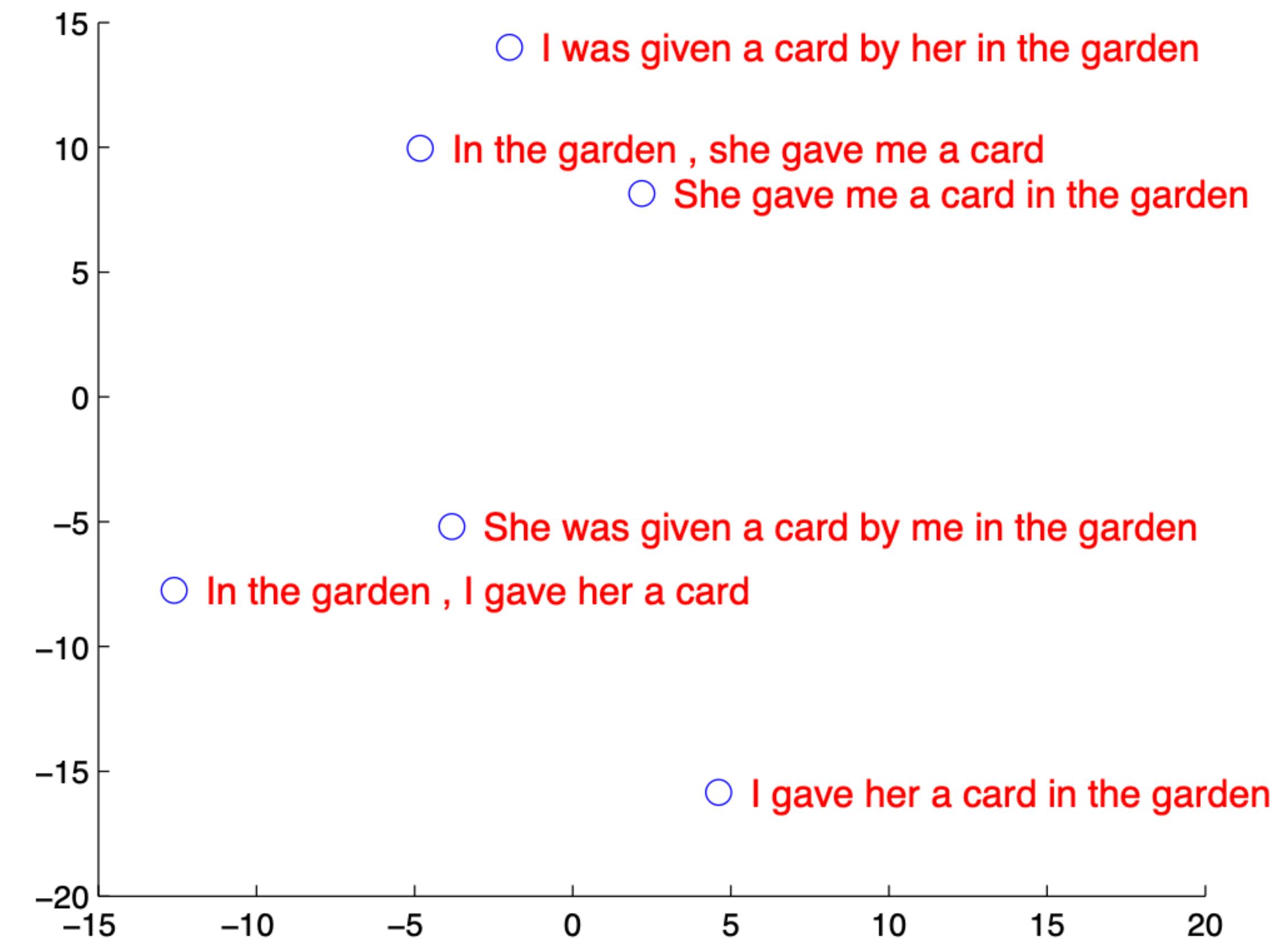
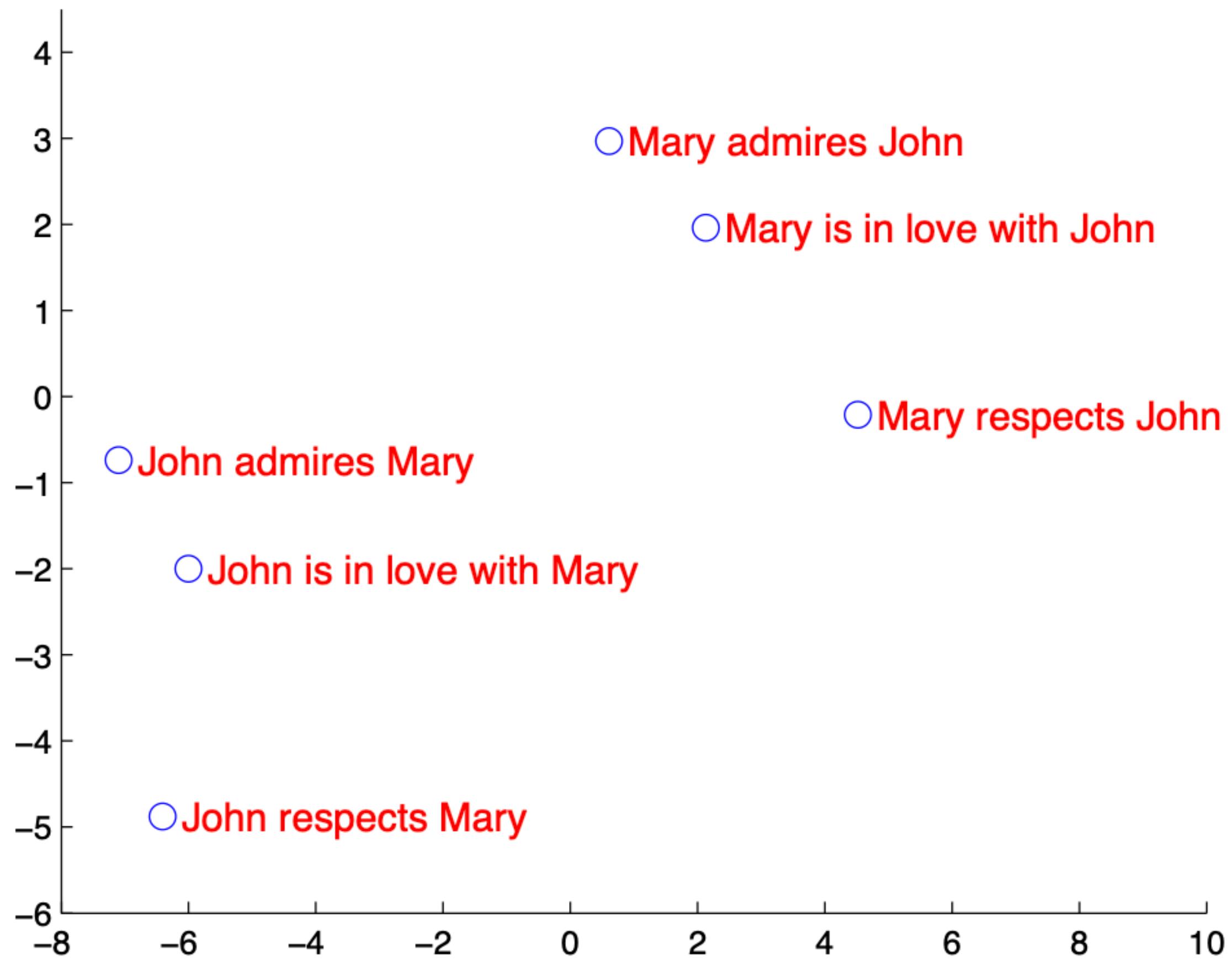


**Ray Mooney**

Invited talk at the ACL 2014 Workshop on  
Semantic Parsing

**2014**

# Recurrent Neural Networks for Translation

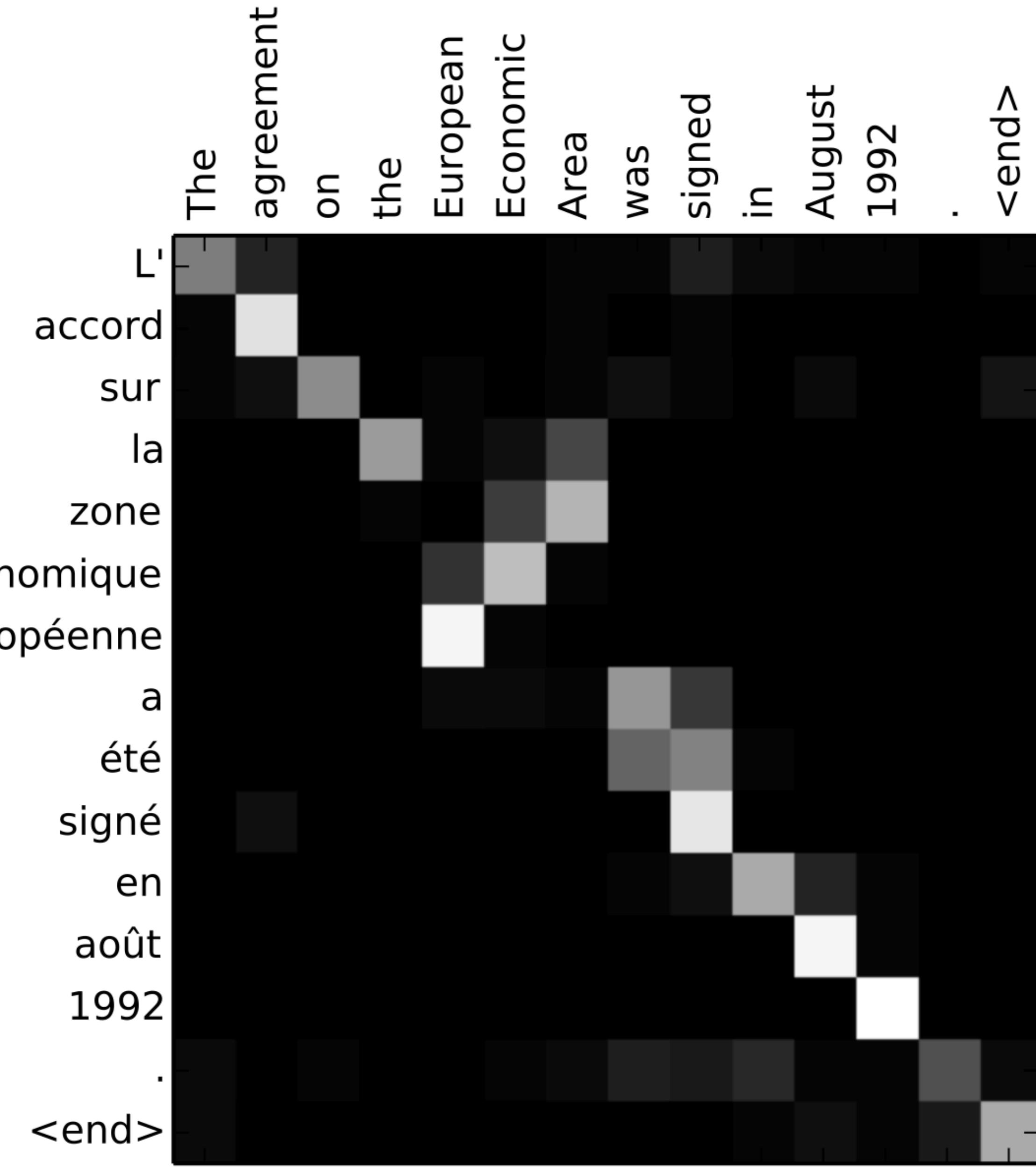


**2015**

# Skype Translator



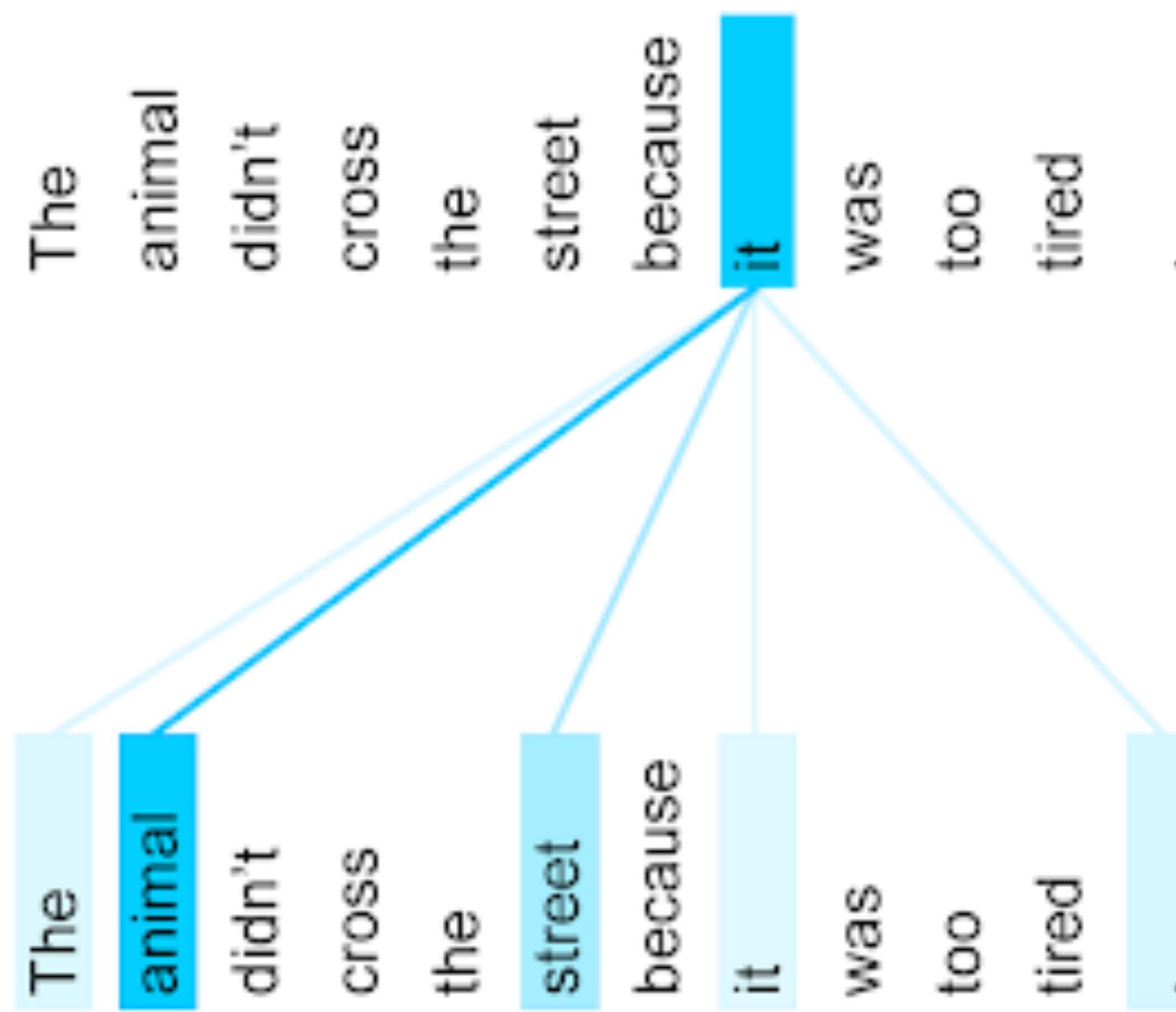
# Attention



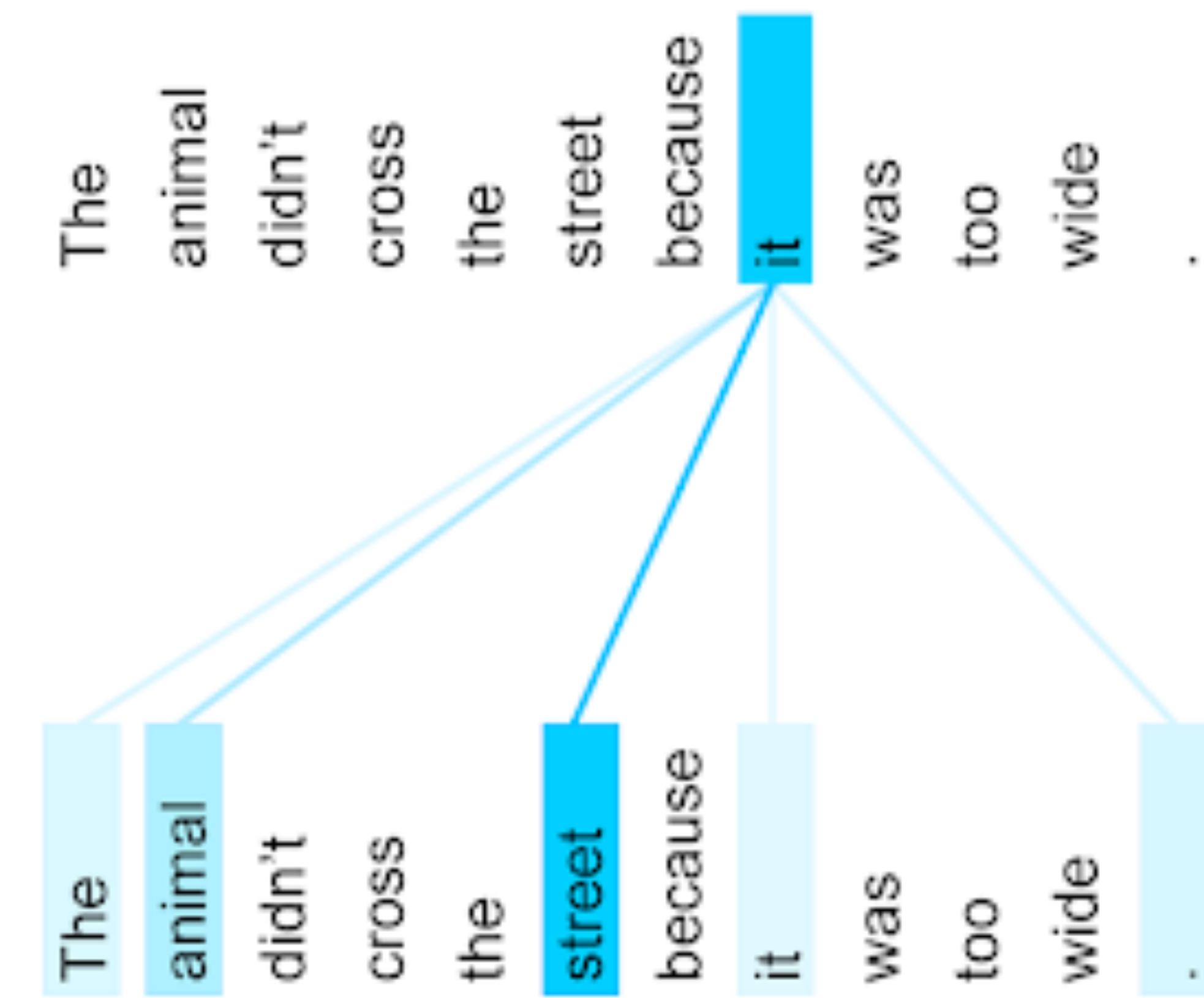
**2017**

# Self-Attention

The animal didn't cross the street because it was too tired .



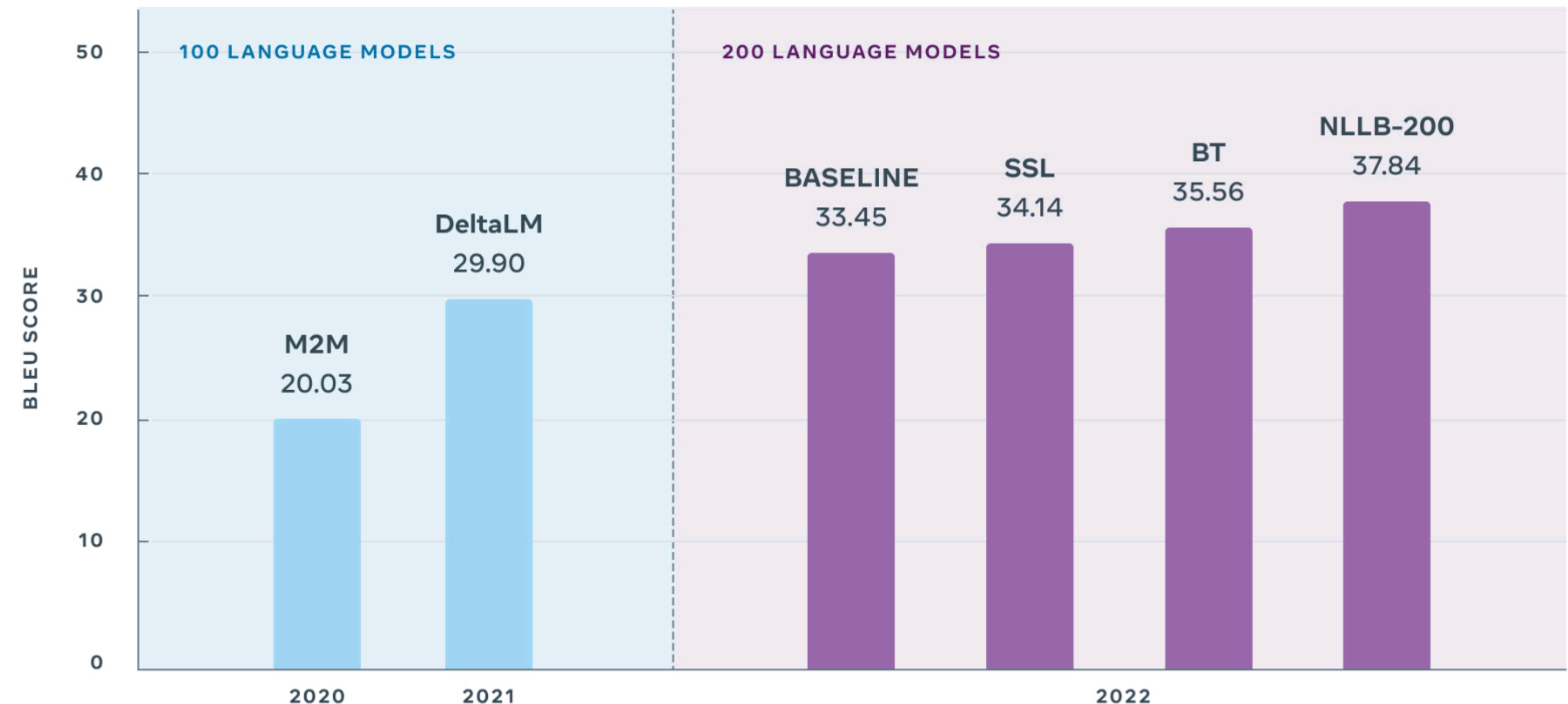
The animal didn't cross the street because it was too wide .



**2022**

# Comparison of NLLB-200 with existing SOTA

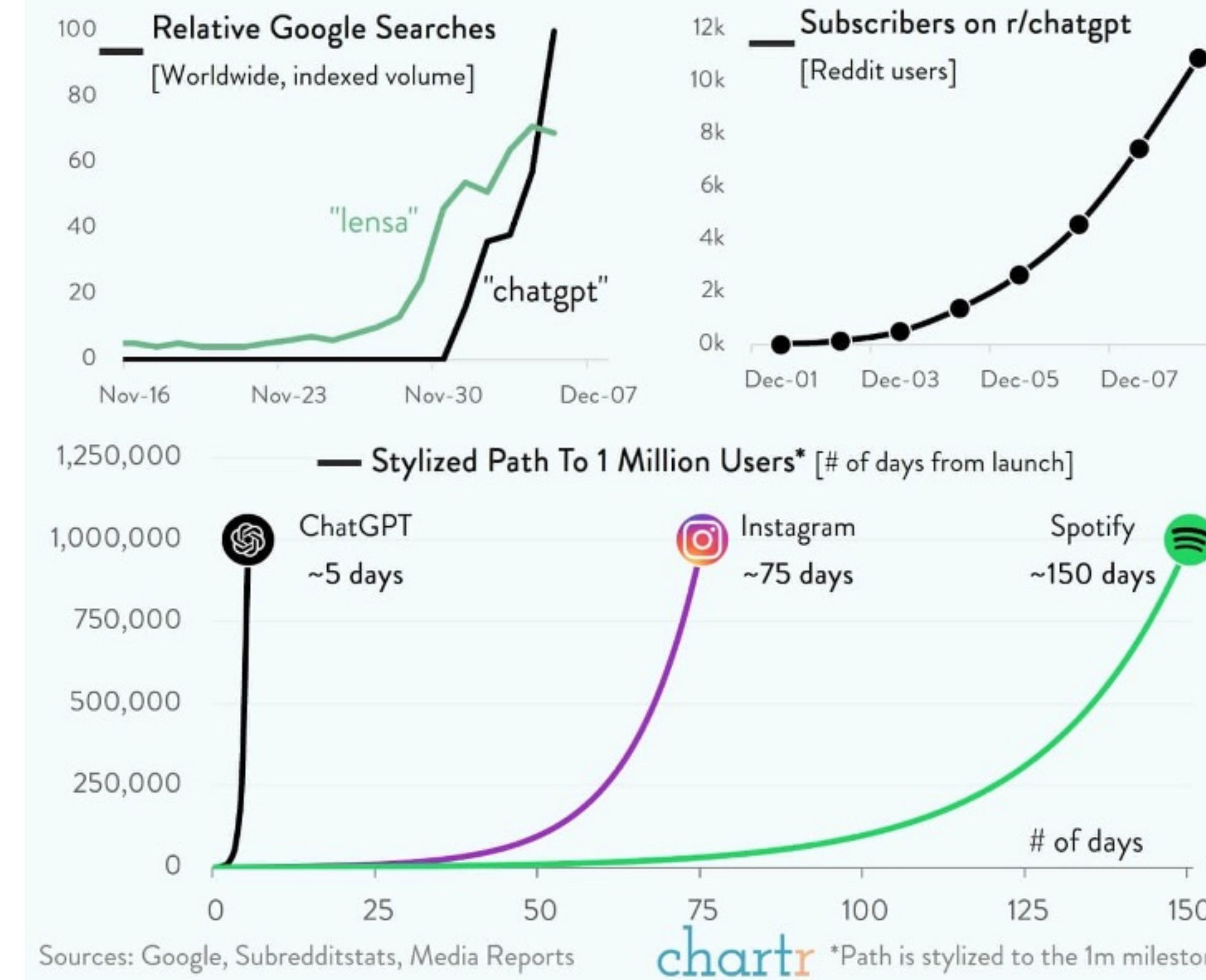
High-quality  
machine  
translation for 200  
languages



This graphic shows average BLEU score on FLORES-101 translations to and from English into 100 languages. On the left there are two published state-of-the-art models, M2M and Delta LM, that support 100 languages. Models on the right support 200 languages: A baseline Transformer model with 3.3B parameters, the baseline model with self-supervised learning (SSL), the baseline model with back translation (BT), and NLLB-200, a large mixture-of-experts based model that leverages both self-supervised learning and back translation.

**Nov 2022**

# ChatGPT From OpenAI Is A Bot Taking The Tech World By Storm

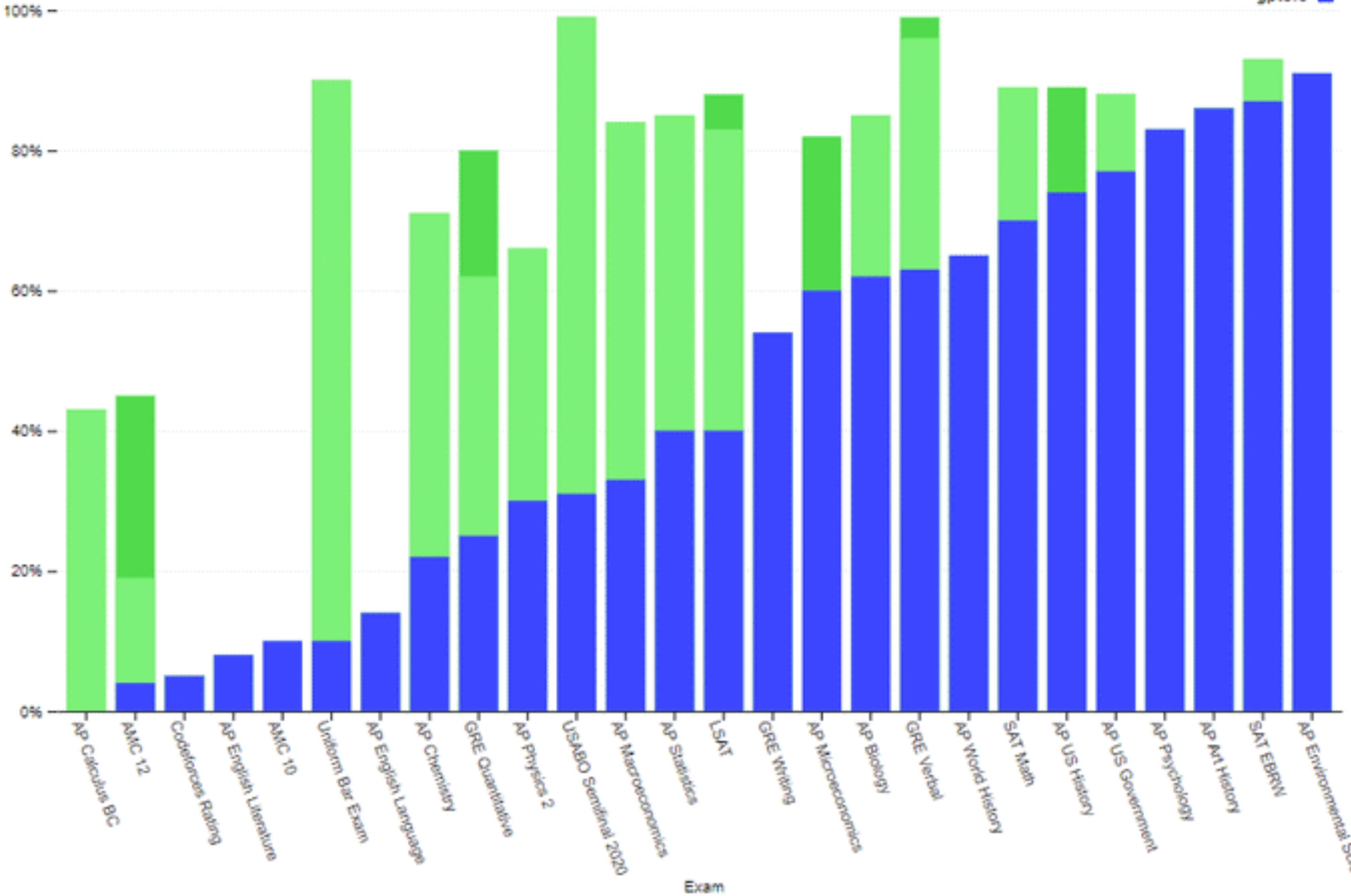


Source: Chartr

## Exam results (ordered by GPT-3.5 performance)

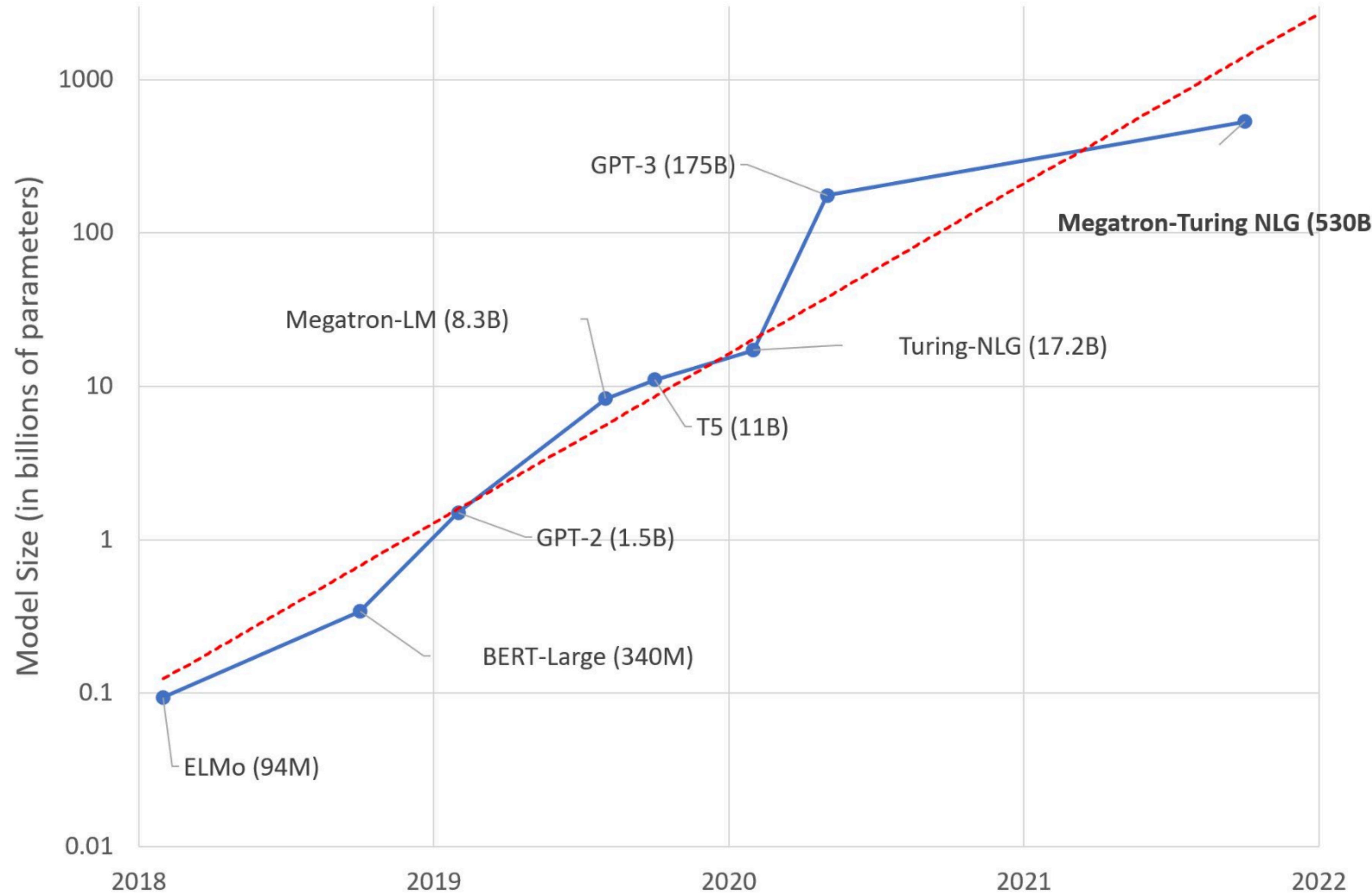
Estimated percentile lower bound (among test takers)

gpt-4 (no vision)  
gpt-4  
gpt3.5



**2023**

# Self-Supervised Pre-trained Language Models



GPT-4 is estimated to have **500B-1000B** parameters; trained using **2.2e25 FLOPs** (estimated)

LOOKS LIKE COMPUTERS  
WILL BEAT HUMANS AT  
GO PRETTY SOON.

WOW.  
THAT'S THE LAST  
OF THE BIG ONES.  
YEAH.



WELL, AT LEAST HUMANS  
ARE STILL BETTER AT, UH,  
COMING UP WITH REASSURING  
PARABLES ABOUT THINGS  
HUMANS ARE BETTER AT?



I MADE A PYTHON SCRIPT  
THAT GENERATES THOUSANDS  
OF REASSURING PARABLES  
PER SECOND.



```
from transformers import pipeline
access_token=os.environ.get("HF_ACCESS_TOKEN")
model_id = "meta-llama/Meta-Llama-3.1-8B-Instruct"
pipe = pipeline(
    "text-generation",
    device=device,
    model=model_id,
    model_kwargs={
        "torch_dtype": torch.bfloat16,
    },
    token=access_token,
)

messages = [
{
    "role": "user",
    "content": 'Finish this reassuring parable: Computers will never ____.'
},
]
for _ in range(1000):
    outputs = pipe(
        messages,
        temperature=0.8,
        max_new_tokens=128,
        do_sample=True,
    )
    print(outputs[0]["generated_text"][-1]["content"])
```

**1951**

# The Turing Test

NEWMAN: I should like to be there when your match between a man and a machine takes place, and perhaps to try my hand at making up some of the questions. But that will be a long time from now, if the machine is to stand any chance with no questions barred?

TURING: Oh yes, at least 100 years, I should say.

'Can digital computers think?'. Interview with Alan Turing.  
BBC Third Programme, 15 May 1951.

- Material for some of these slides comes from ideas borrowed from Kevin Knight, Angel Chang, Danqi Chen, Karthik Narasimhan and others on the Slack group for teaching NLP