89688: Statistical Machine Translation

March 2020

Roee Aharoni
Computer Science Department
Bar Ilan University

👽 vs. 👾

 **vs.** 

```
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 .
```

👽 vs. 👾

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 .

👽 vs. 👾

👽 vs. 👾

Translation dictionary:

anok - pippat
erok - total
ghirok - hilat
hihok - arrat
izok - vat
ok-drubel - at-drubel
ok-yurp - at-yurp
ok-voon - at-voon
ororok - bichat
plok - rrat
sprok - dat
zanzanok - zanzanat

Translation Model

👽 vs. 👾

Translation dictionary:

anok - pippat
erok - total
ghirok - hilat
hihok - arrat
izok - vat
ok-drubel - at-drubel
ok-yurp - at-yurp
ok-voon - at-voon
ororok - bichat
plok - rrat
sprok - dat
zanzanok - zanzanat

Translation Model

Word pair counts:

1 . erok
7 . lalok
2 . ok-drubel
2 . ok-voon
3 . wiwok
1 anok drok
1 anok ghirok
1 hihok yorok
1 izok enemok
2 izok hihok
1 izok jok
1 izok kantok
1 izok stok
1 izok vok

Language Model

Why am I teaching this?

Why am I teaching this?

Why am I teaching this?

- Military Service (2008-2015, statistical machine translation)

Why am I teaching this?

- Military Service (2008-2015, statistical machine translation)
- CS Masters @ Bar Ilan (2011-2015, machine translation evaluation)

Why am I teaching this?

- Military Service (2008-2015, statistical machine translation)

- CS Masters @ Bar Ilan (2011-2015, machine translation evaluation)

- CS Phd @ Bar Ilan (2016-2020, neural machine translation)

Why am I teaching this?

- Military Service (2008-2015, statistical machine translation)
- CS Masters @ Bar Ilan (2011-2015, machine translation evaluation)
- CS Phd @ Bar Ilan (2016-2020, neural machine translation)
- Google (2018-present, multilingual machine translation, domain adaptation)

Why am I teaching this?

- Military Service (2008-2015, statistical machine translation)
- CS Masters @ Bar Ilan (2011-2015, machine translation evaluation)
- CS Phd @ Bar Ilan (2016-2020, neural machine translation)
- Google (2018-present, multilingual machine translation, domain adaptation)
- Let’s collaborate (after the course)!

Introduction

Introduction

Introduction

- Name?

Introduction

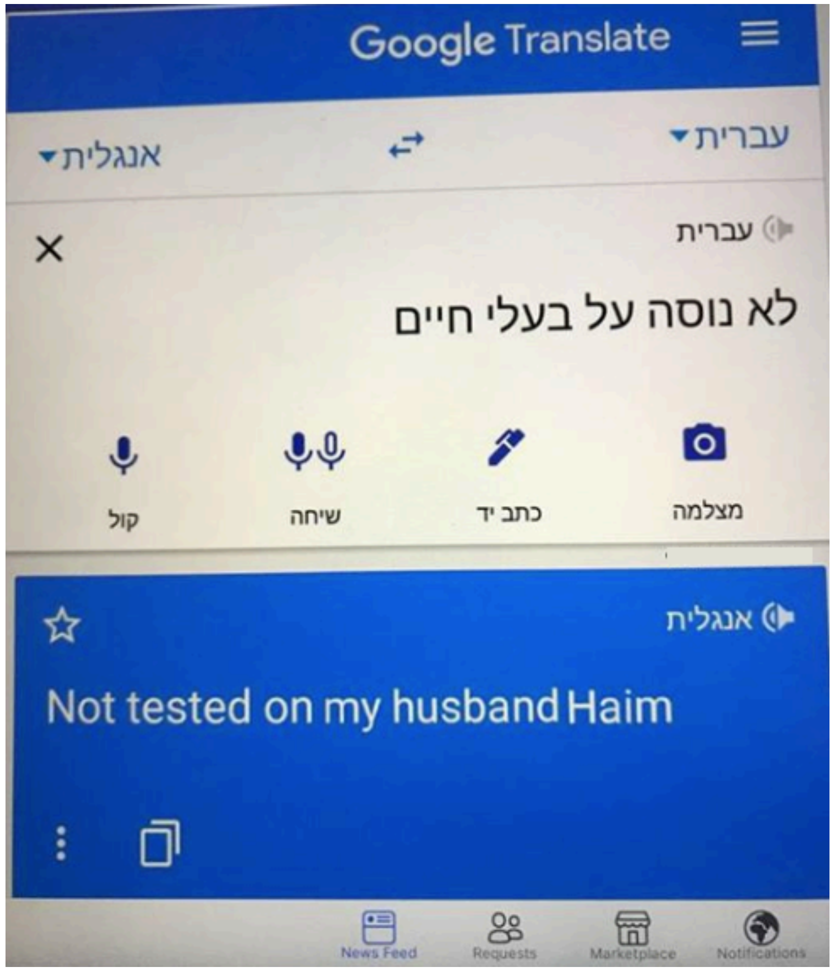
- Name?
- Background?

Introduction

- Name?
- Background?
- Which languages do you speak?

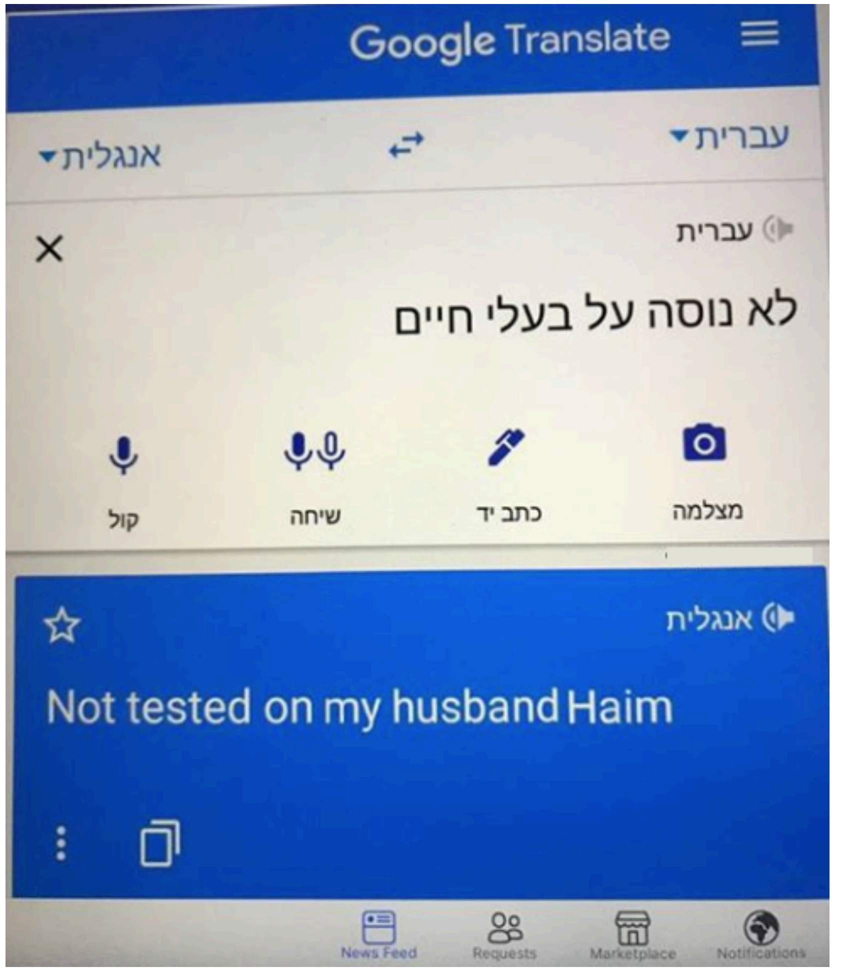
About the Course

About the Course



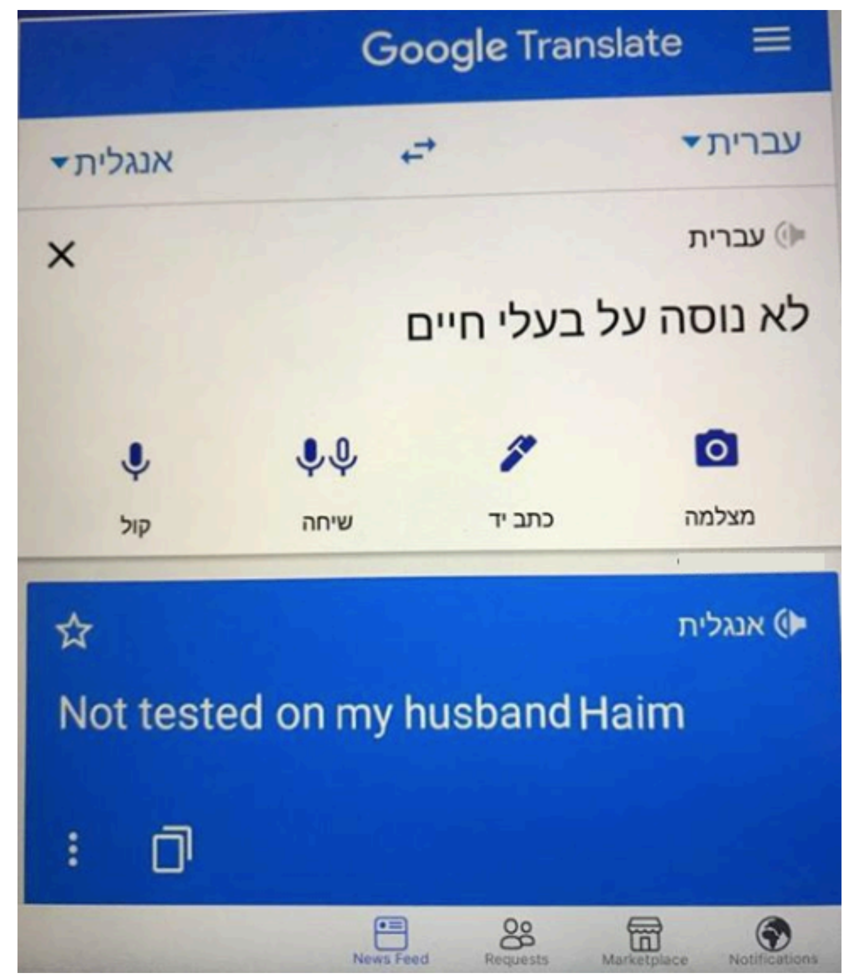
About the Course

- **Experimental** 🧪



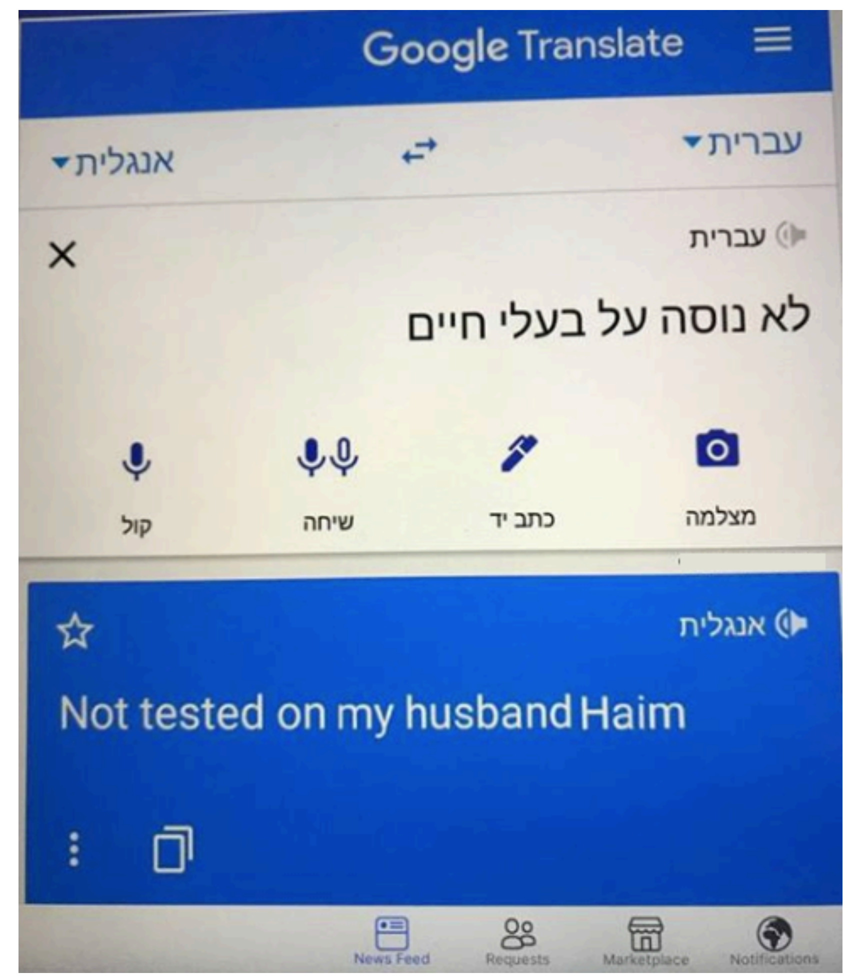
About the Course

- **Experimental** 🧪
- New **syllabus** 📖



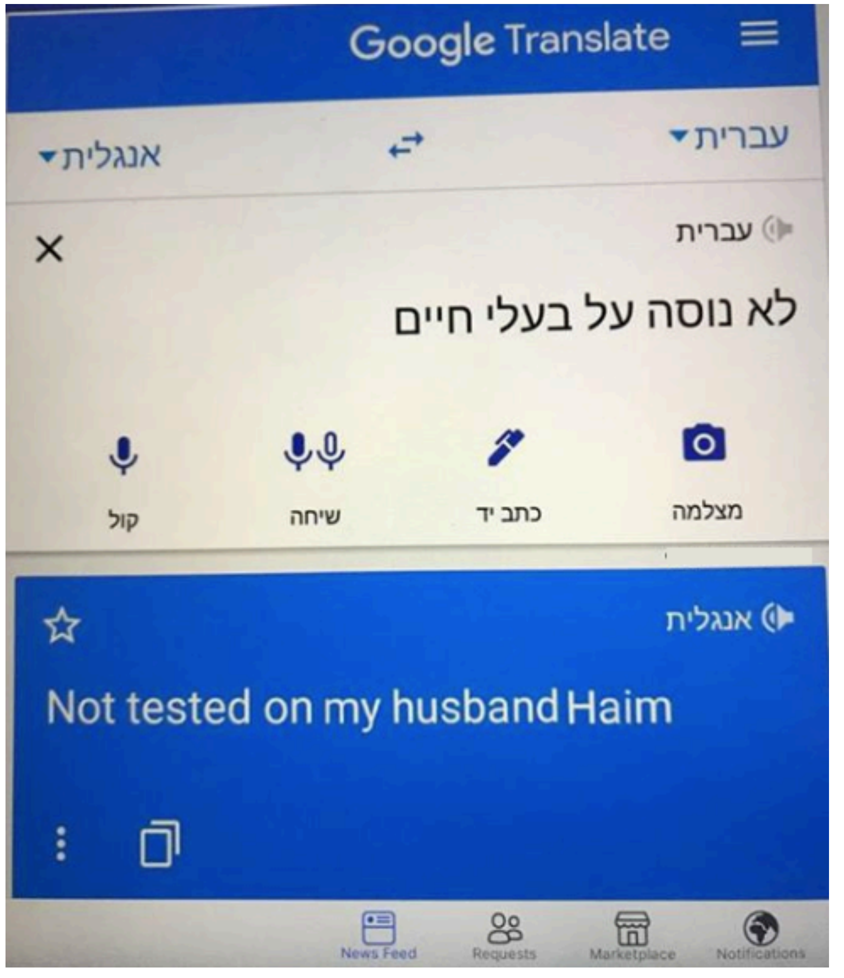
About the Course

- **Experimental** 🧪
- New **syllabus** 📖
- New **format** 💻



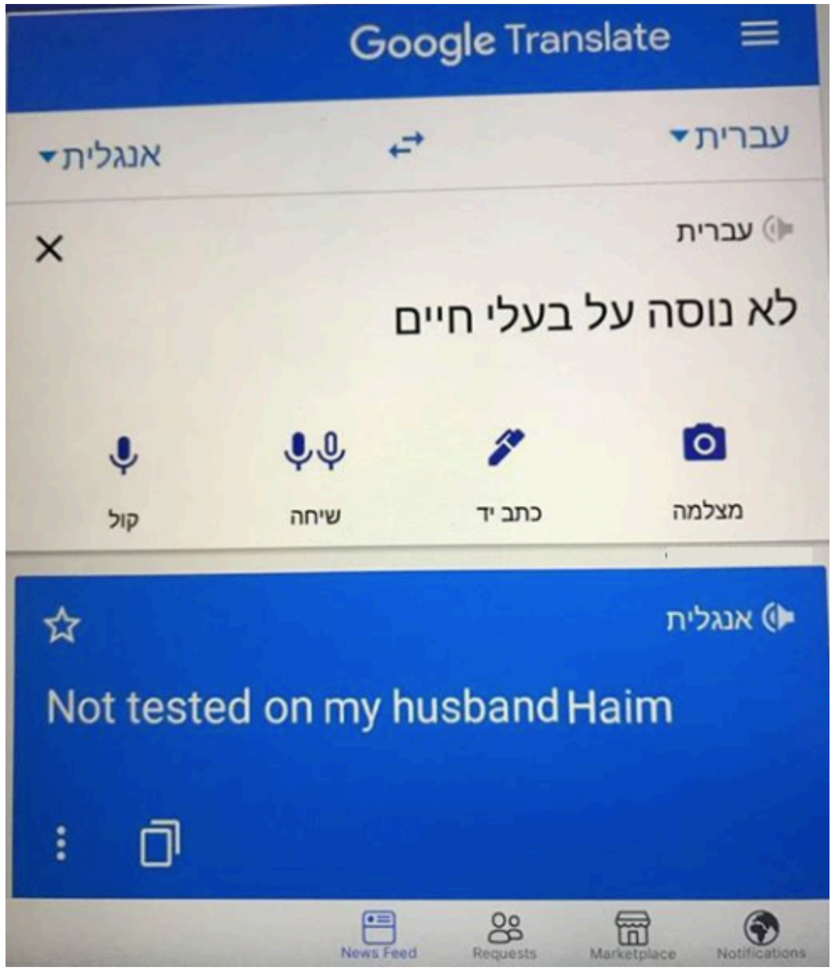
About the Course

- **Experimental** 🧪
- New **syllabus** 📖
- New **format** 💻
- New **lecturer** 😊



About the Course

- **Experimental** 🧪
- New **syllabus** 📖
- New **format** 💻
- New **lecturer** 😊
- **Feedback!** 👍👎



Course Objectives

Course Objectives

- **Understand and describe** state-of-the-art models and algorithms for machine translation.

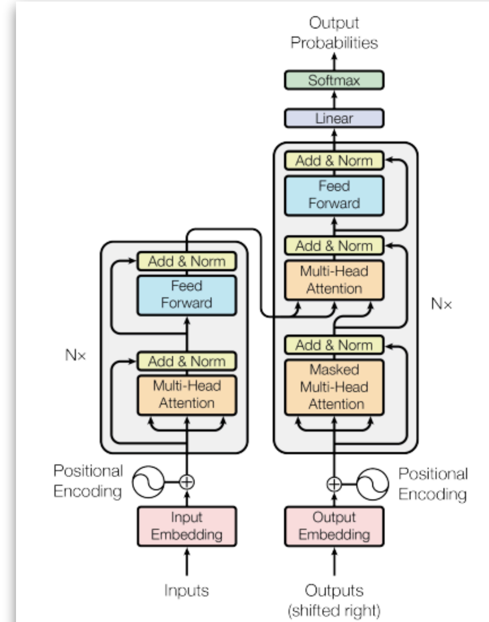


Figure 1: The Transformer - model architecture.

Course Objectives

- **Understand and describe** state-of-the-art models and algorithms for machine translation.
- **Implement and apply** such methods using real-world tasks.

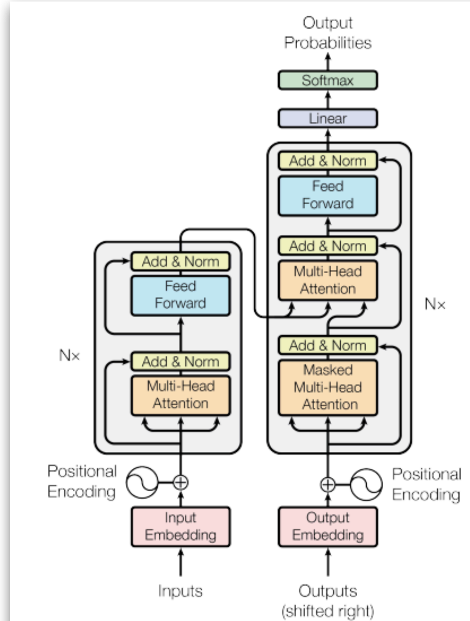


Figure 1: The Transformer - model architecture.

```python
def attention(query, key, value, mask=None, dropout=None):
    "Compute 'Scaled Dot Product Attention'"
    d_k = query.size(-1)
    scores = torch.matmul(query, key.transpose(-2, -1)) \
             / math.sqrt(d_k)
```

Course Objectives

- **Understand and describe** state-of-the-art models and algorithms for machine translation.

- **Implement and apply** such methods using real-world tasks.

- **Evaluate and analyze** the quality of machine translation systems.

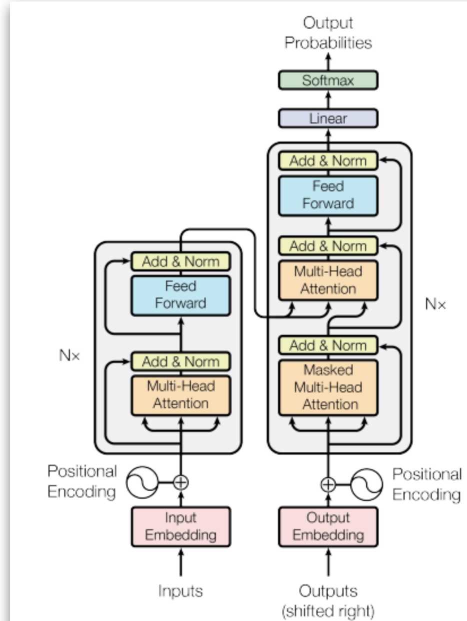
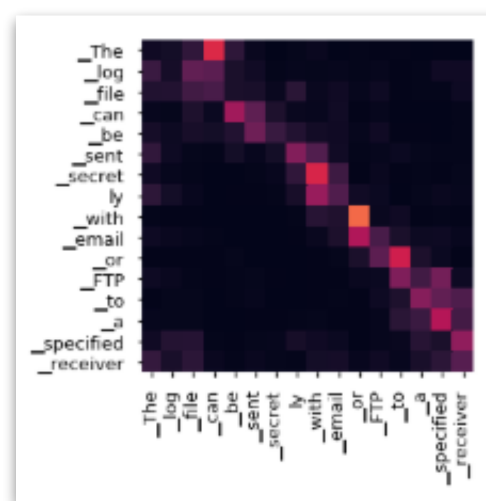
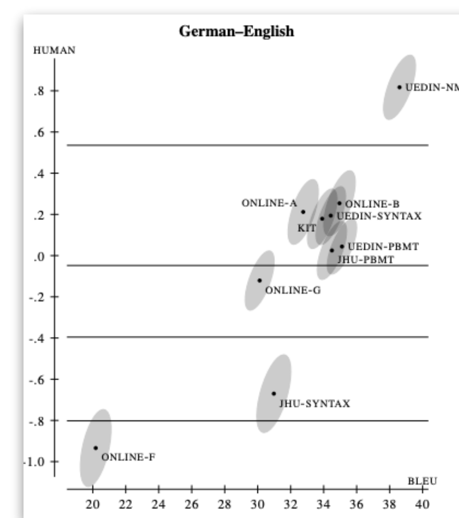


Figure 1: The Transformer - model architecture.

```
def attention(query, key, value, mask=None, dropout=None):
    "Compute 'Scaled Dot Product Attention'"
    d_k = query.size(-1)
    scores = torch.matmul(query, key.transpose(-2, -1)) \
             / math.sqrt(d_k)
```





Logistics

Logistics

- 3 Home assignments, **individual** submissions

Logistics

- 3 Home assignments, **individual** submissions
  - Will require using a deep learning framework (pytorch/ other)

Logistics

- 3 Home assignments, **individual** submissions
  - Will require using a deep learning framework (pytorch/ other)
- Grade: **50%** assignments, **50%** exam

Logistics

- 3 Home assignments, **individual** submissions
  - Will require using a deep learning framework (pytorch/other)
- Grade: **50%** assignments, **50%** exam
- Visiting hours - after class, schedule in advance

Logistics

- 3 Home assignments, **individual** submissions
  - Will require using a deep learning framework (pytorch/other)
- Grade: **50%** assignments, **50%** exam
- Visiting hours - after class, schedule in advance
- To succeed - attend, do the assignments, prepare for the exam

Logistics

- 3 Home assignments, **individual** submissions
  - Will require using a deep learning framework (pytorch/other)
- Grade: **50%** assignments, **50%** exam
- Visiting hours - after class, schedule in advance
- To succeed - attend, do the assignments, prepare for the exam
- My email: roee.aharoni@gmail.com

Other things to note

Other things to note

- Advanced topics - **attendance** is important

Other things to note

- Advanced topics - **attendance** is important
- Diverse group - use it, ask questions, be patient

Other things to note

- Advanced topics - **attendance** is important
- Diverse group - use it, ask questions, be patient
- First time in its current structure - **give feedback**!

What is the problem?

What is a good translation?

What is a good translation?

What is a good translation?

- Transitions from one language to another

What is a good translation?

- Transitions from one language to another
- Preserves the meaning

What is a good translation?

- Transitions from one language to another
- Preserves the meaning
- Fluent output (?)

What is a good translation?

- Transitions from one language to another
- Preserves the meaning
- Fluent output (?)
- Preserves style (?)

What is a good translation?

- Transitions from one language to another
- Preserves the meaning
- Fluent output (?)
- Preserves style (?)
- And many more…

Why is it hard?

Why is it hard?

Why is it hard?

“I sat on the bank”   1

Why is it hard?

“I sat on the bank” 1

“ראיתי איש קרח” 2

Why is it hard?

1 “I sat on the bank”

2 ”ראיתי איש קרח“

3 ”ספר עזר לרופא בהוצאת כתר“

Why is it hard?

1. “I sat on the bank”
2. ”ראיתי איש קרח“
3. ”ספר עזר לרופא בהוצאת כתר“
4. ”פיתה עם לבנה”

Why is it hard?

1 “I sat on the bank”

2 ”ראיתי איש קרח“

3 ”ספר עזר לרופא בהוצאת כתר“

4 ”פיתה עם לבנה″

5 ”תה חזק“

Why is it hard?

1. “I sat on the bank”
2. “ראיתי איש קרח”
3. “ספר עזר לרופא בהוצאת כתר”
4. “פיתה עם לבנה”
5. “תה חזק”
6. “Out of sight, out of mind” - ”Invisible Idiot”

What causes ambiguity?

What causes ambiguity?

- Complex **morphology**

Finnish: ostoskeskuksessa

ostos#keskus+N+Sg+Loc:in

shopping#center+N+Sg+Loc:in

English: ‘in the shopping center’

What causes ambiguity?

- Complex **morphology**
  - Part-of-speech

Finnish: ostoskeskuksessa

ostos#keskus+N+Sg+Loc:in

shopping#center+N+Sg+Loc:in

English: ‘in the shopping center’

What causes ambiguity?

- Complex **morphology**
  - Part-of-speech
  - Number

Finnish: ostoskeskuksessa

ostos#keskus+N+Sg+Loc:in

shopping#center+N+Sg+Loc:in

English: ‘in the shopping center’

What causes ambiguity?

- Complex **morphology**
  - Part-of-speech
  - Number
  - Gender

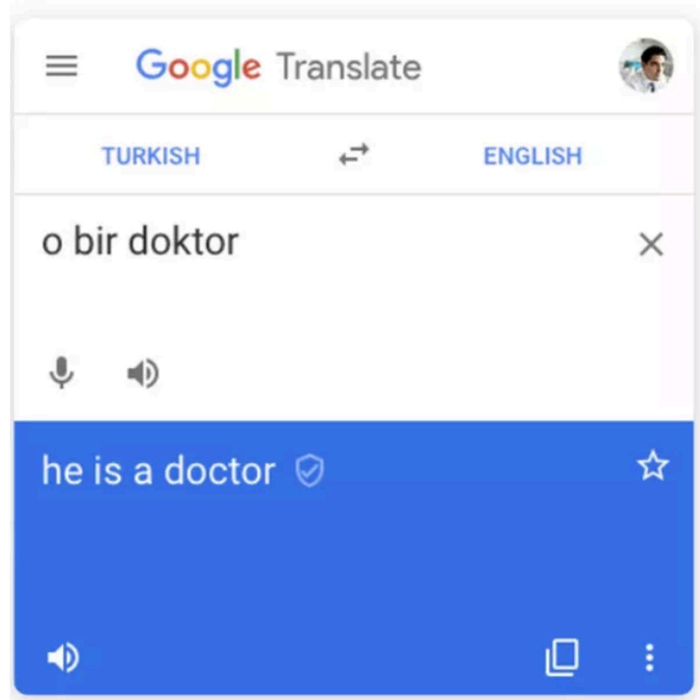
Finnish: ostoskeskuksessa

ostos#keskus+N+Sg+Loc:in

shopping#center+N+Sg+Loc:in

English: ‘in the shopping center’

Before



After

What causes ambiguity?

- Complex **morphology**
  - Part-of-speech
  - Number
  - Gender
  - Tense

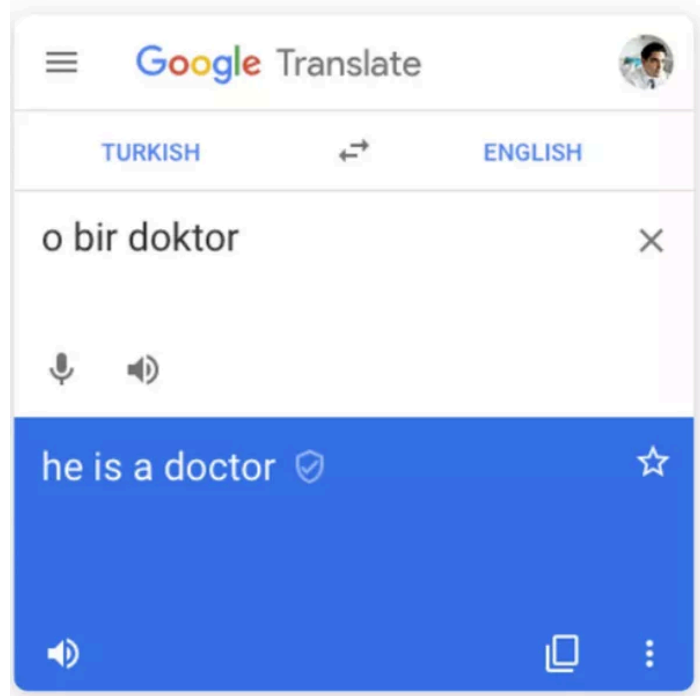
Finnish: ostoskeskuksessa

ostos#keskus+N+Sg+Loc:in

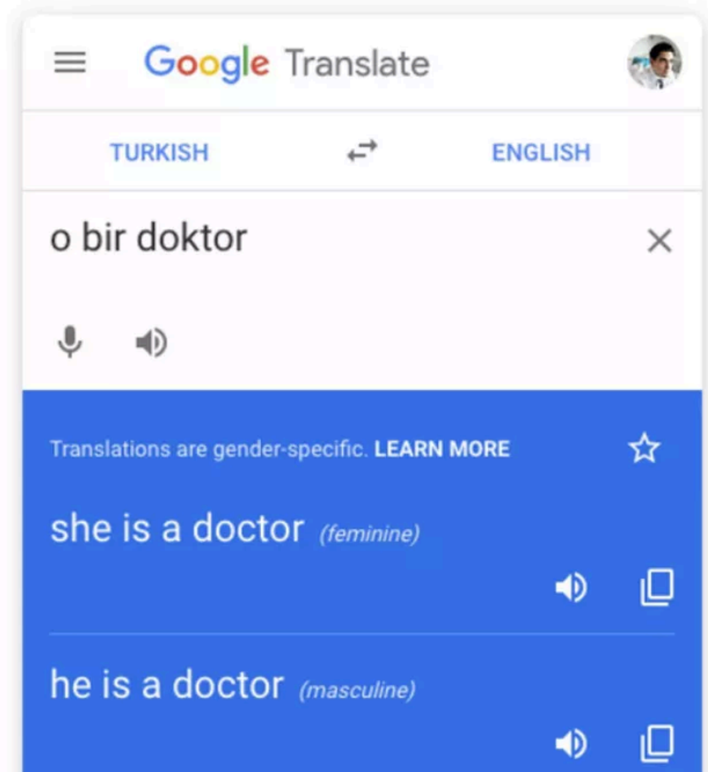
shopping#center+N+Sg+Loc:in

English: ‘in the shopping center’

Before



After



What causes ambiguity?

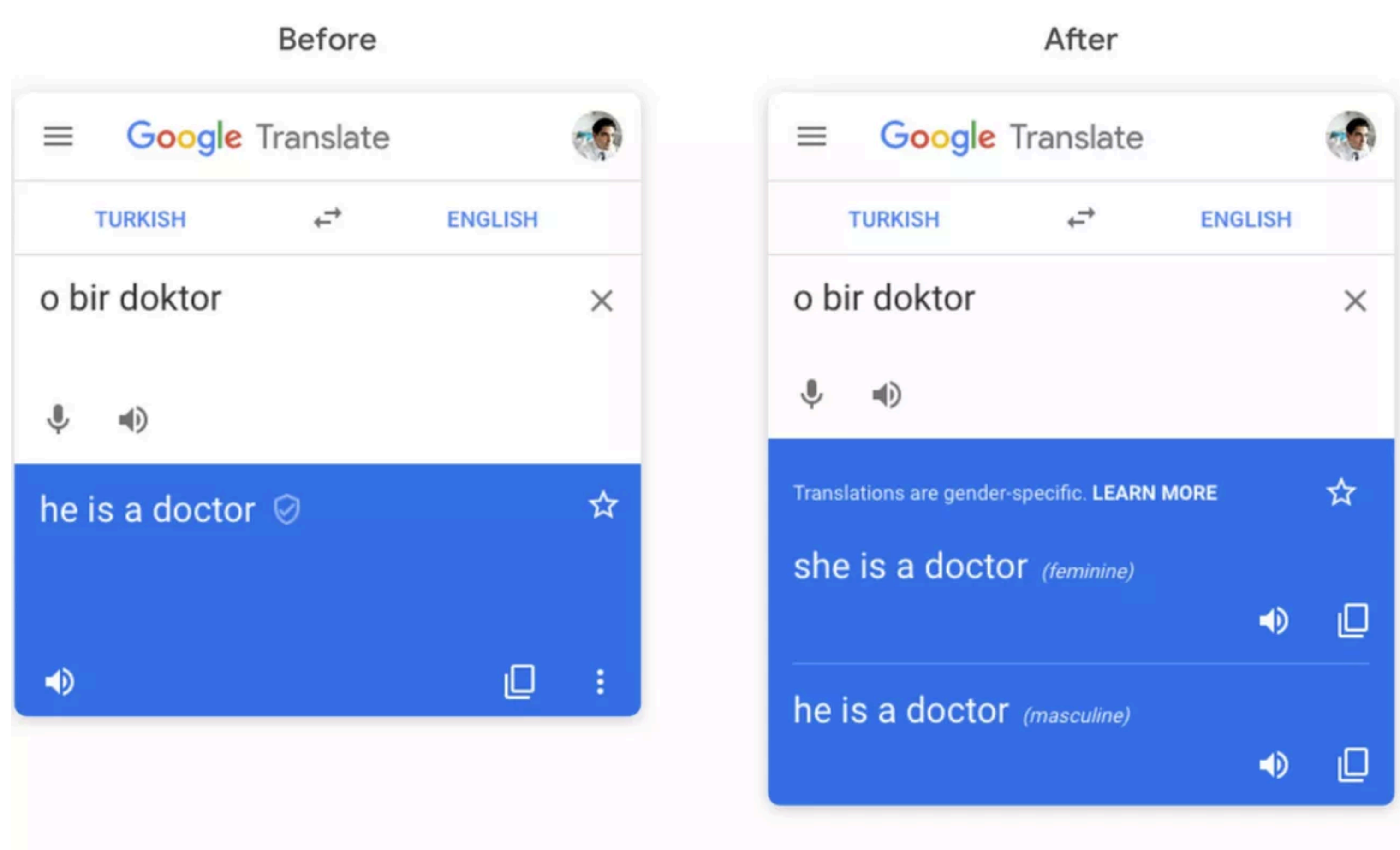
- Complex **morphology**
    - Part-of-speech
    - Number
    - Gender
    - Tense
    - Many more…

Finnish: ostoskeskuksessa

ostos#keskus+N+Sg+Loc:in

shopping#center+N+Sg+Loc:in

English: ‘in the shopping center’



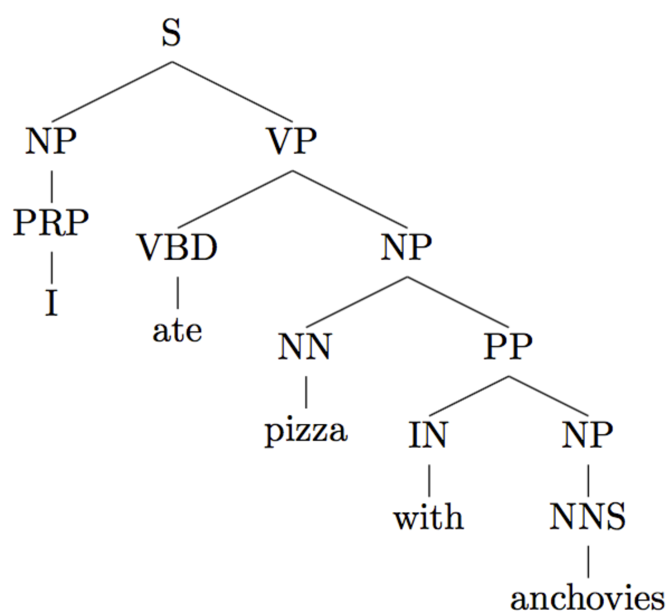
What causes ambiguity?

What causes ambiguity?

- Complexity of **Syntax** - several ways to parse a sentence:

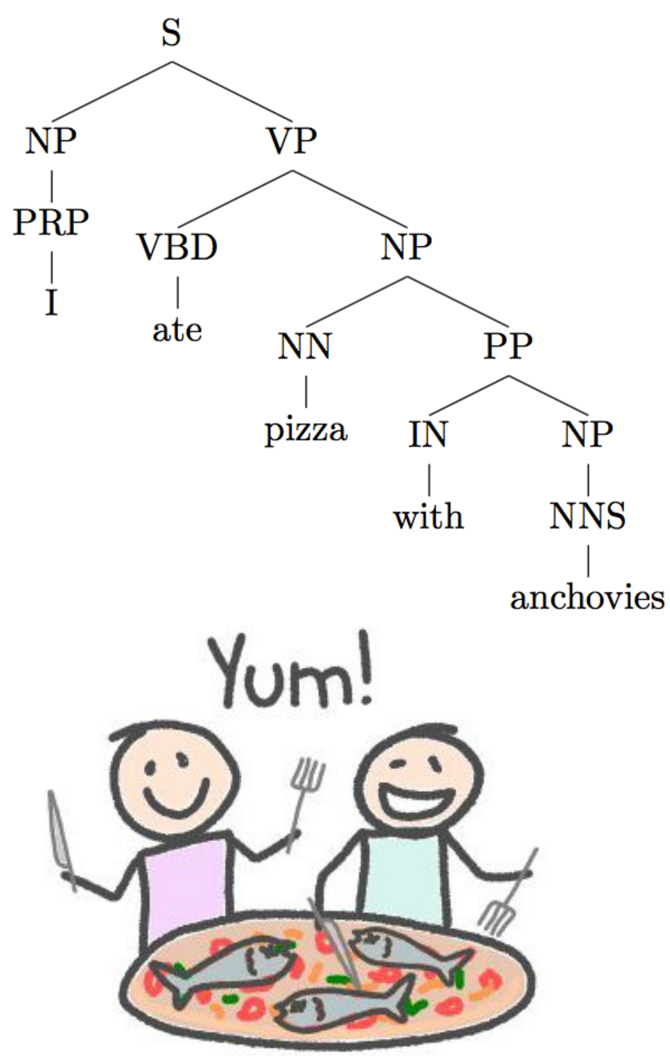
What causes ambiguity?

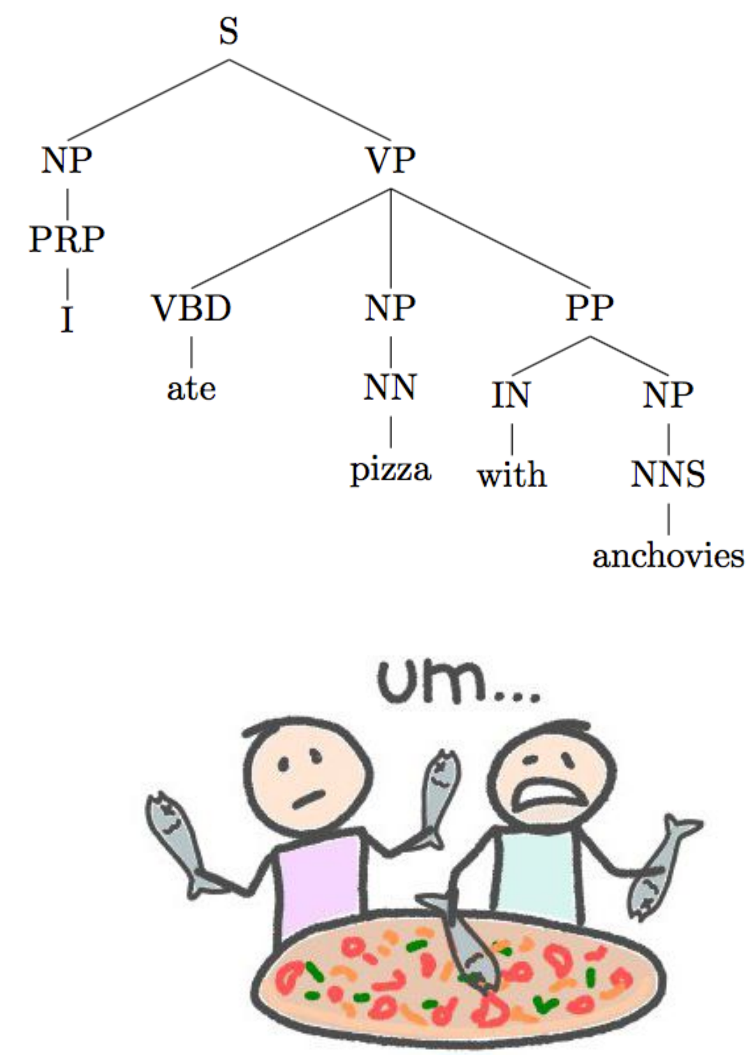
- Complexity of **Syntax** - several ways to parse a sentence:



What causes ambiguity?

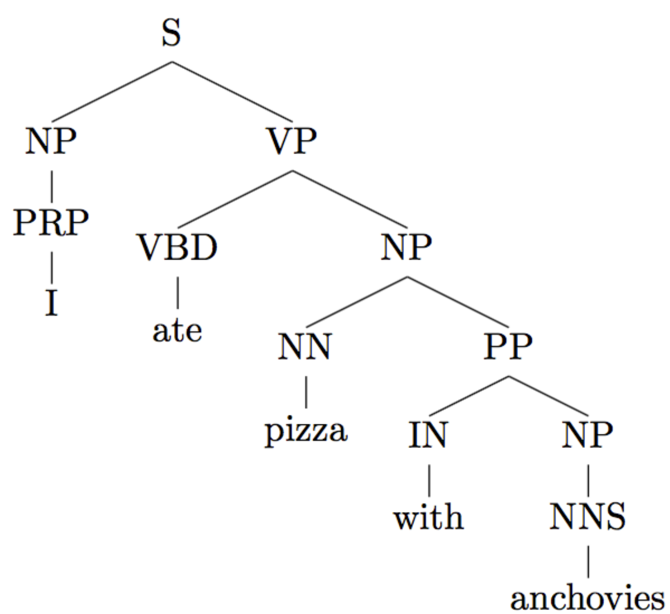
- Complexity of **Syntax** - several ways to parse a sentence:

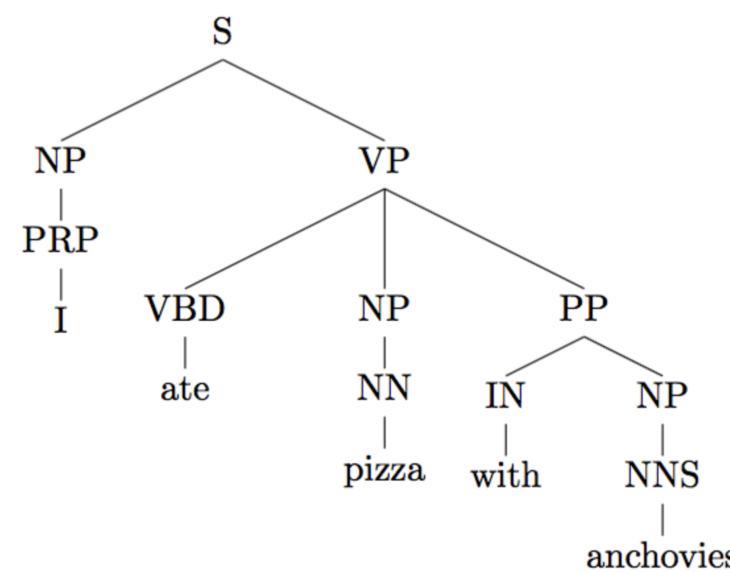


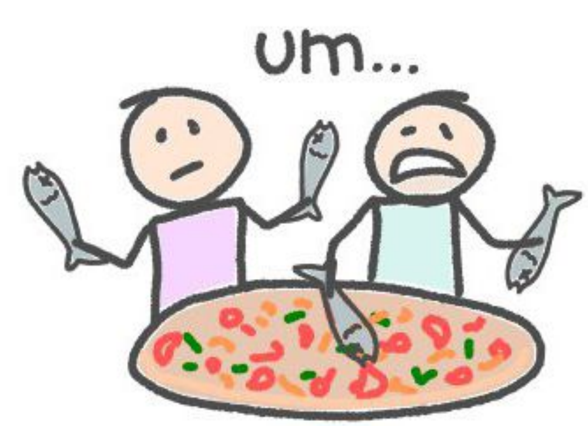


What causes ambiguity?

- Complexity of **Syntax** - several ways to parse a sentence:





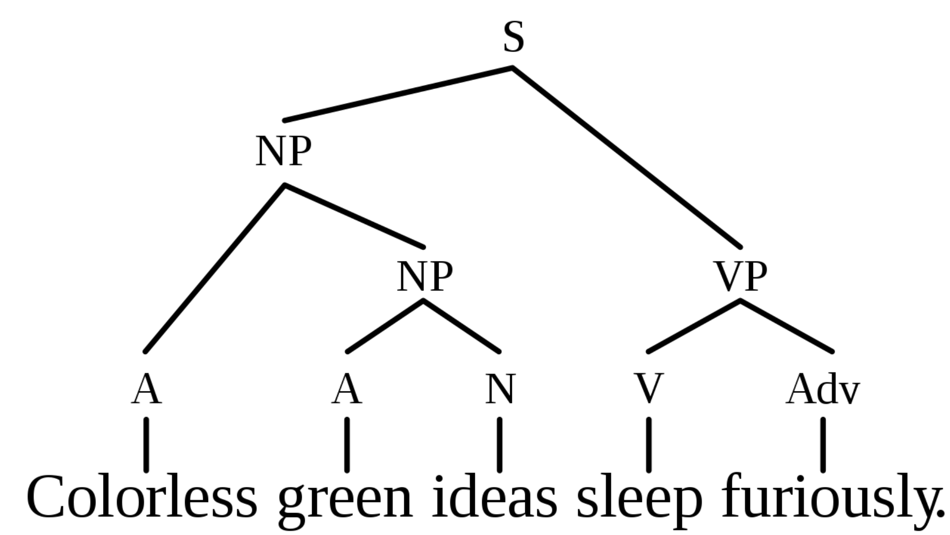


- More generally - **lack of context**

How do we handle this?

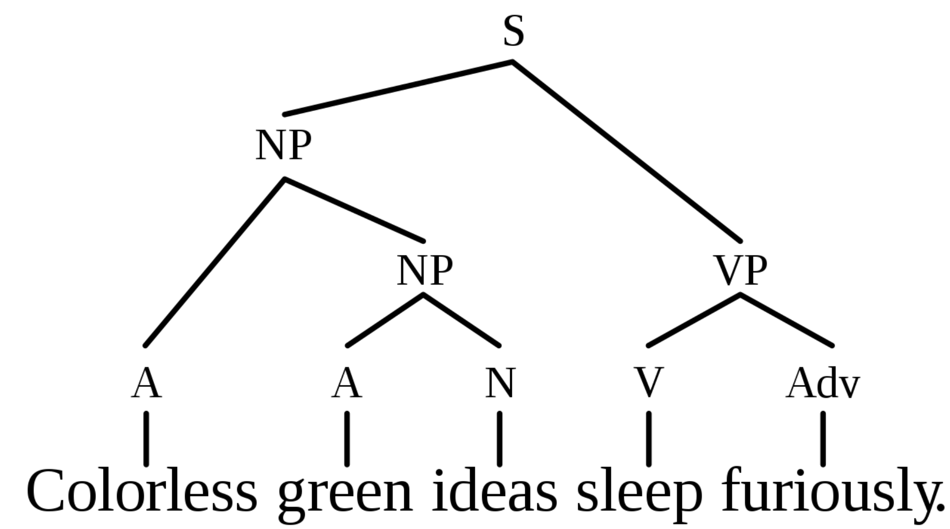
How do we handle this?

- Understand Language/ Linguistics



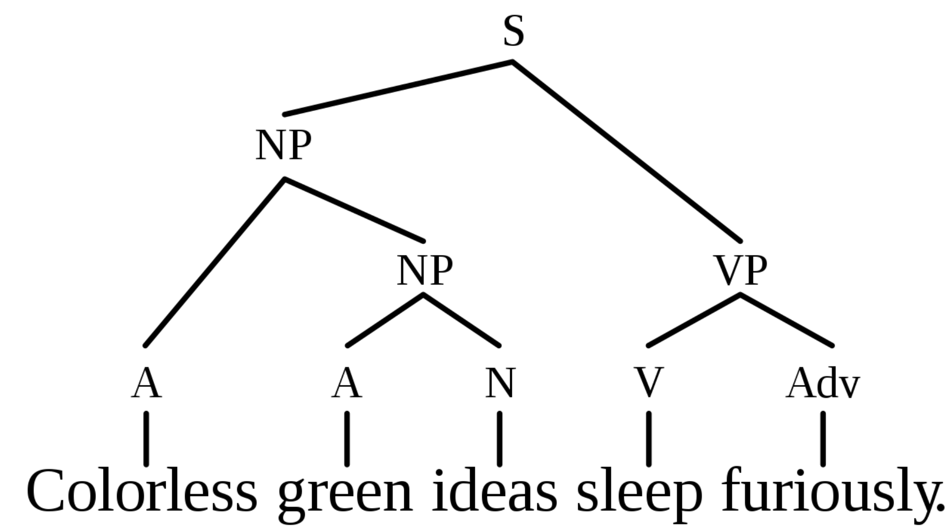
How do we handle this?

- Understand Language/ Linguistics
  - Syntax, Morphology, Typology…



How do we handle this?

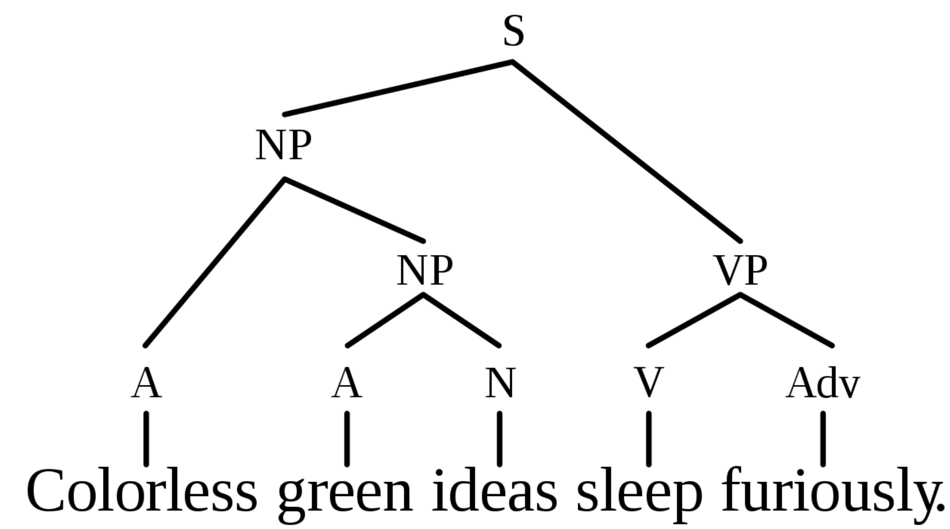
- Understand Language/ Linguistics
  - Syntax, Morphology, Typology…
- Probability/Statistics



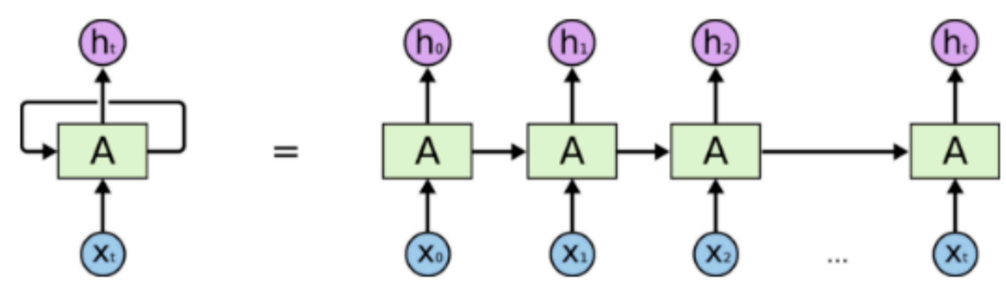
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

How do we handle this?

- Understand Language/ Linguistics
    - Syntax, Morphology, Typology…
- Probability/Statistics
- Machine Learning
    - Neural Networks (“Deep Learning”)



$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$



An unrolled recurrent neural network.

Why is it important?
A Crash Course in History

The Tower of Babel

Pieter Brueghel the Elder (1563)

1940s

First NLP problem: the German Enigma

1957

"I think we are forced to conclude that ... probabilistic models give no particular insight into some of the basic problems of syntactic structure."

–Noam Chomsky, *"Syntactic structures"*

"[I]t must be recognized that the notion of 'probability of a sentence' is an entirely useless one, under any known interpretation of this term."

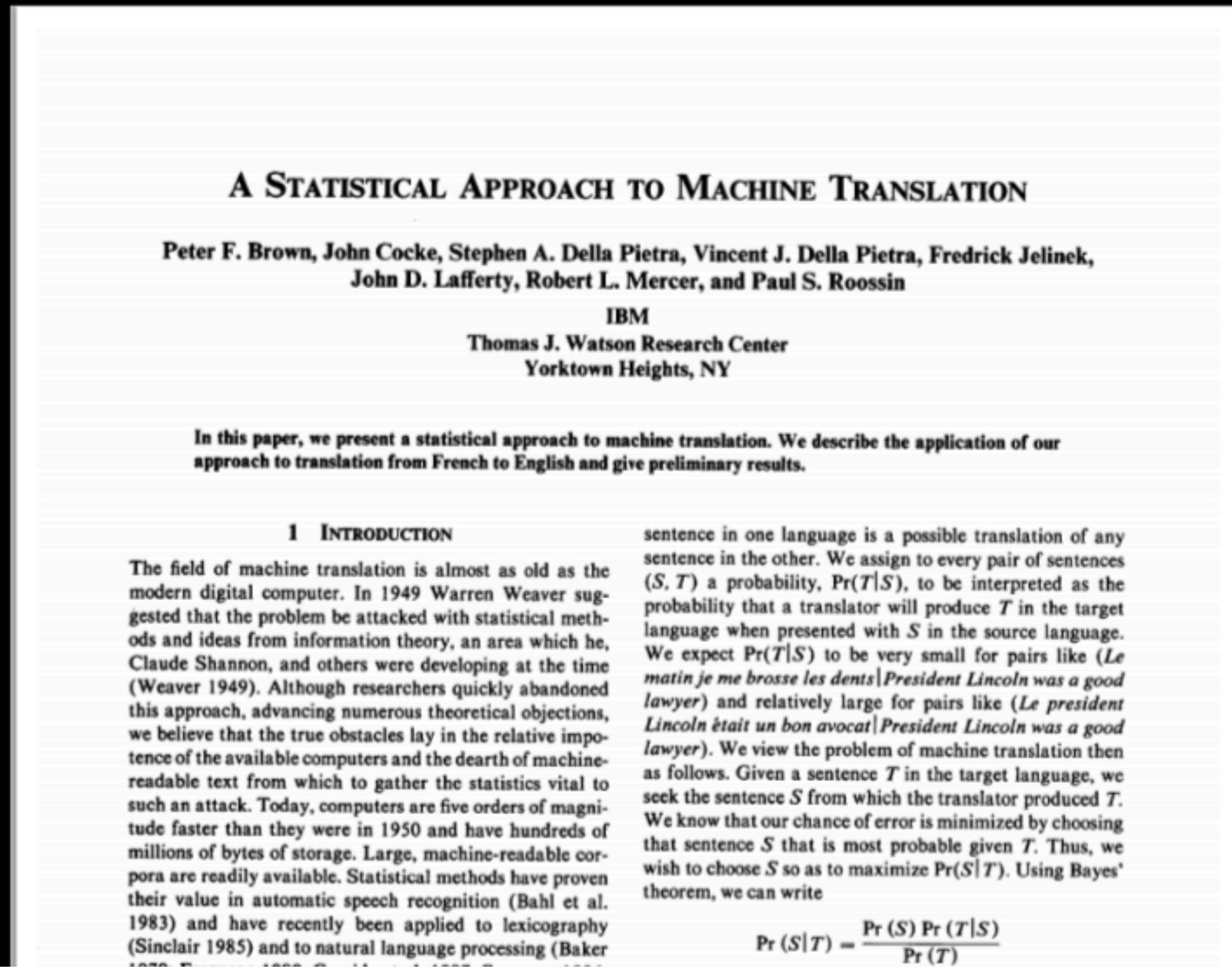
–in *"Challenges to empiricism"* (1969)

1962

Association for Machine Translation and Computational Linguistics

founded

1990



A STATISTICAL APPROACH TO MACHINE TRANSLATION

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin

IBM
Thomas J. Watson Research Center
Yorktown Heights, NY

In this paper, we present a statistical approach to machine translation. We describe the application of our approach to translation from French to English and give preliminary results.

1 INTRODUCTION

The field of machine translation is almost as old as the modern digital computer. In 1949 Warren Weaver suggested that the problem be attacked with statistical methods and ideas from information theory, an area which he, Claude Shannon, and others were developing at the time (Weaver 1949). Although researchers quickly abandoned this approach, advancing numerous theoretical objections, we believe that the true obstacles lay in the relative impotence of the available computers and the dearth of machine-readable text from which to gather the statistics vital to such an attack. Today, computers are five orders of magnitude faster than they were in 1950 and have hundreds of millions of bytes of storage. Large, machine-readable corpora are readily available. Statistical methods have proven their value in automatic speech recognition (Bahl et al. 1983) and have recently been applied to lexicography (Sinclair 1985) and to natural language processing (Baker

sentence in one language is a possible translation of any sentence in the other. We assign to every pair of sentences (S, T) a probability, $\Pr(T|S)$, to be interpreted as the probability that a translator will produce T in the target language when presented with S in the source language. We expect $\Pr(T|S)$ to be very small for pairs like (*Le matin je me brosse les dents*|*President Lincoln was a good lawyer*) and relatively large for pairs like (*Le president Lincoln était un bon avocat*|*President Lincoln was a good lawyer*). We view the problem of machine translation then as follows. Given a sentence T in the target language, we seek the sentence S from which the translator produced T. We know that our chance of error is minimized by choosing that sentence S that is most probable given T. Thus, we wish to choose S so as to maximize $\Pr(S|T)$. Using Bayes' theorem, we can write

$$\Pr(S|T) = \frac{\Pr(S)\Pr(T|S)}{\Pr(T)}$$

IBM's statistical MT paper published in Computational Linguistics

2006

28 April: Google translate goes live

2014

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its

First papers on (working) neural machine translation

2016

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.

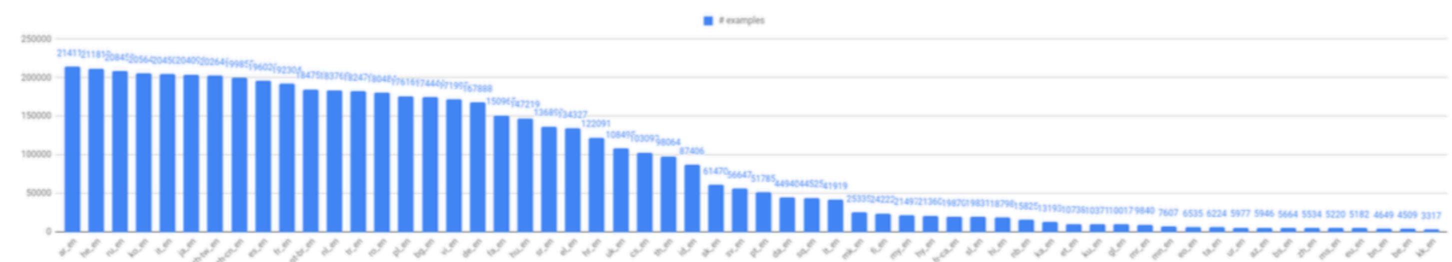
Today we announce the Google Neural Machine Translation system (GNMT), which utilizes state-of-the-art training techniques to achieve the largest improvements to date for machine translation quality. Our full research results are described in a new technical report we are releasing today: "*Google's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

Google Translate launches the world's first neural machine translation system

Challenges in 2020

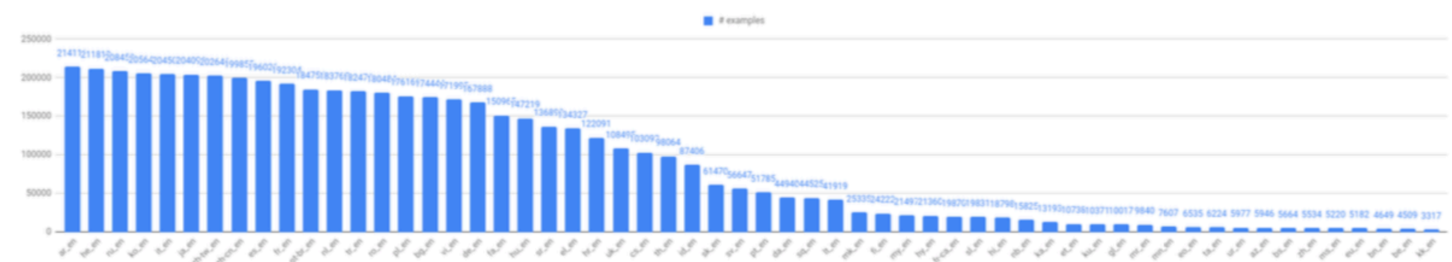
Challenges in 2020

- Low-resource languages



Challenges in 2020

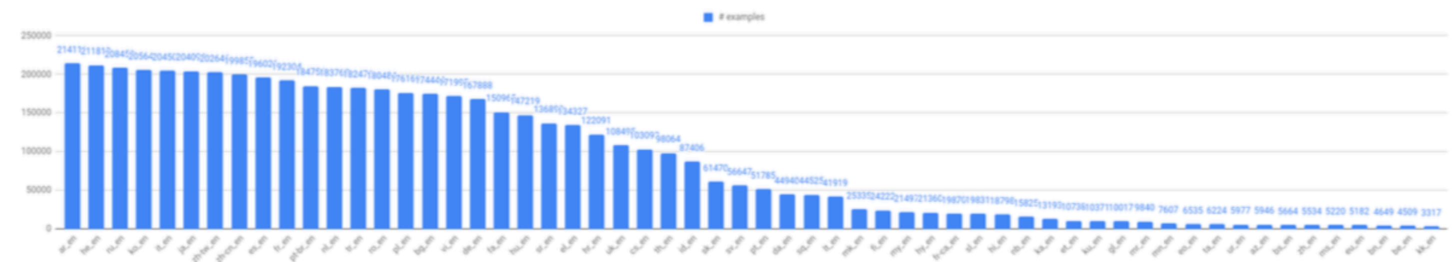
- Low-resource languages

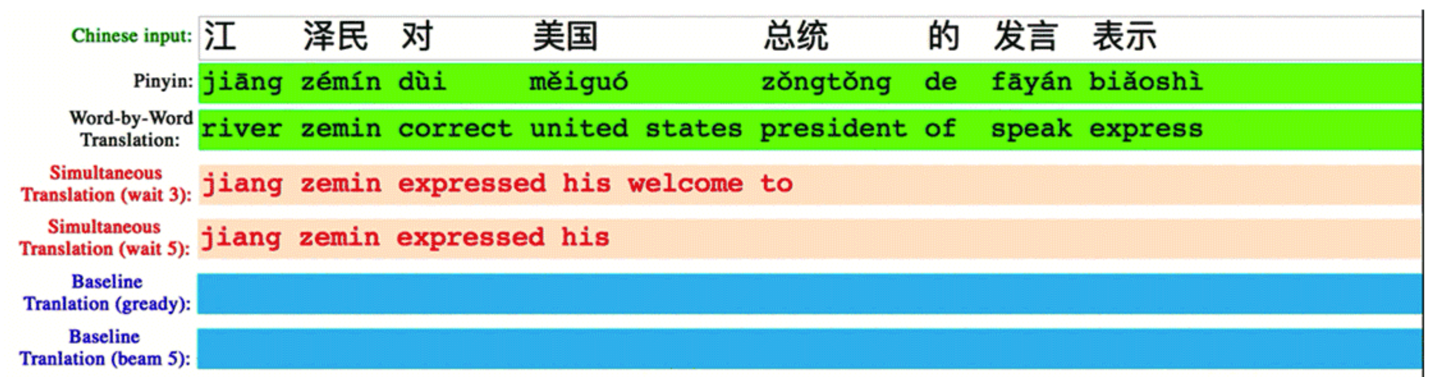


- Speech Translation

Challenges in 2020

- Low-resource languages
- Speech Translation
- Document Translation





Keanu Reeves

From Wikipedia, the free encyclopedia

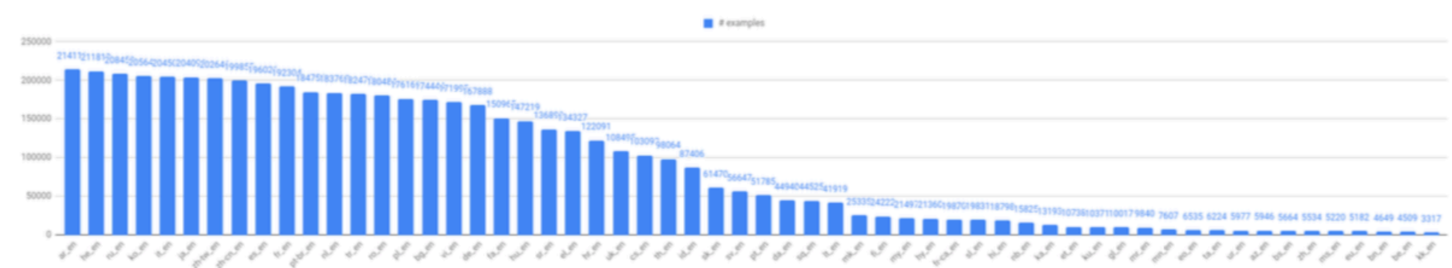
"Keanu" redirects here. For other uses, see Keanu (disambiguation).

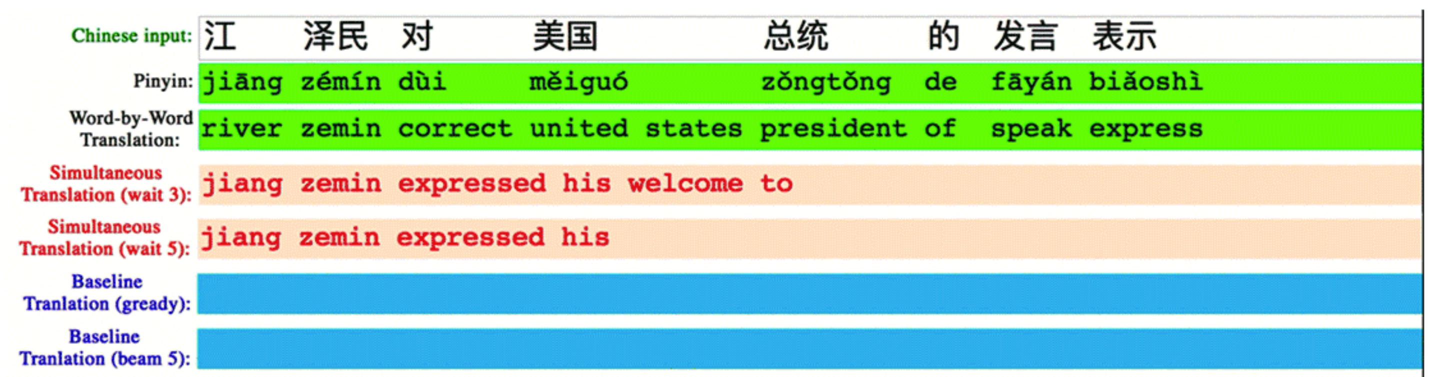
Keanu Charles Reeves (/kiˈɑːnuː/ *kee-AH-noo*;[2][3][4] born September 2, 1964) is a Canadian[a] actor and musician. He gained fame for his starring roles in several blockbuster films, including comedies from the *Bill and Ted* franchise (1989–2020); action thrillers *Point Break* (1991), *Speed* (1994), the *John Wick* franchise (2014–present); psychological thriller *The Devil's Advocate* (1997); supernatural thriller *Constantine* (2005); and science fiction/action series *The Matrix* (1999–2003). He has also appeared in dramatic films such as *Dangerous Liaisons* (1988), *My Own Private Idaho* (1991), and *Little Buddha* (1993), as well as the romantic horror *Bram Stoker's Dracula* (1992).

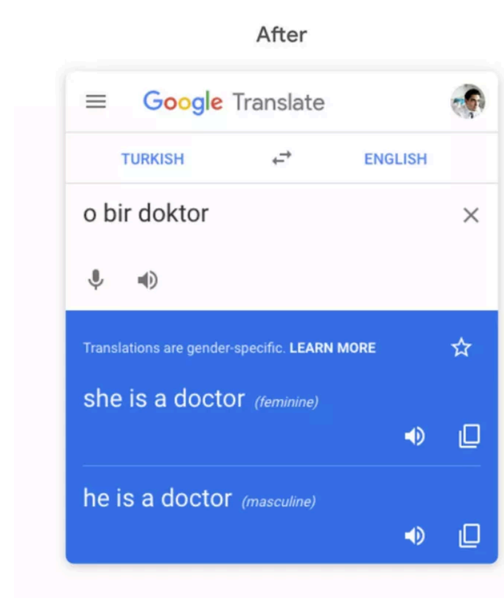
Contents [hide]

Challenges in 2020

- Low-resource languages
- Speech Translation
- Document Translation
- Gender Bias








Keanu Reeves

From Wikipedia, the free encyclopedia

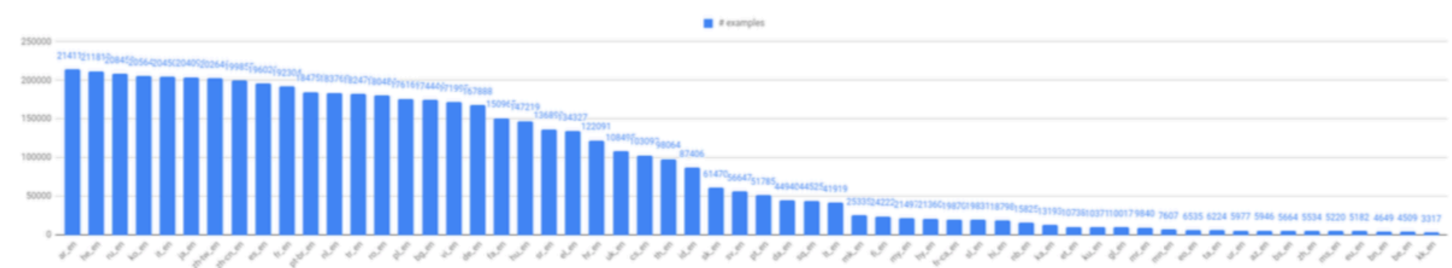
"Keanu" redirects here. For other uses, see Keanu (disambiguation).

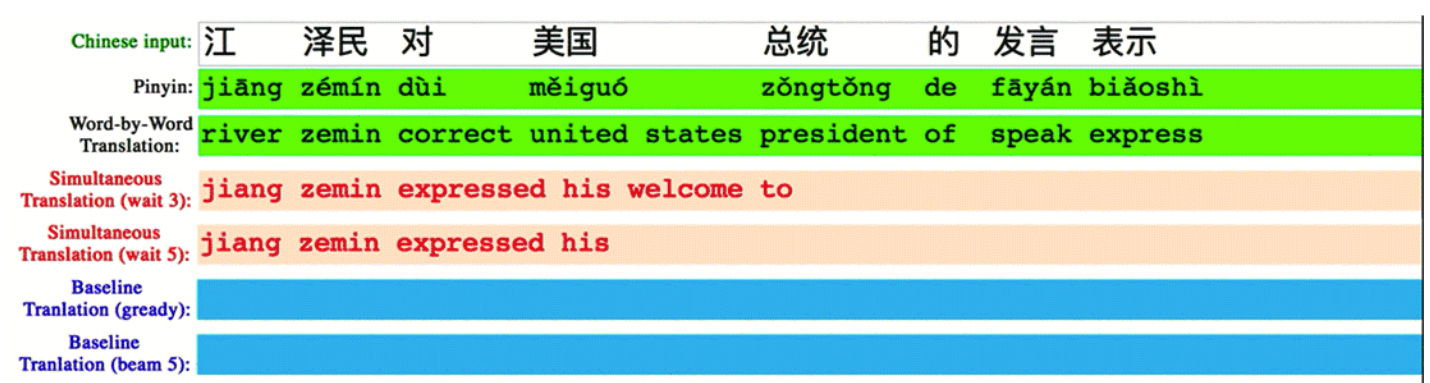
Keanu Charles Reeves (/ki ɑ nu / kee-AH-noo;[2][3][4] born September 2, 1964) is a Canadian[a] actor and musician. He gained fame for his starring roles in several blockbuster films, including comedies from the Bill and Ted franchise (1989–2020); action thrillers Point Break (1991), Speed (1994), the John Wick franchise (2014–present); psychological thriller The Devil's Advocate (1997); supernatural thriller Constantine (2005); and science fiction/action series The Matrix (1999–2003). He has also appeared in dramatic films such as Dangerous Liaisons (1988), My Own Private Idaho (1991), and Little Buddha (1993), as well as the romantic horror Bram Stoker's Dracula (1992).

Contents [hide]

Challenges in 2020

- Low-resource languages
- Speech Translation
- Document Translation
- Gender Bias
- Robustness






MOTHERBOARD
TECH BY VICE

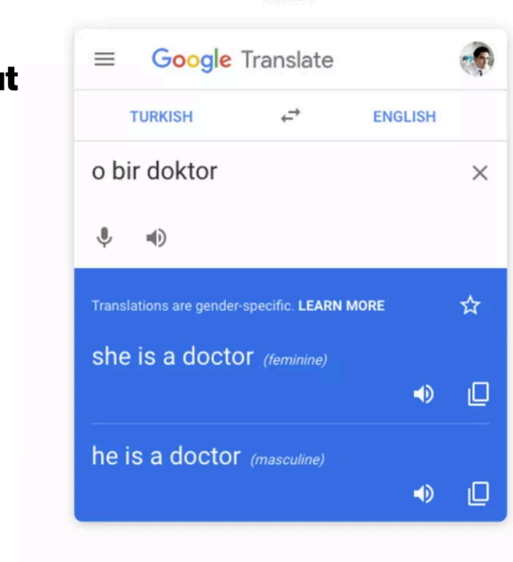
Why Is Google Translate Spitting Out Sinister Religious Prophecies?

Google Translate is moonlighting as a deranged oracle – and experts say it's likely because of the spooky nature of neural networks.

By Jon Christian

Jul 20 2018, 8:04pm Share Tweet Snap

After




Keanu Reeves

From Wikipedia, the free encyclopedia

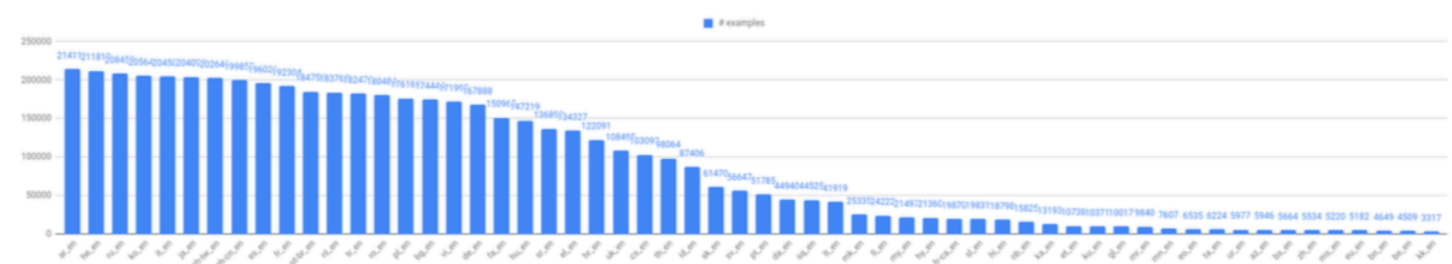
"Keanu" redirects here. For other uses, see Keanu (disambiguation).

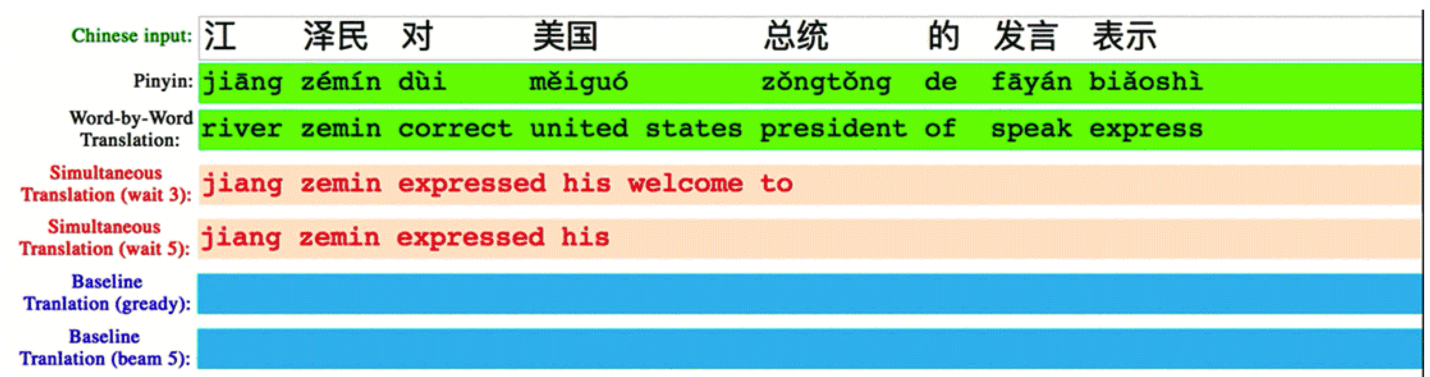
Keanu Charles Reeves (/kiˈɑːnuː/ *kee-AH-noo*;[2][3][4] born September 2, 1964) is a Canadian[a] actor and musician. He gained fame for his starring roles in several blockbuster films, including comedies from the *Bill and Ted* franchise (1989–2020); action thrillers *Point Break* (1991), *Speed* (1994), the *John Wick* franchise (2014–present); psychological thriller *The Devil's Advocate* (1997); supernatural thriller *Constantine* (2005); and science fiction/action series *The Matrix* (1999–2003). He has also appeared in dramatic films such as *Dangerous Liaisons* (1988), *My Own Private Idaho* (1991), and *Little Buddha* (1993), as well as the romantic horror *Bram Stoker's Dracula* (1992).

Contents [hide]

Challenges in 2020

- Low-resource languages
- Speech Translation
- Document Translation
- Gender Bias
- Robustness
- Many more!





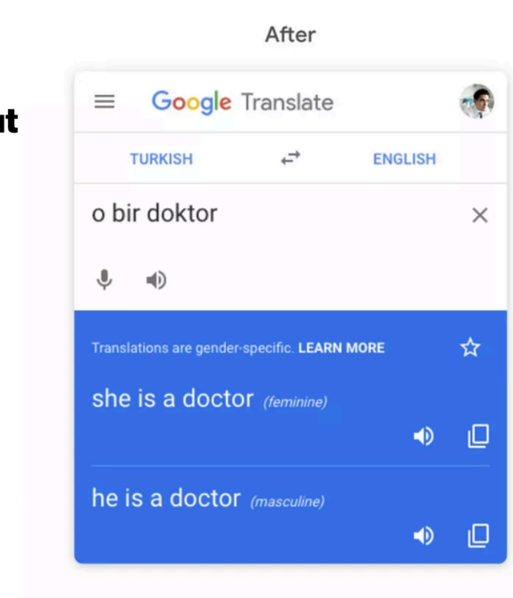

MOTHERBOARD
TECH BY VICE

Why Is Google Translate Spitting Out Sinister Religious Prophecies?

Google Translate is moonlighting as a deranged oracle—and experts say it's likely because of the spooky nature of neural networks.

By Jon Christian

Jul 20 2018, 8:04pm Share Tweet Snap




Keanu Reeves

From Wikipedia, the free encyclopedia

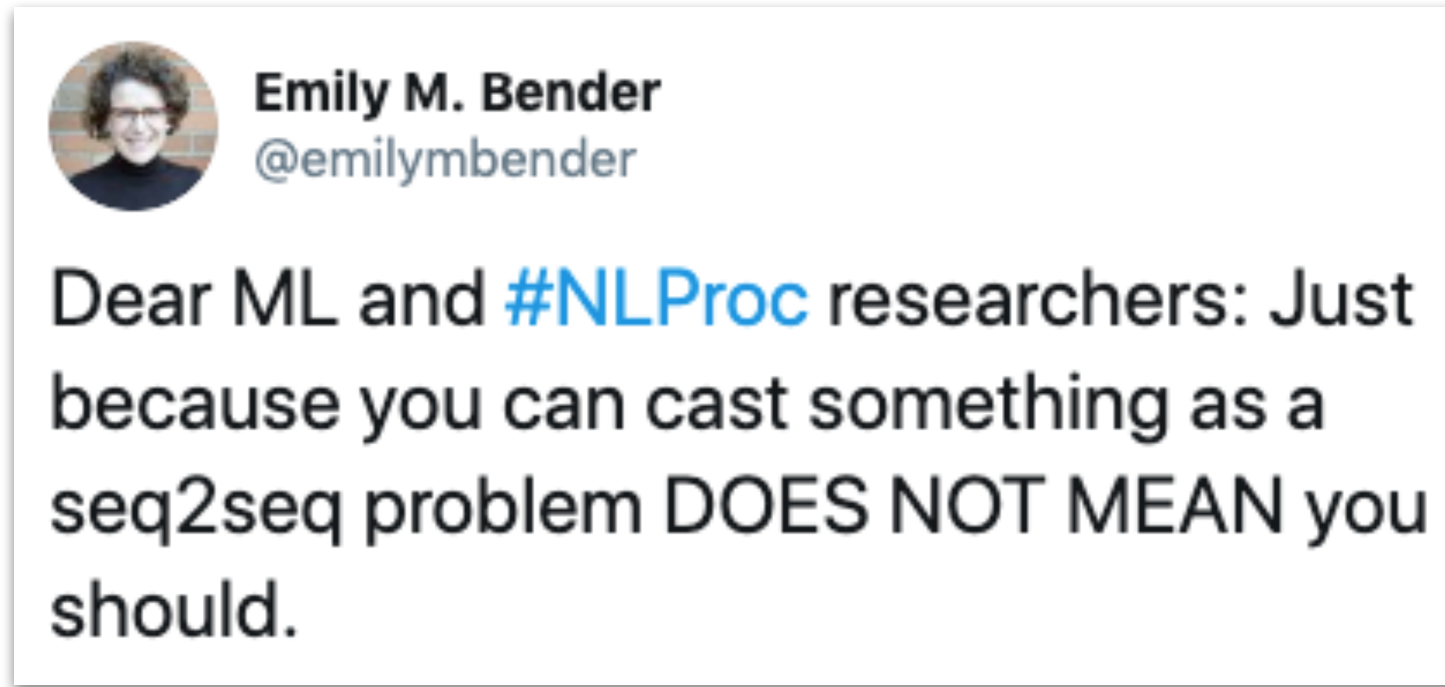
"Keanu" redirects here. For other uses, see Keanu (disambiguation).

Keanu Charles Reeves (/ki ɑ nu / kee-AH-noo;[2][3][4] born September 2, 1964) is a Canadian[a] actor and musician. He gained fame for his starring roles in several blockbuster films, including comedies from the Bill and Ted franchise (1989–2020); action thrillers Point Break (1991), Speed (1994), the John Wick franchise (2014–present); psychological thriller The Devil's Advocate (1997); supernatural thriller Constantine (2005); and science fiction/action series The Matrix (1999–2003). He has also appeared in dramatic films such as Dangerous Liaisons (1988), My Own Private Idaho (1991), and Little Buddha (1993), as well as the romantic horror Bram Stoker's Dracula (1992).

Contents [hide]

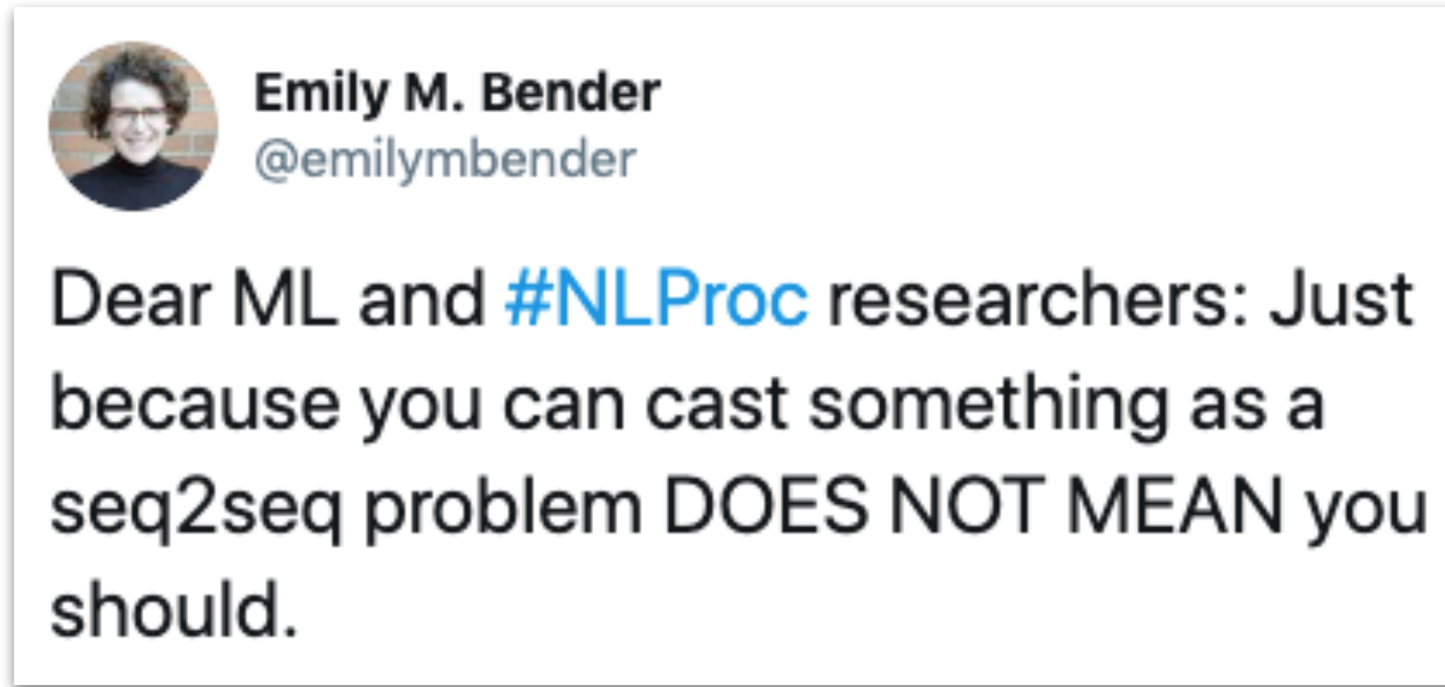
Additional Applications

Additional Applications



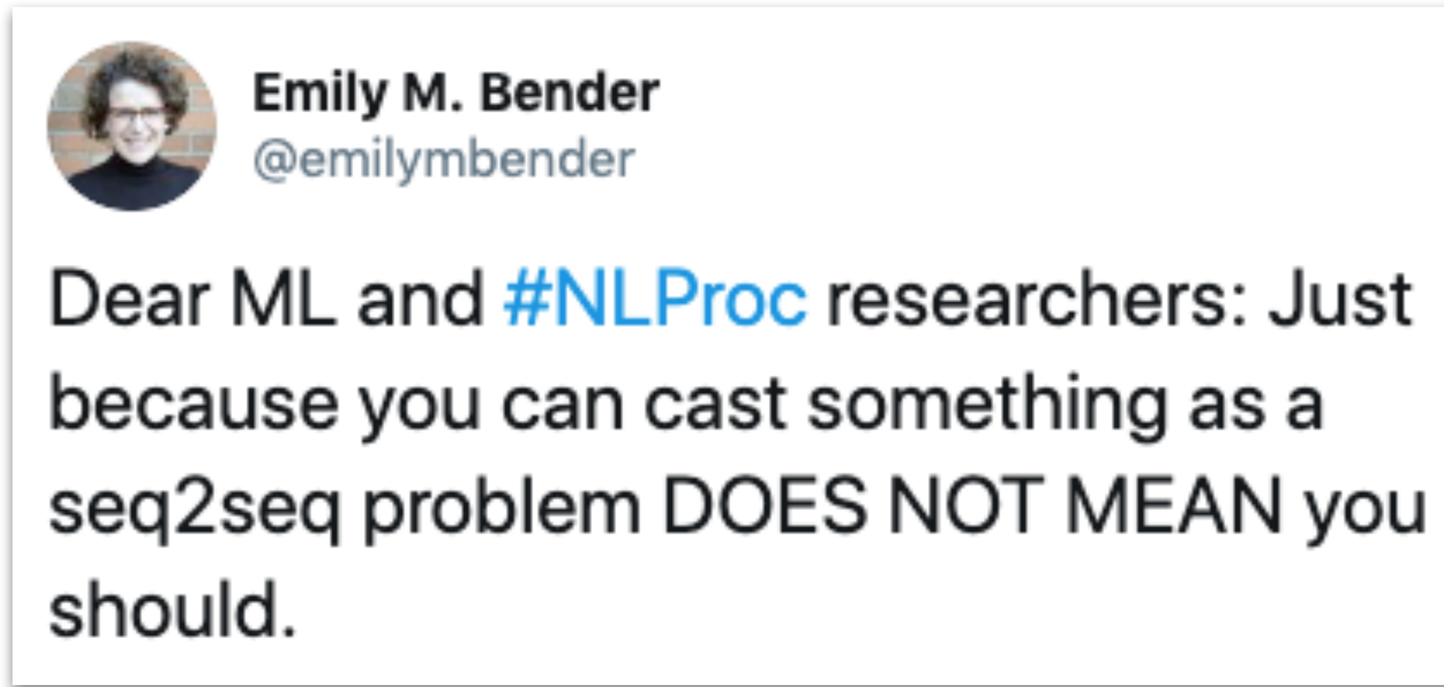
Additional Applications

- Speech recognition



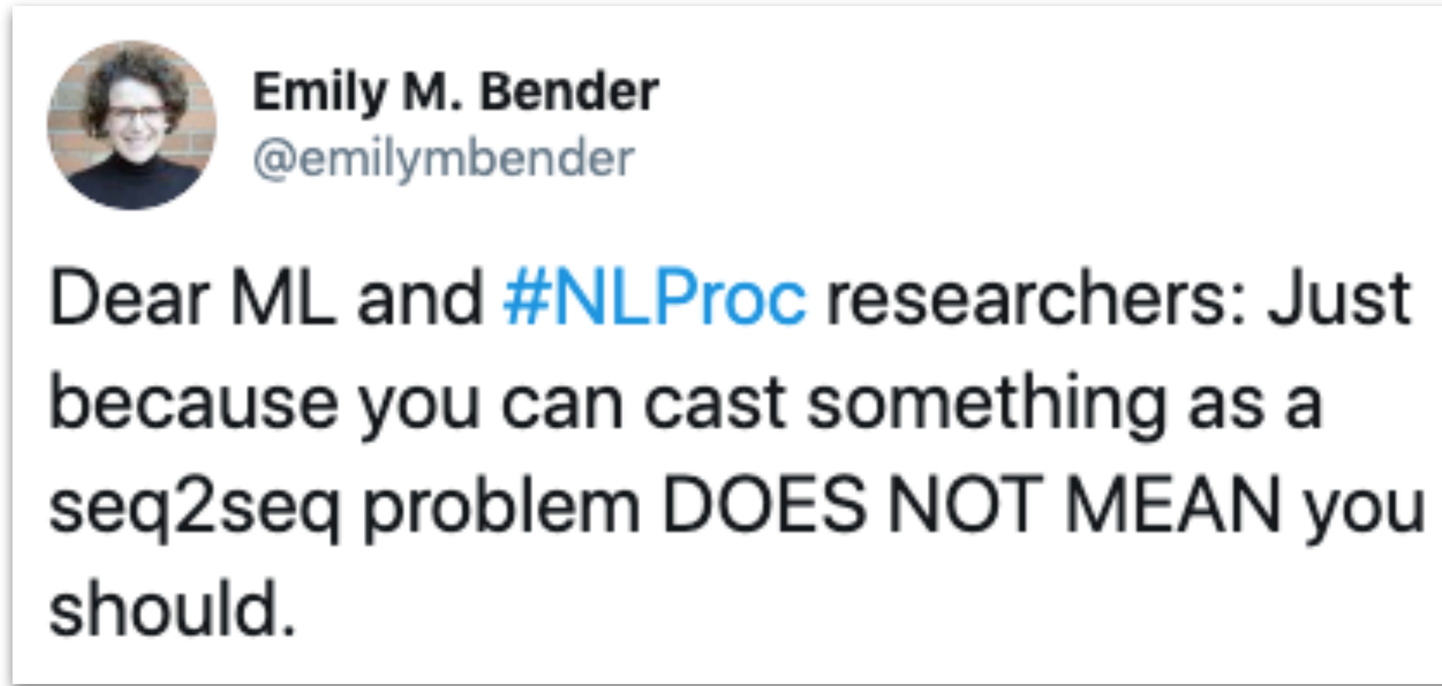
Additional Applications

- Speech recognition
- Summarization



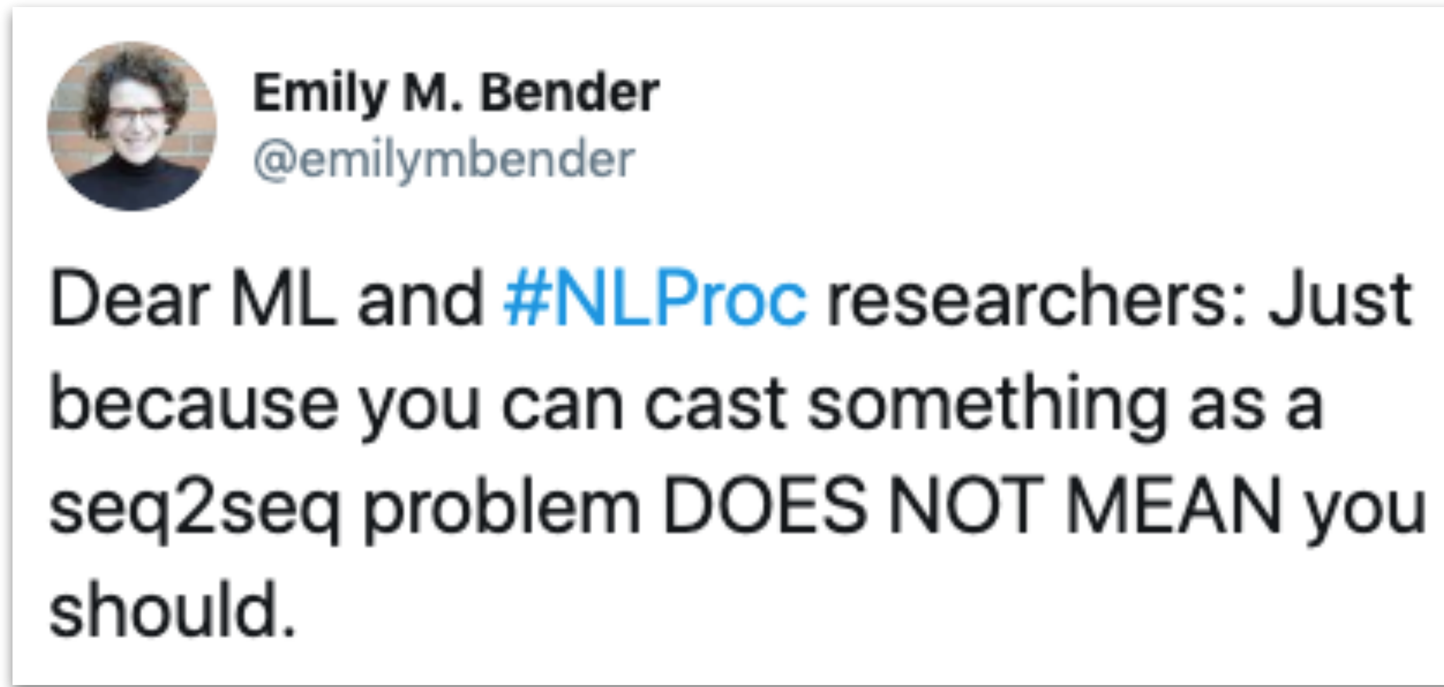
Additional Applications

- Speech recognition
- Summarization
- Text style-transfer



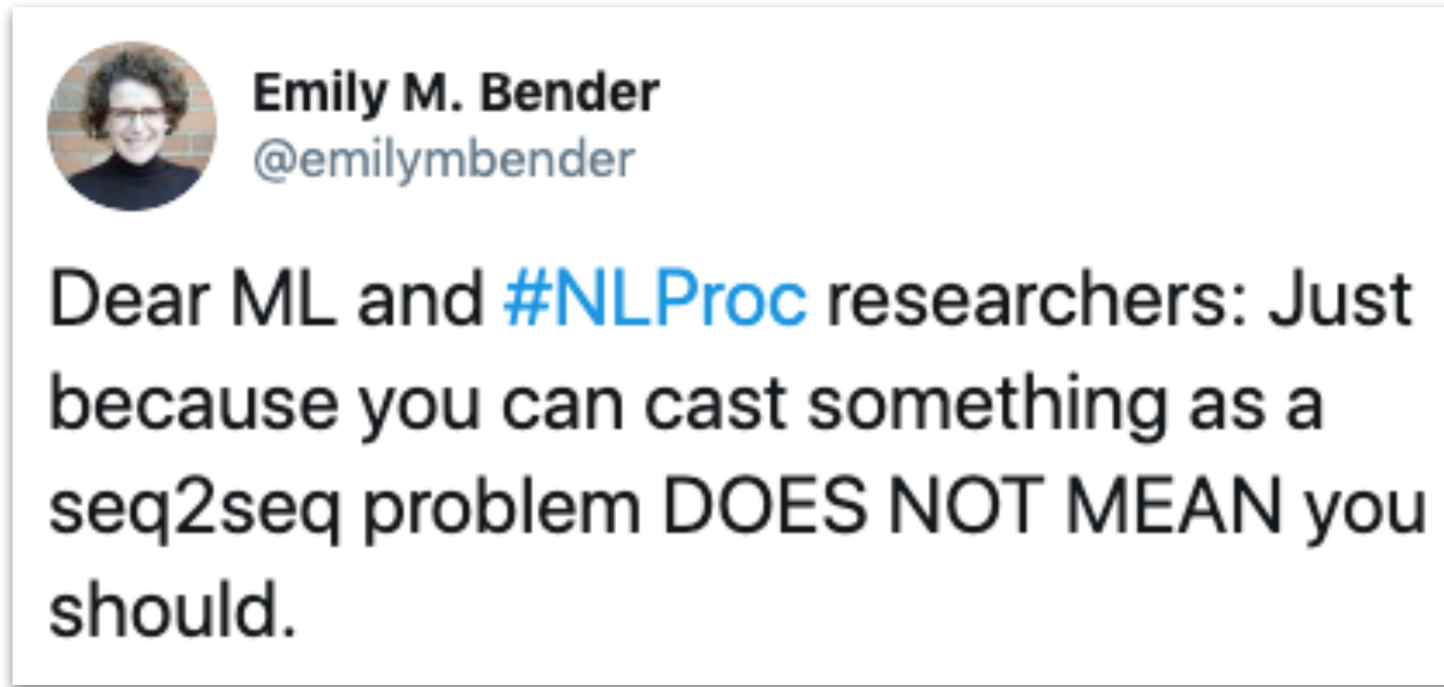
Additional Applications

- Speech recognition
- Summarization
- Text style-transfer
- Paraphrasing



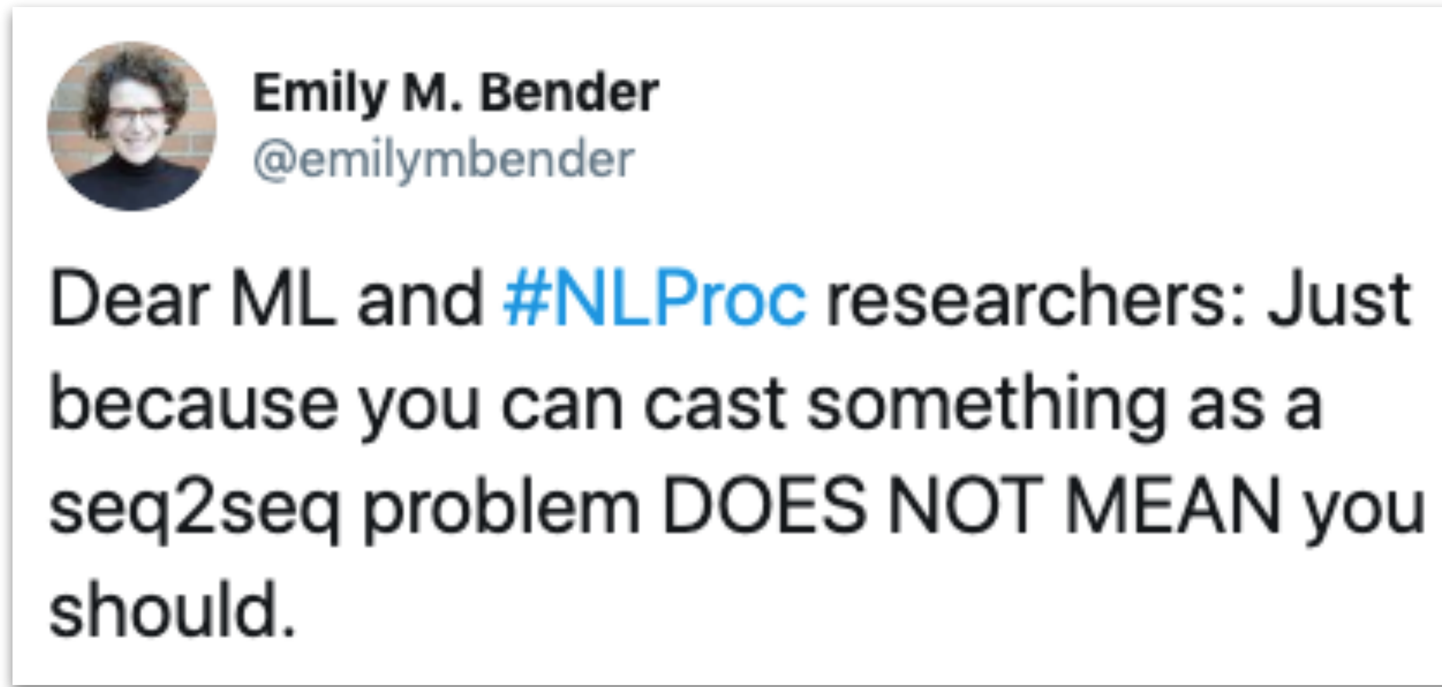
Additional Applications

- Speech recognition
- Summarization
- Text style-transfer
- Paraphrasing
- Parsing



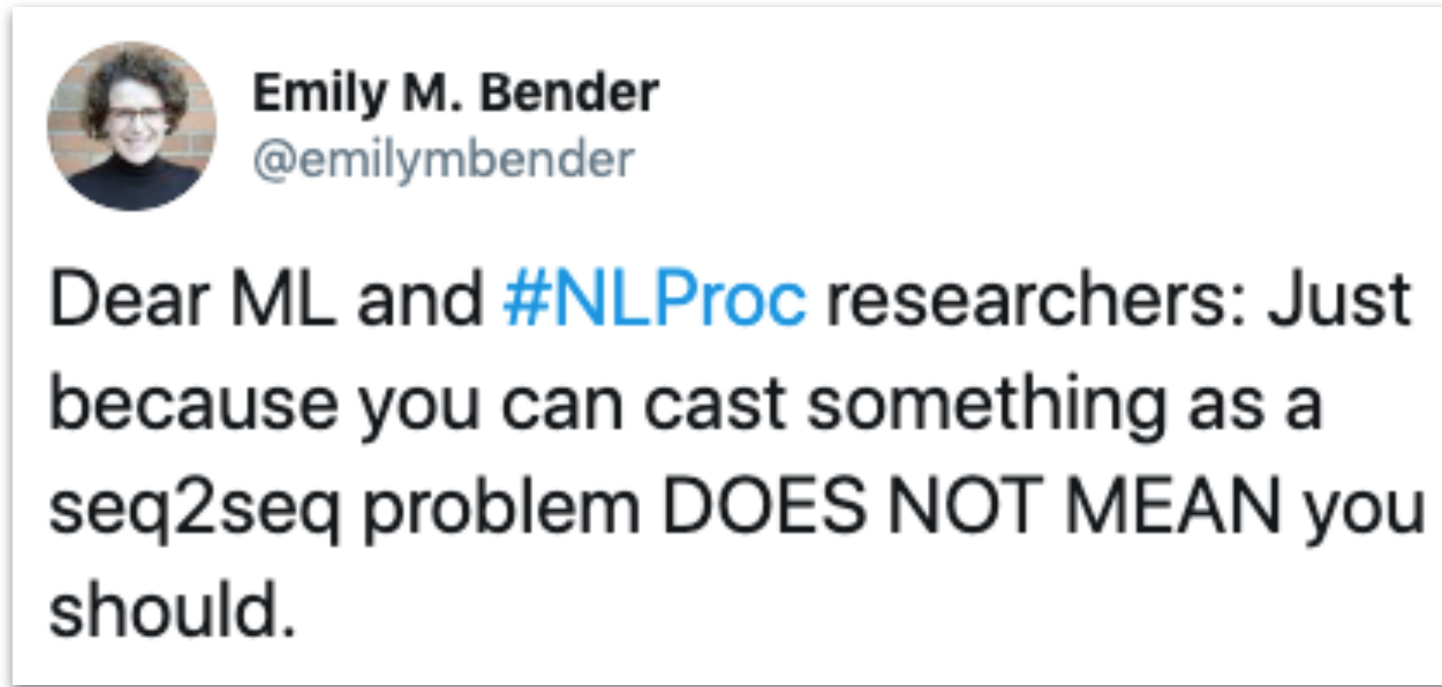
Additional Applications

- Speech recognition
- Summarization
- Text style-transfer
- Paraphrasing
- Parsing
- Dialogue



Additional Applications

- Speech recognition
- Summarization
- Text style-transfer
- Paraphrasing
- Parsing
- Dialogue
- …



The translation industry today

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

amazon

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

The translation industry today

"…the global language services and technology industry, which, according to Slator was **a 23.2 billion $** market in 2018 and projected to grow to **28.2 billion $** by 2022."

SYSTRAN

amazon

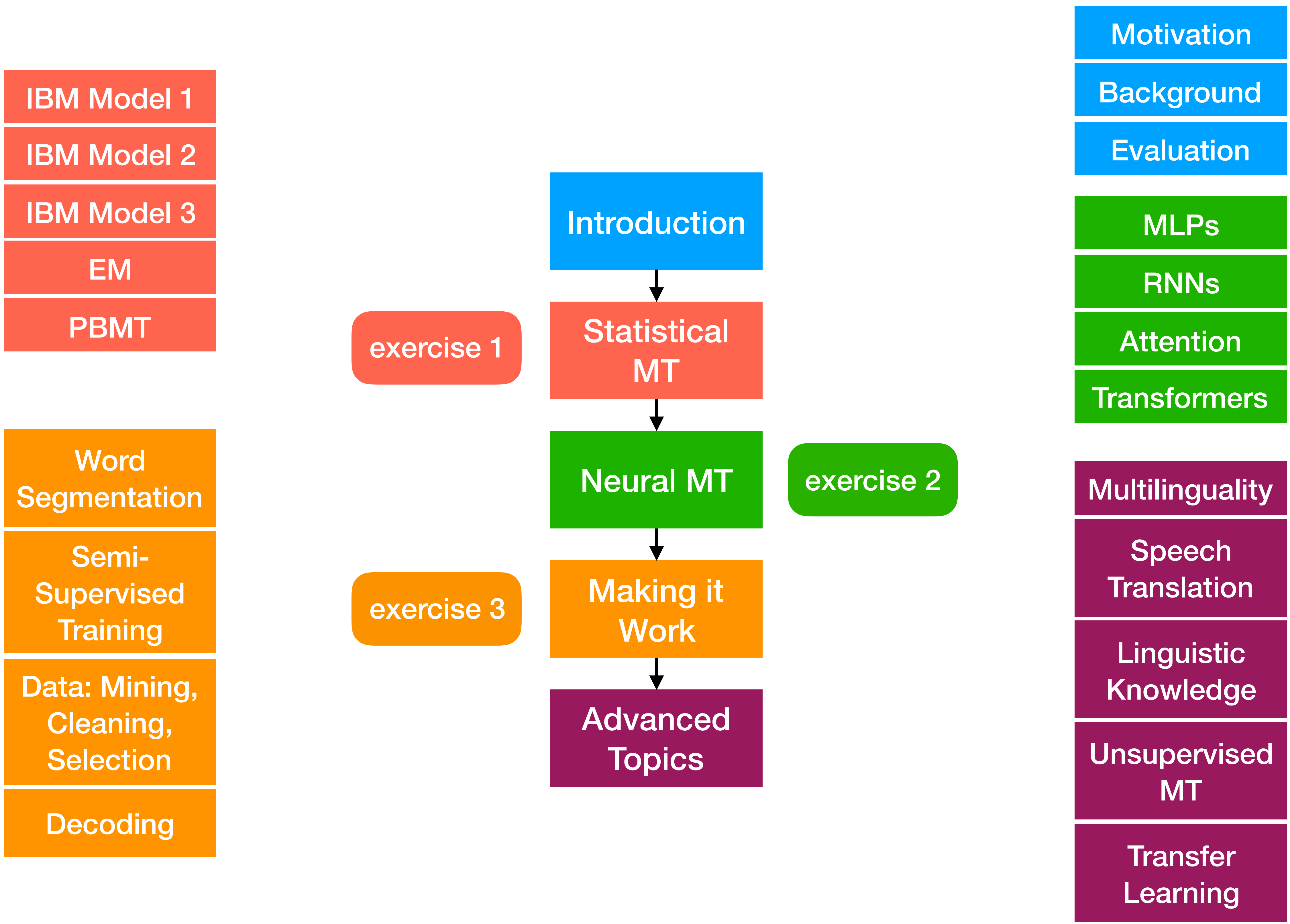
facebook

Google

SDL

Unbabel

Syllabus


IBM Model 1
IBM Model 2
IBM Model 3
EM
PBMT
Word Segmentation
Semi-Supervised Training
Data: Mining, Cleaning, Selection
Decoding
exercise 1
exercise 3
Introduction
Statistical MT
Neural MT
Making it Work
Advanced Topics
exercise 2
Motivation
Background
Evaluation
MLPs
RNNs
Attention
Transformers
Multilinguality
Speech Translation
Linguistic Knowledge
Unsupervised MT
Transfer Learning

Part I: Introduction

Part I: Introduction

- The statistical MT framework

$$\hat{E} = \underset{E}{\operatorname{argmax}}\ P(E \mid F)$$

Part I: Introduction

- The statistical MT framework
- N-gram language models

$$\hat{E} = \underset{E}{\text{argmax}}\ P(E \mid F)$$

P(|E| = 3, e_1="she", e_2="went", e_3="home") =

P(e_1="she")

* P(e_2="went" | e_1="she")

* P(e_3="home" | e_1="she", e_2="went")

* P(e_4="</s>" | e_1="she", e_2="went", e_3="home")

Part I: Introduction

- The statistical MT framework
- N-gram language models
- Evaluation
  - BLEU
  - Human

$$\hat{E} = \underset{E}{\operatorname{argmax}} \; P(E \mid F)$$

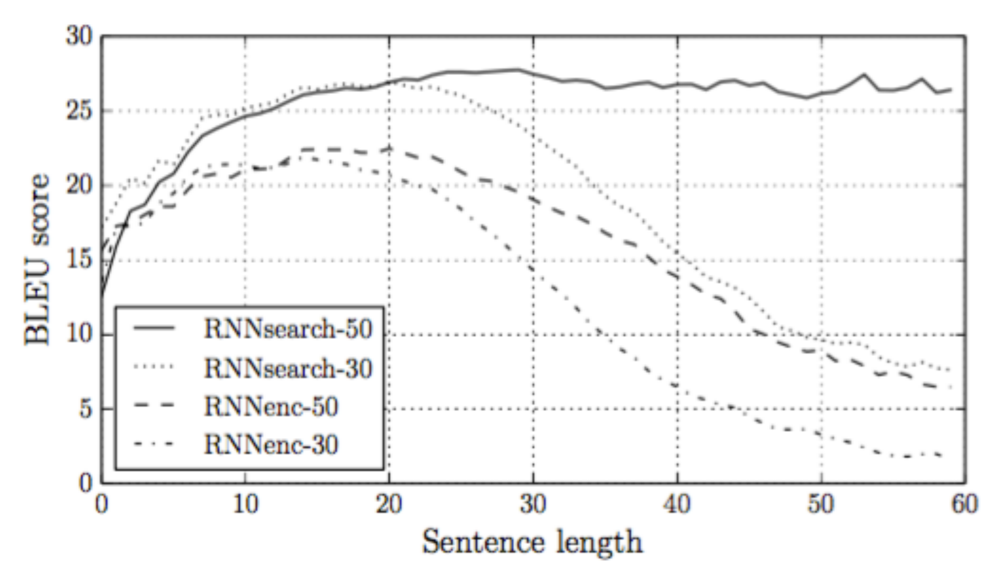
P(|E| = 3, e_1="she", e_2="went", e_3="home") =

P(e_1="she")

* P(e_2="went" | e_1="she")

* P(e_3="home" | e_1="she", e_2="went")

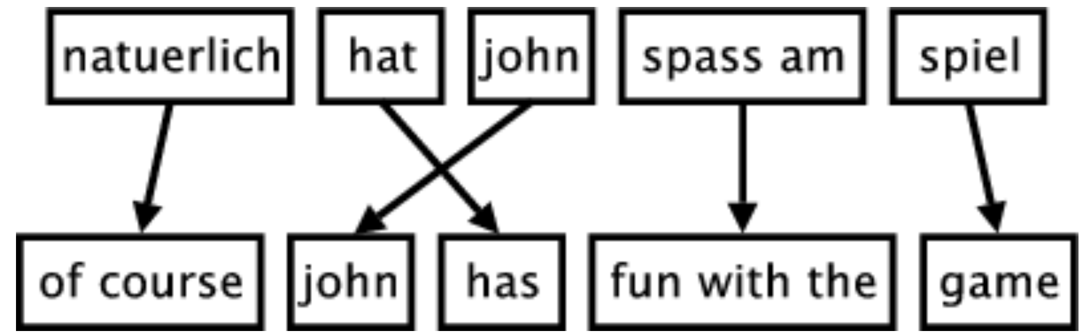
* P(e_4="</s>" | e_1="she", e_2="went", e_3="home")



Part II: Statistical Machine Translation

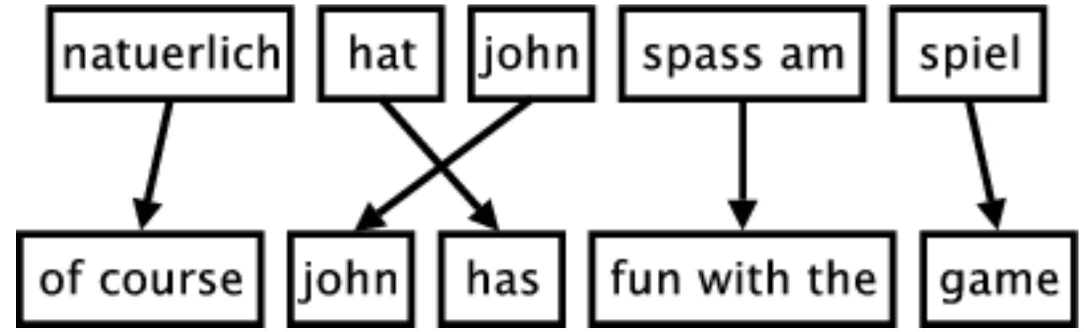
Part II: Statistical Machine Translation

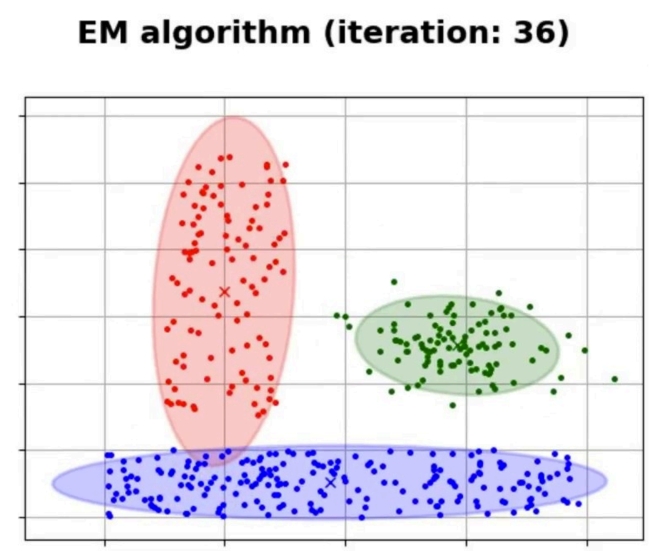
- IBM models



Part II: Statistical Machine Translation

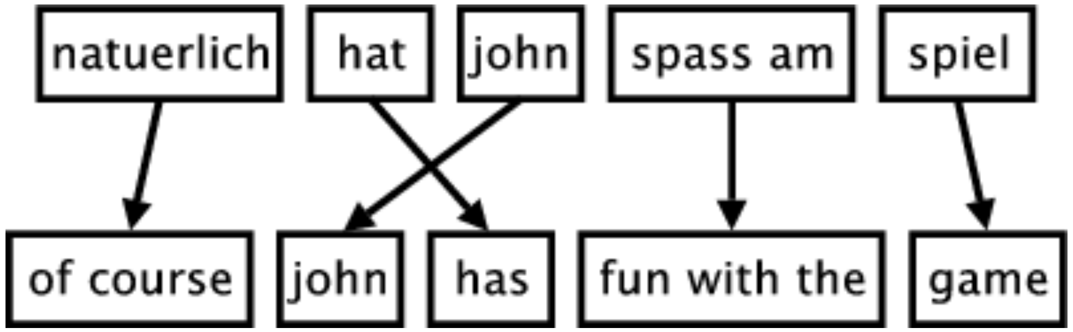
- IBM models
- The EM algorithm

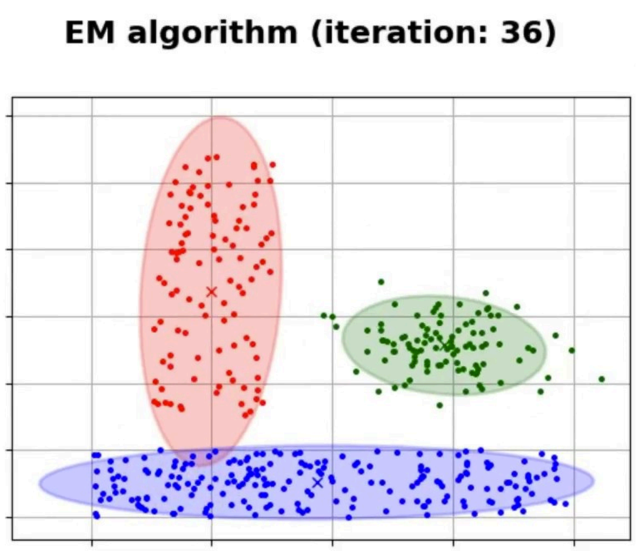




Part II: Statistical Machine Translation

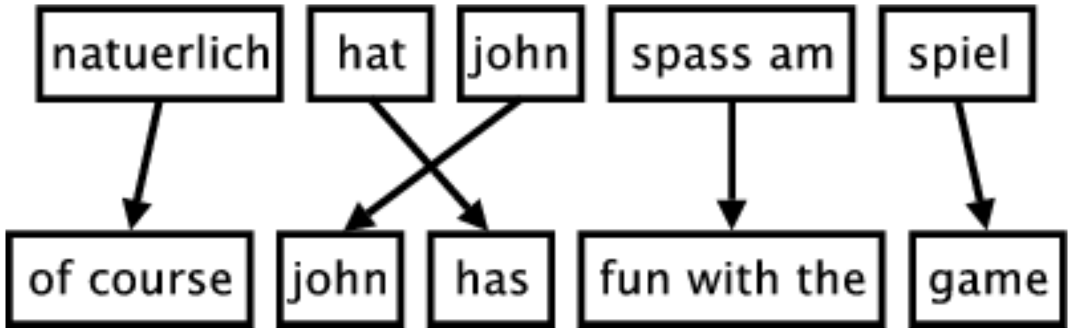
- IBM models
- The EM algorithm
- Phrase-based translation

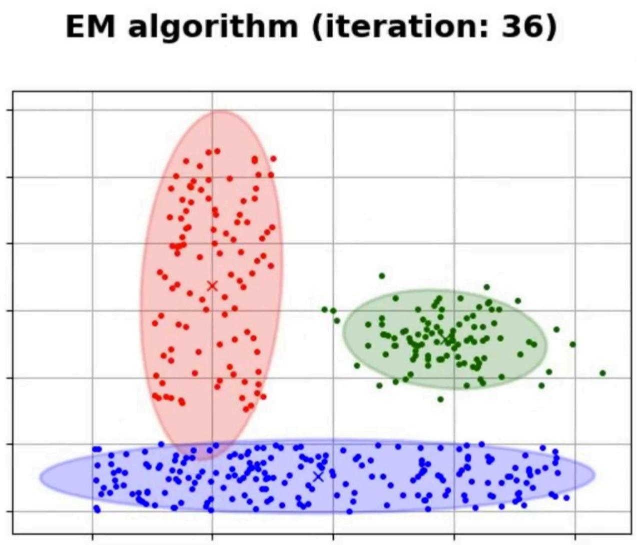


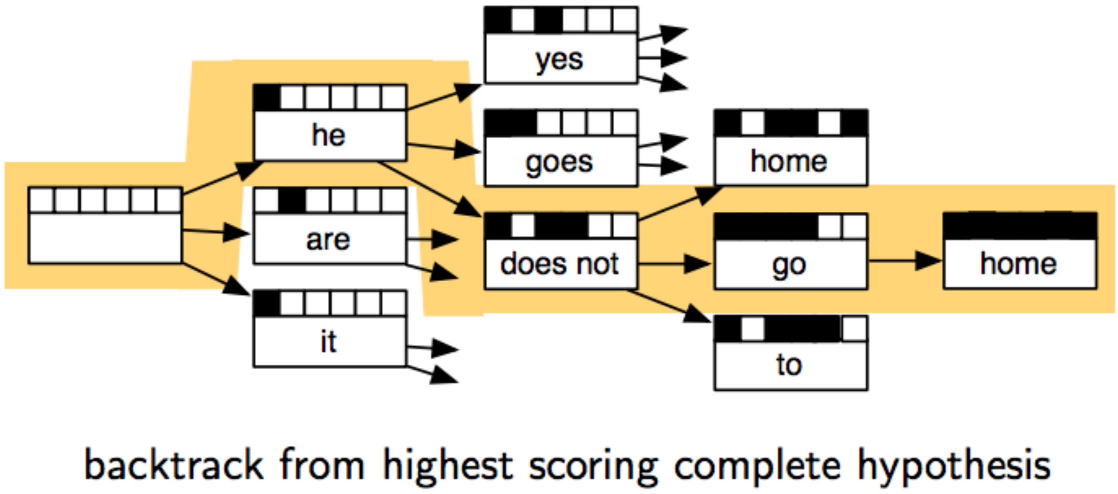


Part II: Statistical Machine Translation

- IBM models
- The EM algorithm
- Phrase-based translation
- Decoding and beam-search



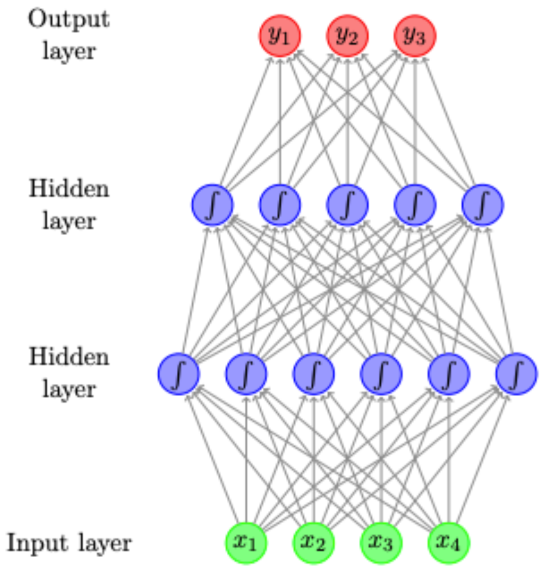




Part III: Neural Machine Translation

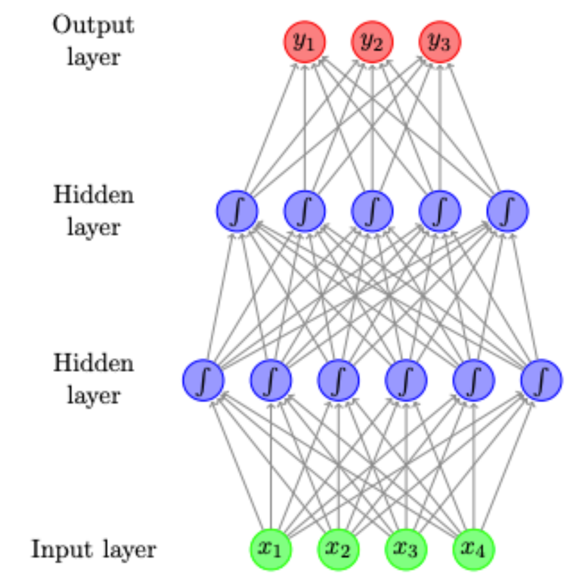
Part III: Neural Machine Translation

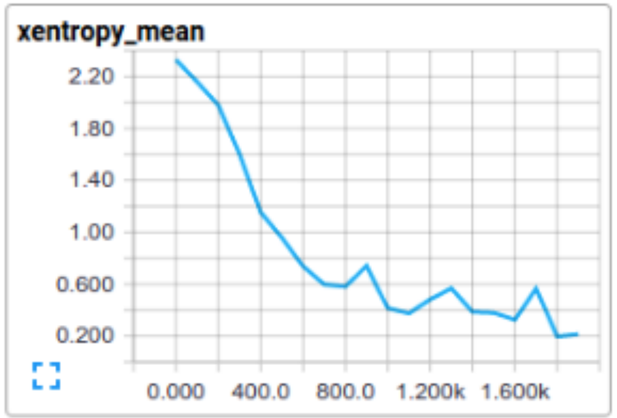
- Introduction to neural networks



Part III: Neural Machine Translation

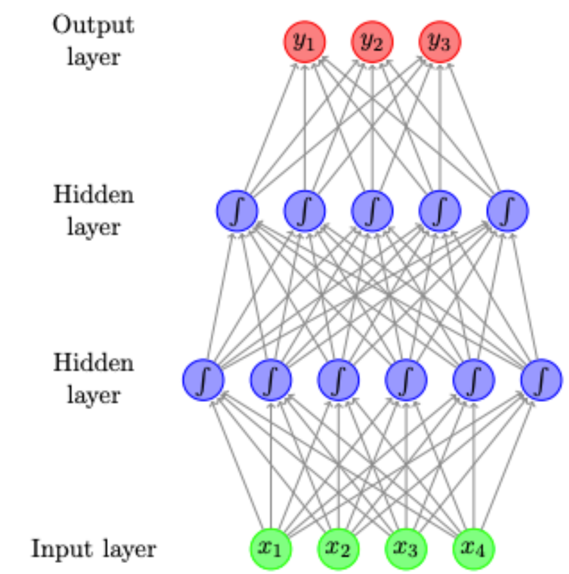
- Introduction to neural networks
    - Optimization

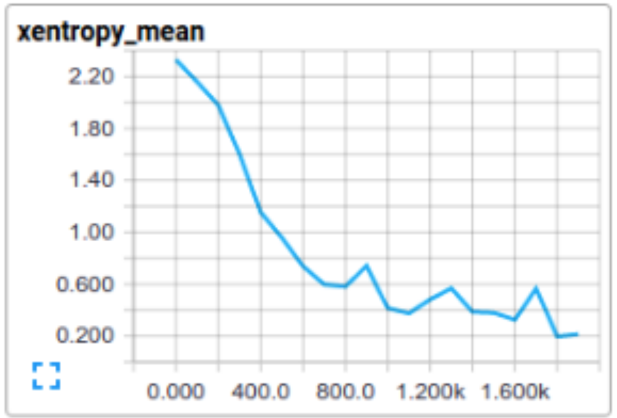




Part III: Neural Machine Translation

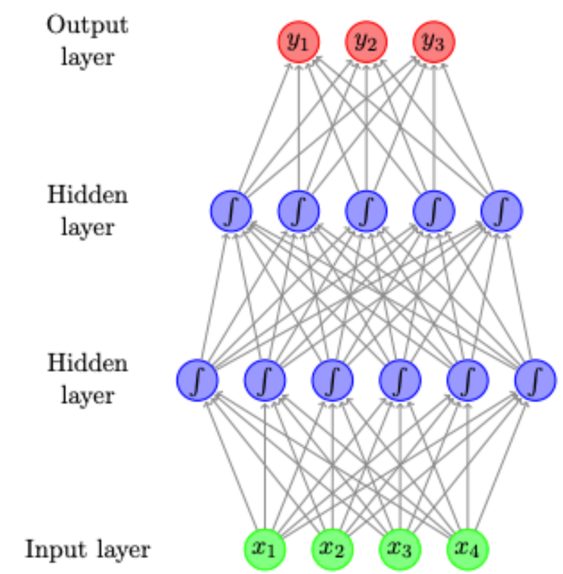
- Introduction to neural networks
  - Optimization
  - Recurrent Neural Networks

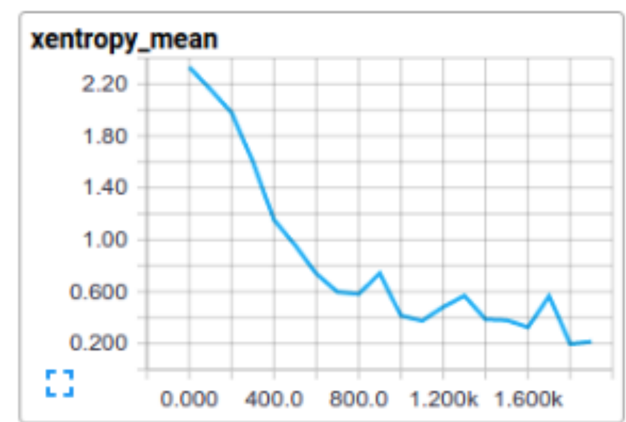




Part III: Neural Machine Translation

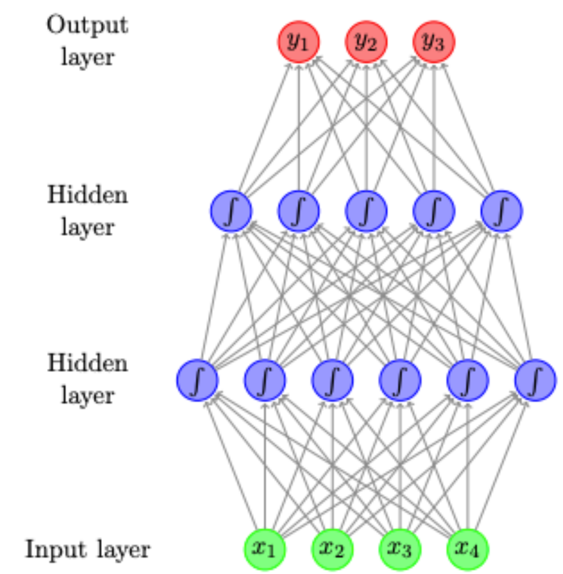
- Introduction to neural networks
  - Optimization
  - Recurrent Neural Networks
  - RNN language models

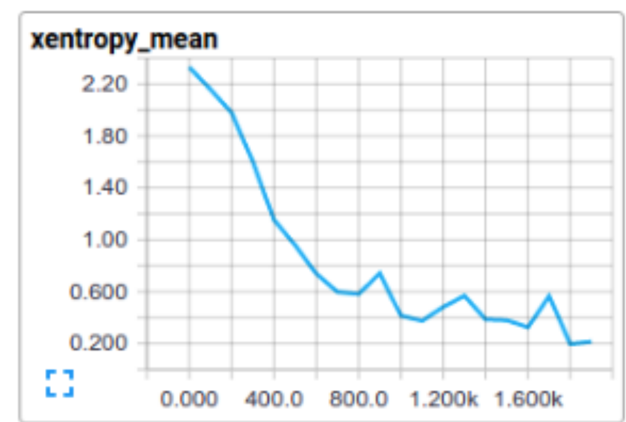


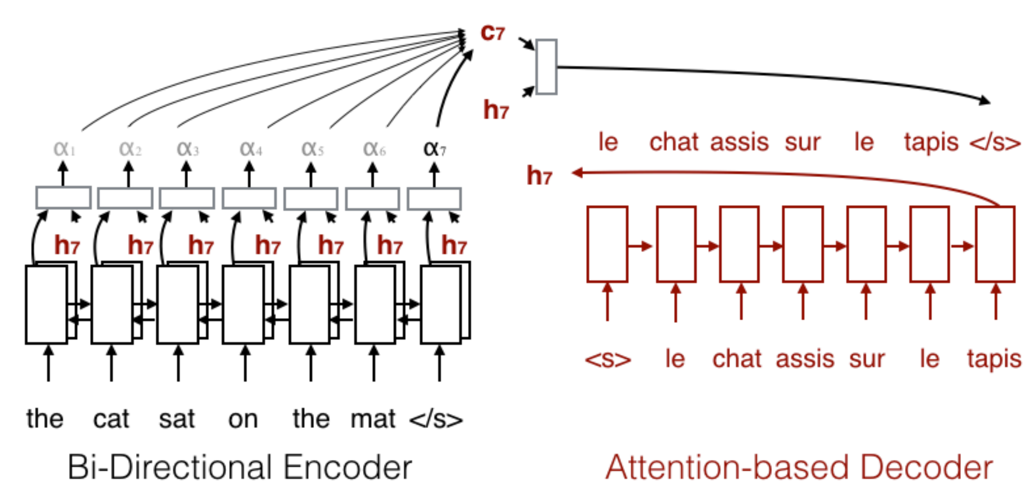


Part III: Neural Machine Translation

- Introduction to neural networks
    - Optimization
    - Recurrent Neural Networks
    - RNN language models
- Encoder-decoder models

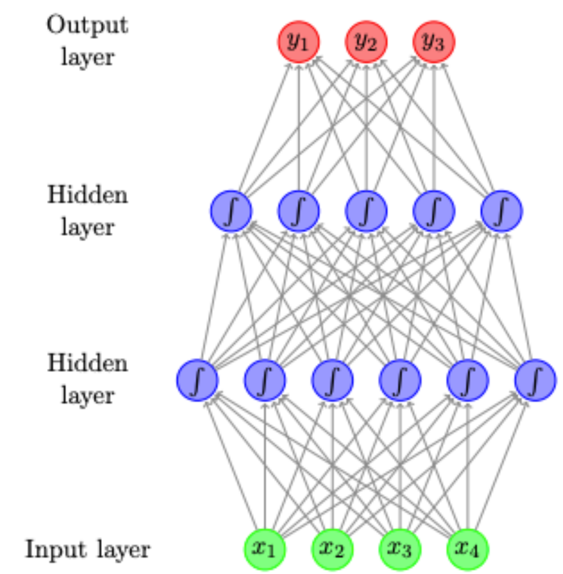


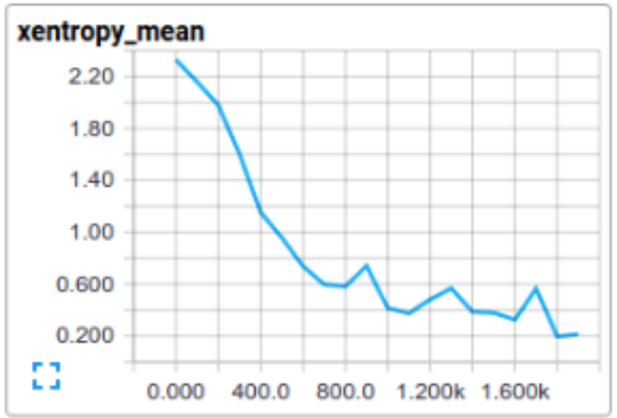


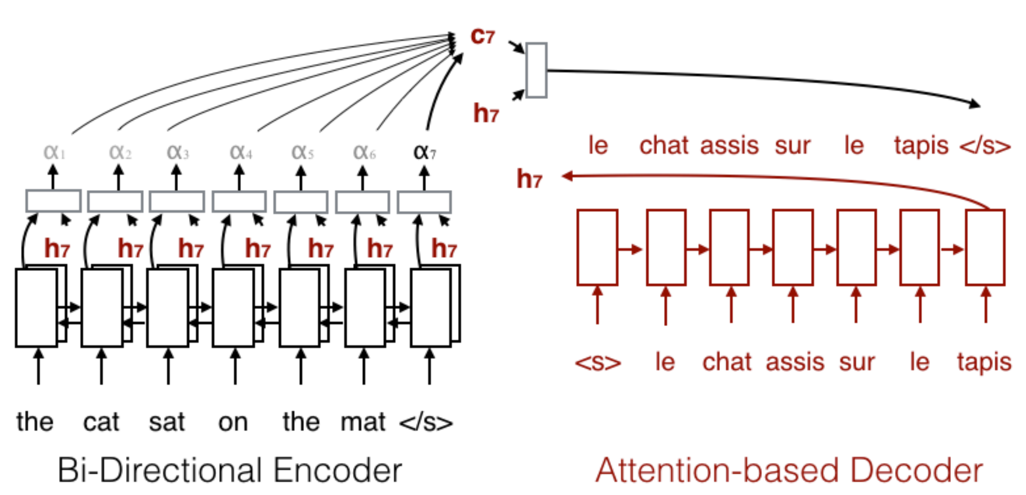


Part III: Neural Machine Translation

- Introduction to neural networks
  - Optimization
  - Recurrent Neural Networks
  - RNN language models
- Encoder-decoder models
  - Attention (is all you need?)







Making it Work

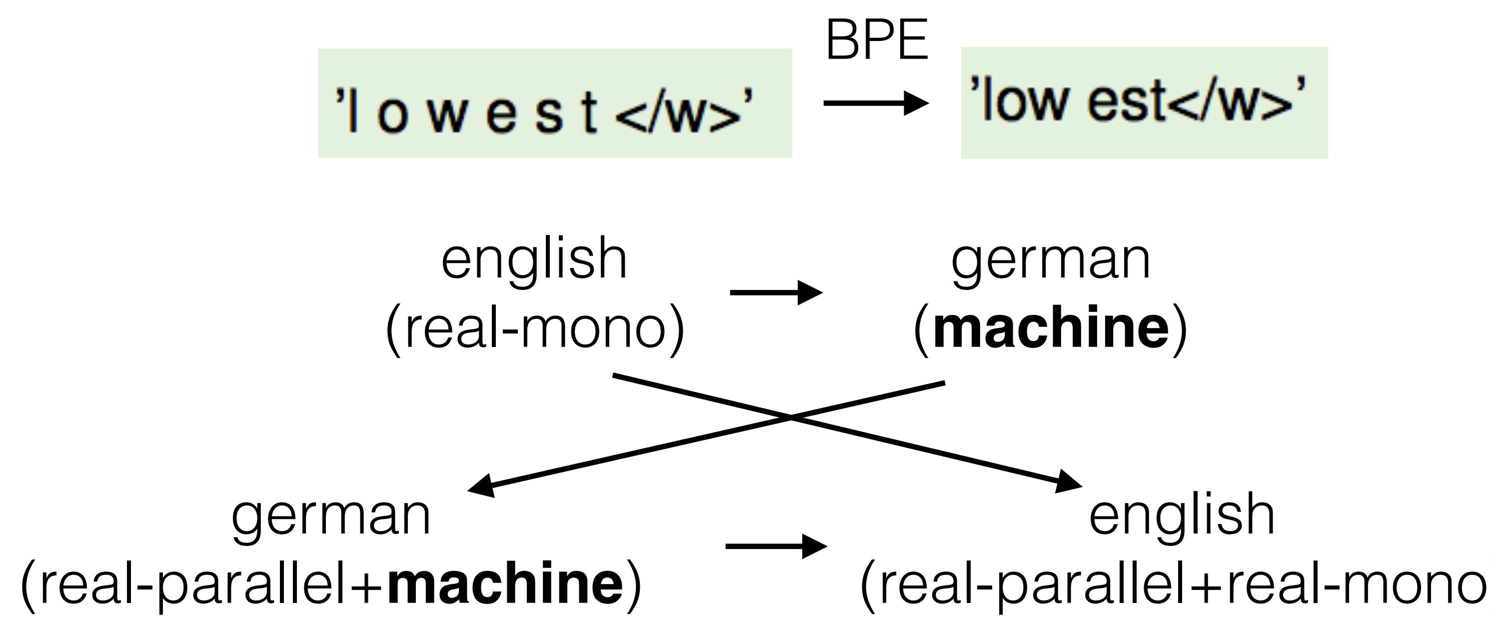
Making it Work

- Word Segmentation

'l o w e s t </w>' $\xrightarrow{\text{BPE}}$ 'low est</w>'

Making it Work

- Word Segmentation
- Semi-supervised training



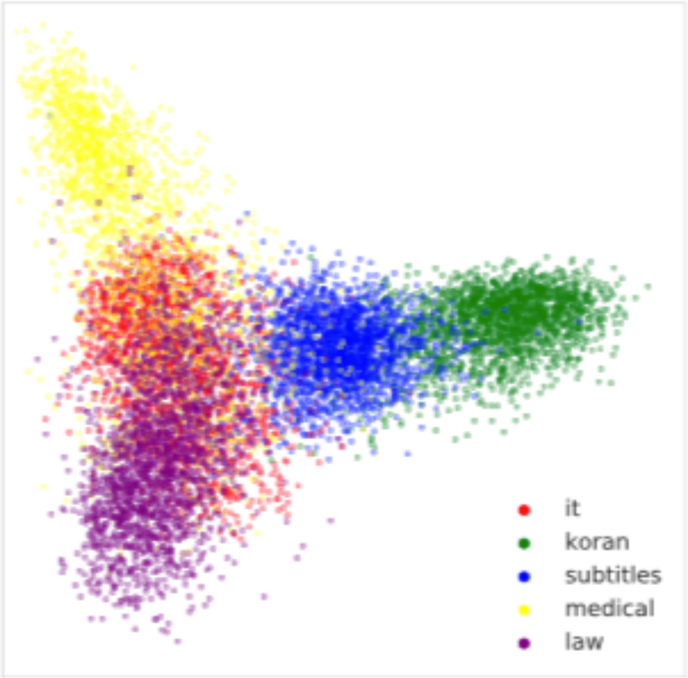
Making it Work

- Word Segmentation
- Semi-supervised training
- Respect the data!

'l o w e s t </w>' $\xrightarrow{\text{BPE}}$ 'low est</w>'

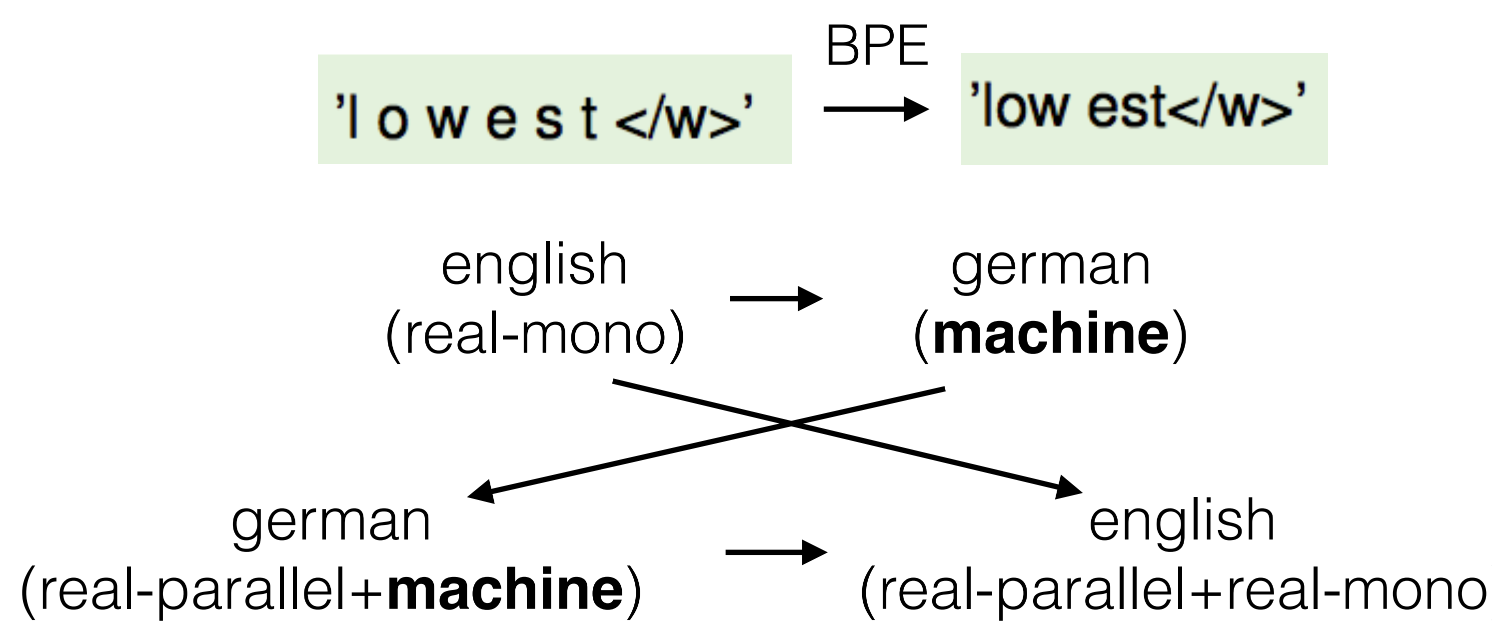
english (real-mono) → german (**machine**)

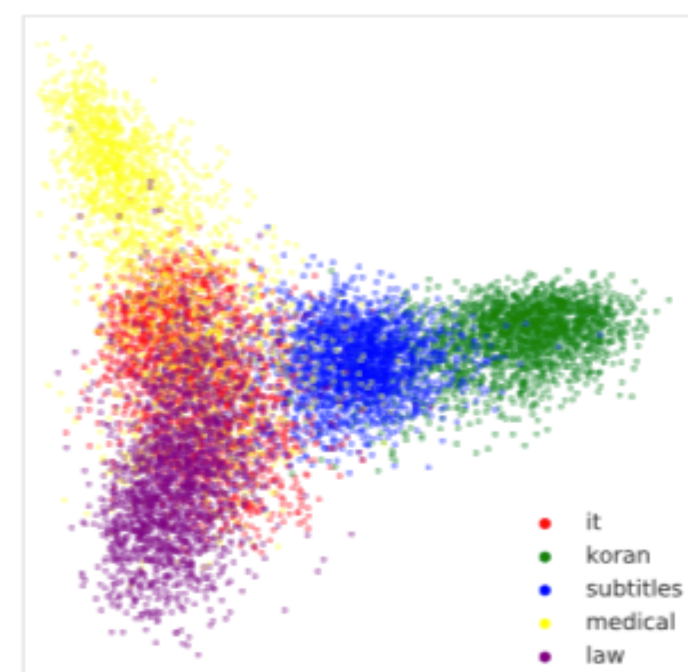
german (real-parallel+**machine**) → english (real-parallel+real-mono)



Making it Work

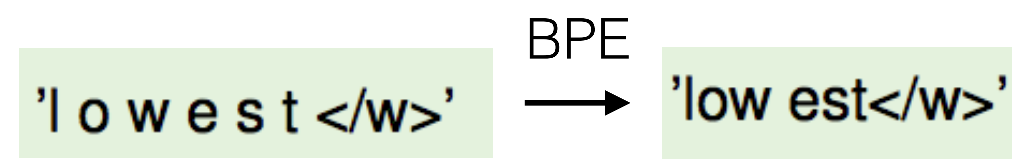
- Word Segmentation
- Semi-supervised training
- Respect the data!
  - Mining for data

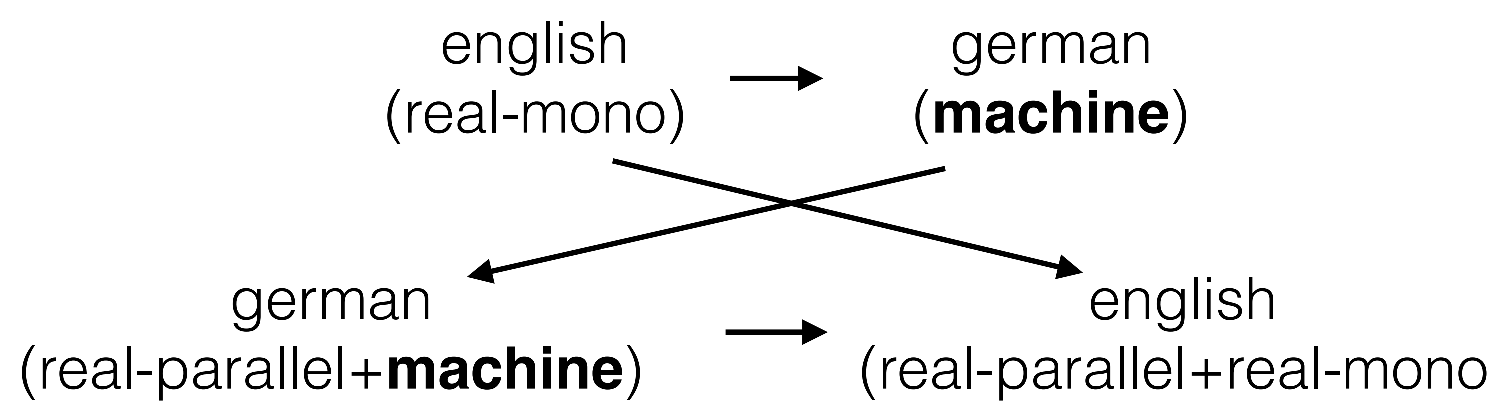


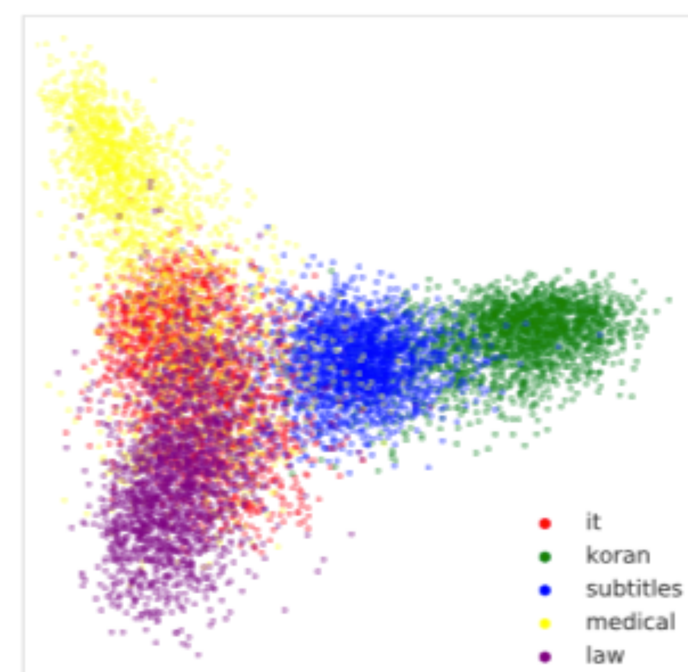


Making it Work

- Word Segmentation
- Semi-supervised training
- Respect the data!
  - Mining for data
  - Data selection







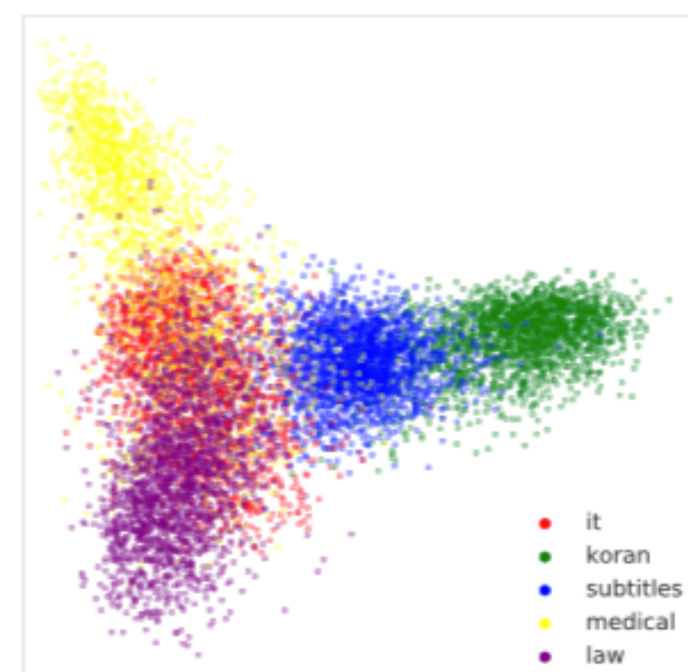
Making it Work

- Word Segmentation
- Semi-supervised training
- Respect the data!
  - Mining for data
  - Data selection
  - Data cleaning

'l o w e s t </w>' $\xrightarrow{\text{BPE}}$ 'low est</w>'

english (real-mono) → german (**machine**)

german (real-parallel+**machine**) → english (real-parallel+real-mono)



Making it Work

- Word Segmentation
- Semi-supervised training
- Respect the data!
  - Mining for data
  - Data selection
  - Data cleaning
- Decoding tricks

'l o w e s t </w>' $\xrightarrow{\text{BPE}}$ 'low est</w>'

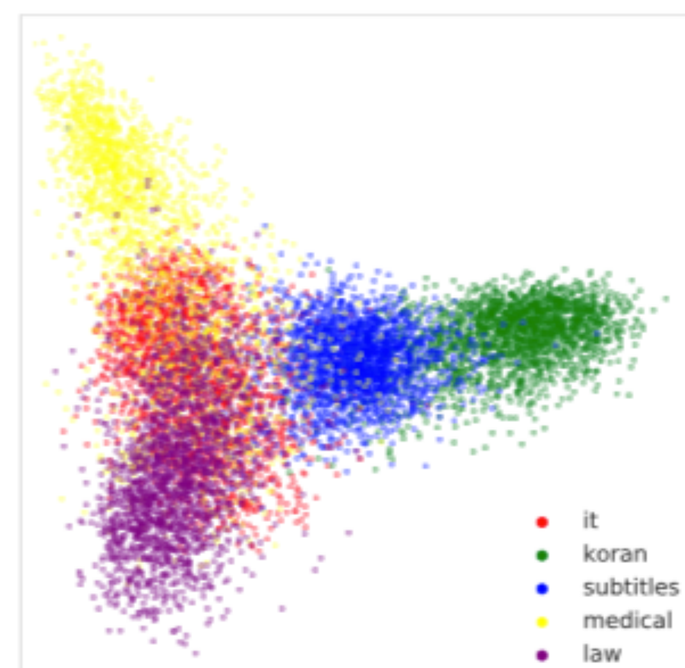
english (real-mono) → german (**machine**)

german (real-parallel+**machine**) → english (real-parallel+real-mono)

$$s(Y,X) = \log(P(Y|X))/lp(Y) + cp(X;Y)$$

$$lp(Y) = \frac{(5+|Y|)^\alpha}{(5+1)^\alpha}$$

$$cp(X;Y) = \beta * \sum_{i=1}^{|X|} \log(\min(\sum_{j=1}^{|Y|} p_{i,j}, 1.0)),$$



Advanced Topics

Advanced Topics

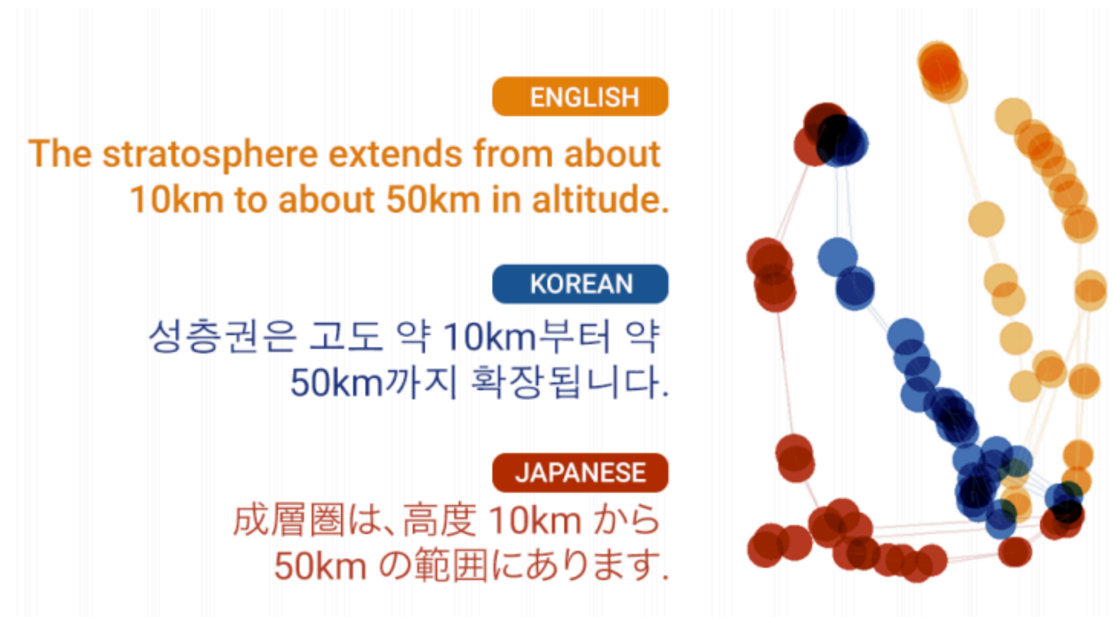
- Multilingual NMT

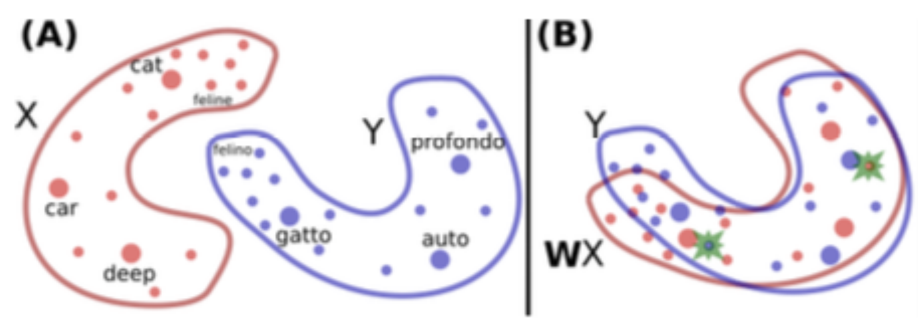
Advanced Topics

- Multilingual NMT
- Pretraining and Transfer Learning

Advanced Topics

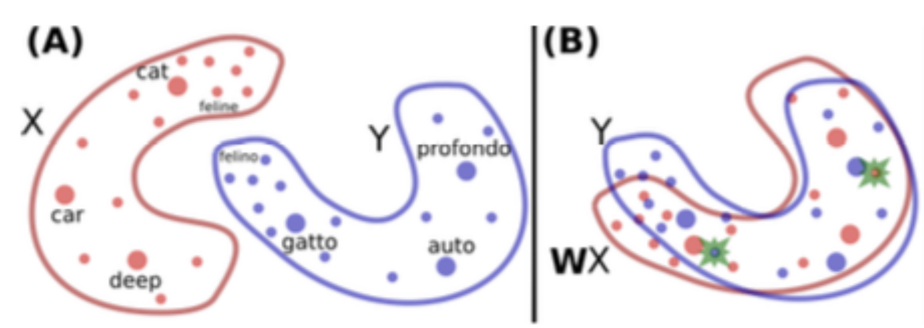
- Multilingual NMT
- Pretraining and Transfer Learning
- Unsupervised MT

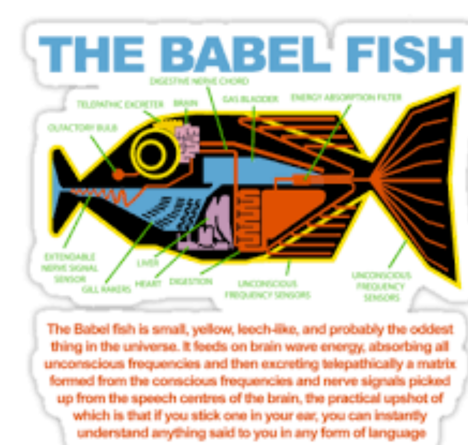




Advanced Topics

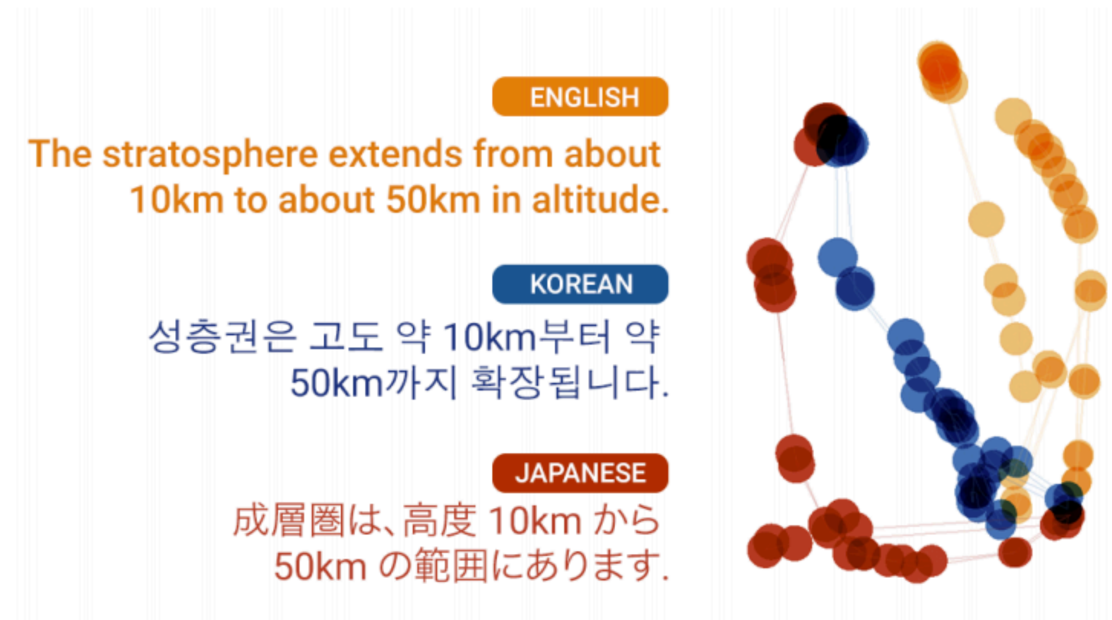
- Multilingual NMT
- Pretraining and Transfer Learning
- Unsupervised MT
- Speech Translation

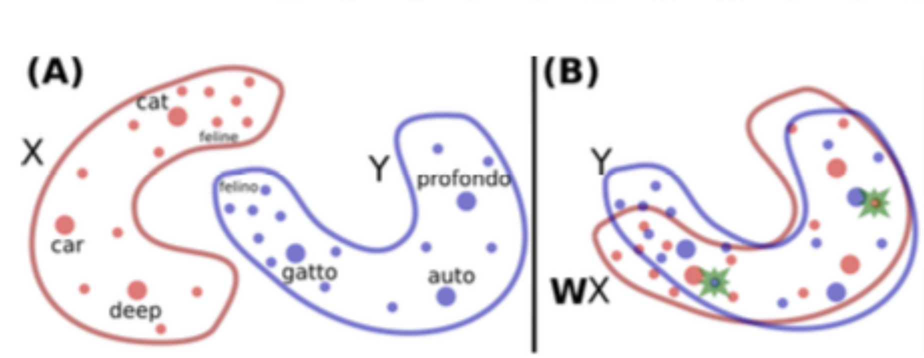


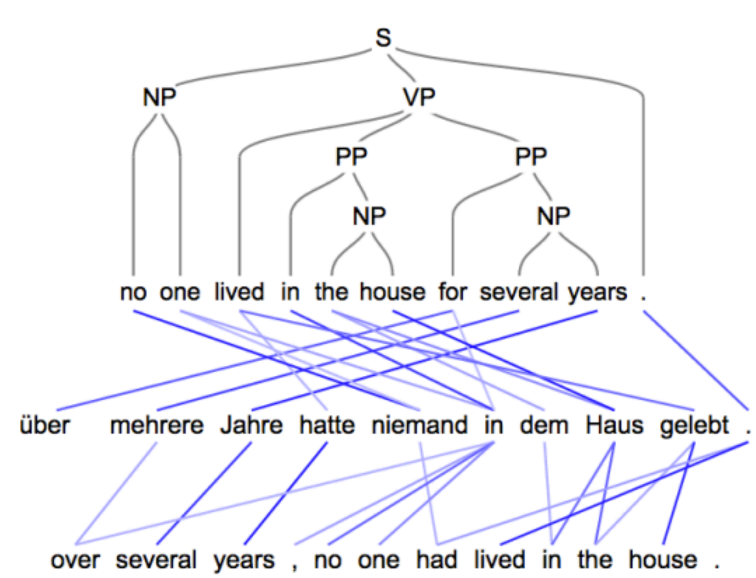


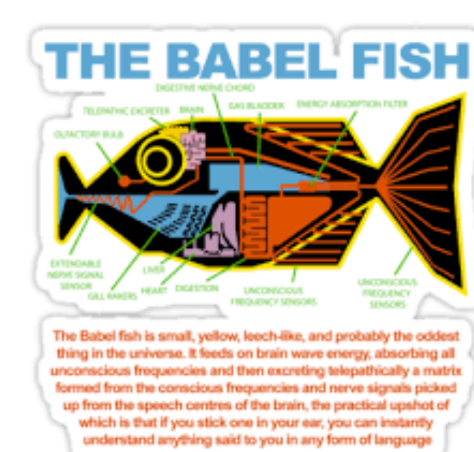
Advanced Topics

- Multilingual NMT
- Pretraining and Transfer Learning
- Unsupervised MT
- Speech Translation
- Integrating Linguistic Knowledge





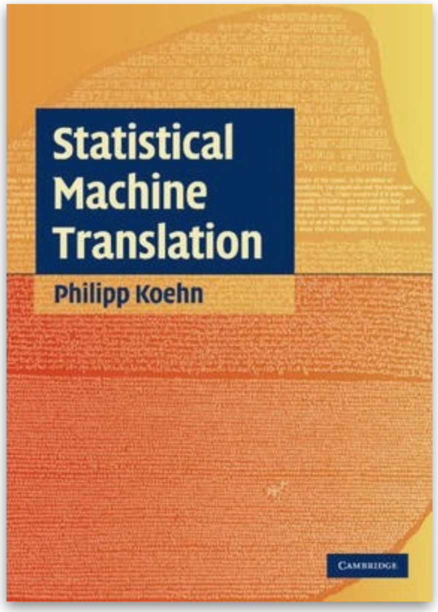




Relevant References

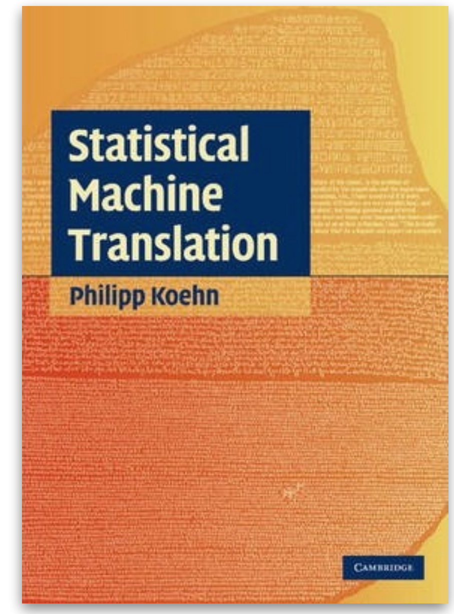
Relevant References

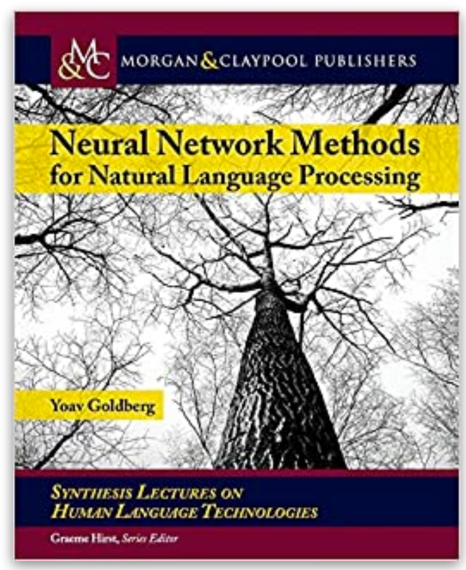
- Phillip Koehn's SMT book (available in the library)



Relevant References

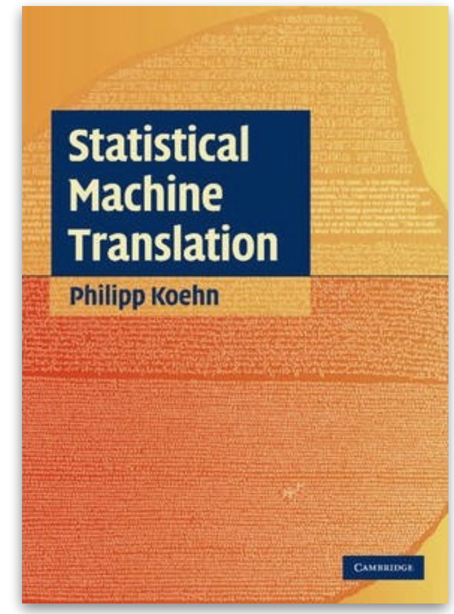
- Phillip Koehn's SMT book (available in the library)
- Yoav Goldberg's primer (free) and book

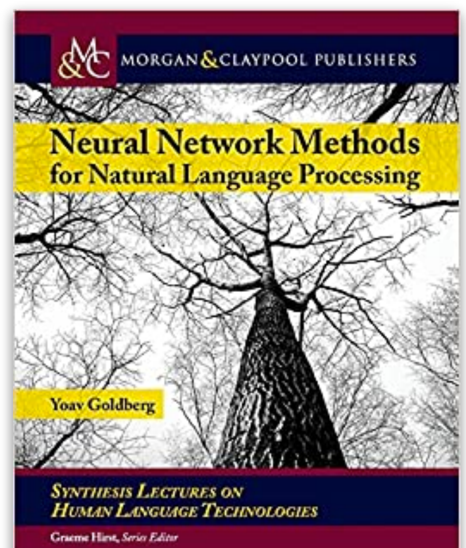


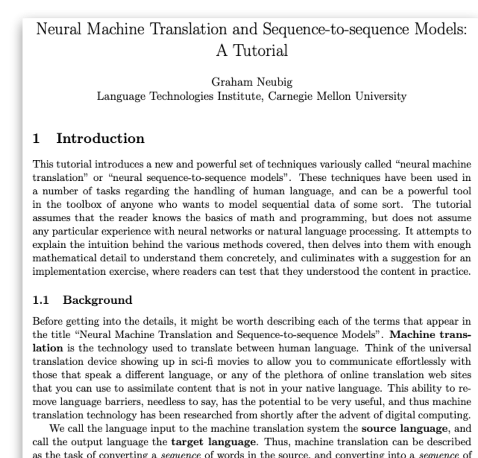


Relevant References

- Phillip Koehn’s SMT book (available in the library)
- Yoav Goldberg’s primer (free) and book
- Graham Neubig’s tutorial on sequence models







Neural Machine Translation and Sequence-to-sequence Models: A Tutorial

Graham Neubig
Language Technologies Institute, Carnegie Mellon University

1 Introduction

This tutorial introduces a new and powerful set of techniques variously called “neural machine translation” or “neural sequence-to-sequence models”. These techniques have been used in a number of tasks regarding the handling of human language, and can be a powerful tool in the toolbox of anyone who wants to model sequential data of some sort. The tutorial assumes that the reader knows the basics of math and programming, but does not assume any particular experience with neural networks or natural language processing. It attempts to explain the intuition behind the various methods covered, then delves into them with enough mathematical detail to understand them concretely, and culiminates with a suggestion for an implementation exercise, where readers can test that they understood the content in practice.

1.1 Background

Before getting into the details, it might be worth describing each of the terms that appear in the title “Neural Machine Translation and Sequence-to-sequence Models”. **Machine translation** is the technology used to translate between human language. Think of the universal translation device showing up in sci-fi movies to allow you to communicate effortlessly with those that speak a different language, or any of the plethora of online translation web sites that you can use to assimilate content that is not in your native language. This ability to remove language barriers, needless to say, has the potential to be very useful, and thus machine translation technology has been researched from shortly after the advent of digital computing.

We call the language input to the machine translation system the **source language**, and call the output language the **target language**. Thus, machine translation can be described as the task of converting a *sequence* of words in the source, and converting into a *sequence* of words in the target. The goal of the machine translation practitioner is to come up with an

Summary

Summary

- Machine translation is **hard**

Summary

- Machine translation is **hard**
- Machine translation is **useful** and **important**

Summary

- Machine translation is **hard**
- Machine translation is **useful** and **important**
- A lot has changed in the recent years…

Summary

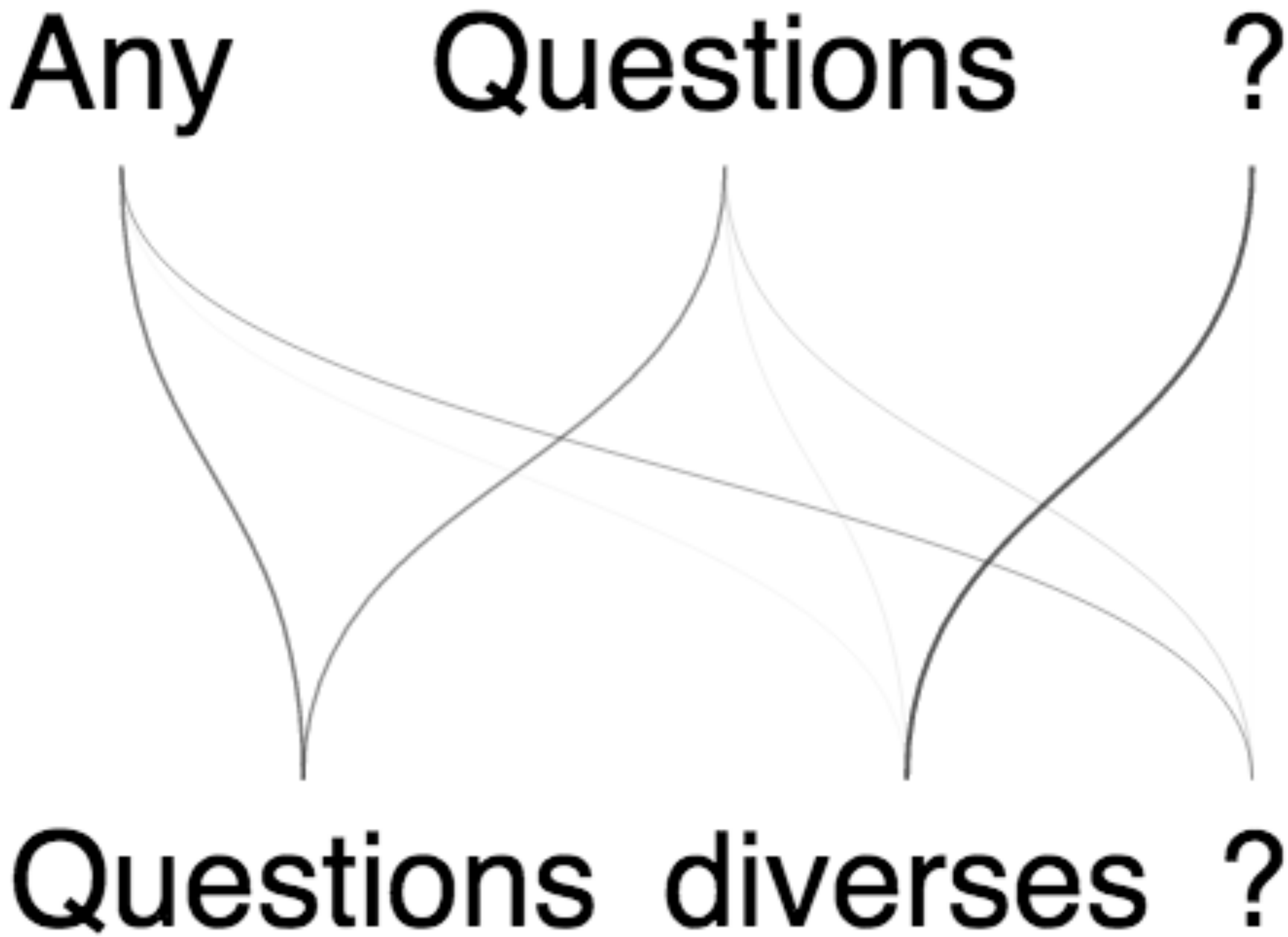
- Machine translation is **hard**
- Machine translation is **useful** and **important**
- A lot has changed in the recent years…
    - We have **a lot to cover**

Summary

- Machine translation is **hard**
- Machine translation is **useful** and **important**
- A lot has changed in the recent years…
  - We have **a lot to cover**
- Your **feedback** is important

Summary

- Machine translation is **hard**
- Machine translation is **useful** and **important**
- A lot has changed in the recent years…
    - We have **a lot to cover**
- Your **feedback** is important
- Looking forward to this semester, stay safe!


Any Questions ?
Questions diverses ?