Christof Monz
Informatics Institute
University of Amsterdam

Natural Language Processing 1

Machine Translation

This Class

- Machine translation

This Class

- Machine translation
- Sequence-to-sequence models

This Class

- Machine translation
- Sequence-to-sequence models
- Neural machine translation

This Class

- Machine translation
- Sequence-to-sequence models
- Neural machine translation
  - encoder-decoder architecture

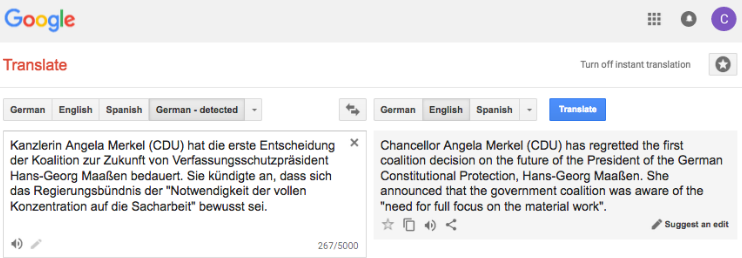
This Class

- Machine translation
- Sequence-to-sequence models
- Neural machine translation
  - encoder-decoder architecture
  - attention mechanism

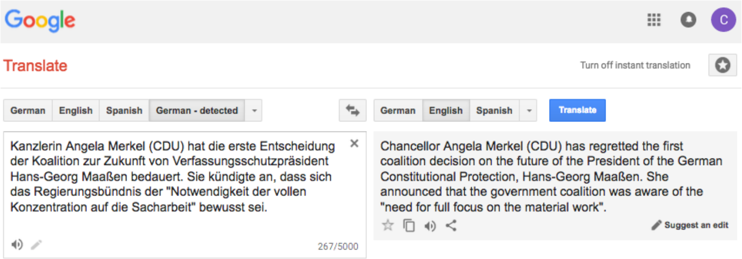
This Class

- Machine translation
- Sequence-to-sequence models
- Neural machine translation
  - encoder-decoder architecture
  - attention mechanism
  - self-attention (Google's Transformer)

Machine Translation

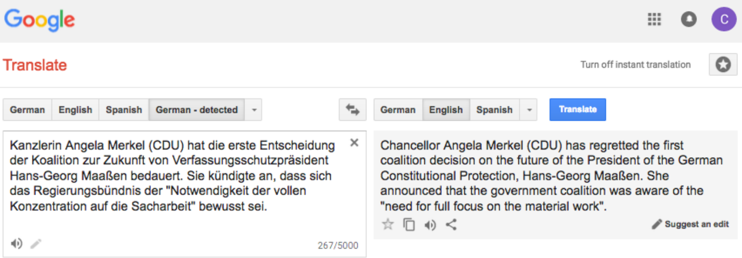


Machine Translation



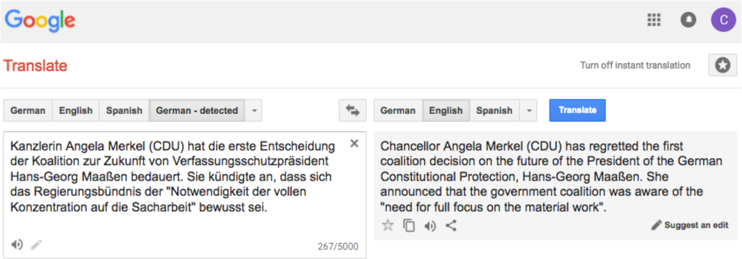
- Active research of AI since its beginnings

Machine Translation



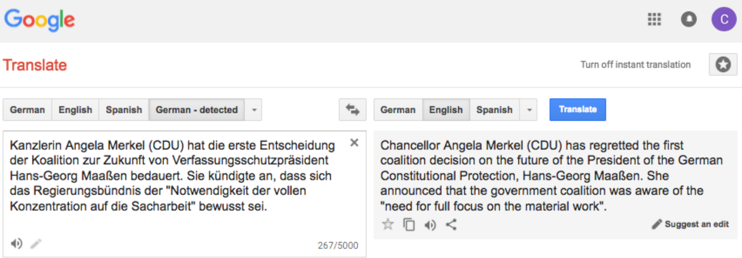
- Active research of AI since its beginnings
- Machine translation (MT) is a nice example of the different paradigm shifts in AI

Machine Translation



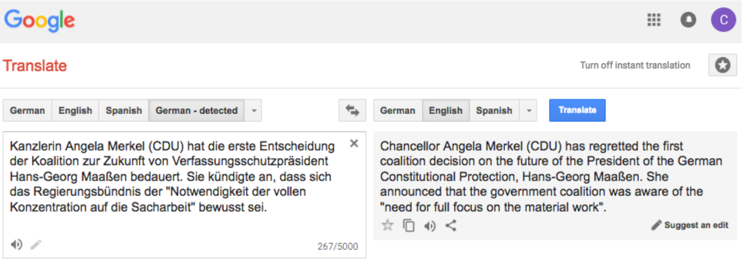
- Active research of AI since its beginnings
- Machine translation (MT) is a nice example of the different paradigm shifts in AI
  - 1950s–1990s: rule-based, symbolic approaches

Machine Translation



- Active research of AI since its beginnings
- Machine translation (MT) is a nice example of the different paradigm shifts in AI
  - 1950s–1990s: rule-based, symbolic approaches
  - 1990s–2016: statistical, data-driven approaches

Machine Translation



- Active research of AI since its beginnings
- Machine translation (MT) is a nice example of the different paradigm shifts in AI
  - 1950s–1990s: rule-based, symbolic approaches
  - 1990s–2016: statistical, data-driven approaches
  - 2014–now: neural, deep learning, data-driven approaches

MT: German to English (high resource)

German source sentence

Die Leitung der für die US-Regierungsgebäude zuständigen Behörde weigert sich laut einem Medienbericht, einen Brief zu unterschreiben, mit dem das Biden-Übergangsteam Zugang zu US-Behörden erhalten und formal diese Woche die Arbeit aufnehmen kann.

MT: German to English (high resource)

German source sentence

Die Leitung der für die US-Regierungsgebäude zuständigen Behörde weigert sich laut einem Medienbericht, einen Brief zu unterschreiben, mit dem das Biden-Übergangsteam Zugang zu US-Behörden erhalten und formal diese Woche die Arbeit aufnehmen kann.

English machine translation anno 2014 (using statistical machine translation)

MT: German to English (high resource)

German source sentence

Die Leitung der für die US-Regierungsgebäude zuständigen Behörde weigert sich laut einem Medienbericht, einen Brief zu unterschreiben, mit dem das Biden-Übergangsteam Zugang zu US-Behörden erhalten und formal diese Woche die Arbeit aufnehmen kann.

English machine translation anno 2014 (using statistical machine translation)

The line of the authority responsible for the US Government buildings refuses according to a medium report signing a letter with which the Biden Übergangsteam entrance to US authorities to receive and formally this week the work take up can.

MT: German to English (high resource)

German source sentence

Die Leitung der für die US-Regierungsgebäude zuständigen Behörde weigert sich laut einem Medienbericht, einen Brief zu unterschreiben, mit dem das Biden-Übergangsteam Zugang zu US-Behörden erhalten und formal diese Woche die Arbeit aufnehmen kann.

English machine translation anno 2014 (using statistical machine translation)

The line of the authority responsible for the US Government buildings refuses according to a medium report signing a letter with which the Biden Übergangsteam entrance to US authorities to receive and formally this week the work take up can.

English machine translation in 2020 (using neural machine translation)

MT: German to English (high resource)

German source sentence

Die Leitung der für die US-Regierungsgebäude zuständigen Behörde weigert sich laut einem Medienbericht, einen Brief zu unterschreiben, mit dem das Biden-Übergangsteam Zugang zu US-Behörden erhalten und formal diese Woche die Arbeit aufnehmen kann.

English machine translation anno 2014 (using statistical machine translation)

The line of the authority responsible for the US Government buildings refuses according to a medium report signing a letter with which the Biden Übergangsteam entrance to US authorities to receive and formally this week the work take up can.

English machine translation in 2020 (using neural machine translation)

According to a media report, the management of the agency responsible for US government buildings is refusing to sign a letter that will allow the Biden transition team to gain access to US authorities and formally start work this week.

MT: Kurdish to English (low resource)

Kurdish source sentence

Hinek werzişvanên Îraqî yên ku li ser destê komên tundrew astender bûne serketin di werzişvanîyê de pêk anîn bi rêya beşdarîkirina di qehremanîyên werzişvanî yên astenderan de.

MT: Kurdish to English (low resource)

Kurdish source sentence

Hinek werzişvanên Îraqî yên ku li ser destê komên tundrew astender bûne serketin di werzişvanîyê de pêk anîn bi rêya beşdarîkirina di qehremanîyên werzişvanî yên astenderan de.

English machine translation in 2020 (using neural machine translation)

MT: Kurdish to English (low resource)

Kurdish source sentence

Hinek werzişvanên Îraqî yên ku li ser destê komên tundrew astender bûne serketin di werzişvanîyê de pêk anîn bi rêya beşdarîkirina di qehremanîyên werzişvanî yên astenderan de.

English machine translation in 2020 (using neural machine translation)

Some Iraqi athletes who have been successful at the hands of extremist groups have achieved success in sports by participating in the sports championships of the demonstrators.

MT: Kurdish to English (low resource)

Kurdish source sentence

Hinek werzişvanên Îraqî yên ku li ser destê komên tundrew astender bûne serketin di werzişvanîyê de pêk anîn bi rêya beşdarîkirina di qehremanîyên werzişvanî yên astenderan de.

English machine translation in 2020 (using neural machine translation)

Some Iraqi athletes who have been successful at the hands of extremist groups have achieved success in sports by participating in the sports championships of the demonstrators.

Human translation (reference or ground truth)

MT: Kurdish to English (low resource)

Kurdish source sentence

Hinek werzişvanên Îraqî yên ku li ser destê komên tundrew astender bûne serketin di werzişvanîyê de pêk anîn bi rêya beşdarîkirina di qehremanîyên werzişvanî yên astenderan de.

English machine translation in 2020 (using neural machine translation)

Some Iraqi athletes who have been successful at the hands of extremist groups have achieved success in sports by participating in the sports championships of the demonstrators.

Human translation (reference or ground truth)

Some Iraqis who suffered debilitating injuries at the hands of extremist groups have gone on to achieve victory in the athletic field through their participation in paralympic sports.

Machine Translation

- Automatically translate: source language $\rightarrow$ target language

Machine Translation

- Automatically translate: source language $\rightarrow$ target language

Arabic $\rightarrow$ English	French $\rightarrow$ Spanish	...	Amharic $\rightarrow$ Vietnamese
Armenian $\rightarrow$ Czech	Armenian $\rightarrow$ Danish	...	Armenian $\rightarrow$ Turkish
$\vdots$	$\vdots$	...	$\vdots$
Uzbek $\rightarrow$ Albanian	Uzbek $\rightarrow$ Hindi	...	Uzbek $\rightarrow$ Ukranian
Vietnamese $\rightarrow$ Azeri	Vietnamese $\rightarrow$ Greek	...	Vietnamese $\rightarrow$ Turkish

Universal Translation

Universal Translation

	ar	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	id	it	ja	ko	ms	nl	no	pl	pt	ru	tr	uk	vi
ar	-	-	10.8	13.8	14.6	15.2	27.7	-	8.0	6.2	-	9.8	10.9	16.0	17.6	7.4	1.9	8.7	14.7	14.4	9.1	19.5	13.5	7.2	6.2	17.7
bg	-	-	15.9	21.3	19.1	20.2	32.3	-	8.4	9.5	25.8	11.4	12.6	18.7	19.8	-	2.2	9.7	19.0	19.0	12.4	22.4	16.5	8.7	10.8	19.4
cs	5.6	18.1	-	18.7	17.9	16.5	25.0	-	7.1	10.6	22.2	8.9	11.4	15.7	16.9	-	2.5	6.7	18.3	19.8	13.1	18.5	15.3	7.9	9.5	16.8
da	5.9	22.4	16.4	-	-	-	42.3	-	8.0	13.5	28.1	11.7	13.9	20.3	22.7	-	2.7	11.7	25.8	27.5	14.7	25.2	17.5	9.2	8.2	18.8
de	7.6	21.3	17.4	-	-	18.9	31.6	-	8.7	11.8	26.8	12.2	16.1	19.9	21.7	8.9	2.9	10.6	24.4	18.6	13.7	23.4	16.6	10.0	10.8	19.8
el	8.1	21.1	13.4	-	18.3	-	31.6	-	-	10.0	26.9	11.4	6.5	19.1	21.4	-	2.1	-	19.8	21.1	-	22.4	15.2	8.9	8.8	-
en	15.7	33.9	23.1	41.2	30.5	32.8	-	39.7	15.2	16.0	41.2	23.1	24.9	32.5	33.4	11.3	4.1	23.4	31.9	41.2	17.0	38.8	20.1	15.8	17.9	28.9
es	-	-	-	-	-	-	38.6	-	10.0	11.7	-	13.8	15.9	22.7	28.6	-	3.2	-	24.2	22.4	14.1	31.5	17.0	11.2	12.3	23.2
fa	6.5	13.6	9.3	13.2	12.9	-	25.1	16.3	-	5.2	18.6	7.2	8.8	15.0	14.8	-	1.9	8.2	13.4	10.4	7.8	16.8	11.4	8.1	5.4	16.8
fi	3.2	10.2	9.6	12.7	10.9	9.4	15.7	12.5	3.0	-	-	5.6	8.7	10.0	10.0	-	1.8	2.2	11.6	9.2	7.1	10.9	8.6	5.6	5.0	12.1
fr	-	24.2	18.8	27.0	23.7	24.6	39.0	-	10.0	-	-	13.8	18.3	23.9	-	10.0	3.5	12.5	25.2	24.1	15.2	29.4	18.5	11.8	12.4	23.6
he	8.5	17.0	12.8	18.2	17.4	17.4	32.5	22.7	6.9	8.1	24.5	-	11.7	17.6	19.1	7.2	2.1	8.3	17.5	16.5	10.5	21.2	14.4	7.7	6.6	17.2
hi	3.5	9.8	7.7	11.2	14.3	10.3	24.2	15.8	3.4	5.0	19.0	6.5	-	12.7	13.3	6.2	1.6	5.4	12.0	8.7	7.1	15.1	12.0	6.2	3.5	15.1
id	7.7	19.9	14.6	20.8	18.9	18.4	32.5	23.8	9.4	9.7	25.3	11.2	16.1	-	-	9.9	3.3	18.9	20.1	21.4	12.6	23.0	15.4	10.6	9.2	23.3
it	9.3	22.4	16.5	24.8	21.9	22.6	34.0	30.4	9.4	11.2	-	12.7	15.8	20.8	-	-	3.1	13.8	22.8	23.7	13.7	28.7	16.4	10.7	11.0	21.7
ja	3.7	7.2	5.8	8.4	7.7	7.8	11.5	-	4.4	4.4	12.3	4.0	8.8	9.4	9.4	-	-	5.2	8.3	7.8	5.5	-	7.3	-	-	-
ko	3.3	7.1	5.6	8.1	8.3	-	13.7	10.9	4.0	4.4	12.3	3.8	8.2	10.0	8.3	-	-	3.9	8.3	7.8	5.3	9.5	7.1	5.0	2.7	12.0
ms	7.4	11.6	8.2	16.5	12.6	-	27.1	-	8.7	5.6	19.5	6.0	11.5	19.8	17.2	-	1.6	-	13.5	10.2	7.8	18.3	12.2	9.2	4.5	23.0
nl	7.8	19.9	16.7	26.8	23.7	-	33.0	25.4	8.7	12.1	28.0	11.5	15.8	20.9	21.9	-	2.9	10.7	-	-	14.3	24.3	-	9.6	8.8	20.3
no	7.9	20.5	18.8	30.4	19.7	21.6	42.9	24.0	5.2	10.4	26.6	11.4	11.3	20.2	24.0	9.4	2.8	10.8	-	-	11.4	23.6	16.9	8.3	9.4	14.0
pl	5.0	13.8	13.0	16.0	13.2	-	17.8	15.6	5.3	8.4	18.4	7.0	10.6	13.5	14.0	-	2.0	7.1	14.3	11.0	-	14.6	12.2	6.3	8.5	14.4
pt	10.0	24.7	17.7	27.1	23.1	24.9	40.6	33.6	10.1	11.4	32.3	13.9	17.4	24.1	29.1	-	3.4	12.6	24.6	22.5	14.6	-	-	11.1	11.8	23.6
ru	6.0	16.9	12.8	15.9	15.3	14.6	20.1	17.6	7.0	7.8	20.6	9.3	12.7	14.5	15.5	7.9	2.0	9.3	-	14.4	11.2	16.8	-	-	17.0	16.1
tr	5.2	11.8	8.7	12.2	12.1	11.2	18.9	14.6	6.1	7.2	17.1	6.5	12.1	13.0	12.6	-	2.2	7.1	12.5	9.9	7.4	13.7	-	-	4.7	14.2
uk	4.0	14.2	10.0	12.2	12.2	10.7	18.6	15.0	4.4	6.4	16.8	4.8	6.5	10.6	12.7	-	1.2	5.2	11.3	10.4	9.3	13.7	19.2	4.5	-	11.7
vi	7.6	16.9	12.9	17.3	17.0	-	27.5	21.8	8.6	9.4	23.3	9.9	15.8	21.4	18.9	-	3.2	16.2	18.1	16.6	11.1	20.7	14.2	10.0	8.7	-

Schwenk et al. (2019)

Universal Translation

How far are we from universal machine translation?

	ar	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	id	it	ja	ko	ms	nl	no	pl	pt	ru	tr	uk	vi
ar	-	-	10.8	13.8	14.6	15.2	27.7	-	8.0	6.2	-	9.8	10.9	16.0	17.6	7.4	1.9	8.7	14.7	14.4	9.1	19.5	13.5	7.2	6.2	17.7
bg	-	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.8	19.4
cs	5.6	18.1	-	18.7	17.9	16.5	25.0	-	7.1	10.6	22.2	8.9	11.4	15.7	16.9	-	2.5	6.7	18.3	19.8	13.1	18.5	15.3	7.9	9.5	16.8
da	5.9	22.4	16.4	-	-	-	42.3	-	8.0	13.5	28.1	11.7	13.9	20.3	22.7	-	2.7	11.7	25.8	27.5	14.7	25.2	17.5	9.2	8.2	18.8
de	7.6	21.3	17.4	-	-	18.9	31.6	-	8.7	11.8	26.8	12.2	16.1	19.9	21.7	8.9	2.9	10.6	24.4	18.6	13.7	23.4	16.6	10.0	10.8	19.8
el	8.1	21.1	13.4	-	18.3	-	31.6	-	-	10.0	26.9	11.4	6.5	19.1	21.4	-	2.1	-	19.8	21.1	-	22.4	15.2	8.9	8.8	-
en	15.7	33.9	23.1	41.2	30.5	32.8	-	39.7	15.2	16.0	41.2	23.1	24.9	32.5	33.4	11.3	4.1	23.4	31.9	41.2	17.0	38.8	20.1	15.8	17.9	28.9
es		-	-	-	-	-	38.6	-	10.0	11.7	-	13.8	15.9	22.7	28.6	-	3.2	-	24.2	22.4	14.1	31.5	17.0	11.2	12.3	23.2
fa	6.5	13.6	9.3	13.2	12.9	-	25.1	16.3	-	5.2	18.6	7.2	8.8	15.0	14.8	-	1.9	8.2	13.4	10.4	7.8	16.8	11.4	8.1	5.4	16.8
fi	3.2	10.2	9.6	12.7	10.9	9.4	15.7	12.5	3.0	-	-	5.6	8.7	10.0	10.0	-	1.8	2.2	11.6	9.2	7.1	10.9	8.6	5.6	5.0	12.1
fr	-	24.2	18.8	27.0	23.7	24.6	39.0	-	10.0	-	-	13.8	18.3	23.9	-	10.0	3.5	12.5	25.2	24.1	15.2	29.4	18.5	11.8	12.4	23.6
he	8.5	17.0	12.8	18.2	17.4	17.4	32.5	22.7	6.9	8.1	24.5	-	11.7	17.6	19.1	7.2	2.1	8.3	17.5	16.5	10.5	21.2	14.4	7.7	6.6	17.2
hi	3.5	9.8	7.7	11.2	14.3	10.3	24.2	15.8	3.4	5.0	19.0	6.5	-	12.7	13.3	6.2	1.6	5.4	12.0	8.7	7.1	15.1	12.0	6.2	3.5	15.1
id	7.7	19.9	14.6	20.8	18.9	18.4	32.5	23.8	9.4	9.7	25.3	11.2	16.1	-	-	9.9	3.3	18.9	20.1	21.4	12.6	23.0	15.4	10.6	9.2	23.3
it	9.3	22.4	16.5	24.8	21.9	22.6	34.0	30.4	9.4	11.2	-	12.7	15.8	20.8	-	-	3.1	13.8	22.8	23.7	13.7	28.7	16.4	10.7	11.0	21.7
ja	3.7	7.2	5.8	8.4	7.7	7.8	11.5	-	4.4	4.4	12.3	4.0	8.8	9.4	9.4	-	-	5.2	8.3	7.8	5.5	-	7.3	-	-	-
ko	3.3	7.1	5.6	8.1	8.3	-	13.7	10.9	4.0	4.4	12.3	3.8	8.2	10.0	8.3	-	-	3.9	8.3	7.8	5.3	9.5	7.1	5.0	2.7	12.0
ms	7.4	11.6	8.2	16.5	12.6	-	27.1	-	8.7	5.6	19.5	6.0	11.5	19.8	17.2	-	1.6	-	13.5	10.2	7.8	18.3	12.2	9.2	4.5	23.0
nl	7.8	19.9	16.7	26.8	23.7	-	33.0	25.4	8.7	12.1	28.0	11.5	15.8	20.9	21.9	-	2.9	10.7	-	-	14.3	24.3	-	9.6	8.8	20.3
no	7.9	20.5	18.8	30.4	19.7	21.6	42.9	24.0	5.2	10.4	26.6	11.4	11.3	20.2	24.0	9.4	2.8	10.8	-	-	11.4	23.6	16.9	8.3	9.4	14.0
pl	5.0	13.8	13.0	16.0	13.2	-	17.8	15.6	5.3	8.4	18.4	7.0	10.6	13.5	14.0	-	2.0	7.1	14.3	11.0	-	14.6	12.2	6.3	8.5	14.4
pt	10.0	24.7	17.7	27.1	23.1	24.9	40.6	33.6	10.1	11.4	32.3	13.9	17.4	24.1	29.1	-	3.4	12.6	24.6	22.5	14.6	-	-	11.1	11.8	23.6
ru	6.0	16.9	12.8	15.9	15.3	14.6	20.1	17.6	7.0	7.8	20.6	9.3	12.7	14.5	15.5	7.9	2.0	9.3	-	14.4	11.2	16.8	-	-	17.0	16.1
tr	5.2	11.8	8.7	12.2	12.1	11.2	18.9	14.6	6.1	7.2	17.1	6.5	12.1	13.0	12.6	-	2.2	7.1	12.5	9.9	7.4	13.7	-	-	4.7	14.2
uk	4.0	14.2	10.0	12.2	12.2	10.7	18.6	15.0	4.4	6.4	16.8	4.8	6.5	10.6	12.7	-	1.2	5.2	11.3	10.4	9.3	13.7	19.2	4.5	-	11.7
vi	7.6	16.9	12.9	17.3	17.0	-	27.5	21.8	8.6	9.4	23.3	9.9	15.8	21.4	18.9	-	3.2	16.2	18.1	16.6	11.1	20.7	14.2	10.0	8.7	-

Schwenk et al. (2019)

Universal Translation

	ar	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	id	it	ja	ko	ms	nl	no	pl	pt	ru	tr	uk	vi
ar	-	-	10.8	13.8	14.6	15.2	27.7	-	8.0	6.2	-	9.8	10.9	16.0	17.6	7.4	1.9	8.7	14.7	14.4	9.1	19.5	13.5	7.2	6.2	17.7
bg	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.8	19.4
cs	5.6	18.1	-	18.7	17.9	16.5	25.0	-	7.1	10.6	22.2	8.9	11.4	15.7	16.9	-	2.5	6.7	18.3	19.8	13.1	18.5	15.3	7.9	9.5	16.8
da	5.9	22.4	16.4	-	-	-	42.3	-	8.0	13.5	28.1	11.7	13.9	20.3	22.7	-	2.7	11.7	25.8	27.5	14.7	25.2	17.5	9.2	8.2	18.8
de	7.6	21.3	17.4	-	-	18.9	31.6	-	8.7	11.8	26.8	12.2	16.1	19.9	21.7	8.9	2.9	10.6	24.4	18.6	13.7	23.4	16.6	10.0	10.8	19.8
el	8.1	21.1	13.4	[illegible]	18.3	[illegible]	31.6	[illegible]	[illegible]	10.0	26.9	11.4	6.5	19.1	21.4	[illegible]	2.1	-	19.8	21.1	[illegible]	22.4	15.2	8.9	8.8	-
en	15.7	33.9	23.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20.1	15.8	17.9	28.9
es	-	-	-	-	-	-	38.6	-	10.0	11.7	-	13.8	15.9	22.7	28.6	-	3.2	-	24.2	22.4	14.1	31.5	17.0	11.2	12.3	23.2
fa	6.5	13.6	9.3	13.2	12.9	-	25.1	16.3	-	5.2	18.6	7.2	8.8	15.0	14.8	-	1.9	8.2	13.4	10.4	7.8	16.8	11.4	8.1	5.4	16.8
fi	3.2	10.2	9.6	12.7	10.9	9.4	15.7	12.5	3.0	-	-	5.6	8.7	10.0	10.0	-	1.8	2.2	11.6	9.2	7.1	10.9	8.6	5.6	5.0	12.1
fr	-	24.2	18.8	27.0	23.7	24.6	39.0	-	10.0	-	-	13.8	18.3	23.9	-	10.0	3.5	12.5	25.2	24.1	15.2	29.4	18.5	11.8	12.4	23.6
he	8.5	17.0	12.8	18.2	17.4	17.4	32.5	22.7	6.9	8.1	24.5	-	11.7	17.6	19.1	7.2	2.1	8.3	17.5	16.5	10.5	21.2	14.4	7.7	6.6	17.2
hi	3.5	9.8	7.7	11.2	14.3	10.3	24.2	15.8	3.4	5.0	19.0	6.5	-	12.7	13.3	6.2	1.6	5.4	12.0	8.7	7.1	15.1	12.0	6.2	3.5	15.1
id	7.7	19.9	14.6	20.8	18.9	18.4	32.5	23.8	9.4	9.7	25.3	11.2	16.1	-	-	9.9	3.3	18.9	20.1	21.4	12.6	23.0	15.4	10.6	9.2	23.3
it	9.3	22.4	16.5	24.8	21.9	22.6	34.0	30.4	9.4	11.2	-	12.7	15.8	20.8	-	-	3.1	13.8	22.8	23.7	13.7	28.7	16.4	10.7	11.0	21.7
ja	3.7	7.2	5.8	8.4	7.7	7.8	11.5	-	4.4	4.4	12.3	4.0	8.8	9.4	9.4	-	-	5.2	8.3	7.8	5.5	-	7.3	-		-
ko	3.3	7.1	5.6	8.1	8.3	-	13.7	10.9	4.0	4.4	12.3	3.8	8.2	10.0	8.3	-	-	3.9	8.3	7.8	5.3	9.5	7.1	5.0	2.7	12.0
ms	7.4	11.6	8.2	16.5	12.6	-	27.1	-	8.7	5.6	19.5	6.0	11.5	19.8	17.2	-	1.6	-	13.5	10.2	7.8	18.3	12.2	9.2	4.5	23.0
nl	7.8	19.9	16.7	26.8	23.7	-	33.0	25.4	8.7	12.1	28.0	11.5	15.8	20.9	21.9	-	2.9	10.7	-	-	14.3	24.3		9.6	8.8	20.3
no	7.9	20.5	18.8	30.4	19.7	21.6	42.9	24.0	5.2	10.4	26.6	11.4	11.3	20.2	24.0	9.4	2.8	10.8	-	-	11.4	23.6	16.9	8.3	9.4	14.0
pl	5.0	13.8	13.0	16.0	13.2	-	17.8	15.6	5.3	8.4	18.4	7.0	10.6	13.5	14.0	-	2.0	7.1	14.3	11.0	-	14.6	12.2	6.3	8.5	14.4
pt	10.0	24.7	17.7	27.1	23.1	24.9	40.6	33.6	10.1	11.4	32.3	13.9	17.4	24.1	29.1	-	3.4	12.6	24.6	22.5	14.6	-	-	11.1	11.8	23.6
ru	6.0	16.9	12.8	15.9	15.3	14.6	20.1	17.6	7.0	7.8	20.6	9.3	12.7	14.5	15.5	7.9	2.0	9.3	-	14.4	11.2	16.8	-	-	17.0	16.1
tr	5.2	11.8	8.7	12.2	12.1	11.2	18.9	14.6	6.1	7.2	17.1	6.5	12.1	13.0	12.6	-	2.2	7.1	12.5	9.9	7.4	13.7	-	-	4.7	14.2
uk	4.0	14.2	10.0	12.2	12.2	10.7	18.6	15.0	4.4	6.4	16.8	4.8	6.5	10.6	12.7	-	1.2	5.2	11.3	10.4	9.3	13.7	19.2	4.5	-	11.7
vi	7.6	16.9	12.9	17.3	17.0	-	27.5	21.8	8.6	9.4	23.3	9.9	15.8	21.4	18.9	-	3.2	16.2	18.1	16.6	11.1	20.7	14.2	10.0	8.7	-

How far are we from universal machine translation?

- 86% of all language directions are of poor quality

Schwenk et al. (2019)

Universal Translation

	ar	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	id	it	ja	ko	ms	nl	no	pl	pt	ru	tr	uk	vi
ar	-	-	10.8	13.8	14.6	15.2	27.7	-	8.0	6.2	-	9.8	10.9	16.0	17.6	7.4	1.9	8.7	14.7	14.4	9.1	19.5	13.5	7.2	6.2	17.7
bg	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.8	19.4
cs	5.6	18.1	-	18.7	17.9	16.5	25.0	-	7.1	10.6	22.2	8.9	11.4	15.7	16.9	-	2.5	6.7	18.3	19.8	13.1	18.5	15.3	7.9	9.5	16.8
da	5.9	22.4	16.4	-	-	-	42.3	-	8.0	13.5	28.1	11.7	13.9	20.3	22.7	-	2.7	11.7	25.8	27.5	14.7	25.2	17.5	9.2	8.2	18.8
de	7.6	21.3	17.4	-	-	18.9	31.6	-	8.7	11.8	26.8	12.2	16.1	19.9	21.7	8.9	2.9	10.6	24.4	18.6	13.7	23.4	16.6	10.0	10.8	19.8
el	8.1	21.1	13.4		18.3		31.6	-		10.0	26.9	11.4	6.5	19.1	21.4	-	2.1	-	19.8	21.1		22.4	15.2	8.9	8.8	-
en	15.7	33.9	23.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20.1	15.8	17.9	28.9
es	-	-	-	-	-	-	38.6	-	10.0	11.7	-	13.8	15.9	22.7	28.6	-	3.2	-	24.2	22.4	14.1	31.5	17.0	11.2	12.3	23.2
fa	6.5	13.6	9.3	13.2	12.9	-	25.1	16.3	-	5.2	18.6	7.2	8.8	15.0	14.8	-	1.9	8.2	13.4	10.4	7.8	16.8	11.4	8.1	5.4	16.8
fi	3.2	10.2	9.6	12.7	10.9	9.4	15.7	12.5	3.0	-	-	5.6	8.7	10.0	10.0	-	1.8	2.2	11.6	9.2	7.1	10.9	8.6	5.6	5.0	12.1
fr	-	24.2	18.8	27.0	23.7	24.6	39.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.0	[illegible]	12.5	22.2	24.1	15.2	29.4	18.5	11.8	12.4	23.6
he	8.5	17.0	12.8	18.2	17.4	17.4	32.5	22.7	6.9	8.1	24.5	-	11.7	17.6	19.1	7.2	2.1	8.3	17.5	16.5	10.5	21.2	14.4	7.7	6.6	17.2
hi	3.5	9.8	7.7	11.2	14.3	10.3	24.2	15.8	3.4	5.0	19.0	6.5	-	12.7	13.3	6.2	1.6	5.4	12.0	8.7	7.1	15.1	12.0	6.2	3.5	15.1
id	7.7	19.9	14.6	20.8	18.9	18.4	32.5	23.8	9.4	9.7	25.3	11.2	16.1	-	-	9.9	3.3	18.9	20.1	21.4	12.6	23.0	15.4	10.6	9.2	23.3
it	9.3	22.4	16.5	24.8	21.9	22.6	34.0	30.4	9.4	11.2	-	12.7	15.8	20.8	-	-	3.1	13.8	22.8	23.7	13.7	28.7	16.4	10.7	11.0	21.7
ja	3.7	7.2	5.8	8.4	7.7	7.8	11.5	-	4.4	4.4	12.3	4.0	8.8	9.4	9.4	-	-	5.2	8.3	7.8	5.5	-	7.3	-	-	-
ko	3.3	7.1	5.6	8.1	8.3	-	13.7	10.9	4.0	4.4	12.3	3.8	8.2	10.0	8.3	-	-	3.9	8.3	7.8	5.3	9.5	7.1	5.0	2.7	12.0
ms	7.4	11.6	8.2	16.5	12.6	-	27.1	-	8.7	5.6	19.5	6.0	11.5	19.8	17.2	-	1.6	-	13.5	10.2	7.8	18.3	12.2	9.2	4.5	23.0
nl	7.8	19.9	16.7	26.8	23.7	-	33.0	25.4	8.7	12.1	28.0	11.5	15.8	20.9	21.9	-	2.9	10.7	-	-	14.3	24.3	-	9.6	8.8	20.3
no	7.9	20.5	18.8	30.4	19.7	21.6	42.9	24.0	5.2	10.4	26.6	11.4	11.3	20.2	24.0	9.4	2.8	10.8	-	-	11.4	23.6	16.9	8.3	9.4	14.0
pl	5.0	13.8	13.0	16.0	13.2	-	17.8	15.6	5.3	8.4	18.4	7.0	10.6	13.5	14.0	-	2.0	7.1	14.3	11.0	-	14.6	12.2	6.3	8.5	14.4
pt	10.0	24.7	17.7	27.1	23.1	24.9	40.6	33.6	10.1	11.4	32.3	13.9	17.4	24.1	29.1	-	3.4	12.6	24.6	22.5	14.6	-	-	11.1	11.8	23.6
ru	6.0	16.9	12.8	15.9	15.3	14.6	20.1	17.6	7.0	7.8	20.6	9.3	12.7	14.5	15.5	7.9	2.0	9.3	-	14.4	11.2	16.8	-	-	17.0	16.1
tr	5.2	11.8	8.7	12.2	12.1	11.2	18.9	14.6	6.1	7.2	17.1	6.5	12.1	13.0	12.6	-	2.2	7.1	12.5	9.9	7.4	13.7	-	-	4.7	14.2
uk	4.0	14.2	10.0	12.2	12.2	10.7	18.6	15.0	4.4	6.4	16.8	4.8	6.5	10.6	12.7	-	1.2	5.2	11.3	10.4	9.3	13.7	19.2	4.5	-	11.7
vi	7.6	16.9	12.9	17.3	17.0	-	27.5	21.8	8.6	9.4	23.3	9.9	15.8	21.4	18.9	-	3.2	16.2	18.1	16.6	11.1	20.7	14.2	10.0	8.7	-

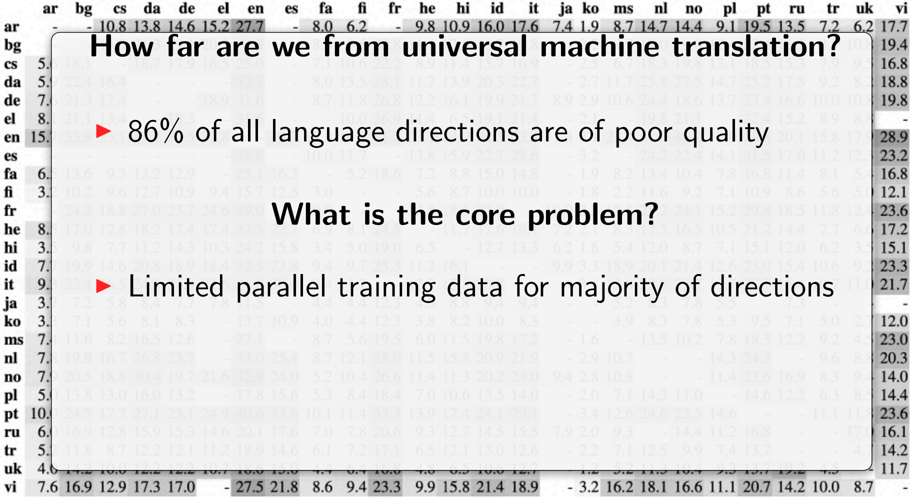
How far are we from universal machine translation?

- 86% of all language directions are of poor quality

What is the core problem?

Schwenk et al. (2019)

Universal Translation



How far are we from universal machine translation?

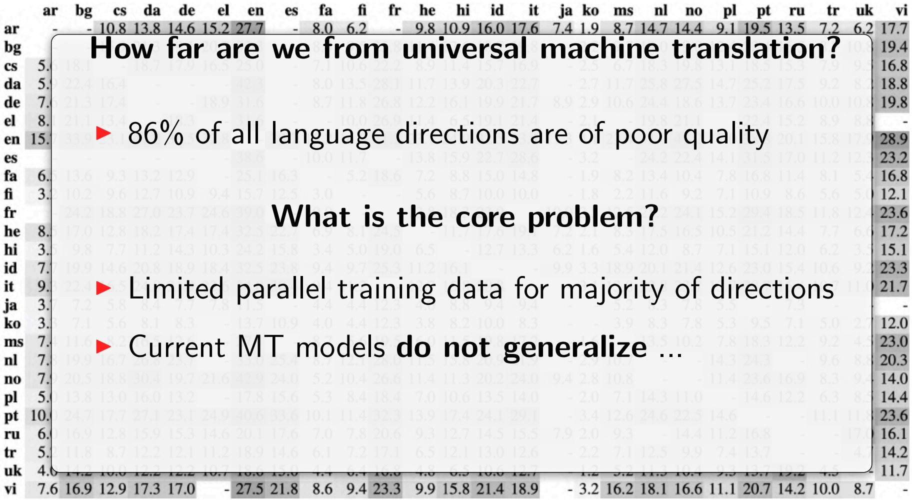
- 86% of all language directions are of poor quality

What is the core problem?

- Limited parallel training data for majority of directions

Schwenk et al. (2019)

Universal Translation



How far are we from universal machine translation?

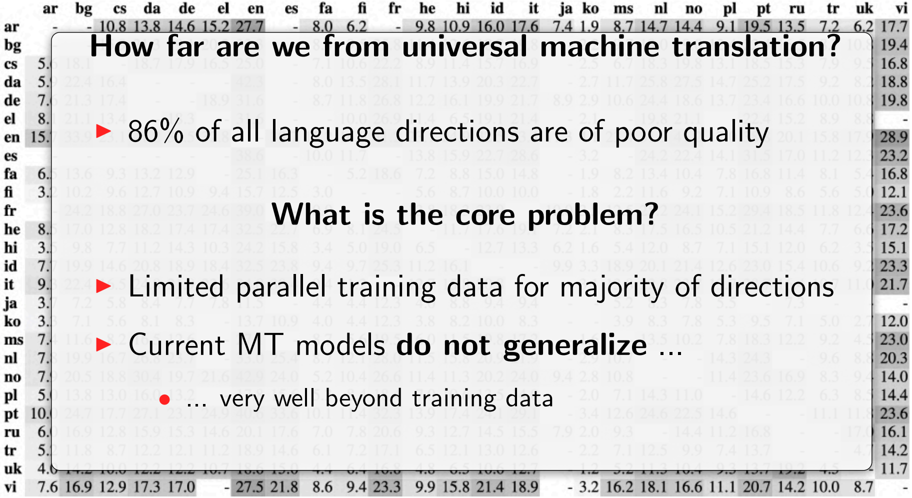
- 86% of all language directions are of poor quality

What is the core problem?

- Limited parallel training data for majority of directions
- Current MT models **do not generalize** …

Schwenk et al. (2019)

Universal Translation



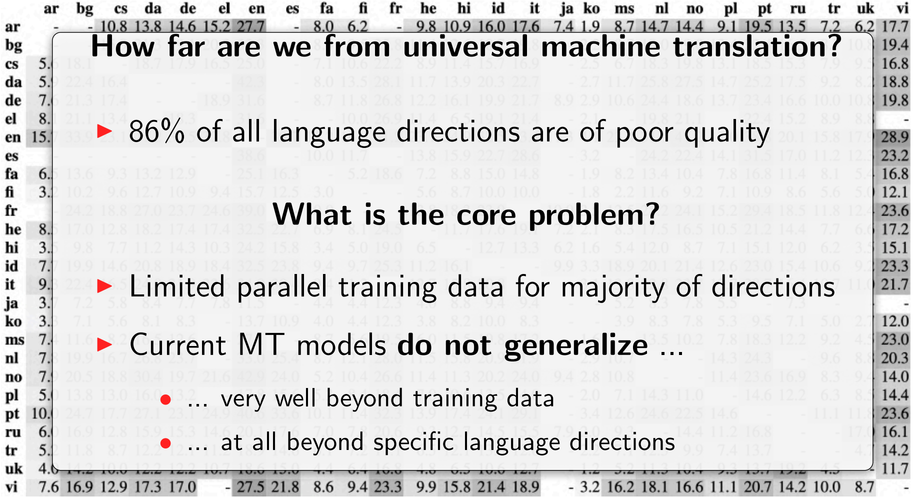
How far are we from universal machine translation?

- 86% of all language directions are of poor quality

What is the core problem?

- Limited parallel training data for majority of directions
- Current MT models **do not generalize** ...
  - ... very well beyond training data

Schwenk et al. (2019)



How far are we from universal machine translation?

- 86% of all language directions are of poor quality

What is the core problem?

- Limited parallel training data for majority of directions
- Current MT models **do not generalize** ...
  - ... very well beyond training data
  - ... at all beyond specific language directions

Schwenk et al. (2019)

Essential Data Component: Bilingual Parallel Corpus

Japan to tighten checks for African swine fever

#Japan #Health & Welfare

Tuesday, Nov. 26, 20:09

The Japanese government plans to give more powers to quarantine officers at airports, as part of its efforts to prevent African swine fever from entering the country.

Outbreaks of the fatal and highly contagious disease have been reported in China, South Korea and other parts of Asia, but no cases have been confirmed in Japan so far.

The agriculture ministry is working on legal amendments to block the entry of the African swine fever virus.

It plans to allow quarantine officers at airports to ask travelers if they have any meat products. They would also be able to inspect luggage without the owner's consent.

日本拟加强口岸检查严防非洲猪瘟病毒

11月 27日 (星期三) 5:24

鉴于非洲猪瘟疫情在亚洲多国不断扩大，为了防止病毒被带入日本国内，农林水产省决定加大在机场等处开展口岸检查工作的防疫官的权限。

非洲猪瘟疫情在中国、韩国等国蔓延。由于目前还没有有效的疫苗，非洲猪瘟的病毒一旦通过猪肉进入日本国内，将给日本的畜牧业等带来沉重打击。鉴于此，农林水产省决定修订相关法律，加强口岸检查工作。

具体措施是，加大在机场等处开展检查工作的家畜防疫官的权限，防疫官可询问入境人员是否携带肉制品，必要时可采取强制措施，检查其行

Essential Data Component: Bilingual Parallel Corpus

Japan to tighten checks for African swine fever

#Japan #Health & Welfare

Tuesday, Nov. 26, 20:09

The Japanese government plans to give more powers to quarantine officers at airports, as part of its efforts to prevent African swine fever from entering the country.

Outbreaks of the fatal and highly contagious disease have been reported in China, South Korea and other parts of Asia, but no cases have been confirmed in Japan so far.

The agriculture ministry is working on legal amendments to block the entry of the African swine fever virus.

It plans to allow quarantine officers at airports to ask travelers if they have any meat products. They would also be able to inspect luggage without the owner's consent.

日本拟加强口岸检查严防非洲猪瘟病毒

11月 27日 (星期三) 5:24

鉴于非洲猪瘟疫情在亚洲多国不断扩大，为了防止病毒被带入日本国内，农林水产省决定加大在机场等处开展口岸检查工作的防疫官的权限。

非洲猪瘟疫情在中国、韩国等国蔓延。由于目前还没有有效的疫苗，非洲猪瘟的病毒一旦通过猪肉进入日本国内，将给日本的畜牧业等带来沉重打击。鉴于此，农林水产省决定修订相关法律，加强口岸检查工作。

具体措施是，加大在机场等处开展检查工作的家畜防疫官的权限，防疫官可询问入境人员是否携带肉制品，必要时可采取强制措施，检查其行

Parallel Corpus

⋮	⋮
李鹏会见新加坡前总统王鼎昌	Li Peng Meets With Former Singapore President Ong Teng Cheong
马来亚、新加坡、沙捞越、沙巴和文莱曾组成联邦,但最後分裂了。	Malaysia, Singapore, Sarawak, Sabah and Brunei once formed a federation, but it also fell apart in the end.
新加坡排行榜首,缅甸则排行榜尾 。	Singapore is at the head of the list, while Burma ranks last.
新加坡则在致力建造一个光纤网环绕的“智能岛”。	Singapore is also devoting itself to building a "intelligence island" embraced by a fiber-optical net.
⋮	⋮

Parallel Corpus

⋮	⋮
李鹏会见新加坡前总统王鼎昌	Li Peng Meets With Former Singapore President Ong Teng Cheong
马来亚、新加坡、沙捞越、沙巴和文莱曾组成联邦,但最後分裂了。	Malaysia, Singapore, Sarawak, Sabah and Brunei once formed a federation, but it also fell apart in the end.
新加坡排行榜首,缅甸则排行榜尾 。	Singapore is at the head of the list, while Burma ranks last.
新加坡则在致力建造一个光纤网环绕的“智能岛” 。	Singapore is also devoting itself to building a "intelligence island" embraced by a fiber-optical net.
⋮	⋮

Encoder-Decoder Architecure

- The general encoder-decoder architecture assumes

Encoder-Decoder Architecure

- The general encoder-decoder architecture assumes
  - a (complex) input that needs to be encoded into a representation

Encoder-Decoder Architecure

- The general encoder-decoder architecture assumes
  - a (complex) input that needs to be encoded into a representation
  - a (complex) output into which the input representation needs to be decoded

Encoder-Decoder Architecure

- The general encoder-decoder architecture assumes
  - a (complex) input that needs to be encoded into a representation
  - a (complex) output into which the input representation needs to be decoded
- Many tasks fall under the general encoder-decoder architecture

Encoder-Decoder Architecure

- The general encoder-decoder architecture assumes
  - a (complex) input that needs to be encoded into a representation
  - a (complex) output into which the input representation needs to be decoded
- Many tasks fall under the general encoder-decoder architecture
- Image captioning
  - input: image (encode using CNNs)
  - output: sentence (decode using an RNN/CNN language model)

Encoder-Decoder Architecure

- The general encoder-decoder architecture assumes
  - a (complex) input that needs to be encoded into a representation
  - a (complex) output into which the input representation needs to be decoded
- Many tasks fall under the general encoder-decoder architecture
- Image captioning
  - input: image (encode using CNNs)
  - output: sentence (decode using an RNN/CNN language model)
- Speech recognition
  - input: audio signal over time (encode using CNN+RNN)
  - output: sentence (decode using an RNN/CNN language model)

Encoder-Decoder Architecure

...

- Sentence Summarization
  - input: sentence (encode using RNN/CNN)
  - output: shorter sentence (decode using an RNN/CNN language model)

Encoder-Decoder Architecure

...

- Sentence Summarization
  - input: sentence (encode using RNN/CNN)
  - output: shorter sentence (decode using an RNN/CNN language model)
- Machine translation
  - input: foreign sentence (encode using RNN/CNN)
  - output: sentence translation (decode using an RNN/CNN language model)

Encoder-Decoder Architecure

...

- Sentence Summarization
  - input: sentence (encode using RNN/CNN)
  - output: shorter sentence (decode using an RNN/CNN language model)
- Machine translation
  - input: foreign sentence (encode using RNN/CNN)
  - output: sentence translation (decode using an RNN/CNN language model)
- When input and outputs are sequences of words/audio we talk about sequence-to-sequence (seq2seq) models

Sequence-to-Sequence Models

- Sequence-to-sequence modeling
  - differs from sequence labeling by

Sequence-to-Sequence Models

- Sequence-to-sequence modeling
  - differs from sequence labeling by
    - not assuming an isomorphic relationship between x_t and y_t

Sequence-to-Sequence Models

- Sequence-to-sequence modeling
  - differs from sequence labeling by
    - not assuming an isomorphic relationship between x_t and y_t
    - by not assuming that $|X| = |Y|$, i.e., they can differ in lenght

Sequence-to-Sequence Models

- Sequence-to-sequence modeling
  - differs from sequence labeling by
    - not assuming an isomorphic relationship between x_t and y_t
    - by not assuming that $|X| = |Y|$, i.e., they can differ in lenght
  - aims to model the complex mapping between X and Y

Sequence-to-Sequence Models

- Sequence-to-sequence modeling
  - differs from sequence labeling by
    - not assuming an isomorphic relationship between x_t and y_t
    - by not assuming that $|X| = |Y|$, i.e., they can differ in lenght
  - aims to model the complex mapping between X and Y
- As sequences are typically modeled using a language model, sequence-to-sequence modeling can be cast a conditional language modeling task: $p(y_t|\mathbf{Y}_{<t}, \mathbf{X})$,

Sequence-to-Sequence Models

- Sequence-to-sequence modeling
  - differs from sequence labeling by
    - not assuming an isomorphic relationship between x_t and y_t
    - by not assuming that $|X| = |Y|$, i.e., they can differ in lenght
  - aims to model the complex mapping between X and Y
- As sequences are typically modeled using a language model, sequence-to-sequence modeling can be cast a conditional language modeling task: $p(y_t|\mathbf{Y}_{<t}, \mathbf{X})$,

  where $\mathbf{X}$ is the output of the encoder (i.e., a representation of the input)

Sequence-to-Sequence Models

- Sequence-to-sequence modeling
  - differs from sequence labeling by
    - not assuming an isomorphic relationship between x_t and y_t
    - by not assuming that $|X| = |Y|$, i.e., they can differ in lenght
  - aims to model the complex mapping between X and Y
- As sequences are typically modeled using a language model, sequence-to-sequence modeling can be cast a conditional language modeling task: $p(y_t|\mathbf{Y}_{<t}, \mathbf{X})$,

  where $\mathbf{X}$ is the output of the encoder (i.e., a representation of the input)
  and $\mathbf{Y}_{<t}$ is a representation of the output of the decoder before time t (the prefix)

Neural Machine Translation

- Sutskever et al. (2014) cast machine translation as a sequence-to-sequence modeling problem where
  - the encoder is an LSTM
  - the decoder is an LSTM

Neural Machine Translation

- Sutskever et al. (2014) cast machine translation as a sequence-to-sequence modeling problem where
  - the encoder is an LSTM
  - the decoder is an LSTM

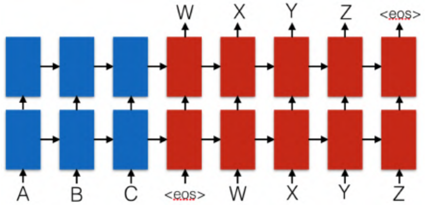


image credit: Zoph et al. (2016)

- How are the encoder and decoder connected?

Neural Machine Translation

- Sutskever et al. (2014) cast machine translation as a sequence-to-sequence modeling problem where
  - the encoder is an LSTM
  - the decoder is an LSTM

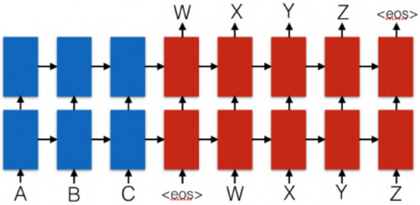


image credit: Zoph et al. (2016)

- How are the encoder and decoder connected?
  - important question!

Neural Machine Translation

- Sutskever et al. (2014) cast machine translation as a sequence-to-sequence modeling problem where
  - the encoder is an LSTM
  - the decoder is an LSTM

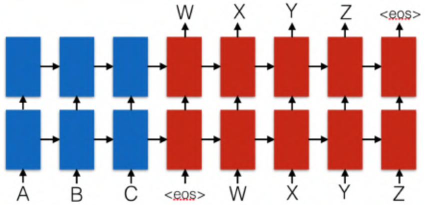
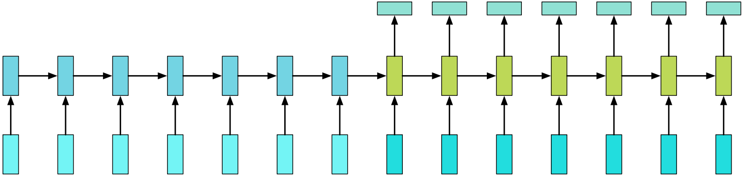


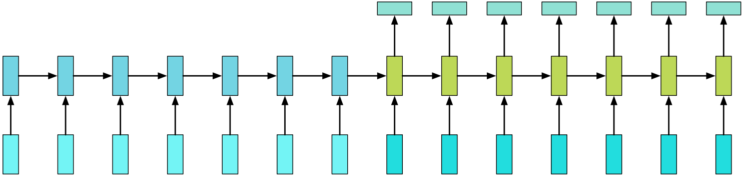
image credit: Zoph et al. (2016)

- How are the encoder and decoder connected?
  - important question!
  - Sutskever et al. (2014) simply initialize the decoder LSTM with the last state of the encoder LSTM

Neural Machine Translation

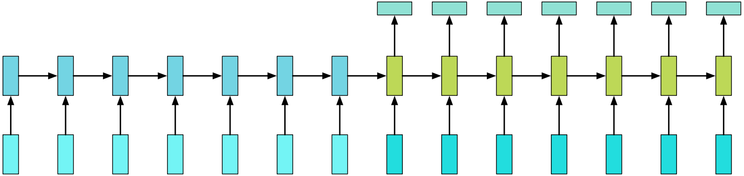


Neural Machine Translation



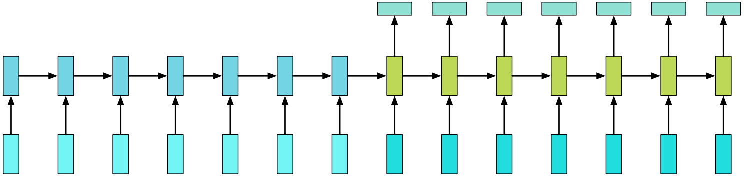
- encoder (left): represents $\mathbf{z}$ as a whole

Neural Machine Translation



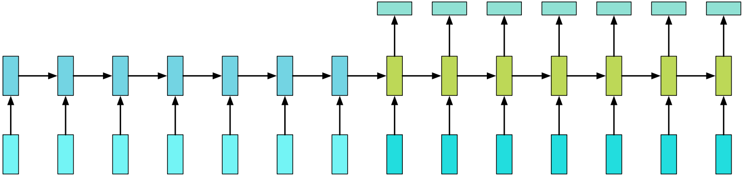
- encoder (left): represents $\mathbf{z}$ as a whole
- decoder (right): reads in $\mathbf{x}$ token by token and learns to predict $\mathbf{y}$

Neural Machine Translation



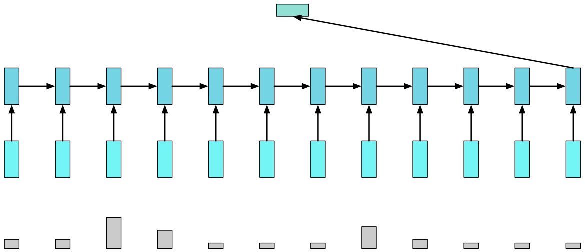
- encoder (left): represents $\mathbf{z}$ as a whole
- decoder (right): reads in $\mathbf{x}$ token by token and learns to predict $\mathbf{y}$
- encoder connects to decoder by setting $\mathbf{h}_0^{dec} = \mathbf{h}_n^{enc}$

Neural Machine Translation

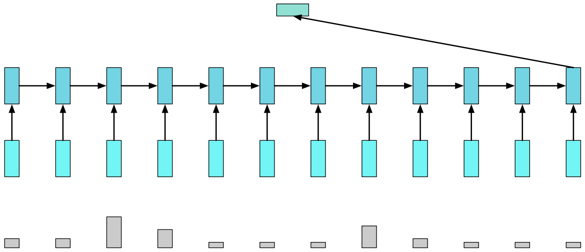


- encoder (left): represents $\mathbf{z}$ as a whole
- decoder (right): reads in $\mathbf{x}$ token by token and learns to predict $\mathbf{y}$
- encoder connects to decoder by setting $\mathbf{h}_0^{dec} = \mathbf{h}_n^{enc}$
- encoder-decoder information bottleneck

Information Distribution

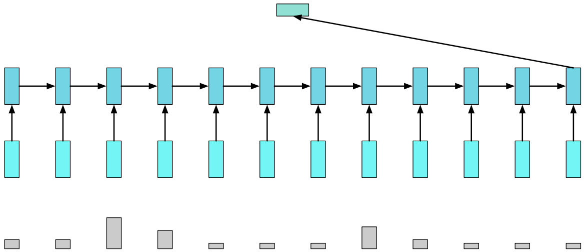


Information Distribution



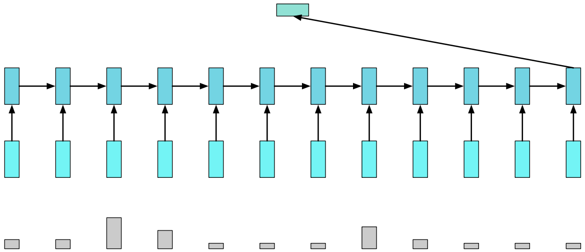
- The hidden layers, including the final hidden layer are of fixed, limited capacity

Information Distribution



- The hidden layers, including the final hidden layer are of fixed, limited capacity
- The final hidden layer cannot represent the full information of a long input sentence

Information Distribution



- The hidden layers, including the final hidden layer are of fixed, limited capacity
- The final hidden layer cannot represent the full information of a long input sentence
- For classification, the sentence representation learns which tokens are important to predict a certain class

Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation

Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation
  - also used in image captioning

Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation
  - also used in image captioning
  - very much related to neural memory networks

Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation
  - also used in image captioning
  - very much related to neural memory networks
- Basic idea:

Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation
  - also used in image captioning
  - very much related to neural memory networks
- Basic idea:
  - don't try to learn one global representation for the source sentence (encoder)

Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation
  - also used in image captioning
  - very much related to neural memory networks
- Basic idea:
  - don't try to learn one global representation for the source sentence (encoder)
  - rather learn context-sensitive token representations

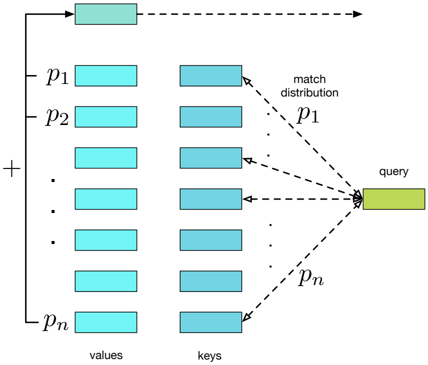
Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation
  - also used in image captioning
  - very much related to neural memory networks
- Basic idea:
  - don't try to learn one global representation for the source sentence (encoder)
  - rather learn context-sensitive token representations
  - when generating a target word, dynamically combine the most relevant source representations

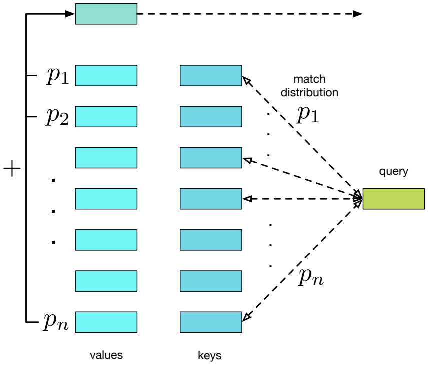
Attention Mechanism

- Bahdanau et al. (2015) introduced the attention mechanism to machine translation
  - also used in image captioning
  - very much related to neural memory networks
- Basic idea:
  - don't try to learn one global representation for the source sentence (encoder)
  - rather learn context-sensitive token representations
  - when generating a target word, dynamically combine the most relevant source representations
- Similar to word alignment, where alignments indicate source-target token translation correspondences
  - attention results in soft (numerical) alignments

Attention Mechanism

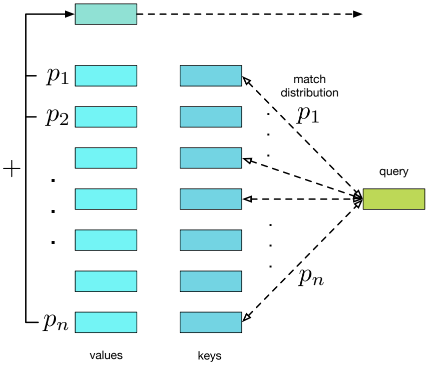


Attention Mechanism



- $\mathbf{c} = \sum_{i=1}^{n} \mathrm{sim}_p(W_k \mathbf{k}^i, W_q \mathbf{q}) W_v \mathbf{v}^i$

Attention Mechanism



- $\mathbf{c} = \sum_{i=1}^{n} \mathrm{sim}_p(W_k \mathbf{k}^i, W_q \mathbf{q}) W_v \mathbf{v}^i$
- $W_q \in \mathbb{R}^{n_m \times m_q}, W_k \in \mathbb{R}^{n_m \times m_k}, W_v \in \mathbb{R}^{n_v \times m_v}$

Attention Mechanism

- How to define $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$?

Attention Mechanism

- How to define $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$?
- Inner product: $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = (W_k\mathbf{k}^i)^\intercal W_q\mathbf{q}$

Attention Mechanism

- How to define $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$?
- Inner product: $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = (W_k\mathbf{k}^i)^\top W_q\mathbf{q}$
- Feed-forward network: $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = \mathbf{w}_s^\top \tanh(W_k\mathbf{k}^i + W_q\mathbf{q} + \mathbf{b}_s)$

Attention Mechanism

- How to define $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$?
- Inner product: $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = (W_k\mathbf{k}^i)^\top W_q\mathbf{q}$
- Feed-forward network: $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = \mathbf{w}_s^\top \tanh(W_k\mathbf{k}^i + W_q\mathbf{q} + \mathbf{b}_s)$
- $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$ results in arbitrary activations

$$p_i = \frac{\exp(\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}))}{\sum_{j=1}^n \exp(\text{sim}(W_k\mathbf{k}^j, W_q\mathbf{q}))}$$

p is also referred to as attention distribution

Attention Mechanism

- How to define $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$?
- Inner product: $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = (W_k\mathbf{k}^i)^\top W_q\mathbf{q}$
- Feed-forward network: $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = \mathbf{w}_s^\top \tanh(W_k\mathbf{k}^i + W_q\mathbf{q} + \mathbf{b}_s)$
- $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$ results in arbitrary activations

  $p_i = \frac{\exp(\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}))}{\sum_{j=1}^{n}\exp(\text{sim}(W_k\mathbf{k}^j, W_q\mathbf{q}))}$

  p is also referred to as attention distribution
- Keys and values can be identical: $\mathbf{k}^i = \mathbf{v}^i$

Attention Mechanism

- How to define $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$?
- Inner product: $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = (W_k\mathbf{k}^i)^\top W_q\mathbf{q}$
- Feed-forward network: $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = \mathbf{w}_s^\top \tanh(W_k\mathbf{k}^i + W_q\mathbf{q} + \mathbf{b}_s)$
- $\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$ results in arbitrary activations

$$p_i = \frac{\exp(\text{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}))}{\sum_{j=1}^{n} \exp(\text{sim}(W_k\mathbf{k}^j, W_q\mathbf{q}))}$$

p is also referred to as attention distribution

- Keys and values can be identical: $\mathbf{k}^i = \mathbf{v}^i$
- W_q, W_k can be identity matrices I_q, I_k if
  - $m_q = m_k$ and similarity is defined as inner product

Attention Mechanism

- How to define $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$?
- Inner product: $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = (W_k\mathbf{k}^i)^\top W_q\mathbf{q}$
- Feed-forward network: $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}) = \mathbf{w}_s^\top \tanh(W_k\mathbf{k}^i + W_q\mathbf{q} + \mathbf{b}_s)$
- $\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q})$ results in arbitrary activations

$$p_i = \frac{\exp(\mathrm{sim}(W_k\mathbf{k}^i, W_q\mathbf{q}))}{\sum_{j=1}^{n}\exp(\mathrm{sim}(W_k\mathbf{k}^j, W_q\mathbf{q}))}$$

p is also referred to as attention distribution

- Keys and values can be identical: $\mathbf{k}^i = \mathbf{v}^i$
- W_q, W_k can be identity matrices I_q, I_k if
  - $m_q = m_k$ and similarity is defined as inner product
- The attention mechanism and thereby the computation of $\mathbf{c}$ is fully differentiable!

Attention in NMT

- Applying attention to NMT
  - what is a query?

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys?

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens: $\mathbf{k}^i = \mathbf{h}_i^{enc}$

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens: $\mathbf{k}^i = \mathbf{h}_i^{enc}$
  - what are the values?

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens: $\mathbf{k}^i = \mathbf{h}_i^{enc}$
  - what are the values? Typically: $\mathbf{v}^i = \mathbf{k}^i$

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens: $\mathbf{k}^i = \mathbf{h}_i^{enc}$
  - what are the values? Typically: $\mathbf{v}^i = \mathbf{k}^i$
  - what is $\mathbf{c}$?

Attention in NMT

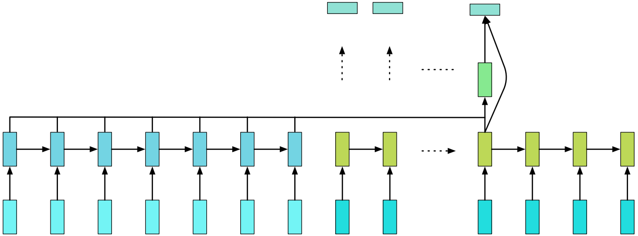
- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens: $\mathbf{k}^i = \mathbf{h}_i^{enc}$
  - what are the values? Typically: $\mathbf{v}^i = \mathbf{k}^i$
  - what is $\mathbf{c}$? The representation of the source focusing on the tokens that are most relevant for generating the next word $j+1$

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens: $\mathbf{k}^i = \mathbf{h}_i^{enc}$
  - what are the values? Typically: $\mathbf{v}^i = \mathbf{k}^i$
  - what is $\mathbf{c}$? The representation of the source focusing on the tokens that are most relevant for generating the next word $j+1$
  - in NMT, $\mathbf{c}$ is often called the context vector

Attention in NMT

- Applying attention to NMT
  - what is a query? Current state of decoder: $\mathbf{q} = \mathbf{h}_j^{dec}$
  - what are the keys? Representations of *all* source tokens: $\mathbf{k}^i = \mathbf{h}_i^{enc}$
  - what are the values? Typically: $\mathbf{v}^i = \mathbf{k}^i$
  - what is $\mathbf{c}$? The representation of the source focusing on the tokens that are most relevant for generating the next word $j+1$
  - in NMT, $\mathbf{c}$ is often called the context vector



Attention in NMT

- Attention leads to significant improvements in translation quality

Attention in NMT

- Attention leads to significant improvements in translation quality
  - in particular for longer source sentences

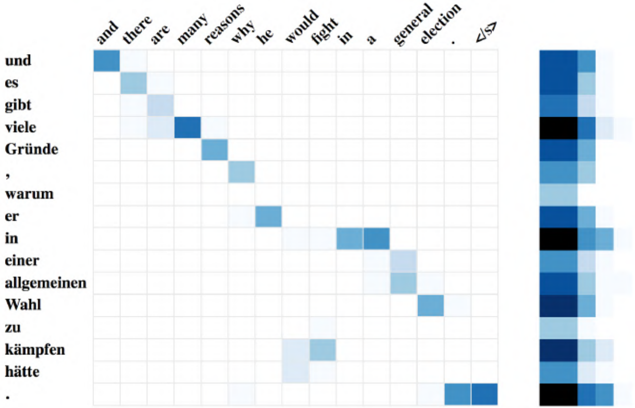
Attention in NMT

- Attention leads to significant improvements in translation quality
  - in particular for longer source sentences
  - it can model complex translation mappings (due to soft alignments)

Attention in NMT

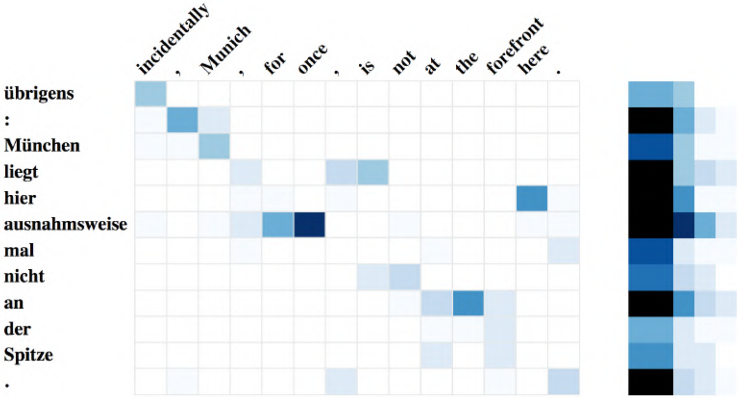
- Attention leads to significant improvements in translation quality
  - in particular for longer source sentences
  - it can model complex translation mappings (due to soft alignments)
- Added benefits:
  - attention can be visualized allowing for some inspection of the model
  - useful for error analysis

NMT Attention Examples



- Attention can model word order differences

NMT Attention Examples



- Attention can model multi-word translations

Credit Assignment Problem

- Language is full of long-distance phenomena

Credit Assignment Problem

- Language is full of long-distance phenomena
  - morphological agreement

Credit Assignment Problem

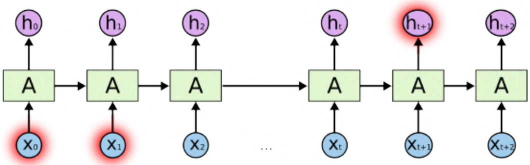
- Language is full of long-distance phenomena
  - morphological agreement
  - topicality

Credit Assignment Problem

- Language is full of long-distance phenomena
  - morphological agreement
  - topicality
  - general grammaticality/fluency

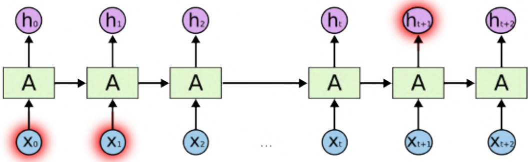
Credit Assignment Problem

- Language is full of long-distance phenomena
  - morphological agreement
  - topicality
  - general grammaticality/fluency



Credit Assignment Problem

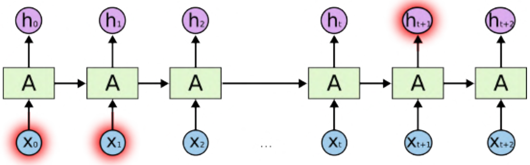
- Language is full of long-distance phenomena
  - morphological agreement
  - topicality
  - general grammaticality/fluency



- RNNs model dependencies along a (long) recurrent path

Credit Assignment Problem

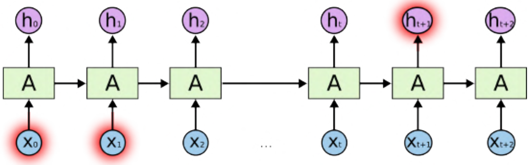
- Language is full of long-distance phenomena
  - morphological agreement
  - topicality
  - general grammaticality/fluency



- RNNs model dependencies along a (long) recurrent path
- Even if the gradient play nice (i.e., don't vanish nor explode) this does not necessarily mean that they model interactions correctly $\longrightarrow$ credit assignment problem

Credit Assignment Problem

- Language is full of long-distance phenomena
  - morphological agreement
  - topicality
  - general grammaticality/fluency

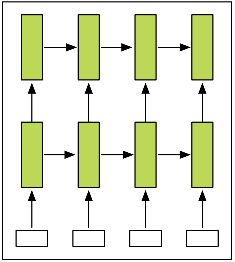


- RNNs model dependencies along a (long) recurrent path
- Even if the gradient play nice (i.e., don't vanish nor explode) this does not necessarily mean that they model interactions correctly $\rightarrow$ credit assignment problem

  should $\mathbf{h}_{t+1}$ really depend on $\mathbf{x}_0$ or $\mathbf{x}_1$ or both or neither?

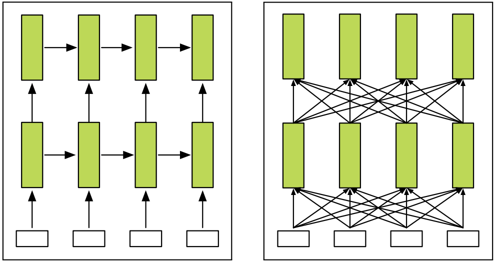
Self-Attention

- Self-attention computes attention between elements of the same sequence
  - can replace RNNs as sequence model
  - shortens paths of credit assignment
  - at the core of Google's Transformer NMT system (Vaswani et al., 2017)



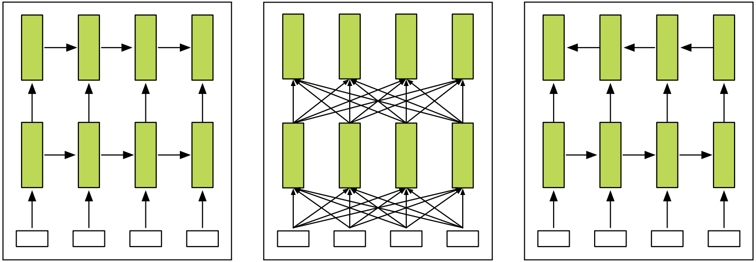
Self-Attention

- Self-attention computes attention between elements of the same sequence
  - can replace RNNs as sequence model
  - shortens paths of credit assignment
  - at the core of Google's Transformer NMT system (Vaswani et al., 2017)



Self-Attention

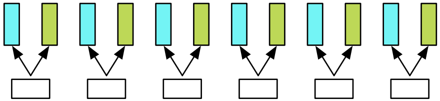
- Self-attention computes attention between elements of the same sequence
  - can replace RNNs as sequence model
  - shortens paths of credit assignment
  - at the core of Google's Transformer NMT system (Vaswani et al., 2017)



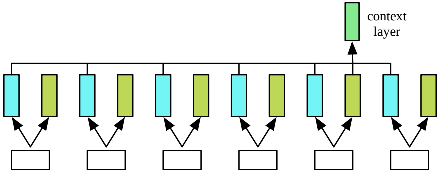
- self-attention is bidirectional (like a biRNN), but no recurrent connections between time steps

Transformer Encoder

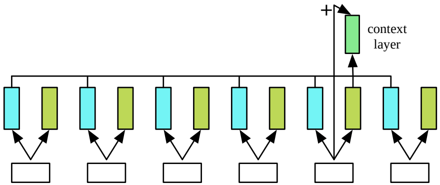
Transformer Encoder



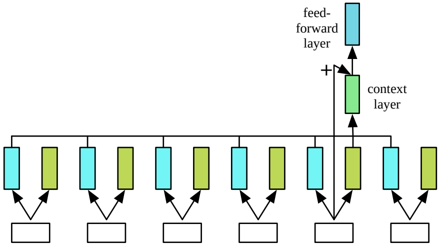
Transformer Encoder



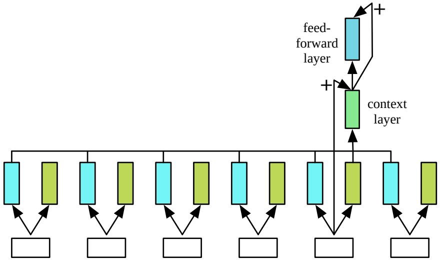
Transformer Encoder



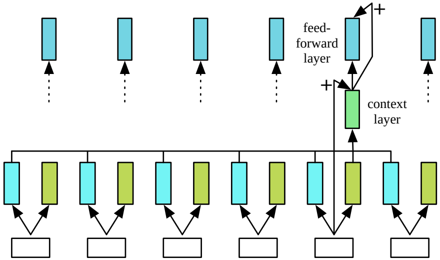
Transformer Encoder



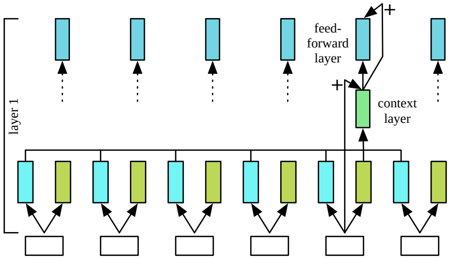
Transformer Encoder



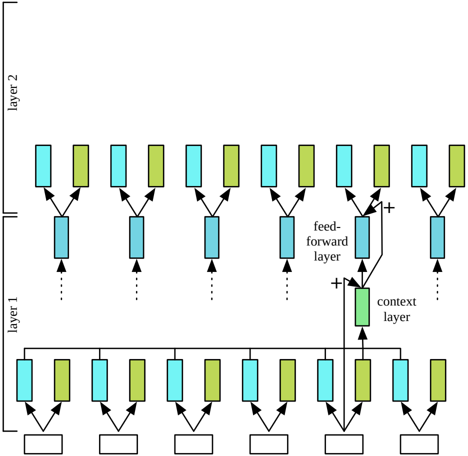
Transformer Encoder



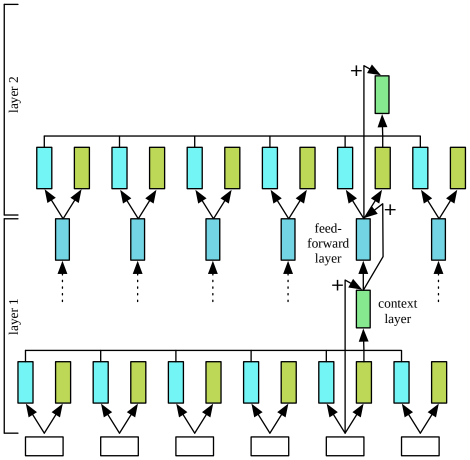
Transformer Encoder



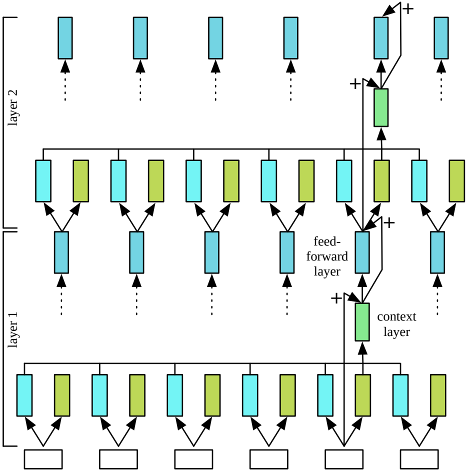
Transformer Encoder



Transformer Encoder



Transformer Encoder



Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) =$

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$ where $\mathbf{d}$ is a (inverted) dropout mask

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$
    where $\mathbf{d}$ is a (inverted) dropout mask
- Residual connections are used for context and feed-forward sub layers

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$ where $\mathbf{d}$ is a (inverted) dropout mask
- Residual connections are used for context and feed-forward sub layers
  - $\mathbf{f}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathrm{ffwd}(\mathbf{c}_{n,t}) + \mathbf{c}_{n,t})$

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$ where $\mathbf{d}$ is a (inverted) dropout mask
- Residual connections are used for context and feed-forward sub layers
  - $\mathbf{f}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathrm{ffwd}(\mathbf{c}_{n,t}) + \mathbf{c}_{n,t})$
  - $\mathbf{c}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathbf{c}_{n,t}) + \mathbf{f}_{n-1,t})$

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$
    where $\mathbf{d}$ is a (inverted) dropout mask
- Residual connections are used for context and feed-forward sub layers
  - $\mathbf{f}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathrm{ffwd}(\mathbf{c}_{n,t}) + \mathbf{c}_{n,t})$
  - $\mathbf{c}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathbf{c}_{n,t}) + \mathbf{f}_{n-1,t})$
    if $n = 1$, $\mathbf{f}_{n-1,t}$ refers to the word embedding at time t

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$
    where $\mathbf{d}$ is a (inverted) dropout mask
- Residual connections are used for context and feed-forward sub layers
  - $\mathbf{f}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathrm{ffwd}(\mathbf{c}_{n,t}) + \mathbf{c}_{n,t})$
  - $\mathbf{c}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathbf{c}_{n,t}) + \mathbf{f}_{n-1,t})$
    if $n = 1$, $\mathbf{f}_{n-1,t}$ refers to the word embedding at time t
- At a given time step t and layer n: $\mathbf{c}_{n,t}$ depends on $\mathbf{f}_{n-1,t}$ which in turn depends on $\mathbf{c}_{n-1,t}$, which depends ...

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$
    where $\mathbf{d}$ is a (inverted) dropout mask
- Residual connections are used for context and feed-forward sub layers
  - $\mathbf{f}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathrm{ffwd}(\mathbf{c}_{n,t}) + \mathbf{c}_{n,t})$
  - $\mathbf{c}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathbf{c}_{n,t}) + \mathbf{f}_{n-1,t})$
    if $n = 1$, $\mathbf{f}_{n-1,t}$ refers to the word embedding at time t
- At a given time step t and layer n: $\mathbf{c}_{n,t}$ depends on $\mathbf{f}_{n-1,t}$ which in turn depends on $\mathbf{c}_{n-1,t}$, which depends ...
  - in neural memory network parlance: multiple-hop attention

Transformer Sub-Layers

- The feed-forward layer is applied point-wise, i.e., at each time step t along a sequence (weights are shared)
- Feed-forward layer at layer n:
  - takes as input the context vector $\mathbf{c}_{n,t}$ of layer n at time t
  - is defined as $\mathrm{ffwd}(\mathbf{c}_{n,t}) = W_n \mathbf{d} \odot (\mathrm{ReLU}(V_n \mathbf{c}_{n,t} + \mathbf{a}_n)) + \mathbf{b}_n$
    where $\mathbf{d}$ is a (inverted) dropout mask
- Residual connections are used for context and feed-forward sub layers
  - $\mathbf{f}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathrm{ffwd}(\mathbf{c}_{n,t}) + \mathbf{c}_{n,t})$
  - $\mathbf{c}_{n,t} = \mathrm{LayerNorm}(\mathbf{d} \odot \mathbf{c}_{n,t}) + \mathbf{f}_{n-1,t})$
    if $n = 1$, $\mathbf{f}_{n-1,t}$ refers to the word embedding at time t
- At a given time step t and layer n: $\mathbf{c}_{n,t}$ depends on $\mathbf{f}_{n-1,t}$ which in turn depends on $\mathbf{c}_{n-1,t}$, which depends ...
  - in neural memory network parlance: multiple-hop attention
- What does LayerNorm do?

Transformer Decoder

- Decoder architecture very similar to encoder, but

Transformer Decoder

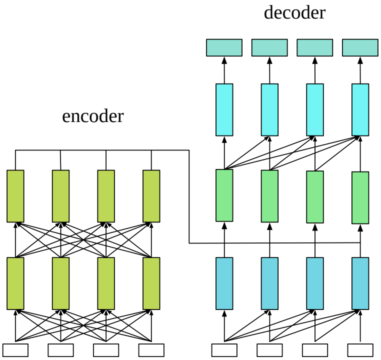
- Decoder architecture very similar to encoder, but
  - needs to integrate decoder-encoder attention

Transformer Decoder

- Decoder architecture very similar to encoder, but
  - needs to integrate decoder-encoder attention
  - self-attention has to be limited to previous time-steps

Transformer Decoder

- Decoder architecture very similar to encoder, but
  - needs to integrate decoder-encoder attention
  - self-attention has to be limited to previous time-steps



The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target
  - complex many-to-many translation mappings

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target
  - complex many-to-many translation mappings
  - long-distance dependencies on the target side

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target
  - complex many-to-many translation mappings
  - long-distance dependencies on the target side
- At this moment Transformer models constitute the state-of-the-art in MT

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target
  - complex many-to-many translation mappings
  - long-distance dependencies on the target side
- At this moment Transformer models constitute the state-of-the-art in MT
  - translations of almost human quality for several language pairs, e.g., French-English, Spanish-English, but also German-English

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target
  - complex many-to-many translation mappings
  - long-distance dependencies on the target side
- At this moment Transformer models constitute the state-of-the-art in MT
  - translations of almost human quality for several language pairs, e.g., French-English, Spanish-English, but also German-English
- However, there are a number of open problems

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target
  - complex many-to-many translation mappings
  - long-distance dependencies on the target side
- At this moment Transformer models constitute the state-of-the-art in MT
  - translations of almost human quality for several language pairs, e.g., French-English, Spanish-English, but also German-English
- However, there are a number of open problems
  - deep learning models are data-hungry and perform less well on **language pairs with limited resources** (e.g., Vietnamese-English, Uzbek-English, Hausa-English, . . .)

The State-of-the-Art in MT

- Neural machine translation (NMT) alleviates many of the problems that have plagued SMT for years
  - long-distance reorderings between source and target
  - complex many-to-many translation mappings
  - long-distance dependencies on the target side
- At this moment Transformer models constitute the state-of-the-art in MT
  - translations of almost human quality for several language pairs, e.g., French-English, Spanish-English, but also German-English
- However, there are a number of open problems
  - deep learning models are data-hungry and perform less well on **language pairs with limited resources** (e.g., Vietnamese-English, Uzbek-English, Hausa-English, . . .)
  - language pairs involving **morphologically rich languages**, such as Finish, Turkish, Arabic (as source and/or target)

Recap

- Machine translation
- Sequence-to-sequence models
- Neural machine translation
  - encoder-decoder architecture
  - attention mechanism
  - self-attention (Google's Transformer)