Machine Translation with Diverse Data Sources

Huda Khayrallah

This talk was presented at NYU Abu Dhabi

CS seminar on September 1, 2019

It is based on the following papers:

https://aclweb.org/anthology/W18-2705

(bibtex: https://aclweb.org/anthology/W18-2705)

https://aclweb.org/anthology/W18-2709

(bibtex: https://aclweb.org/anthology/W18-2709.bib)

Machine Translation with Diverse Data Sources

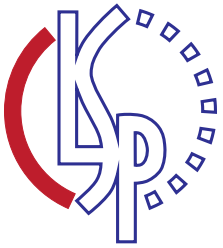
Huda Khayrallah

huda@jhu.edu

Work with:

Brian Thompson, Kevin Duh & Philipp Koehn

JOHNS HOPKINS
UNIVERSITY



Overview

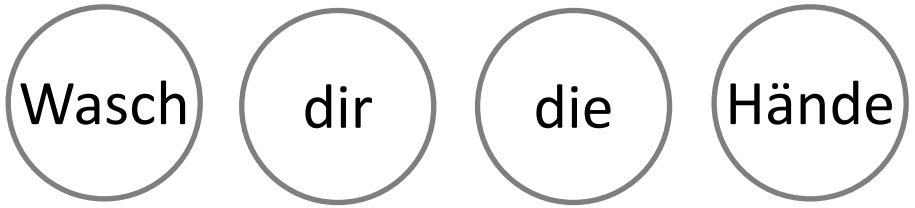
- Overview of Neural Machine Translation (NMT)
- Overview of Domain Adaptation
- Improving Domain Adaptation
  - Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
- Analysis of Noisy Corpora
  - On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]

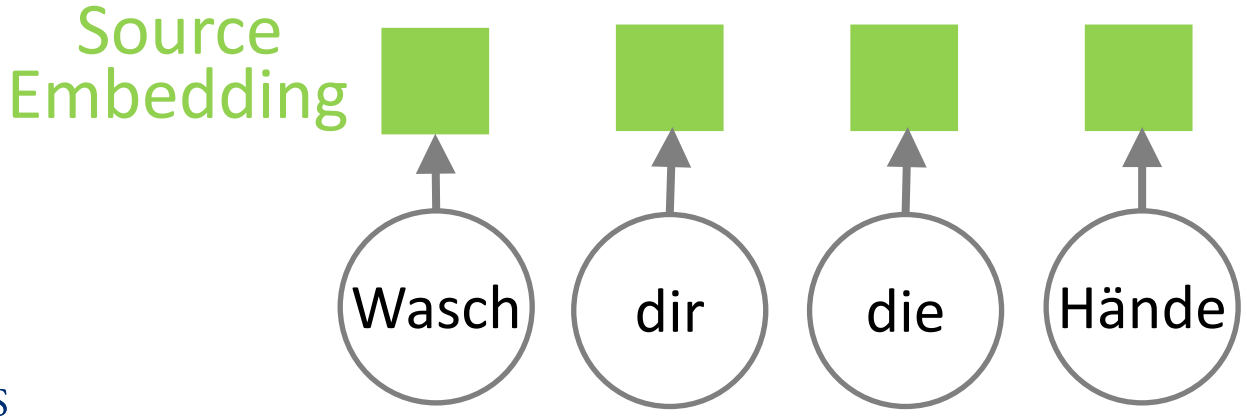
Machine Translation

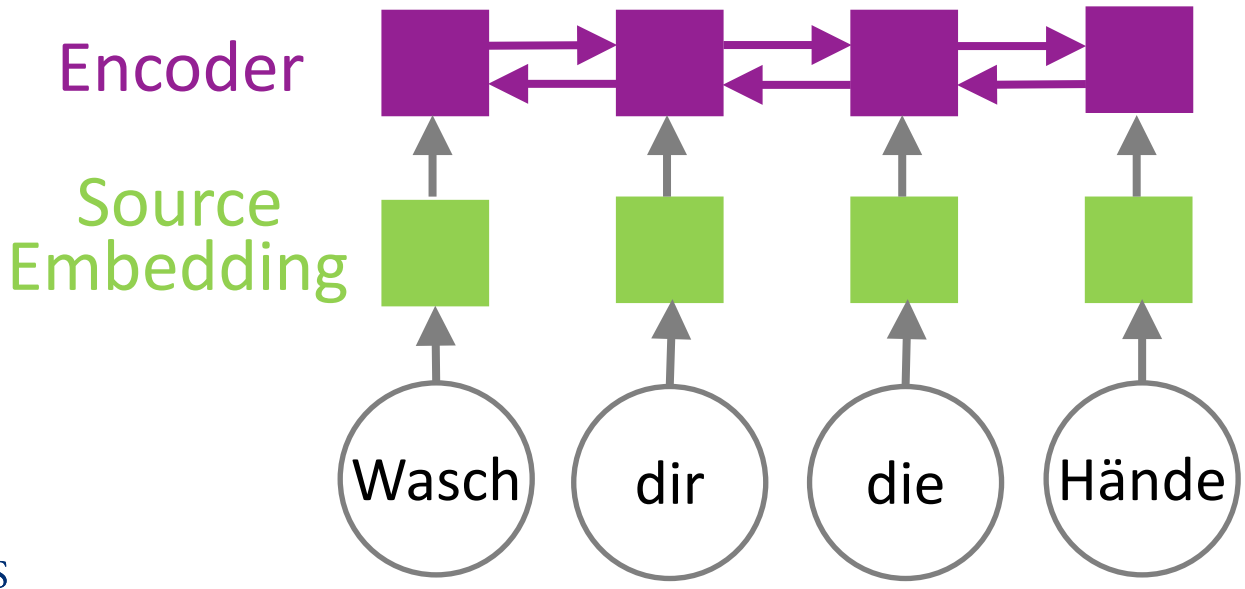
Neural Machine Translation

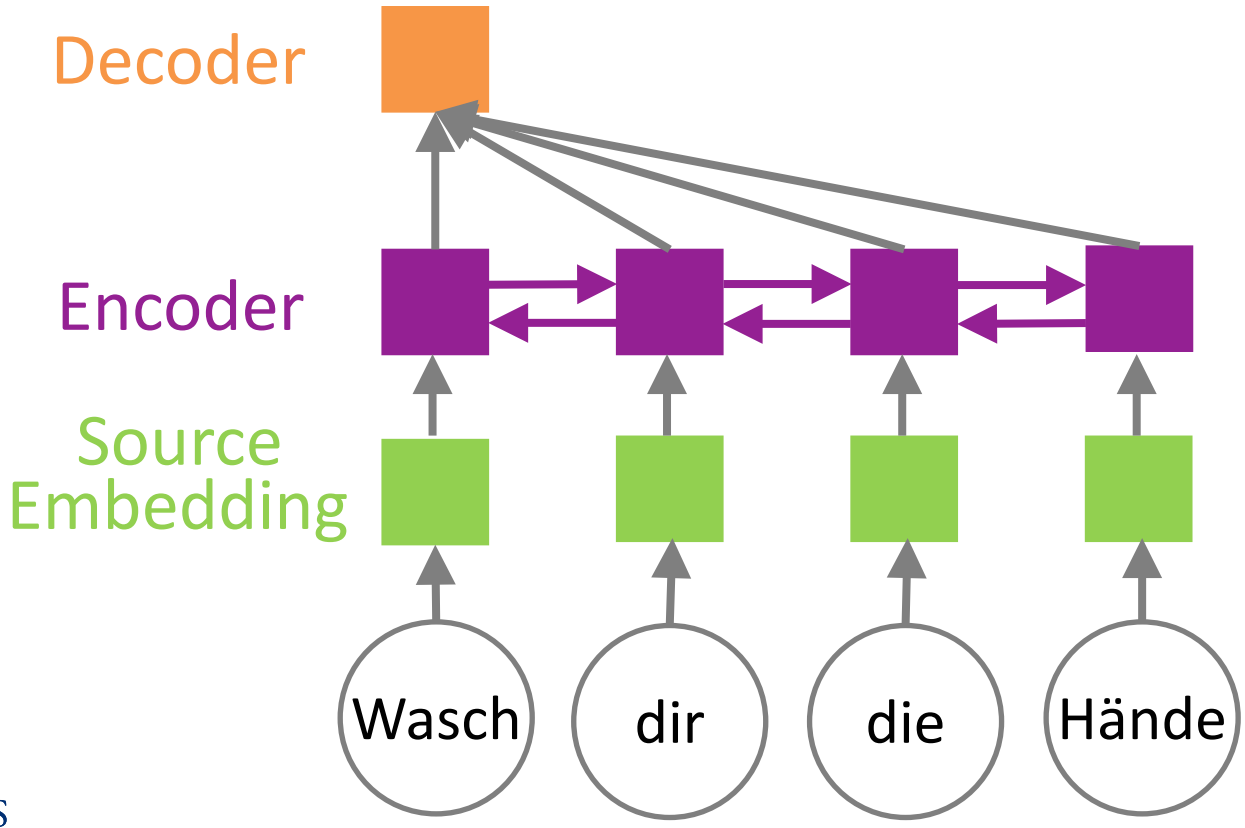
Parallel Text

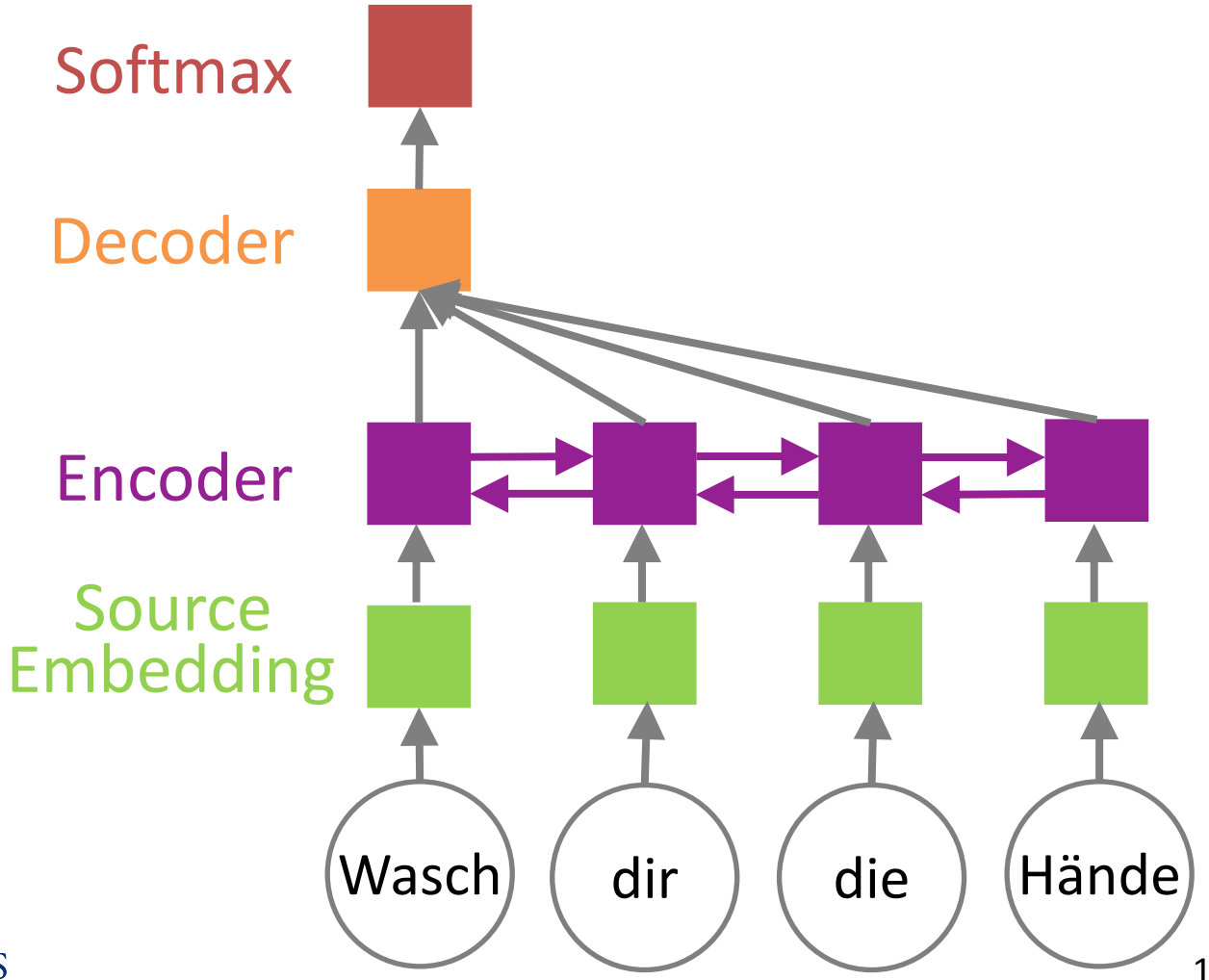
Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

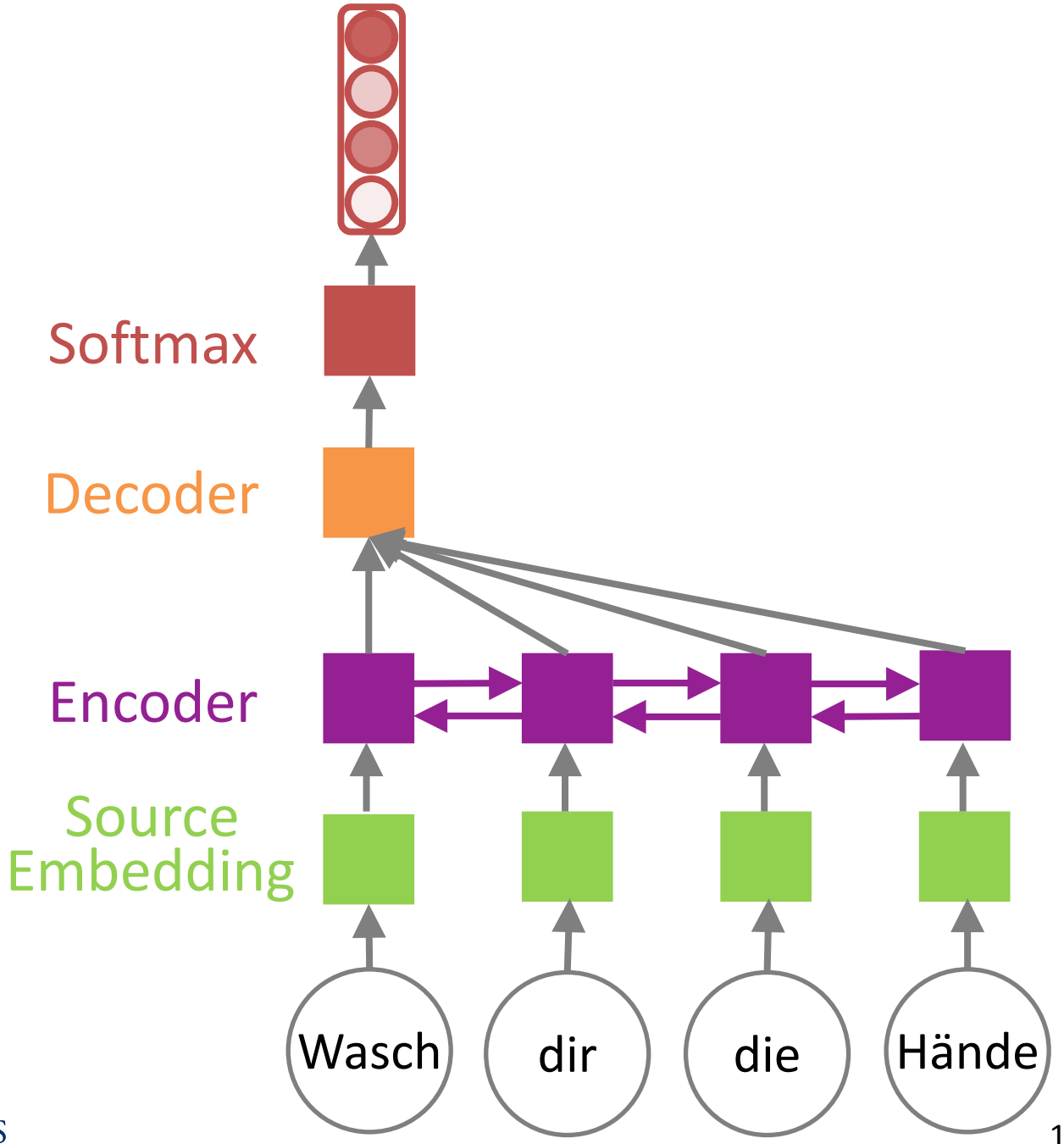

Wasch
dir
die
Hände

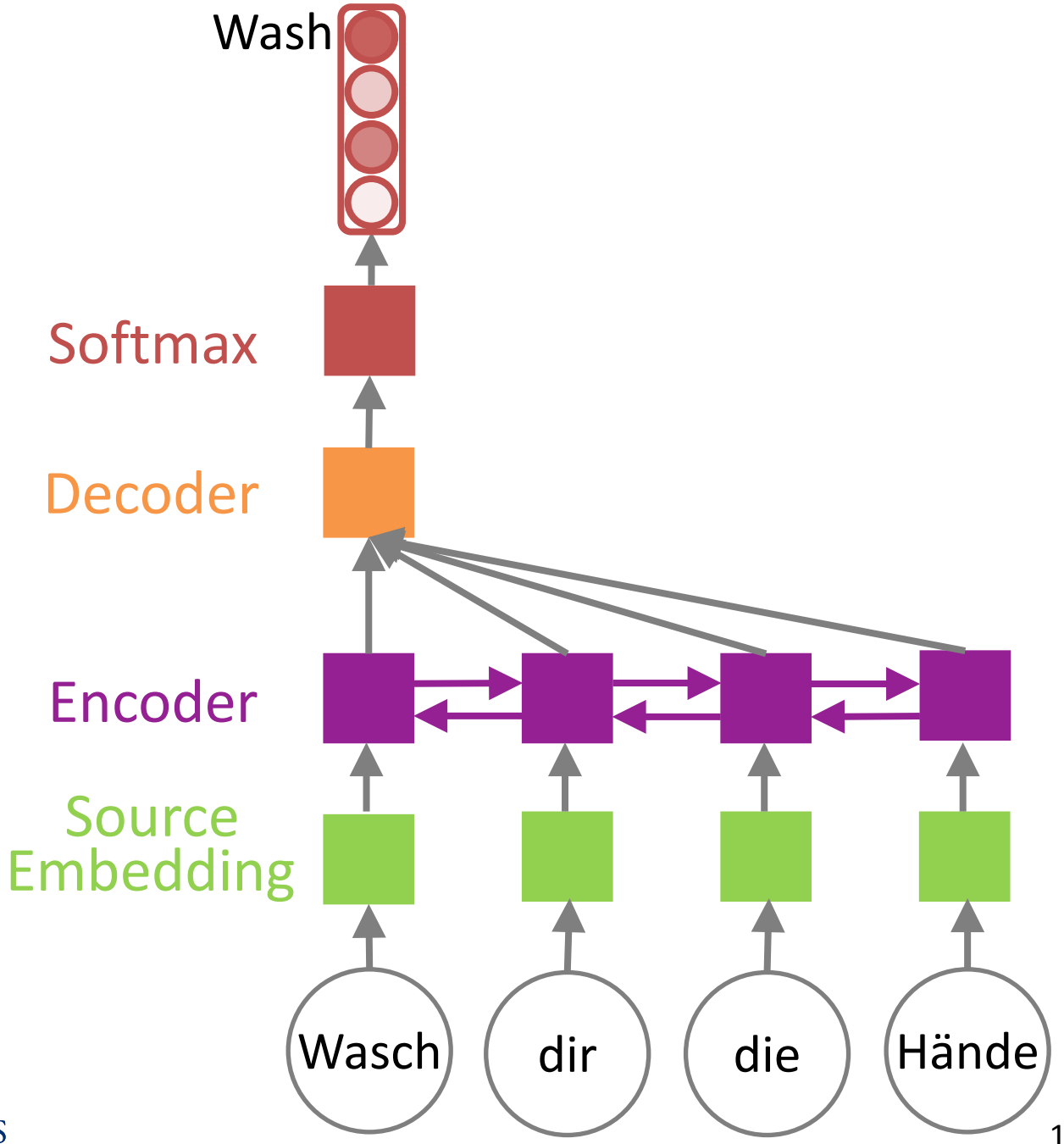

Source
Embedding
Wasch
dir
die
Hände

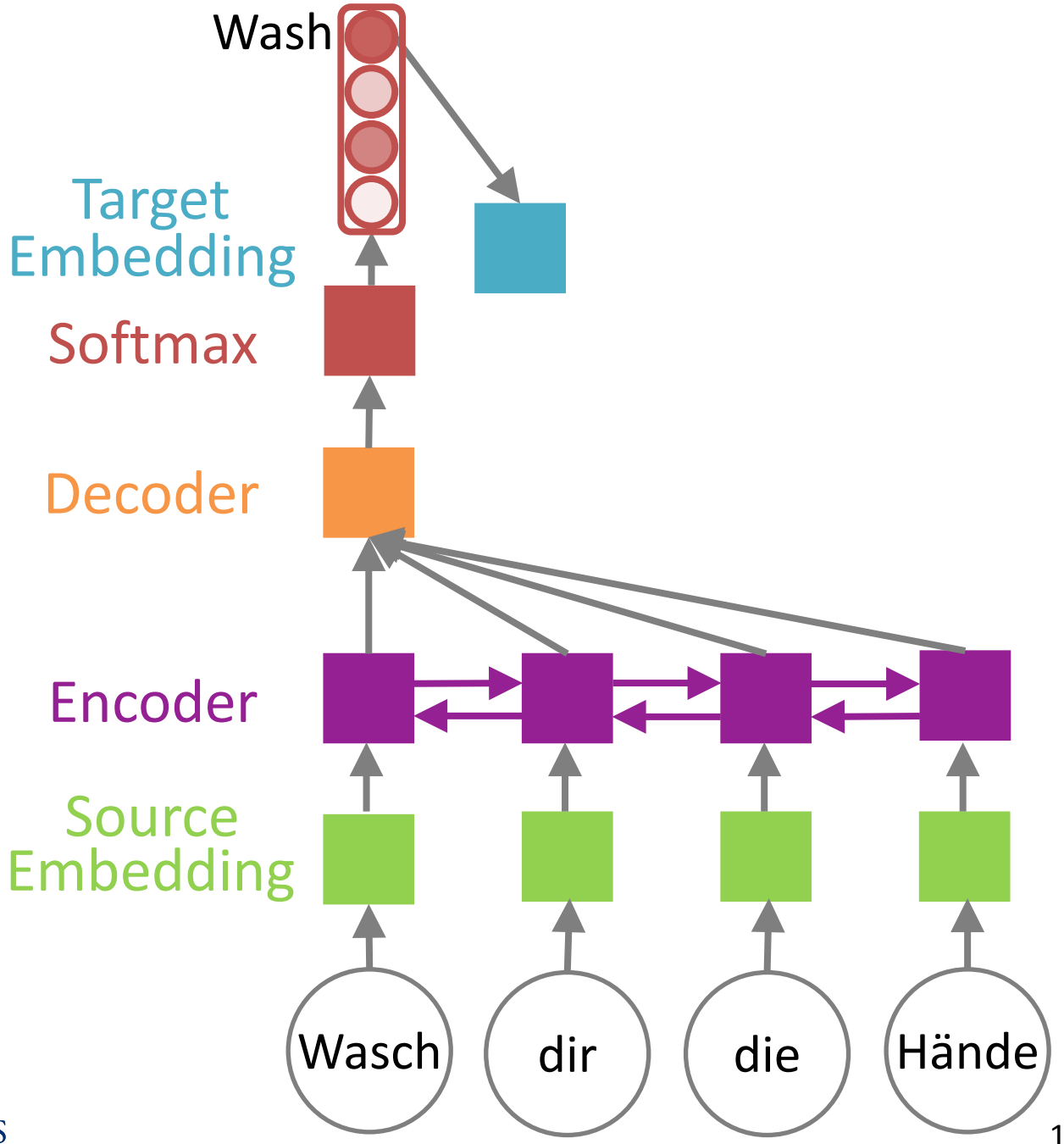

Encoder
Source
Embedding
Wasch
dir
die
Hände

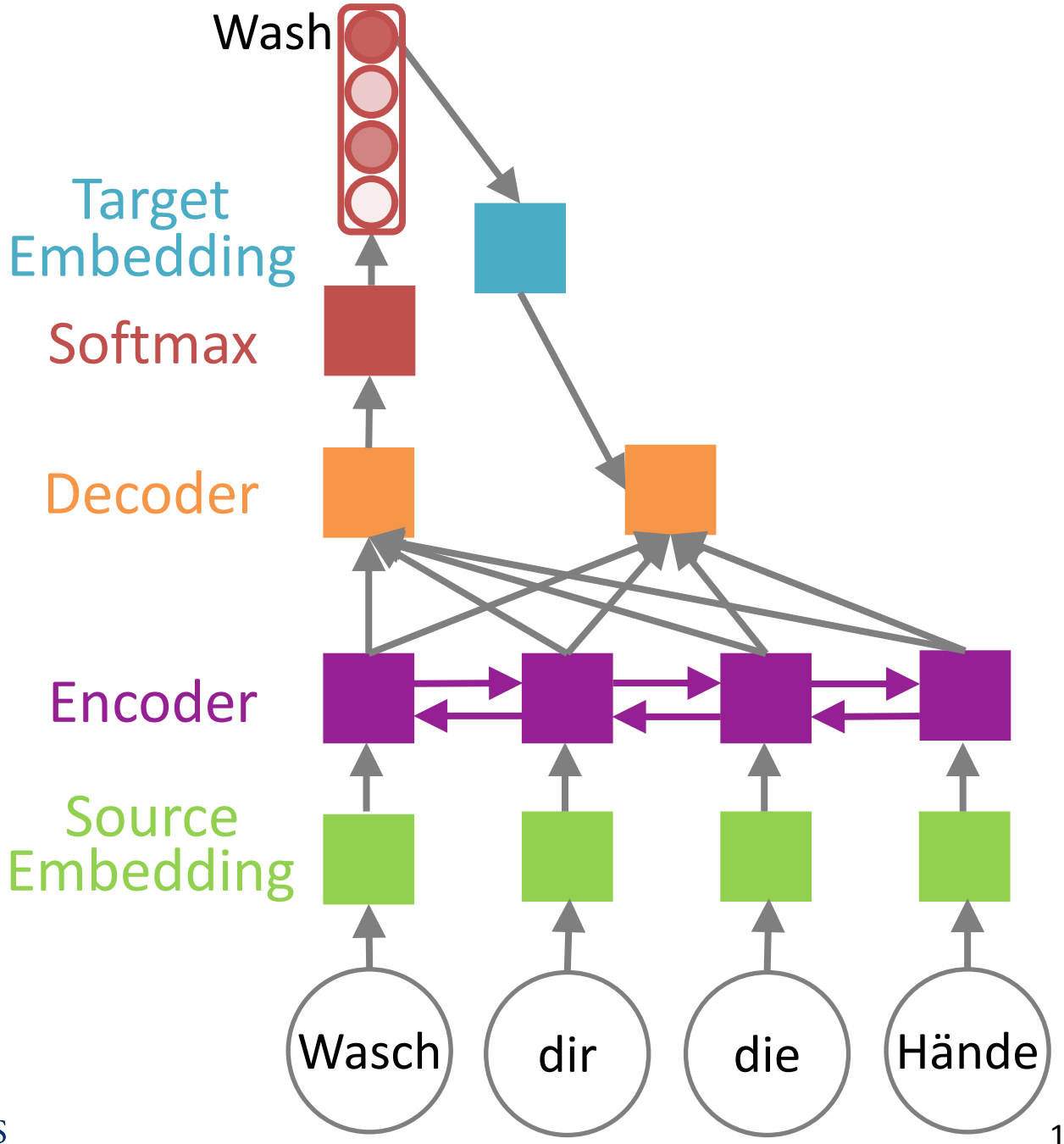

Decoder
Encoder
Source
Embedding
Wasch
dir
die
Hände

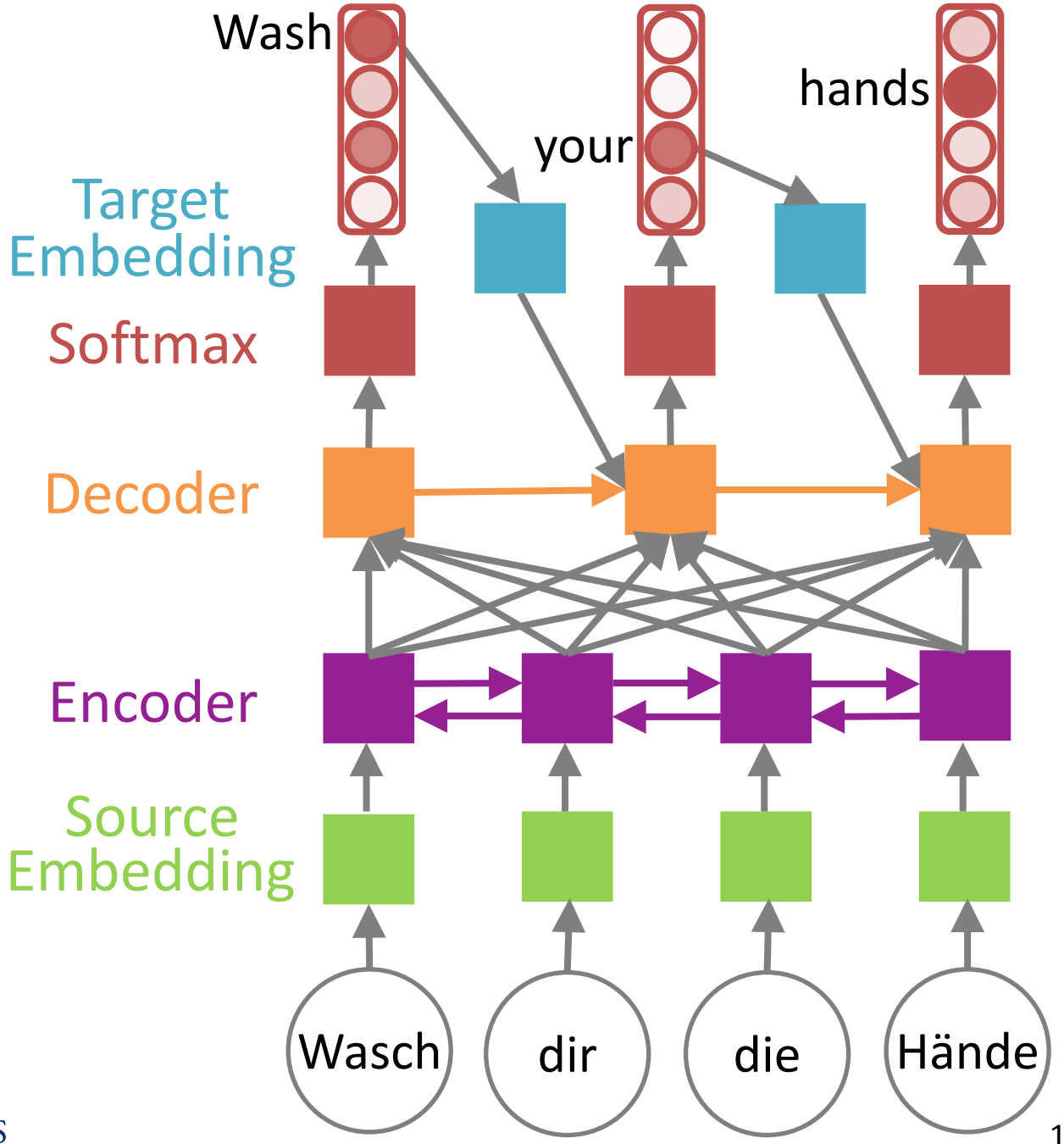

Softmax
Decoder
Encoder
Source
Embedding
Wasch
dir
die
Hände


Softmax
Decoder
Encoder
Source
Embedding
Wasch
dir
die
Hände


Wash
Softmax
Decoder
Encoder
Source
Embedding
Wasch
dir
die
Hände


Wash
Target Embedding
Softmax
Decoder
Encoder
Source Embedding
Wasch
dir
die
Hände


Wash
Target Embedding
Softmax
Decoder
Encoder
Source Embedding
Wasch
dir
die
Hände

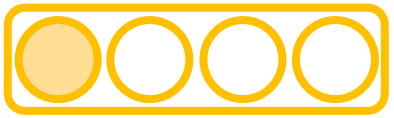
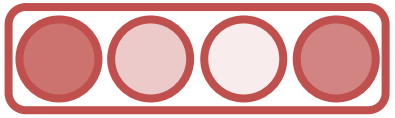

Wash
your
hands
Target Embedding
Softmax
Decoder
Encoder
Source Embedding
Wasch
dir
die
Hände

NMT loss function

$$\mathcal{L}_{\text{NLL}}(\theta) = -\sum_{v \in \mathcal{V}} \left(\mathbb{1}\{y_i = v\} \times \log p(y_i = v \mid x; \theta; y_{j<i}) \right)$$

Gold Target

Model output

Cross Entropy( , )

Gold Target Model output

BLEU

- Weighted *n*-grams precision

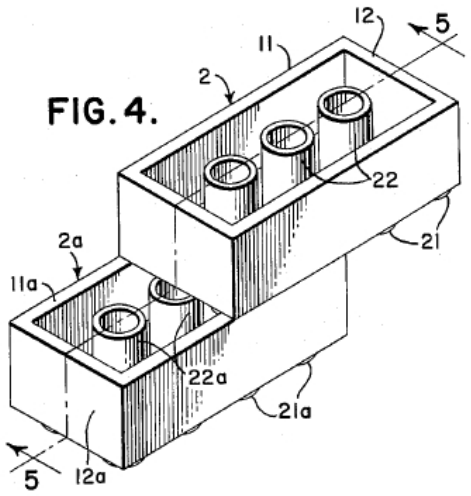
$$min\left(1, \frac{output\ length}{reference\ length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}$$

- Between 0 and 1
  - (**often scaled to be 0-100**)
- **Higher is better**
- Imperfect…
- But… not bad

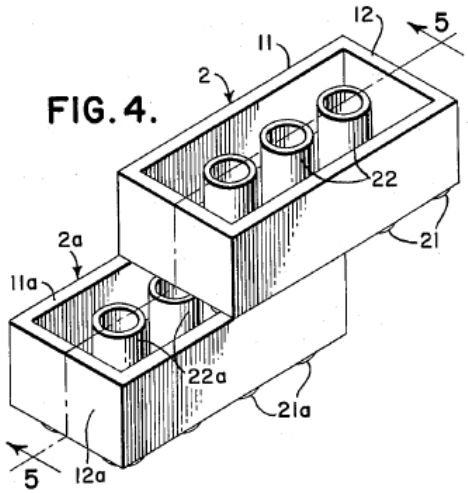
Overview

- Overview of Neural Machine Translation (NMT)
- **Overview of Domain Adaptation**
- Improving Domain Adaptation
  - Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
- Analysis of Noisy Corpora
  - On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]

What do we want to translate?

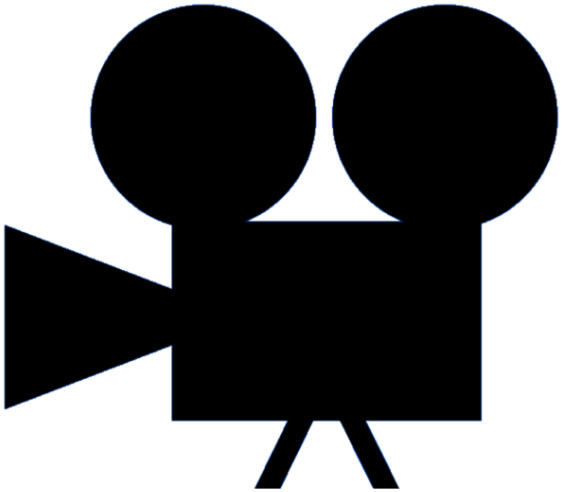


Developmental toxicity, including dose-dependent delayed foetal ossification and possible teratogenic effects, were observed in rats at doses resulting in subtherapeutic exposures (based on AUC) and in rabbits at doses resulting in exposures 3 and 11 times the mean steady-state AUC at the maximum recommended clinical dose.



The films coated therewith, in particular polycarbonate films coated therewith, have improved properties with regard to scratch resistance, solvent resistance, and reduced oiling effect, said films thus being especially suitable for use in producing plastic parts in film insert molding methods.

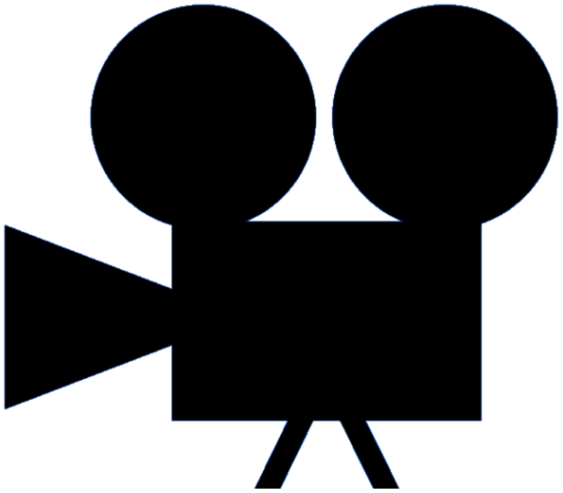
General Domain Data



General Domain Data

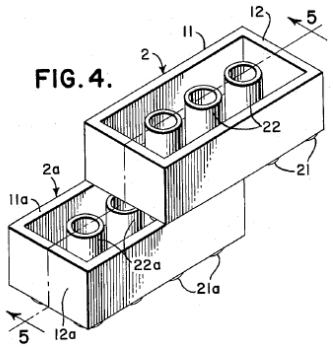
Would it not be beneficial, in the short term, following the Rotterdam model, to inspect according to a points system in which, for example, account is taken of the ship's age, whether it is single or double-hulled or whether it sails under a flag of convenience.

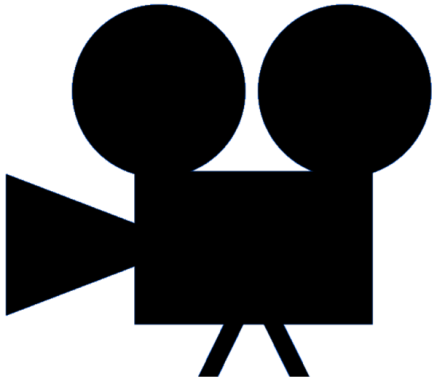
General Domain Data



Mama always said there's an awful lot you can tell about a person by their shoes.

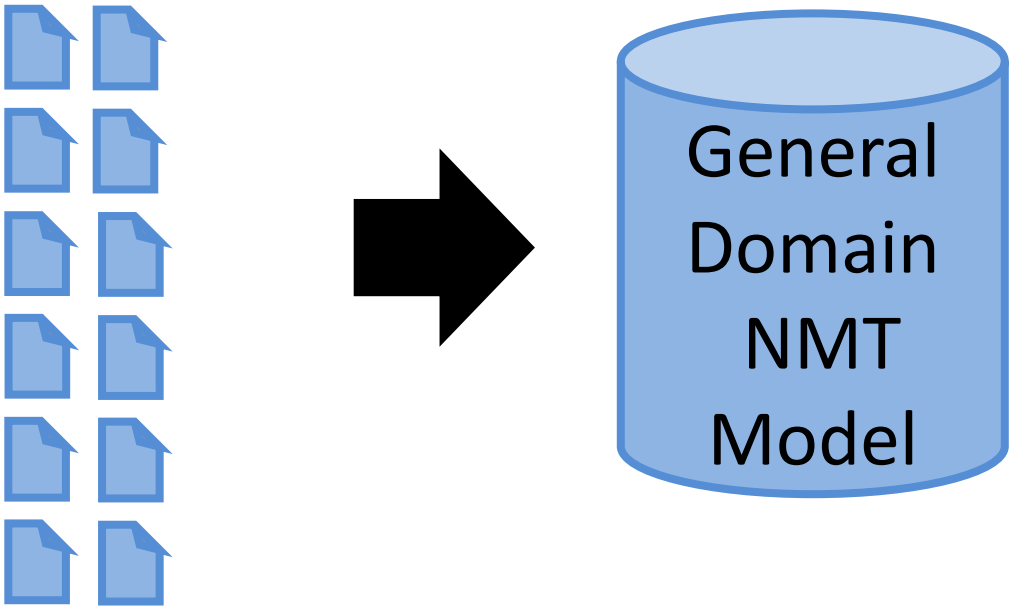
Domain Mismatch





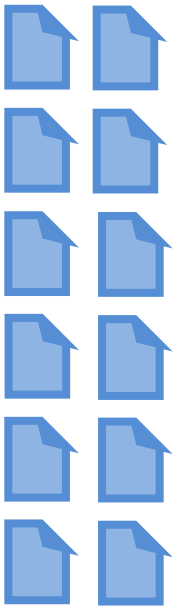
Case Study:
Translating Russian Patents

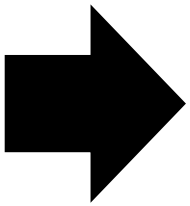
General Domain NMT

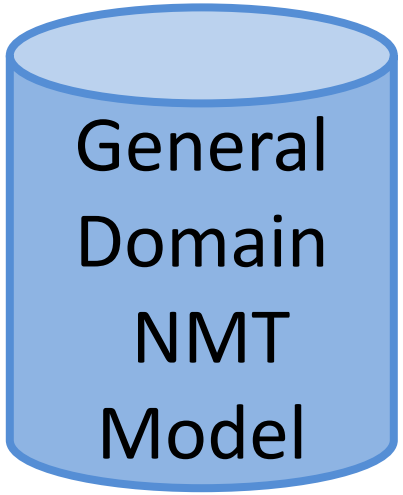


50m General Domain sentence pairs

General Domain NMT







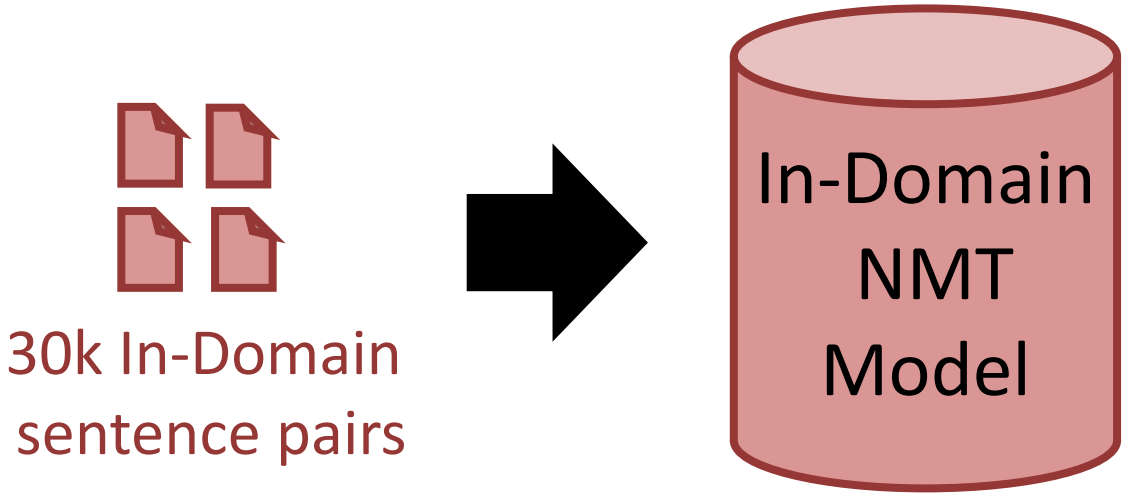
50m General Domain sentence pairs

дверной замок повышенной степени защищенности от взлома

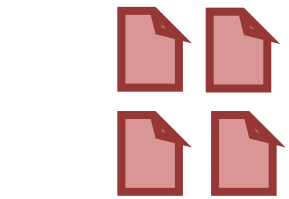
Human: door lock with increased degree of security against burglary

System: door security door security door

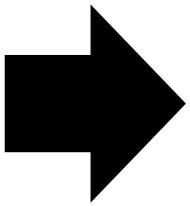
In-Domain NMT

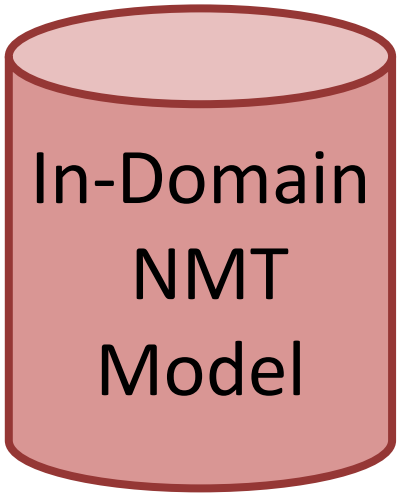


In-Domain NMT



30k In-Domain sentence pairs

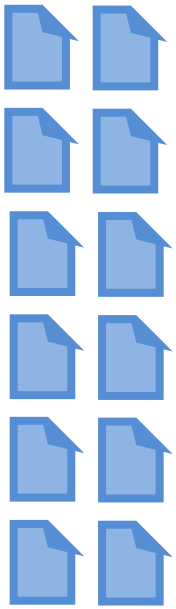




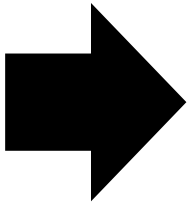
дверной замок повышенной степени защищенности от взлома
Human: door lock with increased degree of security against burglary
System: door lock for a high degree of protection against coke

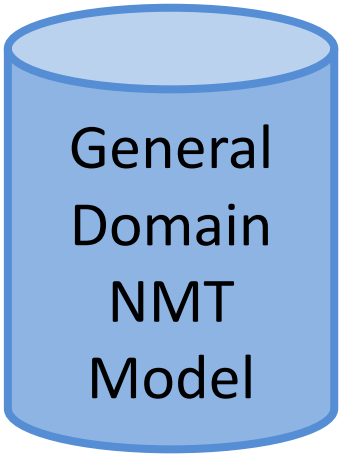
Domain Adaptation

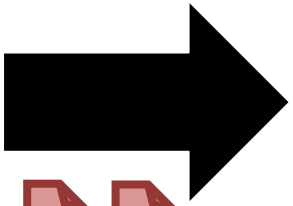
Continued Training

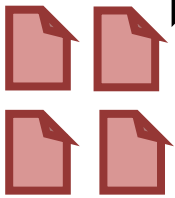


50m General Domain sentence pairs

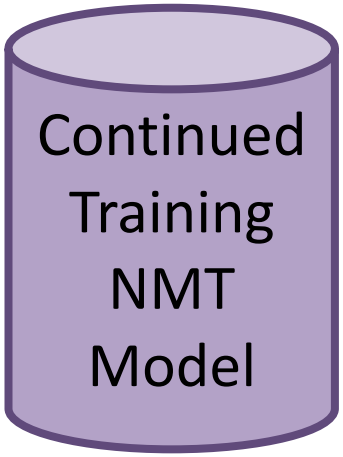




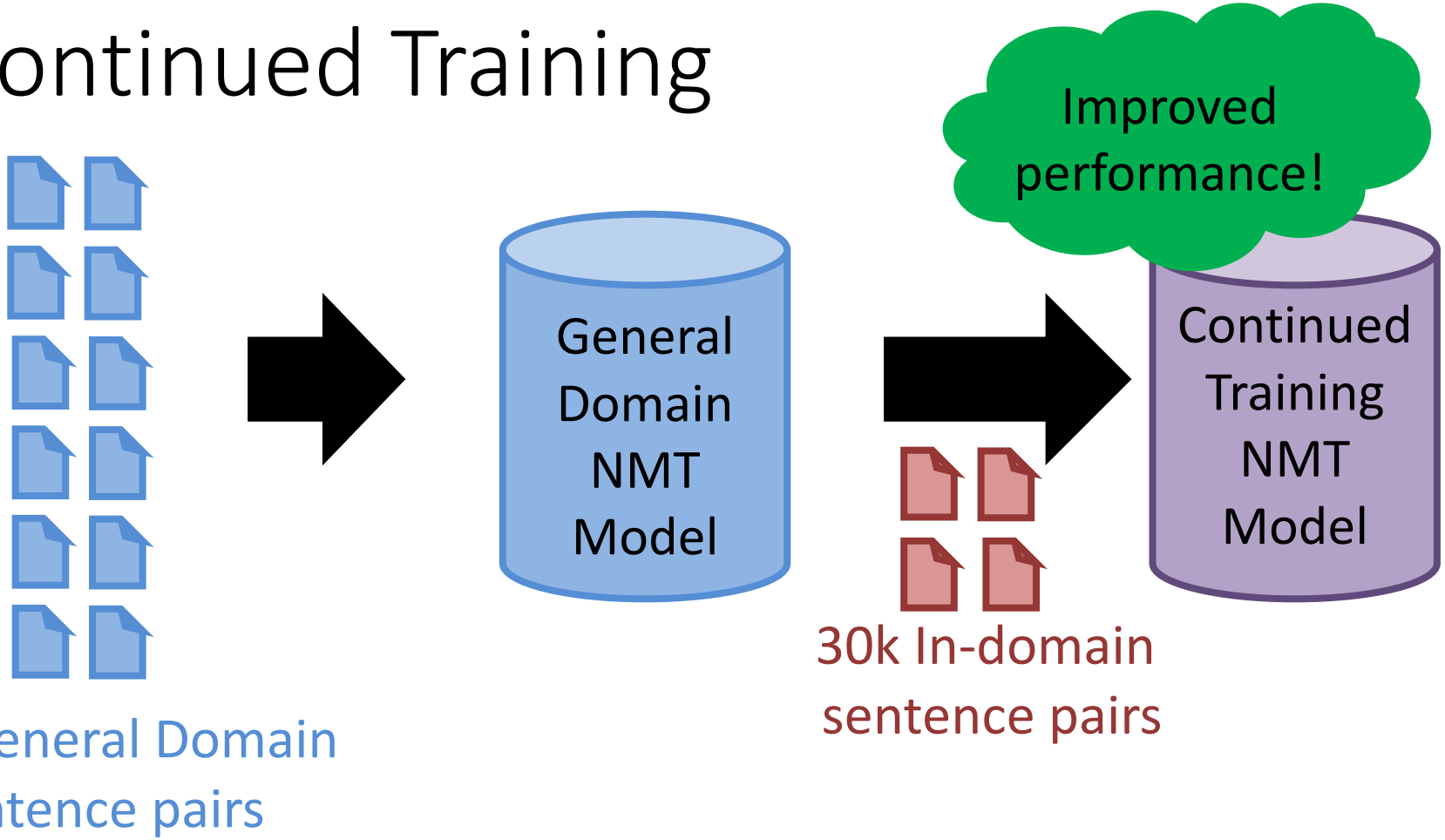




30k In-domain sentence pairs

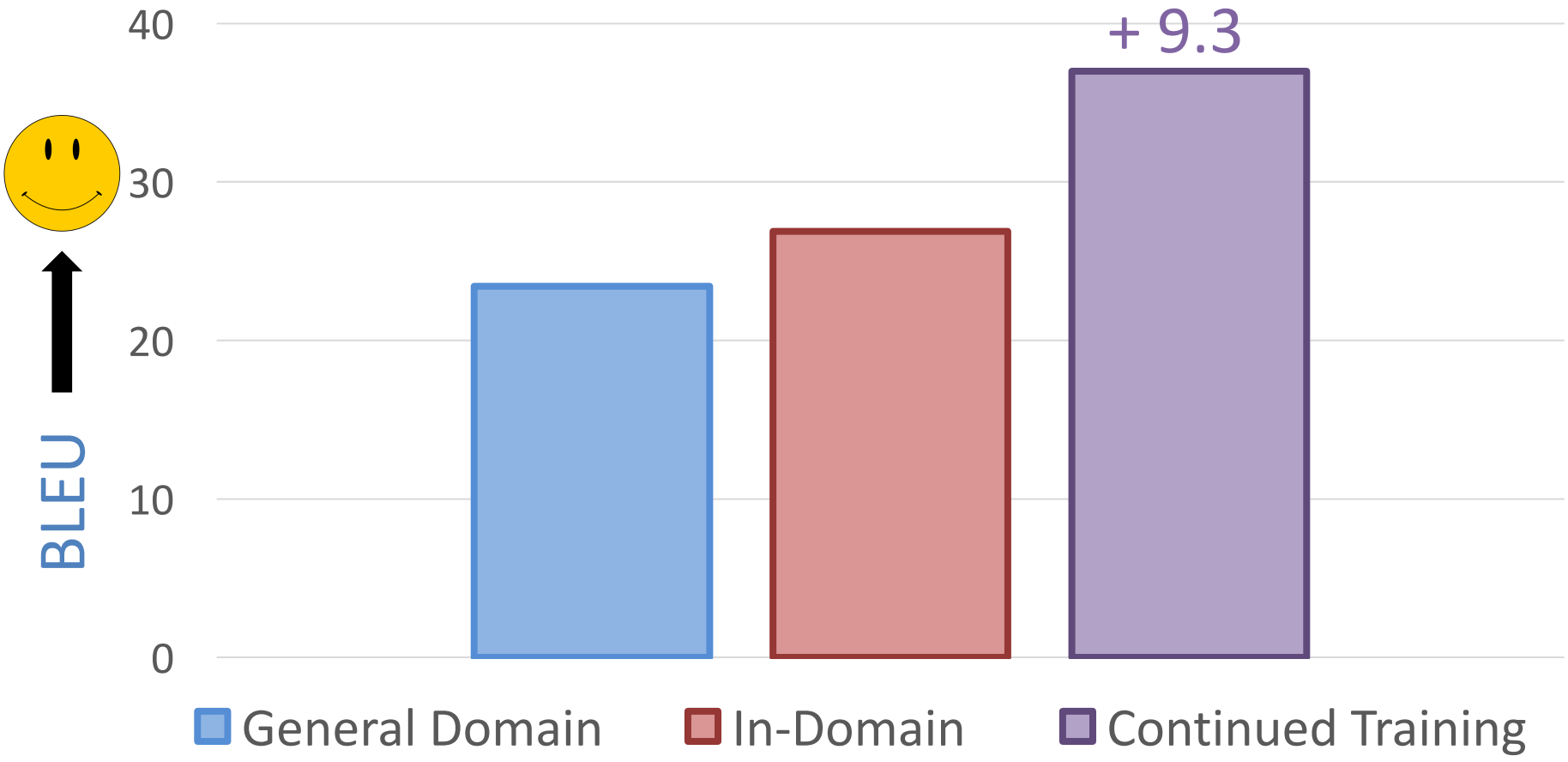


Continued Training

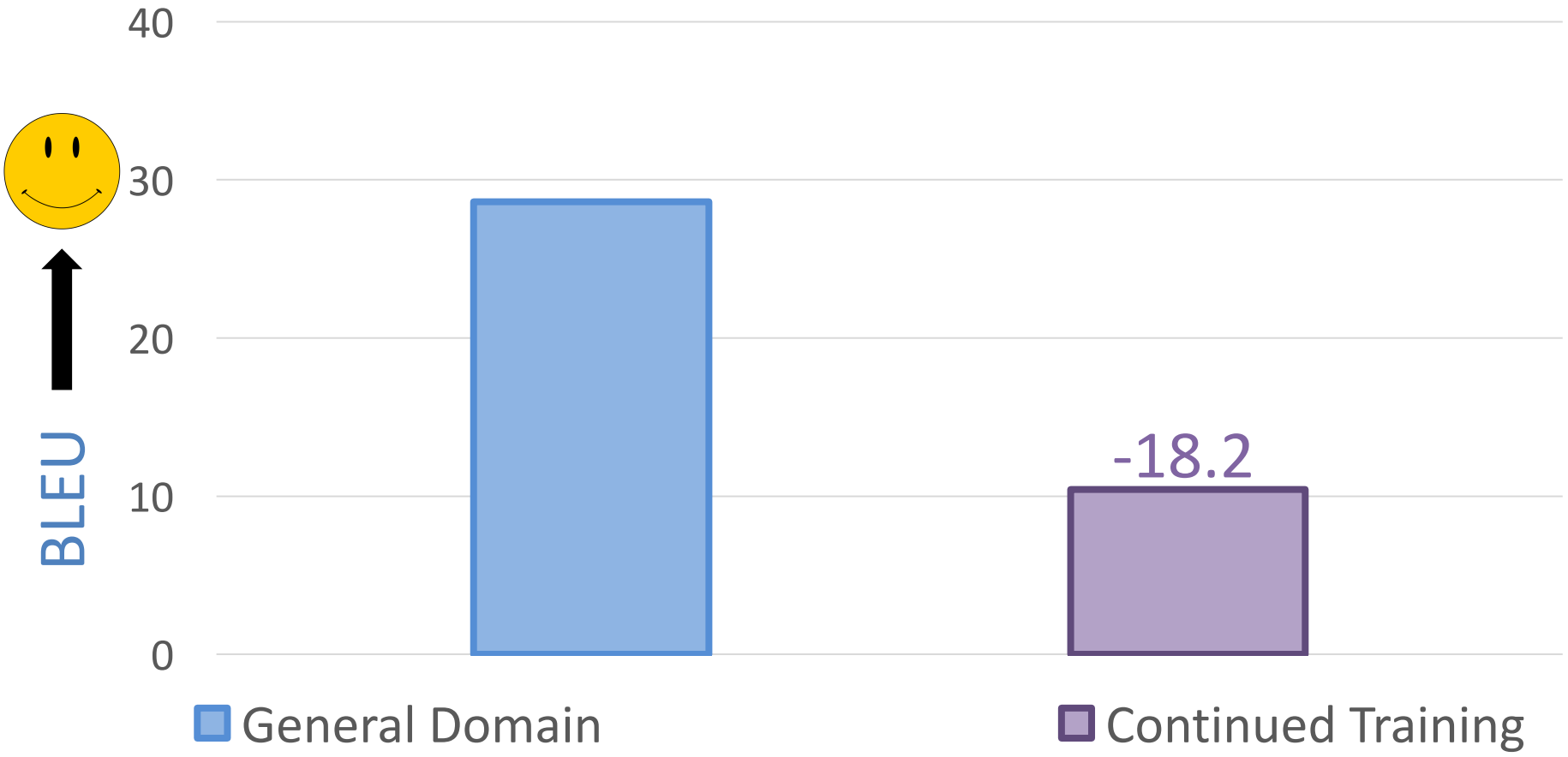


дверной замок повышенной степени защищенности от взлома
Human: door lock with increased degree of security against burglary
System: door lock with increased penetration protection

Russian → English Patents



Russian → English General



Overview

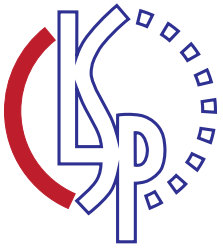
- Overview of Neural Machine Translation (NMT)
- Overview of Domain Adaptation
- **Improving Domain Adaptation**
  - Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
- Analysis of Noisy Corpora
  - On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]

Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation

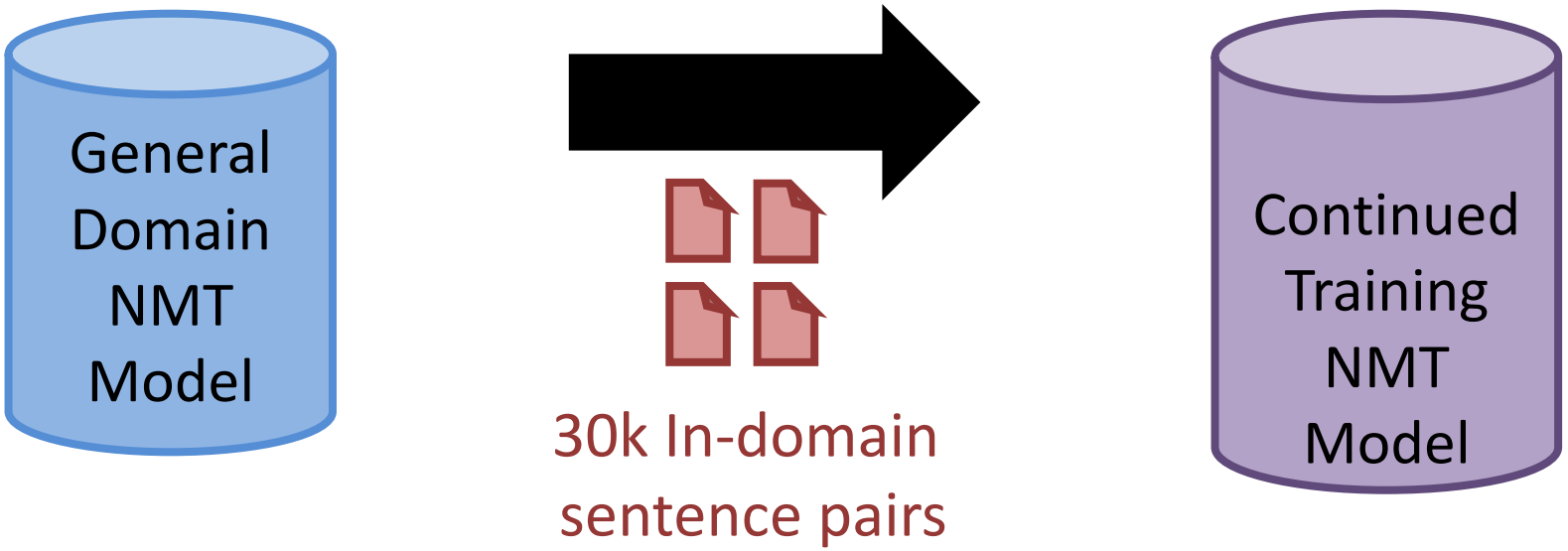
Huda Khayrallah, Brian Thompson,
Kevin Duh & Philipp Koehn

WNMT at ACL 2018

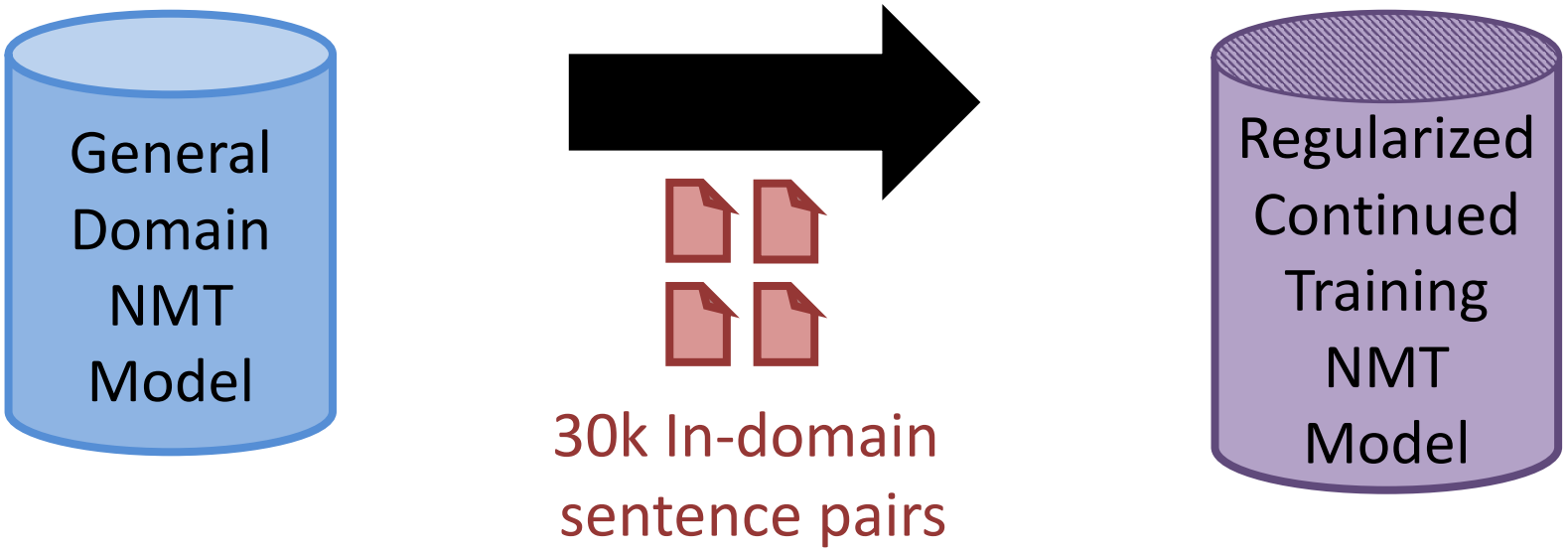
JOHNS HOPKINS
UNIVERSITY



Continued Training



Regularized Continued Training

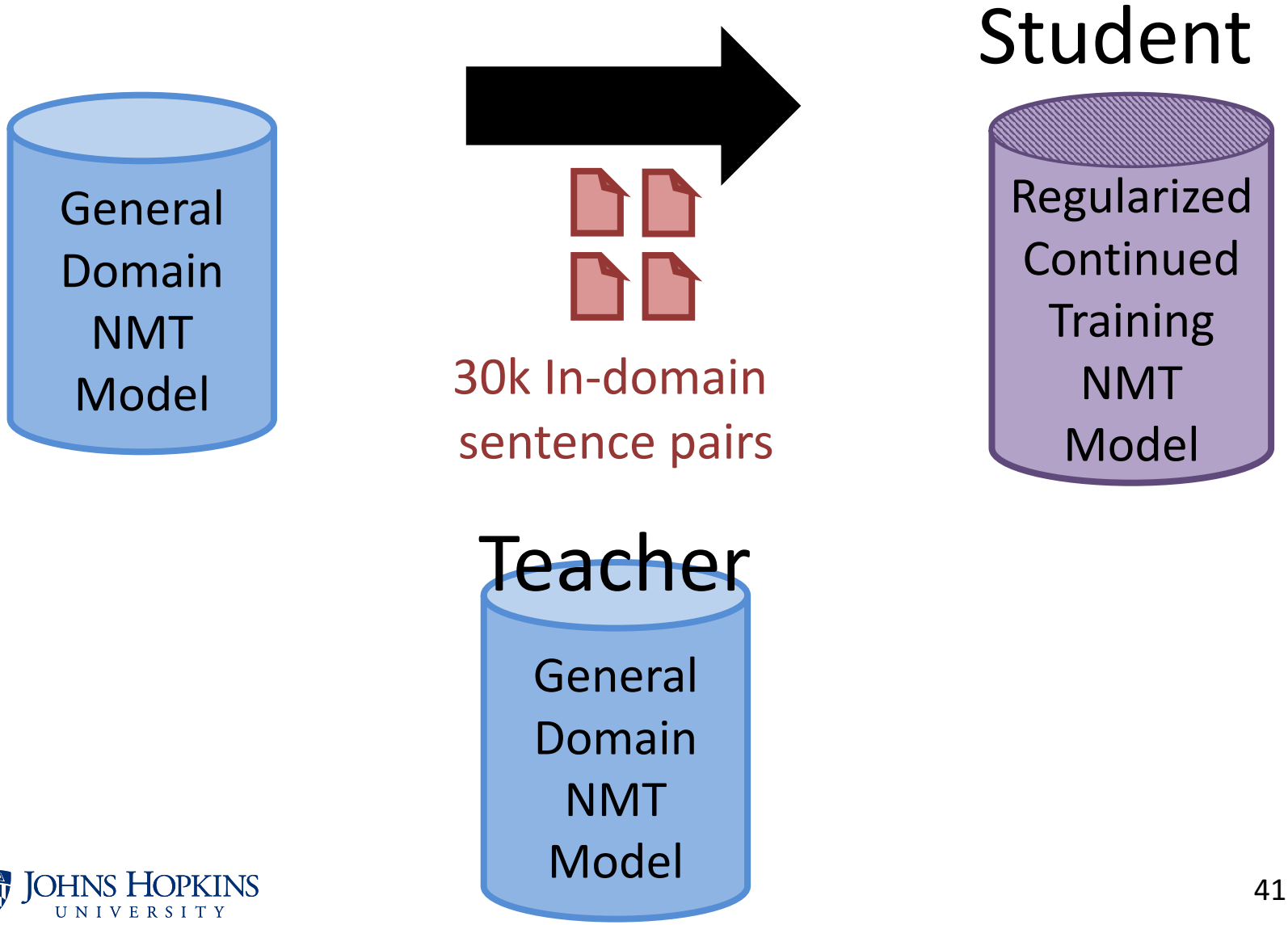


Teacher/Student Models

- Word Level Knowledge distillation

- Often used to make smaller/faster models
- Train one model; use it to ‘teach’ another

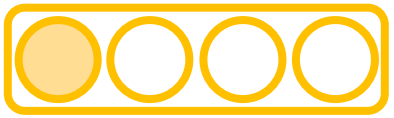
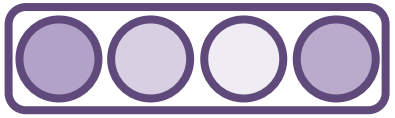
Regularized Continued Training



NMT loss function

$$\mathcal{L}_{\text{NLL}}(\theta) = -\sum_{v \in \mathcal{V}} \left(\mathbb{1}\{y_i = v\} \times \log p(y_i = v \mid x; \theta; y_{j<i}) \right)$$

Gold Target **CT Model output**

Cross Entropy( , )

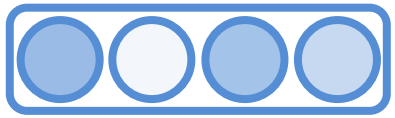
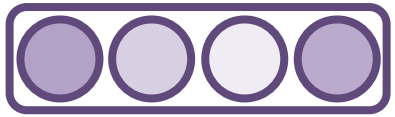
Gold Target **CT Model output**

Teacher/Student Loss Function

$$-\sum_{v \in \mathcal{V}} \left(p_{aux}(y_i = v \mid x; \theta_{aux}; y_{j<i}) \times \log p(y_i = v \mid x; \theta; y_{j<i}) \right)$$

General Model Output (teacher)

CT Model output (student)

Cross Entropy( , )

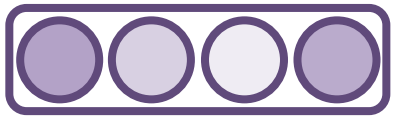
General Model Output (teacher)

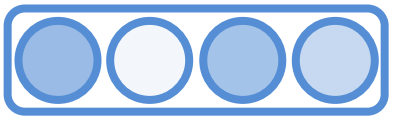
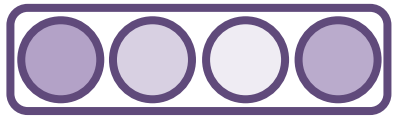
CT Model output (student)

This work: Combine Both

$$(1 - \alpha) \times \left(-\sum_{v \in \mathcal{V}} \left(\mathbb{1}\{y_i = v\} \times \log p(y_i = v \mid x; \theta; y_{j<i}) \right) \right) +$$

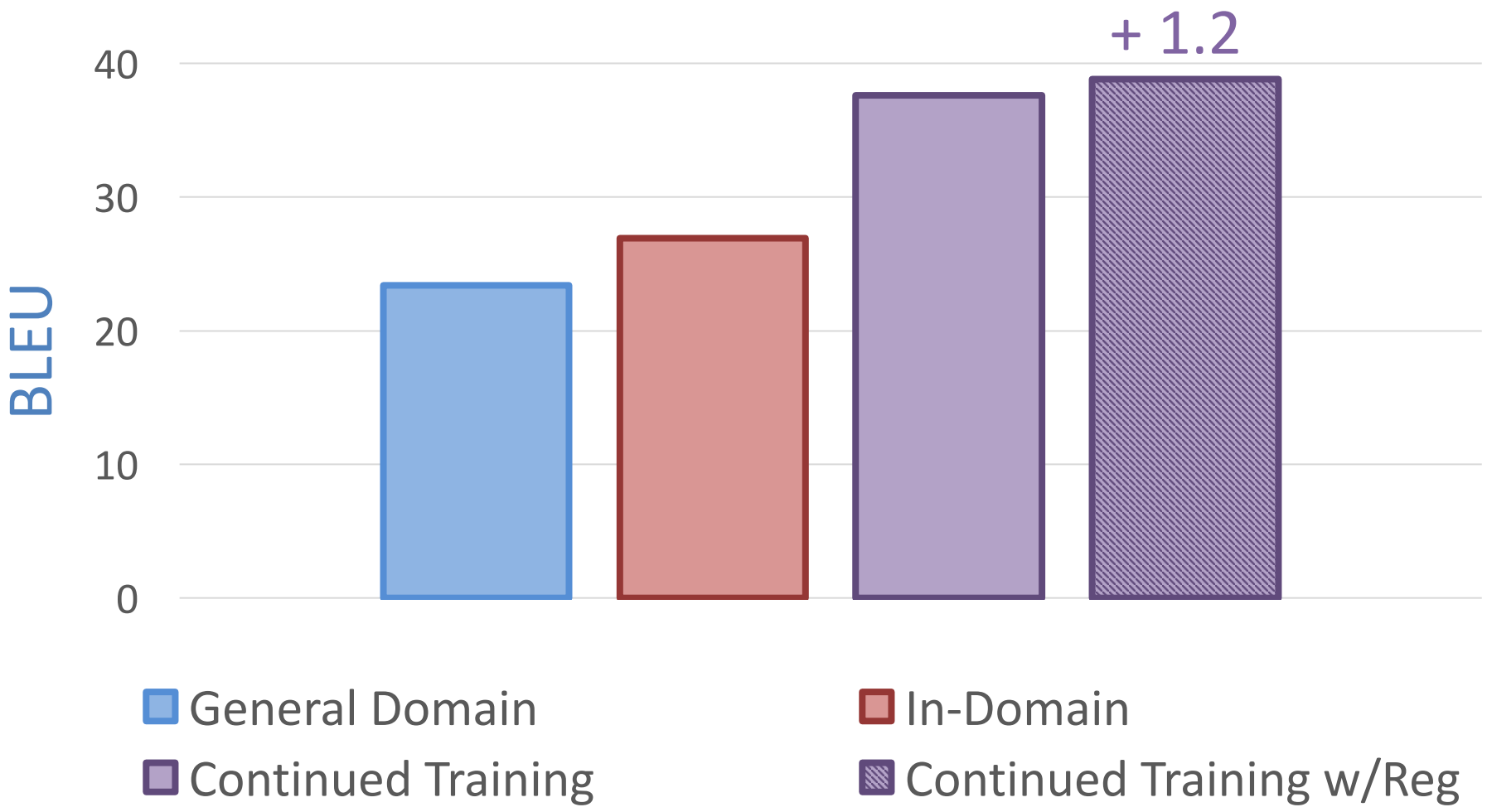
$$\alpha \times \left(-\sum_{v \in \mathcal{V}} \left(p_{aux}(y_i = v \mid x; \theta_{aux}; y_{j<i}) \times \log p(y_i = v \mid x; \theta; y_{j<i}) \right) \right)$$

$(1 - \alpha) \times$ Cross Ent ( , ) +

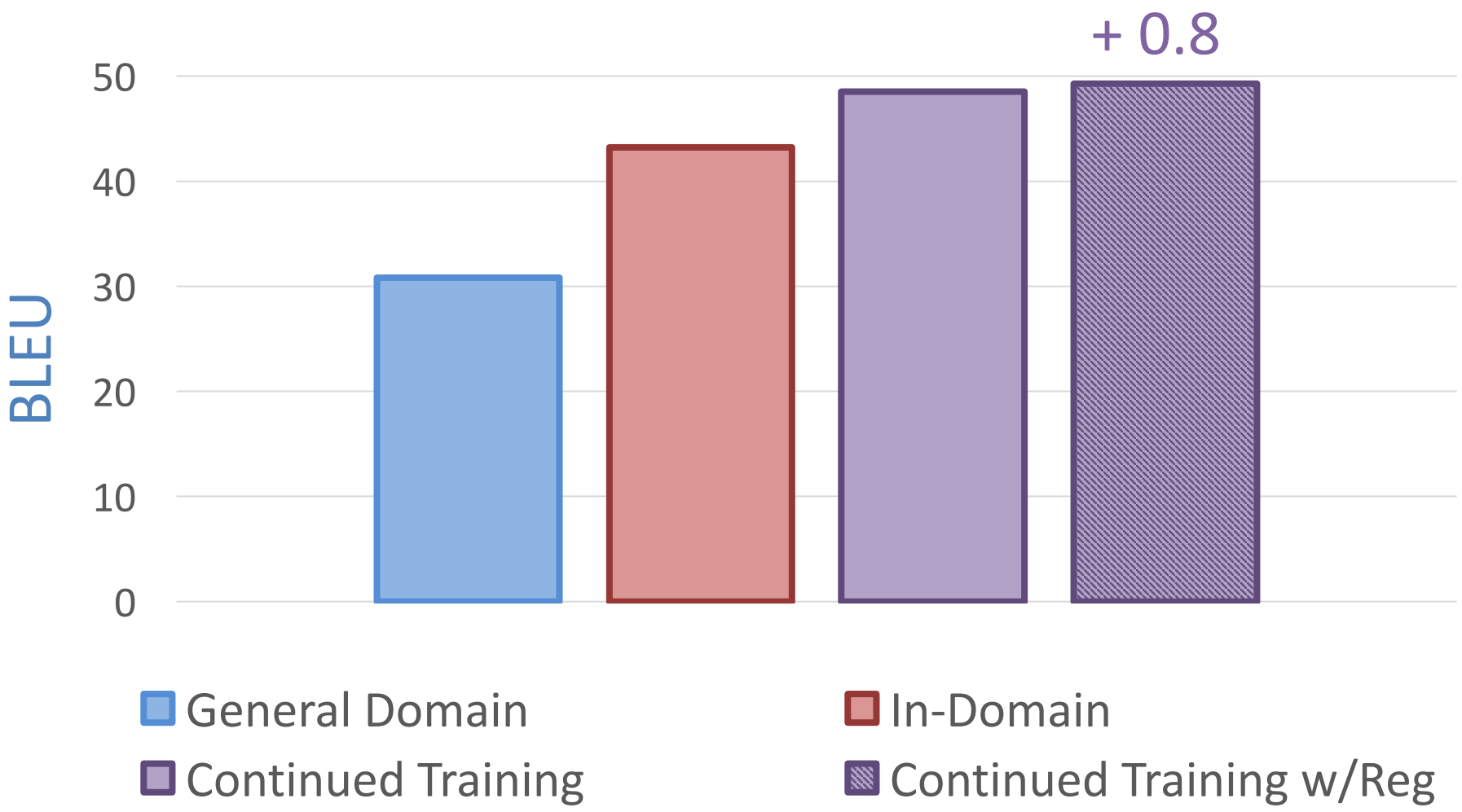
$\alpha \times$ Cross Ent ( , )

Results

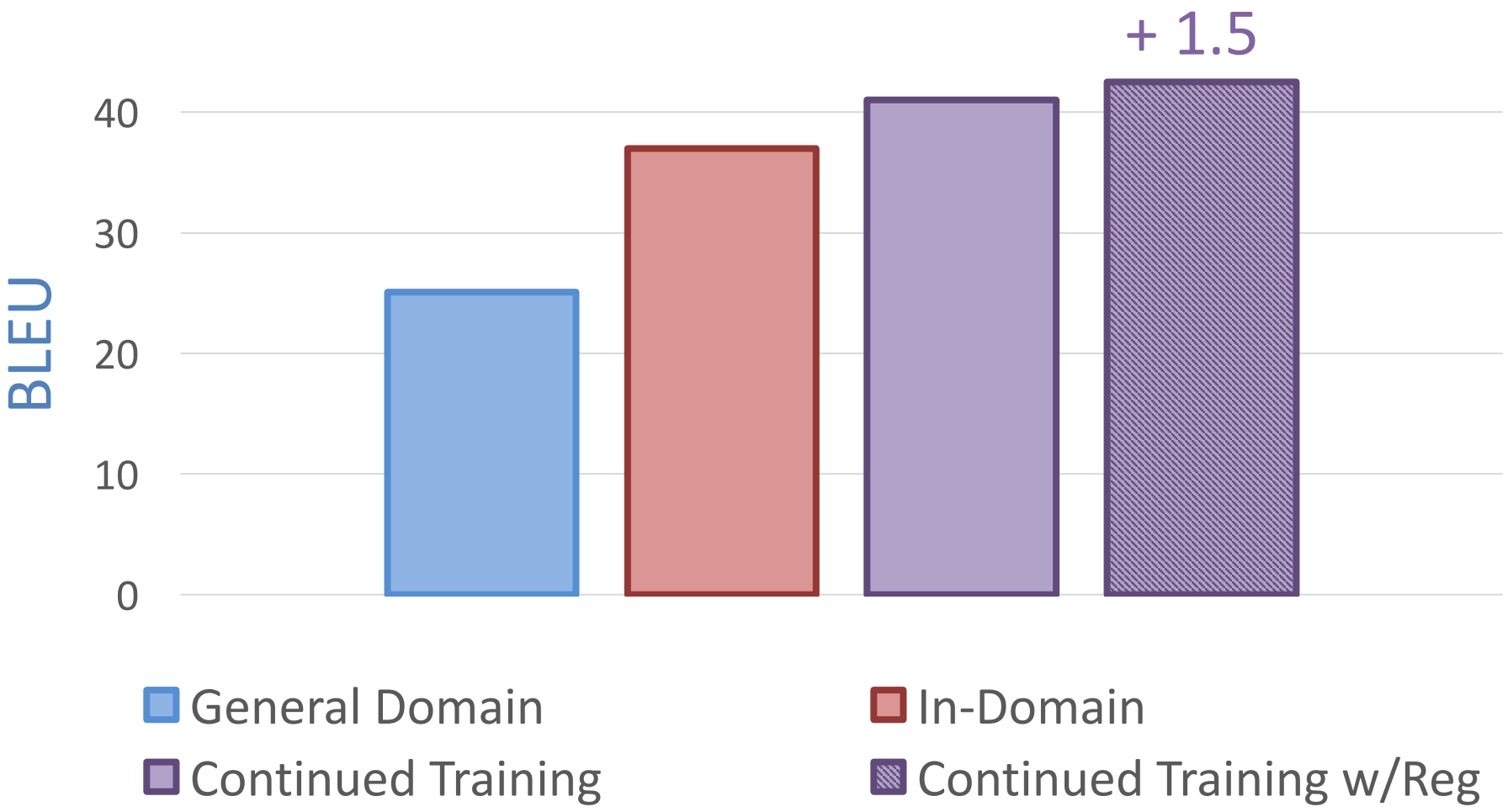
Russian → English Patents



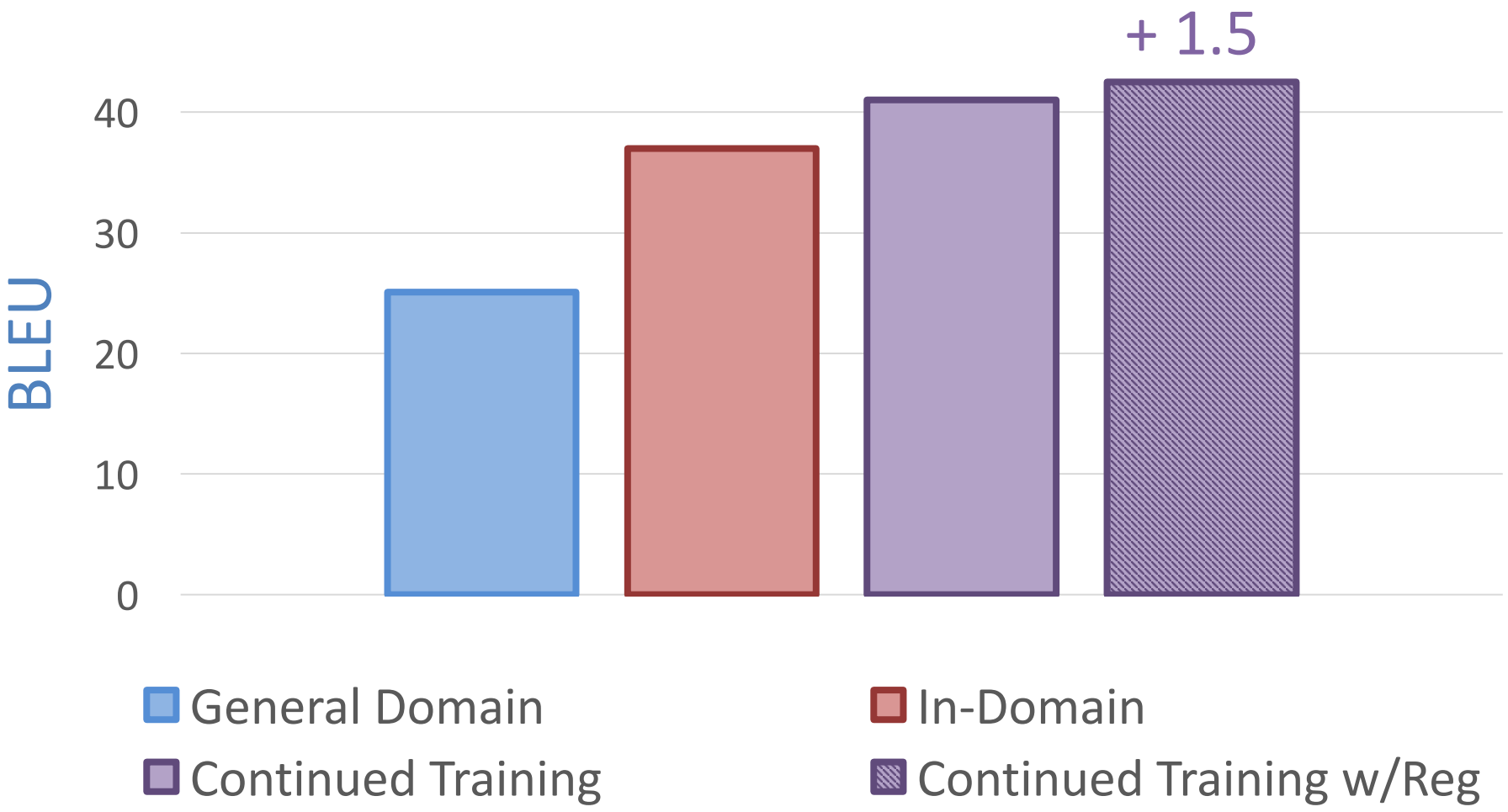
German → English Medical



English → German Medical

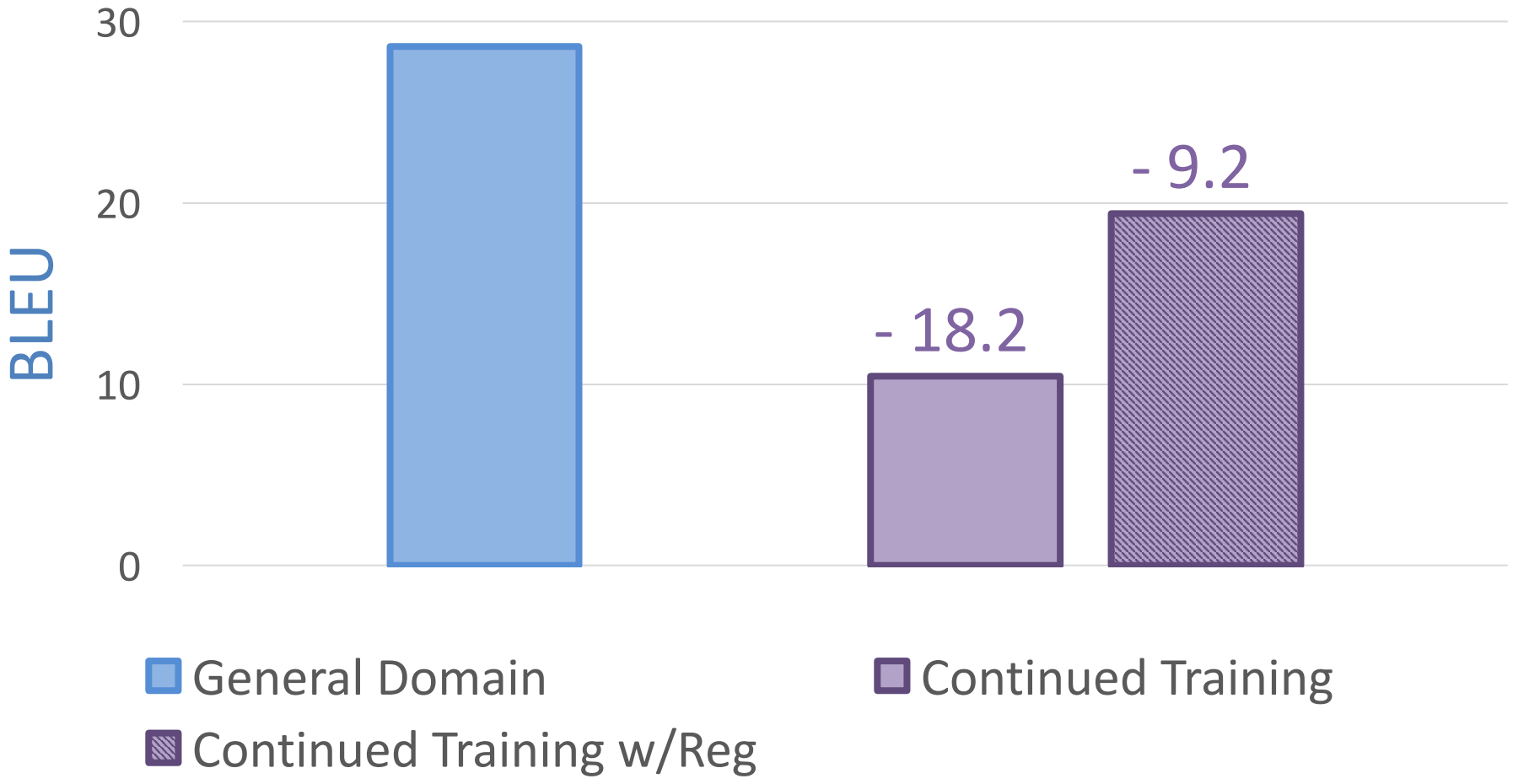


English → German Medical



Analysis

Russian → English General (patents)

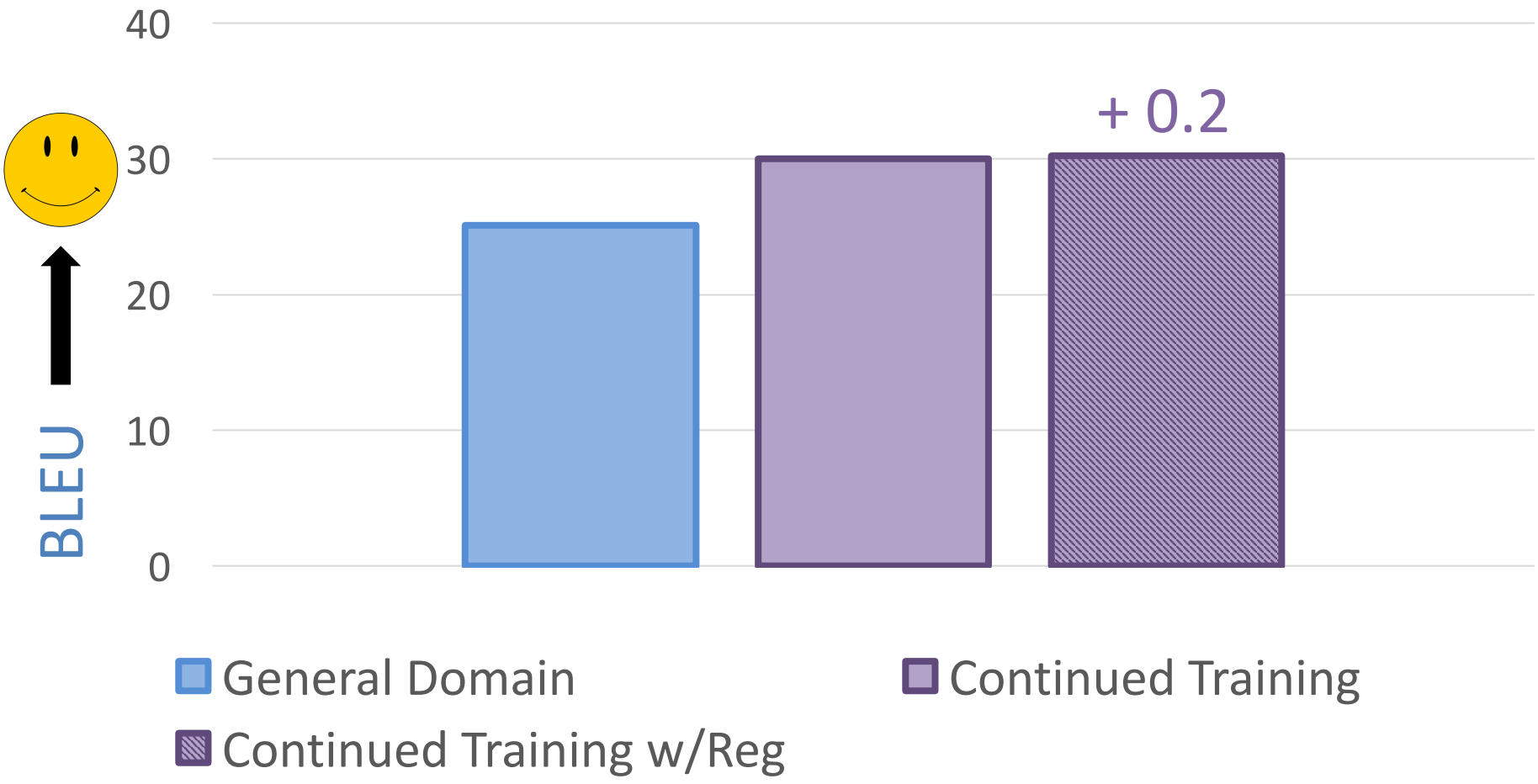


NAACL 2019

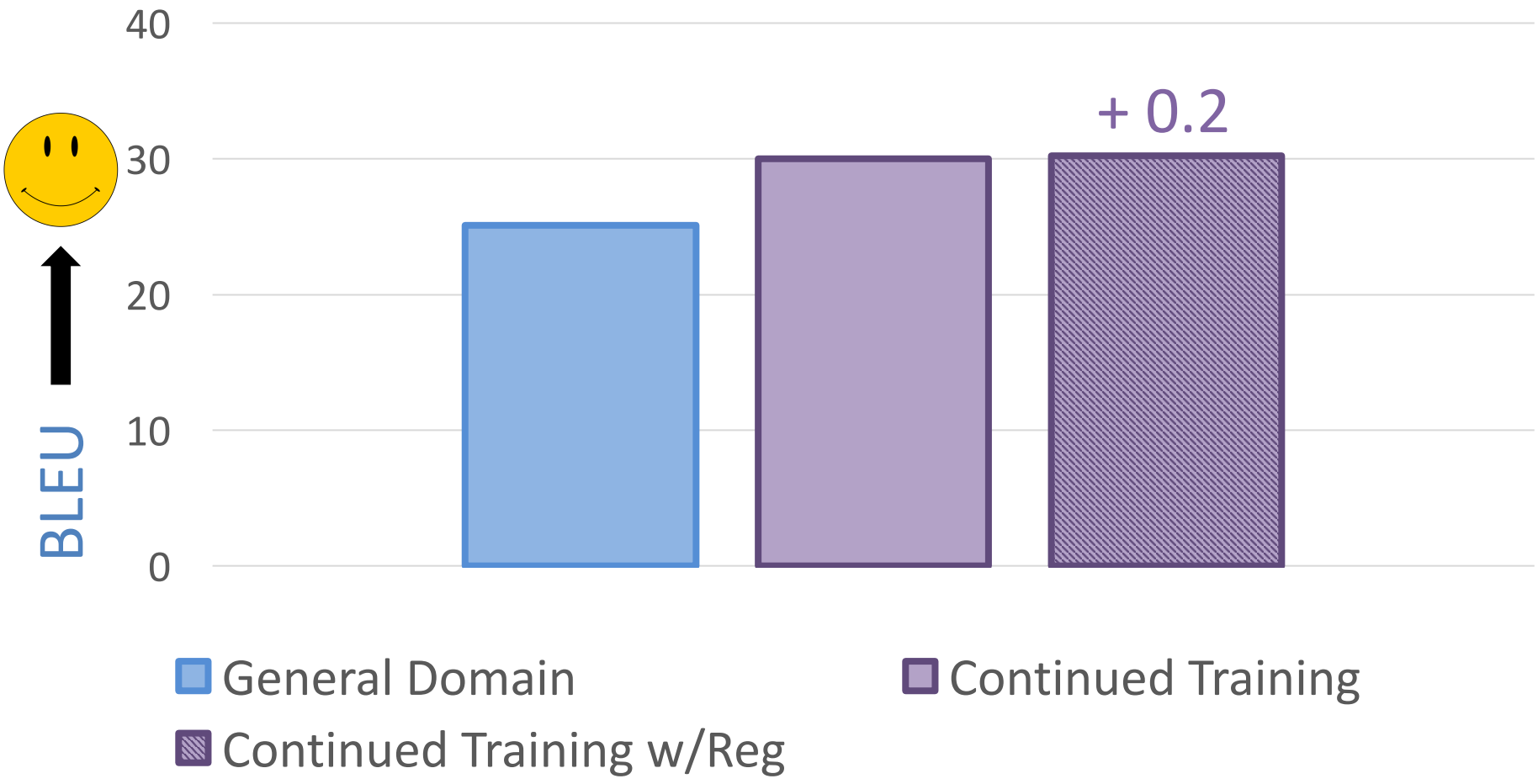
Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation

Brian Thompson[†] **Jeremy Gwinnup**[°] **Huda Khayrallah**[†] **Kevin Duh**[†] **Philipp Koehn**[†]
[†]Johns Hopkins University, [°]Air Force Research Laboratory
{brian.thompson, huda, phi}@jhu.edu,
kevinduh@cs.jhu.edu,
jeremy.gwinnup.1@us.af.mil

German-English Medical – Small



English-German Medical – Small



Overview

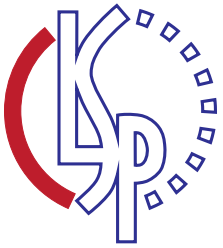
- Overview of Neural Machine Translation (NMT)
- Overview of Domain Adaptation
- Improving Domain Adaptation
  - Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
- **Analysis of Noisy Corpora**
  - On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]

On the Impact of Various Types of Noise on Neural Machine Translation

Huda Khayrallah & Philipp Koehn

WNMT at ACL 2018 [Outstanding Contribution Award]

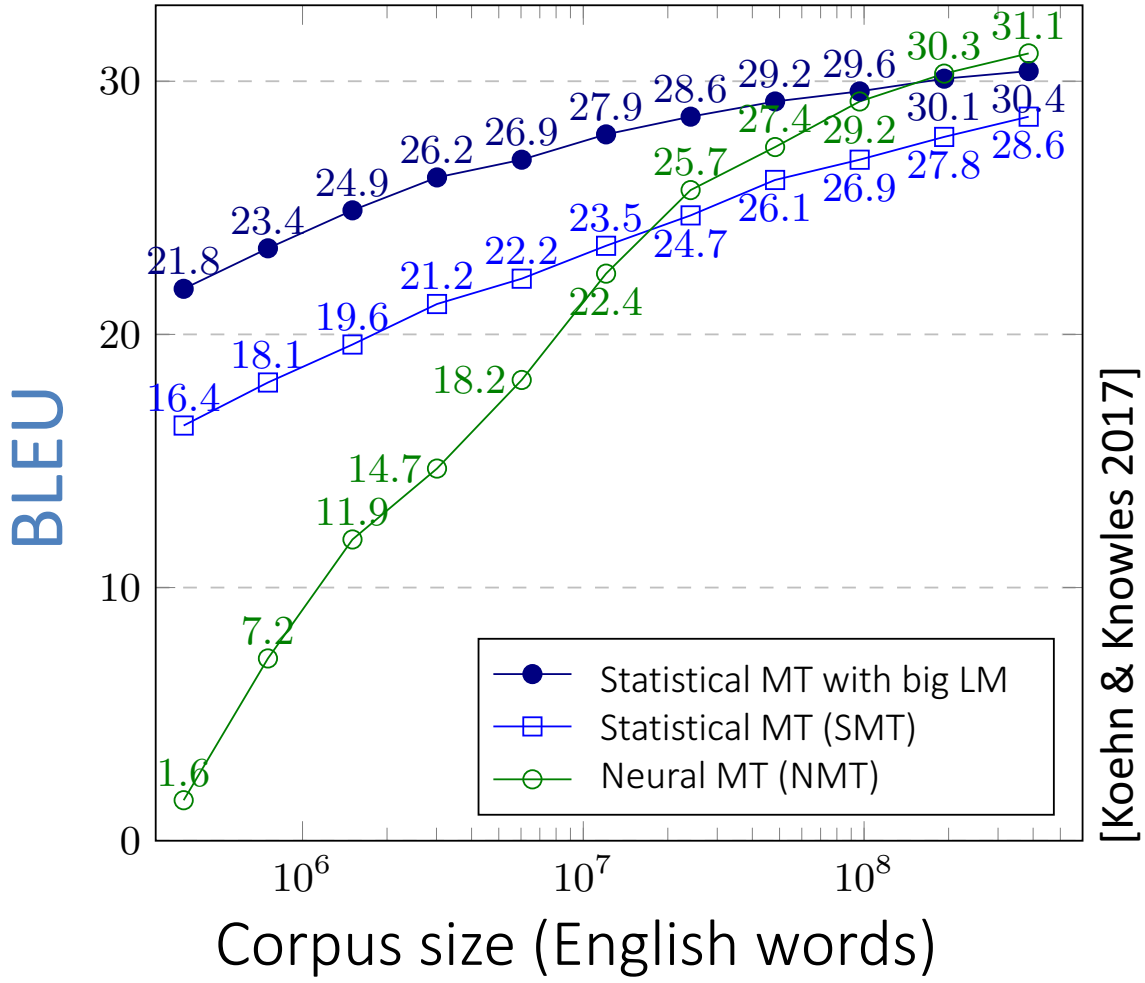
JOHNS HOPKINS
UNIVERSITY



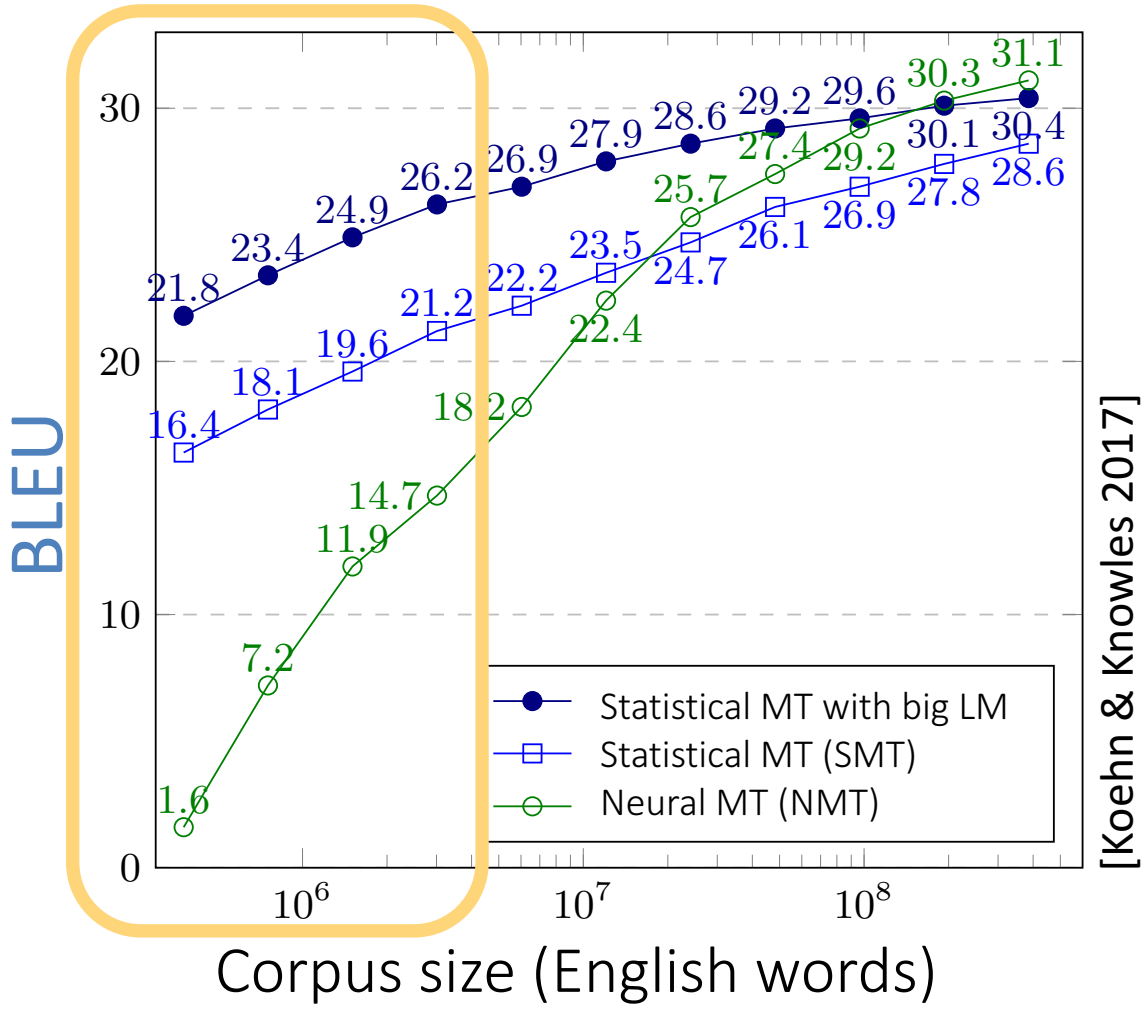
De→En translation

	NMT	SMT
WMT17	27.2	24.0

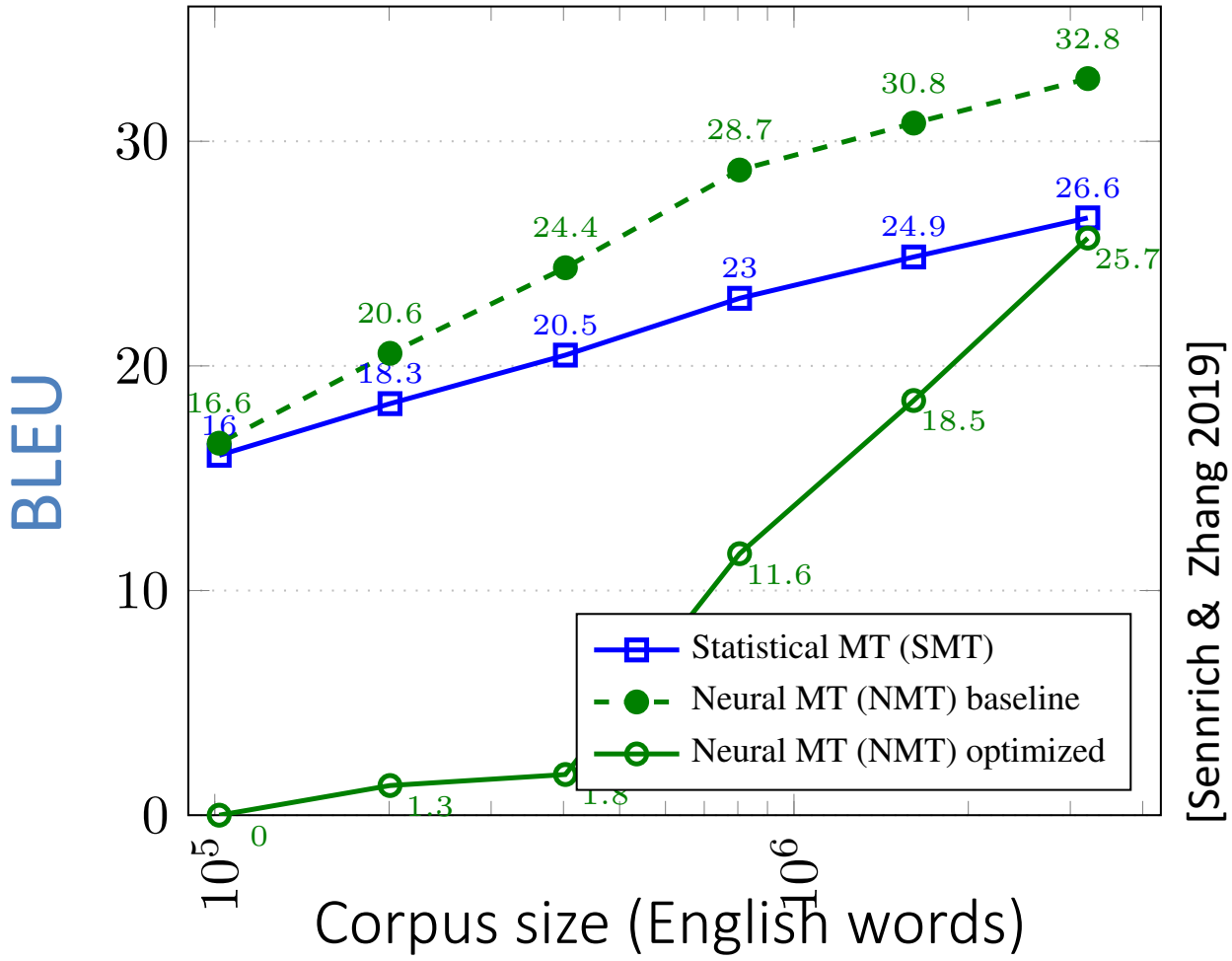
More data is better!



More data is better!



More data is better!



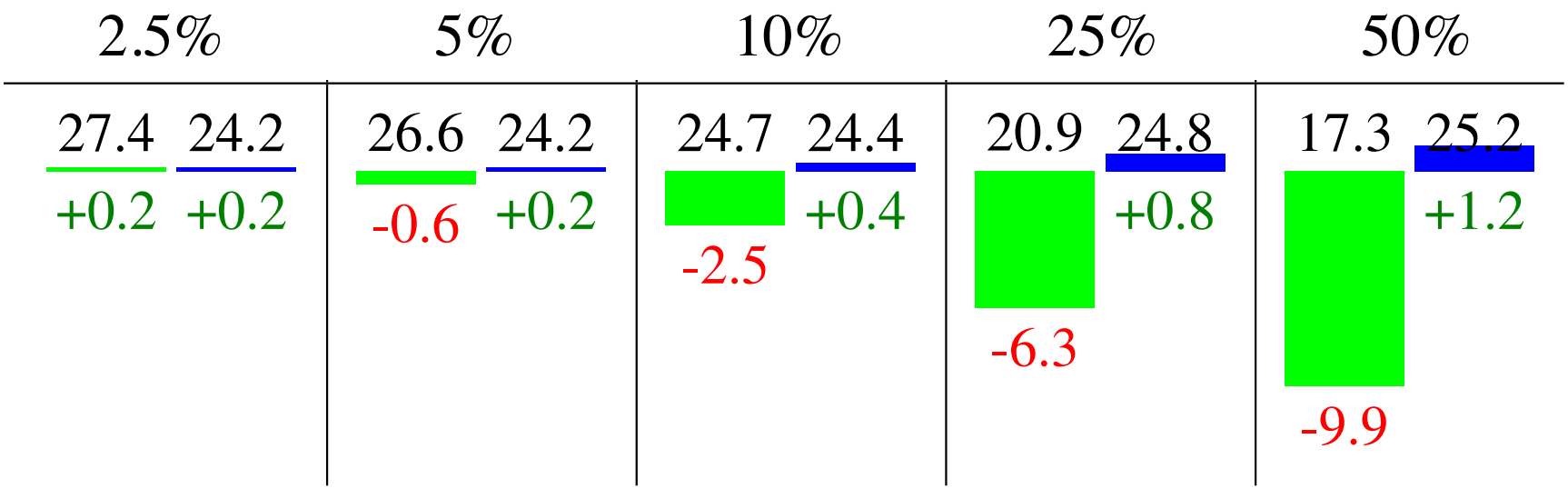
Let’s go get more data!

Annual growth in prices came in at 10.9 per cent, more than double the gain of the 12 months to August 2013, but the gains were not evenly spread across the country.

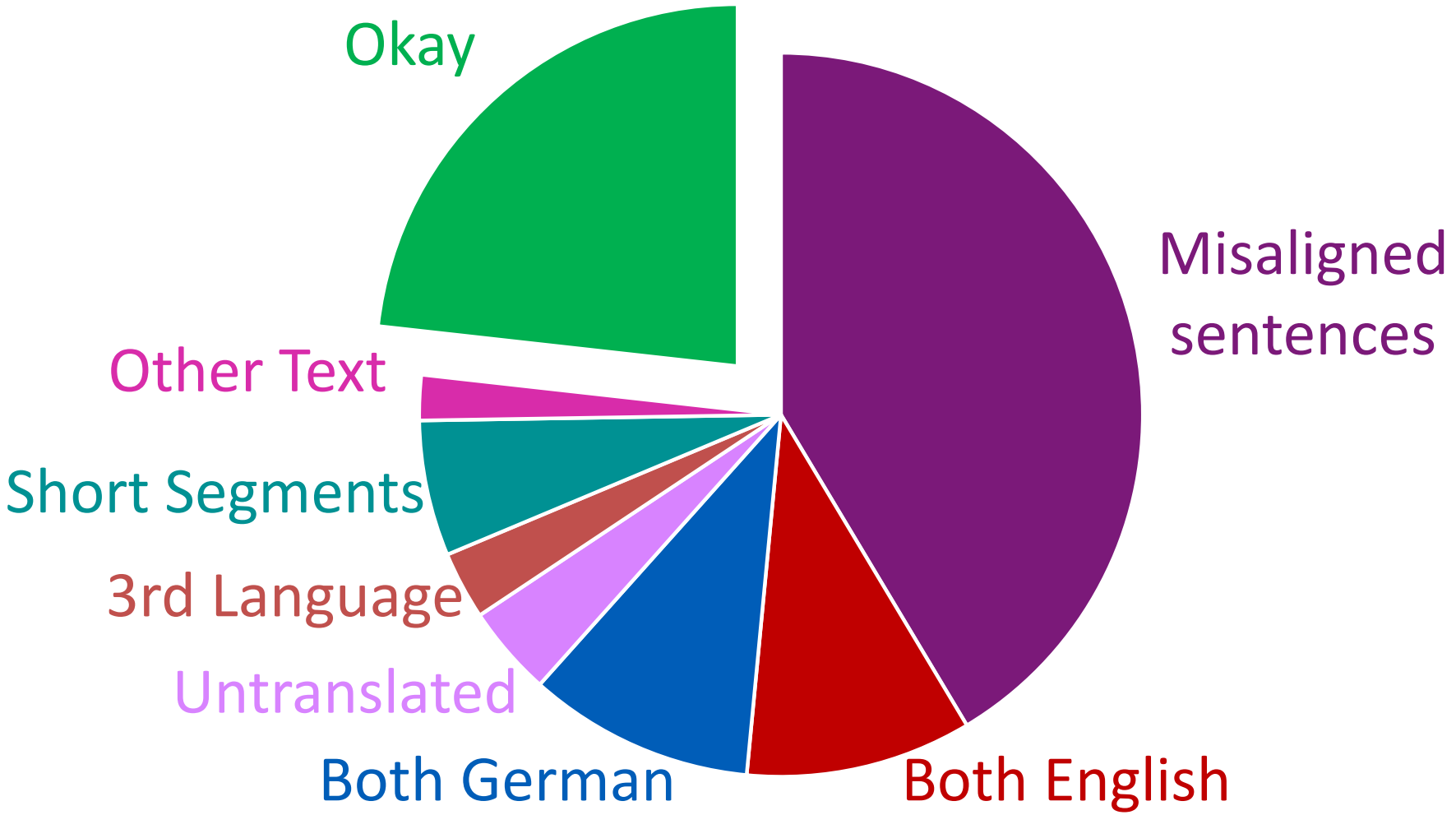
De→En translation

	NMT	SMT
WMT17	27.2	24.0
+ raw paracrawl	17.3 (-9.9)	25.2 (+1.2)

Raw Paracrawl



Manual Analysis



Noise Types

- Misaligned Sentences
- Misordered words
- Wrong Language
- Untranslated Sentences
- Short Segments

Misaligned Sentences

Misaligned Sentences

Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

Misaligned Sentences

Die Koalas sind süß

Die Kängurus springen

Der Koala ist weich

Das Känguru ist schnell

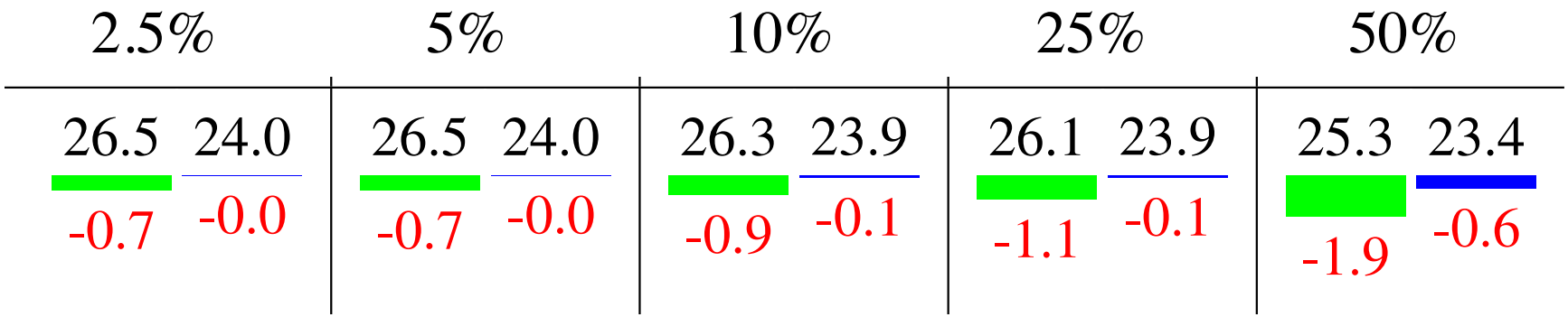
The kangaroos jump

The koala is soft

The kangaroo is fast

The koalas are cute

Misaligned Sentences



NMT SMT

Misordered Words

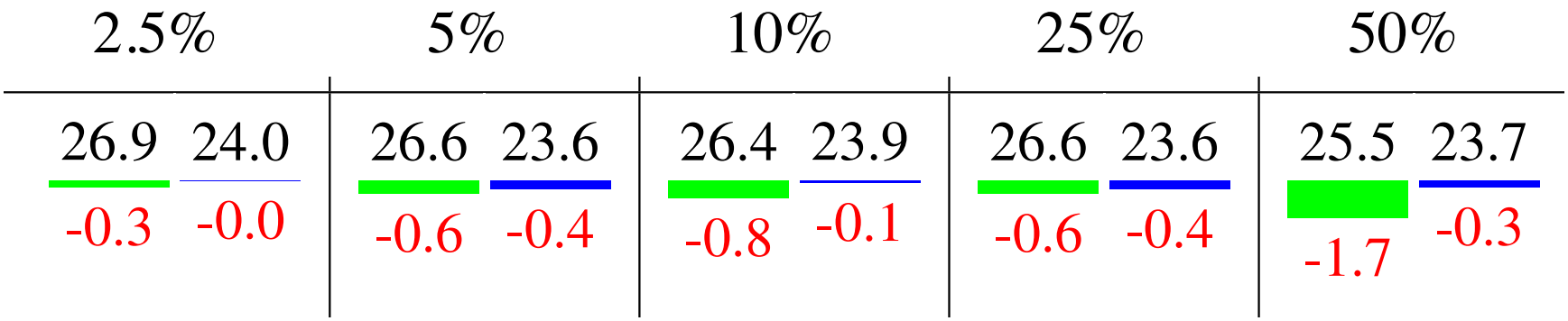
Misordered Words (source)

Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

Misordered Words (source)

Koalas Die sind süß	The koalas are cute
Kängurus springen Die	The kangaroos jump
ist Der weich Koala	The koala is soft
schnell Känguru ist Das	The kangaroo is fast

Misordered Words (source)



NMT SMT

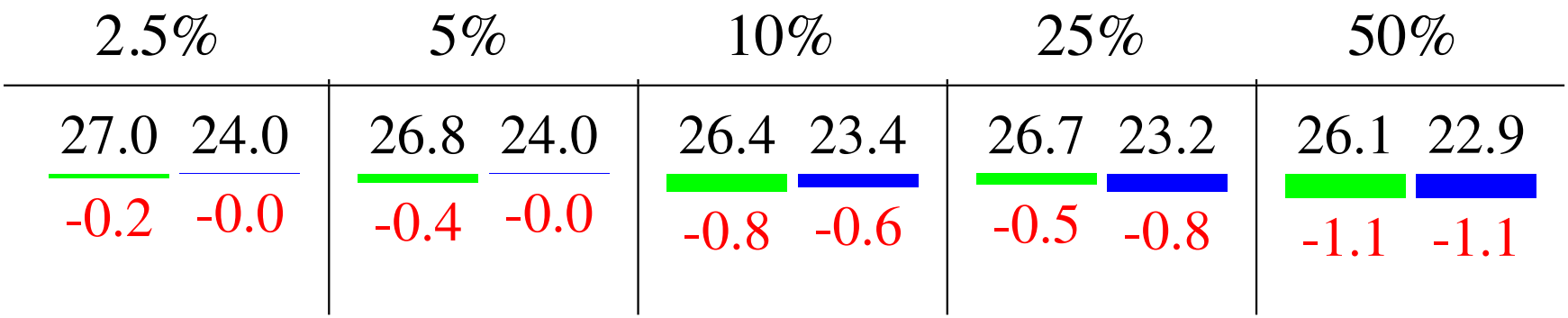
Misordered Words (target)

Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

Misordered Words (target)

Die Koalas sind süß	koalas cute are The
Die Kängurus springen	kangaroos The jump
Der Koala ist weich	is The soft koala
Das Känguru ist schnell	fast The is kangaroo

Misordered Words (target)



NMT SMT

Wrong Language

Wrong Language (French source)

Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

Wrong Language (French source)

Les koalas sont mignons	The koalas are cute
Les kangourous sautent	The kangaroos jump
Le koala est doux	The koala is soft
Le kangourou est rapide	The kangaroo is fast

Wrong Language (French source)

2.5%		5%		10%		25%		50%	
26.9	24.0	26.8	23.9	26.8	23.9	26.8	23.9	26.8	23.8
-0.3	-0.0	-0.4	-0.1	-0.4	-0.1	-0.4	-0.1	-0.4	-0.2

NMT SMT

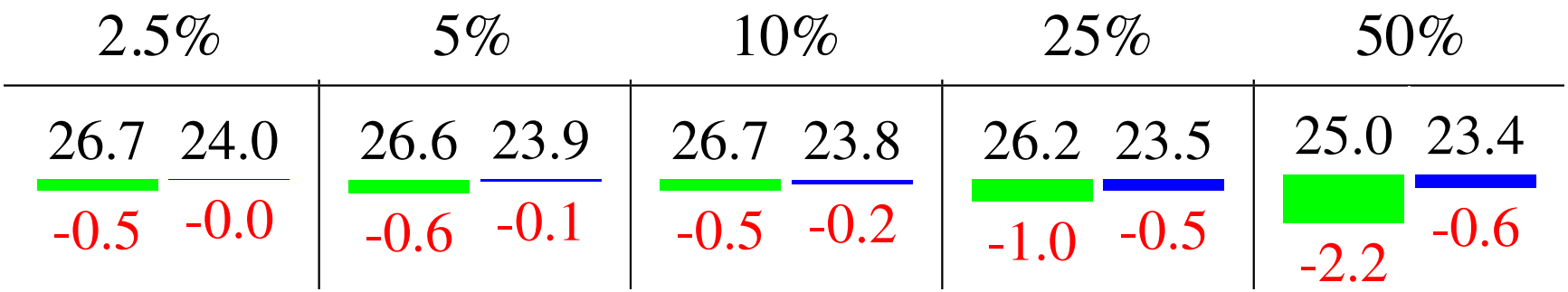
Wrong Language (French target)

Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

Wrong Language (French target)

Die Koalas sind süß	Les koalas sont mignons
Die Kängurus springen	Les kangourous sautent
Der Koala ist weich	Le koala est doux
Das Känguru ist schnell	Le kangourou est rapide

Wrong Language (French target)



NMT SMT

Untranslated

Untranslated (English Source)

Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

Untranslated (English source)

The koalas are cute

The kangaroos jump

The koala is soft

The kangaroo is fast

The koalas are cute

The kangaroos jump

The koala is soft

The kangaroo is fast

Untranslated (English source)

2.5%		5%		10%		25%		50%	
27.2	23.9	27.0	23.9	26.7	23.6	26.8	23.7	26.9	23.5
-0.0	-0.1	-0.2	-0.1	-0.5	-0.4	-0.4	-0.3	-0.3	-0.5

NMT SMT

Untranslated (German target)

Die Koalas sind süß	The koalas are cute
Die Kängurus springen	The kangaroos jump
Der Koala ist weich	The koala is soft
Das Känguru ist schnell	The kangaroo is fast

Untranslated (German target)

Die Koalas sind süß

Die Kängurus springen

Der Koala ist weich

Das Känguru ist schnell

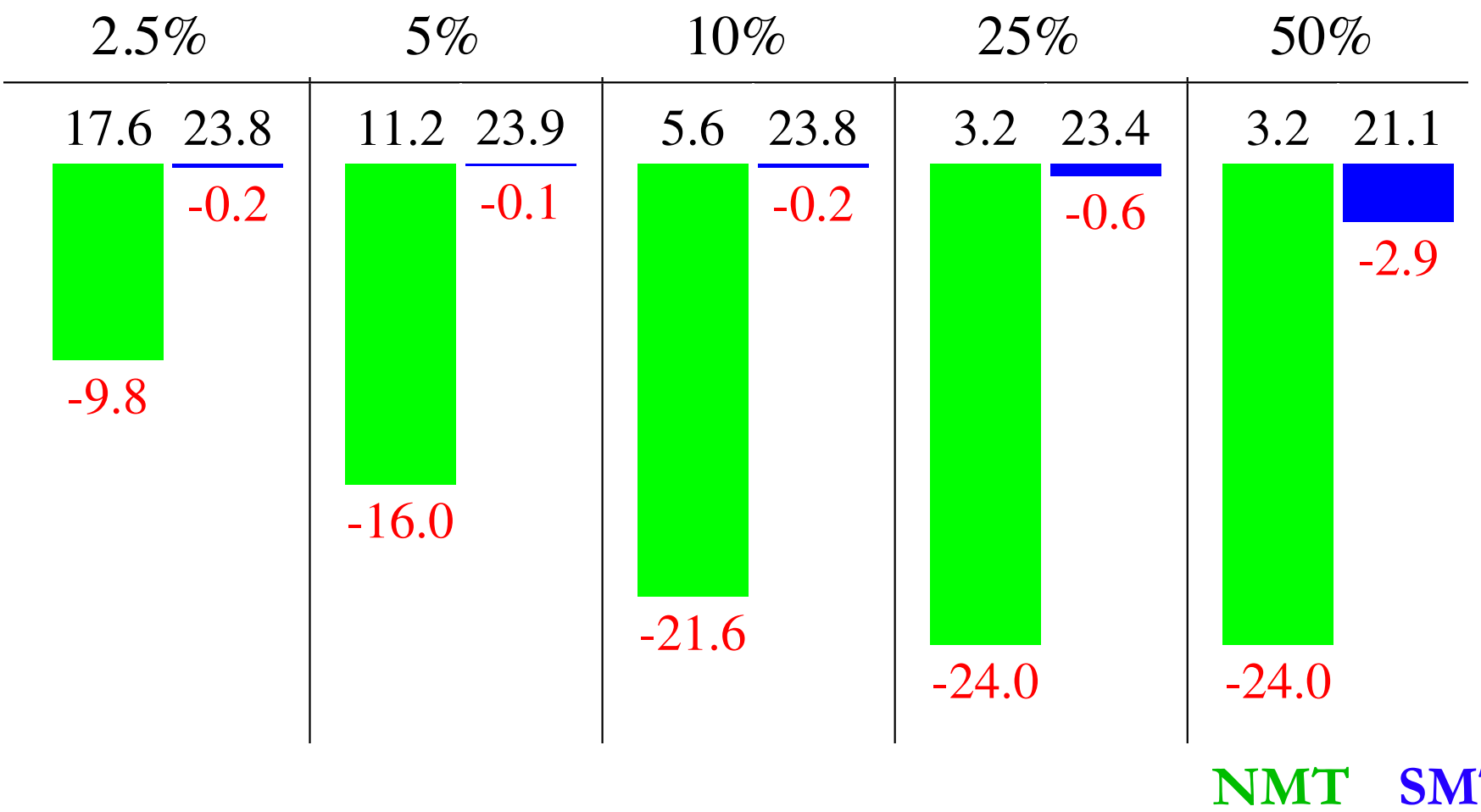
Die Koalas sind süß

Die Kängurus springen

Der Koala ist weich

Das Känguru ist schnell

Untranslated (German target)



Short Segments

Short Segments

Die	The
süß	cute
Känguru	Kangaroo
schnell	fast

Short Segments

≤ 2 words

2.5%		5%		10%	
27.1	24.1	26.5	23.9	26.7	23.8
-0.1	+0.1	-0.7	-0.1	-0.5	-0.2

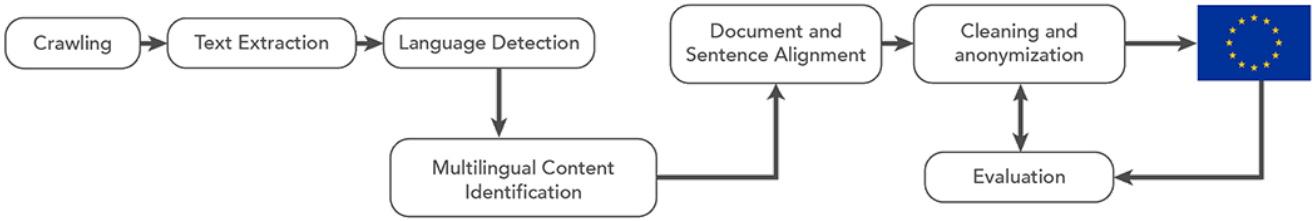
3-5 words

2.5%		5%		10%		25%	
27.8	24.2	27.6	24.5	28.0	24.5	26.6	24.2
+0.6	+0.2	+0.4	+0.5	+0.8	+0.5	-0.6	+0.2

	5%		10%		20%		50%		100%	
MISALIGNED SENTENCES	26.5	24.0	26.5	24.0	26.3	23.9	26.1	23.9	25.3	23.4
	-0.7	-0.0	-0.7	-0.0	-0.9	-0.1	-1.1	-0.1	-1.9	-0.6
MISORDERED WORDS (SOURCE)	26.9	24.0	26.6	23.6	26.4	23.9	26.6	23.6	25.5	23.7
	-0.3	-0.0	-0.6	-0.4	-0.8	-0.1	-0.6	-0.4	-1.7	-0.3
MISORDERED WORDS (TARGET)	27.0	24.0	26.8	24.0	26.4	23.4	26.7	23.2	26.1	22.9
	-0.2	-0.0	-0.4	-0.0	-0.8	-0.6	-0.5	-0.8	-1.1	-1.1
WRONG LANGUAGE (FRENCH SOURCE)	26.9	24.0	26.8	23.9	26.8	23.9	26.8	23.9	26.8	23.8
	-0.3	-0.0	-0.4	-0.1	-0.4	-0.1	-0.4	-0.1	-0.4	-0.2
WRONG LANGUAGE (FRENCH TARGET)	26.7	24.0	26.6	23.9	26.7	23.8	26.2	23.5	25.0	23.4
	-0.5	-0.0	-0.6	-0.1	-0.5	-0.2	-1.0	-0.5	-2.2	-0.6
UNTRANSLATED (ENGLISH SOURCE)	27.2	23.9	27.0	23.9	26.7	23.6	26.8	23.7	26.9	23.5
	-0.0	-0.1	-0.2	-0.1	-0.5	-0.4	-0.4	-0.3	-0.3	-0.5
UNTRANSLATED (GERMAN TARGET)	17.6	23.8	11.2	23.9	5.6	23.8	3.2	23.4	3.2	21.1
	-9.8	-0.2	-16.0	-0.1	-21.6	-0.2	-24.0	-0.6	-24.0	-2.9
SHORT SEGMENTS (max 2)	27.1	24.1	26.5	23.9	26.7	23.8				
	-0.1	+0.1	-0.7	-0.1	-0.5	-0.2				
SHORT SEGMENTS (max 5)	27.8	24.2	27.6	24.5	28.0	24.5	26.6	24.2		
	+0.6	+0.2	+0.4	+0.5	+0.8	+0.5	-0.6	+0.2		
RAW CRAWL DATA	27.4	24.2	26.6	24.2	24.7	24.4	20.9	24.8	17.3	25.2
	+0.2	+0.2	-0.6	+0.2	-2.5	+0.4	-6.3	+0.8	-9.9	+1.2

Filtering methods

ParaCrawl



- BiCleaner [Espla-Gomis & Forcada 2009]
- Zipporah [Xu & Koehn 2017]
- WMT shared task [Koehn, Khayrallah, Heafield & Forcada 2018]
  - Dual Conditional Cross-Entropy Filtering [Junczys-Dowmunt 2018]
  - Zipporah [Khayrallah, Xu & Koehn 2018]

De→En translation

	NMT	SMT
WMT17	27.2	24.0
+ raw paracrawl	17.3 (-9.9)	25.2 (+1.2)

De→En translation

	NMT	SMT
WMT17	27.2	24.0
+ raw paracrawl	17.3 (-9.9)	25.2 (+1.2)
WMT19 + filtered paracrawl	32.4 (+5.2)	25.8 (+1.8)

Overview

- Overview of Neural Machine Translation (NMT)
- Overview of Domain Adaptation
- Improving Domain Adaptation
  - Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation [Khayrallah, Thompson, Duh & Koehn 2018]
- Analysis of Noisy Corpora
  - On the Impact of Various Types of Noise on Neural Machine Translation [Khayrallah & Koehn 2018]

Takeaways

- new methods→ improved performance
- models consist of methods + **data**
- methods pick up patterns in the data
  - noise [Khayrallah & Koehn 2018]
  - annotation artifacts [Poliak et al., 2018]
  - Bias [Ethics in NLP 2017, 2018; Gender Bias in NLP 2019]

Questions?

huda@jhu.edu

cs.jhu.edu/~huda