Continued Training Algorithms

This talk was presented at the final presentation for 2018 JHU SCALE workshop

on Domain adaptation in Machine Translation

These slides were presented by Huda Khayrallah;

the full presentation can be found here:

cs.jhu.edu/~kevinduh/t/scale2018/slides/

Continued Training Algorithms

Huda Khayrallah

SCALE Readout

August 9, 2018

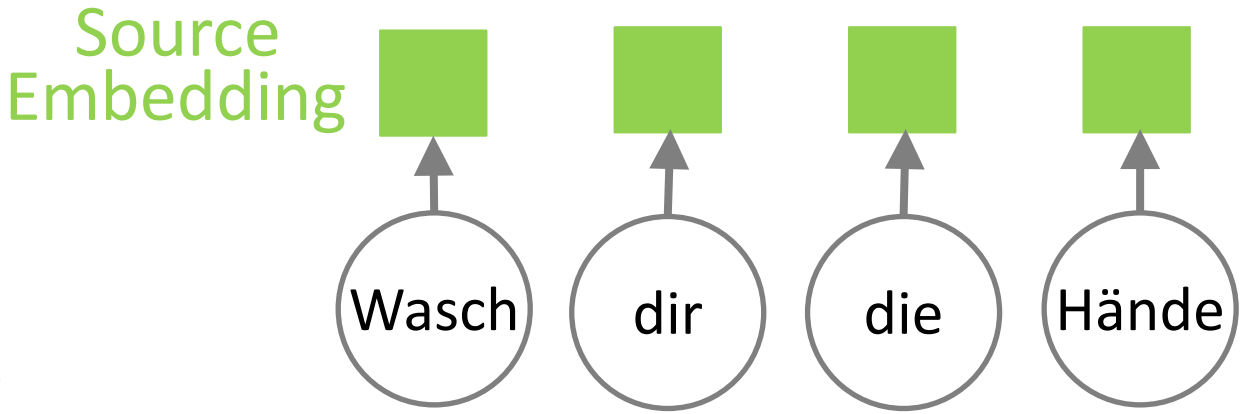
JOHNS HOPKINS
UNIVERSITY

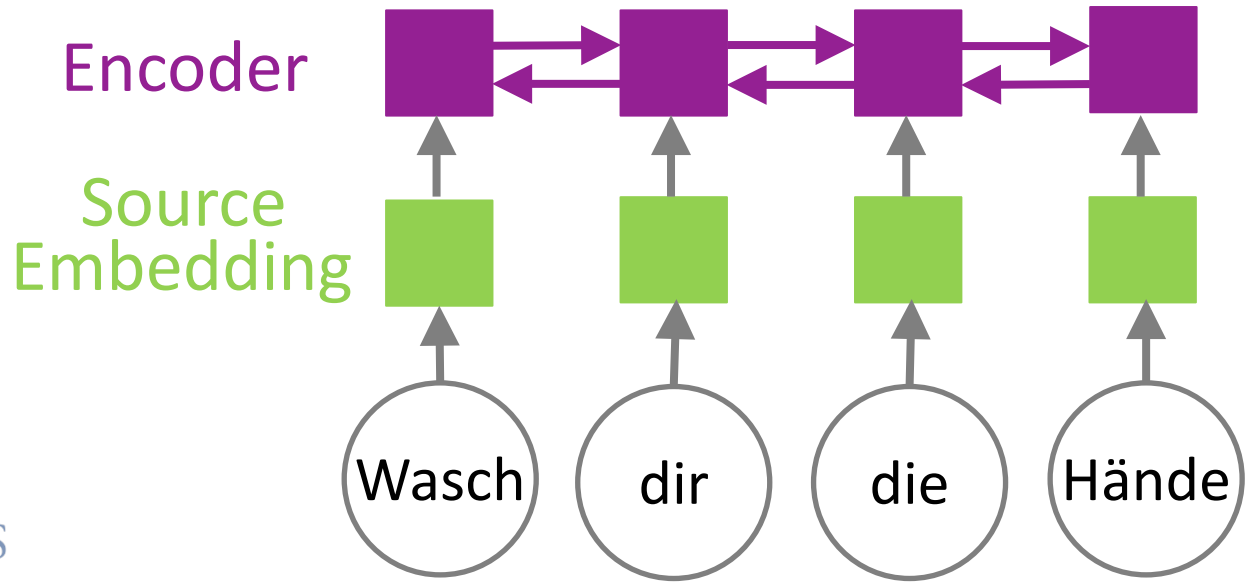
Wasch

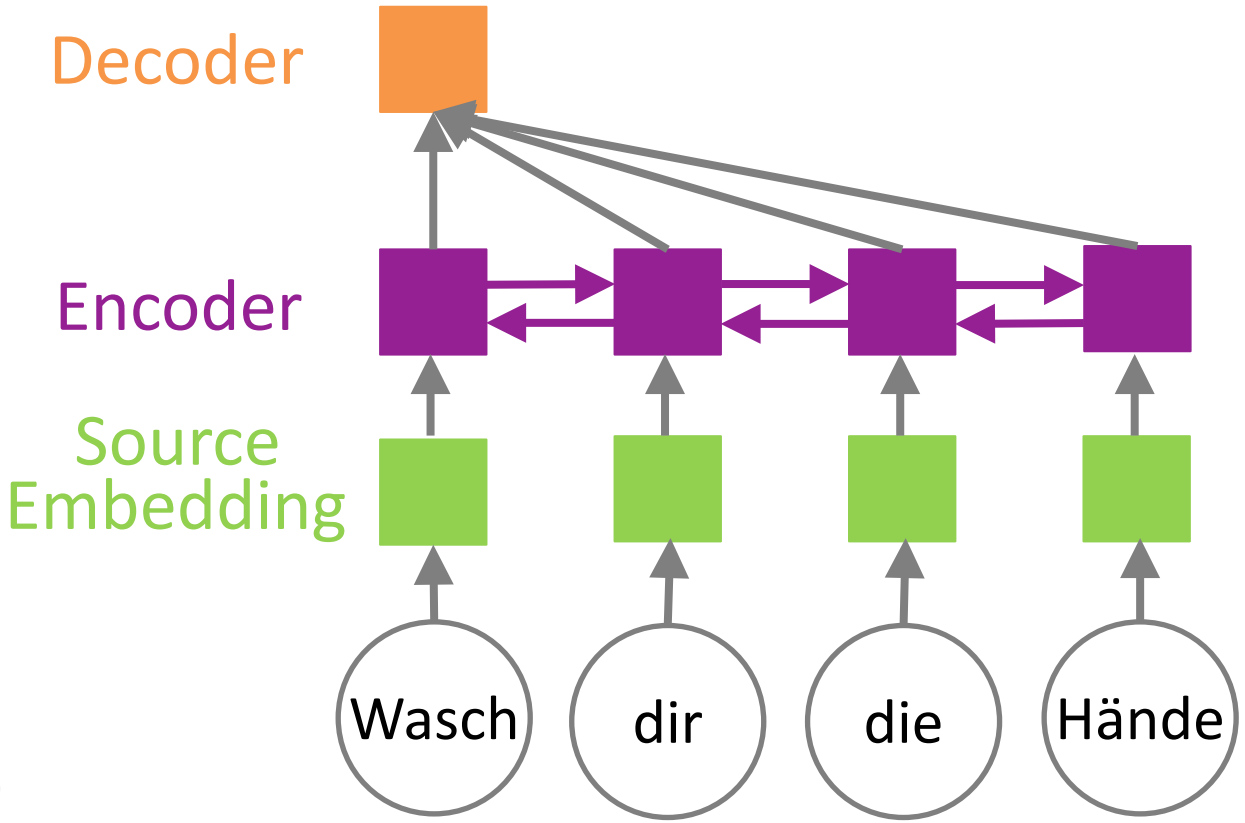
dir

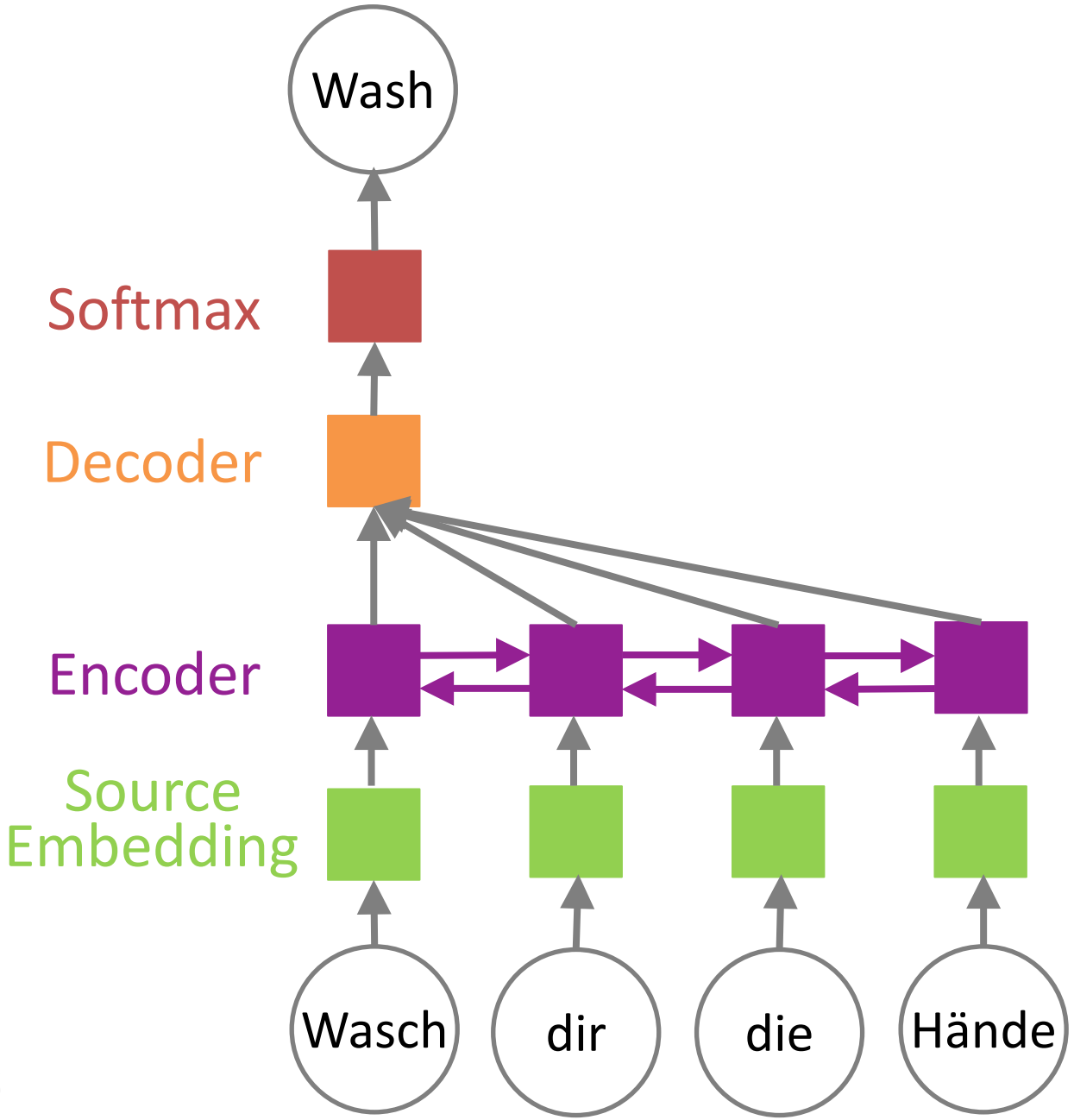
die

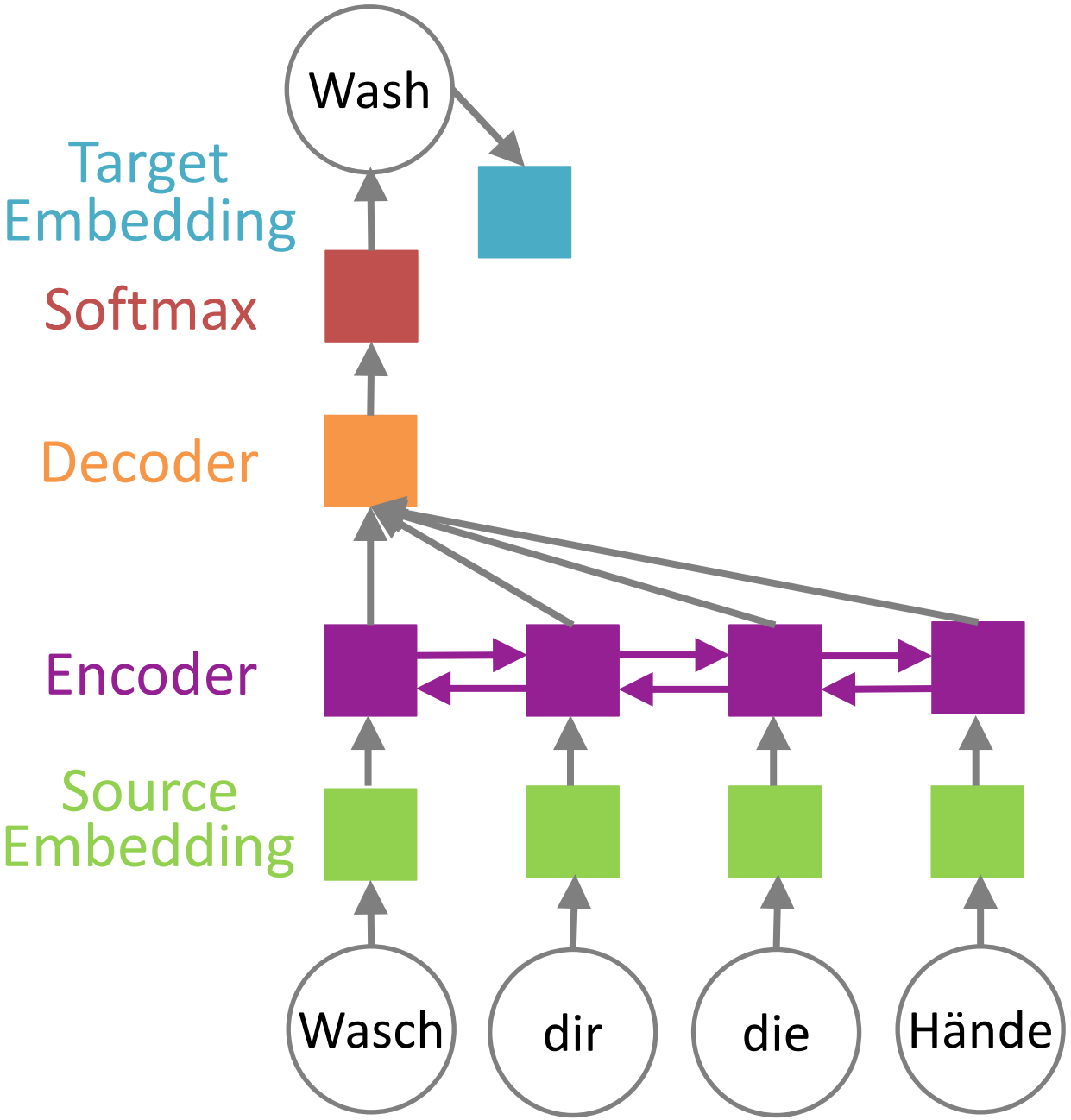
Hände

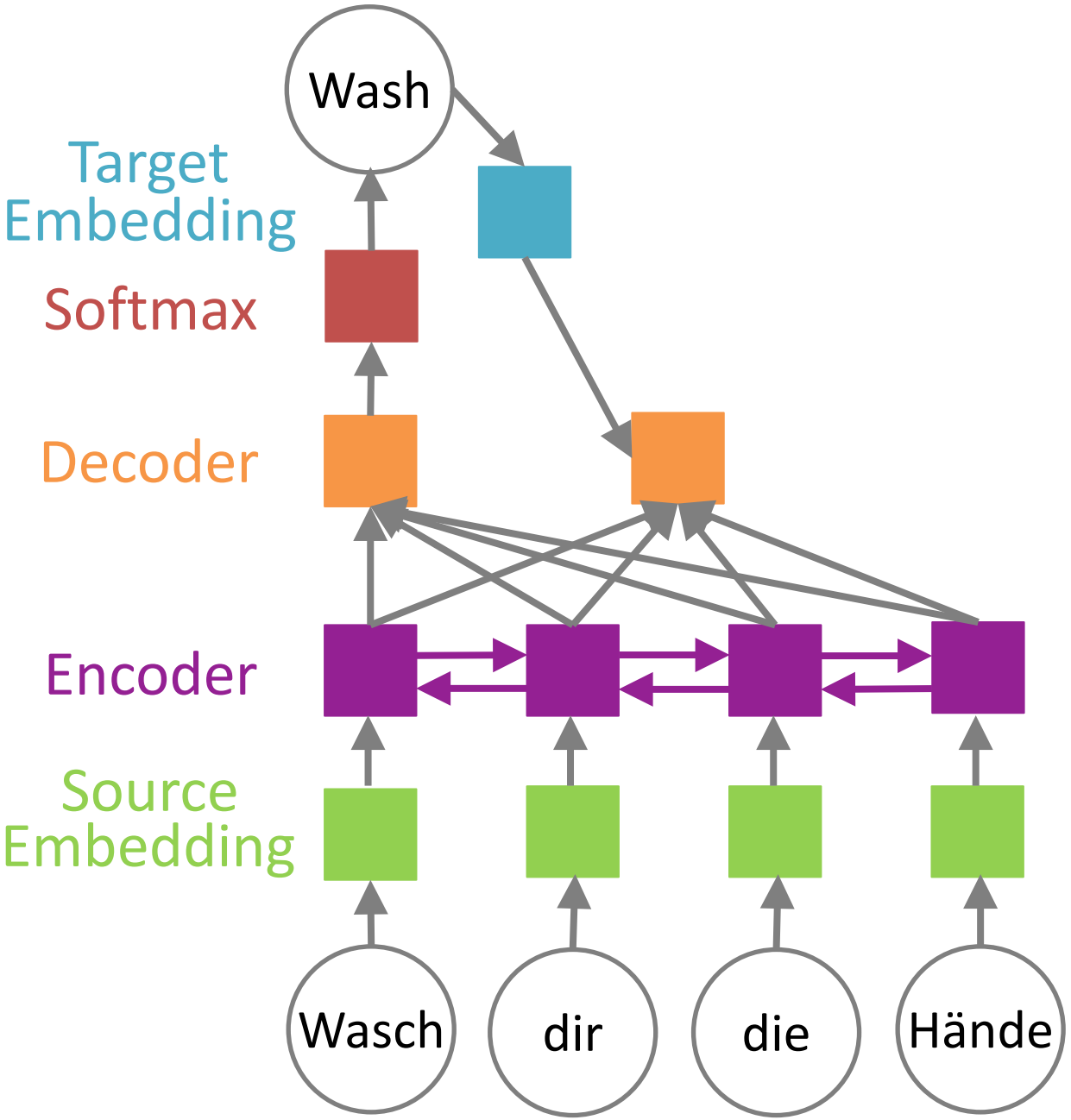

Source
Embedding
Wasch
dir
die
Hände

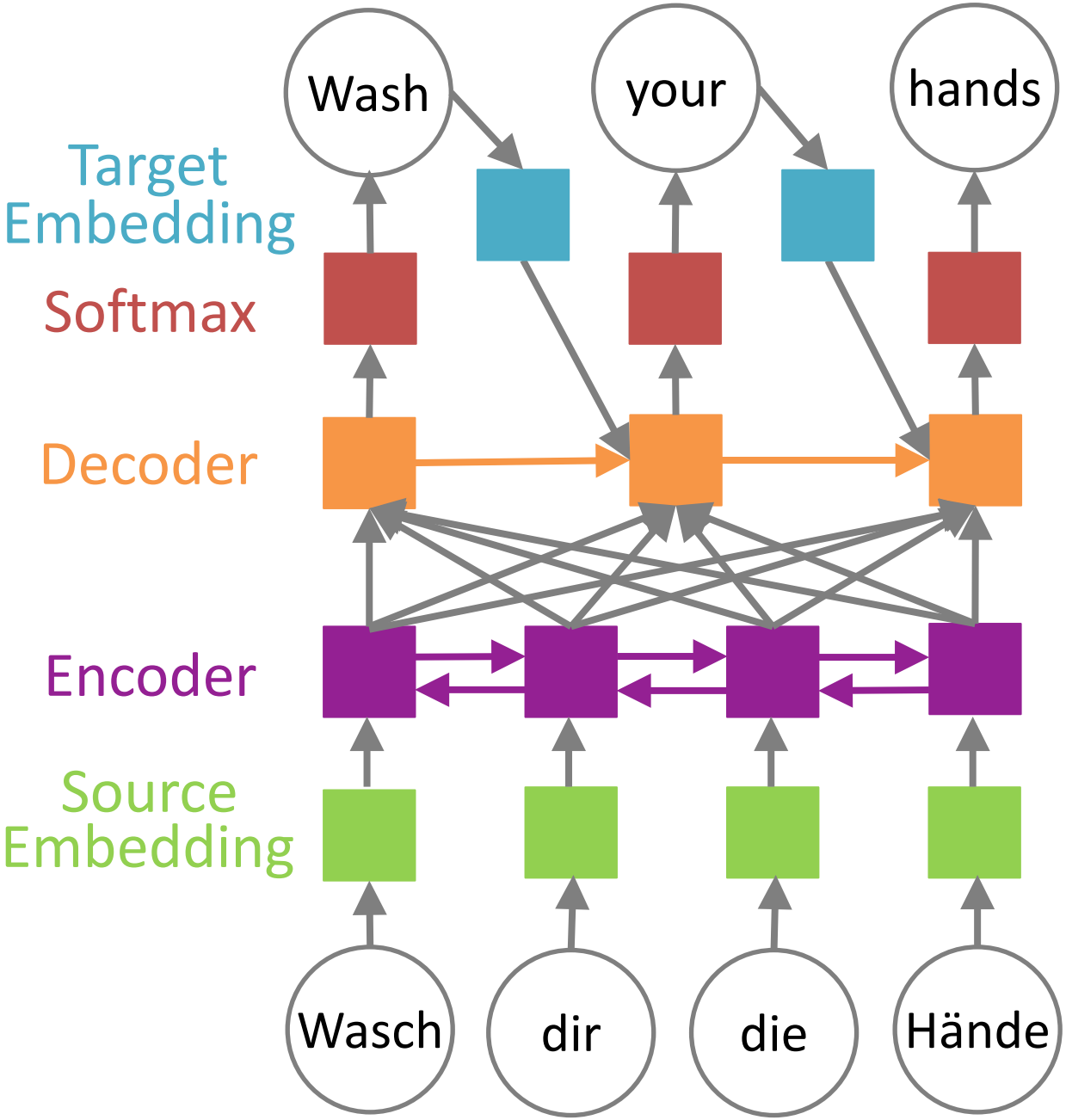

Encoder
Source Embedding
Wasch
dir
die
Hände


Decoder
Encoder
Source
Embedding
Wasch
dir
die
Hände


Wash
Softmax
Decoder
Encoder
Source
Embedding
Wasch
dir
die
Hände


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

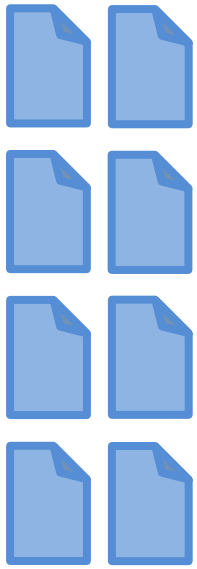

Wash
Target Embedding
Softmax
Decoder
Encoder
Source Embedding
Wasch
dir
die
Hände

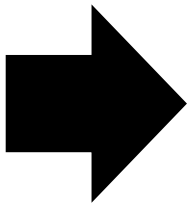


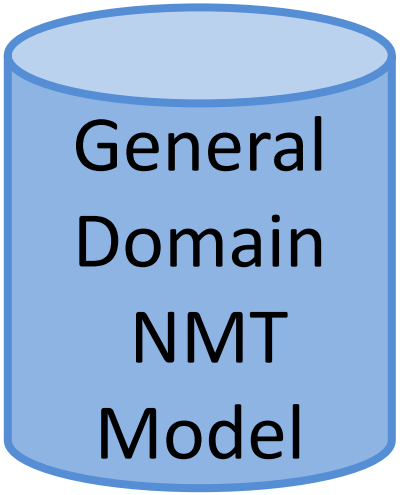
[Bahdanau et al. 2015]

MT Training

General Domain NMT

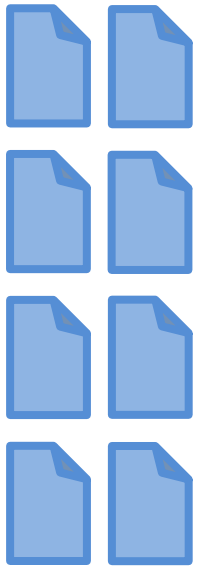


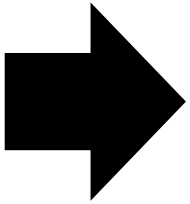


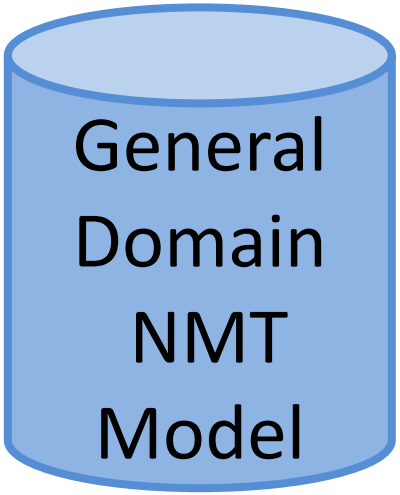


50m General Domain sentence pairs

General Domain NMT





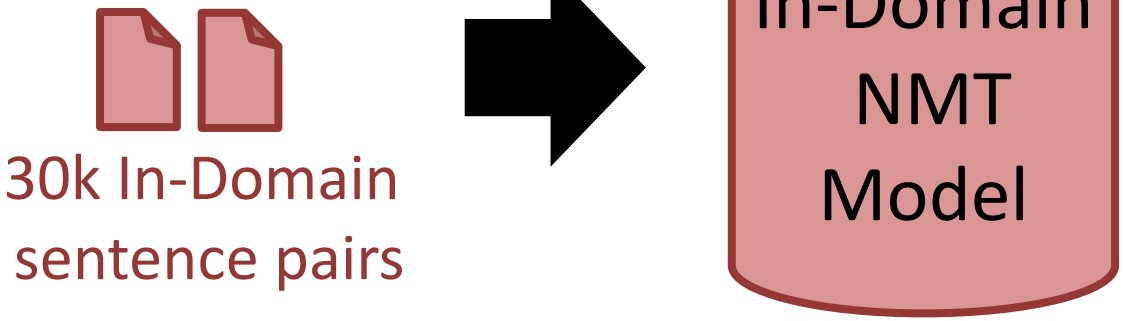


дверной замок повышенной степени защищенности от взлома

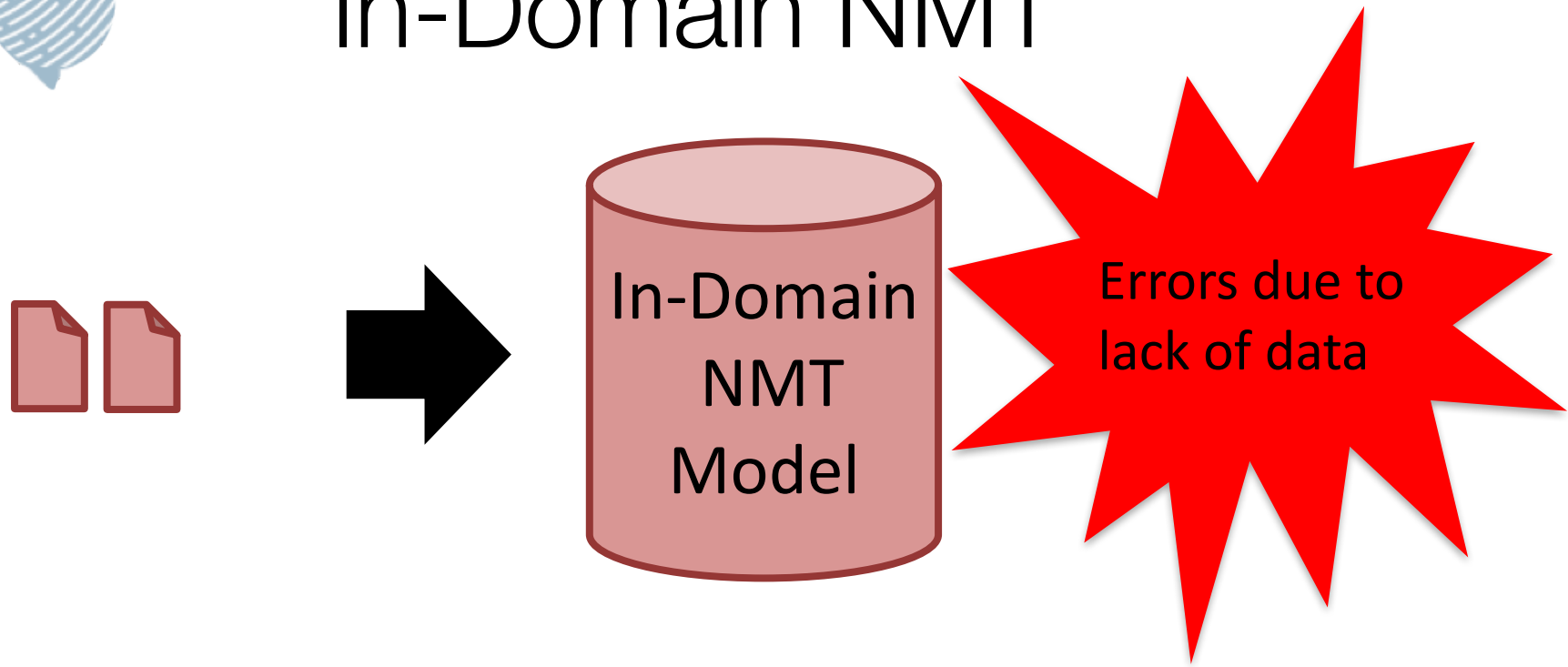
Human: door lock with increased degree of security against burglary

System: door security door security door

In-Domain NMT



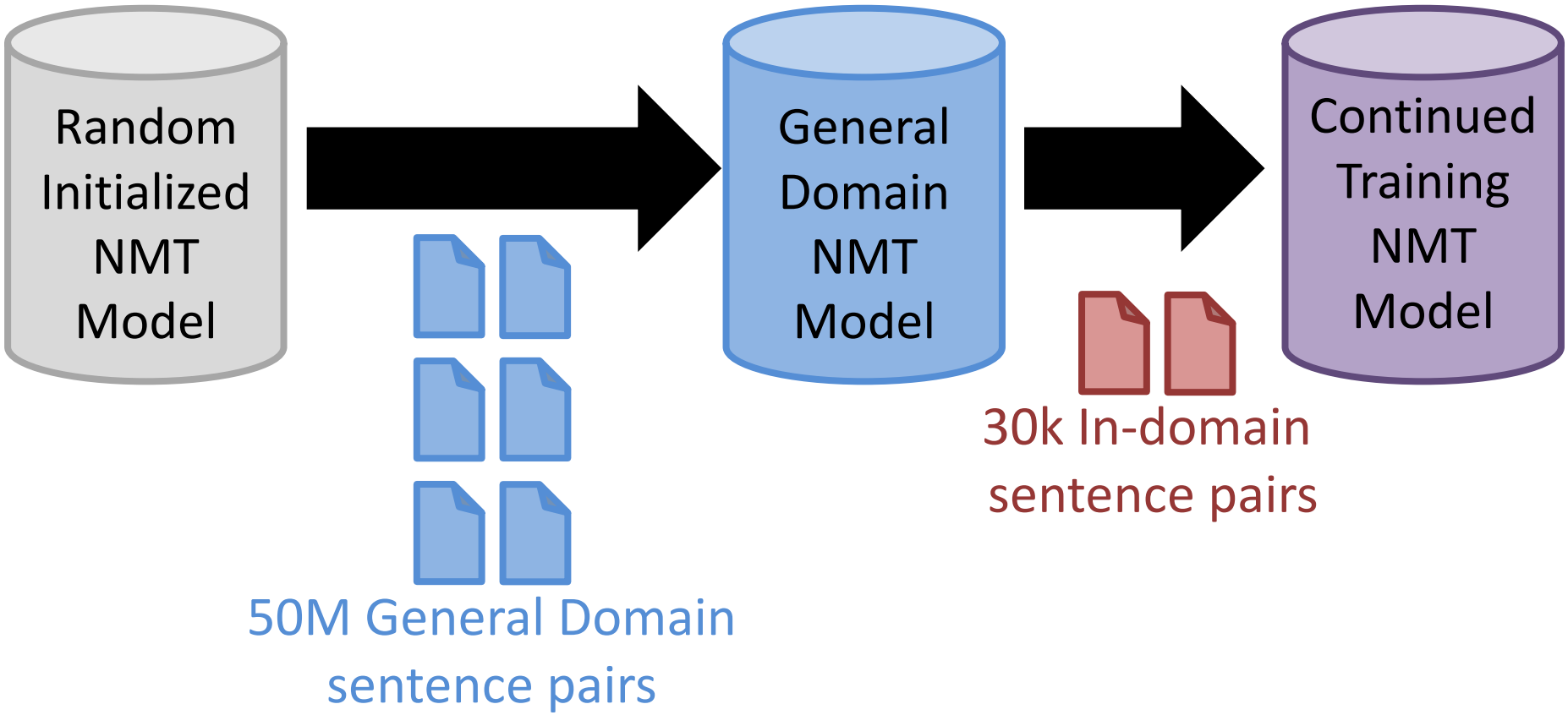
In-Domain NMT



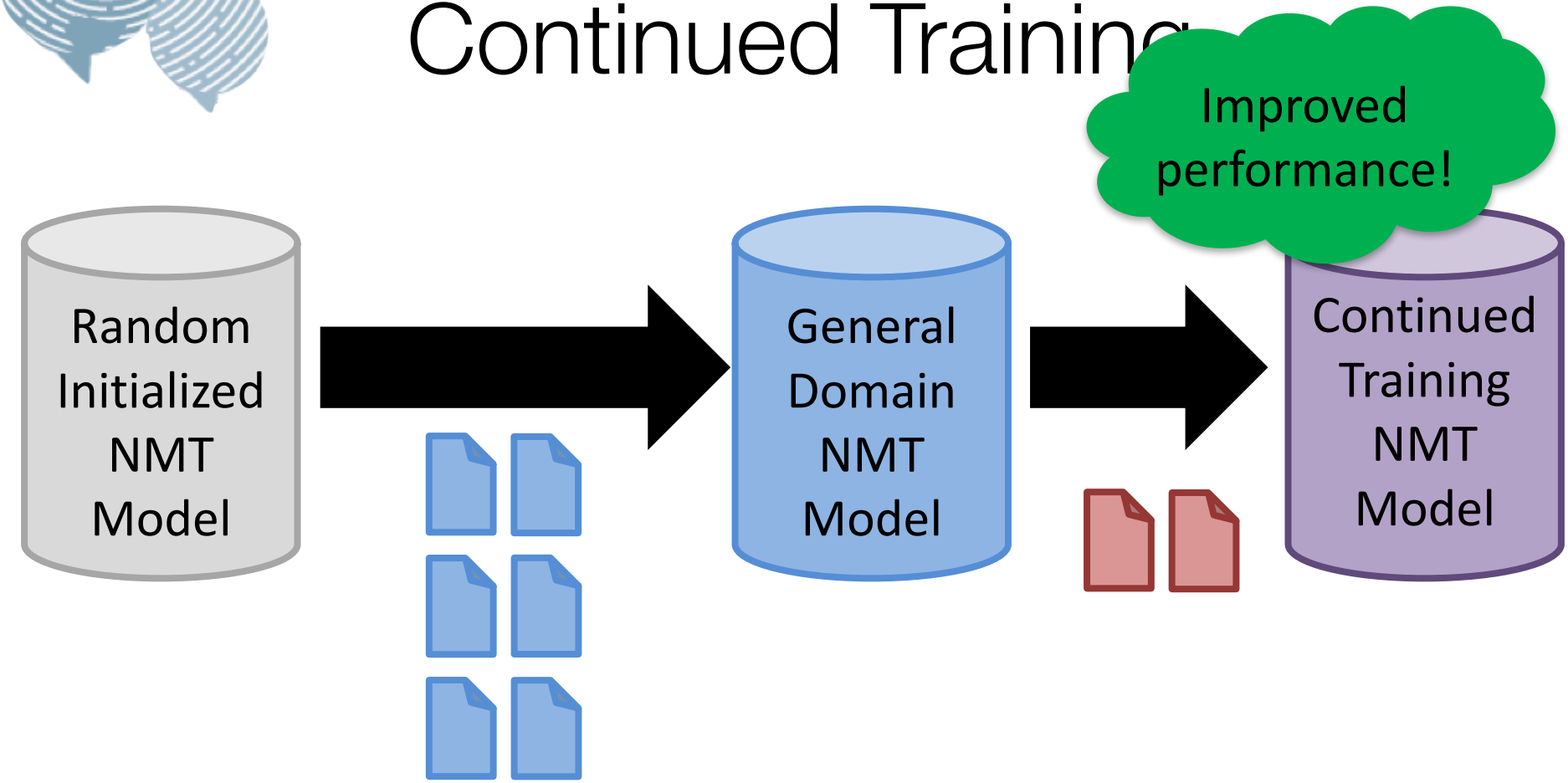
дверной замок повышенной степени защищенности от взлома
Human: door lock with increased degree of security against burglary
System: door lock for a high degree of protection against coke

Domain Adaptation

Continued Training



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

Results

BLEU

Weighted *n*-gram 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

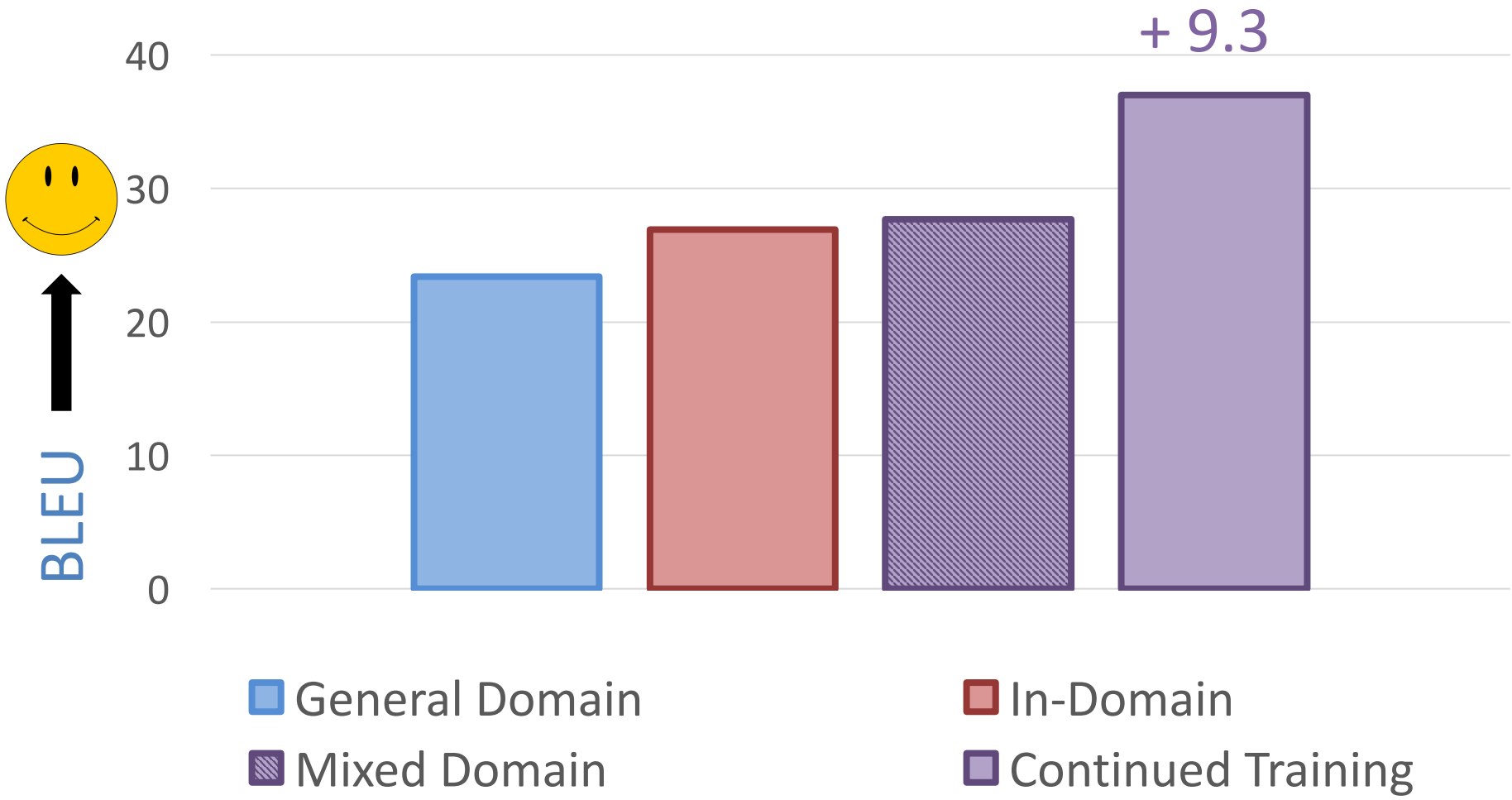
(often scaled to be 0-100)

Higher is better

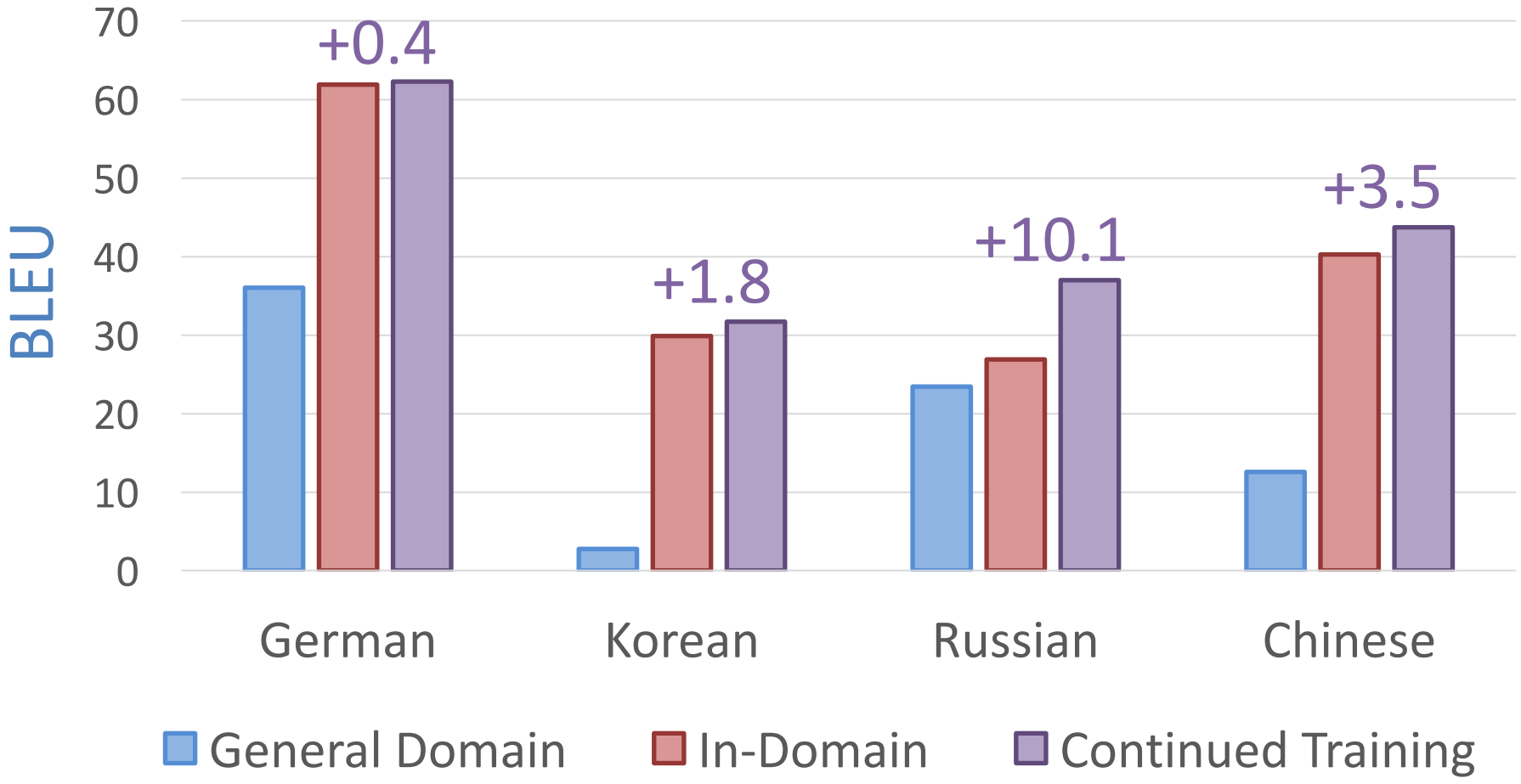
Imperfect…

But… cheap & correlates with humans

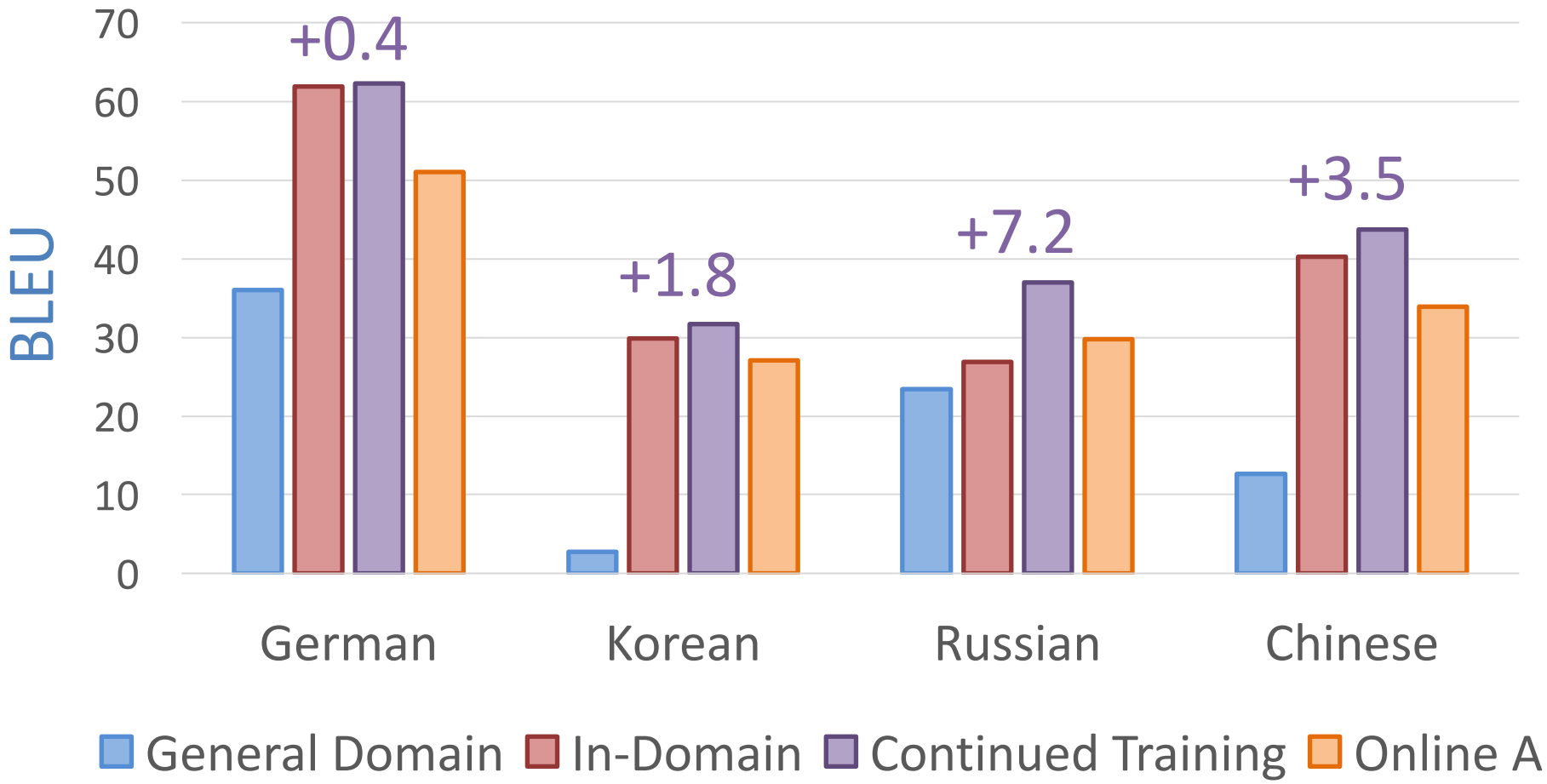
Russian Patent



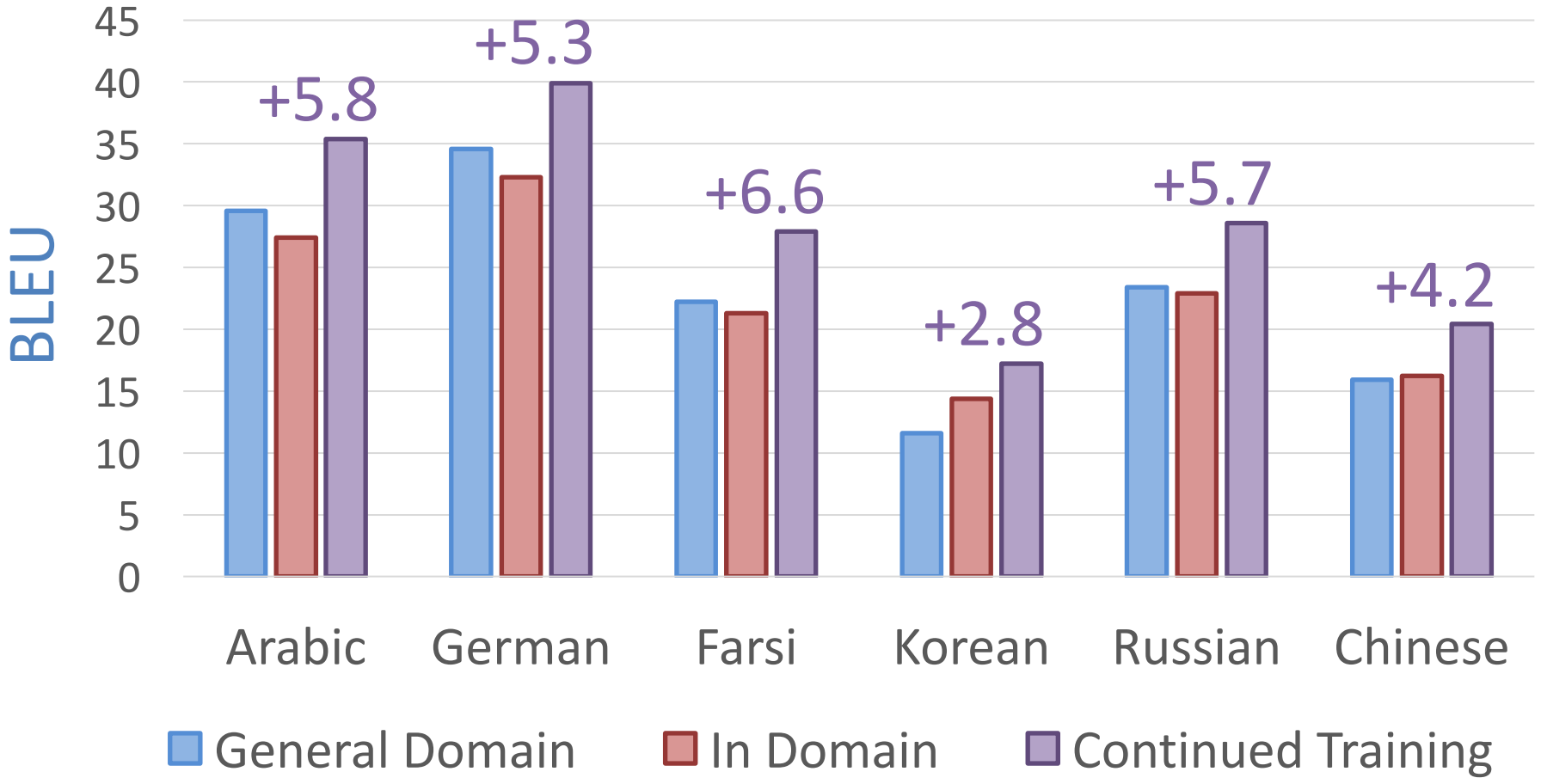
Patent Results



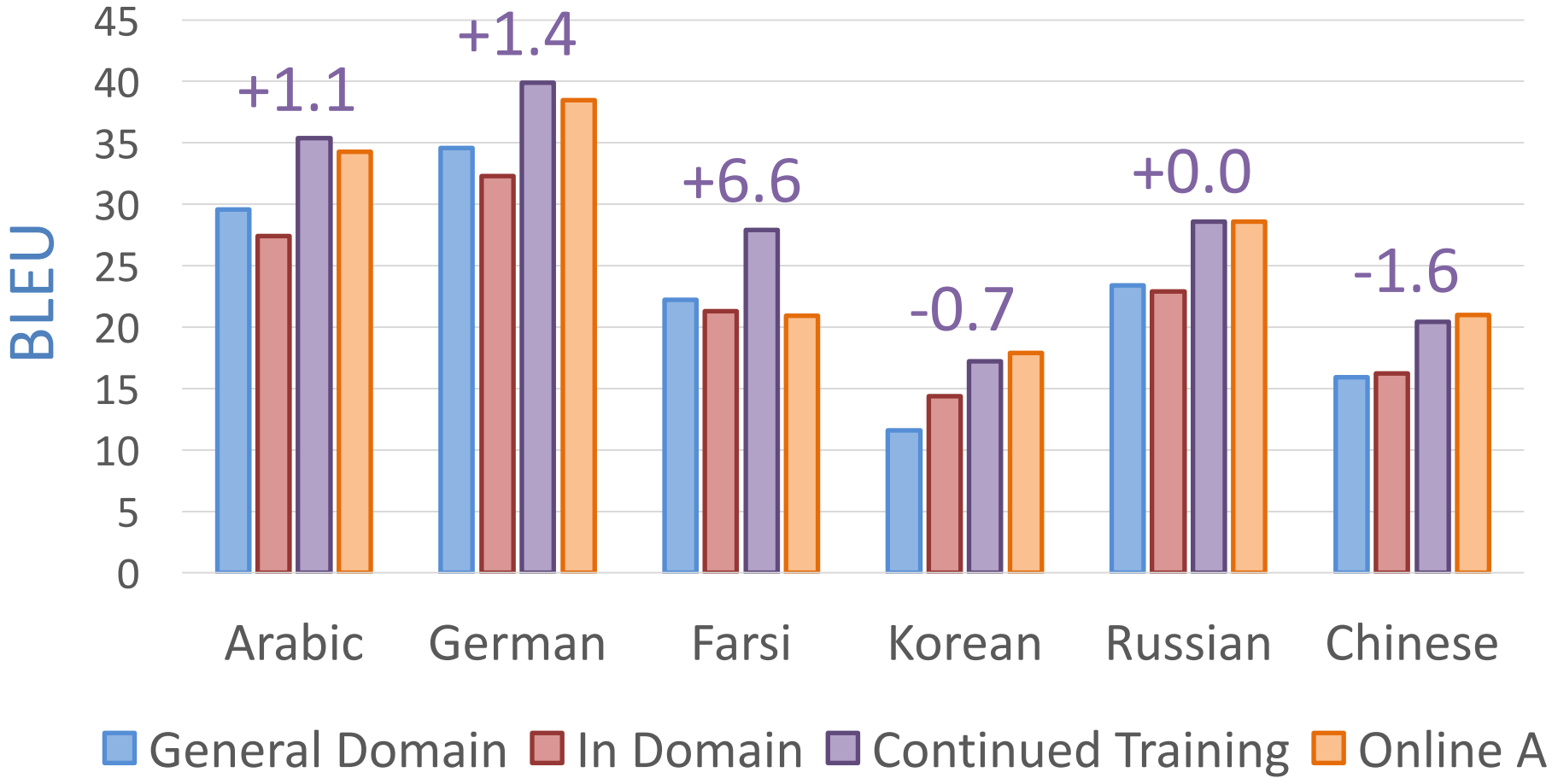
Patent Results



TED Results

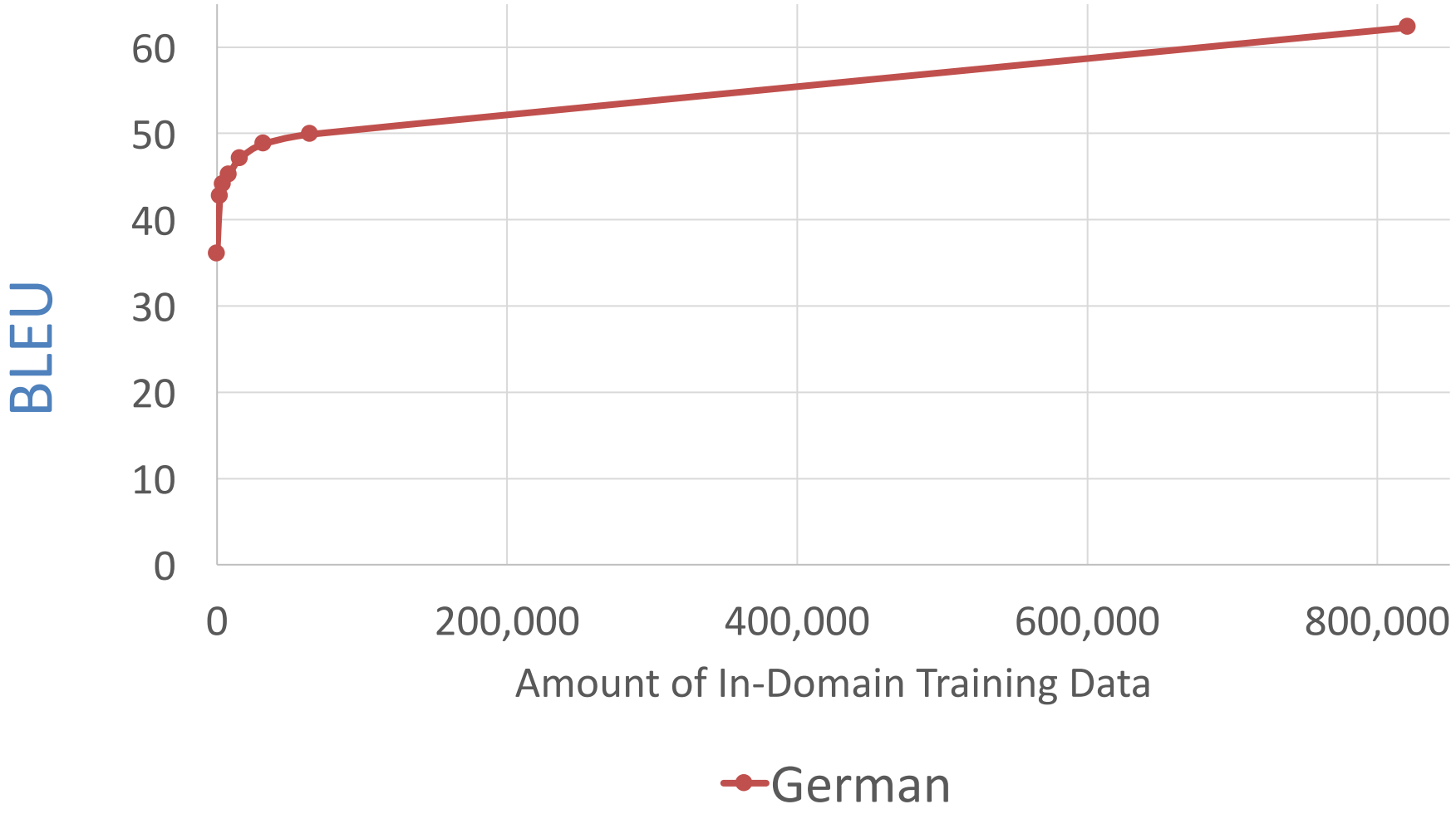

BLEU
45
40
35
30
25
20
15
10
5
0
+5.8
+5.3
+6.6
+2.8
+5.7
+4.2
Arabic
German
Farsi
Korean
Russian
Chinese
General Domain
In Domain
Continued Training

TED Results

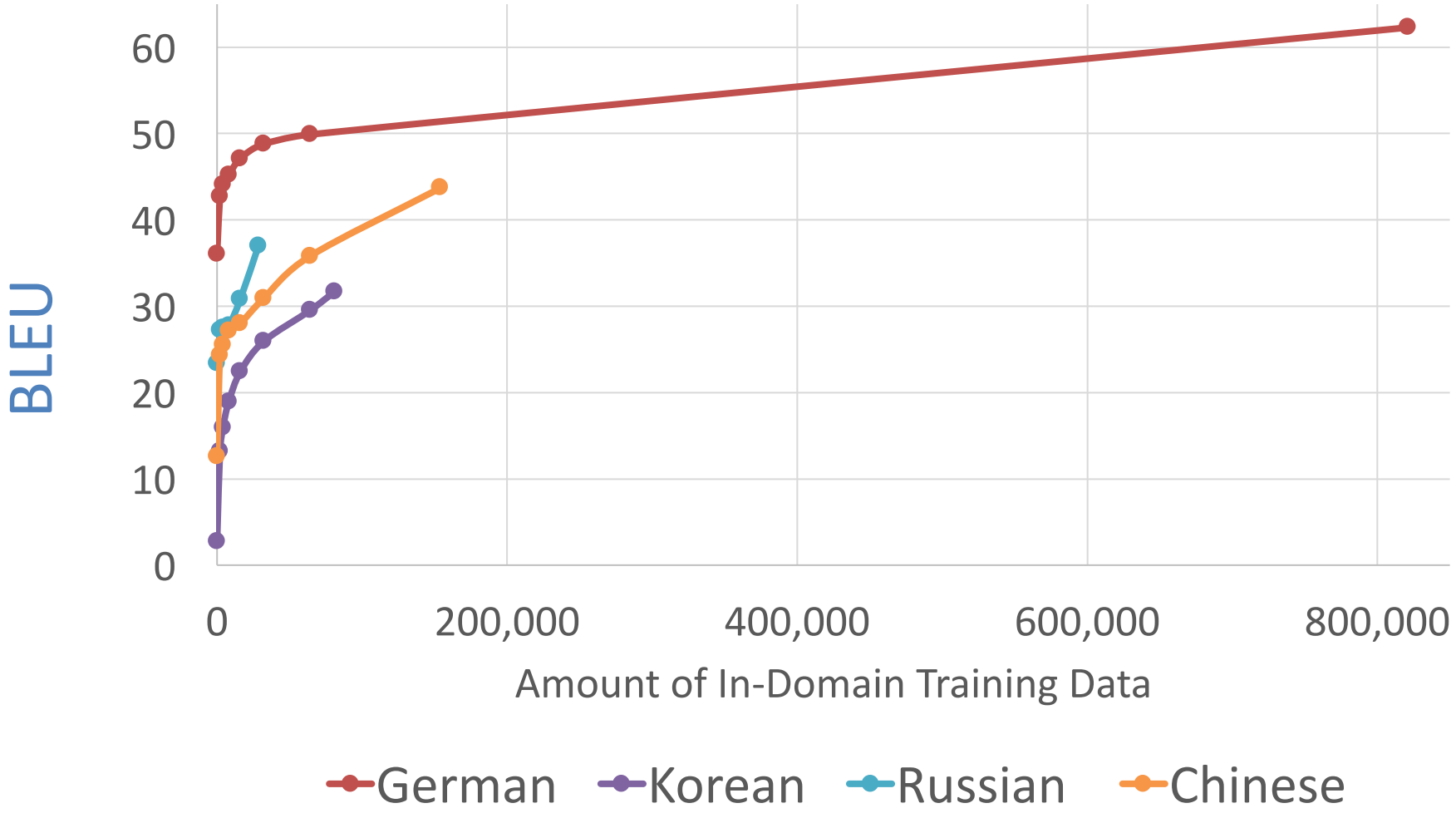

BLEU
45
40
35
30
25
20
15
10
5
0
+1.1
+1.4
+6.6
-0.7
+0.0
-1.6
Arabic
German
Farsi
Korean
Russian
Chinese
General Domain
In Domain
Continued Training
Online A

How much data do we need?

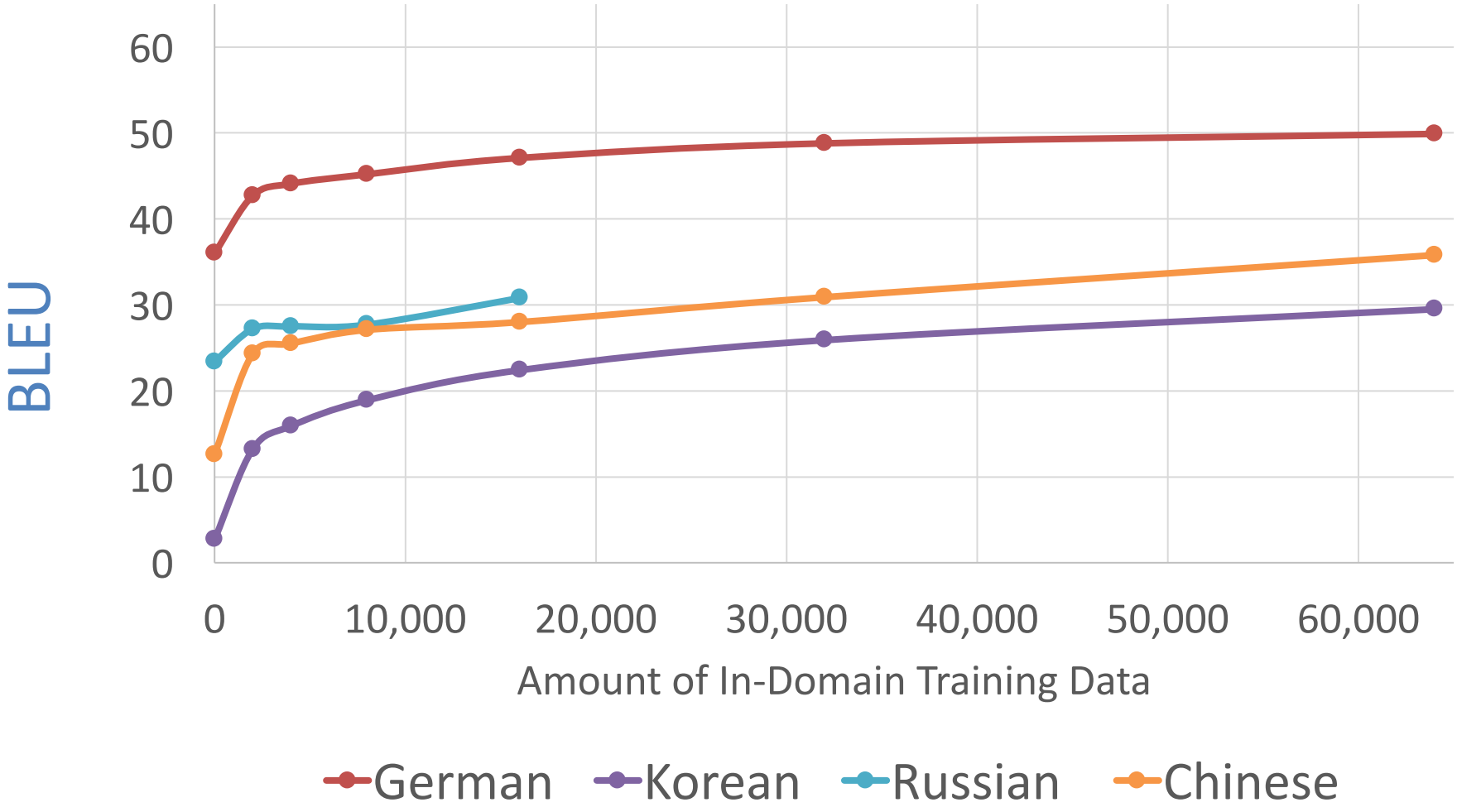
Patent


BLEU
60
50
40
30
20
10
0
0
200,000
400,000
600,000
800,000
Amount of In-Domain Training Data
German

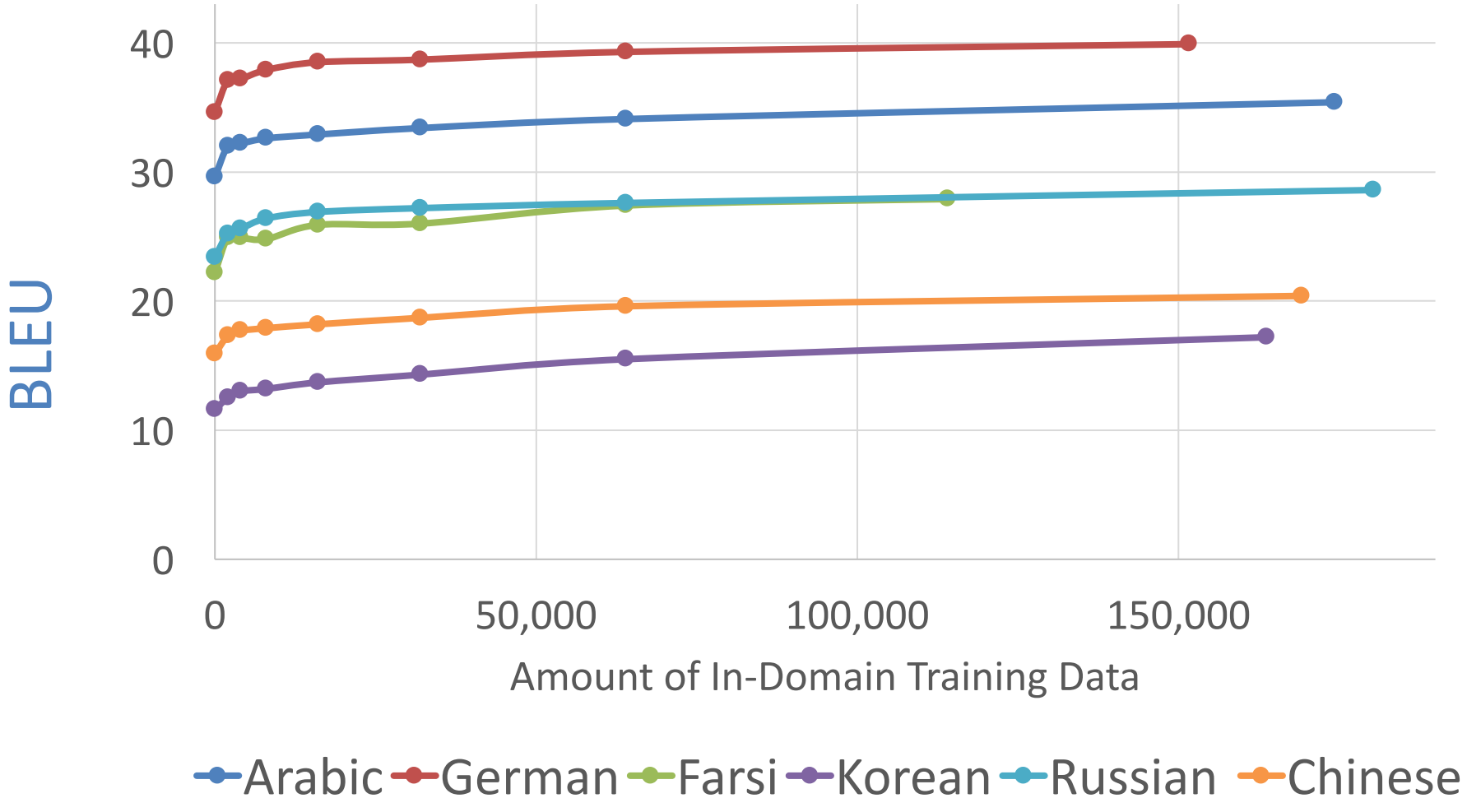
Patent


BLEU
60
50
40
30
20
10
0
0
200,000
400,000
600,000
800,000
Amount of In-Domain Training Data
German
Korean
Russian
Chinese

Patent


BLEU
60
50
40
30
20
10
0
0
10,000
20,000
30,000
40,000
50,000
60,000
Amount of In-Domain Training Data
German
Korean
Russian
Chinese

TED

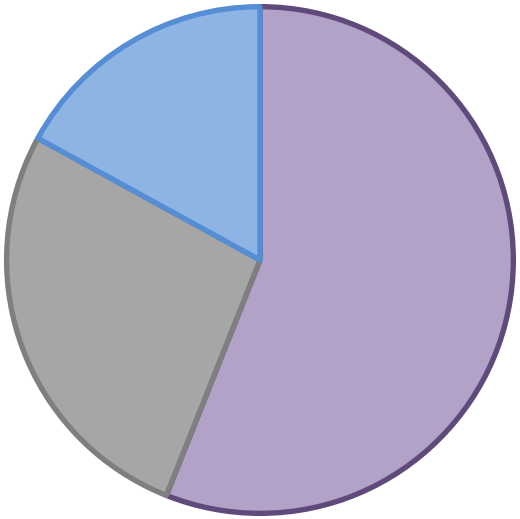

BLEU
40
30
20
10
0
0
50,000
100,000
150,000
Amount of In-Domain Training Data
Arabic
German
Farsi
Korean
Russian
Chinese

Human Evaluation

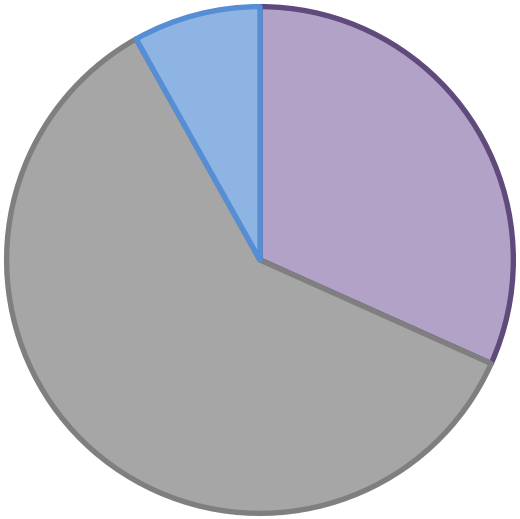
Source	Output 1	Output 2	Ranking
等 了 十个月， 我 终于 见到 了 他 - 将近 一年 啊。	waiting for 10, i finally met him- nearly a year .	after 10 months, i finally saw him- nearly a year.	Output 2 is better
这 就是 免费 的 代 价 。	that's the price of free.	that's the cost of free.	Output 1 is better
我 是 说， 我 已经 够 紧张 的 了	i mean, i'm nervous enough.	i mean, i'm nervous enough.	Both translations are about the same

Continued Training vs General

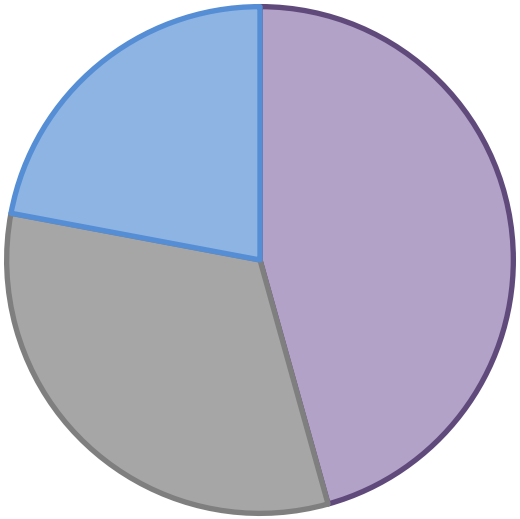
Arabic



Korean



Chinese

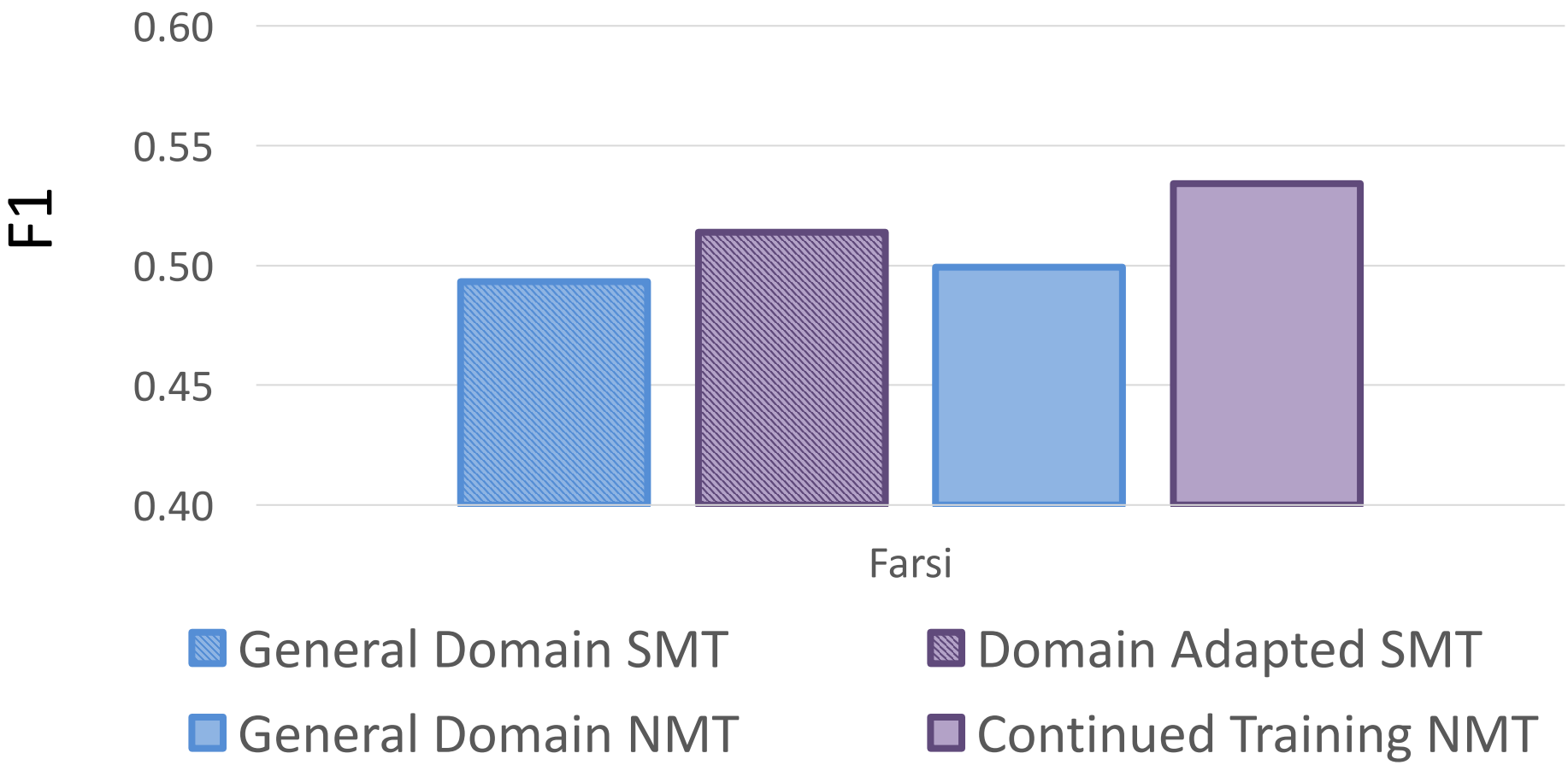


Continued Training Tie General

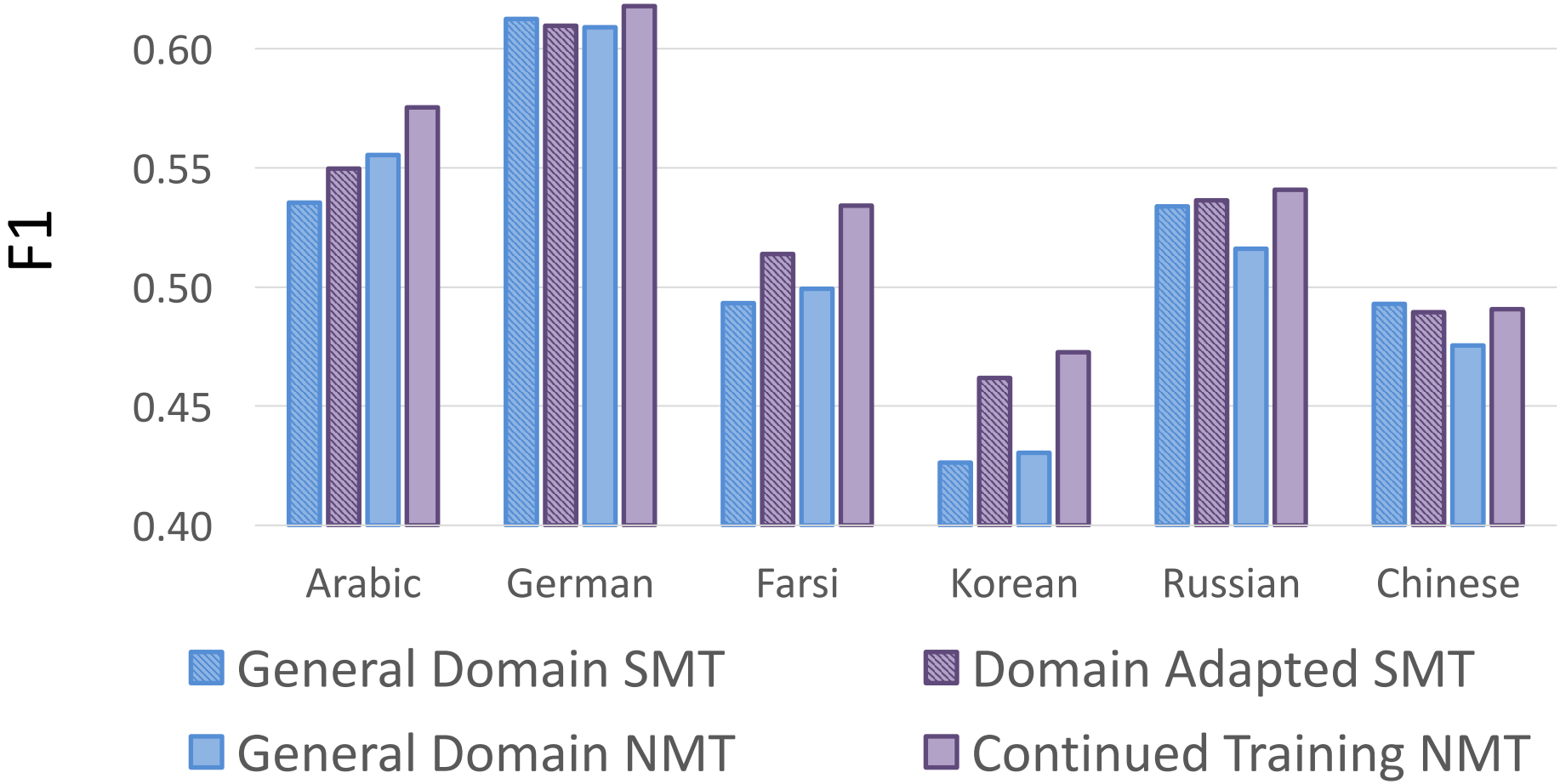
JOHNS HOPKINS UNIVERSITY

Keyword Search

Farsi TED talk NER MicroAvg F1



TED talk NER MicroAvg F1



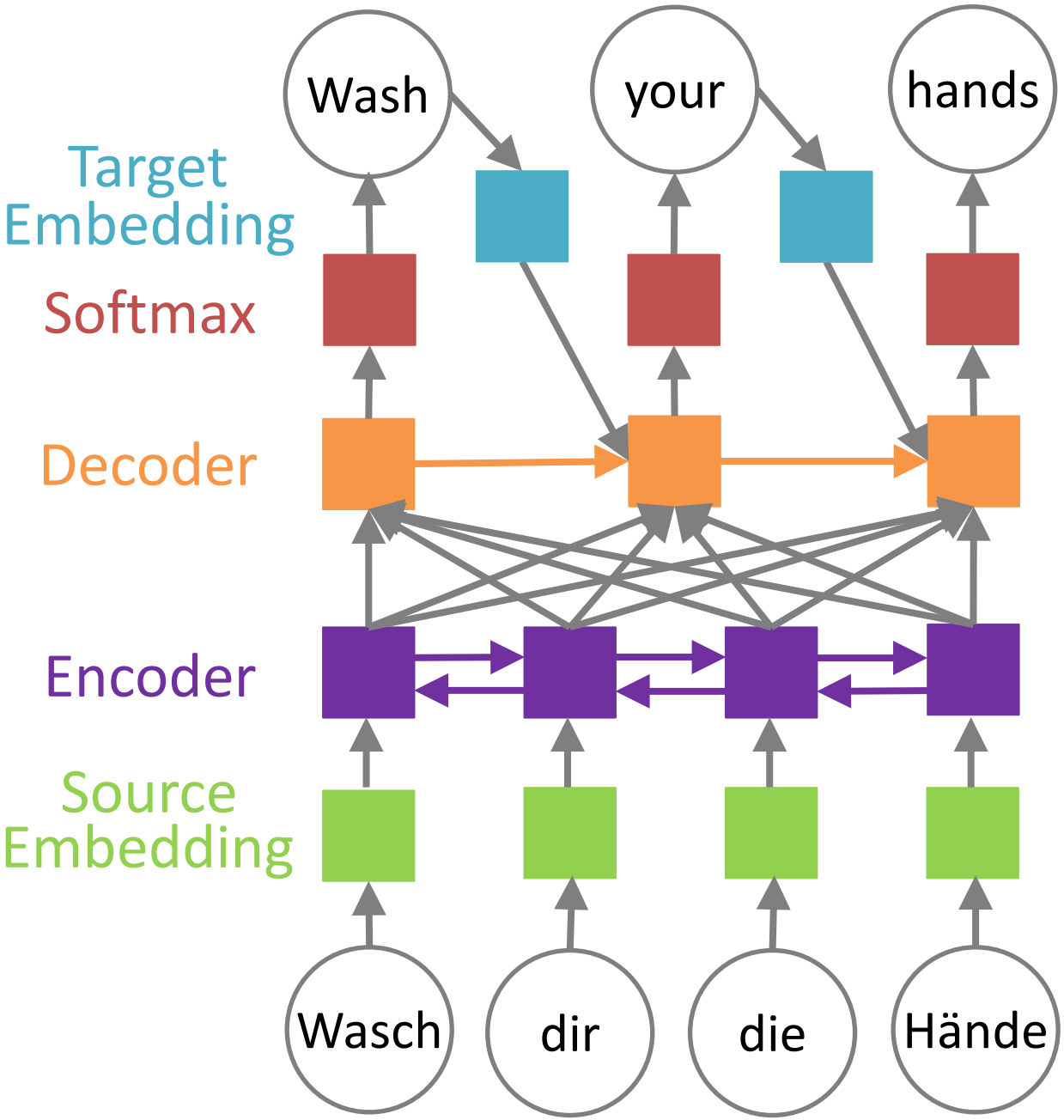
Research Directions

Analysis of Continued Training

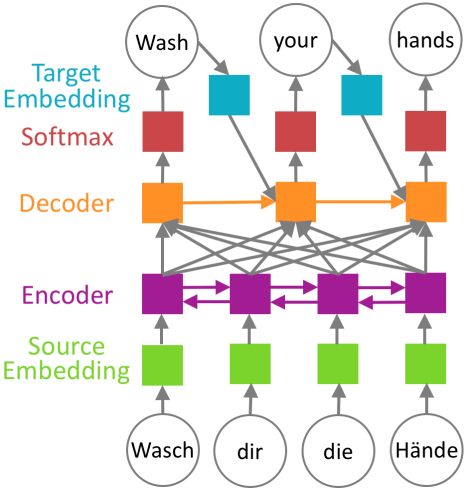
Accepted at EMNLP workshop on

Analyzing and Interpreting Neural Networks for NLP

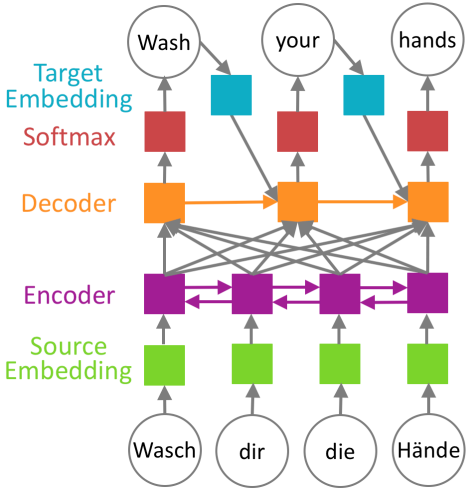
Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson and Philipp Koehn

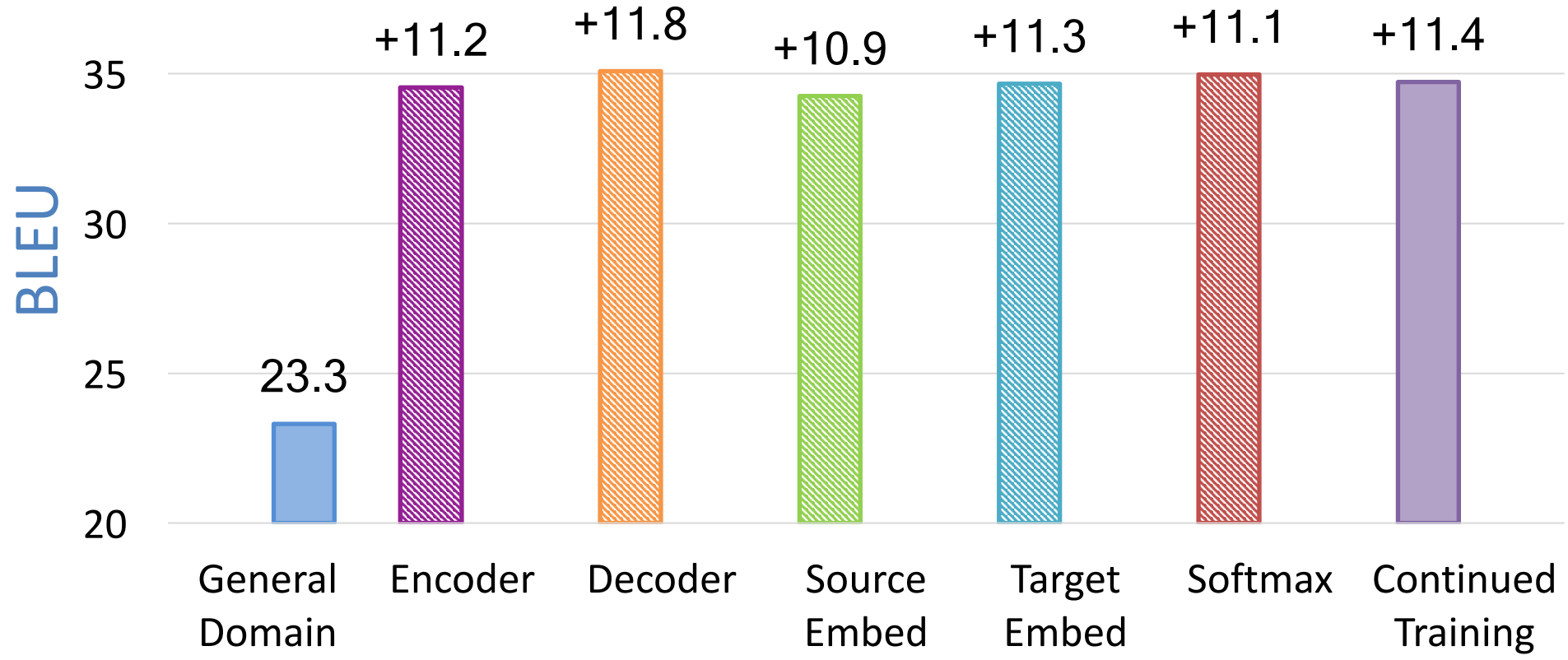

Wash
your
hands
Target Embedding
Softmax
Decoder
Encoder
Source Embedding
Wasch
dir
die
Hände

Selective Training of Components

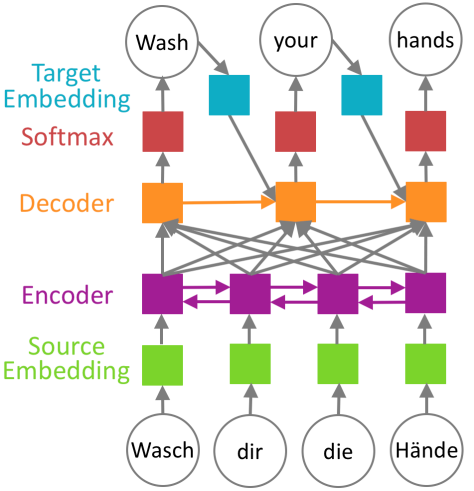


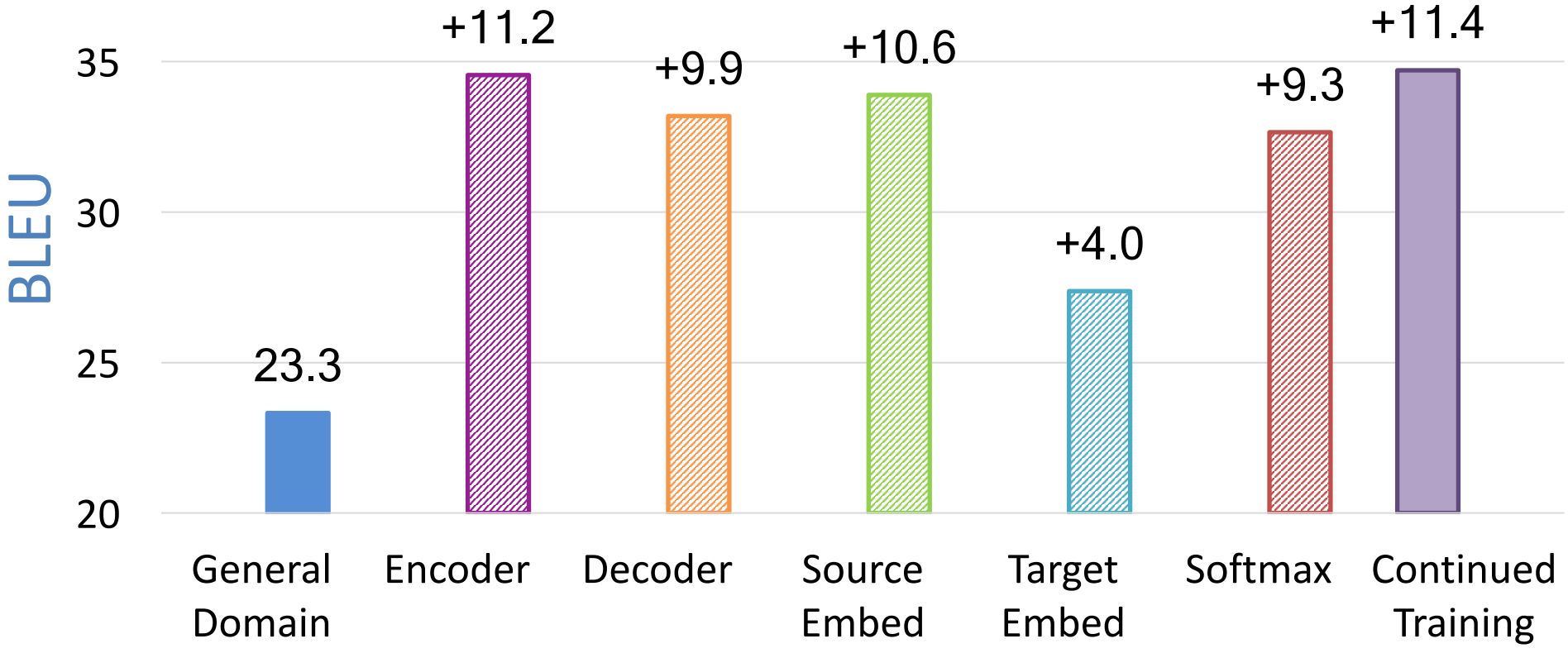
Selective Training of Components



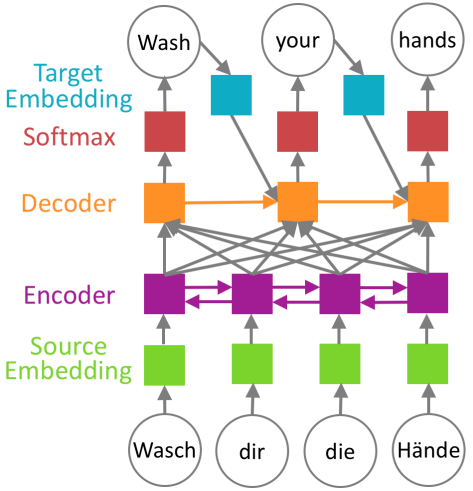


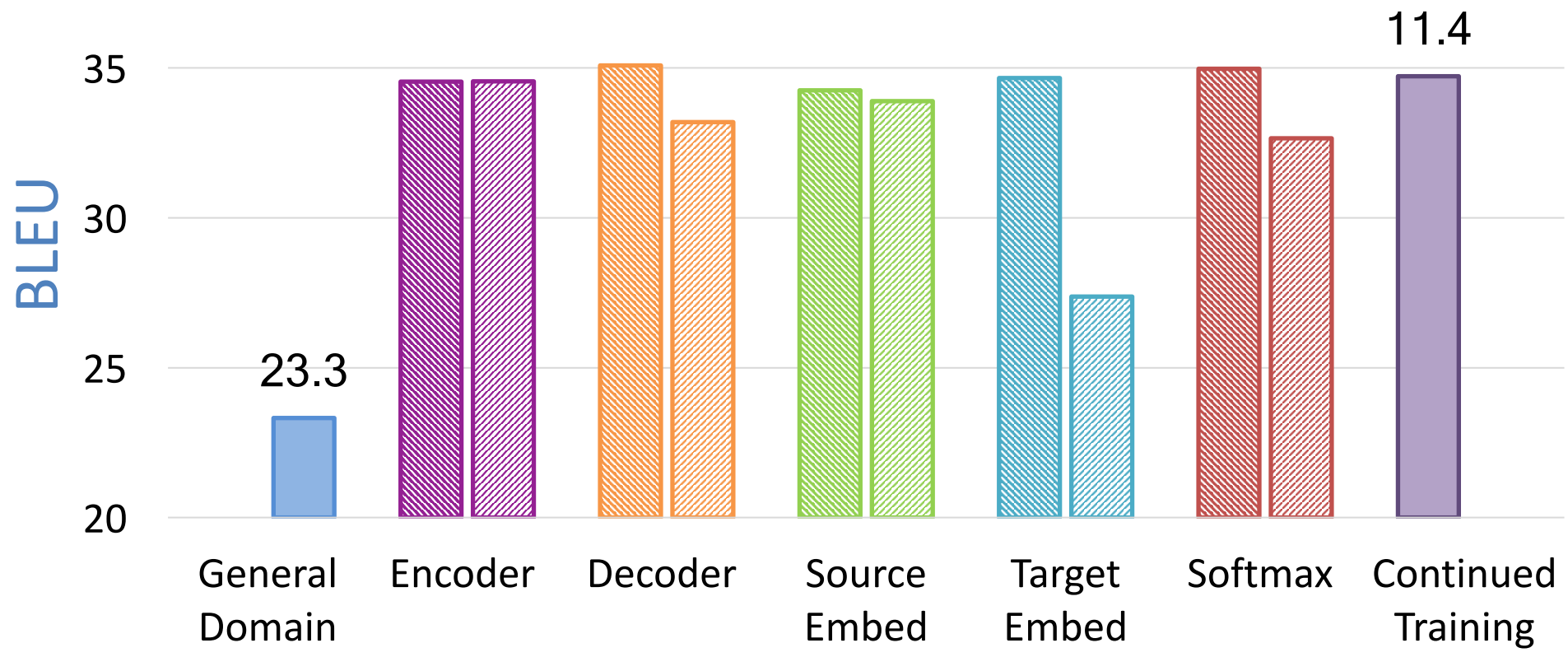
Selective Training of Components





Selective Training of Components

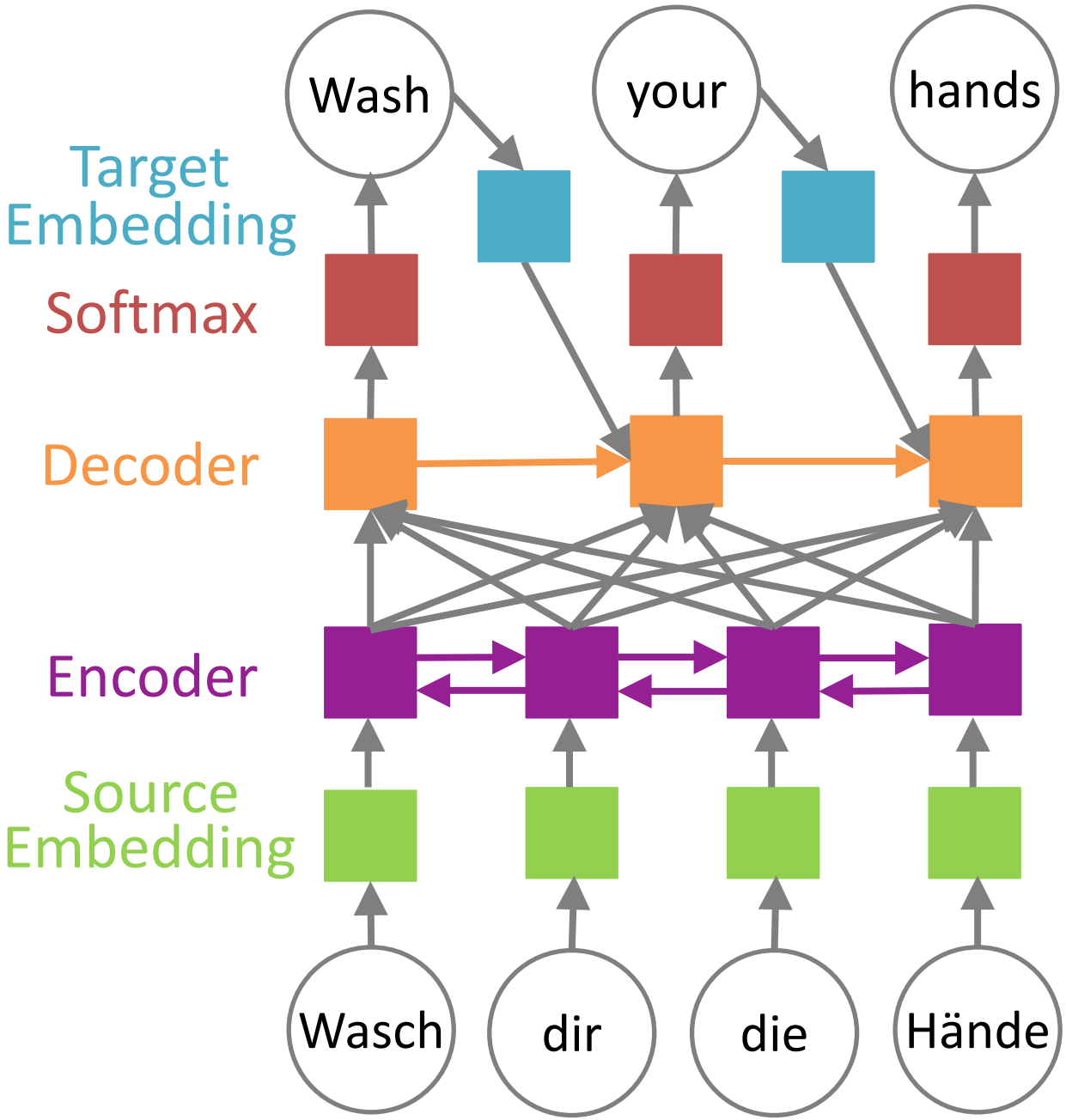


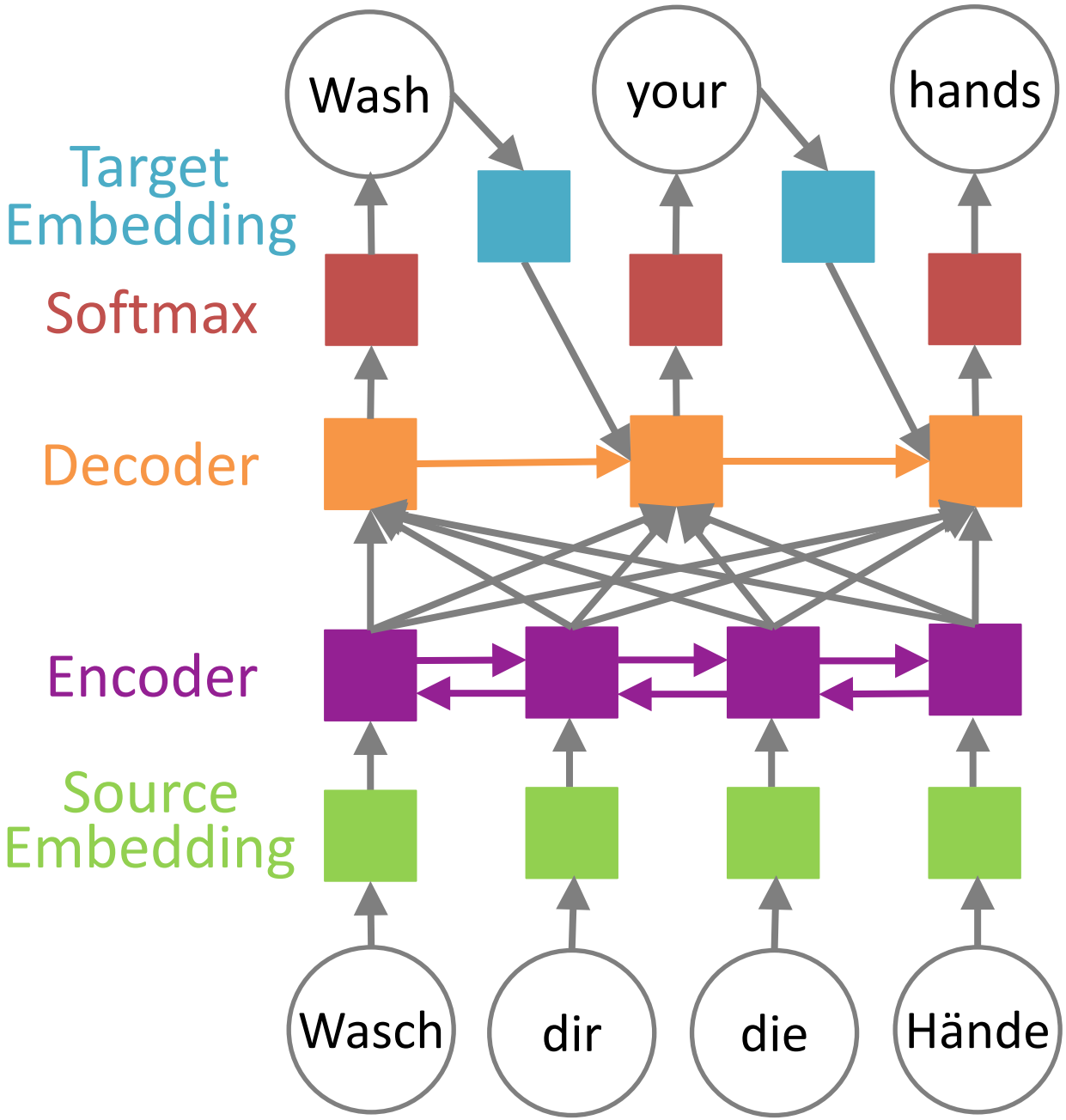


JOHNS HOPKINS UNIVERSITY


JOHNS HOPKINS
UNIVERSITY

Extra Slides


Wash
your
hands
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

Hyperparameters

Model architecture

- num_embed="512:512"
- rnn_num_hidden=512
- rnn_attention_type="dot"
- num_layers=2
- rnn_cell_type="lstm"

Regularization

- embed_dropout=0.0
- rnn_dropout=0.1
- label_smoothing=0.1

Vocabulary

- BPE on Source and Target
- num_words=30k:30k
- word_min_count="1:1"
- max_seq_len="100:100”

Training configuration

- batch_size=4096
- optimizer=adam
- initial_learning_rate=0.0003
- learning_rate_reduce_factor=0.7
- loss="cross-entropy”
- checkpoint_frequency=4000

Alternate MT explanation

Case Study

Our office needs to translate a lot of Russian patents.

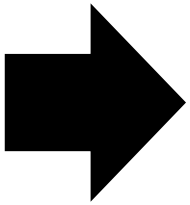
We have a few translators, but they can only process a small fraction of our data.

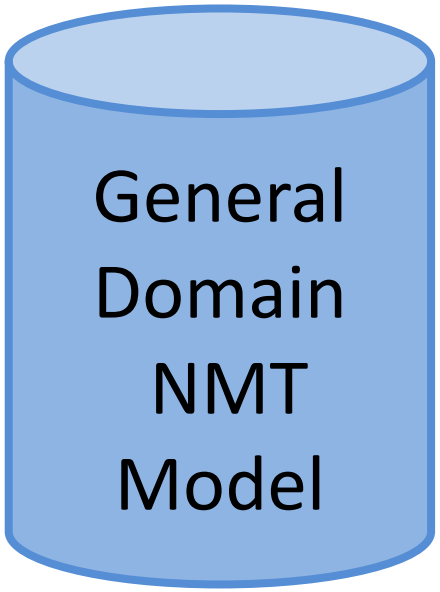
We would like to use machine translation find the most interesting documents and let our translators focus on those.

We know neural machine translation has state-of-the-art performance, so we decide to build a Neural system…

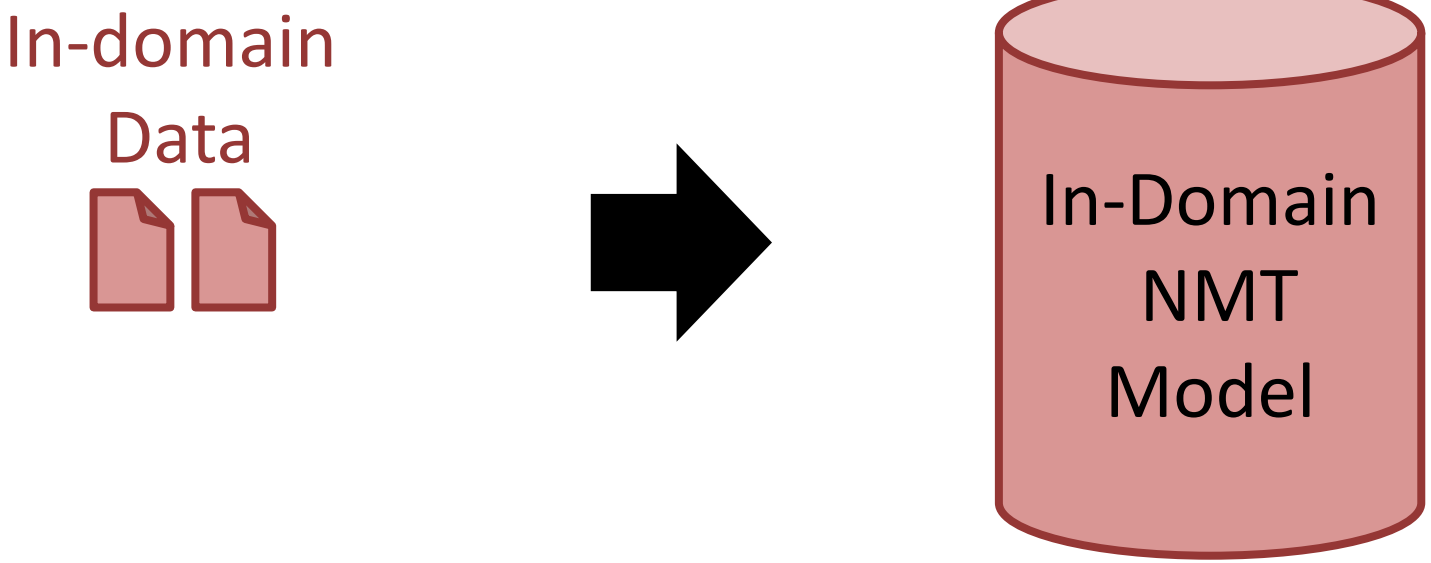
MT training

General Domain Data

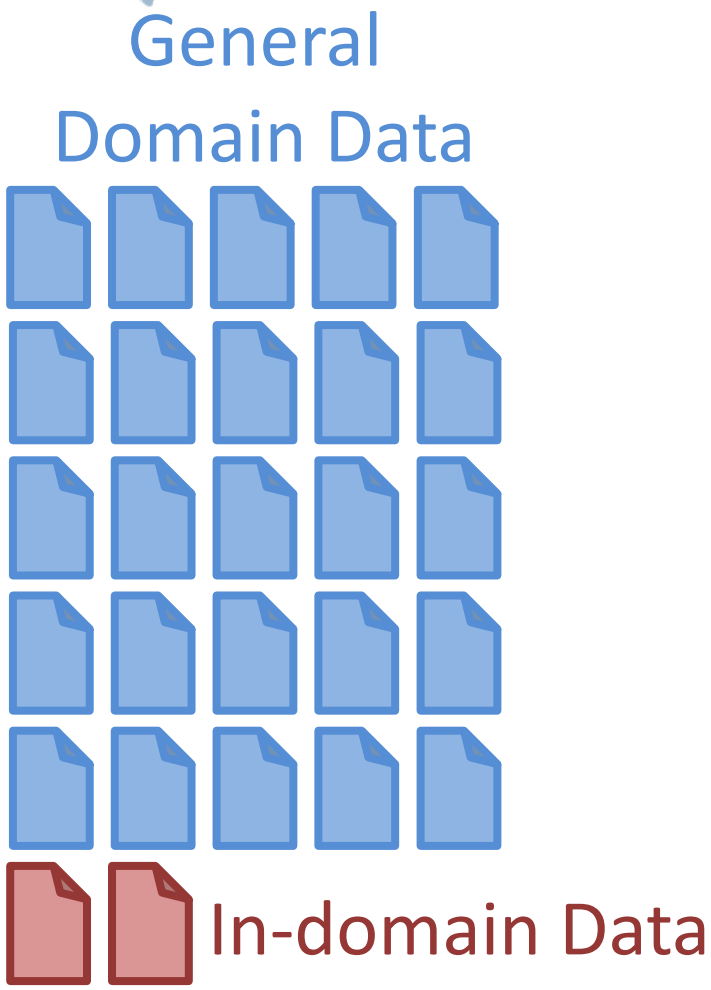


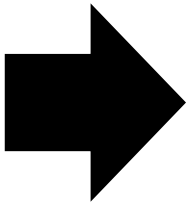


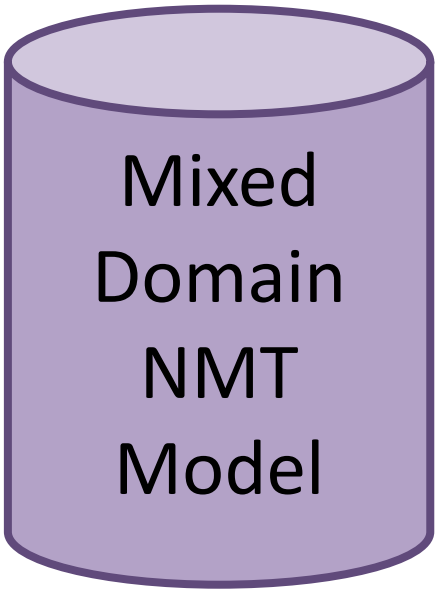
MT training



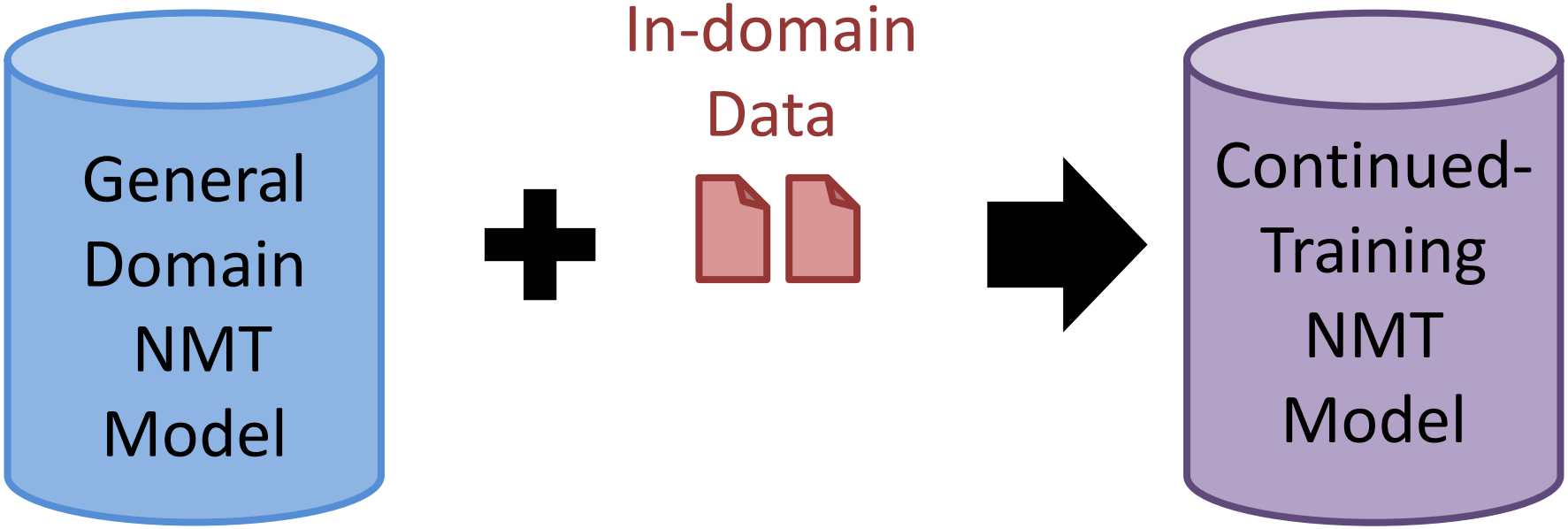
MT training



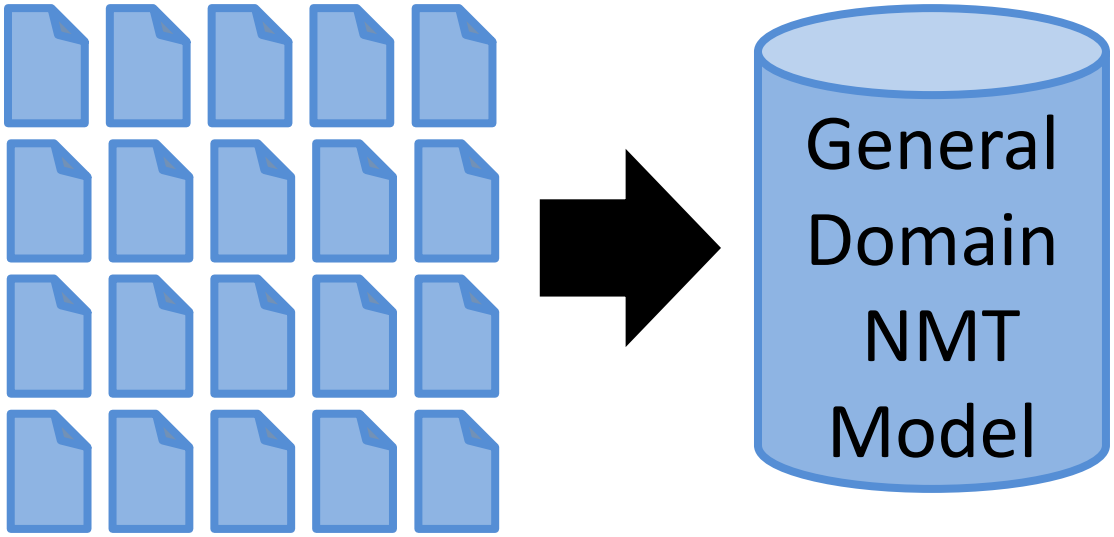




MT training

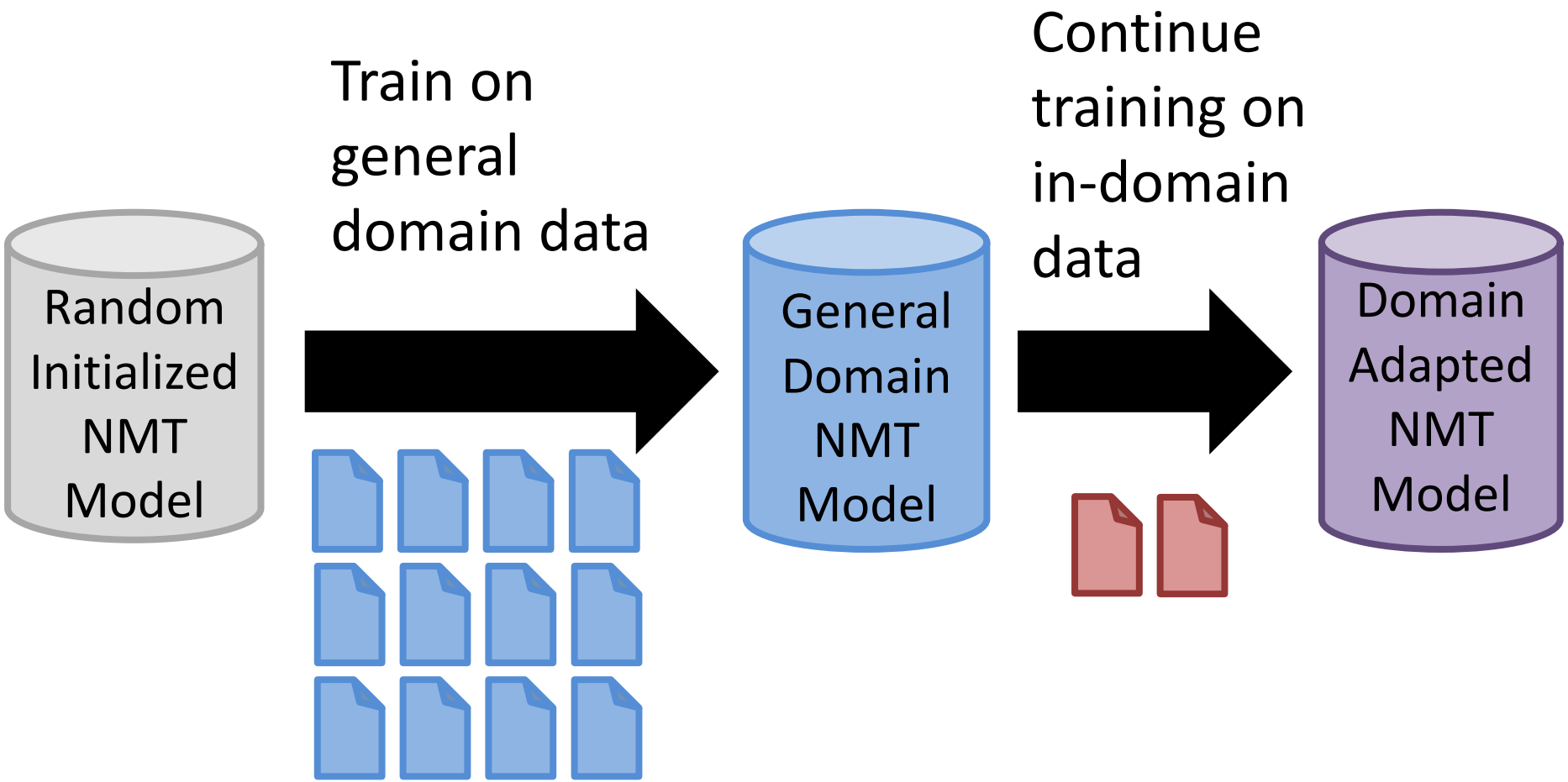

General Domain NMT Model
In-domain Data
Continued-Training NMT Model

MT training

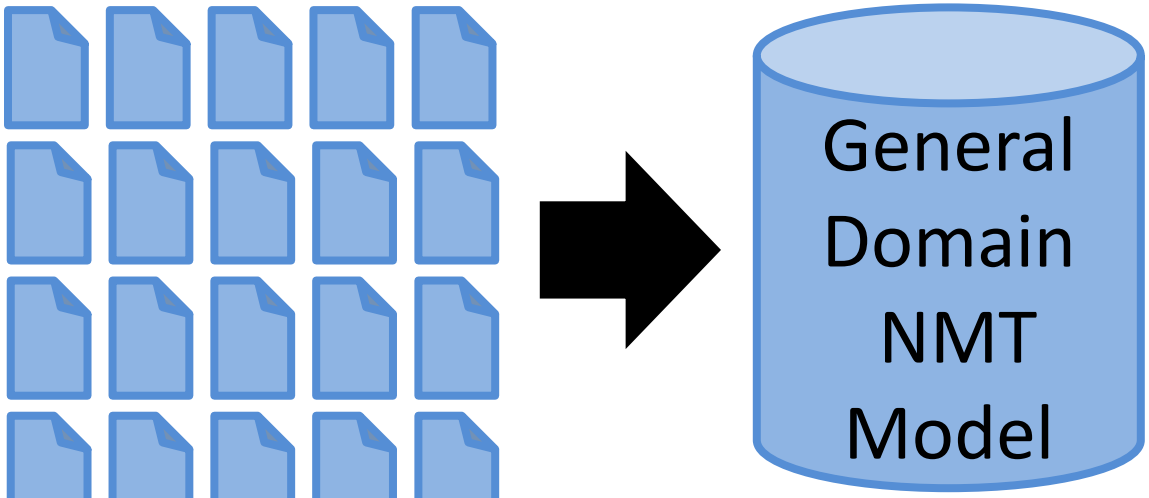


50M General Domain sentence pairs

Continued Training

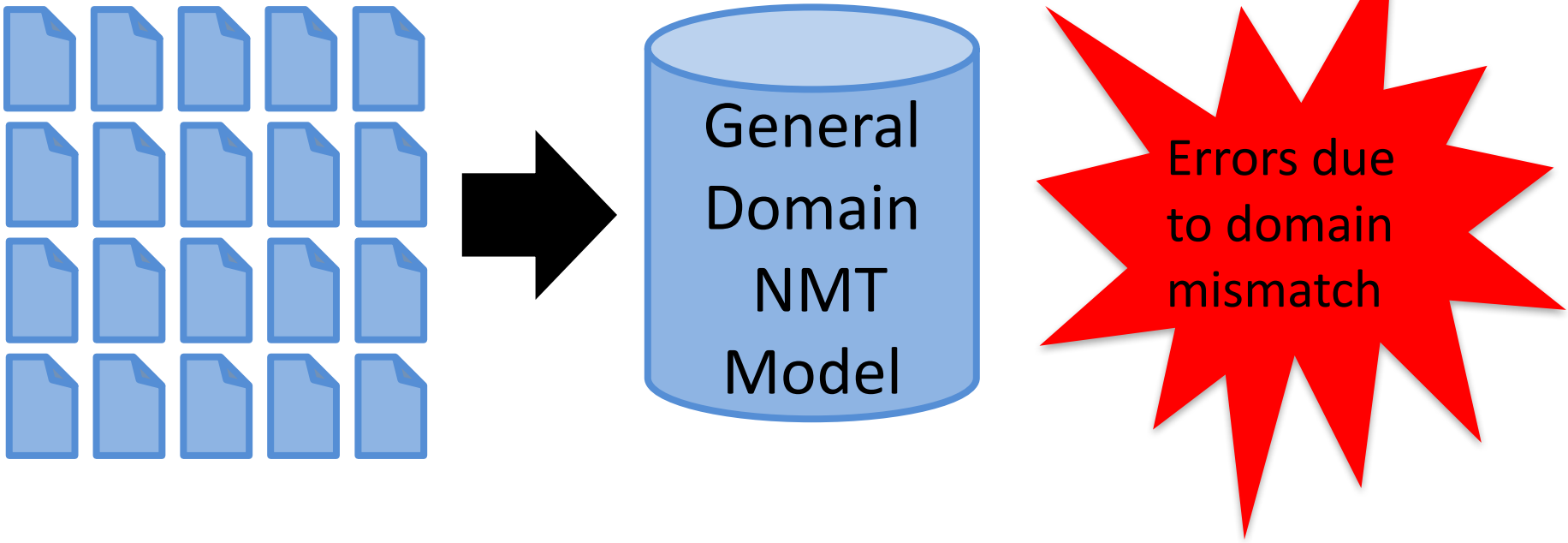


General Domain NMT



50M General Domain sentence pairs

General Domain NMT



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

Keyword Search

Keyword Search (sort of)

Extrinsic measure of MT Output quality based on ability to retrieve (i.e., match) words or phrases

[Insert cartoon]

Human assigned categories

Keyword	venture capitalist, zero gravity, hydrogen
Sentiment	fantastic, messy, bad, happy
Person	Heidi, Chris, Leonardo da Vinci, Aristotle
Organization	Toyota, UNESCO, Ikea, Swedish Army
Geo-Political Entity	Egypt, San Francisco, Haiti
Location	Arctic, Africa, hospital, ER, lobby
Date	Friday, 1980s, last March, today
Temporal Expression	4:00 am, 30-second, six weeks
Numeric Expression	20 percent, 27 kilometers, one-fifth, two nurses

This metric is pessimistic

Inexact matches count as failure

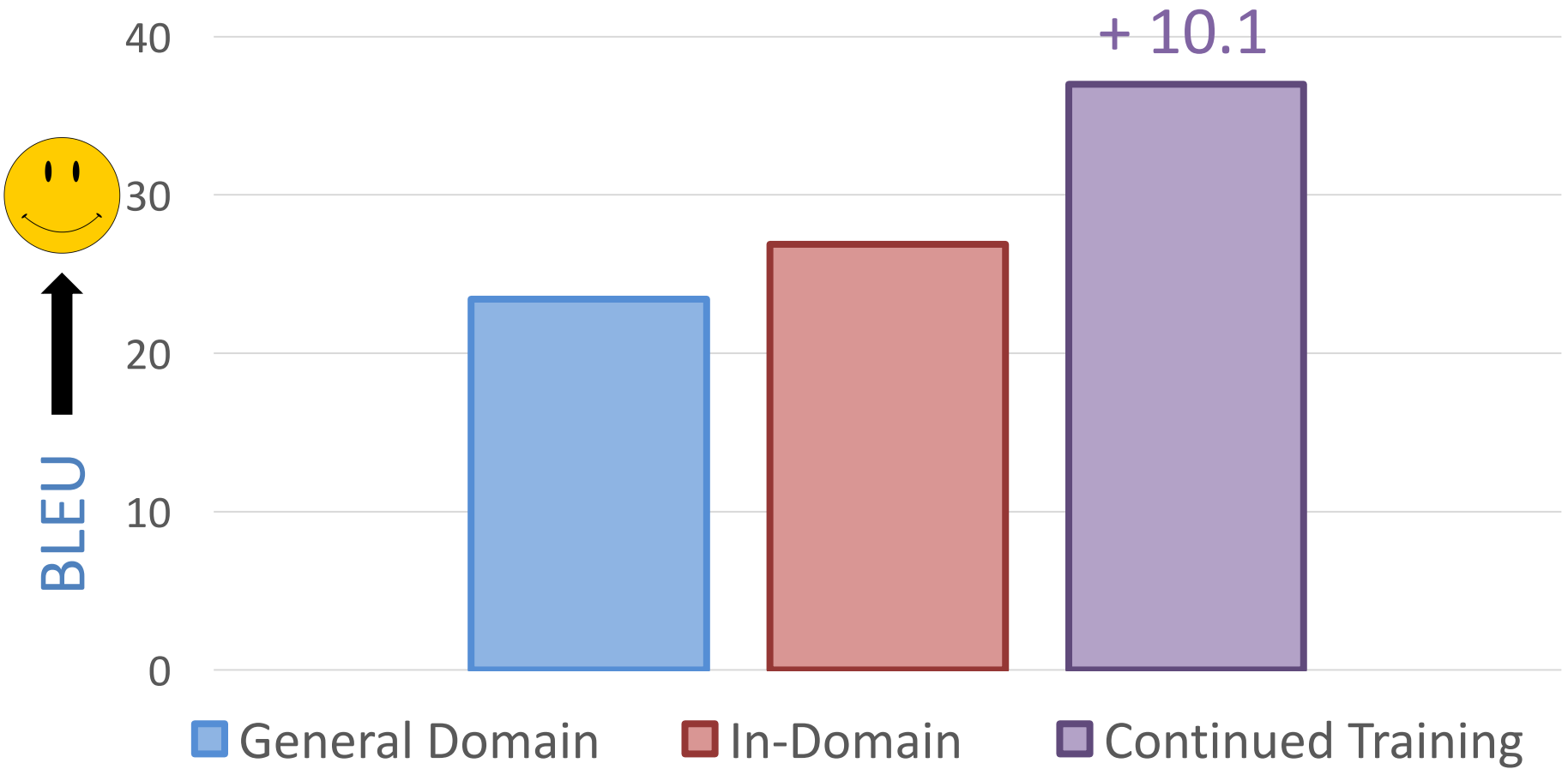
Tokenization issues exacerbate measures

- 70 year old vs. 70-year-old

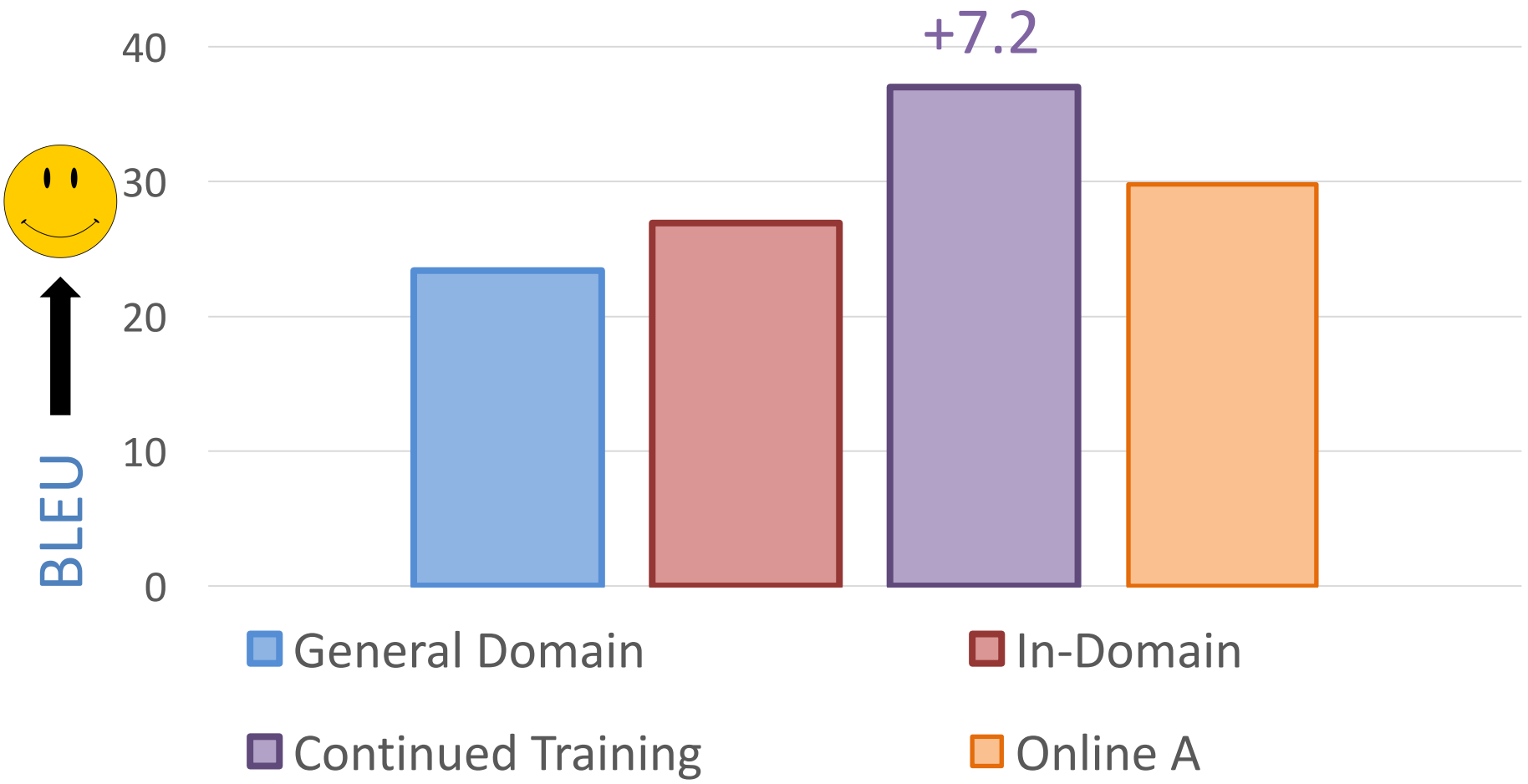
Alternative (very acceptable) translations can count as failure

Results

Russian Patent

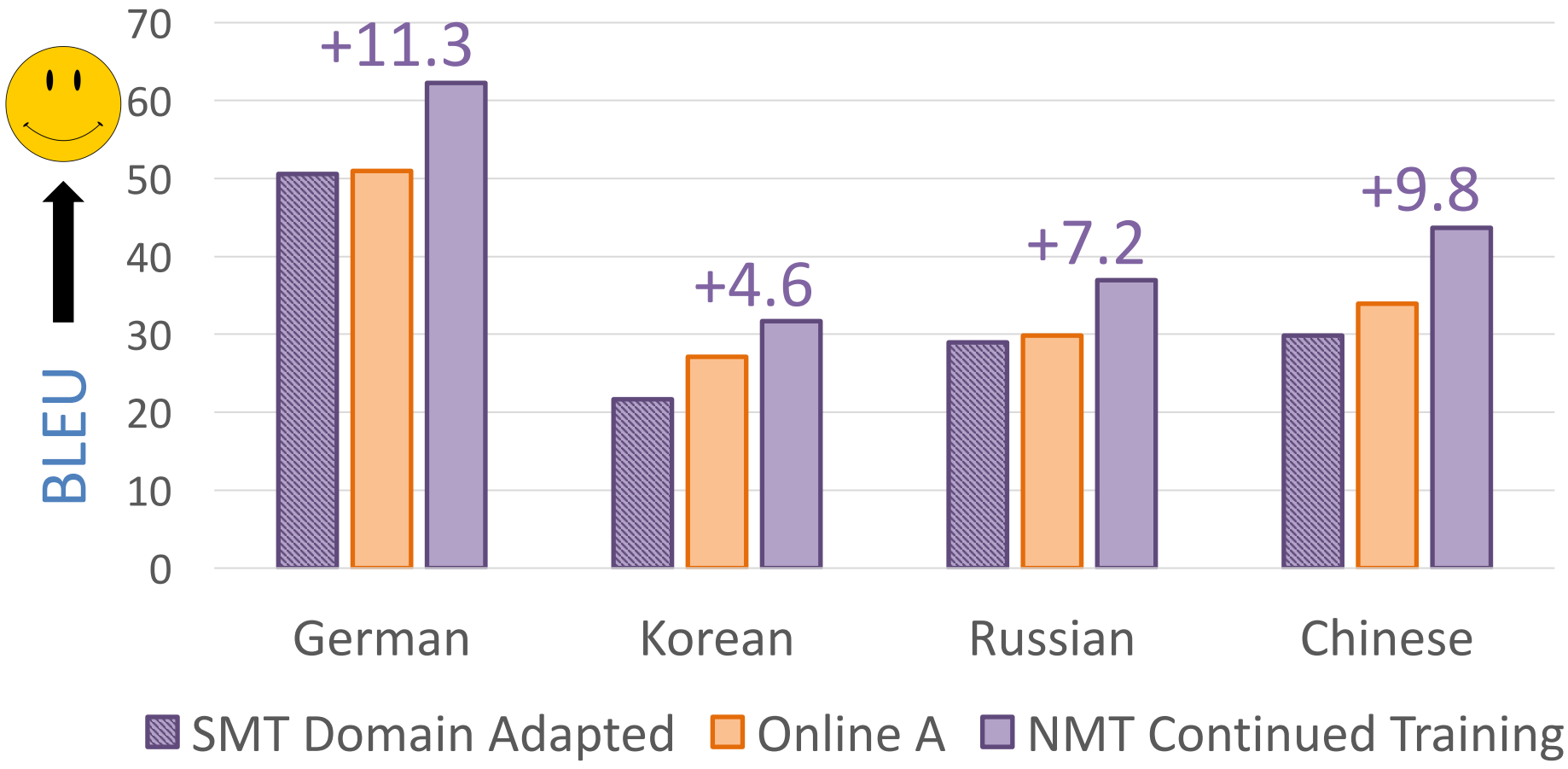


Russian Patent

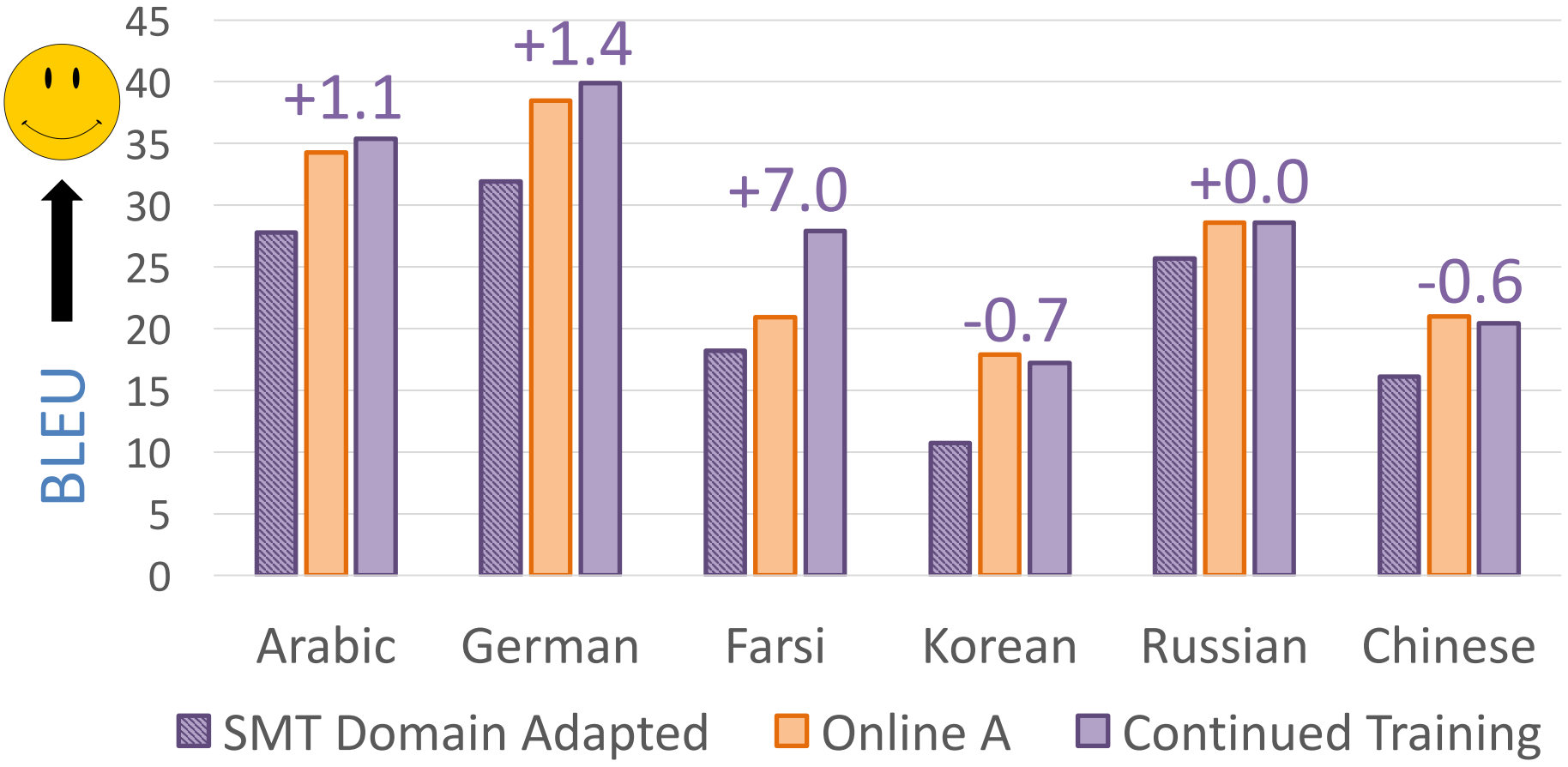


JOHNS HOPKINS UNIVERSITY

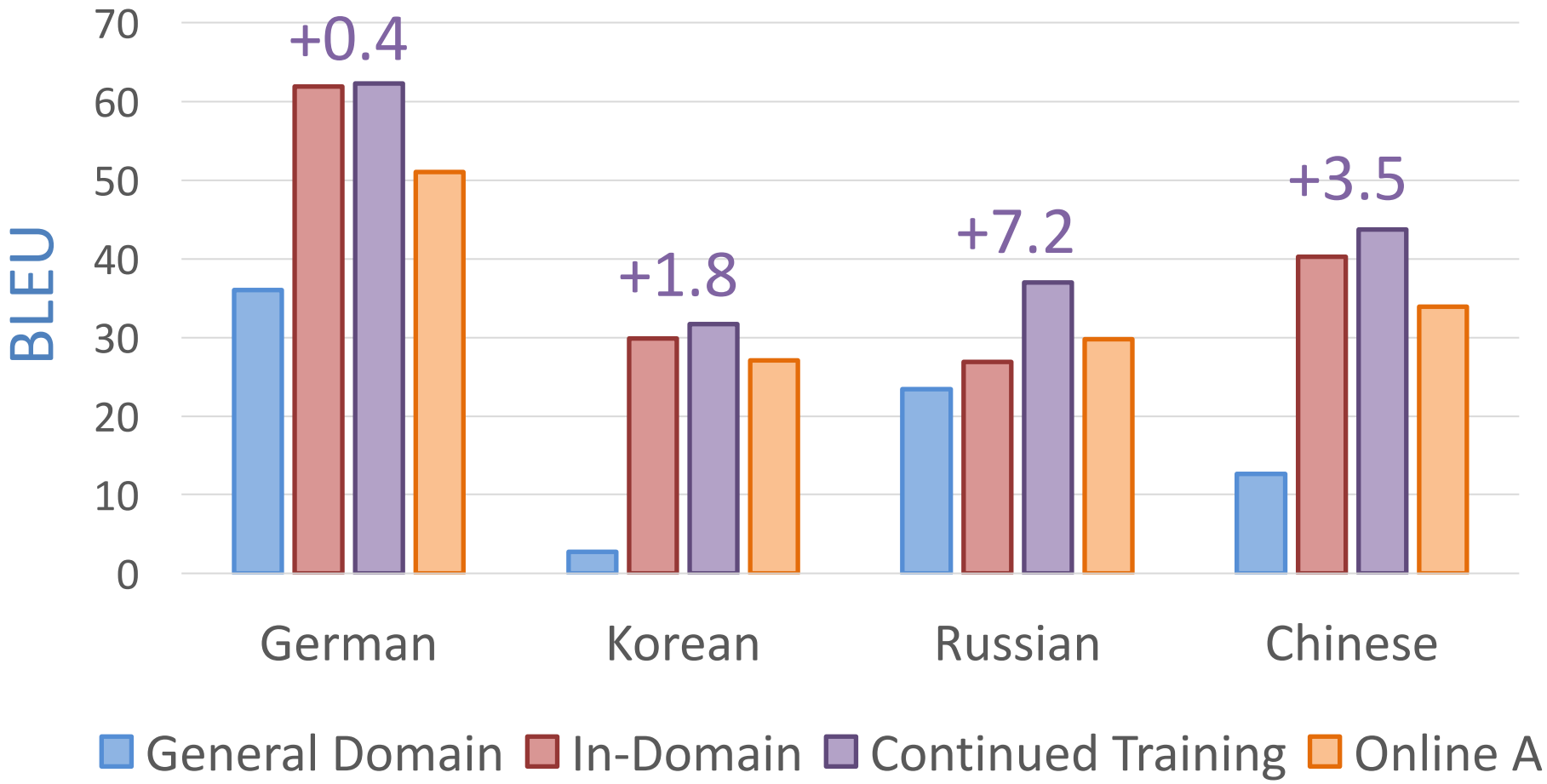
Patent Results



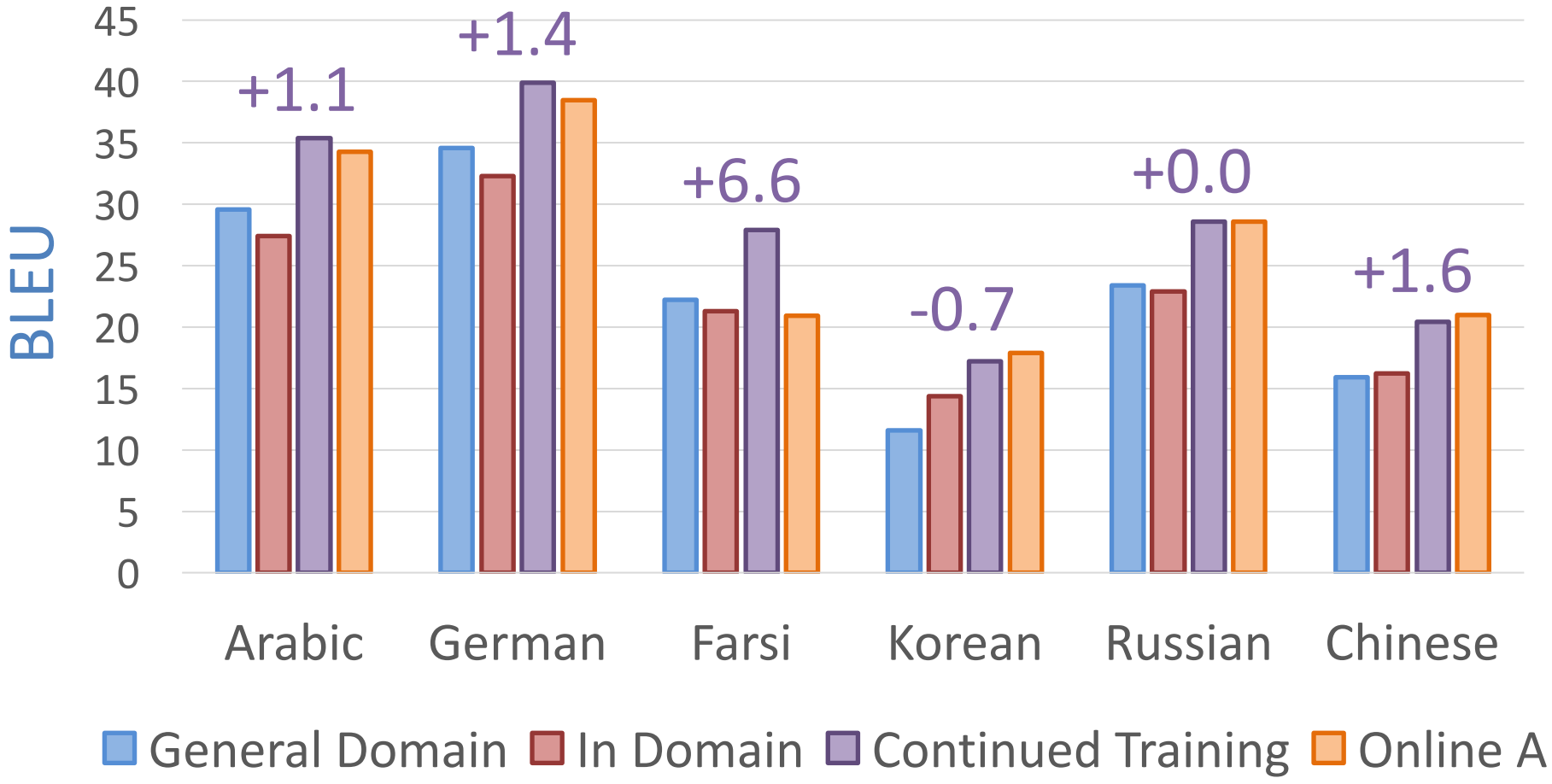
TED Results


BLEU
45
40
35
30
25
20
15
10
5
0
+1.1
+1.4
+7.0
-0.7
+0.0
-0.6
Arabic
German
Farsi
Korean
Russian
Chinese
SMT Domain Adapted
Online A
Continued Training

Patent Results



TED Results


BLEU
45
40
35
30
25
20
15
10
5
0
+1.1
+1.4
+6.6
-0.7
+0.0
+1.6
Arabic
German
Farsi
Korean
Russian
Chinese
General Domain
In Domain
Continued Training
Online A

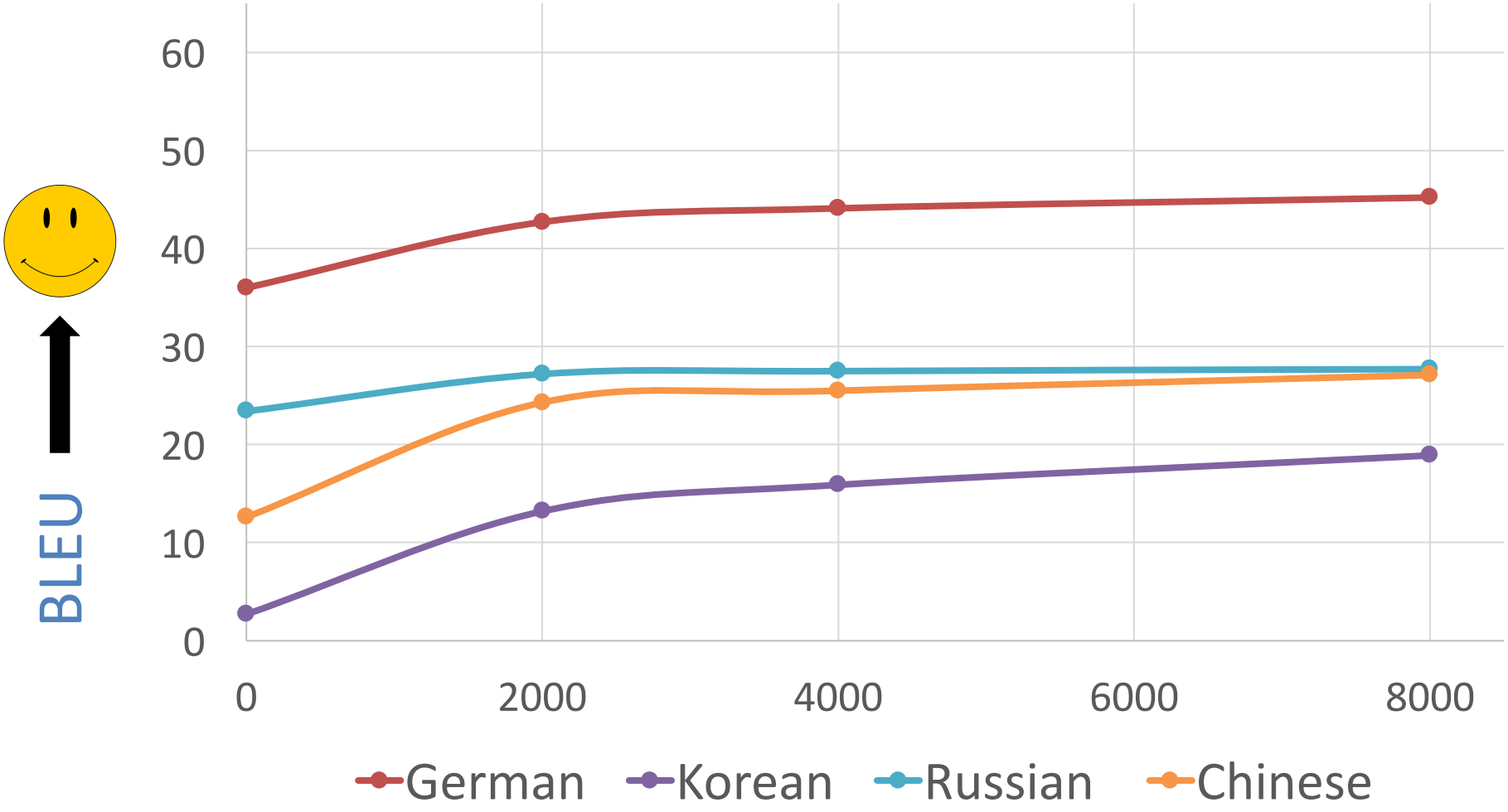
TED results

	Training data	Ar	De	Fa	Ko	Ru	Zh
SMT	General Domain	24.0	31.0	13.9	6.7	25.0	15.2
	Mixed Domain	27.8	31.9	18.2	10.7	25.7	16.1
NMT	General Domain	29.6	34.6	22.2	11.6	23.4	15.9
	In Domain (TED)	27.4	32.3	21.3	14.4	22.9	16.2
	Mixed Domain	---	35.6	---	---	24.5	17.8
	Continued Training	35.4	39.9	27.9	17.2	28.6	20.4
	Microsoft Translator	34.3	38.5	20.9	17.9	28.6	21.0

Patent results

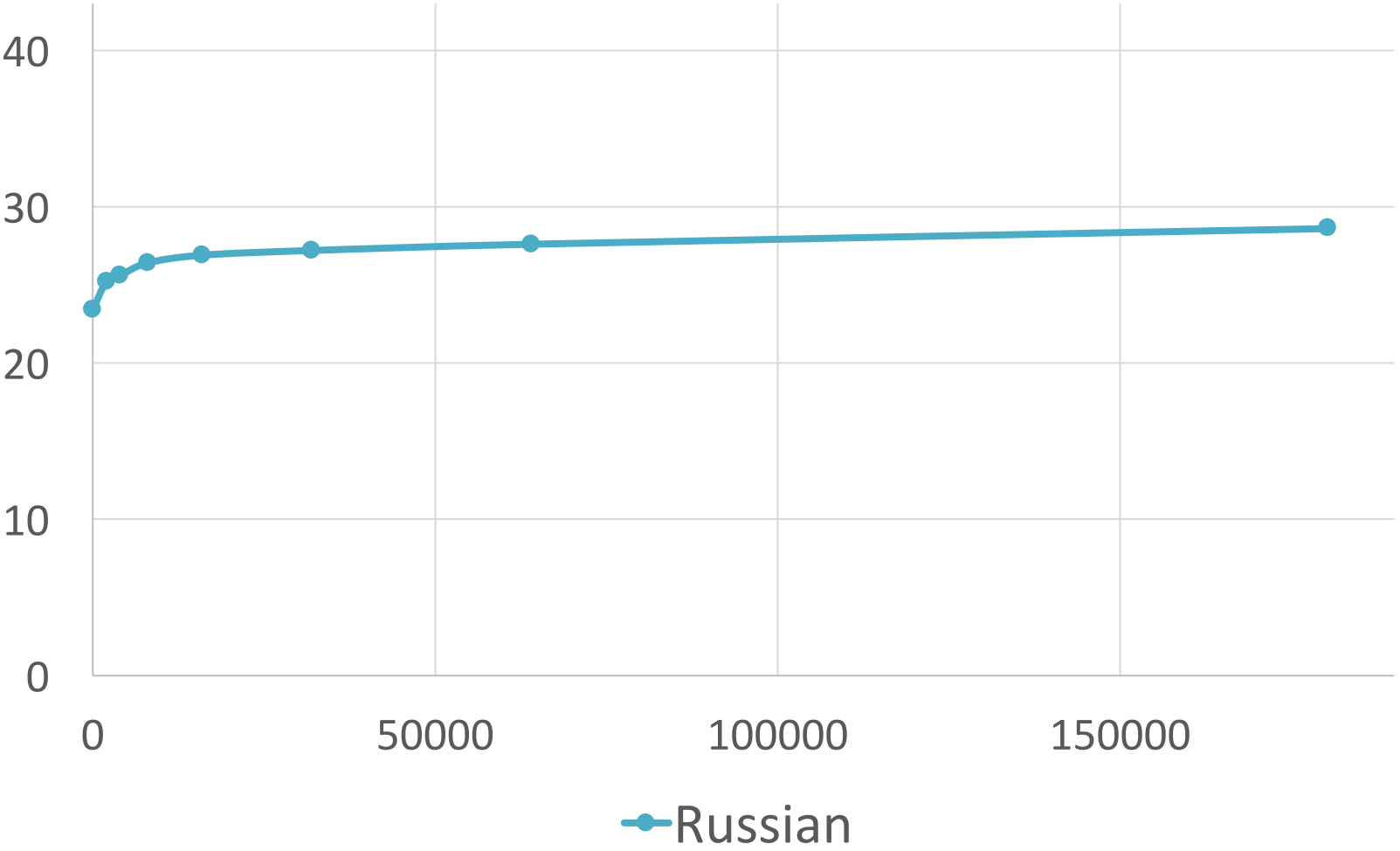
	Training data	De	Ko	Ru	Zh
SMT	General Domain	26.6	2.4	21.4	13.7
	Mixed Domain	50.6	21.7	29.0	29.8
NMT	General Domain	36.0	2.7	23.4	12.6
	In Domain (TED)	61.9	29.9	26.9	40.2
	Mixed Domain	58.4	---	27.7	33.7
	Continued Training	62.3	31.7	37.0	43.7
	Microsoft Translator	51.0	27.1	29.8	33.9

Patent


60
50
40
30
20
10
0
BLEU
0
2000
4000
6000
8000
German
Korean
Russian
Chinese

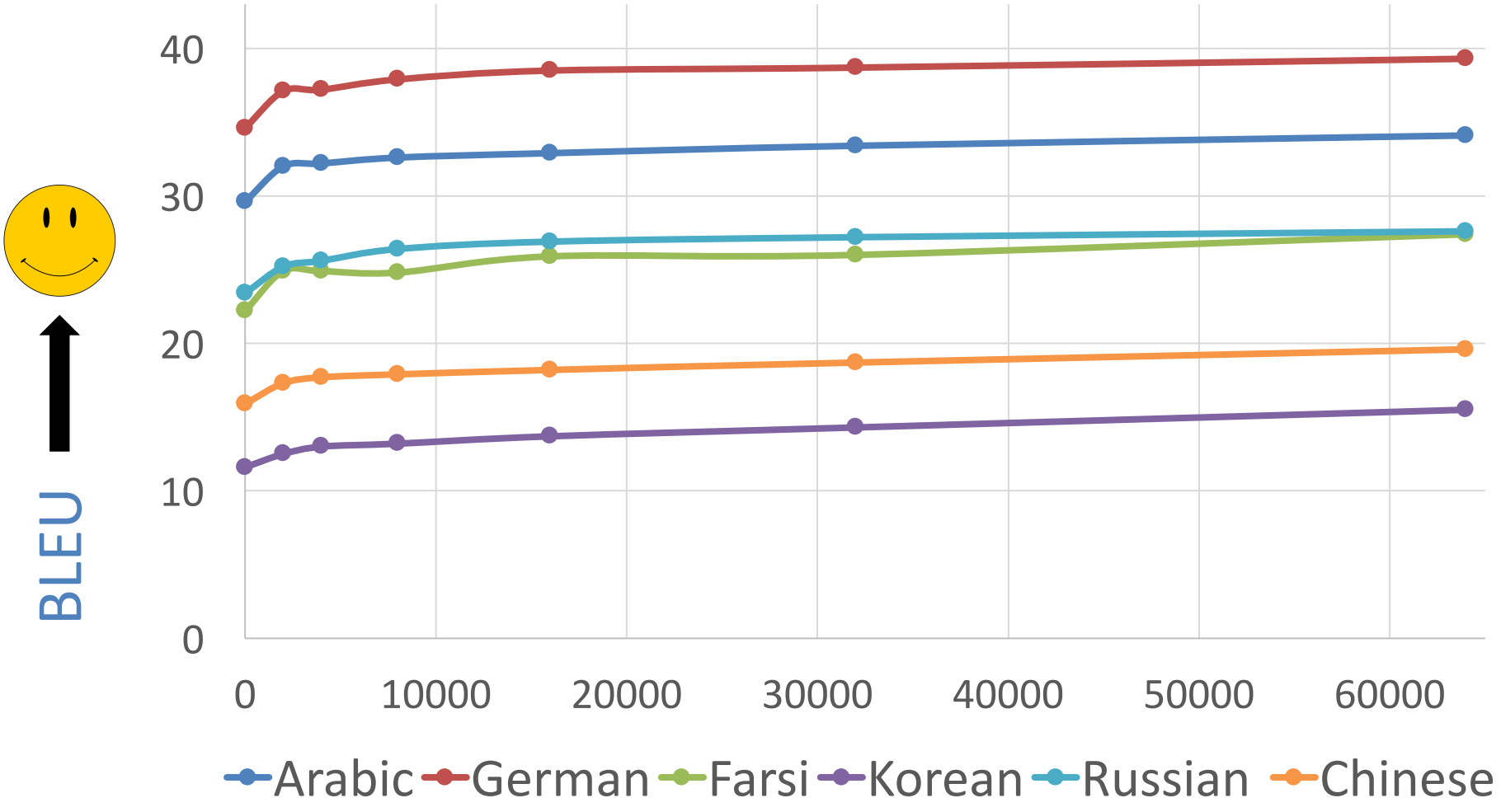
TED

BLEU

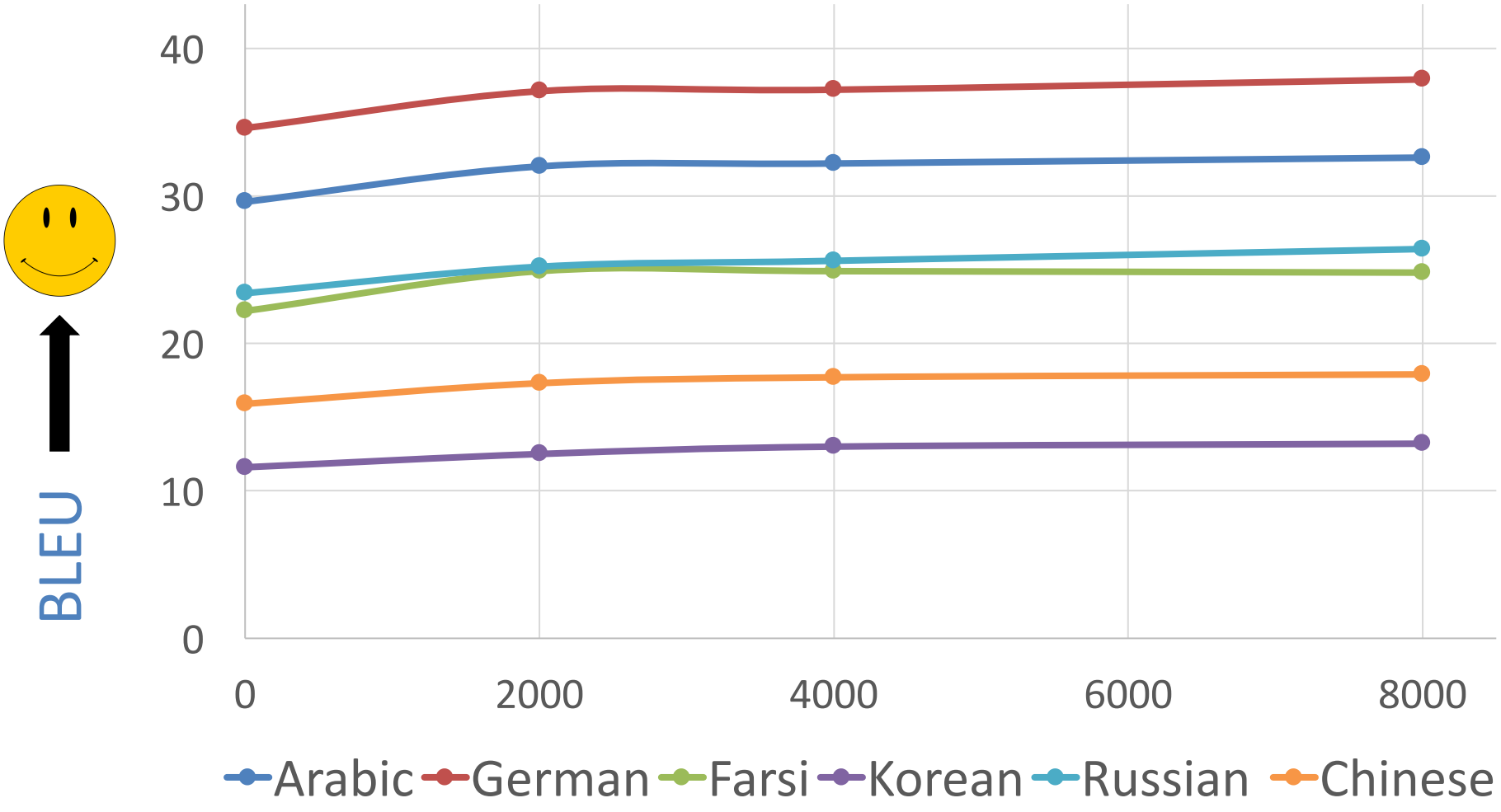


JOHNS HOPKINS
UNIVERSITY

TED


BLEU
40
30
20
10
0
0
10000
20000
30000
40000
50000
60000
Arabic
German
Farsi
Korean
Russian
Chinese

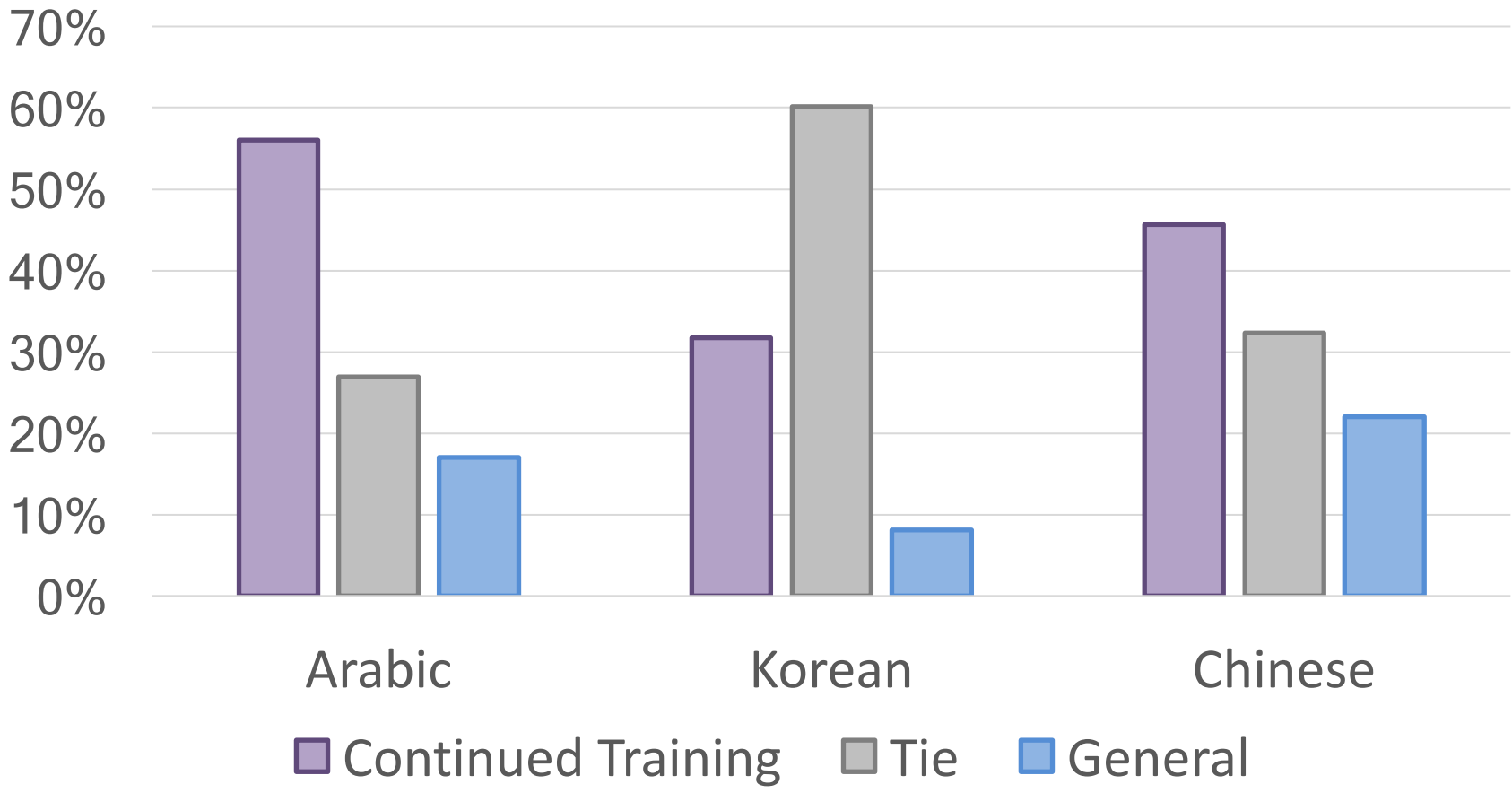
TED


BLEU
40
30
20
10
0
0
2000
4000
6000
8000
Arabic
German
Farsi
Korean
Russian
Chinese

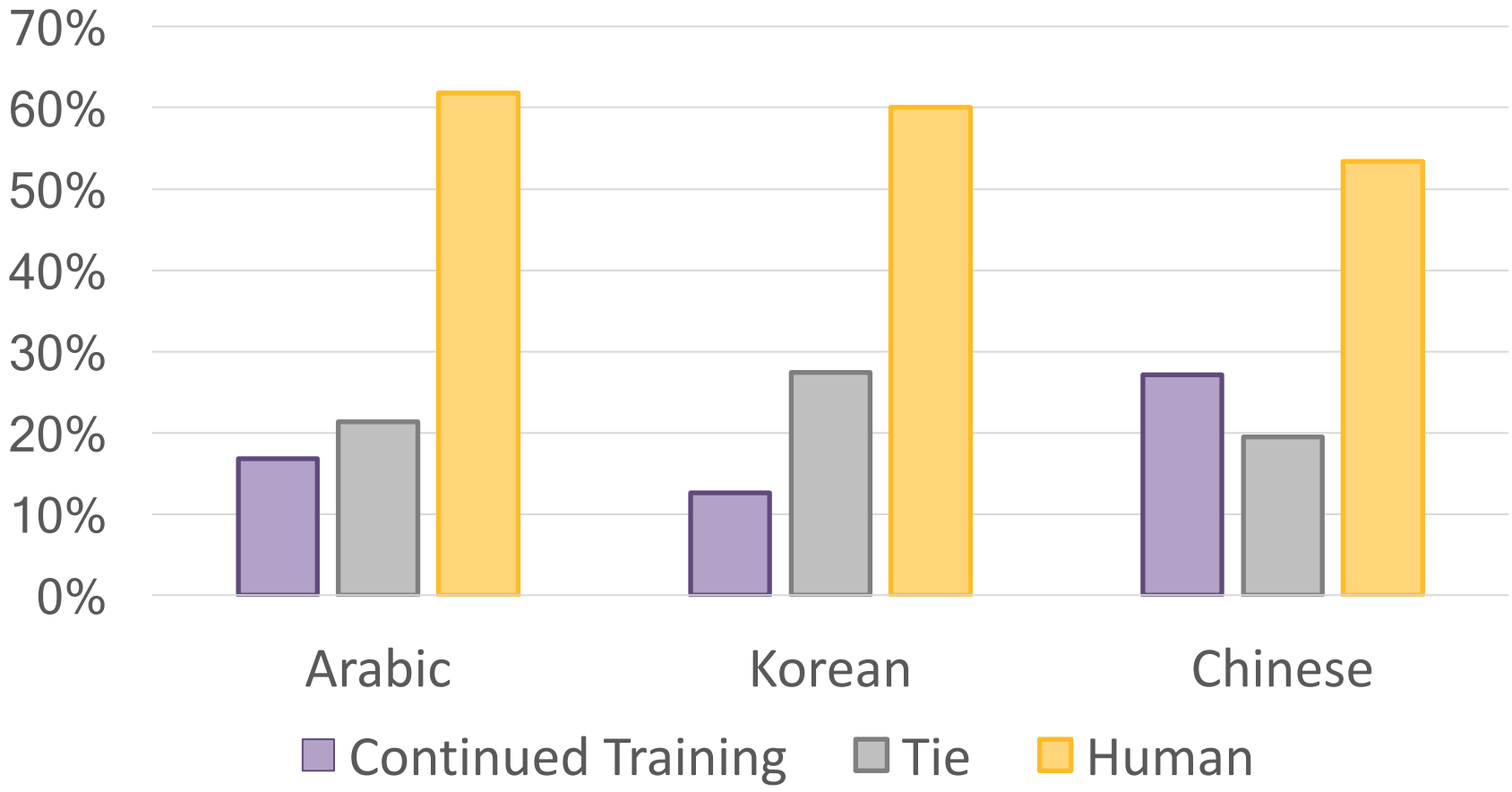
Human Eval

Source Language	System 1 (each line may be different system)	System 2 (each line may be different system)	Ranking	Comments
全医院的员工都认识这	all the staff of the hospital knew this brave young man .	the hospital staff will know this brave young volunteers	System 2 is be	Poor plural agreement,
七个孩子的老幺必须想	the old age of seven children has to figure out how to fit ir	seven children in two crowded must find ways to go to th	Both translati	Equally bad.
克里斯·安德森 ： 没有。	ca : no . do n't start the clock .	chris anderson : no , you ca n't start the three minutes .	System 1 is better	
重设时间， 负责对她不	reset the time for being fair with her .	reset the three minutes , that 's just not fair .	System 2 is better	
有三种可以影响脑部的	there are three ways to influence the brain : through psyc	there are three methods can affect brain : the therapist tr	System 2 is be	Though poor grammar.
因此他说 ： " 艾莉森， 我	so he said , " jason , we 've got to get you surgery . "	so he said , " arson , we have to help you . "	Both translati	Neither gets the name r
不过我已经没有五周前那	but i am not as nervous as i was five weeks ago .	but i did n't feel nervous about five weeks ago .	System 1 is better	
五周以前我做了全髋骨	five weeks ago i had a hip replacement surgery .	five weeks ago i had total hip replacement surgery .	System 2 is better	
你知道这个手术吗 ？	do you know this surgery ?	do you know that surgery ?	Both translations are about the same	
电锯， 电钻， 超可怕 除	electric saw , power drill , totally disgusting unless you 're	chain saw , electric drill , david , unless you 're super scary	System 1 is better	
当然， 大卫， 只要不是	sure david , if it 's not your hip , it 's truth and beauty .	of course , david , as long as it 's not your bones , is real ai	System 1 is better	
总之， 这件事给了我很	anyway , this gave me a big revelation , chris invited me to	anyway , this gave me a lot of insights , chris invited me to	Both translations are about the same	
但首先我有两件事需要	but first of all , i have a list of things that need to be said .	but first you need to know two things about me .	System 2 is better	
就两件	on two .	two .	System 2 is better	
第一， 我是加拿大人。	first , i 'm canadian . second , i 'm the youngest of seven ch	first , i 'm canadian . second , i was the youngest of seven	System 1 is better	
在加拿大， 我们有很好	now , in canada , we have that great healthcare system .	in canada , we have a very good healthcare system .	System 2 is better	
那意味著置换髋骨是免	that means we get our new hips for free .	that means that the bone changes the bone for free .	System 1 is better	
而身为七个小孩中的老	and being the youngest of seven , i have never been at the	and as the seven children in two everything 's the last one	System 1 is better	
我的髋骨已经折磨了我	my legs suffered for many years .	my bones have been torturing me for years .	System 1 is better	
我终于去看了医生， 那	i finally went to the doctor , which was free .	i finally went to the doctor , it 's free .	System 1 is better	

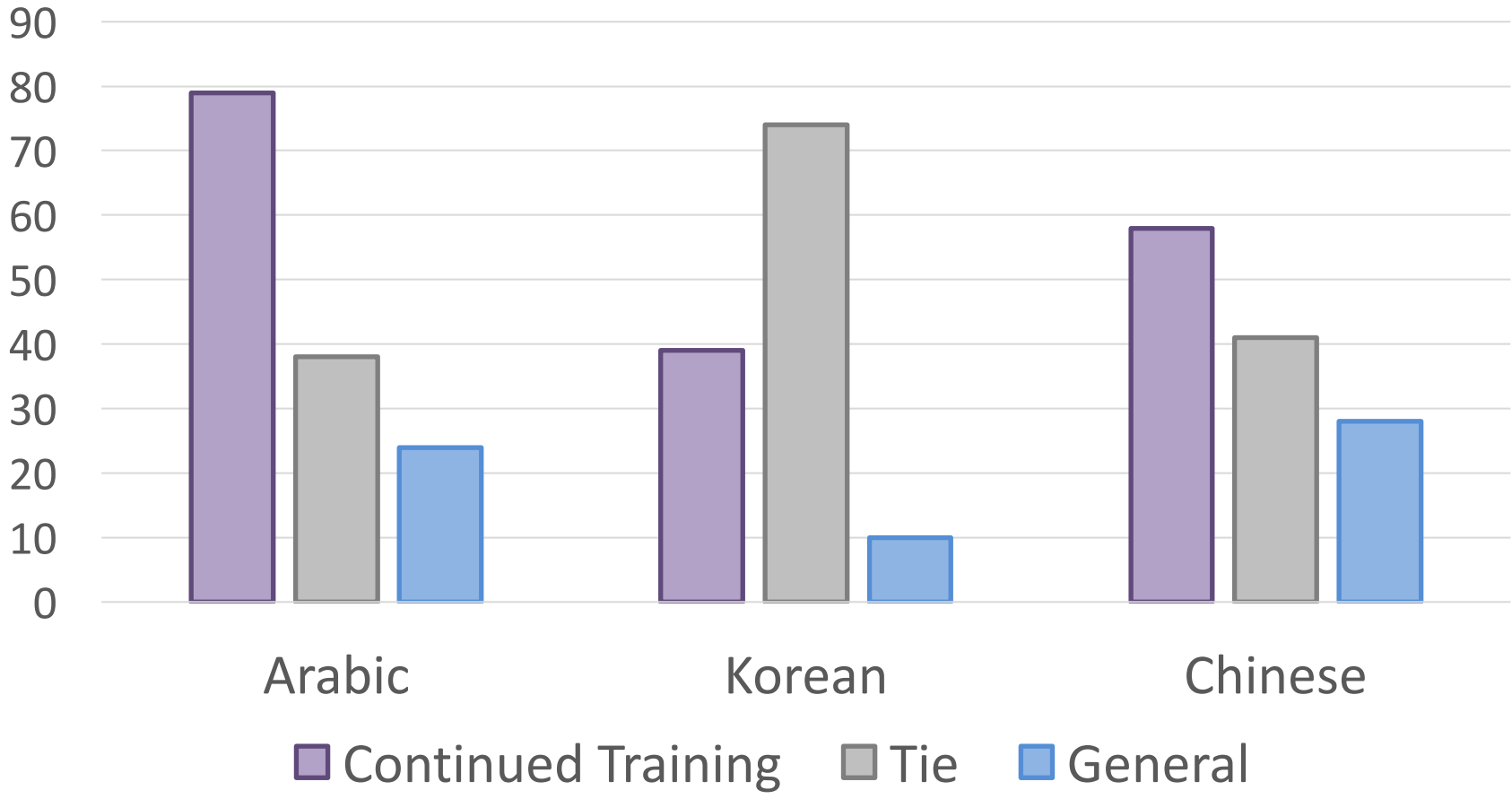
Continued Training vs General



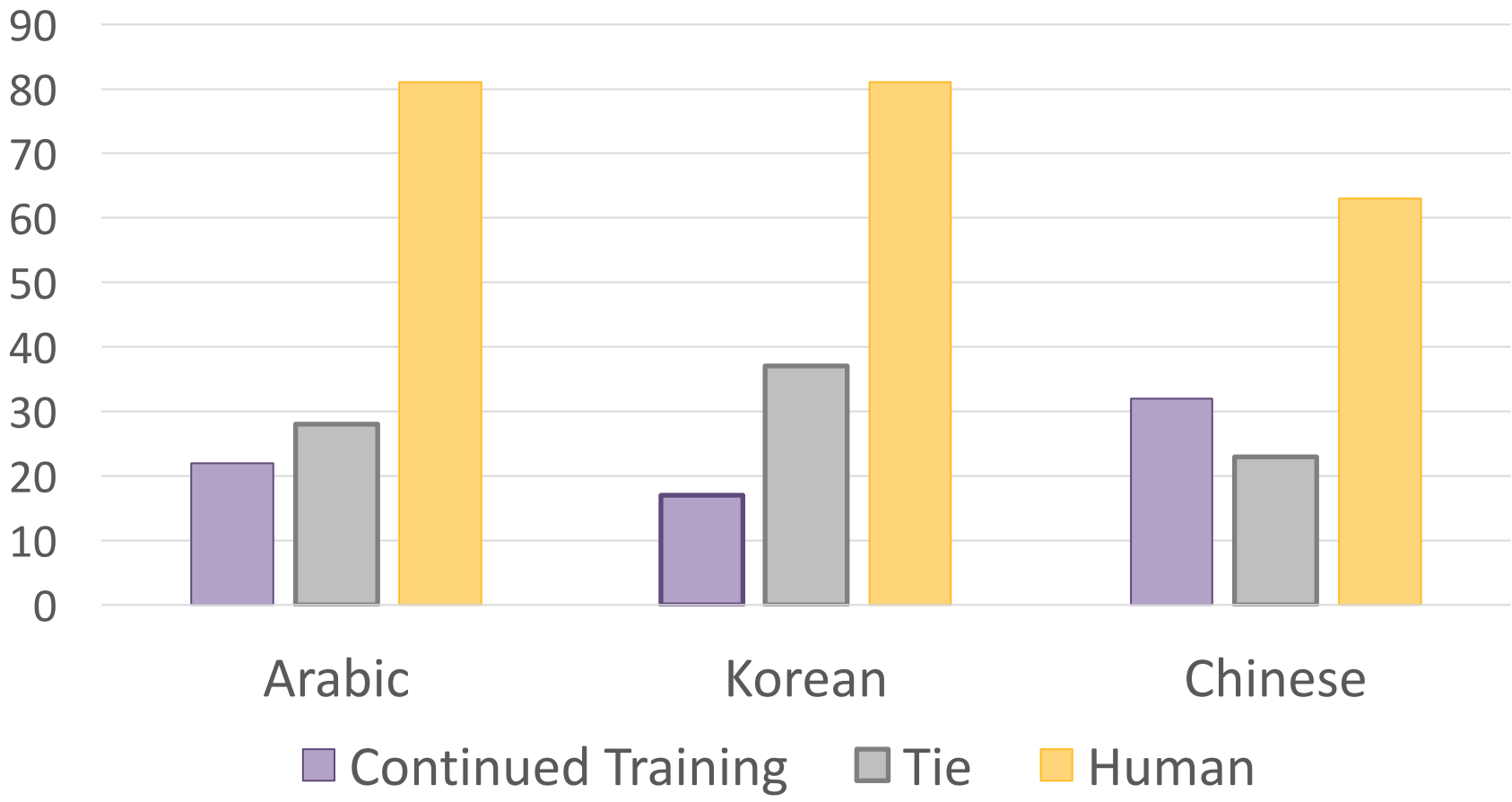
Continued Training vs Human



Continued Training vs General



Continued Training vs Human



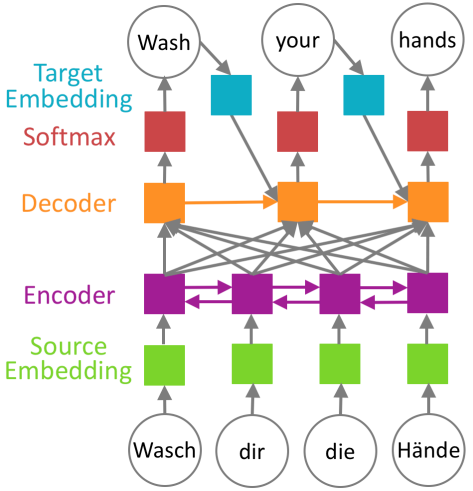
Human Evaluation

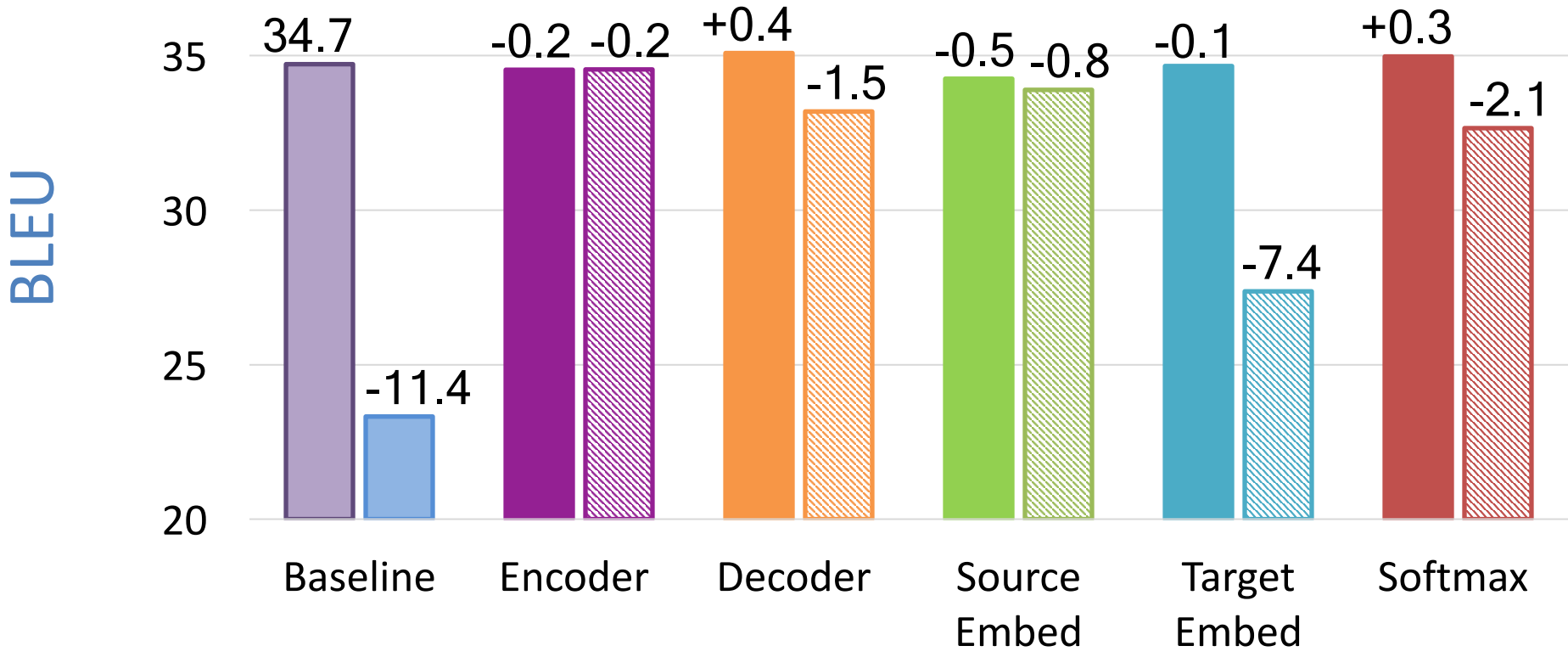
Trends similar across three languages

- System differences consistent with BLEU
- Human reference (unsurprisingly) better

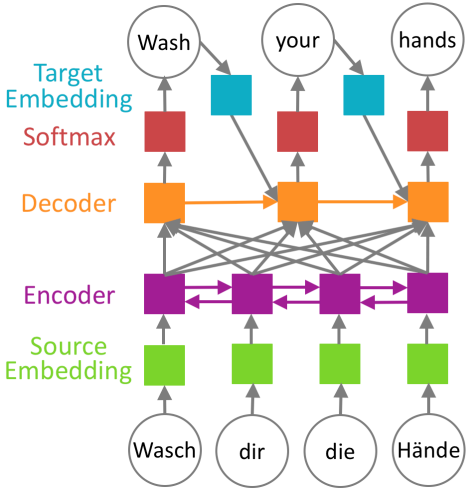
Research Extensions

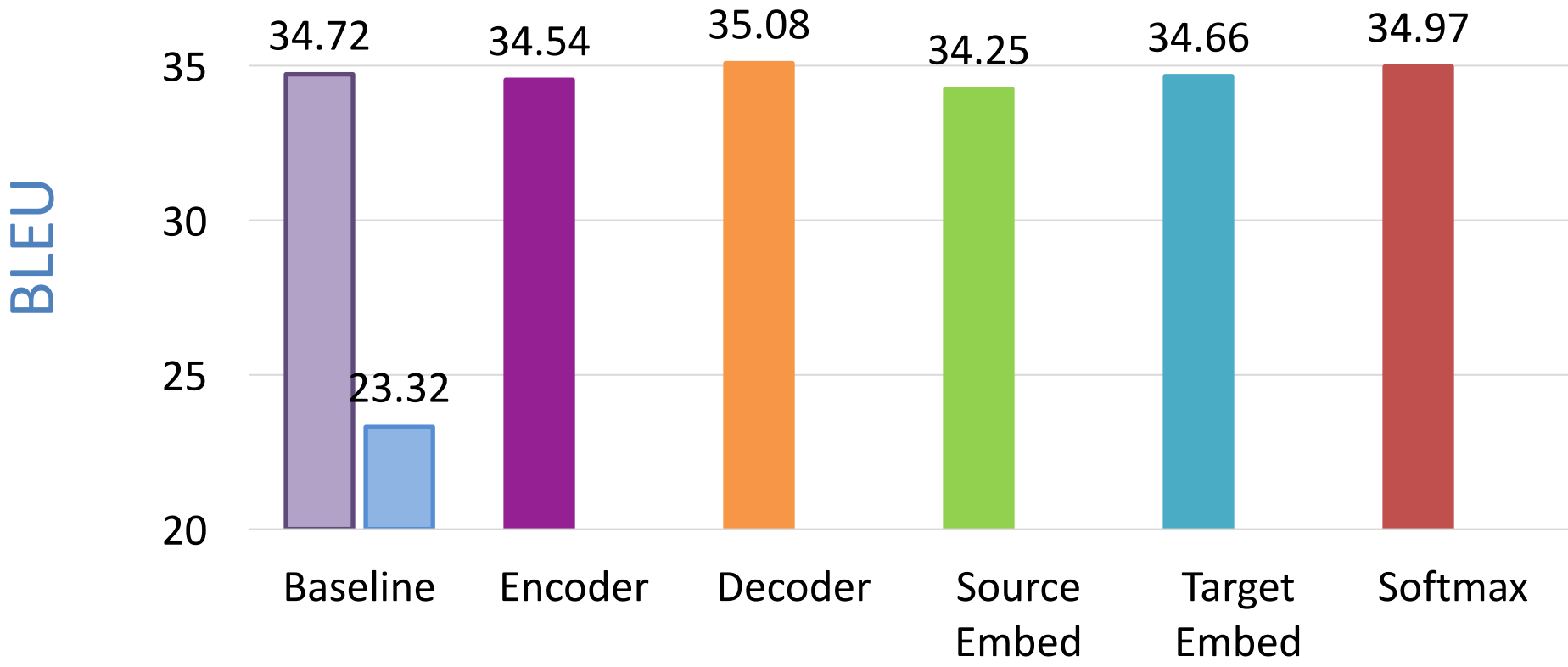
Parameter Freezing



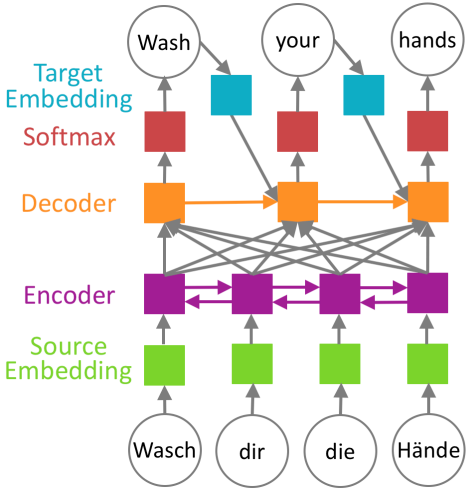


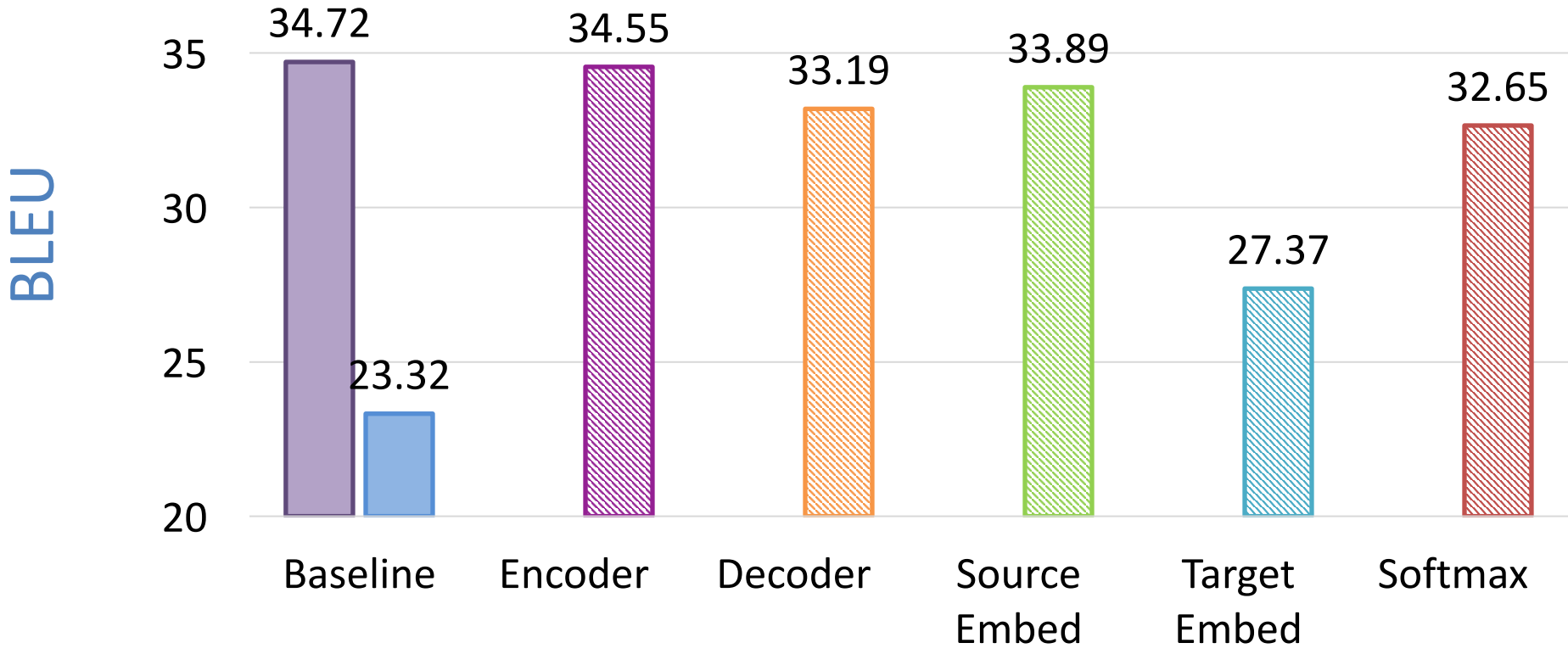
Parameter Freezing



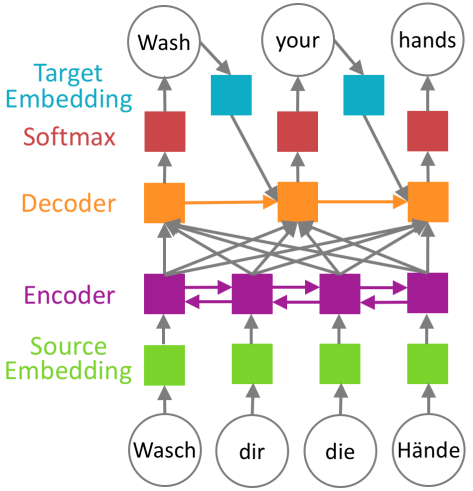


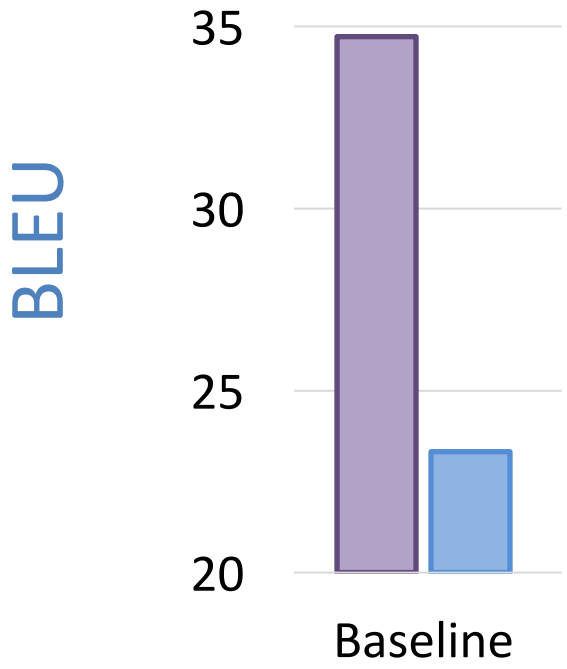
Parameter Freezing



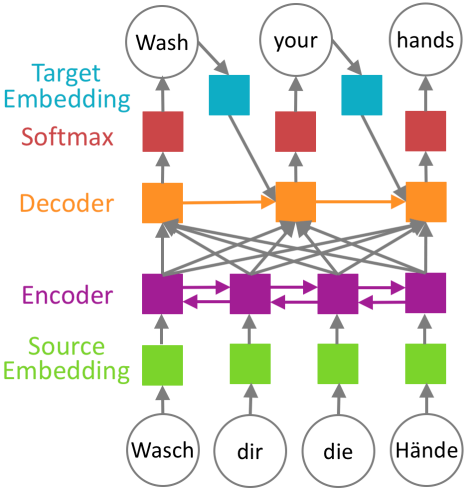


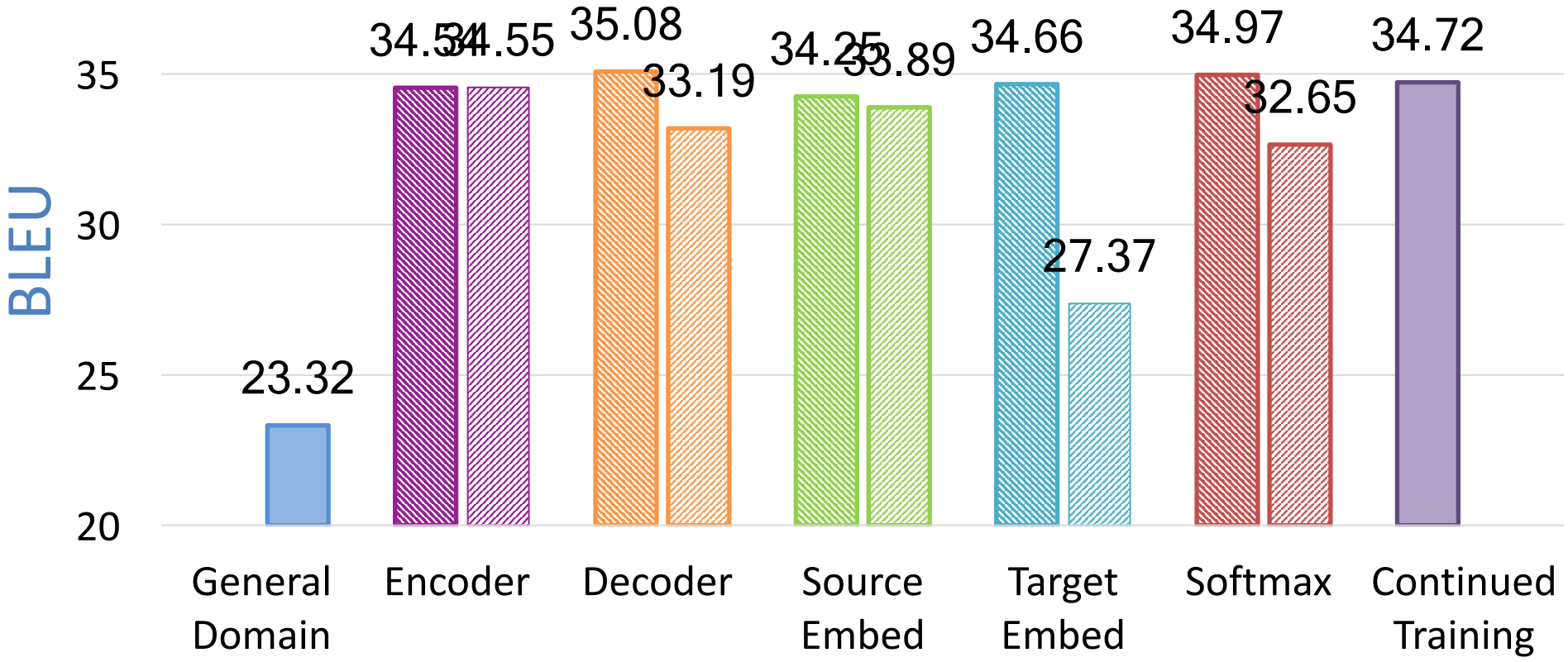
Selective Training



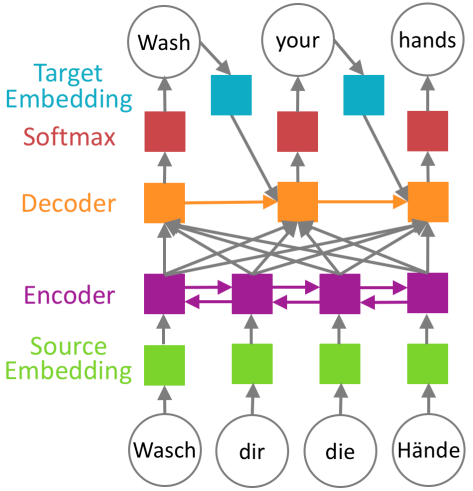


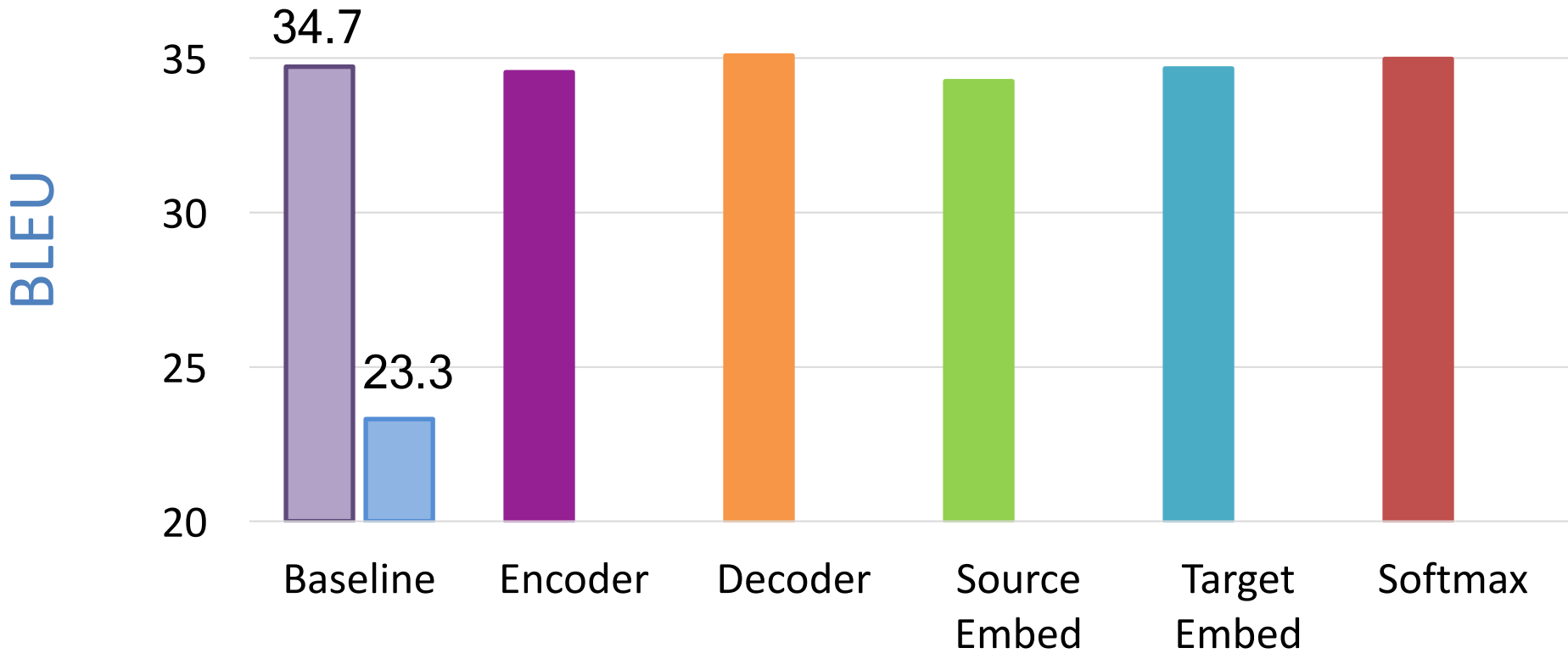
Selective Training of Components





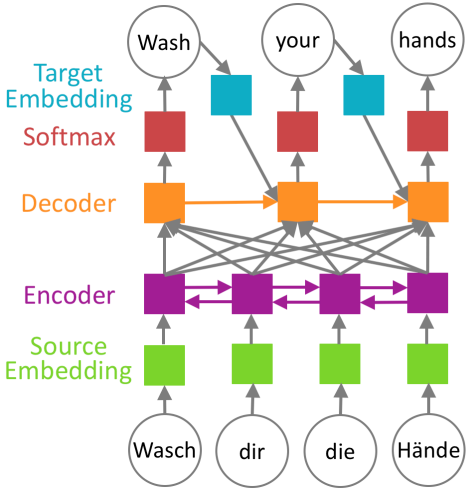
Selective Training

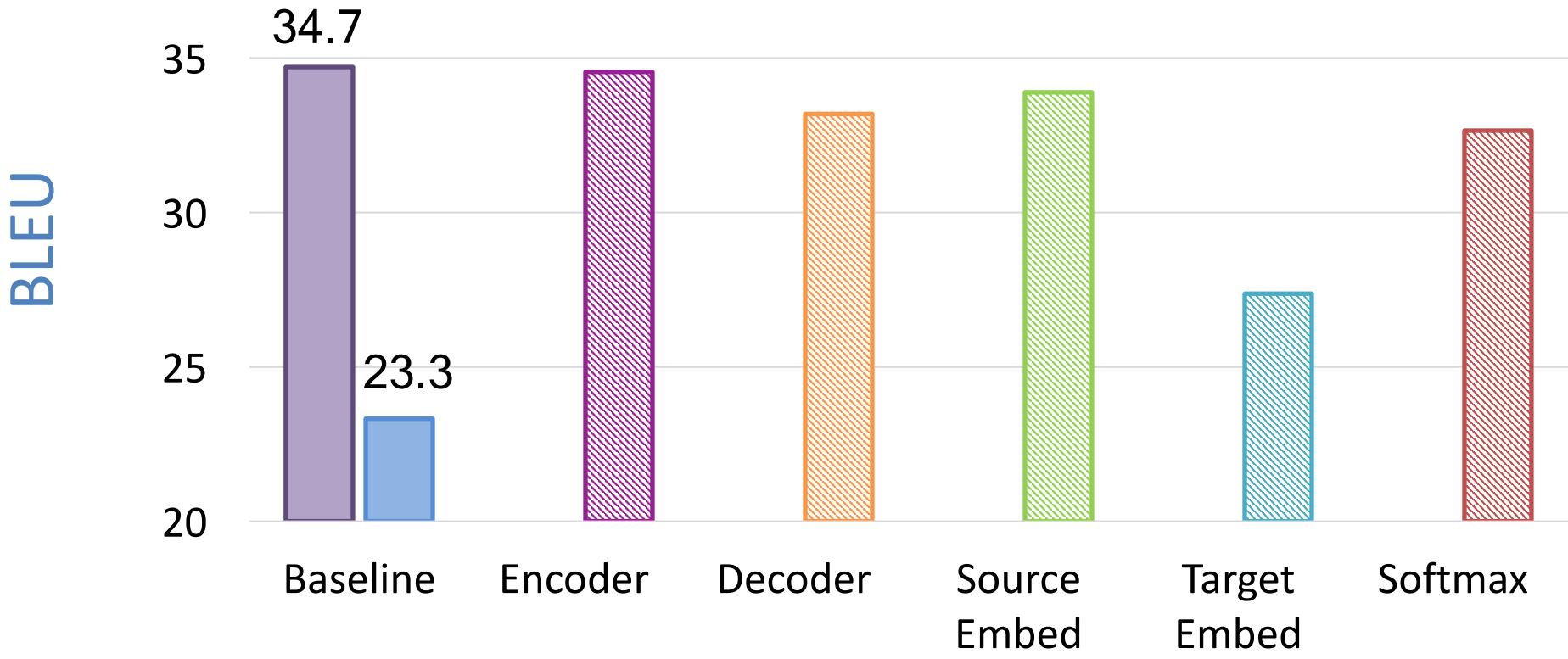




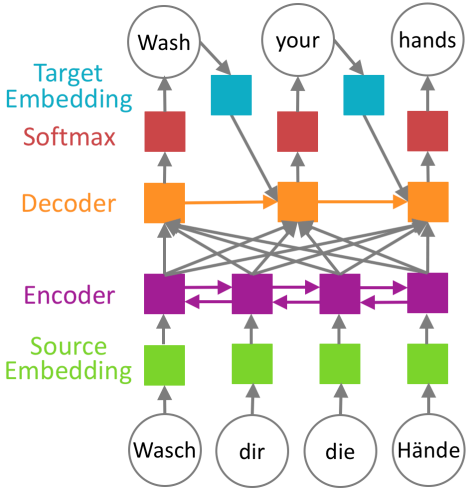
JOHNS HOPKINS UNIVERSITY

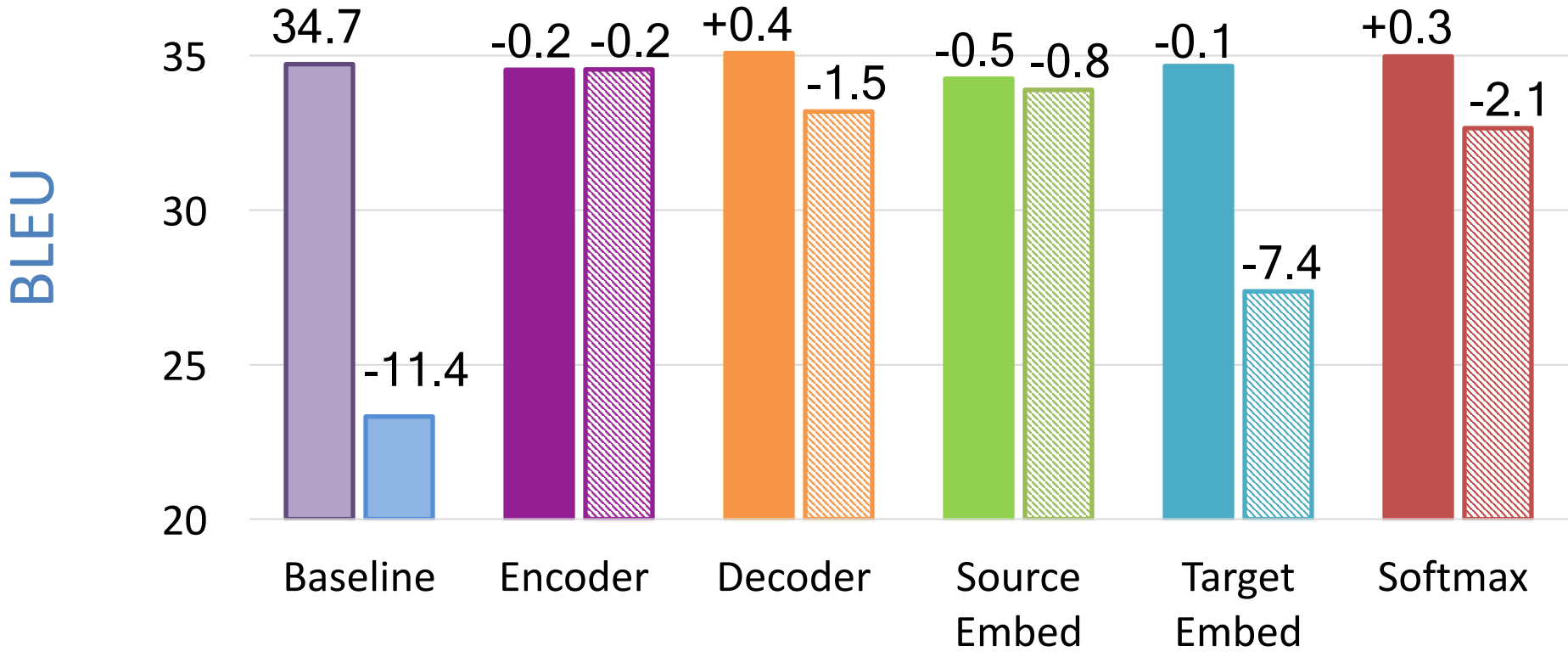
Selective Training





Selective Training





Data Sizes

Datasets

 # sentence/segment pairs

	Ar-En	De-En	Fa-En	Ko-En	Ru-En	Zh-En
Large General Domain	49M Subtitle, UN, LDC	28M Subtitle, WMT	6M Subtitle	1M Subtitle	51M Subtitle, WMT	36M Subtitle,WMT
TED Talks	175k	152k	114k	164k	180k	170k
Patent (WIPO)		821k		81k	39k	154k

in-domain sets

Goal: Improve test results on TED/Patent using both Large General Domain and some In-Domain data

Data

Training data	Arabic	German	Farsi	Korean	Russian	Chinese
General Domain	49M Subtitle, UN, LDC	28M Subtitle, WMT	6M Subtitle	1M Subtitle	51M Subtitle, WMT	36M Subtitle, WMT
TED	175k	152k	114k	164k	180k	170k
Patent	---	821k	---	81k	39k	154k

TED Data

Training data	Arabic	German	Farsi	Korean	Russian	Chinese
General Domain	49M Subtitle, UN, LDC	28M Subtitle, WMT	6M Subtitle	1M Subtitle	51M Subtitle, WMT	36M Subtitle, WMT
In Domain (TED)	175k	152k	114k	164k	180k	170k

So, um... she 's kidding...

Resumption of the session

The European Union supports humanitarian action.

Allison Hunt: My three minutes hasn't started yet, has it?

Patent Data

Training data	German	Korean	Russian	Chinese
General Domain	28M Subtitle, WMT	1M Subtitle	51M Subtitle, WMT	36M Subtitle, WMT
Patent	821k	81k	39k	154k

So, um... she 's kidding...
Resumption of the session
The European Union supports humanitarian action.

The tablets exhibit improved bioavailability of the active ingredient.

OOV rates

TED OOVs (type count)

Training data	Arabic	German	Farsi	Korean	Russian	Chinese
General Domain	133	204	445	225	140	45
In Domain (TED)	745	700	758	249	813	422
Both domains	126	193	329	133	132	43
Total types	8248	5837	6261	4989	7954	5760

TED OOVs (token count)

Training data	Arabic	German	Farsi	Korean	Russian	Chinese
General Domain	176	235	597	316	153	47
In Domain (TED)	840	809	956	327	933	536
Both domains	168	221	418	187	143	45
Total tokens	28636	35209	39223	45715	31575	33397

TED OOVs (type %)

Training data	Arabic	German	Farsi	Korean	Russian	Chinese
General Domain	1.61%	3.49%	7.11%	4.51%	1.76%	0.78%
In Domain (TED)	9.03%	11.99%	12.11%	4.99%	10.22%	7.33%
Both domains	1.53%	3.31%	5.25%	2.67%	1.66%	0.75%

TED OOVs (token %)

Training data	Arabic	German	Farsi	Korean	Russian	Chinese
General Domain	0.61%	0.67%	1.52%	0.69%	0.48%	0.14%
In Domain (TED)	2.93%	2.30%	2.44%	0.72%	2.95%	1.60%
Both domains	0.59%	0.63%	1.07%	0.41%	0.45%	0.13%

Patent OOVs (type count)

Training data	German	Korean	Russian	Chinese
General Domain	5290	2098	1508	495
In Domain (TED)	2331	986	4286	1085
Both domains	2100	594	1262	339
Total types	14566	7939	15964	8627

Patent OOVs (token count)

Training data	German	Korean	Russian	Chinese
General Domain	10264	7748	1980	1171
In Domain (TED)	3864	1724	5715	2061
Both domains	3528	1045	1617	681
Total types	132208	186832	81911	135591

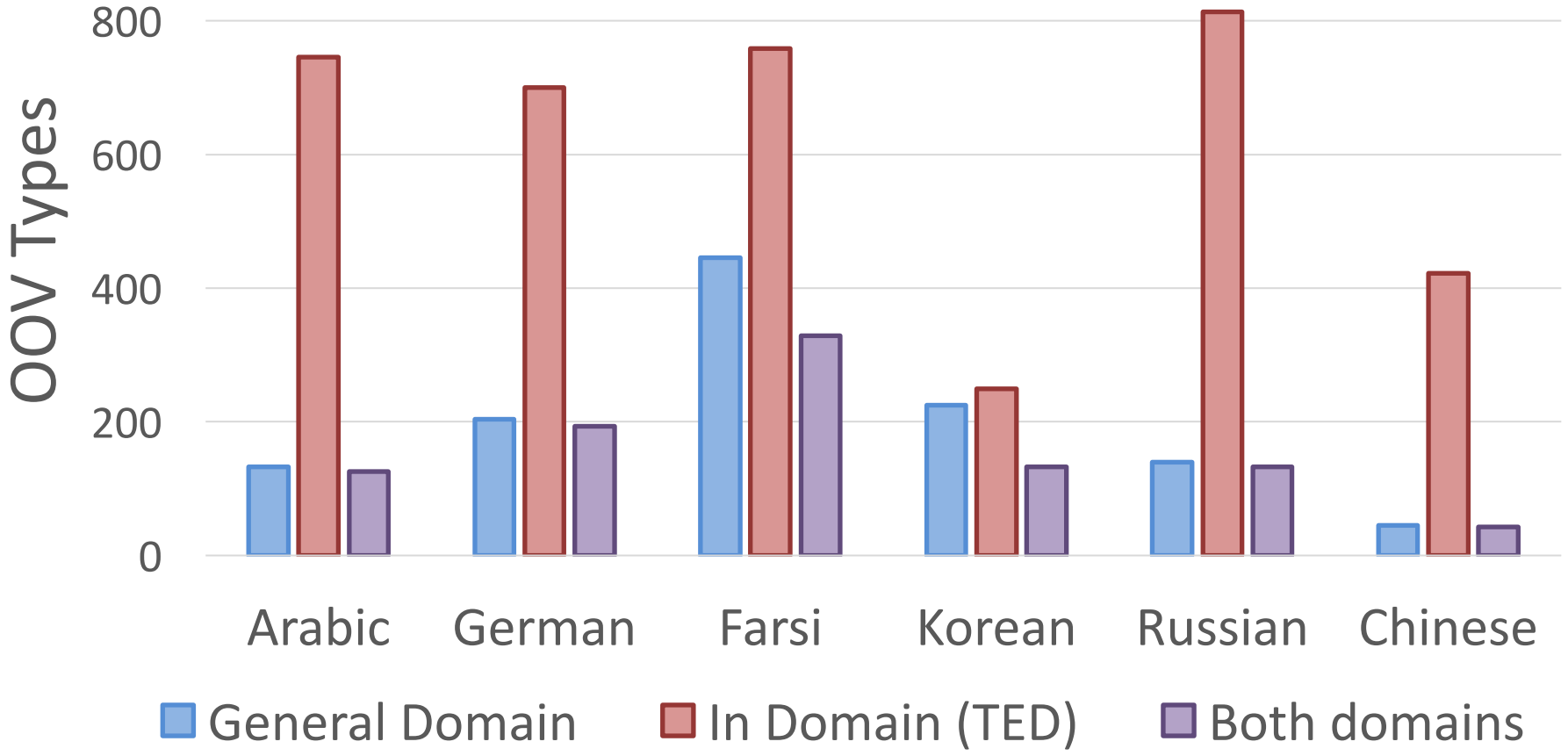
Patent OOVs (type %)

Training data	German	Korean	Russian	Chinese
General Domain	36.32%	26.43%	9.45%	5.74%
In Domain (Patent)	16.00%	12.42%	26.85%	12.58%
Both domains	14.42%	7.48%	7.91%	3.93%

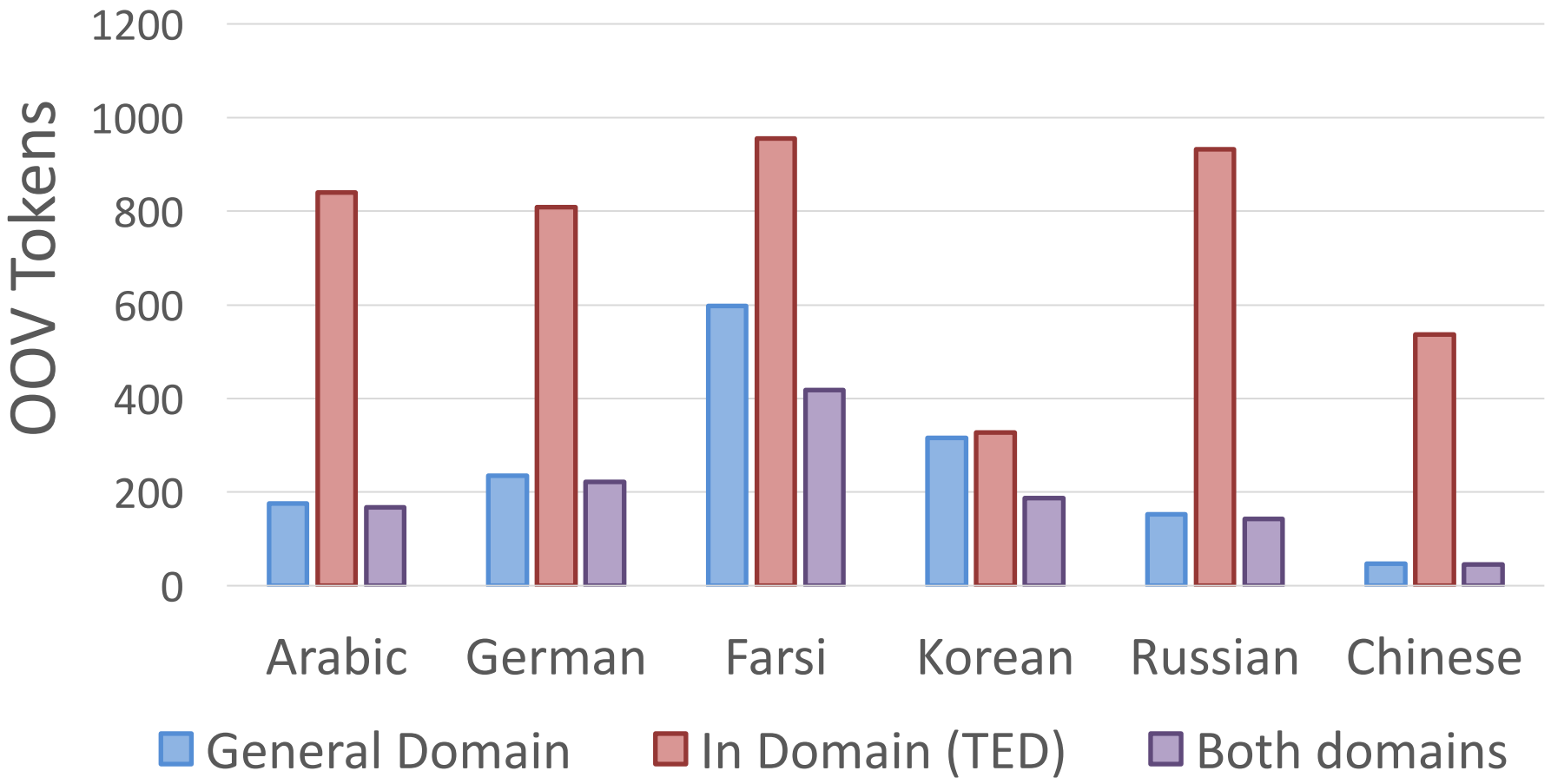
Patent OOVs (type %)

Training data	German	Korean	Russian	Chinese
General Domain	7.76%	4.15%	2.42%	0.86%
In Domain (Patent)	2.92%	0.92%	6.98%	1.52%
Both domains	2.67%	0.56%	1.97%	0.50%

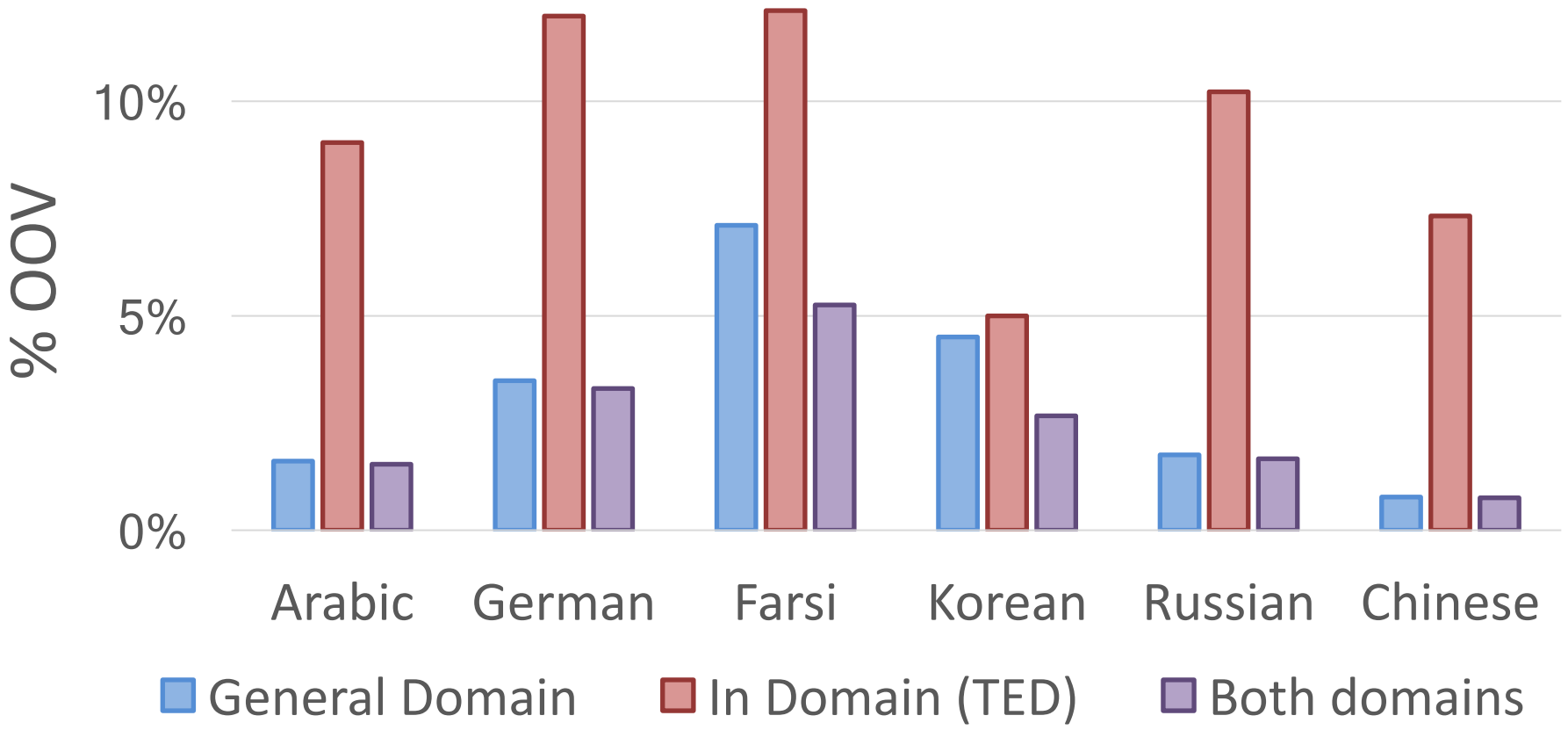
TED OOVs (Type Count)



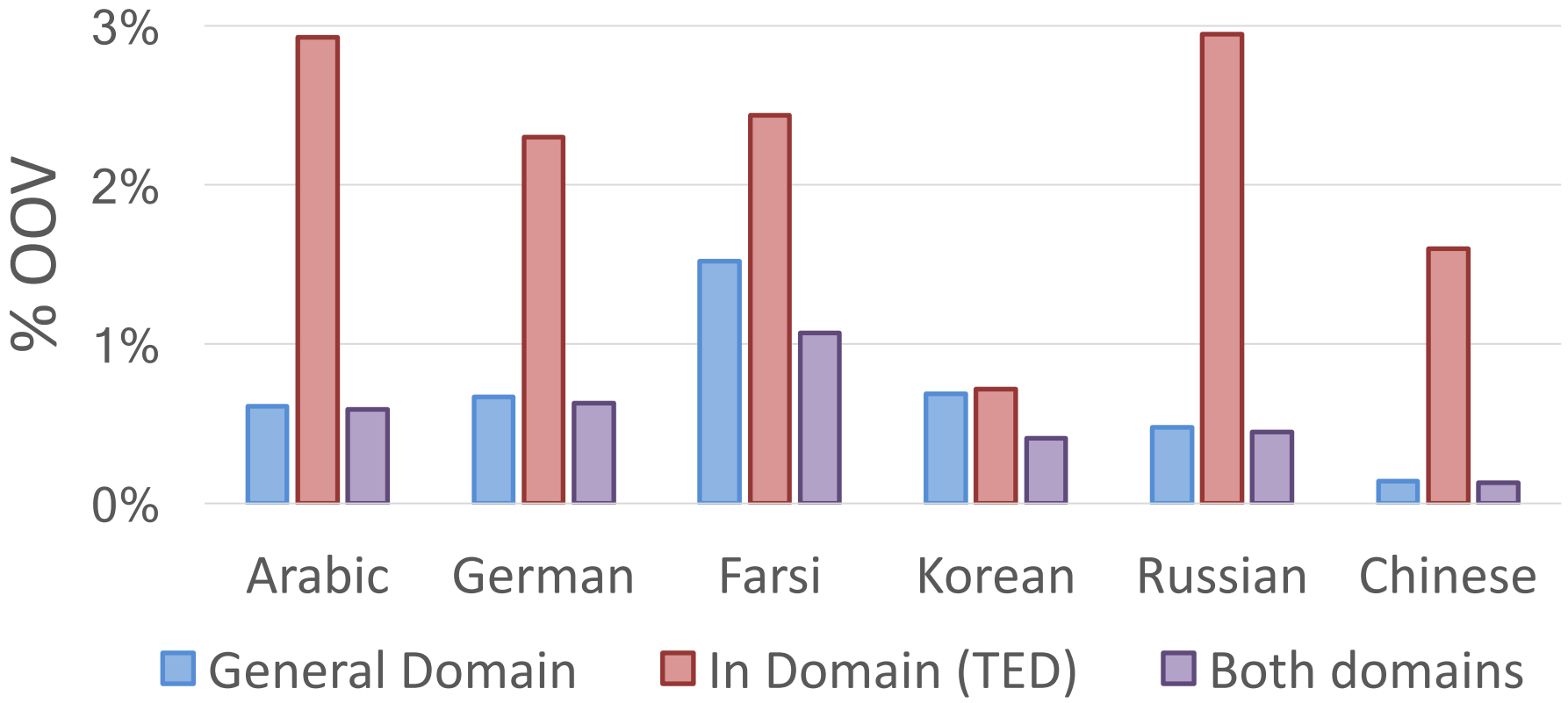
TED OOVs (Token Count)



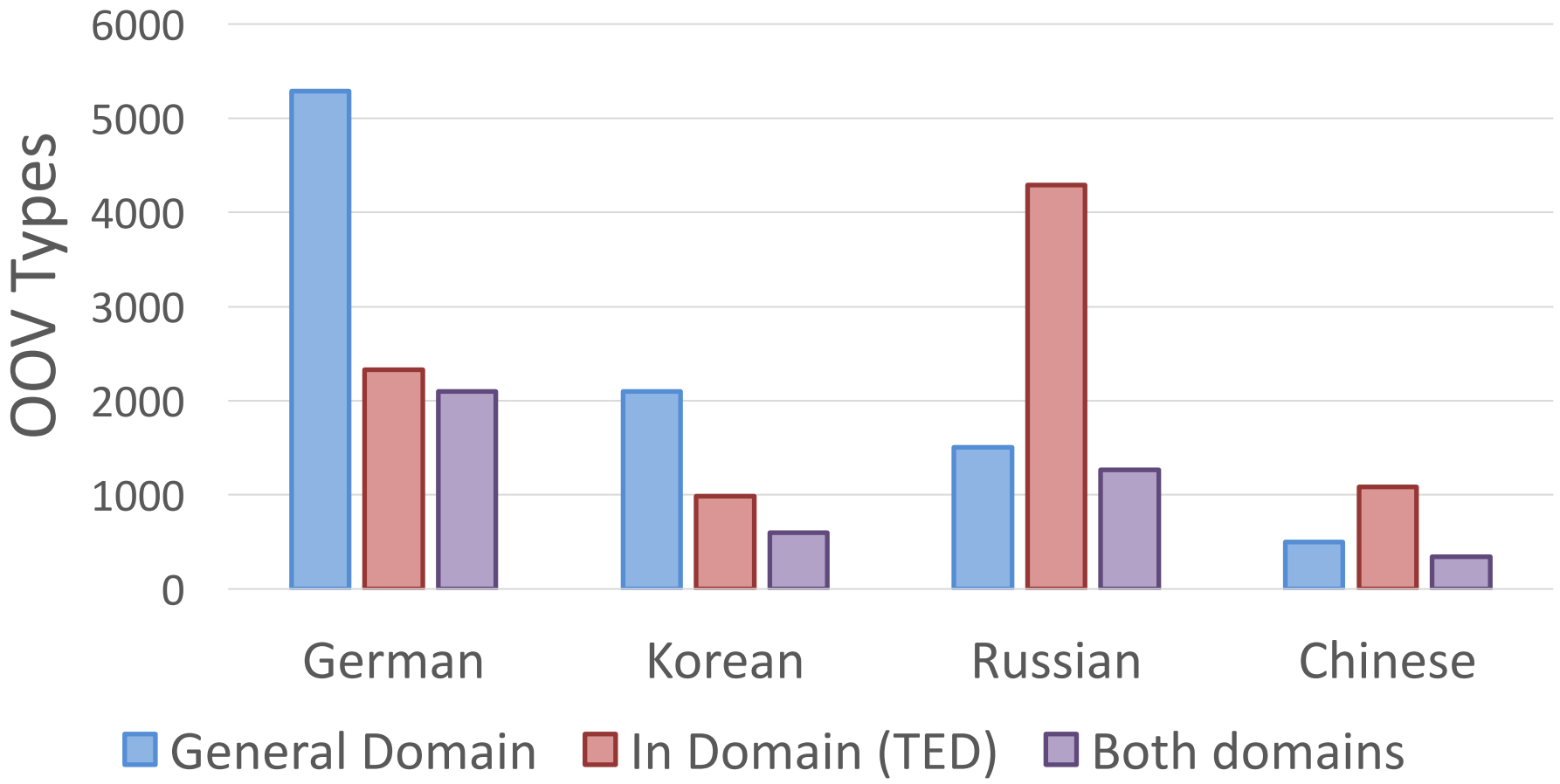
TED OOVs (Type %)



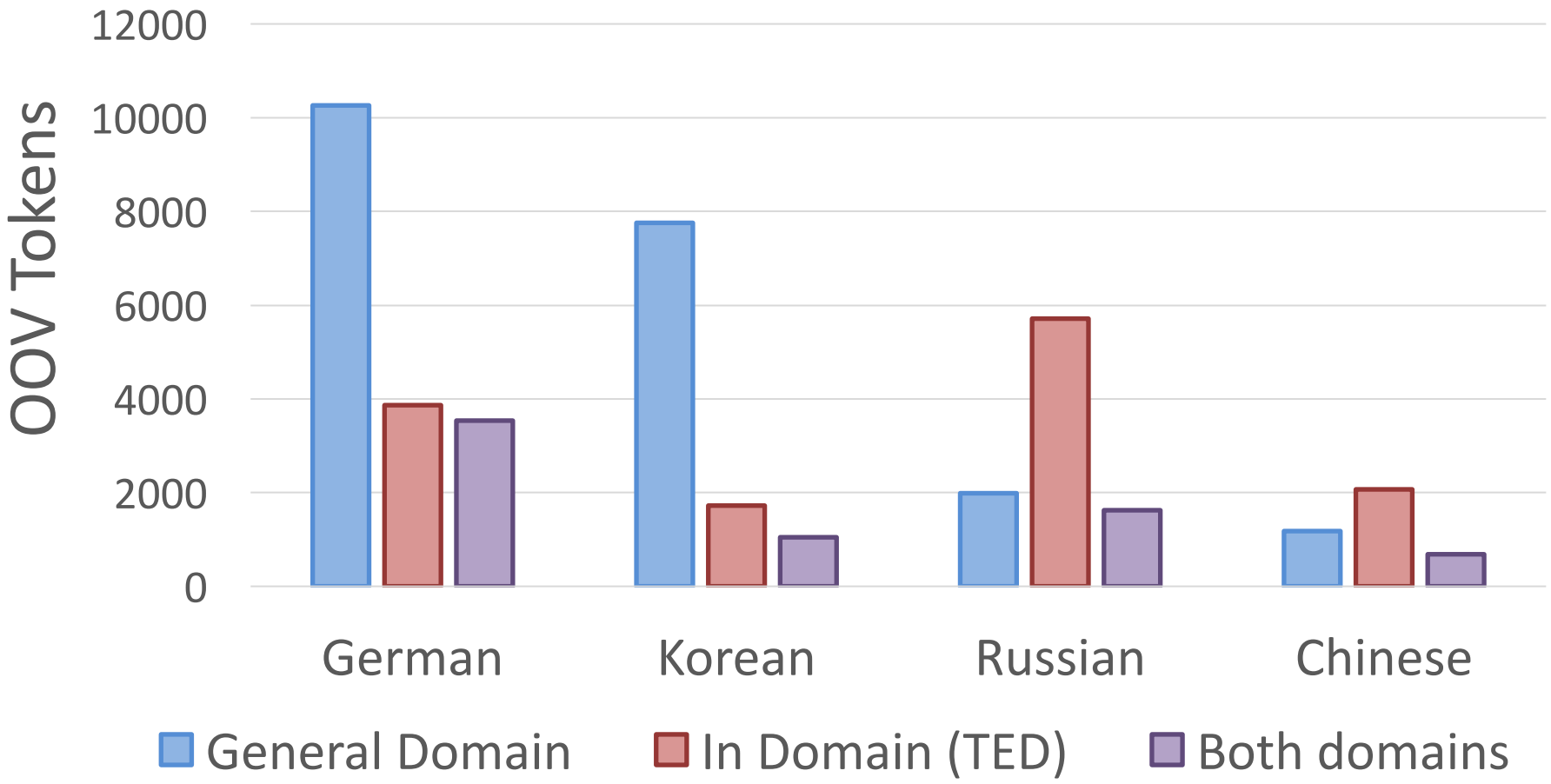
TED OOVs (Token %)



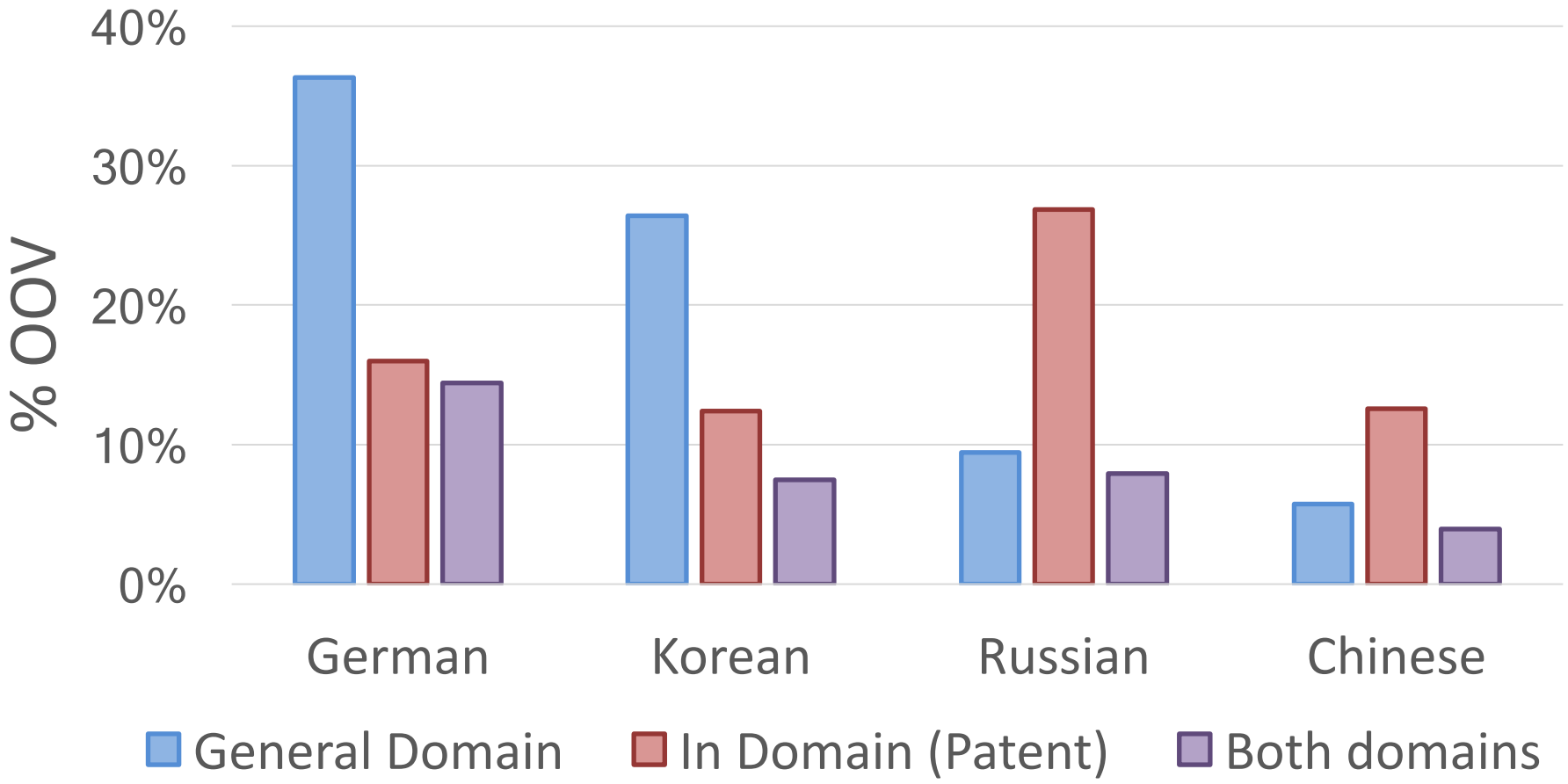
Patent OOVs (Type Count)



Patent OOVs (Token Count)



Patent OOVs (Type %)



Patent OOVs (Token %)