OPUS-MT

Democratizing Machine Translation

Jörg Tiedemann
University of Helsinki
jorg.tiedemann@helsinki.fi

https://github.com/Helsinki-NLP/OPUS-MT

Recent update:
MultiParaCrawl v9b
including Ukrainian!

OPUS

Training data for MT!

OPUS ... the open parallel corpus

https://opus.nlpl.eu/

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

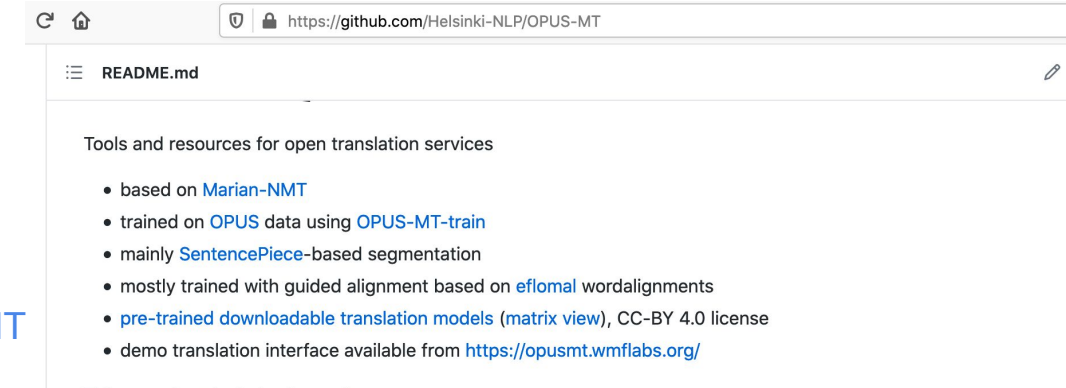
The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@helsinki.fi >

Search & download resources: fi (Finnish) sv (Swedish) all show all versions

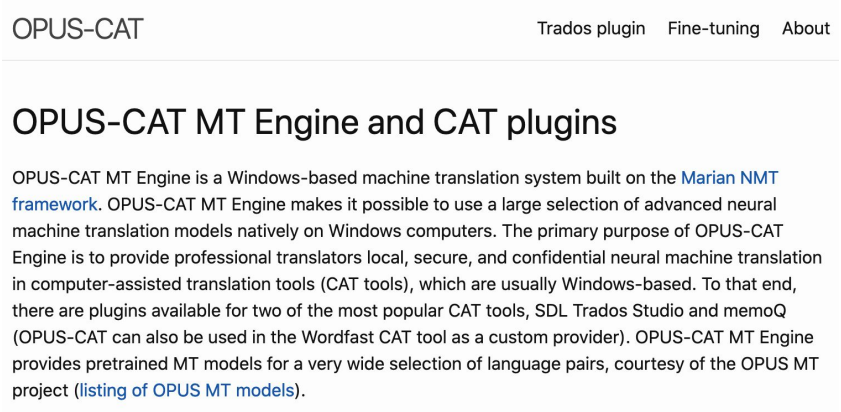
Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase ta

corpus	doc's	sent's	fi tokens	sv tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq			other files
MultiParaCrawl v7.1	1	3.1M	103.2M	187.5M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv				fi sv		sample	
MultiCCAligned v1	1	4.1M	93.7M	151.0M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv				fi sv		sample	
WikiMatrix v1	1	0.9M	79.8M	121.5M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv				fi sv		sample	
OpenSubtitles v2018	19565	15.3M	86.6M	108.9M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv	query	sample	xces/alt
DGT v2019	38586	4.9M	79.3M	92.0M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
EUbookshop v2	3925	1.9M	60.1M	68.1M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv			dic	fi sv	query	sample	moses/strict
Europarl v8	10188	1.9M	40.6M	50.6M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
Finlex v2018	56481	3.3M	47.0M	36.8M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
JW300 v1b	22730	1.7M	24.4M	30.9M	xces fi sv	fi sv			fi sv	fi sv				fi sv		sample	
fiskmo v2	42	2.1M	22.8M	26.1M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
EMEA v3	1962	1.2M	12.6M	13.9M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv	query	sample	
XLEnt v1	1	1.3M	5.9M	8.6M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv				fi sv		sample	
GNOME v1	1675	0.7M	2.9M	4.1M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt		fi sv		sample	
wikimedia v20210402	1	6.3k	1.3M	2.0M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv				fi sv		sample	
QED v2.0a	409	86.6k	0.9M	1.1M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
JRC-Acquis v3.0	203	21.8k	0.7M	0.9M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv				fi sv		sample	
bible-uedin v1	1	31.0k	0.7M	0.9M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
infopankki v1	279	53.3k	0.4M	0.5M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
Tatoeba v2021-03-10	4	10.0k	0.6M	0.2M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv				fi sv		sample	
Ubuntu v14.10	393	0.1M	0.4M	0.4M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
TED2020 v1	220	25.2k	0.3M	0.4M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
KDE4 v2	757	0.1M	0.6M	0	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv	query	sample	
EUconst v1	47	10.2k	0.1M	0.1M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg		fi sv	query	sample	
TildeMODEL v2018	1	4.7k	0.1M	0.1M	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
PHP v1	3283	28.6k	57.6k	75.9k	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt		fi sv	query	sample	
ELRC_416 v1	1	0.7k	3.8k	4.4k	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
ELRA-W0305 v1	1	18	0.2k	0.3k	xces fi sv	fi sv	tmx	moses	fi sv	fi sv		alg smt	dic	fi sv		sample	
MPC1 v1					xces fi sv	fi sv			fi sv	fi sv					query		
total	**160758**	**42.9M**	**665.0M**	**906.9M**	**42.9M**		**41.2M**	**41.2M**									

https://github.com/Helsinki-NLP/OPUS-MT



https://helsinki-nlp.github.io/OPUS-CAT/

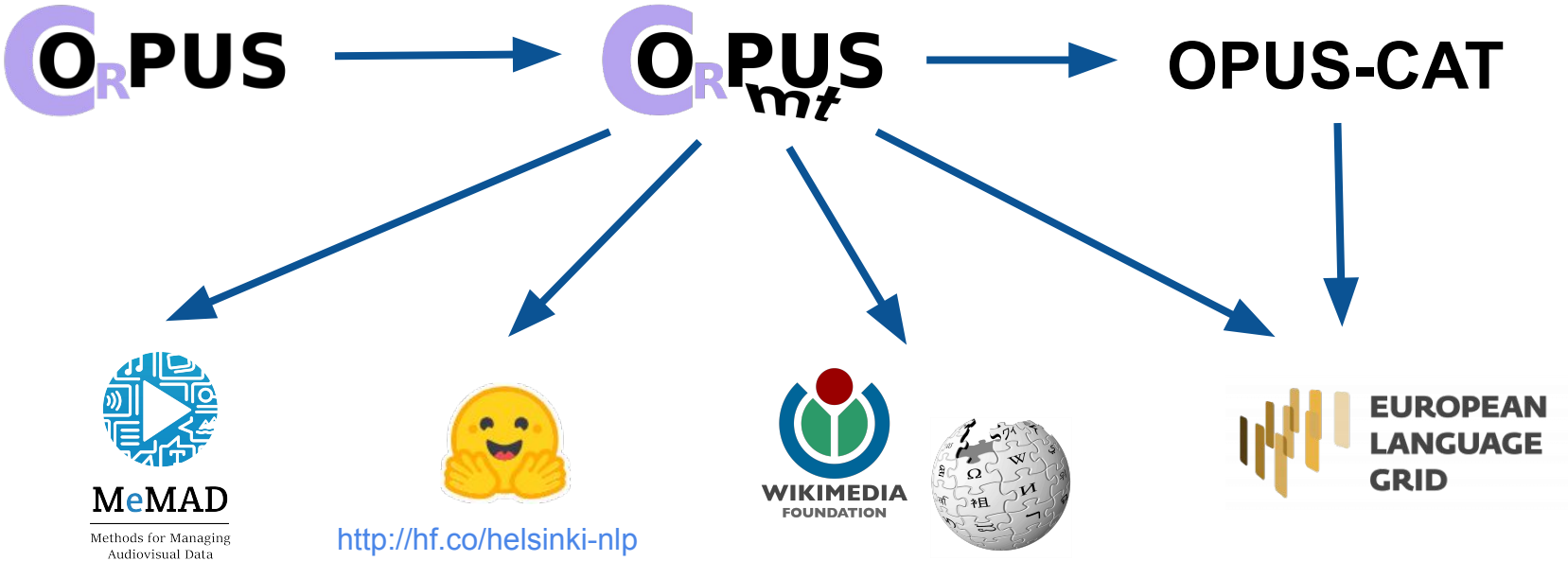


OPUS-CAT

Trados plugin Fine-tuning About

OPUS-CAT MT Engine and CAT plugins

OPUS-CAT MT Engine is a Windows-based machine translation system built on the Marian NMT framework. OPUS-CAT MT Engine makes it possible to use a large selection of advanced neural machine translation models natively on Windows computers. The primary purpose of OPUS-CAT Engine is to provide professional translators local, secure, and confidential neural machine translation in computer-assisted translation tools (CAT tools), which are usually Windows-based. To that end, there are plugins available for two of the most popular CAT tools, SDL Trados Studio and memoQ (OPUS-CAT can also be used in the Wordfast CAT tool as a custom provider). OPUS-CAT MT Engine provides pretrained MT models for a very wide selection of language pairs, courtesy of the OPUS MT project (listing of OPUS MT models).


OPUS
OPUS mt
OPUS-CAT
MeMAD
Methods for Managing Audiovisual Data
http://hf.co/helsinki-nlp
WIKIMEDIA FOUNDATION
EUROPEAN LANGUAGE GRID

Quickstart

Setting up:

```
git clone https://github.com/Helsinki-NLP/OPUS-MT-train.git
git submodule update --init --recursive --remote
make install
```

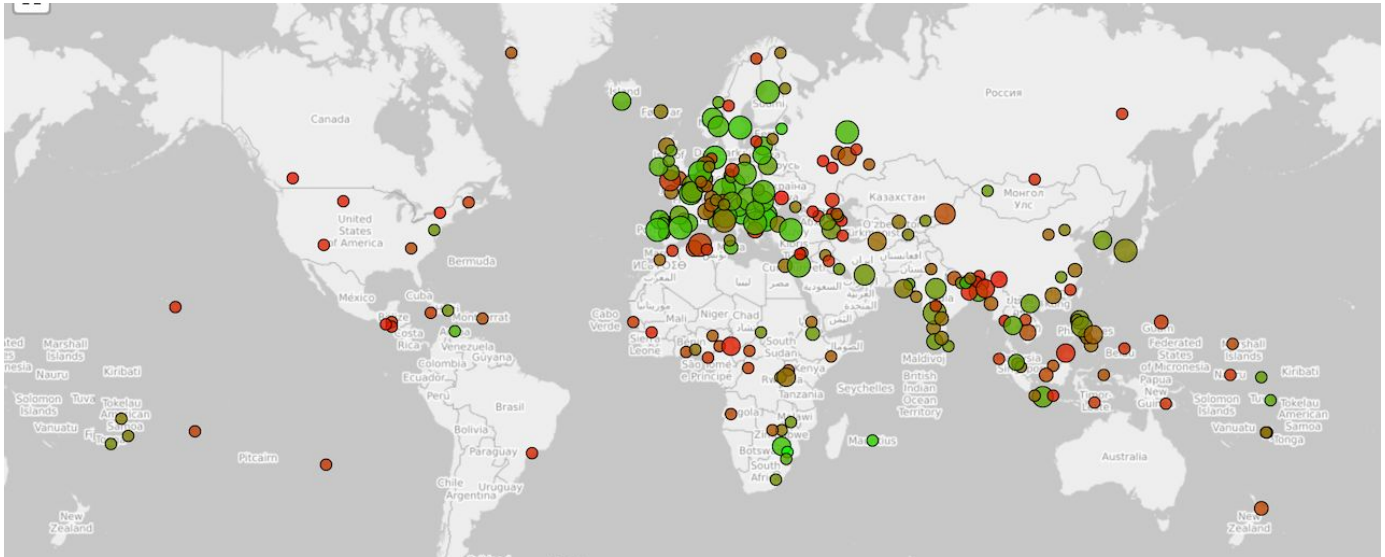
Training a multilingual NMT model (Finnish and Estonian to Danish, Swedish and English):

```
make SRCLANGS="fi et" TRGLANGS="da sv en" train
make SRCLANGS="fi et" TRGLANGS="da sv en" eval
make SRCLANGS="fi et" TRGLANGS="da sv en" release
```

OPUS-MT: Open NMT tools and models

Pre-trained models:

- https://opus.nlpl.eu/Opus-MT/
- https://github.com/Helsinki-NLP/Tatoeba-Challenge



Various ways to use OPUS-MT models

```
from transformers import pipeline
pipe = pipeline("translation", model="Helsinki-NLP/opus-mt-en-fi")
print(pipe('Please, translate this text.'))
```

http://hf.co/helsinki-nlp

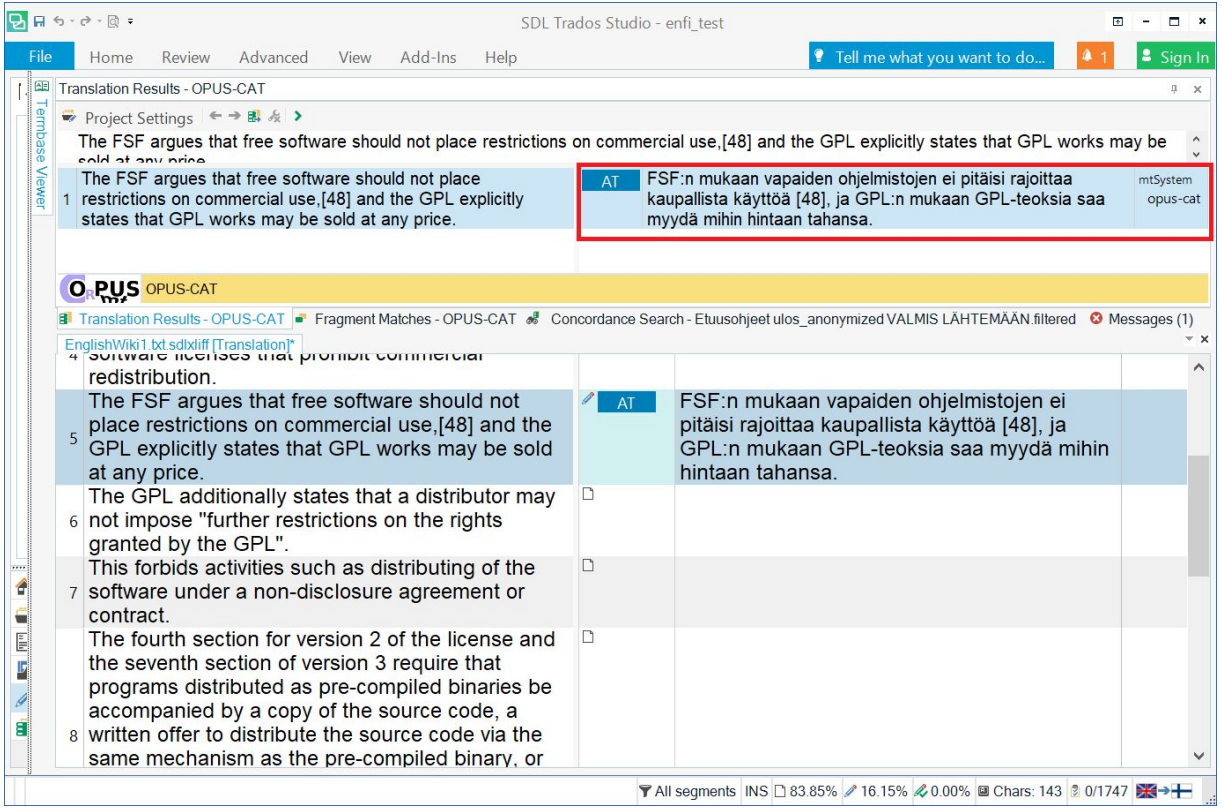
EUROPEAN LANGUAGE GRID

```
from elg import Service
service = Service.from_id(4863)
result = service(request_input="translate this!", request_type="text")
```

```
docker run  -p 8888:8888 helsinkinlp/opus-mt-elg-fin-eng:1.0
curl -X POST -H "Content-Type: application/json" -d
'{"content":"Translate this!"}' "localhost:8888/elg/translate/en/fi"
```

docker hub

OPUS-CAT: Local translation engines on your desktop



Plugins for:

- Trados Studio
- memoQ
- OmegaT

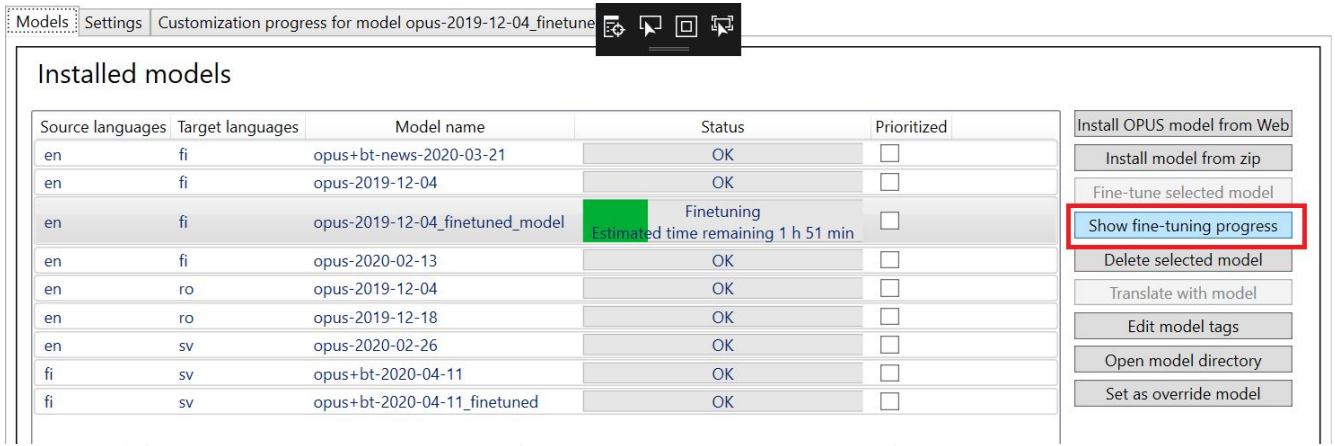
Local MT engine

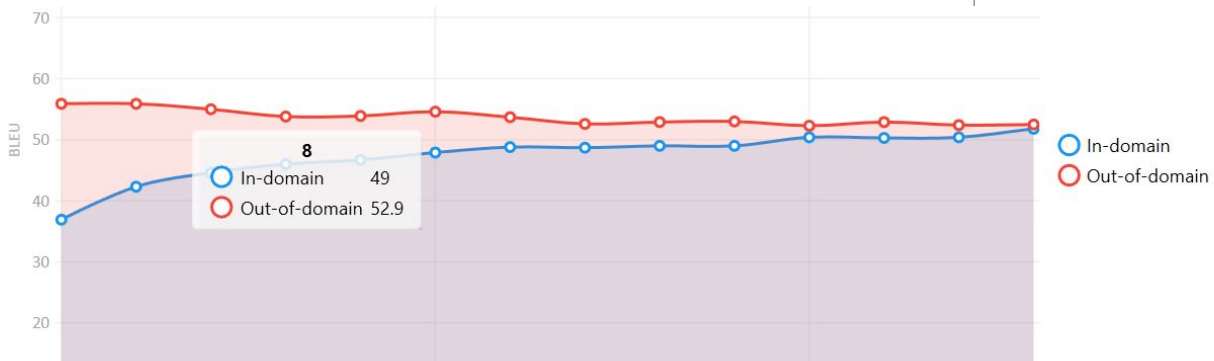
- Data security
- Low latency
- Off-line mode

Extensible and open

- All of OPUS-MT
- No costs

OPUS-CAT: Customized machine translation





Fine-tuning:

- Personal TMX
- Runs locally
- Any number of fine-tuned models

Optimizes:

- Terminology
- Placeholder tags
- Domain-specific translations

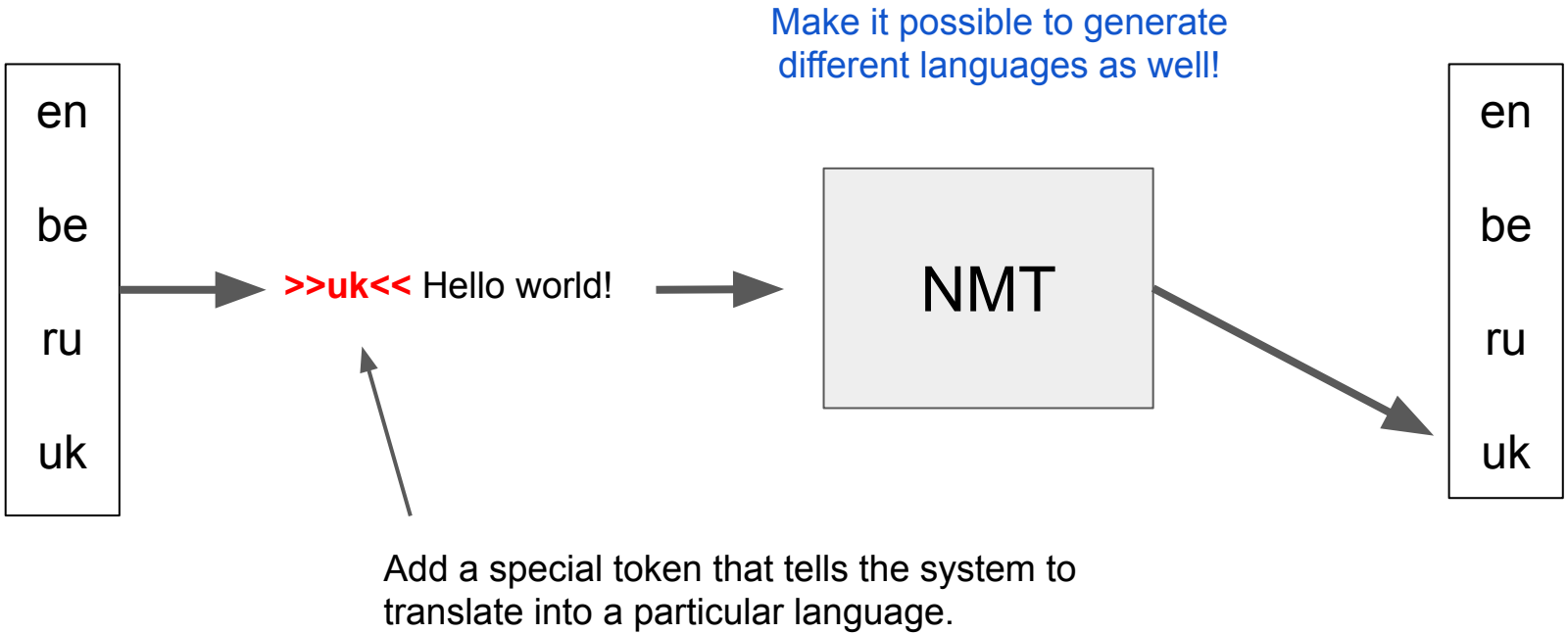
OPUS-MT and Ukrainian LT

Improved MT models for Ukrainian

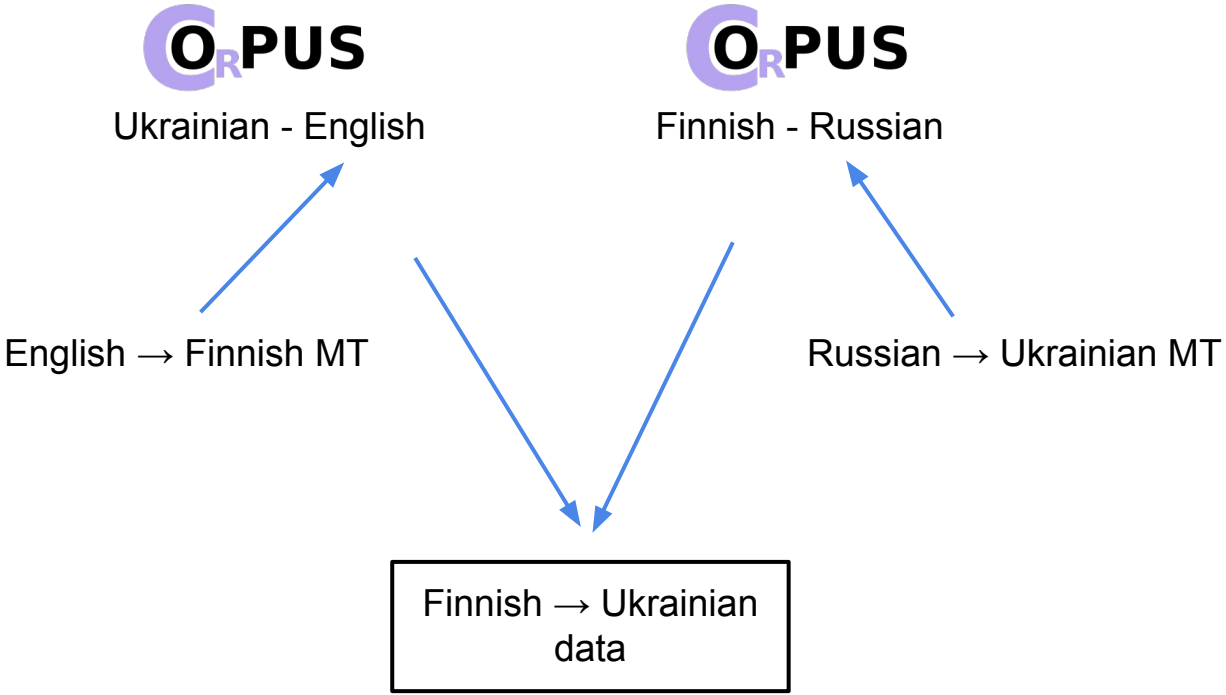
- Data augmentation
- Multilingual models and pivot-based models
- Various model sizes and extensive training

https://github.com/Helsinki-NLP/UkrainianLT/

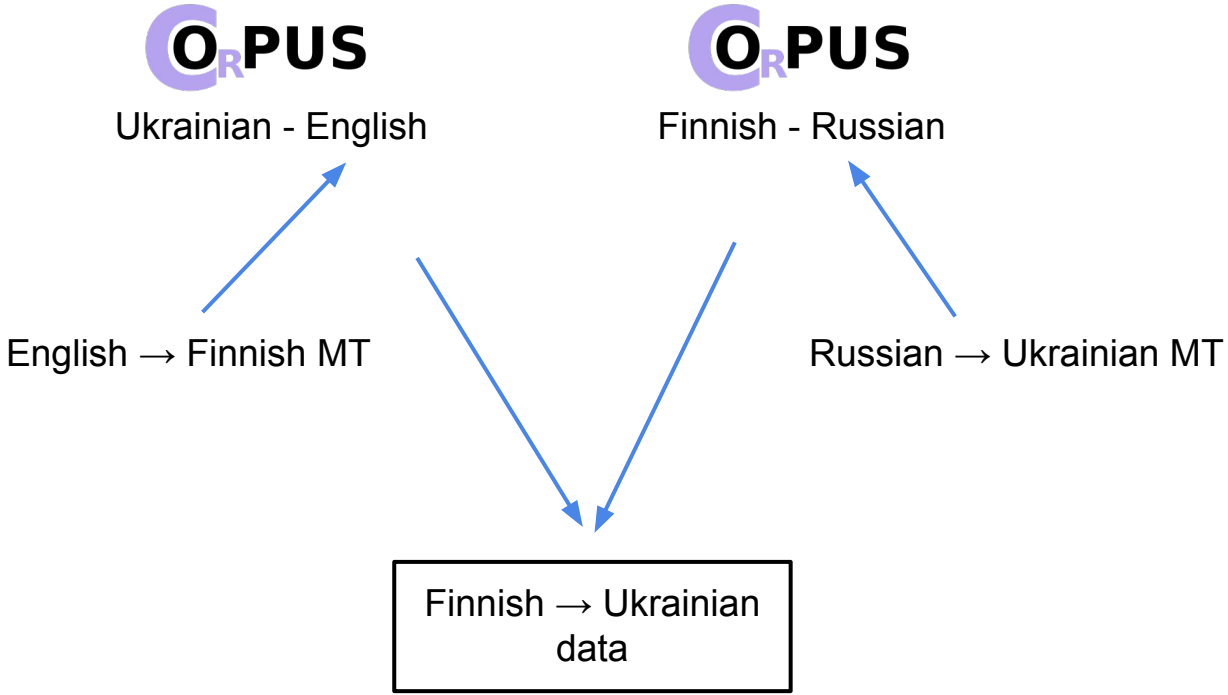
Extension to multilingual machine translation



Data augmentation



Data augmentation



ParaCrawl

English-centric data

MultiParaCrawl:
Finnish → Ukrainian data through English as a pivot language

Data augmentation

English-centric data

Ukrainian - English

Finnish - Russian

English → Finnish MT

Russian → Ukrainian MT

MultiParaCrawl:
Finnish → Ukrainian data through English as a pivot language

Finnish → Ukrainian data

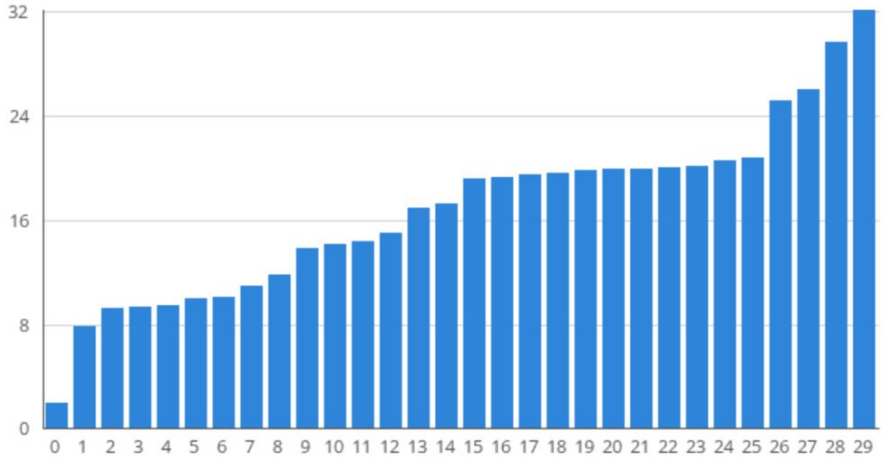
Ukrainian → Finnish MT

Ukrainian Wikipedia

Tracking progress with the OPUS-MT leaderboard

OPUS-MT leaderboard

- language pair: eng - ukr
- benchmark: flores101-devtest [all benchmarks]
- metrics: bleu | chrf



ID	bleu	other	output	model
29	32.1	scores	zip-file	eng-zle/opusTCv20210807+bt_transformer-big_2022-03-13.zip
28	29.6	scores	zip-file	eng-ukr/opusTCv20210807+nopar+ft95-sepvoc_transformer-tiny11-align_2022-03-23.zip
27	25.9	scores	zip-file	eng-ukr/opus+bt-2021-04-14.zip
26	25.1	scores	zip-file	eng-ukr/opusTCv20210807+nopar+ft95-sepvoc_transformer-tiny11-align_2022-03-05.zip
25	20.7	scores	zip-file	eng-sla/opus2m-2020-08-01.zip
24	20.5	scores	zip-file	eng-zle/opus2m-2020-08-02.zip
23	20.1	scores	zip-file	eng-ukr/opus-2021-02-19.zip
22	20.0	scores	zip-file	eng-zle/opus-2020-06-28.zip
21	19.9	scores	zip-file	eng-sla/opus-2020-07-27.zip
20	19.8	scores	zip-file	eng-sla/opus-2020-07-06.zip
19	19.7	scores	zip-file	eng-zle/opus1m+bt-2021-04-10.zip
18	19.5	scores	zip-file	gmw-zle/opus1m-2021-02-12.zip
17	19.4	scores	zip-file	zle-zle/opus-2020-09-26.zip
16	19.2	scores	zip-file	eng-sla/opus-2020-07-14.zip
15	19.1	scores	zip-file	zle-zle/opus-2020-10-04.zip
14	17.2	scores	zip-file	sla-sla/opus-2020-10-04.zip
13	16.8	scores	zip-file	sla-sla/opus-2020-09-26.zip
12	14.9	scores	zip-file	eng-ine/opus2m-2020-08-01.zip
11	14.3	scores	zip-file	eng-ine/opus-2020-07-27.zip
10	14.1	scores	zip-file	eng-ine/opus-2020-07-19.zip
9	13.7	scores	zip-file	eng-ine/opus-2020-07-14.zip
8	11.7	scores	zip-file	ine-ine/opus-2020-07-27.zip
7	10.9	scores	zip-file	ine-ine/opus-2020-07-21.zip
6	10.0	scores	zip-file	eng-mul/opus2m-2020-08-01.zip
5	9.9	scores	zip-file	tatoeba-zero/opus-2020-06-19.zip
4	9.4	scores	zip-file	eng-mul/opus-2020-07-20.zip
3	9.3	scores	zip-file	eng-mul/opus-2020-07-27.zip
2	9.1	scores	zip-file	tatoeba-lowest/opus-2020-06-15.zip
1	7.8	scores	zip-file	eng-mul/opus-2020-07-14.zip
0	1.9	scores	zip-file	eng-zle/opus1m+bt-tuned4eng2bel-2021-04-16.zip

Knowledge Distillation

Size and speed (fin-eng)

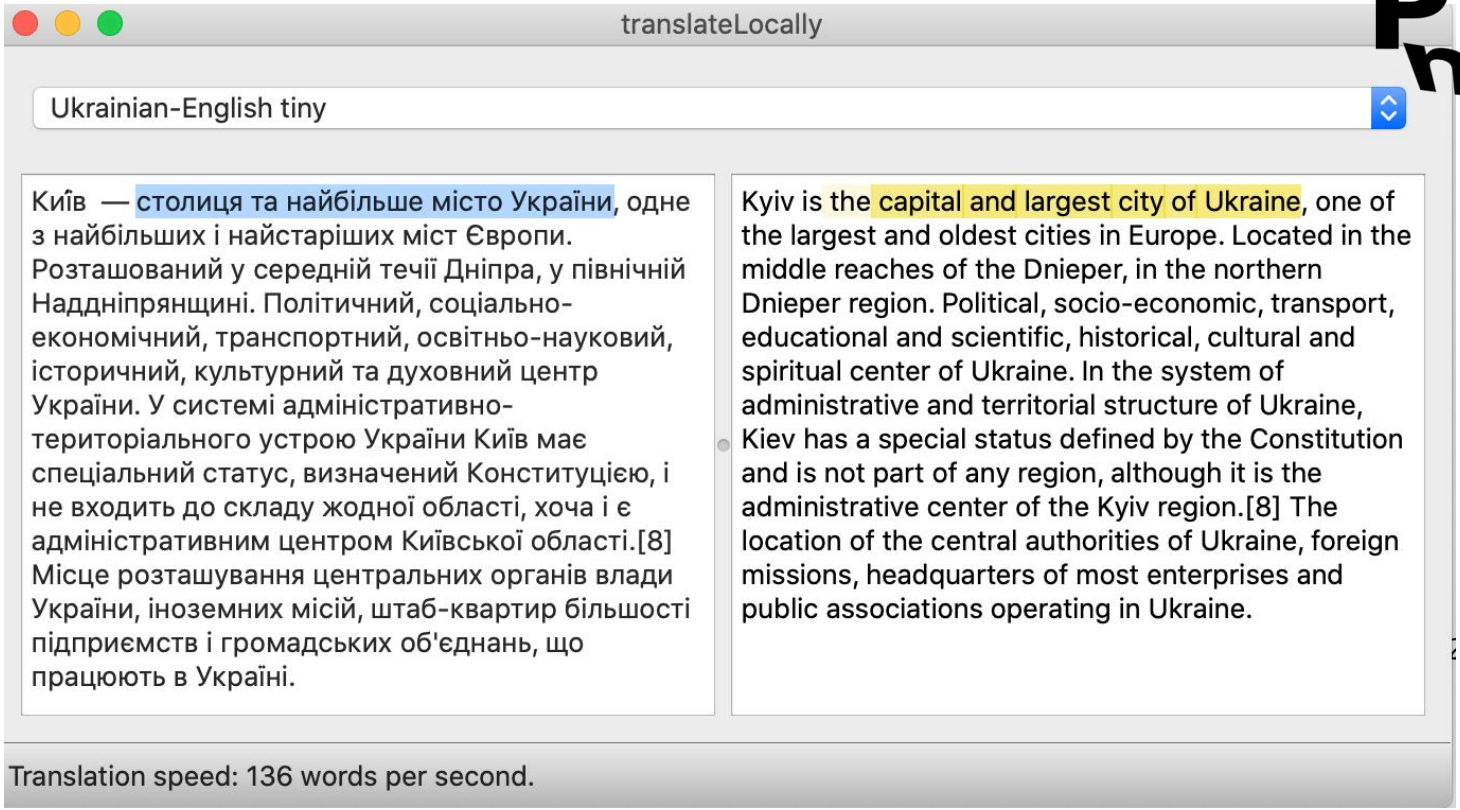
model	size original	size quantized
big	891 MB	224 MB
base	294 MB	74 MB
small	226 MB	57 MB
tiny	89 MB	23 MB

Size and speed (fin-eng)

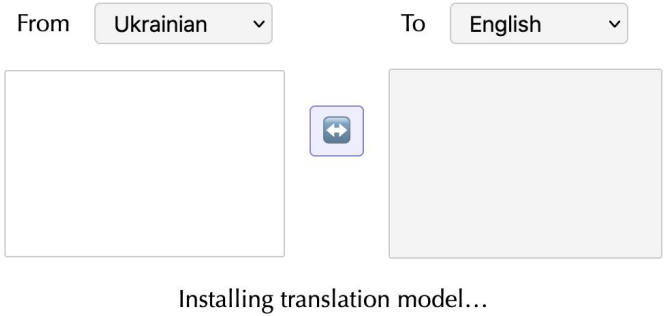
model	size original	size quantized	speed 4 CPUs	speed (+shortlist)
big	891 MB	224 MB		
base	294 MB	74 MB	46.36s	40.47s
small	226 MB	57 MB	24.07s	17.89s
tiny	89 MB	23 MB	9.90s	6.22s

Tatoeba-test: 10,000 sentences, 48,684 words

OPUS-MT-app - based on translateLocally



OPUS-MT and Ukrainian LT



Improved MT models for Ukrainian

- Data augmentation
- Multilingual models and pivot-based models
- Various model sizes and extensive training

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825303

Releases and integration

- Downloads as Marian-NMT, pytorch / HF-transformers, docker
- Web interfaces, in-browser MT, ELG API
- Telegram translation bot

opusmt

@opusmt_bot

SEND MESSAGE

https://github.com/Helsinki-NLP/UkrainianLT/

Thanks!

Especially also thanks for all the contributions
to OPUS and the various tools!

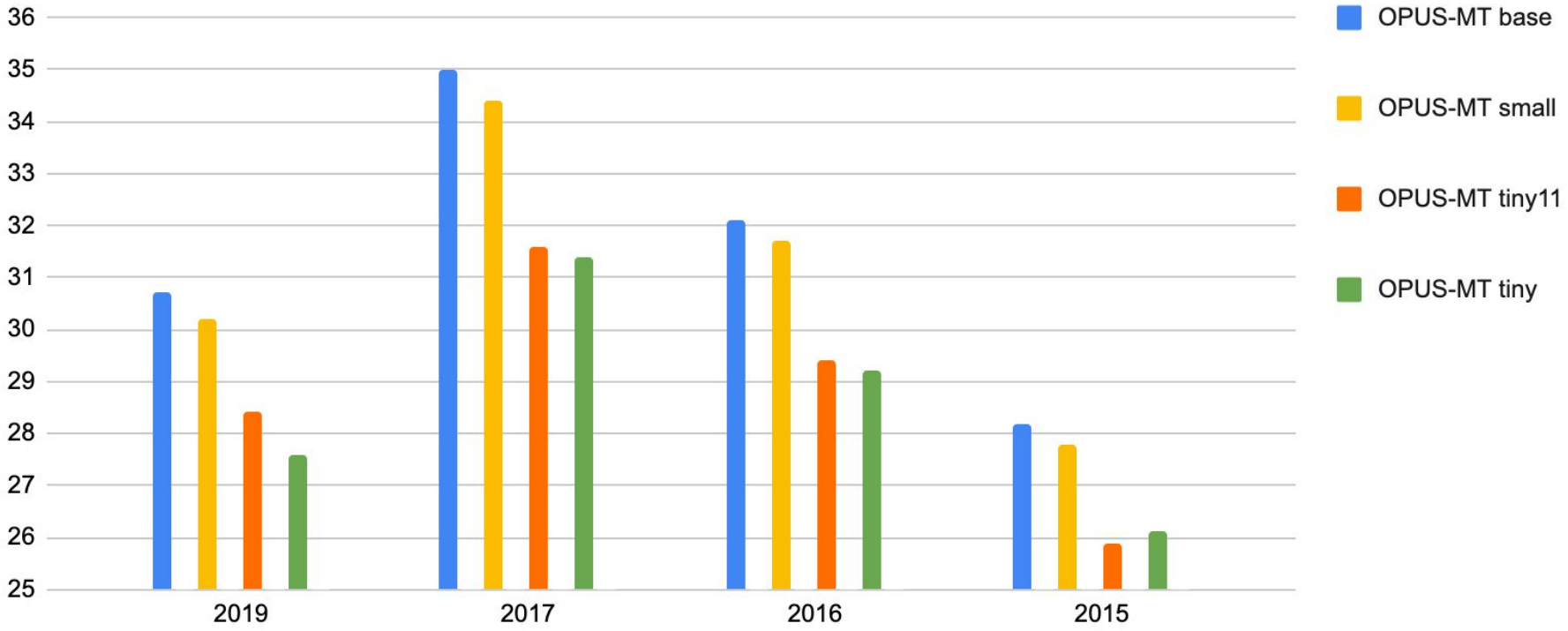
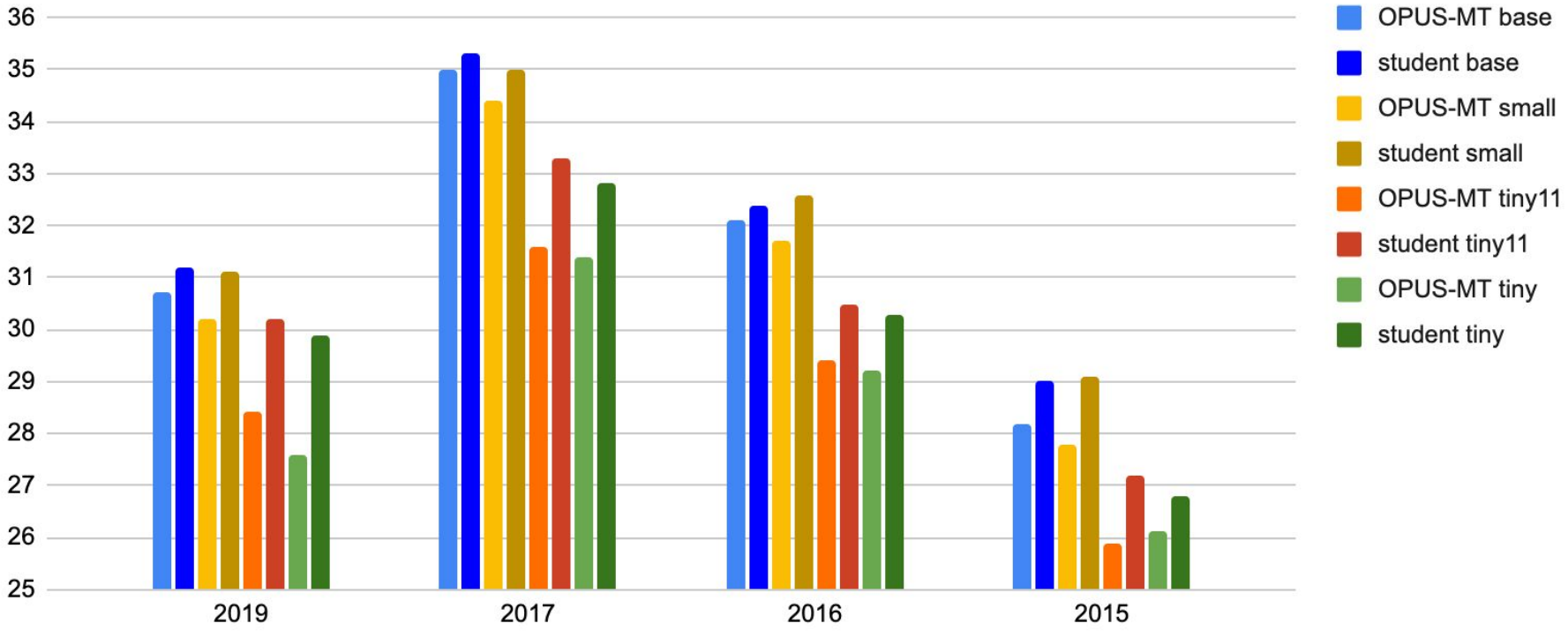
(a long list of names should be added here …)

base	small	tiny11	tiny
6 x encoder 6 x decoder	6 x encoder 2 x SSRU dec	6 x encoder 2 x SSRU dec emb. 256, ffn 1536	3 x encoder 2 x SSRU dec emb. 256

Knowledge Distillation

Simpler Simple Recurrent Units (SSRU; Kim et al., 2019

Performance of different model types (WMT fin-eng)



Performance of different model types (WMT fin-eng)