Putting it All Together

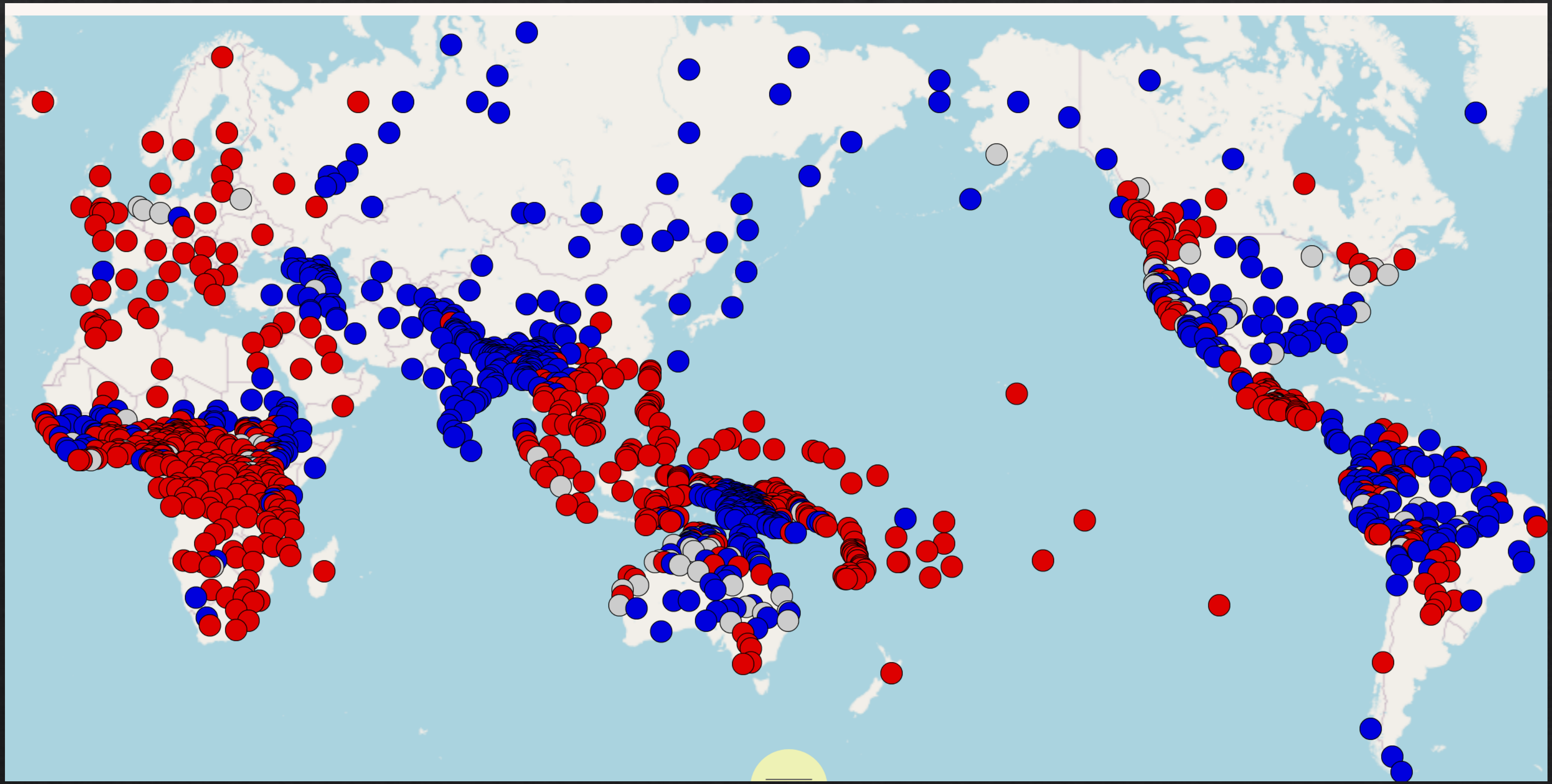
601.764

4/25/23

Coverage

How many languages on Earth?

5,000-7,000

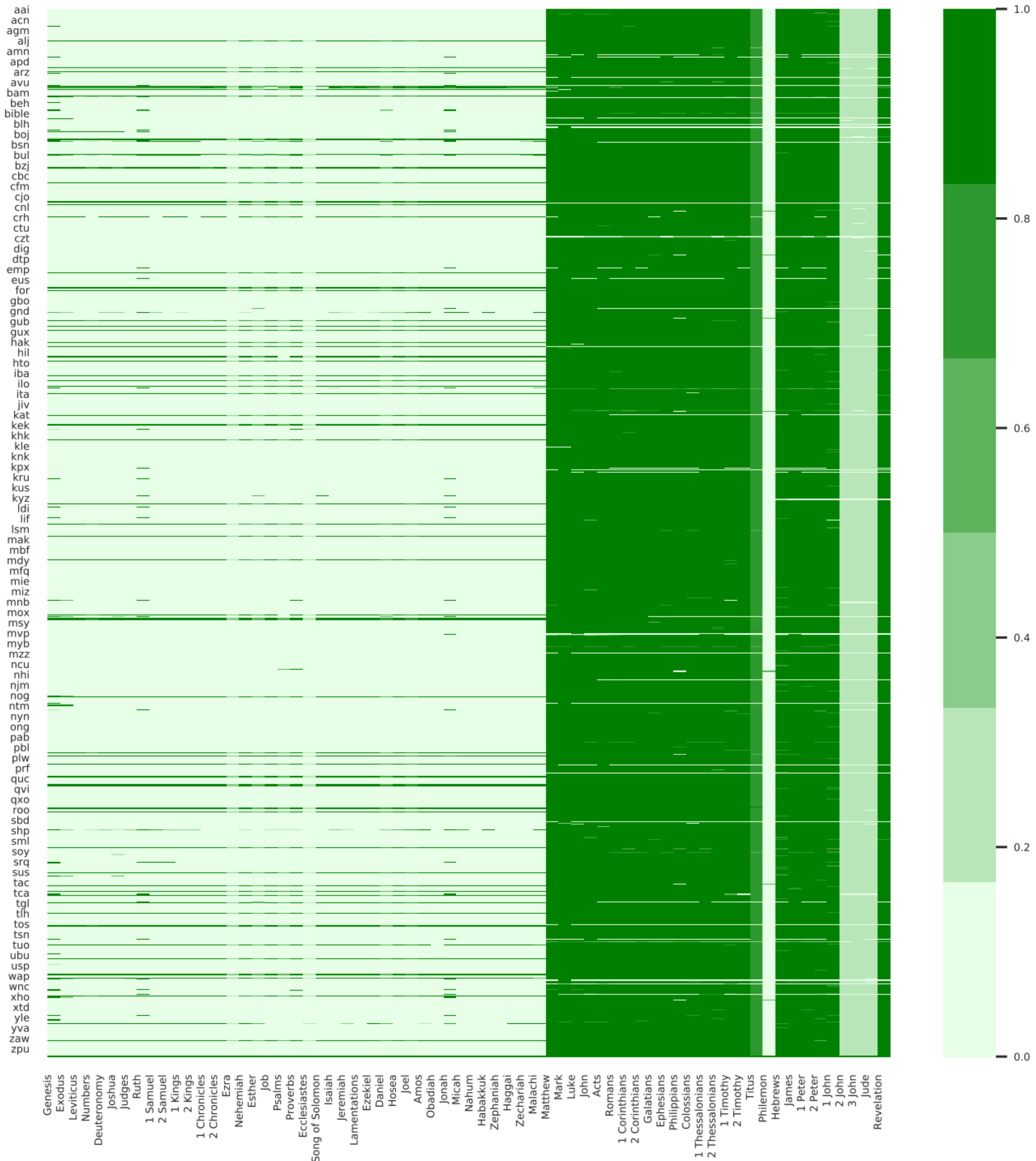


Coverage

Largest Corpora?

The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and **David Yarowsky**
Center for Language and Speech Processing
Johns Hopkins University
(arya, rewicks, dlewis77, amueller, wswu, oadams, gnicola2, yarowsky)@jhu.edu, post@cs.jhu.edu



aai
acn
agm
alj
amn
apd
arz
avu
bam
beh
bible
blh
boj
bsn
bul
bzj
cbc
cfm
cjo
cnl
crh
ctu
czt
dig
dtp
emp
eus
for
gbo
gnd
gub
gux
hak
hil
hto
iba
ilo
ita
jiv
kat
kek
khk
kle
knk
kpx
kru
kus
kyz
ldi
lif
lsm
mak
mbf
mdy
mfq
mie
miz
mnb
mox
msy
mvp
myb
mzz
ncu
nhi
njm
nog
ntm
nyn
ong
pab
pbl
plw
prf
quc
qvi
qxo
roo
sbd
shp
sml
soy
srq
sus
tac
tca
tgl
tlh
tos
tsn
tuo
ubu
usp
wap
wnc
xho
xtd
yle
yva
zaw
zpu
Genesis
Exodus
Leviticus
Numbers
Deuteronomy
Joshua
Judges
Ruth
1 Samuel
2 Samuel
1 Kings
2 Kings
1 Chronicles
2 Chronicles
Ezra
Nehemiah
Esther
Job
Psalms
Proverbs
Ecclesiastes
Song of Solomon
Isaiah
Jeremiah
Lamentations
Ezekiel
Daniel
Hosea
Joel
Amos
Obadiah
Jonah
Micah
Nahum
Habakkuk
Zephaniah
Haggai
Zechariah
Malachi
Matthew
Mark
Luke
John
Acts
Romans
1 Corinthians
2 Corinthians
Galatians
Ephesians
Philippians
Colossians
1 Thessalonians
2 Thessalonians
1 Timothy
2 Timothy
Titus
Philemon
Hebrews
James
1 Peter
2 Peter
1 John
2 John
3 John
Jude
Revelation
1.0
0.8
0.6
0.4
0.2
0.0

JW300: A Wid ce Languages

Departm
IT Universit dom
com

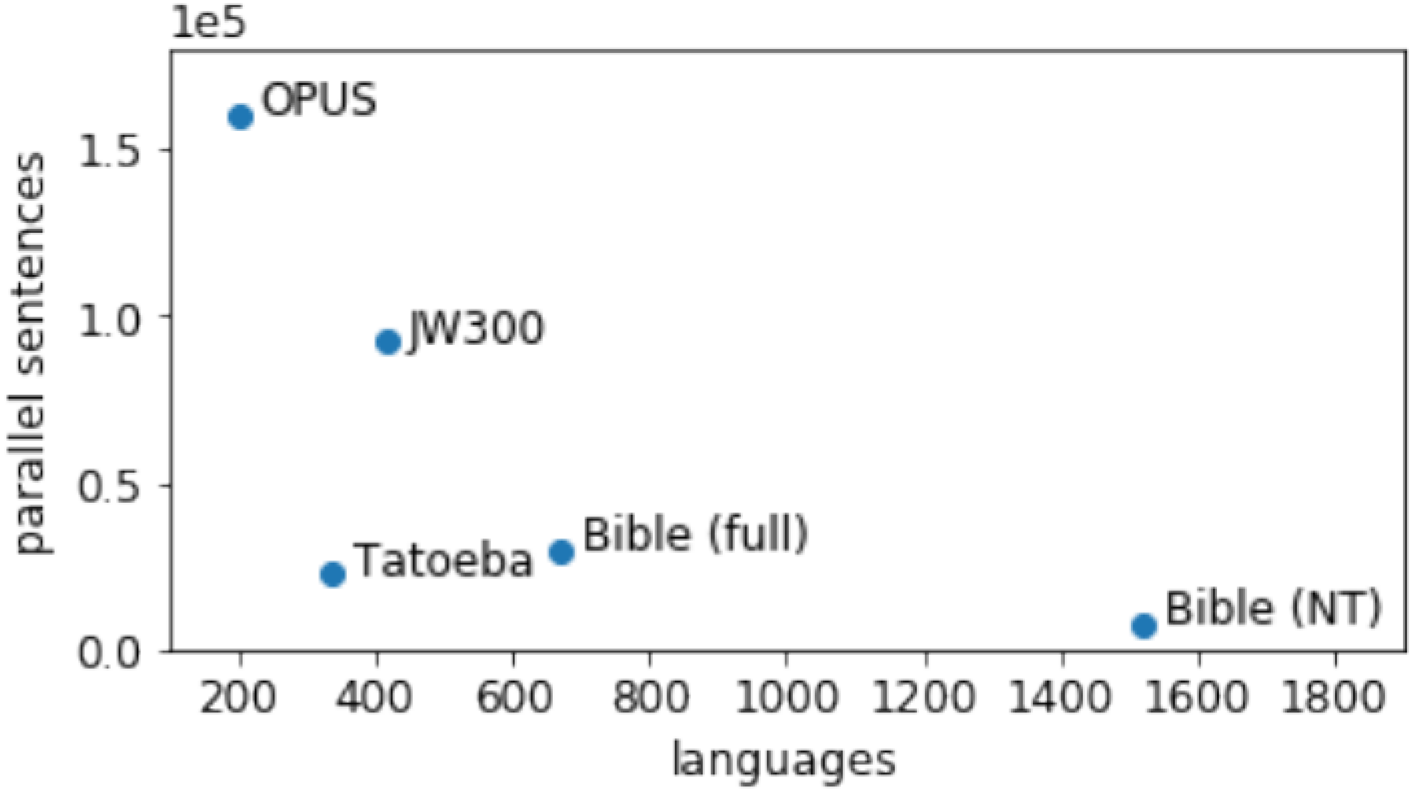


Figure 1: Our dataset JW300 in comparison to other massive parallel text collections with respect to multilingual breadth and volume of parallel sentences. The y-axis depicts the mean number of parallel sentences per language pair.

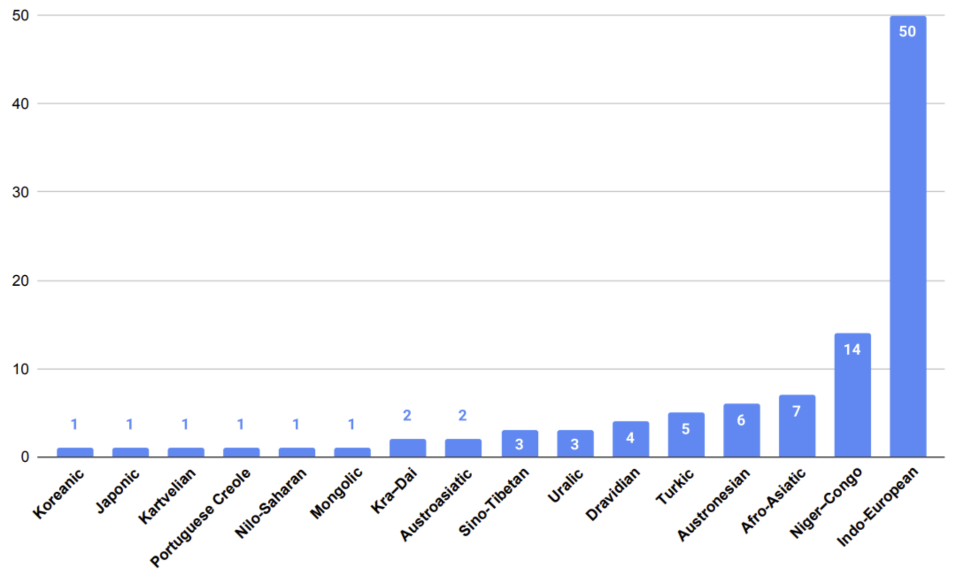


Figure 1: *Distributions of language families in FLEURS (y-axis is the count).*

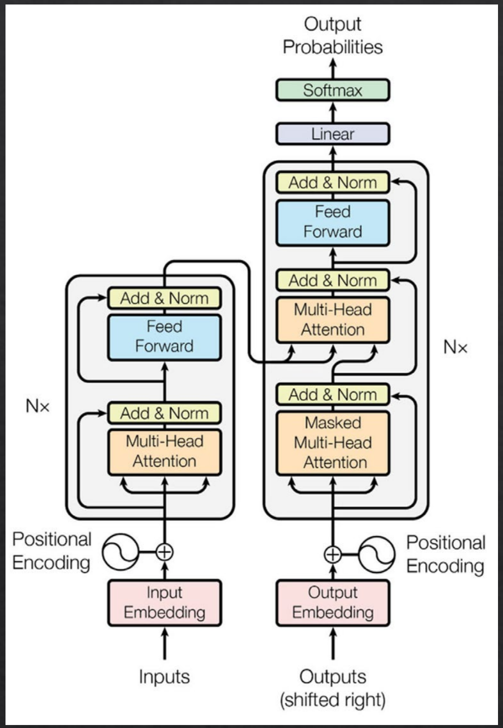
No Language Left Behind: Scaling Human-Centered Machine Translation

NLLB Team, Marta R. Costa-jussà[*], James Cross[*], Onur Çelebi[*], Maha Elbayad[*], Kenneth Heafield[*], Kevin Heffernan[*], Elahe Kalbassi[*], Janice Lam[*], Daniel Licht[*], Jean Maillard[*], Anna Sun[*], Skyler Wang[*,§], Guillaume Wenzek[*], Al Youngblood[*]
Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran
Pierre Andrews[†], Necip Fazil Ayan[†], Shruti Bhosale[†], Sergey Edunov[†], Angela Fan[†,‡], Cynthia Gao[†], Vedanuj Goswami[†], Francisco Guzmán[†], Philipp Koehn[†,¶], Alexandre Mourachko[†], Christophe Ropers[†], Safiyyah Saleem[†], Holger Schwenk[†], Jeff Wang[†]

Meta AI, [§]UC Berkeley, [¶]Johns Hopkins University

Cross-Lingual vs. Multilingual

Pre-Trained Models



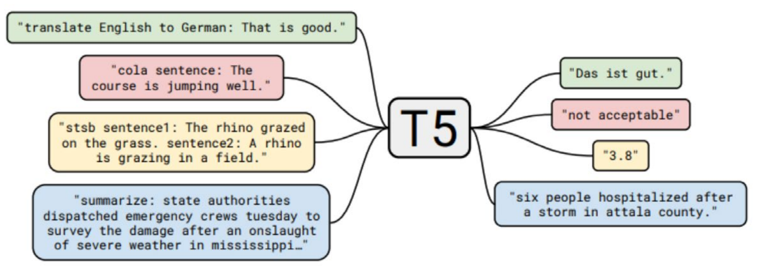
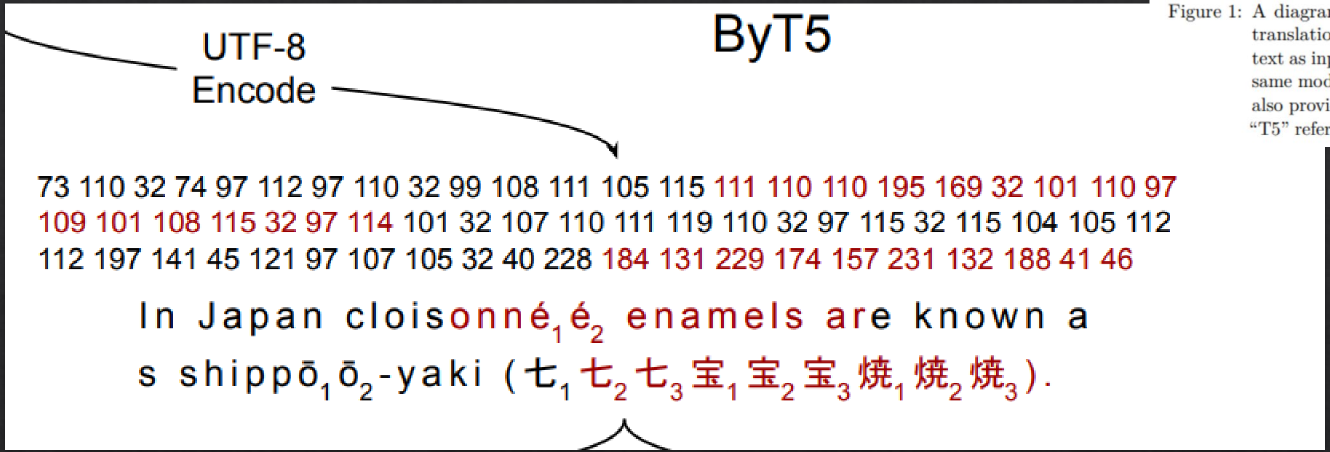
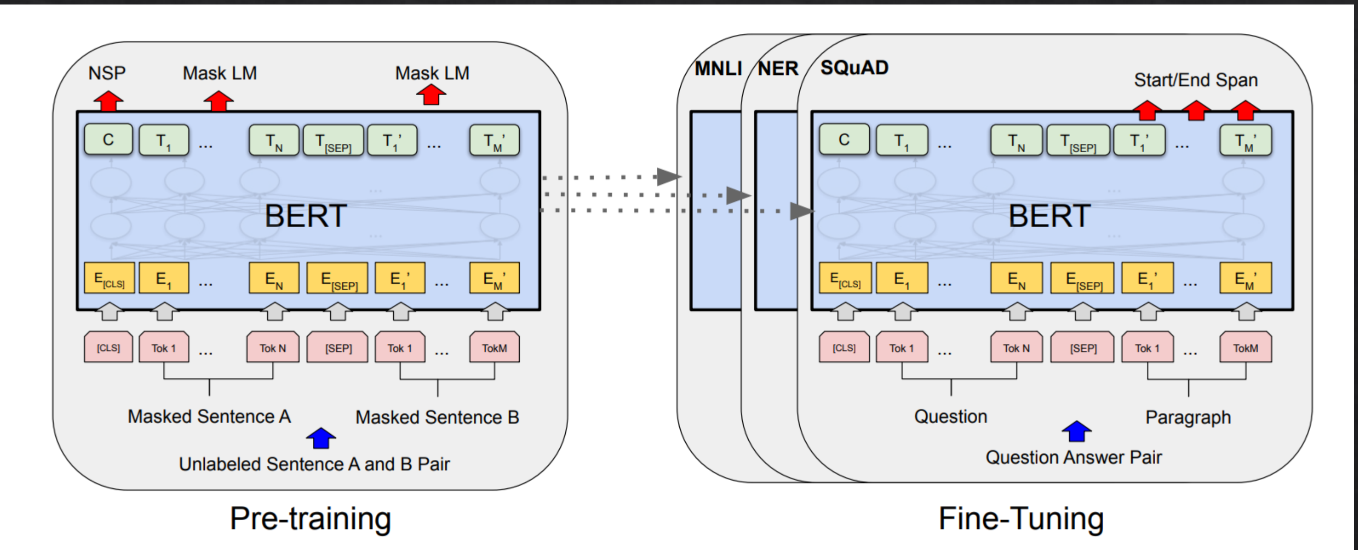
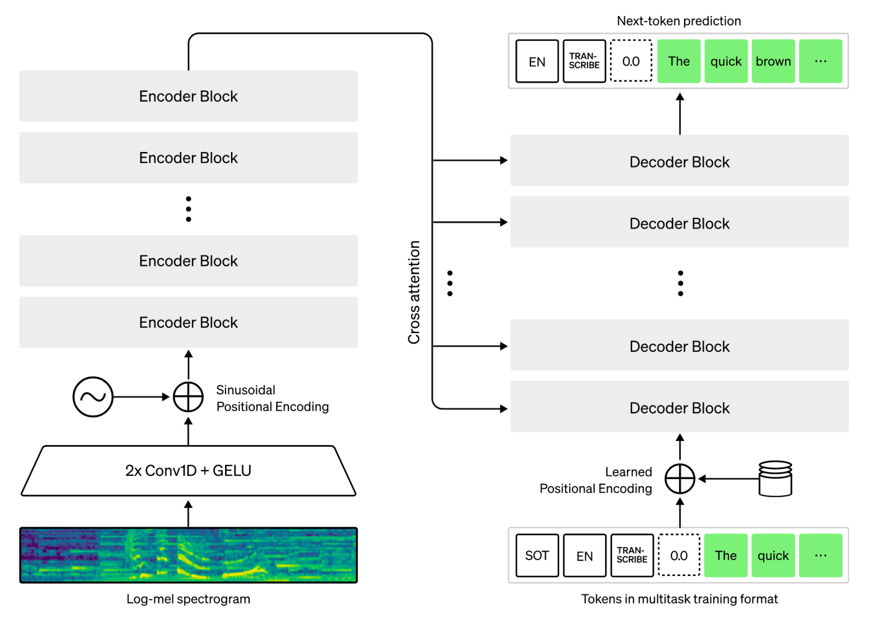


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "Text-to-Text Transfer Transformer".







Typologically Diverse

TyDi QA

Typologically **Di**verse **Q**uestion **A**nswering

A benchmark for information-seeking question answering in typologically diverse languages

The Effect of Translationese in Machine Translation Test Sets

Mike Zhang
Information Science Programme
University of Groningen
The Netherlands
j.j.zhang.1@student.rug.nl

Antonio Toral
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

Translationese

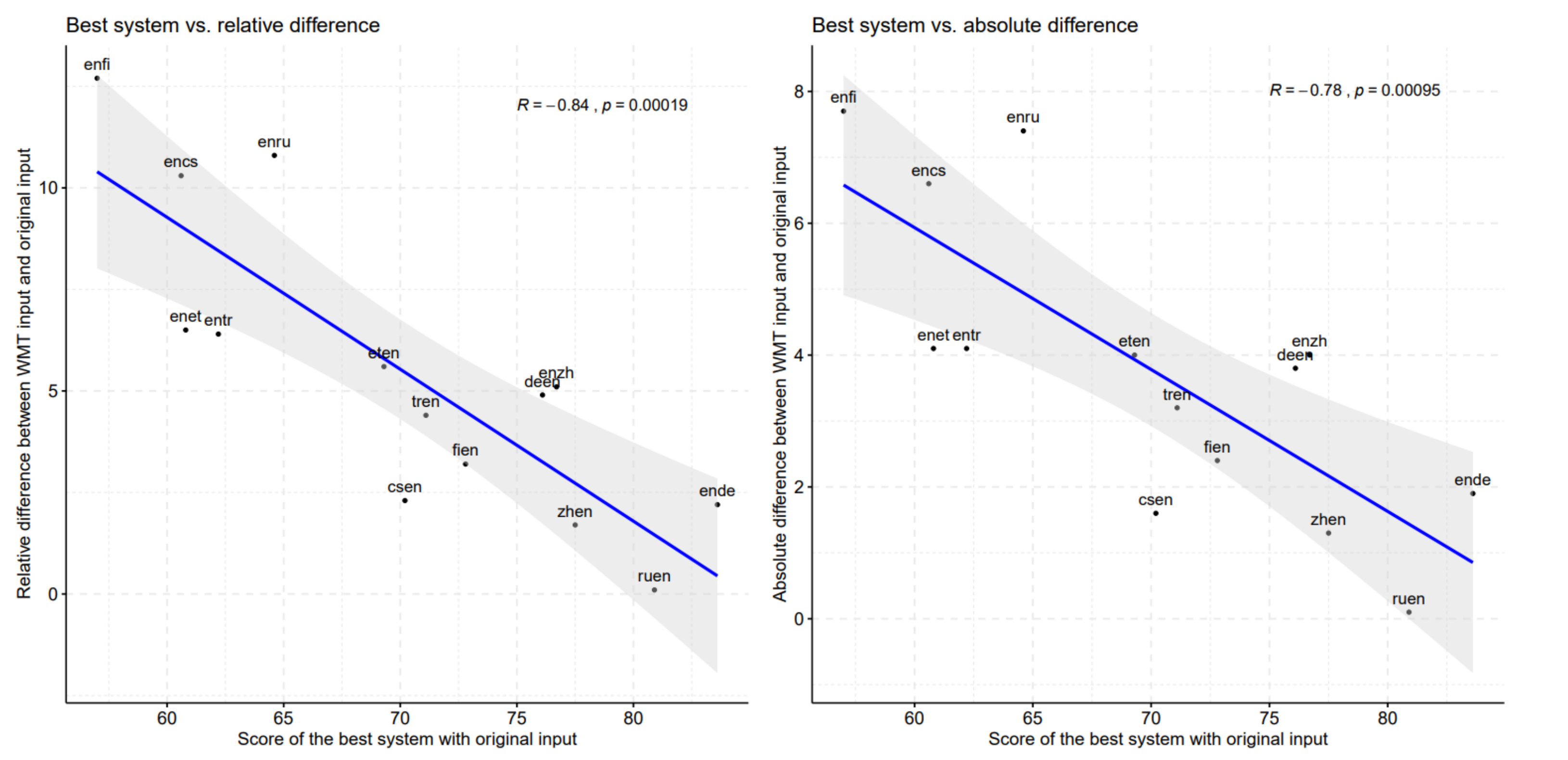


Figure 2: Pearson correlation between the DA scores of the best system for each translation direction at WMT18 and the relative (left) and absolute (right) difference in DA score (%) of comparing WMT input and ORG input. The languages are abbreviated into ISO 639-1 codes (Byrum, 1999).

Language Direction	WMT16			WMT17			WMT18		
	WMT	ORG	TRS	WMT	ORG	TRS	WMT	ORG	TRS
Chinese→English				73.2	-1.5	+3.9	78.8	-1.3	+2.0
English→Chinese				73.2	-4.1	+5.0	80.7	-4.0	+2.3
Czech→English	75.4	-5.8	+5.7	74.6	-4.3	+4.2	71.8	-1.6	+1.6
English→Czech				62.0	-5.8	+7.4	67.2	-6.6	+7.2
Estonian→English							73.3	-4.0	+4.0
English→Estonian							64.9	-4.1	+3.9
Finnish→English	66.9	-3.2	+3.0	73.8	-2.1	+2.2	75.2	-2.4	+2.3
English→Finnish				59.6	-5.1	+5.6	64.7	-7.7	+8.0
German→English	75.8	-4.1	+4.1	78.2	-2.4	+2.2	79.9	-3.8	+4.3
English→German				72.9	-5.1	+4.4	85.5	-1.9	+1.9
Latvian→English				76.2	-0.4	+0.6			
English→Latvian				54.4	-11.2	+11.7			
Romanian→English	73.9	-0.4	+0.5						
Russian→English	74.2	-1.2	+1.8	82.0	-0.7	+0.6	81.0	-0.1	0.0
English→Russian				75.4	-5.8	+5.8	72.0	-7.4	+7.4
Turkish→English	57.1	-1.6	+1.6	68.8	-3.8	+3.9	74.3	-3.2	+3.9
English→Turkish				53.4	-13.4	+11.8	66.3	-4.1	+5.5

Table 2: DA scores for the best MT system for each translation direction of WMT's 2016–2018 news translation shared task. Columns ORG and TRS show the absolute difference of the DA scores in those subsets compared to the whole test set (WMT).

Silver Dataset Creation

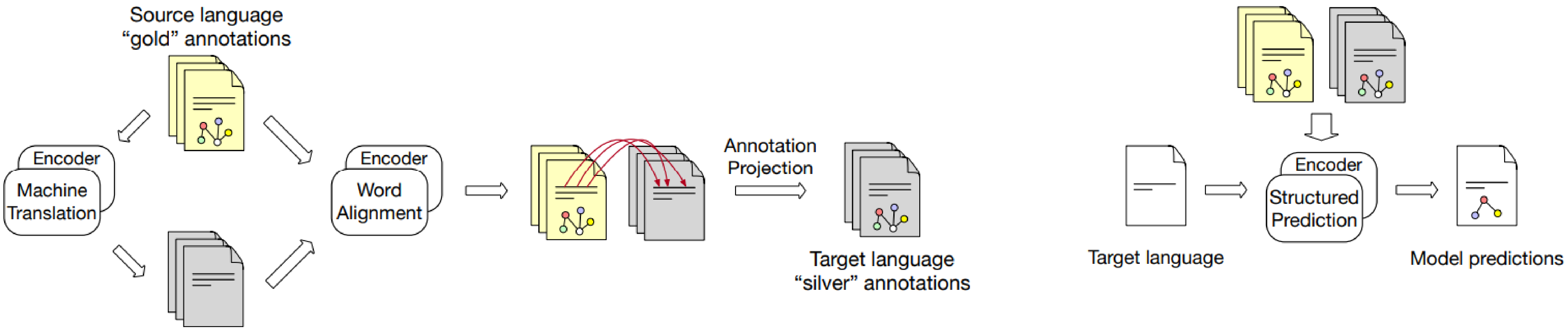


Figure 1: Process for creating projected "silver" data from source "gold" data (left). Downstream models are trained on a combination of gold and silver data (right). Components in boxes have learned parameters.

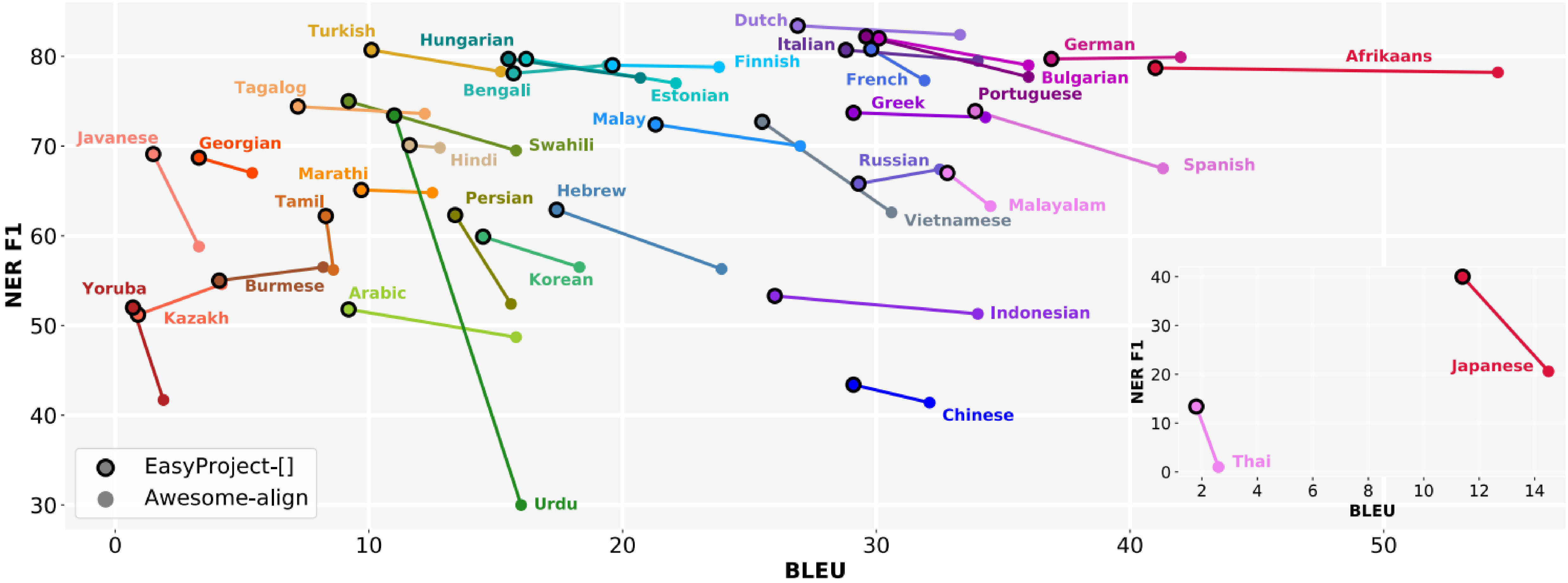


Figure 2: Comparison of translation quality and end-task performance for different label projection methods on the WikiANN dataset. EasyProject (§3.3) outperforms the alignment-based approach on F_1 scores for most languages, although inserting span markers degrade translation quality. The detailed experimental setting is in §4.1.

Tasks

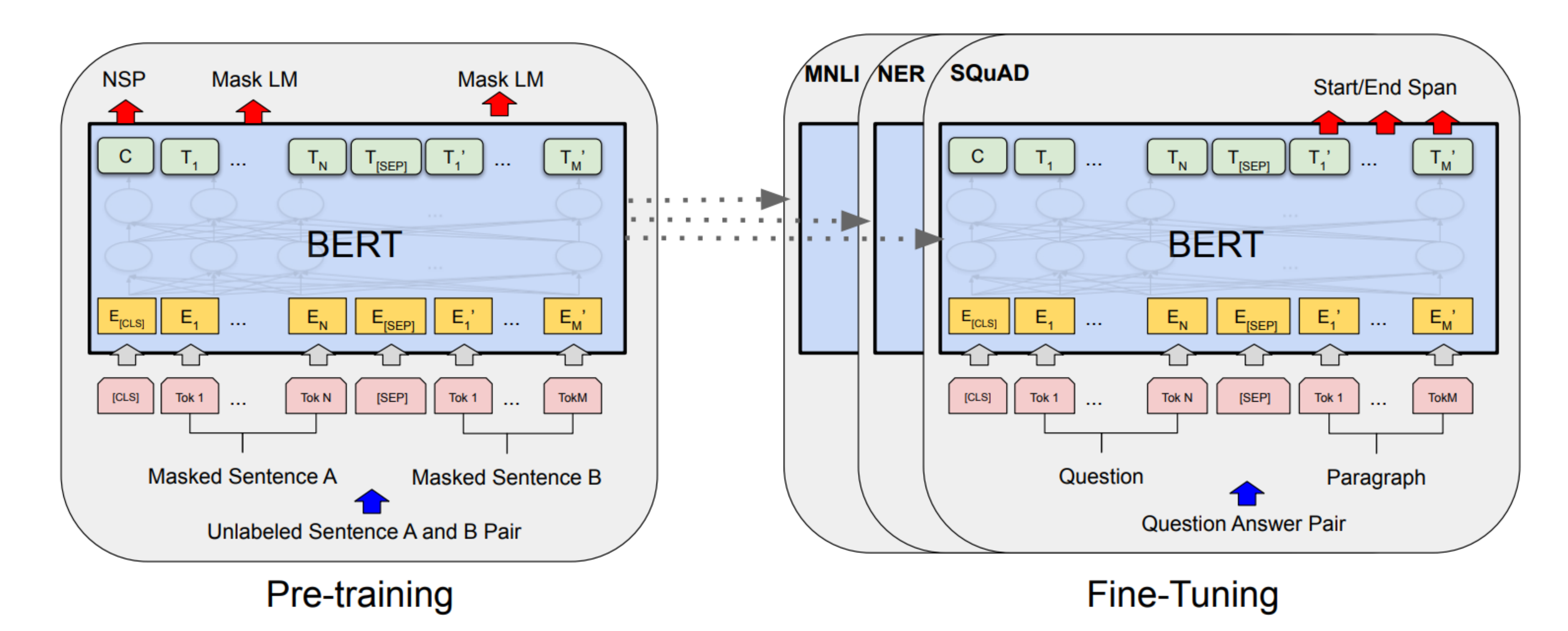
- Information Extraction
- Information Retrieval
- CLIR
- MLIR
- Cross-Lingual Semantics
- SLU
- Question Answering
- Bilingual Lexicon Induction
- Code-Switching/Mixing
- Dialogue Systems
- Speech Recognition
- Speech Synthesis
- Representation Learning
- Many many more…

Low-Resource

Models

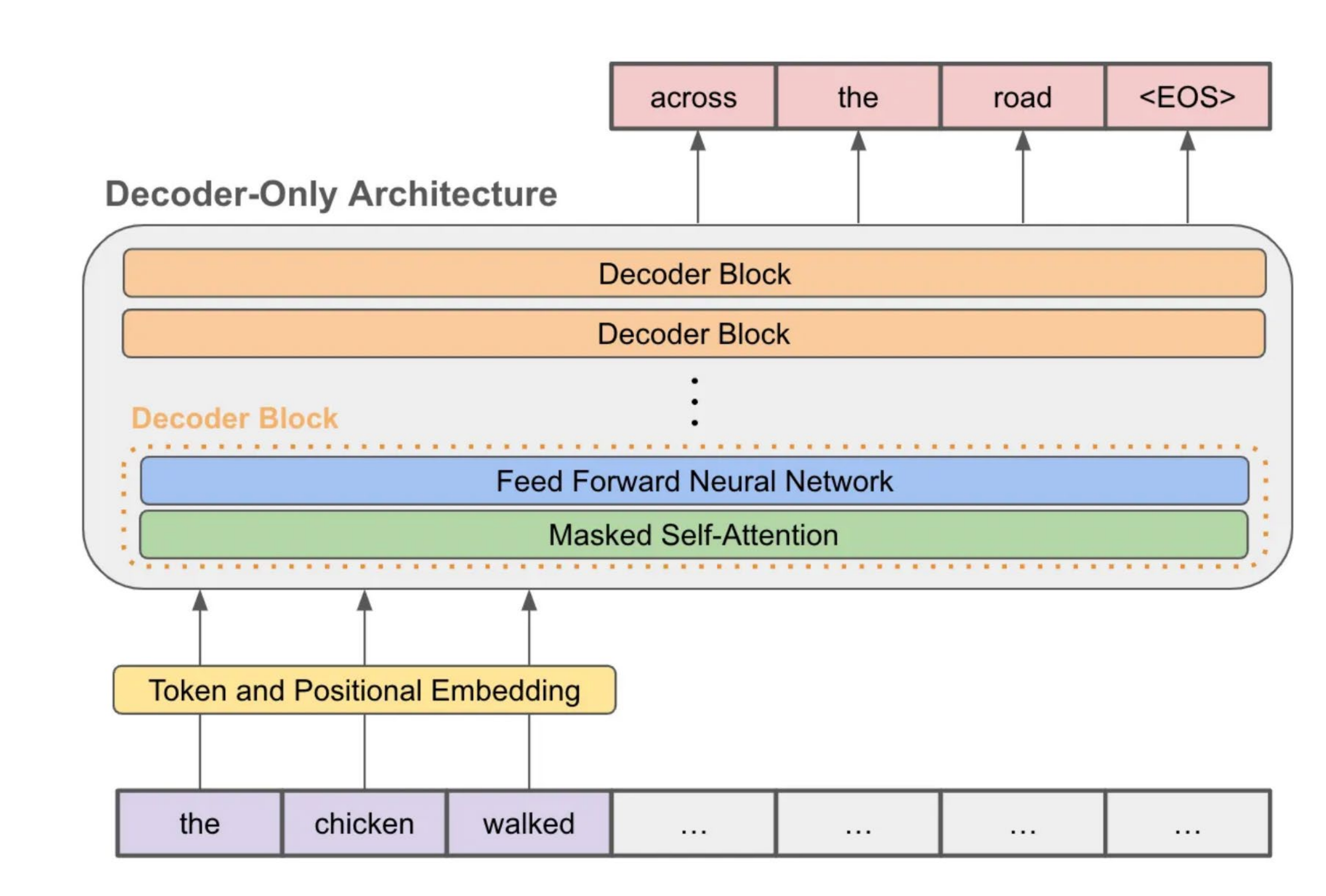
- Pretty much all neural models today
- Mostly transformer based
- Pre-trained models frequently help a lot
- Lots and lots of data

Encoder Models



Devlin et al., 2018

Decoder Models



Depiction of a decoder-only language modeling architecture (created by author)

https://towardsdatascience.com/understanding-the-open-pre-trained-transformers-opt-library-193a29c14a15
Accessed 4/25/23

Encoder-Decoder Models

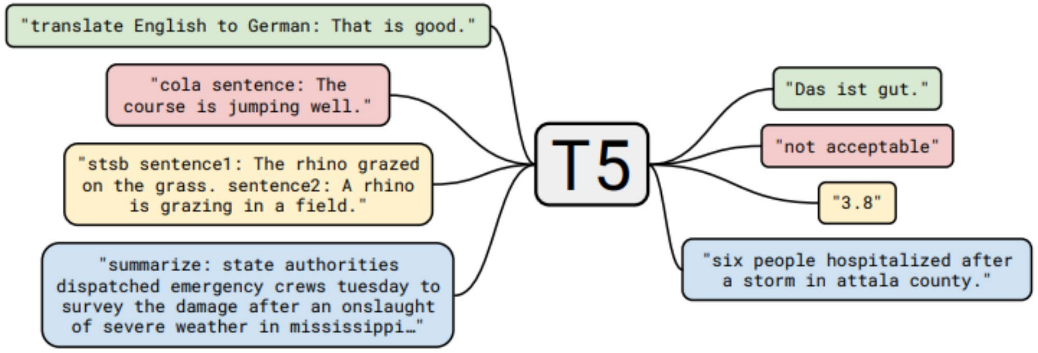
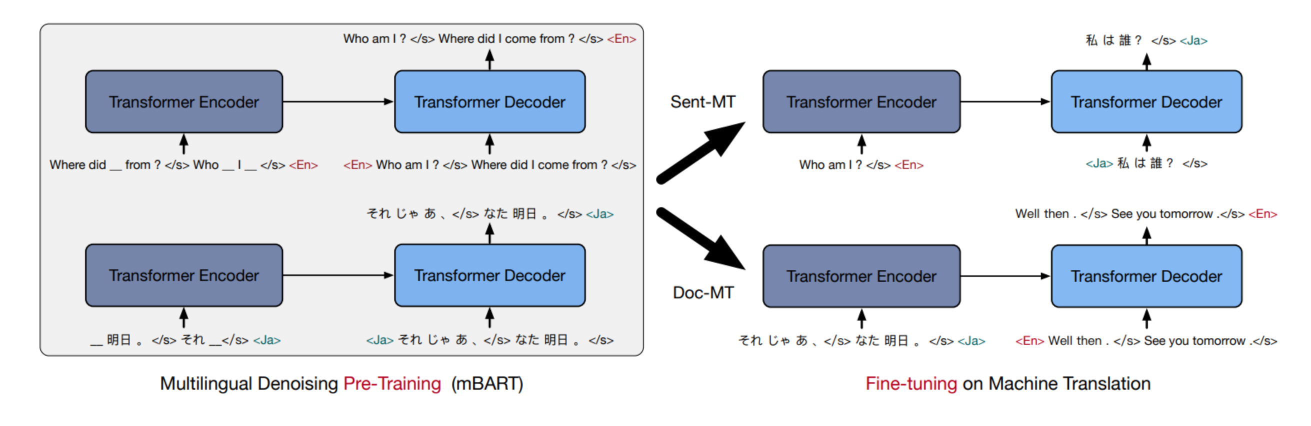


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".



Raffel et al., 2019

Liu et al., 2020

Question Answering

Methods

Rank	Model	EM	F1	Paper	Code	Result	Year	Tags
1	**ByT5** (fine-tuned)	81.9		ByT5: Towards a token-free future with pre-trained byte-to-byte models			2021	fine-tuned
2	**U-PaLM 62B** (fine-tuned)	78.4	88.5	Transcending Scaling Laws with 0.1% Extra Compute			2022	fine-tuned
3	**Flan-U-PaLM 540B** (direct-prompting)	68.3		Scaling Instruction-Finetuned Language Models			2022	
4	**Flan-PaLM 540B** (direct-prompting)	67.8		Scaling Instruction-Finetuned Language Models			2022	
5	**ByT5 XXL**	60.0	75.3	ByT5: Towards a token-free future with pre-trained byte-to-byte models			2021	
6	**U-PaLM-540B** (CoT)	54.6		Transcending Scaling Laws with 0.1% Extra Compute			2022	chain-of-thought
7	**PaLM-540B** (CoT)	52.9		PaLM: Scaling Language Modeling with Pathways			2022	chain-of-thought

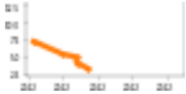
Multilingual LibriSpeech (MLS)

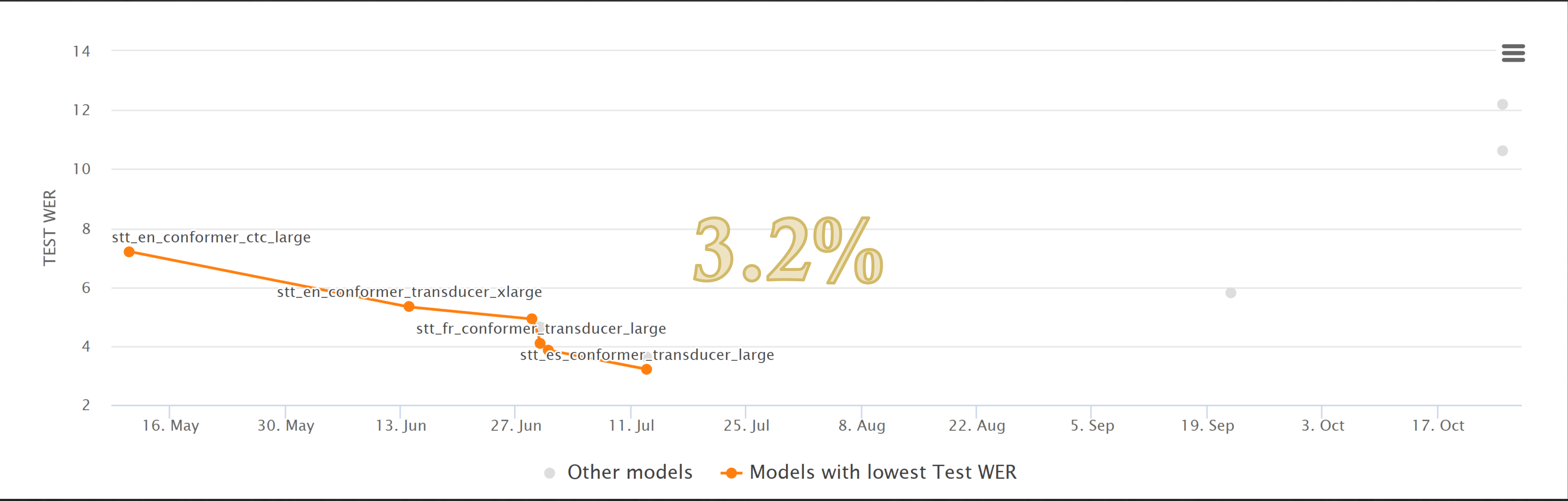
Introduced by Pratap et al. in MLS: A Large-Scale Multilingual Dataset for Speech Research

Multilingual LibriSpeech is a large multilingual corpus suitable for speech research. The dataset is derived from read audiobooks from LibriVox and consists of 8 languages - English, German, Dutch, Spanish, French, Italian, Portuguese, Polish. It includes about 44.5K hours of English and a total of about 6K hours for other languages.

Benchmarks

Edit

Trend	Task	Dataset Variant	Best Model	Paper	Code
	**Speech Recognition**	Multilingual LibriSpeech	stt_es_conformer_transducer_large		
	**Automatic Speech Recognition**	Multilingual LibriSpeech	openai/whisper-medium		

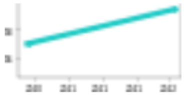

14
12
10
8
6
4
2
TEST WER
stt_en_conformer_ctc_large
stt_en_conformer_transducer_xlarge
stt_fr_conformer_transducer_large
stt_es_conformer_transducer_large
3.2%
16. May
30. May
13. Jun
27. Jun
11. Jul
25. Jul
8. Aug
22. Aug
5. Sep
19. Sep
3. Oct
17. Oct
Other models
Models with lowest Test WER


40
20
0
WER
Whisper Small Es - Sanchit Gandhi

4.019

Benchmarks

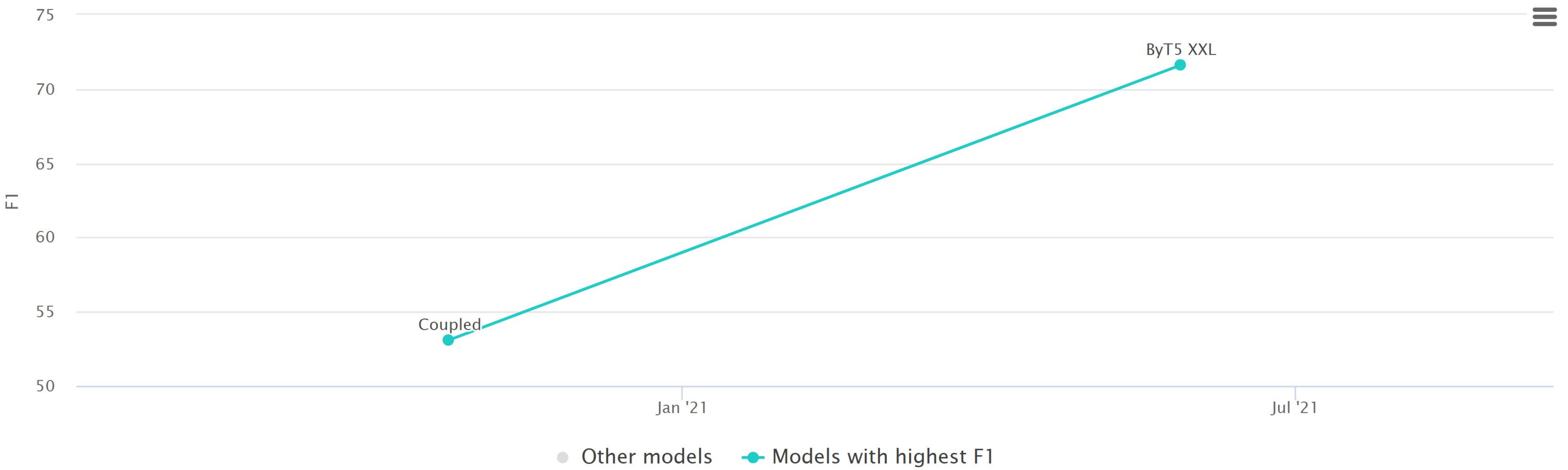
Edit

Trend	Task	Dataset Variant	Best Model	Paper	Code
	**Sequence-to-sequence Language Modeling**	MLSUM	mt5-small-test-ged-mlsum_max_target_length_10		
	**Abstractive Text Summarization**	mlsum-es	marimari-r2r-mlsum		
	**Abstractive Text Summarization**	MLSum-it	mBART		
	**Summarization**	mlsum tu	mt5-base-turkish-sum		
	**Summarization**	MLSUM de	t5-seven-epoch-base-german		

MLQA (MultiLingual Question Answering)

Introduced by Lewis et al. in MLQA: Evaluating Cross-lingual Extractive Question Answering

MLQA (MultiLingual Question Answering) is a benchmark dataset for evaluating cross-lingual question answering performance. MLQA consists of over 5K extractive QA instances (12K in English) in SQuAD format in seven languages - English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese. MLQA is highly parallel, with QA instances



MLDoc (Multilingual Document Classification Corpus)

Edit

Introduced by Schwenk et al. in A Corpus for Multilingual Document Classification in Eight Languages

Multilingual Document Classification Corpus (**MLDoc**) is a cross-lingual document classification dataset covering English, German, French, Spanish, Italian, Russian, Japanese and Chinese. It is a subset of the Reuters Corpus Volume 2 selected according to the following design choices:

- uniform class coverage: same number of examples for each class and language,
- official train / development / test split: for each language a training data of different sizes (1K, 2K, 5K and 10K

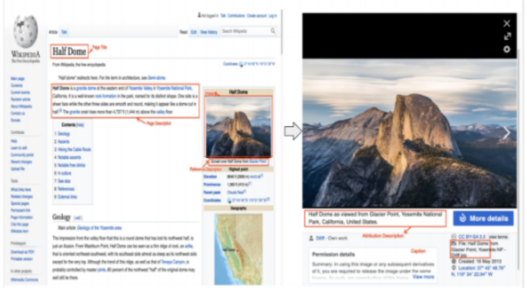
Usage

Rank	Model	Accuracy	Paper	Code	Result	Year	Tags
1	**XLMft UDA**	96.05	Bridging the domain gap in cross-lingual document classification			2019	
2	**MultiFiT, pseudo**	89.42	MultiFiT: Efficient Multi-lingual Language Model Fine-tuning			2019	
3	**Massively Multilingual Sentence Embeddings**	77.95	Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond			2018	
4	**BiLSTM** (UN)	74.52	A Corpus for Multilingual Document Classification in Eight Languages			2018	LSTM
5	**BiLSTM** (Europarl)	72.83	A Corpus for Multilingual Document Classification in Eight Languages			2018	LSTM

WIT (Wikipedia-based Image Text)

Edit

Introduced by Srinivasan et al. in WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning



Wikipedia-based Image Text (**WIT**) Dataset is a large multimodal multilingual dataset. WIT is composed of a curated set of 37.6 million entity rich image-text examples with 11.5 million unique images across 108 Wikipedia languages. Its size enables WIT to be used as a pretraining dataset for multimodal machine learning models.

Key Advantages

A few unique advantages of WIT:

- The largest multimodal dataset (time of this writing) by the number of image-text examples.
- A massively multilingual (first of its kind) with coverage for over 100+ languages.
- A collection of diverse set of concepts and real world entities.
- Brings forth challenging real-world test sets.

Homepage

Usage

Rank	Model	R@1	R@5	Paper	Code	Result	Year	Tags
1	**WIT-ALL**	0.346	0.642	WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning			2021	
2	**CC** (Conceptual Captions)	0.048	0.122	WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning			2021	

WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning

Krishna Srinivasan
Google
krishnaps@google.com

Karthik Raman
Google
karthikraman@google.com

Jiecao Chen
Google
chenjiecao@google.com

Michael Bendersky
Google
bemike@google.com

Marc Najork
Google
najork@google.com

2021

"…the text encoder, we used a bag of words model (with ngrams of size 1 and 2). Each ngram was mapped to an a one amongst a million vocabulary buckets using a hash-function to get a 200D embedding. These ngram embeddings were then summed and passed through a simple FFNN and projected to a final 64D embedding, to match the size of the image encoder embedding. The final activation function we used was ReLU."

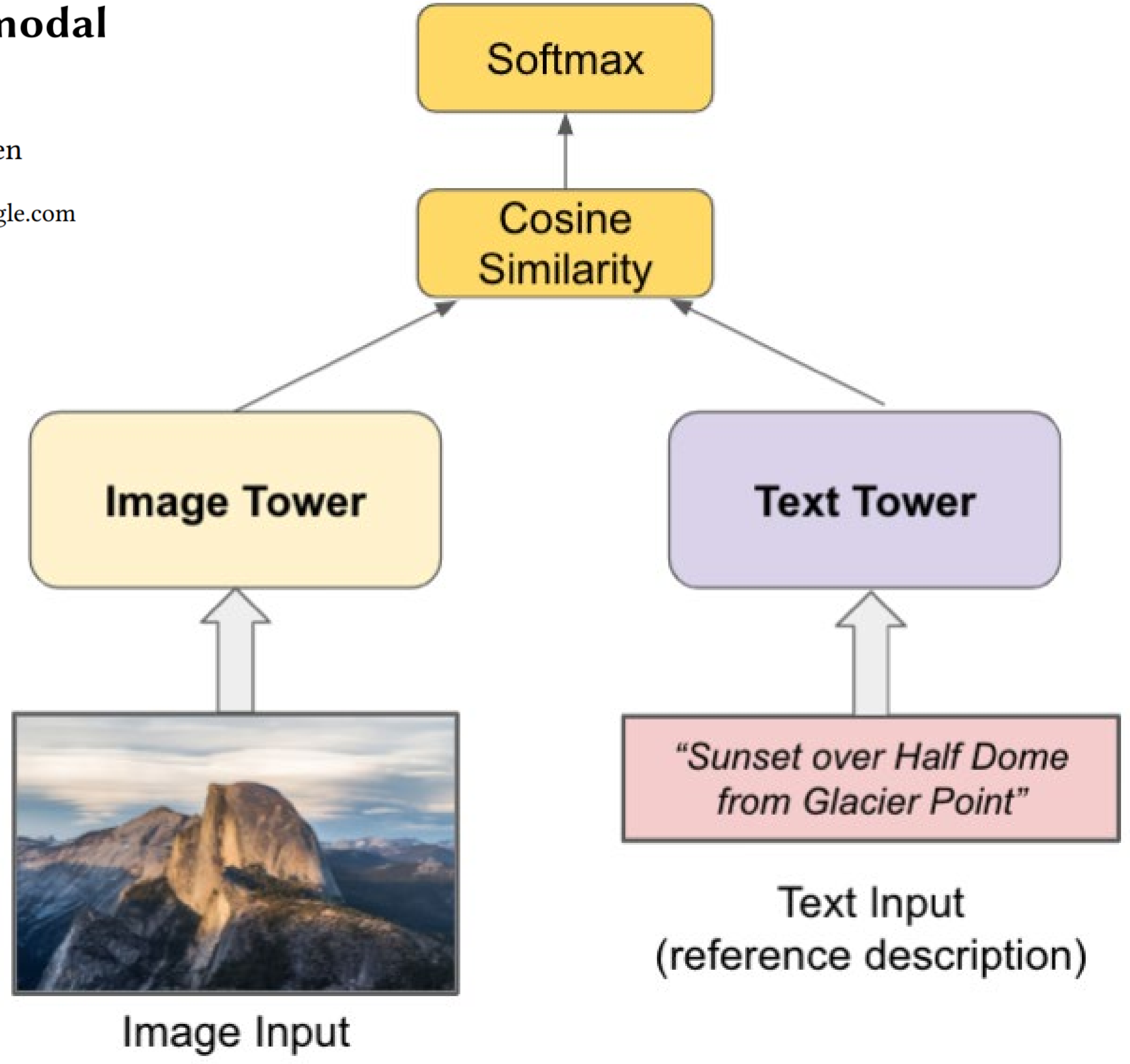


Figure 4: WIT Dual Encoder Model for Training.

FooDI-ML (Food Drinks and groceries Images Multi Lingual)

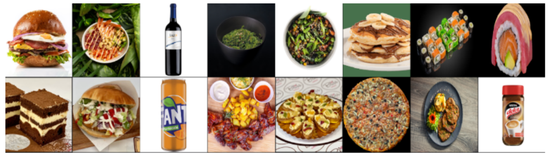
Edit

Introduced by Olóndriz et al. in FooDI-ML: a large multi-language dataset of food, drinks and groceries images and descriptions

Food Drinks and groceries Images Multi Lingual (FooDI-ML) is a dataset that contains over 1.5M unique images and over 9.5M store names, product names descriptions, and collection sections gathered from the Glovo application. The data made available corresponds to food, drinks and groceries products from 37 countries in Europe, the Middle East, Africa and Latin America. The dataset comprehends 33 languages, including 870K samples of languages of countries from Eastern Europe and Western Asia such as Ukrainian and Kazakh, which have been so far underrepresented in publicly available visiolinguistic datasets. The dataset also includes widely spoken languages such as Spanish and English.

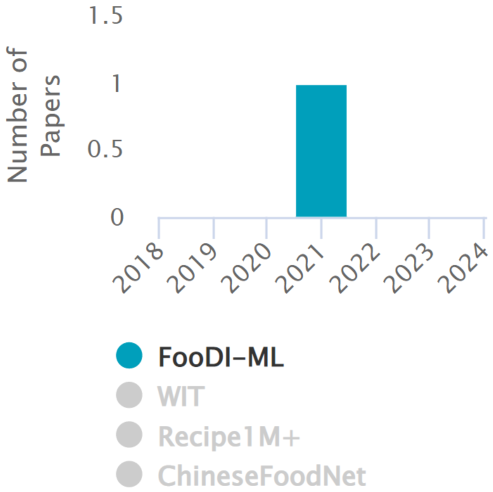
Description from: FooDI-ML: a large multi-language dataset of food, drinks and groceries images and descriptions

Homepage



Source: https://github.com/Glovo/foodi-ml-dat...

Usage



Benchmarks

Edit

“In this work we use two existing SotA approaches: CLIP [19], and an adaptation of the dual-tower method used in the WIT dataset (henceforth, WIT) [23], to provide benchmark metrics for our dataset. CLIP and WIT are very similar in that both depend on fine-tuning previously independently trained encoders (such as ResNet [8] or a Transformer encoder). The main difference between the two is that CLIP is trained over a very large private dataset (400M samples). In addition, CLIP is trained maximising a symmetric binary cross-entropy loss while WIT uses only the first component of the loss (corresponding to image to text retrieval). We use CLIP in a zero-shot manner as proposed by the authors. WIT’s implementation is not publicly available, so we rewrite it and offer it publicly. Note that in our implementation we use a transformer model instead of bag of words to reflect recent advances in sentence encoding.”

GLAMI-1M (A Multilingual Image-Text Fashion Dataset)

Edit

Introduced by Kosar et al. in GLAMI-1M: A Multilingual Image-Text Fashion Dataset

We introduce GLAMI-1M: the largest multilingual image-text classification dataset and benchmark. The dataset contains images of fashion products with item descriptions, each in 1 of 13 languages. Categorization into 191 classes has high-quality annotations: all 100k images in the test set and 75% of the 1M training set were human-labeled. The paper presents baselines for image-text classification showing that the dataset presents a challenging fine-grained classification problem: The best scoring EmbraceNet model using both visual and textual features achieves 69.7%

Rank	Model	Top 1 Accuracy %	Top 5 Accuracy %	Extra Training Data	Paper	Code	Result	Year	Tags
1	**EmbraceNet** (image+text)	69.7	94.0	✓	GLAMI-1M: A Multilingual Image-Text Fashion Dataset			2022	multilingual, Multi-modal, CNN+Transformer, mT5, Transformer, ResNeXt
2	**CLIP** (zero-shot image+text)	32.3	74.5	✓	GLAMI-1M: A Multilingual Image-Text Fashion Dataset			2022	

EmbraceNet: A robust deep learning architecture for multimodal classification

Jun-Ho Choi, Jong-Seok Lee*

School of Integrated Technology, Yonsei University, 85 Songdogwahak-ro, Yeonsu-gu, Incheon, Korea

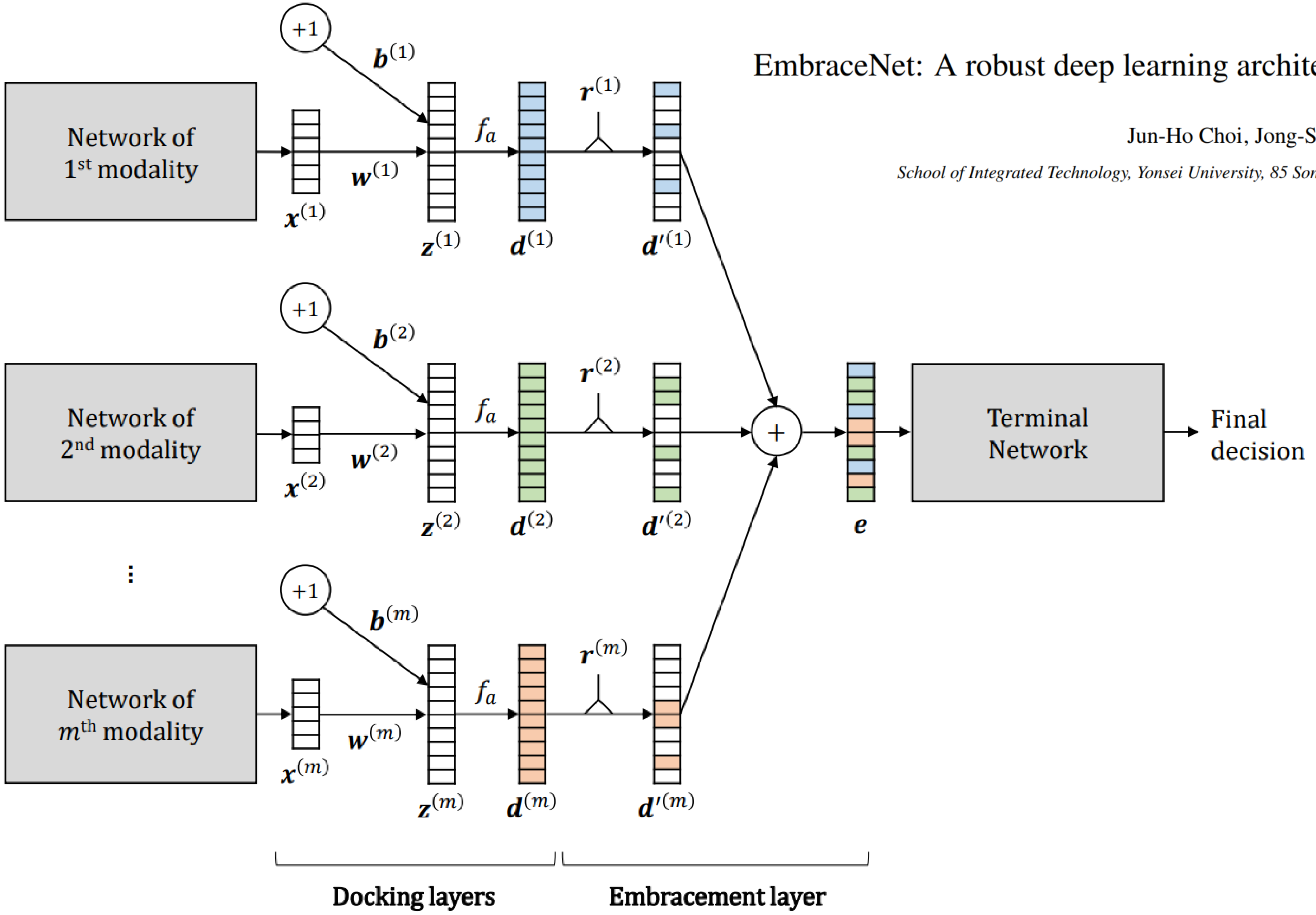


Figure 1: Overall structure of the proposed EmbraceNet model.

XTREME (Cross-Lingual Transfer Evaluation of Multilingual Encoders)

Edit

Introduced by Hu et al. in XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation

The **Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME)** benchmark was introduced to encourage more research on multilingual transfer learning,. XTREME covers 40 typologically diverse languages spanning 12 language families and includes 9 tasks that require reasoning about different levels of syntax or semantics.

The languages in XTREME are selected to maximize language diversity, coverage in existing tasks, and availability of training data. The languages in XTREME are selected to maximize language diversity, coverage in existing tasks, and availability of training data. Among these are many under-studied languages, such as the Dravidian languages Tamil (spoken in southern India, Sri Lanka, and Singapore), Telugu and Malayalam (spoken mainly in southern India), and the Niger-Congo languages Swahili and Yoruba, spoken in Africa.

Table 1. Characteristics of the datasets in XTREME for the zero-shot transfer setting. For tasks that have training and dev sets in other languages, we only report the English numbers. We report the number of test examples per target language and the nature of the test sets (whether they are translations of English data or independently annotated). The number in brackets is the size of the intersection with our selected languages. For NER and POS, sizes are in sentences. Struct. pred.: structured prediction. Sent. retrieval: sentence retrieval.

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

Usage

16	**T-ULRv2 + StableTune**	80.7	88.8	75.4	72.9	89.3	InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training	2020
17	**Anonymous3**	79.9	88.2	74.6	71.7	89.0		2021
18	**FILTER**	77.0	87.5	71.9	68.5	84.4	FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding	2020
19	**Creative**	76.5	86.3	90.8	59.7	77.5		2021
20	**X-STILTs**	73.5	83.9	69.4	67.2	76.5	English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too	2020
21	**RemBERT**	56.1	84.1	73.3	68.6	NA		2020
22	**Anonymous5**	53.1	75.3	66.9	52.5	18.0		2021
23	**mT5**	40.9	89.8	NA	73.6	NA	mT5: A massively multilingual pre-trained text-to-text transformer	2020
24	**Anonymous6**	39.3	44.2	0.0	65.5	34.5		2022
25	**mBERT**	59.6	73.7	66.3	53.8	47.7	XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation	2019

6	**Turing ULR v5** (XLM-E)	83.7	90.0	81.4	74.3	93.7				2021
7	**InfoXLM-XFT**	82.2	89.3	75.5	75.2	92.4				2021
8	**Ensemble-Distil-XFT** (ED-XFT)	82.0	89.2	74.6	75.2	92.4				2022
9	**VECO**	82.0	89.0	76.7	73.4	93.3				2021
10	**VECO + HICTL**	82.0	89.0	76.7	73.4	93.3				2021
11	**Polyglot**	81.7	88.3	80.6	71.9	90.8				2021
12	**Unicoder+ZCode**	81.6	88.4	76.2	72.5	93.7				2021
13	**Unicoder + ZCode**	81.6	88.4	76.2	72.5	93.7				2021
14	**ERNIE-M**	80.9	87.9	75.6	72.3	91.9	ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora			2020
15	**HiCTL**	80.8	89.0	74.4	71.9	92.6				2021

1	**Turing ULR v6**	85.5	91.0	83.8	77.1	94.4	XLM-E: Cross-lingual Language Model Pre-training via ELECTRA			2022
2	**MShenNonG**	85.0	90.4	83.1	76.3	94.4				2022
3	**MShenNonG+TDT**	85.0	90.4	83.1	76.3	94.4				2022
4	**Turing ULR v5**	84.5	90.3	81.7	76.3	93.7	XLM-E: Cross-lingual Language Model Pre-training via ELECTRA			2021
5	**CoFe**	84.1	90.1	81.4	75.0	94.2				2021

ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

Kevin Clark
Stanford University
kevclark@cs.stanford.edu

Minh-Thang Luong
Google Brain
thangluong@google.com

Quoc V. Le
Google Brain
qvl@google.com

Christopher D. Manning
Stanford University & CIFAR Fellow
manning@cs.stanford.edu

2020

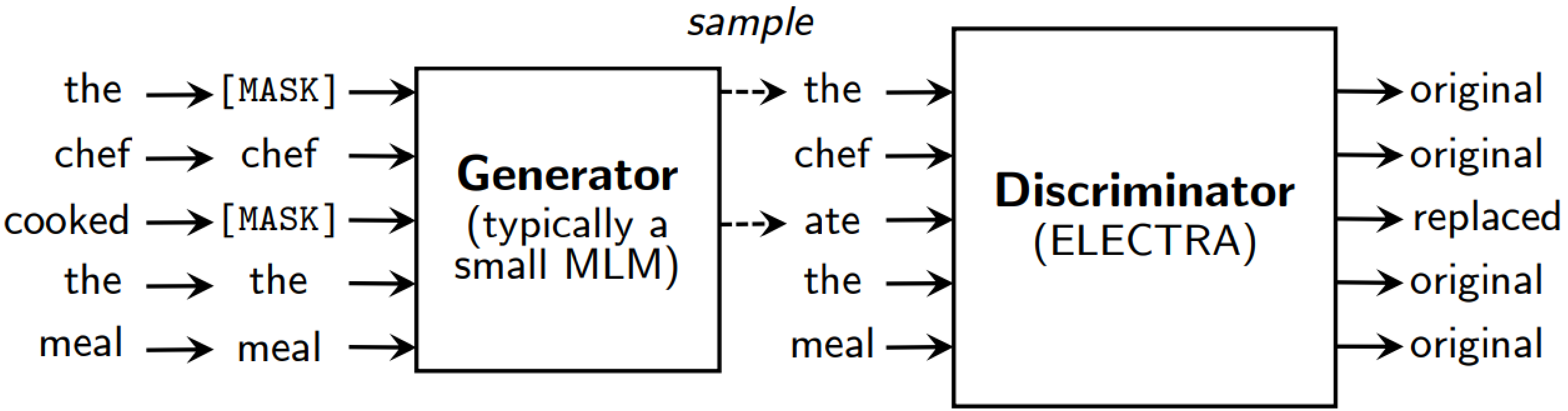


Figure 2: An overview of replaced token detection. The generator can be any model that produces an output distribution over tokens, but we usually use a small masked language model that is trained jointly with the discriminator. Although the models are structured like in a GAN, we train the generator with maximum likelihood rather than adversarially due to the difficulty of applying GANs to text. After pre-training, we throw out the generator and only fine-tune the discriminator (the ELECTRA model) on downstream tasks.

XLM-E: Cross-lingual Language Model Pre-training via ELECTRA

Zewen Chi[†‡]*, Shaohan Huang[‡]*, Li Dong[‡], Shuming Ma[‡], Bo Zheng[‡], Saksham Singhal[‡]
Payal Bajaj[‡], Xia Song[‡], Xian-Ling Mao[†], Heyan Huang[†], Furu Wei[‡]
[†] Beijing Institute of Technology
[‡] Microsoft Corporation
https://github.com/microsoft/unilm

Apply ELECTRA-style tasks to cross-lingual language model pre-training

X-Fact

Edit

Introduced by Gupta et al. in X-FACT: A New Benchmark Dataset for Multilingual Fact Checking

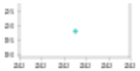
X-FACT is a large publicly available multilingual dataset for factual verification of naturally existing real-world claims. The dataset contains short statements in 25 languages and is labeled for veracity by expert fact-checkers. The dataset includes a multilingual evaluation benchmark that measures both out-of-domain generalization, and zero-shot capabilities of the multilingual models.

Claim	*Muslimische Gebete sind Pflichtprogramm an katholischer Schule.* Muslim prayers are compulsory in Catholic schools.
Label	Mostly-False (*Grösstenteils Falsch*)
Claimant	Freie Welt
Language	German
Source	de.correctiv.org
Claim Date	March 16, 2018
Review Date	March 23, 2018
Claim	*Temos, hoje, a despesa de Previdência Social representando 57% do orçamento.* Today, we have Social Security expenses representing 57% of the budget.
Label	Partly-True (*Exagerado*)
Claimant	Henrique Meirelles
Language	Portuguese (Brazilian)
Source	pt.piaui.folha.uol.com.br
Claim Date	None
Review Date	May 2, 2018

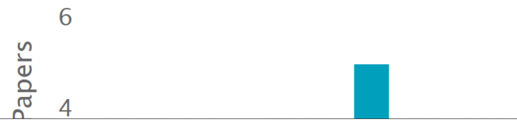
Homepage

Benchmarks

Edit

Trend	Task	Dataset Variant	Best Model	Paper	Code
	**Zero-shot Cross-lingual Fact-checking**	X-Fact	CONCRETE		

Usage



Rank	Model	F1	Paper	Code	Result	Year	Tags
1	**CONCRETE**	19.83	CONCRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval			2022	

2022

CONCRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval

Kung-Hsiang Huang **ChengXiang Zhai** **Heng Ji**
Department of Computer Science, University of Illinois Urbana-Champaign
{khhuang3, czhai, hengji}@illinois.edu

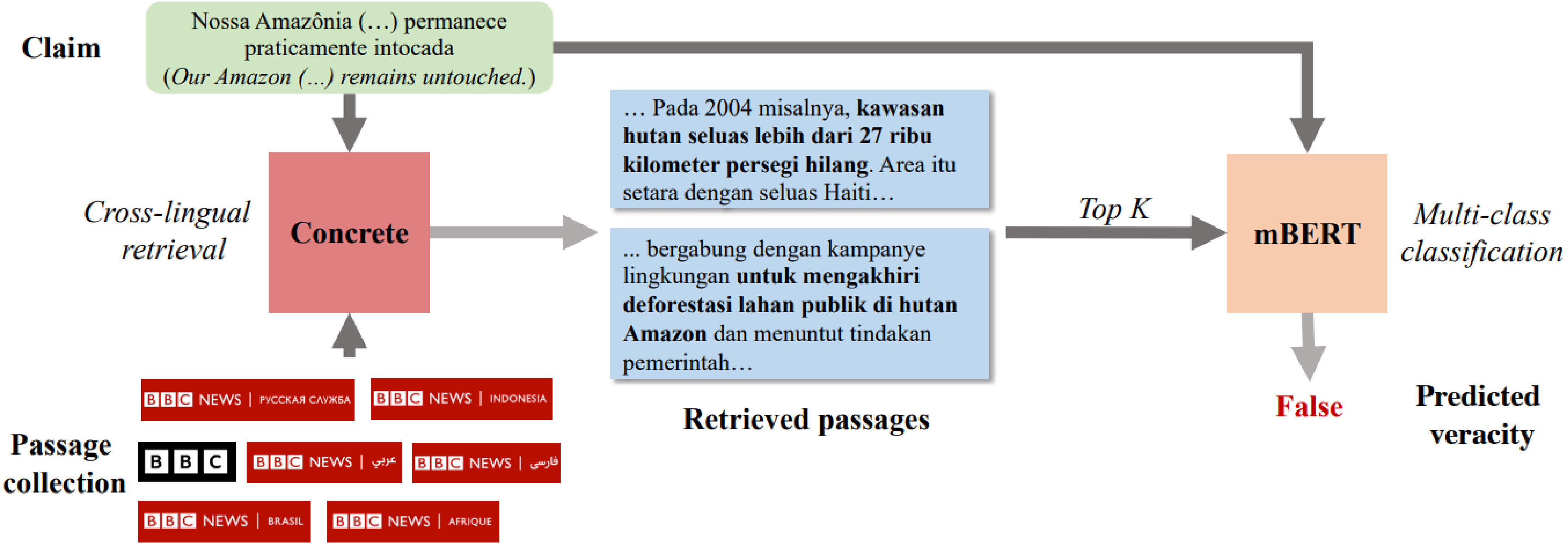


Figure 1: An overview of the proposed framework. Given a claim in arbitrary language, a cross-lingual retriever, CONCRETE retrieves relevant passages in *any languages*. The top-k relevant passages and the claim are then passed to our multilingual reader, mBERT, to predict the veracity of the claim.

VoxPopuli

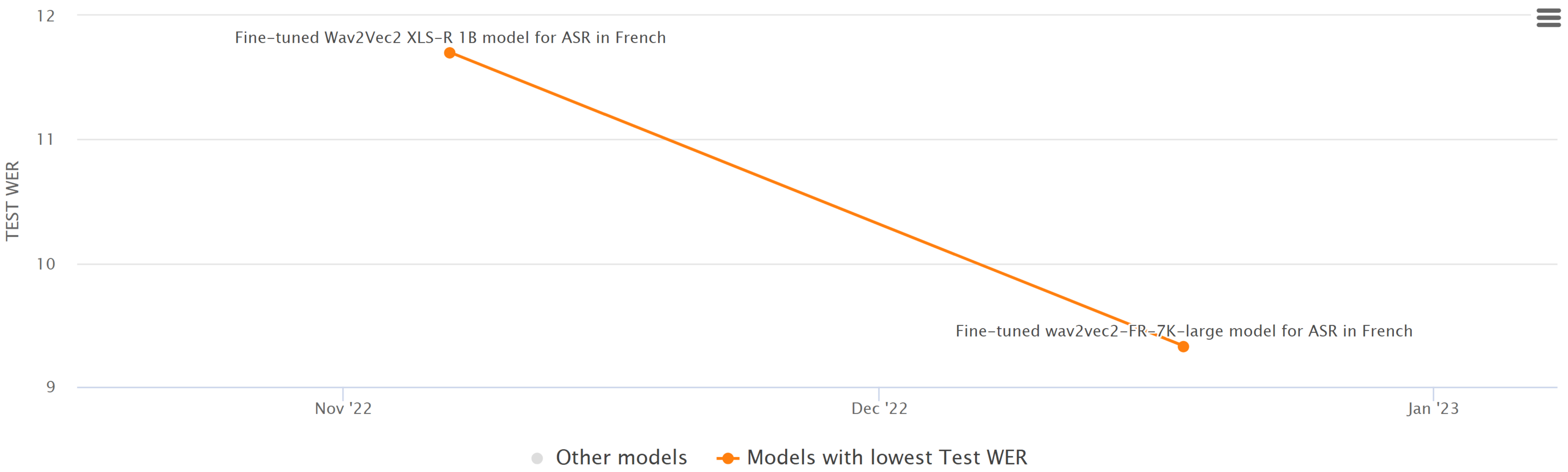
Edit

Introduced by Wang et al. in VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation

VoxPopuli is a large-scale multilingual corpus providing 100K hours of unlabelled speech data in 23 languages. It is the largest open data to date for unsupervised representation learning as well as semi-supervised learning. VoxPopuli also

Community Models   Dataset

View Test WER by Date



XNLI (Cross-lingual Natural Language Inference)

Edit

Introduced by Conneau et al. in XNLI: Evaluating Cross-lingual Sentence Representations

The **Cross-lingual Natural Language Inference** (**XNLI**) corpus is the extension of the Multi-Genre NLI (MultiNLI) corpus to 15 languages. The dataset was created by manually translating the validation and test sets of MultiNLI into each of those 15 languages. The English training set was machine translated for all languages. The dataset is composed of 122k train, 2490 validation and 5010 test examples.

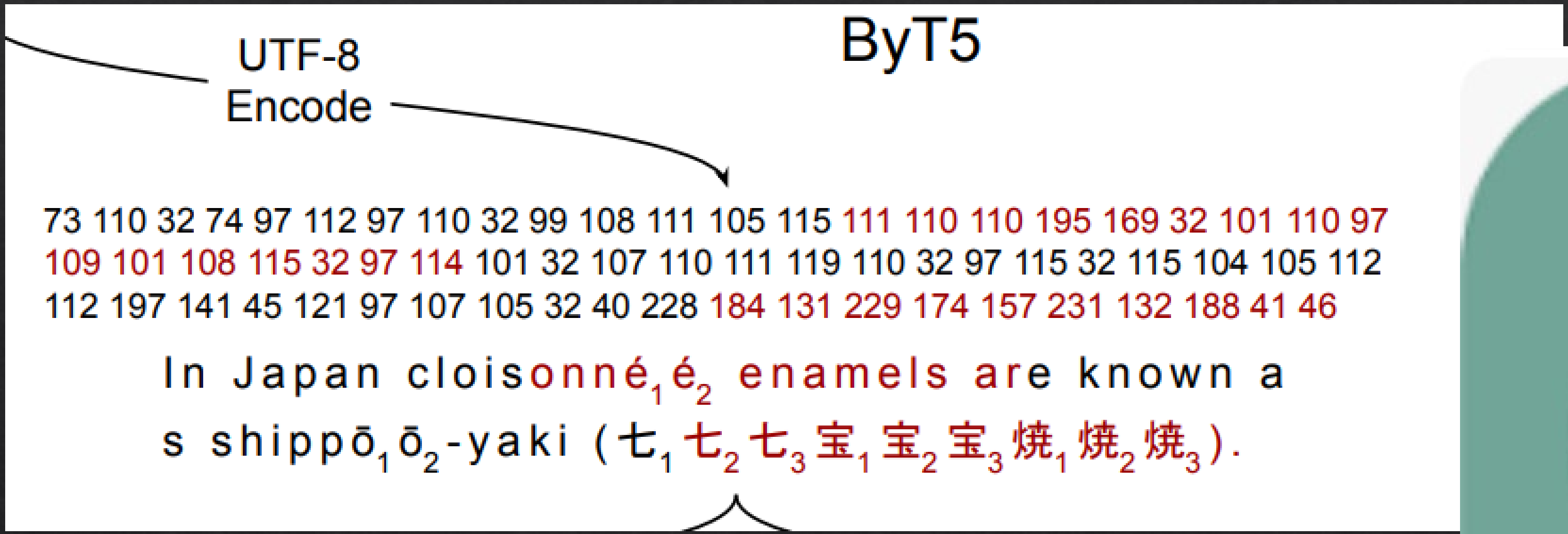
Source: CamemBERT: a Tasty French Language Model

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction

Source: https://github.com/facebookresearch/X...

Rank	Model	Accuracy	Paper	Code	Result	Year	Tags
1	**ByT5 XXL**	83.7	ByT5: Towards a token-free future with pre-trained byte-to-byte models			2021	
2	**Decoupled**	71.3	Rethinking embedding coupling in pre-trained language models			2020	
3	**Coupled**	70.7	Rethinking embedding coupling in pre-trained language models			2020	
4	**ByT5 Small**	69.1	ByT5: Towards a token-free future with pre-trained byte-to-byte models			2021	
5	**mGPT**	40.6	mGPT: Few-Shot Learners Go Multilingual			2022	

ByT5 & GPT-….


ByT5
UTF-8
Encode
73 110 32 74 97 112 97 110 32 99 108 111 105 115 111 110 110 195 169 32 101 110 97
109 101 108 115 32 97 114 101 32 107 110 111 119 110 32 97 115 32 115 104 105 112
112 197 141 45 121 97 107 105 32 40 228 184 131 229 174 157 231 132 188 41 46
In Japan cloisonné$_1$é$_2$ enamels are known a
s shippō$_1$ō$_2$-yaki (七$_1$七$_2$七$_3$宝$_1$宝$_2$宝$_3$焼$_1$焼$_2$焼$_3$).

What’s the Future?

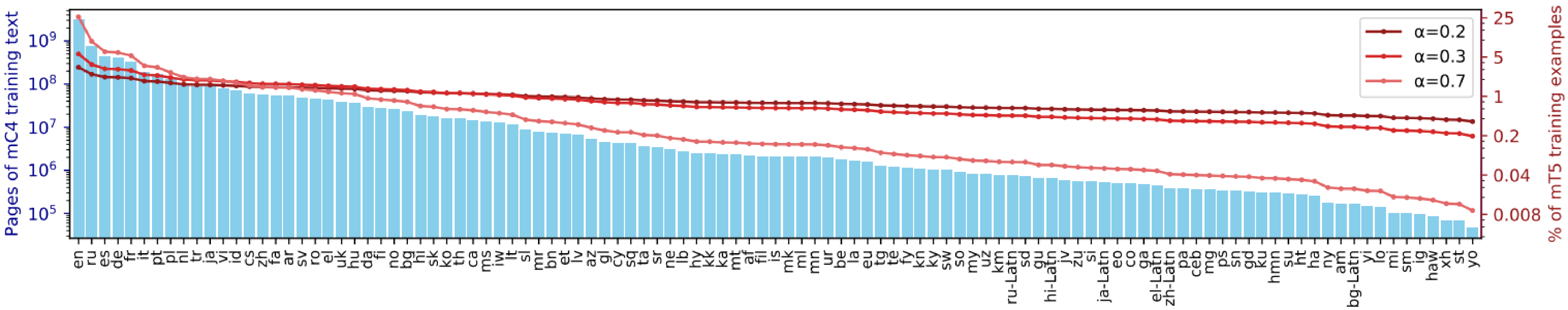


Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

Xue et al, 2020

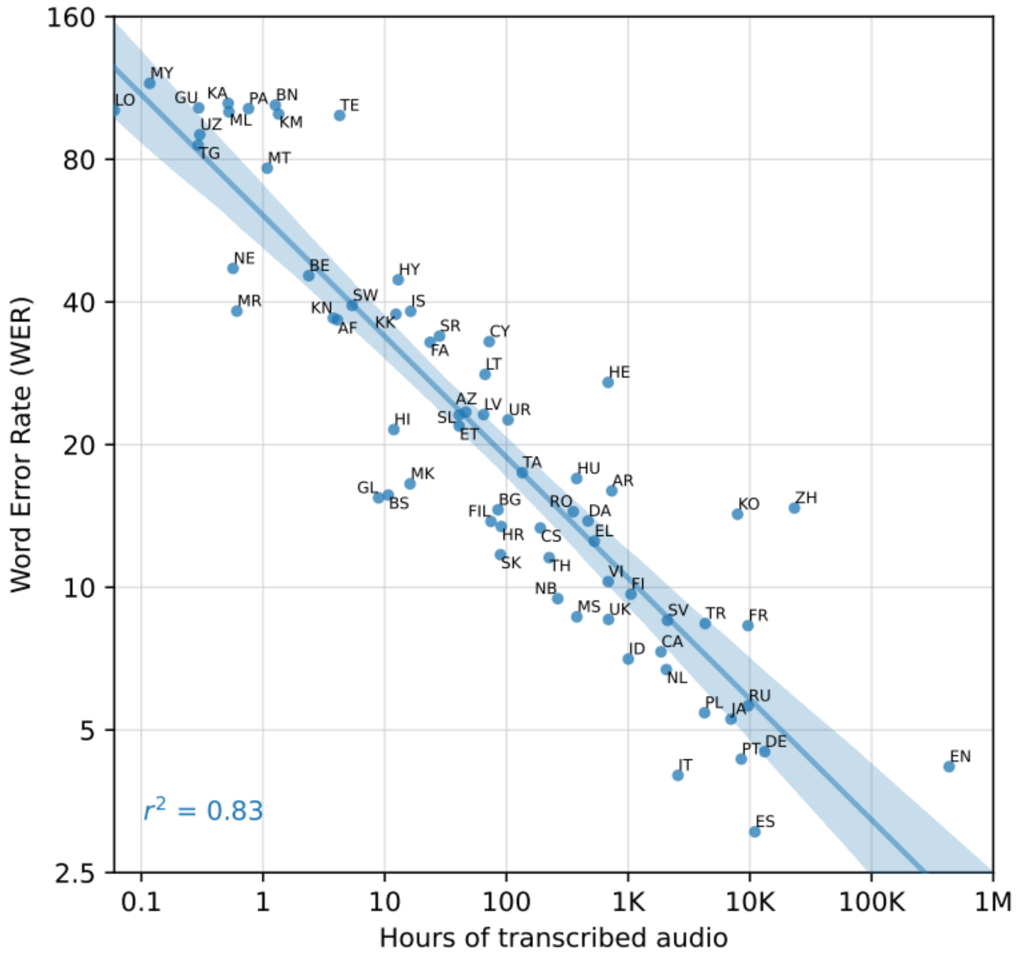


Figure 3. **Correlation of pre-training supervision amount with downstream speech recognition performance.** The amount of pre-training speech recognition data for a given language is very predictive of zero-shot performance on that language in Fleurs.

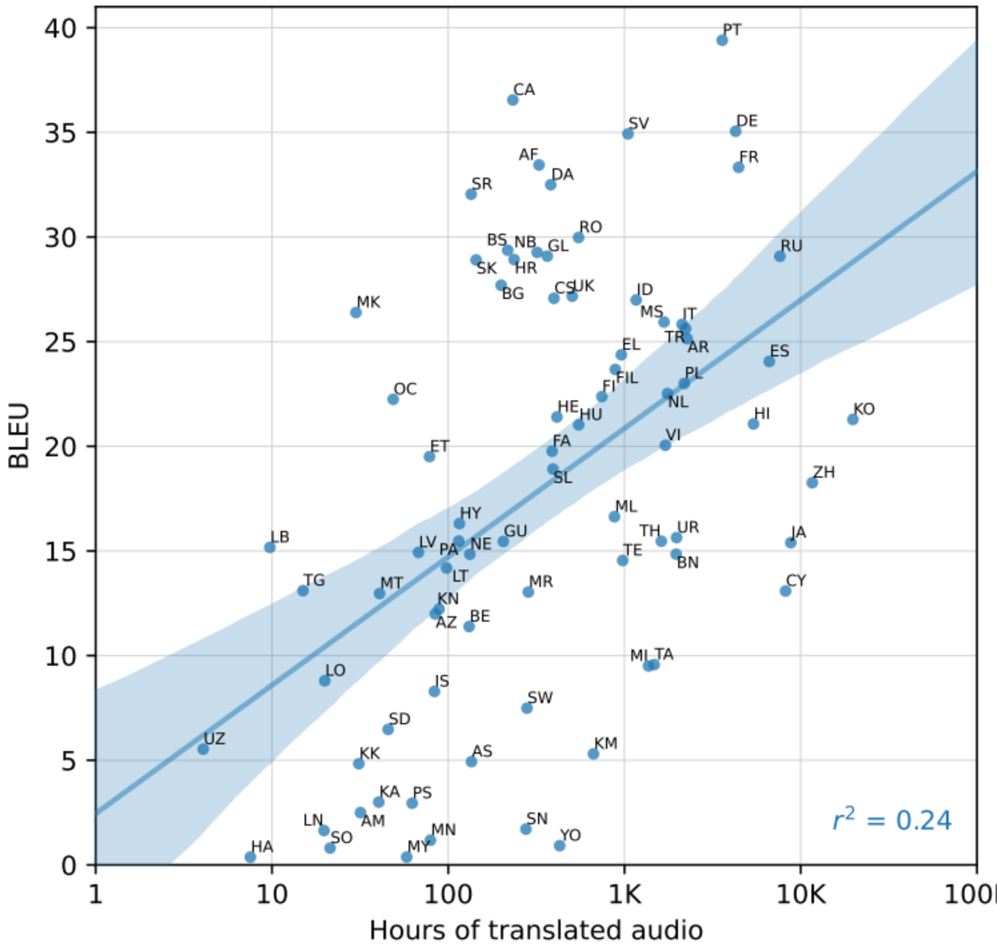


Figure 4. **Correlation of pre-training supervision amount with downstream translation performance.** The amount of pre-training translation data for a given language is only moderately predictive of Whisper's zero-shot performance on that language in Fleurs.

2 Years?

5 Years?

10 Years?