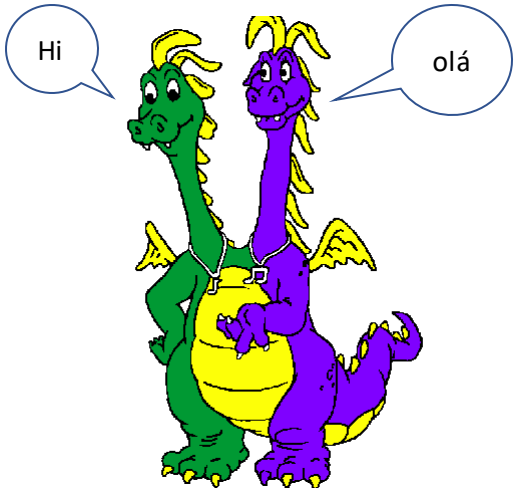
FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding

Yuwei Fang*, Shuohang Wang*, Zhe Gan, Siqi Sun, Jingjing Liu

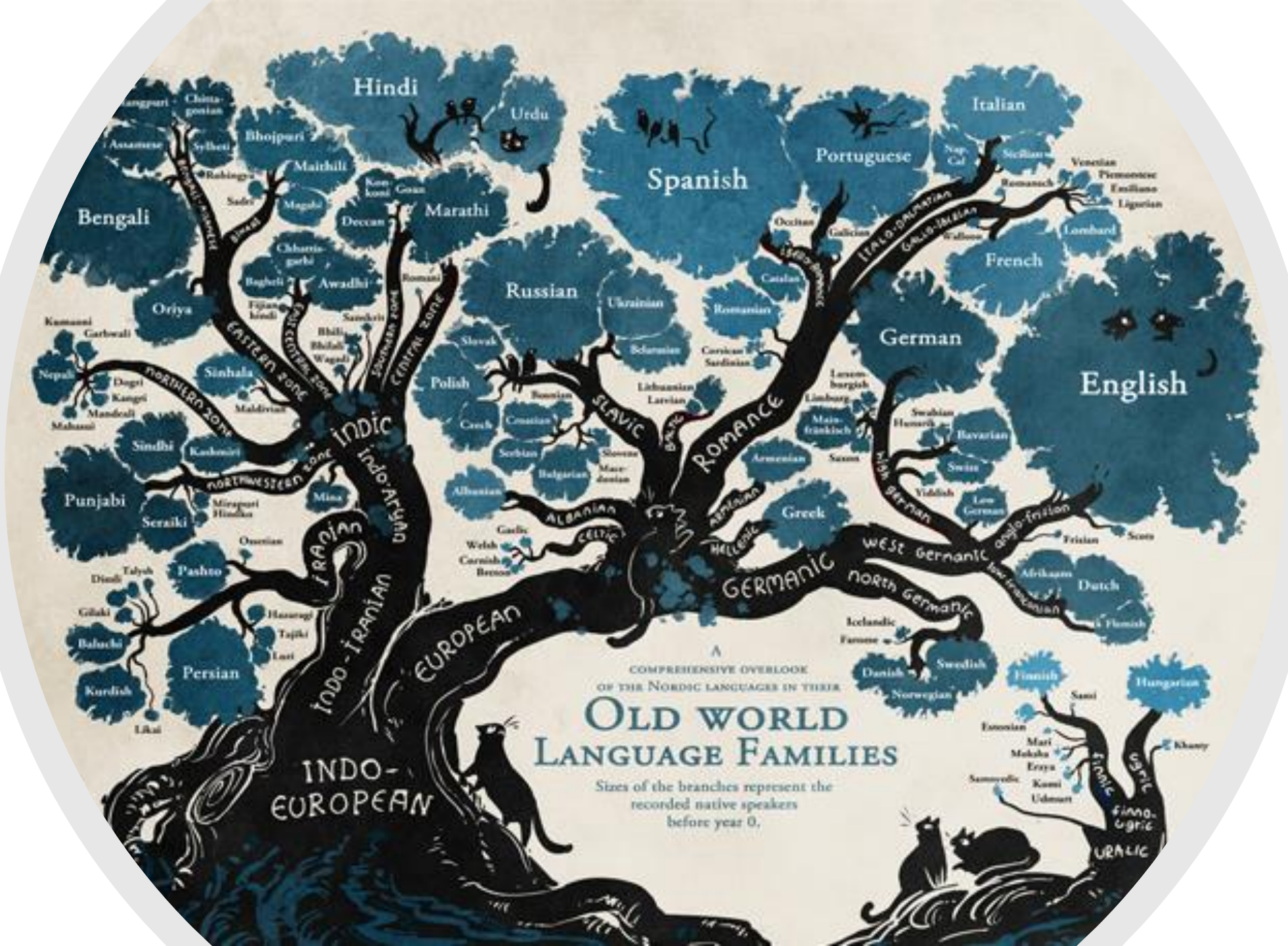


Microsoft Dynamics 365 AI Research

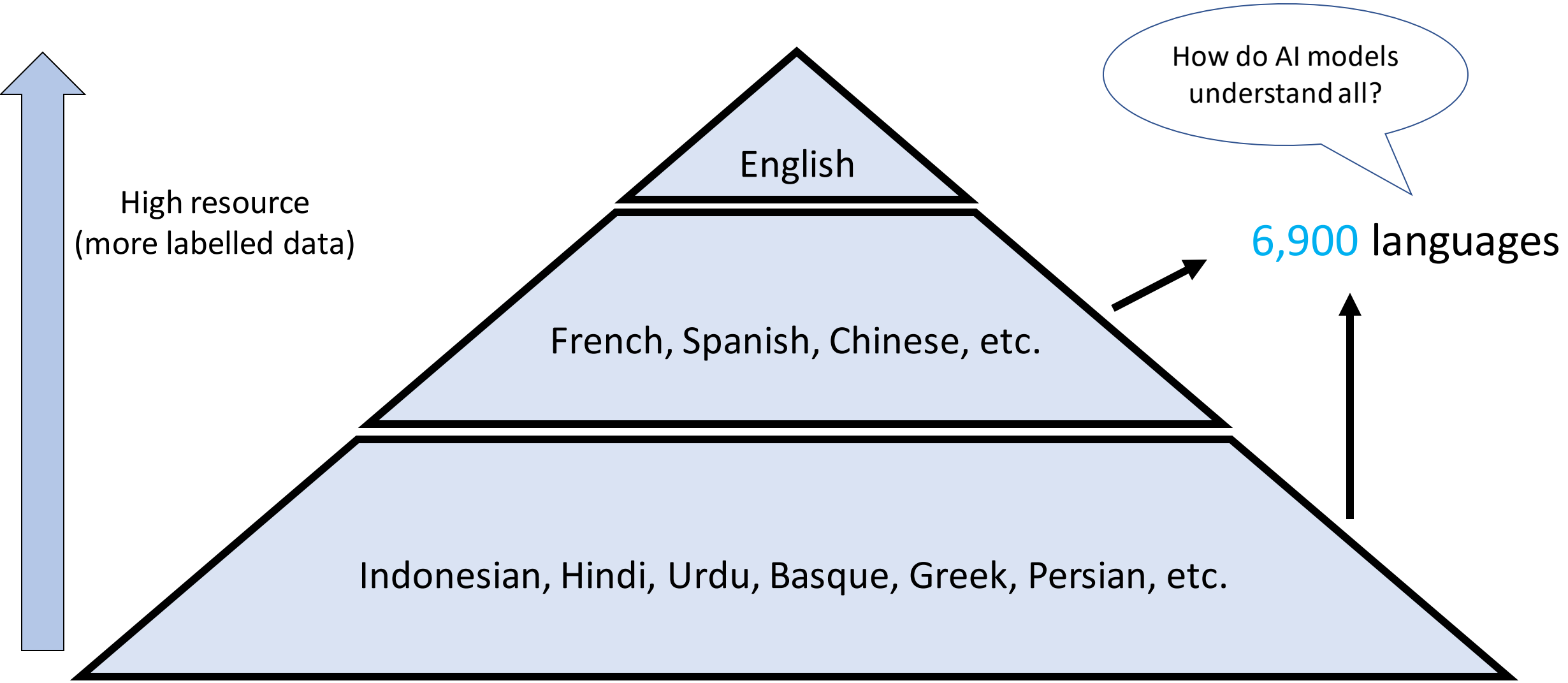
(*: Equal contribution)

Agenda

- Background on Cross-Lingual Tasks
- Motivation of FILTER
- Our FILTER Framework
  - Fusion in the Intermediate Layers of Transformer
  - A general framework for cross-lingual language model finetuning
- Experiments and Analysis
  - State-of-the-arts on XTREME and XGLUE benchmarks
  - Model Analysis


A
COMPREHENSIVE OVERLOOK
OF THE NORDIC LANGUAGES IN THEIR
OLD WORLD
LANGUAGE FAMILIES
Sizes of the branches represent the
recorded native speakers
before year 0.
INDO-EUROPEAN
INDO-IRANIAN
EUROPEAN
INDIC
IRANIAN
ROMANCE
GERMANIC
SLAVIC
BALTIC
ALBANIAN
CELTIC
HELLENIC
ARMENIAN
WEST GERMANIC
NORTH GERMANIC
ANGLO-FRISIAN
URALIC
FINNIC
UGRIC
FINNO-UGRIC
Hindi
Urdu
Bengali
Spanish
Portuguese
Italian
French
German
English
Russian
Ukrainian
Polish
Punjabi
Persian
Pashto
Greek
Albanian
Armenian
Dutch
Icelandic
Danish
Swedish
Norwegian
Finnish
Hungarian
Estonian
Sami

Distribution of Labelled Data



Kas yra Turingas?

Turing quis est?

Cine este Turing?

チューリングは誰ですか？

Kuka on Turing?

Quale hè Turing?

Sino si Turing?

Quem é Turing?

Qui est Turing?

Nor da Turing?

Кто такой Тьюринг?

Who is Turing?

Turing kimdir?

Wie is Turing?

من هو تورينج

谁是图灵？

ಟ್ಯೂರಿಂಗ್ ಯಾರು?

Turing là ai?

튜링은 누구입니까?

Kush është Turing?

Ко је Туринг?

कौन है ट्यूरिंग?

Kdo je Turing?

Turing ఎవరు?

ట్యూరింగ్ ఎవరు?

Kas yra Turingas?

チューリングは誰ですか？

Quem é Turing?

Cine este Turing?

Turing quis est?

Kuka on Turing?

Quale hè Turing?

Sino si Turing?

ಟ್ಯೂರಿಂಗ್ ಯಾರು?

Qui est Turing?

Nor da Turing?

Who is Turing?

Кто такой Тьюринг?

Turing kimdir?

Machine translation

Wie is Turing?

من هو تورينج

谁是图灵？

튜링은 누구입니까?

Turing là ai?

Kdo je Turing?

Ко је Туринг?

ట్యూరింగ్ ఎవరు?

Turing ცნობილია?

कौन है ट्यूरिंग?

Turing ცნობილია?

Kush është Turing?

Kas yra Turingas?

チューリングは誰ですか？

Quem é Turing?

Cine este Turing?

Turing quis est?

Kuka on Turing?

Quale hè Turing?

Sino si Turing?

ట్యూరింగ్ యారు?

Qui est Turing?

Nor da Turing?

Кто такой Тьюринг?

Who is Turing?

Wie is Turing?

من هو تورينج

谁是图灵？

튜링은 누구입니까?

Turing là ai?

Kdo je Turing?

Ко је Туринг?

ట్యూరింగ్ ఎవరు?

Turing ಯಾರು?

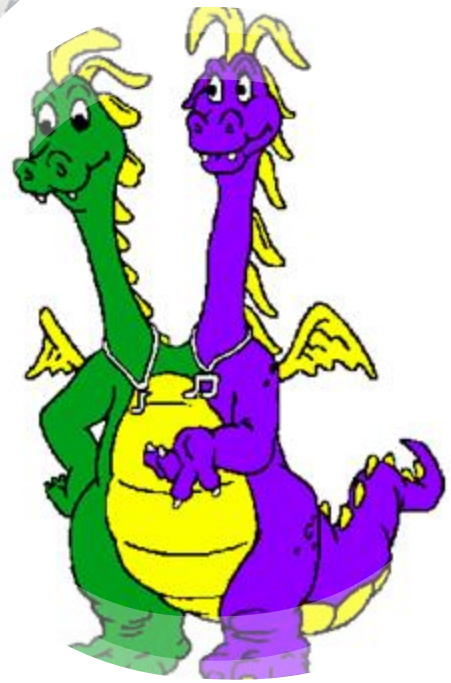
कौन है ट्यूरिंग?

Turing யார்?

Kush është Turing?

XLM

(Map into similar high dimensional space)

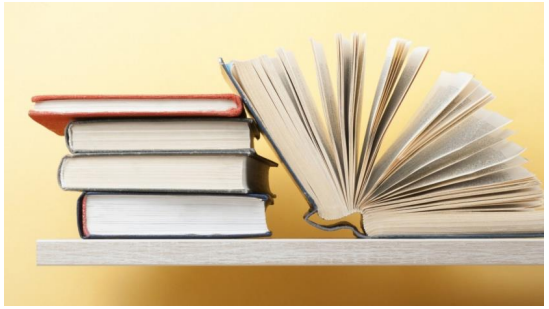


Two mountains to Climb

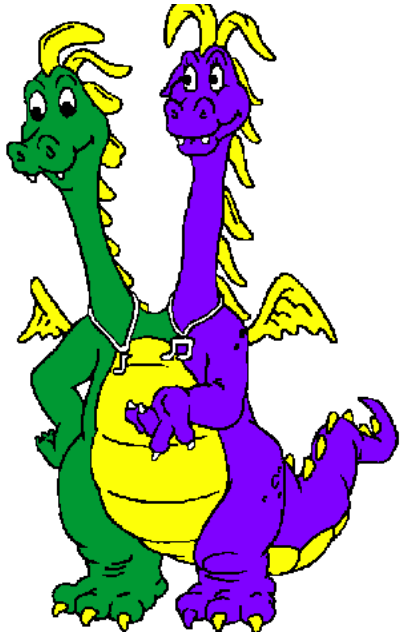
- Machine Translation (MT)
- Cross-lingual language modeling (XLM)

Cross-lingual Language Understanding

- What is Cross-lingual Language Understanding?
  - Models trained in one language can solve tasks in other languages



English



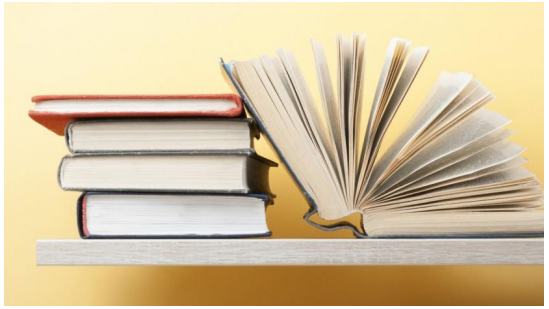
Cross-lingual Language Understanding

- What is Cross-lingual Language Understanding?
  - Models trained in one language can solve tasks in other languages

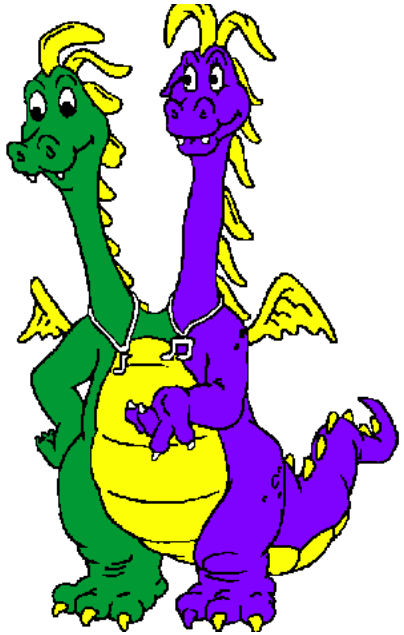
- What tasks need Cross-lingual models?
  - All tasks with text
  - NLP tasks:
    - Question Answering (QA), Information Retrieval (IR)
    - Name Entity Recognition (NER), Part-of-Speech tagging (POS)
    - Natural Language Inference (NLI), Paraphrase Identification
  - Benchmarks: XGLUE & XTREME

Cross-lingual Language Understanding

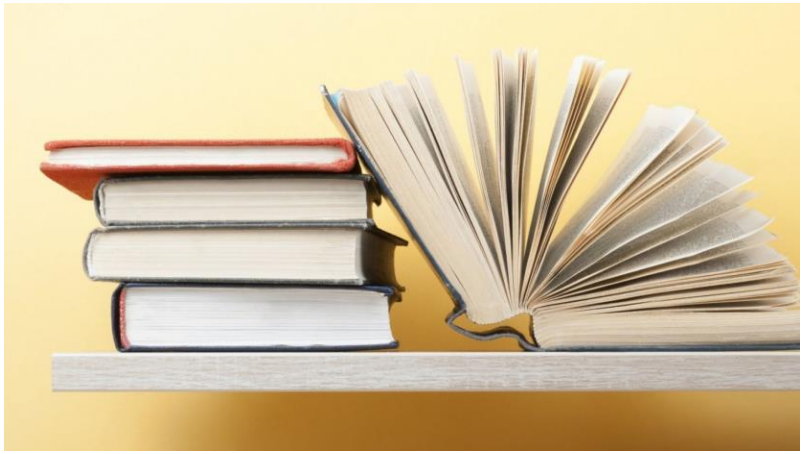
- What is Cross-lingual Language Understanding?
  - Models trained in one language can solve tasks in other languages

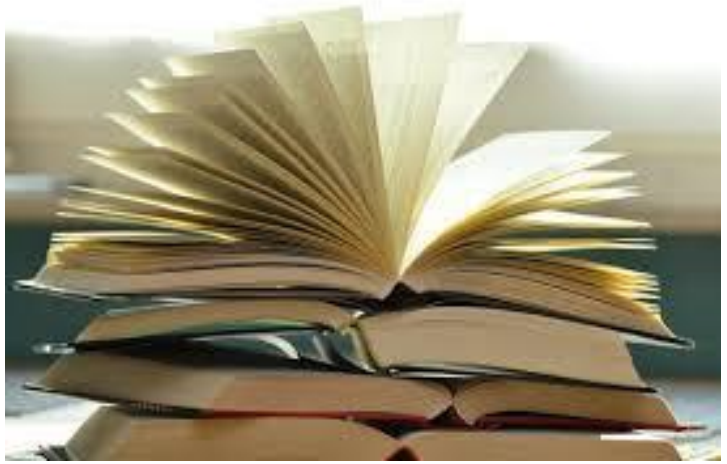


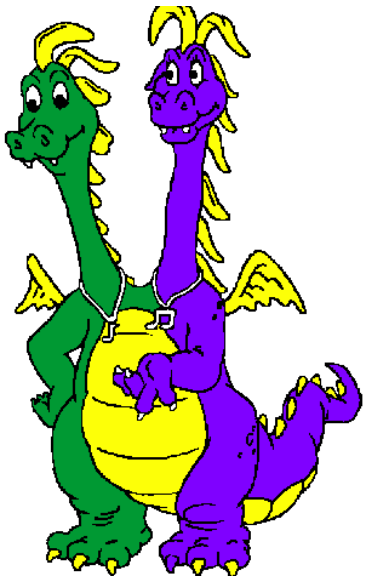
English



Preliminary: Translate-train







English (labeled)

Korean

Preliminary: XLM

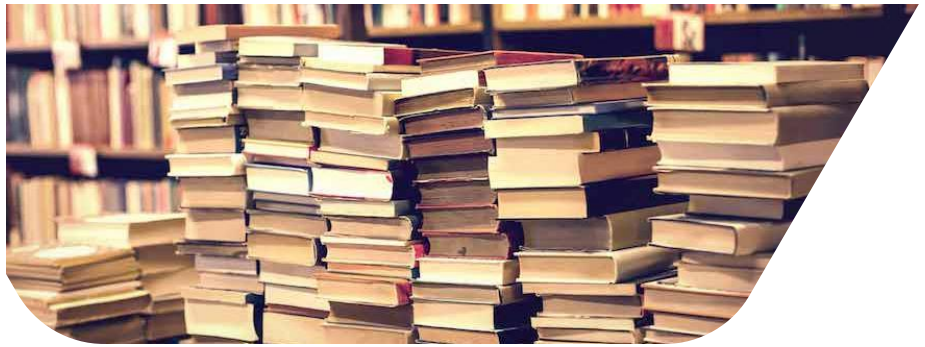
- Read **more** books in **all** languages
- Cross-lingual language model pre-training

FILTER: Fusion in the Intermediate Layers of Transformer

- A general, simple and effective finetuning method based on machine translator and pretrained cross-lingual language model (XLM)

machine translator

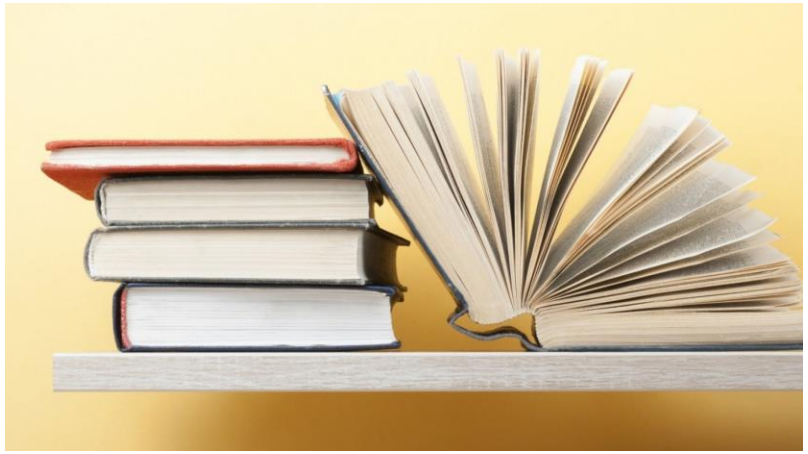
+

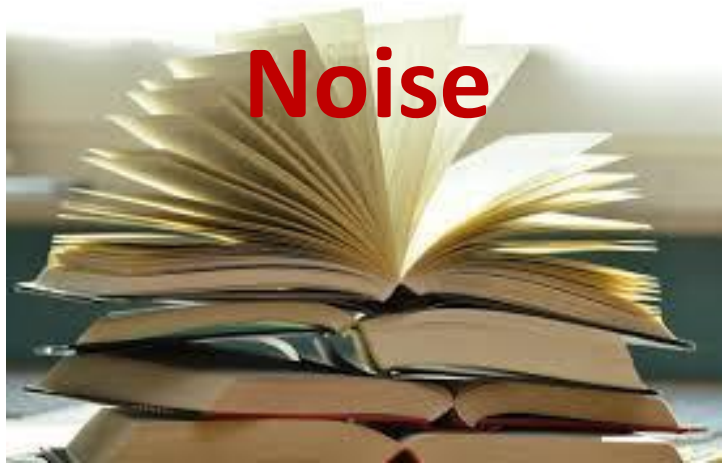


XLM

Why do We Need FILTER?

- Machine translator is still not good enough



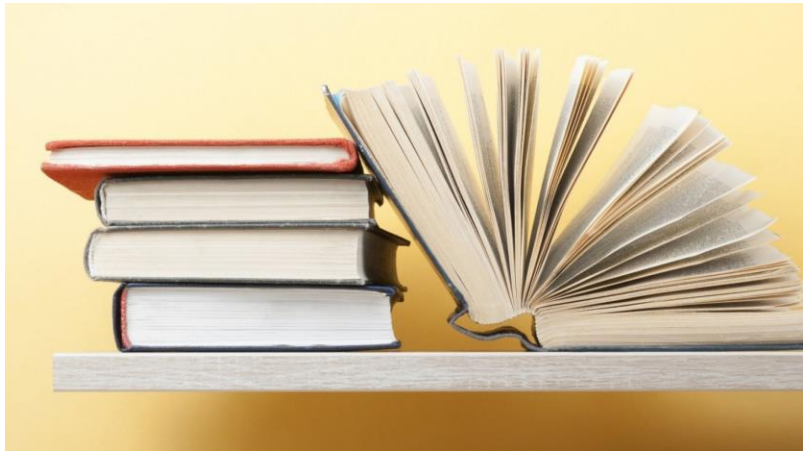


English (labeled)

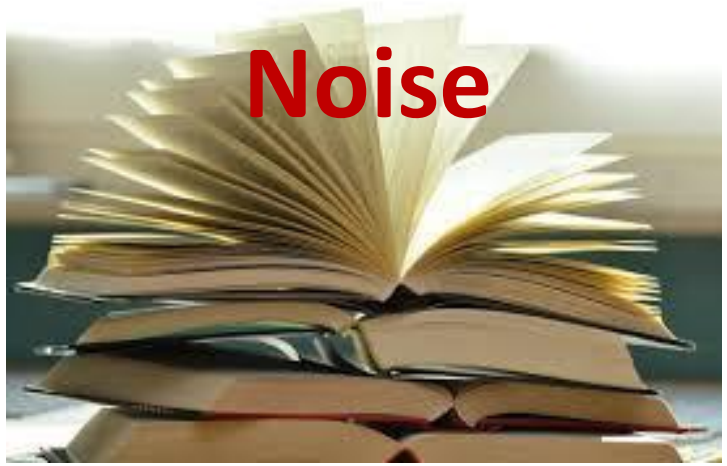
Korean

Why do We Need FILTER?

- Find the balance between pivot language (English) and other languages



refer



English (labeled)

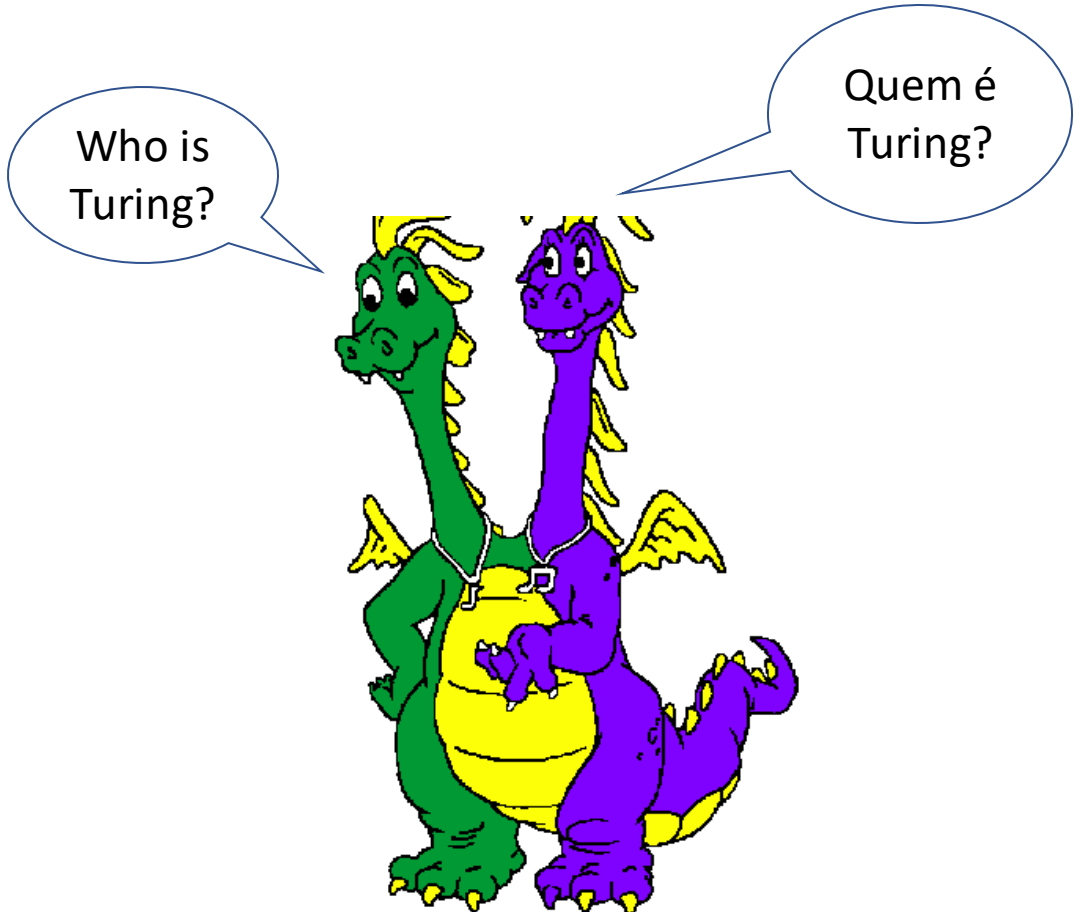
Korean

How does FILTER Work?

"Contrastive learning" on translated text pairs

Q: Who is Turing? A: Alan Turing was a British scientist and a pioneer in computer science.

Q: Quem é Turing? A: Alan Turing foi um cientista britânico e um pioneiro na ciência da computação.



Agenda

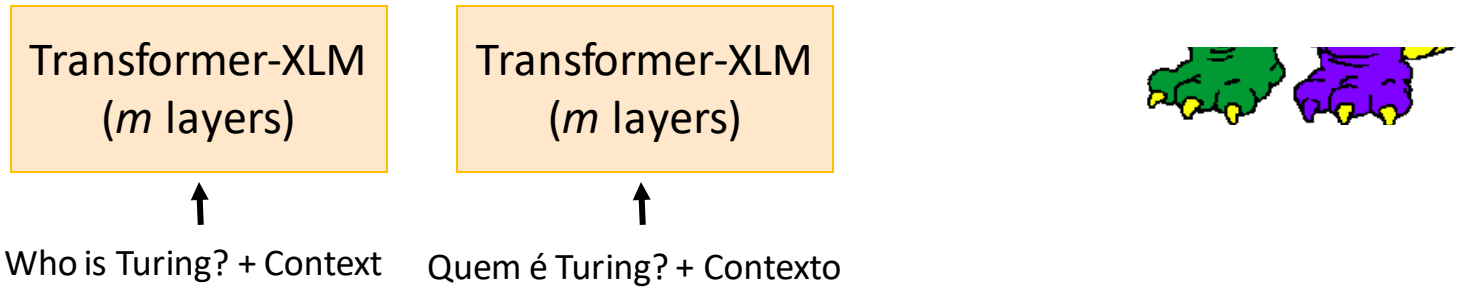
- Background on Cross-Lingual Tasks
- Motivation of FILTER
- Our FILTER Framework
  - Fusion in the Intermediate Layers of Transformer
  - A general framework for cross-lingual language model finetuning
- Experiments and Analysis
  - State-of-the-arts on XTREME and XGLUE benchmarks
  - Model Analysis

FILTER Architecture

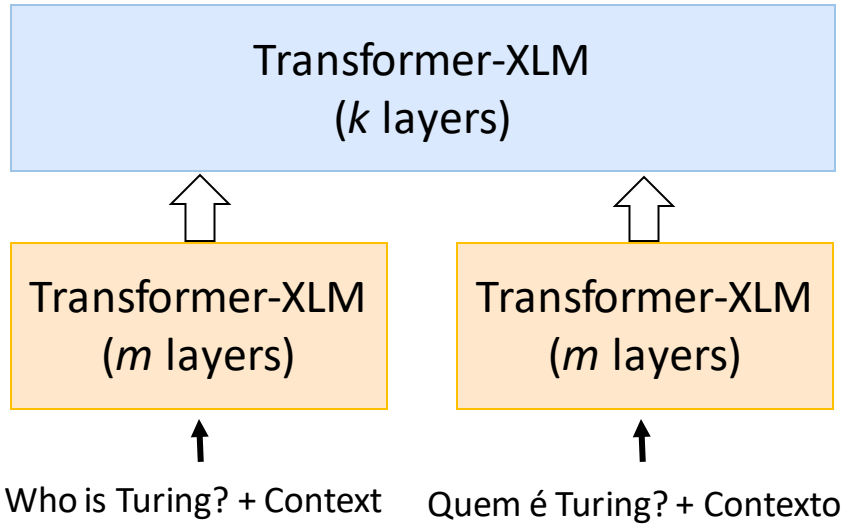
Who is Turing? + Context

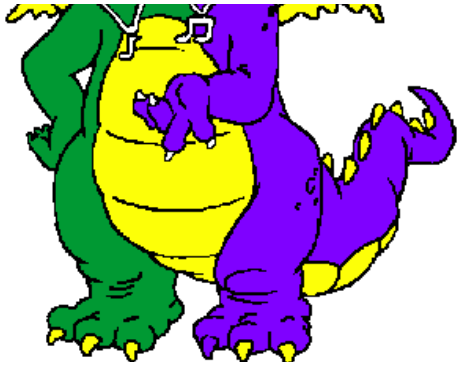
Quem é Turing? + Contexto

FILTER Architecture

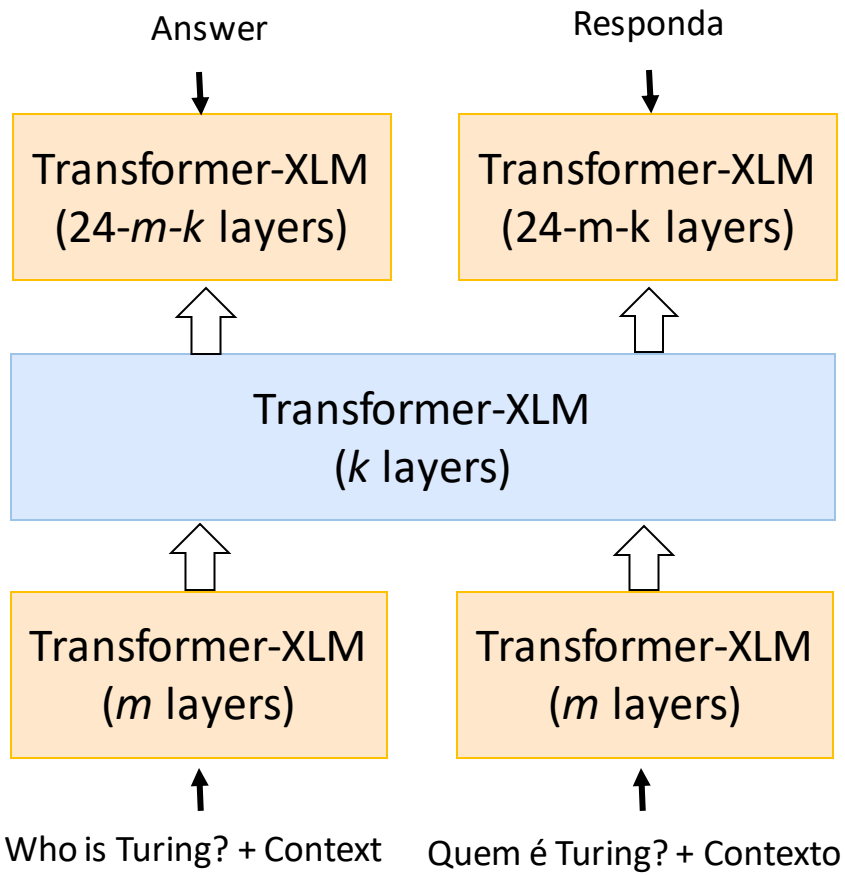


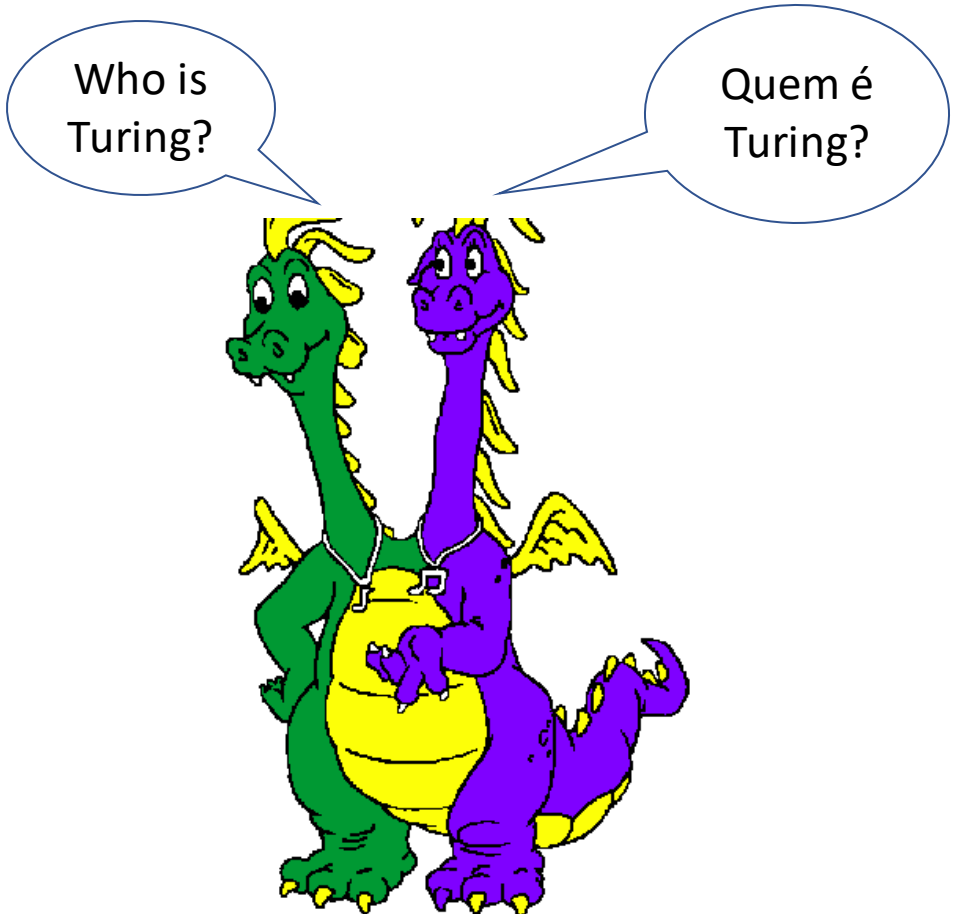
FILTER Architecture

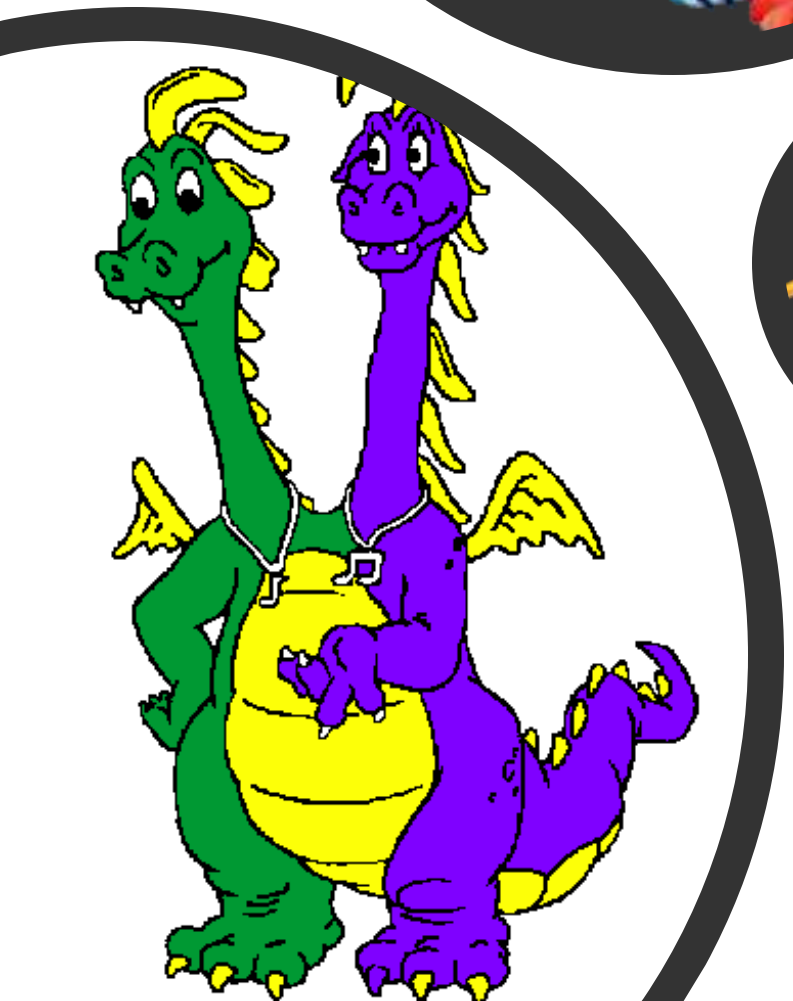




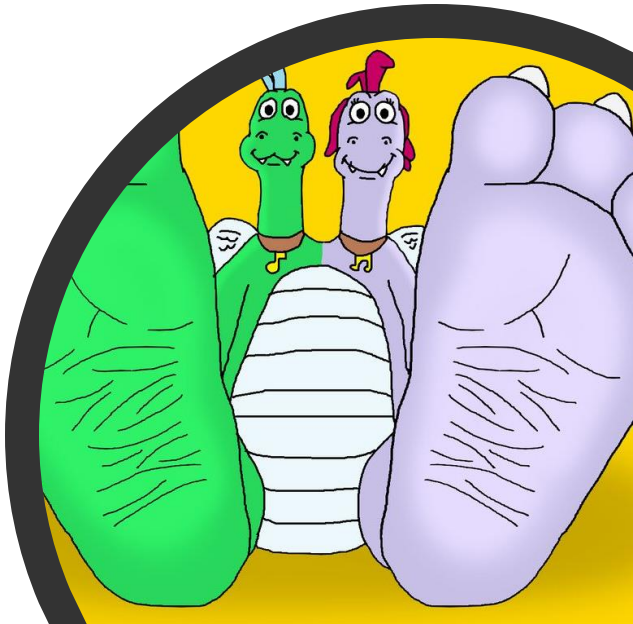
FILTER Architecture



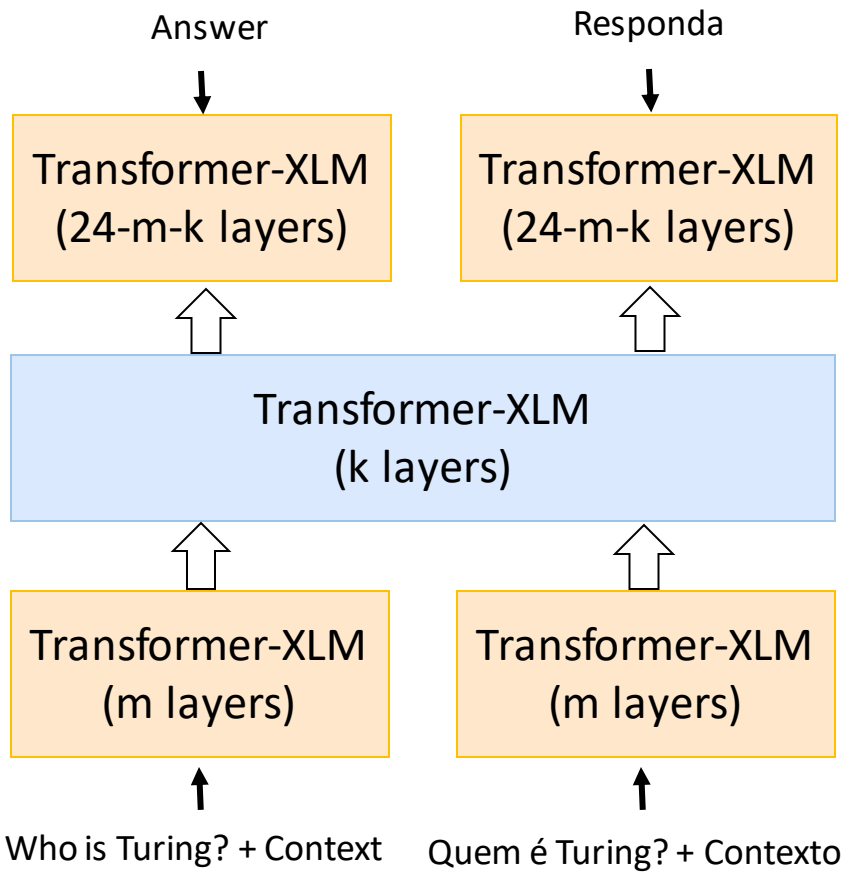


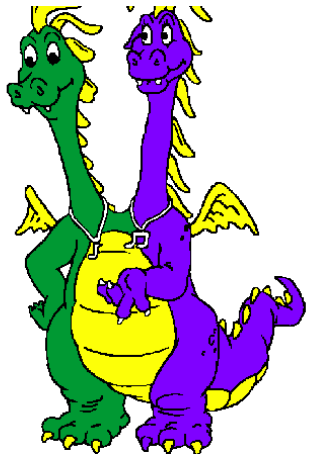


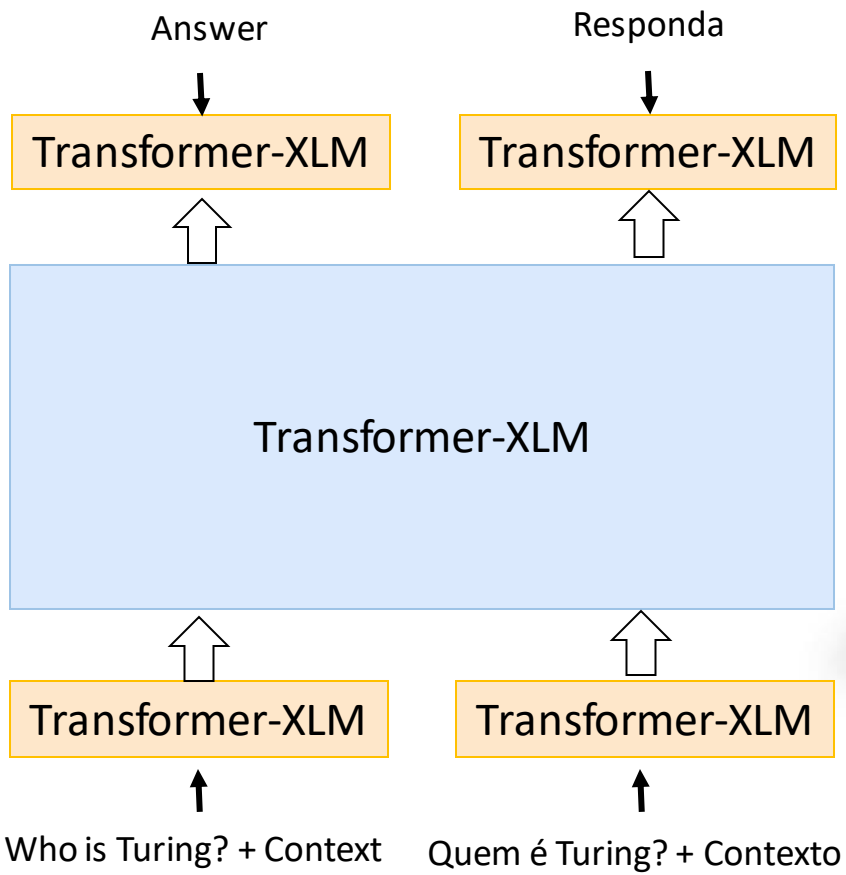
FILTER Zoo



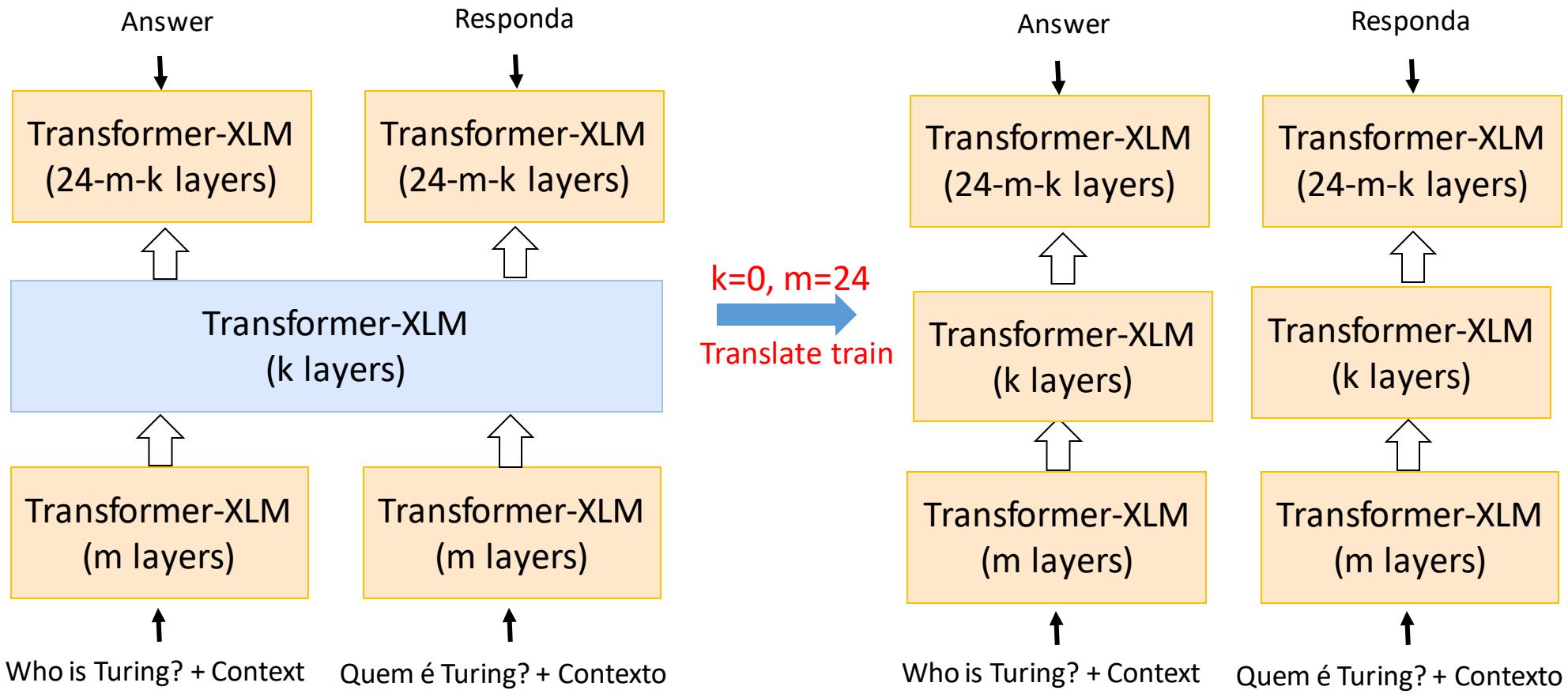
FILTER Architectures



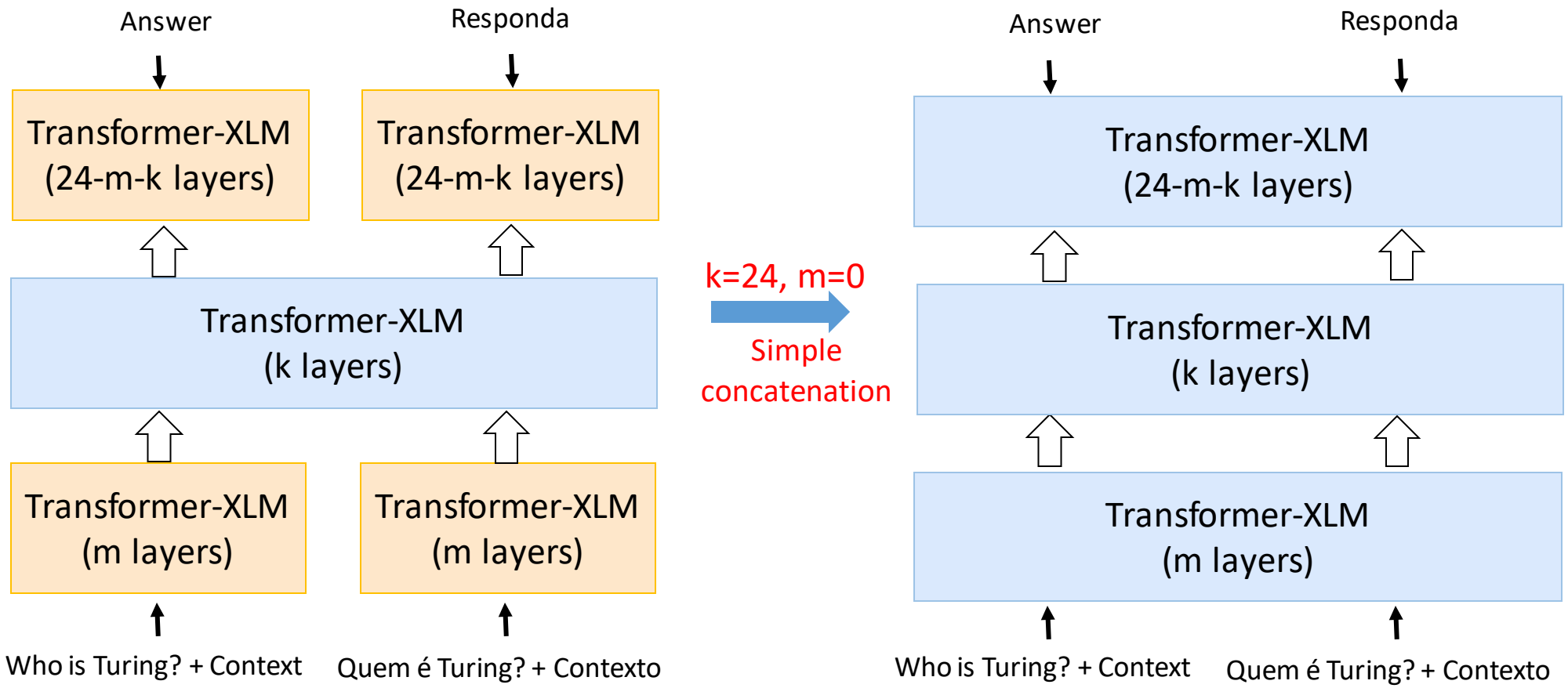




A General Framework for Cross-lingual Fine-tuning

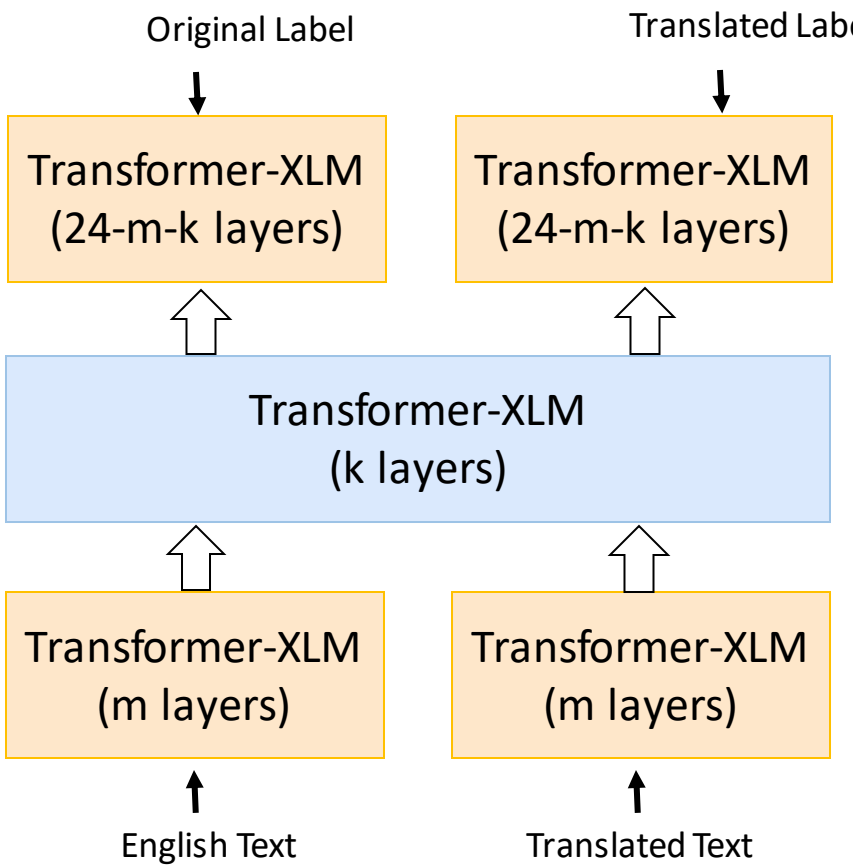


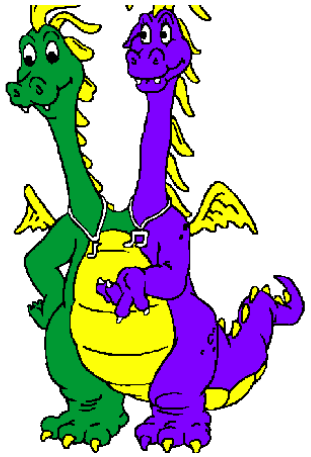
A General Framework for Cross-lingual Fine-tuning



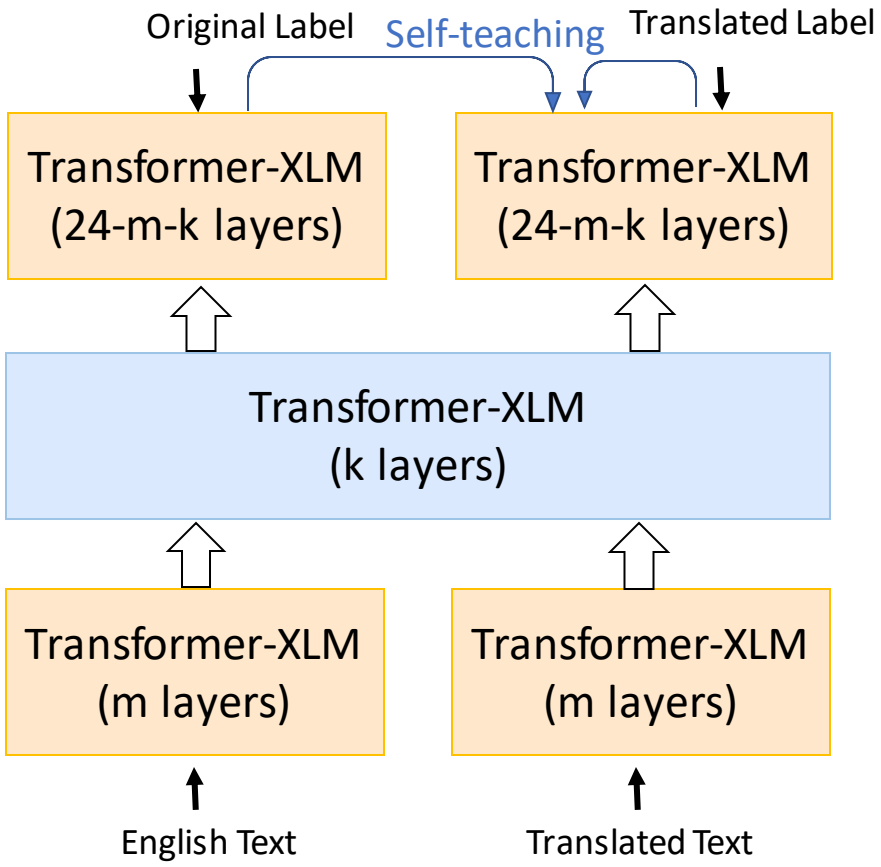
How To Train?

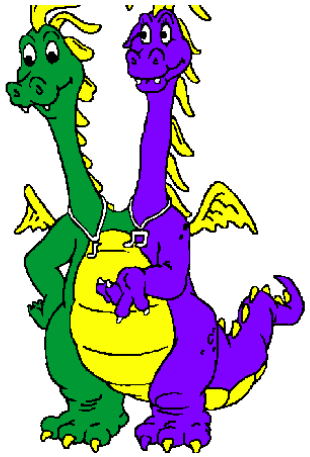
This label may not be accurate for QA task and not available for structure prediction task





How To Train?





Algorithm 1 FILTER Training Procedure.

```
# Teacher model training
# S, l^s: text and label in the source language
# T, l^t: text and label in the target language
for all S, l^s do
    T = Translation (S);
    l_t = Transfer from l_s if available;
    Train FILTER_tea with (S, l^s) and (T, l^t);
end for

# Self-teaching, i.e., student model training
for all S, l^s, T, l^t do
    p^s_tea, p^t_tea = FILTER_tea (S, T)
    Train FILTER_stu with (S, l^s), (T, l^t) and (T, p^t_tea)
end for
```

Experiments and Results

XTREME Leaderboard Results

- Established new SOTA on all tasks in XTREME benchmark
- Improvement: +2.8 on Classification, +1.5 on Structure Prediction, +1.3 on Question Answering, +3.9 on Sentence Retrieval

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
2	VECO	DAMO NLP Team	Alibaba	Aug 31, 2020	74.8	84.7	70.4	67.2	80.5
3	Anonymous1	Anonymous1	Anonymous1	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
4	XLM-R (large)	XTREME Team	Alphabet, CMU	-	68.2	82.8	69.0	62.3	61.6
5	mBERT	XTREME Team	Alphabet, CMU	-	59.6	73.7	66.3	53.8	47.7
6	MMTE	XTREME Team	Alphabet, CMU	-	59.3	74.3	65.3	52.3	48.9
7	XLM	XTREME Team	Alphabet, CMU	-	55.8	75.0	65.6	43.9	44.7

- XTREME : https://arxiv.org/pdf/2003.11080.pdf

XGLUE Leaderboard Results

- Established new SOTA on all tasks in XGLUE benchmark

Rank	Model	Submission Date	NER	POS	NC	MLQA	XNLI	PAWS-X	QADSM	WPR	QAM
1	FILTER (Microsoft Dynamics 365 AI Research)	2020-09-14	82.6	81.6	83.5	76.2	83.9	93.8	71.4	74.7	73.4
			+2.9	+2.0		+10.2	+8.6	+3.7	+3.0	+0.8	+4.5
2	**Unicoder Baseline** (XGLUE Team)	2020-05-25	79.7	79.6	83.5	66.0	75.3	90.1	68.4	73.9	68.9

- XGLUE: https://arxiv.org/pdf/2004.01401.pdf

Results on Different Tasks

- Both FILTER and self-teaching loss further boost the performance

Model	Pair sentence		Structured prediction		Question answering		
	XNLI	PAWS-X	POS	NER	XQuAD	MLQA	TyDiQA-GoldP
Metrics	Acc.	Acc.	F1	F1	F1 / EM	F1 / EM	F1 / EM
Cross-lingual zero-shot transfer (models are trained on English data)							
mBERT	65.4	81.9	70.3	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9
XLM	69.1	80.9	70.1	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1
XLM-R	79.2	86.4	72.6	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0
InfoXLM	81.4	-	-	-	- / -	73.6 / 55.2	- / -
Phang et al. (2020)	80.4	87.7	74.4	63.4	77.2 / 61.3	72.3 / 53.5	- / -†
Translate-train (models are trained on English training data and its translated data on the target language)							
mBERT	74.0	86.3	-	-	70.0 / 56.0	65.6 / 48.0	55.1 / 42.1
mBERT, multi-task	75.1	88.9	-	-	72.4 / 58.3	67.6 / 49.8	64.2 / 49.3
XLM-R, multi-task (Ours)	82.6	90.4	-	-	80.2 / 65.9	72.8 / 54.3	66.5 / 47.7
FILTER (Ours)	83.6	91.2	75.5	66.7	82.3 / 67.8	75.8 / 57.2	68.1 / 49.7
FILTER + Self-Teaching (Ours)	**83.9**	**91.4**	**76.2**	**67.7**	**82.4 / 68.0**	**76.2 / 57.7**	**68.3 / 50.9**

A Closer Look at XNLI for Each Language

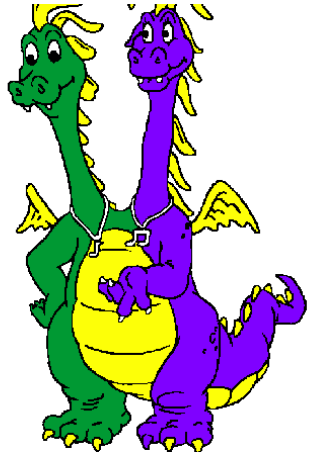
- FILTER outperforms all baselines on each language

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
mBERT	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
MMTE	79.6	64.9	70.4	68.2	67.3	71.6	69.5	63.5	66.2	61.9	66.2	63.6	60.0	69.7	69.2	67.5
XLM	82.8	66.0	71.9	72.7	70.4	75.5	74.3	62.5	69.9	58.1	65.5	66.4	59.8	70.7	70.2	69.1
XLM-R	88.7	77.2	83.0	82.5	80.8	83.7	82.2	75.6	79.1	71.2	77.4	78.0	71.7	79.3	78.2	79.2
XLM-R (*translate-train*)	88.6	82.2	85.2	84.5	84.5	85.7	84.2	80.8	81.8	77.0	80.2	82.1	77.7	82.6	82.7	82.6
FILTER	**89.7**	83.2	86.2	85.5	85.1	**86.6**	85.6	80.9	83.4	78.2	**82.2**	83.1	77.4	83.7	83.7	83.6
FILTER + Self-Teaching	89.5	**83.6**	**86.4**	**85.6**	**85.4**	**86.6**	**85.7**	**81.1**	**83.7**	**78.7**	81.7	**83.2**	**79.1**	**83.9**	**83.8**	**83.9**

Table 2: XNLI accuracy scores for each language. Results of mBERT, MMTE, XLM and XLM-R are from XTREME (Hu et al. 2020). mtl denotes translate-train in multi-task version.

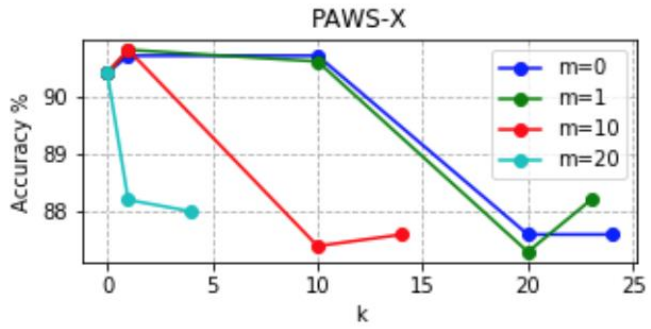
Analysis in FILTER Zoo

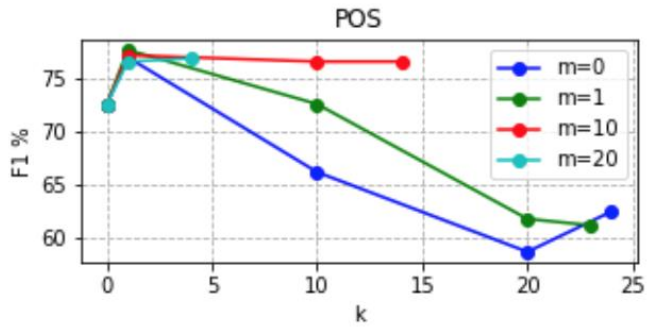
- Which FILTER to use towards different tasks?
  - The number of intermediate fusion layers (k)
  - The number of local transformer layers (m)

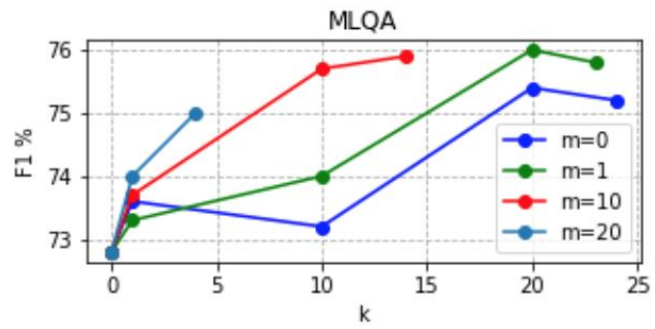



Analysis in FILTER Zoo

- Which FILTER to use towards different tasks?
  - The number of intermediate fusion layers (k)
  - The number of local transformer layers (m)
- Effect of Intermediate Fusing Layers
  - For PAWS-X and POS, the performance **drops** significantly when the number of intermediate fusion layers increases. (e.g., k from 1 to 20, m=1)
  - For MLQA, performance is consistently **improved** with the number of intermediate fusion layers increasing (e.g. k from 1 to 20, m=1)

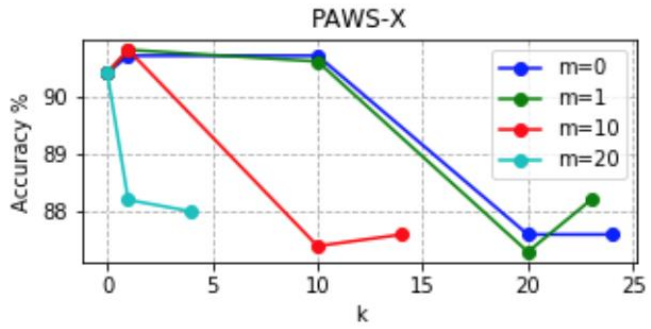


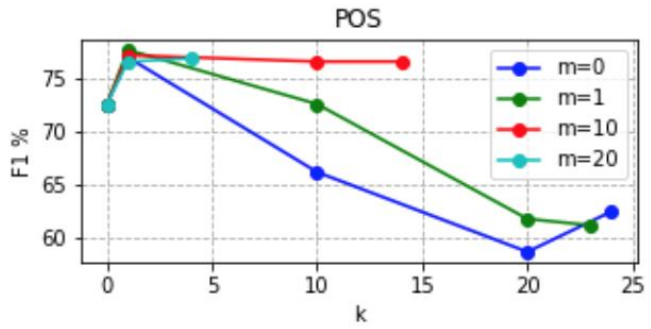


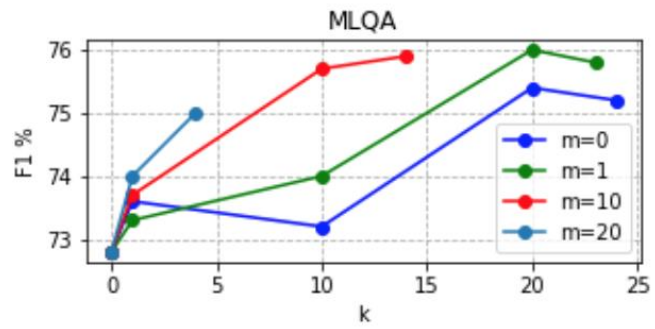


Analysis in FILTER Zoo

- Which FILTER to use towards different tasks?
  - The number of intermediate fusion layers (k)
  - The number of local transformer layers (m)
- Effect of Local Transformer Layers
  - For PAWS-X, FILTER performs better when using **less** local transformer layers (e.*g.*, *m* from 0 to 10, *k*=10)
  - For POS and MLQA, FILTER performs better when using **more** local transformer layers (*e.g.*, *m* from 0 to 10, *k*=10)







Key Observations

- Different tasks need different numbers of “local” transformer layers (*m*) and intermediate fusion layers (*k*)
- Use more local layers for complex tasks such as QA and structured prediction;
- Use fewer local layers for classification tasks

Analysis

- Cross-Lingual Transfer Gap
  - Calculating the difference between the performance on English test set and the average performance of other target languages
  - FILTER reduces the cross-lingual gap significantly compared to the baseline
  - Still a large gap for structure prediction tasks demands stronger cross-lingual transfer

Model	XNLI	PAWS-X	XQuAD	MLQA	TyDiQA-GoldP	Avg	POS	NER
mBERT	16.5	14.1	25.0	27.5	22.2	21.1	25.5	23.6
XLM-R	10.2	12.4	16.3	19.1	13.3	14.3	24.3	19.8
Translate-train	7.3	9.0	17.6	22.2	24.2	16.1	-	-
FILTER	**6.0**	**5.2**	**7.3**	**15.7**	**9.2**	**8.7**	**19.7**	**16.3**

Table 3: Analysis on cross-lingual transfer gap of different models on XTREME benchmark. Note that a lower gap indicates a better cross-lingual transfer model. For QA datasets, we compare EM scores. The average score(Avg) is calculated on all classification and QA tasks. Results on mBERT, XLM-R and Translate-train are from Hu et al. (2020).

Conclusions

- FILTER is a general framework for fine-tuning cross-lingual tasks
- Self-teaching loss is helpful on all tasks, especially on tasks lacking labels in the target languages.
- FILTER can achieve less than 6.0 for cross lingual transfer gap on classification tasks indicates zero-shot can also achieve comparable performance on these tasks.
- FILTER achieves SOTA performance on XTREME and XGLUE benchmark

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
2	VECO	DAMO NLP Team	Alibaba	Sep 29, 2020	77.2	87.0	70.4	68.0	88.1
3	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
4	X-STILTs	Phang et al.	New York University	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
5	XLM-R (large)	XTREME Team	Alphabet, CMU	-	68.2	82.8	69.0	62.3	61.6
6	mBERT	XTREME Team	Alphabet, CMU	-	59.6	73.7	66.3	53.8	47.7
7	MMTE	XTREME Team	Alphabet, CMU	-	59.3	74.3	65.3	52.3	48.9
8	RemBERT	Anonymous2	Anonymous2	-	56.1	84.1	73.3	68.6	-
9	XLM	XTREME Team	Alphabet, CMU	-	55.8	75.0	65.6	43.9	44.7

Congrats @Turing team!

Thanks @Microsoft Azure Translator team!

- Microsoft Translator: https://azure.microsoft.com/en-us/services/cognitive-services/translator/

Thank You

Multimodal AI Group: http://aka.ms/mmai