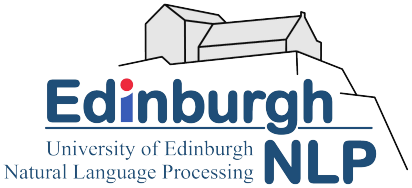
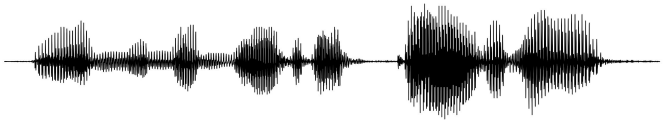
Pre-training on high-resource speech recognition improves low-resource speech-to-text translation

Sameer Bansal
Herman Kamper
Karen Livescu
Adam Lopez
Sharon Goldwater



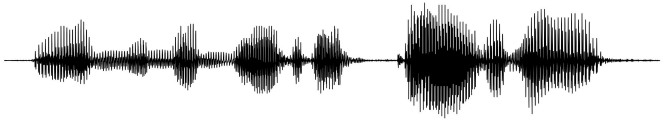
Current systems

Spanish Audio: 

English text: *?*

Current systems

Spanish Audio:



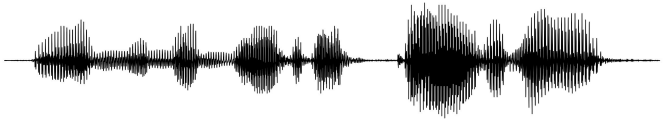
Spanish text: *ola mi nombre es hodor*

English text: *?*

Automatic Speech Recognition

Current systems

Spanish Audio:



Spanish text: *ola mi nombre es hodor*

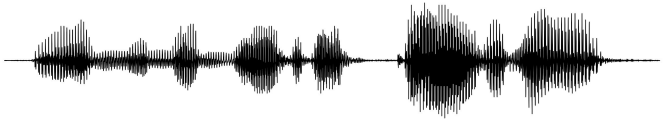
Automatic Speech Recognition

English text: *hi my name is hodor*

Machine Translation

~100 languages supported by Google Translate ...

Unwritten languages

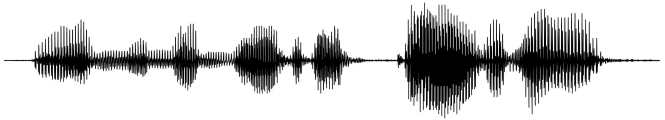
Mboshi: 

Bantu language, Republic of Congo, ~160K speakers

~3000 languages with no writing system

~~Mboshi text:~~ ~~*not available*~~ ~~**Automatic Speech Recognition**~~

Unwritten languages

Mboshi: 

paired with French translations (Godard et al. 2018)

~3000 languages with no writing system

Efforts to collect speech and translations using mobile apps

- *Aikuma*: Bird et al. 2014, *LIG-Aikuma*: Blachon et al. 2016

Haiti Earthquake, 2010

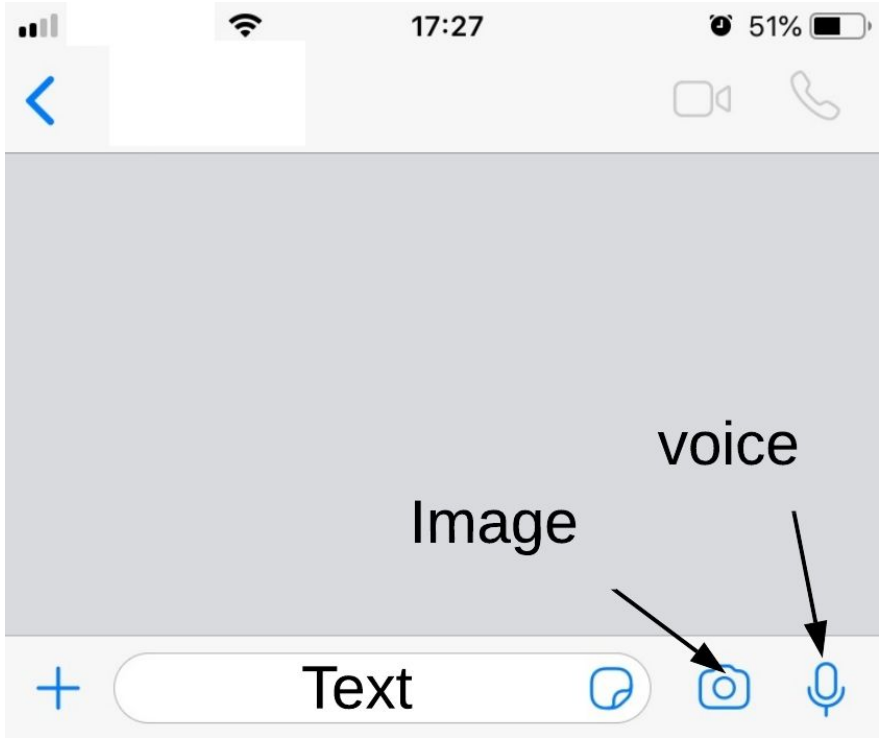
Survivors sent text messages to helpline

Moun kwense nan Sakre Kè nan Pòtoprens

People trapped in Sacred Heart Church, PauP

- International rescue teams face language barrier
- No automated tools available
- Volunteers from global Haitian diaspora help create parallel text corpora in short time
  [Munro 2010]

Are we better prepared in 2019?



Moun kwense nan Sakre Kè nan Pòtoprens

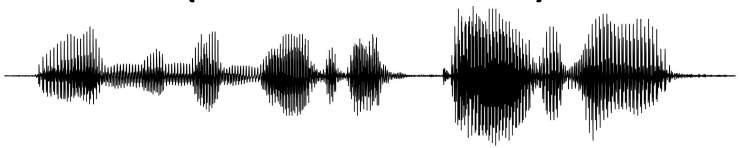
People trapped in Sacred Heart Church, PauP

Voice messages

Can we build a speech-to-text translation (**ST**) system?

… given as training data:

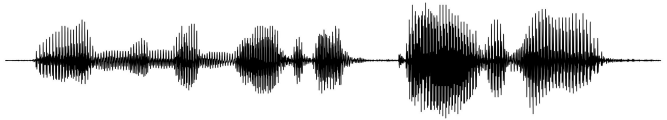
(source audio)



paired with translations

- Tens of hours of speech paired with text translations
- No source text available

Neural models ...

Spanish Audio: 

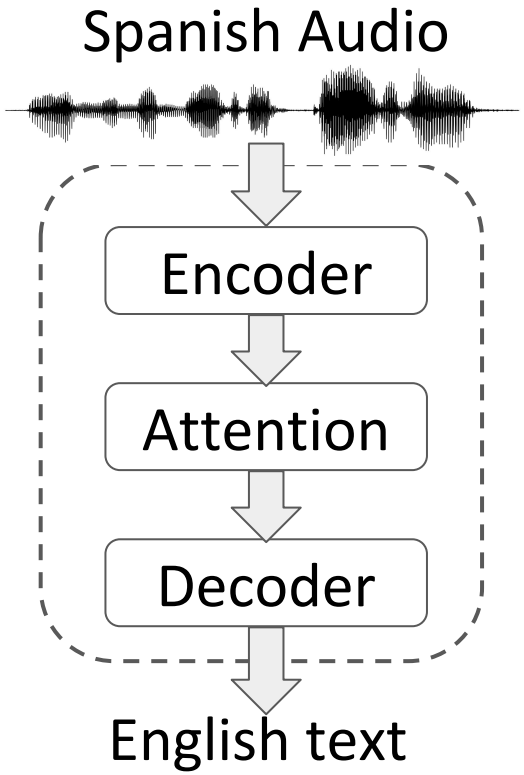
Sequence-to-Sequence

Weiss et al. (2017)

English text: *hi my name is hodor*

Directly translate speech

Spanish speech to English text



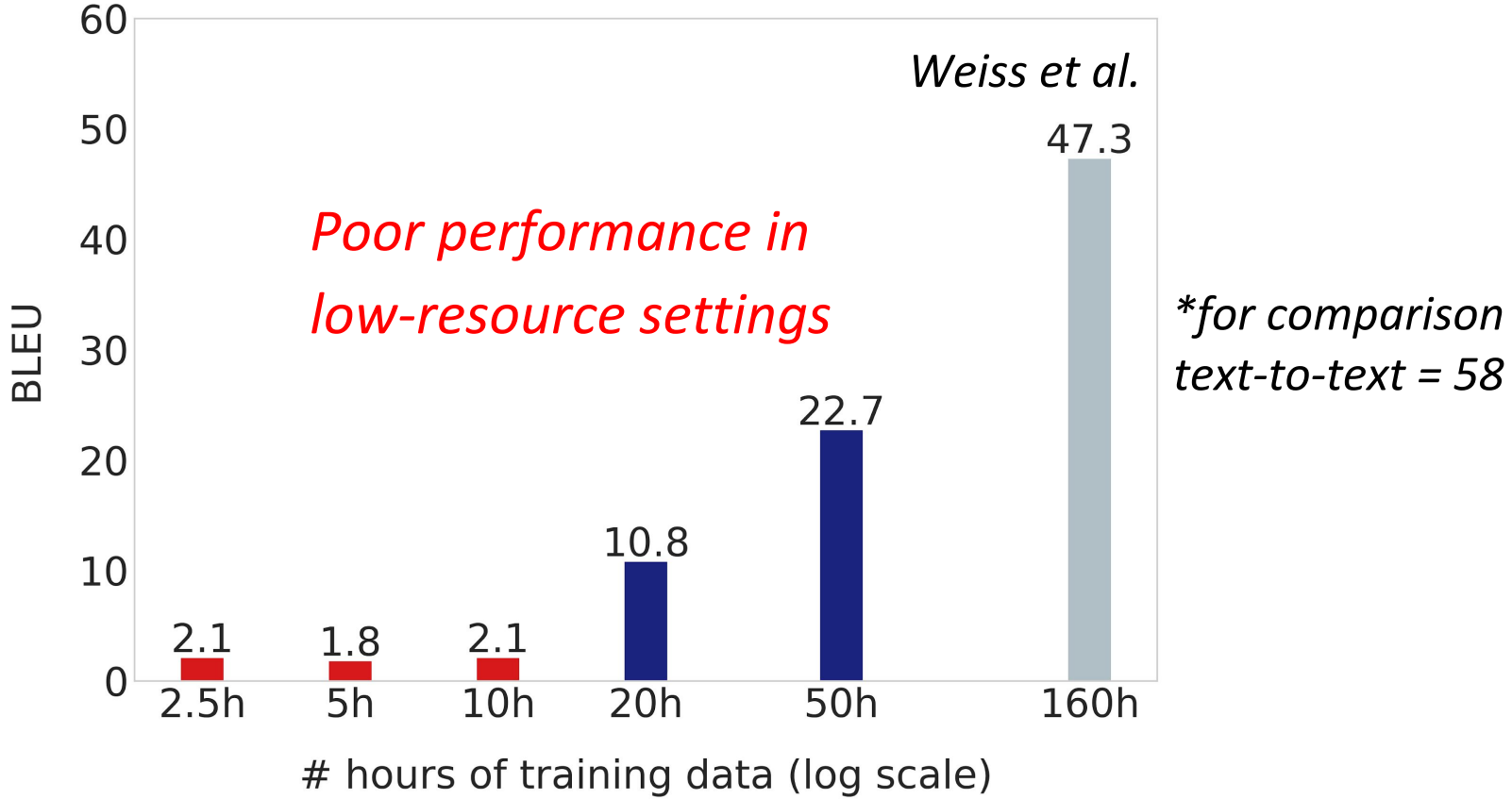
- telephone speech (unscripted)
- realistic noise conditions
- multiple speakers and dialects
- crowdsourced English text translations

Closer to real-world conditions

Spanish speech to English text

*for comparison
text-to-text = 58

But ...


60
50
40
30
20
10
0
BLEU
Weiss et al.
47.3
Poor performance in
low-resource settings
22.7
10.8
2.1
1.8
2.1
2.5h
5h
10h
20h
50h
160h
hours of training data (log scale)
*for comparison
text-to-text = 58

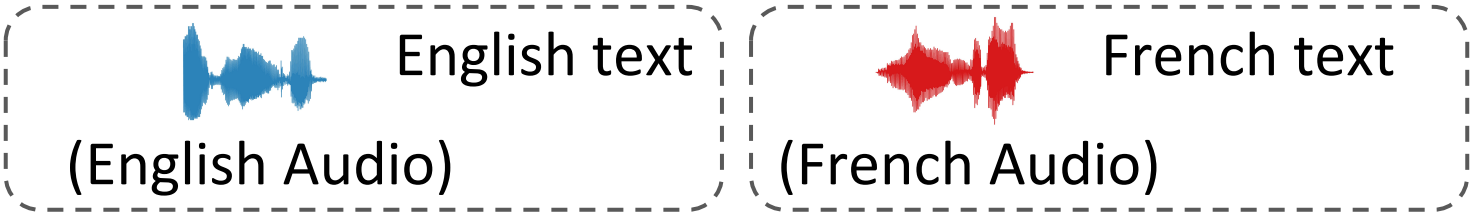
Goal: to improve translation performance

Goal: to improve translation performance

… without labeling more low-resource speech

100s of hours of **monolingual** speech paired with text available

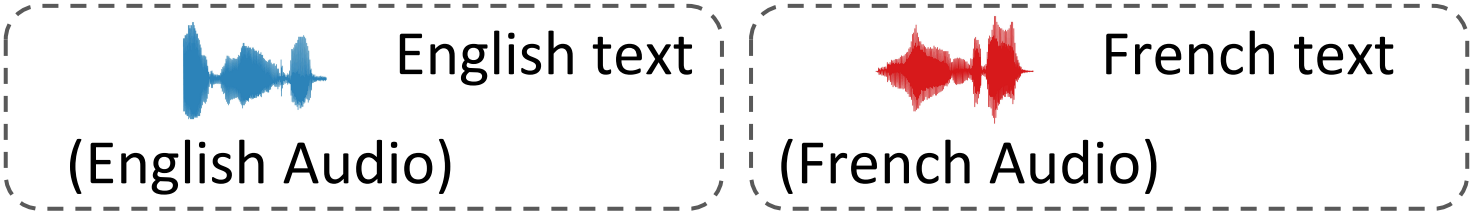
… typically used to train ASR systems



Key idea: leverage monolingual data from a different high-resource language

100s of hours of **monolingual** speech paired with text available

… typically used to train ASR systems



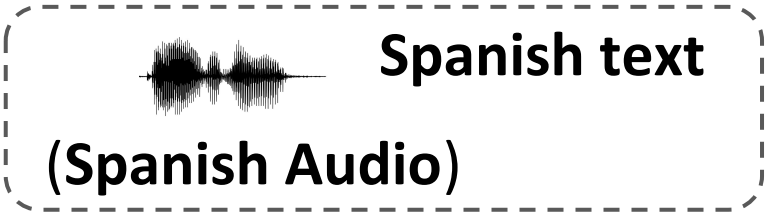
Spanish Audio → Sequence-to-Sequence → **English** text

~20 hours of Spanish-English

?

100s of hours of **monolingual** speech paired with text available

… typically used to train ASR systems

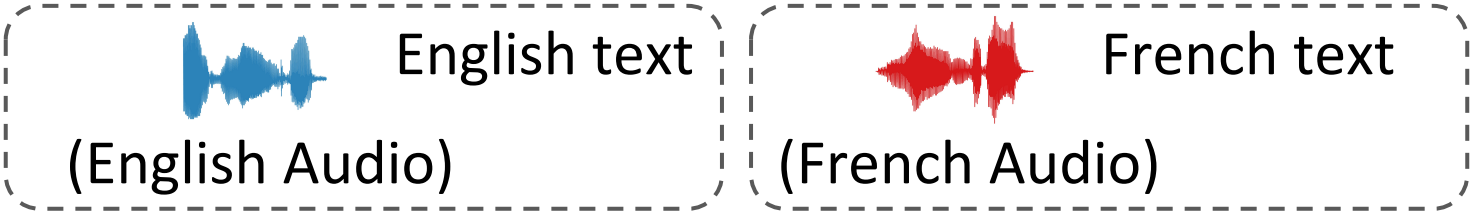


Weiss et al. 2017
Anastasopoulos and Chiang 2018
Bérard et al. 2018
Sperber et al. 2019

~20 hours of Spanish-English

100s of hours of **monolingual** speech paired with text available

… typically used to train ASR systems



Spanish Audio → Sequence-to-Sequence → **English** text

~20 hours of Spanish-English

?

Why Spanish-English?

Why Spanish-English?

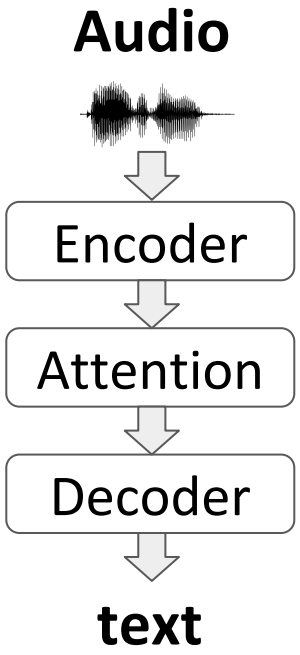
simulate low-resource settings and test our method

Why Spanish-English?

simulate low-resource settings and test our method

Later: results on truly low-resource language --- Mboshi to French

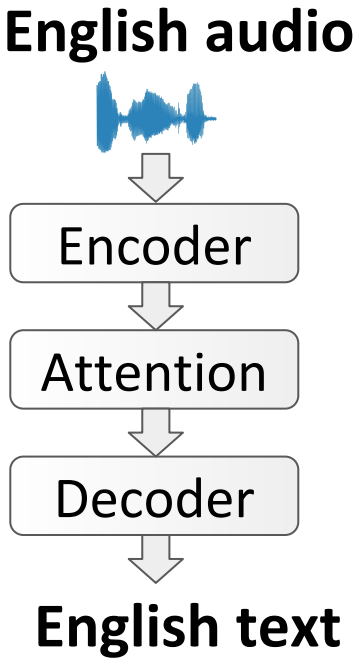
Method



Same model architecture for ASR and ST

*randomly initialized parameters

Pretrain on high-resource

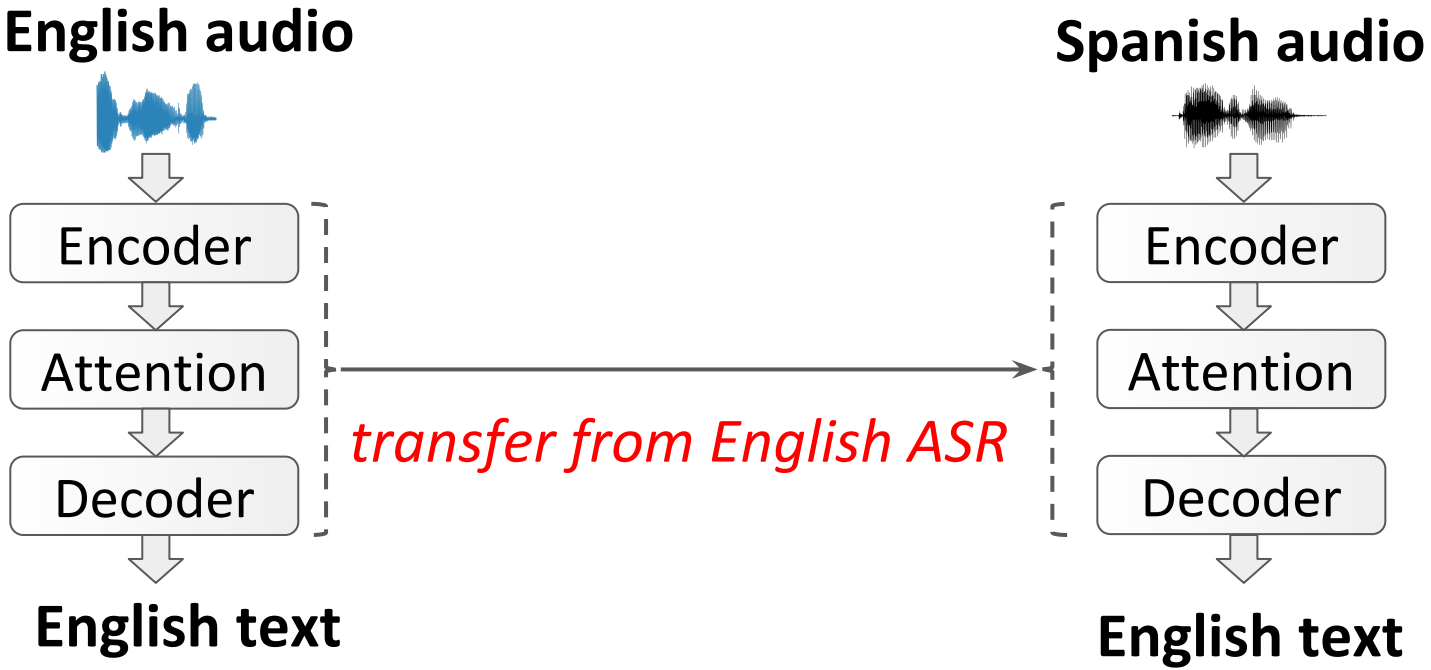


300 hours of English audio and text

*train until convergence

Fine-tune on low-resource

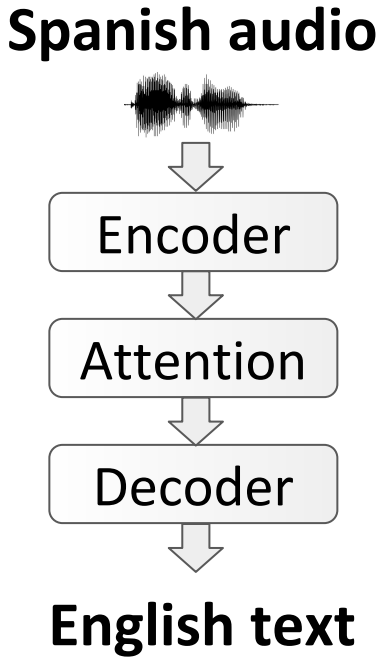
20 hours Spanish-English



Fine-tune on low-resource

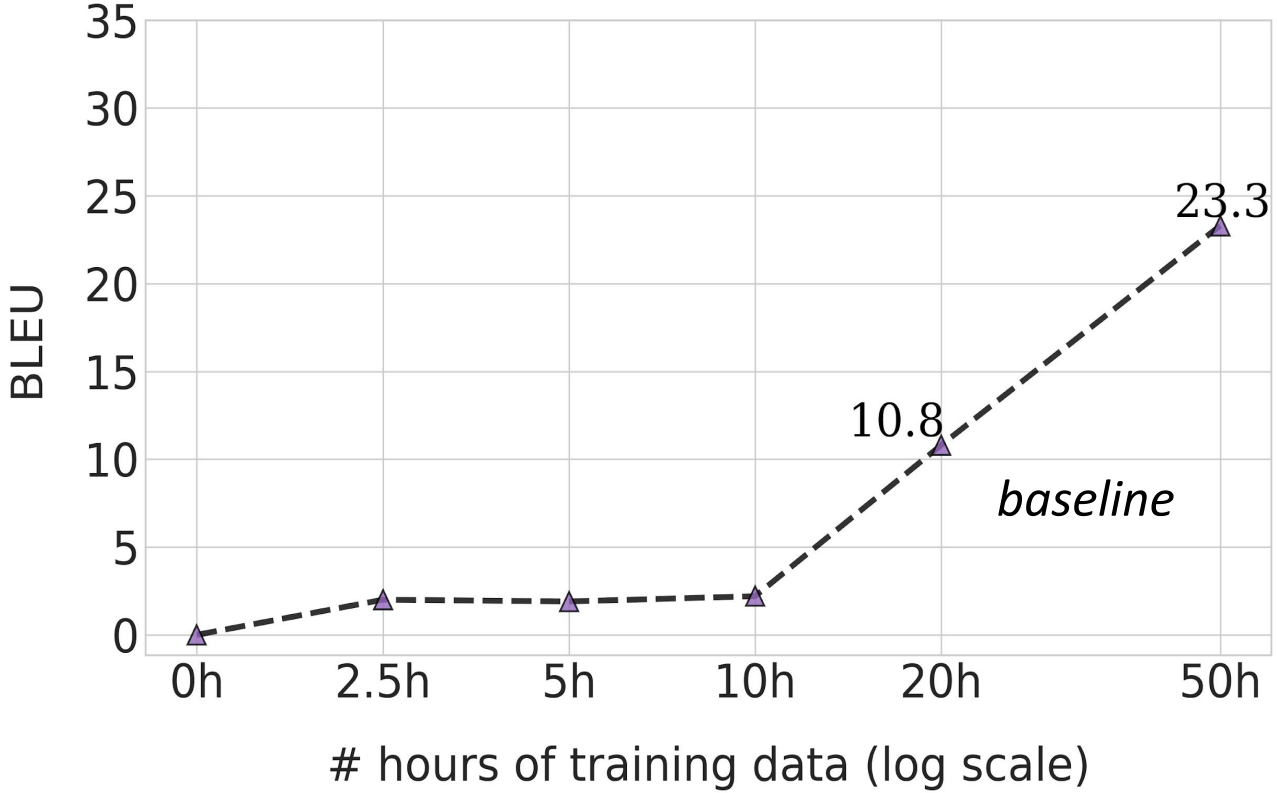
20 hours Spanish-English

*train until convergence



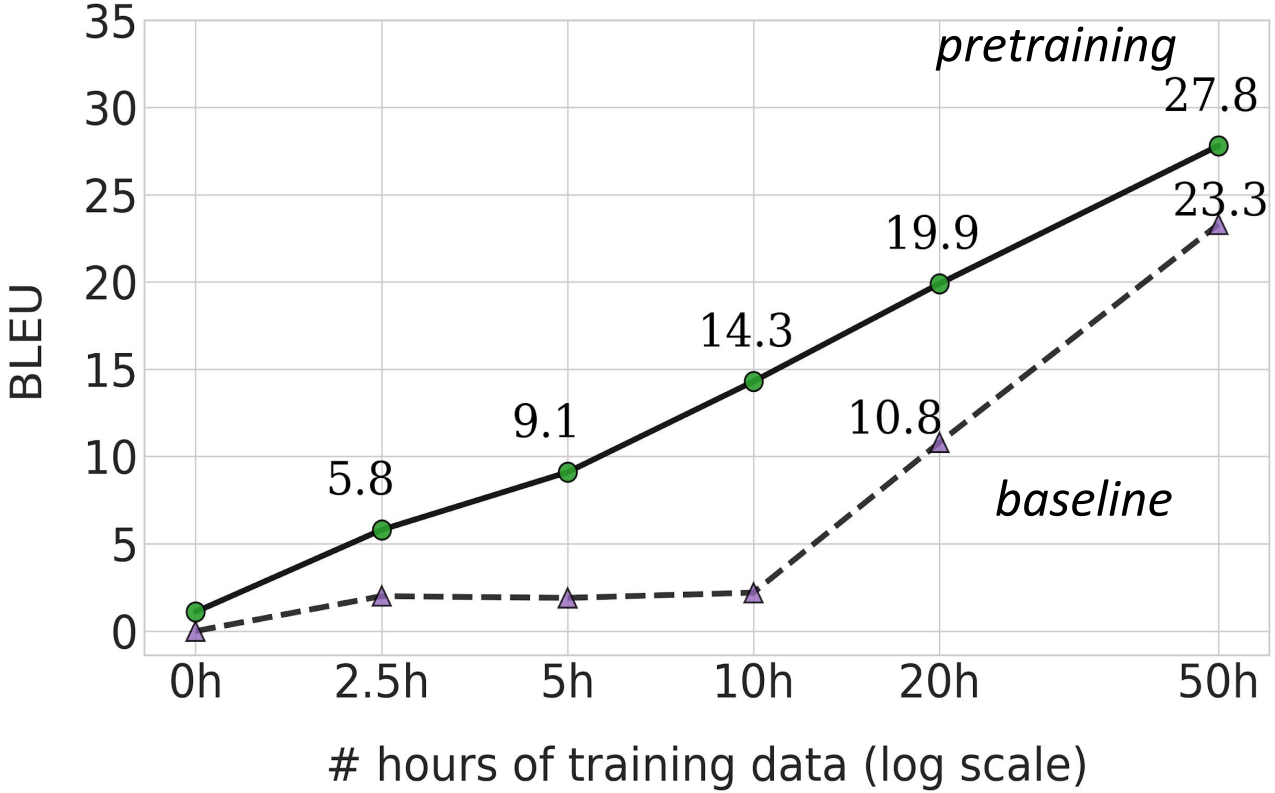
Will this work?

Spanish-English BLEU scores



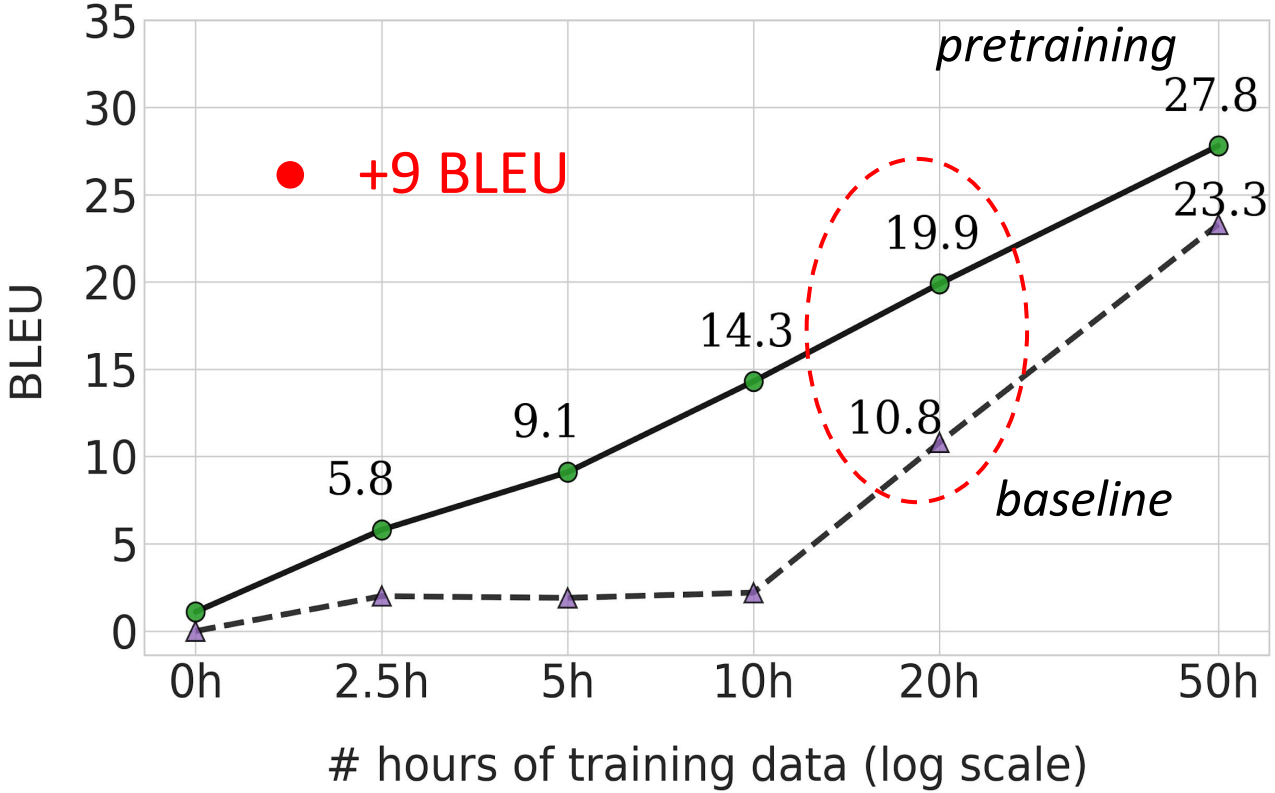
**for comparison*
Weiss et al. = 47.3

Spanish-English BLEU scores



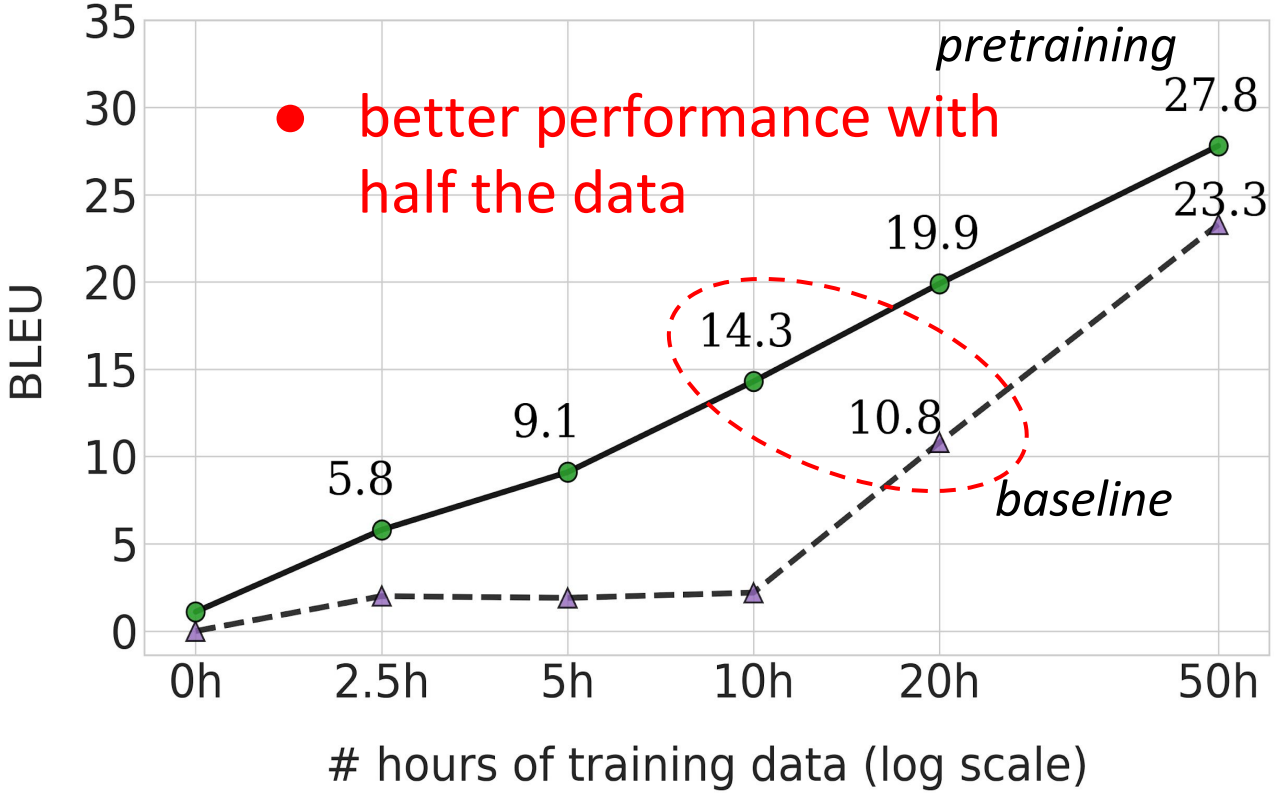
**for comparison*
Weiss et al. = 47.3

Spanish-English BLEU scores



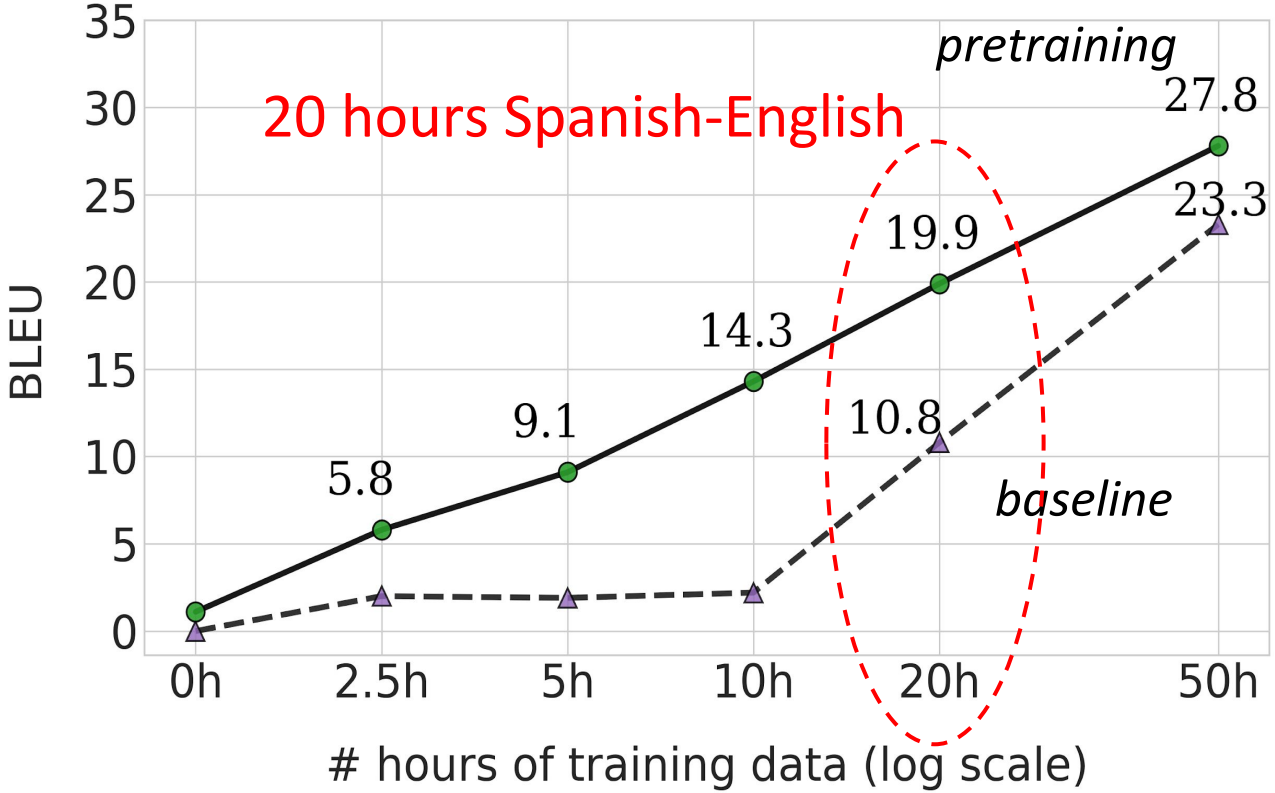
*for comparison
Weiss et al. = 47.3

Spanish-English BLEU scores



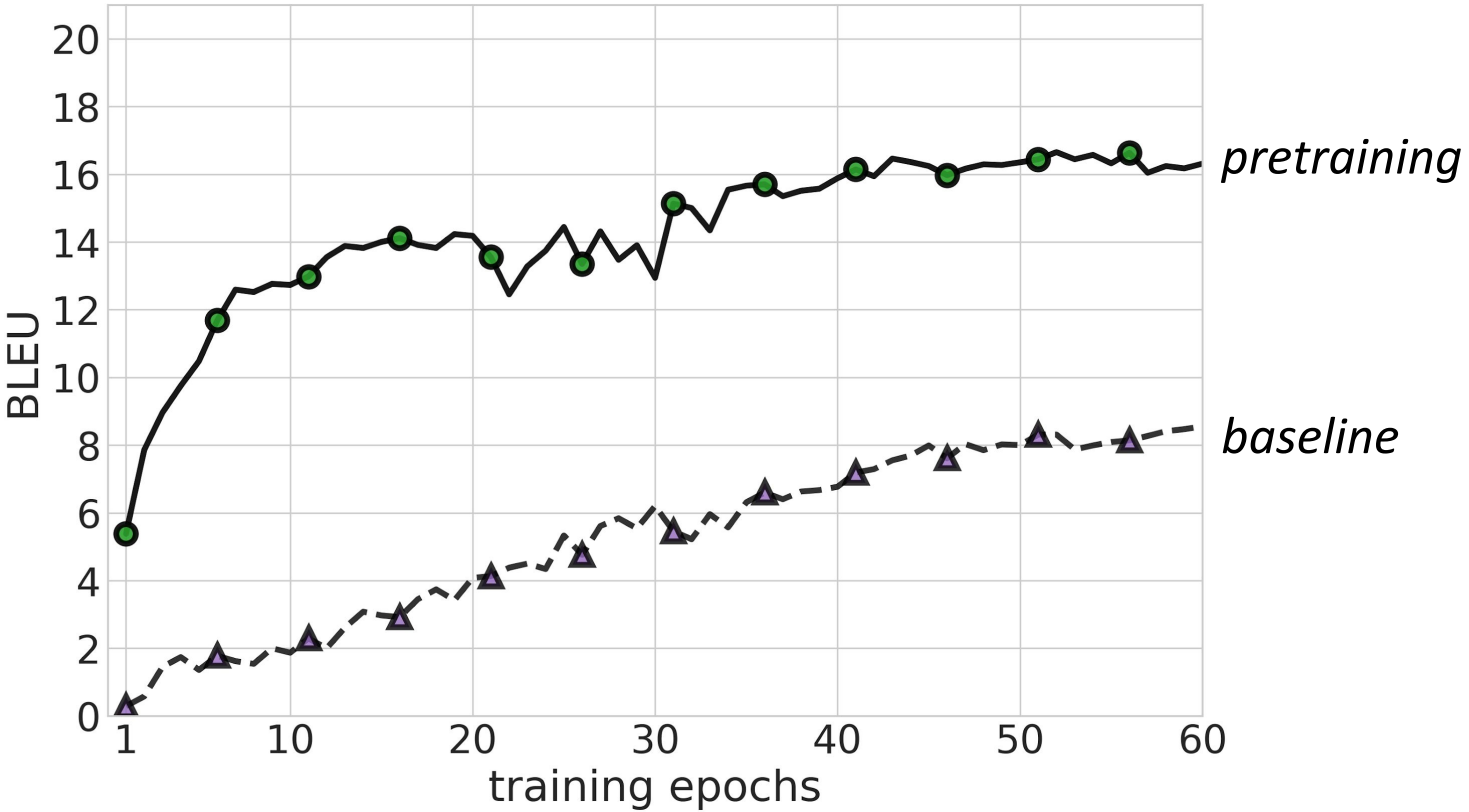
**for comparison*
Weiss et al. = 47.3

Further analysis

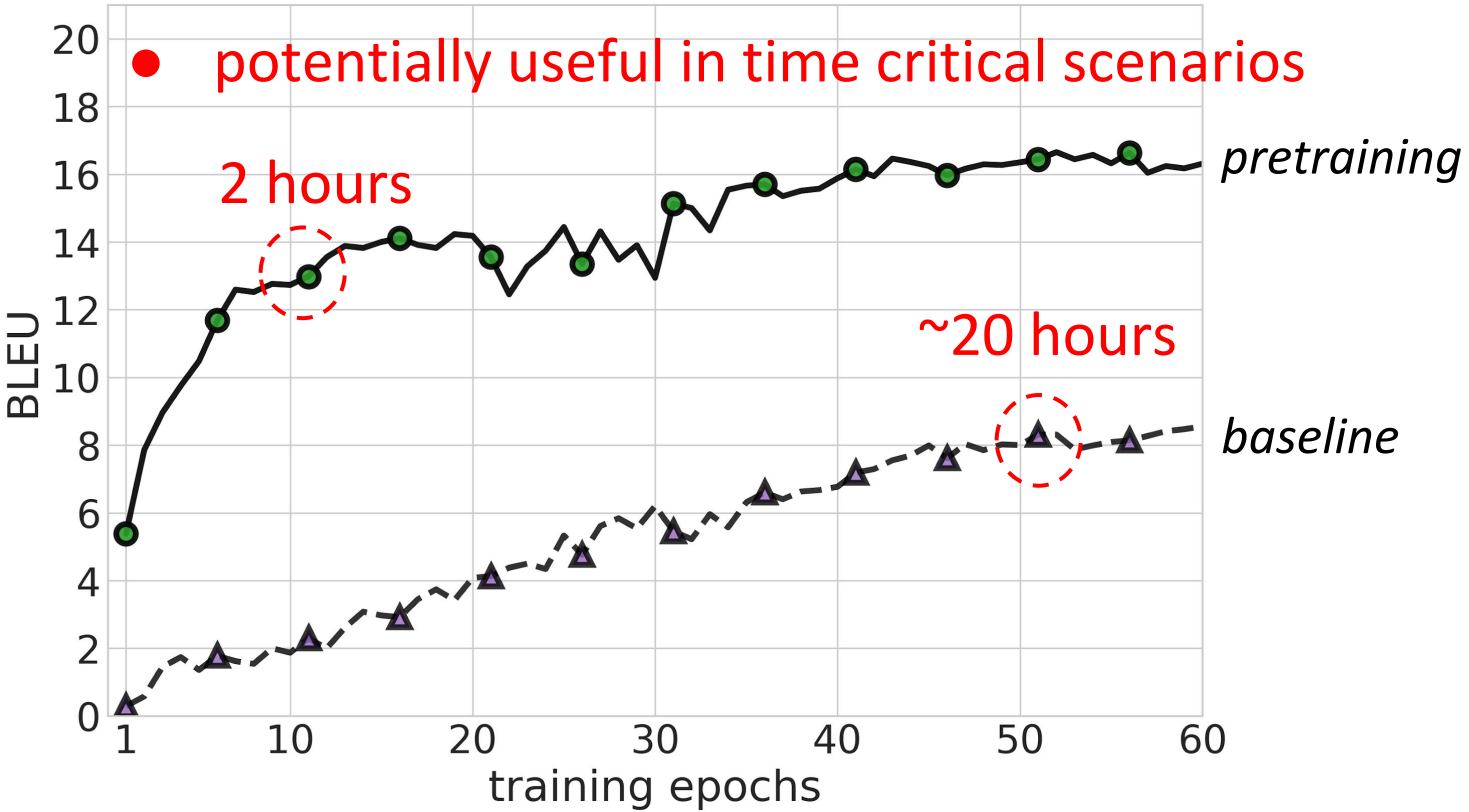


*for comparison
Weiss et al. = 47.3

Faster training time



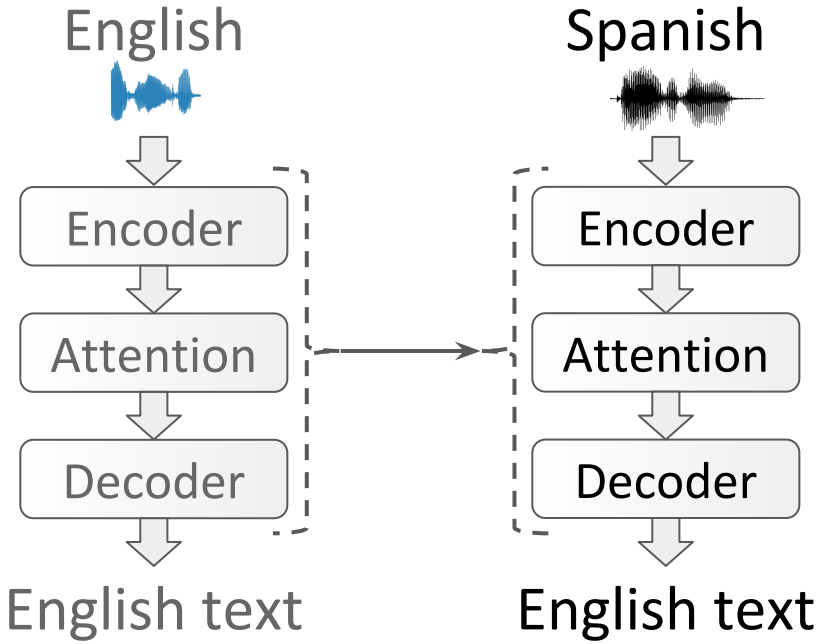
Faster training time



Ablation: model parameters

Spanish to English, *N* = 20 hours

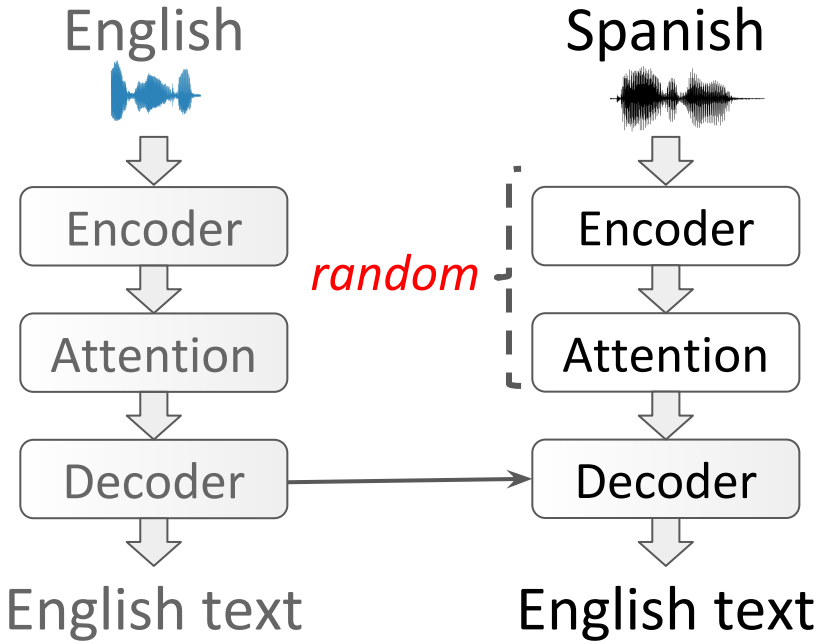
	BLEU
baseline	10.8
+English ASR	19.9



Ablation: model parameters

Spanish to English, *N* = 20 hours

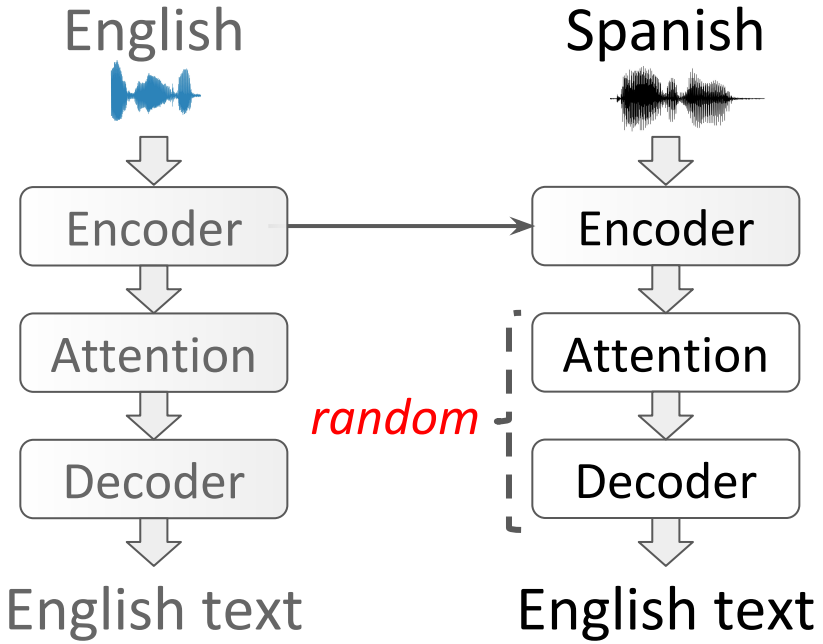
	BLEU
baseline	10.8
+English ASR	19.9
+English ASR: decoder	10.5



Ablation: model parameters

Spanish to English, *N* = 20 hours

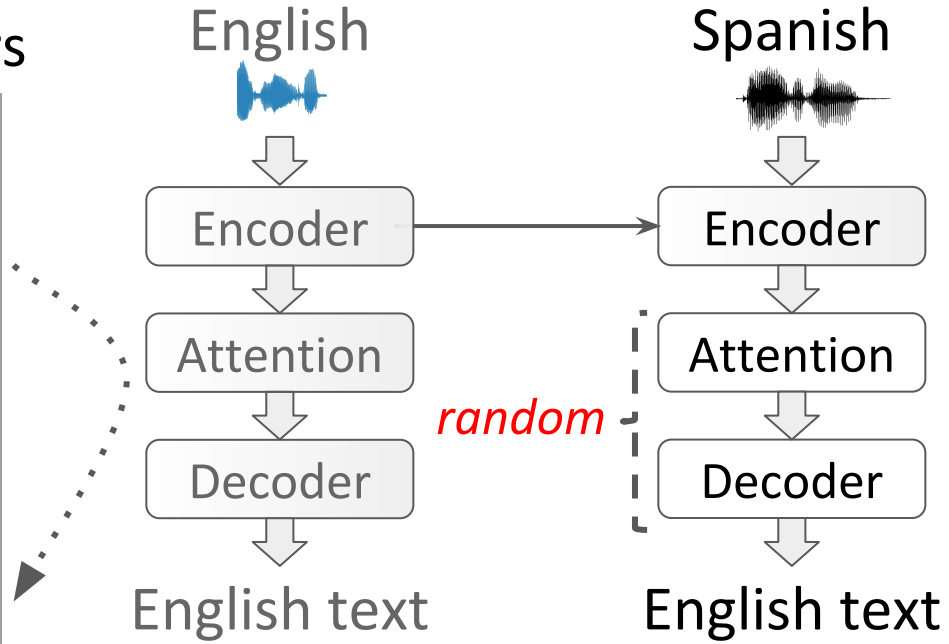
	BLEU
baseline	10.8
+English ASR	19.9
+English ASR: decoder	10.5
+English ASR: encoder	16.6



Ablation: model parameters

Spanish to English, *N* = 20 hours

	BLEU
baseline	10.8
+English ASR	19.9
+English ASR: decoder	10.5
+English ASR: encoder	16.6

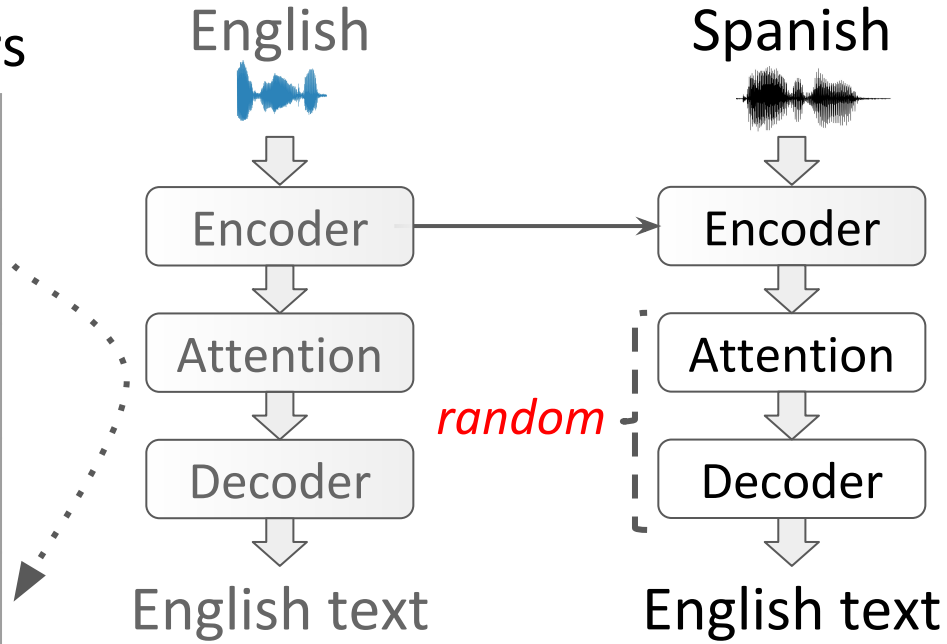


… transferring encoder only parameters works well!

Ablation: model parameters

Spanish to English, *N* = 20 hours

	BLEU
baseline	10.8
+English ASR	19.9
+English ASR: decoder	10.5
+English ASR: encoder	16.6



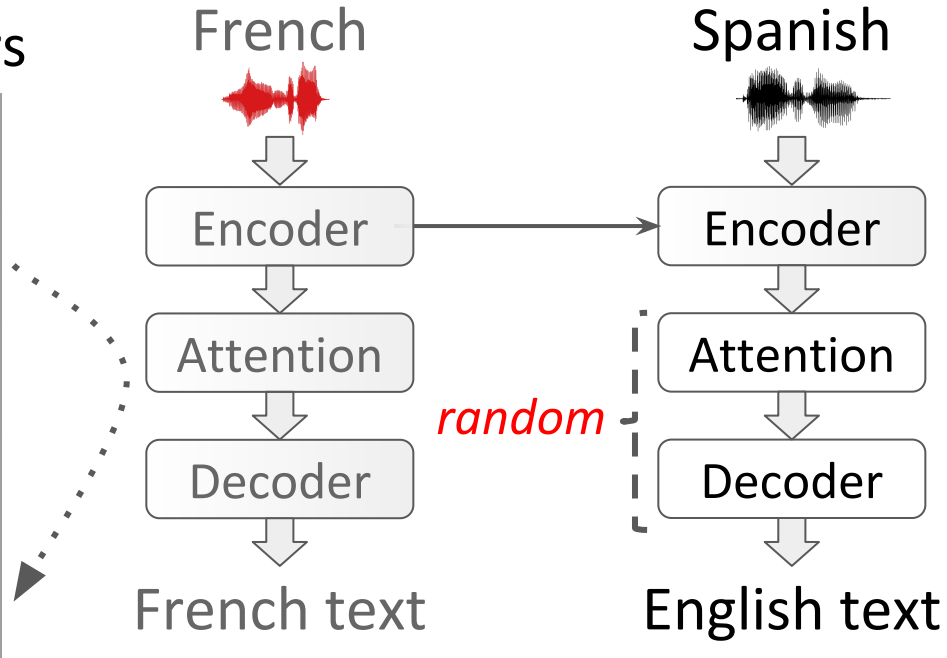
… can pretrain on a language different from both source and target in ST pair

Pretraining on French

Spanish to English, *N* = 20 hours

	BLEU
baseline	10.8
+English ASR	19.9
+English ASR: encoder	16.6
+French ASR: encoder	?

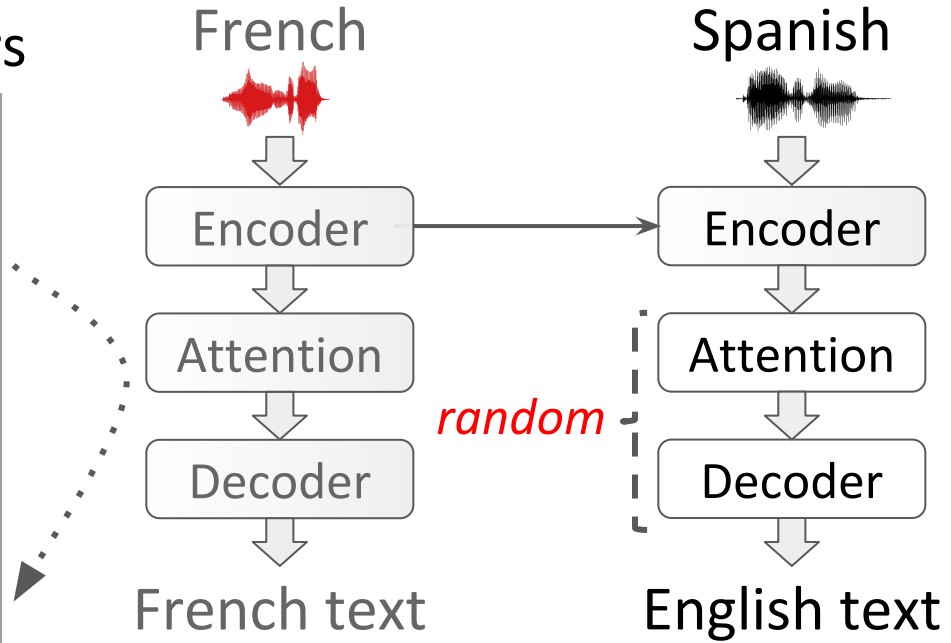
*only 20 hours of French ASR



Pretraining on French

Spanish to English, *N* = 20 hours

	BLEU
baseline	10.8
+English ASR	19.9
+English ASR: encoder	16.6
+French ASR: encoder	**12.5**



French ASR helps Spanish-English ST

Takeaways

- Pretraining on a different language helps
- transfer all model parameters for best gains
- encoder parameters account for most of these

… useful when target vocabulary is different

… Mboshi-French ST

Mboshi-French ST

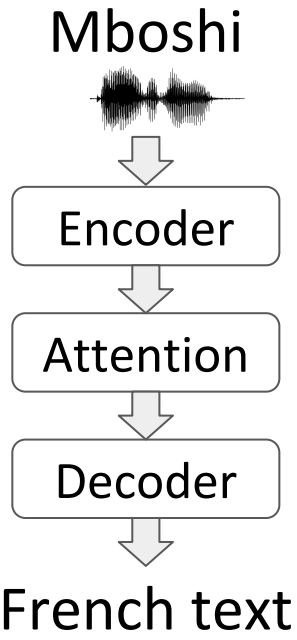
- ST data by Godard et al. 2018
  - **~4 hours** of speech, paired with French translations

- Mboshi
  - Bantu language, Republic of Congo
  - Unwritten
  - ~160K speakers

Mboshi-French: Results

Mboshi to French, *N* = 4 hours

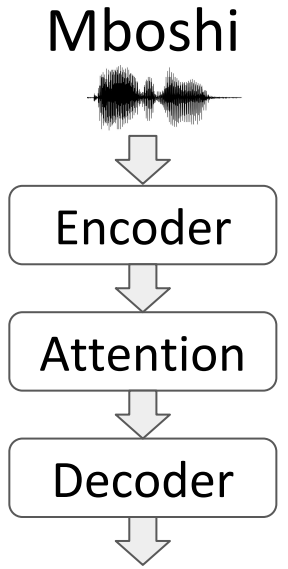
	BLEU
baseline	?



Mboshi-French: Results

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5

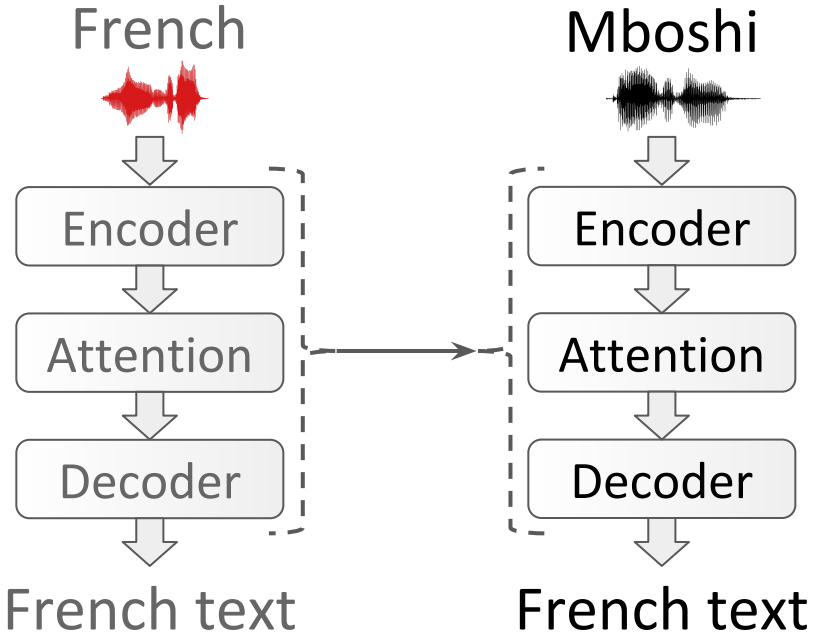


*outperformed by a naive baseline

Pretraining on French ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	?

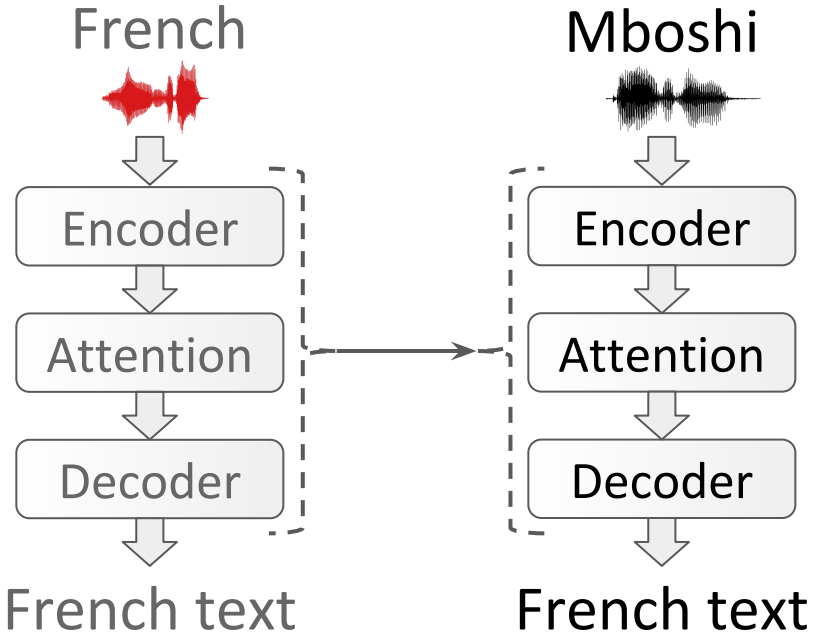


transfer all parameters

Pretraining on French ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	5.9

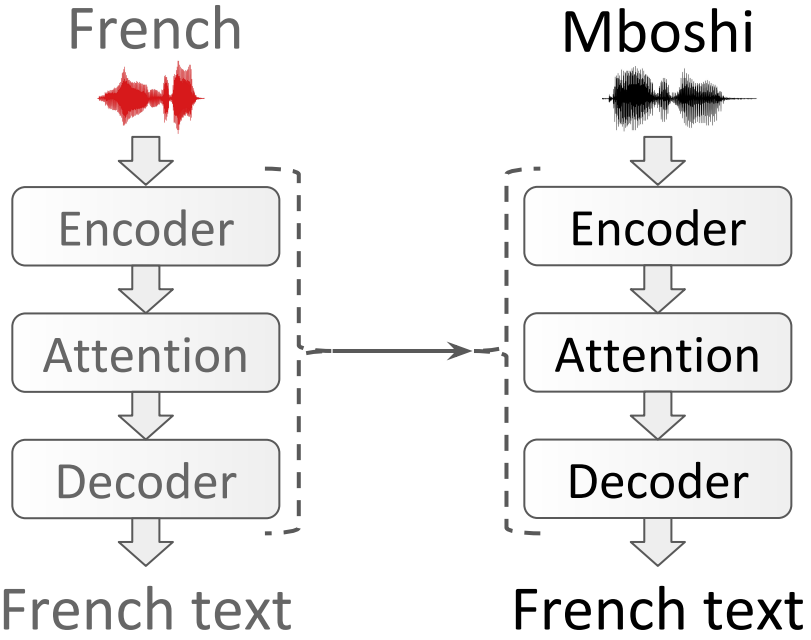


French ASR helps Mboshi-French ST

Pretraining on French ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	5.9

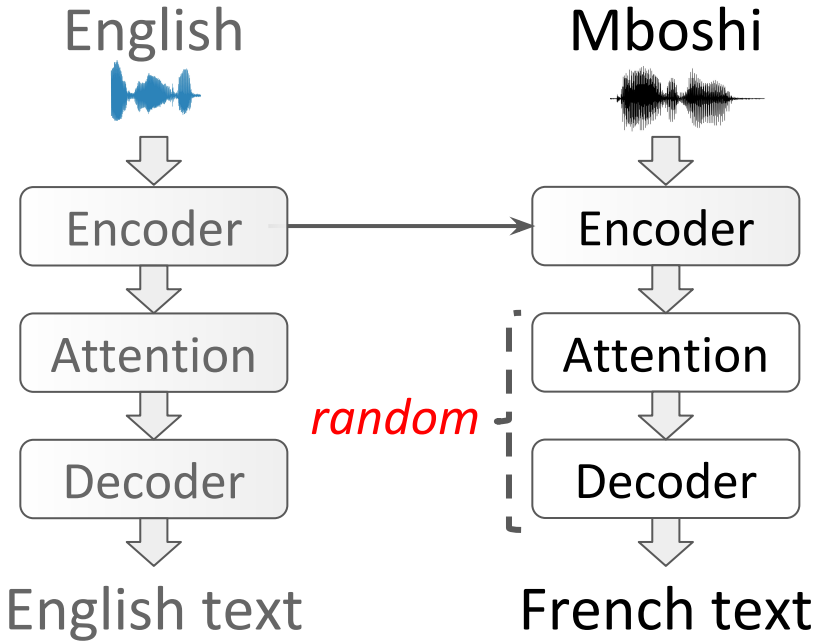


French ASR helps Mboshi-French ST

Pretraining on English ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	5.9
+English ASR: encoder	?

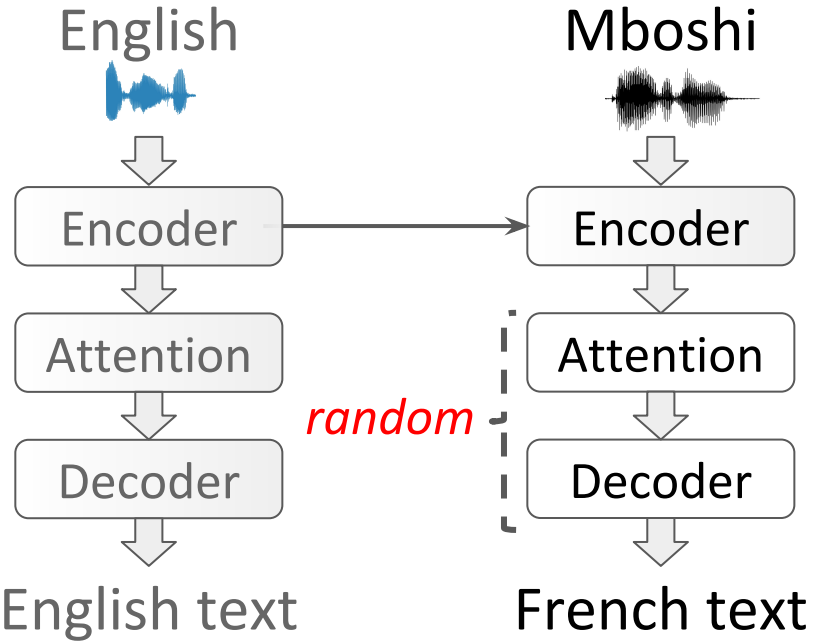


using encoder trained on a lot more data

Pretraining on English ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	5.9
+English ASR: encoder	5.3



English ASR helps Mboshi-French ST

Pretraining on French ASR: can transfer all parameters

… but only 20 hours of data

Pretraining on English ASR: trained on a lot more data (300 hours)

… but can only transfer encoder parameters

Pretraining on French ASR: can transfer all parameters

… but only 20 hours of data

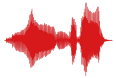
Pretraining on English ASR: trained on a lot more data (300 hours)

… but can only transfer encoder parameters

… combine both?

Pretraining on French and English ASR

French



Encoder

Attention

Decoder

French text

20 hours

English

Encoder

Attention

Decoder

English text

300 hours

Pretraining on French and English ASR

French

Encoder

Attention

Decoder

French text

20 hours

Mboshi

Encoder

Attention

Decoder

French text

4 hours

English

Encoder

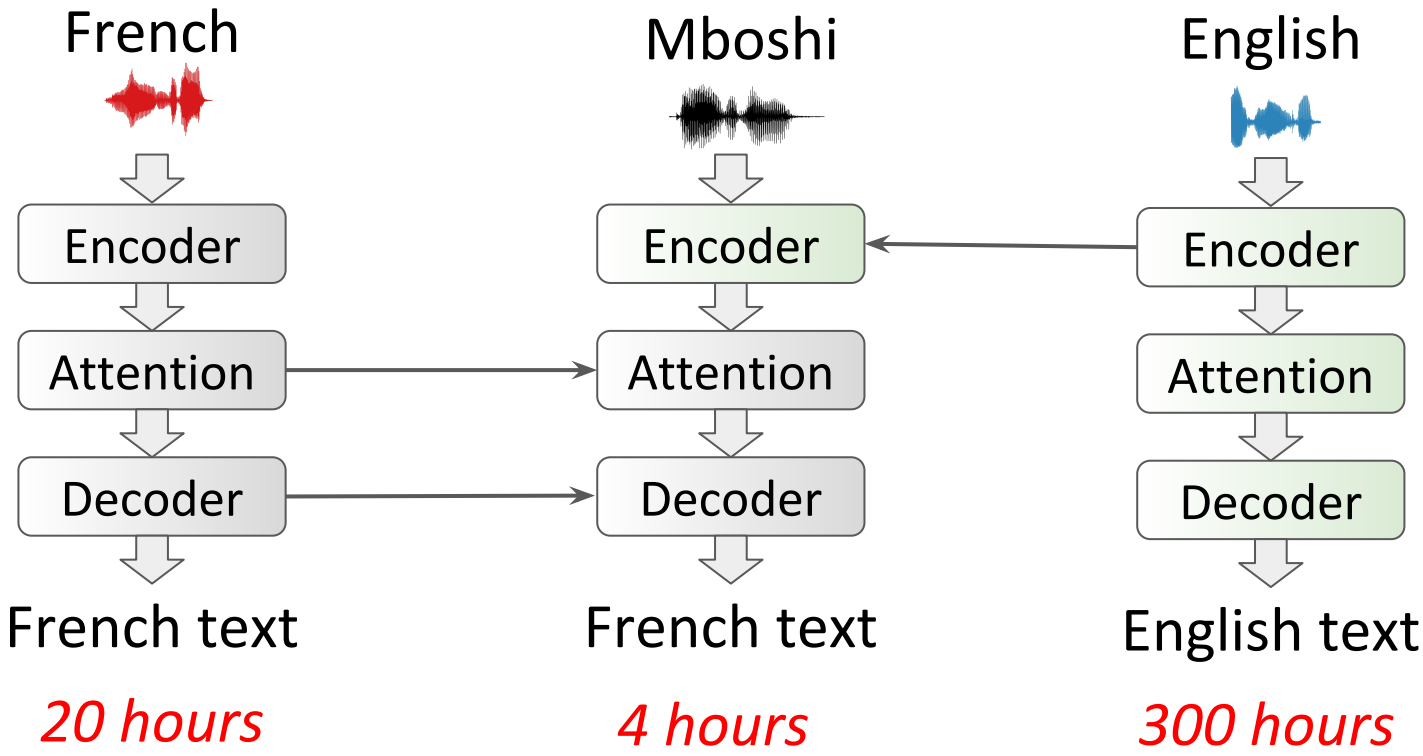
Attention

Decoder

English text

300 hours

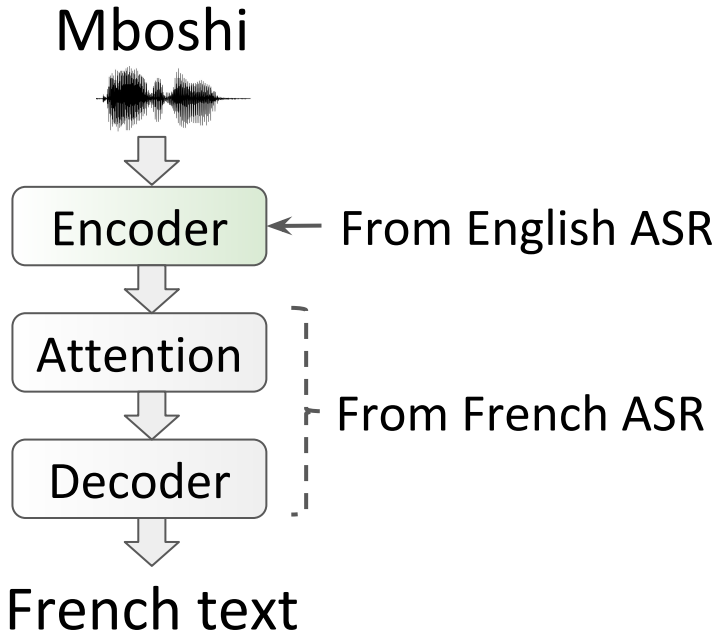
Pretraining on French and English ASR



Pretraining on English ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	5.9
+English ASR: encoder	5.3
+English ASR: encoder +French ASR: remaining	?

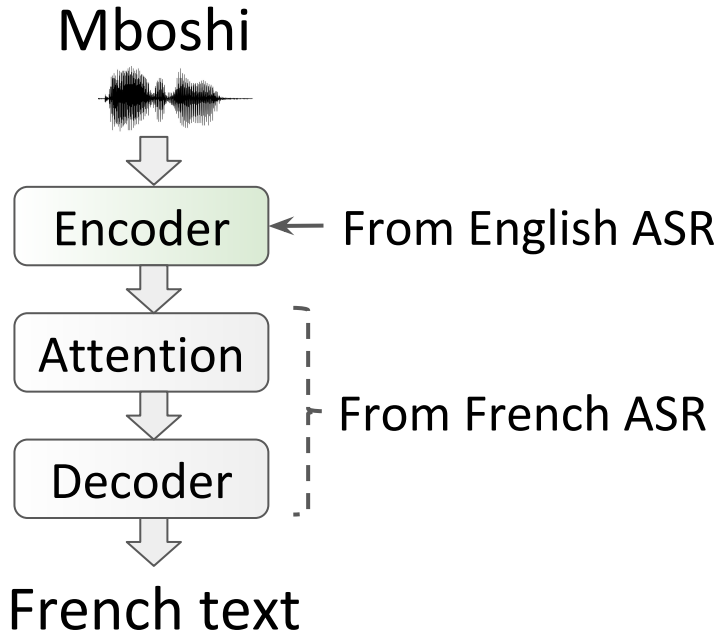


Pretraining on English ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	5.9
+English ASR: encoder	5.3
+English ASR: encoder +French ASR: remaining	7.1

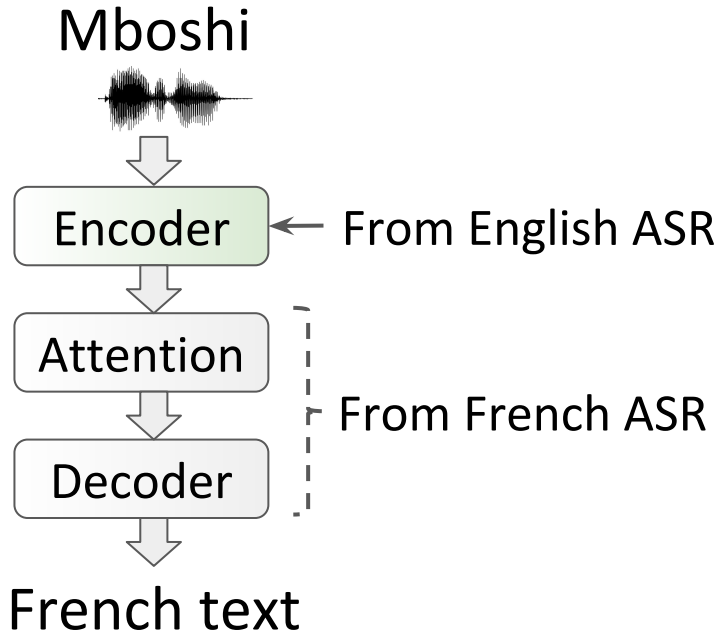
combining gives the best gains



Pretraining on English ASR

Mboshi to French, *N* = 4 hours

	BLEU
baseline	3.5
+French ASR: all	5.9
+English ASR: encoder	5.3
+English ASR: encoder +French ASR: remaining	7.1



BLEU score is still low … but above naive baseline

Conclusions

- Pretraining on high-resource ASR improves low-resource ST
- Potentially useful for endangered and/or unwritten languages
- Bootstrap ST in time-critical scenarios
- Future work: experiments on more languages, multilingual training with joint vocabulary

Thanks

- Anonymous reviewers, Edinburgh NLP members
- Source code available at: https://github.com/0xSameer/ast

I am looking for full-time positions starting November 2019!

- 4th June, 3:30-5 pm - “Fluent Translations from Disfluent Speech in End-to-End Speech Translation”, Salesky et al.
- 5th June, 10:30-10:48 am - “Neural Machine Translation of Text from Non-Native Speakers”, Anastasopoulos et al.

Backup

Mboshi-French naive baseline

model	pretrain	BLEU	Pr.	Rec.
fr-top-8w	–	0	23.5	22.2
fr-top-10w	–	0	20.6	24.5
en-300h	–	0	0.2	5.7
fr-20h	–	0	4.1	3.2
mb-fr-4h	–	3.5	18.6	19.4
	fr-20h	5.9	23.6	20.9
	en-300h	5.3	23.5	22.6
	en + fr	**7.1**	**26.7**	**23.1**

Why does pretraining help?

- Speaker invariance
  - ASR data contains audio from 100s of speakers
- Learning to factor out background noise (?)

BLEU	**Baseline**	**+English ASR**
50 speakers	7.2	17.5 (+143 %)
136 speakers	10.8 (+ 50%)	19.9 (+14%)

Spanish-English ST

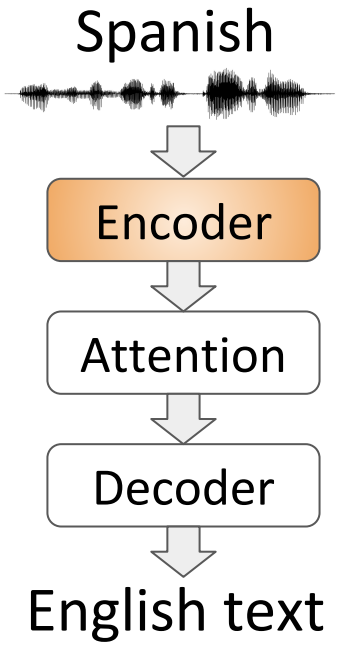
***N* hrs**	**2.5h**	**5h**	**10h**	**20h**	**50h**	**160h Weiss**
baseline	2.1	1.8	2.1	10.8	22.7	47.3
+ASR	5.7	9.1	14.5	20.2	28.3	---
	+3.6	+7.3	+12.4	+9.4	+5.5	

**results on Fisher test set ...*

Spanish-English ST

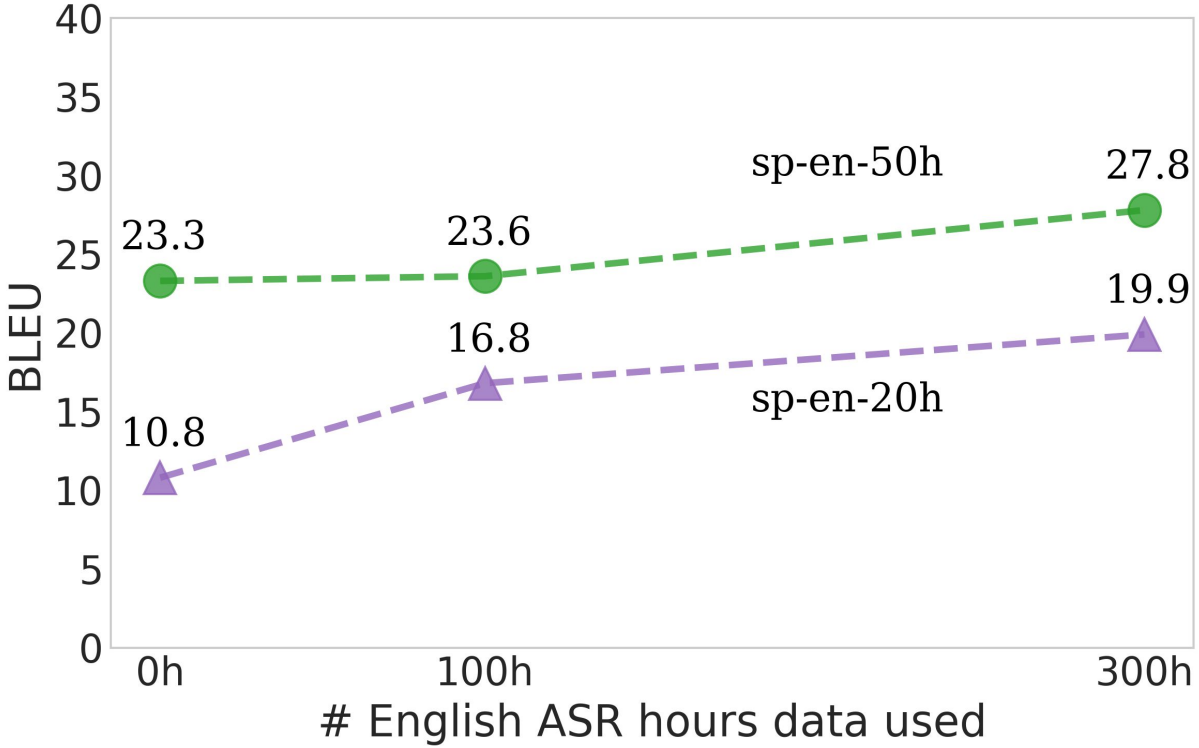
Spanish to English, *N* = 20 hours

	BLEU
baseline	10.8
+En ASR: 300h	16.6
+Fr ASR:20h	**12.5**
+En ASR: 20h	**13.2**

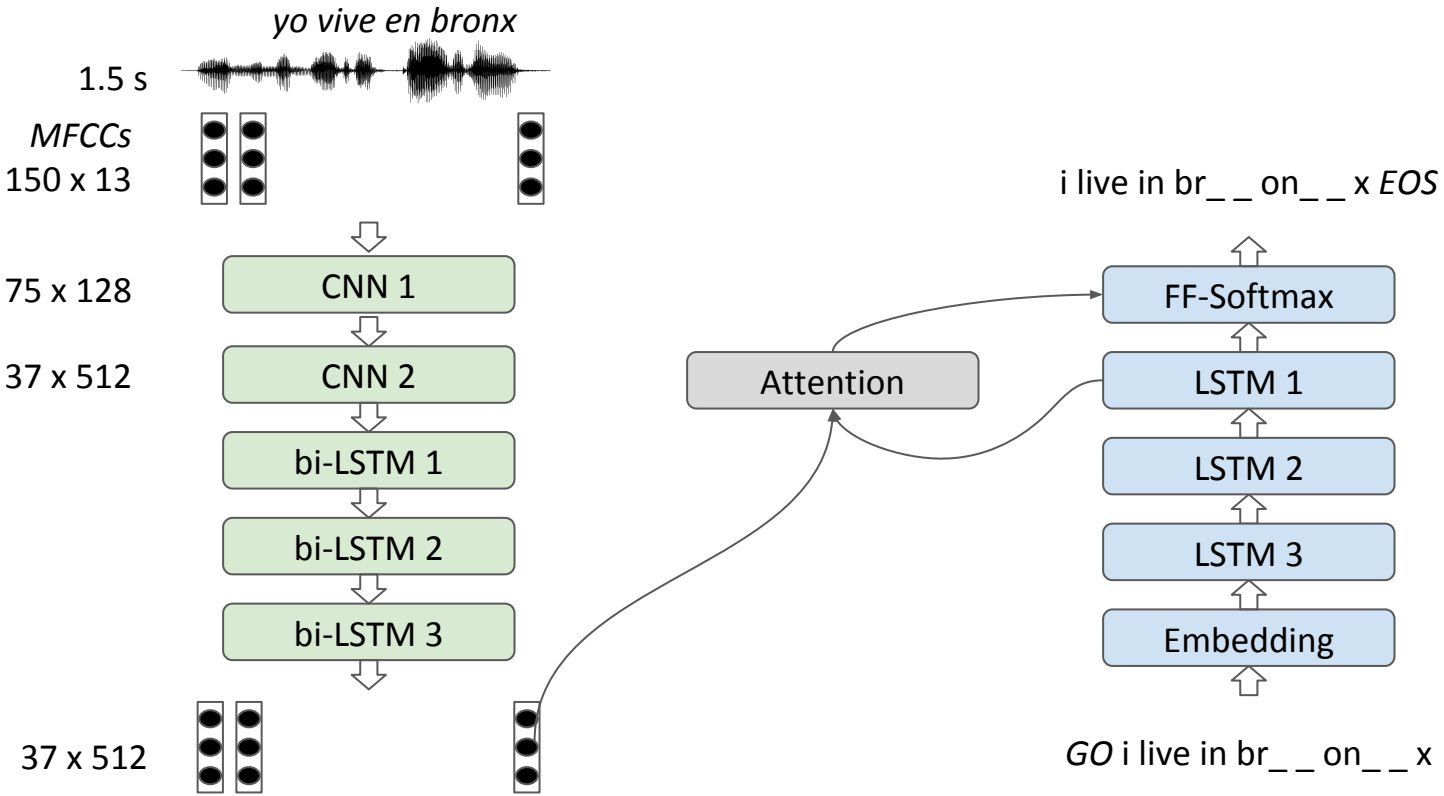


… French ASR helps improve Spanish-English ST

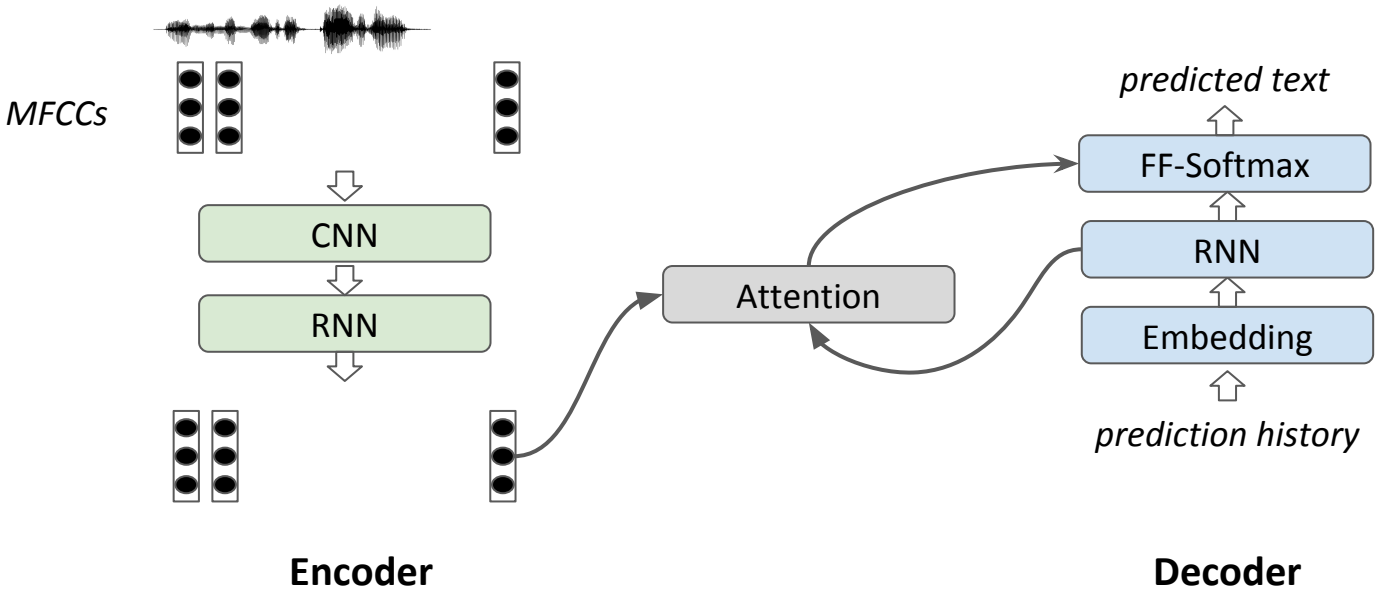
Spanish-English ST



Neural model



Neural model



100s of hours of **monolingual** speech paired with text available

… typically used to train ASR systems

Gülçehre et al., 2015
Toshniwal et al., 2018

English text

Spanish Audio  Sequence-to-Sequence **English** text

~20 hours of Spanish-English