Sec 3:

Leveraging Data Sources

Available data

Techniques

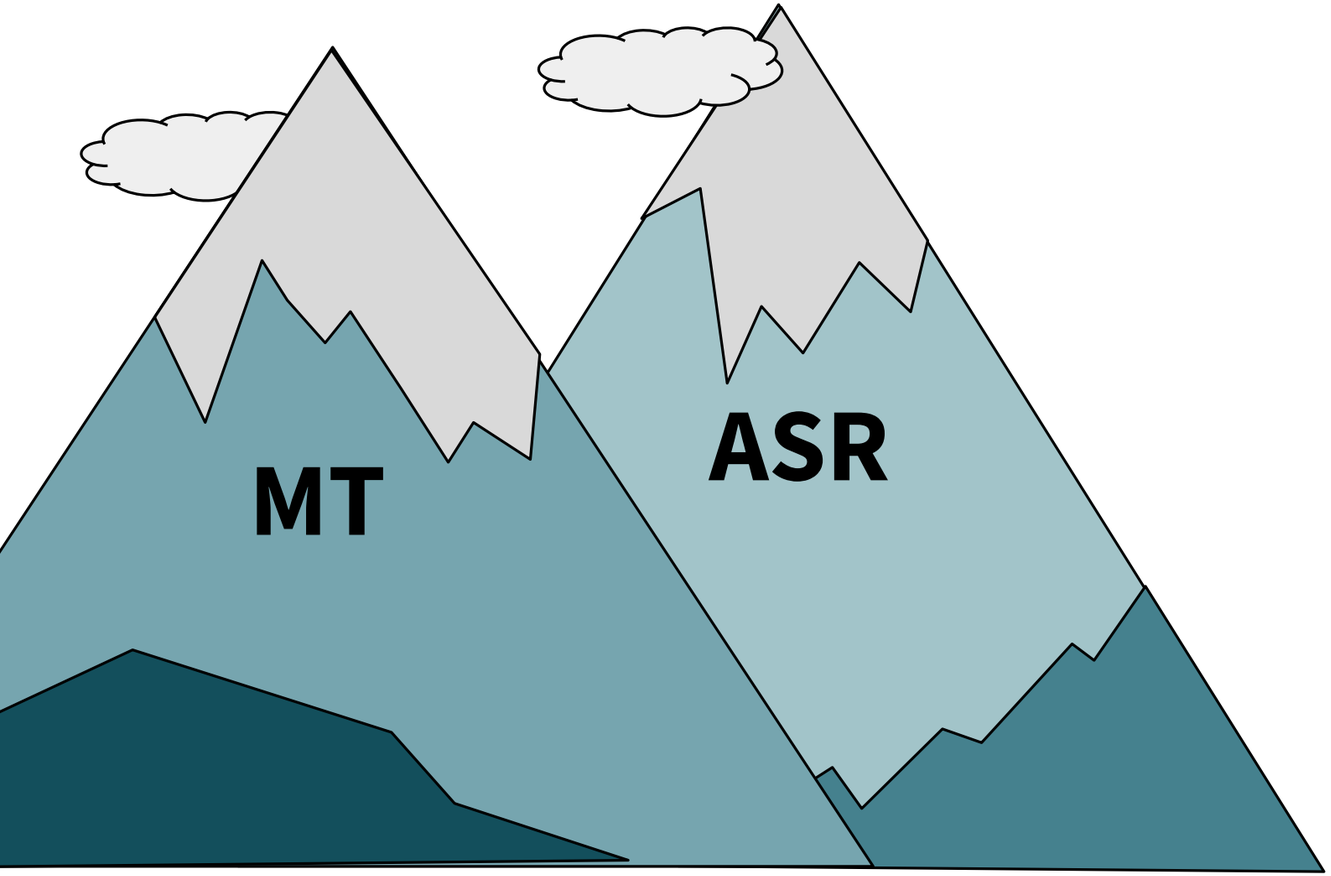
- Multi-task learning
- Transfer learning and pretraining
- Knowledge distillation

Alternate data representations

Sec 3.1

Available Data

Available data



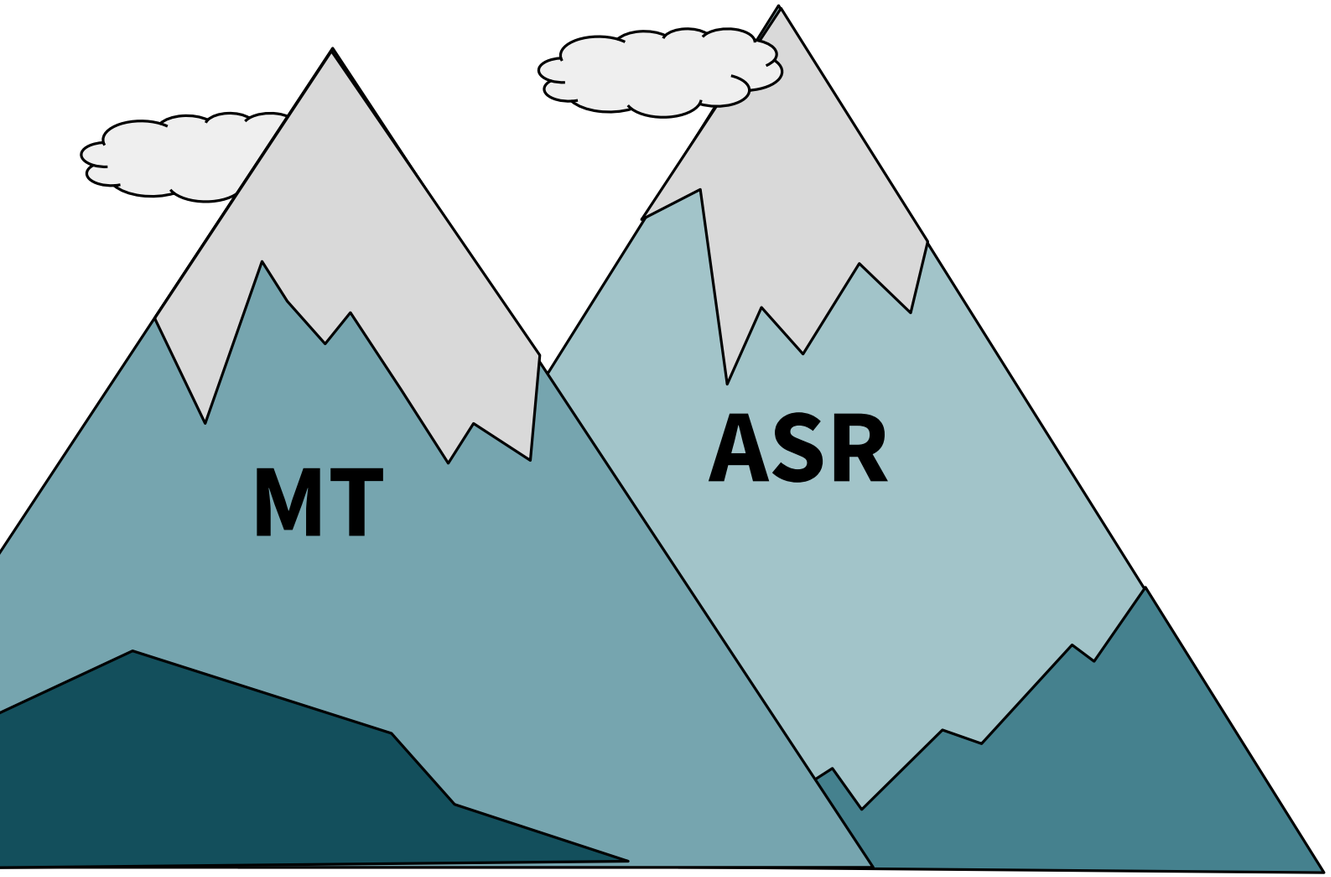
(text, translation)

(audio, transcript)

ST

(audio, transcript, translation)

Available data



(text, translation) (audio, transcript)

Question: Why so few data?

Answer: High creation costs!

1. Find good data (e.g. audio+transcr+transl., free)
2. Download and clean
3. Segment transcripts and translations
4. Align transcripts and translations
5. Align transcripts and audio
6. Filter wrong/poor alignments
7. Pack in suitable format, extract features

MuST-C (Cattoni et al., 2021)

ST

(audio, transcript, translation)

Available data (≥ 20 hrs of speech)

(*no name*)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(*no name*)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Available data (≥ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Half of these corpora were built in the last 2 years

Available data (≥ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Trend (1): increasing data size (>200 hours of translated speech)

Available data (≥ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Trend (2): more language directions

Available data (≥ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Trend (3): multilinguality + non-English speech

Available data (≥ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Trend (4): same segmentation across datasets

Available data (≥ 20 hrs of speech)

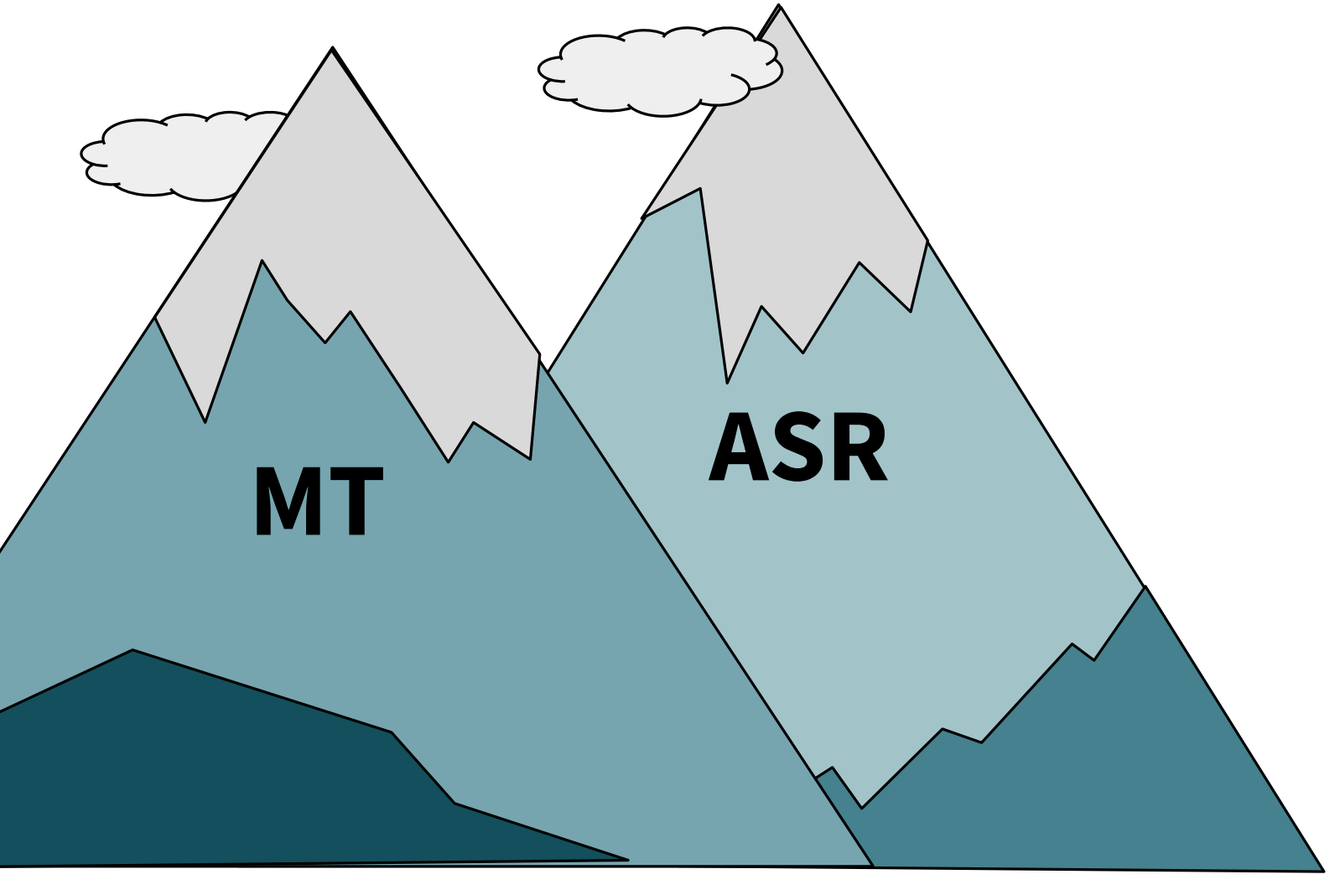
(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Trend (5): common test data across language pairs

Sec 3.2

Techniques

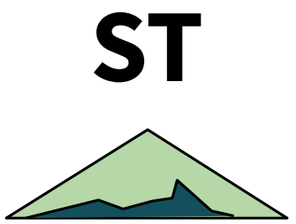
Recap: Available data



(text, translation)

(audio, transcript)

Can we make use of this large amount of data?



(audio, transcript, translation)

Multi-task learning

Definition:

“Multi-task learning improves generalization by leveraging the domain-specific information contained in the training signals of related tasks”

— Caruana, R. (1998)

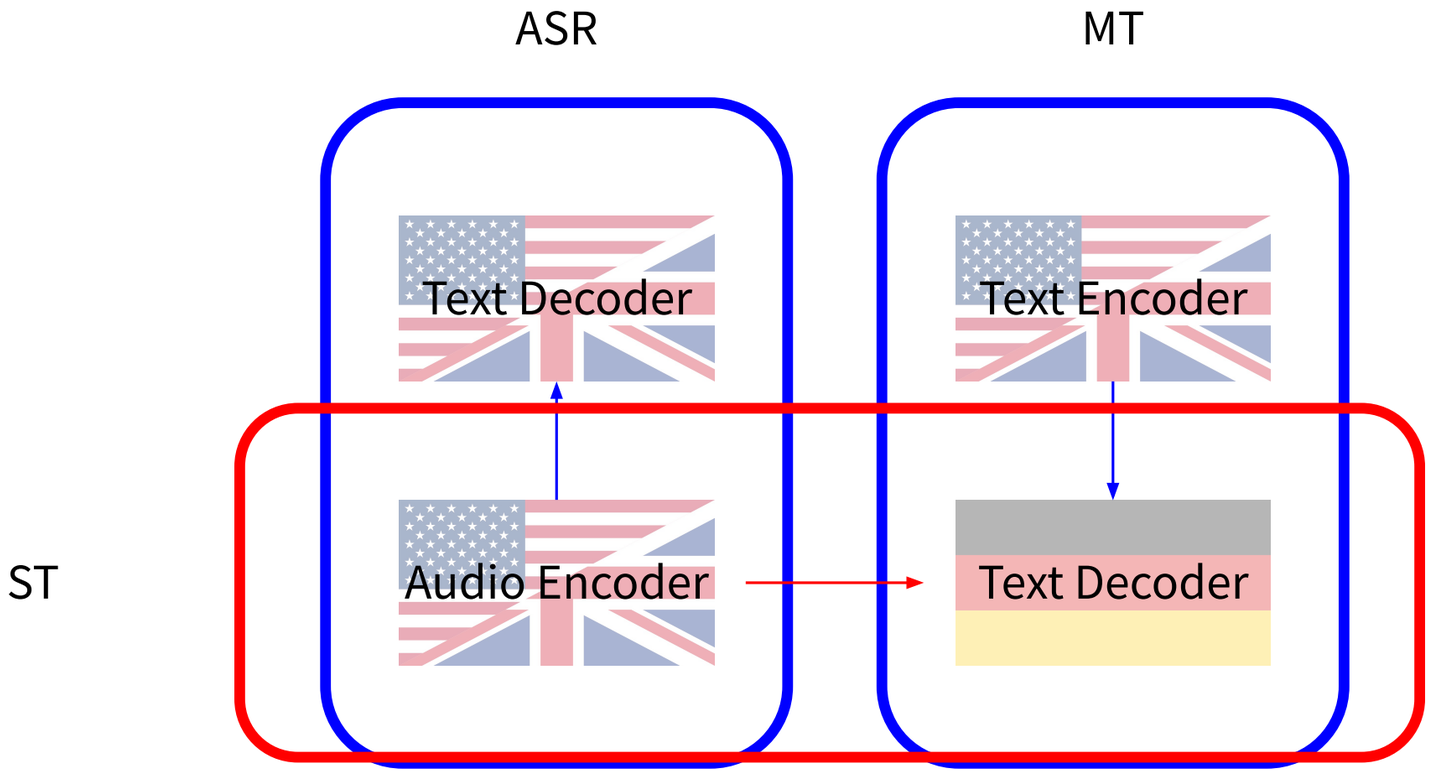
Transfer Learning

Definition:

"Transfer learning and domain adaptation refer to the situation where what has been learned in one setting … is exploited to improve generalization in another setting"

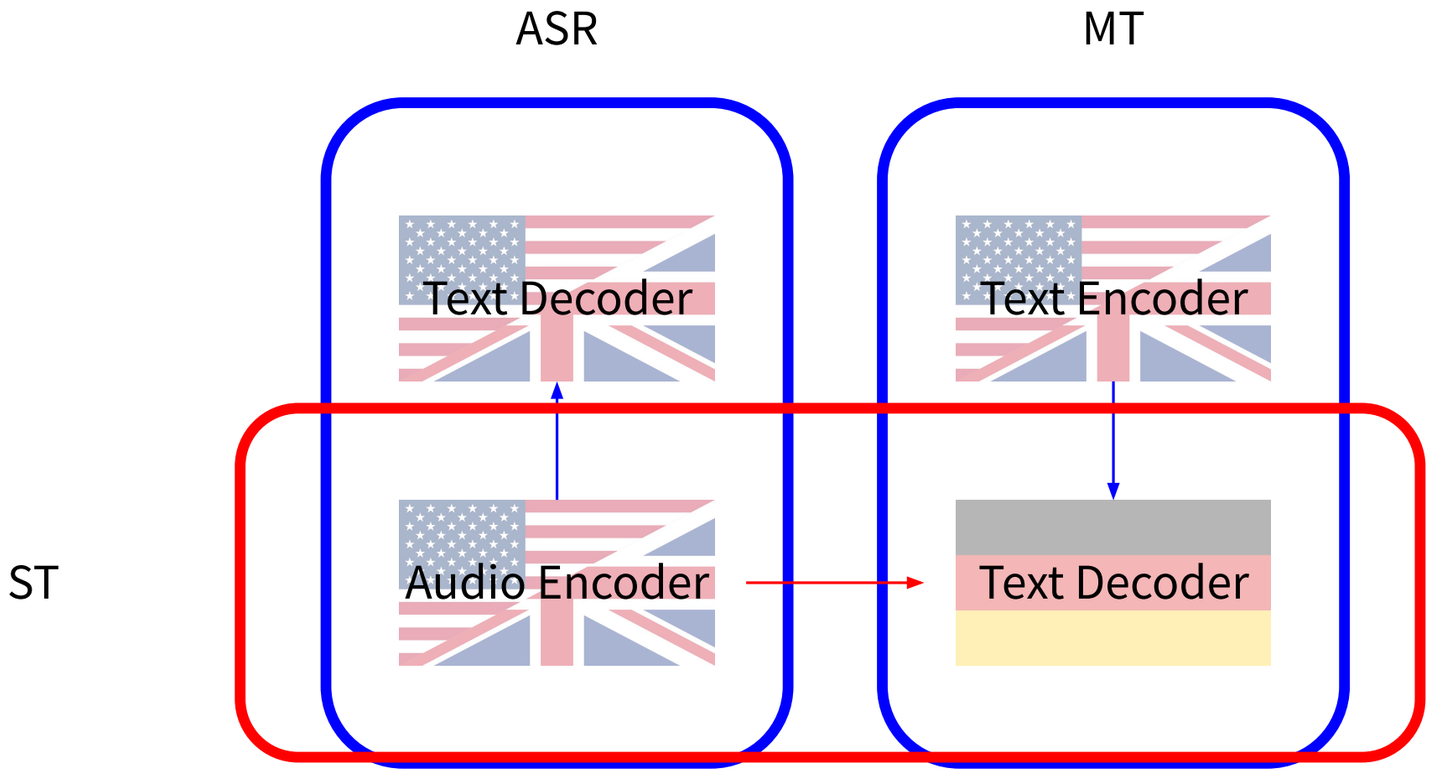
— Page 526, Deep Learning, 2016.

Setting


ASR
MT
Text Decoder
Text Encoder
ST
Audio Encoder
Text Decoder

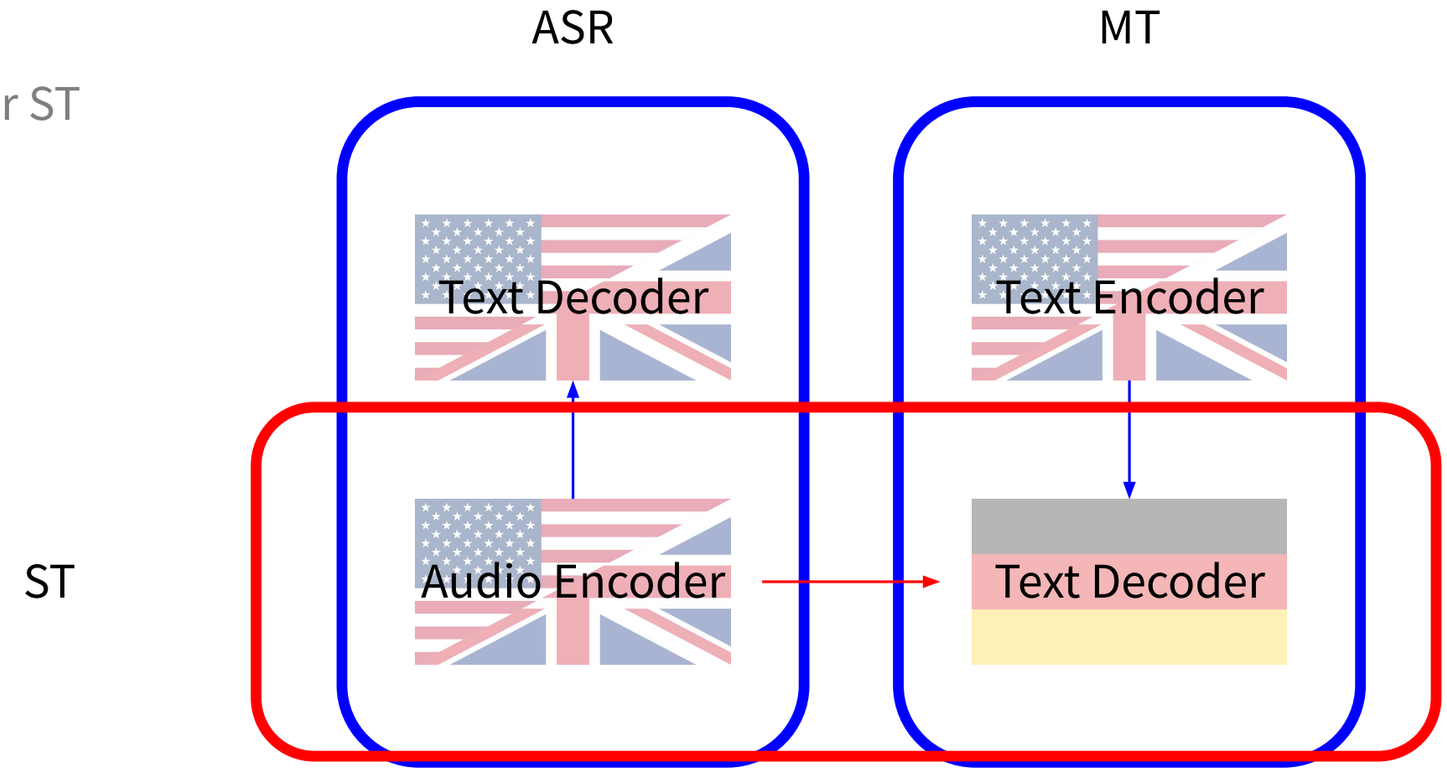
Setting

- Multi-task
  - Train all three tasks jointly



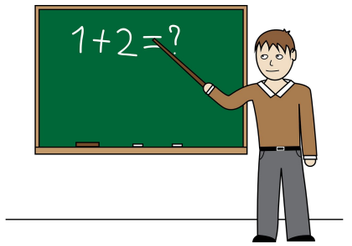
Setting

- Multi-task
- Pre-training
  - Train ASR and MT
  - Reuse part of the model for ST



Setting

- Multi-task
- Pre-training
- Knowledge distillation
  - Take MT model
  - Train ST based on training signal from MT



ST