Neural Network Models for …

Paraphrase Identification

Semantic Textual Similarity

Natural Language Inference

Question Answering

Wuwei Lan and Wei Xu

The Ohio State University
Department of Computer Science and Engineering

• • •

Paraphrase Identification

Semantic Textual Similarity

Natural Language Inference

Question Answering

…
Paraphrase Identification
Semantic Textual Similarity
Natural Language Inference
Question Answering

Sentence Pair Modeling

The General Neural Framework

...

The General Neural Framework

Type I: The Sentence Encoding-based Models

The General Neural Framework …

Type I: The Sentence Encoding-based Models

- Gated recurrent average network [Wieting and Gimpel, 2017]
- Directional self-attention network [Shen et al., 2017]
- **InferSent** BiLSTM with max-pooling [Conneau et al., 2017]
- Gumbel Tree-LSTM [Choi et al., 2017]
- **SSE** Shortcut-stacked BiLSTM [Nie and Bansal, 2017]
- and many others …

The General Neural Framework

Type I: The Sentence Encoding-based Models

- Gated recurrent average network [Wieting and Gimpel, 2017]
- Directional self-attention network [Shen et al., 2017]
- **InferSent** BiLSTM with max-pooling [Conneau et al., 2017]
- Gumbel Tree-LSTM [Choi et al., 2017]
- **SSE** Shortcut-stacked BiLSTM [Nie and Bansal, 2017]
- and many others …

Type II: The Word Interaction-based Models

The General Neural Framework …

Type I: The Sentence Encoding-based Models

- Gated recurrent average network [Wieting and Gimpel, 2017]
- Directional self-attention network [Shen et al., 2017]
- **InferSent** BiLSTM with max-pooling [Conneau et al., 2017]
- Gumbel Tree-LSTM [Choi et al., 2017]
- **SSE** Shortcut-stacked BiLSTM [Nie and Bansal, 2017]
- and many others …

Type II: The Word Interaction-based Models

- **PWIM** Pairwise word interaction [He and Lin, 2016]
- Subword-based pairwise word interaction [Lan and Xu, 2018]
- Attention based CNN [Yin et al., 2016]
- **DecAtt** Decomposable attention [Parikh et al., 2017]
- **ESIM** Enhanced LSTM for NLI [Chen et al., 2017]
- and many others …

Motivation for this Work

		SNLI	MNLI	Quora	URL	PIT	STS14	WikiQA	TrecQA
Type I	**InferSent**	0.845	-	-	-	-	0.700	-	-
	SSE	0.860	0.746	-	-	-	-	-	-
Type II	**DecAtt**	0.863	-	0.865	-	-	-	-	-
	ESIM_seq	0.880	0.723	-	-	-	-	-	-
	ESIM_tree	0.878	-	-	-	-	-	-	-
	ESIM_seq+tree	0.886	-	-	-	-	-	-	-
	PWIM	-	-	-	0.749	0.667	0.767	0.709	0.759

- Previous systems only reported results on a few selected datasets.

Reproduced Results for Sentence Pair Modeling

		SNLI	MNLI	Quora	URL	PIT	STS14	WikiQA	TrecQA
Type I	**InferSent**	0.846	0.705	0.866	0.746	0.451	0.715	0.287	0.521
	SSE	0.855	0.740	**0.878**	0.650	0.422	0.378	0.624	0.628
Type II	**DecAtt**	0.856	0.719	0.865	0.652	0.430	0.317	0.603	0.660
	ESIM_seq	0.870	0.752	0.850	0.748	0.520	0.602	0.652	**0.771**
	ESIM_tree	0.864	0.736	0.755	0.740	0.447	0.493	0.618	0.698
	ESIM_seq+tree	**0.871**	**0.753**	0.854	0.759	0.538	0.589	0.647	0.749
	PWIM	0.822	0.722	0.834	**0.761**	**0.656**	**0.743**	**0.706**	0.739

- We filled in the blanks and systematically compared 7 models on 8 datasets.

Reproduced Results for Sentence Pair Modeling

		SNLI	MNLI	Quora	URL	PIT	STS14	WikiQA	TrecQA
Type I	**InferSent**	0.846	0.705	0.866	0.746	0.451	0.715	0.287	0.521
	SSE	0.855	0.740	**0.878**	0.650	0.422	0.378	0.624	0.628
Type II	**DecAtt**	0.856	0.719	0.865	0.652	0.430	0.317	0.603	0.660
	ESIM_seq	0.870	0.752	0.850	0.748	0.520	0.602	0.652	**0.771**
	ESIM_tree	0.864	0.736	0.755	0.740	0.447	0.493	0.618	0.698
	ESIM_seq+tree	**0.871**	**0.753**	0.854	0.759	0.538	0.589	0.647	0.749
	PWIM	0.822	0.722	0.834	**0.761**	**0.656**	**0.743**	**0.706**	0.739

- We filled in the blanks and systematically compared 7 models on 8 datasets.
- No model consistently performs well across all tasks!

• • •

• • •

Paraphrase Identification

paraphrase non-paraphrase

Dataset: Quora (400k), URL (51k), PIT (16k)

...

Paraphrase Identification

paraphrase non-paraphrase

Dataset: Quora (400k), URL (51k), PIT (16k)

Semantic Textual Similarity

score[0,5]

Dataset: STS14 (11k)

…

Paraphrase Identification

paraphrase | non-paraphrase

Dataset: Quora (400k), URL (51k), PIT (16k)

Semantic Textual Similarity

score[0,5]

Dataset: STS14 (11k)

Natural Language Inference

entailment | neutral | contradiction

Dataset: SNLI (570k), MNLI (432k)

…

Paraphrase Identification

paraphrase | non-paraphrase

Dataset: Quora (400k), URL (51k), PIT (16k)

Semantic Textual Similarity

score[0,5]

Dataset: STS14 (11k)

Natural Language Inference

entailment | neutral | contradiction

Dataset: SNLI (570k), MNLI (432k)

Question Answering

true | false

Dataset: WikiQA (12k), TrecQA (56k)

• • •

Question Answering [1]

true

Q: How much is 1 tablespoon of water?

A: In Australia one tablespoon (measurement unit) is 20 mL

[1] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. (EMNLP 2015).
[2] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A Continuously Growing Dataset of Sentential Paraphrases (EMNLP 2017).

…

Question Answering [1]

true

false

Q: How much is 1 tablespoon of water?

A: In Australia one tablespoon (measurement unit) is 20 mL

A: It is abbreviated as t, tb, tbs, tbsp, tblsp, or tblspn.

[1] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. (EMNLP 2015).
[2] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A Continuously Growing Dataset of Sentential Paraphrases (EMNLP 2017).

• • •

Question Answering [1]

true

false

Q: How much is 1 tablespoon of water?

A: In Australia one tablespoon (measurement unit) is 20 mL

A: It is abbreviated as t, tb, tbs, tbsp, tblsp, or tblspn.

Paraphrase Identification [2]

paraphrase

CO2 levels haven't been this high for 3 to 5 million years.

CO2 levels mark 'new era' in the world's changing climate.

[1] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. (EMNLP 2015).
[2] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A Continuously Growing Dataset of Sentential Paraphrases (EMNLP 2017).

…

Question Answering [1]

true

false

Q: How much is 1 tablespoon of water?

A: In Australia one tablespoon (measurement unit) is 20 mL

A: It is abbreviated as t, tb, tbs, tbsp, tblsp, or tblspn.

Paraphrase Identification [2]

paraphrase

non-paraphrase

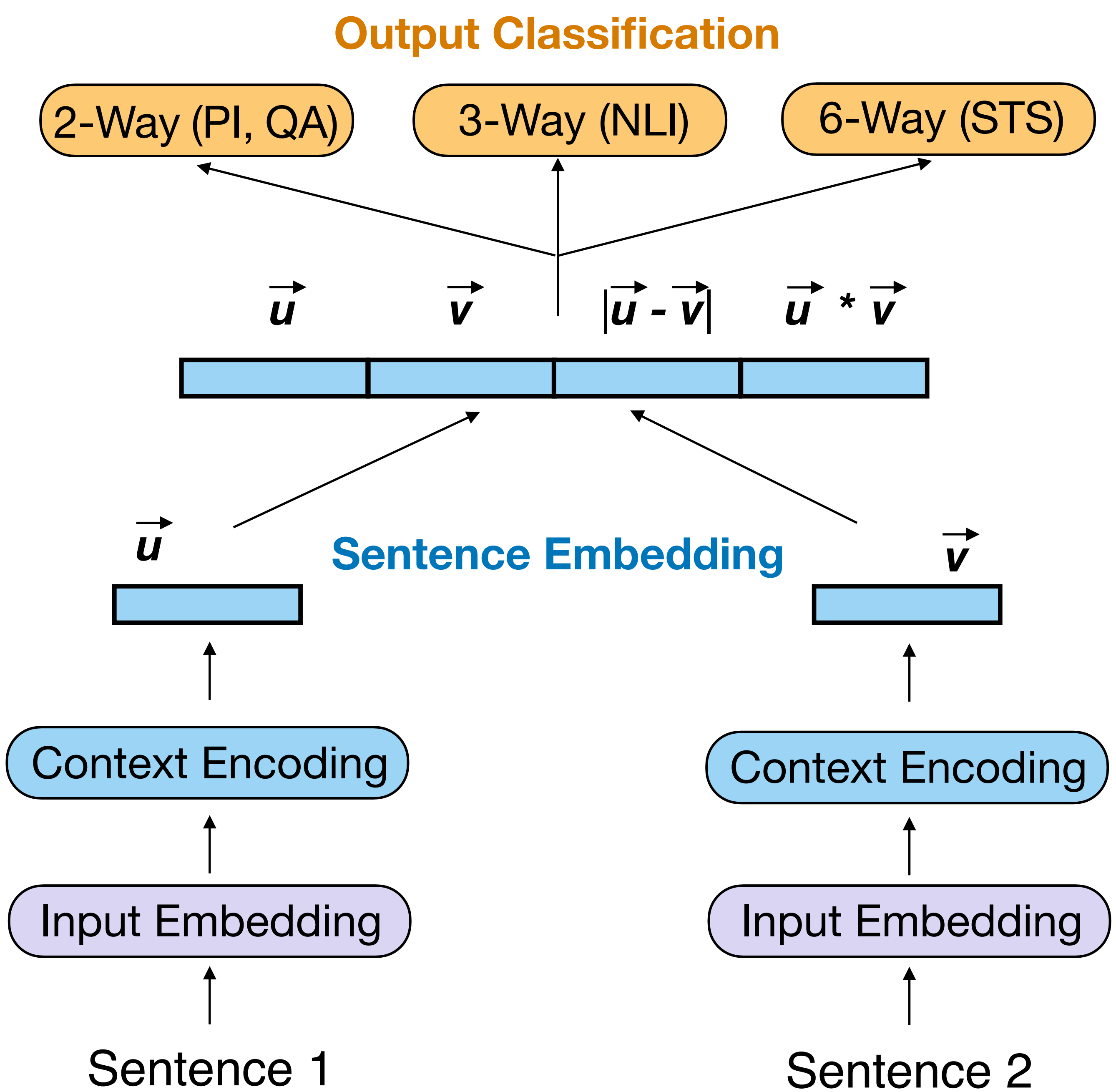
CO_2 levels haven’t been this high for 3 to 5 million years.

CO_2 levels mark ‘new era’ in the world’s changing climate.

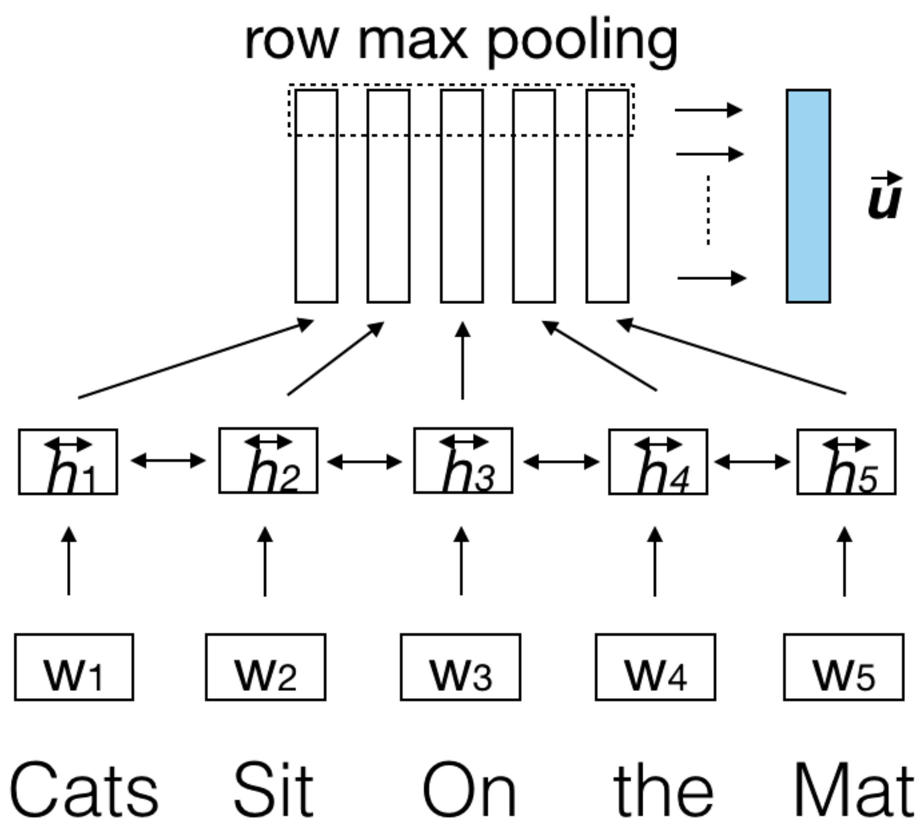
First whole year over 400ppm. We are too complacent with this news .

[1] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. (EMNLP 2015).
[2] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A Continuously Growing Dataset of Sentential Paraphrases (EMNLP 2017).

Type I: Sentence Encoding-based Models



Type I: Sentence Encoding-based Models

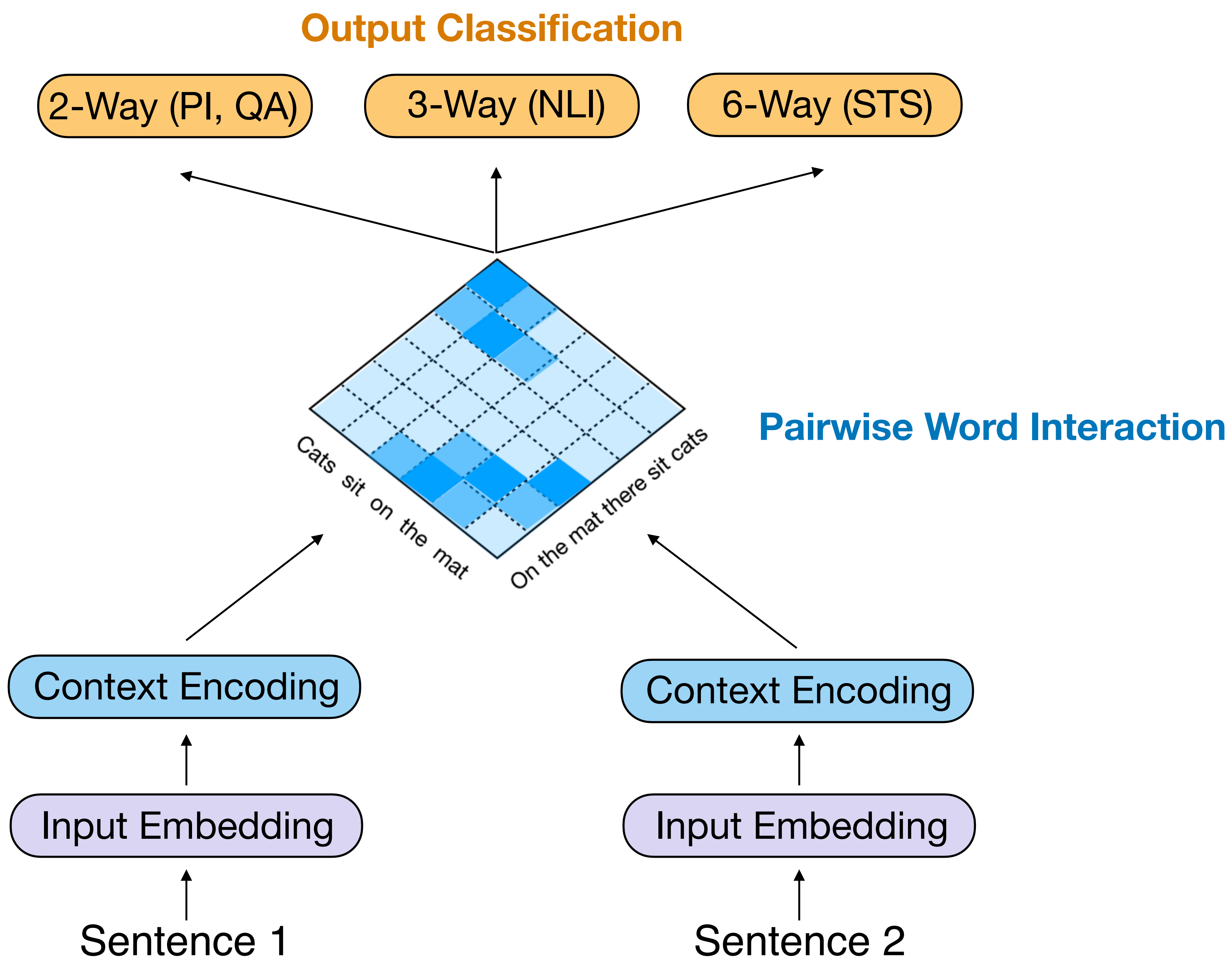


Sentence Embedding

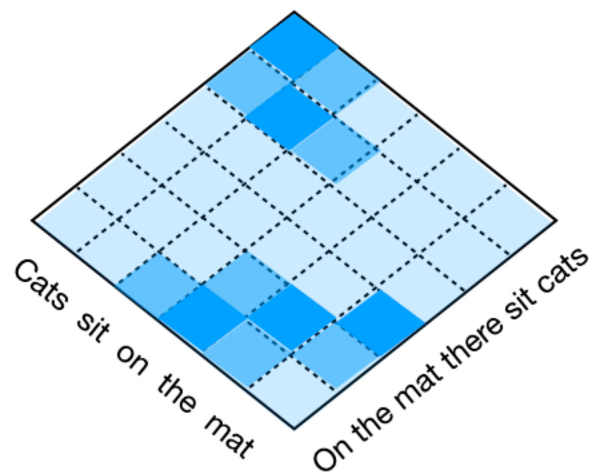
InferSent: 1-layer Bi-LSTM.[3]
SSE: 3-layer Bi-LSTM with skip connection.[4]

[3] Jihun Choi, Kang Min Yoo, and Sang-goo Lee: Unsupervised learning of task-specific tree structures with tree-LSTMs. (EMNLP 2017).
[4] Yixin Nie and Mohit Bansal. Shortcut-stacked sentence encoders for multi-domain inference. (RepEval 2017)

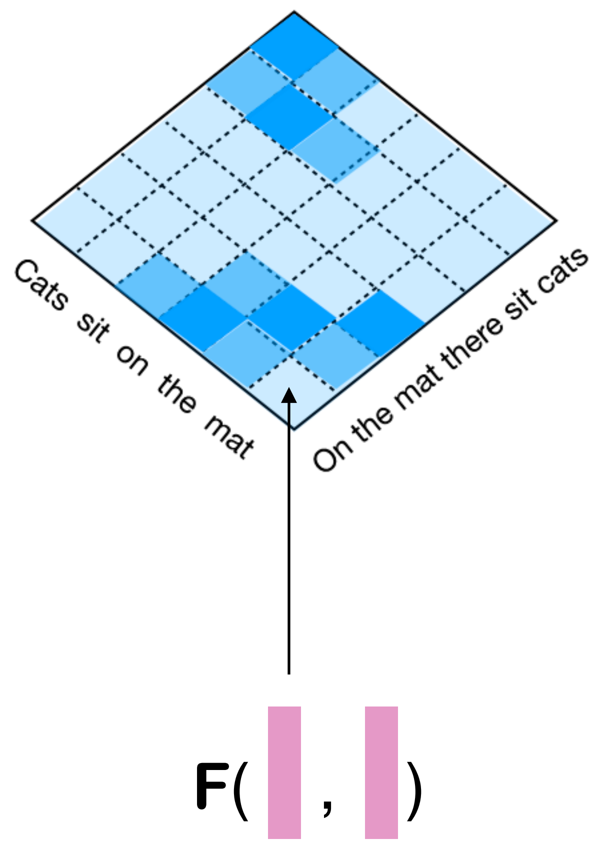
Type II: Word Interaction-based Models



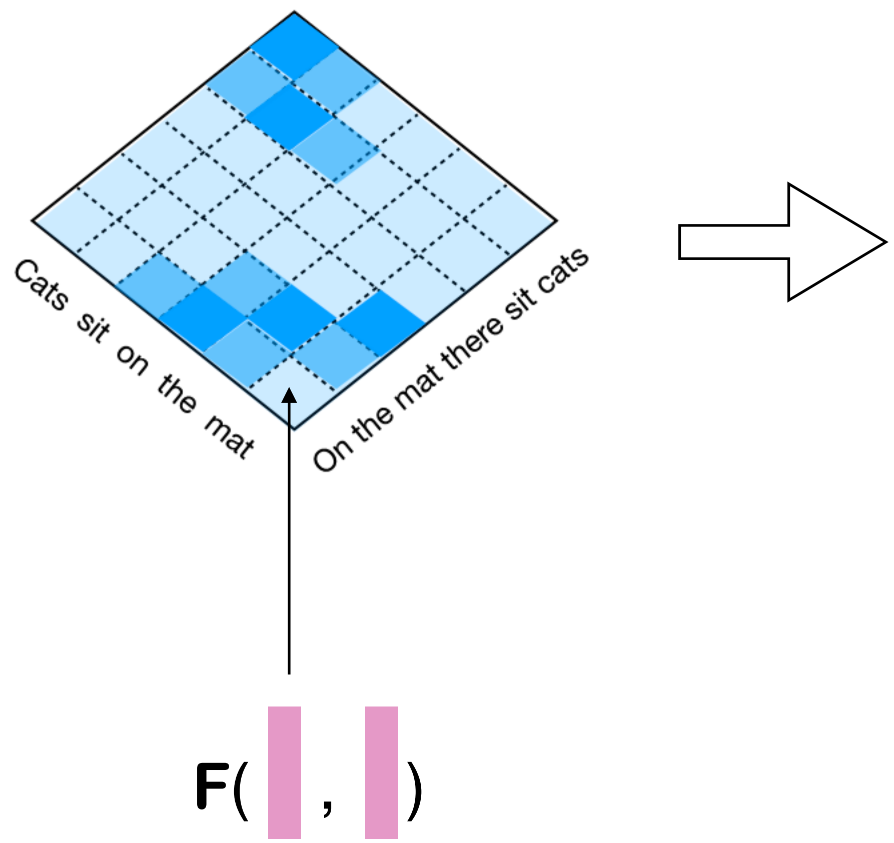
- semantic relation between two sentences depends largely on aligned words/phrases



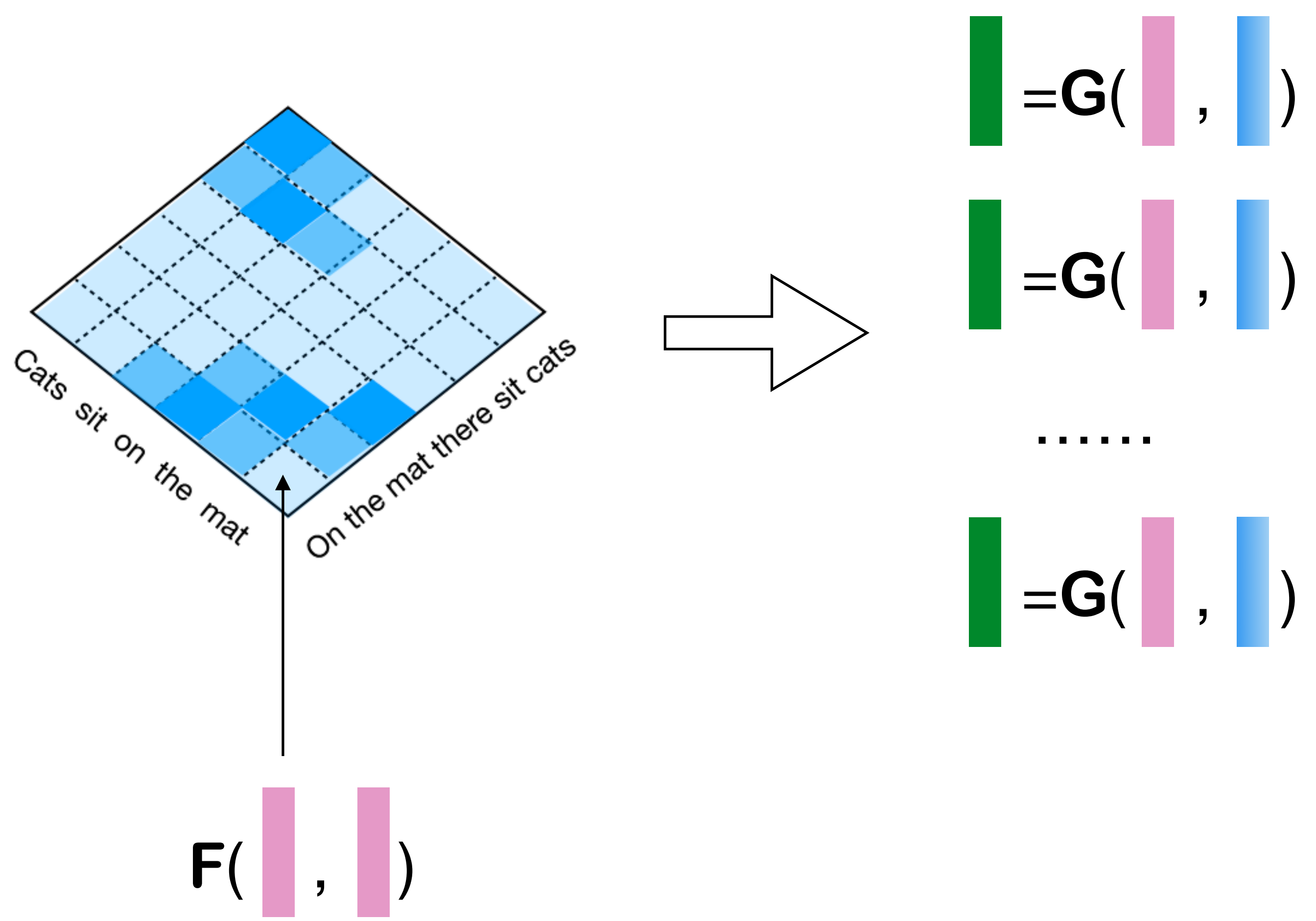
Cats sit on the mat
On the mat there sit cats



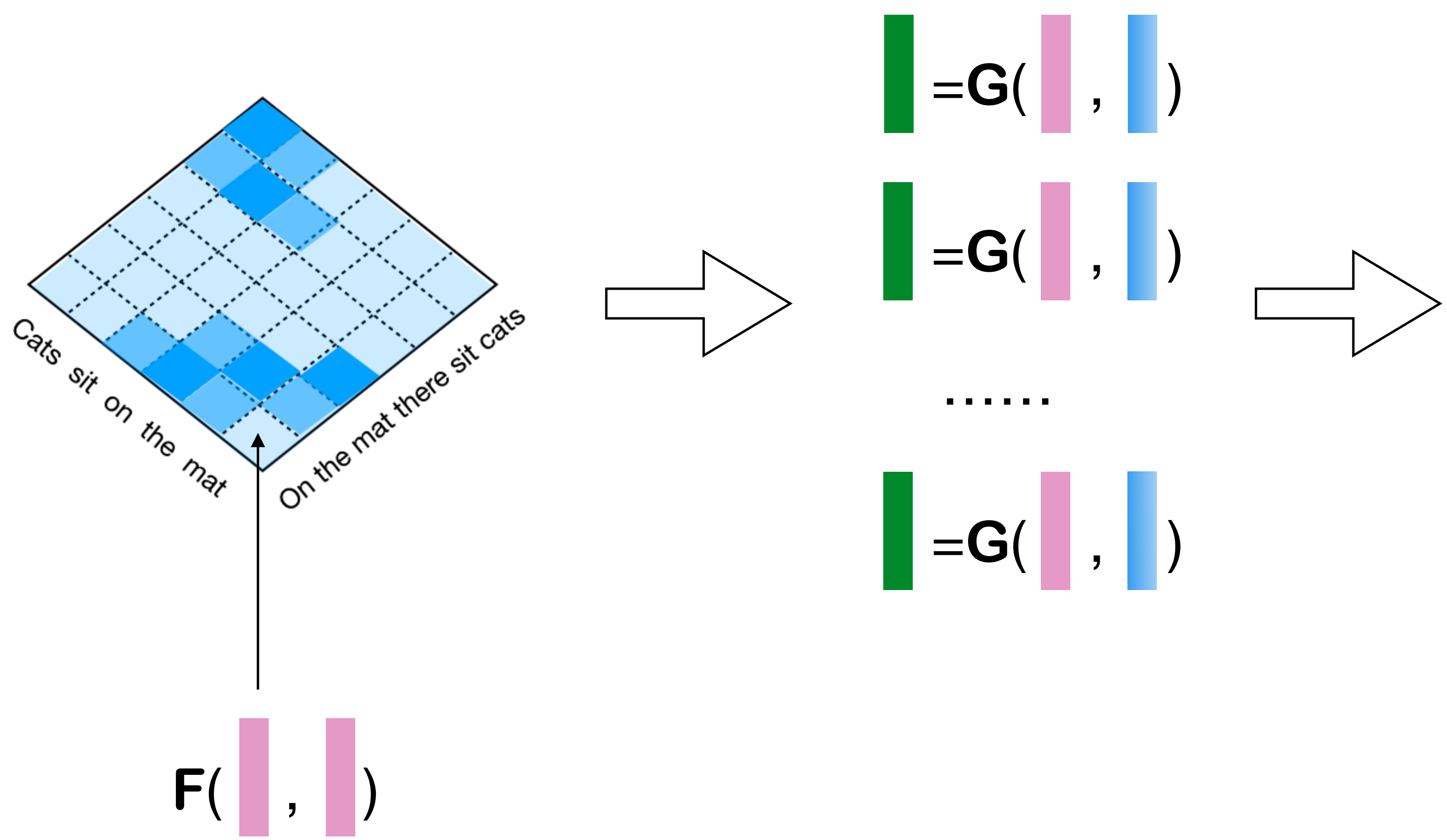
Cats sit on the mat
On the mat there sit cats
F(,)

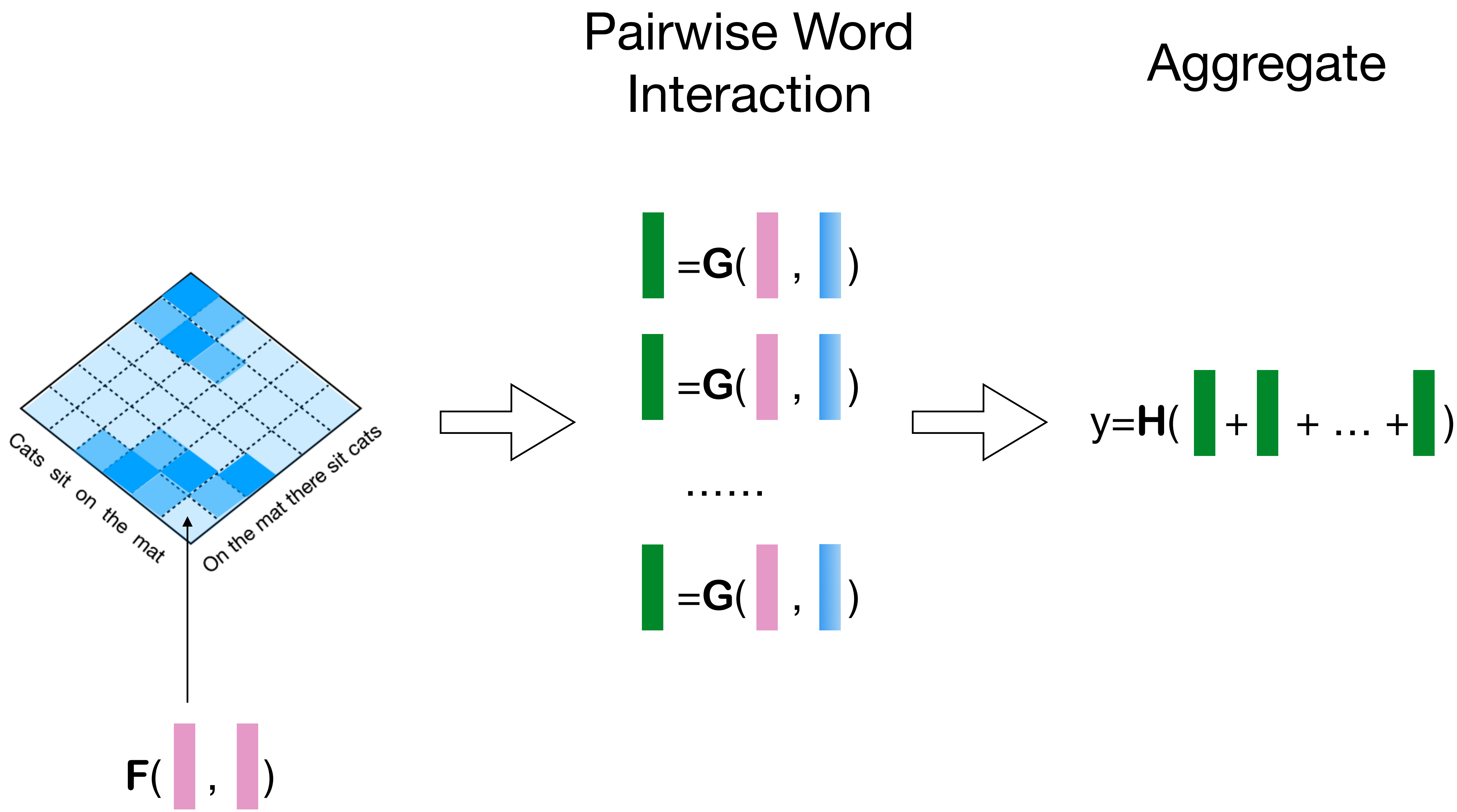

Cats sit on the mat
On the mat there sit cats
F(,)

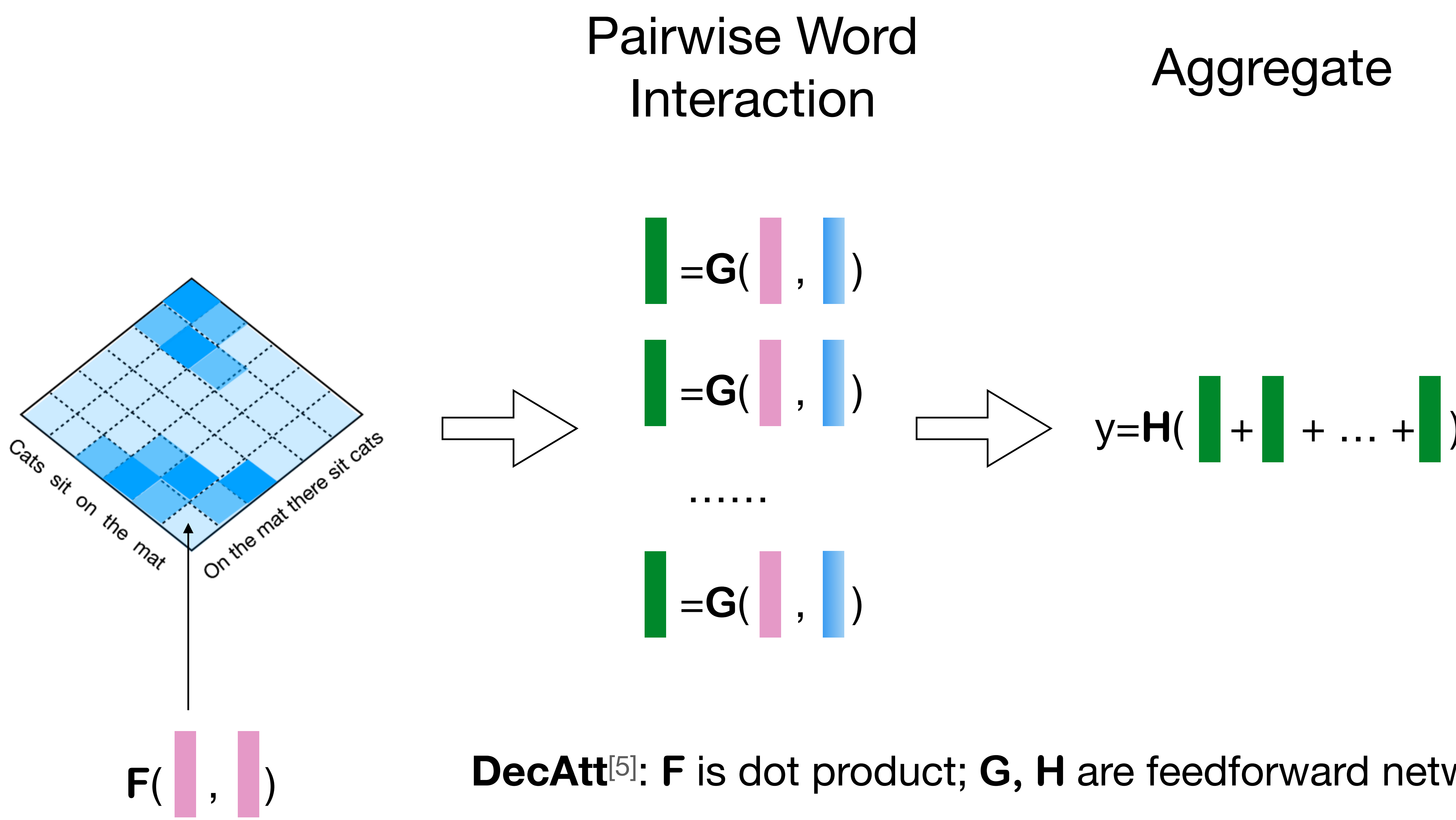
Pairwise Word Interaction



Pairwise Word Interaction




Pairwise Word
Interaction
Aggregate
Cats sit on the mat
On the mat there sit cats
=G(,)
=G(,)
......
=G(,)
y=H(+ + ... +)
F(,)

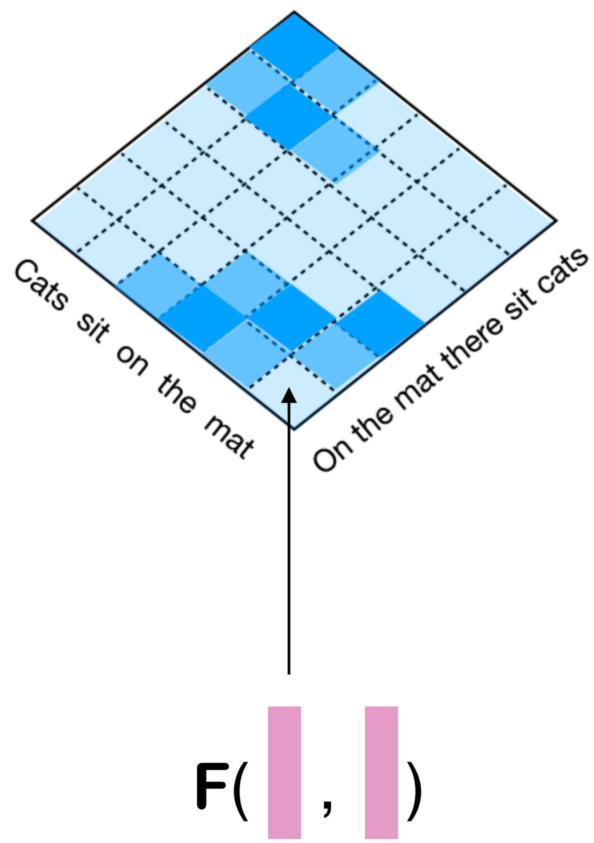


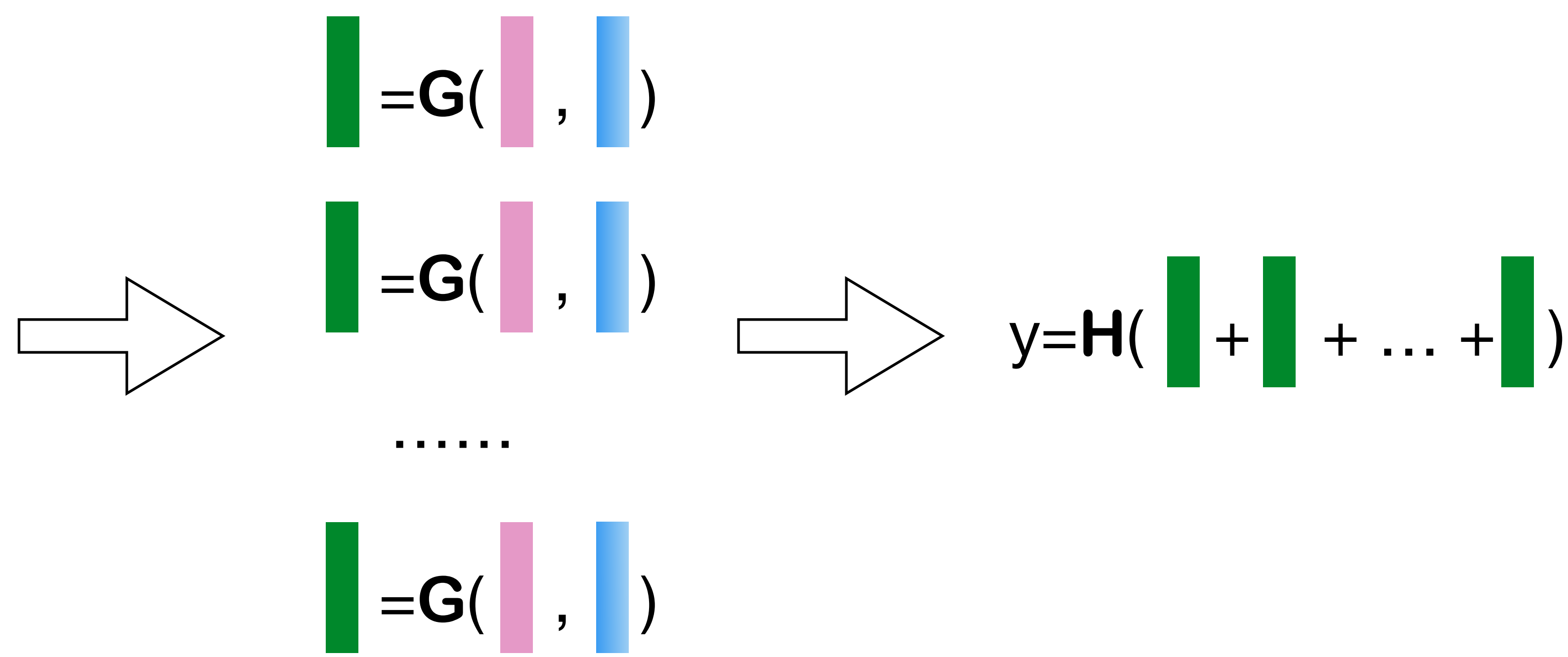
DecAtt[5]: **F** is dot product; **G, H** are feedforward networks.

[5] Ankur Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkorei. A decomposable ¨ attention model for natural language inference. (EMNLP 2016)

Pairwise Word Interaction

Aggregate





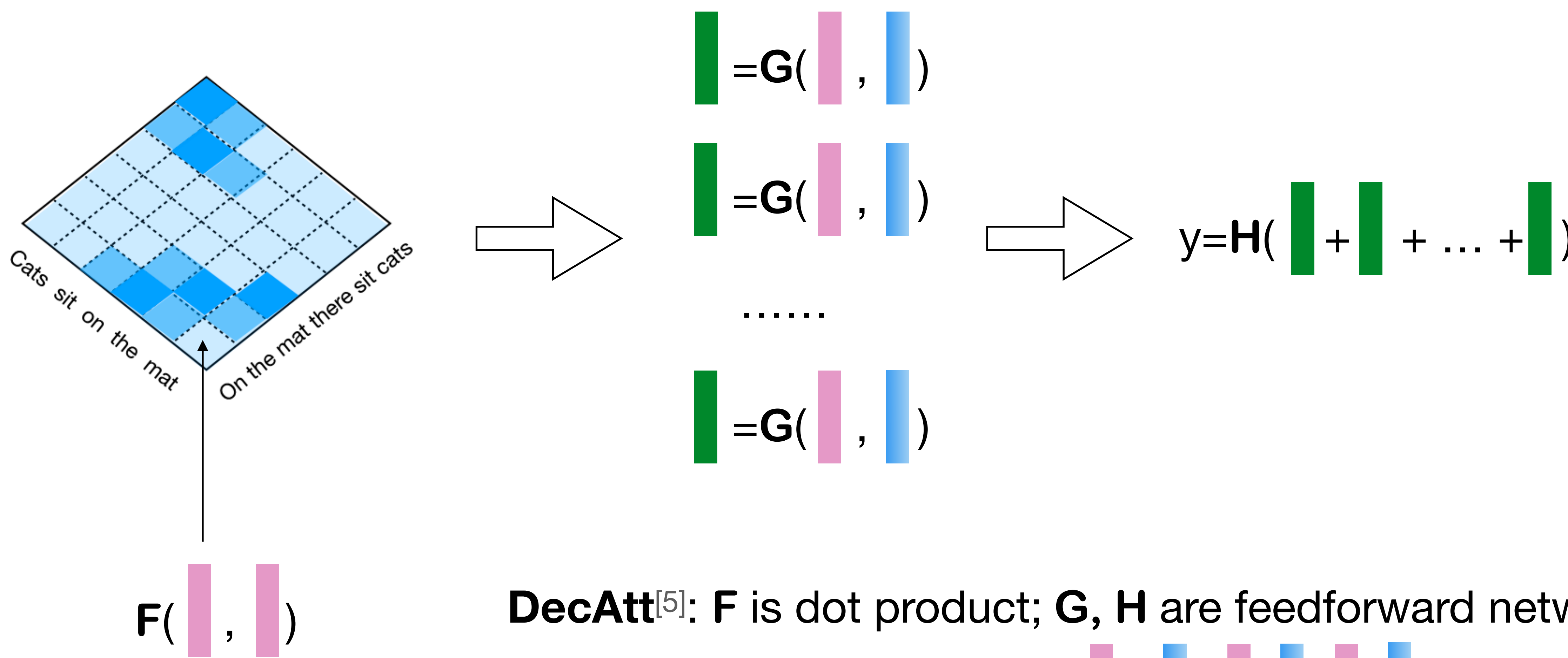
DecAtt[5]: **F** is dot product; **G, H** are feedforward networks.

ESIM[6]: more features in **G**(, , −, ⊙), and **G** is replaced with Bi-LSTM/Tree-LSTM.

[5] Ankur Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkorei. A decomposable ¨ attention model for natural language inference. (EMNLP 2016)
[6] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. (ACL 2017)

Pairwise Word Interaction

Aggregate



DecAtt[5]: **F** is dot product; **G, H** are feedforward networks.

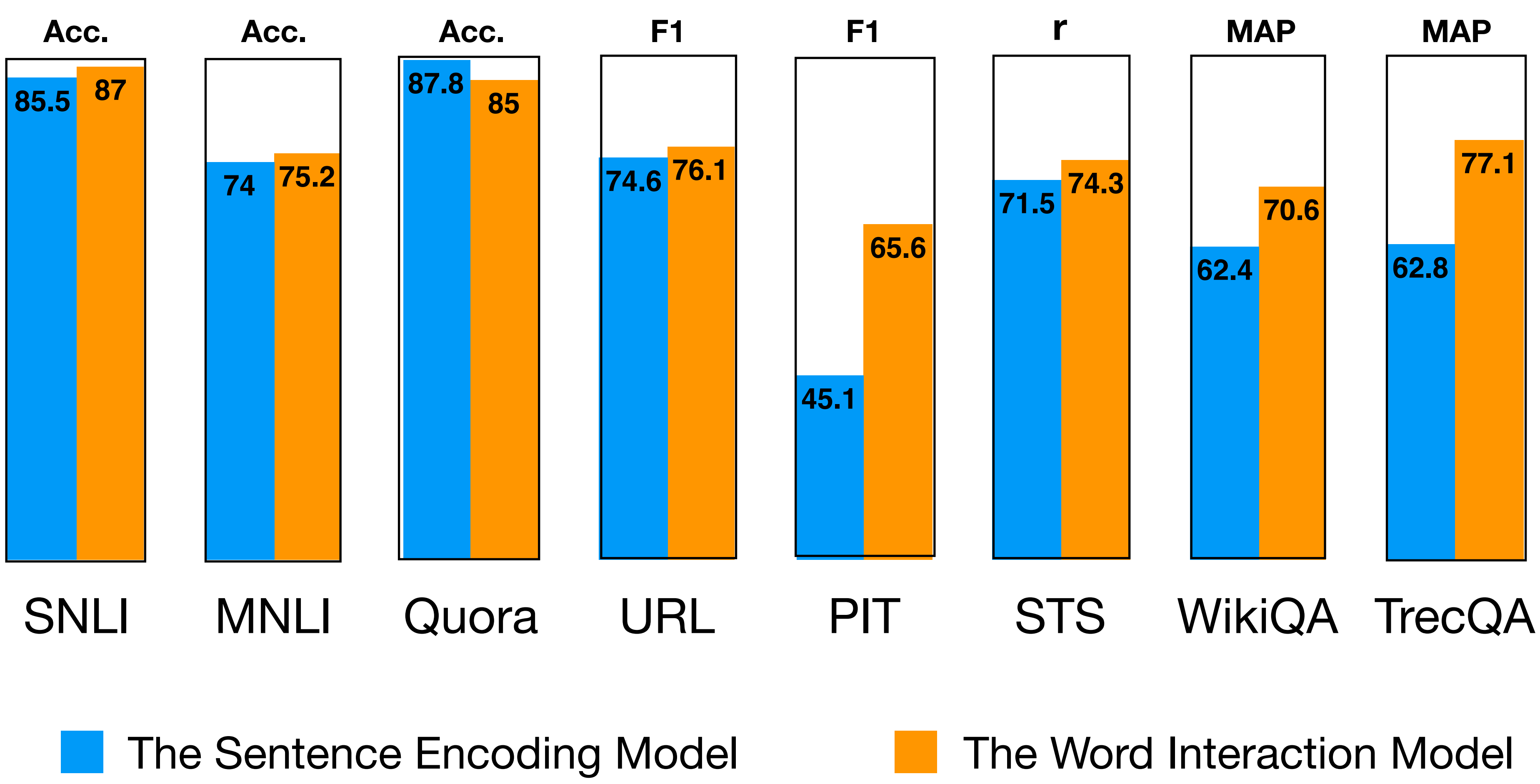
ESIM[6]: more features in **G**(, , −, ⊙), and **G** is replaced with Bi-LSTM/Tree-LSTM.

PWIM[7]: **F** uses cosine, L2 and dot product; **G** (,) is "hard" attention; **H** is deep CNN.

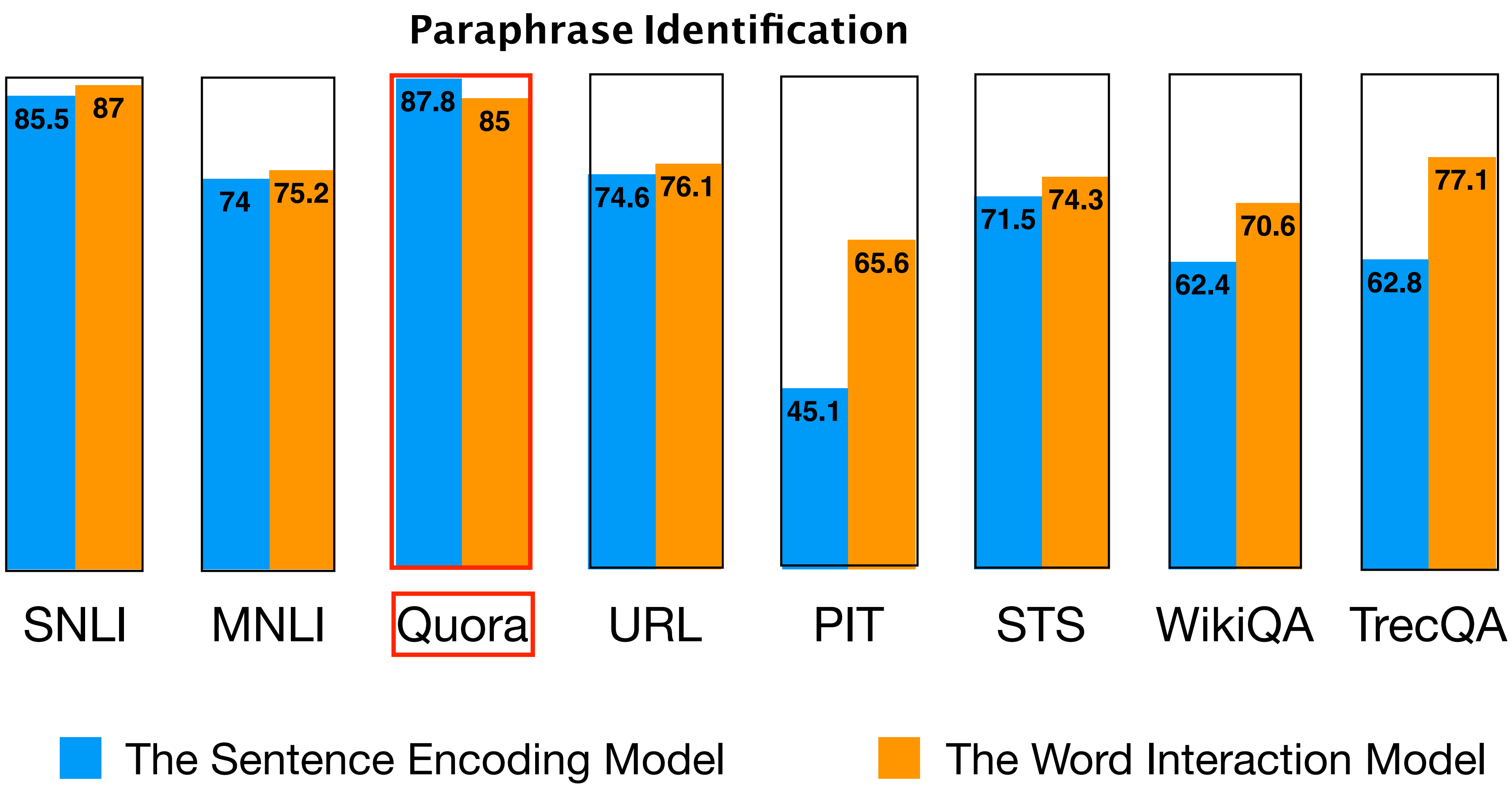
[5] Ankur Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkorei. A decomposable ¨ attention model for natural language inference. (EMNLP 2016)
[6] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. (ACL 2017)
[7] Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. (NAACL 2016)

What Type of Model performs better ?

What Type of Model performs better ?



What Type of Model performs better?



- Word Interaction-based Models perform much better (except Quora).

Why is Quora an exception ?

Why is Quora an exception ?

paraphrase

How can I be a great public speaker?

How can I learn to be a great public speaker?

Why is Quora an exception ?

paraphrase

How can I be a great public speaker?

How can I learn to be a great public speaker?

Longest Common Sequence / Sentence Length (%)

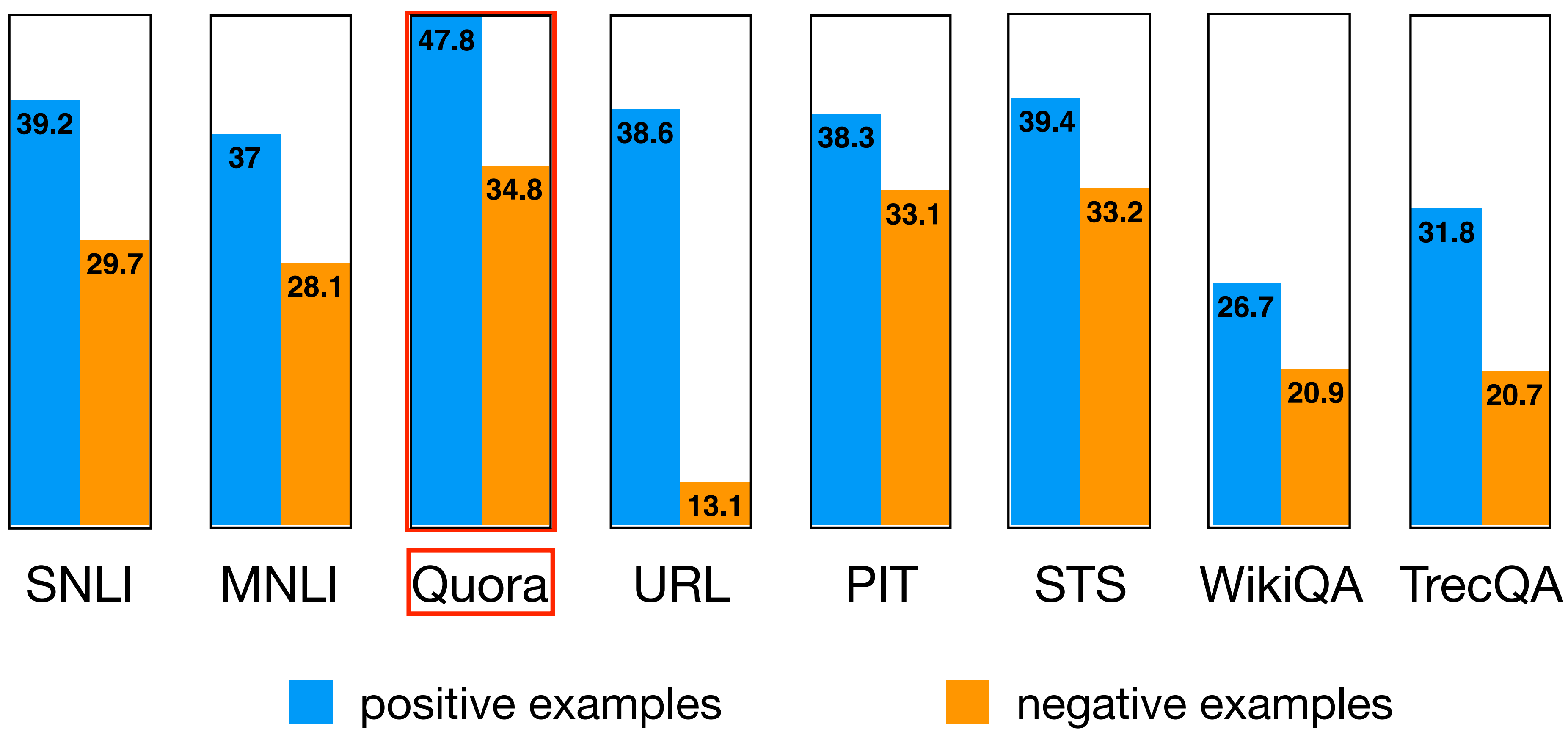
Why is Quora an exception ?

paraphrase

How can I <u>be a great public speaker?</u>

How can I learn to <u>be a great public speaker?</u>

Longest Common Sequence / Sentence Length (%)



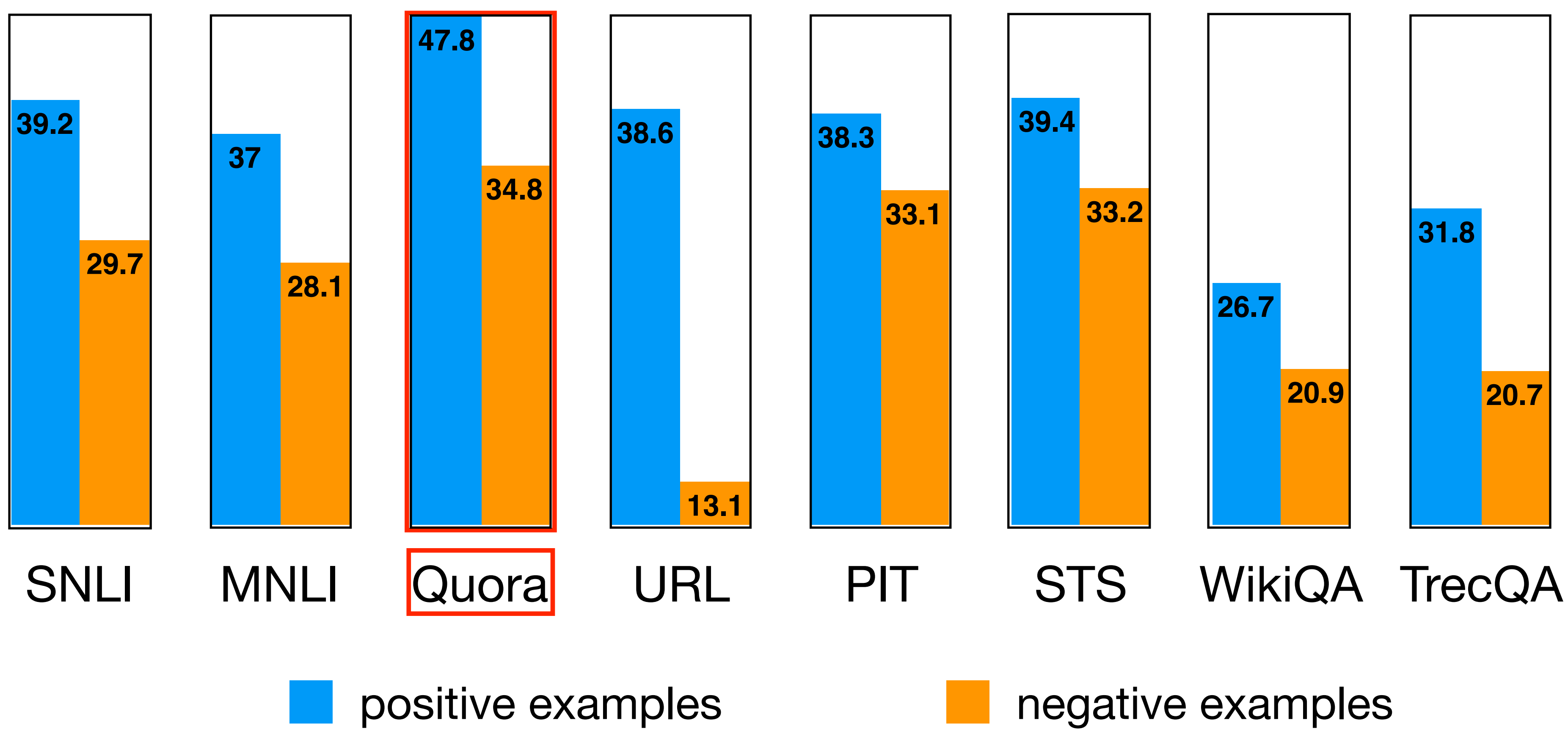
Why is Quora an exception ?

paraphrase

How can I *<u>be a great public speaker?</u>*

How can I learn to *<u>be a great public speaker?</u>*

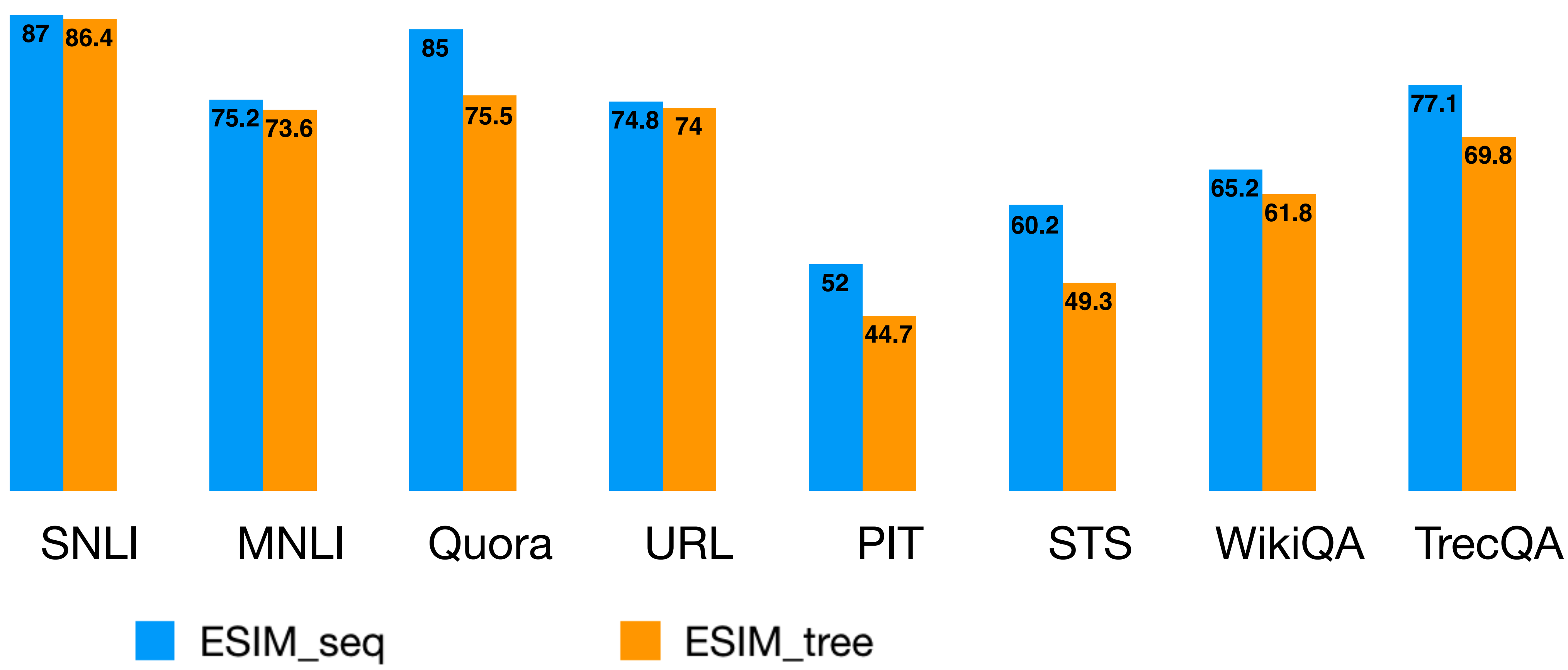
Longest Common Sequence / Sentence Length (%)



- Longer common sequences results in similar (RNN-based) sentence embeddings.

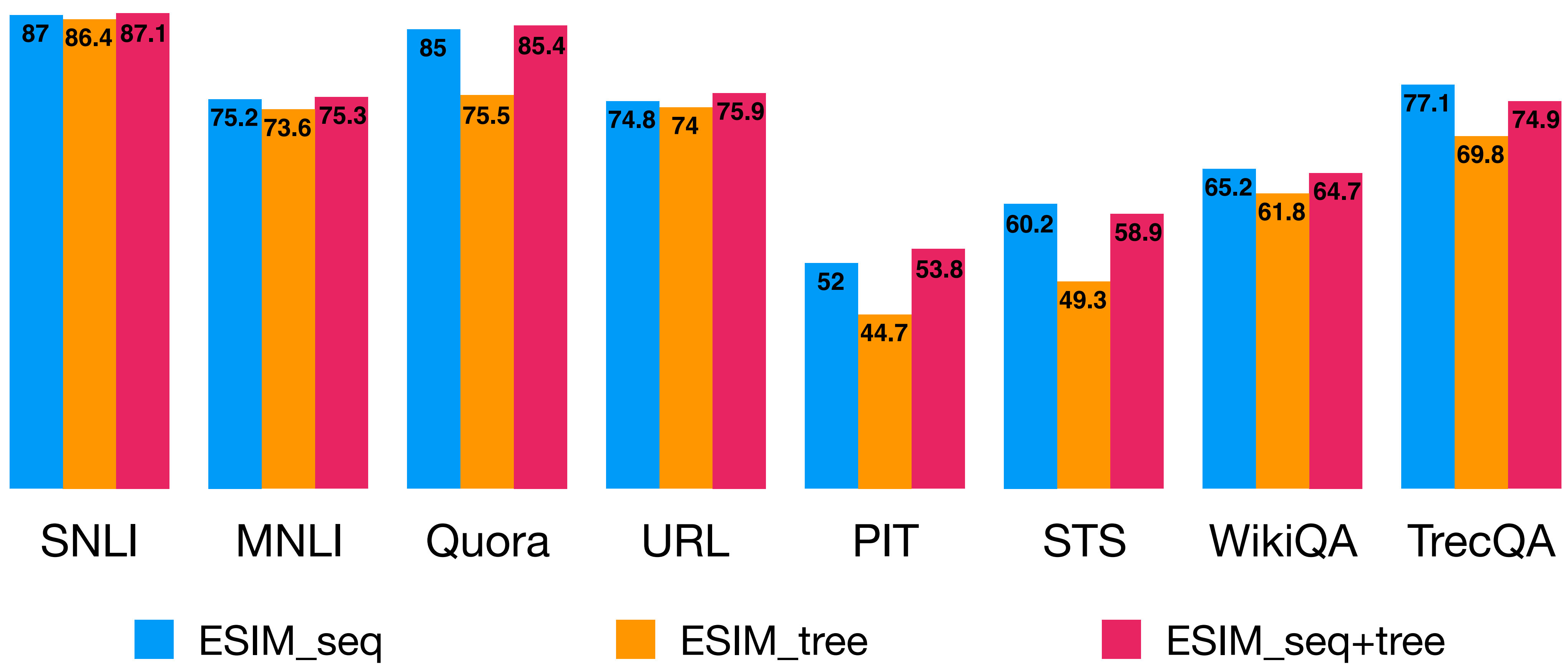
Bi-LSTM or Tree-LSTM?

Bi-LSTM or Tree-LSTM?



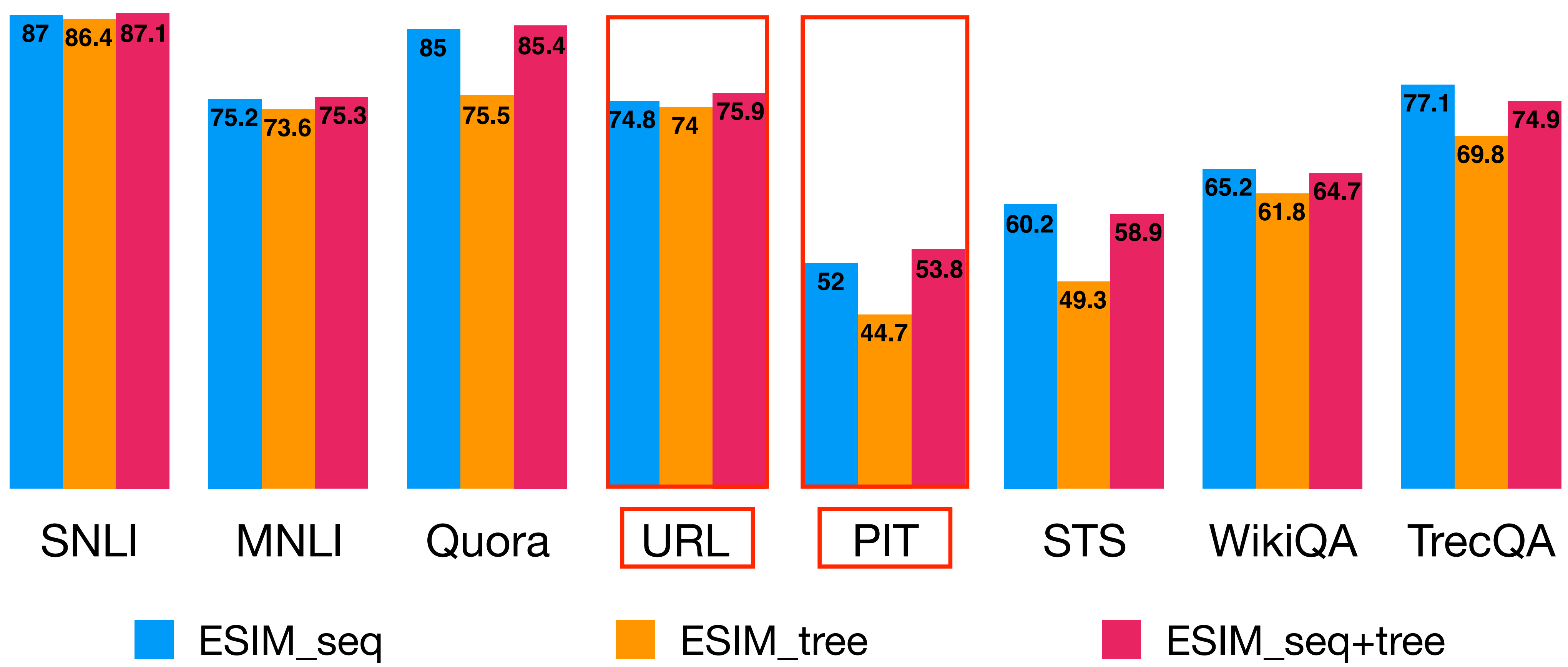
- ESIM_seq (Bi-LSTM) performs better than ESIM_tree (Tree-LSTM) on every dataset.

Bi-LSTM or Tree-LSTM?



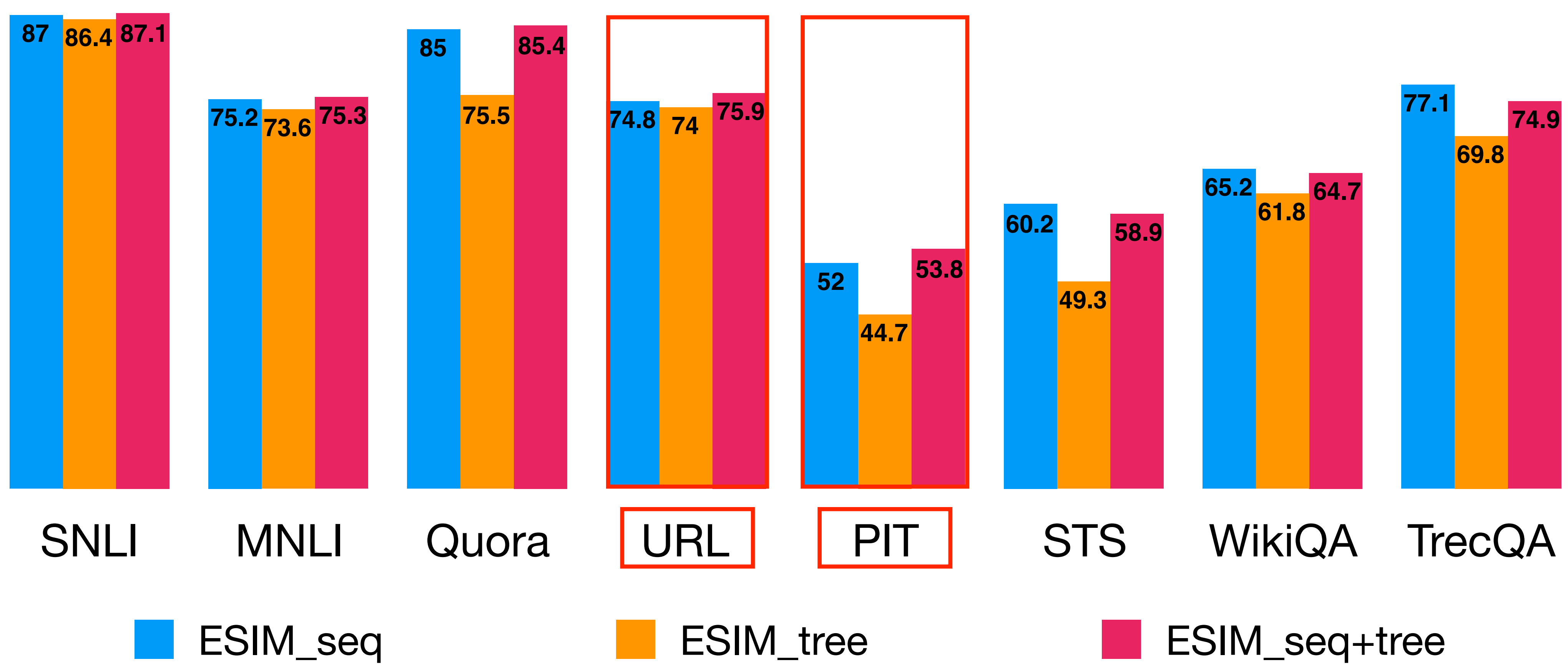
- ESIM_seq (Bi-LSTM) performs better than ESIM_tree (Tree-LSTM) on every dataset.

Bi-LSTM or Tree-LSTM?

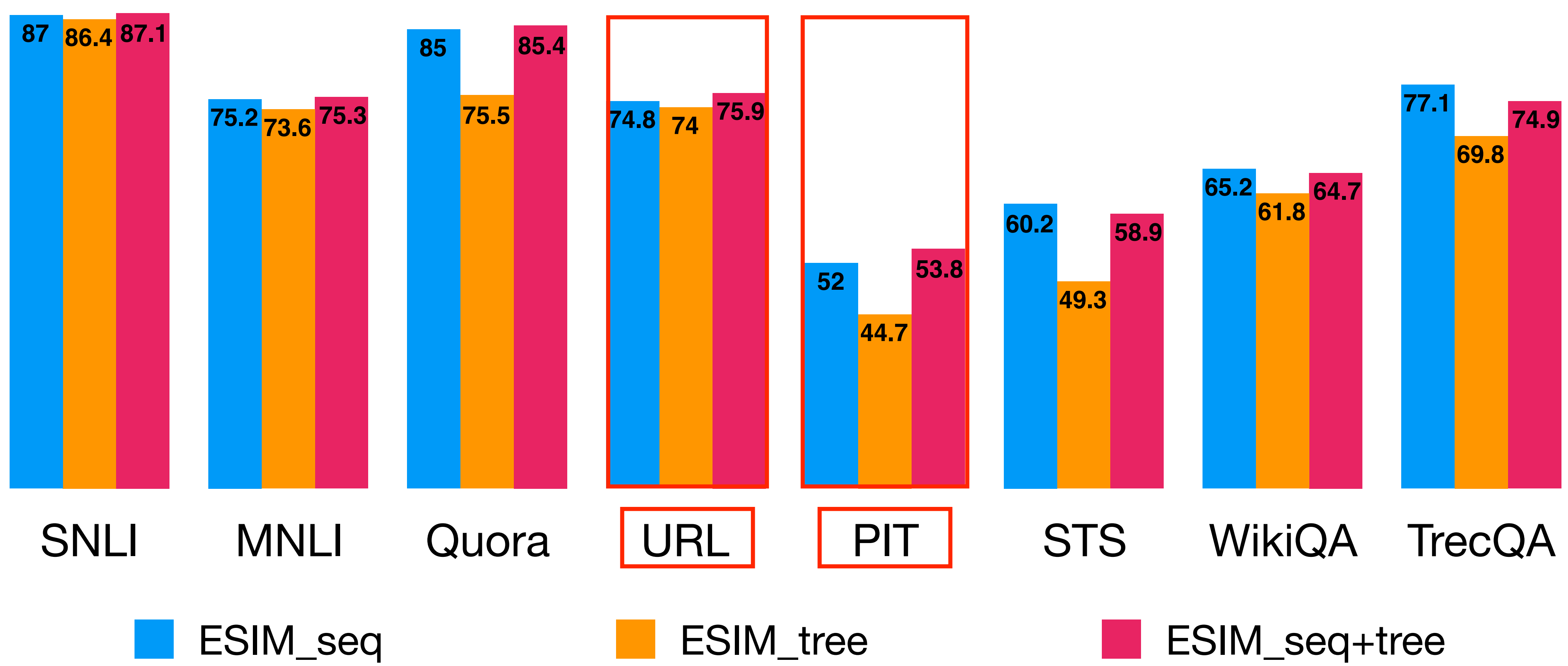


- ESIM_seq (Bi-LSTM) performs better than ESIM_tree (Tree-LSTM) on every dataset.
- Adding Tree_LSTM (ESIM_seq+tree) helps on Twitter data (URL and PIT).

Why Tree-LSTM helps with Twitter data ?



Why Tree-LSTM helps with Twitter data ?



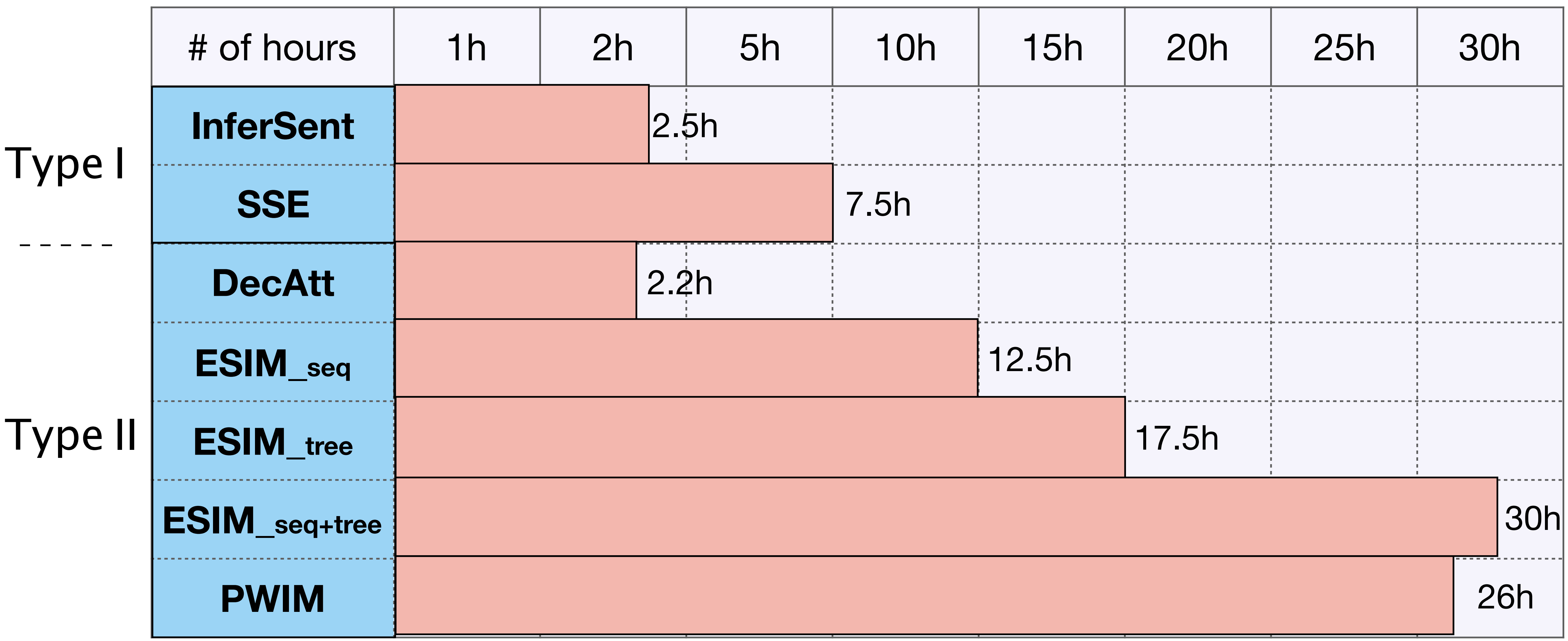
Paraphrase

<u>ever wondered,</u> why your recorded #voice sounds weird to you?

why do our recorded voices sound so weird to us?

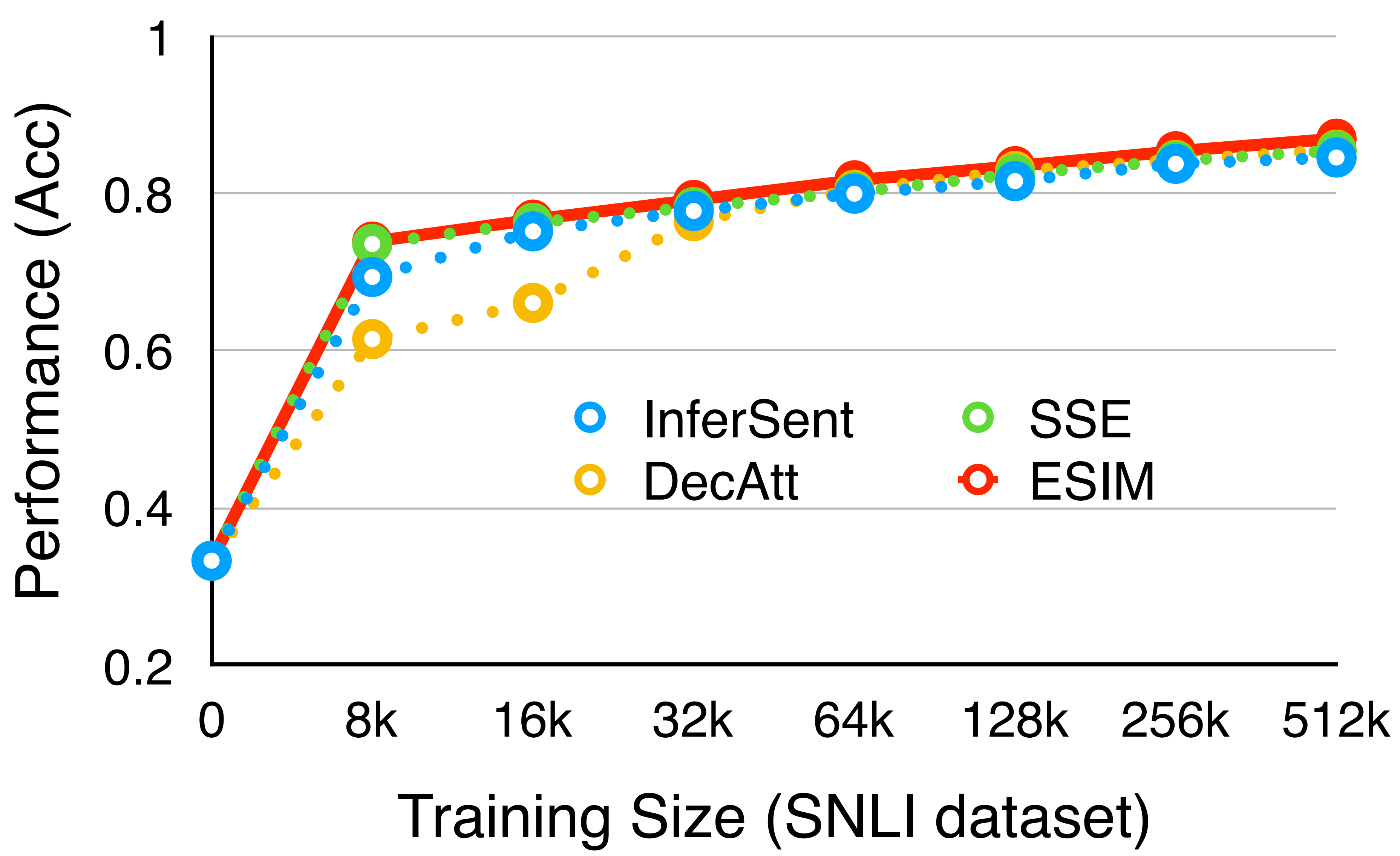
- Disruptive context can be put into less important position in Tree-LSTM.

Training Time on SNLI

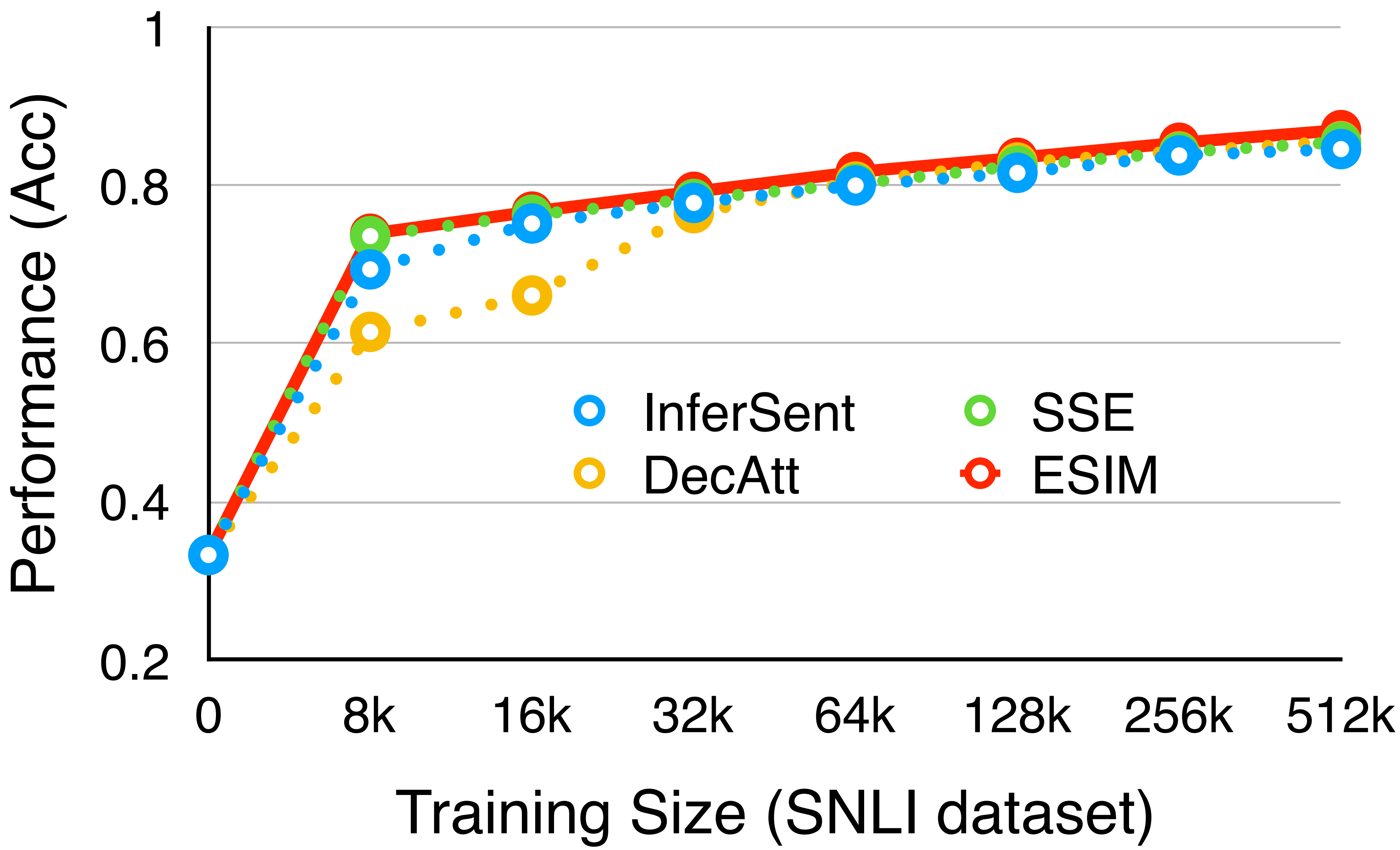


- Training time comparison across different models on SNLI dataset (550k sent pairs).

Do we need more data?



Do we need more data?



- The learning curves are still increasing. More data can help!

We also need more **natural** data!

We also need more **natural** data!

- Natural data — two sentences are written independently and have no label bias.

We also need more **natural** data!

- Natural data — two sentences are written independently and have no label bias.

  SNLI is large but contains data annotation artifacts. [8]

[8] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data (NAACL 2018).

We also need more **natural** data!

- Natural data — two sentences are written independently and have no label bias.

  SNLI is large but contains data annotation artifacts. [8]

  Twitter data contains natural paraphrases in large quantity, though can be noisy. [9]

[8] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data (NAACL 2018).
[9] Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, and Yangfeng Ji. Extracting Lexically Divergent Paraphrases from Twitter (TACL 2014).

We also need more **natural** data!

- Natural data — two sentences are written independently and have no label bias.

SNLI is large but contains data annotation artifacts. [8]

Twitter data contains natural paraphrases in large quantity, though can be noisy. [9]

paraphrase ← *Ezekiel Ansah is wearing real3D glasses with the lenses punched out*

Ezekiel Ansah wearing 3D glasses wout the lens

non-paraphrase ← *I wore the 3D glasses wout lenses before Ezekiel Ansah*

[8] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data (NAACL 2018).
[9] Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, and Yangfeng Ji. Extracting Lexically Divergent Paraphrases from Twitter (TACL 2014).

We also need more **natural** data!

- Natural data — two sentences are written independently and have no label bias.

  SNLI is large but contains data annotation artifacts. [8]

  Twitter data contains natural paraphrases in large quantity, though can be noisy. [9]

  more for future work!

paraphrase

non-paraphrase

Ezekiel Ansah is wearing real3D glasses with the lenses punched out

Ezekiel Ansah wearing 3D glasses wout the lens

I wore the 3D glasses wout lenses before Ezekiel Ansah

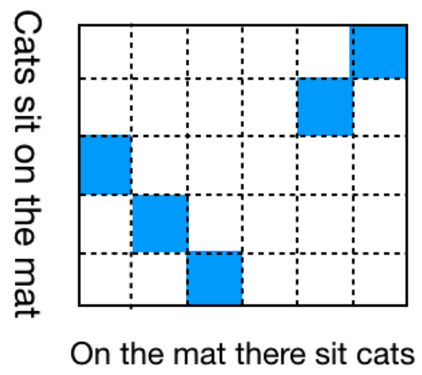
[8] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data (NAACL 2018).
[9] Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, and Yangfeng Ji. Extracting Lexically Divergent Paraphrases from Twitter (TACL 2014).

Takeaways

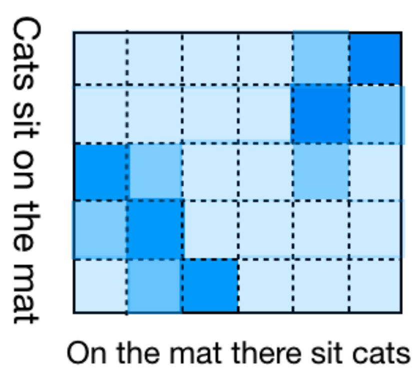
- Systematic comparison of **5** representative models on **8** datasets
- **Large**, **clean**, and **more natural** data is needed for studying semantics!
- Code is available: **https://github.com/lanwuwei/SPM_toolkit**

Ohio Supercomputer Center

Backup slides: word alignment



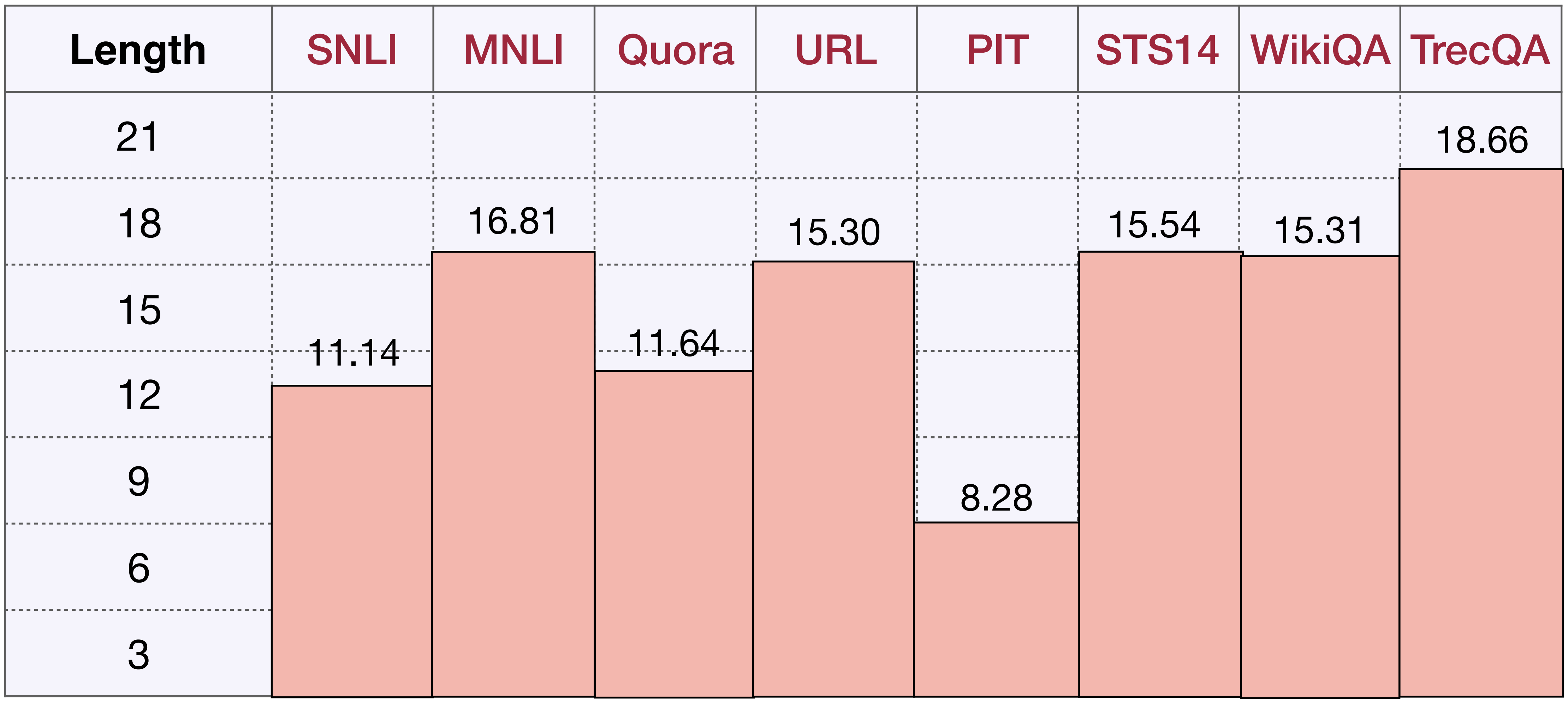
PWIM[8]**:** hard alignment.



DecAtt[9] **ESIM**[10]**:** soft alignment.

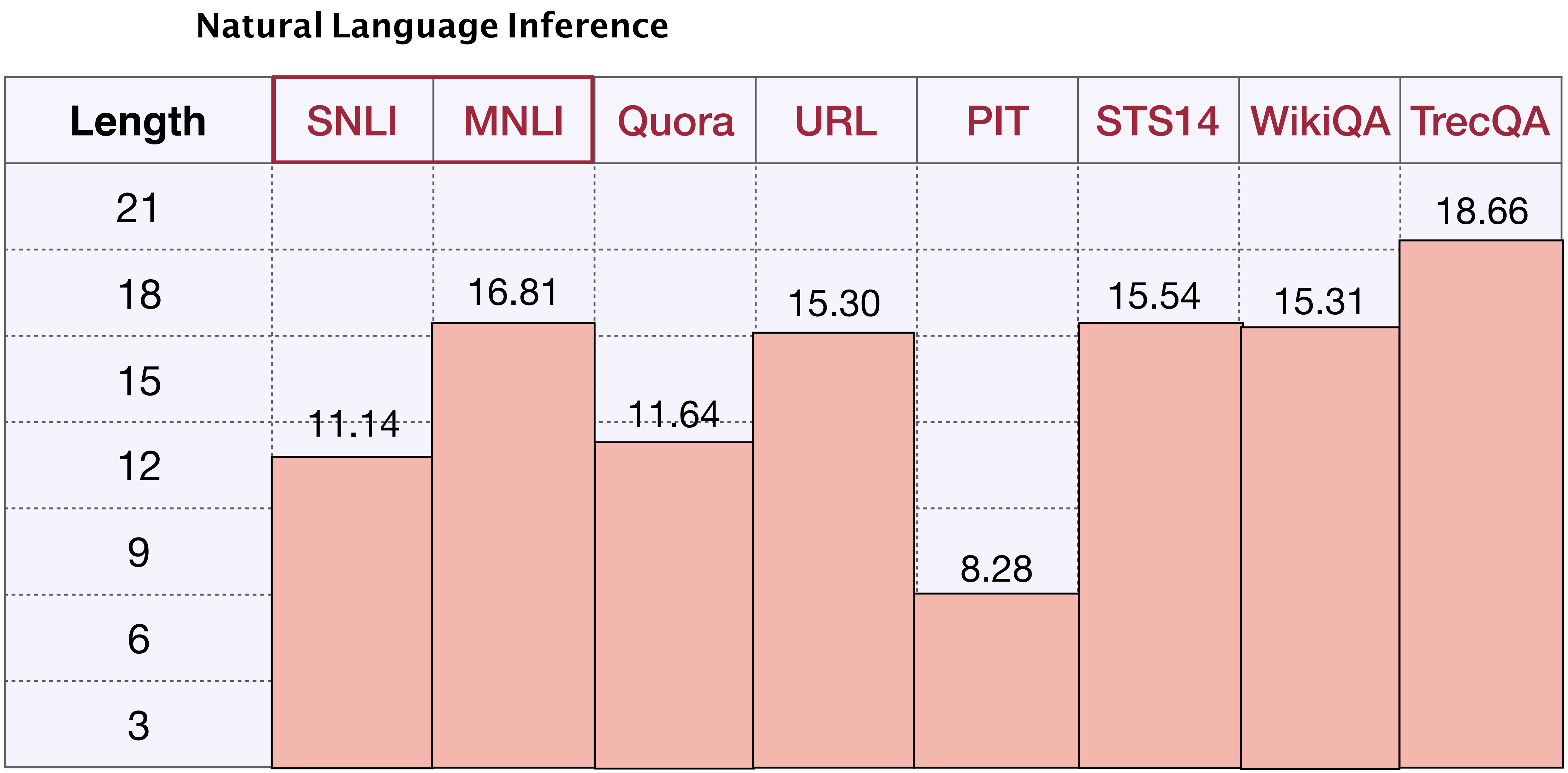
[8] Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. (NAACL 2016)
[9] Ankur Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkorei. A decomposable ¨ attention model for natural language inference. (EMNLP 2016)
[10] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. (ACL 2017)

Backup slides: sentence length Statistics



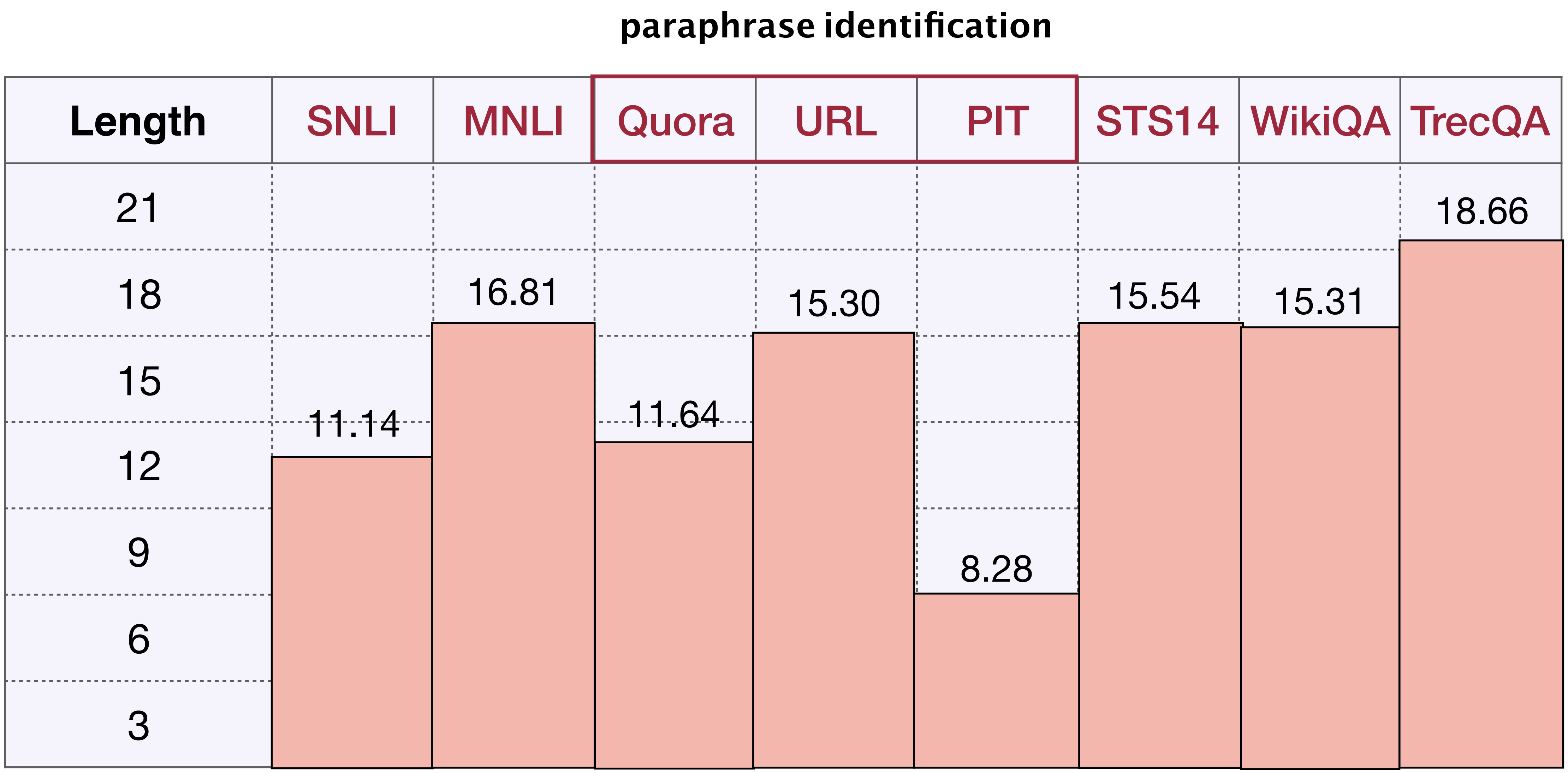
- Sentence length comparison in different datasets (training set).

Backup slides: sentence length Statistics



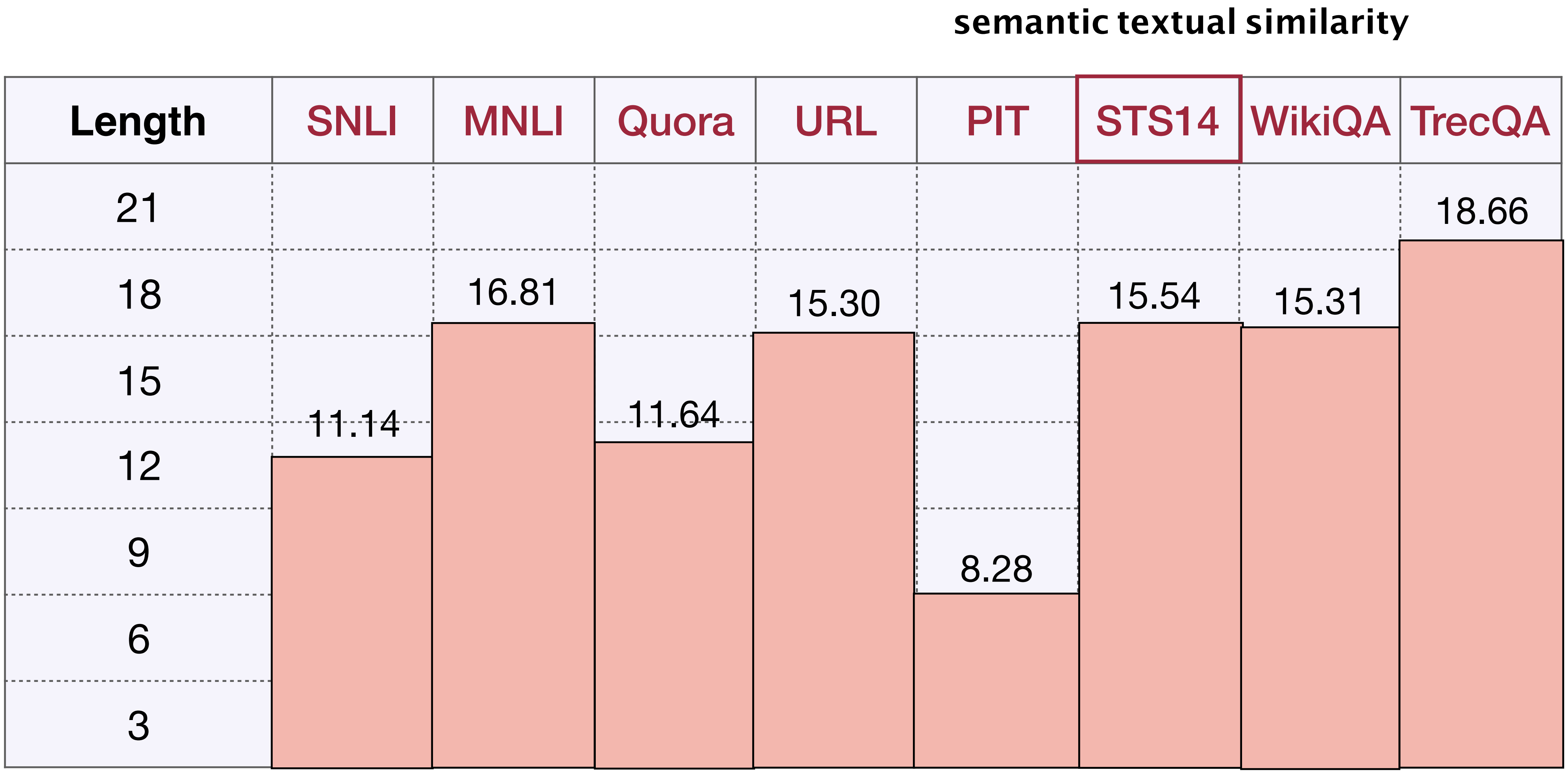
- Sentence length comparison in different datasets (training set).

Backup slides: sentence length Statistics



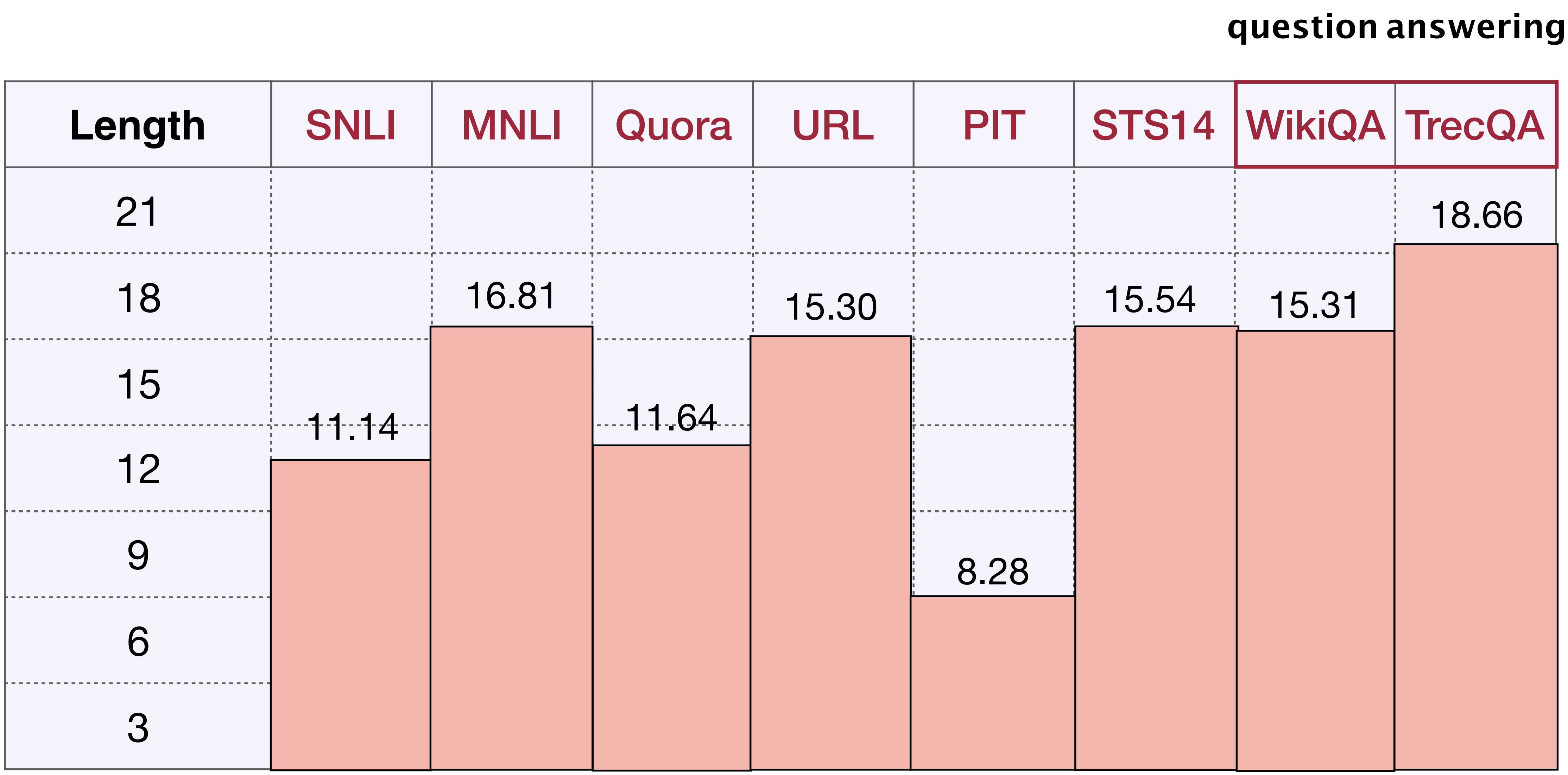
- Sentence length comparison in different datasets (training set).

Backup slides: sentence length Statistics



- Sentence length comparison in different datasets (training set).

Backup slides: sentence length Statistics



- Sentence length comparison in different datasets (training set).

Backup slides: experiment settings

Word Embedding: Glove Twitter 200d vectors for PIT and URL; Glove Common Crawl (840B tokens) 300d vectors for other datasets.

Hyper-parameters: the same settings as in the original papers/ implementations. Check appendix in arXiv paper for more details.

Fine tuning: No. Because we want to test their generalization ability, fine tuning can make models overfit on specific datasets.