Why is sentence similarity benchmark not predictive of application-oriented task performance?

Kaori Abe[1], Sho Yokoi[1,2], Tomoyuki Kajiwara[3], and Kentaro Inui[1,2]

1. Tohoku University 2. RIKEN 3. Ehime University

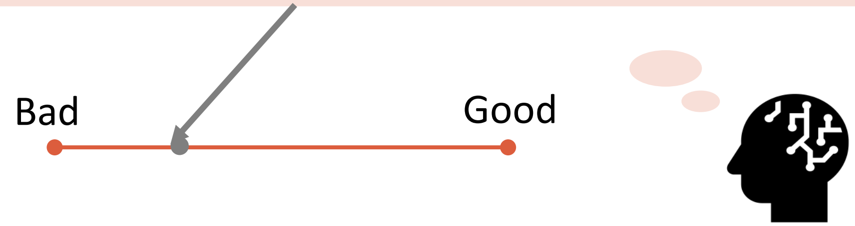
Predicting similarity is required in various NLP application tasks

- Many NLP application-oriented tasks needs **prediction similarity between two sentences**

Examples of NLP application-oriented tasks

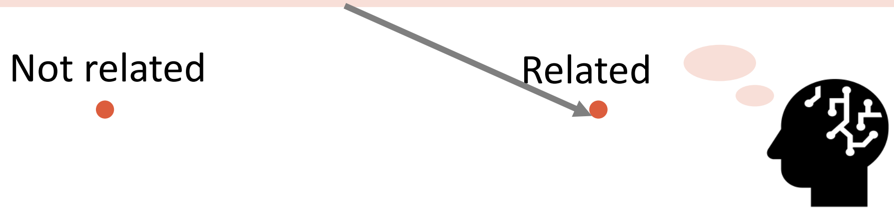
MT Metrics (MTM)

hyp	*Fresh fruit was replaced with cheaper dried fruit.*
ref	*Fresh fruit is cheap dried fruit instead.*



Passage Retrieval (PR)

query	*botulinum definition*
passage	*medical Definition of botulinum toxin : a very …*



Predicting similarity is required in various NLP application tasks

- **STS** is **a de-facto standard for prediction similarity**
  - Designed for applications [Aggire+’12; Cer+’17]
  - Used in many studies [Reimers&Gurevych+’19; Zhang+’20; Gao+’21; etc.]

Semantic Textual Si

s1	*A man is riding a mec…*
s2	*A man rode a mechanical bull.*

Different — Similar

A … n:

“better on STS → bet … ation-oriented tasks”

Examples of NLP application-oriented tasks

MT Metrics (MTM)

hyp	*Fresh fruit was replaced with cheaper dried fruit.*
ref	*Fresh fruit is cheap dried fruit instead.*

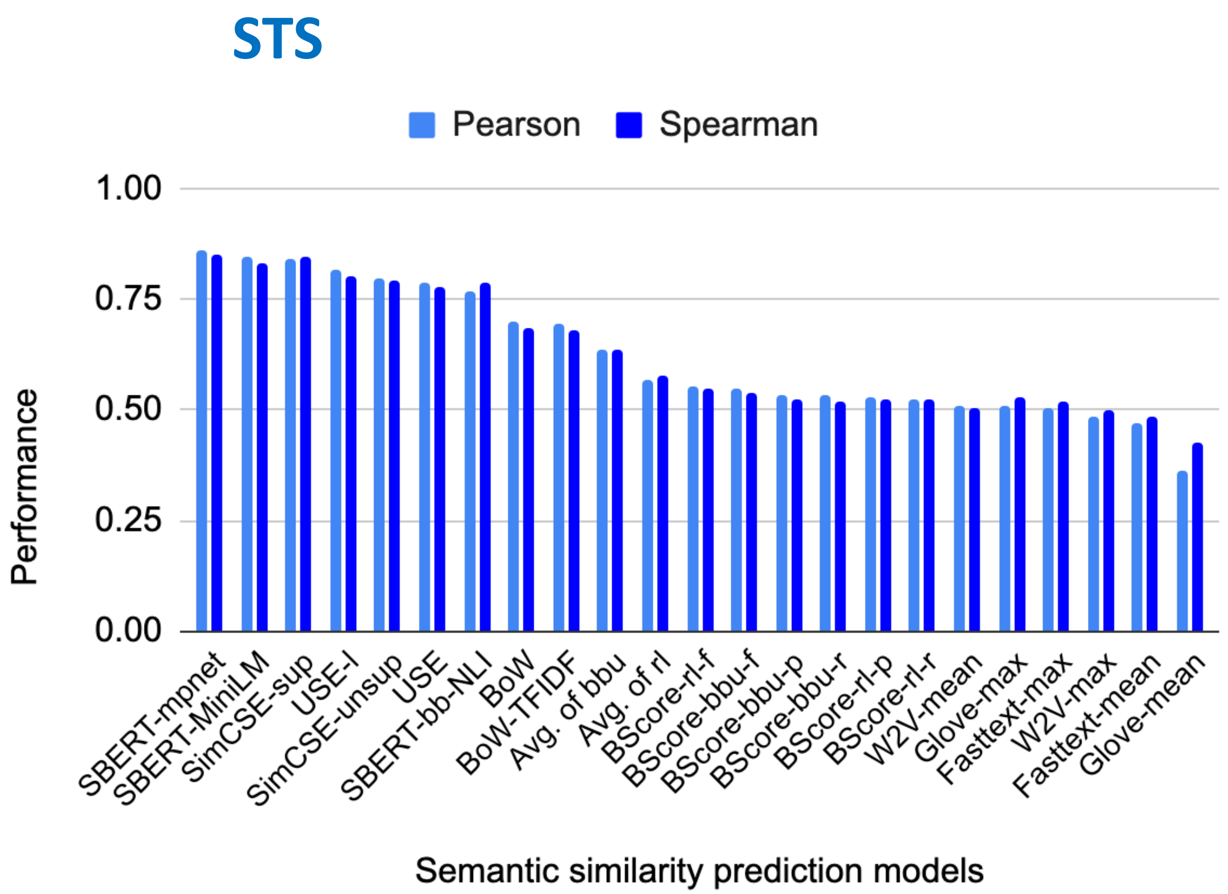
Good

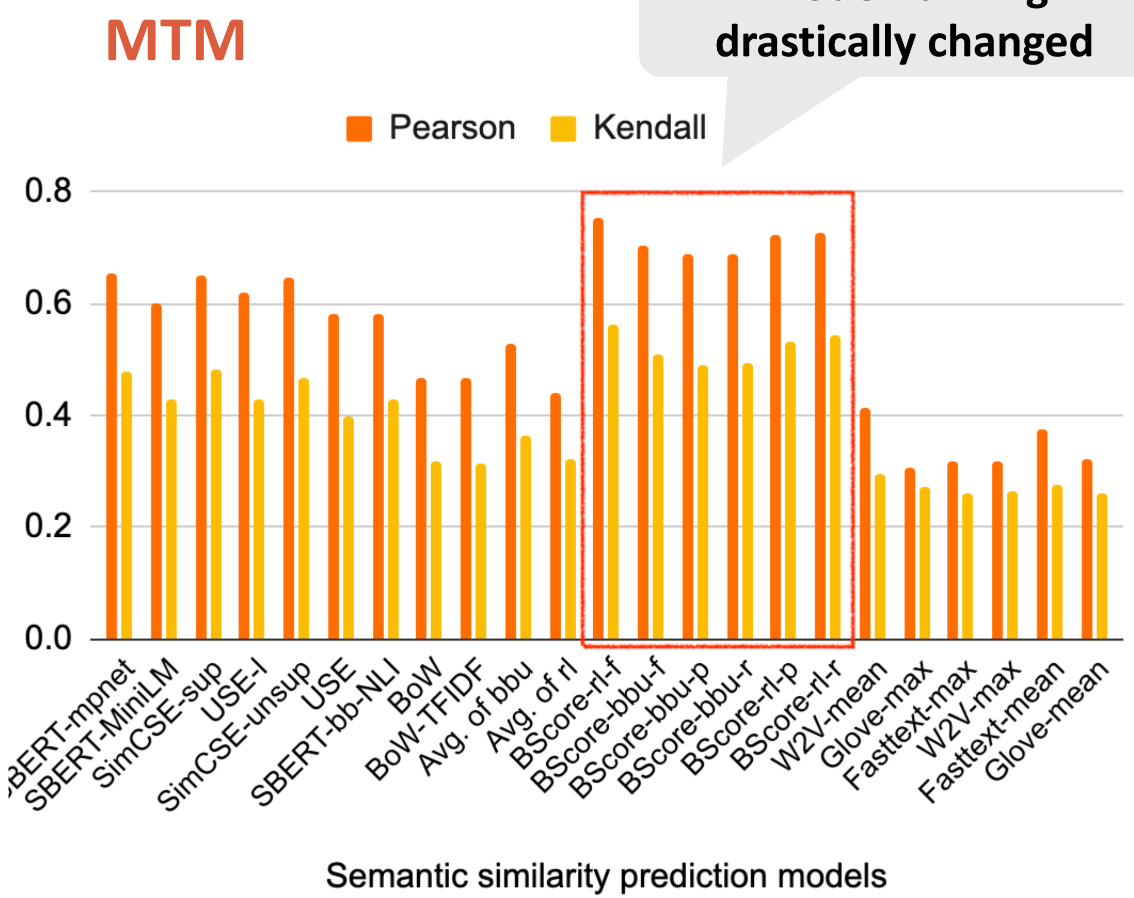
query	*botulinum definition*
passage	*medical Definition of botulinum toxin : a very …*

Not related — Related

Evaluation gap between STS and application-oriented tasks (e.g., MT Metrics)

Model ranking drastically changed



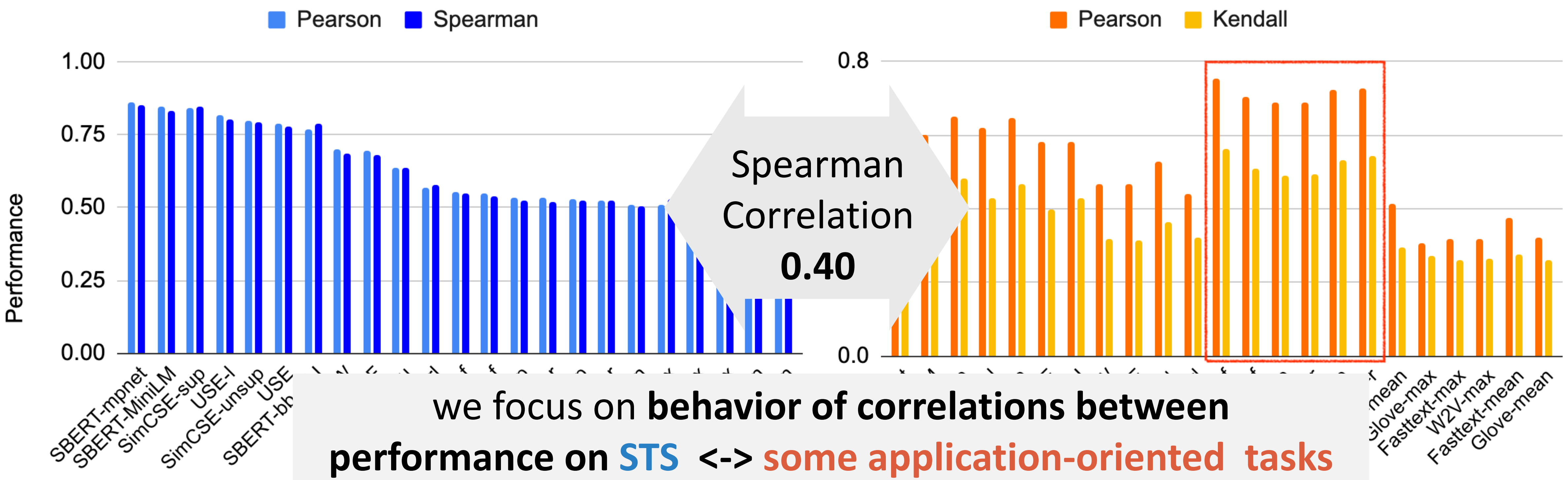


SBERT: [Reimers&Gurevych'+19], SimCSE: [Gao+'21], BERTScore: [Zhang+'19]

Evaluation gap between STS and application-oriented tasks (e.g., MT Metrics)

STS

MTM



we focus on **behavior of correlations between performance on STS <-> some application-oriented tasks**

Semantic similarity prediction models

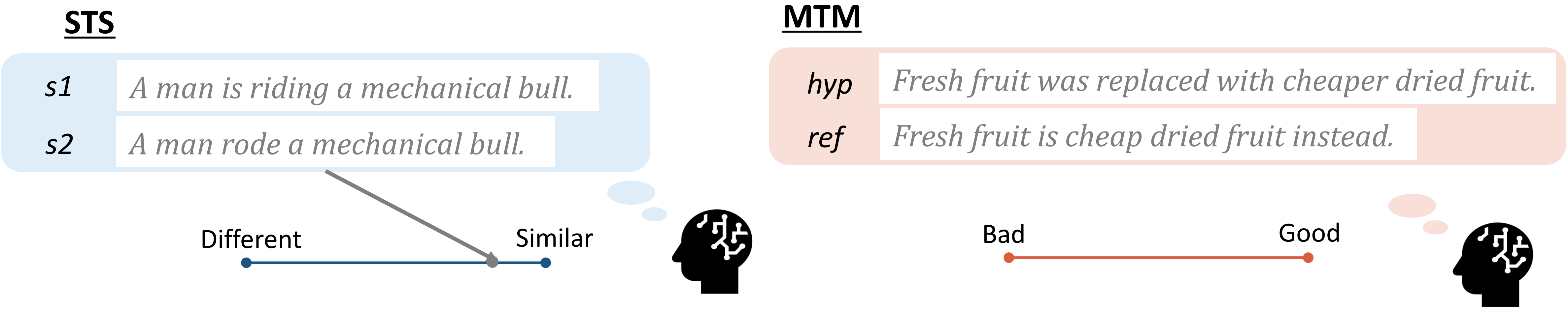
Semantic similarity prediction models

SBERT: [Reimers&Gurevych'+19], SimCSE: [Gao+'21], BERTScore: [Zhang+'19]

RQ. Gap of some factors in datasets → evaluation gap?

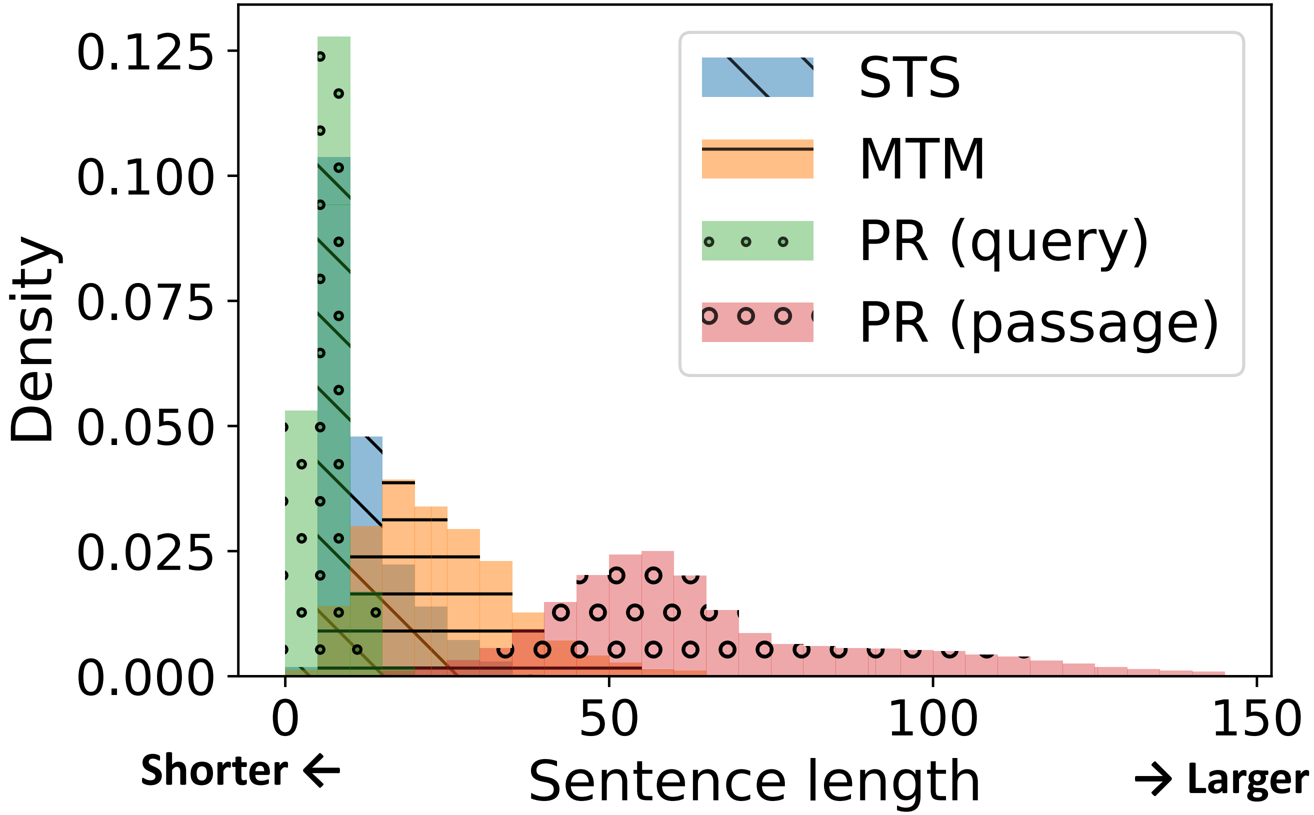
RQ: what causes evaluation gap between STS and application-oriented tasks?

- We expose **three factors**:
  1. Sentence length
  2. Vocabulary (domain)
  3. Granularity of golden similarity scores



Experiment 1: Sentence Length gap → Evaluation gap?

- STS's sentence length is **shorter** than application tasks' one
  - **STS** < **MTM**
  - **STS** << **PR (passage)**

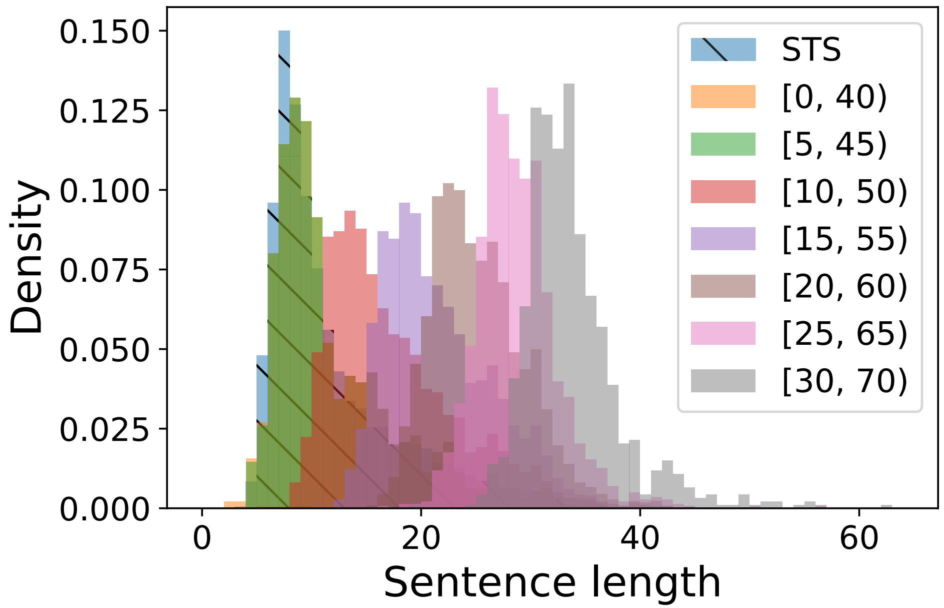


Experiment 1: Sentence Length gap → Evaluation gap?

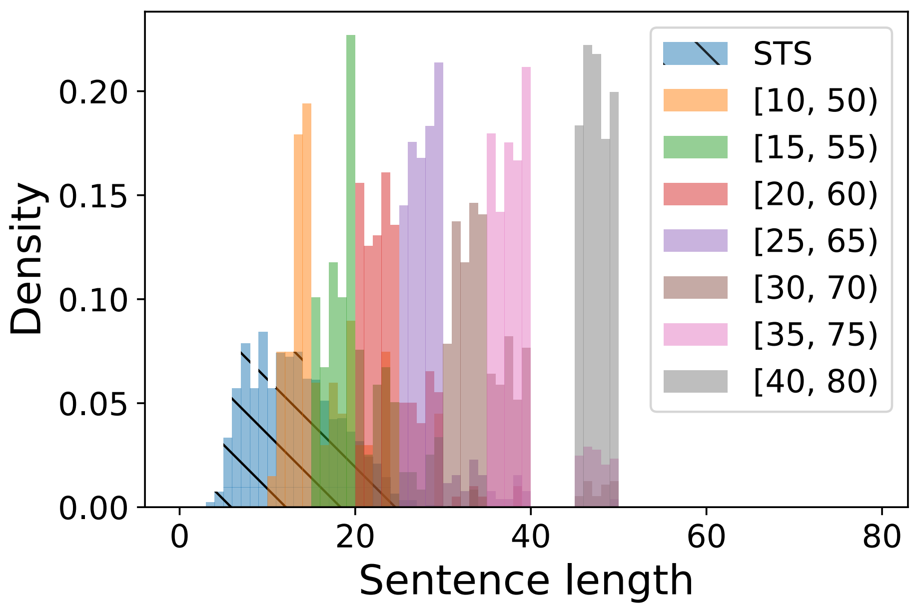
*Hypothesis: **Longer** sentence length subsets → **Large** Evaluation gap*

- We made **subsets** according to the STS sentence length distribution
  - **MTM-[x, y)** : examples of sentence length [x, y) in MTM dataset
  - **PR-[x, y)** : 〃 in PR dataset

MTM-[x, y)



PR-[x, y)



Experiment 1: Sentence Length gap → Evaluation gap?

*Hypothesis: **Longer** sentence length subsets → **Large** Evaluation gap (**Low** correlation)*

Shorter ↑ … ↓ Longer

STS <-> MTM

MTM-[0, 40)	0.351
MTM-[5, 45)	0.351
MTM-[10, 50)	0.390
MTM-[15, 55)	0.407
MTM-[20, 60)	0.317
MTM-[25, 65)	0.284
MTM-[30, 70)	0.313

STS <-> PR

PR-[10, 50)	0.613
PR-[15, 55)	0.792
PR-[20, 60)	0.752
PR-[25, 65)	0.777
PR-[30, 70)	0.779
PR-[35, 75)	0.770
PR-[40, 80)	0.814

<-> MTM : as hypothesized
<-> PR : The trend is not observed

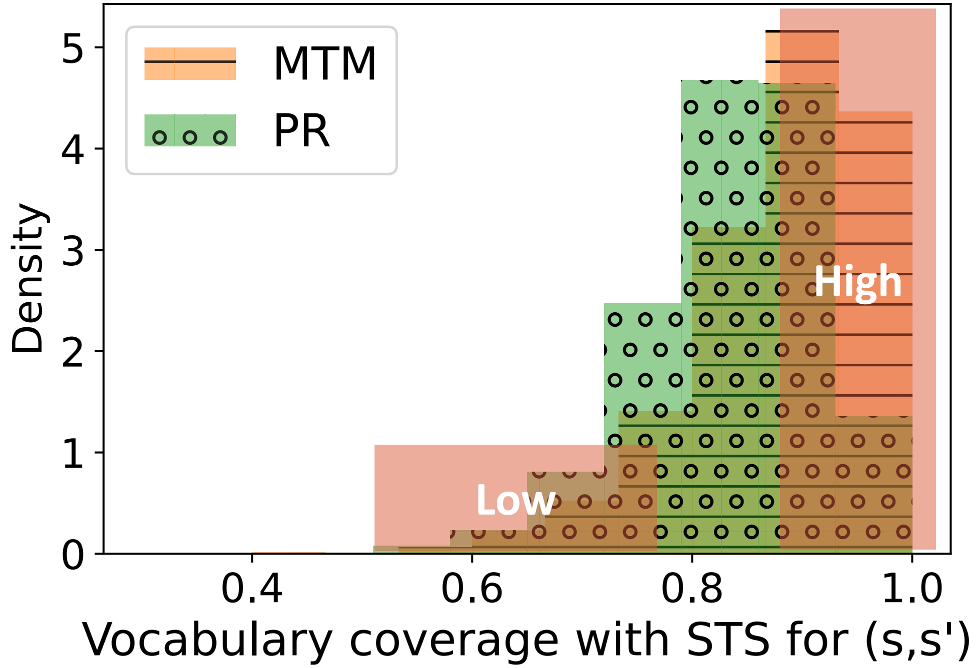
※ Spearman correlation between STS pearson corr. <-> {MTM: pearson corr., PR: MRR@10}
※ Darker color represents lower correlation

Experiment 2: Vocabulary gap → Evaluation gap?

- STS Vocabulary (V_{STS}) **could not cover** the application-oriented tasks' one

*Hypothesis: **Different** vocabulary dist. → **Large** Evaluation gap*

- We made **High/Low subsets** according to the vocabulary coverage
  - **High**: top 100 examples
  - **Low**: bottom 100 examples



(= ratio of $w \in V_{\mathrm{STS}}$)

Experiment 2: Vocabulary gap → Evaluation gap?

*Hypothesis: **Different** vocabulary dist. → **Large** Evaluation gap (**Low** correlation)*

	<-> STS domain	Vocab coverage High		〃 Low
MTM (News)	News (in-domain)	0.438	>	**0.373**
	Image caption	**0.046**	<	0.177
	Forum	0.779	>	**0.046**
PR (QA)	(all)	0.851	>	**0.673**

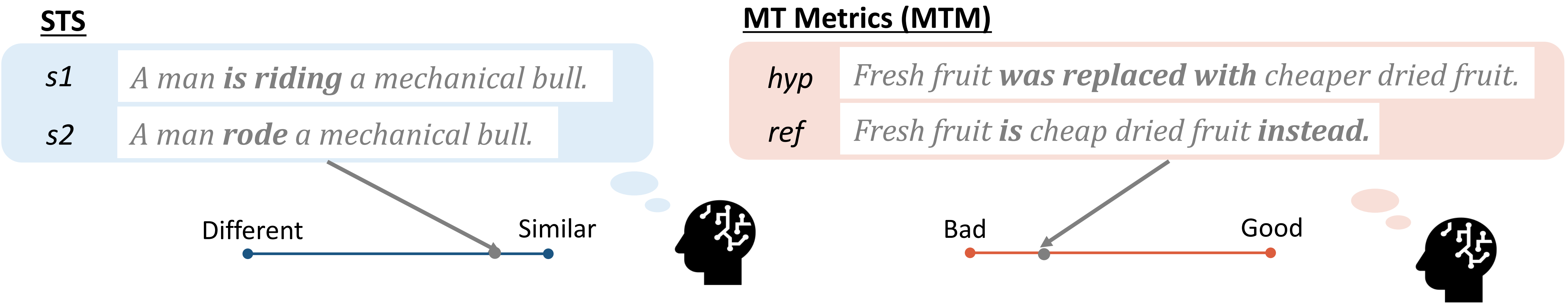
※ Spearman correlation between STS pearson corr. <-> {MTM: pearson corr., PR: MRR@10}

STS <-> both tasks (MTM, PR) : as hypothesized except for STS image caption domain

- In the image caption domain, the correlation values are lower for both the subsets

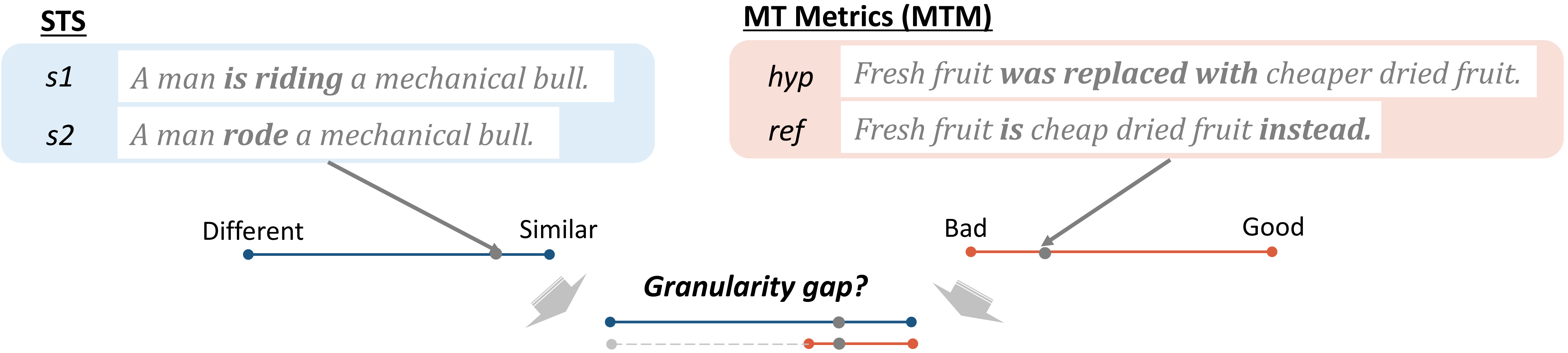
Experiment 3: Similarity granularity gap → Evaluation gap?

- **Gap of golden label criteria** between STS and MTM
  - STS: sharing most elements, different tense → **4** (**higher**)
  - MTM: sharing most elements, different tense, difficult to make sense → **-0.83** (**lower**)



Experiment 3: Similarity granularity gap → Evaluation gap?

- **Golden label criterion gaps** between STS and MTM
  - STS: sharing most elements, different tense → **4** (**higher**)
  - MTM: sharing most elements, different tense, difficult to make sense → **-0.83** (**lower**)



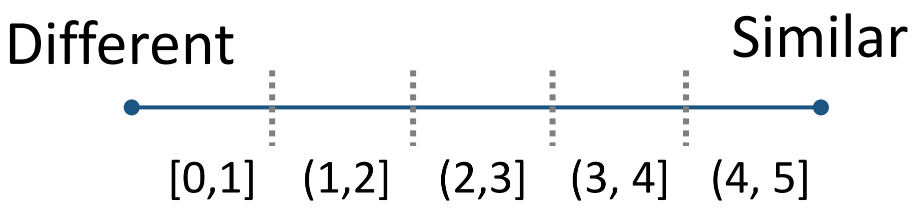
In MTM, we should capture more fine-grained & high similarity sentence pairs [Ma+, 2019]
→ *Hypothesis: Is STS's granularity **insufficient for fine-grained evaluation**?*

Experiment 3: Similarity granularity gap → Evaluation gap?

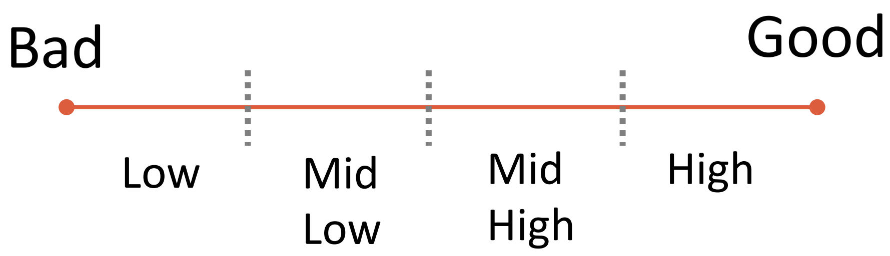
Hypothesis: STS's granularity is ***insufficient for fine-grained evaluation***

- We made subsets according similarity scores for STS and MTM
  - STS: 5 subsets (based on label definition)
  - MTM: 4 subsets (based on quantiles)

STS



MTM



Experiment 3: Similarity granularity gap → Evaluation gap?

*Hypothesis: STS's granularity is **insufficient for fine-grained evaluation***

	STS-[0, 1]	STS-(1,2]	STS-(2,3]	STS-(3,4]	STS-(4,5]
MTM-Sim-Low	0.101	-0.001	-0.008	0.627	0.643
MTM-Sim-MidLow	0.065	-0.046	-0.172	0.708	0.690
MTM-Sim-MidHigh	-0.097	-0.214	-0.330	0.639	0.592
MTM-Sim-High	-0.088	-0.267	-0.387	0.533	0.529

※ Spearman correlation between STS pearson corr. <-> MTM pearson corr.
※ Darker color represents lower correlation

only the high-similarity subsets of STS were highly correlated with MTMs
→ STS granularity does not capture fine-grained similarity

Conclusions & Future work

- We alert that the potentially-common assumption for STS benchmark

  "better on STS → better on application-oriented tasks"

- We expose **three factors** contribute to the evaluation gap between STS and application-oriented tasks
  - Factor 1: **Sentence length gap**
  - Factor 2: **Vocabulary coverage gap**
  - Factor 3: **Similarity granularity gap**
- Future work
  - Make a reliable benchmark for prediction similarity model
  - Investigate other factors, tasks, and domains
  - Causal inference

Appendix

Dataset: STS benchmark

- **STS dataset: STS-b [Cer+, 2017]**
  - Data: (s1, s2, human_label)
  - Human workers annotated the similarity label (5~6 grades) per instance (s1, s2)
  - Evaluation metric: pearson or spearman correlation

Application-oriented task datasets: MTM, PR

- MTM dataset: WMT17 [Bojar+, 2017]*1
  - Evaluate hypothesis (model output) with references
  - We used segment-level Direct Assesement dataset
  - Data: (hyp, ref, human label)
    - Human workers annotated the similarity label (100 grades) per segment (hyp, ref)
  - Evaluation metric: pearson or kendall correlation
- Passage Retrieval dataset: MS-MARCO [Bajaj+, 2016]*2
  - Search most related passage with query
  - We used Passage Re-ranking dataset
  - Data : (query, [1,000 passages list], related_passage)
    - Search related_passage from 1,000 passages using query
  - Evaluation metric: Mean Reciprocal Rank (MRR)@10

*1 https://www.statmt.org/wmt17/results.html
*2 https://microsoft.github.io/msmarco/

15 Model descriptions

- **BoW, BoW+TFIDF** : 2 models
  - Pooling: sum
- **BoV-{Pre-trained Vectors}-{Pooling}** : 6 models*
  - Pre-trained Vectors: word2vec, Glove, fasttext
  - Pooling: {max, mean}
  - * in MS-MARCO, remove word2vec models due to computational order
- **BERTScore-{Pre-trained LM}-{Scores}** : 6 models
  - Pre-trained LM: {BERT-base-uncased, RoBERTa-large}
  - Scores: {precision, recall, F1-score}
- **SimCSE (unsupervised model)** : 1 model

→ We calculate **correlation between performance on STS <-> application tasks (MTM, PR)** on these models in each subset

STS is one of the representative benchmark tasks in NLP

- **GLUE**: a collection of benchmark dataset in NLP
  - Aims generalization model for dataset size, text genres, degrees of difficulty

- **Semantic Textual Similarity (STS)** is one of GLUE tasks

GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

https://gluebenchmark.com/

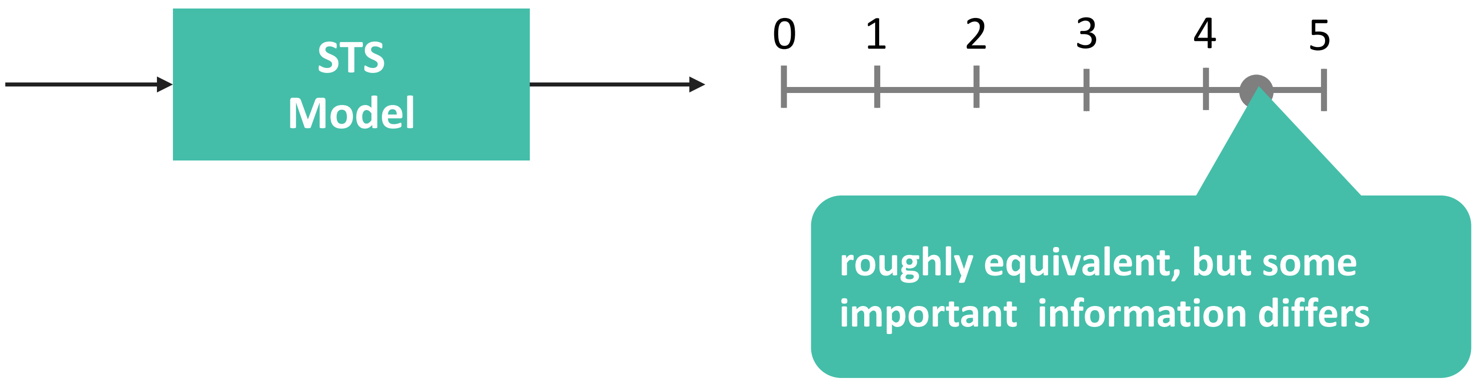
Task Definition: Semantic Textual Similarity (Agirre+, 2012)

Input

s1: *They flew out of the nest in groups.*
（彼らは集団で巣から飛び出した。）

s2: *They flew into the nest together.*
（彼らは一緒に巣へ飛び込んだ。）

Output



- Judge similarity of two sentences with gradation score
  - 0 -> "no relation", 5 -> completely same
- Benchmark dataset: STS-b[Cer+, 2017]