文間意味的類似度のベンチマークタスクと実応用タスクの乖離

Gap between Semantic Textual Similarity Benchmark Task and Downstream Tasks

阿部香央莉 [*1] 横井祥 [*1*2] 梶原智之 [*3] 乾健太郎 [*1*2]
Kaori Abe Sho Yokoi Tomoyuki Kajiwara Kentaro Inui

[*1]東北大学 Tohoku University
[*2]理化学研究所 RIKEN
[*3]愛媛大学 Ehime University

The Semantic Textual Similarity (STS) task measures the ability to evaluate the similarity between two sentences, which is necessary for downstream tasks such as machine translation evaluation and related passage retrieval. Several NLP researchers discuss the performance of this ability on benchmark dataset. However, there is a possibility that a system that is highly evaluated on the benchmark dataset may not be able to demonstrate appropriate effectiveness in actual downstream tasks. In this study, we examined this gap between STS and downstream tasks, clarified what factors are important in evaluating the similarity between two sentences in the downstream tasks, and discussed a policy for improving the benchmark dataset.

1. はじめに

2 文間の意味的類似度の計算は，言語処理における多くの後段タスクにおいて必要となる．例えば，機械翻訳，画像キャプション生成，要約，平易化などのテキスト生成モデルの出力を評価する際，モデルから出力された文を毎回人手で評価するにはコストがかかるため，基本的には参照文という正解の文を用意して，参照文と出力文を照らし合わせて評価を行う．このとき，文字列による表層的なマッチングでは適切な評価が得られないことが問題となっており [Sulem 18, Freitag 20]，出力文と参照文の意味的な類似度を測ることができる自動評価への期待が高まっている [Zhang 20, Yuan 21]．また，質問応答における関連文検索においても，クエリとなる質問文と検索対象の文をそれぞれ文表現に落とし込み，その 2 つの文表現の類似度を比較することで，関連度の高い文を抽出するモデルが主流となっている [Chen 17, Karpukhin 20]．また，剽窃検出や記述式問題の自動採点などのタスクにおいても，原文と意味が全く同じで表現を変えたのみの剽窃文や，模範解答と学生の様々な記述文の間で，意味的類似度の計算が必要となる．

2 文間の類似度予測能力を測るタスクとして，Semantic Textual Similarity (STS; 文間意味的類似度タスク) が言語処理分野のデファクトとなっている [Agirre 12]．この STS のベンチマークデータセットである STS-b [Cer 17] での精度を元に，これまで様々な文間類似度予測システムの性能が報告されている [Reimers, Giorgi 21, Gao 21]．また，STS のためのモデルを直接機械翻訳評価に組み込むことで，モデル性能を向上させた例もある [Wieting 19, Yasui 19]．したがって，STS のベンチマークデータを解けるモデルを作っていくことが言語処理の広範な応用タスクの性能向上につながる，という仮定が分野のコンセンサスとなっている．

しかし，実際のところこの仮定は正しいのだろうか？本研究では，STS と後段タスクではデータの性質が異なること，またこれが原因となり STS における評価結果と後段タスクでの評価結果にギャップが生まれ得ることを示す．具体的には (i) 各文の長さ (ii) 比較する 2 文の長さの差のふたつの観点で調査をおこない，ひとつめの観点が評価ギャップに直結することを述べる．

連絡先: 阿部香央莉 (abe-k@tohoku.ac.jp)
東北大学大学院 情報科学研究科 システム情報科学専攻 乾研究室
〒 980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-05
東北大学工学研究科 電子情報システム・応物系 1 号館 6 階

2. STS・後段タスク設定

本論文では，文表現獲得のベンチマークとして扱われている STS と，文表現を扱う代表的な後段タスクである 2 種類のタスク（機械翻訳評価, 関連文検索）での評価のギャップの要因を検証する．本論文で取り扱う 3 種類のタスクおよびデータセットについて説明する．

2.1 STS (STS-b)

STS とは，文表現の獲得を目的として SemEval において 2012 年に提案されて以降 2017 年まで開催されたタスクである．この STS は言語処理分野において主要なタスクの一つとなっており，GLEU データセットにも採用されている [Wang 19]．1 インスタンスは (文 1，文 2，スコア) で表され，スコアは文 1，文 2 の意味的類似度を表す．STS におけるベンチマークデータ STS-b [Cer 17] は，STS12-17 間で収集されたデータから，画像キャプション・ニュース記事・ フォーラム関連のデータを用いて作成されたものである．評価時は，正解のスコアと類似度予測モデルが予測したスコアのピアソン相関またはスピアマン相関で評価を行う．

2.2 機械翻訳評価（WMT17）

STS で獲得されるような文表現の代表的な応用先の一つとして，MT Metric（機械翻訳評価）タスクがあげられる．Metric タスクは，機械翻訳ワークショップの中でも最大規模である WMT において 2008 年から開催されているタスクの一つである．本タスクでは，ワークショップに投稿された機械翻訳モデルの出力結果に対して行われた人手評価との相関が高くなるような自動評価モデルを提案することを目的としている．2016 年から，参照文と出力文の近さをクラウドワーカーが 100 段階評価する Direct Assessment (DA) という評価形式が採用されている．本論文では，WMT17 の segment-level DA の x-en (cs-en, de-en, fi-en, lv-en, ru-en, tr-en, zh-en) データ [*1] を用いて分析を行う．1 インスタンスは（参照訳, 翻訳モデル出力文, スコア）で表され，スコアは参照訳と出力文の類似度を表

*1 https://www.statmt.org/wmt17/results.html

	STS-b	WMT17	MS-MARCO
インスタンス数	8,628	3,793	6,890
文数 (s1+s2)	17,256	4,261	13,337,934

表 1: 本論文で使用するデータセットの統計量.

	文長	文長差
STS-b	10.173 ± 5.370	1.995 ± 1.988
WMT17 (x-en)	20.455 ± 10.008	2.893 ± 2.756
MS-MARCO	31.493 ± 30.492	48.489 ± 21.337
MS-MARCO (query)	6.033 ± 2.527	-
MS-MARCO (passage)	56.952 ± 23.597	-

表 2: 各データセットにおける文長および文長差の標本平均および標本標準偏差（文長は単語数で計算）.

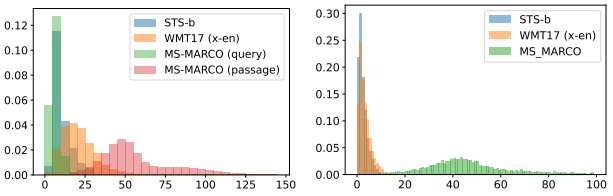


図 1: 各データセットにおける文長分布および文長差のヒストグラム（左：文長，右：文長差）.

し，0-100 の値で人手評価されたものを正規化した値とする．評価時は，人手評価のスコアと自動評価モデルが予測したスコアのピアソン相関またはケンドールの順位相関で評価を行う．

2.3 関連文検索（MS-MARCO）

質問応答における関連文検索は，Google 検索など幅広いユーザが用いる検索システムの性能向上のために必要とされる重要なタスクの一つである．関連文検索の主なデータセットとして， 本論文では Microsoft から提供されている MS-MARCO [Bajaj 18] の Passage Re-ranking のデータを用いる．Passage Re-ranking は，あるクエリに対し 1000 の関連文候補が与えられ，その候補をクエリに対する関連度順でリランキングするタスクである．本論文では，説明の簡略化のために 1 インスタンスを (クエリ, 関連文候補, 正解となる関連文リスト) で表すこととする．2 文間類似度予測システムへの入力としては, (クエリ，関連文候補中の文）のペア 1000 件が与えられ，候補中の文ごとにクエリに対する類似度を予測し，この類似度を本タスクにおける関連度と見なしてリランキングする．リランキング後の上位に，正解となる関連文リスト中の関連文が出現するほど，良いモデルであると言える．評価時は Mean Reciprocal Rank（MRR）で評価を行う．

3. STS・後段タスクデータ間のギャップ

現在の言語処理モデルは一般的に単語を処理の基本単位としており，テキストにおける類似度を測る際も同様である．ここで，類似度を測定する対象のテキストに含まれる単語の数に大きな分散があることに注意したい．例えば，MS-MARCO においてはごく短い検索用のスニペットや，また非常に長い文書が扱われることもある．このようなタスクによる文長の違いは評価結果のギャップに繋がるのではないだろうか？本論文では，以後，文長に着目した分析および実験を行う．各タスクのデータセット (STS-b，WMT17, MS-MARCO) に関するデータの統計量を表 1 に示す．以降，本論文では「文長」といった時，各文の単語数を表すものとする．

まず， STS のデータセット (STS-b) および後段タスクのデータセット (WMT17, MS-MARCO) の文長分布を確認する．3 種類のデータセットの文長についての標本平均および標本標準偏差を表 2 に示す．また，同じく 3 種類のデータセットの文長分布のヒストグラムを図 1 左に示す．表 2 および図 1 左より，STS-b と WMT17 および MS-MARCO には，類似度を評価する対象となる文の長さの分布に大きなギャップがあることが読み取れる．例えば，文長分布のヒストグラムにおける STS-b は 10 付近に尖った山が突出する形となっているのに対し，WMT17 や MS-MARCO の関連文 (passage) の山はそれぞれ 20 付近，50 付近に頂点が来る比較的なだらかな山となっている．MS-MARCO のクエリ（query）の分布は STS-b に近い分布となっているが，このクエリは実際の文書検索時のもの（例：“foods and supplements to lower blood sugar” など）であり，非文も多く含まれている．これらの図表より，STS-b の文長分布は後段タスクデータの文長分布と異なる形をしていることがわかる．

また，図 1 の通り，MS-MARCO は特に類似度を比較する 2 文（クエリ，関連文）間の文長差が大きい．そこで，文長差についての分布についてヒストグラムを示したのが図 1 右である．STS-b と WMT17 の間では文長差の分布に大きな違いはないが，その 2 種類と MS-MARCO では文長差の分布に極めて大きな違いがあることが図 1 右から読み取れる．

これらの分布の違いを元に，我々は STS と後段タスク間での評価のギャップが生じる原因として，(i) データセット間の**文長分布のギャップ**が評価のギャップを引き起こしている (ii) データセット間の**文長差分布のギャップ**が評価のギャップを引き起こしているという二つの仮説を考えた．以降では，これらの仮説に関して実験を行う．

4. STS･後段タスクデータ間のギャップの影響

4.1 実験設定

4.1.1 仮説 (i) 文長分布ギャップが評価ギャップの原因

後段タスクのデータセット（WMT17）を，文長に応じて複数のサブセットに分割し，それぞれのサブセットでの評価と STS-b での評価との相関を測る．サブセットの作成にはビン分布近似を用いる．文長が短いビンであるほど，サブセットは STS-b の文長分布に近づき，反対に文長が長いビンであればあるほど，STS-b の文長分布から平均がずれたサブセットとなる．ビンの基準には，WMT17 は参照文と出力文の平均値，MS-MARCO は関連文の長さを使用した．

各サブセットのサイズおよび標本平均・標本標準偏差を表 3 に，文長のヒストグラムを図 2(a) および (b) に示す．表や図に示した通り，文長が短いビン（WMT-(0,40) など）は STS-b の文長分布に近い分布を再現できていることが確認できる．

4.1.2 仮説 (ii) 文長差分布ギャップが評価ギャップの原因

後段タスクのデータセット（WMT17, MS-MARCO）を，文ペアの文長差に応じて 3 つのサブセット（WMT-diff-short, WMT-diff-long, MS-MARCO-diff-long）に分割し，それぞれのサブセットでの評価 と STS-b での評価との相関を測る．サブセットの作成は，仮説 (i) と同様の手順で行い，WMT-diff-short は (0, 20)，WMT-diff-long は (4, 24), MS-diff-long は (10, 30) の範囲でそれぞれ文長差を元にサブセットを作成した．

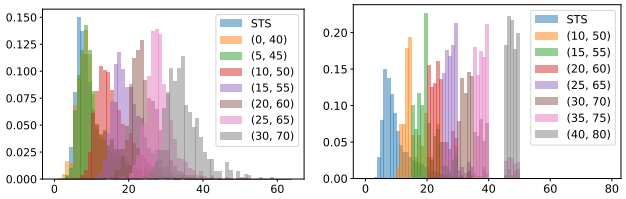
出来上がった文長差サブセット（WMT-diff-short, WMT-diff-long）の文長差分布（標本平均・標本標準偏差）を表 4,

	WMT		MS-MARCO	
ビン	サイズ	文長平均・分散	サイズ	文長平均・分散
(0, 40)	657	10.463 ± 5.206	-	-
(5, 45)	637	10.681 ± 5.134	-	-
(10, 50)	1413	15.858 ± 4.985	67	16.045 ± 4.420
(15, 55)	1327	19.778 ± 4.414	119	19.849 ± 3.759
(20, 60)	951	23.838 ± 3.980	199	23.704 ± 3.285
(25, 65)	602	27.879 ± 3.905	262	28.000 ± 2.980
(30, 70)	430	34.280 ± 4.816	561	34.526 ± 3.855
(35, 75)	-	-	690	38.323 ± 3.549
(40, 80)	-	-	932	46.987 ± 1.390

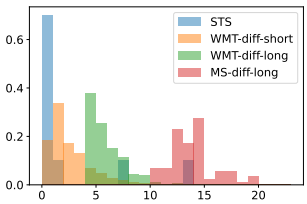
表 3: WMT17，MS-MARCO の文長サブセットのサイズ（インスタンス数）および標本平均，標本標準偏差．

	サイズ	文長差
WMT-diff-short	2766	2.041 ± 1.961
WMT-diff-long	1030	5.450 ± 1.753
MS-diff-long	87	13.414 ± 2.071

表 4: WMT17，MS-MARCO の文長差サブセットのサイズ（インスタンス数）および標本平均，標本標準偏差．



(a) WMT17 文長サブセット　(b) MSMARCO 文長サブセット



(c) 文長差サブセット

図 2: 各サブセットの文長分布．

ヒストグラムを図 2(c) に示す．仮説 (i) と同じく，こちらも WMT-diff-short は STS-b の文長差分布に近い分布を再現できていることが確認できる．

4.1.3　2 文間類似度予測モデル

仮説 (i), (ii) の影響を調べるために，複数の類似度予測モデルを用意し，それらモデルの STS-b での評価結果と後段タスクでの評価結果の相関を調べる．もし STS-b に近い分布を持つ WMT-(0, 40) や WMT-diff-short サブセットでの評価と STS-b での評価の相関が，STS-b から平均がずれた分布となる WMT-(30, 70)，MS-(40, 80) や WMT-diff-long での評価と STS-b での評価の相関よりも高ければ，仮説が支持される．すなわち，文長や文長差の違いが評価のギャップに影響を与えている可能性が高いということになる．

STS での評価と後段タスクでの評価の相関を測るために，複数の類似度予測モデルでの結果が必要となる．本実験で扱うモデル 11 種類を以下に示す．

- BoW, BoW-TFIDF (sum pooling, cos 類似度)
- BoV ({Word2Vec, GloVe, fastText} × {mean, max pooling} × cos 類似度)
- BERTScore (precision, recall, f1-score)[*2] [Zhang 20]
- SimCSE[*3] [Gao 21]

BoV モデルの構築には `pymagnitude` モジュールを用い，Word2Vec, GloVe, fastText の事前学習済みモデルには 300 次元の Light モデル（Google-word2vec-GoogleNews-100B, Stanford-GloVe-CommonCrawl-840B, Facebook-fastText-CommonCrawl-600B）を使用した．BERTScore のモデルはデフォルトの `roberta-large` を使用した．また，SimCSE のモデルは `princeton-nlp/sup-simcse-bert-base-uncased`（教師あり学習済モデル）を使用した．

STS-b での評価には，ピアソンの相関係数 r およびスピアマンの相関係数 ρ の 2 種類を用いる．また，WMT17 およびそのサブセットでの評価には，ピアソンの相関係数 r およびケンドールの順位相関係数 τ の 2 種類を用いる．MS-MARCO での評価には，MS-MARCO のガイドラインに則り MRR@10 で評価した結果を用いる．実験では，STS-b での評価 2 種類と WMT17 での評価 2 種類，MS-MARCO での評価 1 種類を組み合わせた，計 4 種類の相関の結果（{STS-r, STS-ρ} × {WMT17-ρ, WMT17-τ, MS-MRR}）を得る [*4]．

4.2　実験結果：仮説 (i), (ii) の検証

作成したサブセットを元に，文長のギャップおよび文長差のギャップが評価の相関に影響を与えているかどうかを確認する．表 5 に，各類似度予測器に対する WMT17 全体およびそのサブセットでの評価結果と STS-b での評価結果のスピアマン相関を示す．また，表 6 に，各類似度予測器に対する MS-MARCO 全体およびそのサブセットでの評価結果と STS-b での評価結果のスピアマン相関を示す．この時，後段タスクでの評価結果として，表 3 のサイズで元データから 5 回サンプリングした結果 (シード値 0-4) の平均値を用いている．

仮説 (i) について，WMT，MS-MARCO 両方の後段タスクの文長サブセット（WMT-(x,y), MS-(x,y)）において，STS-b の平均文長から離れていけばいくほど，STS-b での評価との相関が低くなっていく傾向が見てとれる．これは，データセット間の**文長ギャップが評価のギャップに繋がっている**ことを示唆する結果である．

仮説 (ii) については，予想に反し，WMT では STS-b の平均から離れたサブセット（WMT-diff-long）の方が STS-b の平均に近づけたサブセット（WMT-diff-short）よりも STS-b での評価との相関が高いという結果となった．また，WMT サブセットよりさらに文長差ギャップが大きい MS-diff-long においても，データ全体の相関よりも STS-b での評価との相関が高い，という結果が得られた．このことより，類似度を比較する 2 文の文長差ギャップが評価のギャップに影響を与えているわけではない，ということがわかった．

4.3　考察：STS と最も相関が高い後段タスクの文長は？

表 5 や表 6 より，STS-b と後段タスク間の相関は文長に応じて単調減少するわけではなく，途中でデータセット全体よりも相関が高くなる箇所が WMT，MS-MARCO 両結果に現れている．WMT では STS-ρ × WMT17-r の結果におけ

*2　`https://github.com/Tiiiger/bert_score`
*3　`https://github.com/princeton-nlp/SimCSE`
*4　計算時間の都合で，MS-MRR の評価は BoV-Word2Vec-{mean, max} を除いたモデル 9 種類での結果を報告している．

	STS-r × WMT17-r	STS-ρ × WMT17-r	STS-r × WMT17-τ	STS-ρ × WMT17-τ
WMT-all	0.671 (p=0.017*)	0.476 (p=0.118)	0.755 (p=0.005*)	0.622 (p=0.031*)
WMT-(0, 40)	0.650 (p=0.022*)	0.434 (p=0.159)	0.643 (p=0.024*)	0.434 (p=0.159)
WMT-(5, 45)	0.657 (p=0.020*)	0.448 (p=0.145)	0.706 (p=0.010*)	0.524 (p=0.080)
WMT-(10, 50)	0.517 (p=0.085)	0.420 (p=0.175)	0.524 (p=0.080)	0.406 (p=0.191)
WMT-(15, 55)	0.643 (p=0.024*)	0.517 (p=0.085)	0.601 (p=0.039*)	0.434 (p=0.159)
WMT-(20, 60)	0.399 (p=0.199)	0.357 (p=0.255)	0.133 (p=0.681)	0.035 (p=0.914)
WMT-(25, 65)	0.413 (p=0.183)	0.385 (p=0.217)	0.469 (p=0.124)	0.469 (p=0.124)
WMT-(30, 70)	0.364 (p=0.245)	0.217 (p=0.499)	0.147 (p=0.649)	-0.007 (p=0.983)
WMT-diff-short	0.713 (p=0.009*)	0.531 (p=0.075)	0.776 (p=0.003*)	0.664 (p=0.018*)
WMT-diff-long	0.769 (p=0.003*)	0.657 (p=0.020*)	0.727 (p=0.007*)	0.657 (p=0.020*)

表 5: WMT17 サブセットにおける評価結果と STS-b での評価結果のスピアマン相関係数（$p < 0.05$ には*を付加）.

	STS-r × MS-MRR	STS-ρ × MS-MRR
MS-all	0.758 (p=0.011*)	0.673 (p=0.033*)
MS-(10, 50)	0.733 (p=0.016*)	0.624 (p=0.054)
MS-(15, 55)	0.855 (p=0.002*)	0.782 (p=0.008*)
MS-(20, 60)	0.879 (p=0.001*)	0.794 (p=0.006*)
MS-(25, 65)	0.782 (p=0.008*)	0.648 (p=0.043*)
MS-(30, 70)	0.745 (p=0.013*)	0.624 (p=0.054)
MS-(35, 75)	0.709 (p=0.022*)	0.636 (p=0.048*)
MS-(40, 80)	0.673 (p=0.033*)	0.600 (p=0.067)
MS-diff-long	0.806 (p=0.005*)	0.697 (p=0.025*)

表 6: MS-MARCO サブセットにおける評価結果と STS-b での評価結果のスピアマン相関係数（$p < 0.05$ には*を付加）.

る (15, 55) においてデータ全体よりも相関が高くなった．また MS-MARCO ではその相関の上昇がより顕著に現れており，{STS-r, STS-ρ} × MS-MRR における (15, 55), (20, 60), (25, 65) の範囲でデータ全体よりも相関が高くなっている．

一部のサブセットで相関が高くなる原因として，WMT-(15, 55) では BoW モデルの評価が他と比べて高くなっていることが挙げられる．各類似度予測モデルの STS-r を見ると，SimCSE が 0.842，次いで BoW が 0.697，BoW-IFIDF が 0.696, その後に BERTScore，BoV モデルが続く形となっており，単純な BoW モデルで二番目に高い性能が出ていることが確認できた．対し，WMT 全体では BERTScore，SimCSE，BoW，BoV の順番になっており，サブセットの相関で見ると WMT-(15, 55) が WMT 全体よりも高いという結果になった．

全体的には，比較的短い文長であるほど STS-b との相関が高くなる傾向が両後段タスクで一貫しており，文長が STS-b と後段タスクでの評価ギャップに影響を与えていることが明らかになったのは事実である．これは，現状の STS ベンチマークデータでは，実際の後段タスク中で出現しうる長い文に対しての 2 文間類似度予測の評価が十分にできていないということにつながる．今後の課題として，今回の結果を元に STS-b のデータの文長分布がどうあるのが適切かを詳しく分析していくこと，また文長以外の観点（例えば語彙やデータセットの難易度など）でさらに分析を深めることが挙げられる．

5. おわりに

各種類似度予測モデルの，STS データでの性能と後段タスクデータでの性能が必ずしも一致しないことを起点に，後段タスクの必要条件としての STS データの良し悪しを議論した．本論文では，主に STS データおよび後段タスクデータの文長分布のギャップが評価のギャップに影響を与えているかどうか，STS-b の分布に近似したサブセットを後段タスクのデータから抽出し，それらの相関を測ることによって検証を行った．結果として，文長のギャップが STS-b での評価と後段タスクでの評価のギャップに影響を与えている可能性を示唆する結果が得られた．今後の方針として，STS-b における適切な文長分布や，文長以上に影響を与えている要因があるかを調べ，STS のベンチマークデータセット改善に努めていきたい．

謝辞 本研究は科研費 JP20J21694 の助成を受けたものです。

参考文献

[Agirre 12] Agirre, E., et al.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity, in **SEM*, pp. 385–393 (2012)

[Bajaj 18] Bajaj, P., et al.: MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, *ArXiv* (2018)

[Cer 17] Cer, D., et al.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, in *SemEval*, pp. 1–14 (2017)

[Chen 17] Chen, D., et al.: Reading Wikipedia to Answer Open-Domain Questions, in *ACL*, pp. 1870–1879 (2017)

[Freitag 20] Freitag, M., et al.: BLEU might be Guilty but References are not Innocent, in *EMNLP*, pp. 61–71 (2020)

[Gao 21] Gao, T., et al.: SimCSE: Simple Contrastive Learning of Sentence Embeddings, in *EMNLP*, pp. 6894–6910 (2021)

[Giorgi 21] Giorgi, J., et al.: DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations, in *ACL-IJCNLP*, pp. 879–895 (2021)

[Karpukhin 20] Karpukhin, V., et al.: Dense Passage Retrieval for Open-Domain Question Answering, in *EMNLP*, pp. 6769–6781 (2020)

[Reimers] Reimers, N.Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in *EMNLP-IJCNLP*

[Sulem 18] Sulem, E., et al.: BLEU is Not Suitable for the Evaluation of Text Simplification, in *EMNLP*, pp. 738–744 (2018)

[Wang 19] Wang, A., et al.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, in *ICLR* (2019)

[Wieting 19] Wieting, J., et al.: Beyond BLEU:Training Neural Machine Translation with Semantic Similarity, in *ACL*, pp. 4344–4355 (2019)

[Yasui 19] Yasui, G., et al.: Using Semantic Similarity as Reward for Reinforcement Learning in Sentence Generation, in *ACL-SRW*, pp. 400–406 (2019)

[Yuan 21] Yuan, W., et al.: BARTScore: Evaluating Generated Text as Text Generation, in *NeurIPS* (2021)

[Zhang 20] Zhang, T., et al.: BERTScore: Evaluating Text Generation with BERT, in *ICLR* (2020)