正準角および部分空間に基づく BERTScore の拡張

石橋 陽一 [1]　横井 祥 [2,3]　須藤 克仁 [1]　中村 哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 東北大　[3] 理研 AIP
{ishibashi.yoichi.ir3, sudoh, s-nakamura}@is.naist.jp
yokoi@tohoku.ac.jp

概要

本研究では生成文自動評価の代表的な手法である BERTScore を拡張し、新たな文類似度 **Subspace-BERTScore** を提案する。我々は文に明示的に含まれない意味も考慮するため、BERTScore における文表現（ベクトル集合）と指示関数（集合への含まれ度合い）に着目し、これらを部分空間表現と正準角を用いて自然に拡張する。生成文評価タスクと文類似度タスクでの実験により提案法が BERTScore の文類似推定性能を安定して改善する事を示す。

1　はじめに

生成文の自動評価は、自然言語処理の重要な課題の 1 つである。自動評価ではシステム訳文と参照文を比較することで意味的な同等性を評価することを目指している [1]。ベクトル集合のマッチングに基づく類似度の中で代表的な手法である BERTScore [2] はシンプルで利便性が高く、最も使用されている手法である。BERTScore は BERT [3] や RoBERTa [4] 等の事前学習済みベクトルを利用して、ベクトル間のコサイン類似度の総和に基づき文類似度を計算する手法である。

BERTScore において技術的に最も重要なものは指示関数である。これは、文を単語集合と考えたときに、ある単語集合中のトークン（*royal*）が、もう一方の単語集合（$B = \{We, are, king, and, queen\}$）にどの程度含まれているか（$royal \in B$）を定量化するものである。我々が着目するのは、この一件自然なアプローチが，文内のシンボルとしては明示的に含まれないが文意には影響する意味を取り逃がす可能があるという点である。例えば *king* と *queen* を含む文が持つ *royal* という意味情報が、*royal* という単語が B に含まれないという理由で BERTScore の類似性計算は反映することができない。

そこで、本研究では、部分空間の考え方 [5] を取り入れることで、BERTScore の指示関数を拡張し、文類似度の改善に取り組んだ。提案法は表現力を向上させるだけでなく、ベクトル集合間のマッチングを Recall、Precision、F-score に帰着させるという BERTScore の基本的な方針を継承しているため、タスクへの非依存性といった BERTScore の利点は損なわれていない。本研究では生成文評価タスクと文類似度タスクでの実験により提案法が BERTScore の文類似推定性能を改善する事を示した。

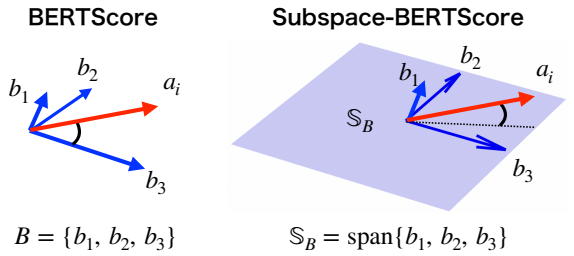

図 1　提案法（Subspace-BERTScore）と BERTScore の比較。提案法は文（B）を部分空間（$\mathbb{S}_\mathbf{B}$）で表現し、部分空間とトークンベクトル（$\boldsymbol{a}_i$）との類似度（≒ 正準角）を計算することで、文類似尺度の表現力を向上させている。

2　準備

問題設定　本研究では 2 つの文に対する類似度の開発を目指している。本手法が対象とするタスク（生成文自動評価・文類似度タスク）では 2 つの文が与えられ、これらの類似性を判定する。したがって本研究では与えられる 2 つの文 A, B を事前学習済みのベクトルで表現した上で、それらの類似度を計算する。

記法　以降の議論を明確にするために、いくつかの記号を定義しておく。2 つの文 A, B のトークンの集合をそれぞれ $A = \{a_1, a_2, ...\}, B = \{b_1, b_2, ...\}$ と表記する。事前学習済み埋め込みによって表現される文脈化トークンベクトルの集合を $\mathbf{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ...\}, \mathbf{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ...\}$ と表記する。ここで $\boldsymbol{a}, \boldsymbol{b}$ をトークンのベクトルとする。また、$\mathbf{A}$ によっ

て張られる部分空間を $\mathbb{S}_{\mathbf{A}}$ と表記する。

3 BERTScore

ここでは生成文のマッチングに基づく文類似度における最も代表的な評価手法である BERTScore [2] における「文の表現方法」と技術的に最も重要な「指示関数」を概説しその限界を指摘する。

離散シンボル集合間の類似度 BLEU [1] や METEOR [6] などの BERTScore 以前のいくつかの評価手法は、Recall (R)、Precision (P) の計算に基づく手法であった[1)]。

$$R = \frac{1}{|A|} \sum_{a_i \in A} \mathbb{1}_{\text{set}}[a_i \in B], \tag{1}$$

$$P = \frac{1}{|B|} \sum_{b_i \in B} \mathbb{1}_{\text{set}}[b_i \in A]. \tag{2}$$

ここで指示関数 $\mathbb{1}_{\text{set}}$ は「集合に含まれる (1) か否か (0)」を離散的に定量化する関数である。

$$\mathbb{1}_{\text{set}}[a_i \in B] = \begin{cases} 1 & \text{if } a_i \in B, \\ 0 & \text{if } a_i \notin B. \end{cases} \tag{3}$$

BERTScore における 3 つの類似度 BERTScore は R, P そして F-score (F) を埋め込みを用いて近似的に計算する手法である。

$$R_{\text{BERT}} = \frac{1}{|A|} \sum_{\boldsymbol{a}_i \in \mathbf{A}} \mathbb{1}_{\text{vectors}}(\boldsymbol{a}_i, \mathbf{B}), \tag{4}$$

$$P_{\text{BERT}} = \frac{1}{|B|} \sum_{\boldsymbol{b}_i \in \mathbf{B}} \mathbb{1}_{\text{vectors}}(\boldsymbol{b}_i, \mathbf{A}), \tag{5}$$

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{6}$$

文はトークンベクトルの集合 $\mathbf{A}, \mathbf{B}$ として表現される。また、$\mathbb{1}_{\text{vectors}}$ は「ベクトル集合」に対する指示関数である。直感的には、この $\mathbb{1}_{\text{vectors}}$ は「ベクトル集合への −1〜1 の含まれ度合い」であり、$\mathbb{1}_{\text{set}}$ よりも柔軟な指示関数となっている。具体的には、$\mathbb{1}_{\text{vectors}}(\boldsymbol{a}_i, \mathbf{B})$ は文 A 中の i 番目のトークンベクトル $\boldsymbol{a}_i$ が文 B に対し意味的にどの程度含まれているかを $\boldsymbol{a}_i$ と B のトークンベクトルとの cos 類似度の最大値として定量化している。

$$\mathbb{1}_{\text{vectors}}(\boldsymbol{a}_i, B) = \max_{\boldsymbol{b}_j \in B} \cos(\boldsymbol{a}_i, \boldsymbol{b}_j) \in [-1, 1] \tag{7}$$

ここで $\cos(\boldsymbol{a}_i, \boldsymbol{b}_j)$ は $\boldsymbol{a}_i$ と $\boldsymbol{b}_j$ の cos 類似度である。

表現力の問題 指示関数 $\mathbb{1}_{\text{vector}}$ は BERTScore の計算において最も重要な演算である。$P_{\text{BERT}}, R_{\text{BERT}}$ は $\mathbb{1}_{\text{vectors}}$ の和であるし、F_{BERT} は $P_{\text{BERT}}, R_{\text{BERT}}$ に基づいて計算される。しかしながら BERTScore の指示関数は表現力の観点で問題がある。図 1 左は $\mathbb{1}_{\text{vectors}}$ の計算を可視化したものである。ここで $\mathbb{1}_{\text{vectors}}$ は B のすべてのベクトルとの類似度を計算し、最大値のみ返す。すなわち、B 中の「特定の単語」との類似度である。したがって、文が持つ非明示的で重要な意味情報との類似性が考慮されない。例えば *The king and queen.* は *royal* の意味を含むが、表層的には含まれない。BERTScore における類似度の計算対象は文中の単語のみであるので、*royal* との類似性は直接計算できない。この問題は文を単なるベクトル集合で表現していること、そしてそのようなベクトル集合のための指示関数を採用している事に起因する。

4 Subspace-BERTScore

我々は部分空間の考え [5] に基づいた文表現と指示関数により BERTScore を拡張した Subspace-BERTScore を提案する。

部分空間を使った文表現 BERTScore では、文がベクトル集合で表現されていたが、我々は文の非明示的な情報を表現するため、文をトークンベクトルが張る部分空間 $\mathbb{S}$ として表現する（図 1）。以前の研究で、単語ベクトルが張る部分空間が単語集合の意味的特徴をよく反映する表現であることがわかっている [7, 5]。

部分空間に対する指示関数 部分空間と単語の類似性を計算するため、部分空間に適用できる指示関数であるソフト帰属度 [5] を導入する。これは直感的には「部分空間とベクトルの類似度」として機能する。この指示関数はベクトル $\boldsymbol{a}$ と部分空間 $\mathbb{S}_{\mathbf{B}}$ とのなす最小の角度（第一正準角）に応じて 0 から 1 の連続値を返す[2)]。

$$\mathbb{1}_{\text{subspace}}(\boldsymbol{a}_i, \mathbb{S}_{\mathbf{B}}) = \max_{\boldsymbol{b}_j \in \mathbb{S}_{\mathbf{B}}} |\cos(\boldsymbol{a}_i, \boldsymbol{b}_j)| \in [0, 1] \tag{8}$$

ここで、$\mathbb{S}_{\mathbf{B}}$ は文 B のベクトルで張られる線形部分空間である（図 1）。

ソフト帰属度は文の全情報を含む部分空間との類似度を計算でき、BERTScore の指示関数の表現力を改善する。また、式を見るとソフト帰属度 $\mathbb{1}_{\text{subspace}}$（式 8）は BERTScore における指示関数 $\mathbb{1}_{\text{vectors}}$（式 7）の自然な拡張となっていることがわかる。

1) 実際には式 1, 式 2の A, B はトークンの n-gram のリストが使われる。

2) ソフト帰属度（式 8）は特異値分解で計算できる [5]。

提案法：P, R, F の拡張　これまでの議論に基づき、我々は文の部分空間表現とソフト帰属度を用いて BERTScore の R, P, F を計算する **Subspace-BERTScore** を提案する。

$$R_{\text{subspace}} = \frac{1}{|A|} \sum_{\boldsymbol{a}_i \in A} \mathbb{1}_{\text{subspace}}(\boldsymbol{a}_i, \mathbb{S}_{\mathbf{B}}), \tag{9}$$

$$P_{\text{subspace}} = \frac{1}{|B|} \sum_{\boldsymbol{b}_i \in B} \mathbb{1}_{\text{subspace}}(\boldsymbol{b}_i, \mathbb{S}_{\mathbf{A}}), \tag{10}$$

$$F_{\text{subspace}} = 2\frac{P_{\text{subspace}} \cdot R_{\text{subspace}}}{P_{\text{subspace}} + R_{\text{subspace}}}. \tag{11}$$

ここで、$R_{\text{subspace}}, P_{\text{subspace}}, F_{\text{subspace}}$ は Subspace-BERTScore の最終的な評価尺度である。

重要度の重み付け　過去の研究で、出現頻度の低い単語は一般的な単語よりも文において文類似性に重要な役割を果たすことが知られている [6, 8]。このことから BERTscore では重要度を重み付けする方法も提案している。そこで我々の手法も BERTScore と同様に以下のように重み付けを行う。

$$R_{\text{subspace}} = \frac{\sum_{\boldsymbol{a}_i \in A} \text{weight}(\text{a}_\text{i}) \mathbb{1}_{\text{subspace}}(\boldsymbol{a}_i, \mathbb{S}_{\mathbf{B}})}{\sum_{\boldsymbol{a}_i \in A} \text{weight}(\text{a}_\text{i})}, \tag{12}$$

$$P_{\text{subspace}} = \frac{\sum_{\boldsymbol{b}_i \in B} \text{weight}(\text{b}_\text{i}) \mathbb{1}_{\text{subspace}}(\boldsymbol{b}_i, \mathbb{S}_{\mathbf{A}})}{\sum_{\boldsymbol{b}_i \in B} \text{weight}(\text{b}_\text{i})}. \tag{13}$$

ここで、weight は重み付け関数である。我々は BERTScore で採用された inverse document frequency (IDF) またはベクトルの L2 ノルム [9] を使用する。

5　定量的評価

提案法の性能を検証するため、生成文の自動評価タスクと、文類似度タスクで評価を類似おこ。ここでは複数のデータセットにおける総合的な結果を述べる。各データセットの個々の結果については付録の表 4、表 5に記載する。

5.1　生成文自動評価タスク

実験設定　先行研究 [2] に習い我々は機械翻訳における自動評価タスクである WMT18 metrics task [10] の評価データセットを使用した。このデータセットには機械翻訳システム訳文と参照訳文、そして人手評価が含まれる。このタスクでは機械翻訳システム訳文と参照訳文の類似度を算出し、人手評価と自動評価のケンドールの順位相関係数（Kendall's τ）で評価する。本実験で我々は 7 言語から英語への翻訳におけるセグメントレベルの人手評価を使用した。事前学習済み言語モデルには BERTScore で高い性能を達成したことが報告されている $\text{RoBERTa}_{\text{large}}$[3] [4] の第 17 層を使用した。

表 1　WMT18 におけるセグメントレベルの人手評価との平均相関係数（Kendall's τ）。

手法	メトリック	重み付け	Avg.
BERTScore	F	-	.365
	P	-	.353
	R	-	.360
Subspace-BERTScore	F	-	**.372**
	P	-	**.358**
	R	-	**.365**
BERTScore	F	IDF	.364
	P	IDF	.353
	R	IDF	.360
Subspace-BERTScore	F	IDF	**.371**
	P	IDF	**.357**
	R	IDF	**.364**

表 2　各 STS データセットにおける人手評価との平均相関係数（Spearman's ρ）

手法	メトリック	重み付け	Avg.
BERTScore	F	-	.506
	P	-	.494
	R	-	.496
Subspace-BERTScore	F	-	**.526**
	P	-	**.511**
	R	-	**.514**
BERTScore	F	L2	.511
	P	L2	.500
	R	L2	.497
Subspace-BERTScore	F	L2	**.531**
	P	L2	**.516**
	R	L2	**.515**

結果　結果を表 1に示す。3 つの類似尺度 F-score、Precision、Recall に着目すると、いずれの類似度でも提案法が最も高い人手評価との相関を達成した。特に、この 3 つの中で F-score が最も高い性能であった。また、我々は 先行研究 [2] と同じく IDF による重み付けを BERTScore と提案法に行ったが、両者とも性能向上は確認されなかった。

5.2　文類似度タスク

実験設定　文類似度タスク（STS）では 2 つの文の類似度を算出し、人手評価と自動評価のスピアマンの順位相関係数（Spearman's ρ）[4] で評価する。

3）https://huggingface.co/roberta-large

4）WMT18 metrics task では τ が使われたが STS では多くの研究が ρ を採用しているため我々も ρ で評価を行っている。

データセットとして我々は SemEval shared task の 2012-2016 [11, 12, 13, 14, 15]、STS-benchmark (STS-B) [16]、そして SICK-Relatedness (SICK-R) [17] を用いた。我々は事前学習済み言語モデル $\text{BERT}_{\text{base}}$[5] [3] の最終層のベクトルを使用した。

結果　結果を表 2に示す。F-score、Precision、Recall のいずれの方法でも生成文自動評価タスクと同様、提案法が人手評価との最も高い相関を達成した。この中で最も高い性能であったメトリックは F-score であった。また、重要度の重みとして STS において有効性が確認されている L2 ノルム [9] で指示関数を重み付けした実験を行った結果、提案法および BERTScore ともに性能向上が確認され、両者ともに重み付けが有効であることがわかった。また、L2 ノルムによる重み付けを行った場合も提案法が既存法より優れていた。

6　指示関数の効果の分析

提案法は BERTScore よりも文の非明示的な情報を活用できているだろうか？ここで我々は非明示的な情報を持つ文に両手法を適用し比較を行う。

6.1　実験設定

ここでは「非明示的だが *royal* の意味を含む文」と「明示的に *royal* の意味を含む文」に対して指示関数を適用し比較する。我々は非明示的な文として *They are the king and queen.* を使用した[6]。この文には *royal* という単語は文中に出現していないが、文の本質的な意味としては *royal* に近い意味を包含している。明示的に含む文は *They are royalty.* である。指示関数が良く機能するならば、「非明示的な文」と「明示的に含む文」に対してどちらも *royal* を意味的に含むと判断するので、両者に対する指示関数の値には大きな差がないはずである。そこで我々は提案法と BERTScore の指示関数をこれらの文に使用し、指示関数の値の差で比較した。

6.2　結果

結果を表 3に示す。提案法は BERTScore よりも指示関数の値の差が小さいことがわかる。提案法は *king* や *queen* が張る空間で文を表現するため、文中に *royal* を含まなくともその空間に *royal* のような意味が含まれている。したがって、この文に対する指示関数の値は *royalty* を含む文のそれと近くなり、両者の差が小さくなったと考えられる。一方で BERTScore の指示関数は提案法のように *king* と *queen* どちらも含むような意味空間との類似性は計算できないため、提案法よりも差が大きくなったと考えられる。以上の結果から、提案法は BERTScore よりも高い表現力を持つことが示唆される。

表 3　指示関数の比較（$\mathbb{1}_{\text{vecs}}$: BERTScore の, $\mathbb{1}_{\text{sub}}$: 提案法の）。両者は文 B に対する単語 *royal* の包含度合いを計算する。

Word ($\boldsymbol{a}$)		Sentence (B)	$\mathbb{1}_{\text{vecs}}(\boldsymbol{a}, B)$	$\mathbb{1}_{\text{sub}}(\boldsymbol{a}, \mathbb{S}_{\mathbf{B}})$
royal	非明示的	They are the king and queen.	0.60	0.69
	明示的	They are royalty.	0.72	0.76
	指示関数の値の差		0.12	**0.07**

7　関連研究

近年の生成文評価には BLEURT [18]、C-SPEC [19, 20]、COMET [21] などがある。これらは人手評価のデータを使用して評価モデルの学習を行う手法である。これは本研究と目的が異なる。我々のタスクはモデルの学習は行わないマッチングベースの評価であり、事前学習された言語モデルによる教師なしの類似度を提案している。

BERTScore と類似した手法として MoverScore [22] がある。両者を比較すると BERTScore は基本的な性能が MoverScore よりも高く、マッチングベースの自動評価において最も使われる手法であることから、本研究では BERTScore を拡張した。

8　結論

本研究では、BERTScore の表現力の限界を指摘し、より表現力の高い方法 Subspace-BERTScore を提案した。これは BERTScore の自然な拡張になっており、類似度としての性能が向上している。実験では、機械翻訳における生成文自動評価タスクと文類似度タスクで BERTScore と提案法を比較した。その結果、全てのタスクと全メトリックで提案法が BERTScore よりも優位であり、部分空間の方法を取り入れることで性能を向上させることが示された。

また、我々が採用した部分空間に基づくアプローチは汎用性が高く、現代の自然言語処理で頻繁に使用される計算（文表現やベクトル集合に対する表現形式・類似性計算）を代替できる可能性があり、将来的にあらゆる場面での応用が期待できる。

5）https://huggingface.co/bert-base-uncased
6）これ以外の例は付録（表 6）に記載する。

謝辞

本研究は JSPS 科研費（JP22H03654, JP22H05106）、JST ACT-X（JPMJAX200S）の支援を受けたものです。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA**, pp. 311–318. ACL, 2002.

[2] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **CoRR**, Vol. abs/1907.11692, , 2019.

[5] Yoichi Ishibashi, Sho Yokoi, Katsuhito Sudoh, and Satoshi Nakamura. Subspace-based set operations on a pre-trained word embedding space. **CoRR**, Vol. abs/2210.13034, , 2022.

[6] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, **Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005**, pp. 65–72. Association for Computational Linguistics, 2005.

[7] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. Representing sentences as low-rank subspaces. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers**, pp. 629–634. Association for Computational Linguistics, 2017.

[8] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In **IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015**, pp. 4566–4575. IEEE Computer Society, 2015.

[9] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word Rotator's Distance. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020**, pp. 2944–2960. Association for Computational Linguistics, 2020.

[10] Qingsong Ma, Ondrej Bojar, and Yvette Graham. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, **Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018**, pp. 671–688. Association for Computational Linguistics, 2018.

[11] Eneko Agirre, Daniel M. Cer, Mona T. Diab, and Aitor Gonzalez-Agirre. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, **Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012**, pp. 385–393. The Association for Computer Linguistics, 2012.

[12] Eneko Agirre, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic Textual Similarity. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, **Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA**, pp. 32–43. Association for Computational Linguistics, 2013.

[13] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In Preslav Nakov and Torsten Zesch, editors, **Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014**, pp. 81–91. The Association for Computer Linguistics, 2014.

[14] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, **Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015**, pp. 252–263. The Association for Computer Linguistics, 2015.

[15] Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, **Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016**, pp. 497–511. The Association for Computer Linguistics, 2016.

[16] Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens, editors, **Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017**, pp. 1–14. Association for Computational Linguistics, 2017.

[17] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014**, pp. 216–223. European Language Resources Association (ELRA), 2014.

[18] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020**, pp. 7881–7892. Association for Computational Linguistics, 2020.

[19] Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020**, pp. 3553–3558. Association for Computational Linguistics, 2020.

[20] Kosuke Takahashi, Yoichi Ishibashi, Katsuhito Sudoh, and Satoshi Nakamura. Multilingual machine translation evaluation metrics fine-tuned on pseudo-negative examples for WMT 2021 metrics task. In Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Tom Kocmi, André Martins, Makoto Morishita, and Christof Monz, editors, **Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021**, pp. 1049–1052. Association for Computational Linguistics, 2021.

[21] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020**, pp. 2685–2702. Association for Computational Linguistics, 2020.

[22] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019**, pp. 563–578. Association for Computational Linguistics, 2019.

付録

本研究で実施した生成文自動評価タスクと文類似度タスクの詳細な結果を表 4と表 5に示す。結果の傾向は実験 5 と同じであり、個々のデータセットに対しても提案法が BERTScore を上回っている。

表 4 英語への機械翻訳自動評価タスクにおけるシステム訳と参照訳のセグメントレベルの人手評価との相関係数（Kendall's τ）。BERTScore の結果は [2] から引用した。

手法	メトリック	重み付け	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Avg.
BERTScore	F	-	.404	.550	.397	.296	.353	.292	**.264**	.365
	P	-	**.387**	.541	.389	.283	.345	.280	**.248**	.353
	R	-	.388	.546	.391	.304	.343	.290	.255	.360
Subspace-BERTScore	F	-	**.411**	**.557**	**.403**	**.309**	**.358**	**.303**	**.264**	**.372**
	P	-	.382	**.548**	**.393**	**.290**	**.352**	**.294**	**.248**	**.358**
	R	-	**.391**	**.547**	**.392**	**.313**	**.358**	**.292**	**.259**	**.365**
BERTScore	F	IDF	.408	.550	.395	.293	.346	.296	.260	.364
	P	IDF	.391	.540	.387	.280	.334	.284	**.252**	.353
	R	IDF	.386	.548	.394	.305	.338	.295	.252	.360
Subspace-BERTScore	F	IDF	**.413**	**.553**	**.403**	**.300**	**.360**	**.302**	**.268**	**.371**
	P	IDF	**.402**	**.541**	**.392**	**.285**	**.345**	**.285**	.250	**.357**
	R	IDF	**.391**	**.549**	**.396**	**.312**	**.356**	**.285**	**.258**	**.364**

表 5 各 STS タスクにおける人手評価との相関係数（Spearman's ρ）

手法	メトリック	重み付け	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
BERTScore	F	-	.312	.546	.450	.602	.636	.446	.553	.506
	P	-	.261	.532	.462	.576	.622	.443	.559	.494
	R	-	.350	.527	.416	.602	.623	.430	.522	.496
Subspace-BERTScore	F	-	**.335**	**.573**	**.476**	**.610**	**.650**	**.479**	**.562**	**.526**
	P	-	**.282**	**.550**	**.488**	**.580**	**.630**	**.475**	**.568**	**.511**
	R	-	**.369**	**.552**	**.436**	**.611**	**.639**	**.462**	**.530**	**.514**
BERTScore	F	L2	.321	.540	.452	.613	.640	.454	.558	.511
	P	L2	.274	.529	.468	.589	.627	.450	.565	.500
	R	L2	.348	.520	.414	.610	.624	.437	.524	.497
Subspace-BERTScore	F	L2	**.342**	**.568**	**.477**	**.621**	**.653**	**.486**	**.568**	**.531**
	P	L2	**.292**	**.547**	**.492**	**.592**	**.634**	**.479**	**.574**	**.516**
	R	L2	**.367**	**.544**	**.434**	**.620**	**.640**	**.468**	**.532**	**.515**

表 6 指示関数（$\mathbb{1}_{\text{vectors}}$: BERTScore, $\mathbb{1}_{\text{subspace}}$: 提案法）の比較。両者は文 B に対する単語 *genius* の包含度合いを計算する。

Word (a)		Sentence (B)	$\mathbb{1}_{\text{vectors}}(\boldsymbol{a}, B)$	$\mathbb{1}_{\text{subspace}}(\boldsymbol{a}, \mathbb{S}_{\mathbf{B}})$
genius	非明示的	He's very skilled at playing music.	0.69	0.76
	明示的	He's a very talented musician.	0.78	0.82
	指示関数の値の差		0.09	**0.05**