E-044

キャプション自動生成における BERTScore の有効性について
The Effectiveness of BERTScore in Automatic Caption Generation

対馬陣 †
Jin Tsushima
松原雅文 ‡
Masafumi Matsuhara

1. はじめに

近年，深層学習を用いた動画におけるキャプション生成技術が向上しており，人間の動作を日本語で説明することが可能になってきている．

機械翻訳やキャプション生成など，自然言語生成における自動評価では，意味的な類似性を評価することが求められる．そのため，生成された候補文と正解となる文章を比較することが必要となる．一般的に用いられる指標として，ROUGE や BLEU，CIDEr などが存在している．しかし，これらの手法は文章の表面的な類似性に依存しており，異音同義表現の多い日本語文章にはあまり適していないと考えられる．

近年では，自然言語生成における自動評価指標として，BERTScore が提案されている [1]．BERTScore では，事前学習された BERT から得られるベクトル表現を利用して，文章間の類似度を評価しており，自然言語生成における既存の評価手法と比べて，人手評価値との相関が高いことが示されている．これについて，英語や中国語などの言語での有効性は示されているが，日本語文での有効性は示されていない．

本研究では，深層学習を用いてキャプション生成モデルを構築し，日常的な動作を記録した動画を用いて学習を行っている [2]．本稿では，日本語学習済みモデルを適用して BERTScore を算出し，提案手法で生成したキャプションにおいて，BERTScore による日本語文章評価が有効であるかどうかを検証する．

2. 関連研究

自然言語生成における自動評価指標として，BERTScore が提案されている [1]．この研究では，事前学習された BERT から得られる，文脈上の埋め込みと生成文の各トークンと正解文の各トークンを利用した自動評価指標を提案している．また，363 種類の機械翻訳と画像キャプション生成システムから得られる出力を用いて評価を行い，BERTScore が人手評価値との相関が高いことから，評価指標として有効であることが示されている．しかし，日本語においては評価されておらず，日本語文章評価における有効性は示されていない．

人間の動作を認識し，日本語で説明するための研究が存在する [3]．この研究では，人間の動作を認識・説明するための日本語キャプションデータセットを構築し，構築されたデータセットが人間の動作を認識・説明するのに有効であることが示されている．

本研究では，このデータセットを学習データとして，キャプション生成モデルに人間の動作を学習させることで，日常的な動作をキャプションとして出力している．本稿では，提案手法で生成される日本語文章に対して，BERTScore による文章評価を行い，BERTScore による日本語文章評価の有効性を検証する．

†岩手県立大学ソフトウェア情報学研究科
‡岩手県立大学ソフトウェア情報学部

3. 提案手法

3.1. キャプション生成モデルの構築

動画データもテキストデータも時系列データであるため，系列データを別の系列データに変換することができる Seq2Seq をベースとして，モデルを構築する．また，本手法では，Seq2Seq のみでは捉えることのできない特徴を学習させるため，Seq2Seq に Attention 機構を付与したモデルを用いて，キャプション生成モデルを構築する．また，Seq2Seq を構成する Recurrent Neural Network(RNN) には，RNN の一種である Long Short Term Memory ネットワークを用いる．

3.2. キャプション生成モデルの学習

学習データには，動画とその内容を説明するキャプションを用いる．また，学習前に，各データを学習に適した形に変換する．

動画データについては，各フレームの特徴量を学習済み CNN モデルを用いて抽出する．各動画のキャプションについては，付与されているキャプションが，どこ・誰・動作の要素に分かれているため，まず，各要素の間に「で」と「が」を補完し，文章として成立させる．補完したキャプションに対して分かち書きを行い，文頭と文末の情報をそれぞれ<sos>，<eos>として付与する．その後，学習に用いるキャプションの語彙数に基づき数値に変換する．

抽出された各フレームの特徴量と，数値に変換されたキャプションを入力としてモデルの学習を行う．

3.3. キャプション生成

学習後のモデルに対して，変換した動画と文頭情報<sos>を入力することで，動画の特徴量をもとに文の最初に出現する単語を予測する．その後，予測された単語の次に出現する単語の予測を再帰的に行い，生成された単語列を入力動画に対するキャプションとして出力する．

本手法では，キャプション生成に Beam Search を使用する．Beam Search を用いたキャプション生成では，単語予測の各ステップにおいて，そこまでの対数尤度が高い単語列を K 個保持しながら単語を選択する．長い系列を見渡して尤度を評価することで，より適切な文を生成することが可能となる．

4. 実験

4.1. 概要

提案手法によって生成されたキャプション (以下，生成キャプション) において，BERTScore による日本語文章評価が有効であるかどうかを検証する．キャプション生成モデルには，STAIR Actions キャプションデータセット [3] を用いて学習を行ったものを用いる．

日本語文章には，異音同義表現が多く存在しており，同じ光景に対する説明文においても，異なる表現が用いられる可能性がある．また，上記データセットの動画には，１つの動画に対して正解キャプションが 5 つ付与されている．そのため，BERTScore と一般的な指

標として用いられている BLEU のそれぞれに対して，正解キャプション同士のスコアを総当たりに算出する．対象となるキャプションが 5 つであるため，10 通りのキャプションの組み合わせに対してスコアを算出し，正解キャプションにおけるスコアリングのぶれを確認する．スコアリングのぶれについては，10 通りのスコアの四分位数における，四分位範囲で算出する．その後，5 つの正解キャプションに対する生成キャプションのスコアをそれぞれ算出し，平均値を求め，その結果について比較・評価を行う．

4.2. 実験条件

学習データとして，データセット内の 100 種類ある各動作から，動画を 100 本ずつ，合計 10,000 本使用した．データセットの各動画には 5 つのキャプションが付与されている．そのため，今回は学習テキストとして，合計 50,000 文のテキストを使用した．また，学習の際に用いる CNN モデルには，VGG16 モデルを使用し，Beam Search における K は 5 とした．

評価用データとして，データセットから学習に使用していない動画を任意で選択した．

BERTScore を算出する際に用いる事前学習済みモデルについては，東北大学の乾・鈴木研究室が開発した日本語学習済みモデル[1]を使用する．

4.3. 実験結果と考察

正解キャプションにおける各スコアリング結果のぶれを図 1 に，使用したキャプションの一部を表 1 に記す．また，図 1 における，箱ひげ図と重なっている点線は，正解キャプションに対する生成キャプションのスコアリング結果の平均値を表している．

表 1 より，正解キャプション同士は意味的に類似しているが，表現にはばらつきが見られることが分かる．図 1 より，表面的な類似性に依存する BLUE ではスコアの四分位範囲が大きいのに対し，BERTScore では四分位範囲が小さい．そのため，BERTScore は，一般的に用いられる指標よりも日本語の異音同義表現における，スコアのぶれが少ないといえる．

表 1 より，生成キャプションが正解キャプションと意味的に類似していることが分かる．図 1 より，BERTScore は，正解キャプションに対する生成キャプションのスコアリング結果の平均値が四分位範囲内に収まっていることが分かる．BLEU は，正解キャプションに対する生成キャプションのスコアリング結果の平均値が四分位範囲内に収まっていない場合もある．そのため，BERTScore は表面的な類似性に依存せず，一般的に用いられる指標よりも文章の意味が考慮された評価指標であるといえる．これらのことから，BERTScore による日本語文章評価が有効であることが示唆された．

表 1: 使用キャプションとそのカテゴリ (一部のみ表記)

動画カテゴリ：lying_on_floor, 床に横たわっている
正解キャプション
部屋でチェックのシャツを着た男性が床にあおむけで寝ている 白い壁の部屋でチェック柄のシャツを着た女性が仰向けになっている カーペットの敷かれた白い壁の部屋で 青いチェックのシャツを着た男性が床に仰向けに寝ている 金属のラックがある部屋で黒い長ズボンをはいた男性が仰向けになっている 白い壁の室内で青色の服を着ている男性が床に寝転がっている
生成キャプション
白い壁の部屋で黒い服を着た男の子が寝そべっている

[1]https://huggingface.co/cl-tohoku/bert-base-japanese

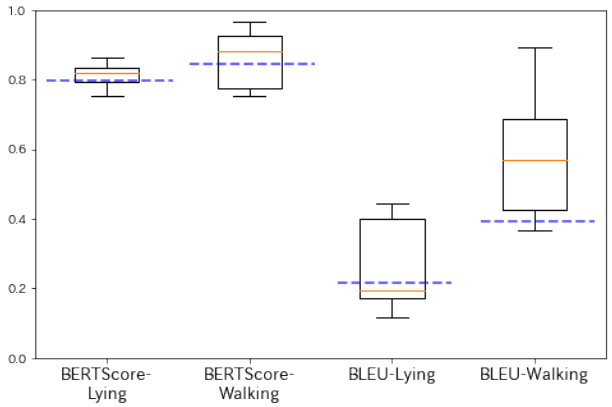

図 1: 各スコアリング結果のぶれ

5. おわりに

本稿では，日常動作を学習させたキャプション生成モデルから生成される日本語文章に対して，BERTScore による文章評価を行い，BERTScore による日本語文章評価の有効性について述べた．日本語文章には，異音同義表現が多く存在しているため，BERTScore と BLEU を用いて，正解キャプション同士のスコアと，正解キャプションに対する生成キャプションのスコアを総当たりに算出し，比較・評価実験を行った．その結果，BERTScore は，日本語の異音同義表現におけるスコアのぶれが少なく，生成キャプションにおけるスコアが四分位範囲内に収まっていたことから，日本語文章評価において有効であることが示唆された．

本稿での実験は，5 つの正解キャプションに対する評価であったため，今後は，同一カテゴリにおけるすべての動画の正解キャプションに対して，BERTScore を算出し，結果を確認する予定である．また，キャプション生成精度向上のため，提案手法に BERTScore を適用することによる，意味を考慮した生成方法についての検討も行う予定である．

謝辞

本研究の一部は JSPS 科研費 21K12611 の助成を受けたものである．

参考文献

[1] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi “BERTScore: Evaluating Text Generation with BERT” International Conference on Learning Representations 2020 (ICLR 2020)

[2] 対馬　陣, 松原雅文, 見守りカメラを用いた異常状態に対するキャプション自動生成手法の提案, 情報処理学会 第 84 回全国大会, 4W-08, pp. 819-820, 2022.

[3] 重藤優太郎, 吉川友也, 蘭佳慶, 竹内彰一, 人間の動作を日本語で説明するためのキャプションデータセット, 言語処理学会第 25 回年次大会, pp. 1173-1176, 2019.