写真を用いた英文記述問題の低コストな自動採点方法の検討

古屋 昭拓† 永田 亮†† James Tomko†
† 甲南大学 †† 甲南大学知能情報学部
E-mail: †{s1771095,f1970001}@ s.konan-u.ac.jp., ††nagata-nlp2020@ ml.hyogo-u.ac.jp.

1. はじめに

記述式問題の採点は時間とコストを要するため，自動採点の研究が盛んに行われている．特に，英語を対象にした研究が盛んである．例えば，日本語文を英文に翻訳する和文英訳問題の自動採点 [6] や自由記述を対象にしたエッセイライティングの自動採点 [1], [2] など多くの研究が存在する．

本稿では，研究例の少ない写真描写問題の自動採点を対象にする．詳細は，2. で説明するが，写真描写問題とは与えられた写真の内容を描写する英文を作成する記述式問題である(注 1)．写真描写問題では，重要なライティング能力の一つである風景や状況を適切に描写するという能力を直接問えるという利点がある（和文英訳やエッセイライティングでは直接問いにくい）．また，写真という制約があるものの記述内容の自由度は比較的高く，使用する語句と表現の多様性という面からもよい点がある．

写真描写問題の採点は大きく二つの観点から行われる．一点目は，写真の内容である．すなわち，写真の内容に即した文を作成しなければならない．二点目は，英文としての正しさである．文の理解を妨げるような誤りは減点対象となる．以上をまとめると，写真描写問題の採点は，写真の内容，英文の内容，および英文の文法的正しさを考慮して行う必要がある．

恐らく，オーソドックスな自動採点方法は，問題ごとに訓練データを作成し，専用の採点器を作成する方法であろう．すなわち，写真と解答を入力すると採点結果（スコア）を出力するような採点器である．しかしながら，この方法は，訓練データの作成に時間とコストを要する．問題ごとに，写真と模範解答が異なるため，毎回，解答を収集し，人手で採点しなければならない．理想的には，問題ごとに，そのような訓練データを作成せずとも自動採点ができることが好ましい．言い換えれば，できるだけ低コストな自動採点の枠組みの実現が望まれる．

そこで，本稿では，低コストな写真描写問題の自動採点方法の検討を行う．基本的なアイデアは，既存の画像キャプション生成モデルを写真描写問題の自動採点に流用するというものである．画像キャプション生成では，ニューラル言語モデルを用いて文生成を行うことが多い．言語モデルを用いる場合，入力画像に対して，任意の文の生成確率を推定することができる．画像（写真）の内容に即した文であれば生成確率は高くなり，逆に画像と関連が低い文であれば生成確率は低くなる．また，言語モデルの性質により，誤りが多く含まれるなど不自然な文の生成確率は低くなる．このように，生成確率の大小は，写真描写問題の採点に深く関連することが期待できる．

本稿では，パイロットスタディとして，画像キャプション生成手法 [4] の写真描写問題の自動採点への流用可能性を調査する．具体的には，写真描写問題 6 問を対象にして，同手法から得られる文の生成確率と写真描写問題のスコアの関係を調査した結果を報告する．また，その結果に基づき，低コストで高性能な自動採点を実現するために必要となる技術と残された課題を考察する．

2. 写真描写問題

写真描写問題は，与えられた写真の内容に合う英文を書く問題である．様々なバリエーションが考えられるが，例えば TOEIC(注 2)のライティングテストでは，与えられた二つの語句を使用して，一文で解答することが求められる．TOEIC ライティングテストの公式ガイド [3] によると，同テストでは，(1) 適切な語句を用いること，(2) 写真に基づいた文であること，(3) 文法的に正しい一文であることが模範解答の条件として挙げられている．一方で，意味が正しく伝わる限り，綴りや句読点の誤りは評価に影響を与えないとされている．具体的な採点スケールは 4 段階であり，各スコアは表 1 のように定められている．

本研究では，この問題形式と採点方法を参考にする．異なる点は次のとおりである．表 1 からわかるように，スコア 0 点は単純な規則で判定可能であるので本研究では対象外とする．したがって，以降では，スコアが 1〜3 点の三段階評価を考える．また，「二つの語句を含むかどうか」と「一文で書かれているかどうか」は採点に大きな影響を与えるが，これも単純な規則で判定可能であるため本研究では考慮の対象外とする．以降，このような設定の写真描写問題を考える．

(注 1)：写真の内容を口述する問題形式も考えられるが，本稿では，記述式問題のみを考える．

(注 2)：https://www.iibc-global.org/toeic.html

表 1： 写真描写問題の採点スケール（公式ガイド [3] より抜粋）.

採点スケール	採点ポイント
3	以下の特徴を持つ 1 文で構成されている ・文法的誤りがない ・与えられた 2 つの語（句）を適切に使っている ・写真と関連する内容が記述されている
2	以下の特徴を持つ 1 文もしくは複数以上の文で構成されている ・文の理解を妨げない程度の文法的誤りが一箇所以上ある ・与えられた 2 つの語（句）を使っている。ただし、1 つの文中でなかったり、語形が正確でない ・写真と関連する内容が記述されている
1	以下の特徴のいずれかを示している ・文の理解を妨げる誤りがある ・与えられた 2 つの語（句）の片方、もしくは両方ともを使っていない ・写真と記述内容の関連性がない
0	無解答。英語以外の言語で書かれている。英文で使われることのない記号が使用されている

3. パイロットスタディ

3.1 使用データ

パイロットスタディで使用するデータとして，公式ガイド [3] に掲載されている写真描写問題の模擬問題 6 問を使用した．スコア 3 点と 2 点のものについては，解答例として収録されているものを使用した．スコア 1 点のものについては，与えられた語句が使用されていない，または二文以上で書かれているなど本研究で対象外としている例しか収録されていないため使用しなかった．代わりに，当該の問題以外の問題で，スコアが 3 点または 2 点である解答文をランダムに一つ選択し，スコア 1 点の例として使用した．このデータセットを以降，問題集データセットと呼ぶことにする．表 2 に，問題集データの統計量を示す．

更に，実際の英語学習者の解答も用いた．大学生 11 人に上述の模擬問題 6 問を解いてもらった．辞書は使わず，コンピュータを用いて記述してもらった．記述時間は 30 分以内とした．この結果を第二著書が採点した．以降，このデータセットを学習者データセットと呼ぶことにする．表 3 に，学習者データセットの統計量を示す (注 3)．

表 2： 公式問題集から得たデータの解答数.

問題番号	スコア 1	スコア 2	スコア 3	合計
0	5	3	8	16
1	5	3	8	16
2	5	3	8	16
3	5	3	8	16
4	5	3	7	15
5	5	3	8	16
合計	30	18	47	95

表 3： 学習者から得たデータの解答数.

問題番号	スコア 1	スコア 2	スコア 3	合計
0	7	4	0	11
1	1	8	2	11
2	0	10	1	11
3	5	5	0	10
4	4	6	1	11
5	9	1	1	11
合計	26	34	5	65

3.2 方　　法

1. で述べたように，画像キャプション生成手法 [4] を流用した．この手法は，CNN でエンコードした画像情報を，注意機構を用いて参照し，LSTM ベースの言語モデルでキャプションを生成する．

通常の画像キャプション問題であれば出力は自然言語の文であるが，自動採点では，採点対象の文に与えられるスコアを推定する．画像キャプション生成手法を写真描写問題の採点に流用するためには，文を生成するのではなくスコアに対応する数値を推定するように出力を変更しなければならない．幸いなことに，手法 [4] のようにデコーダが言語モデルであるような手法であれば，画像が与えられたときの文の生成確率を任意の文に対して推定することができる．この生成確率は，文が画像の内容に合致していれば高くなり，逆に，画像に無関係であれば低くなるはずである．

以上の考え方に基づき，上述のデータセットに対する生成確率を推定した．推定には，訓練済みの画像キャプション生成モデル (注 4) を用いた．解答文は，全て小文字に変換し，トークンに分割した．また，カンマとピリオドは削除した．解答文の長さからの影響を低減するために，生成確率の幾何平均をとった．更に，問題ごとに生成確率の最大値で除することで，生成確率を正規化した．以上の手順で，問題集データセットと学習者データセットそれぞれについて，正規化した生成確率のヒストグラムを作成した．なお，上述の画像

(注 3)：問題番号 3 番の問題については，解答なしの答案があったため解答数合計が他より一つ少ない．

(注 4)：https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

キャプション生成モデルで，各問題の写真に対して，キャプションを生成したところ，3 枚については関連したキャプションが生成できた．残り 3 枚については関連が低いキャプションを生成した．後者では，写真をうまく認識できていない可能性が高いため，以降，両者を分けて議論する．

3.3 結　　果

図 1 と図 2 に，問題集データセットの解答文に対するヒストグラムを示す．前者がキャプション生成に成功した問題，後者がキャプション生成に失敗した問題に対応する．

両図における大きな傾向として，スコア 1 点の解答の大部分は，生成確率が低い区間に集中することがわかる．また，キャプション生成に成功した問題では，生成確率 0.4 より大きい区間には，スコア 1 点の解答はないこともわかる．以上のことから，キャプション生成に成功する写真については，生成確率により，写真と関連がない解答をある程度の精度で排除できるといえる．

一方で，スコア 2 点と 3 点の解答については，一部，生成確率が高い区間に分類されるものもあるが，区間 0〜0.3 に大部分が位置している．これらの区間の解答については，スコアがうまく弁別できていないことになる．特に，キャプション生成に失敗した写真ではこの傾向が強い．したがって，正しくキャプションが生成できる写真を問題に用いることが重要であることがわかる．幸い，正しくキャプションが生成できるかどうかは事前（問題作成前）に確認できるため，大きな問題にはなりにくい．

図 3 と図 4 に，学習者データセットの解答文に対するヒストグラムを示す．以前と同様に，前者がキャプション生成に成功した問題，後者がキャプション生成に失敗した問題に対応する．

両図より，学習者データセットでは，スコアの弁別がより難しくなっていることがわかる．スコア 3 点については（そもそも数が極端に少ないが），ある程度高い生成確率が得られているが，スコア 1 点と 2 点については，生成確率が低いものも高いものもあり，全体的にうまく弁別できていない．特に，スコア 2 点の解答文の多くが区間 0〜0.1 に位置していることは重大である．この理由については，次節で考察する．

4. 考　　察

前節の結果より，生成確率により，スコアの弁別ができる部分はあるが，スコア 1 点と 2 点の部分については課題が残ることを確認した．特に，生成確率が低い区間では弁別がうまくいかないことを確認した．

弁別がうまくいかない理由の一つに，綴り誤りを挙げることができる．学習者データセットは多数の綴り誤りを含む．また，問題集データセットも綴り誤りを含む．今回使用した画像キャプション生成モデルでは，綴り誤りがある単語は，未知語として扱われる．一方で，使用した採点基準は，「意味

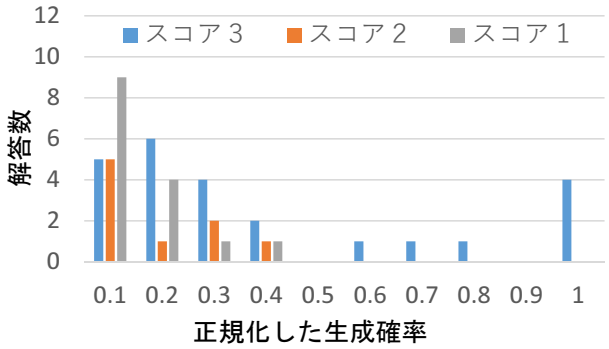

図 1： 問題集データセットのヒストグラム（キャプション生成に成功した問題）．

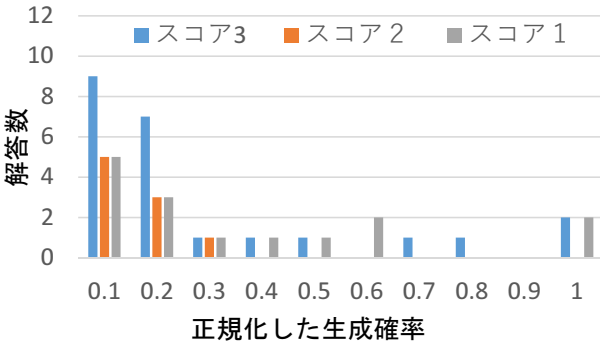

図 2： 問題集データセットのヒストグラム（キャプション生成に失敗した問題）．

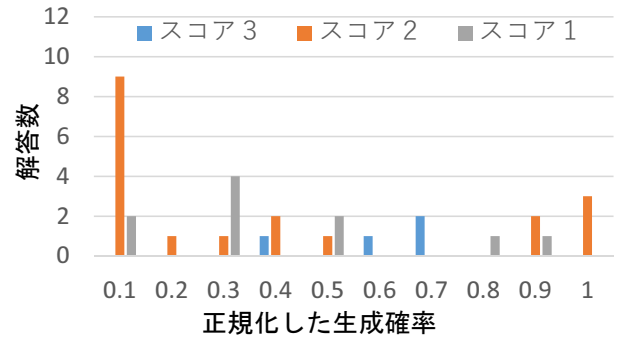

図 3： 学習者データセットのヒストグラム（キャプション生成に成功した問題）．

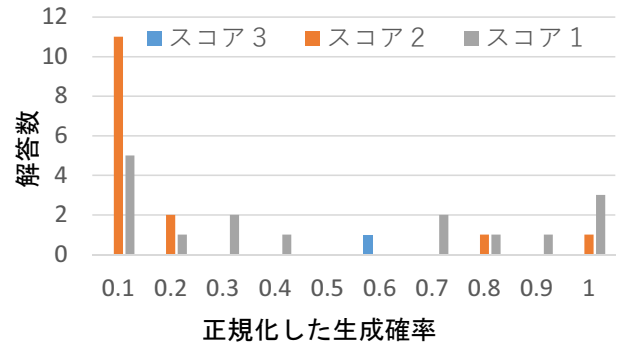

図 4： 学習者データセットのヒストグラム（キャプション生成に失敗した問題）．

が正しく伝わる限り，綴りや句読点の誤りは評価に影響を与えない」と規定している．したがって，綴り誤りのある単語が出現しても，意味が正しく伝わる限り，正しい綴りの単語と認識して採点が行われる．実際，問題集データセットでも tennice や tenis などの綴り誤りが出現したが，採点には影響を与えていなかった(注 5)．この部分で，画像キャプション生成モデルと採点基準に不一致がある．実際，図 5 に示すように，綴り誤りを修正して，学習者データセットの生成確率のヒストグラムを求めると，スコア 1 点の解答については，うまく弁別できることがわかる．

正しい綴りの単語でも未知語となることがあり，綴り誤りと同じ問題を引き起こす．例えば，stroller や forgets などの単語は語彙に存在せず，未知語扱いとなっていた．また，過去形の単語や現在分詞も未知語扱いとなる傾向がみられた．これは，画像キャプションでは，過去形や進行形が使用されることが少ないことに起因する．

綴り誤りを含む未知語の問題は，未知語の分散表現を得る手法を利用して緩和できる．例えば，文献 [5] などを用いて未知語に対する分散表現を得て，(画像キャプション生成モデル内の) 既知語の分散表現と合わせて，画像キャプション生成の言語モデルへの入力とすることで，実質，未知語のない状態で，解答文の生成確率を計算できる．

別の問題として，軽微な文法誤りがある．例えば，問題集データセットでは，“near a building” とすべきところを “at near building” とした誤り（前置詞の余剰と冠詞の抜け）は，軽微な文法誤り（文の理解を妨げない程度の文法的誤りが一箇所以上ある）と判定され，スコア 2 点と採点されていた．一方で，英語としては不自然であるので，言語モデルは生成確率を低く見積もる傾向にある．これも，画像キャプション生成モデルと採点基準の不一致となる．この不一致のため，スコア 1 点と 2 点の弁別がうまくいかないと分析できる．軽微な文法誤りについては，文法誤り訂正技術を用いて予め修正してから自動採点を行うという対応策が考えられる．

以上のように改善すべき点はいくつもあるが，画像キャプション生成手法の自動採点への流用には，学習者へのフィードバックという面でよい点もある．言語モデルから得られる各単語の生成確率を利用することで，誤っている箇所を同定できる可能性がある．また，改善のための提案として生成確率が高い単語列を学習者に与えることも考えられる．更に，注意機構から得られる重みを利用することで，写真のどの部分をうまく描写できていないか，どの単語が写真の描写として適切でないかを認識できる可能性がある．それらの情報を用いて，学習者に対話的にフィードバックを与えることも考えられる．

(注 5)：今回使用した公式ガイドには，採点例と採点理由が記載されている．

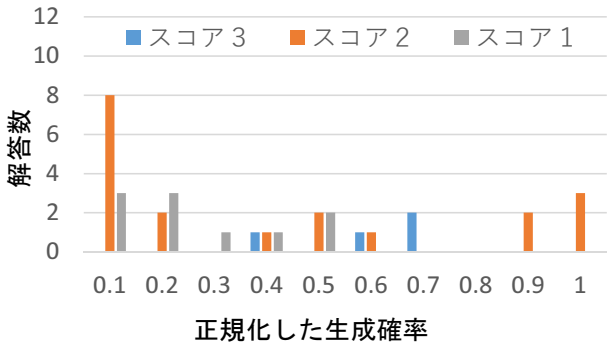

図 5： 綴り誤りを訂正した学習者データセットのヒストグラム（キャプション生成に成功した問題）．

5. おわりに

本稿では，低コストな写真描写問題の自動採点の実現を目指して，画像キャプション生成手法の自動採点への流用についてパイロットスタディを行った．具体的には，写真描写問題 6 問を対象にして，画像キャプション生成手法から得られる文の生成確率と写真描写問題のスコアの関係を調査した．その結果，生成確率でスコアを弁別できる部分もあるが，高精度の自動採点を実現するためには，さらなる改善が必要なことを確認した．特に，綴り誤りを含む未知語と軽微な文法誤りの認識について改善が必要なことを示した．未知語については，分散表現の作成を工夫することで改善できる可能性があることを示した．また，画像キャプション生成を利用した学習者へのフィードバック情報の生成についても述べた．

今後は，データセットの拡充を行い，調査の範囲を拡大する予定である．現在，独自の写真描写問題 75 問に対して，40 人分の解答を収集中である．このデータを用いて，低コストで高性能な自動採点技術を開発する予定である．

参考文献

[1] D. Alikaniotis, H. Yannakoudakis, and M. Rei, “Automatic text scoring using neural networks,” Proc. of 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.715–725, 2016.

[2] Y. Attali and J. Burstein, “Automated essay scoring with E-rater v.2.0,” The Journal of Technology, Learning, and Assessment, vol.4, no.3, pp.3–30, 2006.

[3] Educational Testing Service, TOEIC ®スピーキングテスト/ライティングテスト公式ガイド, 国際ビジネスコミュニケーション協会, 2010.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” Proc. of 32nd International Conference on Machine Learning, pp.2048–2057, 2015.

[5] 五十川 真生，梶原 智之，荒瀬 由紀，“大域的な類似度と部分文字列を用いた未知語分散表現の生成手法，” 言語処理学会第 25 回年次大会発表論文集, pp.1049–1052, 2019.

[6] 西村 則久，明関 賢太郎，安村 通晃，“英作文における自動添削システムの構築と評価，” 情報処理学会論文誌, vol.40, no.12, pp.4388–4395, 1999.