DEIM2022 F43-2

自動生成された商品説明文評価手法の提案

武内　里紗†　　福本　健二†　　寺田　浩之††　　馬頭　正文†††　　灘本　明代†

† 甲南大学知能情報学部　〒 658-8501 兵庫県神戸市東灘区岡本 8-9-1
†† おちゃのこネット株式会社　〒 651-0096 兵庫県神戸市中央区雲井通 7-1-1
††† 株式会社コンタクト　〒 651-0096 兵庫県神戸市中央区雲井通 7-1-1
E-mail: †{s1871056,m2124005}@s.konan-u.ac.jp, ††terada@ocnk.net, †††bato@contact.co.jp,
††††nadamoto@konan-u.ac.jp

あらまし　我々は深層学習を用いた商品説明文の自動生成の研究を行っている．深層学習によって生成された文の評価方法として一般には BLEU と ROUGE が知られている．しかしながら，これらの評価方法では購買意欲が湧く商品説明文を評価することが困難である．そこで本研究では，自動生成された購買意欲が湧く商品説明文評価手法として商品説明文評価軸を提案する．さらに，その商品説明文評価軸を用いて自動生成された商品説明文の評価を行う．具体的には，商品説明文評価軸として破綻軸，一貫軸，正確軸，魅力軸，丁寧軸からなる 5 つの商品説明文評価軸の提案を行う．そしてこれらの商品説明文評価軸を用いて，これまで我々の提案してきた LSTM，GPT-2，GPT-2 りんなモデルの 3 種類の深層学習を用いた商品説明文生成手法の評価を行う．

キーワード　商品説明文，深層学習，テキスト生成評価

1　は じ め に

近年の EC サイトの利用者は増加の傾向にある[1]．特にコロナ禍の状況により EC サイトの利用率も増加している．また，EC サイトには企業が出品するだけでなく，メルカリ[2]やヤフオク[3]に代表されるように，一般のユーザが気軽に出品することが可能なフリーマーケット型の EC サイトも普及している．このフリーマーケット型の CtoC 市場が拡大しており，一般ユーザでも気軽にインターネット上に商品を出品する機会が増えている．

一般に商品をインターネット上に出品する際には，商品の名称や画像，サイズなどの商品の基本的な情報のみならず，商品の購入を促すような商品を説明する文が必要になる．本研究では，商品を説明する文を商品説明文と呼ぶ．EC サイトの商品出品ページにおける商品の画像には商品の色や大まかな外見を含めたサイズ感などの視覚的な情報を EC サイトの利用者に伝えられる．しかしながら，実際に触れた時の材質の質感や実寸を行った具体的な数値としての寸法や，出品者の感じる商品の使用感等の視覚的な情報だけでは説明できない特徴を，商品説明文では記載することができる．そのため，商品説明文により消費者の購入を促す効果が期待されている．

しかしながら，一般の出品者にとって商品説明文を書くことは容易ではない．なぜなら，商品の種類名や実寸値，色合い，使用材質などの，購入を決める際には必要不可欠と言っていいような基本情報や状態を正確に伝えなければならないからである．また，出品している商品が購入を考えている人にとって，魅力的にかつ購入へと気持ちを傾けさせるように購入者に印象付けなければならない．つまり商品説明文は購入者に対して，自らの商品の購買意欲を湧かせる必要がある．そのため一般の出品者にとってこのような購買意欲を湧かせる商品説明文を書くことは容易ではない．そこで，我々は商品説明文の自動生成の研究を行っている [1], [2].

一方，商品説明文はその商品を見ている人（以下，閲覧ユーザ）に購買意欲を湧かせることが重要である．しかしながら，自動生成された商品説明文が常に閲覧ユーザに購買意欲を湧かせる文であるとは限らない．そこで我々は，自動生成された商品説明文が閲覧ユーザに購買意欲を湧かせる文であるかどうかの評価を行う必要があると考えた．一般に，深層学習により自動生成された文章を評価する指標として，BLEU [3] と ROUGE [4] があげられる．しかしながら，これらの指標は正解データとの比較で評価されるため，生成された文章により閲覧ユーザが購買意欲の湧く商品説明文かどうかを評価することは困難である．

そこで本研究では，閲覧ユーザの購買意欲を湧かせる商品説明文自動生成のための評価手法の提案を行う．具体的には，まず商品説明文をどのような視点で評価するかを決める評価項目を決定する．そしてその評価項目を類型化し，商品説明文の評価をするための評価軸を決定する．この評価軸を商品説明文評価軸と呼ぶ．商品説明文評価軸とは生成された商品説明文を評価するにあたり，その文章が商品の魅力を述べそして閲覧ユーザにとって購買意欲を湧かせるかどうかの判定の基準となるものである．現在様々な評価軸が提案されており [5]，どのような基準で評価するかによって結果は大きく異なる．その為に評価軸は生成された商品説明文を評価するにあたり重要であると考える．さらに，提案する商品説明文評価軸の有用性を計るために，先行研究 [1] にて自動生成した商品説明文を用いて，人手

1 : https://www.makeshop.jp/main/know-how/knowledge/ec-msarket.html
2 : https://jp.mercari.com/
3 : https://auctions.yahoo.co.jp/

による評価実験を行う．ここで用いた先行研究で自動生成した商品説明文は，Long Short-term Memory（LSTM）と GPT-2 を用いた商品説明文に加え， rinna 社[4]が提案を行った GPT-2 りんなモデル[5]の計 3 種類の文章を比較し，提案する商品説明文評価軸の有用性を示す．

以下，2．章で関連研究について述べ，3．章で商品説明文評価軸について述べ，4．章で実験とその考察について述べる．最後に 5．章でまとめと今後の課題について述べる．

2 関 連 研 究

近年深層学習の発展に伴い，様々な文章の生成手法の提案が行われている．それに伴い．自動生成された文章に対して様々な評価軸が提案されている．Chris ら [5] は，自動生成された文章を人手評価するための最適な方法を提案している．本研究で提案する商品説明文評価軸は，Chris らの提案する評価軸を参考にしている．澤井 [6] らは，自然言語生成分野における生成結果の品質評価について議論し，広告に適した評価用ツールに必要な機能を提案している． 本研究では商品説明文に特化した評価手法の提案をしているところが異なる．

Tao ら [7] は，電子商取引におけるパターン制御された商品説明文の自動生成を行っている．Tao らが生成した商品説明文の評価は完全性，一貫性，読みやすさ，魅力度を用いている．一方，本研究では魅力度の定義を詳細にし，評価指標の提案をしているところが異なる．

また，Fei ら [8] は，商品名要約のためのマルチソースポインタネットワークの提案を行っている．ここでは精度，商品名保持率，読みやすさ，情報性の評価を用いている．本研究では新たな評価指標魅力度の提案をしているところが異なる．

Yongzhen [9] らは，進化的商品説明文生成のためのユーザーのクリック行動を活用した動的な微調整のアプローチの提案を行っている．ここでの商品説明文の評価は貞操観念，多様性，有効性，人気，信頼性，流暢性，妥当性，関連性を用いている．本研究では新たな評価指標魅力度の提案をしているところが異なる．

3 商品説明文評価軸

商品説明文評価軸とは自動生成された商品説明文が閲覧ユーザにとって購買意欲を湧かせる商品説明文になっているかどうかを評価する際に，どのような視点で評価を行うかを明確にするための軸である．

商品説明文評価軸の決定方法は以下の通りである．

（1） 商品説明文評価項目候補の決定．

ここでいう評価項目とは商品説明文評価軸の候補になるものを示す．

（2） 商品説明文評価項目を検討する．

（3） 検討した商品説明文評価項目を人手により類型化する．

（4） (3) のグループ毎に商品説明文評価軸を決定する．

3.1 商品説明文評価項目

商品説明文を評価するにあたり，どのような視点で評価するかが重要である．そこで本研究は第 2 章で紹介した論文を参考に以下の 17 項目を提案する．

この商品説明文評価項目を用いて商品説明文評価軸を決定する．

完全性: データが全て揃っていて欠損や不整合がないことを保証すること．

一貫性: 最初から最後まで主張がぶれていなく同じであること．

可読性: 読み取れる度合い．

魅力度: 人の心をひきつけて夢中にさせる度合い．

忠実性: 内容を誤魔化したり省略したりせずそのまま示すこと．

多様性: 幅広く性質の異なる事が多数存在すること．

人気度: どれだけ注目を浴びているかの度合い．

信頼性: 期待される役割を果たすことのできる能力．

流暢性: ヒトに対して，情報（主に言語情報）を適切に，素早く，数多く処理及び出力させることのできる能力や特性．

適切性: 物事の目的や趣旨にぴったり合っていること．適当であり正しいこと．

関連性: 事柄と事柄の間に繋がりがあること．

正確性: 真の価値に近い性質．

読みやすさ: 容易に読み進められるか．(漢字 3 割平仮名 7 割程度)

情報性: 伝えられる内容に関する特徴・性質．

妥当性: ある検査などにおいてどれだけ的確に行えているかを表すもの．

破綻性: 文が，修復しようがないほどうまくいかなくなること．行き詰まること．

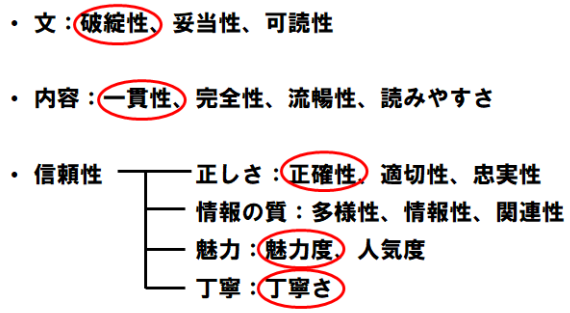
丁寧さ: 言動が礼儀正しいこと．細かいところまで気配り，配慮すること．注意深く入念すること．

3.2 商品説明文評価軸の決定

我々は列挙した評価項目の中でも類似性がいくつか見られると考え，商品説明文の評価視点である評価項目の類型化を行う．3.1 節で挙げた定義をもとに 12 人で人手による類型化を行った．類型化の結果を図 1 に示す．

図 1 より，商品説明文評価軸を以下の 6 つに決定した．

図 1 類型化の結果



4 : https://rinna.co.jp/

5 : https://prtimes.jp/main/html/rd/p/000000009.000070041.html

- 破綻軸

文に関しての意味を持つものとし，破綻性，妥当性，可読性を評価項目とする．

- 一貫軸

内容に関しての意味を持つものとし，一貫性，完全性，流暢性，読みやすさを評価項目とする．

- 正確軸

信頼性のうち，正しさに関しての意味を持つものであり，正確性，適切性，忠実性を評価項目とする．

- 多様軸

信頼性 (情報) に関しての意味を持つものであり，多様性，情報性，関連性を評価項目とする．

- 魅力軸

文に関しての意味を持つものであり魅力度，人気度を評価項目とする．

- 丁寧軸

文に関しての意味を持つものであり丁寧さを評価項目とする．

以上 6 つに分類した中で多様軸を省いた破綻軸，一貫軸，正確軸，魅力軸，丁寧軸の 5 つの軸を今回の商品説明文評価軸と設定する．多様軸を省いた理由は，人手を要することなく機械的に評価できると考えたからである．

4 実　　験

提案した商品説明文評価軸が実際に自動生成された商品説明文の評価に対して有効かどうかを計るために実際に先行研究 [1] にて自動生成した商品説明文を用いて 3 種類の実験を行った．

4.1 実 験 条 件

実験データ

実験に用いた商品説明文は LSTM，GPT-2，GPT-2 りんなモデルにより生成された商品説明文から無作為に抽出した各々40 文，合計 120 文である．表 1 に実験に用いた商品説明文の例を示す．被験者は 10 名であり，実験はクラウドソーシングを用いた．実験を行う際には，この 3 つの手法各々から無作為に 8 文ずつ抽出し，合計 24 文を 1 セットとして合計 5 セット用意した. 被験者に表示する順番は，GPT-2 りんなモデルと GPT-2 と LSTM の 3 つのモデルで生成した商品説明文を無作為に表示した．

軸毎の判定方法

実験において被験者は生成された商品説明文を読んで，軸毎に評価を行う．その軸毎の評価の判定条件を以下に示す．実験では被験者は以下の判定条件を読んでから，実験を行う．

（1） 破綻軸：

生成した商品説明文が文としておかしくないかどうかを評価する．具体的には，同一の単語を繰り返し意味が通らないものや，文が途中で途切れてしまう等の文法的に破綻しているどうかを評価する．文が破綻していれば評価は 1 とし，文が破綻していなければ評価は 5 とする．例えば「シンプルなデザインで，シンプルなデザインになっています.」は「シンプルなデザイン」が繰り返されており，この 商品説明文は破綻していると言える．

（2） 一貫軸：

生成した商品説明文の内容に矛盾が生じていないかを評価する．具体的には，商品説明文に一貫性があるかどうかを評価する．商品説明文の内容が一貫していなければ評価は 1 とし，一貫していれば評価は 5 とする．例えば,「シンプルなデザインであるため，ゴージャスなお部屋 にぴったりです.」はシンプルなデザインと前半で説明しているのに対し，ゴージャスな部屋にぴったりといったように説明文の前半と後半とで内容に矛盾を生じている．このような商品説明文は一貫性がないと言える．

（3） 正確軸：

生成した商品説明文に商品の基本情報を含んでいるかどうかを評価する．具体的には，ソファの色，種類，デザイン，座面素材，販売対象者，表面素材に関する情報が商品説明文に含まれているかどうかを評価する．商品情報が含まれていなければ評価は 1 とし，多く含まれていれば評価は 5 とする．例えば,「カラーはブラウンで，大きめのサイズであるため家族向けに人気のソファになっています.」はユーザの入力したカラーの値であるブラウンが含まれているため，正確性が高いと言える．

（4） 魅力軸:

生成された商品説明文を読んでその商品に対する購買意欲がどの程度湧いたかどうかを評価する．具体的には，商品説明文を読んだ際に商品を購入したいと感じるかの評価であり，商品説明文を読み，商品を購入したいと感じなければ評価は 1 とし，購入したいと感じれば評価は 5 とする．

（5） 丁寧軸:

生成した商品説明文の表現方法がいかに丁寧に商品を説明しているかを評価する. 具体的には似た意味，言い回し表現が含まれている，例を挙げている，根拠を明確に示している，実際に使用する場面が思い浮かぶなどの条件を満たしていれば評価を 5 とする．

各生成手法の特徴

表 1 に実験で使用したデータの一部を示す．以下に各生成手法で生成された商品説明文の特徴を述べる．

- LSTM

文の長さは短く，文末が「この商品は〜です。」のような特徴を言い切る形となっている．

- GPT-2

文の長さは中間で，LSTM と同じく文末が「この商品は〜です。」のような特徴を言い切る形となっている．

- GPT-2 りんなモデル

文の長さは長く，LSTM，GPT-2 とは異なり文末が「この商品は〜のため魅力的です。」「この商品は〜のためオススメです。」のような相手に紹介する形となっている．また，商品の項目ごとに文章を生成しているため，商品の基本情報を多く含んでいる．具体的な項目を挙げると，ソファの色，種類，デザイン，座面素材，販売対象者，表面素材が挙げられる．

4.2 実　験　1

購入したい商品が具体的に決まっていないユーザにとって，

表 1　生成手法毎の生成文例

生成手法	生成文
LSTM	背もたれの角度をつけることで体が自然にくつろげるように設計されています
	圧迫感のないくつろぎの空間を演出します
GPT-2	ソファーのように座面はポケットコイルを使用していて柔らかく沈み込む心地で、座面幅も広々としているので、ゆったりと座っていても座り続けていてもおやすみのソファーです。
	さらに、 クッション〇個が付属ですので、ご自宅にある家具との相性もよろしくないので、またソファと同色や別の商品と組み合わせて使うこともできるので、お部屋の間取りやテイストによって様々な組み合わせができます。
GPT-2 りんなモデル	ソファのお色はホワイトをご用意しているので、ナチュラルテイストのローテーブルや、ナチュラル系のテーブルとコーディネートすればおしゃれカフェ風コーディネートができます。フロアンファとなっているので、レイアウトの変更が自由に行えます。モダンデザインのソファとなっているので、北欧スタイルを中心にく pnoun¿やjpnoun¿など様々なインテリアに合わせていただけます。ソファの座面の素材には綿を使用しているので、年中快適にお使い頂けます。お子様のいるご家庭におすすめのソファので、家族みんなで仲良く、思い出を作っていけるソファです。ソファ表面の素材はファブリックなので、カバーは取り外し可能です。
	ソファのお色はホワイトをご用意しているので、 お部屋の雰囲気を軽やかに彩ってくれます。リクライニングソファとなっているので、 ごろんとくつろぐことが出来ます。モダンなデザインのソファとなっているので、シンプルなデザインなので、様々な雰囲気の家具との相性が良いですし、お部屋に配置する時の置き方により様々表情を見せるソファとなります。ソファの座面の素材には綿を使用しているので、一年を通してお使いいただけます。ファミリーにおすすめのソファなので、 是非ご検討下さい。ソファ表面の素材はファブリックなので、お手入れも楽観的です。

提案した商品説明文評価軸が有用かどうかを確認することを目的として実験 1 を行う．

実験手順

被験者に 3 つのモデルで生成した商品説明文を 1 セット 24 文を無作為に表示する．被験者は提示された各文について 5 つの商品説明文評価軸をそれぞれ 1 から 5 までの 5 段階に評価する．以上の手順を 5 回繰り返す．

結果と考察

実験 1 の結果を表 2 に示す．破綻軸の結果は LSTM が平均 3.54 と最も高い評価を得られた．理由としては， LSTM は短い 1 文からなっているため，文として破綻している箇所が少なく，破綻度が高い評価になっていると考える．

また，一貫軸も同様に LSTM が平均 4.01 と最も高い評価を得られた．理由としては LSTM で生成した商品説明文は短い 1 文でできているのに対し，GPT-2 で生成した商品説明文の方は複数の文で構成されており，1 文あたりに含まれる単語数が多い．さらに，GPT-2 で生成された商品説明文は文の前半部分と後半部分との間で主張が変化している説明文がある場合がある．これにより．一貫軸では商品説明文中の内容に矛盾が生じていれば一貫軸の評価が低くなることが分かった．以上より，破綻軸と一貫軸は共に文章が短い方が高い評価になりやすいと考える．

次に，正確軸の結果は，GPT-2 りんなモデルが平均 4.55 と最も高い評価を得られた．これは，GPT-2 りんなモデルが商品の種類，色，材質といった様々な項目ごとに説明文を生成し，それら全てを組み合わせているため商品の基本情報を豊富に含んでいることによると考える．したがって正確軸では商品説明文中での使用単語のカテゴリーの豊富さを正確に反映し，高い評価になるということが分かった．

魅力軸の結果は GPT-2 りんなモデルが平均 3.00 と 3 つの手法の中で最も高い評価を得られた．その理由は，LSTM および GPT-2 生成した商品説明文では，文末が「〜です。」や「〜である。」のような断言する形の文末が多くあった．それに対して，GPT-2 りんなモデルで生成される商品説明文の文末は，「〜オススメです。」や「〜魅力的です。」のように相手に商品を推薦する文末が多く見られた．その結果，それぞれに違いが見られたと考える．したがって魅力軸は文中や文末の伝え方を正確に反映し，高い評価になるということが分かった．

丁寧軸においては，GPT-2 りんなモデルが平均 4.03 と高い評価を得られた．この結果となった理由は GPT-2 りんなモデルが商品の種類，色，材質といった様々な項目ごとに説明文を生成し，それらの文の中には具体的な使用シーンや商品のイメージについて詳細に述べる文が多く見られた．したがってそれら全てを組み合わせたときには商品のイメージを的確に閲覧者に抱かせ，その結果丁寧な商品説明文であるという印象になったと考える．したがって丁寧軸では商品の具体的なイメージを抱いていない被験者にとって，商品説明文中での具体案を抱かせる説明となっており，高い評価になったと考える．

また，魅力度と丁寧さは関係性があり，商品購入のきっかけとなるのが文章自体の魅力度のみだけではなく文章中の表現の仕方にも影響があることが考えられる．以上より提案した商品説明文評価軸が商品の具体的なイメージを抱いていない閲覧者にとって商品説明文に適していることを確認することができた．また，提案した各々の商品説明文評価軸が各手法の特徴を捉えていたことから，提案した商品説明文評価軸は有効であると言える．

相関係数

実験 1 で得られた値から相関係数を算出した結果を表 3 に示す．破綻軸と一貫軸の相関に関しては，アンケートの質問の違いを理解できていない可能性が大きいと考える．実際に対面で予備実験を行った際にも，2 つの軸の違いが分かりにくいという意見があった．破綻軸と魅力軸，一貫軸と魅力軸にはとても強い

相関が得られた．しかし，正確軸と魅力軸の相関は 0.19 とかなり弱かった．このことから情報量が多くても購買意欲が湧かない可能性があると考えられる．

表 2　実験 1 の結果

生成手法	破綻軸	一貫軸	正確軸	魅力軸	丁寧軸
LSTM	3.54	4.01	2.16	1.77	2.32
GPT-2	2.62	3.11	2.28	1.57	2.22
GPT-2 りんなモデル	2.80	3.01	4.55	3.00	4.03

表 3　相 関 係 数

評価軸 1	評価軸 2	生成手法	相関係数
破綻軸	一貫軸	LSTM	0.80
		GPT-2	0.72
		GPT-2 りんなモデル	0.82
破綻軸	正確軸	LSTM	0.04
		GPT-2	0.15
		GPT-2 りんなモデル	0.23
破綻軸	魅力軸	LSTM	0.54
		GPT-2	0.62
		GPT-2 りんなモデル	0.80
一貫軸	正確軸	LSTM	0.08
		GPT-2	0.32
		GPT-2 りんなモデル	0.30
一貫軸	魅力軸	LSTM	0.34
		GPT-2	0.52
		GPT-2 りんなモデル	0.91
正確軸	魅力軸	LSTM	0.53
		GPT-2	0.62
		GPT-2 りんなモデル	0.19

4.3 実　験　2

被験者に対して，状況設定を行うことによって実際の購入時のイメージが湧き，購買意欲がより湧くのではないかと考え，本実験を行う．そこで，購入したい商品の詳細もしくはある程度のイメージが決まっているユーザにとって，自動生成された商品説明文に対して，提案した商品説明文評価軸の有用性を示すことを目的として実験を行う．具体的には，実験 1 に被験者に対して状況設定を提示することにより，商品説明文評価軸の値の変化を確認する．

実験条件

実験 2 では，実験 1 に加えて状況設定をしたうえで実験を行った．状況設定は,「あなたは社会人 3 年目で初めて東京で一人暮らしを始めることになりました．部屋の大きさは 1K です．それにあたり一人暮らしにぴったりな，サイズが小さめのソファを購入しようと考えています.」である．実験データは実験 1 と同様である．

実験手順

被験者には，3 つのモデルで生成した商品説明文を 1 セット 24 文を無作為に表示し, 各文について 5 つの商品説明文評価軸をそれぞれ 1 から 5 までの 5 段階に評価してもらった．以上の手順を 5 回繰り返す．

結果と考察

実験 2 の結果を表 4 に示す．破綻軸の結果は LSTM が平均 3.46 で最も高かった．実験 1 と同様，破綻軸は LSTM が短い 1 文からなっているという特徴を反映した結果高い評価になったと考える．

一貫軸でも実験 1 と同様に LSTM が平均 3.94 と最も高い評価が得られた．これは商品説明文に記述された文の数が GPT-2 で生成された商品説明文のほうが多いため，説明文の前半部分と後半部分で主張が変化している可能性を示し，その結果を一貫軸は反映していると考える．

正確軸の結果も GPT-2 りんなモデルが平均 4.46 で実験 1 と同様に最も高い評価を得られた．これは正確軸が，GPT-2 りんなモデルは商品の色や種類といった属性毎に商品説明文を生成し組み合わせているという特徴を反映した結果高い評価となったと考える．

また魅力度軸においても，GPT-2 りんなモデルが平均 2.82 と実験 1 と同様 3 つの手法の中で最も高い評価を得られた．これは魅力軸が文中や文末の表現方法を踏まえていたため高い評価となったと考える．しかし，魅力軸の値においては実験 1 よりも少し低い評価が得られた．その理由としては，状況設定を行うことによって購入時のイメージは付きやすくなったものの，状況設定に共感が持てない場合や設定から少しでも外れている文章に低い評価を付けたと考える．

丁寧軸においては，実験 1 と同様に GPT-2 りんなモデルが平均 4.03 と高い評価を得られた．このような結果となった理由としては GPT-2 りんなモデルが商品の様々な項目ごとに説明文を生成し，それらの文の中には閲覧者に寄り添った表現の文が多く見られた．その結果それらすべてを組み合わせた GPT-2 りんなモデルは商品に対する多くの想像を閲覧者に抱かせ，高い評価になったと考える．

以上より，実験 1 と比べ実験 2 では魅力軸の結果のみ変化が見られたことから，閲覧者の状況設定を行ったことによる負の干渉があったと考える．したがって，提案した商品説明文評価軸は，購入したい商品の詳細やある程度のイメージが決まっているユーザが対象であっても有効的に評価できると言える．

表 4　実験 2 の結果

生成手法	破綻軸	一貫軸	正確軸	魅力軸	丁寧軸
LSTM	3.46	3.94	2.30	1.91	2.58
GPT-2	2.75	3.24	2.34	1.72	2.46
GPT-2 りんなモデル	2.95	3.07	4.46	2.82	4.03

4.4 実　験　3

実験 1 と実験 2 では 3 つの手法で生成した商品説明文を無作為に混ぜたデータセットを用いて実験を行った．この場合，被験者は純粋に商品説明文を評価しているのではなく手法の比較を行ってしまう可能性がある．そこで，実験 3 では 3 つの手法で生成した商品説明文の評価を手法毎に評価し，我々の提案する商品説明文評価軸の有用性を示すことを目的とする．

実験条件

実験 3 では，GPT-2 りんなと GPT-2 と LSTM の結果各々を提示して実験を行った．生成された商品説明文はそれぞれのモデル毎に無作為に抽出した各 40 文である．実験を行う際には，この 3 つの手法の内 1 つから生成された商品説明文を無作為に 10 文抽出し，それを 1 セットとして 1 手法あたり 3 セット, 合計 9 セット用意した. 被験者は各文について 4 つの商品説明文評価軸をそれぞれ 1 から 5 までの 5 段階に評価する．この手順をモデル毎に 3 回, 計 9 回繰り返す.

結果と考察

実験 3 の結果を表 5 に示す．破綻軸の結果は LSTM が平均 3.88 で最も高かった．これは実験 1 と同様，破綻軸は LSTM が短い 1 文からなっているという特徴を反映した結果高い評価になったと考える．一貫軸でも実験 1 と同様に LSTM が平均 4.30 と最も高い評価が得られた．これは商品説明文に記述された文の数が GPT-2 で生成された商品説明文のほうが多いため，説明文の前半部分と後半部分で主張が変化している可能性を示し，その結果を一貫軸は反映していると考える．正確軸の結果は GPT-2 りんなモデルが平均 4.00 で実験 1 と同様に最も高い評価を得られた．これは正確軸が，GPT-2 りんなモデルは商品の色や種類といった属性毎に商品説明文を生成し組み合わせているという特徴を反映した結果高い評価となったと考える．しかし魅力度軸においては，LSTM が平均 3.06 と実験 1 とは異なる結果が得られた．理由として，生成手法を分類して評価することにより他の手法と比較することなく，その手法本来の評価を魅力軸の値として得ることができたからであると考える．

表 5　実験 3 の結果

生成手法	破綻軸	一貫軸	正確軸	魅力軸
LSTM	3.88	4.30	3.72	3.06
GPT-2	2.94	3.52	3.32	2.40
GPT-2 りんなモデル	2.62	2.70	4.00	2.48

5　まとめと今後の課題

本研究では BLEU や ROUGE では評価できない購買意欲に関して，購買意欲が沸く商品説明文評価手法として商品説明文評価軸を提案し，その商品説明文評価軸を用いて商品説明文の評価を行った．具体的には，評価の着目点である評価項目を同じ分類ごとにまとめ，破綻軸，一貫軸，正確軸，魅力軸，丁寧軸という 5 つの商品説明文評価軸を設定した．設定した商品説明文評価軸と自動生成された 3 種類の商品説明文を使用して商品説明文評価軸の有効性を複数の手法で確認した．実験結果より，商品説明文がどの程度購買意欲を沸かせているのかを定量化して評価することができた．無作為に 3 種類を並べて行った実験 1 では GPT-2 りんなモデルが最も購買意欲を沸かせた．また、購買意欲を現わす魅力度と丁寧さは関係性があり，商品購入のきっかけとなるのが文章自体の魅力度のみだけではなく文章中の表現の仕方にも影響があると考える．得られた値をもとに算出した相関係数より，各軸同士の相関を確認した．それと共に商品説明文内の情報量が多いからといって購買意欲には直結しないこと，商品の種類名や実寸値，色合い，使用材質などの基本情報であっても購入者の求めている情報とは限らないということが考えられる．さらに，評価する人手の状況を設定した実験 2 より，状況を設定したための悪影響が見られたが，実験 1 とは多少異なる結果となったため提案する商品説明文評価軸が評価するユーザのバックグラウンドが異なっても有効であることを示した．また，生成モデルごとに自動生成文を評価した実験 3 より，各自動生成文が評価に影響し合っていることを確認し，LSTM を用いた自動生成文が最も購買意欲をわかせている事が分かった．

今後の課題として多様軸についての評価を行う．また，試行回数を増やし，各実験でのばらつきを減らすとともに評価を行う人手についても条件を加えて実験を行っていきたい．さらに，丁寧さについては文章の表現の仕方について着眼し仮説を立てたが文章量についても関係があるのではないかと考えるためその点についての実験を行っていく．

謝辞

論文の一部は JSPS 科研費 19H04218，19H04221，20K12085，及び私学助成金（大学間連携研究助成金）の助成によるものである．ここに記して謹んで感謝の意を表する．

文　　献

[1] 福本健二，武内里紗，寺田浩之，馬頭正文，灘本明代，“GPT-2 を用いた商品属性データ構造に基づく家具説明文の自動生成”第 12 回ソーシャルコンピューティングシンポジウム (SoC 2021)，社団法人 電子情報通信学会，信学技報，2021.

[2] Kenji Fukumoto, Rinji Suzuki, Hiroyuki Terada, Masafumi Bato, Akiyo Nadamoto, “Comparison of Deep Learning Models for Automatic Generation of Product Description on E-commerce site” Woodstock ’18, June 03—05, 2018, Woodstock, NY, iiWAS 2021.

[3] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[4] Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries” 2004.

[5] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, Emiel Krahmer, “Best practices for the human evaluation of automatically generated text” Proceedings of The 12th International Conference on Natural Language Generation, pages 355—368, Tokyo, Japan, 28 Oct - 1 Nov, 2019.

[6] 澤井悠，張培楠，吉本暁文，“自動生成された広告文の人手評価における評価指標と支援ツールの提案”The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020.

[7] Tao Zhang, Jin Zhang, Chengfu Huo, Weijun Ren, “Automatic Generation of Pattern-controlled Product Description in E-commerce” Creative Commons CC-BY 4.0, 2019 IW3C2 (International World Wide Web Conference Committee).

[8] Fei Sun, Peng Jiang, Hanxiao Sun, Changhua Pei, Wenwu Ou, Xiaobo Wang, “Multi-Source Pointer Network for Product Title Summarization” Association for Computing Machinery, 2018.

[9] Yongzhen Wang, Jian Wang, Heng Huang, Hongsong Li,

Xiaozhong Liu, “Evolutionary Product Description Generation: A Dynamic Fine-Tuning Approach Leveraging User Click Behavior” Session 1B: Knowledge and Explainability, SIGIR ’20, July 25—30, 2020, Virtual Event, China.