LP-to-Text: マルチモーダル広告文生成

村上聡一朗 [1,2] 星野翔 [1] 張培楠 [1] 上垣外英剛 [2] 高村大也 [3] 奥村学 [2]
[1] 株式会社サイバーエージェント [2] 東京工業大学 [3] 産業技術総合研究所
{murakami_soichiro,hoshino_sho,zhang_peinan}@cyberagent.co.jp
kamigaito@lr.pi.titech.jp takamura.hiroya@aist.go.jp oku@pi.titech.ac.jp

概要

検索連動型広告では, 広告効果を高めるために, ランディングページ (LP) と広告文の内容に高い関連性が求められる. また, 広告データは, 広告主によって扱う商品数や予算が異なるため, 配信数に偏りが生じやすい. 上の課題に対し本研究では, LP のマルチモーダル情報と広告データの不均衡性を考慮した広告文生成手法を提案する. 実験では, LP のレイアウトの考慮による生成品質の向上および不均衡データに対する頑健性の向上を確認した.

1 はじめに

検索連動型広告とは, ユーザの検索クエリに関連する広告文を検索結果画面に提示する広告である. 図 1 のように, 広告文をクリックした遷移先にはランディングページ (LP) が設定されており, LP はサービスの魅力をユーザへ訴求することで購入や申込等の行動を促すことを目的としている. 従って, 広告文には, LP と関連性の高い内容をユーザへ訴求し, クリックを促す重要な役割があるため, 一般的に広告制作者は, LP 等を参考に広告文を作成する. しかし, 近年のデジタル広告の需要拡大に伴う制作者の負担増加により, 広告文作成の自動化が求められている.

本研究では LP からの広告文生成に取り組む. 広告文には様々な要件や課題がある. (1) 広告文では, 広告効果を高めるため, LP と関連した重要な情報を含む必要がある. (2) 広告文は, 広告主により扱う商品数やサービス数, 広告予算が異なるため, 配信事例数に偏りが生じやすい. さらに, テンプレートが多用されることも多く語彙の偏りが生じやすい [1]. この不均衡性は機械学習に基づく文生成で大きな障壁となる [2]. これらに対し本研究では, (1) LP の画像やレイアウト, テキストを考慮した上で, LP と関連性の高い広告文の生成手法を探求する. (2) さらに, 不均衡データに対して頑健な生成モデルを提案する.

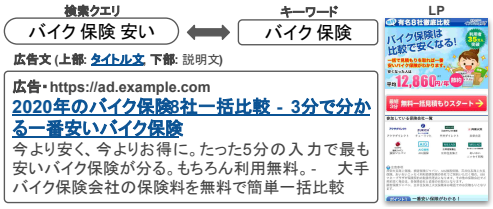

図 1 検索連動型広告および LP の例

2 関連研究

文書画像に対する言語生成 本研究では, LP を画像やテキストが配置された文書画像として捉え, タスクを文書画像からの言語生成として考える. 従って, 文書画像理解 [3, 4] や文書画像の機械読解と関連性が高い [5]. これらの研究では一般的なビジネス文書や Web ページを対象としているが, 本研究で扱う LP では, 割引等の訴求を強調するためのテキストの大きさや配置等のレイアウト情報が特徴的であり, より重要であると考えられる. そこで本研究ではレイアウト情報を考慮した生成モデルを提案する.

広告文生成 広告文生成はこれまで多くの研究があるが, 近年では, ニューラル言語生成による手法が主流となっている [6, 7, 8, 9, 10]. 特に最近では, 広告効果を報酬として強化学習を用いる手法 [11, 12], 外部知識の導入により生成文の多様性を向上する手法 [1] 等が提案されている. これらの研究では, キーワードや LP 概要文[1)], 外部知識等の言語情報を主に利用しているが, 本研究では LP の画像情報やレイアウトも活用する. これにより, LP の特徴的なデザインにより強調された割引やサービス内容等の重要な情報を明示的に考慮できるため, より LP と関連性の高い広告文生成が期待できる.

3 提案手法

検索連動型広告は, 図 1 のように, 検索クエリと広告主が予め設定したキーワードが一致した場合, 検

1) 各 LP の<meta>タグに含まれるテキスト

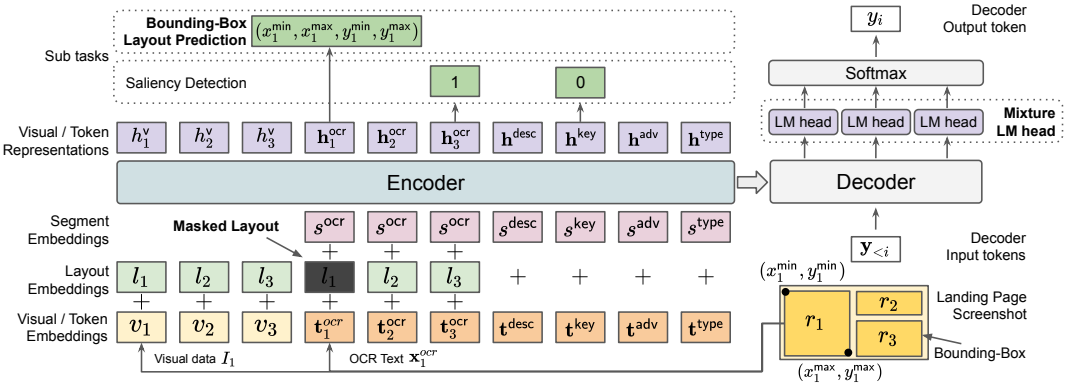

図 2　提案モデル

索結果に提示される. また, 各広告は, タイトル文と説明文の 2 種類から構成されており, それぞれの内容は LP やキーワードと関連性が高いことが求められる. 加えて, 生成モデルの運用コスト軽減の観点から, タイトル文や説明文といった広告種別や広告主ごとにモデルを作成するのではなく, これらに幅広く対応した統一的な生成モデルが必要とされる. f 本研究では, LP を画像やテキストが配置された文書画像として捉え, LP からの広告文生成を文書画像からの言語生成タスクとして考える. さらに, 前述の要件を満たすために, LP だけでなく, LP 概要文やキーワード, 広告主, 広告種別といったメタ情報も用いる. 本研究では, 先行研究 [3, 4] と同様に, 文書画像中のテキストを活用するために, LP に対して文字認識 (OCR) を行い, 検出された矩形領域 (Bounding-Box) のテキストとレイアウトを使用する.

3.1　広告文生成モデル

提案モデルの概要を図 2 に示す. 本研究では, 先行研究 [5] に倣い, 文書コーパスで事前学習された Transformer[13] を使用し, LP の画像やレイアウト, テキストを考慮した広告文生成モデルを探求する. 入力のテキスト情報として, LP 概要文 $\mathbf{x}^{\text{desc}}$, キーワード $\mathbf{x}^{\text{key}}$, 広告主名 $\mathbf{x}^{\text{adv}}$, 広告文の種別 $\mathbf{x}^{\text{type}}$, LP から OCR で取得した矩形領域集合 $\mathbf{R} = \{r_i\}_{i=1}^{|\mathbf{R}|}$ の各 OCR トークン系列 $\mathbf{x}_i^{\text{ocr}}$ を含む, 5 種類のトークン系列集合 $\mathbf{X} = \{\mathbf{x}^{\text{desc}}, \mathbf{x}^{\text{key}}, \mathbf{x}^{\text{adv}}, \mathbf{x}^{\text{type}}, \{\mathbf{x}_i^{\text{ocr}}\}_{i=1}^{|\mathbf{R}|}\}$ を用いる. ここで, 各トークン系列 $\mathbf{x}^*$ は, $\mathbf{x}^* = (x_i^*)_{i=1}^{|\mathbf{x}^*|}$ とする. さらに, 矩形領域集合 $\mathbf{R}$ に対応するレイアウト (座標情報) $\mathbf{C} = \{c_i\}_{i=1}^{|\mathbf{R}|}$ および画像情報 $\mathbf{I} = \{I_i\}_{i=1}^{|\mathbf{R}|}$ を用いる. ここで, $c_i = (x_i^{\text{min}}, x_i^{\text{max}}, y_i^{\text{min}}, y_i^{\text{max}}) \in \mathbb{R}^4$ であり, 例えば x_i^{min} は矩形領域 r_i の x 座標の最小値を表す[2]. また, 学習時には, 一般的な系列生成問題と同様, 参照文 $\mathbf{y} = (y_i)_{i=1}^{|\mathbf{y}|}$ の負の対数尤度の最小化を行う.

ここで, 図 2 の各 Embedding について説明する.

Token Embedding　各トークン系列 $\mathbf{x}^*$ は Embedding 系列 $\mathbf{t}^* \in \mathbb{R}^{D \times |\mathbf{x}^*|}$ に変換後, エンコーダへ入力する. ここで, D は Embedding の次元を表す. 例えば, 矩形領域 r_i の OCR トークン系列 $\mathbf{x}_i^{\text{ocr}} = (x_{i,j}^{\text{ocr}})_{j=1}^{|\mathbf{x}_i^{\text{ocr}}|}$ に対して, $t_{i,j}^{\text{ocr}} \in \mathbb{R}^D$ からなる系列 $\mathbf{t}_i^{\text{ocr}} \in \mathbb{R}^{D \times |\mathbf{x}_i^{\text{ocr}}|}$ を得る.

Segment Embedding　エンコーダでは, 各トークン系列 $\mathbf{x}^*$ の領域を区別する. 例えば, トークン系列 $\mathbf{x}^{\text{desc}}$ に対して, $s^{\text{desc}} \in \mathbb{R}^D$ を導入する.

Visual Embedding　LP の文字色やフォントなどの視覚情報を活用するために, 各矩形領域 r_i に対応する画像 I_i を導入する. 具体的には, 取得した画像 I_i を 128×32 (Width×Height) へリサイズし, CNN による特徴抽出により視覚特徴 $v_i \in \mathbb{R}^D$ を作成する.

Layout Embedding　LP では, 文字の配置位置や大きさは重要な要素である. 矩形領域 r_i のレイアウト c_i を MLP へ入力し, $l_i \in \mathbb{R}^D$ を獲得する.

上の Embeddings を用いて, エンコーダへの入力を作成する (図 2). 例えば, 矩形領域 r_i における j 番目の OCR トークン $x_{i,j}^{\text{ocr}}$ に対応する入力 $e_{i,j}^{\text{ocr}} \in \mathbb{R}^D$ は, 次式により作成する: $e_{i,j}^{\text{ocr}} = \text{LayerNorm}(t_{i,j}^{\text{ocr}} + s^{\text{ocr}} + l_i)$.

3.2　マルチタスク学習

LP と関連性が高い内容の広告文を生成するために, 本研究ではレイアウトや内容を考慮するためのサブタスクを導入し, マルチタスク学習を実施する.

Bounding-Box Layout Prediction　重要な情報を伝えるために, LP においてテキストの表示位置や大きさ等のレイアウト情報は重要である. しかし, 文書で事前学習されたモデルに対して, 唐突にレイアウト情報を与えるだけでは, レイアウト情報の意味を早期に獲得することは困難と予想される. そこで, 近年の Masked Language Modeling [14, 9] の成功に倣い,

2)　座標 $(x_i^{\text{min}}, x_i^{\text{max}}, y_i^{\text{min}}, y_i^{\text{max}})$ は LP の幅と高さで正規化する.

周辺情報から対象矩形領域のレイアウトを予測するサブタスクを新たに提案する. 全てのレイアウト情報は学習時および推論時に参照可能ではあるが, 意図的に対象レイアウトをマスクし, 自身の OCR テキストを含む周辺情報から対象レイアウトを予測するよう学習することで, OCR テキストと対象レイアウトの対応関係を効率的に獲得することを期待する. 具体的には, 矩形領域 r_i に対するレイアウト l_i をマスクし, エンコードされた矩形領域 r_i の表現ベクトル $\mathbf{h}_i^{\text{ocr}}$ から OCR テキスト $\mathbf{x}_i^{\text{ocr}}$ に対するレイアウト c_i の各座標 $(x_i^{\text{min}}, x_i^{\text{max}}, y_i^{\text{min}}, y_i^{\text{max}})$ を MLP で予測する. 例えば, 図 2 では, まず, 矩形領域 r_1 に対応するレイアウト l_1 を用いずに, 入力 $\mathbf{e}_1^{\text{ocr}}$ を作成する. 次に, エンコーダより獲得した $\mathbf{h}_1^{\text{ocr}}$ を MLP へ入力し, レイアウト c_1 の各座標 $(x_1^{\text{min}}, x_1^{\text{max}}, y_1^{\text{min}}, y_1^{\text{max}})$ を予測する[3]. 学習の際は, MLP の予測値と正解 (例えば, $\hat{x}_i^{\text{min}}$ と x_i^{min}) の二乗誤差 $|\hat{x}_i^{\text{min}} - x_i^{\text{min}}|^2$ を最小化する. 提案手法では, 矩形領域集合 $\mathbf{R}$ のうち, 30%の対象矩形領域のレイアウトをマスクし, 本タスクを学習する.

Saliency Detection LP と関連性の高い内容をモデルで考慮し, 広告文の品質を向上することを狙い, Tanaka らが提案した Saliency loss [5] を導入する. 本研究では, トークン系列集合 $\mathbf{X}$ の中で参照文 $\mathbf{y}$ に出現するトークン x に対して 1, それら以外には 0 を予測する二値分類を学習する. これは, 入力から言及すべき内容を選択する内容選択に相当する [8].

3.3 Mixture LM Head

広告文は, 広告主によって語彙の多様性や事例数が大きく異なる不均衡データであり, 機械学習に基づく文生成において大きな障壁となる [2]. この課題に対し, 本研究では, Mixture of Experts (MoE) [15] に基づくモデルを提案する. MoE を用いたエンコーダデコーダモデルでは, 1 つの入力に対して異なる Expert (デコーダ) を割り当てることで多様な文生成が可能になることが知られている [16]. また, MoE により不均衡データに頑健な機械学習モデルを構築する取り組みもある [17]. そこで本研究では, MoE を導入することで, 様々な広告主の広告文からなる不均衡データに対して頑健な文生成モデルを目指す.

これまでの多くの研究 [16, 18, 19] では, デコーダ全体を Expert として用いているが, MoE を Transformer などの大規模モデルに適用する場合, パラメータ数が大幅に増加し, 学習が困難になることが懸念される. そこで本研究では, Language modeling head (LM head)[4] を Expert として用いる Mixture LM head を提案する (図 2). これにより, モデル全体のパラメータ数を抑えつつ[5], Expert (LM head) ごとに異なる語彙特徴を捉えることを狙う. また, 提案モデルでは, Multiple Choice Learning (MCL) [20] に基づいて MoE を学習する. MCL では, 複数の Expert の中で, 最も loss が低い Expert の損失関数の勾配に基づいて, パラメータを更新する [18]. これにより, 様々な広告主の広告データが写像される隠れ状態を, モデル内でより明確に領域分割することで, 不均衡データに対する頑健性が向上することを期待する [16].

4 実験設定

データセット 実験では, 金融や EC, 不動産等の全 13 業種の広告主 (全 41 社分) の広告データをそれぞれ 39,166 件, 5,254 件, 5,035 件の学習, 開発, 評価データに分割し使用する. 各広告に紐づく LP として, LP ファーストビュー[6] のスクリーンショット画像を用い, Cloud Vision API の OCR 機能[7] により, 矩形領域集合 $\mathbf{R}$ に対する画像 $\mathbf{I}$, レイアウト $\mathbf{C}$, テキスト $\{\mathbf{x}_i^{\text{ocr}}\}_{i=1}^{|\mathbf{R}|}$ を取得した.

評価指標 生成文の品質を測るために, LP 等から人手で作成された参照文との N-gram の一致率に基づく BLEU-4 (B-4) [21], ROUGE-1 (R-1), ROUGE-2 (R-2) [22] を使用する. また, 生成文の多様性を測るための指標として, Distinct-1 (D-1), Distinct-2 (D-2) [23], Pairwise-BLEU (P-B) [16], Self-BLEU (S-B)[24] を用いる[8]. 加えて, LP と広告文の関連性を測るために, 指定したキーワードが生成文に含まれる事例の割合を表すキーワード挿入率 (Kwd), 入力トークン $\mathbf{X}$ と生成文 $\mathbf{y}$ の ROUGE-1 (Precision) により算出する Fidelity スコア (Fid) を用いる.

実装 モデルには, 事前学習済みの T5[9] [9] を使用する. また, 特筆の無い限り, Mixture LM head による文生成時は評価文に対して最も loss が低い Expert を使用し, 貪欲法に基づき 1 文を生成する. その他の詳細な実験設定は付録 A に記載する.

3) 簡略化のために, 図 2 では MLP と $\mathbf{e}_1^{\text{ocr}}$ を省略している.

4) デコーダの出力状態を語彙次元サイズへ変換する MLP 層

5) Transformer (パラメータ数は 223M) のデコーダを Expert とした場合, パラメータ数は 711M だが, LM head を Expert とした場合は 297M となる. ここで, Expert 数は 4 とする.

6) LP をスクロールせずに最初に見える範囲

7) https://cloud.google.com/vision/docs/ocr

8) P-B, S-B の算出には B-4 を用いる.

9) https://huggingface.co/sonoisa/t5-base-japanese

表 1　各 Embedding およびサブタスクの比較

Model	B-4	R-1	R-2	D-1	D-2	S-B	P-B	Fid	Kwd
T5 (K=1)	11.1	23.5	13.4	1.8	4.1	94.1	–	52.4	32.7
$+v$	10.7	19.6	10.6	1.2	2.4	91.7	–	57.0	35.1
$+v, l$	11.4	19.8	11.4	1.0	2.2	94.8	–	57.6	37.0
$+v, l, s$	9.8	19.6	10.5	1.3	2.7	94.4	–	57.9	35.0
T5 (K=4)	8.6	19.0	10.1	1.5	3.5	**89.8**	87.3	56.6	33.8
$+v, s$	8.3	17.9	9.2	1.3	3.2	95.3	82.6	53.8	35.9
$+v, s$ [B]	9.2	23.1	12.5	**1.9**	4.4	93.6	86.5	59.7	37.5
$+v, l, s$	9.1	21.1	11.5	1.8	4.5	92.3	**81.3**	54.9	34.7
$+v, l, s$ [B]	12.5	24.0	13.6	**1.9**	5.1	94.3	83.8	54.7	34.6
$+v, l, s$ [S]	10.1	20.0	10.6	1.5	3.5	95.4	82.6	**61.0**	**38.5**
$+v, l, s$ [B][S]	**15.6**	**27.1**	**16.3**	**1.9**	**5.0**	93.2	86.5	54.3	35.0

表 2　Expert 数ごとの生成文, N-gram の異なり数

Model	#Text	#N-1	#N-2	#N-3	#N-4
T5 (K=1) $+v, l, s$	228	608	1,520	2,603	3,208
T5 (K=2) $+v, l, s$	503	762	2,160	4,231	5,559
T5 (K=4) $+v, l, s$	426	739	2,074	4,002	5,201

5　実験結果

各 Embedding の有用性　表 1 に各 Embedding (v, l, s) を用いたモデルの比較を示す[10]．ここで, K は Expert 数, [B] および [S] は Bounding-Box Layout Prediction, Saliency Detection を導入したモデルを表す．K=1 において, 各 Embedding の導入により, LP と広告文の関連性を表す指標 (Fid, Kwd) の向上を確認した．K=4 では, Kwd の改善を確認した．

マルチタスク学習の導入効果　T5 (K=4) において, 各サブタスク ([B], [S]) を導入することで, 生成文の品質が向上することを確認した．OCR テキストに対してレイアウト情報を用いない場合 $(+v, s)$ であっても, [B] を導入することでスコア向上を確認した．また, $+v, l, s$ に対して [B] を導入することでスコアが大幅に向上した．このことから, 文書コーパスで事前学習された T5 に対するレイアウト情報の追加学習が効果的だったことが分かる．また, 2 つのサブタスクを同時に導入した場合, さらにスコアが向上した．

Mixture LM head による多様性の改善　表 2 に Expert 数を 1, 2, 4 とした場合の生成文 (#Text), 生成文に含まれる N-gram (#N-1,#N-2,#N-3,#N-4) の異なり数を示す．表 2 より, Expert 数を増やしたことでこれらの異なり数が増加したことから, Mixture LM head による多様性の改善が推察できる．また, K=2 において, 最も異なり数が多い結果が得られた．さらに, 表 3 に Mixture LM head および Beam search により生成した広告文の多様性および生成品質の比較を示す．ここで, B は Beam 幅, S は Top_k および Top_p sampling [25] の有無であり, 各モデルは生成時に 1 つの入力に対し 4 文ずつ生成する[11]．例えば, K=4 の場合, 各 Expert で貪欲法に基づき 1 文ずつ (計 4 文) を生成する．K=4 では, 貪欲法に基づいて生成するものの, Beam search を用いる K=1 と比べて, 多様性が匹敵した上で, 生成品質が劣化しないことを確認した．

10)　ベースラインとして, T5 (K=1, 4) では v, l, s を用いない．また, v に対する l はいずれのモデルにおいても必ず導入する．

表 3　Mixture LM head と Beam search の比較 (T5 $+v, l, s$)

K	B	S	#Text	#N-1	#N-2	#N-3	#N-4	R-1	R-2
1	4	✓	484	766	2,157	4,355	5,831	20.4	11.1
1	4	–	**827**	768	2,203	4,523	6,076	19.5	10.4
4	1	–	668	**793**	**2,341**	**4,900**	**6,609**	**21.1**	**11.5**

表 4　広告主ごとの ROUGE-1 の比較 (T5 $+v, l, s$)

Adv.	#Train	#Test	K=1	K=4
(1)	2,265	280	**20.6**	20.1
(2)	1,988	338	13.0	**19.0**
(3)	1,906	328	**23.6**	22.1
(4)	642	42	16.5	**19.1**
(5)	620	12	20.4	**36.8**
(6)	609	133	21.0	**25.3**
(7)	506	49	11.4	11.4
(8)	397	31	34.3	**41.4**

不均衡データに対する頑健性　不均衡データに対する頑健性の検証のために, 学習データにおいて事例が少ない広告主の評価データに対する性能を確認する．表 4 に学習データの事例数 (#Train) が上位 3 社, 下位 5 社の広告主 (Adv.) の評価データ (#Test) に対する評価結果を示す．複数の Expert (K=4) の導入により, 学習事例が少ない広告主のデータに対しても, 品質の高い広告文が生成できることを確認した．

6　おわりに

本研究では, LP からの広告文生成を文書画像からの言語生成タスクとして取り組んだ．実験では, 文書コーパスで事前学習された T5 に対して, 追加学習時に新たにレイアウト情報を効率的に学習するためのマルチタスク学習を導入することで, レイアウト情報を考慮した際の生成品質が大幅に向上することを示した．また, 広告データにおける語彙やデータ事例数の不均衡性に対して, MoE に基づく Mixture LM head を提案し, 生成文の多様性および少量事例 (広告主) に対する生成品質の頑健性の向上を示した．

11)　Sampling 有りの場合, Top_k = 50, Top_p = 1.0 とした．

参考文献

[1] Siyu Duan, Wei Li, Jing Cai, Yancheng He, and Yunfang Wu. Query-variant advertisement text generation with association knowledge. In **Proceedings of the 30th ACM International Conference on Information & Knowledge Management**, p. 412–421, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. **Journal of Big Data**, Vol. 6, No. 1, pp. 1–54, March 2019.

[3] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '20, p. 1192–1200, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 2579–2591, Online, August 2021. Association for Computational Linguistics.

[5] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 13878–13888, May 2021.

[6] Kevin Bartz, Cory Barr, and Adil Aijaz. Natural language generation for sponsored-search advertisements. In **Proceedings of the 9th ACM conference on Electronic commerce**, EC '08, pp. 1–9, New York, NY, USA, July 2008. Association for Computing Machinery.

[7] Atsushi Fujita, Katsuhiro Ikushima, Satoshi Sato, Ryo Kamite, Ko Ishiyama, and Osamu Tamachi. Automatic generation of listing ads by reusing promotional texts. In **Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business**, ICEC '10, pp. 179–188, New York, NY, USA, August 2010. Association for Computing Machinery.

[8] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research**, Vol. 61, No. 1, p. 65–170, jan 2018.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[10] Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. Reinforcing pretrained models for generating attractive text advertisements. In **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining**, KDD '21, pp. 3697–3707, New York, NY, USA, August 2021. Association for Computing Machinery.

[11] J. Weston Hughes, Keng-hao Chang, and Ruofei Zhang. Generating better search engine text advertisements with deep reinforcement learning. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '19, p. 2269–2277, New York, NY, USA, 2019. Association for Computing Machinery.

[12] Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. An empirical study of generating texts for search engine advertising. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers**, pp. 255–262, Online, June 2021. Association for Computational Linguistics.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, p. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. **Neural Computation**, Vol. 3, No. 1, pp. 79–87, 1991.

[16] Tianxiao Shen, Myle Ott, Michael Auli, and Marc'aurelio Ranzato. Mixture models for diverse machine translation: Tricks of the trade. In **Proceedings of the 36th International Conference on Machine Learning**, Vol. 97, pp. 5719–5728, 2019.

[17] SB Kotsiantis and PE Pintelas. Mixture of expert agents for handling imbalanced data sets. **Annals of Mathematics, Computing & Teleinformatics**, Vol. 1, No. 1, pp. 46–55, 2003.

[18] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In **Advances in Neural Information Processing Systems**, Vol. 29. Curran Associates, Inc., 2016.

[19] Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander Rush. End-to-end content and plan selection for data-to-text generation. In **Proceedings of the 11th International Conference on Natural Language Generation**, pp. 46–56, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics.

[20] Abner Guzmán-rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In **Advances in Neural Information Processing Systems**, Vol. 25. Curran Associates, Inc., 2012.

[21] 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**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[23] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[24] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**, SIGIR '18, p. 1097–1100, New York, NY, USA, 2018. Association for Computing Machinery.

[25] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. **Computing Research Repository**, Vol. abs/1904.09751, , 2019.

A 付録

付録では, モデルのハイパーパラメータや学習器等の詳細な実験設定について説明する.

A.1 実装詳細

実験で使用した T5 [9] は, Wikipedia [12], OSCAR[13], CC-100[14] の日本語文書コーパス (約 100GB) を用いて事前学習されている. 本モデルは, Hugging Face の transformers ライブラリ[15] で利用可能であり, 各種パラメータは設定ファイル[16] に記載されている.

マルチタスク学習で導入する Bounding-Box Layout Prediction および Saliency Detection では, 1 層の MLP を用いる. 視覚特徴 v_i を作成するための CNN では, 入力チャネルは 3, 出力チャネルは 32, ストライドは 3, カーネルサイズは 5 としており, 出力に対して Batch Normalization, ReLU および 1 層の MLP の適用し, v_i を獲得した.

A.2 モデル学習

モデルパラメータの最適化手法には Adam を使用し, 学習率は 3×10^{-5}, ミニバッチサイズは 8 とした. 学習時の epoch 数は 20 とし, 開発データに対する loss が 3 epoch 連続で劣化した場合には早期終了した. Mixture LM head を用いる場合は, モデルの過学習の抑制を目的として, 3 epoch まではモデルパラメータ全体を学習し, 4 epoch 以降は LM head のパラメータのみ更新した.

また, Shen ら [16] は, MoE の学習において, Expert を選択する際 (E-step 時) に Dropout を無効化することで, Expert 選択の一貫性が向上し, 各 Expert が異なる領域に特化するため, 生成文の多様性が向上することを明らかにした. そこで, 本研究でも同様の手順でモデル学習を実施している.

12) https://ja.wikipedia.org/
13) https://oscar-corpus.com/
14) http://data.statmt.org/cc-100/
15) https://github.com/huggingface/transformers
16) https://huggingface.co/sonoisa/t5-base-japanese/blob/main/config.json