質問応答の教師なしドメイン適応：機械読解と言語モデルのマルチタスク学習

西田　光甫†　　西田　京介†　　斉藤いつみ†　　浅野　久子†　　富田　準二†

† 日本電信電話株式会社 NTT メディアインテリジェンス研究所　〒 239–0847 神奈川県横須賀市光の丘 1–1
E-mail: †{kosuke.nishida.ap,kyosuke.nishida.rx,itumi.saito.df,hisako.asano.fe,junji.tomita.xe}@hco.ntt.co.jp

あらまし　機械読解はテキストを読み解いて質問応答をするタスクであるが，訓練データに含まれないドメイン（out-domain）では精度が悪いことが知られている．本研究では，out-domain のテキストを理解する能力を言語モデルから獲得することでターゲットドメインに適応することを目指した．ターゲットドメインにおける機械読解の教師データなしにドメイン適応を行うタスクに取り組み，ターゲットドメインでの言語モデルとソースドメインでの機械読解のマルチタスク学習を行う手法を提案した．評価実験にて提案手法がドメイン適応をしないモデルの精度を上回り，特に医療ドメインにおいて EM/F1（回答範囲の完全/部分一致) で 4.3/4.2%精度を向上することを確認した．

キーワード　機械読解，ドメイン適応，教師なし学習

1 はじめに

機械読解はテキスト資源を参照し，テキストを読み解きながら質問応答をするタスクである．機械読解は SQuAD [21] に代表される数々のデータセットの公開及び BiDAF [23] に代表される深層学習モデルの提案に伴い飛躍的な発展を遂げた．近年では BERT [3] に代表される事前学習済み言語モデルの Fine-Tuning が機械読解を含めた多くの自然言語処理タスクで State-of-the-Art の精度を達成している．

しかしながら，大規模事前学習に基づくモデルでさえも汎用的な言語理解能力を獲得しているとは言い難い．テキスト資源のドメインが違うなど，機械読解の訓練データと評価データが異なる確率分布に従う場合，モデルの性能が悪化することが知られている [5]．このドメイン依存性は実利用時の課題となっている．例えば機械読解アプリケーションをサービスに導入する際は，ドメイン毎に数万件規模の（参照テキスト，質問，回答）の 3 つ組から成る教師データが要求される [33]．しかし，数万件規模の教師データの作成は金銭的・時間的コストが大きい．特にドメインが個人情報や専門知識を含む場合は作成にクラウドソーシングを利用できず，膨大なコストが必要となる．

本研究では，機械読解における教師なしドメイン適応（Unsupervised Domain Adaptation of Reading Comprehension, UDARC）に取り組む．UDARC のタスク設定を表 1 に示す．モデルの学習で用いることができる（参照テキスト，質問，回答）タプルの教師データはソースドメインのデータのみである．加えて学習には，ターゲットドメインの教師なし参照テキストが利用可能である．ターゲットドメインでは機械読解の教師データ（質問・回答ペア）を利用することができない．UDARC のタスク設定は汎用的な言語理解能力獲得への興味に留まらず，実利用時の課題に応える設定となっている．なぜなら，機械読解アプリケーションのプロバイダはターゲットドメインの参照テキストとして利用可能なテキスト資源（例：商品の説明書など）を所持していることが想定されるためである．UDARC は，そのようなテキスト資源を持っているがドメインに応じた機械読解コーパスを持っていない事業者にも，機械読解のサービス導入を可能にすることを目標にしたタスクである．

	訓練			評価		
	入力		出力	入力		出力
	参照テキスト	質問	回答	参照テキスト	質問	回答
ソース	✓	✓	✓			
ターゲット	✓			✓	✓	✓

表 1　機械読解における教師なしドメイン適応（UDARC）のタスク設定．訓練時に用いることができるターゲットドメインのデータは教師なしの参照テキストのみである．UDARC では，質問・回答ペアに関しては zero-shot でのドメイン適応が必要となる．

我々は，機械読解のドメイン依存性を，訓練データにないドメイン（out-domain) の言語理解能力の不足に起因すると仮定した．即ち，out-domain のテキスト資源を用いた言語モデルの訓練によって，out-domain の機械読解教師データを使わずに精度を向上できると考えた．比較手法として，2 つの機械読解モデルを採用した．1 つ目のモデルはドメイン適応を行わないモデルであり，事前学習済みの BERT をソースドメインの機械読解教師データで Fine-Tuning し，ターゲットドメインで評価する．2 つ目のモデルは教師なしドメイン適応で通常用いられる手法であり，まずターゲットドメインで BERT の言語モデルを追加訓練した後，ソースドメインの機械読解教師データで Fine-Tuning し，ターゲットドメインで評価する．

本研究では，ターゲットドメインにおける言語モデルとソースドメインにおける機械読解のマルチタスク学習を提案した．マルチタスク学習によって，機械読解への Fine-Tuning 時に起きるターゲットドメインの知識忘却を防ぐことが期待できる．また，UDARC の有効性と上記モデルの性能の評価を行った．

本研究の貢献を以下に挙げる．

- **モデルの提案**：我々は機械読解における教師なしドメイン適応（UDARC）を解くため，out-domain の言語理解能力を機械読解の教師データなしで獲得することが可能であると仮定した．この仮定を基に，ターゲットドメインの言語モデルとソースドメインの機械読解のマルチタスク学習を提案した．提案モデルによって，質問に答える能力をソースドメインから，テキストを理解する能力をターゲットドメインから知識忘却を緩和して獲得できる．
- **タスクの実現可能性の評価**：機械読解の教師なしドメイン適応の実現性を検証するため，5 つのドメインで実験を行った．結果，ドメイン適応ありのモデルはドメイン適応なしのモデルの精度を上回った．特に提案モデルは最大の精度向上を示し，Wikipedia ドメインから医療ドメインへの適応において EM/F1（回答範囲の完全/部分一致）で 4.3/4.2%の向上を達成した．また，ソースドメインの教師データで学習したモデルからの教師なしドメイン適応が，ターゲットドメインでの教師あり学習の精度を上回る場合があることを確認した．

2 タスク定義

本研究では，機械読解の中で最も用いられる抽出型機械読解に着目した．抽出型機械読解の定義を以下に示す．

定義 1 (抽出型機械読解)．抽出型機械読解は（参照テキスト，質問）の組を入力し，参照テキストの中から 1 つの区間を抜き出すことで質問に答えるタスクである．モデルは参照テキスト中の区間，つまり回答の始端位置と終端位置，を予測することで回答を抽出する．

抽出型機械読解モデルは，ベクトル $s, e \in \mathbb{R}^l$ を出力する．ここで，l は参照テキストの長さ，s, e は次元に対応する位置のトークンが回答の始端・終端であることのスコアを表す．抽出型機械読解の目的関数は，真の回答の始端位置を i，終端位置を j とした以下の CrossEntropy 損失である

$$L_{\mathrm{RC}} = -\log \frac{\exp(s_i)}{\sum_k \exp(s_k)} - \log \frac{\exp(e_j)}{\sum_k \exp(e_k)}. \qquad (1)$$

本研究は抽出型機械読解の教師なしドメイン適応を行う．タスクの定義を以下に示す．

定義 2 (UDARC: 機械読解における教師なしドメイン適応)．I^S_{RC} をソースドメインにおける（参照テキスト，質問，回答区間）の 3 つ組の集合であるとする．I^T_{LM} をターゲットドメインにおける参照テキストの集合であるとする．UDARC のタスクは I^S_{RC} と I^T_{LM} で訓練した抽出型機械読解モデルを用いてターゲットドメインの質問に回答するタスクである．

本研究は，マスク化言語モデル（Masked Language Model, MLM）[3] によってターゲットドメインの知識を獲得することで UDARC に取り組む．マスク化言語モデルの定義を以下に示す．

定義 3 (マスク化言語モデル)．語彙を V とする．コーパス中のトークン系列を $\bar{X} \in V^l$ とする．ただし l はトークン系列長である．マスク化言語モデルの入力は $\bar{X}$ に摂動（一部トークンのマスクなど）を与えた系列 $X \in V^l$ である．マスク化言語モデルは摂動前のトークン系列 $\bar{X}$ を X から予測する．出力は行列 $H(X) \in \mathbb{R}^{|V| \times l}$ であり，$h_{vt}(X)$ は $\bar{X}$ の第 t トークンが $v \in V$ であることのスコアである．

マスク化言語モデルの目的関数は，摂動を与えたトークン位置の集合を T として，以下の損失である

$$L_{\mathrm{LM}} = -\sum_{t \in T} \log \frac{\exp(h_{\bar{x}_t t}(X))}{\sum_v \exp(h_{vt}(X))}. \qquad (2)$$

3 関 連 研 究

3.1 機 械 読 解

UDARC に取り組んだ先行研究に [7], [30] がある．彼らはターゲットドメインで擬質問・回答ペアを生成することで機械読解の教師なしドメイン適応を行った．我々は彼らと異なり，ターゲットドメインの知識を機械読解データを用いずに獲得することを目標としている．我々のアプローチの利点の 1 つは，計算コストの大きい擬質問生成を行わないことである．

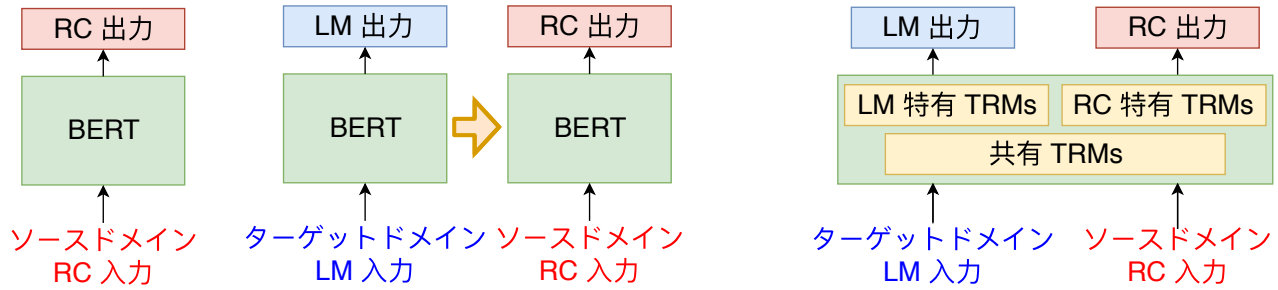
我々に近い目標を持つタスクに MRQA 2019 shared task [5] がある．彼らは抽出型機械読解がテスト時の分布に関して汎化性能を持ち，テスト時の摂動に対して頑健になることを目指した．このタスクは 6 つの in-domain 訓練データと 12 の out-domain 評価データで構成されており，多対多の設定で汎化能力を測るタスクであると言える．我々の目標は，教師なしテキストから out-domain の言語理解能力を獲得することであり，1 対 1 のドメイン適応に取り組む．この目標は実利用時の課題に対応している．shared task の結果，機械読解の out-domain での性能は背後で用いた事前学習済み言語モデルの性能に大きく依存するという結論が得られた．この結論は，我々の言語モデルの訓練を用いて機械読解の性能向上を図るアプローチの妥当性を示唆している．

機械読解は抽出型 [21]，選択肢型 [13]，穴埋め型 [8]，生成型 [18] の大きく 4 つのタイプに分かれる．本研究では抽出型に着目したが，我々のアプローチは他のタイプの機械読解にも有効であると考えられる．

近年多くの機械読解データセットが公開されているが，ドメインの閉じたデータは限られている．例として，医療 [25], 科学 [2], [10], ソフトウェア [4] のドメインのデータがある．データセットの少なさは，我々が UDARC に取り組む動機の 1 つである．なぜなら，専門知識が必要なドメインでは質問対応の自動化への需要が大きいが，教師データ作成の難しさが機械読解の導入障壁となっていると解釈できるためである．

3.2 ドメイン適応

教師なしドメイン適応は，ソースドメインの教師ありデータ（入力と真の出力のデータ）とターゲットドメインの教師なしデータ（入力データのみ）を用いてモデルをターゲットドメインに適応させるタスクである．教師ありドメイン適応では初め



(a) No-adaptation (b) Sequential (c) Multi-Task

図 1　モデルの概要．言語モデル（LM）のサンプルはターゲットドメインから，機械読解（RC）のサンプルはソースドメインから抽出する．“TRMs” は Transfomer 層を示す．

にソースドメインでモデルを訓練した後ターゲットドメインの入出力を使ってモデルを適応するが，教師なしドメイン適応では別のアプローチを取る．自然言語処理では，[35] が初めに表現を学習した後分類問題を学習する 2 段階アルゴリズムを採用した．[17] は分類問題と structural correspondence learning [1] の特徴選択問題のマルチタスク学習を提案した．また，[6] はドメイン敵対的モデルを用いることで表現学習と分類問題を組み合わせて学習した．

しかし，UDARC では通常の教師なしドメイン適応手法を適用できない．表 1 が示すサンプルの入力（機械読解では（参照テキスト，質問）のペア）と出力（機械読解では回答）のうち，教師なしドメイン適応はターゲットドメインの入力データを必要とするためである．一方 UDARC は，サンプルの出力（回答）に加え，入力の一部（質問）も用いることができない．

Zero-shot 学習は訓練中に現れないクラスの予測を可能にするために，モデルを未知クラスに適応させるタスクである [24]．特に近年は，訓練中に現れないドメインに対して適応する zero-shot ドメイン適応 [19], [31] のタスクが提案されている．UDARC はターゲットドメインの完全な入力を訓練中に知ることができないという点で，一種の zero-shot ドメイン適応の問題であると解釈できる．

Zero-shot ドメイン適応は教師なしドメイン適応よりも困難なタスクであり，取り組んでいる研究は少ない．[19] は画像の分類問題において，タスクに無関係なターゲットドメインのデータを用いることでドメイン適応を行った．[31] と [16] は訓練データが複数のドメインから構成されることを仮定した研究である．[9] はソースドメインとターゲットドメインの分布の違いについて事前知識を持つことを仮定した研究である．

4　手　　法

本節では，ドメイン適応をしないモデル（No-adaptation Baseline），連続的にドメイン適応を行うモデル（Sequencial Model），マルチタスク的にドメイン適応を行う提案モデル（Multi-task Model）の 3 つのモデルを説明する．全てのモデルは，事前学習済みの $\text{BERT}_{\text{base}}$ モデルから学習した．$\text{BERT}_{\text{base}}$ モデルは大規模コーパスから学習された言語モデルであり，機械読解を含む個別のタスクへ Fine-Tuning することで高い精度を達成できる．

4.1　事 前 学 習

本小節では $\text{BERT}_{\text{base}}$ が行った事前学習について説明する．詳細は [3] を参照されたい．

事前学習中の入力は，コーパス中からランダムにサンプリングされた 2 つの文章を連結したトークン系列 [‘[CLS]’; 文章 1; ‘[SEP]’; 文章 2; ‘[SEP]’] である．‘[CLS]’，‘[SEP]’ は特殊トークンを，‘;’ は連結を示す．また，0–1 変数（segment id）の系列 $[0, \cdots, 0, 1, \cdots, 1]$ も入力する．この系列は，トークン系列と同じ長さの系列であり，トークン系列の ‘[CLS]’，文章 1，1 つ目の ‘[SEP]’ に当たる箇所では 0，文章 2 と 2 つ目の ‘[SEP]’ に当たる箇所では 1 の値が入る．segment id の系列によって，モデルが各トークンの種別を識別し，文章 1・2 の相互関係をモデリングすることが可能となる．

事前学習に用いたコーパスは BookCorpus （8 億単語）[34] と英語版 Wikipedia （25 億単語）である．事前学習のタスクは Next Sentence Prediction （NSP）とマスク化言語モデル（MLM）である．NSP では 2 つの文章がコーパス中で連続する文章であるか，別の箇所からサンプリングした文章であるかの 2 クラス分類を行う．MLM は入力トークンの一部をランダムに ‘[MASK]’ トークンなどに置き換える摂動を与える．

4.2　No-adaptation Baseline

本研究では，ベースラインモデルとして BERT をソースドメインの機械読解教師データで Fine-Tuning したモデルを採用した．本モデルはドメイン適応を行わないモデルであり，精度は UDARC の下界に相当すると考えられる．

図 1 (a) にモデルの概要を示す．Fine-Tuning の手法は [3] にある抽出型機械読解の手法に従った．つまり，BERT の上に 2 次元の線形変換層を重ねた

$$[s_t; e_t]^\top = W_{\text{RC}} z_{\cdot t} + b_{\text{RC}} \in \mathbb{R}^2 \qquad (3)$$

によって始端・終端スコアを出力し，目的関数 L_{RC}（式 1）を用いて抽出型機械読解の学習を行う．ここで、BERT の最終層の出力を $Z \in \mathbb{R}^{768 \times l}$ とする．$W_{\text{RC}} \in \mathbb{R}^{2 \times 768}, b_{\text{RC}} \in \mathbb{R}^2$ は学習パラメータである．モデルへの入力は文章 1 に質問を，文章 2 に参照テキストを挿入したトークン系列 [‘[CLS]’; 質問;

‘[SEP]’; 参照テキスト; ‘[SEP]’] と segment id の系列である．

4.3 Sequential Model

Sequential Model は初めに事前訓練済み BERT をターゲットドメインの教師なしテキストに適応させる．次に，ソースドメインの機械読解教師データへの Fine-Tuning を行う．

図 1 (b) にモデルの概要を示す．初めの教師なしドメイン適応では，ターゲットドメインでの追加学習を事前学習時同様の手法で行う．つまり，線形変換層を BERT の上に重ね，BERT と線形変換層を NSP，MLM で学習する．NSP では pooling と線形変換によって 2 次元ベクトルを出力し，CrossEntropy 損失によって学習する．MLM では

$$h_{\cdot t} = W_{\mathrm{LM}} z_{\cdot t} + b_{\mathrm{LM}} \in \mathbb{R}^{|V|} \tag{4}$$

によってスコアを出力し，目的関数 L_{LM}（式 2）を用いて学習する．$W_{\mathrm{LM}} \in \mathbb{R}^{|V| \times 768}, b_{\mathrm{LM}} \in \mathbb{R}^{|V|}$ は学習パラメータである．その後は No-adaptation Baseline 同様の手法でソースドメインの機械読解データへの Fine-Tuning を行う．

ドメイン適応を本研究で用いた順序とは逆の順序で，即ちソースドメインの機械読解モデルへの Fine-Tuning 後にターゲットドメインの言語モデルに適応する順序で学習することも可能であるように思える．しかしその場合は，機械読解の能力が言語モデルへの適応時に失われる，破壊的忘却が起きてしまう．本研究の順序は [35] の先行研究と同じ順序を採用している．

4.4 提案モデル

Sequential Model では，ターゲットドメインの知識が機械読解への Fine-Tuning 中に忘却される恐れがある．また，事前学習済み言語モデルのマルチタスク学習による Fine-Tuning は個々のシングルタスク学習による Fine-Tuning の性能を上回ることが知られている [14]．これらを背景に，我々はマルチタスク学習によるアプローチを提案する．提案モデルでは，BERT の上に機械読解用の線形層（式 3）と言語モデル用の線形層（式 4）を重ね，機械読解のサンプルでは機械読解層を，言語モデルのサンプルでは言語モデル層を用いる．図 1 (c) にモデルの概要を示す．我々の提案モデルは 2 つの特徴を持つ．

4.4.1 Transformer 層の分割

BERT は多層化された Transformer 層 [29] で構成されている．提案手法では，このうち上部 n 層をタスク特有の層としてタスク別のパラメータを用意する．タスク特有層の初期値は元の事前学習済み BERT と同一であり，コピーすることで 2 つのタスク特有層を用意する．その他の層はタスクで共有の層として，同一のパラメータを用いる．よって，機械読解のサンプルはタスク共有層，機械読解特有層，機械読解出力層を順に通り，言語モデルのサンプルはタスク共有層，言語モデル特有層，言語モデル出力層を順に通る．この着想は [26] の観察に基づいている．彼らは，BERT において品詞分類などの基本的統語情報は低層で，共参照解析などの高度な意味情報は高層で捉えられていることを実験的に示した．我々は，基本的統語情報はタスク共通の表現，高度な意味情報はタスク依存の表現であると仮定した．そこで，ターゲットドメインの言語モデルをタスク共有層を使って学習することで，モデルがターゲットドメインの基本的統語情報を理解することを可能にした．また，高層部はタスク特有の層とすることで，機械読解における高度な意味情報の表現を言語モデルにおける表現が阻害しないようにした．

Algorithm 1 Multi-task learning approach

Input: source RC samples I^S_{RC}, target LM samples I^T_{LM}, num. of steps N, the RC training ratio k

1: **for all** i in $1, \cdots, N$ **do**
2: Select mini-batch $b_S \sim I^S_{RC}$
3: Train the shared layers, RC-specific layers, and RC output layer with b_S by minimizing L_{RC}.
4: **if** $i\%k == 0$ **then**
5: Select mini-batch $b_T \sim I^T_{LM}$
6: Train the shared layers, LM-specific layers, and LM output layer with b_T by minimizing L_{LM}.
7: **end if**
8: **end for**

4.4.2 1-Segment 言語モデル

事前学習時の言語モデルの入力は [‘[CLS]’; 文章 1; ‘[SEP]’; 文章 2; ‘[SEP]’] であるが，提案手法では入力を [‘[CLS]’; ‘[LM]’; 文章 1; ‘[SEP]’] とした．同様に，segment id は全て 0 とした．よって，言語モデルの訓練で NSP は用いず，MLM のみを用いて学習を行った．これは，segment 1 を含む入力を機械読解に限定することで，NSP における文章 1・文章 2 の相互作用が機械読解の質問・参照テキストの相互作用を阻害しないことを意図している．また，‘[LM]’ はサンプルが言語モデルのサンプルであることを意味する特殊トークンである．

4.4.3 訓　　練

訓練時は機械読解の訓練・言語モデルの訓練を交互に行った．それぞれの訓練の目的関数は L_{RC}（式 1），L_{LM}（式 2）である．機械読解の訓練は言語モデルの訓練の k 倍の回数行う．アルゴリズムは Algorithm1 に示す．

4.4.4 計算コスト

提案手法の利点に，計算コストの小ささがある．推論時は言語モデル特有層を用いないため，モデルサイズ・時間計算量ともに元々の BERT と同じである．訓練時も，step 数は $1 + 1/k$ 倍になるだけなので，通常の BERT の Fine-Tuning から時間計算量に大きな変化はない．また，言語モデルの事前学習には膨大な計算コストがかかるため，ターゲットドメインに対して都度事前学習を行うことは現実的ではない．

5 実　　験

5.1 データセット

本研究では UDARC タスクを 5 つのデータセットで評価した．訓練データへの要請として，データサイズが大きく，広いトピックを網羅していることがある．評価データへの要請として，ドメインが閉じていることがある．以上の要請から，我々は下記のデータセットで実験を行った．統計値は表 2 に示す．

データセット	ドメイン	訓練データサイズ	開発データサイズ	教師なし参照テキスト数
SQuAD	Wikipedia	87599	10570	19047
NewsQA	ニュース	107064	5988	95933
BioASQ	医療	0	1504	55148
DuoRC	映画	69524	15591	5137
Natural Questions	HTML Wikipeadia	104071	12836	12222

表 2　使用したデータセットの統計値.

一部のデータセットはテストデータが非公開であるため，評価は全て開発データで行った．表中の教師なし参照テキスト数はサンプルの数を表しており，全てのデータを実験で使ったとは限らない．実際に使用した教師なしテキストの数は設定によって異なり，訓練データサイズ，エポック数，k の値に依存して決定する．以下，使用したデータセットについて記述する．

a） SQuAD1.1

SQuAD1.1 は Wikipedia から作成された [21]．ソースドメインとターゲットドメインで利用した．ターゲットドメインとして利用する際は，訓練データ中の参照テキストを教師なしテキストとした．

b） NewsQA

NewsQA は CNN news から作成された [27]．ソースドメインとターゲットドメインで利用した．ターゲットドメインとして利用する際は，CNN news scripts [8] から NewsQA 作成時同様の手法で収集したテキストを教師なしテキストとした．

c） BioASQ

BioASQ は a biomedical semantic indexing and question answering challenge[1] で用いられた [28]．我々はこのデータセットを MRQA 2019 shared task が抽出型機械読解の評価データに加工したものをターゲットドメインとして利用した．訓練中の教師なしテキストとして Pubmed から論文の概要を収集した．

本研究は教師なしドメイン適応が目的であり，BioASQ が最も重要なデータセットである．なぜなら，BioASQ はドメイン（医療）がソースドメイン（Wikipedia もしくはニュース）・BERT の事前訓練コーパス（BookCorpus と Wikipedia）に含まれない唯一のデータセットであるからである．

d） DuoRC

DuoRC は Wikipedia, IMDb から収集した同一の映画についての 2 種のあらすじの一方を参照テキストとして利用する [22]．我々は，DuoRC の ParaphraseRC タスクをターゲットドメインとして利用した．ParaphraseRC は，クラウドワーカーが Wikipedia のあらすじを読んで作成した質問に，IMDb のあらすじを参照テキストに利用して回答するタスクである．そのため，質問と参照テキストの共有する文字列が少なく，通常の機械読解より難しいデータセットになっている．

映画ドメインはソースドメインとは異なるが，BERT の事前訓練で用いられた BookCorpus には多くの物語が含まれている．そのため，DuoRC が要求する知識は BERT の事前学習である程度獲得されていると考えられる．

e） Natural Questions (NQ)

NQ は HTML フォーマットで書かれた Wikipedia を参照テキストとする [12]．参照テキストはトークンと HTML タグの系列として表現されている．我々は MRQA 2019 shared task が抽出型機械読解の訓練・評価データに加工したものをターゲットドメインとして利用した．参照テキストは前処理前の NQ の参照テキストを用いた．

Wikipedia ドメインは SQuAD，BERT の事前訓練双方に含まれるドメインであるが，NQ の評価によってドメイン適応手法がプレーンテキストから HTML フォーマットへの適応を可能にするかを調査できる．

5.2 実験設定

我々は No-adaptation Baseline，Sequential Model，提案手法を上記データセットで比較した．

実験には，BERT の PyTorch 実装 [2]を利用した．モデルの訓練は NVIDIA Tesla P100 GPU 4 枚を用いた．最適化アルゴリズムには Adam [11] を用いた．Warm-up Proportion を 0.1，学習率を 0.00005，バッチサイズを 32，エポック数を 3 とした．入力長を 384 に制限し，入力長よりも長い系列はストライド幅を 128 として分割した．その他のハイパーパラメータは $\text{BERT}_{\text{base}}$ の値と同一である．機械読解訓練と言語モデル訓練の比率 k は 10 とした．タスク特有層の数 n は 3 とした．以上のハイパーパラメータは全ての機械読解の Fine-Tuning で同じ値を使用した．Sequential Model の言語モデル訓練で利用するハイパーパラメータは，入力長を最大の 512 に設定し，他の値はデフォルト値を利用した．評価指標は完全一致（EM）と部分一致（F1）であり，これらは SQuAD の公式評価指標である．

5.3 議　　論

5.3.1 UDARC が有効となる条件は何か

表 3 に SQuAD からのドメイン適応のターゲットドメインにおける評価結果を，表 4 に NewsQA からの結果を示す.「ターゲットでの教師あり学習」の行はターゲットドメインの教師データを用いて訓練・評価を行った結果である．ターゲットでの教師あり学習の行は UDARC のタスク設定ではなく，教師なしドメイン適応手法の上界に相当することが期待される．初めに，各ターゲットドメインにおける性能を評価する．

a） BioASQ

BioASQ は本研究の目的に最も合致したデータセットである．両ドメインからの適応において，ドメイン適応モデルは，No-adaptation Baseline の性能を上回った．特に提案手法は SQuAD からの適応で 4.3/4.2 ポイント上回った．

BioASQ は BERT の事前学習コーパスに含まれていない医療ドメインのデータセットである．よって，本結果は UDARC が事前学習・Fine-Tuning 時に未知のドメインにおいて効果的であることを示す．また，本結果は UDARC における我々の

1 : Task 7b, Biomedical Semantic QA は ECML PKDD 2019 と共催された．詳細は http://BioASQ.org/ にある．

2 : https://github.com/huggingface/pytorch-transformers

	SQuAD での訓練	ドメイン適応	ターゲット NewsQA	BioASQ	DuoRC	NQ
ターゲットでの教師あり学習			41.5/56.0	—	20.2/27.2	58.9/72.2
No-adaptation Baseline	✓		35.2/50.7	41.1/53.6	24.5/33.0	**44.4**/**57.5**
Sequential Model	✓	✓	35.2/51.0	44.5/57.1	25.4/33.8	—
Multi-Task Model	✓	✓	**35.9**/**51.4**	**45.4**/**57.8**	**25.5**/**34.1**	43.8/56.7

表 3　ソースデータセットに SQuAD を用いた実験結果．セルの左側は EM，右側は F1.「ターゲットでの教師あり学習」の行はターゲットデータセットで教師あり訓練を行った結果であり，教師なしドメイン適応手法の上界に相当する．BioASQ は訓練データを持たないためターゲットでの教師あり学習が空欄になっている．NQ の Sequential Model が空欄となっているのは，NQ が HTML フォーマットであるため文章の区切りを定義できず，Sequential Model が必要とする NSP が行えないためである．

	NewsQA での訓練	ドメイン適応	ターゲット SQuAD	BioASQ	DuoRC	NQ
ターゲットでの教師あり学習			80.9/88.4	—	20.2/27.2	58.9/72.2
No-adaptation Baseline	✓		59.8/73.9	34.5/48.3	22.5/31.2	39.0/52.7
Sequential Model	✓	✓	59.7/75.3	36.6/**50.4**	23.7/**32.7**	—
Multi-Task Model	✓	✓	**60.6**/**75.8**	**36.8**/50.3	**23.8**/32.3	**42.0**/**56.2**

表 4　ソースデータセットに NewsQA を用いた実験結果．

仮定が正しいことを示唆している．即ち，BERT の事前学習は巨大なコーパスで行われているにも関わらず，ターゲットドメインにおける言語モデルの追加学習は，教師なしテキストから out-domain の参照テキストを理解する能力を獲得できる．

b） DuoRC

教師なしドメイン適応のモデルはターゲットドメインで訓練をしたモデルの性能を上回った [3]．この結果は，本実験で用いた DuoRC の ParaphraseRC タスクの難しさに起因していると考えられる．ParaphraseRC タスクには質問と参照テキストが共有する文字列が少ない，参照テキストが長い，といった通常の機械読解にない難しさがある．そのため，ParaphraseRC タスクの教師データは機械読解の能力を獲得するためには難しすぎたと考えられる．以上のことから，UDARC はターゲットドメインの教師データから直接機械読解の能力を獲得することが難しい場合にも有用であると考えられる．

c） Natural Questions

NQ では，4 つの設定（ソースドメインを SQuAD・NewsQA とした No-adaptation Baseline・提案手法）の実験を行った．結果，SQuAD からのドメイン適応では精度の向上がなかった．NewsQA からのドメイン適応では精度が向上した．提案手法はニュースドメインから Wikipedia ドメインへの適応には効果があるが，プレーンテキストから HTML フォーマットへの適応ができないと解釈できる．HTML フォーマットにはタグで区切られた区間に構造的な依存関係が存在するため，HTML フォーマットへの適応はドメイン適応よりも難しい課題であり，言語モデルのタスク設計を新たに考える必要があると言える．

3 つの閉じたドメインによる検証結果をまとめる．UDARC は，ターゲットドメインがソースドメイン・BERT の事前学習コーパスに含まれない未知ドメインである場合（BioASQ）及びターゲットドメインの教師データから機械読解の能力を学習することが困難な場合（DuoRC）に効果的であった．一方，HTML フォーマットへの適応は今後の課題となった（NQ）．

5.3.2　教師なしドメイン適応手法は精度を向上するか

ここでは，表 3，表 4 を用いて 3 つのモデルを比較する．NQ 以外では，ドメイン適応モデルが No-adaptation Baseline の性能を上回った．EM の指標に関しては提案手法は Sequential Model を全てのソース・ターゲットドメインの組み合わせで上回った．この傾向は，Sequential Model が out-domain の知識を Fine-Tuning 中に忘却している可能性を示唆している．しかしながら，提案手法と Sequential Model に有意な差はなかった．Sequential Model は提案手法同様に効果的であり，BERT の事前学習後にターゲットドメインで言語モデルの追加学習をすることは効果的であると言える．この知見は他の自然言語処理タスクにも適用可能である．

先行研究と比較すると，[30] は BERT を用いて SQuAD から NewsQA の適応の評価をした．彼らのドメイン適応手法での No-adaptation Baseline に比べた精度向上は 0.6/0.5 ポイントであり，我々の 0.7/0.7 の向上と同程度である．彼らの手法は擬質問・回答ペアを生成するため計算コストが大きく，我々の手法は低コストで同水準の精度向上を達成した．

5.3.3　ドメイン適応はソースドメインでの性能に影響を与えるか

ドメイン適応によってソースドメインでの性能がどの程度下がるかを評価した．表 5 と表 6 に結果を示す.「教師あり学習」の行はソースドメインのみで訓練を行った結果である．

結果として，提案手法と Sequential Model の性能はソースドメインでの教師あり学習の性能を上回る傾向が見られた．訓練データと評価データではデータ点の確率分布が同一であっても，サンプリングに起因する違いが存在する．本結果から，ド

3：ターゲットドメインでの教師あり学習の性能 20.2/27.2 は [22] で報告されている精度 19.7/27.6 と同等の精度を達成している．

	NewsQA	BioASQ	DuoRC	NQ
教師あり学習		80.9/88.4		
Sequential	81.2/88.6	80.6/88.4	81.0/88.4	—
Multi-Task	81.1/88.5	81.1/88.5	80.7/88.3	80.9/88.4

表 5　ソースドメイン（SQuAD）における評価結果．ターゲットドメインを各データセットにしてドメイン適応を行った．「教師あり学習」の行の結果は SQuAD で訓練をしてドメイン適応を行わなかったモデルである．

	SQuAD	BioASQ	DuoRC	NQ
教師あり学習		41.5/56.0		
Sequential	41.8/56.9	42.0/57.1	42.7/58.0	—
Multi-Task	42.6/57.6	42.0/57.0	42.8/57.8	42.3/57.4

表 6　ソースドメイン（NewsQA）における評価結果．

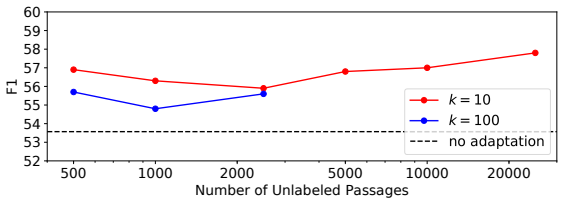


図 2　横軸を教師なしテキストの数，縦軸をターゲットドメインでの F1 とした SQuAD から BioASQ への適応結果．線の右端は利用可能なテキストの最大数である．言語モデルの訓練回数は機械読解の $1/k$ であるため，最大数は k と教師データ数に依存する．

メイン適応には一種の汎化の効果があり，訓練データのサンプリングへの過学習を抑えたと考えられる．UDARC は追加の教師ありデータを必要としないため，他の教師あり学習タスクに容易に適応可能であり，同様の効果によって精度の向上を期待することができる．

5.3.4　ドメイン適応に必要な教師なしテキストの数は幾つか

教師なしテキストの数の観点から提案手法の性能を評価した．実験は，本研究の主要データである BioASQ への適応で行った．図 2 に結果を示す．

教師なしテキストが 500 個しかない場合であっても，提案手法は No-adaptation Baseline の性能を上回った．精度の上昇は 3.7/3.3 ポイントである．機械読解と言語モデルの比率 k については $k = 10$ が高い精度を達成したが，事前の実験においてはデータセットに依存して $k = 100$ の方が高い精度を出すことが確認できており，一概には言えない．医療ドメインは BERT の事前学習コーパスと大きく異なるため，他のドメインに比べて言語モデルの向上の余地が大きく，言語モデルの訓練を多く行うことで精度が向上すると考えられる．モデルの性能は教師なしテキストの数が増えるにつれて向上するとは限らないが，最もよい精度は，利用可能な最大数である 26280 個のテキストを利用した場合に達成した．

5.3.5　ドメイン適応に適当なソースドメインは何か

ソースドメインの嗜好性に関して評価を行うため，訓練データの数を均等に制限してソースドメインを比較する評価を行った．図 3 と図 4 に提案手法で BioASQ と DuoRC への適応を行った結果を示す．

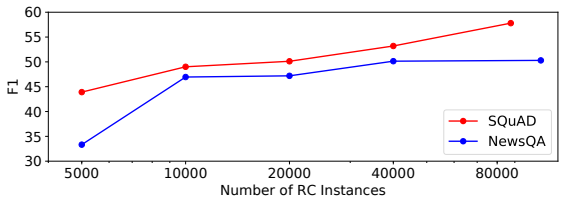


図 3　横軸を教師ありデータの数，縦軸をターゲットドメインでの F1 とした SQuAD，NewsQA から BioASQ への適応結果．

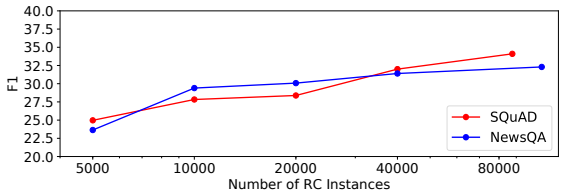


図 4　横軸を教師ありデータの数，縦軸をターゲットドメインでの F1 とした SQuAD，NewsQA から DuoRC への適応結果．

BioASQ では SQuAD からの適応が NewsQA からの適応を上回った．DucRC では，SQuAD と NewsQA からの適応は同程度の性能であった．よって，提案手法の性能はソースドメインの選択に依存するが，その選択指標は未解決である．

5.3.6　ソースドメインの教師データはどの程度必要か

図 3 と図 4 ではソースドメインの教師データの数に関して評価できる．結果，教師データの数が多いほど精度が高く，特に教師データが 10000 個に達するまでは精度の向上が著しいことがわかった．この結果は [33] の観察と一致する．彼らは BERT の Fine-Tuning に数万件規模のデータが必要であることを実験的に示した．UDARC においても同じ傾向が成り立つ．

6　おわりに

本研究は機械読解を教師データなしでターゲットドメインに適応させるタスク UDARC に取り組んだ．

本研究の独自性　本研究は質問に回答する能力を機械読解から，参照テキストを理解する能力をターゲットドメインの言語モデルから獲得することに初めて着目した研究である．この着眼点は，ターゲットドメインの擬質問・回答ペアを生成する先行研究と大きく異なる．また，機械読解と言語モデルのマルチタスク学習を行う手法を提案し，Fine-Tuning 中にターゲットドメインの知識を忘却することを緩和した．

本研究の重要性　5 つのデータセットを用いて，機械読解における教師なしドメイン適応が有効であることを示した．特に，医療ドメインでの提案手法は Wikipedia ドメインからの適応でドメイン適応なしモデルの性能を EM/F1 で 4.3/4.2 ポイント向上した（5.3.1.a 節）．またターゲットドメインの教師データが不十分な場合は，提案手法は通常の教師あり学習の性能を EM/F1 で 5.3/6.9 ポイント上回った（5.3.1.b 節）．先行研究と比較すると，提案手法は擬質問生成の計算コストなしに同程度の精度向上を達成した（5.3.2 節）．

これらの結果により，機械読解の学習データの作成コストが高い専門ドメインにおいても，テキストコーパスさえあれば機械読解モデルを学習可能であることを示した．本研究の成果は，チャットボットやコンタクトセンタの問い合わせ対応といった，産業上重要な課題に適用可能である．

また，事前学習済み言語モデル [3], [15], [20], [32] とそれらの Fine-Tuning は機械読解を含む多くの自然言語処理タスクで State-of-the-Art の性能を達成している．事前学習済み言語モデルの教師なしドメイン適応に取り組んだ本研究は，教師なしドメイン適応が適用されていない言語処理のあらゆるタスクに貢献できる．

文　献

[1] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, pp. 120–128, 2006.

[2] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.

[4] B. Dhingra, K. Mazaitis, and W. W. Cohen. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*, 2017.

[5] A. Fisch, A. Talmor, M. Seo, E. Choi, and D. Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *MRQA@EMNLP*, pp. 1–13, 2019.

[6] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[7] D. Golub, P.-S. Huang, X. He, and L. Deng. Two-stage synthesis networks for transfer learning in machine comprehension. In *EMNLP*, pp. 835–844, 2017.

[8] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NIPS*, pp. 1693–1701, 2015.

[9] M. Ishii, T. Takenouchi, and M. Sugiyama. Zero-shot domain adaptation based on attribute information. *arXiv preprint arXiv:1903.05312*, 2019.

[10] N. F. L. Johannes Welbl and M. Gardner. Crowdsourcing multiple choice science questions. In *W-NUT@EMNLP*, pp. 94–106, 2017.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *TACL*, 7:453–466, 2019.

[13] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale reading comprehension dataset from examinations. In *EMNLP*, pp. 785–794, 2017.

[14] X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding. In *ACL*, pp. 4487–4496, 2019.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[16] M. Mancini, S. R. Bulò, B. Caputo, and E. Ricci. Adagraph: Unifying predictive and continuous domain adaptation through graphs. In *CVPR*, pp. 6568–6577, 2019.

[17] T. Miller. Simplified neural unsupervised domain adaptation. In *NAACL-HLT*, pp. 414–419, 2019.

[18] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, 2016.

[19] K.-C. Peng, Z. Wu, and J. Ernst. Zero-shot deep domain adaptation. In *ECCV*, pp. 764–781, 2018.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *Technical Report OpenAi*, 2019.

[21] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, pp. 2383–2392, 2016.

[22] A. Saha, R. Aralikatte, M. M. Khapra, and K. Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *ACL*, pp. 1683–1693, 2018.

[23] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. In *ICLR*, 2017.

[24] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pp. 935–943, 2013.

[25] S. Šuster and W. Daelemans. CliCR: A dataset of clinical case reports for machine reading comprehension. In *NAACL-HLT*, pp. 1551–1563, 2018.

[26] I. Tenney, D. Das, and E. Pavlick. BERT rediscovers the classical nlp pipeline. In *ACL*, pp. 4593–4601, 2019.

[27] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. NewsQA: A machine comprehension dataset. In *Rep4NLP@ACL*, pp. 191–200, 2017.

[28] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, and G. Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

[30] H. Wang, Z. Gan, X. Liu, J. Liu, J. Gao, and H. Wang. Adversarial domain adaptation for machine reading comprehension. In *EMNLP-IJCNLP*, pp. 2510–2520, 2019.

[31] Y. Yang and T. Hospedales. Zero-shot domain adaptation via kernel regression on the grassmannian. *DIFF-CV@BMVC*, 2015.

[32] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[33] D. Yogatama, C. d. M. d'Autume, J. Connor, T. Kocisky, M. Chrzanowski, L. Kong, A. Lazaridou, W. Ling, L. Yu, C. Dyer, and P. Blunsom. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 2019.

[34] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pp. 19–27, 2015.

[35] Y. Ziser and R. Reichart. Task refinement learning for improved accuracy and stability of unsupervised domain adaptation. In *ACL*, pp. 5895–5906, 2019.