利用規約における QA データセットの作成及び検証

高野海里 [1]　プタシンスキ・ミハウ [1]　桝井文人 [1]
[1] 北見工業大学 テキスト情報処理研究室
f1912290090@std.kitami-it.ac.jp {michal, f-masui}@mail.kitami-it.ac.jp

概要

本研究では利用規約が読まれていないことで起こるユーザと企業の双方の問題やトラブルを指摘し，その問題を軽減するため，冗長な利用規約をわかりやすく自動要約する方法が有効であると考えられる．しかし，自動要約の評価には，人間評価者の応用が必要となり，近年公開されている数多くの言語モデルの内，どのモデルが最適しているかの評価が困難となっている．そこで，自動要約の評価を効率化させるために，文書を要約する前にその文書を理解しなければならないという考えのもと，質問応答のタスクを用いて最適な言語モデルを選抜する．その中で，3 つの質問応答データセットを用いて単言語・多言語両方種類の複数の言語モデルを微調整を行うことで，日本語の質問応答タスクのおいて優れた言語モデルと，優れたパラメーター（学習率等）について検証し，その結果について報告する．

1　はじめに

インターネットの普及に伴いさまざまなサービスが配信されており，それらの多くは利用開始時に利用規約に同意することが求められている．サービスの内容や社会的役割の増大に併せて文章量が増加・複雑化しているためか，利用規約を読まないユーザは多くいることが報告されている [1]．例として Twitter の利用規約[1] は 11,432 文字，Linkedin の利用規約[2] は 13,233 文字，Facebook の利用規約[3] は 15,852 文字と原稿用紙にして 30 枚以上もの文量である．これらの冗長な利用規約はサービスを利用するユーザだけでなく，サービスを提供する企業にとっても問題を生む可能性がある．ユーザにとっては利用規約を読まないことで詐欺などの危険や不利益を被る可能性があり，企業側にとっては禁止事項の伝達漏れ等によるトラブルを誘発する可能性がある．

このような問題を軽減する手法の一つとして，ユーザにとって読みやすくするために，冗長な利用規約をわかりやすく要約し，ユーザによる読解率を上げることである．しかし，企業の作業員が行う人手のみによる利用規約を要約する作業は，時間及び人件費がかかり，さらに利用規約の内容が法律に合わせて定期的に更新しなければならないことからすると，人手にて行うことは極めて非効率的である．その課題を解決するためには，文章の自動要約の技術を使うという方法が挙げられる．要約とは元の意味が変わらないように文章を短くすることであり，正確な要約を行うためにはその文章の意味や内容を正しく理解する必要がある．

そこで，我々はこのような要約時における入念な理解の重要性に着目し，言語モデルにあらかじめ要約対象文を理解[4] させることで自動要約の精度を向上できると考えた．そこで本研究では，要約対象文を理解した状態を「要約対象文に関する質問に正しく答えられる状態」と見なし，質問応答のタスクで学習させた言語モデルがどの程度の精度で自動要約を行えるかを調査する．本論文では具体的に質問応答の部分に焦点を当て，最も高い精度を出せる事前学習済み言語モデルを調査した結果を報告する．

2　関連研究

ユーザが利用規約を読まないことにより生じるリスクを低減するために，利用規約の読解を促進させる研究は過去にも行われてきた．例えば，竹ノ内ら [2] は，利用規約の表示方法に工夫を加えたが，ユーザの理解度に大きな改善が得られなかったことを報告している．さらに，抽出型要約の観点からは，野村ら [3] は，個人の属性に左右される重要文と一般に共通する重要文を利用規約から抽出している．ま

1） https://twitter.com/ja/tos
2） https://jp.linkedin.com/legal/user-agreement
3） https://ja-jp.facebook.com/legal/terms

4） ここの “理解”とは人間と同じ意味の言語理解のことではなく言語モデルにおいて文書を正確に処理を行うこと指す．

た，Doc2vec や SentenceBERT を用いて文章をベクトル化し，既知の利用規約との類似度が低い未知条項を抜き出すことで不利益をもたらしうる文章を抽出する研究も行われている [4, 5]．これらに対して我々のアプローチは言語モデルを用いた抽象型要約を視野に入れているため，より理解しやすい要約文の生成が期待できる．

3　基本的な概念の説明

本研究では Question Answering（質問応答，略：QA) と Automatic Summarization（自動要約，略：AS) の技術を組み合わせることで利用規約の文章量を減少させ，ユーザに読解を促進させることを最終目標とする．そこで具体的には，質問応答の技術を最適なモデルの選抜に用い，自動要約の技術における人手による評価を効率化させる．

質問応答とは，ある文章を入力に，その文章に関する質問への回答を出力するタスクである．タスクの応用例として，顧客から寄せられた質問に対する企業の回答の自動化 [6] や対災害システム [7] などが挙げられる．自動要約とは，ある文章において段落の意味を失わない程度に短くするタスクである．自動要約の応用例として，研究論文の要約による論文選択時間の短縮 [8] などが挙げられる．

そこで本研究は，図 1 に示したフローチャートのように以下 5 段階で研究を進行する．

1. 利用規約の文章に対して，要約対象文である「コンテキスト」「質問」「回答」からなる質問応答データセットを作成する．
2. 作成したデータセットを用いて複数の言語モデルを学習させ，QA タスクで評価し比較する．
3. 評価値が低い場合，データセットの調節（データの追加，質問再作成など）後 2. に戻ってプロセスを繰り返す．
4. 評価値の高い言語モデルを用いて，別に用意した利用規約の文章を要約させ，さらに SA タスクに応用し評価する．
5. 評価値が低い場合は，改めてデータセットの調節後 2. に戻ってプロセスを繰り返す．

4　実験

本実験では，8 つの言語モデルを 3 種類のデータセットで学習後，評価実験を行うことで，日本語の質問応答のタスクにおいて優れたモデルを調査する．また学習率を 2e-5 と 1e-4 に変化させ学習を行うことで，日本語の質問応答タスクにおいてどちらの学習率が優れているか調査する．また本実験で使用した実行環境は OS: Ubuntu 20.04.03 LTS，CPU: Intel Xeon E5-2687W v2 @ 3.40GHz，メモリ: 188GB，GPU: GeForce RTX 3060 である．

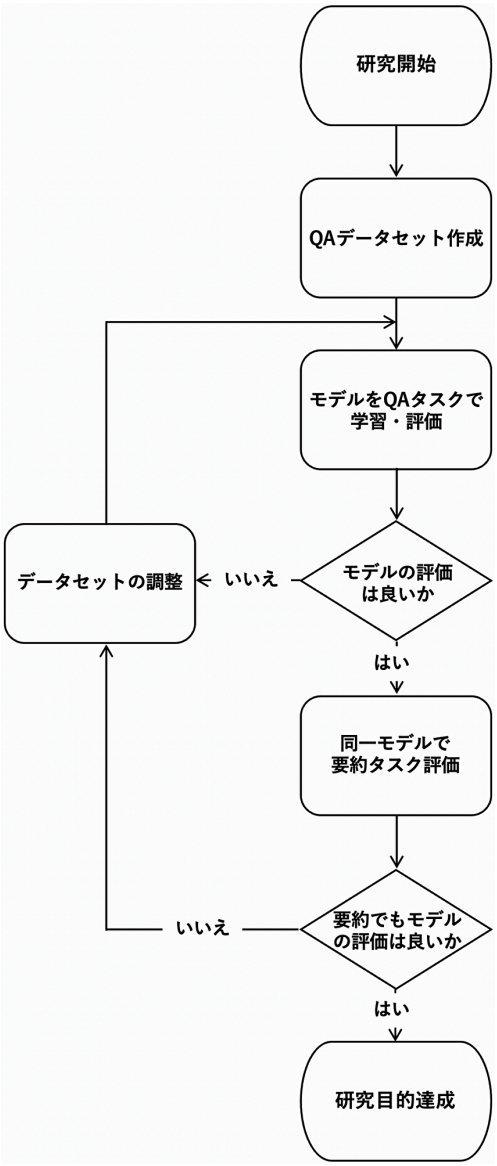

図 1　研究進行のフローチャート

4.1　言語モデル

本実験ではデータを Tokenize（トークン化）する際に，文中における文字の位置と範囲を示している return_offsets_mapping を使用でき，HuggingFace[5] にて公開されていて日本語に対応している 8 つの言語モデルを用いて質問応答タスクの学習を行っ

5）https://huggingface.co

た．使用した言語モデルの種類は BERT，DeBERTa，RoBERTa，DistilBERT の 4 つであり，多言語で事前学習された言語モデルを 4 つ，日本語のみで事前学習された言語モデルを 4 つ使用した．

4.2 データセット

JaQuAD JaQuAD (Japanese Question Answering Dataset)[9] は日本語機械読解のために作成された質問応答データセットであり，日本語 Wikipedia の記事から収集されたコンテキストに対し 39,696 の質問と答えで構成されている．コンテキストは哲学，歴史をはじめとした，さまざまなドメインにおける質の高い記事[6)7)]から取集されている．

JSQuAD JSQuAD は JGLUE (Japanese General Language Understanding Evaluation)[10] に構築されている質問応答データセットである．2021 年 11 月 1 日時点の日本語版 Wikipedia の記事から収集されたコンテキストに対し 72,721 の質問と答えで SQuAD 形式で構成されているまた本実験では SQuAD 形式から DatasetDict 型への変換を行うことで，他のデータセットと限りなく近い条件で使用している．

QADfT QADfT (Question Answering Dataset for Terms of Use)[8)] は JaQuAD を参考に，筆者自ら作成した利用規約特化の factoid 型質問応答データセットである．Twitter(147 個，参照日:2022/7/13)，Facebook(69 個, 参照日:2022/8/2)，ニコニコ (61 個, 参照日:2022/8/12)[9)]，LINE(100 個，参照日:2022/8/12)[10)]，Nintendo Switch(54 個，参照日:2022/8/12)[11)]，Nintendo WiiU(89 個，参照日:2022/8/16)[12)]，Nintendo 3DS(288 個, 参照日:2022/8/31)[13)]，Nintendo ネットワーク (192 個，参照日:2022/8/25)[14)]，の利用規約から収集したコンテキストに対し，合計 1,000 の質問と答えで構成されている．作成したデータセットの一例を 1 に示す．

表 1 作成したデータセットの一例

ID	FB-064
コンテキスト	2. ウイルスもしくは悪意のあるコードをアップロードすること、スパムを送信するためにサービスを利用すること、または弊社のサービス、システム、もしくは製品の正常な機能、保全、運用、もしくは表示を停止させる、妨げる、損なう、もしくはこれらに過負荷をかける可能性のあるその他一切の行為は禁止されています。
質問	何をアップロードすることは禁止されてるか？
回答	ウイルスもしくは悪意のあるコード

4.3 評価指標:SQuAD

SQuAD は質問応答における評価指標の一つである．評価に EM と F1 の二つのパラメータを用いる．EM は Exact Match の略で，言語モデルによって生成された答えと正解が完全一致しているかを示すものである．F1 は言語モデルによって生成された答えと正解がどれたけ近似しているかを示すものである．本指標における EM のスコア範囲は 0-100 であり，F1 のスコア範囲は 0-1 である．

4.4 結果と考察

各データセットを用いた評価実験の結果を表 2〜5 に示す．また，表の可読性向上のためスコアの右隣の丸括弧はスコアが高い順に 1,2,3... としており，上位 4 つのスコアを太字で示している．加えて，付録 A にて可読性向上のため表 2〜5 を棒グラフで表したものを図 2〜5 に示す．

双方の学習率，各データセットでスコアの比較をすると，**KoichiYasuoka/bert-base-japanese-wikipedia-ud-head**，**bert-base-multilingual-cased**，**bert-base-multilingual-uncased** の 3 つの言語モデルは常に上位 4 つのスコアに含まれることから，これらの言語モデルは日本語の質問応答タスクにおいて優れたモデルであるといえる．加えて，これら 3 つの言語モデルの特徴として，モデルの種類が BERT であることから，BERT モデルは他モデルより日本語の質問応答タスクにおいて優れていると考察できる．また，軽微な差はあれど全体的にスコアが高いことから，学習率は 1e-4 の方が優れているといえる．ただし，例外として **xlm-roberta-base** を QADfT

6) https://ja.wikipedia.org/wiki/Wikipedia:良質な記事
7) https://ja.wikipedia.org/wiki/Wikipedia:秀逸な記事
8) https://huggingface.co/datasets/umisato/QADfT
9) https://account.nicovideo.jp/rules/account
10) https://terms.line.me/line_terms?lang=ja
11) https://www.nintendo.co.jp/support/switch/eula/usage_policy.html
12) https://www.nintendo.co.jp/support/wiiu/pdf/wiiu_eula.pdf
13) https://www.nintendo.co.jp/support/information/2022/pdf/Nintendo3DS_EULA_20220830.pdf
14) https://www.nintendo.co.jp/support/nintendo_network/eula/index.html

で学習させた場合のみ，学習率が 2e-5 でのスコアが大幅に高くなった．これに関して詳しい原因は不明であるが，他のデータセットと比べて QADfT のデータ数が 1000 件と少なく，結果においてはっきりとしたパターンが現れにくかったことが考えられる．

さらに，単言語のモデルの中，青空文庫で事前学習したモデルは全て結果が低く，Wikipedia を元に事前学習したモデルは，Wikipedia の QA データセットで比較的に高い結果となっていることから，事前学習言語モデルにおけるドメイン依存性のことを示していると考えられ，従来研究でも指摘されている現象である [11][12][13].

そして，単言語のモデルと比べて，多言語のモデルの方が大幅結果が高く，事前学習言語モデルでは言語間の知識共有のことを示しており，従来研究で指摘されていることが再確認されている [12][13].

総合的に判断した結果，研究の次段階では利用規約の自動要約に用いるモデルを選抜した．そこで，まず，総合的に結果が高かった **bert-base-multilingual-cased** を用いることにした．同じモデルの「-uncased」バージョンも結果が高かったが，モデルの種類が似ているため，最高結果のもののみを用いることにした．さらに，複数の種類のモデルの性能を比較するため，**xlm-roberta-base** も用いることにした．そして，前者の 2 モデルが多言語のモデルのため，単言語のモデルの性能も比較するためには **KoichiYasuoka/bert-base-japanese-wikipedia-ud-head** も用いることにした．

5 おわりに

本研究では利用規約が読まれていない社会背景に対して，質問応答と自動要約を合わせて用いることで自動要約の精度を上げつつ，ユーザに利用規約の読解を促せるかを考えた．その初期段階として，3 つの質問応答データセットを用いて，8 種類の言語モデルの微調整を行うことで，日本語の質問応答タスクのおいて優れた言語モデルと，優れた学習率を選抜することができた．具体的な評価実験の結果として，8 つの言語モデルから日本語の質問応答タスクにおいて優れたものを得ることができ，ほとんどの言語モデルの学習率は 1e-4 の方が優れていることが分かった．しかし一部の言語モデルとデータセットの組み合わせでは，学習率が 2e-5 の方が優れていた．結果から BERT モデルは他モデルに比べ日本語の質問応答タスクにおいて優れていると考察できる．また，一部の言語モデルに関しては，組み合わせたデータセットのデータ数が 1000 件と少なかったがため，学習率が 2e-5 の方が優れた結果になったのではないかと考察できる．今後の課題として，学習率が 2e-5 の方が適していた言語モデルに組み合わせたデータセットのデータ数を増やした後，変化が見られるかを再度検証する．また，4.4 で述べた 3 つの言語モデルを用いて自動要約を行う．

表 2 評価実験の結果 (EM, learning late=2e-5)

MODEL NAME	QADfT	JaQuAD	JSQuAD
KoichiYasuoka/bert-base-japanese-wikipedia-ud-head	**60.0(4)**	**71.8(2)**	**74.1(2)**
KoichiYasuoka/deberta-base-japanese-aozora-ud-head	1.0(8)	52.0(8)	67.6(8)
KoichiYasuoka/deberta-base-japanese-wikipedia-ud-head	4.0(7)	61.6(6)	70.8(5)
KoichiYasuoka/roberta-base-japanese-aozora-ud-head	14.0(6)	54.7(7)	67.4(7)
bert-base-multilingual-cased	**76.0(1)**	**74.3(1)**	**74.4(1)**
bert-base-multilingual-uncased	**73.0(2)**	**71.6(3)**	**72.5(3)**
distilbert-base-multilingual-cased	41.0(5)	64.7(5)	69.2(6)
xlm-roberta-base	**71.0(3)**	**66.8(4)**	**71.6(4)**

表 3 評価実験の結果 (EM, learning late=1e-4)

MODEL NAME	QADfT	JaQuAD	JSQuAD
KoichiYasuoka/bert-base-japanese-wikipedia-ud-head	**75.0(1)**	**75.7(1)**	**76.8(1)**
KoichiYasuoka/deberta-base-japanese-aozora-ud-head	1.0(8)	61.2(8)	70.7(7)
KoichiYasuoka/deberta-base-japanese-wikipedia-ud-head	33.0(5)	67.8(5)	**73.2(3)**
KoichiYasuoka/roberta-base-japanese-aozora-ud-head	20.0(6)	66.1(7)	71.5(6)
bert-base-multilingual-cased	**75.0(1)**	**73.6(2)**	**74.1(2)**
bert-base-multilingual-uncased	**72.0(3)**	**71.8(3)**	**72.9(4)**
distilbert-base-multilingual-cased	**66.0(4)**	**68.7(4)**	70.5(8)
xlm-roberta-base	20.0(6)	67.1(6)	71.6(5)

表 4 評価実験の結果 (F1, learning late=2e-5)

MODEL NAME	QADfT	JaQuAD	JSQuAD
KoichiYasuoka/bert-base-japanese-wikipedia-ud-head	**0.815(4)**	**0.809(3)**	**0.857(3)**
KoichiYasuoka/deberta-base-japanese-aozora-ud-head	0.147(8)	0.631(8)	0.793(8)
KoichiYasuoka/deberta-base-japanese-wikipedia-ud-head	0.224(7)	0.710(6)	0.820(6)
KoichiYasuoka/roberta-base-japanese-aozora-ud-head	0.342(6)	0.692(7)	0.804(7)
bert-base-multilingual-cased	**0.921(1)**	**0.837(1)**	**0.867(2)**
bert-base-multilingual-uncased	**0.921(1)**	**0.816(2)**	**0.849(4)**
distilbert-base-multilingual-cased	0.660(5)	0.758(5)	0.826(5)
xlm-roberta-base	**0.888(3)**	**0.799(4)**	**0.868(1)**

表 5 評価実験の結果 (F1, learning late=1e-4)

MODEL NAME	QADfT	JaQuAD	JSQuAD
KoichiYasuoka/bert-base-japanese-wikipedia-ud-head	**0.890(1)**	**0.836(1)**	**0.878(1)**
KoichiYasuoka/deberta-base-japanese-aozora-ud-head	0.131(8)	0.716(8)	0.821(8)
KoichiYasuoka/deberta-base-japanese-wikipedia-ud-head	0.539(5)	0.765(6)	0.841(5)
KoichiYasuoka/roberta-base-japanese-aozora-ud-head	0.419(6)	0.759(7)	0.838(6)
bert-base-multilingual-cased	**0.890(1)**	**0.828(2)**	**0.863(3)**
bert-base-multilingual-uncased	**0.886(3)**	**0.812(3)**	**0.852(4)**
distilbert-base-multilingual-cased	**0.834(4)**	0.791(5)	0.837(7)
xlm-roberta-base	0.419(6)	**0.801(4)**	**0.867(2)**

参考文献

[1] 公正取引委員会. デジタル広告の取引実態に関する中間報告書. Technical report, 公正取引委員会, 4 2020.

[2] 竹ノ内朝陽, 矢谷浩司. サービス利用規約の読解促進を目指した表示手法の比較検討. **FIT**, Vol. 3, pp. 57–64, 2020.

[3] 佳太野村, 司前川, 晃三水谷, 正之荒井. 利用規約等における重要文の抽出手法の検討. 第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 575–576, 03 2016.

[4] 竹ノ内朝陽, 矢谷浩司. 利用規約中の特異な文章を強調するインタフェース. 電子情報通信学会全国大会. The Institute of Electronics, Information and Communication Engineers, 3 月 2021.

[5] 大晟森口, 健二中村. 機械学習を用いた利用規約からの未知条項の抽出に関する研究. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 34, No. 0, pp. 861–862, 2018.

[6] 佐藤理, 太田竜男, 佐藤雪乃, 松本桂子, 田澤和彦. 事例データベースを利用した質問応答システムの構築. 全国大会講演論文集, 第 49 回, ソフトウェア, pp. 175–176, 09 1994.

[7] 後藤淳, 大竹清敬, 橋本力, 川田拓也, 鳥澤健太郎. 質問応答に基づく対災害情報分析システム. 自然言語処理, Vol. 20, No. 3, pp. 367–404, 2013.

[8] Jinghong LI. 観点を反映した深層学習および強化学習による学術論文の自動要約生成. 2021.

[9] ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. Jaquad: Japanese question answering dataset for machine reading comprehension. **arXiv preprint arXiv:2202.01764**, 2022.

[10] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 29, No. 2, pp. 711–717, 2022.

[11] 柴田祥伍, プタシンスキミハウ, エロネンユーソ, ノヴァコフスキカロル, 桝井文人. 日本語大規模ブログコーパス yacis に基づいた electra 事前学習済み言語モデルの作成及び性能評価. 言語処理学会第 28 回年次大会 (NLP2022), pp. 285–289, 2022.

[12] Juuso Eronen, Michal Ptaszynski, Fumito Masui, Masaki Arata, Gniewosz Leliwa, and Michal Wroczynski. Transfer language selection for zero-shot cross-lingual abusive language detection. **Information Processing & Management**, Vol. 59, No. 4, p. 102981, 2022.

[13] Juuso Eronen, Michal Ptaszynski, and Fumito Masui. Zero-shot cross-lingual transfer language selection using linguistic similarity. **Information Processing & Management**, Vol. 60, No. 3, p. 103250, 2023.

A 付録

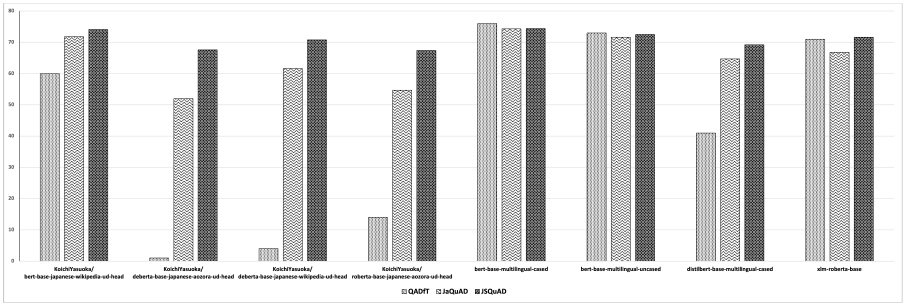

図 2 EM (learning late=2e-5)

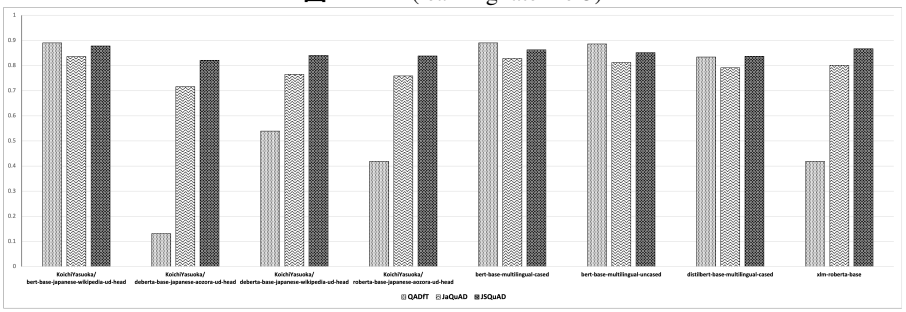

図 3 EM (learning late=1e-4)

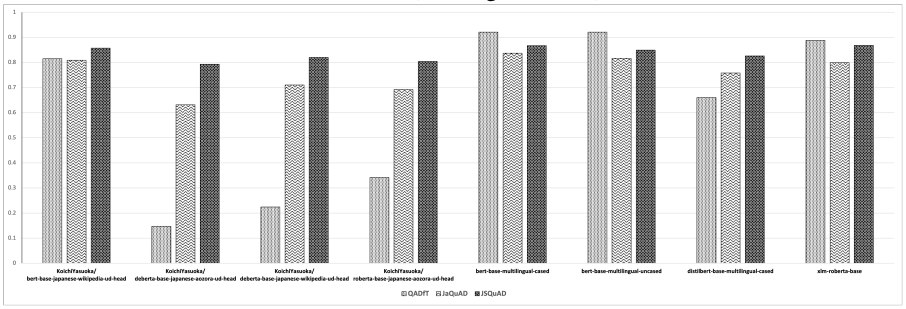

図 4 F1 (learning late=2e-5)

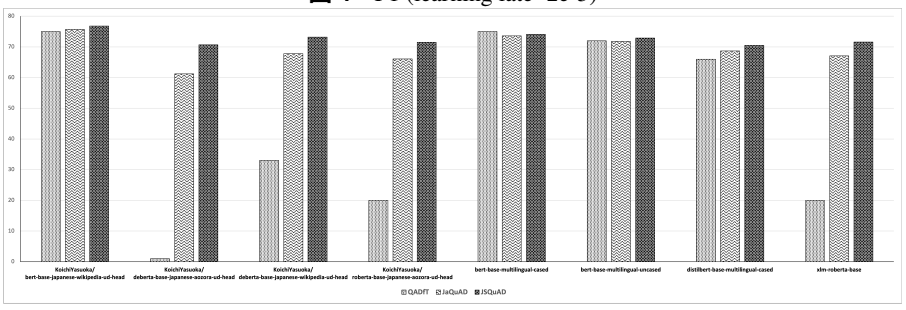

図 5 F1 (learning late=1e-4)