実行時機械翻訳による多言語機械読解

浅井明里 †, 江里口瑛子 ¤, 橋本和真 ‡, 鶴岡 慶雅 †
† 東京大学 ¤Microsoft Research ‡Salesforce Research
†akari-asai@g.ecc.u-tokyo.ac.jp,tsuruoka@logos.t.u-tokyo.ac.jp
¤Akiko.Eriguchi@microsoft.com
‡k.hashimoto@salesforce.com

1 序論

与えられた段落を読み解き質問に回答する機械読解 (reading comprehension) は検索エンジンや AI アシスタントを始め幅広い応用可能性があり，SQuAD [11] 等の大規模なデータセットを活用したニューラルネットワークによる機械読解 (ニューラル機械読解) モデルが目覚ましい発展を遂げている [4, 10, 12]. 一方で，これらの大規模データセットはしばしば英語に限定され，図 1 で示されるような，非英語言語で機械読解を行うニューラルモデルの学習を困難にしている. 各言語で新たにデータセットを作成する努力がなされてきたが [3, 7]，追加的な大規模データセットの作成コスト，及び英語における既存の言語資源やモデルを十分に活用できない等の点に課題がある.

本研究では，機械読解学習データセットの存在しないターゲット言語において機械読解システムを構築するため，ニューラル機械読解モデルとニューラル機械翻訳 (Neural Machine Translation, 以下 NMT) モデルを組み合わせた「実行時機械翻訳による多言語機械読解」を考案する. 提案手法はターゲット言語での機械読解学習データを一切必要とせず，また既存の英語における機械読解モデルと柔軟に組み合わせることができるという点に特徴がある．評価実験では，機械読解の代表的なタスクである SQuAD に焦点を当て，日仏の 2 言語で新たに評価データセットを作成した．両言語において，提案手法が回答の再翻訳を用いたベースライン手法を大幅に上回る性能を達成することを確認した. コード及び作成したデータセットは全て公開されている[1].

2 実行時機械翻訳による多言語機械読解

図 2 に提案手法の全体像を示す．提案手法は (1) ターゲット言語 L で与えられた入力を，アテンション付き NMT モデルを用いて，大規模な学習データセットが存在するピボット言語 P に翻訳し，(2) 事前に学習されたピボット言語 P におけるニューラル機械読解モデルに入力する．その後，(3) 予測された回答を NMT の重み付きアラインメント情報を用いて言語 L の段落内のフレーズと対応付け，回答を抽出する.

2.1 ピボット言語への翻訳

提案手法では，翻訳モデルにおける内部情報を機械読解モデルに効果的に活用するため，既存の翻訳サービス (e.g., Google 翻訳) を用いるのではなく，アテンション付き NMT [8] により，ターゲット言語 L で与えられた質問文及び段落をピボット言語 P に翻訳する. アテンション付き NMT は双方向の Recurrent Neural Network (RNN) によるエンコーダと，アテンション機構を持つ単方向の RNN によるデコーダで構成される.

[1] https://github.com/AkariAsai/extractive_rc_by_runtime_mt

ワルシャワで生まれた最も有名な人の一人は、放射能に関する研究 で国際的に認められ、**ノーベル賞**を受賞した最初の女性である マリー・キュリーだった。ワルシャワ出身の**有名な音楽家**には ウワディスワフ・シュピルマンとフレデリック・ショパンが 含まれる。ショパンはワルシャワから約60kmのŻelazowaWolaで 生まれたが、彼は**7ヶ月**の時に家族と共に市内に移住した。

マリー・キュリーが女性で最初の受賞者となったのは何か？
ノーベル賞
フレデリック・ショパンとは誰か？
有名な音楽家
ショパンは何歳の時に家族とワルシャワに引っ越したか？
7ヶ月

図 1: 日本語 SQuAD データセットの段落での質問文及び回答ペアの一例.

系列長 T の入力系列が与えられたとき，i 番目の単語に対応するエンコーダの隠れ状態を $h_i \in \mathbb{R}^{d_1}$ とし，同様に j 番目の出力語を生成するデコーダの隠れ状態を $\tilde{h}_j \in \mathbb{R}^{d_2}$ とする．ここで d_1, d_2 はそれぞれエンコーダ及びデコーダの隠れ状態のサイズを表す．提案手法では bilinear attention を用い入力系列の i 番目と出力系列の j 番目の単語間のアテンションスコア α_{ij} を以下のように計算する:

$$\alpha_{ij} = \frac{\exp\left(h_i W \tilde{h}_j\right)}{\sum_{k=1}^{T} \exp\left(h_k W \tilde{h}_j\right)}, \tag{1}$$

ここで W は学習されるパラメータ行列である．これにより，ターゲット言語における段落 C_L と質問文 Q_L はピボット言語 P で記述された段落 C_P 及び質問文 Q_P にマップされる.

2.2 ピボット言語における機械読解

ピボット言語 P に段落及び質問文ペアを翻訳することにより，言語 P で存在する大規模コーパスで学習された機械読解モデルを適用することが可能になる．抽出型の機械読解タスクにおいては，回答は段落中の任意の連続した文字範囲 (スパン) とされ，モデルは回答を抽出するために，回答の開始位置及び終了位置を予測する．段落中の i 番目の単語が回答開始・終了位置である確率分布 $p_s(i)$ 及び $p_e(i)$ としたとき，$p_s(i)$ と $p_e(i)$ の同時確率が最大となる開始位置 s 及び終了位置 e を以下の式に則って計算する:

$$(s, e) = \underset{(m,n),\, m \leq n}{\arg\max}\ p_s(m) p_e(n). \tag{2}$$

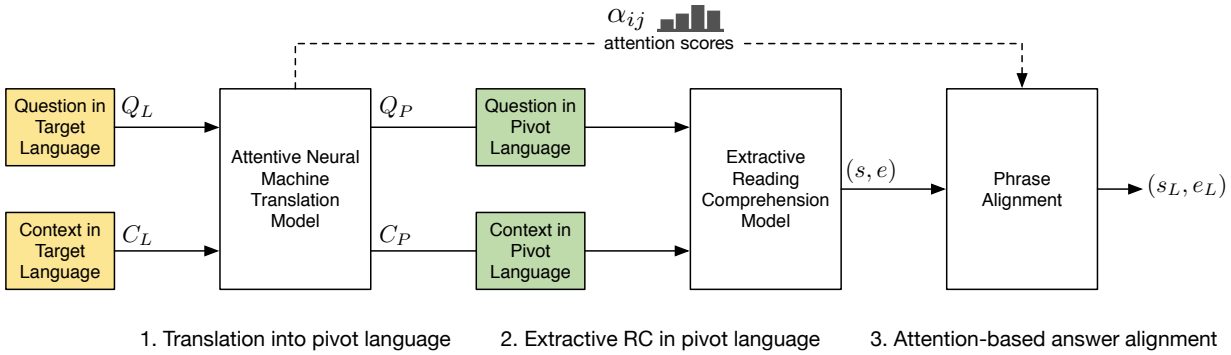

図 2: 提案手法の全体像. $\boldsymbol{\alpha}_{ij}$ は NMT のアテンションウェイトであり，(s, e) 及び $(s_{\mathrm{L}}, e_{\mathrm{L}})$ はピボット言語 P (e.g., English) とターゲット言語 L における回答の開始・終了位置をそれぞれ示している．

2.3 ターゲット言語への回答のアラインメント

言語 P で与えられた回答から言語 L における回答を獲得するため，言語 P から言語 L へ翻訳を行う新たな翻訳システムを用い，予測された回答を再翻訳する方法が考えられる．しかしこの手法は言語 L で与えられた段落及び質問文の情報にグラウンドされない翻訳を行うために，元の段落 C_p における表記等と異なる翻訳結果を生成してしまうことが懸念される．

NMT のアテンションの重み α_{ij} は予測された j 番目の単語にとって，入力系列の i 番目の隠れ状態がどの程度影響を与えているかの推定値を示すことが知られている [1, 8]．提案手法はこのアテンションの重みを活用する．言語 P に翻訳された段落 C_P から，抽出された回答に含まれる j 番目の単語それぞれを，式 3 を用いて，ターゲット言語 L で記述された段落 C_L における対応する $\ell(j)$ 番目の単語にアラインすることで言語 L での回答を選択する．

$$\ell(j) = \underset{i,\, 1 \le i \le T}{\arg\max}\, \alpha_{ij}. \tag{3}$$

最終的にターゲット言語 L で記述された段落 C_L での回答の開始位置及び終了位置である (s_P, e_P) は以下のように計算される:

$$s_L = \min\{\ell(s), \ell(s+1), \cdots \ell(e)\}, \tag{4}$$
$$e_L = \max\{\ell(s), \ell(s+1), \cdots \ell(e)\}. \tag{5}$$

3 日仏 SQuAD データセットの作成

提案手法の有効性を検証するため，日本語及びフランス語の 2 言語において，公開されている SQuAD v1.1[2]の開発用データセット[3]を基に，SQuAD 形式の評価データセットをクラウドソーシングにより新たに作成した．

具体的な評価データセットの作成手順としては以下の通りである．SQuAD V1.1 の開発用データセットは 48 の Wikipedia 記事に含まれる 2,067 段落に基づき，合計で 10,570 の質問文及び段落のペアにより構成されている．本研究では，まず 48 の記事のそれぞれ最初の段落及びそれに付随する質問文を抽出し，質問文及び段落については Amazon Mechanical Turk[4] 上でバイリンガル話者により，英語からターゲット言語 (i.e., 日本語及びフランス語) に翻訳し，その後翻訳結果の正当性の検証及び対応する回答のスパンの抽出をバイリンガル話者により行った．この結果，一般名詞 (e.g., 教師) から科学技術的な概念 (e.g., 葉緑体) まで広範なトピックを扱う，327 の段落・質問文ペアから構成される日仏 SQuAD データセットを新たに構築した．

4 多言語機械読解のための機械翻訳実験

まず，提案手法において非常に重要な役割を担う，言語 L から言語 P への翻訳を行う NMT モデルを，多言語機械読解への応用を見据えた方法で学習させ，評価結果を報告する．以降，ピボット言語 P は英語，またターゲットとする言語 L については日本語及びフランス語を想定する．

4.1 Wikipedia に基づく対訳コーパスの自動作成

予備実験において，日英対訳科学技術対訳コーパスである ASPEC [9] を用いて学習された NMT モデルでは，取り扱うドメインの相違等により，日本語 SQuAD データセットでの翻訳において極めて低い翻訳性能を示すことを確認した．また，広範なドメインを取り扱う高品質な対訳コーパスの存在を仮定することは，これらの対訳コーパスの存在しない言語への提案手法の適用可能性を著しく低下させてしまう点も懸念される．

本研究では，ターゲットとする言語の Wikipedia 記事, 及び Wikipedia の言語間リンク[5]を用いて取得された対応する英語の Wikipedia 記事に含まれる文に対し，文単位のアラインメントツール[6]を用いて大規模な対訳コーパスを自動構築した．日本語及びフランス語それぞれで，英語記事に含まれる文とのアラインメントスコアが高い上位 1,002,000 文対を抽出し，これを 1,000,000 文対の学習データセット及び 2,000 文の評価データセットに分割した．

4.2 質問文の翻訳

NMT モデルは汎化のために大量の学習データを必要とし，低頻度の入力に対して適切に学習できないことが知られている [14]．Wikipedia 対訳コーパスにおいて，質問文の占める割合は 0.1% と著しく低いため，このコーパスのみを用いて NMT モデルを学習させた場合，質問文が平叙文として翻訳されてしまう等，正しく翻訳できない傾向が確認された．本研究では，以下の 2 つのアプローチで質問文翻訳性能の改善を試みた．

[2] https://github.com/rajpurkar/SQuAD-explorer.
[3] SQuAD の評価データセットは一般に公開されていないため，本研究では開発データセットより，評価データを作成した．
[4] https://www.mturk.com/
[5] https://en.wikipedia.org/wiki/Help:Interlanguage_links
[6] https://github.com/danielvarga/hunalign

Translation method	Ja-En		Fr-En	
	Wiki	**Question**	**Wiki**	**Question**
Our NMT	23.95	22.75	45.64	40.47
Google Translate	24.09	37.98	41.08	50.91
Bing Translator	23.61	30.47	47.41	55.88

表 1: {Japanese, French}-to-English NMT の Wikipedia 評価データセット (Wiki) 及び SQuAD の質問文翻訳 (Question) の BLEU スコア.

Translation method	Ja-En		Fr-En	
	Wiki	**Question**	**Wiki**	**Question**
Our NMT	23.95	22.75	45.64	40.47
w/o beam search	20.78	23.06	41.93	36.21
w/o question oversampling	20.76	16.94	42.05	35.03
w/o questions	20.36	10.68	41.37	22.75

表 2: NMT の BLEU スコアに対する ablation study の結果．順に 1) ビームサーチを除き, 2) 質問文対訳対をオーバーサンプリングせず利用し, 3) 質問文対訳対を全く利用しない.

人手による少量の質問文対訳ペアの追加　SQuAD の学習データセットから 200 文をランダムに抽出し，Amazon Mechanical Turk を用いて日本語及びフランス語の翻訳を付与し，この少量の質問文翻訳対を対訳コーパスに追加した．

オーバーサンプリング　学習リソースの限られたドメインにおいて，少量の学習データをオーバーサンプリングすることで NMT の翻訳性能が向上することが確認されている [2, 6]．本研究ではこの手法を適用し，上述の少量の質問文翻訳対を l 回重複させ，Wikipedia に基づ居く対訳コーパスと混合させることにより，学習用対訳コーパスを作成した．

4.3 段落及び質問文翻訳に関する実験結果

表 1 に，上述の学習コーパスを用いて学習した NMT モデル，Google Translate[7] 及び Microsoft Translator Text API v3 (Bing Translator)[8] のそれぞれの BLEU スコアを示す. 表で示された結果より，Wikipedia 記事の翻訳に関しては，Wikipeida から自動生成された対訳コーパスで学習された NMT モデルが，2 つの商用翻訳エンジンと比較し遜色ない翻訳性能を示していることが確認できる．一方で，疑問文の翻訳に関しては，両言語とも，より多くの質問文を含むと予想される大規模なコーパスで学習された商用翻訳エンジンが我々の NMT モデルと比較し一貫して高い性能を示している．

次に，NMT モデルに対して ablation study を行った結果を表 2 に示す．Wikipedia の段落翻訳性能はビームサーチの有無を除きほぼ一貫した性能を示している一方で，質問文の翻訳に関しては，僅か 200 文の手動で作成された翻訳結果を追加しない場合 (表中の “w/o questions”)，BLEU スコアがおよそ半減することが確認できる．この結果は最小限の追加的な人手での翻訳データを適切に活用することにより，重要な質問文翻訳の性能を大きく改善できることを示唆している．

[7]翻訳結果は 2018 年 8 月時点の Google Translate API `https://translate.google.com` を介して取得した.

[8]翻訳結果は 2018 年 12 月時点の Microsoft Translator Text API v3 `https://docs.microsoft.com/en-us/azure/cognitive-services/translator/` を介して取得した.

Method	Japanese		French	
	F1	**EM**	**F1**	**EM**
Our method	**52.19**	**37.00**	**61.88**	**40.67**
Back-translation by				
using our NMT models	22.02	7.65	42.94	19.57
using Google Translate	42.60	24.77	44.02	23.54
using Bing Translator	35.35	13.15	50.76	18.65

表 3: 日仏 SQuAD データセットにおける，提案手法及び比較手法の機械読解の結果.

5 多言語機械読解実験

4.3 節で最も高い性能を示した NMT モデルを用いた機械読解の日本語及びフランス語 SQuAD での実験の結果を示す．本稿では，BiDAF [12] 及び BiDAF + Self Attention + ELMo [10] の 2 つの機械読解モデルを予め SQuAD の学習データセットで学習させた．評価指標は Exact Match (EM) 及び文字単位での適合率及び再現率の調和平均である F1 を用いた [11]．BiDAF については F1 77.1, EM 67.2 を，BiDAF + Self Attention + ELMo については F1 83.2 及び EM 74.7 を達成した．

比較手法　本研究はターゲットとする言語において学習データセットを使用しない問題設定において機械読解システムを構築する初めての試みであるため，直接的に性能比較が可能な手法等が存在しない．そのため，本稿における比較手法として，予測された回答を言語 P から言語 L に直接再翻訳し，回答を行う back-translation を用いる．回答の再翻訳には，4.1 節の Wikipedia 対訳コーパスで学習した NMT モデル，Google Translate 及び Bing Translator を用いた．

5.1 実験結果

表 3 に提案手法及び比較手法の日仏 SQuAD 評価データセットにおける実験結果を示す．提案手法が，翻訳性能で上回る翻訳システムを用いた比較手法を大幅に上回り，最も高い F1 及び EM スコアを獲得している．これは NMT モデルの内部情報を効果的に活用することが，学習データが存在しない状況下での多言語機械読解システムの構築に有効であることを示している．

また，表 4 に提案手法の ablation study の結果を示す．日仏 SQuAD において，Wikipedia に基づく対訳コーパスのみで NMT を学習させた場合 (表中の “w/o questions”)，機械読解システムの性能は日本語において F1 スコアが 26.99，EM スコアが 22.37，フランス語において F1 スコアが 20.25，EM スコアが 14.07 と大幅に悪化していることが確認できる．これは質問文翻訳性能が多言語での機械読解システムの構築に大変重要であり，最小限の追加的なアノテーションにより，多言語機械読解システムの性能向上が達成できることを再度強調している．一方で，質問文翻訳において我々の NMT モデルを上回る性能を示した Google Translate を用いて質問文を翻訳した場合 (表中の “w/ Google Translate for question translation”)，性能が却ってやや悪化した．これはドメインに特化した方法で質問文翻訳の性能を向上させることの重要性を示唆する．

5.2 エラー分析

次に，提案手法の日本語及びフランス語 SQuAD データセットにおけるエラー分析の結果を示す[9]．表 5 は

[9]エラー分析にあたって，機械読解モデルの性能に拠るエラーを除外するため，評価データセットのうち，英語で評価した際にモデルが回答に失敗した 41 質問ペア (評価データセット全体の 13%)

Method	Japanese		French	
	F1	EM	F1	EM
Our method	52.19	37.00	61.88	40.67
w/o self attention	50.08	35.47	57.56	37.61
w/o ELMo	50.59	34.55	55.14	36.69
w/o beam search	33.97	20.48	49.28	29.66
w/o question oversampling	25.20	14.63	41.63	26.60
w/o questions	51.52	36.09	61.17	40.67
w/ Google Translate for question translation				

表 4: 日仏 SQuAD データセットにおける，提案手法の ablation study 結果．順にベストのモデルから 1) 機械読解モデルを BiDAF に変更し, 2) NMT よりビームサーチを除き, 3) オーバーサンプリングを行わず質問文を追加し, 4) 質問文を追加しない．また質問文の翻訳にのみ Google Translate を用いた場合も検証した．

Type of Errors	Japanese # (%)	French # (%)
Wrong question translation	29 (59%)	15 (54%)
Wrong context translation	27 (55%)	11 (39%)
Others	6 (12%)	6 (21%)

表 5: 日本語及びフランス語データセットにおける提案手法のエラー分析結果．

ランダムに抽出された 100 の段落質問文ペアに対し，人手評価により，エラーを (1) 誤った質問文翻訳に由来するエラー (Wrong question translation), (2) 誤った段落翻訳に由来するエラー (Wrong context translation), (3) それ以外 (Others) の 3 つのカテゴリに重複を許容して分類した結果である．

SQuAD モデルは段落と質問文の間の n グラムマッチや質問文の疑問詞等，表層的なヒューリスティクスに依存することが報告されており [4, 5, 13]，本研究においても，質問文に含まれる単語一つが誤って翻訳される等が致命的なエラーに繋がるケースが複数観測され，これはタイプ (1) のエラーに分類される．

またタイプ (2) のエラーに関しては，主に回答の含まれる段落中の一部分が翻訳時に翻訳されず失われてしまう「訳抜け」により，回答となる部分が翻訳後の段落 C_p に存在しないために機械読解モデルが正しい回答を抽出できない例が確認された．

タイプ (3) に分類されたエラーの多くは機械読解モデル言い換え表現に対する頑健性の欠如に起因するものが複数観測された．図 3 はこの一例を示す．この例では，日本語 SQuAD から英語への翻訳に際して，質問文を意味的･文法的に誤って翻訳しているにも関わらず，機械読解モデルは正しい回答の抽出に成功している．一方で，フランス語 SQuAD では，質問文及び段落共に正しく翻訳しているが，機械読解モデルは回答において手掛かりとなる “spread” が翻訳時に “diffusion” と言い換えられたことにより，誤った回答を抽出している．

6 結論

本研究では，学習データセットの豊富なピボット言語での機械読解モデル及びアテンション付き機械翻訳モデルを組み合わせた，学習データセットの存在しない言語における機械読解システムを提案した．新たに作成した日仏 SQuAD における実験結果は，提案手法が最先端の翻訳システムを用いた back-translation 手法より大幅な性能向上を達成することを示した．

[The Original SQuAD Dataset]
Paragraph: It has also allowed for the rapid spread of technologies and ideas.
Question: Imperialism is responsible for the rapid spread of what?
Answer: technologies and ideas

[Translated results of French SQuAD]
Paragraph: This also allowed the quick diffusion of technologies and ideas .
Question: what imperialism spread ?
Answer: colonisation , l' utilisation de la force militaire ou d' autres moyens (colonization, use of military force, or other means)

[Translated results of Japanese SQuAD]
Paragraph: it has also allowed the technology and ideas to spread rapidly .
Question: What did imperialism allow the cause of imperialism to spread rapidly ?
Answer: 技術とアイディア (technology and ideas)

図 3: 英語，フランス語，日本語 SQuAD における結果の比較．

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In *ACL*, 2017.

[3] Wei He, Kai Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *Workshop on Machine Reading for Question Answering*. ACL, 2017.

[4] Minghao Hu, Yuxing Peng, and Xipeng Qiu. Reinforced mnemonic reader for machine reading comprehension. In *IJCAI*, 2018.

[5] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, 2017.

[6] Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 2017.

[7] Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. Semi-supervised training data generation for multilingual question answering. In *LREC*, 2018.

[8] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[9] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In *LREC*, 2016.

[10] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[11] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

[12] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *ICLR*, 2017.

[13] Junbei Zhang, Xiaodan Zhu, Qian Chen, Lirong Dai, and Hui Jiang. Exploring question understanding and adaptation in neural-network-based question answering. In *ICCC*, 2017.

[14] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *EMNLP*, 2016.

については分析対象から除外した．