様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　５ 月 ２４ 日現在

機関番号：１４４０１
研究種目：研究活動スタート支援
研究期間：2017～2018
課題番号：１７Ｈ０６８２２

研究課題名（和文）マルチリソース適応によるローリソースニューラル機械翻訳の高度化

研究課題名（英文）Multiple resource adaptation for low resource neural machine translation

研究代表者
チョ　シンキ（CHU, CHENHUI）

大阪大学・データビリティフロンティア機構・特任助教（常勤）

研究者番号：７０７８４８９１

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：外国人観光客の急増や2020年東京オリンピック開催などのため、翻訳の需要が急速に高まっており、機械翻訳（MT）は不可欠である。MTでは翻訳知識が対訳コーパス（文単位のバイリンガルテキスト）から獲得される。しかし、日本語とほとんどの言語の間（例えばインドネシア語）及び分野（例えば医療）において対訳コーパスは少ないため、翻訳の品質が低い。そのような低資源の場合にいかに翻訳精度を上げるかは挑戦的かつ未解決な問題である。我々は資源豊富な言語対（例えば英語ーフランス語）や分野（例えば議会）の対訳コーパス及び単言語コーパスといったマルチリソースを用いて低資源MTの翻訳品質を大幅に向上した。

研究成果の学術的意義や社会的意義
深層学習に基づくニューラル機械翻訳（NMT）の発展により、大規模な対訳コーパスが入手できる場合に最先端の翻訳精度を達成したが、対訳コーパスが少量な場合に翻訳精度が低いことが知られている。しかし、特定言語対や分野の対訳コーパスが大量に存在しない場面が数々ある。例えば、2020年東京オリンピックでは、日本語から東南アジア言語へのスポーツ分野でのMTサービスが非常に重要だと思われるが、それらの言語対や分野において対訳コーパスは少量かほとんど存在しない。本研究で提案したマルチリソース適用はそのような低資源のNMTの翻訳精度向上に成功し、MTの実用化をさらに前進させた。

研究成果の概要（英文）：In Japan, because of the rapid increase of foreign tourists and the host of the 2020 Tokyo Olympic Games, translation needs are rapidly growing, making machine translation (MT) indispensable. In MT, the translation knowledge is acquired from parallel corpora (sentence-aligned bilingual texts). However, as parallel corpora between Japanese and most languages (e.g., Japanese-Indonesian) and domains (e.g., medical domain) are very scarce (only tens of thousands of parallel sentences or fewer), the translation quality is not satisfied. Improving MT quality in this low-resource scenario is a challenging unsolved problem. Our core idea is adapting knowledge from multiple resources, including parallel corpora of resource rich-languages (such as French-English) and domains (such as the parliamentary domain), and large-scale monolingual web corpora to improve low-resource NMT. Experiments show that we significantly improved low-resource MT with multi-resource adaptation.

研究分野：機械翻訳

キーワード：ニューラル機械翻訳　分野適応　低資源

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

国際化が進む中、翻訳の需要はさまざまな場面で急速に高まっており、すべてのテキストを人手で翻訳することは不可能である。機械翻訳（MT）は翻訳のコストを削減して効率を高めるための強力なツールとして、国際化を促進するのに重要な役割を果たしている。日本では、外国人観光客の急増（2018 年は 3000 万人以上）や 2020 年東京オリンピック開催などのため、MT は不可欠である。

深層学習に基づくニューラル機械翻訳（NMT）の発展により、MT の精度は大きく向上してきた＜引用文献①＞。NMT は大規模な対訳コーパス（文単位のバイリンガルテキスト）が入手できる場合に最先端の翻訳精度を達成したが、対訳コーパスが少量な場合、翻訳精度が低いことが知られている＜引用文献②＞。しかし、特定言語対や分野の対訳コーパスが大量に存在しない場面が数々ある。例えば、2020 年東京オリンピックでは、日本語から東南アジア言語への医療やスポーツ分野での MT サービスが非常に重要だと思われるが、それらの言語対や分野において対訳コーパスは少量かほとんど存在しない。そのような低資源の場合にいかに翻訳精度を上げるかは挑戦的かつ未解決な問題である。

２．研究の目的

本研究では、コンピュータビジョンなどで成功を収めている分野適応に着目し、低資源 NMT のための分野適応手法を考案する。分野適応は資源豊富な分野から学習した知識を低資源な分野に適応するという手法である。MT において英語を中心に資源豊富な言語対や分野は多数ある。例えば、英語-フランス語の議会分野の対訳コーパスは 2.5M 文対存在する。そこで、他言語や分野の資源豊富な対訳コーパスで学習した翻訳知識を低資源 NMT に適用する。また、ウェブの普及に伴い、大規模な単言語コーパス（数億文規模）が多くの言語で入手できるようになっている。本研究では単言語コーパスも低資源 NMT に適用する。最終的には他言語や分野の対訳コーパス及び単言語コーパスといったマルチリソースを統合的に適用し、特定言語対や分野の翻訳品質の向上を目指す。図１に本研究の目的を示す。

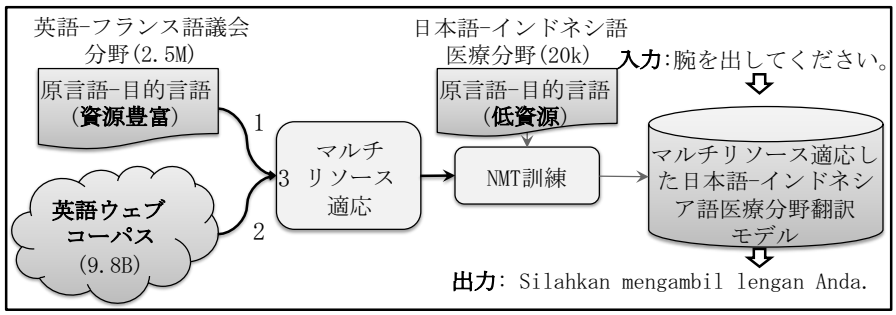

図１：研究目的

３．研究の方法

（１）他言語や分野の対訳コーパスによる適応

他言語や分野の資源豊富な対訳コーパスを利用するために、図２の mixed fine-tuning の手法を提案する。従来の移転学習手法である fine-tuning は資源豊富な対訳コーパスから NMT モデルを収束するまで訓練してから、低資源な対訳コーパスで NMT モデルのパラメタを fine-tuning することによって分野適応を実現する。しかし、低資源な対訳コーパスはデータ量が少ないため fine-tuning が overfit しやすい傾向がある。それを改善するために我々は図２の mixed fine-tuning 手法を提案する。

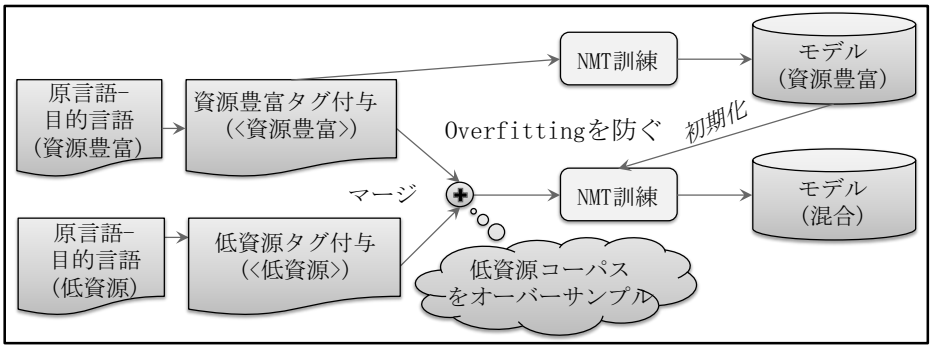

図２：Mixed fine-tuning

- 資源豊富な対訳コーパスで NMT モデルを収束するまでに訓練する。
- 資源豊富と低資源な対訳コーパスを混合したコーパスで前のステップで訓練した NMT モデルを fine-tuning する。データを混合する時は原言語文に<資源>タグを付与することによって NMT のデコーダがそれぞれの資源の文を生成しやすくようにコントロールする。また、低資源なコーパスをオーバーサンプリングし、それぞれの資源に対して等しく訓練することを狙う。

（２）単言語コーパスとマルチリソース適応

NMT 自体が言語モデルを学習できるため、目的言語の単言語コーパスを原言語に逆翻訳して擬似コーパスを作成すればそれを用いて NMT を訓練することが可能である（引用文献③）。図 3 に単言語コーパスの逆翻訳による適用手法を示す。擬似コーパス作成は低資源 NMT に有効であることは示されているが、逆翻訳システムの性能との関係や他分野コーパス適用との統合は研究されていない。本研究では mixed fine-tuning による他分野コーパス適用後のシステムで擬似コーパスを作成し、単言語コーパスによる適用を行う。さらに、作成した擬似コーパスも用いて mixed fine-tuning を行うマルチリソース適応手法を提案する。

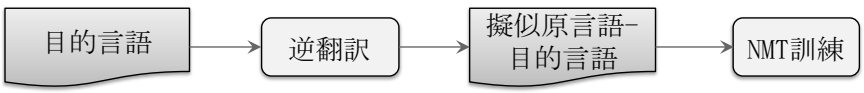

図 3：単言語コーパスの逆翻訳による適用

４．研究成果

他言語や分野の対訳コーパスによる適応は〔学会発表⑩、⑪〕、単言語コーパスとマルチリソース適応は〔雑誌論文②〕で発表した。さらに NMT における分野適用についてのサーベイ論文を〔学会発表⑥〕で発表し、日本通訳翻訳学会で分野適応の最先端についての招待講演をした〔学会発表⑤〕。以下、他言語や分野の対訳コーパスによる適応、単言語コーパスとマルチリソース適応のそれぞれの効果について実験を通して説明する。

（１）他言語や分野の対訳コーパスによる適応

表 1：他言語や分野の対訳コーパスによる適応の実験データ

コーパス（単位：文対）	訓練	開発	テスト
ALT 日英（Wikinews 分野）	18k	1,000	1,018
KFTT 日英（Wikipedia 京都分野）	440k	1,166	1,160
IWSLT 日英（対話分野）	223k	871	1,549
IWSLT 中英（対話分野）	209k	887	1,570

他言語や分野の対訳コーパスによる適応の有効性を検証するために、翻訳実験を行った。表 1 に実験で用いたデータを示す。なお、ALT 日英は in-domain タスクで、訓練データは 18k 文対しかなく低資源な設定である。他言語や分野の対訳コーパスとして KFTT 日英（Wikipedia 京都分野）、IWSLT 日英（対話分野）及び IWSLT 中英（対話分野）を利用する。それぞれ 440k、223k、209k 文対の訓練データがある。

表 2：他言語や分野の対訳コーパスによる適応実験結果（最も性能の高いものと有意差がないもの（$p < 0.05$）を太字で表す）

システム	ALT 日英（BLEU-4）
ベースライン：ALT 日英 NMT	8.47
KFTT 日英適用	21.74
IWSLT 日英適用	19.76
IWSLT 中英適用	19.10
KFTT 日英+IWSLT 日英適用	**24.29**
IWSLT 日英+IWSLT 中英適用	19.35
KFTT 日英+IWSLT 日英+IWSLT 中英適用	**24.04**

表 2 に他言語や分野適応実験の翻訳結果を示す。適用結果として、資源豊富なコーパスをそれぞれ利用した時の精度とそれらを組み合わせて利用時の精度を示している。評価には BLEU スコア＜引用文献④＞を用いた。BLEU スコアは参照訳との一致度合いを表す評価尺度であり、高い方が良い。太字は最も良いシステムとそれらと統計的有意差がないシステムを表す。表 2 に示した通り、適応することによって in-domain タスクの ALT 日英の BLEU スコアが 8 ポイントから 24 ポイントまで大幅くに上がった。BLEU スコア 8 ポイントは人手 5 段階評価中の 1 ぐらいで、24 ポイントは 5 段階評価中の 3 ぐらいである。よって、他言語や分野適応により翻訳精度が大幅に向上したことが分かる。

（２）単言語コーパスとマルチリソース適応

表 3：他言語や分野の対訳コーパスによる適応の実験データ

コーパス（単位：文）	訓練	開発	テスト
WIKI 中日（Wikipedia 分野）	136k	198	198
ASPEC 中日（科学技術分野）	672k	2,090	2,107
日本語 WIKI（Wikipedia 分野）	3M	N/A	N/A

単言語コーパスとマルチリソース適応の有効性を検証するために、翻訳実験を行った。表 3 に実験で用いたデータを示す。なお、WIKI 中日は in-domain タスクで、訓練データは 136k 文対しかなく低資源な設定である。他分野の対訳コーパスとして ASPEC 中日(科学技術分野)を使用した。ASPEC 中日には 672k 文対の訓練データが含まれる。単言語コーパスは日本語 Wikipedia から抽出した 3M 文を利用した。

表 4 に単言語コーパスとマルチリソース適応実験結果を示す。ASPEC 中日適用は mixed fine-tuning を使った他分野対訳コーパス適用の結果で、日本語 WIKI 適用は単言語コーパスの適用結果で、ASPEC 中日+日本語 WIKI 適用はマルチリソースの適応結果である。表 4 に示した通り、単

言語コーパス適用は他分野対訳コーパス適用に匹敵した精度向上を果たした上、マルチリソース適用の方がそれぞれ適用時より性能が優れていることが分かる。

表 4：単言語コーパスとマルチリソース適応実験結果（最も性能の高いものと有意差がないもの（$p < 0.05$）を太字で表す）

システム	WIKI 中日(BLEU-4)
ベースライン：WIKI 中日 NMT	17.43
ASPEC 中日適用	37.57
日本語 WIKI 適用	37.30
ASPEC 中日+日本語 WIKI 適用	**41.37**

＜引用文献＞

① Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, (2015).
② Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation, pp. 28-39, (2017).
③ Rico Sennrich, Barry Haddow and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 86-96, (2016).
④ Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, (2002).

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① 瓦祐希，Chenhui Chu，荒瀬由紀．統計的機械翻訳のための Recursive Neural Network による事前並び替えと分析．自然言語処理，査読有，vol. 26(1)，pp. 155-178，(2019.3)．
② Chenhui Chu，Raj Dabre and Sadao Kurohashi. A Comprehensive Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. 情報処理学会論文誌，査読有，vol. 26(1)，pp. 1-10，(2018.6)．

〔学会発表〕（計 11 件）

① 瓦祐希，Chenhui Chu，荒瀬由紀．ニューラル機械翻訳における事前並び替えの影響分析．言語処理学会 第 25 回年次大会，pp.1455-1458，(2019.3)．
② Chenhui Chu，梶原 智之，中島 悠太，長原 一，渡辺 理和，大久保 規子．多国間法律の比較と統計分析のための多言語機械翻訳． 第 119 回人文科学とコンピュータ研究会発表会，(2019.2)．
③ Yuki Kawara，Yuto Takebayashi，Chenhui Chu and Yuki Arase. Osaka University MT Systems for WAT 2018: Rewarding, Preordering, and Domain Adaptation. In Proceedings of the 5th Workshop on Asian Translation，(2018.12)．
④ Yuto Takebayashi，Chenhui Chu，Yuki Arase and Masaaki Nagata. Word Rewarding for Adequate Neural Machine Translation. In Proceedings of the 15th International Workshop on Spoken Language Translation，pp. 14-22，(2018.10)．
⑤ Chenhui Chu．ニューラル機械翻訳における分野適応の最先端.日本通訳翻訳学会第 19 回年次会,招待講演,(2018.9)
⑥ Chenhui Chu and Rui Wang. A Survey of Domain Adaptation for Neural Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics，pp. 1304-1319，(2018.8)．
⑦ Yuki Kawara，Chenhui Chu and Yuki Arase. Recursive Neural Network Based Preordering for English-to-Japanese Machine Translation. In Proceedings of the ACL 2018 Student Research Workshop，pp. 21-27，(2018.7)．
⑧ 竹林 佑斗，Chenhui Chu，荒瀬 由紀，永田 昌明．ニューラル機械翻訳における単語予測の重要性について．2018 年度人工知能学会全国大会，(2018.6)．
⑨ 瓦祐希，Chenhui Chu，荒瀬由紀．Recursive Neural Network を用いた事前並び替えによる英日機械翻訳．言語処理学会 第 24 回年次大会，pp.33-36，岡山，(2018.3)．
⑩ Chenhui Chu and Raj Dabre. Multilingual and Multi-Domain Adaptation for Neural Machine Translation. In Proceedings of the 24th Annual Meeting of the Association for Natural Language Processing (NLP 2018)，pp.909-912，Okayama，Japan，(2018.3)．
⑪ Chenhui Chu，Raj Dabre and Sadao Kurohashi. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In Proceedings of the 55th

Annual Meeting of the Association for Computational Linguistics (ACL 2017 short), Vancouver, Canada, (2017.7).

〔図書〕(計　1　件)

① Using Comparable Corpora for Under-Resourced Areas of Machine Translation. Inguna Skadiņa, Robert Gaizauskas, Bogdan Babych, Nikola Ljubešić, Dan Tufiș, Andrejs Vasiļjevs (edited) (担当:分担執筆, 範囲: Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi, Chapter 7.3: Chinese-Japanese Parallel Sentence Extraction from Quasi-Comparable and Comparable Corpora, pp. 255-290), Springer, (2018.11).

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕

ホームページ等
https://researchmap.jp/chu/

６．研究組織

(1)研究分担者

研究分担者氏名：

ローマ字氏名：

所属研究機関名：

部局名：

職名：

研究者番号（8桁）：

(2)研究協力者

研究協力者氏名：Raj Dabre

ローマ字氏名：Raj Dabre

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。