特許文を対象としたMulti-Source ニューラル機械翻訳

Multi-Source Neural Machine Translation of Patent Sentences

筑波大学システム情報系知能機能工学域教授
宇津呂　武仁

1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。京都大学等を経て、2012 年より筑波大学システム情報系知能機能工学域教授。自然言語処理、機械翻訳、ウェブマイニングの研究に従事。電子情報通信学会、情報処理学会、人工知能学会、言語処理学会、ACL 各会員。

筑波大学大学院システム情報工学研究科知能機能システム専攻
磯部　僚也

2019 年芝浦工業大学システム理工学部電子情報システム学科卒業。現在、筑波大学大学院システム情報工学研究科知能機能システム専攻 博士前期課程在学中。機械翻訳の研究に従事。

NTT コミュニケーション科学基礎研究所上席特別研究員
永田　昌明

1987 年京都大学大学院工学研究科修士課程修了。同年、日本電信電話株式会社入社。現在、コミュニケーション科学研究所 上席特別研究員。工学博士。自然言語処理の研究に従事。電子情報通信学会、情報処理学会、人工知能学会、言語処理学会、ACL 各会員。

1 はじめに

多言語 NMT において、複数原言語の対訳文を参照して目的言語の一文に翻訳を行う（n-to-1）Multi-Source 翻訳手法[7] においては、モデルの訓練および評価においては全言語の対訳文が揃うことが前提とされるが、実世界のコーパスにおいてこの条件が満たされる状況は極めて限られる。一方、複数翻訳方向の 1-to-1 翻訳が混在した Google の多言語翻訳[1] に用いられるモデルにおいても、三言語以上の対訳関係を無視し独立な二言語対とみなすため、評価時に複数の原言語の文が存在したとしても、その利点をいかすことができない点が弱点である。そこで本論文では、三言語対訳コーパスが利用できる場合を対象として、これを有効活用する Multi-Source Many-to-One 手法を提案する。

本論文では、複数の原言語文を文区切り記号連結したものを入力とする文連結型 Multi-Source 翻訳と通常の Single-Source 翻訳（1-to-1 翻訳）を、入力に言語タグを付与することにより区別して一つのモデルとして訓練する方法を Multi-Source Many-to-One 翻訳と定義する。具体的には、英中日の三言語対訳、英日対訳、中日対訳があるとき、英語文と日本語文を連結した入力には“<ENZH2JA>”タグを、英語文の入力には“<EN2JA>”タグを、中国文の入力には“<ZH2JA>”タグを、それぞれ付与して、日本語文を出力する一つのモデルを訓練する。

また、本論文の比較対象として、Nishimura ら[3] の方法では、そもそも複数エンコーダで複数言語入力を実現しているので、単一言語入力の翻訳モデルとパラメータを共有することは不可能である。しかし、単

一言語入力の場合には、Nishimura ら[3] の方法において三言語対訳の欠落を表現するのに使われている“<NULL>”トークンを使用することにより、文連結型 Multi-Source 翻訳において Nishimura ら[3] の方法を近似したものを、本研究の一つの比較対象とする。一方、Google 多言語翻訳のうちの Many-to-One 翻訳は、出力言語が同一である複数の言語対に対して、一つのモデルとして訓練する方法である。しかし、この方法では、英中日の三言語対訳データのように、三つ以上の言語が互いに対訳になっている場合に、これらが互いに対訳になっていることを利用することができない。そこで、本論文では、この Google の Many-to-One 翻訳を参考にした提案法の非 Multi-Source 版（すなわち、Many-to-One 翻訳）をもう一つの比較対象とする。以上の二つの比較実験を行うために、特許の三言語対訳データから表 1 のデータを作成し、図 1 に示す形でモデルの比較を行う。

表 1　特許三言語対訳コーパスの欠落設定（各数字は欠落のない文数を示す。“X”は欠落した部分を示す）

文番号	英	中	日
1-200,000	20 万	20 万	20 万
200,001-600,000	40 万	X	40 万
600,001-1,000,000	X	40 万	40 万

2　多言語 NMT の関連研究

2.1　Multi-Source 翻訳

Zoph ら[7] はマルチエンコーダの再帰型ニューラルネットワーク（RNN）モデルを用いて Multi-Source 翻訳を実現した。各エンコーダは各言語の原言語文を処理する。1-to-1 翻訳モデルよりも翻訳精度が改善しているが、全言語の対訳訓練文を用意する必要がある。実世界のコーパスにおいてこの条件が満たされる状況は、非常に限定される。この問題に対して、Nishimura らは、欠落した言語の部分を、トークン[3] や自動生成した疑似対訳文[2] で置き換える手法を提案した。

2.2　Google 多言語翻訳

Johnson ら[1] は、英語を含む三言語コーパスを中心に、異なる翻訳方向の 1-to-1 文対を混在させた翻訳手法を提案した。この手法では、三言語以上の文対応関係が含まれる場合もこれを利用せず、全ての言語対を 1-to-1 の対応に変換して翻訳モデルの訓練を行う。原言語を英語とし、目的言語を他二言語とする One-to-Many 翻訳、原言語を他二言語とし、目的言語を英語とする Many-to-One 翻訳、原言語と目的言語両方を三言語とする Many-to-Many 翻訳の三種類の設定に分けら

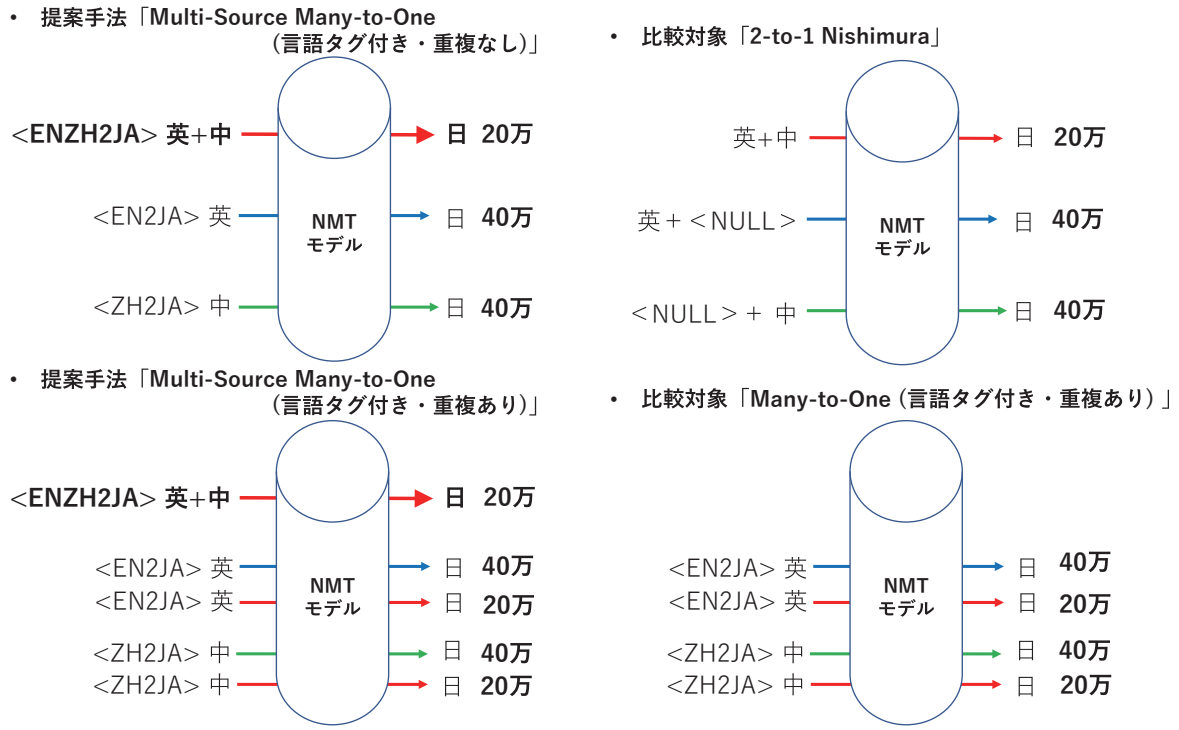

図 1　提案手法と先行研究のモデルの比較

SOURCE	*TARGET*
The present invention relates to a cover glass . <CONCAT> 本发明涉及覆盖玻璃 。	本発明は、カバーガラスに関する 。
The sample S has a known shape . <CONCAT> 样品S为已知的形状 。	サンプル S は既知の形状である 。
The process may terminate thereafter . <CONCAT> 该过程此后可以结束 。	その後 、プロセスは終了してもよい 。

図２　二文連結による {英 , 中} -to- 日の訓練データ例
（実際の訓練過程では、訓練前に 4 節の前処理を適用する）

れている。出力時の目的言語を指定するため、入力文の先頭に、該当する言語タグを付与する。本論文では、上述の三種類の設定の中でも、特に Many-to-One 翻訳を参考にして提案法の非 Multi-Source 版（すなわち、Many-to-One 翻訳）を作成し、比較評価を行う。

3 特許三言語対訳コーパス

本論文では、日本、中国、米国の 2005 ～ 2017 年特許データからパテントファミリーに基づいて対訳文書対を抽出し、文献[5] の手法により文対応を付与し、英語と中国語を原言語、日本語を目的言語として [1]、日中英三言語対訳コーパスを作成した。その結果、100 万訓練文対、２万開発文対、および、２万評価文対の三言語対訳コーパスが作成された。

4 提案手法

4.1　文連結型 Multi-Source 翻訳

本論文では、文連結手法を用いて三言語対訳コーパスを有効利用する手法を提案する。本論文では、“<CONCAT>” トークンを用いて複数の言語の原言語文を連結する。連結後の対訳文対の例を図２に示す。このように、連結された二文を一文として扱い、翻訳モデルの訓練・評価を行う。このアプローチによれば、単エンコーダモデルによって 2-to-1 翻訳が実現可能となるため、マルチエンコーダの実装が不要となる。

さらに、本論文では、Nishimura ら[3, 2] が提案した手法において、同様の文連結手法を導入する方式の翻訳精度の評価も行った。文献[2] に従う場合には、No.1 ～ No.200,000 の文を用いて {英、日} -to- 中と {中、日} -to- 英の翻訳モデルを訓練した後、表 1 の欠落部分 “X” に相当する翻訳文を自動生成し全体を疑似対訳文として扱い、この手法を “2to1 Nishimura（疑似対訳）” と呼ぶ。一方、文献[3] に従う場合には、表 1 の欠落部分 “X” を “<NULL>” トークンに置き換え、この手法を “2to1 Nishimura（NULL）” と呼ぶ。

4.2　Multi-Source Many-to-One 翻訳

本論文では、欠落を含む三言語対訳コーパス（表 1）において、日本語を目的言語とした Many-to-One 翻訳に着目し、Multi-Source Many-to-One 翻訳の方式を提案する。本論文の各提案手法、および、比較対象の手法との間の関係を図 1 に示す。

この手法においては、まず、4.1 節の文連結 2-to-1 翻訳手法を No.1 ～ No.200,000 の三言語対訳文に適用し、{英、中} -to- 日翻訳訓練用対訳文対を生成し、連結後の原言語文対の先頭に “<ENZH2JA>” タグを付与する。次に、No.200,001 ～ No.600,000 の英日対訳文から、1-to-1 翻訳用に英 -to- 日翻訳訓練用対訳文対を作成し、各原言語文の先頭に “<EN2JA>” タグを付与する。さらに、No.600,001 ～ No.1,000,000 の中日対訳文から、1-to-1 翻訳用に中 -to- 日翻訳訓練用対訳文対を作成し、各原言語文の先頭に “<ZH2JA>” タグを付与する。以上の訓練用対訳文対を用いて訓練した翻訳モデルを “Multi-Source Many-to-One 翻訳（言語タグ付き・重複なし）” モデルと呼ぶ。このモデルの直接の比較対象は、Nishimura ら[3] が提案した “2to1 Nishimura (NULL)” モデルである。

さらに、No.1 ～ No.200,000 の三言語対訳文から、1-to-1 翻訳用の英 -to- 日翻訳訓練用対訳文対と中 -to- 日翻訳訓練用対訳文対をそれぞれ生成し、“Multi-Source Many-to-One 翻訳（言語タグ付き･重複なし）” モデルの訓練用対訳文対に追加して訓練を行った翻訳モ

1　英語の Tokenization は Moses Tokenizer（https://github.com/moses-smt/mosesdecoder/）、中国語の形態素解析は Jieba（https://github.com/fxsjy/jieba）、日本語は文字単位に分割した。英語文と中国語文の合計文長が 1,000 トークン以上となる対訳文対は除外した。

表2　特許コーパスの BLEU 評価における欠落の有無の比較評価（単一モデルにおいて複数の翻訳方向を評価する場合を＊で示す。1-to-1 翻訳ベースラインに対して有意差がある（$p<0.01$）場合を † で示す。2-to-1 翻訳ベースラインに対して有意差がある（$p<0.01$）場合を ‡ で示す。）

（a）　欠落なし

設定	英 -to- 日	中 -to- 日	{英、中} -to- 日
欠落なし（100 万対訳文対）	66.42	66.99	71.18
欠落なし（20 万対訳文対）	62.59	62.36	66.69 （2-to-1 翻訳ベースライン）
欠落なし（60 万対訳文対） （1-to-1 翻訳ベースライン）	65.89	64.99	N/A

（b）　欠落あり（訓練事例：100 万対訳文対）

設定		英 -to- 日	中 -to- 日	{英、中} -to- 日
2to1 Nishimura（疑似対訳）[2]		N/A	N/A	70.63 ‡
＊ 2to1 Nishimura（NULL）[3]		64.60	62.68	70.04 ‡
＊ Many-to-One 翻訳 （言語タグ付き・重複あり）		66.31 †	66.03 †	N/A
＊ Multi- Source Many-to-One 翻訳 （言語タグ付き）	重複なし	64.32	62.70	69.84 ‡
	重複あり	66.29 †	66.71 †	70.63 ‡

デルを、“Multi-Source Many-to-One 翻訳（言語タグ付き・重複あり）”モデルと呼ぶ。このモデルの直接の比較対象は、Google の Many-to-One 翻訳を参考にした提案法の非 Multi-Source 版（すなわち、Many-to-One 翻訳）である。このモデルの訓練用対訳文対は、“Multi-Source Many-to-One 翻訳（言語タグ付き・重複あり）”モデルの訓練用対訳文対から {英、中} -to- 日翻訳訓練用対訳文対を除外したもの（図 1 参照）であり、また、Johnson ら[1] のモデルと比べて原言語を識別するタグを付与している点が異なる[2]

5　評価

評価結果を表2に示す[3]。

欠落なしの訓練文規模 100 万文対において、2-to-1 翻訳の BLEU スコアが 1-to-1 翻訳より 4.19 ポイント有意に改善した。次に、欠落を考慮した特許コーパスにおける BLEU 評価結果では、欠落なしの場合のベースライン翻訳モデルとして、(i) 2-to-1 翻訳モデルでは、表 1 の No.1 ～ No.200,000 の文を用いて訓練した「欠落なし（20 万対訳文対）」のモデル、(ii) 1-to-1 翻訳モデルでは、英 -to- 日翻訳において No.1 ～ No.600,000 の文を用いて訓練した「欠落なし（60 万対訳文対）」モデル、および、中 -to- 日翻訳において No.1 ～ No.200,000 の文と No.600,001 ～ No.1,000,000 の文を用いて訓練した「欠落なし（60 万対訳文対）」モデル、をそれぞれ評価した。2-to-1 翻訳においては、本論文で提案した文連結型 2-to-1 翻訳モデルは、Nishimura らの手法[2, 3] に対しても適用可能であることが示された。さらに、本論文の提案手法である “Multi-Source Many-to-One 翻訳（言語タグ付き）”の「重複あり・なし」両モデルとも、ベースラインモデルの BLEU を有意に改善した[4]。一方、1-to-1 翻訳においては、“Many-to-One 翻訳（言語タグ付き・重複あり）”モデル、および、提案手法の “Multi-Source Many-to-One 翻訳（言語タグ付き・重複あり）”モデルは、ベースラインの BLEU を有意に改善した。“2to1

2　図 1 中の赤色の 1-to-1 翻訳用の英 -to- 日翻訳訓練用対訳文対 20 万文対と中 -to- 日翻訳訓練用対訳文対 20 万文対における日本語文が重複している点に注意する。

3　本論文では、fairseq ツールキット[4] を用いて実装された Transformer モデル[6] を用いた。Head の数を 4、エンコーダとデコーダを各 6 層、単語分散表現を 512 次元、隠れ層を 1,024 次元、学習率を 0.0003 とし、Adam optimizer を使用した。ドロップアウトを 0.3 として、50 エポックの訓練を行う。ハードウェアとして、NVIDIA Tesla P100 16GB GPU 1 枚を使用した。評価時、モデルの日本語出力文は MeCab の IPADic 辞書の形態素単位に分割して評価を行った。BLEU スコアの評価においては、Moses デコーダー（https://github.com/moses-smt/mosesdecoder）のスクリプト（multi-bleu.perl）を使用した。BLEU スコアの有意差検定においては、mteval Toolkit（https://github.com/odashi/mteval）を使用した。

4　「重複あり・なし」両モデル間の BLEU の差においても有意差がある（$p<0.01$）。“Multi-Source Many-to-One 翻訳（言語タグ付き・重複あり）”モデルにおいては、擬似対訳を用いなくても、疑似対訳を用いる “2to1 Nishimura（疑似対訳）”モデルと同等の翻訳精度を達成できており、訓練効率の点からも優位性が大きいと言える。

Nishimura（NULL）”モデルとの比較においては、“Multi-Source Many-to-One 翻訳（言語タグ付き・重複あり）”モデルによる 2-to-1 翻訳の BLEU 70.63 は“2to1 Nishimura（NULL）”モデルの BLEU70.04 を有意（$p<0.01$）に改善し、1-to-1 翻訳の BLEU 66.29 および 66.71 は“2to1 Nishimura（NULL）”モデルの BLEU64.60 および 62.68 を有意（$p<0.01$）に改善した。一方、“Many-to-One 翻訳（言語タグ付き・重複あり）”モデルの 1-to-1 翻訳との比較においては、“Multi-Source Many-to-One 翻訳（言語タグ付き・重複あり）”モデルにより同等の BLEU を達成しており、それに加えて、単一のモデルにより 2-to-1 翻訳も行うことができている。

6 実例分析

{英、中}-to- 日翻訳モデルによる Multi-Source 翻訳の例を表 3 に示す。英単語“breaker”の日本語訳は「器」と「装置」の可能性があるが、中国語では日本語と同じ漢字「装置」が使われているので、中日方向の翻訳においては訳語の曖昧性の問題がない。このため、“breaker”は、中 -to- 日翻訳モデル、および、{英、中}-to- 日翻訳モデルにおいては、正しく日本語訳された。

Multi-Source 翻訳過程における自己注意ヒートマップを図 3 に示す。この図において、注意の値は 4 つの Head の平均値となる。“Normally”-「通常」、および、“breaker”-「切断装置」等、英語と中国語の原言語文の間での注意の対応が適切にとれていることが分かっ

表 3　特許データにおける翻訳例（{英 , 中}-to- 日翻訳）

原言語文・参照文・モデル	文	BLEU
原言語文（英） 原言語文（中）	Normally , the IGBT 14 of the current breaker 10 is off . 在通常时、电流切断 装置 10 的 IGBT14 断开 。	N/A N/A
参照文	通常時は、電流遮断 装置 10 の IGBT14 はオフしている。	N/A
英 -to- 日 中 -to- 日 {英、中}-to- 日	通常、電流遮断 器 10 の IGBT14 はオフである。 通常時には、電流遮断 装置 10 の IGBT14 がオフする。 通常時には、電流遮断 装置 10 の IGBT14 はオフしている。	37.79 54.41 85.79

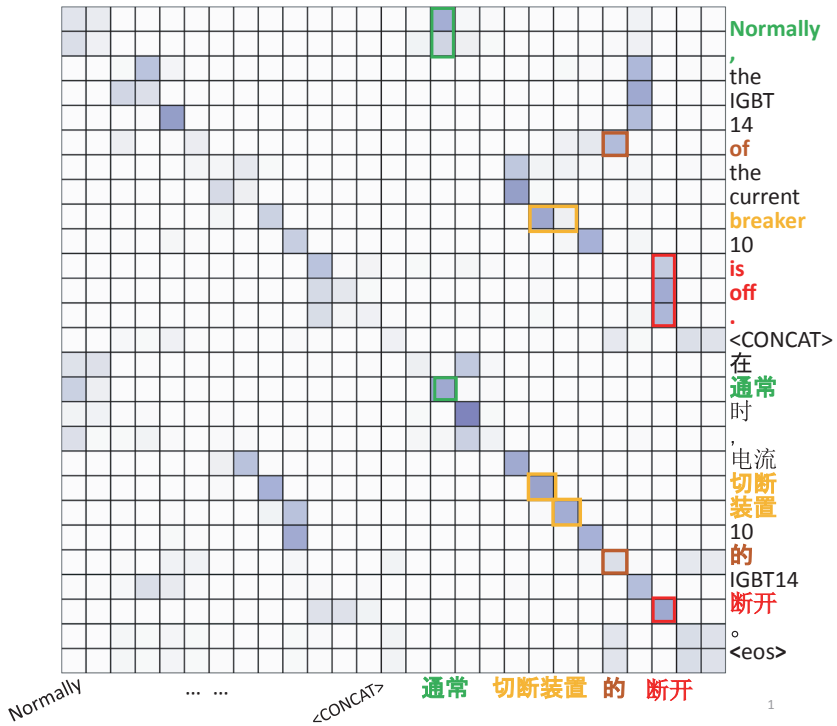

図 3　エンコーダ 6 層目の自注意機構におけるヒートマップの例

た。提案手法である文連結型 2-to-1 翻訳手法においては、この例のように、連結された Multi-Source 文において、このように自己注意が適切に機能することによって、高い翻訳精度を達成できていると考えられる。

7 おわりに

本論文では、出力言語が同じである複数の言語対を一つのモデルにする Many-to-One 翻訳の考え方を拡張した Multi-Source Many-to-One 翻訳モデルを提案した。評価実験においては、Nishimura ら[3] のマルチエンコーダを用いた Multi-Source 翻訳モデルを文連結により実現したモデルと比較して、2-to-1 翻訳・1-to-1 翻訳とも BLEU を有意に改善することを示した。この方法は見方を変えると、Many-to-One 翻訳において、三言語以上の対訳が存在している場合には、三言語対訳データであることを活用する拡張となっている。さらに、評価実験においては、1-to-1 翻訳においては Google の Many-to-One モデルを参考にした”Many-to-One 翻訳（言語タグ付き・重複あり）”モデルと同等の BLEU を達成し、かつ、単一のモデルにより 2-to-1 翻訳も行うことができることを示した。

謝辞：日英中特許データを提供して頂いた日本特許情報機構（Japio）の関係者各位に深謝の意を表する。

参考文献

[1] M. Johnson, M. Schuster, Q. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viegas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of ACL, Vol. 5, pp. 339–351, 2017.

[2] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura. Multi-source neural machine translation with data augmentation. In Proc. 15th IWSLT, pp. 48–53, 2018.

[3] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura. Multi-source neural machine translation with missing data. In Proc. 2nd WNGT, pp. 92–99, 2018.

[4] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. Fairseq: A fast, extensible toolkit for sequence modeling. In Proc. NAACL-HLT: Demonstrations, pp. 48–53, 2019.

[5] M. Uchiyama and H. Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In Proc. 41th ACL, pp. 72–79, 2003.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Proc. 30th NIPS, pp. 5998–6008, 2017.

[7] B. Zoph and K. Knight. Multi-source neural translation. In Proc. NAACL-HLT, pp. 30–34, 2016.