多言語機械翻訳への言語類型論特徴の導入

平野 颯　　大内 啓樹　　渡辺 太郎
奈良先端科学技術大学院大学 先端科学技術研究科
{hirano.hayate.hc2, hiroki.ouchi, taro}@is.naist.jp

概要

本研究は多言語機械翻訳モデルの推論に対する説明性向上を目的として，言語類型論特徴をモデルに導入する方法を提案する．言語類型論は，各言語が共通してもつ特徴を探る言語学の一分野であり，研究の蓄積はデータベースにまとめられている．原言語文に言語類型論データベースから獲得した言語特徴を導入し，IWSLT17 の多言語翻訳データセット上で学習を行うことで，導入した言語特徴の寄与を明らかにした．

1　はじめに

近年のニューラル機械翻訳モデルは主にエンコーダ・デコーダの構成を取る．すなわち，原言語文の各単語をベクトル空間に埋め込み，中間表現にエンコード，目的言語の文にデコードするようにモデルを構成する．

多言語機械翻訳の目標の 1 つは，全言語の単語を同一の空間に埋め込むことにより，2 言語間における機械翻訳モデルと同程度のパラメータ数のみで高い翻訳性能を発揮するモデルを構成することである．すべての翻訳方向に対して同一のパラメータを学習することで，言語相互の知識を転移でき，2 言語間における機械翻訳モデルを上回る性能が得られることが報告されている [1, 2]．それだけでなく，学習時に見られなかった言語対に対する翻訳，zero-shot 翻訳が可能となることが示されている．

ここで，以上の有望な性質が言語の特性にどの程度影響されるのかは明らかにされていない．本研究は目的言語を示すタグを原言語文に挿入することで，多言語機械翻訳を可能にした Johnson et al. [2] に着想を得て，本研究は(1) 言語特性を擬似的に再現するようなタグの挿入方法を提案し（3.1 節），翻訳性能に悪影響を及ぼさないことを確認し; (2) 言語類型論データベースから得られた表現をタグに導入し（3.2 節），導入した言語特徴の寄与を明らかにした．

2　関連研究

多言語機械翻訳のモデル設定は，翻訳言語間におけるパラメータの共有度合いによって分類できる．Firat et al. [3] は浅いパラメータ共有モデル，すなわち言語ごとに個別の単語埋め込み，エンコーダおよびデコーダを持ち，全言語に対して共通した注意機構を持つモデルを提案した．このようなモデル構成は，2 言語間の機械翻訳モデルと同様に言語に固有のエンコーダ，デコーダを持ち，後述する深いパラメータ共有モデルと比べ各言語に対して翻訳性能を向上させることが容易である．しかしながら，モデルの構成上多数のモデルパラメータを持たざるを得ず，翻訳方向の増加に耐えうる構成でないことが課題である．Ha et al. [4] や Johnson et al. [2] は，全ての翻訳方向に対してパラメータを共有する深いパラメータ共有モデルを提案し，2 言語間における機械翻訳モデルと同等のパラメータ数で多言語機械翻訳を実現した．ここで，50 を超える言語を対象とした研究 [5] も進められているが，翻訳言語対の数に対して性能の向上が小さいことなど，翻訳データとモデルの関係について更なる調査が必要である．

言語類型論では言語の横断的な調査を行うことで，それらが類似性を持つことがあるのはなぜか，類似性はどの程度一般化が可能かを研究する．言語現象は construction および strategy という 2 軸で分析される [6]．construction は，特定の意味もしくは機能を表現するために利用される任意の言語の形態統語構造のことを指す．strategy は特定の形態統語構造の実現方法のことであり，言語横断的に有効な方法で形式的に定義される．例えば，指示対象のカテゴリを叙述する方策として predicate nominal construction が定義できる．英語では copura strategy が用いられ，copura は言語横断的に有効な定義が与えられている．言語類型論研究の蓄積として，各

strategy と各言語での実現値はデータベースの形で整理が進んでいる．本研究では，データベースで整理されている各言語における strategy の実現値のことを言語特徴と呼ぶ．

3 提案手法

3.1 タグによる擬似的な言語特徴の導入

Johnson et al.[2] は，目的言語に対応するタグを原言語文に挿入した．

<2de> How are you? -> Wie geht's?

上記の例は英語からスペイン語への翻訳を表しており，ドイツ語への翻訳であることを意味する<2de>がタグに当たる．この<2de>タグは，同時に学習を行っている他の言語，例えばスペイン語と共通した言語特徴を持っていないという仮定を置く．実際にドイツ語では SOV 語順を取りうるが，スペイン語は取らず，「SOV 語順を取る」という状態を学習しうると仮定できる．このような仮定のもとで，同時に学習を行う他の言語と共通する特徴を表現するタグを挿入する．例えば，英語 → ドイツ語（en→de），英語 → スペイン語（en→es），英語 → フランス語（en→fr）の 3 方向での多言語翻訳を行う場合，以下のようなタグを挿入する．

<2de> <de-es> <de-fr> How are you? -> Wie geht's?

<2es> <de-es> <es-fr> How are you? -> ¿Cómo estás?

<2fr> <es-fr> <de-fr> How are you? -> Ça va?

この設定において，英語 → ドイツ語の翻訳方向に関して，<2de>タグは他の翻訳方向には見られないものであり，対象とする翻訳言語のうち特定の言語にのみ出現する特徴をシミュレートするタグである．これとは別に<de-es>および<de-fr>タグはそれぞれ英語 → スペイン語，英語 → フランス語の翻訳方向にも挿入されているタグであり，複数の言語で同一の特徴をシミュレートするタグである．すなわち，前述の仮説の上で<2de>は「SOV 語順を取る」という状態を学習でき，<de-es>はドイツ語およびスペイン語にあり，フランス語に無い特徴，例えば「主格・対格型の格標識を用いる」という状態を学習しうると考えることができる．このシミュレーションでは，Johnson et al., [2] と同様に原言語文にタグを挿入する方法でも，各タグが他の翻訳方向にも挿入されているために，翻訳方向の決定という観点では曖昧なものになる．

3.2 タグへの言語特徴の導入

本節では，多言語機械翻訳モデルに言語特徴を導入する方法を説明する．3.1 節と同様にタグを用いる点は同じであるが，

<base> How are you? -> Wie geht's?

上記の例のようにすべての翻訳方向に対して同じタグ（ここでは<base>）を挿入する．そして，このベースとなるタグに対して言語類型論データベースから導出されたベクトル表現を組み込む．このベクトル表現の各要素は言語特徴に対応し，例えば英語を例に取ると主要語順が SVO であることに対応する要素が 1 となり，主要語順が SOV であることに対応する要素は-1 となるような対応が付けられている．また，言語類型論データベースには欠損値が存在し，そのような要素は 0 と対応づけられる．実際に埋め込み表現としてモデルに入力されるベクトルは以下の式 (1) で計算される．

$$\boldsymbol{t}_{typ} = \boldsymbol{v}_{tgt}^{\mathrm{T}} W + \boldsymbol{t}_{base} \tag{1}$$

埋め込みベクトルの次元数を h，言語類型論データベースから導出されるベクトルの次元数を t とする．$\boldsymbol{v}_{tgt} \in \mathbb{R}^t$ は目的言語の言語特徴の組に対応するベクトル表現であり，$\boldsymbol{t}_{base} \in \mathbb{R}^h$ は<base>タグに対応する埋め込みベクトル，$W \in \mathbb{R}^{t \times h}$ は学習パラメータであり，$\boldsymbol{t}_{typ} \in \mathbb{R}^h$ が目的言語の言語特徴が導入されたタグのベクトルである．

4 実験および考察

4.1 実験設定

実験には，系列モデリングのフレームワークである fairseq [7] を利用し， Vaswani et al. [8] における base の設定に対応する Transformer を用いた．

翻訳モデルの学習および，評価はすべて IWSLT17 多言語翻訳データセット[1)]のイタリア語（it），英語（en），オランダ語（nl），ルーマニア語（ro）が含まれる対訳文上で行った．データセットに含まれる文数を表 1 に示す．4 言語すべてに共通して SentencePiece [9] によるサブワードへの分割を行い，語彙数は 8,000 とした．モデルのハイパーパラメータの設定は Vaswani et al. [8] の base の設定に準拠しているが，パラメータのドロップアウトは 0.3 とし，

1） https://sites.google.com/site/iwsltevaluation2017/TED-tasks

表 1 IWSLT17 多言語翻訳データセットの対訳文数

言語ペア	訓練データ	検証データ	テストデータ
en-it	235,423	2,495	1,147
en-nl	240,850	2,780	1,181
en-ro	224,162	2,592	1,129
it-en	235,423	2,495	1,147
nl-en	240,850	2,780	1,181
ro-en	224,162	2,592	1,129

学習はステップ数を決めるのではなく早期終了を実施した．3 エポック連続で検証データでの損失が下がらなかった場合に学習を打ち切った．翻訳の生成時にはビーム幅 5 のビームサーチを用いた．

本研究では，以下のような設定で実験を行った．**ThreeDirections** は，英語 →{イタリア語，オランダ語，ルーマニア語}の 3 方向で学習し，同一の方向で翻訳モデルの評価を行うものである．**SixDirections** は英語 ↔{イタリア語，オランダ語，ルーマニア語}の 6 方向で学習し，イタリア語 → オランダ語のように学習時に見られなかった zero-shot 翻訳方向について評価を行う設定である．また，擬似的な言語特徴を検証するために<2de> <de-xx> <de-yy>の全てのタグを挿入する設定を **Tsim** とし，言語を特定するタグ<2de>を省いたものを **TsimNoTarget** と呼ぶ．比較のために行う，Johnson et al. [2] と同様のタグ，すなわち<2de>を挿入する設定は **Target** と呼ぶ．以上の 3 設定で示したタグはいずれもドイツ語への翻訳の場合であり，xx および yy は同時に学習を行うドイツ語以外の言語を表す．Tsim および TsimNoTarget は，いずれも言語特徴のシミュレーションを意図する点で共通しているが，TsimNoTarget は，特定の言語のみに対応する特徴を学習できない状況での実験設定である．また，<base>タグへの言語特徴の導入実験は **Typ-insert** と呼ぶ．Typ-insert において言語類型論データベースからのベクトル表現の導出には lang2vec [10] の wals_syntax 要素を利用した．wals_syntax 要素は，データベース World Atlas of Language Structures [11] から導出されたものであり，含まれる言語特徴は 103 種類である．実験の精度評価はすべて BLEU で報告した．

4.2 実験結果

表 2 は ThreeDirections の実験結果を示したものである．3 言語すべてにおいて TsimNoTarget で最も高い翻訳性能が得られた．3.1 節で述べたように，

表 2 ThreeDirections の実験結果

	Tsim	**TsimNoTarget**	**Target**
en → it	29.3	30.8	29.3
en → nl	26.4	27.8	25.8
en → ro	21.6	23.0	22.1

表 3 SixDirections の実験結果

		Tsim	**TsimNoTarget**	**Target**
zero-shot	it → nl	9.0	8.4	3.6
	it → ro	9.0	10.0	9.1
	nl → it	9.2	9.1	7.1
	nl → ro	7.6	8.4	7.1
	ro → it	9.9	10.1	7.8
	ro → nl	8.2	7.3	3.5
supervised	en → it	29.9	29.1	27.9
	it → en	33.6	33.2	31.3
	en → nl	25.6	25.7	24.9
	nl → en	28.0	28.1	27.3
	en → ro	21.7	21.2	21.1
	ro → en	28.1	28.2	26.2

Target と比較して Tsim および TsimNoTarget で挿入されるタグは，翻訳方向の決定という観点からは不利な設定であるため，これは直感に反する．また，表 3 に SixDirections の実験結果を示す．en→xx もしくは xx→en と英語が含まれる翻訳方向は学習時に与えたものと同じ翻訳方向（supervised）であり，参考値として載せている．zero-shot，supervised のどちらにおいても Tsim，TsimNoTarget が Target と比較して高い翻訳性能を示した．表 2 および，表 3 の結果は，言語特徴をモデルに導入することで翻訳性能を向上することができる可能性を示している．ここで，表 3 から zero-shot 翻訳方向の性能が supervised 翻訳方向と比較して大きく劣ることが見て取れる．BLEU における 1-gram の一致率を確認した結果，Supervised 翻訳方向では参照訳との 1-gram 一致率は平均で 60.7%であったのに比べ，zero-shot 翻訳方向では平均で 35.7%と低いことがわかった．ここで，文字 n-gram ベースの言語判別ツールである langdetect[2] を利用して，本来の目的言語とは異なる単語をどれだけ出力してしまったかを検証した結果を表 4 に示す．この結果は，本研究と同様のデータセットで実験を行っている Wu et al. [12] の結果と同様であり，zero-shot 翻訳方向でも Supervised 方向に匹敵するような翻訳性能を得るためには学習に用

2) https://github.com/Mimino666/langdetect

表 4　翻訳方向の誤り割合（%）

	Tsim	TsimNoTarget	Target
it → nl	18.9	19.2	66.4
it → ro	10.4	7.85	9.96
nl → it	12.8	14.8	25.2
nl → ro	5.80	5.15	8.90
ro → it	11.3	9.33	20.7
ro → nl	18.1	17.6	64.6

表 5　類型論特徴の導入実験

	Typ-insert	Target
en → it	28.3	29.3
en → nl	25.2	25.8
en → ro	20.9	22.1

いる翻訳データセットのサイズが重要な要素であると考える．

最後に，3.2 節で説明した類型論特徴の導入実験の結果を表 5 に示した．言語特徴を導入した場合でも，Johnson et al.[2] のものと同等の翻訳性能が得られた．

5　おわりに

本研究は，言語に内在する共通性を探る言語学の一分野である言語類型論にて蓄積されたデータを多言語機械翻訳モデルに導入する方法を示した．言語特徴をモデルに導入することで，説明性だけでなく翻訳性能を向上させうる可能性を示した．今後は，モデルが言語特徴を推論により効果的に利用できる表現力の高い導入方法を検討したい．

謝辞

本研究は JSPS 科研費 JP20K23325 の助成を受けたものである．

参考文献

[1] Barret Zoph and Kevin Knight. Multi-source neural translation. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 30–34, San Diego, California, June 2016. Association for Computational Linguistics.

[2] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 339–351, 10 2017.

[3] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 866–875, San Diego, California, June 2016. Association for Computational Linguistics.

[4] Thanh-Le Ha, Jan Niehues, and Alex Waibel. Toward multilingual neural machine translation with universal encoder and decoder. In **Proceedings of the 13th International Workshop on Spoken Language Translation**, pp. 1–7, 2016.

[5] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. Linguistic typology meets universal dependencies. In **TLT**, pp. 63–75, 2017.

[7] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of NAACL-HLT 2019: Demonstrations**, 2019.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[9] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[10] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.

[11] Matthew S. Dryer and Martin Haspelmath, editors. **WALS Online**. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

[12] Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. Language tags matter for zero-shot neural machine translation. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 3001–3007, Online, August 2021. Association for Computational Linguistics.