ニューラル機械翻訳における逆翻訳へのタグ付与の効果

Benjamin Marie[†] · Raphael Rubino[†] · 藤田 篤[†]

本稿では、2020 年 7 月に開催された The 58th Annual Meeting of the Association for Computational Linguistics (ACL) において我々が発表したショート論文 “Tagged Back-translation Revisited: Why Does It Really Work?” (Marie, Rubino, and Fujita 2020) について、日本語で紹介する。

1 本論文で取り組んだ問題

ニューラル機械翻訳 (NMT) システムの性能を向上させるアプローチとして、対訳データに加えて単言語データを利用する手法がいくつか存在する。その中で、訓練済みの NMT システムを用いて目標言語の単言語データを起点言語に翻訳することで疑似対訳データを生成し、新たな NMT システムの訓練に用いる『逆翻訳法』(Sennrich, Haddow, and Birch 2016a) は、NMT システムの特にデコーダを強化し、訳文の流暢性を向上させられることが知られている。特に、(正規の) 対訳データが小規模にしか存在しない状況下で有用である。

このようにして生成された疑似対訳データには、流暢でない起点言語文や正しい翻訳ではない文対が含まれる。したがって、逆翻訳法で NMT システムを訓練すると、疑似対訳データの特徴が学習され、それらが翻訳時 (デコード時) にも反映されると予想される。NMT システムの訓練時に疑似対訳データと正規の対訳データを区別するために、Caswell, Chelba, and Grangier (2019) は、疑似対訳データの起点言語側の各文の先頭にタグを付与する手法を提案した。この手法は、元々は多言語 NMT システムの訓練時に異なる言語対の対訳データを区別するために考案された (Johnson, Schuster, Le, Krikun, Wu, Chen, Thorat, Viégas, Wattenberg, Corrado, Hughes, and Dean 2017) ものであるが、単一言語対の分野適応にも有用であり (Chu, Dabre, and Kurohashi 2017)、逆翻訳法においても翻訳性能を向上できることが経験的に示された。しかしながら、実際に異なる対訳データを区別できているか、どのような違いが学習されているかは解明されていない。

人間が直接作成したテキスト (以下、原テキスト) と比べると、人間によって翻訳されたテキスト (以下、翻訳テキスト) は、より平易で、文長が短く、語彙が小さいなどの特徴を持つ (Laviosa-Braithwaite 1998)。これらの特徴は、逆翻訳法で用いられる疑似対訳データの起点言

[†] 国立研究開発法人情報通信研究機構 (NICT)

語文の特徴 (Burlot and Yvon 2018) とも重なる。そこで、逆翻訳へのタグ付与 (Caswell et al. 2019) によって正規の対訳データと疑似対訳データの違いを学習できるか否かを、原テキストと翻訳テキストに対する NMT システムの性能比較を通じて検証した。

2 実験設定

WMT19 の英語・独語翻訳 (en-de)¹, WMT15 の英語・仏語翻訳 (en-fr)² の対訳データを用いて NMT システムを訓練、評価した。逆翻訳の生成元となる単言語データとしては、英語、独語、仏語の各々について WMT が提供するすべての News Crawl から無作為に抽出した 2,500 万文を用いた。モデル選択のための開発用データには、英語・独語翻訳については newstest2016 を、英語・仏語翻訳については newstest2013 を用いた。これらには原テキストと翻訳テキストの両方が比較的均等に含まれているためである。評価用データには、英語・独語翻訳については WMT10 から WMT19 まで、英語・仏語翻訳については WMT08 から WMT15 までのすべての newstest のうち、上記の開発用データ以外を用いた。Moses³ のツールを用いてすべてのデータをトークナイズし、文字の大小を統制 (truecasing) した。

NMT システムとしては、Transformer Base モデル (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017) を Marian⁴ を用いて訓練した。各言語の語彙は、対訳データの各言語部分のみに基づいて、バイト対符号化 (Sennrich, Haddow, and Birch 2016b) で 32,000 語を定めた。その他の設定の詳細は Marie et al. (2020) を参照されたい。

まず正規の対訳データのみを用いて NMT システムを訓練し、次にそのシステムを用いて目標言語の単言語データを翻訳することで疑似対訳データを作成した。その際、ビーム幅は 12、長さに対する正規化重みは 1.0 とした。そして、正規の対訳データと疑似対訳データを混合して用い、新たな NMT システムを訓練した。データの混合前に疑似対訳データの各起点言語文にタグを付与して得たシステムを “T-BT”、タグを用いずに得たシステムを “BT” と記す。

すべてのシステムを、BLEU スコア (Papineni, Roukos, Ward, and Zhu 2002) で評価した。その際、SacreBLEU⁵ の “--origlang” オプションで “L1” および “non-L1” (L1 は起点言語のコード) を指定することにより、原テキストと翻訳テキストに対する翻訳結果を区別して評価した。

¹ <http://www.statmt.org/wmt19/translation-task.html>, ParaCrawl を除く 520 万文対。

² <http://www.statmt.org/wmt15/translation-task.html>, 3,280 万文対。

³ <https://github.com/moses-smt/mosesdecoder>

⁴ <https://marian-nmt.github.io/>, 1.7.6 (1d4ba73)

⁵ <https://github.com/mjpost/sacreBLEU>, パラメタの signature は、BLEU+case.mixed+lang.L1-L2+numrefs.1+{origlang.L1,origlang.non-L2}+smooth.exp+test.XXX+tok.13a+version.1.4.2。ただし、“L1”と“L2”は各々起点言語と目標言語の言語コード (en, de, fr のいずれか)、“XXX”は評価用データセットの名前である。

3 実験結果

表 1 に, BT と T-BT の各々による BLEU スコア, および正規の対訳データのみを用いて訓練した NMT システムによる BLEU スコアとの差分を示す. BT は, 翻訳テキストについては多くの場合に BLEU スコアを向上させたものの, 原テキストに対する BLEU スコアを一貫して低下させた. 一方 T-BT は, 原テキストに対する BT の負の効果を打ち消し, 多くの場合に BLEU スコアを向上させた. BT と T-BT の結果の比較を通じて, 訓練時に正規の対訳データと疑似対訳データをタグで区別することの有用性, およびタグを用いずに訓練した NMT システムは疑似対訳データに対して過度に適合することが確認できた.

訓練時と同様に翻訳時に入力文にタグを付与する場合の効果についても調査した. 表 2 に示すように, 原テキストにタグを付与して T-BT を誤誘導した場合, BLEU スコアが低下した. 一方, 翻訳テキストに対してタグを付与した場合, T-BT は翻訳テキストの翻訳に適した処理を行うことができ, 8 個の評価データのうち 6 個において BLEU スコアを向上させた.

4 考察と今後の課題

実験を通じて, 逆翻訳法による NMT システムの訓練は疑似対訳データの特徴に過度に適合すること, その結果翻訳テキストに対する翻訳品質は向上するものの, 原テキストに対する翻訳品質は一貫して低下することを確認した. 後者は, 「最先端の NMT システムは原テキストよりも翻訳テキストの翻訳が得意である」(Graham, Haddow, and Koehn 2019; Zhang and Toral 2019) という報告とも符号する. ただし, 疑似対訳データに対してタグを付与し, 正規の対訳データと区別して訓練することより, 原テキストに対する翻訳の大幅な品質劣化を回避することができる. また, 翻訳テキストの翻訳時に同じタグを付与することで翻訳品質をさらに改善できる可能性がある.

Marie et al. (2020) では, 正規の対訳データを 20 万文対に減らした少資源設定の実験と, より大規模な疑似対訳データ(en→fr では 7,660 万文対, それ以外では 1 億 5,000 万文対)を用いた実験についても述べている. 少資源の実験では, 表 1 とは異なる結果が観察された. 具体的には, 逆翻訳法は原テキストの翻訳, 翻訳テキストの翻訳の両方において常に翻訳品質を改善し, タグ付与の効果は希薄であった. 大規模な疑似対訳データを用いた実験では, 逆翻訳の品質の影響がより濃く出るため, BT の原テキストに対する翻訳品質が大幅に劣化し, その分だけタグ付与の効果も大きかった. ただし, 最終的な BLEU スコアは, 表 1 の T-BT の結果を大きく改善するほどではなかった. 以上より, 逆翻訳法で NMT システムの性能向上を図る際は, 疑似対訳データに対してタグを付与し, 正規の対訳データと区別すべきである.

今後は, 翻訳テキストの自動検出に関する試み(Rabinovich and Wintner 2015; Rubino, Lapshinova-

表 1 逆翻訳 (BT) およびタグ付き逆翻訳 (T-BT) を用いて訓練した NMT システムによる BLEU スコア

システム	newstest	de→en			en→de		
		全体	原テキスト	翻訳テキスト	全体	原テキスト	翻訳テキスト
BT	2010	28.9 (+0.5)	33.2 (−0.9)	27.9 (+0.7)	21.8 (−2.3)	24.6 (−5.7)	21.0 (−1.2)
	2011	25.3 (−0.3)	29.9 (−1.0)	24.2 (−0.2)	19.9 (−1.4)	23.8 (−1.9)	19.0 (−1.1)
	2012	27.1 (+0.3)	27.9 (−1.6)	27.0 (+0.7)	20.4 (−1.2)	24.5 (−4.6)	19.3 (−0.2)
	2013	30.3 (+0.3)	34.7 (−1.6)	29.2 (+0.6)	23.8 (−1.9)	25.1 (−2.8)	23.6 (−1.7)
	2014	32.8 (+2.2)	27.4 (−2.5)	36.8 (+7.0)	25.4 (−0.5)	23.2 (−3.3)	27.9 (+2.7)
	2015	33.8 (+2.4)	22.5 (−1.9)	39.5 (+5.5)	27.2 (−1.1)	28.1 (−2.9)	24.7 (+1.9)
	2017	35.5 (+3.0)	27.2 (−1.1)	42.8 (+7.4)	26.4 (−0.1)	26.3 (−3.6)	25.5 (+3.3)
	2018	43.9 (+4.6)	32.0 (−1.0)	53.8 (+10.4)	38.0 (−1.4)	38.9 (−5.9)	35.0 (+3.8)
	2019	—	33.1 (−1.5)	—	—	31.4 (−4.8)	—
T-BT	2010	29.5 (+1.1)	34.4 (+0.3)	28.4 (+1.2)	25.0 (+0.9)	30.5 (+0.2)	23.4 (+1.2)
	2011	26.4 (+0.8)	31.7 (+0.8)	25.2 (+0.8)	22.1 (+0.8)	25.8 (+0.1)	21.0 (+0.9)
	2012	28.1 (+1.3)	30.2 (+0.7)	27.7 (+1.4)	22.8 (+1.2)	30.0 (+0.9)	20.9 (+1.4)
	2013	30.8 (+0.8)	36.0 (−0.3)	29.6 (+1.0)	26.4 (+0.7)	28.1 (+0.2)	26.1 (+0.8)
	2014	32.4 (+1.8)	29.6 (−0.3)	33.8 (+4.0)	27.9 (+2.0)	26.7 (+0.2)	29.4 (+4.2)
	2015	33.9 (+2.5)	24.9 (+0.5)	37.7 (+3.7)	29.9 (+1.6)	32.1 (+1.1)	25.6 (+2.8)
	2017	35.5 (+3.0)	28.1 (−0.2)	41.2 (+5.8)	28.7 (+2.2)	30.7 (+0.8)	26.0 (+3.8)
	2018	43.2 (+3.9)	33.0 (+0.0)	50.4 (+7.0)	41.8 (+2.4)	45.6 (+0.8)	35.5 (+4.3)
	2019	—	35.0 (+0.4)	—	—	37.6 (+1.4)	—
システム	newstest	fr→en			en→fr		
		全体	原テキスト	翻訳テキスト	全体	原テキスト	翻訳テキスト
BT	2008	22.9 (−1.7)	27.9 (−2.6)	22.2 (−1.5)	23.2 (−0.2)	21.2 (−3.3)	23.6 (+0.5)
	2009	26.5 (−2.3)	41.1 (−5.3)	23.9 (−1.6)	27.7 (+1.1)	22.7 (−2.0)	28.4 (+1.4)
	2010	29.3 (−1.4)	27.4 (−7.8)	29.5 (+0.5)	28.2 (−0.5)	22.5 (−11.1)	29.8 (+2.5)
	2011	29.4 (−1.9)	29.3 (−4.7)	29.4 (−1.1)	30.9 (+0.0)	36.7 (−8.2)	29.3 (+2.1)
	2012	29.7 (−1.4)	34.3 (−4.3)	28.6 (−0.6)	28.4 (+1.1)	26.3 (−4.1)	29.0 (+2.5)
	2014	36.6 (+0.6)	31.4 (−4.7)	40.3 (+5.6)	32.9 (−3.1)	26.1 (−12.1)	39.6 (+6.1)
	2015	36.2 (+0.0)	40.9 (−3.1)	29.8 (+3.5)	35.7 (+1.7)	25.1 (−4.4)	44.9 (+6.5)
T-BT	2008	24.5 (−0.1)	29.5 (−1.0)	23.7 (+0.0)	23.8 (+0.4)	25.1 (+0.6)	23.5 (+0.4)
	2009	28.9 (+0.1)	46.4 (+0.0)	25.7 (+0.2)	27.3 (+0.7)	25.1 (+0.4)	27.7 (+0.7)
	2010	31.2 (+0.5)	35.1 (−0.1)	29.6 (+0.6)	30.0 (+1.3)	34.1 (+0.5)	28.9 (+1.6)
	2011	31.8 (+0.5)	33.3 (−0.7)	31.4 (+0.9)	31.6 (+0.7)	45.3 (+0.4)	28.0 (+0.8)
	2012	31.8 (+0.7)	38.3 (−0.3)	30.1 (+0.9)	28.9 (+1.6)	31.9 (+1.5)	28.1 (+1.6)
	2014	37.3 (+1.3)	36.1 (+0.0)	37.2 (+2.5)	38.2 (+2.2)	39.7 (+1.5)	36.5 (+3.0)
	2015	36.6 (+0.4)	43.2 (−0.8)	27.9 (+1.6)	36.0 (+2.0)	30.7 (+1.2)	41.2 (+2.8)

括弧内の数値は、正規の対訳データのみを用いて訓練した NMT システムによる BLEU スコアとの差分。

表 2 評価用データにもタグを付与した場合の T-BT モデルの BLEU スコアの増減

翻訳対象	de→en		en→de		fr→en		en→fr	
	2017	2018	2017	2018	2012	2015	2012	2015
原テキスト	−2.0	−2.6	−5.9	−9.6	−7.5	−4.9	−10.1	−11.1
翻訳テキスト	+1.6	+3.4	+0.8	+1.6	−3.1	+1.4	−0.3	+3.6

Koltunski, and van Genabith 2016) をふまえ, 正規の対訳データ中における原テキストと翻訳テキストの識別, それに基づく NMT システムの訓練技術について研究を進める予定である.

参考文献

- Burlot, F. and Yvon, F. (2018). “Using Monolingual Data in Neural Machine Translation: A Systematic Study.” In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, pp. 144–155, Brussels, Belgium. Association for Computational Linguistics.
- Caswell, I., Chelba, C., and Grangier, D. (2019). “Tagged Back-Translation.” In *Proceedings of the 4th Conference on Machine Translation (Volume 1: Research Papers)*, pp. 53–63, Florence, Italy. Association for Computational Linguistics.
- Chu, C., Dabre, R., and Kurohashi, S. (2017). “An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 385–391, Vancouver, Canada. Association for Computational Linguistics.
- Graham, Y., Haddow, B., and Koehn, P. (2019). “Translationese in Machine Translation Evaluation.” *arXiv preprint arXiv:1906.09833*.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation.” *Transactions of the Association for Computational Linguistics*, **5**, pp. 339–351.
- Laviosa-Braithwaite, S. (1998). “Universals of Translation.” In *Routledge Encyclopedia of Translation Studies*, pp. 288–291. Routledge.
- Marie, B., Rubino, R., and Fujita, A. (2020). “Tagged Back-translation Revisited: Why Does It Really Work?” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5990–5997, Online. Association for Computational Linguistics.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “Bleu: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, USA. Association for Computational Linguistics.
- Rabinovich, E. and Wintner, S. (2015). “Unsupervised Identification of Translationese.” *Transactions of the Association for Computational Linguistics*, **3**, pp. 419–432.
- Rubino, R., Lapshinova-Koltunski, E., and van Genabith, J. (2016). “Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classifica-

- tion.” In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 960–970, San Diego, USA. Association for Computational Linguistics.
- Sennrich, R., Haddow, B., and Birch, A. (2016a). “Improving Neural Machine Translation Models with Monolingual Data.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany. Association for Computational Linguistics.
- Sennrich, R., Haddow, B., and Birch, A. (2016b). “Neural Machine Translation of Rare Words with Subword Units.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). “Attention is All you Need.” In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pp. 5998–6008. Curran Associates, Inc.
- Zhang, M. and Toral, A. (2019). “The Effect of Translationese in Machine Translation Test Sets.” In *Proceedings of the 4th Conference on Machine Translation (Volume 1: Research Papers)*, pp. 73–81, Florence, Italy. Association for Computational Linguistics.

略歴

Benjamin Marie: 2016 年フランス・Paris-Sud University で Ph.D. を取得。同年より NICT 先進的音声翻訳研究開発推進センター研究員。主に機械翻訳を含む多言語処理の研究に従事。

Raphael Rubino: 2011 年フランス・Avignon University で Ph.D. を取得。2019 年より NICT 先進的音声翻訳研究開発推進センター研究員。主に機械翻訳、テキスト品質評価、機械学習の研究に従事。

藤田 篤: 2005 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。博士(工学)。2014 年より NICT 先進的音声翻訳研究開発推進センター主任研究員。主に言い換え表現の生成と認識、機械翻訳の研究に従事。