機械翻訳のためのチェックリスト 📝

星野　翔
株式会社みらい翻訳 *
shhshn.shohoshino@gmail.com

1　はじめに

ニューラル機械翻訳（NMT）の進歩は翻訳精度の向上に必ずしも結びついていない。一般に NMT の出力は流暢性 fluency が高いが、正確性 adequacy に難があると言われており [1]、統計的機械翻訳 (SMT) の出力より正確性が悪化する場合すらある。その原因のうちに、過剰生成 over-generation や過少生成 under-generation と呼ばれる現象がある。NMT の出力には過不足があり、入力にない情報を過剰に出力したり、入力にある情報を欠落させることがある。SMT では、入力を組み替えて出力を構成していたため、これら現象は原理上発生しにくかった。

こうした膠着状態からの脱却を目指して、我々は新しいタスク「機械翻訳のためのチェックリスト」を提案する。このタスクの目的は、参照訳を用いず、機械翻訳の入出力から文レベルの自明な誤りを検出することで、翻訳精度を底上げすることである。具体的には、NMT で数多く報告される過剰生成・過少生成を検出するが、誤訳 mistranslation は検出しない。このチェックリストには二通りの利用方法があり、1. NMT に適用することで最低限の質を担保できるし、2. 対訳データに適用することで対訳ノイズを発見できる。またブラックボックスの NMT を対象として、特定のアーキテクチャに依存しない。後者の利用方法のように対象システムが NMT でなくとも良い。出力のスコア化はチェックリストの対象外である。

従来、機械翻訳の誤り検出のタスクとしては、主に自動評価、誤り分析、品質推定 (QE) の三種類が用いられている。1. 自動評価は、出力と参照訳の近さをスコア化するタスクで、BLEU [2] が代表的な尺度である。2. 誤り分析 [3, 4, 5] は、参照訳に基づき出力の誤りを分類するタスクで、スコア化を目的としない点で提案手法に最も近い。3. QE [6, 7] は、参照訳を用いず出力を評価するタスクで、文レベルでは HTER [8] に基づき出力の後編集に必要な回数をスコア化する。このように、従来研究とチェックリストは似て異なる目的を持つため、新しいタスクとして提案する。

2　提案手法

ここから我々によるチェックリストの実装案である、単語アライメントに基づくチェックリストを説明する。提案手法では、NMT の入出力の他に、単語アライメントが利用可能であると仮定する。単語アライメント利用の是非は後述するが、基本的に対訳データで単語アライメントを教師なし学習し、force align によって予測する。単語はサブワード [9] またはバイト [10] でも良い。

つまり NMT の入力または対訳データの原言語文（以下、入力）を $\mathcal{X} = [x_1, \ldots, x_X \mid X \in \mathbb{Z}]$、NMT の出力または対訳データの目的言語文（以下、出力）を $\mathcal{Y} = [y_1, \ldots, y_Y \mid Y \in \mathbb{Z}]$、さらに入出力間の単語アライメントを $\mathcal{A} = \{(f, e) \mid 1 \leq f \leq X \wedge 1 \leq e \leq Y\}$ と定義する。提案手法には三つ組 $(\mathcal{X}, \mathcal{Y}, \mathcal{A})$ が与えられ、0 または 1 を出力して終了する。文レベルで何らかの誤りを検出できれば 1 を出力し、誤りを全く検出できなければ 0 を出力する。

提案手法は、基本的な考え方として、なるべく再現率が高くなる検出方法を目指す。そうでなければチェックリストとしての信頼性を損うためである。そこで具体的には、過剰生成・過少生成の検出を 1. 反復、2. 中断、3. 過剰生成、4. 過少生成の四つのサブタスクに細分化する。そして個々の性能を測定・改善することで、チェックリスト全体としての性能の向上を試みる。各サブタスクへの取り組み方を以下順に説明する。

* 本稿は著者個人の見解であり、株式会社みらい翻訳公式の見解ではない。

2.1 反復

過剰生成の特殊ケースで、あるフレーズが繰り返し出力される現象を仮に反復 repetition と呼ぶ。反復は、入力に繰り返しが含まれない限り、出力に繰り返しが含まれるか否かで検出できる。[1)]

つまり入出力について、繰り返しが含まれる出力 $\mathcal{Y}' = [y_1, \ldots, y_i, \ldots, y_j, y_i, \ldots, y_j, \ldots, y_{Y'} \mid 1 \leq i, j \leq Y']$ が空集合にならず、一方で繰り返しが含まれる入力 $\mathcal{X}' = [x_1, \ldots, x_k, \ldots, x_l, x_k, \ldots, x_l, \ldots, x_{X'} \mid 1 \leq k, l \leq X']$ が空集合になり、同時に成り立たないものとする。このとき 1 を出力する。

2.2 中断

過少生成の特殊ケースで、入力のごく一部分に相当する情報しか出力されない現象を仮に中断 termination と呼ぶ。中断は、入出力の長さ比が極端に変化しない限り、入出力の長さ比に閾値を設けることで検出できる。[2)]

つまり入力の単語数 X と 出力の単語数 Y を見比べて、長さ比の閾値 $r \in \mathbb{R}$ について $\frac{Y}{X} \leq r$ ならば中断とみなし 1 を、$\frac{Y}{X} > r$ ならば 0 を出力する。閾値 r は提案手法のハイパーパラメータであり、本研究では事前実験で性能が良かった $r = 0.5$ を用いる。

2.3 過剰生成と過少生成

入力に含まれない情報の出力を過剰生成、入力に含まれる情報の出力での欠落を過少生成と呼ぶ。例えば、湧き出しは過剰生成であり、訳抜けは過少生成である。過剰生成・過少生成の検出は難しいため、まず反復・中断を取り除く。その上で、信頼性の高い単語アライメントが利用可能であると仮定し、単語アライメントを用いて判定する。

2.3.1 過剰生成

過剰生成では、入力に含まれない出力情報の有無を単語アライメントで判定する。しかし、単語アライメントが null の単語を誤判定する可能性がある。例えば、英語の冠詞 the などストップワードが該当する。そこで安全幅 $n \in \mathbb{Z}$ を設けて、前後 n 単語を含むフレーズ全体が null かどうかで判定する。

つまり $\mathcal{Y}'' = [y_1, \ldots, y_e, \ldots, y_{Y''} \mid 1 \leq e \leq Y'' \wedge (f, e') \notin \mathcal{A} \wedge e - n \leq e' \leq e + n]$ が空集合でなければ 1 を出力する。安全幅 n も提案手法のハイパーパラメータであり、本研究では事前実験で性能が良かった $n = 3$ を用いる。

2.3.2 過少生成

過少生成では、出力に含まれない入力情報の有無を単語アライメントで判定する。過少生成でも、過剰生成と同じく共通の安全幅 n を用いて、前後 n 単語を含むフレーズ全体が null かどうかで判定する。

つまり $\mathcal{X}'' = [x_1, \ldots, x_f, \ldots, x_{X''} \mid 1 \leq f \leq X'' \wedge (f', e) \notin \mathcal{A} \wedge f - n \leq f' \leq f + n]$ が空集合でなければ 1 を出力する。

2.4 単語アライメント利用の是非

このような利用に耐えうる信頼性の高い単語アライメントの入手は難しい。NMT で一般的なアテンションは単語アライメントではない。そのため NMT と単語アライメントをジョイントないしマルチタスクで学習する研究 [13, 14] がある。提案手法では、特定の手法に依存しないために、別途学習・予測された単語アライメントが与えられると仮定する。

こうしてチェックリストの準備が整ったため、実世界のデータを用いてその性能と有効性を明らかにする。

3 実験

3.1 実験設定

提案手法を NMT の出力および対訳データへそれぞれ適用し、チェックリストとしての利用に耐えうるか検証する。単語アライメントの信頼性の観点で、NMT の出力は対訳データより難易度が高くなると予想される。そのため、適用対象を 1. 単語アライメントの学習に使用した対訳データ、 2. その他の対訳データ、 3. NMT の出力 と三種別に細分化し有効性を検証する。提案手法の評価方法としては、性能の定量化が難しいため、人手で定性的に分析する。

NMT は、`fairseq v0.10.2` [15] と英独翻訳モデル [16] を組み合わせて使用する。WMT2018 [17] のニュース分野で良い成績を収めた訓練済みの翻訳モデルが一般公開されており、再現実験が容易なため

1) OpenNMT [11] に似た機能が実装されているが、翻訳精度低下の恐れがある NMT ビーム探索中の検出ではなく、出力後の検出を目的としたい。

2) 目的が異なるが、入出力の長さ比を用いて NMT ビーム探索を補正する手法 [12] が提案されている。

区分	wmt18 学習用	news12 開発用	news18 評価用
種別	対訳データ	対訳データ	NMT 出力
force align	無し	有り	有り
データ量	5,223,425	3,003	2,998
反復	13	0	0
中断	0	13	5
過剰生成	35,472	14	3
過少生成	38,034	18	2
検出せず	5,149,766	2,958	2,988

表 1 提案手法によるチェックリストの対象と実験結果

対象に選定する。もちろん他の機械翻訳システム・翻訳モデルでも構わない。

単語アライメントは、`fast_align` [18] を使用して WMT2018 の学習データから教師なし学習する。性能向上のため、英独と独英の両方向で学習し、最後に `grow-diag-final-and` ヒューリスティックを使用して多対多 many-to-many の単語アライメントに仕上げる。学習データ以外については force align で予測し、同様の手順で多対多の単語アライメントを出力する。

実験データは、同じく WMT2018 の学習データ（以下 wmt18）と評価データ（以下 news18）に加えて、開発データとして WMT2012 [19] の評価データ（以下 news12）を使用する。データ作成方法など実験設定の詳細は付録 A で説明する。

3.2 実験結果

表 1 に、提案手法によるチェックリストの出力結果をデータ種別とサブタスク別に示す。四つのサブタスクの何れでも誤りを検出できなかった場合は「検出せず」に分類した。提案手法は約 1%のデータで文レベルの誤りを検出した。

まずデータ種別による違いとして、反復・中断の有無があった。本研究では NMT の出力である news18 に反復が発生しなかったが、対訳データ中に反復の対訳ノイズが存在していた wmt18 では検出に成功した。反対に、news18 では中断が発生したが、wmt18 では検出できなかった。

次にサブタスクによる違いとして、wmt18 での反復・中断の少なさに対して、過剰生成・過少生成の多さが際立った。単語アライメントの利用により対訳ノイズの大半を過剰生成・過少出力に分類できたことが要因と考えられるが、一方でストップワードの誤検出も多く、過剰生成・過少生成をさらに細分化する必要性も示唆される。

図 1 に、対訳データでの実際の検出例を示す。例えば、wmt18 の反復では、文全体が繰り返されている明らかな対訳ノイズを検出できた。また news12 での中断では、英語入力に対して独語出力が不自然に短い対訳ノイズを検出できた。また wmt18 と news12 の両方で過剰生成・過少生成を検出できている。特に news12 の過少生成では、一見「Poland」を「Pole」と訳せているが、「Polen」と訳すべきだったと考えられる。このように、提案手法の有効性を確認できたものの、チェックリストとしての性能にはまだまだ改善の余地がある。

3.3 議論

チェックリストの性能を地道に改善していくことで当初の目的である NMT の翻訳精度の底上げを実現できるだろうか。提案手法は単語アライメントが利用可能であるという強い前提を置いているが、この問題設定にいささか無理があるとも言える。仮に完璧な単語アライメントを利用できれば、NMT の自動評価は事実上完成するし、また NMT そのものを学習し直すことができるため、機械翻訳と同等以上に難しい部分問題を前提としている疑いがある。

このような疑問には以下のように応えたい。まず機械翻訳の評価そのものは人間にとっても難しい [20] が、単語アライメントを人手で作成するタスクの一致率は比較的高く、教師なし学習による機械化もある程度進んでいる。そのため、単語アライメントは機械翻訳やその評価より難易度が低い部分問題と言える。また実世界では、NMT の誤訳検出にも増して、NMT の信頼性に直結する過剰生成・過少生成の検出が喫緊の課題となっている。そのため我々は、過剰生成・過少生成の検出を目的とする単語アライメントの利用を妥当な落とし所と考える。

4 関連研究

機械翻訳の自動評価には、n-gram の適合率に基づく BLEU [2] や SacreBLEU [21] が広く用いられている。また順位相関係数に基づく Birch と Osborne [22] や Isozaki ら [23] の研究がある。単語アライメントに基づく手法としては、松尾ら [24] の研究がある。

機械翻訳の誤り分類では、人手分類のための Vilar

データ	サブタスク	入出力	検出対象（下線部）
wmt18	反復	独語出力	Neue Bilder von DOFUS 2.0 ! Neue Bilder von DOFUS 2.0 !
	過剰生成	英語入力	Later , he became Commissioner for the Internal Market , Customs Union , Industrial Innovation , Environment and Consumers and Vice-President for Industrial Policy , Research and Innovation .
		独語出力	Als Kommissar gehörte er der Kommission , zunächst als Kommissar für Binnenmarkt , Zollunion , industrielle Innovation , Umwelt und Verbraucherfragen an sowie von 1984 bis 1988 als Vizepräsident für Industriepolitik , Forschung und Innovation .
	過少生成	英語入力	(ES) Mr President , ladies and gentlemen , firstly I must thank this House for its serious and detailed work .
		独語出力	Als erstes möchte ich dem Parlament für die geleistete Arbeit danken . Ich glaube , daß ernsthaft und gründlich gearbeitet wurde
news12	中断	英語入力	How do you explain this change ?
		独語出力	Warum ?
	過剰生成	英語入力	Mayor Bloomberg told reporters that , because of that court order , the city had suspended the reopening of the public space and protesters were informed , however , that local laws do not allow them to re-install with camping shops and sleeping bags .
		独語出力	Bürgermeister Bloomberg stellt vor der Presse klar , das aufgrund dieser richterlichen Anordnung die erneute Öffnung des Platzes für den Publikumsverkehr und die Demonstranten aufgehoben worden sei . Die Demonstranten wies er darauf hin , dass die Stadtgesetze ihnen nicht erlaubten , sich erneut mit Zelten und Schlafsäcken an diesem Ort einzurichten .
	過少生成	英語入力	” Believe me that when I meet someone from Poland and Hungary , and a Slovak is coming to join us , the Frenchman next to me says : ” Hey , what ’s that gathering about , is that the Visegrad Group again ? ’ and he is not happy about it in the slightest . ’
		独語出力	” Glauben Sie mir , wenn eine Pole oder Ungar zu mir kommt und ein Slowake in der Nähe ist , dann meint der Franzose neben mir : ” He , was steckt ihr da die Köpfe zusammen - ist das schon wieder eure V4 ? ” und dabei macht eine ziemlich saure Miene .

図 1 提案手法によるチェックリストの文レベル誤り検出例

ら [3] の研究がある。また自動的な誤り分類のための Popović ら [4] や Akabe ら [5] の研究がある。

機械翻訳における参照訳を用いない QE [6, 7] には、単語レベルと文レベルの 2 つのタスクがあるが、文レベルでは HTER [8] を予測する。具体的な手法としては predictor-estimator に基づく POSTECH [25] などが用いられる [26]。参照訳を用いず評価する点において、我々のチェックリストも志は同じである。

NMT における誤り検出の取り組みとしては、訳抜けの検出に NMT のアテンション及び逆翻訳を利用する後藤と田中 [27] の研究がある。さらに対訳データの訳抜けを NMT のアテンションや単語アライメントから検出し、反復回数を人手で検出する後藤ら [28] の研究がある。

5 おわりに

本研究は機械翻訳のためのチェックリストという新しいタスクを考案し、その実装案として、単語アライメントに基づく手法を提案した。実世界での英独 NMT 翻訳モデルと対訳データを用いた実験で、提案手法のチェックリストとしての有効性を示した。提案手法の限界として、自明ではない誤訳を検出できず、また NMT への対処療法となっている欠点があり、今後の研究が待たれる。このチェックリストが有効活用されることで、将来的に NMT の翻訳精度が底上げされることを願う。

参考文献

[1] 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.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-

Jing Zhu. 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, 2002.
[3] David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error analysis of statistical machine translation output. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, pp. 697–702, 2006.
[4] Maja Popović and Hermann Ney. Towards automatic error analysis of machine translation output. Computational Linguistics, Vol. 37, No. 4, pp. 657–688, 2011.
[5] Koichi Akabe, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Discriminative language models as a tool for machine translation error analysis. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1124–1132, 2014.
[6] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In Proceedings of the 20th International Conference on Computational Linguistics, pp. 315–321, 2004.
[7] Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. Estimating the sentence-level quality of machine translation systems. In Proceedings of the 13th Annual conference of the European Association for Machine Translation, pp. 28–35, 2009.
[8] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation.
[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1715–1725, 2016.
[10] Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, pp. 9154–9160, 2020.
[11] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 67–72, 2017.
[12] Yilin Yang, Liang Huang, and Mingbo Ma. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3054–3059, 2018.
[13] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Neural machine translation with supervised attention. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, pp. 3093–3102, 2016.
[14] Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. Jointly learning to align and translate with transformer models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 4453–4462, 2019.
[15] 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 48–53, 2019.
[16] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 489–500, 2018.
[17] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pp. 272–303, 2018.
[18] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 644–648, 2013.
[19] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, editors. Proceedings of the Seventh Workshop on Statistical Machine Translation, 2012.
[20] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 33–41, 2013.
[21] Matt Post. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186–191, 2018.
[22] Alexandra Birch and Miles Osborne. LRscore for evaluating lexical and reordering quality in mt. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 327–332, 2010.
[23] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 944–952, 2010.
[24] 松尾潤樹, 小町守, 須藤克仁. 単語分散表現を用いた単語アライメントによる日英機械翻訳の自動評価尺度. 情報処理学会研究報告, Vol. 2016-NL-229, No. 20, pp. 1–7, 2016.
[25] Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In Proceedings of the Second Conference on Machine Translation, pp. 562–568, 2017.
[26] Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An open source framework for quality estimation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 117–122, 2019.
[27] 後藤功雄, 田中英輝. ニューラル機械翻訳での訳抜けした内容の検出. 言語処理学会 第 23 回年次大会 発表論文集, pp. 1018–1021, 2017.
[28] 後藤功雄, 美野秀弥, 山田一郎. 訳抜けを含む訓練データと訳抜けのない出力とのギャップを埋めるニューラル機械翻訳. 言語処理学会 第 26 回年次大会 発表論文集, pp. 1412–1415, 2020.

	news18
著者ら報告値	
baseline	42.38
+BT	44.94
+ensemble	46.39
+filter copies	46.53
再現実験値	44.80

表 2　WMT2018 英独ニュース翻訳での再現実験結果

A　付録: 実験設定詳細

翻訳モデルとして、Facebook の公開データ[3)]を使用した。このデータにはアンサンブル用モデルと、翻訳結果修正用 source copy filtering モデルが含まれているが、このうち `wmt18.model1.pt` のみ使用した。前処理やハイパーパラメータは同梱の手順 `wmt18.sh` に従った。

参考までに、表 2 に WMT2018 評価データでの SacreBLEU 値を報告する。我々の再現実験以外の値は、翻訳モデル [16] 著者らの報告結果[4)]を借用した。本研究は実験の都合でアンサンブルと source copy filtering を使用しなかったが、著者らの SacreBLEU 値 44.94 に対して 44.80 と、ほぼ同等の性能を再現できたと考える。

単語アライメントの学習では、`fast_align` の最新版[5)]を使用して、基本的にジョイントモデル用の手順[6)]に従った。しかしながら事前実験でチェックリストへの悪影響があったため、サブワード（BPE）を使用しなかった。

3）`https://github.com/pytorch/fairseq/tree/master/examples/backtranslation`
4）`https://arxiv.org/abs/1808.09381v2`
5）`https://github.com/clab/fast_align/commit/cab1e9aac8d3bb02ff5ae58218d8d225a039fa11`
6）`https://github.com/pytorch/fairseq/tree/master/examples/joint_alignment_translation`