Decoder のみを用いた機械翻訳モデルの分析

木山朔　金輝燦　平澤寅庄　岡照晃　小町守
東京都立大学
{kiyama-hajime,kim-hwichan,hirasawa-tosho}@ed.tmu.ac.jp
{teruaki-oka, komachi}@tmu.ac.jp

概要

現在の機械翻訳モデルの主流は Encoder–Decoder モデルである．一方で，Encoder を使わず Decoder のみの翻訳モデル (Decoder-only モデル) が提案されており， Encoder–Decoder モデルと同程度の BLEU が報告されている．しかし既存研究が実験に用いた評価指標や言語対は限定的であり，十分な検証が行われているとは言えない．そこで本研究では Encoder–Decoder モデルと Decoder-only モデル間の詳細な比較・分析を報告する．特に英日方向について人手評価の結果，流暢性と妥当性で Decoder-only モデルの有用性が確認された．

1　はじめに

自然言語処理の様々なタスクにおいて Encoder–Decoder モデルは広く用いられている [1, 2]．特に機械翻訳や文書要約などの生成タスクで Encoder–Decoder が主流になっている．その一方 GPT [3, 4, 5] を代表する Decoder-only モデルの高精度な言語生成も注目を集めている．

Wang ら [6] は GPT-3 の few-shot learning での翻訳が教師なしモデルと同程度の BLEU を獲得したことから, Decoder のみを使った翻訳モデル (Decoder-only モデル) を提案した．原言語文と目的言語文を結合させ，言語モデリングタスクで Decoder を学習し，翻訳モデルとする．図 1 に Decoder-only モデルの推論時の概要図を示す．原言語文とタグ（<en2ja>など）を入力とし目的言語文を生成している．また Gao ら [7] は Wang らのモデルに対し原言語文全体を self-attention で見れるように拡張することで Encoder–Decoder モデルと同等の BLEU を達成できることを報告した．これより Gao らは Encoder–Decoder モデルは冗長であると主張している．しかし彼らの実験では BLEU や TER [8] による自動評価しか行われておらず，Encoder–Decoder, Decoder-only モデル間の翻訳結果の詳細な分析がなされていない．

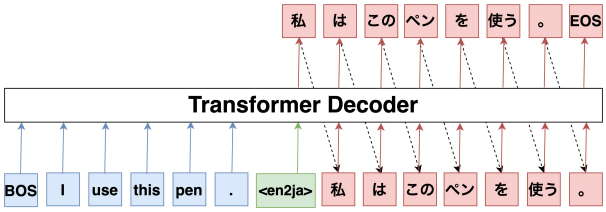

図 1: Decoder-only モデルの推論時の概要図

そこで本研究では自動評価，人手評価の 2 つの観点から Decoder-only モデルの詳細な分析を行った．自動評価では流暢性を測るために perplexity を，妥当性を測るために BLEU に加えて COMET [9] と WMD [10] を追加した．人手評価では定量的な評価として流暢性と妥当性を，定性的な評価として翻訳の質の評価（誤訳の分類）を行なった．本研究での主な発見は以下の通りである：

- 自動評価指標 BLEU，COMET，WMD では Encoder–Decoder モデルの方がスコアが高いが，Perplexity では Deocder-only モデルが Encoder–Decoder モデルを上回った．
- 人手評価において流暢性と妥当性は Decoder-only モデルが良く，Mis-translation も Encoder–Decoder モデルと比較して少ない．

2　関連研究

2.1　Encoder–Decoder

機械翻訳とは，原言語文 $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ を目的言語文 $\mathbf{y} = \{y_1, y_2, ..., y_M\}$ に翻訳するタスクである．ここで $x_i \in \mathbf{x}$, $y_j \in \mathbf{y}$ は各言語の文のトークンを示す．ニューラル機械翻訳では以下の条件付き確率が最大になるように目的言語文を生成する．

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{M} P(y_t|\mathbf{x}, \mathbf{y}_{<t})$$

ここで $\mathbf{y}_{<t}$ は $t-1$ 番目までの目的言語文のトークンである．Sutskever ら [1] は，条件付き確率 P を LSTM [11] で構成される Encoder と Decoder という二つのニューラルネットワークでモデル化した．学習時には，対訳データ $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_j\}, \mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_j\}$ を使い，以下の損失が最小になるようにパラメータを更新する．

$$L_{\mathrm{MT}} = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} -\log P(\mathbf{y}|\mathbf{x})$$
$$= \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \sum_{t=1}^{M} -\log P(y_t|\mathbf{x}, \mathbf{y}_{<t})$$

Encoder–Decoder モデルは機械翻訳や文書要約などの様々なタスクで広く用いられている．特に self-attention を用いた Encoder–Decoder モデルである Transformer [2] は現在の手法のベースとなっている．

2.2 Decoder-only

Decoder-only モデルとして GPT が挙げられる [3, 4, 5]．GPT は Transformer の Decoder のみを用いたモデルであり非常に高精度な言語生成が可能である．Wang ら [6] は GPT-3 の few-shot learning での翻訳が教師なしモデルと同程度の BLEU を獲得したことから，LM4MT という Decoder-only の翻訳モデルを提案した．Decoder-only モデルでは，原言語文 $\mathbf{x}$ にタグ（<en2ja>など）を結合させたものを入力とし，目的言語文 $\mathbf{y}$ を生成する．なお，原言語文 $\mathbf{x}$ は教師強制的 (Teacher forcing) に処理を行い，目的言語文 $\mathbf{y}$ は自己回帰により生成を行う．

Decoder-only モデルの学習時には，原言語側には自己符号化の損失 L_{AE}，目的言語側には L_{MT} を用いる．ハイパーパラメータ λ で重み付けした L_{AE} と L_{MT} の和である L_{LM} が最小になるようにパラメータを学習する．

$$L_{\mathrm{AE}} = \sum_{\mathbf{x}\in\mathbf{X}} \sum_{s=1}^{N} -\log P(x_s|\mathbf{x}_{<s})$$
$$L_{\mathrm{LM}} = \lambda L_{\mathrm{AE}} + L_{\mathrm{MT}}$$

Wang らは学習が進むにつれ λ を小さくなるようにスケジューリングし，目的言語文側の損失を重視するように学習している．

Gao ら [7] は上記 Decoder-only モデルの改良を行った．Gao らは原言語部分の Attention mask を原言語全体が見れるように変更し，原言語文に対してノイズを加え Decoder-only モデルを学習することで，Encoder–Decoder と同様の BLEU を達成した．このことから，Encoder–Decoder モデルは冗長であり，Decoder のみで十分であると主張した．

このように既存研究では，Decoder-only モデルの有用性が報告されているが，BLEU や TER のみでしか評価がされておらず，その他の自動評価や人手評価を用いた詳細な分析がなされていない．そこで，本研究では，BLEU に加え Perplexity，COMET，WMD を用いて自動評価を行い，また，英日対の人手評価として流暢性，妥当性，翻訳の質を分析した．人手評価では Decoder-only モデルが流暢性と妥当性がともに高いことがわかった．

3 実験設定

データセット en-de，en-ja の 2 つの言語対に実験を行った．en-de では WMT16 [12] を，en-ja では ASPEC [13] を用いた．データサイズは順に 4.5M, 3.0M である．トークナイズには英語とドイツ語では moses [1)] [14] を，日本語は MeCab [2)] + ipadic を利用した．その後，en-de 対では joint-BPE [3)] [15] モデル，en-ja 対では言語ごとに BPE の SentencePiece [4)] [16] モデルを学習し，サブワード分割を行った．分割回数はそれぞれ 32,000 と 4,000 とした．

モデル Encoder–Decoder，Decoder-only モデルは共に Transformer ベースのモデルを用いる．Encoder–Decoder モデルをベースラインとして Decoder-only モデルとの比較を行う．Encoder–Decoder モデルは，Vaswani ら [2] を参考に，各層を 6 層にした．Decoder-only モデルは，層数を 12 層に設定した．また Decoder の入力と出力の単語分散表現を共有した．損失の計算では $\lambda = 1$ とする．詳細なハイパーパラメータは付録 A に示す．実装は共に fairseq [5)] [17] を利用した．

自動評価 自動評価においては，Perplexity，BLEU，COMET [6)] [9]，WMD [10] を用いた．Perplexity は XGLM-1.7B [7)] [18] を用いており，流暢性を測る目的で計測した．BLEU は sacreBLEU [8)] [19] を用いて計測した．COMET は複数言語で学習された単語分散表現，WMD は目的言語に対応する単

1) https://github.com/moses-smt/mosesdecoder
2) https://taku910.github.io/mecab/
3) https://github.com/rsennrich/subword-nmt
4) https://github.com/google/sentencepiece
5) https://github.com/facebookresearch/fairseq
6) https://github.com/Unbabel/COMET
7) https://huggingface.co/facebook/xglm-1.7B
8) https://github.com/mjpost/sacrebleu

表 1: 各モデルの翻訳の自動評価

	BLEU (↑)		Perplexity (↓)		COMET (↑)		WMD (↓)	
Model	en-de	en-ja	en-de	en-ja	en-de	en-ja	en-de	en-ja
Encoder–Decoder	**25.99**	**42.57**	122.26	94.373	**0.4332**	**0.4277**	**0.6970**	**0.3137**
Decoder-only	23.86	40.31	**114.70**	**88.573**	0.4091	0.4259	0.7254	0.3338

表 2: en-ja 対における pair-wise の流暢性と妥当性の人手評価．win は Decoder-only の方が良く，tie は同程度，lose は Decoder-only の方が悪いことを示す．

	fluency			adequacy		
	win	tie	lose	win	tie	lose
annotator 1	**25**	59	16	**22**	63	15
annotator 2	**26**	51	23	**28**	51	21
annotator 3	18	63	**19**	21	57	**22**
annotator 4	**30**	50	20	**18**	74	8

表 3: 英日対における翻訳の質の人手評価．U は Under-translation，M は Mis-translation，O は Over-translation を示す．

	Enc-Dec			Dec-only		
	U	M	O	U	M	O
annotator 1	10	22	**0**	**8**	**11**	2
annotator 2	**8**	30	**2**	10	**16**	3
annotator 3	**9**	15	**0**	14	**2**	1
annotator 4	7	13	2	**6**	**5**	**1**

語分散表現 [9] [10] [20] を用いて計算を行い，それぞれ原文，または参照文と出力文との意味的な妥当性を評価する目的で計測した．COMET のモデルは wmt20-comet-qa-da を用いた．

英日モデルの人手評価　日本語話者 4 名による，en-ja モデルの人手評価を行った．ASPEC のテストデータからランダムに 100 文をサンプリングし，pair-wise による流暢性と妥当性の評価，翻訳の質に関する評価を行なった．翻訳の質の評価として，Under-translation（不十分な翻訳），Mis-translation（誤訳），Over-translation（過翻訳）の 3 つの側面に関して，人手で計測した．なお評価の際には参考のため参照訳を併記し，モデル名は伏せた．

4 実験結果

自動評価　表 1 に各モデルの翻訳の自動評価結果を示す．Perplexity は Decoder-only が良く，流暢性という観点では Decoder-only が Encoder–Decoder を上回った．一方で BLEU や COMET，WMD などの自動評価指標では Encoder–Decoder が良い．

英日モデルの人手評価　表 2 に英日対における pair-wise による流暢性と妥当性の人手評価を示す．流暢性に関してはアノテータ 1，2，4 は Decoder-only の方が良いと評価している．これは Perplexity の自動評価と一致する結果である．妥当性に関してもアノテータ 1，2，4 は Decoder-only の方が良いと評価している．これは COMET や WMD の自動評価とは相反する結果である．

表 3 に en-ja 対における翻訳の質の人手評価を示す．Under-translation と Over-translation については各モデル間で大きな差異がみられなかった．しかし Mis-translation については Deocder-only の方が Encoder–Decoder より大幅に少ない．この結果から見ても，人手評価では自動評価と反対に Decoder-only の方が妥当性が良いと評価されることがわかった．

自動・人手評価の考察　ここでは自動評価手法として COMET を使い，自動評価と人手評価での妥当性の不一致を分析する．翻訳の質に問題がある例で不一致が起きやすいことがわかった．表 4 に Encoder–Decoder は Mis-translation しているが，Decoder-only は正確に訳せている例を示す．Encoder–Decoder の出力では「放電」という単語が存在し，「スクラバする」といった誤った訳が出力されている．本来は「discharges」が「放出する」と訳されるはずであるが，「discharges」を名詞として捉えてしまったため「放電」と訳されたと考えられる．一方で Decoder-only の出力は参照訳に近く流暢かつ妥当である．COMET による自動評価では Encoder–Decoder は 0.9152，Decoder-only は 0.9055 であり Encoder–Decoder が良い．人手評価では 4 人全員が Decoder-only のほうが妥当性は高く不一致が起こっている．同様の例が Under-translation の時にも見られた（付録 B）．これより COMET では Mis-translation や Under-translation といった翻訳の質の問題に対応しきれていないと考えられる．

人手評価において Decoder-only は妥当性が良く，Mis-translation が Encoder–Decoder と比べ減ることが

9) https://fasttext.cc/docs/en/pretrained-vectors.html
10) http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

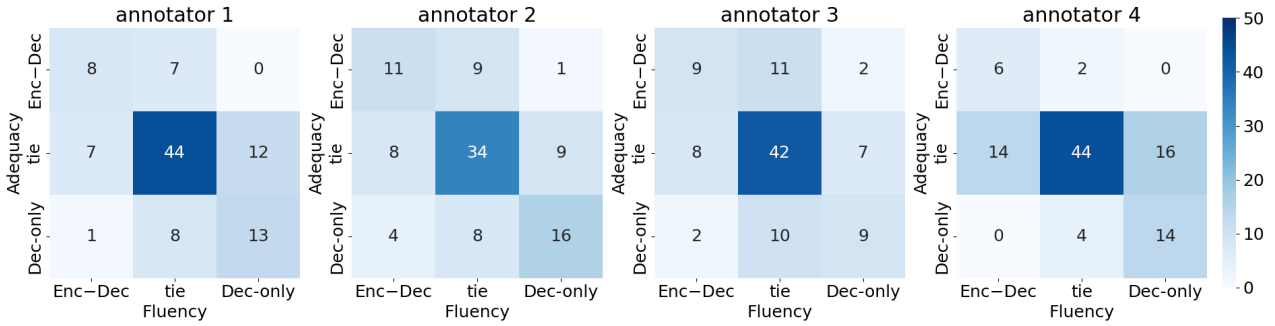

図 2: 各アノテータごとの流暢性と妥当性の関係

表 4: Encoder–Decoder で Mis-translation があり，Decoder-only で Mis-translation が見られなかった例

source	The epitaxial system **scrubber** in the last stage **discharges** the exhaust in the atmosphere.
reference	最終段のエピ系スクラバーが排気を大気中に放出する。
Encoder–Decoder	エピタキシャル系スクラバは、最終段階の**放電**で大気中の排気を**スクラバする**。
Decoder-only	最終段階のエピタキシャル系スクラッバは大気中の排気を排出する。

表 5: 英日対における推論速度の比較

	tokens	token/s	time
Encoder–Decoder	70,574	5537.08	12.7s
Decoder-only	135,539	2936.39	**46.2s**

わかった．多言語モデルの学習に関する既存研究 [21] では，パラメータのシェアが言語横断的なモデルを学習する上で重要であることが示されている．Decoder-only は原言語，目的言語間でパラメータをシェアしており，Transformer より言語横断的な内部表現を学習していると考えられ，これが妥当性の改善につながったと考察する．

流暢性と妥当性の人手評価の分布を調査した．図 2 に各アノテータごとの流暢性と妥当性の関係を示す．横が流暢性を，縦が妥当性の評価結果を示している．どのアノテータも Encoder–Decoder と Decoder-only の流暢性と妥当性を同程度と判断しているものが最も多い．一方，Encoder–Decoder では流暢性が良いが Decoder-only では妥当性が高い，Encoder–Decoder では妥当性が良いが Decoder-only では流暢性が高いといった極端に性質が分かれる例は少なかった．これは評価の際に原文と参照訳を併記したためと考えられる．既存研究 [22] では原文と翻訳結果，参照訳と翻訳結果による人手評価よりも，原文と参照訳と翻訳結果のすべての情報を持った人手評価の方が良いと示されている．

推論速度 Decoder-only モデルの問題点として推論速度が挙げられる．表 5 に英日対における推論速度の比較結果を示す．推論時に使用した GPU は NVIDIA A6000 である．ASPEC のテストデータ 1,812 文を推論に用いた．1 秒あたりの処理するトークン数が Encoder–Decoder では 5537.08，2936.39 であり，推論速度は Encoder–Decoder と比べて 4 倍程度遅くなった．原因としては，原言語文と目的言語文をタグで結合させているためトークン数が 2 倍になっていること，causal に処理する部分が 2 倍になっていることで 1 トークンあたりの処理時間が 2 倍になっていることが挙げられる．この点において Decoder-only は Encoder–Decoder よりも推論速度が遅く，改善が必要である．

5 おわりに

本論文では Encoder–Decoder と Decoder-only を比較し Decoder の性質を調査した．流暢性は自動評価と人手評価とともに Decoder-only が高く，妥当性においても人手評価では Decoder-only の方が良い．これは自動評価とは相反する結果であり翻訳の質を自動評価では考慮しきれていないと考えられる．Decoder-only で妥当性が上がった理由としては原言語と目的言語間でのパラメータシェアが挙げられる．また人手評価において，流暢性と妥当性との関連を示唆する結果となった．一方で Decoder-only の問題点として推論速度の低下がある．

今後の方針としては Decoder-only モデルに基づいた新たな機械翻訳モデルの開発，機械翻訳タスク以外での Decoder-only モデルの再検討，翻訳の質を考慮した自動評価手法の検討，人手評価の流暢性と妥当性との関連の分析が挙げられる．

謝辞

人手評価を行なった東京都立大学の榎本大晟さん，佐藤郁子さん，中島京太郎さんに感謝いたします．

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, **NeurIPS**, Vol. 27. Curran Associates, Inc., 2014.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **NeurIPS**, Vol. 30. Curran Associates, Inc., 2017.

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **NeurIPS**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[6] Shuo Wang, Zhaopeng Tu, Zhixing Tan, Wenxuan Wang, Maosong Sun, and Yang Liu. Language models are good translators. **arXiv preprint arXiv:2106.13627**, Vol. abs/2106.13627, , 2021.

[7] Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney. Is encoder-decoder redundant for neural machine translation? In **AACL**, pp. 562–574, Online only, November 2022. Association for Computational Linguistics.

[8] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In **AMTA**, pp. 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[9] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **EMNLP**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[10] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In **ICML**, pp. 957–966. PMLR, 2015.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[12] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In **WMT**, pp. 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics.

[13] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In **LREC**, pp. 2204–2208, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **ACL**, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **ACL**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[16] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **EMNLP**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[17] 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 **NAACL**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[18] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models. **EMNLP**, 2022.

[19] Matt Post. A call for clarity in reporting BLEU scores. In **WMT**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[20] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. **TACL**, Vol. 5, pp. 135–146, 2017.

[21] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In **ACL**, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics.

[22] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. **TACL**, Vol. 9, pp. 1460–1474, 2021.

A　ハイパーパラメータ

表 6: Encoder–Decoder と Decoder-only のハイパーパラメータの相違点

	Encoder–Decoder	Decoder-only
Encoder	6	0
Decoder	6	12
Task	Translation	Language Modeling
Encoder embedding dim	512	0
Dncoder embedding dim	512	768
Encoder fnn embedding dim	2048	0
Dncoder fnn embedding dim	2048	3072
Encoder attention heads	8	0
Decoder attention heads	8	12
attention drop	0	0.1
activation function	relu	gelu
share decoder input output embedding	×	○

B　Under-translation での妥当性の不一致

表 7 に Under-translation で妥当性の不一致が起きた例を示す．COMET による自動評価では Encoder–Decoder は 0.1101，Decoder-only は 0.1342 であり Decoder-only が良い．人手評価では 4 人全員が Encoder–Decoder が良いと判断しており，Decoder-only の翻訳を Under-translation と判断している．「ゴムを用いた」という部分が Decoder-only ではないため妥当性が低いと判断された．

表 7: Under-translation で妥当性の不一致が起きた例

source	An experiment to form a branch pipe on a part of the main pipe of an aluminum alloy pipe **with rubber** was conducted.
reference	**ゴムを用いて**アルミニウム合金管の胴部の一部に枝管を成形する実験を行った。
Encoder–Decoder	**ゴムを用いた**アルミニウム合金管の主管の一部に分枝管を形成する実験を行った。
Decoder-only	アルミ合金管の主管の一部に分岐管を形成する実験を行った。