文脈文アノテーションによる ドキュメント機械翻訳の精度向上に関する研究

安本 玄樹　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{yasumoto.genki.ye1, sudoh, s-nakamura}@is.naist.jp

1　はじめに

ニューラル機械翻訳の台頭により，機械翻訳の精度は高まってきており，文単位の翻訳においては人間とほとんど同等のレベルであるという報告もある [1]．しかしながら，ドキュメント単位での翻訳においては人間翻訳が上回っていると報告されている [2]．これは前後文の文脈や文章全体の話題を考慮した翻訳が，現在のニューラル機械翻訳では困難であることを意味する．以上の背景から，文脈を利用してニューラル機械翻訳の精度を高める研究が近年盛んに行われており，周辺文の情報を活用するモデル [3, 4] やドキュメント全体を参照して文脈情報を利用するモデル [5, 6] などが提案されている．

周辺文の情報を利用する場合，その中に翻訳に必要な情報が含まれていれば翻訳をより良いものにできると考えられるが，含まれていない場合は文脈を活用することはできず，却ってノイズとなってしまうこともある．一方でドキュメント全体を参照する場合，必要な情報を特定の文や単語に絞ることが難しいという問題が生じる．

このような問題を解消するために，周辺文の文脈を用いるモデルでは，使用する文脈文を選択することで [7]，ドキュメント全体の文脈を用いるモデルでは，アテンションをスパースに当てることで [6]，精度を向上させている．結局のところ，文脈情報を必要なときに必要なだけ利用することができれば，これらの問題は解消されるはずである．

そこで本研究では，文脈文をアノテーションで作成し，必要だと思われる文脈情報を活用できた場合，ドキュメント機械翻訳モデルが十分な精度を出すことができるのかを検証する．翻訳の言語対としては，日英を選択した．これは日英翻訳において，日本語の省略がよく問題になるため，それを補う形でアノテーションを行うことを想定したためである．なお，作成したアノテーションデータは公開予定である[1)]．

2　文脈文アノテーション

各ドキュメントに対して以下の２種類の文脈文アノテーションを行う．

- ドキュメントから文脈として最適であると考えられる文を１文抜き出す場合
- ドキュメントから文脈として最適であると考えられる文をアノテータが書き出す場合

アノテーションを行う際は，原言語のドキュメントのみを参照し，目的言語のドキュメントは参照しないこととする．アノテーションに関する作業は全て筆者が行った．文脈文アノテーションの詳細は次に示す通りである．

2.1　文脈文を１文抜き出す場合

ドキュメントが与えられたとき，それぞれの文を翻訳するのに必要だと思われる文脈をドキュメントの中の他の文から選択して文脈文とする．文脈を必要としないと判断した場合は，空文を表す “_blank” トークンを文脈文の代わりとする．このアノテーションを施した例を表 1 に示す．

2.2　文脈文を１文書き出す場合

文脈文を書き出す場合は，ドキュメントの情報から自由に文脈文を作成することができる．このアノテーションを施した例を表 2 に示す．文脈文を 1 文抜き出す場合（表 1）と比較して，１文目の文脈文が異なっていることがわかる．これは１文目を翻訳時に主語の情報が必要になると考え，主語に相当する情報を書き出したことによる．このように，状況に応じて必要な情報を文脈文として書き出す．文脈

1） https://github.com/YasumotoGenki/annotated_context_for_OpenSubtitles_ja-en

表 1　選択した文脈例（一部抜粋）

文番号	選択した文脈文	ドキュメント（原言語の文章）
1	これは聖戦の 始める準備ができたという意味か？	もっと練習用の ダミーが必要になるな
2	もっと練習用の ダミーが必要になるな	私が大量に注文する
3	_blank	これは聖戦の 始める準備ができたという意味か？
4	これは聖戦の 始める準備ができたという意味か？	準備はできてる どこから始めたらいいか分からない

表 2　書き出した文脈例（一部抜粋）

文番号	書き出した文脈文	ドキュメント（原言語の文章）
1	我々には	もっと練習用の ダミーが必要になるな
2	もっと練習用の ダミーが必要になるな	私が大量に注文する
3	_blank	これは聖戦の 始める準備ができたという意味か？
4	これは聖戦の 始める準備ができたという意味か？	準備はできてる どこから始めたらいいか分からない

が不要な場合は，同様に “_blank” トークンを文脈文の代わりとする．

3　実験

3.1　データ

本実験では，OpenSubtitles2018 [8] の日英対訳を用いる．翻訳は日本語から英語へと行う．全てのデータを作品 ID ごとに分けた後，train, dev, test に分割した．分割は dev, test において，それぞれの文数が 50k 以上になるように，作品をランダムにサンプリングし，残ったものを train セットとした．分割後のドキュメント数，文数については表 3 の通りである．

表 3　データセット

OpenSubtitles2018 (ja-en)	train	dev	test
ドキュメント数	2,617	69	70
文数	1,982,514	50,053	51,033

トークナイズには sentencepiece [9] を用い，語彙サイズは日本語，英語ともに 30k とした[2)]．特殊トークンである，“_blank” などは 30k の語彙に含まない．

3.2　モデル

実験に使用するモデルは，Dual Encoder Transformer [10] および CADec [4] とする．

3.2.1　Dual Encoder Transformer

Li ら [10] の提案した Dual Encoder Transformer は 2 種類あり，Decoder の内部で目的言語と文脈文のアテンションを利用するもの（inside-context）と，Decoder の外部で原言語文と文脈文のアテンションをとり，それを Decoder に入力するというもの（outside-context）である．実装は Li ら [10] のものを用い，ハイパーパラメータはデフォルト値を利用した[3)]．

3.2.2　CADec (Context-Aware Decoder)

CADec (Context-Aware Decoder) では，一度文脈文，原言語をそれぞれ Transformer に入力して原言語および目的言語側の embedding を得る．そして CADec にそれらを入力して文脈情報を汲み取るモデルとなっている．著者による公開実装を利用し[4)]，ハイパーパラメータもデフォルトの設定に従った．

3.3　学習

これらのモデルはどちらも Transformer [11] をベースとしており，2 段階で学習される．最初にそれぞれベースとなる Transformer を文脈なしの状態で学習し，その Transformer をベースとして文脈を処理する部分を含めた全体の学習を行う．ベースとなる Transformer の学習は，全ての train データを利用して行った．学習の際，train, dev データそれぞれに “_blank” トークンを 1 組ずつ追加し，“_blank” が “_blank” と対応していることも学習させる．

2 段階目の学習は，文脈文と原言語を入力として行う．Dual Encoder Transformer には，直前の文を文脈文として与え，CADec には，翻訳する文の直前に述べられている文を最大 3 文まで文脈文として与える．ドキュメントの最初の文を翻訳する時のみ，“_blank” トークンを文脈文とした．この 2 段階目の学習にも，全ての train データを用いた．

3.4　文脈文アノテーション

文脈文アノテーションは test セットのみに対して行う．OpenSubtitles の日英対訳の中には，アライメ

2）　日本語の character coverage は 0.9995, 英語の character coverage は 1 とした．

3）　https://github.com/libeineu/Context-Aware

4）　https://github.com/lena-voita/good-translation-wrong-in-context

表 4　実験結果（全ての test データ：全 70 ドキュメント／51033 文）

文脈		なし				周辺文			
モデル		BLEU	BERTScore			BLEU	BERTScore		
			P	R	F1		P	R	F1
Dual Encoder Transformer	base Transformer	15.43	53.77	44.79	49.19	-	-	-	-
	Dual_{inside}	-	-	-	-	15.63	53.76	44.79	49.17
	$\text{Dual}_{outside}$	-	-	-	-	15.57	53.67	44.63	49.05
CADec	base Transformer	15.67	52.28	45.19	48.66	-	-	-	-
	CADec	-	-	-	-	15.70	52.41	45.15	48.70

表 5　実験結果（アノテーションしたドキュメントデータ：全 3 ドキュメント／1785 文）

文脈	周辺文				選択した文				書き出した文			
モデル	BLEU	BERTScore			BLEU	BERTScore			BLEU	BERTScore		
		P	R	F1		P	R	F1		P	R	F1
Dual_{inside}	27.45	64.20	57.39	60.71	27.30	64.17	57.15	60.58	27.31	64.19	57.22	60.62
$\text{Dual}_{outside}$	27.07	63.65	57.03	60.25	26.76	63.81	56.66	60.15	26.77	63.77	56.71	60.15
CADec	26.57	64.12	58.27	61.14	26.57	64.03	58.19	61.05	26.59	64.07	58.23	61.09

ントが十分にとれていないものも多いため，CADec のベース Transformer を用いて一度 test セットの各ドキュメントを翻訳し，BLEU [12] を閾値としてスコア 20 以上のドキュメントのみをアノテーション対象とした．今回の実験では，３つのドキュメントに対して文脈文アノテーションを行った．各ドキュメントの文数は表 6 の通りである．

表 6　アノテーションを施したドキュメント

ドキュメント番号	1	2	3	合計
文数	526	381	878	1,785

3.5　実験結果

BLEU，BERTScore [13] を評価指標として用いる．BLEU を測定する際のトークナイザは，mosestokenizer を利用し[5]，BLEU の実装は multi-bleu.perl を利用した[6]．BERTScore は，著者の公開実装を利用し[7]，fine-tuning は行わず，rescaling をした結果を示す．基本的なモデルの性能を示すため，アノテーションした文脈文を用いずに，test セットを翻訳した結果を表 4 に示す．アノテーションしたドキュメントを用いた場合と直前の文脈を用いた場合の結果を表 5 に示す．BLEU，BERTScore 共に百分率表示とする．

3.6　分析および考察

BLEU や BERTScore で評価した結果については，有意な差は見て取れず，モデルによる明らかな性能の違いは認められなかった．実際の翻訳例を表 7 に示す．

翻訳が改善している例では，文脈文にある “ヤツら” と言う表現をもとに，主語が “They” と正しく翻訳されている．ただしこの翻訳例において，Dual Encoder Transformer の訳出に変化は見られなかった．このように，文脈文は十分だと考えられる場合であっても，上手く訳出されない例は他にも見受けられた．

翻訳に変化が見られない例について述べる．目的言語の “they” にあたる “鹿” という単語は，選択した文脈文，書き出した文脈文共に含まれていたが，翻訳では “I” と，誤って訳出されてしまっている．翻訳に変化を与えることができなかった理由としては，

- 学習時に文脈文を参照して翻訳を行わなければならないケースが割合として少なく，文脈文を翻訳に利用する学習ができていない可能性
- 原言語を参照するよりも，目的言語の言語モデルに強く依存した翻訳になっている可能性

が考えられる．前者に関しては，OpenSubtitles の日英対訳ドキュメント自体がデータとして不十分，もしくはクリーニングが必要であることが示唆される．後者に関しては，train に使用したデータ量が多いほど，翻訳を行う時に原言語の情報を利用することが報告されている [14]．文脈を活かしきれなかったともいえる．

翻訳が悪くなった例では，アノテーション時にドキュメントを誤読しており，書き出した文脈文の “私たち” という表現を受けて “We” と誤って訳出さ

5）https://github.com/luismsgomes/mosestokenizer
6）https://github.com/OpenNMT/OpenNMT-py/blob/master/tools/multi-bleu.perl
7）https://github.com/Tiiiger/bert_score

れてしまっている．OpenSubtitles は字幕データであるが，映像を含まないテキストのみのデータであるため，日本語の曖昧性によって多少の誤読が発生し得る．その影響を受けた例といえる．書き出した文脈文を “私は” と修正したところ，主語は正しく “I” と翻訳された．この他にも，“_blank” を文脈文としたときに，訳抜けが発生してしまった例などがあった．文脈文が翻訳に悪い影響を与えてしまった理由としては，

- アノテーションした文脈文はドキュメント内で使用されている日本語と分布が異なるため，言語分布が異なる入力を与えると翻訳に悪影響を及ぼす可能性

が考えられる．特に “_blank” トークンに関しては，これに当てはまる．

その他の翻訳例は付録とした．翻訳例を見る限り，CADec の方が文脈に即した翻訳を行うことが多かった．これは学習時に CADec は最大 3 文の文脈を用いており，その中で推論に必要なものが含まれていた割合が高いと考えられる．Dual Encoder Transformer は直前の 1 文のみを文脈として学習させたため，その文脈から推論する例が CADec と比較して少なく，学習が不十分だった可能性が残る．学習データに関する，より詳細な分析が必要である．

4 おわりに

本研究では，アノテーションした文脈文を利用して，翻訳結果が改善されるのかを検証した．結果としては，アノテーションした文脈文をテスト時に利用しても，翻訳結果が十分に改善されることはなかった．その理由として学習時の工夫（データクリーニングやデータ量の増大，“_blank” トークンの扱い方など）が必要とされる可能性が示唆された．

今回のアノテーションでは，目的言語のドキュメントを参照しなかったが，参照しても良いという条件下で，より質の良いアノテーションデータを作成してみること，学習データに対してもアノテーションデータを作成し，モデルの学習が十分に行われることを保証することは今後の課題としたい．

5 謝辞

本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものである．

表 7 文脈文の変化による機械翻訳出力の違い

翻訳が改善している例

原言語文	壁の内側にいる
目的言語文	I think they’re inside the walls.
周辺文 1	大丈夫・・・
周辺文 2	イヤ
周辺文 3	パトリック
選択した文脈文	ヤツらがいる ヤツらよ
書き出した文脈文	ヤツら
Dual_{inside}	Inside the walls.
$\text{Dual}_{outside}$	Inside the walls.
CADec + 周辺文	**He’s** inside the walls.
CADec + 選択した文脈文	**They’re** inside the walls.
CADec + 書き出した文脈文	**They’re** inside the walls.

翻訳に変化が見られない例

原言語文	自分でやったと思います
目的言語文	I think they did it to themselves.
周辺文 1	健康な鹿が溺れただと
周辺文 2	意味が分からん
周辺文 3	たしかに　おかしい でも...
選択した文脈文	健康な鹿が溺れただと
書き出した文脈文	健康な鹿が溺れただと
Dual_{inside}	I think I did it myself.
$\text{Dual}_{outside}$	I think I did it myself.
CADec	I think I did.

翻訳が悪くなった例

原言語文	飲むために来たのよ
目的言語文	I’m here to drink.
周辺文 1	病気なの
周辺文 2	大丈夫？
周辺文 3	今はレナの事 話したくない
選択した文脈文	今はレナの事 話したくない
書き出した文脈文	私たちは
Dual_{inside}	I’m here to drink.
$\text{Dual}_{outside}$	I’m here to drink.
CADec + 周辺文	**I’m** here to drink.
CADec + 選択した文脈文	**I’m** here to drink.
CADec + 書き出した文脈文	**We are** here to drink.

注 1 ：文脈によって出力に変化がある場合のみ，モデル名＋文脈文の種類を記し，出力に変化がなかった場合は，モデル名のみの記載としている
注 2 ：CADec の周辺文は 3 文全て，Dual Encoder Transformer の周辺文は最後の 1 文となる

参考文献

[1]Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation, 2018.

[2]Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[3]Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2947–2954, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4]Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1198–1212, Florence, Italy, July 2019. Association for Computational Linguistics.

[5]Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1275–1284, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[6]Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Selective attention for context-aware 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. 3092–3102, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7]Ryuichiro Kimura, Shohei Iida, Hongyi Cui, Po-Hsuan Hung, Takehito Utsuro, and Masaaki Nagata. Selecting informative context sentence by forced back-translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pp. 162–171, Dublin, Ireland, August 2019. European Association for Machine Translation.

[8]Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[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]Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and changliang li. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3512–3518. Association for Computational Linguistics, July 2020.

[11]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, pp. 5998–6008. Curran Associates, Inc., 2017.

[12]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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[13]Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, Vol. abs/1904.09675, , 2019.

[14]Elena Voita, Rico Sennrich, and Ivan Titov. Analyzing the source and target contributions to predictions in neural machine translation. *arXiv preprint arXiv:2010.10907*, 2020.

付録

表 8　文脈文の変化による機械翻訳出力の違い

翻訳が改善している例

原言語文	体の一部が問い続けてる
目的言語文	I guess part of me kept asking,
周辺文 1	何が？
周辺文 2	分からない
周辺文 3	信じられない事が 起きたんだ
選択した文脈文	_blank
書き出した文脈文	ボクの体の一部が
Dual_{inside}	Part of me keeps asking.
$\text{Dual}_{outside}$	Part of me keeps asking.
CADec + 周辺文	Part of **him** still asks.
CADec + 選択した文脈文	Part of **him** still asks.
CADec + 書き出した文脈文	Part of **me** still asks.

翻訳の一部は改善しているが，一部悪くなった例

原言語文	メールの返事が 来ない
目的言語文	She isn't answering any of my texts.
周辺文 1	24 って誰？
周辺文 2	アリスの元彼
周辺文 3	レナはどうしてる？
選択した文脈文	レナはどうしてる？
書き出した文脈文	彼女から
Dual_{inside}	I didn't get an email.
$\text{Dual}_{outside}$	I didn't get an email.
CADec + 周辺文	**He** never responded to **my** e-mail.
CADec + 選択した文脈文	**He** never responded to **my** e-mail.
CADec + 書き出した文脈文	**She** never responded to **her** e-mail.

翻訳が悪くなった例

原言語文	代表者は SCPD の努力を賞賛しています 病院を閉鎖から救いました とは言え一部の関係者は自警団が 関与してるかもしれないと言っています
目的言語文	Representatives praise the efforts of the SCPD in saving the hospital from shutting down, though some sources say the Vigilante may have been involved.
周辺文 1	すなわち 英雄になる事は決してないわ
周辺文 2	この街が安全な限り
周辺文 3	どうでもいい 長い追跡の後 警察は チェンナ・ナ・ウェイを逮捕しました 地元の中国の トライアドの高位の一員です
選択した文脈文	_blank
書き出した文脈文	_blank
Dual_{inside}	The delegates are applauding the efforts of the SCPD, but some of the vigilantes are saying that they may be involved.
$\text{Dual}_{outside}$ + 周辺文	The delegates are commending the efforts of the SCPD, but **they've managed to save the hospital from shutting down** some of the vigilantes may be involved.
$\text{Dual}_{outside}$ + 選択した文脈文	The delegates are applauding the efforts of the SCPD, but they're saying that some of the vigilantes may be involved.
$\text{Dual}_{outside}$ + 書き出した文脈文	The delegates are applauding the efforts of the SCPD, but they're saying that some of the vigilantes may be involved.
CADec	The representatives of the SCPD have managed to save the hospital from closed, although some of the vigilantes may be involved.

注 1：文脈によって出力に変化がある場合のみ，モデル名＋文脈文の種類を記し，
出力に変化がなかった場合は，モデル名のみの記載としている
注 2：CADec の周辺文は 3 文全て，Dual Encoder Transformer の周辺文は最後の 1 文となる