対話翻訳における長距離文脈の利用

今村 賢治　　隅田 英一郎
国立研究開発法人 情報通信研究機構
{kenji.imamura,eiichiro.sumita}@nict.go.jp

1 はじめに

機械翻訳では，長らく 1 文単位の翻訳方式が主流であったが，近年のニューラル機械翻訳では，文外の文脈を利用した翻訳が研究されてきている (Wang et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Zhang et al., 2018) .

たとえば，Bawden et al. (2018) は，英仏翻訳を対象に，原言語文脈，目的言語文脈も利用した翻訳実験を行い，翻訳品質や翻訳の一貫性評価などを行った．その結果，目的言語文脈を使うことによって，翻訳品質が若干向上することを確認している．Voita et al. (2018) は，原言語文脈を利用するエンコーダーを提案し，指示代名詞の照応解析に有効であると報告している（対象は英露翻訳）．しかし，照応や一貫性に影響する現象は言語によって異なるため，日本語を対象とした翻訳では，結果が異なる可能性がある．

そこで本稿では，文脈の必要性が比較的高い対話文を対象に，文脈を利用した翻訳実験を行う．本稿で着目する項目は，以下のとおりである．

- 英日，日英翻訳を対象とする．
- 文脈の必要性が比較的高いと考えられる対話文を対象とする[1]．ただし今回は，シナリオライターが作文した疑似対話を使用する．
- Bawden et al. (2018) の実験では，直前の 1 文のみを文脈として扱っていたが，参照範囲を広くしたときの影響を調査する．

2 文脈利用翻訳

2.1 疑似対話データ

今回用いるコーパスは，GCP コーパス (Imamura and Sumita, 2018) で，日本語，英語，中国語などの 10 言語を対象にした多言語コーパスである．日本を訪れた外国人が，現地の日本人と対話するシチュエーションを想定し，シナリオライターが疑似対話を作成した．そして，日本語文を他の言語に翻訳することで多言語化している．GCP コーパスの例を表 1 に示す．疑似対話であるため，言いよどみ，言い直しは含まれていない．

GCP コーパスの各発話（本稿では文と同義）には，発話者情報が付与されているが，今回は，発話者が 2 名（日本人と外国人 1 名ずつ）の対話だけを対象とした．今回使用した疑似対話の統計量は，表 2 のとおりである．

2.2 モデル及びデータの加工

今回の文脈利用法は，Bawden et al. (2018) と同様に，Tiedemann and Scherrer (2017) の方法を使用する．これは，複数の文を結合したデータを用いることにより，文脈を考慮した翻訳を実現するものである．翻訳器自体は，1 文翻訳用のものをそのまま利用する．彼らは，Sequence-to-Sequence モデルを使用したが，今回の実験は，翻訳品質が高い Transformer (Vaswani et al., 2017) モデルを使用する．

図 1 は，本稿の文脈利用法の概要である．

- 入出力ともに，各文の先頭に発話者タグ sp を付与する．発話者タグは，<Foreigner>（外国人），<Japanese>（日本人）の 2 種類である．対話の場合，一人称，二人称主語がゼロ代名詞化されることが多いため，発話者タグにより外界照応解決の補助をさせることを狙っている．
- 原言語に関しては，発話者タグ付きの文を複数接続して，翻訳器に入力する．たとえば原言語文脈長を 2 文とした場合，当該文 S_i に対する原言語文脈は直前 2 文となり、入力は $\langle sp_{i-2}, S_{i-2}, sp_{i-1}, S_{i-1}, sp_i, S_i \rangle$ となる．発話者タグが文区切り記号を兼ねている．
- 目的言語も原言語と同様に，発話者タグ付きの文を複数接続して学習を行う．テスト時には，出力の

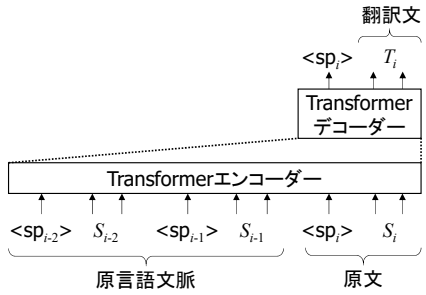
[1] 日本語の新聞記事と雑談対話を比較した文献 (今村他, 2015) によると，雑談対話は，参照先が別の文に存在するゼロ代名詞と，参照先が外界に存在する外界照応の割合が，新聞記事に比べて高かった．

表 1: 医療分野における疑似対話例

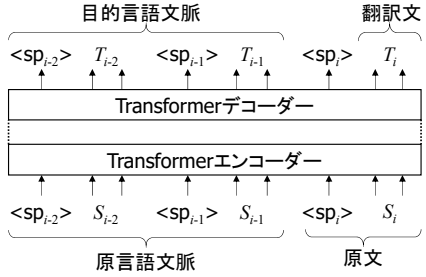
大分類	中分類	小分類	発話者種別	文（発話）
医療	病気・けがになったら	緊急時の対応をする	外国人	このあたりに病院はありませんか？
			日本人	あのコンビニを右に曲がったところに内科の医院があります。
			日本人	具合が悪いんですか？
			外国人	めまいがします。
			日本人	日本語は得意ではないんですか？
			外国人	得意ではありません。
			日本人	あそこの医院だと英語は通じないかもしれません。
			日本人	隣の駅前の総合病院なら英語のできるスタッフがいます。

表 2: 疑似対話データの統計量

	訓練	開発	テスト
対話数	157,808	160	159
文数	2,003,902	2,010	2,003
文数/対話	12.7	12.6	12.6
英単語数/文	13.2	13.6	13.1



(a) X-to-1 (X=3 の場合)



(b) X-to-X (X=3 の場合)

図 1: 入出力データとモデルの構成

発話者タグを走査し，最終発話者タグ以降の単語列を翻訳文として評価する．したがって，テスト時の目的言語文脈は，翻訳器によって生成されたものになる．

今回は，原言語文脈を使用するが目的言語文脈を使用しない方法（X-to-1．X が当該文を含む文数を表す）と，原言語文脈と目的言語文脈の両方を使用する方法 (X-to-X) を実験する．X-to-1 の実験により原言語文脈の影響を調査し，X-to-X と X-to-1 を比較することによって目的言語文脈の影響を調査する．

2.3 他の文脈モデルとの関連

文脈を考慮したニューラル機械翻訳モデルが提案されてきている．

Wang et al. (2017) は Sequence-to-Sequence を基にしたモデルを提案した．このモデルでは，原言語文脈を文単位に隠れ状態にエンコードし，さらに文の系列をエンコードする．そして，原文のエンコード結果と混合して翻訳を行う．

Voita et al. (2018) は，Transformer をベースに，エンコーダーの最終レイヤーに文脈を混合する multi-head アテンション機構を追加している．Zhang et al. (2018) も同様な方式であるが，原言語文脈の隠れ状態をエンコーダー，デコーダー双方に入力している点が Voita et al. (2018) と異なっている．これらの方式は，いずれも原言語文脈のみを考慮した翻訳モデルである．

今回用いた方式は，モデルを変更していないにも関わらず，原言語文脈，目的言語文脈ともに扱えるのが特徴である．原言語文脈は，Transformer エンコーダーの self-attention で原文と混合され，デコーダーの context attention でも翻訳文と混合される．目的言語文脈はデコーダーの self-attention で翻訳文と混合される．つまり，本稿で用いた方式は，機能的には最新の文脈モデルを包含している．ただし，文脈も含めて翻訳を行うため，翻訳時間とメモリ使用量の観点では明らかに非効率である[2]．

3 実験

3.1 その他の実験条件

本節では，英日翻訳，日英翻訳を対象に実験を行う．2 節で述べた以外の実験条件は，以下のとおりである．

[2]無限長の文脈を扱う実験も行ってみたが，メモリ不足で学習できなかった．

表 3: Marian NMT の設定

項目	設定
モデル	6 層 Transformer，ヘッド数 8，モデル幅 512 次元，FFN 幅 2,048 次元，ドロップアウト 0.1，label smoothing 0.1.
学習	Adam 最適化．学習率 0.0004，ウォームアップを 5 エポックしたのちに 10 エポック以降指数的に減衰．開発セットのパープレキシティが最小となった時点で停止．ミニバッチサイズ約 250．最大データ長 1,000 単語.
テスト	ビーム幅 6，開発セットの BLEU スコアが最高となった時点の長さ正規化定数を使用.

データ　表 2 のコーパスは，日本語は MeCab（辞書は IPA-dic 2.7.0）で，英語は Moses ツールキットの tokenizer を使って単語分割した．さらに，バイトペア符号化法 (Sennrich et al., 2016) で，各言語約 1 万 6 千のサブワードに分割した.

翻訳システム　翻訳器は，Marian NMT (Junczys-Dowmunt et al., 2018) を使用した．モデルはエンコーダー，デコーダーともに 6 層の Transformer モデルである．詳細な設定を表 3 に示す．なお，文脈長を長くするにしたがい入出力データも長くなるが，翻訳器が許容するデータ長を 1,000 単語とし，すべてのデータが学習できるように設定した.

比較対象　発話者タグがない 1 文翻訳をベースラインとし，原言語および目的言語の文脈長を 0 から 4 まで変化させた翻訳を比較した.

評価　評価は，開発セットのパープレキシティ(PPL)と，テストセットの BLEU スコアで行った．開発セット PPL は，モデルの良さを表し，BLEU スコアは翻訳品質を表す．BLEU スコアの検定は MultEval ツール[3] で行い，危険率を 5%とした ($p < 0.05$).

3.2　結果

表 4 と図 2 は実験結果である．両者は同一データである．表は詳細な分析用、図は全体傾向を把握するため、両方を示した。なお、ここでの文脈長 (X) は，当該文を含む.

まず，ベースラインと，1-to-1 翻訳を比較すると，英日翻訳，日英翻訳ともに BLEU スコアは低下している（有意差なし）．両者の差異は発話者タグの有無であるので，このデータでは発話者タグは，翻訳品質には影響していない.

[3] https://github.com/jhclark/multeval

次に，X-to-1 について着目する．英日翻訳では，$X > 2$ の BLEU スコアはすべてベースラインより有意に高かったが，日英翻訳では，$X = 3$ を除き，ベースラインより BLEU スコアは低かった（ただし，$X = 3$ も含めて有意差はない）．この結果からは，原言語文脈は英日翻訳には有効だったが，日英翻訳には効果がなかった．実際，開発セットパープレキシティをみても，英日は文脈長を長くするにしたがい，パープレキシティが低下するが，日英については，ほぼ変化がない.

次に，X-to-X と X-to-1 を比較する．英日翻訳，日英翻訳ともに，有意差はほとんどなかったが，X-to-X の BLEU スコアが X-to-1 のスコアを上回る傾向がみられた（英日，日英ともに，X-to-X が 3 データポイント上回ったのに対して，X-to-1 は 1 データポイントのみ）．開発セットパープレキシティでは，すべてのデータポイントについて，X-to-X のパープレキシティが低く，目的言語文脈は，英日，日英翻訳ともに若干の効果があった．また，文脈長は，$X \leq 5$ の範囲では長い方が翻訳品質が高い傾向があった。

3.3　議論

Bawden et al. (2018) は，英仏翻訳を対象に，文脈翻訳実験を行った。彼らは、原言語側には指示代名詞が含まれており、その照応解析のため、原言語文脈は有効と分析した。また、フランス語は男性名詞、女性名詞の一貫性が必要なので、目的言語文脈も必要であるとした。

今回の実験では、翻訳品質（とモデルの品質）しか評価していないため、どの文脈に含まれるどの言語現象が影響したのか、不明である。しかし、原言語、目的言語にかかわらず，英語側の文脈の効果があったのは，フランス語と同様に、単複の一致など，文法的一貫性が必要だったからではないかと推測される。日本語については、原言語文脈の効果がなく，目的言語文脈は有効な傾向があったが、この理由は不明である。いずれにしても、今後詳細な分析が必要と考える。

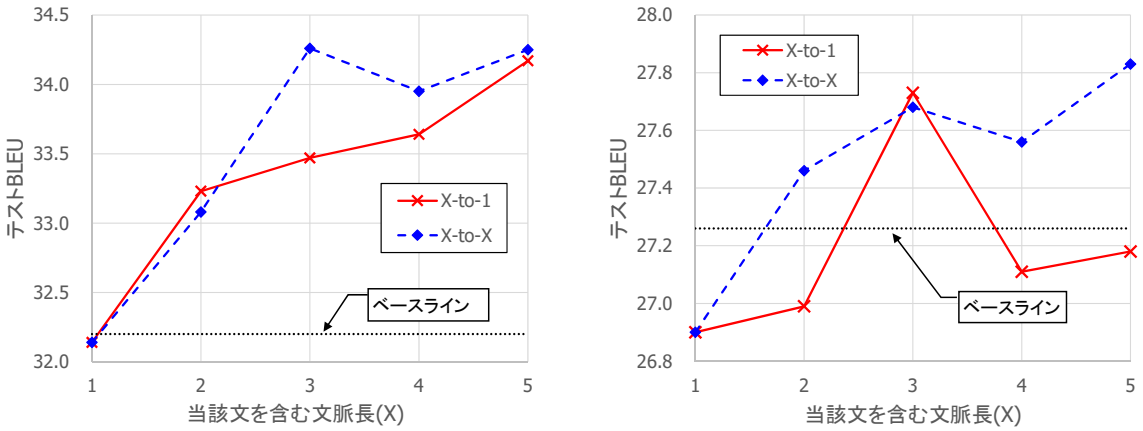
4　まとめ

本稿では、英日・日英の疑似対話データを対象に、文脈を含んだ翻訳実験を行った。その結果、言語によって文脈の効果は異なるが、条件によっては原言語文脈も目的言語文脈も翻訳品質向上に影響する傾向が得られた。

表 4: 実験結果

† はベースラインと有意差がある（$p < 0.05$）ことを表す．$*$ は，同じ文脈長の X-to-1 と X-to-X と比べ，有意差があることを示す．

	英日				日英			
当該文を含む文脈長 (X)	開発セット PPL↓		テストセット BLEU↑		開発セット PPL↓		テストセット BLEU↑	
	X-to-1	X-to-X	X-to-1	X-to-X	X-to-1	X-to-X	X-to-1	X-to-X
ベースライン	3.36		32.20		4.01		27.26	
1	3.17		32.14		3.74		26.90	
2	3.11	3.08	33.23 †	33.08 †	3.77	3.68	26.99	27.46
3	3.07	3.03	33.47 †	34.26 †*	3.69	3.61	27.73	27.68
4	3.07	3.02	33.64 †	33.95 †	3.67	3.61	27.11	27.56
5	3.04	3.02	34.17 †	34.25 †	3.72	3.61	27.18	27.83



(a) 英日翻訳の BLEU スコア　　(b) 日英翻訳の BLEU スコア

図 2: 文脈長毎の BLEU スコア

今回は主に翻訳品質のみの測定を行ったが、BLEU スコア以外の分析は必須である。また，今回はモデルを変更せずに、データ加工で文脈を扱うようにしたが、文脈を毎回翻訳しているので、効率は悪い。目的言語文脈のみの影響を調査するには、モデルの改造も必要である。今後は、詳細な分析とモデル改造の検討を行いたいと考えている。

謝辞

本研究は総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進-多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の一環として行われました．

参考文献

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proc. of NAACL-HLT 2018 (Volume 1: Long Papers)*, pages 1304–1313.

Kenji Imamura and Eiichiro Sumita. 2018. Multilingual parallel corpus for global communication plan. In *Proc. of LREC 2018*, pages 3453–3458.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proc. of ACL 2018, System Demonstrations*, pages 116–121.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL-2016 (Volume 1: Long Papers)*, pages 1715–1725.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proc. of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proc. of ACL-2018 (Volume 1: Long Papers)*, pages 1264–1274.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proc. of EMNLP-2017*, pages 2826–2831.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proc. of EMNLP-2018*, pages 533–542.

今村 賢治, 東中 竜一郎, 泉 朋子. 2015. 対話解析のためのゼロ代名詞照応解析付き述語項構造解析. **自然言語処理**, 22(1):3–26, 3 月.