同時機械翻訳のための文脈を考慮したセグメントコーパス

中林 明子
東京大学 大学院総合文化研究科
akinkbys@phiz.c.u-tokyo.ac.jp

加藤 恒昭
東京大学 大学院総合文化研究科
kato@boz.c.u-tokyo.ac.jp

1 はじめに

発話を音声認識によってテキスト化し、文の完了を待たずに翻訳を開始する、同時機械翻訳を検討している．特に講演などにおいては一文が長くなる傾向があり、文の完了を待たずに翻訳を開始することが期待される．しかし一般的な翻訳システムでは一文が入力されることを前提としているため、翻訳の単位が小さくなれば、適切な翻訳は困難となる．このような同時機械翻訳のタスクにおいては、発話からの遅延を抑えると同時に、翻訳精度を維持することが求められる．

同時通訳者は、同時性を保ちつつ適切な通訳を行うため、Salami Technique と呼ばれる手法を用いている．一文を複数のセグメントに分割し、セグメント単位で前から順に処理する．その際、再構成、簡素化、省略することで自然な訳出を実現する [1]．

同時機械翻訳においても、入力をセグメントに分割し、セグメント単位に翻訳する方式が研究の初期から検討されている．しかしこの手法は、ニューラル機械翻訳の枠組みにおいて以下の課題がある．
(1) セグメント単位の翻訳は、それまでの文脈が反映されず、発話の中で適切なものとならない．(2) 一般的な文単位のパラレルコーパス（以下、文コーパス）から、セグメント単位での適切な翻訳を学習することは限界がある．あわせて、同時機械翻訳における大きな課題として、(3) 同時通訳コーパス資源が限られているため、学習に用いられる文コーパスは翻訳コーパスであり、同時機械翻訳にとって最適な語順となっていないことがある．

本論文では、これらの問題に対応するために、すでに現れたセグメントを文脈とみなして、これを参照してより適切な通訳結果を出力するようなセグメント単位のコーパス（以下、セグメントコーパス）を作成し、それを利用して、文コーパスで学習した翻訳モデルのチューニングを行うことを提案する．

以下、関連研究を紹介した後、セグメントコーパスとその作成について述べ、実験結果を報告する．実験の結果、単純なセグメント単位の翻訳に比べて、BLUE スコアが大きく向上した．また翻訳結果より、前方のセグメントを参照した翻訳が可能であることを確認した．

2 関連研究

1 節で述べた (1) の問題に対して、Gu, et al. [2] は強化学習により、翻訳システムの READ/WRITE を制御することで対応した．その後、(2) に対応するため、文ではなく prefix 単位で学習する Wait-k [3] が提案されたが、適切なセグメント分割は考慮されず、常に一定の遅延で出力するため、特に語順が異なる言語間では自然な翻訳を出力できていない [4]．遅延の policy をより柔軟に学習する方法 [5] [6] や、文コーパスを分割してセグメントコーパスを作成する方法 [7] も提案されているが、これらは翻訳コーパスに基づいて行われる．

本提案は、入力文をセグメントに分割し、セグメント単位で翻訳を行うという基本的な枠組みに従うが、(1)(2) の問題に加えて (3) の問題を視野に入れ、同時通訳者の用いる手法を参考に同時翻訳結果として望ましいセグメントコーパスを作成する．翻訳コーパスから同時機械翻訳向けのコーパスを作成する提案として [8] [9] があるが、セグメントを扱うニューラル機械翻訳の枠組みの中での活用は述べられていない．コーパスの作成において、文単位の翻訳コーパスで学習したベースラインの翻訳モデル（以下、ベースモデル）を利用するという提案も独自のものである．

3 文脈を考慮したセグメントコーパス

以下の対からなるセグメントコーパスを構築した．入力言語側は、一文をセグメントに分割し、すでに現れたセグメントを文脈とみなして文脈記号で接続する．出力言語側は、分割されたセグメントに同時通訳者の用いる手法を参考にして必要な情報を付加し、これをベースモデルで翻訳した結果を用いる．前方のセグメントを参照して、必要な情報を追加するよう学習することを期待している．

3.1 セグメント

本論文では、一文を翻訳する単位となるセグメントに分割し、セグメントごとに翻訳を行うことを想定する．セグメントをどこで分割すべきかについては、様々な研究がなされてきた [10] [11] が、本論文では、セグメント長の最大値を 6 単語とし、その範囲で、最も大きい構成要素の境界でセグメントを分割した．本論文では、このようなセグメント分割が同時通訳者が行っているものであるかは議論しないが、構成要素の単位をセグメントとすることは妥当な仮定であると考えている．また、このようなセグメント分割には、漸進的パージングが必要となるが、本論文ではこれが可能であると仮定している．

3.2 同時通訳者の手法

同時通訳者は、特に語順が大きく異なるような言語対では、セグメント単位の翻訳をより自然な出力とするため、様々な工夫を行う．よく見られる手法の一つとして、以下のような繰り返しがある．なお、“/”はセグメントの分割位置を示す．

名詞句（NP）の繰り返し：

- 原文：So being a Russian girl / of fairly strong being able to drink liquor very strongly ,
- 通訳：ロシア人の**女性**ですけれども / 非常にお酒が強い**人**なので

動詞の繰り返し：

- 原文：thus producing a sound / from the chanter.
- 通訳：そして今度また音を**出し**ます/ チャンターから**出す**んですね

（同時通訳データベース（SIDB）[12] より）

同時機械翻訳においても、第一の例の二つ目のセグメントを訳出する際は、前方の名詞句（NP）を繰り返すことでより自然な翻訳となる．これを可能とするために前方のセグメントを参照し、名詞句の繰り返しを学習できるよう、コーパスを作成する．

3.3 文脈を考慮したセグメントコーパスの構築

以下の手順で文脈を考慮したセグメントコーパスを構築した．

1. 入力言語の Training データをセグメントに分割する．
   - 入力言語：they explained the technology / of the dna analysis .
2. 入力言語のセグメントが前置詞“of”で始まる場合、それが修飾する前方の名詞句（NP）を付加する．
   - 入力言語：they explained the technology / the technology of the dna analysis .
3. Step1-2 のセグメントをベースモデルで翻訳し、対訳ペアを作成する．
   - 出力言語：技術 について 解説 し た / dna 解析 の 技術
4. 入力言語のセグメントに、一文を一つの単位としてすでに現れたセグメントを付加し、間に文脈符号（_BREAK_）を付与する．
   - 入力言語：they explained the technology / they explained the technology _BREAK_ of the dna analysis .
   - 出力言語：技術 について 解説 し た / dna 解析 の 技術
5. Step3 の結果のうち、翻訳結果に同じ単語が 3 回以上現れるペアは誤訳である可能性が高いため、コーパスから削除する．
   - 入力言語：at present ,
   - 出力言語：現在 , 現在 で は , 現在 で ある

Step2 では 3.2 節に示した同時通訳者の名詞句の繰り返しを取り入れている．Step4 は、機械翻訳において文脈情報を付与する手法の一つ [13] で ある．

4 実験

文コーパスで学習したベースモデルに対して、文脈を考慮したセグメントコーパスを用いてチューニングを行った．同時通訳者の手法に見られるように、前方のセグメントの情報が用いた出力が可能で

表 1　データセット

		文数	セグメント数
Original	Training	1,543,762	
	Dev	995	3,553
	Test	1,029	3,674
チューニング	Training	518,516	1,870,349

表 2　実験結果

モデル	文	セグメント
ベースモデル	33.1	13.8
セグメント	24.0	20.4
セグメント+NP	23.3	19.4
セグメント+NP+文脈	10.1	21.7

あるかを確認した．

4.1　モデル

PyTorch ベースの OpenNMT[1] を使用して英日翻訳モデルを構築した．6 層の Transformer のモデルで、表現の次元数は 512、中間層の次元数は 2,048、ヘッド数は 8 とした．ドロップアウトは 0.1、バッチサイズは 4,096 とした．英語と日本語の入力には BPE [14] を適用し、英語と日本語の語彙は共有した．語彙数は 40,543 であった．最適化アルゴリズム には Adam を使用した．次節のコーパスを用いてベースモデルを構築した．このベースモデルを 3.3 節のとおり作成した、文脈を考慮したセグメントコーパスでチューニングした．

4.2　データ

データセットとして、ASPEC コーパス [15] を利用した．データセットの概要は表 1 のとおりである．

チューニングに使用するセグメントコーパスのデータは、ベースモデルで使用した Training データと同等のボリュームとなるよう、ベースモデルで使用した Training データより抽出した．

4.3　実験結果

定量評価には ASPEC コーパスの評価セットを利用し、指標として BLEU[2] を使用した．セグメント単位の翻訳では、それぞれの翻訳結果を一文ごとに結合し、参照翻訳と比較した．その結果を表 2 に示す．

1 節で論じているように、ASPEC コーパスの参照翻訳は必ずしも本提案が目指す結果ではない．その位置付けからすると、本評価は、ベースモデルをセグメント単位で翻訳した場合と比較して、句や節などの比較的狭いスパンでより適切な結果が得られているかに着目していることになる．より大きな構造のレベルで同時通訳結果としてより適切となっているか否かの評価は今後の課題である．

モデルに対して、文単位の評価セットとセグメント単位の評価セットで評価を行った．**ベースモデル**は、セグメントコーパスでチューニングする前の翻訳モデルを指す．**セグメント**は 3.3 節の Step1+3+5 を実施したコーパスでチューニングしたモデル、**セグメント+NP** は Step1+2+3+5 を実施したコーパスでチューニングしたモデル、**セグメント+NP+文脈**は、本論文の提案である Step1+2+3+4+5 を実施したコーパスでチューニングしたモデルを指す．このモデルでは、セグメント単位の評価セットにも文脈を付与している．

ベースモデルはセグメント単位の評価で大きく劣る．単純なセグメントコーパスによるチューニングでも、セグメント単位の評価は大きく改善するが、今回提案するコーパスによってさらなる向上が見られている．

5　考察

実際の翻訳例を表 3 に示す．翻訳例 1 において、セグメント“of 100 hours”は「100 時間の連続走行」と翻訳されており、同時通訳者の手法で見られるように、それが修飾する前方の名詞句を参照し、これを繰り返して翻訳することができている．翻訳例 2 も同様に、前置詞“of”で始まるセグメントは、前方の名詞句を参照し、これを繰り返して翻訳している．

本論文では 3.3 節の Step2 において名詞句の付与だけを行ったが、3.2 節の動詞の反復を含めるなど、これをより一般化することで、より受け手にとって負担のない翻訳を出力できるようになることが期待される．例えば翻訳例 2 において、セグメント“to explain the basic concept（本来は in order to explain the basic concept）”の翻訳と直前のセグメントの翻訳を見ると、セグメント間の関係が見えづらいが、前方のセグメントの動詞を参照してこれを繰り返すことで、より自然な翻訳を出力できるようになる．

1）　https://opennmt.net
2）　https://github.com/mjpost/sacrebleu

表 3 翻訳例

翻訳例 1

入力	the 1.0 t coil / succeeded in continuous running / of 100 hours .
参照訳	1 . 0 t コイル は 100 時間 の **連続 運転** に 成功 し た
ベースモデル	1 . 0 t コイル **連続 走行** に 成功 し た 100 時間
チューニング後	1 . 0 t コイル, **連続 走行** に 成功 し た, 100 時間 の **連続 走行**

翻訳例 2

入力	the problem of reflection and refraction / of cylindrical wave in plane boundary / was taken up in order / to explain the basic concept .
参照訳	基本 概念 を 説明 する ため , 平面 境界 における 円筒 波 の **反射 ・ 屈折** の 問題 を 取り上げ た
ベースモデル	**反射 と 屈折** の 問題, 平面 境界 における 円筒 波, 次 に 取り上げ た, 基本 概念 を 説明 し た
チューニング後	**反射 と 屈折** の 問題, 平面 境界 における円筒波 の **反射 と 屈折**, 順 に 取り上げ た, 基本 概念 を 説明 し た

6 まとめ

本論文では、入力文をセグメントに分割し、セグメント単位で翻訳を行うという基本的な枠組みのもと、すでに現れたセグメントを文脈とみなして、これを参照してより適切な通訳結果を出力するようなセグメント単位コーパスを作成し、これを利用して翻訳モデルのチューニングを行うことで、同時機械翻訳の精度向上を図った．チューニングを行うことで、ベースモデルにおけるセグメント単位の翻訳に比べて、BLUE スコアが大きく向上した．また翻訳結果より、前方のセグメントを参照した翻訳が行われていることを確認した．

今後の取り組みとして、セグメント間の接続方法を検討したい．現在はセグメントごとに翻訳された結果を接続しているが、例えば翻訳例 1 の最初と 2 つ目のセグメントの翻訳結果において、2 つ目のセグメントの翻訳時に前方のセグメントの翻訳結果を考慮し「1.0t コイル〈は〉連続走行に成功した」のように〈は〉を挿入することができれば、より自然な翻訳を得ることができると考える．

参考文献

[1] Roderick Jones. *Conference interpreting explained.* Routledge., 1998.

[2] Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1053–1062, Valencia, Spain, April 2017.

[3] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3025–3036, Florence, Italy, July 2019.

[4] Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. Dutongchuan: Context-aware translation model for simultaneous interpreting. *CoRR*, Vol. abs/1907.12984, , 2019.

[5] Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1313–1323, Florence, Italy, July 2019.

[6] Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. Simultaneous translation with flexible policy via restricted imitation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5816–5822, Florence, Italy, July 2019.

[7] Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. Low-latency neural speech translation. In *Proceedings of Interspeech, 2018*, pp. 1293–1297, 2018.

[8] He He, Alvin Grissom II, John Morgan, Jordan Boyd-Graber, and Hal Daumé III. Syntax-based rewriting for simultaneous machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 55–64, Lisbon, Portugal, September 2015.

[9] 二又航介, 須藤克仁, 中村哲. 英日同時通訳システムのための疑似同時通訳コーパス 自動生成手法の提案. 言語処理学会第 26 回年次大会, 2020.

[10] Hideki Kashioka, Takehiko Maruyama, and Hideki Tanaka. Building a parallel corpus for monologue with clause alignment. In *Proceedings of the Ninth Machine*

Translation Summit, pp. 216–223, 2003.

[11] Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 551–556, Baltimore, Maryland, June 2014.

[12] Shigeki Matsubara, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. Bilingual spoken monologue corpus for simultaneous machine interpretation research. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May 2002.

[13] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 82–92, Copenhagen, Denmark, September 2017.

[14] 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 (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016.

[15] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2204–2208, Portorož, Slovenia, May 2016.