教師なし機械翻訳に基づく話し言葉翻訳へのドメイン適応の検討

福田 りょう　　須藤 克仁　　中村 哲
奈良先端科学技術大学院大学
{fukuda.ryo.fo3, sudoh, s-nakamura}@is.naist.jp

1 はじめに

我々が会話で日常的に用いる, いわゆる「話し言葉」の機械翻訳は難しく, その理由は学習に必要な対訳データの少なさにある. 利用可能な言語資源の多くは「書き言葉」であり, 話し言葉の書き起こしデータは非常に少ないと言える. このような低資源下における機械翻訳の学習手法としてドメイン適応が知られている. 講義翻訳の従来研究では, 論文抄録対訳コーパスを用いて翻訳モデルを学習した. これは, 話し言葉の翻訳学習に書き言葉対訳を利用したドメイン適応学習といえる. この手法において, 発話スタイルへの適応が課題の一つである. 講義の書き起こしは言い淀みや口語表現など話し言葉特有の表現を含む. それに対し論文抄録は書き言葉であり, 文体が大きく異なるためドメイン適応が難しい.

そこで本研究では, 高品質な話し言葉の機械翻訳の学習を目的とし, 書き言葉を擬似的な話し言葉に変換することによる効果的なドメイン適応手法の検討を行なった. NAIST 授業アーカイブ [1] の日英翻訳において, 論文抄録対訳コーパス ASPEC[2] を日本語話し言葉コーパス CSJ[3] 調に変換した擬似話し言葉によるドメイン適応学習は, 擬似話し言葉を用いなかった場合と比較して最大+1.54 ポイント BLEU が向上した.

2 提案手法

提案手法は, 書き言葉から話し言葉への言語内翻訳器 (2.1) と, 話し言葉の言語間翻訳器 (2.2) で構成される (図 1).

2.1 書き言葉から話し言葉への翻訳

書き言葉から話し言葉への文体変換を学習するにあたり, はじめにスタイル変換タスクの 2 手法 ([4][5]) を実験・検討した. しかし, 変換結果の多くが非文であり有望な変換はほぼ見られなかった. スタイル変換のタスクには, ポジティブからネガティブへの文の感情変換や, 話し手の属性変換などがある. これらのタスクは, 文構造を維持して内容語のみを対話に置き換える, といった簡単な変換で達成される場合が多い. そのため, 書き言葉から話し言葉変換の際に求められる語順変化や間投詞の挿入といった, 複雑な言い換えの学習が困難であると考えられた.

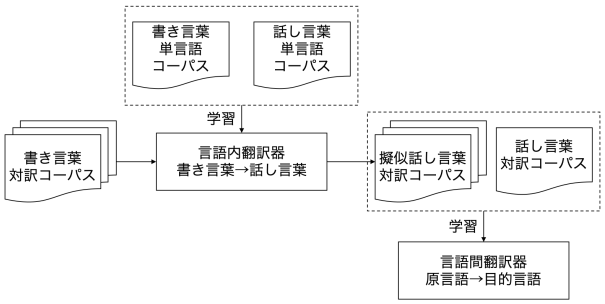

図 1: 提案方式のフロー

そこで我々は, 書き言葉から話し言葉への変換を翻訳タスクと捉え, 既存の機械翻訳手法の適用を検討した. 翻訳は, 語順変化のような複雑な言い換えを含むため, 書き言葉から話し言葉への変換も十分学習が可能であると期待した. 通常, ニューラル機械翻訳 (NMT) [6][7] の学習は対訳データを必要とするが, 書き言葉と話し言葉の対訳データは入手困難である. Lample ら [8] は, 対訳関係にない 2 言語の単言語コーパスを用いて翻訳を学習する教師なしニューラル機械翻訳 (UNMT) を提案した. 本稿では, 対訳関係にない書き言葉データと話し言葉データを用いて UNMT を学習することにより, 書き言葉から話し言葉への翻訳器を作成した.

2.2 話し言葉の機械翻訳

低資源な話し言葉対訳データのみから, 高品質な翻訳器を作成することは困難である. またドメイン適応学習においても, ドメイン外データとして書き言葉を用いた場合, ドメイン内データである話し言葉との差異が大きく, 効果的な学習は望めないことが先行研究により示されている. そこで本稿では, 擬似話し言葉を用いた話し言葉翻訳のドメイン適応学習を提案する. 2.1 で作成した擬似話し言葉データをドメイン外データとして利用することで, 効果的なドメイン適応学習が可能であると考えた.

機械翻訳において, ドメイン適応の手法は”Data centric”と”Model centric”に大別できる [9]. Data centric は学習データに着目した手法であり, 既存の機械翻訳モデルへの適用が容易である. Model centric は, モデルの構造や学習方法などに着目した手法である. 今回, data centric な手法である Multi-domain 学習 [10] と, model centric な手法である Fine-tuning[11] の 2 つに

表 1: 教師なし機械翻訳による擬似話し言葉文生成例

書き言葉 (ASPEC)	擬似話し言葉 (CSJ-like ASPEC)
代替フロン中には可燃性のものがあるので注意が必要である。	代替フロン中には、可燃性のものがあるので注意が必要であるということが言えます。
3）消化管内 pH 変化	三番に消化管内 pH 変化です。
超伝導トンネル接合（ＳＴＪ）を用いた標題検出器を開発した。	で超伝導トンネル接合ですね、ＳＴＪを用いた標題検出器を開発しました。
ゴーグル機能として使用するだけでなく，シースルー機能を持たせた。	でゴーグル機能として使用するだけでなくて、ルーシース機能を持たせます。

表 2: 書き言葉から話し言葉への翻訳器の学習データ

単言語データ		文数
書き言葉	ASPEC-JE (日本語)	1,003,602
話し言葉	CSJ	134,477
	CEJC	128,668
	NAIST 授業アーカイブ	22,251

よるドメイン適応学習を検討した．Multi-domain 学習は, ドメイン外データとドメイン内データを混合して学習に用いる手法である．文頭にドメインラベルを付加する手法やドメイン間のデータ数を揃える手法などが提案されているが, 今回は書き言葉ドメインと話し言葉ドメインのデータを連結するだけの最も単純な手法を使用した．Fine-tuning は, 大規模ドメイン外データでモデルを事前学習後, ドメイン内コーパスで追加学習を行う手法である．

3 実験

3.1 実験: 書き言葉から話し言葉への翻訳

3.1.1 実験設定

コーパス 本節で用いた学習データのサイズを表 2 に示す．書き言葉データは, ASPEC-JE の日本語文を使用した．話し言葉データは, CSJ, 日本語日常会話コーパス CEJC[12], NAIST 授業アーカイブを使用した．

翻訳システム Lample らによる UNMT の実装[1]を利用した．共有エンコーダ, デコーダは 3 層 Transformer で構成し, 埋め込みベクトルおよび隠れベクトルの次元数は 512 とした．最適化には Adam を使用した．学習は, ASPEC 評価データ 1,790 文の折り返し BLEU スコアの停滞が 10 エポック連続するまで行ない, 最も高いスコアを得たモデルをテストに使用した．学習データに対し BPE によるサブワード化を行なった．サブワード語彙は共有し, 語彙サイズは 16,000 とした．

3.1.2 実験結果

ASPEC テストデータ 1,812 文に対する, 折り返し翻訳の BLEU と perplexity を表 3 に示す．話し言葉データに CSJ を用いたモデルが最も高い評価を得たため, 以降の実験にはこのモデルを用いる．表 1 は擬似話し言葉の生成例である．文体の変化やフィラーの挿入, 段落番号や括弧の除去などが見られ, 話し言葉らしさが獲得できたと言える．一方で, 時制の変化や, 語順の入れ替えによる単語の崩れなど, 望ましくない変換も見られた．

[1]https://github.com/facebookresearch/UnsupervisedMT

表 3: 各話し言葉データを用いて学習した翻訳モデルの BLEU と perplexity

話し言葉データ	折り返し翻訳	
	BLEU	perplexity
CSJ	**80.98**	**1.617**
CEJC	15.14	15.98
NAIST 授業アーカイブ	17.02	20.54
CSJ+CEJC+授業アーカイブ	14.54	17.46

表 4: 言語モデルの授業アーカイブに対する perplexity

学習データ	perplexity	未知語
ASPEC-JE (日本語)	1210.7	47,757
CSJ	107.4	35,542
CSJ-like ASPEC	360.7	37,561
NAIST 授業アーカイブ	29.6	0

続いて, 擬似話し言葉を話し言葉翻訳器の学習データとして用いることの妥当性を調べるために, 言語モデルを構築し話し言葉の perplexity を測定した [13]. perplexity が低いほど, 言語モデルの学習データが話し言葉らしいと考えられる．表 4 は, 単言語データを用いて構築した 3-gram 言語モデルの, NAIST 授業アーカイブ 22,251 文への perplexity と未知語の数である．書き言葉コーパス ASPEC-JE の日本語文を話し言葉調に変換 (CSJ-like ASPEC) することにより, 大幅に perplexity や未知語が減少し, 話し言葉らしさを高めることができたと言える．

3.2 実験: 話し言葉の機械翻訳

3.2.1 実験設定

コーパス 本節で用いた学習データのサイズを表 6 に示す．ドメイン内データは, 話し言葉コーパスである NAIST 授業アーカイブの日英対訳データを使用

表 5: 日英機械翻訳モデルの書き言葉と話し言葉に対する BLEU

適応手法	ASPEC-JE	授業アーカイブ
ASPEC-JE (適応なしベースライン)	**27.52**	6.16
CSJ-like ASPEC	23.86	5.58
ASPEC-JE & 授業アーカイブ (Multi-domain 学習ベースライン)	17.13	6.61
CSJ-like ASPEC & 授業アーカイブ	24.28	8.15
ASPEC-JE + 授業アーカイブ (Fine-tuning ベースライン)	23.99	12.71
CSJ-like ASPEC + 授業アーカイブ	20.93	12.81
ASPEC-JE & CSJ-like ASPEC + 授業アーカイブ	24.18	12.55
ASPEC-JE + CSJ-like ASPEC + 授業アーカイブ	23.19	**12.82**

表 6: 話し言葉翻訳器の学習データ

対訳データ		対訳数
ドメイン内	NAIST 授業アーカイブ	7,031
ドメイン外	ASPEC-JE	1,003,602
	CSJ-like ASPEC	1,003,602

した．ドメイン外データは, 書き言葉コーパスである ASPEC-JE, および ASPEC-JE の日本語側を擬似話し言葉に変換して作成した擬似話し言葉コーパス (CSJ-like ASPEC) を使用した．

翻訳システム　オープンソースの NMT システムである OpenNMT-py[2]を使用した．エンコーダ, デコーダは Transformer で構成し, 埋め込みベクトル次元を 512, 隠れベクトル次元を 2048 とした．最適化には Adam を使用した．学習データに対し BPE によるサブワード化を行なった．サブワード語彙は日英で共有し, 語彙サイズは 16,000 とした．

3.2.2　実験結果

ASPEC-JE と授業アーカイブのテストデータ各 1,812 文に対する, 各手法の BLEU スコアを表 5 に示す．表において, & (アンド) 記号は Multi-domain 学習を, + (プラス) 記号は Fine-tuning を意味している．例えば, "ASPEC-JE & CSJ-like ASPEC + 授業アーカイブ"は「ASPEC と CSJ-like ASPEC の混合データで事前学習後, 授業アーカイブによる追加学習」である．

ドメイン適応なしの学習では, CSJ-like ASPEC を用いることで, ASPEC-JE テストデータに対する翻訳精度が-3.66 ポイント, 授業アーカイブに対しては-0.58 ポイントと低下した．Multi-domain 学習では, ASPEC-JE は+7.15 ポイント, 授業アーカイブは+1.54 ポイントと向上した．ドメイン外データとして ASPEC-JE を用いたベースラインでは, 書き言葉と話し言葉の差異が大きく効果的な学習が困難であったのに対し, ドメイン外データとして擬似話し言葉を使うことで, 2 つのドメイン間距離が近くなりドメイン適応が容易になったと考えられる．Fine-tuning を行うことで, ドメイン外データに ASPEC-JE を用いた場合と CSJ-like ASPEC を用いた場合共に, ドメイン適応学習無しのモデルと比較して, 授業アーカイブに対する翻訳精度が大きく向上した．しかし ASPEC-JE に対する精度は約 3 ポイント低下しており, 対象ドメインに過適合し汎化性能が下がった結果であると考えられる．また CSJ-like ASPEC を用いることで, 授業アーカイブのスコアが+0.1 ポイントと向上したが, 有意な差があるとは言えない．ドメイン外データとして ASPEC-JE と CSJ-like ASPEC 両方を用い, Multi-domain 学習と Fine-tuning の組み合わせ学習や 2 段階に渡る Fine-tuning などを検討したが, Fine-tuning ベースラインを有意に上回る結果は見られなかった．

表 7 は, 同一の入力文に対する各手法の出力文の一例である．例 1 の入力文は「…という話」という口語表現を含む．ドメイン適応を行わない場合, "becomes the talk"や"becomes a dialogue"のように直接的な訳出をし, 出力文に違和感が残る．Multi-domain 学習や Fine-tuning を行うことで, より滑らかな文を生成できるようになった．例 2 では, 文頭の「あ」という間投詞を訳出するかしないかという部分にドメイン適応の効果が表れている．このように, ドメイン適応を行う場合と行わない場合で訳出が大きく異なる例が多く見られた．しかし, ドメイン外データとして, 書き言葉である ASPEC-JE を使用するか擬似話し言葉である CSJ-like ASPEC を使用するかという違いによる, 明確な変化は見て取れなかった．

4　おわりに

本研究では, 話し言葉機械翻訳の精度向上を目的とし, 教師なし機械翻訳による書き言葉から話し言葉への変換 (2.1), 擬似話し言葉を用いた話し言葉翻訳のドメイン適応学習 (2.2) を行なった．2.1 では, 文体の変化やフィラーの挿入といった話し言葉らしさを与える変換モデルを作成し, 作成した擬似話し言葉コーパスが話し言葉翻訳への利用に有望であることを示した．2.2 では, Multi-domain 学習においては擬似話し言葉

[2]https://github.com/OpenNMT/OpenNMT-py

表 7: 各手法による話し言葉の日英機械翻訳例

例 1

Input	同じように、こういう場合でも、そういうグラフというのはつくれますよという話になると。
Reference	Similarly, in such a case as well, we can develop such a graph.
ASPEC-JE	Similarly, it **becomes the talk** of that this graph is more and more.
CSJ-like ASPEC	In the same way, it **becomes a dialogue** which is a graph, a song.
ASPEC-JE & 授業アーカイブ	Similarly, we can create a graph even in such a case.
CSJ-like ASPEC & 授業アーカイブ	Similarly, even in the case of such a case, the graph is connected.
ASPEC-JE + 授業アーカイブ	In the same way, even in such a case, we can create such a graph.
CSJ-like ASPEC + 授業アーカイブ	In the same way, even in such a case, we can make such a graph.

例 2

Input	あ、違うな。
Reference	Oops, that's not it.
ASPEC-JE	They are different.
CSJ-like ASPEC	It is different.
ASPEC-JE & 授業アーカイブ	**Oh**, it is different.
CSJ-like ASPEC & 授業アーカイブ	**Oh**, it is different.
ASPEC-JE + 授業アーカイブ	**Oh**, it is different.
CSJ-like ASPEC + 授業アーカイブ	**Ah**, it is different.

が有効であることを示せたが, Fine-tuning による学習においては有意差が見られなかった．フィラーの有無や文体の違いは Fine-tuning で十分適応できるため, 擬似話し言葉の有用性が失われてしまったと考えられる．Fine-tuning のみでは対応が難しい話し言葉の特徴を探すことや, 表面的な変換だけではなく語順や文長が大きく変わるような変換を行うことなどが今後の課題である．

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．

参考文献

[1] 須藤克仁, 林輝昭, 西村優汰, 中村哲. 授業アーカイブの翻訳字幕自動作成システムの試作. 情報処理学会研究報告自然言語処理 (NL), Vol. 2019-NL-240, No. 15, pp. 1–4.

[2] 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, 2016.

[3] K. MAEKAWA. Corpus of spontaneous japanese : its design and evaluation. *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7–12, 2003.

[4] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 866–876, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[5] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[8] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[9] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1304–1319, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[10] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 385–391, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[12] 小磯花絵, 伝康晴. 『日本語日常会話コーパス』データ公開方針：法的・倫理的な観点からの検討を踏まえて. 国立国語研究所論集, No. 15, pp. 75–89, jul 2018.

[13] 小橋優矢, 西村良太, 北岡教英. Sequence-to-sequence model を用いた話し言葉音声認識用言語モデルのための書き言葉から話し言葉へのテキスト変換. 日本音響学会 2019 年秋季研究発表会 (ASJ), 滋賀, sep 2019.