2022年度AAMT/Japio特許翻訳研究会
報 告 書

機械翻訳及び機械翻訳評価に関する研究
並びに
第7回特許情報シンポジウム開催報告

2023 年 3 月

一般財団法人　日本特許情報機構

目次

1.はじめに・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・1

辻井　潤一　AAMT/Japio 特許翻訳研究会委員長

産業技術総合研究所 人工知能研究センター センター長

2.年間報告書

2.1 ニューラル機械翻訳の強化学習における様々な評価指標報酬の調査・・・・・・・・4

中谷　祐貴　愛媛大学

梶原　智之　愛媛大学

二宮　　崇　愛媛大学

2.2 文書単位の翻訳品質改善に向けた取り組み・・・・・・・・・・・・・・・・・・・・・・13

鈴木　　潤　東北大学

2.3 非自己回帰モデルによるポストエディットを用いた

高速な語彙制約付き機械翻訳・・・・・・・・・・・・・・・・・・・・・・・・17

小町　　守　東京都立大学

3.サーベイ論文報告：特許機械翻訳の課題解決に向けたサーベイ論文の執筆・・・・・・28

須藤　克仁　奈良先端科学技術大学院大学

4.シンポジウム開催報告：第 7 回特許情報シンポジウム・・・・・・・・・・・・・・・・・30

綱川　隆司　静岡大学

AAMT/Japio 特許翻訳研究会委員名簿

（敬称略・順不同）

委員長	辻井　潤一	国立研究開発法人産業技術総合研究所 人工知能研究センター センター長 東京大学大学院 名誉教授
副委員長	須藤　克仁	奈良先端科学技術大学院大学 情報科学研究科 准教授
	綱川　隆司	静岡大学学術院 情報学領域 講師
委員	荒瀬　由紀	大阪大学大学院 情報科学研究科 准教授
	今村　賢治	国立研究開発法人 情報通信研究機構 ユニバーサルコミュニケーション研究所 先進的音声翻訳研究開発推進センター 先進的翻訳技術研究室
	越前谷　博	北海学園大学大学院 教授
	江原　暉将	元・山梨英和大学 教授
	岡崎　直観	東京工業大学 教授
	菊井玄一郎	国立研究開発法人 科学技術振興機構
	黒橋　禎夫	京都大学大学院 教授
	後藤　功雄	NHK 放送技術研究所
	小町　　守	東京都立大学 教授
	鈴木　　潤	東北大学 データ駆動科学・AI 教育研究センター 大学院情報科学研究科 教授
	田村　晃裕	同志社大学 准教授
	中澤　敏明	東京大学大学院 特任研究員
	二宮　　崇	愛媛大学 教授
	渡辺　太郎	奈良先端科学技術大学院 教授
オブザーバー	岡　　俊行	有限会社アジア産業 研究開発部長
	園尾　　聡	東芝デジタルソリューションズ株式会社
	王　　向莉	株式会社ディープランゲージ

オブザーバー（(一般）日本特許情報機構）：

大塩　只明	特許情報研究所 調査研究部 研究企画課
木下　　聡	特許情報研究所 調査研究部
小林　　明	専務理事/特許情報研究所 所長
塩澤　如正	特許情報研究所 調査研究部 研究企画課 課長
西出　隆二	特許情報研究所 調査研究部 部長
塙　　金治	特許情報研究所 研究管理部 研究管理課
舩戸さやか	特許情報研究所 調査研究部 研究企画課 主任
三橋　朋晴	特許情報研究所 研究管理部 研究管理課 課長

事務局	株式会社インターグループ

2022 年度 AAMT/Japio 特許翻訳研究会・活動履歴

2022 年 4 月 26 日
第 1 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2022 年 6 月 27 日
第 2 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2022 年 9 月 12 日
第 3 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2022 年 11 月 17 日
第 4 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2022 年 12 月 22 日
第 5 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2023 年 2 月 20 日
第 7 回特許情報シンポジウム
（於オンライン開催）

2023 年 3 月 6 日
第 6 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

1.　はじめに

AAMT/Japio 特許翻訳研究会委員長
産業技術総合研究所人工知能研究センター　センター長
辻井　潤一

ここ数年間、自然言語の処理技術は、人工知能研究の中心的な位置を占めてきている。この傾向は、ChatGPT の出現が社会問題化することでさらに加速している。

2015 年には、世界的に著名な碁の専門棋士を破った AlphaGO が人工知能のブームを引き起こした。ChatGPT の衝撃は、AlphaGO のそれを上回る社会的な関心事となっている。AlphaGO が、碁という、いわば閉じた、限定されたゲームの世界での成果であったのに対して、ChatGPT は、膨大な知識を内在化することで、一見、非常に多様な分野に関して人間以上の知識を持っているかのような対話を行う。分野限定でない、汎用の知能体が出現したという印象を与えている。

人間と区別できない会話能力をもつシステムが完成したら、人工知能が実現したとみなそうという、前世紀中葉に A.チューリングが提唱したテストに合格するシステムが完成したことになる。A.チューリングは、20 世紀末か 21 世紀初頭には、このテストにパスする人工知能が出現すると予言したが、この予言が 20 年の遅れで実現したことになる。

ChatGPT は、間違った内容のテキストを際限なく生成するという Hallucination(幻覚)現象という厄介な欠点を持つが、生成されるテキストは、きわめて自然で人間が書いたテキストと区別することはむつかしい。この ChatGPT は、ニューラル翻訳システムの研究から生み出されたトランスフォーマーの技術を基盤にしていが、この Hallucination 現象は機械翻訳においても観測され、問題とされてきた。一見、自然な翻訳文が生成されるが、生成された翻訳文が、原文の内容とは違っていること、これが人間のポストエディットの作業をより困難にする、という問題である。

翻訳は、原文の持つ情報をできるだけ正確に翻訳文に反映する作業であるが、「原文の持つ情報を正確に」翻訳文に反映するとはどういうことかは、翻訳論の基本的な問題となっている。自然さや翻訳文のもつ聞き手への効果を重視すると、翻訳者の解釈が混入した意訳となり、原文への忠実度は劣化する。ニューラル翻訳が文単位からテキスト単位へと広がることで、より自然な翻訳が ChatGPT と同じような Hallucination 現象を引き起こすことにもなる。翻訳の正しさをどのように保障し、誤訳を防ぐかは大きな課題となろう。

また、GPT のような事前学習モデル（流行の用語では、基盤モデル）の有効性や正当性は、それを訓練するのにつかわれた訓練データに強く依存する。基盤モデルの研究の大きな部分が、優れた訓練データとは何か、それをどのように獲得するかに向けられている。これは、ニューラル翻訳のための訓練データをどのように収集するのか、また、文単位の

翻訳からテキスト翻訳に向かう場合の訓練データとはどのようなものであるべきか、という問題に向き合うことになる。

ChatGPT の基礎となったトランスフォーマーの技術は、言語処理の基盤技術だけでなく、画像認識での Vision Transformer、生命科学の研究を大きく変革させたタンパク質の 3 次元構造の予測を行う AlphaFold2、計算量が爆発する巡回セールスマン問題のようなプランニング研究にも適用され、大きな技術変革を引き起こしている。

このトランスフォーマー技術の生みの親である機械翻訳においても、次の新たな技術課題を解いていく新たな技術が生み出されるわくわくする時代を迎えたことになる。

本報告書が、この新たな時代を切り開く一助となることを願っている。

2. 年間報告書

2.1 ニューラル機械翻訳の強化学習における様々な評価指標報酬の調査

愛媛大学　中谷　祐貴
梶原　智之
二宮　崇

2.1.1　はじめに

深層学習に基づく系列変換モデル（Vaswani et al. 2017）は流暢な文生成が可能であり、機械翻訳やテキスト平易化など、多くのテキスト生成タスクで成功を収めている。テキスト生成に関する先行研究の多くは、出力文と参照文の間のクロスエントロピー損失を用いて、トークン単位での最尤推定によって系列変換モデルを訓練している。クロスエントロピー損失の微分可能性は、教師あり学習の枠組みでの勾配計算に有効ではあるが、柔軟性に欠ける。つまり、意味的に妥当な出力文が、参照文との表層的な不一致によって、不当に低評価を受ける場合がある。

このような Loss-Evaluation Mismatch 問題（Ranzato et al. 2016; Wiseman & Rush 2016）は、強化学習（Williams 1992）による評価指標への直接的な最適化によって対処できる。強化学習の報酬には、系列変換モデルのパラメタで微分不可能な関数を用いることができるため、単語 N-gram 単位の評価指標である BLEU（Papineni et al. 2002）や文単位の評価指標である BLEURT（Sellam et al. 2020）など、任意の評価指標を採用できる。強化学習を用いることで、機械翻訳（Ranzato et al. 2016; Hashimoto & Tsuruoka 2019; Yasui et al. 2019）やテキスト平易化（Zhang & Lapata 2017; Nakamachi et al. 2020）などの深層学習に基づくテキスト生成の性能改善が報告されている。

機械翻訳においては、多くの先行研究（Ranzato et al. 2016; Wu et al. 2018; Hashimoto & Tsuruoka 2019; Kiegeland & Kreutzer 2021）が強化学習の報酬計算に BLEU を用いているが、BLEU は人手評価との相関が充分に高いわけではない。機械翻訳の自動評価タスク（Bojar et al. 2017）では、表層マッチングに基づく chrF（Popović 2017）や BERT（Devlin et al. 2019）に基づく手法（Shimanaka et al. 2019; Zhang et al. 2020; Sellam et al. 2020）など、BLEU よりも高い人手評価との相関を持つ評価指標が提案されている。そのため、これらの評価指標を用いた報酬計算によって、機械翻訳の強化学習を更に改善できる可能性がある。

本稿は、国際ワークショップ「The 9th Workshop on Asian Translation（WAT 2022）」において我々が提案したニューラル機械翻訳強化学習のための様々な評価指標報酬に関する研究（Nakatani et al. 2022）について報告する。本研究では、図 2.1.1 に示す強化学習の枠組みで Transformer ベースの機械翻訳モデル（Vaswani et al. 2017）を訓練する。ただし、機械翻訳の強化学習では、数万トークンからなる語彙を扱うために、行動空間が非常に大きい。そのため、先行研究（Ranzato et al. 2016; Hashimoto & Tsuruoka 2019）と同様に、クロスエントロピー損失最小化の事前訓練を経た機械翻訳モデルに対しての再訓練として強化学習を適用する。そして、報酬計算と性能評価の両方において複数の評価指標を検討し、機械翻訳の強化学習に適した報酬関数について調査する。

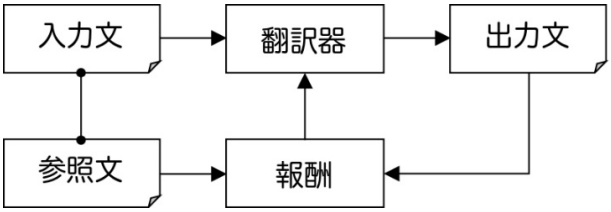


図 2.1.1: 強化学習に基づく機械翻訳

IWSLT-2014 の独英翻訳タスク（Cettolo et al. 2014）における評価実験の結果、BLEU を報酬とする強化学習では BLEU および chrF の表層マッチングに基づく評価指標しか改善できないことが明らかになった。一方で、文間の意味的類似度推定（STS: Semantic Textual Similarity）タスク（Cer et al. 2017）において訓練した BERT (Yasui et al. 2019) や BLEURT (Sellam et al. 2020) など、BERT に基づく一部の評価指標を報酬とする機械学習では、様々な評価指標が改善された。

2.1.2 機械翻訳の強化学習

本研究では、様々な評価指標を報酬とする深層強化学習によって、事前訓練済みの機械翻訳モデルを再訓練する。まず 2.1.2.1 節で機械翻訳モデルの事前訓練について説明し、次に 2.1.2.2 節で強化学習による再訓練について述べ、最後に 2.1.2.3 節で強化学習に用いる報酬としての機械翻訳の評価指標を概説する。

2.1.2.1 事前訓練

ニューラル機械翻訳モデルは、入力文を符号化するエンコーダと出力文を生成するデコーダからなる系列変換モデルとして構成される。 エンコーダは、入力文のトークン列 $x = (x_0, x_1, \dots, x_N)$ が与えられ、隠れ状態 $h = (h_0, h_1, \dots, h_N)$ を出力する。 デコーダは、エンコーダによって生成された隠れ状態 h が与えられ、出力文のトークン列 $y = (y_0, y_1, \dots, y_M)$ を 1 つずつ出力する。トークン y_t の生成確率は、 x および $y_{<t} = (y_1, \dots, y_{t-1})$ を条件として最大化され、 N 個の対訳文 $(x^i, y^i)_{i=1}^N$ が与えられると、出力予測の対数尤度を次のように計算する。

$$\sum_{i=1}^{N} \log p(y^i|x^i) = \sum_{i=1}^{N}\sum_{t=1}^{M} \log p(y_t^i|y_1^i, \dots, y_{t-1}^i, x^i)$$

事前訓練では、入力文 x および長さ M 以下の出力文 y からなるデータ集合 $D = (x^1, y^1), \dots, (x^N, y^N)$ に対して、次のクロスエントロピー損失を最小化する。

$$L_X = -\sum_{i=1}^{N}\sum_{t=1}^{M} y_t^i \log p(y_t^i|y_1^i, \dots, y_{t-1}^i, x^i)$$

2.1.2.2 再訓練

強化学習に基づく機械翻訳モデルの再訓練のために、REINFORCE (Williams 1992) を用いる。

REINFORCE は方策勾配アルゴリズムの一種であり、機械翻訳モデルが報酬の期待値を最大化するように訓練される。

再訓練の損失関数は、対数尤度を報酬で重み付けすることで求められる。

$$L_R = \sum_{t=1} \log \pi_\theta(\hat{y}_t|\hat{y}_{t-1}, s_t)\,(R(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T) - R_b)$$

ここで、 s_tは時刻 t におけるデコーダの隠れ状態、 R は報酬関数、 R_b はベースライン報酬、$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T$ はデコーダからの出力文である。本研究では、ベースライン報酬としてミニバッチ内の平均報酬を使用する。

また、訓練を安定させるために、先行研究（Hashimoto & Tsuruoka 2019）と同様に強化学習の際に以下の損失関数を使用する。

$$L = \lambda L_X + (1 - \lambda) L_R$$

2.1.2.3 強化学習の報酬

本研究では、以下の評価指標を強化学習の報酬として用いる。

- BLEU[1] (Papineni et al. 2002): 単語 N-gram の一致率を用いて、出力文と参照文の表層的な類似度を評価する。
- chrF[1] (Popović 2017): 文字 N-gram と単語 N-gram の F 値を用いて、出力文と参照文の表層的な類似度を評価する。
- STS BERT (Yasui et al. 2019): STS タスク（Cer et al. 2017）において再訓練した BERT (Devlin et al. 2019) を用いて、出力文と参照文の意味的な類似度を評価する。
- Sentence BERT[2] (Reimers & Gurevych 2019): 自然言語推論（NLI: Natural Language Inference）タスク（Bowman et al. 2015）において再訓練した BERT を用いて、出力文と参照文の意味的な類似度を評価する。
- SimCSE[3] (Gao et al. 2021): NLI コーパス中の含意関係にある文対を正例として対照学習した BERT を用いて、出力文と参照文の意味的な類似度を評価する。
- BERTScore[4] (Zhang et al. 2020): RoBERTa (Large) から得られる文脈化トークン埋め込みの最大マッチングを用いて、出力文と参照文の意味的な類似度を評価する。
- BERT Regressor (Shimanaka et al. 2019): 機械翻訳の自動評価タスク（Bojar et al. 2017）において再訓練した BERT を用いて、出力文と参照文の意味的な類似度を評価する。
- BLEURT[5]（Sellam et al. 2020）: 折返翻訳などによって自動生成した単言語パラレルコーパス上で再訓練し、さらに機械翻訳の自動評価タスクにおいて再訓練した BERT を用いて、出力文と参照文の意味的な類似度を評価する。

[1] https://github.com/mjpost/sacrebleu
[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[3] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large
[4] https://github.com/Tiiiger/bert_score
[5] https://storage.googleapis.com/bleurt-oss/bleurt-large-512.zip

表 2.1.1: IWSLT-2014 De→En タスクにおける機械翻訳の強化学習の性能
（太字は強化学習による改善，下線は最高値を示す）

報酬	BLEU	Sent. BERT	BERT Reg.	SimCSE	chrF	BERT Score	BLEURT	STS BERT
なし	33.73	75.66	0.0478	82.10	54.27	58.47	0.0639	3.654
BLEU	**34.26**	74.91	0.0202	81.93	**54.39**	58.01	0.0234	3.641
Sent. BERT	**33.78**	**75.79**	**0.0513**	**82.24**	**54.38**	**58.72**	**0.0649**	**3.656**
BERT Reg.	33.47	**75.80**	**0.0557**	**82.32**	54.25	**58.64**	**0.0681**	3.650
SimCSE	33.73	**75.84**	**0.0512**	**82.25**	**54.37**	**58.76**	**0.0669**	**3.659**
chrF	**33.90**	**75.81**	**0.0517**	**82.24**	**54.45**	**58.69**	**0.0671**	**3.657**
BERTScore	**33.96**	**75.80**	**0.0511**	**82.30**	**54.48**	**58.80**	**0.0677**	**3.658**
BLEURT	**33.85**	**75.90**	**0.0572**	**82.33**	**54.44**	**58.92**	**0.0759**	**3.660**
STS BERT	**34.09**	**76.11**	**0.0528**	**82.52**	**54.62**	**59.10**	**0.0700**	**3.684**

2.1.3 評価実験

2.1.3.1 実験設定

事前訓練と強化学習の再訓練の両方で IWSLT-2014 の独英タスク（Cettolo et al. 2014）を用いた。訓練用データは 159,392 文対、検証用データは 7,245 文対、評価用データは 6,750 文対である。

機械翻訳モデルには Transformer (Vaswani et al. 2017) を使用し、レイヤ数を 6、ヘッド数を 4、次元数を 256、ドロップアウト率を 0.3 とした。事前訓練では、最適化手法を Adam (Kingma & Ba 2015) (学習率は 0.0003)、バッチサイズを 2,048 とし、検証用データにおける BLEU による early stopping によって訓練を停止した。強化学習では、最適化手法を Adam（学習率は 0.00001)、$\lambda = 0.3$、バッチサイズを 512 とし、報酬として用いる評価指標による early stopping によって訓練を停止した。実装には Reinforce-Joey[6]（Kiegeland & Kreutzer 2021）を用いた。

報酬計算および性能評価のための評価指標には、2.1.2.3 節のツールを用いた。ただし、STS BERT (Yasui et al. 2019) および BERT Regressor (Shimanaka et al. 2019) については、Hugging Face Transformers[7]（Wolf et al. 2020）の $\text{BERT}_{\text{BASE}}$[8]を用いて実装した。

2.1.3.2 実験結果

表 2.1.1 に実験結果を示す。1 行目の“報酬なし”は、強化学習を行わずに事前訓練のみを行ったベースラインである。このベースラインと 2 行目以降の強化学習の比較から、報酬に用いるのと同じ評価指標で性能を評価した際には、どの手法でも強化学習によって性能が向上することがわかる。

[6] https://github.com/samuki/reinforce-joey
[7] https://github.com/huggingface/transformers
[8] https://huggingface.co/bert-base-uncased

表 2.1.2：WMT-2017 Metrics タスクにおける人手評価とのピアソン相関（太字は最高値）

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	平均
BLEU	0.412	0.413	0.565	0.393	0.460	0.531	0.524	0.471
chrF	0.517	0.531	0.671	0.525	0.599	0.607	0.591	0.577
STS BERT	0.535	0.597	0.667	0.637	0.611	0.589	0.608	0.606
Sent. BERT	0.632	0.621	0.692	0.685	0.690	0.657	0.635	0.659
SimCSE	0.696	0.628	0.684	0.696	0.713	0.660	0.672	0.678
BERTScore	0.710	0.745	0.833	0.756	0.746	0.751	0.775	0.759
BERT Reg.	0.712	0.732	0.858	0.804	0.775	0.789	0.765	0.776
BLEURT	**0.845**	**0.845**	**0.870**	**0.865**	**0.861**	**0.846**	**0.860**	**0.856**

報酬として BLEU を用いた場合には、強化学習によって BLEU および chrF の表層マッチングに基づく評価指標しか改善されておらず、その他の BERT ベースの評価指標では性能が悪化している。一方で、同じく表層マッチングに基づく chrF を報酬として用いた場合には、強化学習によって全ての評価指標が改善されている。

BERT ベースの報酬のうち、Sentence BERT による強化学習は全体的にベースラインモデルとの差分が少なく、効果が少ないことがわかる。また、SimCSE を報酬とする強化学習では BLEU の改善が見られず、BERT Regressor を報酬とする強化学習ではベースラインモデルよりも BLEU が悪化した。

BERT ベースの報酬のうち、BERTScore、BLEURT、STS BERT を用いることで、今回検証した全ての評価指標において性能が向上することを確認できた。特に、STS BERT は過半数の評価指標において最高性能を達成しており、機械翻訳の強化学習に最も適した報酬関数であると言える。

2.1.4 分析

2.1.4.1 評価指標のメタ評価

本節では、表 2.1.1 の実験において強化学習の報酬として有効であった評価指標が、機械翻訳の人手評価と高い相関を持つのかを検証する。本分析では、WMT-2017 の自動評価タスク（Bojar et al. 2017）における to-English 言語対を対象に、評価指標と人手評価のピアソン相関を調査する。本タスクは、cs-en・de-en・fi-en・lv-en・ru-en・tr-en・zh-en の 7 言語対が対象で、各 560 文対（出力文と参照文）に人手評価が付与されている。

分析の結果を表 2.1.2 に示す。BERT ベースの評価指標が、BLEU および chrF の表層マッチングの評価指標よりも人手評価との高い相関を持つことがわかる。特に、BLEURT が全ての言語対において人手評価との最高の相関を示している。しかし、想定とは異なり、強化学習の報酬として最適であった STS BERT は、人手評価との相関は低かった。

表 2.1.3：評価指標同士でのピアソンの相関係数

	BLEU	STS BERT	chrF	SimCSE	Sent. BERT	BERT Reg.	BLEURT	BERT Score	平均
BLEU	-	0.449	0.788	0.417	0.428	0.517	0.496	0.641	0.534
STS BERT	0.449	-	0.671	0.772	0.788	0.648	0.665	0.636	0.661
chrF	0.788	0.671	-	0.616	0.635	0.608	0.613	0.715	0.664
SimCSE	0.417	0.772	0.616	-	0.856	0.653	0.717	0.664	0.671
Sent. BERT	0.428	0.788	0.635	0.856	-	0.674	0.712	0.662	0.679
BERT Reg.	0.517	0.648	0.608	0.653	0.674	-	0.866	0.798	0.681
BLEURT	0.496	0.665	0.613	0.717	0.712	0.866	-	0.805	0.696
BERTScore	0.641	0.636	0.715	0.664	0.662	0.798	0.805	-	0.703

2.1.4.2 評価指標間の相関関係

本節では、評価指標間の相関関係が強化学習の性能評価に影響を与えているのかを検証する。2.1.4.1 節と同様に WMT-2017 の自動評価タスクにおける to-English 言語対を対象として、本節では評価指標間のピアソン相関を調査する。

分析の結果を表 2.1.3 に示す。まず、BLEU と他の評価指標の相関が低いことがわかる。単語 N-gram を扱う chrF やトークン埋め込みのマッチングに基づく BERTScore との相関は比較的高いものの、文単位で評価を行う他の評価指標との相関が低いことから、BLEU が文単位での大域的な評価に適していない可能性が示唆される。この特性が、表 2.1.1 および表 2.1.2 において BLEU の性能が低いことに影響を与えている可能性がある。

STS BERT は、他の評価指標との相関が比較的低い傾向が見られた。つまり、表 2.1.1 の実験において STS BERT を報酬とする強化学習が多くの評価指標から高評価を得たことは、報酬と評価指標の相性の問題ではないと考えられる。

2.1.5 おわりに

本研究では、報酬計算と性能評価の両方において複数の評価指標を検討することで、機械翻訳の強化学習に適した報酬について調査した。IWSLT-2014 の独英翻訳における実験の結果、STS タスクにおいて訓練した BERT を報酬とする強化学習により、多くの評価指標において性能を改善できることが明らかになった。ただし、STS BERT は WMT-2017 の自動評価タスクにおける人手評価との相関は比較的低く、この観点からは良い評価指標とは言えない。また、STS BERT と他の評価指標との相関も比較的低いため、報酬と評価指標の相性が良いとも言えない。なお、BERTScore および BLEURT は、人手評価との相関も良好で、他の評価指標との相関も比較的高く、かつ強化学習の報酬にも適した評価指標であると言える。

謝辞

本稿は、国際ワークショップ「 The 9th Workshop on Asian Translation (WAT 2022)」に採択された論文（Nakatani et al. 2022）に基づいて、それらの論文を再構成し、解説したものである。

これらの研究成果は JSPS 科研費（若手研究、課題番号：JP20K19861）および国立研究開発法人情報通信研究機構の委託研究（課題番号: 225）により得られたものである。ここに謝意を表する。

参考文献

O. Bojar, Y. Graham and A. Kamran. (2017). Results of the WMT17 Metrics Shared Task, in Proceedings of the Second Conference on Machine Translation, pp. 489–513.

S. R. Bowman, G. Angeli, C. Potts and C. D. Manning. (2015). A large annotated corpus for learning natural language inference, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.632–642.

D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio and L. Specia. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, in Proceedings of the 11th International Workshop on Semantic Evaluation, pp. 1–14.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli and M. Federico. (2014). Report on the 11th IWSLT evaluation campaign, in Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign, pp.2–17.

J. Devlin, M.-W. Chang, K. Lee and K. Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

T. Gao, X. Yao and D. Chen. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6894–6910.

K. Hashimoto and Y. Tsuruoka. (2019). Accelerated Reinforcement Learning for Sentence Generation by Vocabulary Prediction, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3115–3125.

S. Kiegeland and J. Kreutzer. (2021). Revisiting the Weaknesses of Reinforcement Learning for Neural Machine Translation, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1673–1681.

D. P. Kingma and J. Ba. (2015). Adam: A Method for Stochastic Optimization, in Proceedings of the 3rd Inter-national Conference on Learning Representations.

A. Nakamachi, T. Kajiwara and Y. Arase. (2020). Text Simplification with Reinforcement Learning Using Supervised Rewards on Grammaticality, Meaning Preservation, and Simplicity, in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference

on Natural Language Processing: Student Research Workshop, pp. 153–159.

Y. Nakatani, T. Kajiwara and T. Ninomiya. (2022). Comparing BERT-based Reward Functions for Deep Reinforcement Learning in Machine Translation, in Proceedings of the 9th Workshop on Asian Translation (WAT 2022), pp.37-43.

K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. (2002). 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.

M. Popović. (2017). chrF++: words helping character n-grams, in Proceedings of the Second Conference on Machine Translation, pp. 612–618.

M. Ranzato, S. Chopra, M. Auli and W. Zaremba. (2016). Sequence Level Training with Recurrent Neural Networks, in Proceedings of the 4th International Conference on Learning Representations.

N. Reimers and I. Gurevych. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3982–3992.

T. Sellam, D. Das and A. Parikh. (2020). BLEURT: Learning Robust Metrics for Text Generation, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892.

H. Shimanaka, T. Kajiwara and M. Komachi. (2019). Machine Translation Evaluation with BERT Regressor, arXiv:1907.12679.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin. (2017). Attention is All you Need, in Advances in Neural Information Processing Systems 30, pp. 5998–6008.

R. J. Williams. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning, Machine Learning, Vol. 8, pp. 229–256.

S. Wiseman and A. M. Rush. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1296–1306.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush. (2020). Transformers: State-of-the-Art Natural Language Processing, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45.

L. Wu, F. Tian, T. Qin, J. Lai and T.-Y. Liu. (2018). A Study of Reinforcement Learning for Neural Machine Translation, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3612–3621.

G. Yasui, Y. Tsuruoka and M. Nagata. (2019). Using Semantic Similarity as Reward for Reinforcement Learning in Sentence Generation, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 400–406.

X. Zhang and M. Lapata. (2017). Sentence Simplification with Deep Reinforcement Learning, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 584–594.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi. (2020). BERTScore: Evaluating Text Generation with BERT, in Proceedings of the 8th International Conference on Learning Representations.

2.2 文書単位の翻訳品質改善に向けた取り組み

東北大学 鈴木　潤

2.2.1 文書単位での機械翻訳の背景

機械翻訳技術の進展により翻訳品質は実用レベルに達し，商用サービスとしても実社会で広く利用される段階に至っている．しかし，一文単位の翻訳（以下「**単文翻訳**」と略記），特にデータが集めやすいドメインの機械翻訳結果は十分に実用レベルと言えるが，例えば，文章全体を翻訳する際などに，表現やスタイルが文書全体で一貫した翻訳結果を得ることは現状の機械翻訳技術でも容易ではない．実際に，機械翻訳システムの翻訳結果を文書全体で評価した時，整合性や一貫性の観点で人間の翻訳に大きく及ばないことが散見される[Laubli et al., 2018]．よって，文書単位で整合性および一貫性のある翻訳（以下「**文書翻訳**」と略記）の実現は，機械翻訳の研究コミュニティにおいても喫緊に取り組むべき課題の一つと認識されている．

一貫性のある文書翻訳結果を得たい場合，安直には文書全体を一文とみなして処理すればよい．原理的には単文翻訳モデルと全く同じ翻訳方式で翻訳することができる．例えば，文献[Junczys-Dowmunt et al., 2019]では，文書全体を入出力とするモデルを提案している．実際にこの翻訳方式は単文翻訳から文書翻訳への自然な拡張といえる．しかし，現在機械翻訳モデルとしてよく用いられる Transformer[Vaswani et al., 2017]は，いくつかの要因により文長が長くなると翻訳精度が低下するという欠点が指摘されている[Ranzato et al., 2016]．そのため，自然な拡張による文書翻訳の翻訳品質を劣化させる別の要因があり，実用上は解決すべき課題として残っている．また，別観点の課題として，高い翻訳品質を達成するには多くのデータが必要となるが，文書単位で対応のついた対訳データは，文単位の対訳データと比べてそれほど多くない．これらの要因から，現在の文単位のニューラル翻訳の方式をそのまま文書翻訳に拡張しても，整合性や一貫性が向上しても全体の翻訳品質は低下するという現象が見られる．

これらの背景から，文書翻訳の性能向上を目標に新たな方法を考案する．具体的には，従来用いられてきた Transformer ベースの単文翻訳モデルに対して，少量の文書対訳データを用いて追加学習をすることで，データ量の問題に対応する．また，単文翻訳モデルが生成する複数の翻訳候補を文書翻訳時の制約と考え，文単位の制約付き生成をすることで長文を生成する際の翻訳品質の劣化を防ぎ文書単位の翻訳を獲得する．

提案法を単文翻訳で各文を独立に翻訳した結果を並べる方式で文書を翻訳する場合と比較し，提案法が BLEU スコアの意味で高い翻訳品質を達成できることを報告する．

2.2.2 提案法

本研究では，文書全体を入力し文書全体を出力するモデル（文書翻訳モデル）をベースとしつつ，通常の 1 文から 1 文へ翻訳するモデル（単文翻訳モデル）も活用する．具体的には，文単位の学習データを用いて翻訳モデルの事前学習をする．次に，事前学習済みモデルに対して文献[Junczys-Dowmunt et al., 2019]の手法を参考に文書翻訳用のデータを用いて追加学習する．また，

事前学習済みモデルに対して単文翻訳の追加学習データを用いて文書翻訳モデルとは独立に追加学習をおこない単文翻訳用モデルも作成する．つまり，同じ事前学習済みモデルから文書翻訳用モデルと単文翻訳用モデルの 2 種類のモデルを作成することになる．モデルの学習に関してはこの 2 種類のモデルを作るという点が従来と違う点と言える．

次に，文書翻訳モデルを用いて文章を生成する際には文書単位の入出力に対して学習した文書翻訳モデルを用いて翻訳をするが，文書中の各文は単文翻訳モデルが生成する Nベスト文を出力候補の制約として生成する．これは，単文翻訳モデルが良いと判断する出力文章群の中から，文書翻訳モデルを用いて文書全体として適切な文の組合せを求める処理とみなすこともできる．

2.2.3 実験設定

データ：前述のように提案法の実験をするためには，文書翻訳用のデータが必要となる．ここでは文書翻訳追加学習用および評価用データとして，WMT-2020 ニュース翻訳タスクの開発/評価セットを使用した[Barrault et al., 2020]．

モデル：単文翻訳の事前学習済み翻訳モデルとして，WMT-2022 で好成績を得た NT5 チームが学習した Transformer[Vaswani et al., 2017]ベースの翻訳モデル[Morishita et al., 2022]を用いた．次に，文書翻訳モデルは，WMT-2020 ニュース翻訳タスクの開発データを文書単位の入出力を用いて追加学習した．一方，単文翻訳モデルは，WMT-2020 ニュース翻訳タスクの開発データを文単位の入出力で追加学習した．翻訳モデルの学習には fairseq ツールキット[Ott et al., 2019]を使用した．また，文書翻訳を実行するために fairseq ツールキットを活用しつつ制約付き翻訳用のコードを独自に作成した．

評価指標：本実験の評価は，WMT-2020 の標準評価ツールである SacreBLEU[Post 2018]を用いて BLEU スコアにて実施した．

2.2.4 実験結果

表 1 に実験結果を示す．提案法を使うことでベースラインである文単位で翻訳する方法より性能が向上した．また，表中の「文書翻訳モデル(w/o Nベスト制約)」とは，文書翻訳を単文翻訳と同様に制約なしで普通に翻訳した際に得られた結果である．このことから，単文翻訳の制約を利用することで，性能が向上することが示唆された．

表 1: 実験結果

	英語->日本語	日本語->英語
単文翻訳モデル(ベースライン)	27.7	26.2
文書翻訳モデル(提案法)	28.1 (+0.4)	26.5 (+0.3)
文書翻訳モデル(w/o Nベスト制約)	26.7 (-1.0)	25.6 (-0.6)

2.2.5 まとめ

ここで示した結果は文書翻訳の取り組みにおける最初の実験結果という位置づけである．よっ

て，今後様々な実験設定を用いて提案法の有効性を検証する必要がある．また提案法自体もより精緻に改良し，さらなる翻訳品質の向上を目指す．

参考文献

[Laubli et al., 2018] Samuel Laubli, 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, October-November 2018.

[Junczys-Dowmunt et al., 2019] Marcin Junczys-Dowmunt. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. InProceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers), pp. 225–233, August 2019.

[Ranzato et al., 2016] MarcAurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In 4th International Conference on Learning Representations, 2016.

[Barrault et al., 2020] Loıc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta R. Costa-jussa, Christian Federmann, Yvette Graham, Roman Grund kiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubesic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 Conference on Machine Translation (WMT20). In Proceedings of the Fifth Conference on Machine Translation, pp. 1–55, November 2020.

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In Advances in Neural Information Processing Systems, Vol. 30, 2017.

[Morishita et al., 2022] Makoto Morishita, Keito Kudo, Yui Oka, Katsuki Chousa, Shun Kiyono, Sho Takase, and Jun Suzuki. NT5 at WMT 2022 General Translation Task. In Proceedings of the Seventh Conference on Machine Translation, pp. 318–325, December 2022.

[Ott et. al., 2019] 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 Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 48–53, June 2019.

[Post, 2018] Matt Post. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the

Third Conference on Machine Translation: Research Papers, pp. 186–191, October 2018.

2.3 非自己回帰モデルによるポストエディットを用いた高速な語彙制約付き機械翻訳

東京都立大学　システムデザイン学部
小町　守

2.3.1　はじめに

翻訳を行う際に特定の表現で翻訳したいケースが存在する．例えば科学や医療などの特定の分野で翻訳する際には，そこで出現する専門用語や固有名詞は特定の表現で翻訳することが求められる．本研究ではこのような翻訳先の言語における特定の表現が事前に 0 個以上指定されているという，語彙に関する制約を含む機械翻訳に取り組む．

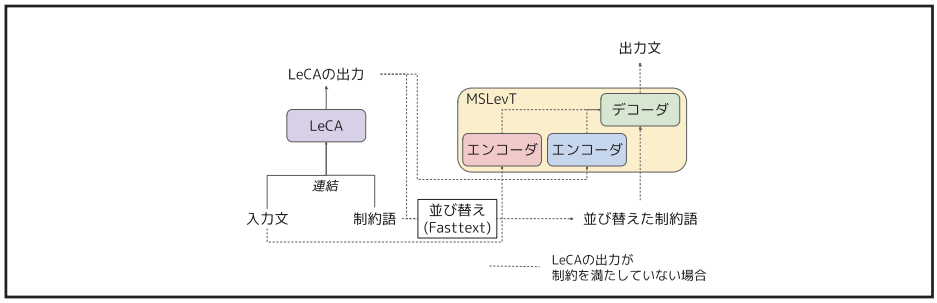
ニューラル機械翻訳はその評価尺度である BLEU [14] において高いスコアを出すことで知られており，現在も広く研究が進められている．一方で，ニューラルモデルは出力語を制御することが難しいことが指摘されている．Hokamp ら [1] は制約語を翻訳文に確実に含めるために，翻訳文の生成時に制約も含めた探索であるグリッドビームサーチ（GBS）を提案した．しかし，この手法は制約を確実に出力できる“ハード” な手法であるものの，制約語による探索の制限によって翻訳精度が損なわれ得ることが問題点である．一方で，Chen ら [2] は制約語を入力文の末尾に連結させるようなデータ拡張を行うことで，制約語の出力を試みた．しかし，この手法は制約語を挿入することによる翻訳精度の低下は比較的少ないものの，出力することを保証していない“ソフト” な手法である．Chousa ら[3] はこれらの手法を組み合わせる手法を用いることで，翻訳精度を維持しながら確実に制約語を出力させることに成功した．しかし，この手法では，ビームサイズが小さい時に GBS の制約により出力が終わらないことがあることが示唆されており，ビームサイズを決定するための予備実験が必要になる．また，それに伴いビームサイズが大きいと推論速度が遅くなることも問題である．

本研究では，Wan ら [4] の提案したポストエディット手法であるMulti Source Levenshtein Transformer を“ソフト” な手法で得られた出力の中で制約を満たしていない文に対して後処理的に利用することで，制約語を含む機械翻訳に取り組む．このモデルは入力文と修正前の翻訳文のための2つのエンコーダと，Gu ら [5] によって提案されたLevenshtein Transformer のデコーダからなるモデルである．Levenshtein Transformer は，削除，挿入，置換の3つの操作を繰り返すことによって出力文の修正を行うモデルである．デコーダの初期値として制約語を入れてそれらに対して削除と制約語内への挿入の操作を禁じた上で，先の3操作を繰り返させることにより制約を満たした出力を生成することが可能となる．また，修正を繰り返すという機構によって制約語を強制的に出力させたとしても文構造が大きく崩れないことが期待される．

また，本タスクの設定と与えた制約語に操作を行わないという手法の制限により，初期値として制約語を与えるにあたってその順序を事前に決定する必要がある．この課題に対して，翻訳文中に制約語に該当する同義語が存在しているという仮定の下，制約語が“ソフト” な手法で得られた

翻訳文に対してどの単語，フレーズに対応するのかを測って語順を決定することで取り組んだ．

日英対訳コーパスを用いて双方向で実験を行った結果，本研究の提案手法によって制約語を確実に出力することが可能であることを確認した．また，翻訳精度も“ソフト”な手法と同等の性能で翻訳することができた．更に，ビームサイズ決定のための予備実験が不要なことに加え，先行研究で利用されたビームサイズと推論速度の比較を行うことで，“ハード”な手法を用いるよりも今回提案したポストエディットを用いた手法が推論時間を大きく削減できることが確認できた．



2.3.2 提案手法

本研究での制約語を含む機械翻訳に対する取り組みを図1に示す．まず，データ拡張を用いた手法である，LeCA [2] を用いて制約語を考慮した翻訳を行う．しかし，この手法は制約語の出力を保証するものではないため，制約を満たせなかった文に対して自動ポストエディットを行う．自動ポストエディットには Wan ら [4] の提案したMulti-source Levenshtein Transformer（MSLevT）を用いるが，そのためには制約語の順序を決定する必要がある．従って，fasttext [6] を用いてLeCA の出力と制約語の単語ベクトルを取得し，コサイン類似度を測ることで制約語の語順を決定する．そして，MSLevT の片方のエンコーダに原言語文を，もう一方のエンコーダに LeCA の出力を入力し，デコーダの初期値に並び替えた制約語を与えることで制約を満たす翻訳を実現する．各手法について以下の節で説明する．

LeCA を用いた語彙制約を考慮した翻訳

LeCA はデータ拡張により，事前に指定された単語を生成文に含めやすくなるよう設計されている．具体的には，制約語を入力文の末尾に連結させた系列を入力とするようなデータを利用する．このようなデータにより，デコーダで翻訳を行う前に原文と制約語の情報を取得することができ，制約語を考慮した上で文生成を始めることができる．さらに，LeCA はポインタネットワークを採用しており，入力文に連結された制約語を適切な場所にコピーしながら翻訳文を生成することが可能となっている．

fasttext による制約語の並び替え

先行研究 [3] より LeCA のみを用いた際の翻訳精度が非常に高いことから，制約語を出力できなかった LeCA の翻訳文中にも制約語に対応する語が同義語あるいは近い表層形の形で存在するという仮定の下で並べ替えを行う．並び替えの手順を以下に示す．

1. 任意の一つの制約語と LeCA の出力両方における各単語の埋め込みベクトルを fasttext を用いて取得する.

2. 複数単語からなるフレーズの埋め込みベクトルは制約語を構成する単語の埋め込みベクトルの平均とする．LeCA に出現する制約語に対応するフレーズは制約語と同じ単語数と仮定して，制約語がn-gram のとき LeCA の出力文における全ての n-gram の埋め込みベクトルを計算する．ここでは以降 n-gram の塊を“単語列” と呼ぶ．また単語列の先頭の単語をその単語列の代表語とみなす．制約語がある単語列に対応するとき，その制約語は代表語に対応するものとする．

3. 制約語と全ての単語列の間でコサイン類似度を計算し，降順にランクをつける．

4. 基本的には最も順位の高い単語列に制約語が対応するものとする．しかし，対応する代表語が他の制約語と重複した場合，コサイン類似度のより高い方を優先して対応づけ，もう一方を棄却する．ここで棄却された制約語は，次点にコサイン類似度の高い単語列に対応づける．

5. 全ての制約語に対して同様の処理を行い，全ての制約語を LeCA の出力文中の単語列に対応づける.

この手法を用いたときの制約語の順番と参照訳に出現する制約語の順番のスピアマンの順位相関係数は英語で 0.664，日本語で 0.718 であった．正の相関が見られることから，本提案手法の有効性が確認できる.

Multi-source Levenshtein Transformer（MSLevT）によるポストエディット

MSLevT は2つのエンコーダと Levenshtein Transformer のデコーダからなるモデルである．エンコーダでは片方に原言語文を，もう一方には LeCA の生成した翻訳文を入力する．先行研究[7] では，両エンコーダの出力を連結したものを Attention 機構の Key としてデコーダに渡すことでより良い表現を獲得でき，自動ポストエディットに効果的であると主張されている．本研究でもそれに従って同様の機構を用いる.

デコーダには初期値として制約語を与えて，Levenshtein Transformerと同様の操作を行うことで制約を満たした文を生成することを実現する．Levenshtein Transformer は，１）トークンの削除，２）プレースホルダーの挿入，３）プレースホルダーをトークンへ置換，の3つの操作を繰り返すことで文を生成する．これらの操作は事前に指定したイテレーション数に達するか，生成される文が変化しなった際に終了する．しかし，制約を満たすために初期値として入力される制約語への編集は好ましくない．従って，初期値への削除の操作と，制約語内部への挿入の操作を行わないように先行研究の実装を変更した.

2.3.3 実験設定

データセット

学習データには WAT 2022 [8] の restricted translation task で公開されているデータを用いる．これは日本語で書かれた科学技術論文の概要とその翻訳である英語論文から作成された対訳コーパスである Asian Scientific Paper Excerpt Corpus （ASPEC）[9] に基づくデータである．学習データが300万文，検証データが1,790文，テストデータが1,812文与えられる．検証デー

タとテストデータに関しては文ごとに制約語がリストとして与えられている．テストデータに関して一文あたりに与えられる制約語の個数の平均は日本語で3.30個，英語で3.25個，最大数は14個であった．

ASPEC のデータは文アライメントを用いて作成されており，学習データは確信度の高い順に並び替えられている．先行研究 [10] では ASPEC の確信度の低いデータを学習に利用することは翻訳精度の悪化を招くことが実験で示されているため，本研究では300万文のうち先頭の200万文を機械翻訳モデルの学習データとして利用する．単言語コーパスとして利用する際には対訳コーパスとしての確信度は関係がないため，fasttext の学習には300万文全てを利用する．

ASPEC を用いた先行研究 [3, 11] と同様に，日本語文は形態素解析器である MeCab [12]（mecab-ipadic-NEologd）を用いてトークナイズを行い，英語文は mosestokenizer を用いて前処理を行い，それぞれ sentencepiece を用いて語彙サイズ4,000でサブワード分割した．加えて，Moses の clean-corpus-n を用いて文長が250を超えるものと，日本語文と英語文の文長比が9倍以上のものは取り除いた．

ASPEC の日本語文に出現する空白と英数字は全角となっているが，sentencepiece の実装内の前処理において半角文字に正規化される．制約語の評価は完全一致に基づいて行われるため，日本語の生成文は後処理的に空白と半角の英数字を全角に変換するように正規化処理を行う．また，ASPEC のテストセットのうち37文は未知語（捏，顆など）を含む制約語が与えられていた．

評価指標

■BLEU score BLEU score はシステム出力と参照訳の単語 n-gram の一致率に基づいてコーパスレベルで計算されるものであり，機械翻訳の分野では広く用いられている評価指標である．スコアの計算には公開されているツールであるSACREBLEU [15] を利用する．

■Consistency score（CS） Consistency score はコーパス中の文のうち，制約語を全て出力できた文の割合を表す．制約語が文中に含まれているかどうかは完全一致によって判定を行う．また，英文の評価を行う際にはシステム出力と参照訳の両方を小文字にしてから空白を含む文字一致によって評価を行う．

■Final score（FS） 最終的な評価としてBLEU score とConsistency score を共に考慮した評価を行う．具体的には，制約を満たしていない文を空文字に置き換えてBLEU score を計算する．

モデル

■LeCA LeCA モデルにはtransformer big [17] モデルを用いる．Chen ら [2] の公開している実装に基づき，Chousa らの論文を参考にハイパーパラメータを決定した．モデルは検証データにおいて最もBLEU score が高かったものを選択した．

■fasttext Python 言語の fasttext モジュールを利用する．本研究では日英両言語ともに初期状態から学習を行った．パラメータはデフォルトのものを用いた．

■MSLevT Wan ら [4] の実装したものとおおよそ同様のモデル，ハイパーパラメータで MSLevT を利用する．しかし，彼らの実装では LevT の文生成の過程において制約語を削除し

たり，制約語内への挿入を行なってしまう可能性があるため，生成される文が必ずしも制約語を含むとは限らない．したがって，Susantoら［18］の実装同様に，初期値として入力される制約語の削除，および制約語内への挿入を禁止するように実装を書き換えた．

一般に非自己回帰モデルは自己回帰モデルからの知識蒸留を行うことで BLEU score が向上することが報告されている［19, 20］．そこで，LeCA の出力文を参照訳とする蒸留データを MSLevT 用に作成する．蒸留データのみを用いた MSLevT の学習では一方のエンコーダへの入力と参照訳が同一のものになるため，モデルは入力をコピーするように学習が行われる．一方で，テストを行う際には LeCA の出力に含まれない制約語が LevT の初期値として与えられるため，学習時の条件との不一致が存在する．したがって本研究では蒸留データを元のデータと組み合わせて扱う．ここでは2つの方法でこれらのデータを組み合わせて利用した．一つ目は MSLevT を蒸留データで10エポック学習し，オリジナルのデータでファインチューニングを行う．二つ目は蒸留データとオリジナルのデータを混合したデータを用いて MSLevT の学習を行う．

また，本研究で用いたテストデータには制約に未知語を含むものが存在している．これらのケースにおいて MSLevT は初期値として制約語を与えても，out-of-vocabulary（OOV）なので未知語として特殊トークンに変換して処理を行う．さらに先行研究の実装ではこの特殊トークンに対して置換の操作が行われてしまい，制約語が別の単語に変換されてしまう問題があった．したがって，本研究では未知語のための特殊トークンを新たに設定し，未知語を特殊トークンのまま出力するように実装に変更を加えた．そして，出力文に特殊トークンが出現した際には文生成時に満たせていない制約語の情報と，特殊トークン前後のマッチングによりスパンを検出して制約語に置換することで対応した．

2.3.4 実験結果

各種スコア

表1は LeCA の出力，LeCA の出力を全て MSLevT でポストエディットしたもの，LeCA の出力に対して制約を満たしていないもののみを MSLevT でポストエディットしたものの翻訳結果である．

表 1: 実験結果．ただし，“dist → orig” は蒸留データで 10 エポック学習した後，元のデータでファインチューニングしたものを，“dist + orig” は蒸留データと元のデータを混合したデータで学習したことを表す．

モデル	英日			日英		
	BLEU	CS	FS	BLEU	CS	FS
LeCA	**52.0**	0.805	36.0	39.0	0.719	19.6
MSLevT	35.8	**1.000**	35.8	32.6	**1.000**	32.6
MSLevT (dist→orig)	37.5	**1.000**	37.5	32.2	**1.000**	32.2
MSLevT (dist+orig)	44.4	**1.000**	44.4	**40.2**	**1.000**	**40.2**
LeCA+MSLevT	50.1	**1.000**	50.1	39.3	**1.000**	39.3
LeCA+MSLevT (dist→orig)	50.4	**1.000**	**50.4**	39.3	**1.000**	39.3
LeCA+MSLevT (dist+orig)	50.2	**1.000**	**50.2**	39.8	**1.000**	39.8

英日翻訳では LeCA が最も高い BLEU を記録したものの，CSが0.805となっている．そのため, FS も36.0と元の BLEU に比べて16.0低くなった. 一方で LeCA の出力を全て MSLevT でポストエディットした結果である “MSLevT” の結果からは BLEUが大きく低下するものの CSが1.0となっていることから全ての制約語を出力できていることが確認できる．そして，制約を満たせていないもののみをポストエディットした結果である “LeCA +MSLevT” では LeCA からの BLEU の減少を1.9に抑えつつCSを1.0にすることができ，LeCA の翻訳精度を維持しながら全ての制約語を出力できることが確認できた．日英翻訳でも英日翻訳同様の傾向が見られた．ただし，日英翻訳では “LeCA + MSLevT” が “LeCA” よりも高い BLEU を達成することができた．

以下，蒸留データを用いた時の結果に関しての結果をまとめる．蒸留データを用いて MSLevT を一定エポック学習した後, 元のデータでファインチューニングする方法（MSLevT（dist → orig））は英日翻訳で+1.7，日英翻訳では-0.4となり，MSLevT の精度向上には一貫した有効性が確認できなかった．また，この方法の問題点として，蒸留データのみで学習を行う際に出力がLevT の出力をそのまま出力するため，検証データによるモデルの評価と決定ができないことがある．したがって, 何エポックを事前学習として蒸留データで回すのかを実験的に決定しなければならず, 汎用性が高くない．一方で蒸留データと元のデータの混合データ（dist+orig）を用いて学習することには，一定の効果が見られることが確認できた．“MSLevT” の結果では英日翻訳では+8.6, 日英翻訳では+7.6 と BLEU の大きな改善が見られる．この要因として，蒸留データの多様性の少なさによる非自己回帰モデルの精度向上のほかに，MSLevT が LeCA の出力が入力されるエンコーダの情報をより利用するようになったことが考えられる．MSLevT の出力結果を見るとLeCA の出力結果と比較して制約語以外の情報が欠落する傾向が見られたため，蒸留データを用いてコピーすることも学習することでこの問題が軽減したことが考えられる．しかし，これらの蒸留データを用いて学習したモデルを用いて，制約を満たせなかった文のみに対して訂正を行うと，“LeCA +MSLevT（dist → orig）” の英日翻訳で+0.3，日英翻訳で+0.0，“LeCA + MSLevT（dist + orig）” の英日翻訳で+0.1，日英翻訳で+0.5と僅かな向上しか見られなかった．この要因としてはポストエディットをした文数の割合が英日で19.5%，日英で28.1%であり，MSLevT の精度向上がコーパス全体に与える影響が少なかったことが原因として考えられる．また，日英翻訳ついては蒸留データと元のデータを組み合わせて MSLevT を学習させると，LeCA の出力のうち制約を満たしていないもののみを選択してポストエディットするよりも, 全 LeCA の出力をポストエディットする方が高い BLEU を示した．この結果から，MSLevT は英文に対しては制約を満たすためのみならず，翻訳品質を向上させるために利用することも有効であることが示唆される．

表 2: GPU での計算速度．モデルの出力には GPU を一台用いた．

モデル	ビームサイズ	英日		日英	
		Sec/sent	BLEU	Sec/sent	BLEU
LeCA	5	0.094	52.0	0.099	39.0
	30	0.221	52.1	0.228	38.9
提案手法	5	0.115	50.1	0.126	39.3

翻訳速度

本研究で用いた手法に関して先行研究との比較として翻訳速度の調査を行った．LeCA と GBS を組み合わせた先行研究では，ビームサイズが30以上でないと制約語を全て生成することができないという報告があった一方，本研究の提案手法はビームサイズを変更することなく，全ての制約語を含む翻訳文を生成することができる．そこで，ここではビームサイズが30のときの LeCA との推論速度の比較を行う．各条件における1文あたりの翻訳速度と，LeCA のビームサイズが5 のときの翻訳速度を1としたときの翻訳速度の比率をまとめた結果が表2である．

結果から，LeCA はビームサイズを5から30に上げると翻訳にかかる時間が2倍以上になることが確認できる．一方で本研究で提案した手法では翻訳文の生成までに1.3倍以内の時間で制約語を確実に含む文を生成できることが確認できる．つまり，MSLevT で翻訳文の訂正を行うのにかかる時間は LeCA のビームサイズを上げることによる遅延時間よりも大幅に短いことがいえる．MSLevT の訂正時間が短い理由として，非自己回帰モデルによる生成であるという点と制約語を含まない文を選択しているため訂正する文数が少ないという点が挙げられる．さらに，Postら [21] の報告では，GBS を用いるとおよそ3倍の計算時間がかかることが指摘されており，この点からも本研究の提案手法は先行研究と比較して計算速度に関して優位性があるといえる．また本手法は，制約語を全て含めることを目的とした，ビームサイズを決めるための予備実験を行う必要がない．

2.3.5 おわりに

本研究では，制約語を含む機械翻訳に対して非自己回帰モデルを用いた自動ポストエディット手法を導入した．実験の結果，LeCA と同程度の翻訳精度を維持したまま，制約語を100%生成文に含めることに成功した．さらに，提案手法はビームサイズを決定するための予備実験を必要とせず，グリッドビームサーチを用いた既存の手法と比較して，制約を満たしながら高速に翻訳文を生成することが可能であることを示した．

参考文献

[1] C. Hokamp and Q. Liu, “Lexically constrained decoding for sequence generation using grid beam search,” In ACL, 2017.

[2] G. Chen, Y. Chen, Y. Wang, and V.O. Li, “Lexical-constraint aware neural machine

translation via data augmentation," In IJCAI, 2020.
[3] G. Chen, Y. Chen, Y. Wang, and V.O. Li, "Input augmentation improves constrained beam search for neural machine translation: NTT at WAT 2021," In WAT, 2021.
[4] D. Wan, C. Kedzie, F. Ladhak, M. Carpuat, and K. McKeown, "Incorporating terminology constraints in automatic postediting," In ACL, 2020.
[5] J. Gu, C. Wang, and J. Zhao, "Levenshtein transformer," In NeurIPS, 2019.
[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov,"Enriching word vectors with subword information," In ACL, 2017.
[7] A. Tebbifakhr, R. Agrawal, M. Negri, and M. Turchi, "Multisource transformer with combined losses for automatic post editing," In WMT, 2018.
[8] T. Nakazawa, H. Mino, I. Goto, R. Dabre, S. Higashiyama, S. Parida, A. Kunchukuttan, M. Morishita, O. Bojar, C. Chu, A. Eriguchi, K. Abe, Y. Oda, and S. Kurohashi, "Overview of the 9th workshop on Asian translation," In WAT, 2022.
[9] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara, "ASPEC: Asian scientific paper excerpt corpus," In LREC, 2016.
[10] M. Morishita, J. Suzuki, and M. Nagata,"NTT neural machine translation systems at WAT 2017," In WAT, 2017.
[11] M. Morishita, J. Suzuki, and M. Nagata,"NTT neural machine translation systems at WAT 2019," In WAT, 2019.
[12] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," In EMNLP, 2004.
[13] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," In EMNLP, 2018.
[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," In ACL, 2002.
[15] M. Post, "A call for clarity in reporting BLEU scores," In ACL, 2018.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, "Attention is all you need," In NeurIPS, 2017.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, "Attention is all you need," In NeurIPS, 2017.
[18] R.H. Susanto, S. Chollampatt, and L. Tan, "Lexically constrained neural machine translation with Levenshtein transformer," In ACL, 2020.
[19] J. Gu, J. Bradbury, C. Xiong, V.O. Li, and R. Socher, "Non-autoregressive neural machine translation," In ICLR, 2018.
[20] C. Zhou, J. Gu, and G. Neubig, "Understanding knowledge distillation in non-autoregressive machine translation," In ICLR,
2020.

[21] M. Post and D. Vilar,“ast lexically constrained decoding with dynamic beam allocation for neural machine translation,” In NAACL, 2018.

3. サーベイ論文報告

3. 特許機械翻訳の課題解決に向けたサーベイ論文の執筆

奈良先端科学技術大学院大学　須藤　克仁

3.1 活動の概要

AAMT/Japio 特許翻訳研究会では、特許機械翻訳に関する調査研究活動の一環として、昨今のニューラル機械翻訳の発展を踏まえ、特許機械翻訳に資する機械翻訳関連技術をまとめたサーベイ論文の執筆を行った。本活動は 2020 年度から研究会委員とオブザーバーで構成される 10 名のメンバーによって進められ、研究会会合や特許情報シンポジウム、研究会年度報告書で随時途中報告を行ってきた。論文は 2022 年 5 月に言語処理学会の論文誌「自然言語処理」の解説論文として投稿され、同 6 月に採録が決定、同 9 月に出版された。「自然言語処理」はオープンアクセス雑誌であり、科学技術振興機構(JST)が運営する電子出版プラットフォームである J-STAGE にて無料で公開されている。書誌情報は以下の通りである（なお、著者順は著者名の 50 音順とした)。

今村賢治, 越前谷博, 江原暉将, 後藤功雄, 須藤克仁, 園尾聡, 綱川隆司, 中澤敏明, 二宮崇, 王向莉, 「特許機械翻訳の課題解決に向けた機械翻訳技術解説」, 自然言語処理 29 巻 3 号 pp. 925-985, 2022. https://doi.org/10.5715/jnlp.29.925

3.2 サーベイ論文の概要

本サーベイ論文では、訳抜け・過剰訳への対策、用語訳の統一、長文対策、低リソース言語対策、評価、翻訳高速化・省メモリ化、の 6 つの実用上の課題について、近年のニューラル機械翻訳関連技術を中心に 200 件あまりの文献調査の結果から関連が認められたものについてまとめている。ニューラル機械翻訳は非常に発展が速く、本活動を進める間にも次々に新しい技術が提案・公開されていく状況であったが、本論文では 2015 年前後のニューラル機械翻訳初期からの発展をカバーしつつ、特許機械翻訳という目標に向けたこれまでの技術の整理を行っている。ただし、ニューラル機械翻訳の基礎的な内容については触れていないため、適宜関連する教科書や基礎的な解説記事などを参照されたい。

3.3 おわりに － 活動のまとめ

本活動は新型コロナウィルス感染症の影響により研究会の活動がオンライン会議に移行する中で進められた。執筆・出版には当初の予定より多くの時間を要することとなったが、執筆メンバーが持つ様々な知見を持ち寄り、一つのサーベイ論文として完成に至ったことは喜ばしい。

最後に、本活動にご尽力いただいた執筆メンバーの皆様、研究会活動を支えていただいている Japio 他関係各機関の皆様に改めて感謝したい。

4. シンポジウム開催報告

4. シンポジウム開催報告：第 7 回特許情報シンポジウム

静岡大学　綱川　隆司
（シンポジウム副実行委員長）

4.1　開催概要

特許情報シンポジウムはアジア太平洋機械翻訳協会および日本特許情報機構の主催で 2010 年に第 1 回が開催されて以来ほぼ隔年で開催されており、今回で第 7 回を迎えた。2021 年に開催された第 6 回はコロナ禍の影響でオンライン形式となり、諸般の事情を考慮して今回もオンライン形式にて 2023 年 2 月 20 日（月）に開催となった。シンポジウムは本研究会の委員をプログラム委員として企画を行い、当日は綱川が副実行委員長として司会進行を務めた。事前登録者数は 175 名、当日の同時最大参加人数も 120 名程度に達し、イベントとしては成功裏に終了した。例年、特許情報に関連する研究者、実務に携わる方、および政府関係者による発表および議論を行っており、今回は合計 7 名の招待講演者からの講演、およびシンポジウム実行委員長の須藤先生からの特許機械翻訳に関する解説論文の紹介を行った。

4.2　開催概要

4.2.1　招待講演（1）

名和 大輔 様（日本特許庁 総務部 総務課 特許情報室 特許情報利用推進班 班長）
講演題目「特許庁における機械翻訳に関する取組」

中国における特許出願件数の増加傾向や日本国内からの海外出願率の増加といった近年の動向を背景に、特許の審査情報等の海外発信や海外出願の検索といった場面で機械翻訳は広く活用されている。日本特許庁における取り組みとして機械翻訳プラットフォームの構築・運用や WIPO、EPO による取り組み、機械翻訳品質の調査報告や、特許庁が提供する機械翻訳文の利用方法についてご紹介いただいた。

4.2.2　招待講演（2）

曹 竟成(Cao Jingcheng)様(中国特許庁 中国**专**利信息中心 (China Patent Information Center))

講演題目「多**语**种机器翻**译**系**统**：助力**专**利**审查**和**检**索 (An Advanced Multilingual MT System: Faciliating Patent Examination and Search)」

中国特許庁において特許情報の管理や機械翻訳などの情報処理を扱う中国特許情報センターの活動の一つとしてなされている多言語機械翻訳システムの技術について、これまで培ってきた特

許翻訳の知見、最新のニューラル翻訳技術、コーパス資源をもとに、外部翻訳メモリ・用語辞書、バッチ翻訳等の機能を備え、高い翻訳品質を達成していることが紹介された。

张 孝飞（Zhang Xiaofei）様（中国特許庁 知识产权出版社（Intellectual Property Publishing House Co., Ltd.））

講演題目「新一代人工智能专利机器翻译研究与应用（Research on New Generation of AI Patent Machine Translation System and Application thereof)」

中国の知識財産出版社における特許情報向けの機械翻訳サービスの最新技術に関してご紹介いただいた。最新技術である汎用のニューラル機械翻訳技術をベースに、豊富な特許翻訳文献とデータ資源をもとに統計翻訳、テンプレートベース翻訳およびルールベースの翻訳を組み込むことで、多言語かつ広範な専門用語を含む特許情報に対して業界トップレベルの翻訳品質を達成している。

4.2.3 招待講演（３）

高橋 博之 様（株式会社クロスランゲージ エンジニアリーダー)

講演題目「技術翻訳のための機械翻訳のカスタマイズ手法」

近年、飛躍的に向上した汎用の機械翻訳エンジンを技術翻訳に応用する際に生じる課題として、専門用語の誤訳や文体・言い回しの不自然さに対応するためのカスタマイズ手法について解説していただいた。代表的な 2 社のクラウド上での機械翻訳サービスを題材に、学習用対訳文の準備および環境設定から翻訳性能の評価までの一連の手順を示し、実際の性能向上効果が示された。

4.2.4 招待講演（４）

正林 真之 様（正林国際特許商標事務所 所長）

講演題目「特許情報の進歩的な活用方法」

特許や知的財産（IP）の情報をビッグデータとして扱い、従来型の特許調査の枠を超えて企業の経営、事業、研究開発等の戦略策定に知的財産やその他の情報を活用し、その企業の現在および未来の位置づけを提言する「IP ランドスケープ」についてご紹介いただき、商品やサービスといったビジネスとの関係など、マトリクス分析を始めマクロ・ミクロ両方からなる多面的な視点で情報を可視化する手法をご紹介いただいた。

4.2.5 招待講演（５）

後藤 功雄 様（NHK 放送技術研究所）

講演題目「ニュースを対象とした日英機械翻訳システムの研究開発」

NHK のテレビ放送やインターネットサービスで提供されている英語ニュースの原稿作成にも機械翻訳システムが用いられており、本発表ではニュース原稿の日英間での特徴の差異を踏まえ、単なる正確な対訳ではなく日本語ニュース記事から英語ニュース記事を生成するために行われた研究開発についてご紹介された。近年の翻訳モデルの性能維持に不可欠な高品質かつ大規模なニュース対訳文対の構築を行うとともに、それだけではカバーし切れない数字の翻訳や文脈利用等の具体的手法について解説された。

4.2.6　招待講演（６）

狩野 芳伸 様（静岡大学情報学部 准教授）

講演題目「自然言語処理の未解決課題～対話、法律、医療、政治分野での研究事例」

自然言語処理技術の応用分野として期待される対話システムや、法律文書処理などの法律分野、電子カルテ処理などの医療分野、議会議事録の分析といった政治分野など様々な研究例をご紹介いただき、現状残されている計算資源、出力結果の説明や根拠の提示、生成の一貫性、常識や世界知識の導入といった課題に対する部分的な取り組みや今後の方向性についての議論がなされた。

4.2.7　解説論文紹介

須藤 克仁 様（シンポジウム実行委員長・AAMT/Japio 特許翻訳研究会 副委員長、奈良先端科学技術大学院大学）

講演題目「特許機械翻訳の課題解決に向けた機械翻訳技術解説」

AAMT/Japio 特許翻訳研究会からの活動報告の一環として、研究会メンバーにより調査・執筆を行い 2022 年に公開された同名の解説論文について内容の紹介を行った。論文内では特許領域における機械翻訳活用の場面で問題になり得る課題として、訳抜け対策、用語訳の統一、長文対策、低リソース言語対策、性能評価、翻訳の高速化・省メモリ化のそれぞれの側面から現状の技術・研究と今後についての議論を展開している。

4.3　所感

本シンポジウムの開催と前後して、幅広い分野の質問に多様な言語で詳細な回答を生成できるチャットボットである ChatGPT が世間で高い注目を集めており、これまで限られた人が利用していたテキスト生成モデルが急速に一般に浸透する時期を迎えている。本シンポジウムで取り上げられた話題ではニューラル機械翻訳を中心にこれらの技術の中核部分が用いられているものの、今後、より大規模な特許情報データにより解決に向かう課題がある一方で、正確性や説明可能性などの課題は依然として残っており、今後の研究開発の進展が期待されるところである。

2022年度AAMT/Japio特許翻訳研究会報告書

発 行 日　2023 年 3 月

発　　行　一般財団法人　日本特許情報機構（Japio）
〒135−0016　東京都江東区東陽町 4 丁目1番 7 号
佐藤ダイヤビルディング
TEL：(03) 3615−5511　FAX：(03) 3615−5521

編　　集　一般社団法人　アジア太平洋機械翻訳協会（AAMT）

印　　刷　株式会社インターグループ