2023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
国際ワークショップPLST2023開催報告

2024 年 3 月

一般財団法人　日本特許情報機構

目次

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 フェロー
		東京大学大学院　名誉教授
		マンチェスター大学 教授
副委員長	須藤　克仁	奈良先端科学技術大学院大学 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
	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(JST)
	黒橋　禎夫	京都大学 特定教授
		国立情報学研究所 所長
	後藤　功雄	NHK 放送技術研究所
	小町　　守	一橋大学大学院ソーシャル・データサイエンス研究科 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	高　　京徹	株式会社高電社
	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 研究企画課
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

2023 年度 AAMT/Japio 特許翻訳研究会・活動履歴

2023 年 4 月 13 日
第 1 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2023 年 5 月 11 日
第 2 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2023 年 7 月 20 日
第 3 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2023 年 9 月 5 日
第 10 回特許・技術文書翻訳ワークショップ（PLST2023）
（於中国・マカオ開催）

2023 年 10 月 12 日
第 4 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2023 年 12 月 14 日
第 5 回 AAMT/Japio 特許翻訳研究会
（於ハイブリット開催）

2023 年 2 月 26 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 フェロー
マンチェスター大学 教授
辻井　潤一

機械翻訳、自然言語処理は、規則に基づく方式から始まり、統計的機械翻訳をへて現在のニューラル翻訳へと発展してきた。その長い機械翻訳の歴史では、言葉の理解とは一応切り離して、２つの言語の表現間の対応をどのようにとるかが研究の焦点となっていた。

この基本的な枠組みが、２０２２年１１月に発表された ChatGPT によって大きく変化しようとしている。ここ２年間足らずの期間で、GPT 以外にも大規模な言語モデルが次々に発表され、これらの大規模言語モデルが言葉による質問応答や対話において、人間と変わらない能力を示し始めている。大規模言語モデルは、初期の人工知能研究が目指したチューリングテストに合格する能力を示し、人間と同じ言語の理解能力をもったかのような動作をしめす。

大規模言語モデルは、さらに画像や音声などほかのモダリティを取り込む研究開発も活発化している。大規模言語モデルから始まったこの種の研究は、特定分野の知能ではなく、幅広い分野での知能を実現するという意味で、AGI（Artificial General Intelligence）の研究と称される。有名な人工知能研究者の何人かは、この AGI の研究が「人間の知能を超える Super-Intelligence を実現する」可能性とそれ内包する危険性を論じて、大きな話題となっている。

私は、AGI の議論は一種のハイプで、人間、あるいはそれを超える知能が出現しつつあるとは考えていないが、大規模言語モデルが人工知能技術の大きな変革点であり、機械翻訳においても、大きな変革をもたらすと考えている。従来の機械翻訳が原文との言語的な対応を保存する直訳であるのに対して、大規模言語モデルは相手言語での自然な表現により重点が置かれた、一意訳的な性質をもった翻訳を行う。同じ原文を読み手や翻訳の目的によって訳文の調子を変化させたり、完全ではないが、複数の翻訳が可能な場合には、訳し分けに必要な状況に関する情報を要求したりする。

一方で、大規模言語モデルは、質問応答において、真実とはかけ離れた、誤ったテキストを生成するという Hallucination という現象を示すことはよく知られている。自然なテキストは生成するが、事実ではないテキストを生成する。翻訳の場合も、自然なテキストではあるが、原文にはない情報を付け加えたり、逆に情報を落としてしまったりする。自然な訳文が出されることで、こういう原文からのずれを発見し修正する人間による後編集がよりコストがかかる作業になる。同じ原文をビジネス用に形式性の高い調子で翻訳したり、友人間の砕けたコミュニケーション用に翻訳したりと、従来の機械翻訳では不可能であったこと

が可能になるが、新たな問題も生じることになる。特に、専門性の高い分野の翻訳や大量のテキスト間の一貫性を保った翻訳の実現などは、今後の課題となっている。

大規模言語モデルとほかの外部リソースとを結合する RAG（Retrieval-Augmented Generation）といった技術も発展しており、大規模言語モデルの翻訳を過去の翻訳例を参考にする Example-based MT に近づけることも可能であろう。今後の数年間で、機械翻訳の分野が大きく変化する兆しがある。今後の本委員会の研究活動がより一層の成果を生むことを期待している。

2. 年間報告書

2.1 確率的生成モデルに基づくバイリンガルサブワード分割

愛媛大学　松井 大樹
愛媛大学　二宮　崇
同志社大学　田村 晃裕

2.1.1 はじめに

ニューラル機械翻訳（Neural Machine Translation、以下 NMT）(Sutskever et al. 2014; Bahdanau et al. 2015; Luong et al. 2015) では、予め指定した語彙に基づいて計算を行うため、翻訳時の原言語文に低頻度語や未知語が表れると翻訳性能が低下する。このような語彙の問題に対処するため、バイトペア符号化（Byte Pair Encoding、以下 BPE）(Sennrich et al. 2016) やユニグラム言語モデル (Kudo 2018) などによるサブワード分割が現在広く用いられている。BPE によるサブワード分割は事前トークナイズを要するのに対し、SentencePiece (Kudo and Richardson 2018) によるユニグラム言語モデルは生文からサブワード列に直接分割するため、日本語や中国語といった分かち書きされていない言語においても形態素解析器を必要としない。

しかしながら、これらの分割法は対訳関係を考慮せず、各言語ごとにサブワード分割を学習するため、機械翻訳タスクに適したサブワード分割になるとは限らない。例として、日英翻訳において、“nonextended”と「延長されなかった」という対訳対があるとする。この場合、“nonextended”は“no next end ed”などよりも“non extend ed”のほうが優れた分割であり、「延長されなかった」は「延　長　され　な　かった」などよりも「延長　されな　かった」のほうが優れた分割である。これは NMT が各サブワードの対訳関係“non”と「されな」、“extend”と「延長」、“ed”と「かった」を対応付けて学習できるためである。これらの問題を解決するために、対訳関係を考慮したバイリンガルサブワード分割（出口ら 2021; Hiraoka et al. 2021）が提案されている。しかし、出口ら (2021) のバイリンガルサブワード分割は原言語サブワード列と目的言語サブワード列のトークン長をそろえるものであり、トークン長が近いとはいえ各トークンがアライメント関係にあるとは必ずしも言えない。Hiraoka ら (2021) のバイリンガルサブワード分割は NMT モデルとサブワード分割モデルが一体化しており、利用する場合には同時に NMT モデルの学習が必要となり、サブワード分割および機械翻訳モデルの学習に大きなコストを要する。

本稿は、言語処理学会第 29 回年次大会において我々が提案したバイリンガルサブワード分割法 (松井ら 2023) について報告する。提案手法は、分かち書きされない言語を含む翻訳性能を改善するため、SentencePiece によるユニグラム言語モデル分割に基づいたサブワード列を得る。バイリンガルサブワード分割のための確率モデルを新たに定義し、原言語のサブワードと目的言語のサブワードが対となる確率を EM アルゴリズムを用いて求める。具体的に、提案手法は、ユニグラム言語モデルによって得られる原言語文と目的言語文それぞれの分割候補の組み合わせを求め、各対のサブワードのアライメント関係を取得し、ユニグラム言語モデルによる生起確率とアライメント確率を掛け合わせ、確率が最も大きくなるサブワード列対を選択する。提案手法を用いることで、原言語文と目的言語文のトークンの対訳関係が整うことになり、言語間でトーク

ンが１対１に対応付けされやすくなる。そのため、従来のサブワード分割法より NMT に適した分割が得られることが期待される。

本手法は原言語文と目的言語文の対訳アライメント関係を利用して分割するため、対訳コーパスが与えられる訓練時には問題は起きないが、原言語文しか与えられない翻訳時にはそのままではサブワード分割ができない。そこで提案手法では、EM アルゴリズムによって求められたアライメント確率の周辺化を行い、ユニグラム言語モデルによる生起確率と各原言語文サブワードの周辺確率を掛け合わせ、確率が最も大きくなるサブワード分割候補を選択する。

WAT Asian Scientific Paper Excerpt Corpus（以下、ASPEC）(Nakazawa et al. 2016) 英日・日英タスクにおいて、従来法と提案手法を用いた翻訳性能を比較したところ、Transformer NMT モデルの性能が改善した。

2.1.2 従来法

本節では提案手法の基礎となるユニグラム言語モデルに基づいたサブワード分割法（Kudo 2018; Kudo and Richardson 2018）について説明する。ユニグラム言語モデルでは、各サブワードが独立に生起すると仮定し、サブワード列の生起確率 $P_U(\boldsymbol{x})$を次式により表す。

（ユニグラム言語モデル）

$$P_U(\boldsymbol{x}) = \prod_{i=1}^{I} P(x_i)$$

$$s.t.\ x_i \in V\ (\forall i), \qquad \sum_{x \in V} P(x) = 1$$

ただし、$\boldsymbol{x} = (x_1, x_2, \cdots, x_I)$はサブワード列であり、$V$は語彙集合（サブワード辞書）である。各サブワードの生起確率 $P(x_i)$ は EM アルゴリズムによって周辺尤度 L_{lm} を最大化することにより推定される。

$$L_{lm} = \sum_{s=1}^{|D|} \log P(X_s) = \sum_{s=1}^{|D|} \log \sum_{\boldsymbol{x} \in S(X_s)} P_U(\boldsymbol{x})$$

ただし、D は対訳コーパスであり、X_s は D の中の s 番目の原言語文または目的言語文であり、$S(X_s)$ は X_s の分割候補集合である。

生起確率が最大となるサブワード列（最尤解）は次式によって得られる。

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in S(X)}{\operatorname{argmax}} P_U(\boldsymbol{x})$$

ただし、X は原言語文または目的言語文である。また、k-best 分割候補も X に対するユニグラム言語モデルによって計算される確率 $P_U(\boldsymbol{x})$ に基づいて得ることが出来る。ただし、サブワード列の生起確率は各サブワードの尤度の積の形で表されるため、系列長の短い（トークン数の少ない）サブワード列が高い確率を持つ傾向がある。

SentencePiece におけるユニグラム言語モデルを用いたサブワード分割は生文から直接学習できるため、日本語や中国語といった分かち書きされない言語においても単語分割器や形態素解析

器を必要とせずに分割できるという特徴がある。

2.1.3 提案手法

本節では、提案手法となるバイリンガルサブワード分割のための EM アルゴリズムについて説明する。まず、サブワード文対の確率モデルの定義を与え、その後、EM アルゴリズムによるパラメータ更新式の導出、対訳コーパスのサブワード分割を行う手法、翻訳時のサブワード分割を行う手法について説明する。提案手法は NMT モデルや訓練法を修正する必要はなく、従来のサブワード分割法を置き換えるだけで適用可能である。

確率モデルは、ユニグラム言語モデルが出力するサブワード列の生起確率と原言語サブワードと目的言語サブワードのアライメント確率の積で与えられる。ただし、各原言語サブワード列と目的言語サブワード列及びそれらのアライメントは明示的に与えられておらず、隠れ状態となっている。そのため、潜在変数付き確率モデルの学習として有名な EM アルゴリズムを用いて、アライメント確率を学習する。次に、最も確率の高いサブワード列のアライメントを選択することで、訓練コーパスのサブワード分割を行う。

NMT の訓練時には対訳コーパスを利用できるが、翻訳時には対訳文が存在しない。そこで、原言語側サブワードのみを参照する周辺確率を求めることで、原言語文のサブワード分割を行う。

2.1.3.1 提案手法の確率モデル

まず原言語文 X と目的言語文 Y が与えられたとき、提案手法における確率モデルを次のように定義する。

$$P(X,Y)=\sum_{\boldsymbol{x}\in S(X)}\sum_{\boldsymbol{y}\in S(Y)}P_M(\boldsymbol{x},\boldsymbol{y})\approx\sum_{k=1}^{K}\sum_{l=1}^{L}P_M\left(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(l)}\right)$$

ただし、X に対するサブワード分割候補 $S(X)$ のうち、サブワード生起確率 $P_U(\boldsymbol{x})$ が高い top-K 個をそれぞれ$\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(k)},\cdots,\boldsymbol{x}^{(K)}$、$Y$ に対するサブワード分割候補 $S(Y)$ のうち、サブワード生起確率 $P_U(\boldsymbol{y})$ が高い top-L 個をそれぞれ$\boldsymbol{y}^{(1)},\cdots,\boldsymbol{y}^{(l)},\cdots,\boldsymbol{y}^{(L)}$とする。また、$P_M$ は原言語文のサブワード列 $\boldsymbol{x}$ と目的言語文のサブワード列 $\boldsymbol{y}$ に対する確率モデルであり、次式で定義する。

$$P_M(\boldsymbol{x},\boldsymbol{y})=P_U(\boldsymbol{x})P_U(\boldsymbol{y})\prod_{u,v\in A(\boldsymbol{x},\boldsymbol{y})}\alpha_{uv}$$

ただし、$A(\boldsymbol{x},\boldsymbol{y})$ は原言語のサブワード列 $\boldsymbol{x}$ と目的言語のサブワード列 $\boldsymbol{y}$ の間のアライメントを返す関数であり、アライメントは対応するサブワード対の集合とする。また、α_{uv} は原言語側サブワード u と目的言語側サブワード v が対応する確率である。

2.1.3.2 アライメント確率 α_{uv} の算出

ユニグラム言語モデル P_U とアライメントを返す関数 A は所与のものとして、EM アルゴリズムを用いてアライメント確率 α_{uv} を求める。

$$P_M^{old}(\boldsymbol{x},\boldsymbol{y}) = P_U(\boldsymbol{x})P_U(\boldsymbol{y}) \prod_{u,v\in A(\boldsymbol{x},\boldsymbol{y})} \alpha_{uv}^{old}$$

$$P_M^{new}(\boldsymbol{x},\boldsymbol{y}) = P_U(\boldsymbol{x})P_U(\boldsymbol{y}) \prod_{u,v\in A(\boldsymbol{x},\boldsymbol{y})} \alpha_{uv}^{new}$$

$$Q = \sum_n \sum_k \sum_l \frac{P_M^{old}\left(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)}\right)}{\sum_{k'}\sum_{l'} p_M^{old}\left(\boldsymbol{x}_n^{(k')},\boldsymbol{y}_n^{(l')}\right)} \log p_M^{new}\left(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)}\right)$$

α_{uv}^{new} に関して Q 関数を最大化することにより、α_{uv}^{new} の更新式を得る。

$$\alpha_{uv}^{new} = \frac{\sum_n \sum_k \sum_l E_{nkluv}}{\sum_{u''}\sum_{v''}\sum_n \sum_k \sum_l E_{nklu''v''}}$$

$$E_{nkluv} = \frac{P_M^{old}\left(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)}\right)}{\sum_{k'}\sum_{l'} p_M^{old}\left(\boldsymbol{x}_n^{(k')},\boldsymbol{y}_n^{(l')}\right)} C_{nkluv}$$

ただし、C_{nkluv} は、n 番目の文対における原言語のサブワード列 $\boldsymbol{x}_n^{(k)}$ と目的言語のサブワード列 $\boldsymbol{y}_n^{(l)}$ に対し、サブワード u と v がアライメント関係となっている回数である。

2.1.3.3 訓練データのサブワード分割

訓練データ D の各文対 X,Y に対して、次式に従ってサブワード列 $\boldsymbol{x}^{(\hat{k})},\boldsymbol{y}^{(\hat{l})}$ を求め、サブワード文対として採用する。

$$\hat{k},\hat{l} = \underset{k,l}{\operatorname{argmax}} P_M\left(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(l)}\right)$$

2.1.3.4 翻訳時のサブワード分割

翻訳時におけるサブワード分割では、アライメント確率を目的言語側サブワードで周辺化することによって原言語側サブワードの確率を求める。テストデータの各文 X に対して、次式に従ってサブワード列 $\boldsymbol{x}^{(\hat{k})}$ を求め、サブワード文として採用する。

$$\alpha'_u = \sum_{v\in V_{target}} \alpha_{uv}$$

$$\hat{k} = \underset{k}{\operatorname{argmax}} P_{M'}\left(\boldsymbol{x}^{(k)}\right)$$

$$P_{M'}\left(\boldsymbol{x}^{(k)}\right) = P_U(\boldsymbol{x}) \prod_{u\in\boldsymbol{x}} \alpha'_u$$

ただし、V_{target} は目的言語側のサブワード集合である。

2.1.4 実験

2.1.4.1 実験設定

提案手法と従来法（ユニグラム言語モデル）の翻訳性能を比較した。ユニグラム言語モデルに

表 2.1.1: ASPEC 英-日における翻訳性能の比較（BLUE（%））

	英日	日英
ユニグラム言語モデル（従来手法）	27.4	26.7
バイリンガルサブワード分割（提案手法）	**27.8**	**27.3**

よるサブワード列候補集合を得るために、SentencePiece[1] (Kudo 2018; Kudo and Richardson 2018) を利用した。原言語側サブワードと目的言語側サブワードのアライメントを得るために、fast_align[2] (Dyer et al. 2013) を利用した。NMT には Fairseq (Ott et al. 2019) を使用し、Transformer base (Vaswani et al. 2017) モデルを利用した。翻訳性能を評価するために、sacreBLEU (Post 2018) を利用した。sacreBLEU の日本語のトークナイズには ja-mecab (Kudo et al. 2004) を、英語のトークナイズには 13a を利用した。

2.1.4.2 データセットとハイパーパラメータ

データセットには WAT ASPEC 英日・日英翻訳タスク [3]を用いた。NMT の訓練には訓練データのうち、100 万文対(train-1.txt)を利用した。開発データとテストデータのデータ数はそれぞれ 1,790、1,812 文対であった。ユニグラム言語モデルの学習は、原言語側と目的言語側で独立して行い、辞書サイズはどちらも 16,000 に設定した。候補数は原言語側と目的言語側それぞれユニグラム言語モデルによるサブワード生起確率が高い上位 10 通り（top-k=top-l=10）とした。すべての Transformer base モデルにおいて、パラメータの最適化には adam (Kingma and Ba 2014)、学習率は 1e-4、バッチサイズは 128 とし、その他のパラメータは Fairseq のデフォルトのままとした。学習は 30 エポックで終了させ、各エポックのモデルのうち、開発データ上で最も性能のよかったものを利用してテストデータの翻訳を行った。実験はランダムシードを変えて 2 度行い、その平均を実験結果とした。

2.1.4.3 実験結果

実験結果の BLEU スコア（Papineni et al. 2002）を表に示す。表から分かる通り、バイリンガルサブワード分割は英日、日英翻訳の両言語方向において、ユニグラム言語モデルより性能が改善されている。バイリンガルサブワード分割を用いることでユニグラム言語モデルに対し、英日・日英翻訳においてそれぞれ 0.4 BLEU ポイント、0.6 BLEU ポイントの性能改善が確認された。

2.1.4.4 考察

提案手法によるサブワード分割について考察する。バイリンガルサブワード分割を用いることで良く翻訳できた例を表に示す。提案手法は従来手法よりも原言語サブワードが出力結果に対応するように分割されていることから、正確に翻訳できたと考えられる。バイリンガルサブワード

[1] https://github.com/google/SentencePiece
[2] https://github.com/clab/fast_align
[3] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

表 2.1.2: 提案手法の良い例

従来法	
原言語サブワード	_A k at uki _disease
出力結果	アキスキー病
提案手法	
原言語サブワード	_A ka tu ki _disease
出力結果	アカ ツ キ病
正解データ	
	アカツキ病

表 2.1.3: 提案手法の悪い例

従来法	
原言語サブワード	_a _expressway
出力結果	高速道路
提案手法	
原言語サブワード	_a _express way
出力結果	急 行 路
正解データ	
	高速道路

分割を用いることで悪くなった例を表に示す。提案手法は従来手法よりも細かく分割されており、“express”と「急行」、“way”と「路」を関連付けるよう学習していることから、“express way”を「急行 路」と不正確に翻訳したと考えられる。

2.1.5 まとめ

本論文では、確率モデルと EM アルゴリズムを用いたニューラル機械翻訳のための新たなサブワード分割法を提案した。アライメント確率を導入したバイリンガルサブワード分割のための確率モデルの定義を与え、そのアライメント確率を求める EM アルゴリズムを導出した。訓練コーパスに対するサブワード分割は、提案する確率モデルを用いて、最も確率の高いサブワード列対を選択することで実現した。翻訳時には、目的言語側のサブワード列が得られないため直接提案モデルを適用することはできないが、アライメント確率を周辺化することで、原言語側だけで定義される確率モデルを与えた。実験の結果、WAT ASPEC 英日・日英翻訳タスクにおいて、Transformer NMT モデルの性能が改善し、提案手法の有効性を確認した。

今後の課題として、条件付き確率を導入することでアライメントの方向性を考慮した確率モデルに拡張することや、英日以外の言語対での実験等が挙げられる。

謝辞

本稿は、言語処理学会 第 29 回年次大会で発表した論文（松井ら 2023）に基づいて、その論文を再構成し、解説したものである。

これらの研究成果は、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）および JSPS 科研費 JP21K12031 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

D. Bahdanau, K. Cho and Y. Bengio. (2015). Neural Machine Translation by Jointly Learning to Align and Translate, in Proceedings of ICLR.

C. Dyer, V. Chahuneau and N. A. Smith. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, pp. 644–648.
T. Hiraoka, S. Takase, K. Uchiumi, A. Keyaki and N. Okazaki. (2021). Joint Optimization of Tokenization and Downstream Model. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 244–255.
D. P. Kingma and J. Ba. (2015). Adam: A Method for Stochastic Optimization, in Proceedings of the 3rd International Conference on Learning Representations.
T. Kudo, K. Yamamoto and Y. Matsumoto. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 230–237.
T. Kudo. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 66–75.
T. Kudo and J. Richardson. (2018). 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.
T. Luong, H. Pham and C. D. Manning. (2015). Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421.
T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi and H. Isahara. (2016). ASPEC: Asian Scientific Paper Excerpt Corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2204–2208.
M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier and M. Auli. (2019). 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.
M. Post. (2018). A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186–191.
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.
R. Sennrich, B. Haddow and A. Birch. (2016). 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.
I. Sutskever, O. Vinyals and Q. V. Le. (2014). Sequence to Sequence Learning with Neural Networks, in Proceedings of NeurIPS.
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.
出口 祥之, 内山 将夫, 田村 晃裕, 二宮 崇, 隅田 英一郎. (2021). ニューラル機械翻訳のためのバイリンガルなサブワード分割. 自然言語処理, Vol. 28, No. 2, pp. 632-650.
松井 大樹, 二宮 崇, 田村 晃裕. (2023). バイリンガルサブワード分割のための EM アルゴリズム, 言語処理学会 第 29 回年次大会 発表論文集, pp. 1469–1473.

2.2 タグ付けモデルと埋め込みモデルのアンサンブルに基づき固有表現を活用するニューラル機械翻訳モデル

同志社大学　南端　尚樹
田村　晃裕
加藤　恒夫

2.2.1 はじめに

ニューラルネットワークに基づく機械翻訳（Neural Machine Translation: NMT）の性能改善を行う研究の流れの一つとして，原言語文や目的言語文の単語の品詞や文構造などの言語学的素性を活用する試みが行われている．その中で，言語学的素性として，人名や地名，組織名といった特定の表現を表す固有表現（Named Entity: NE）に着目し，NMT において NE 情報を活用する研究が行われている[6, 12, 13, 16]．NE には複合語が多く存在するため，NE の情報を NMT に与えることで単語のチャンク情報を翻訳に活用できる．また，NE の種類の情報は，多義語を翻訳する際の語義曖昧性解消に役立つことが報告されている[13]．

NMT において NE 情報を活用する方法として，NE の種類と開始／終了情報を含む NE タグを文中の NE の前後に挿入する「タグ付けモデル」[16]や，NE 埋め込みを単語埋め込みに組み込む「埋め込みモデル」[12, 13]が提案されている．NE 情報を活用する NMT の初期の研究では原言語文の NE 情報が活用され，目的言語文の NE 情報は活用されていなかったが，近年では，原言語文の NE に加えて目的言語文の NE の情報も活用することで翻訳性能が向上することが報告されている[12, 16]．

NE を活用する NMT において，推論時に複数のモデルの出力を統合するアンサンブルによって翻訳性能が改善することが報告されている．Nowakowski ら[6]は，原言語文の NE 情報を活用する 3 つの埋め込みモデルをアンサンブルするモデルを提案し，翻訳性能を改善している．しかし，NE を活用する NMT における従来のアンサンブルでは目的言語文の NE は活用されておらず，また，埋め込みモデル以外を用いたアンサンブルは研究されていない．

そこで本研究では，NE を活用する NMT の翻訳性能を改善するため，原言語文と目的言語文の両方の NE を活用するタグ付けモデルと埋め込みモデルのアンサンブルに基づく NMT モデルを提案する．提案モデルの概要を図 1 に示す．提案モデルではタグ付けモデル及び埋め込みモデルを独立に学習する．そして推論時に，図 1 のように，学習した二つのモデルによる出力確率を平均した確率に基づき目的言語文を生成する．

WMT2014 の英語とドイツ語間の翻訳タスク[2]及び WMT2020 の英語と日本語間の翻訳タスク[1]において提案モデルを評価した結果，アンサンブルにより英独翻訳では最大 1.13 ポイント，独英翻訳では最大 1.62 ポイント，英日翻訳では最大 2.08 ポイント，日英翻訳では最大 1.24 ポイント BLEU が向上することを確認した．

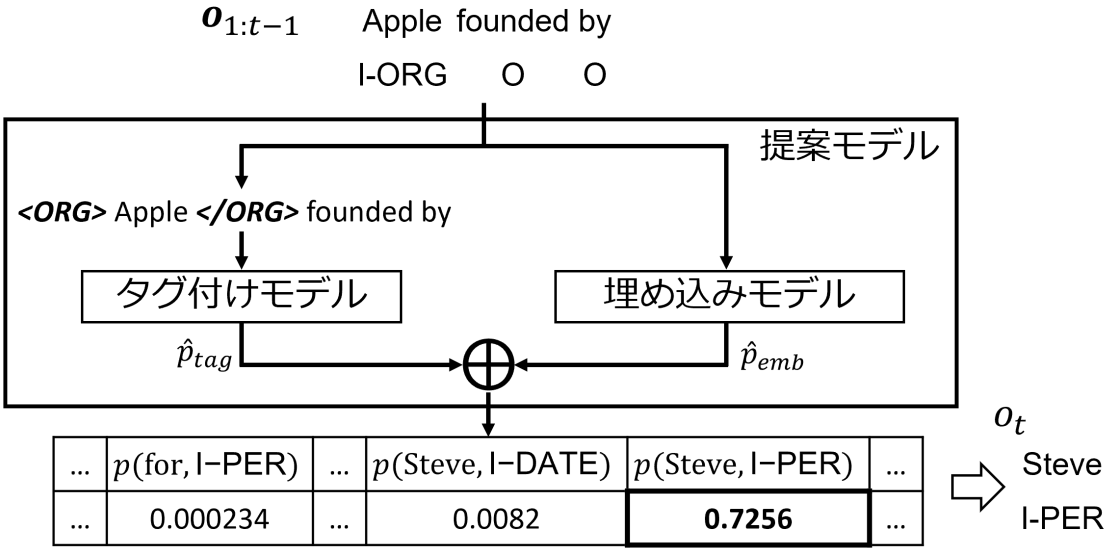


図 1: 提案モデルの概要図

2.2.2 従来研究

本節では，NE 情報を活用する従来の NMT モデルのうち，提案モデルのベースとなるタグ付けモデル[16]と埋め込みモデル[12]について説明する．これら二つの NMT モデルは Transformer[14]を改良したモデルであり，原言語文と目的言語文の両方の NE を活用している．

2.2.2.1 タグ付けモデル

タグ付けモデルは，原言語文の NE と目的言語文の NE の前後に，NE の種類と開始／終了の情報を含む NE タグを挿入し，NE タグが挿入された文に基づいて学習や翻訳を行う．タグ付けモデルの概要を図 2 に示す.

学習時には，まず，教師データの原言語文及び目的言語文に対して NER を行い，文中の NE を特定する．そして，特定した NE の前に「<c>」という NE タグを挿入し，後に「</c>」という NE タグを挿入する．ここで，cは NE の種別を表す NE クラスである．図 2 では，原言語文である日本語文中の「スティーブ・ジョブズ」が人名（PER）の NE,「アップル」が組織名（ORG）の NE と認識され,「スティーブ・ジョブズ」と「アップル」の前後に，それぞれ,「<PER>」と「</PER>」,「<ORG>」と「</ORG>」という NE タグが挿入されている．また，目的言語文である英語文も同様に,「Steve Jobs」と「Apple」が，それぞれ人名と組織名の NE として認識され,「Steve Jobs」の前後には「<PER>」と「</PER>」,「Apple」の前後には「<ORG>」と「</ORG>」が挿入されている．そして，NE タグが挿入された対訳文対を用いて NMT モデルを学習する．

推論時には，まず原言語文$\boldsymbol{x}$に対して，学習時と同様に NER を行い，原言語文内の NE の前後に NE タグを挿入する．そして，NE タグが挿入された原言語文$\boldsymbol{x}'$を，学習済みの NMT モデルで翻訳する．具体的には，各時刻tにおいて Transformer デコーダの出力層で単語及び NE タグの出力確率$p_{tag}(y'_t|\boldsymbol{x}', \boldsymbol{y}'_{1:t-1})$を算出し，算出した出力確率に基づいて目的言語文を生成する．ここで，y'_tは時刻tでのタグ付けモデルの出力であり，$y'_t \in V \cup T_{tag}$である．ただし，Vは目的言語の単語集合，T_{tag}は目的言語の NE タグの集合である．つまり，y'_tは目的言語の単語あるいは NE

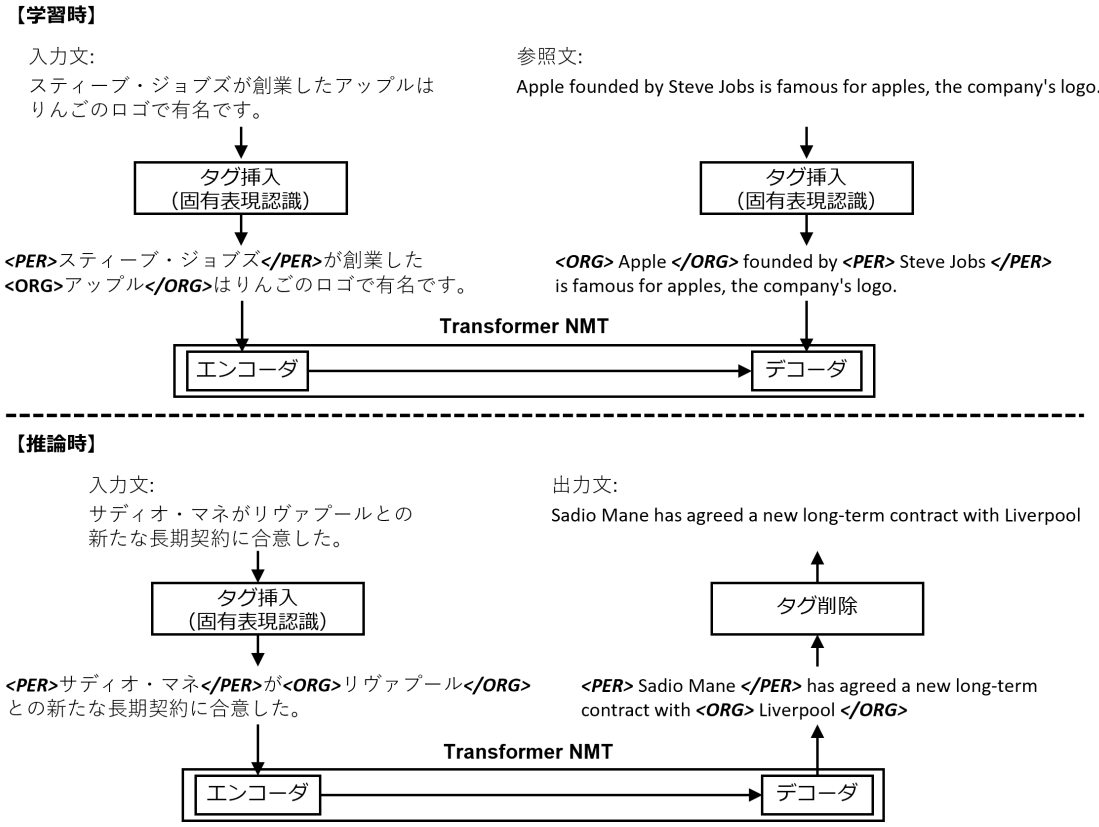


図 2: タグ付けモデルの概要図

タグである．また，$\boldsymbol{y}'_{1:t-1}$は時刻1から$t-1$までの出力系列を表す．翻訳の結果，生成文が NE タグを含む場合は NE タグを削除し，NE タグのない文を目的言語文として出力する．生成文が NE タグを含まない場合は，生成文をそのまま目的言語文として出力する．

2.2.2.2 埋め込みモデル

埋め込みモデルでは，通常の Transformer NMT モデル[14]のエンコーダとデコーダに，NE の情報を埋め込む「NE 埋め込み層」と，単語埋め込み層の出力に NE 埋め込み層の出力を足し合わせる「加算層」が追加されている．さらに，デコーダには，出力文の単語の NE を予測する「NE 出力層」が追加されている．埋め込みモデルの概要を図 3 に示す．図 3 の埋め込みモデルでは，NE 情報として各単語の IO タグを用いている．原言語側の IO タグは原言語文の NER の結果から獲得し，目的言語側の IO タグは NE 出力層の出力から獲得する．

埋め込モデルでは，まず，単語と IO タグを単語埋め込み層と NE 埋め込み層に送り，単語埋め込みと NE 埋め込みを生成する．そして，加算層で単語埋め込みと NE 埋め込みを足し合わせ，単語と NE の両方の情報を含む埋め込みに基づいて翻訳を行う．具体的には，各時刻tにおいて，単語出力層で単語y_tの出力確率$p_{emb}(y_t|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_{1:t-1})$を算出すると共に，NE 出力層で IO タグ$l_t$の出力確率$p_{emb}(l_t|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_{1:t-1})$を算出し，時刻$t$における出力単語とその IO タグのペア$\tilde{y}_t=(y_t,l_t)$を，以下の確率に基づいて予測する．ただし，$\tilde{\boldsymbol{x}}$は原言語側の単語とその IO タグのペアの系列を表す．

$$p_{emb}(y_t,l_t|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_{1:t-1})=p_{emb}(y_t|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_{1:t-1})\times p_{emb}(l_t|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_{1:t-1}) \quad (1)$$

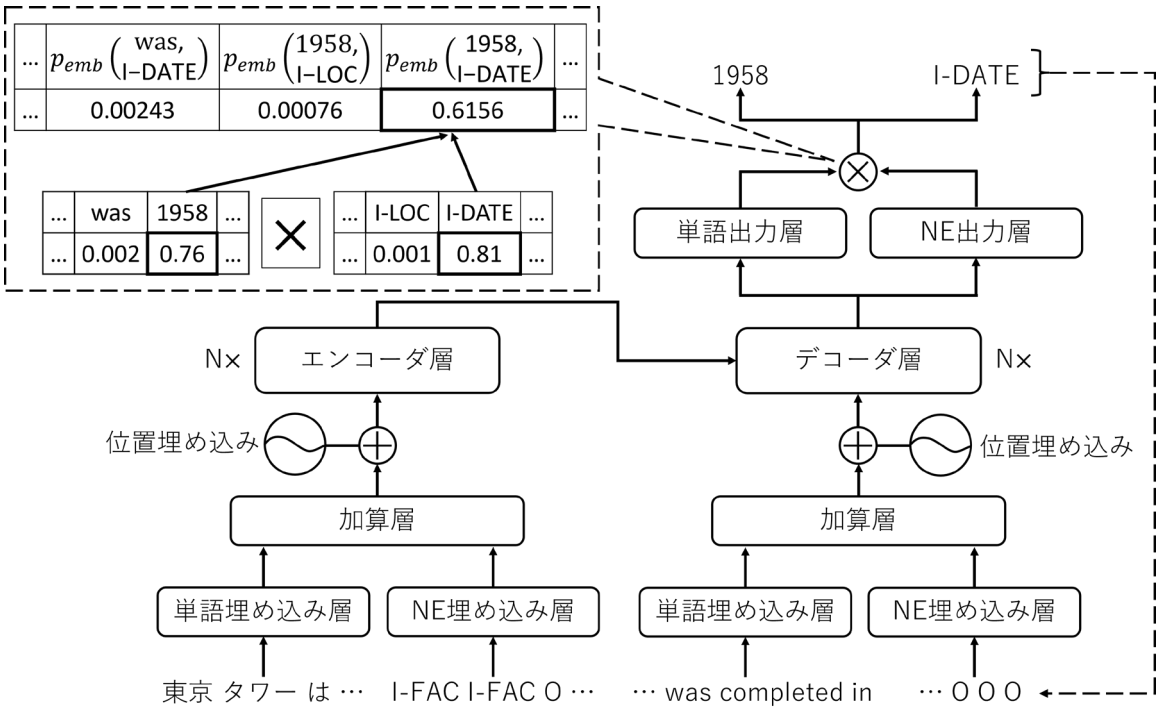


図 3: 埋め込みモデルの概要図

2.2.3 提案モデル：アンサンブルモデル

本節では，原言語文と目的言語文の両方の NE を活用するタグ付けモデルと埋め込みモデルのアンサンブルに基づく NMT モデルを提案する．提案モデルは，推論時に，独立に学習したタグ付けモデルと埋め込みモデルを用いて目的言語文を生成する．目的言語文の生成は，図 1 に示すように，タグ付けモデルの出力確率と埋め込みモデルの出力確率を平均した確率に基づいて目的言語の単語と NE クラスを予測することで行う．具体的には，各時刻tにおいてタグ付けモデルと埋め込みモデルを用いて，目的言語の単語（$y_t \in V$）と NE（$c_t \in C \cup \{O\}$）のペア$o_t = (y_t, c_t)$の出力確率$p(o_t|\boldsymbol{x}', \tilde{\boldsymbol{x}}, \boldsymbol{o}_{1:t-1})$を式（2）の通りに求める．ただし，$C$は NE クラスの集合であり，「O」は NE ではないことを表すクラスである．

$$p(o_t|\boldsymbol{x}', \tilde{\boldsymbol{x}}, \boldsymbol{o}_{1:t-1}) = \frac{\hat{p}_{emb}(o_t|\tilde{\boldsymbol{x}}, \boldsymbol{o}_{1:t-1}) + \hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})}{2} \quad (2)$$

ここで，$\hat{p}_{emb}(o_t|\tilde{\boldsymbol{x}}, \boldsymbol{o}_{1:t-1})$と$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$はそれぞれ，埋め込みモデルに基づく$o_t$の出力確率とタグ付けモデルに基づく$o_t$の出力確率を表す．そして，求めた出力確率に基づき，式（3）の通りに目的言語文$\hat{\boldsymbol{y}}$を生成する．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(o_t|\boldsymbol{x}', \tilde{\boldsymbol{x}}, \boldsymbol{o}_{1:t-1}) \quad (3)$$

ここで，Tは目的言語文の文長である．以降では，2.2.3.1 節で埋め込みモデルに基づく$\hat{p}_{emb}$の算出方法を説明し，2.2.3.2 節でタグ付けモデルに基づく$\hat{p}_{tag}$の算出方法について説明する．

2.2.3.1 $\hat{p}_{emb}$の算出方法

2.2.2.2 節で説明した通り，埋め込みモデルは各時刻において目的言語の単語とその NE の IO タグのペア$\tilde{y}_t = (y_t, l_t)$に対する確率を予測する．NE の IO タグl_tと NE クラスc_tは 1 対 1 に対応するので，提案モデルでは，埋め込みモデルを用いて時刻tにおける目的言語の単語と NE のペア$o_t = (y_t, c_t)$の出力確率$\hat{p}_{emb}(o_t|\tilde{\boldsymbol{x}}, \boldsymbol{o}_{1:t-1})$を式（4）の通りに求める．

$$\hat{p}_{emb}(o_t|\tilde{\boldsymbol{x}}, \boldsymbol{o}_{1:t-1}) = p_{emb}(y_t, l_t|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_{1:t-1}) \quad (4)$$

2.2.3.2 $\hat{p}_{tag}$の算出方法

タグ付けモデルは，NE タグと目的言語の単語からなる系列を逐次的に生成する．したがって，$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$の算出の際には，まず，時刻$t-1$までの目的言語の単語とその NE クラスのペアの系列$\boldsymbol{o}_{1:t-1}$を，タグ付けモデルの入力となる目的言語の単語と NE タグからなる系列$\boldsymbol{y'}_{1:t'-1}$へ変換する（ステップ 1）．その後，$\boldsymbol{y'}_{1:t'-1}$に続く各要素（NE タグと目的言語の単語）の出力確率を算出し，算出した確率に基づき$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$を求める（ステップ 2）．以降で，各ステップについて説明する．

ステップ 1：$\boldsymbol{o}_{1:t-1}$から$\boldsymbol{y'}_{1:t'-1}$への変換

本ステップでは，$\boldsymbol{o}_{1:t-1}$中の目的言語の単語系列$\boldsymbol{y}_{1:t-1}$において，隣接する単語間で NE が異なる箇所に NE タグを挿入することで$\boldsymbol{y'}_{1:t'-1}$に変換する．つまり，$c_1 \neq 0$または$c_{i-1} \neq c_i (i = 2,3,\dots,t-1)$となるところに NE タグを挿入する．具体的には，次の（ⅰ）から（ⅳ）のタグ挿入により$o_{1:t-1}$を$y'_{1:t'-1}$に変換する．

（ⅰ）$c_1 \neq 0$の時：開始タグ<c_1>をy_1の直前に挿入．

（ⅱ）$c_{i-1} = 0, c_i \in C$の時：開始タグ<c_i>をy_{i-1}とy_iの間に挿入．

（ⅲ）$c_{i-1} \in C, c_i = 0$の時：終了タグ</c_{i-1}>をy_{i-1}とy_iの間に挿入．

（ⅳ）$c_{i-1} \in C, c_i \in C, c_{i-1} \neq c_i$の時：終了タグ</$c_{i-1}$>と開始タグ<$c_i$>を$y_{i-1}$と$y_i$の間に挿入．

ステップ 2：$\hat{p}_{tag}$の計算

本ステップでは，$\boldsymbol{y'}_{1:t'-1}$をタグ付けモデルへ入力し，$\boldsymbol{y'}_{1:t'-1}$に続く要素に対するタグ付けモデルの出力確率$p_{tag}(y'_{t'}|\boldsymbol{x}', \boldsymbol{y'}_{1:t'-1})$に基づいて$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$を計算する．その際，時刻$t'$におけるタグ付けモデルの出力要素が NE タグである場合，o_tの出力単語を決定することができない．そこで，タグ付けモデルの出力要素が初めて目的言語の単語になるまで予測を継続し，$\hat{p}_{tag}$を計算する．以降では，t'時点において全ての NE タグが閉じている場合（NE タグが出現していない，あるいは全ての開始タグに対応する終了タグが出現している場合）と，t'時点において NE タグが開いている場合（開始タグに対応する終了タグが出現していない場合）に分けて，$\hat{p}_{tag}$の計算方法を詳細に説明する．

なお，タグ付けモデルにおいて NE タグを含む出力系列は，同じクラスの開始タグと終了タグが対となる（条件 1），対となる開始タグと終了タグの間は NE タグを含まない単語列となる（条件 2）の 2 つの条件を満たす必要がある．そこで，これらの条件を満たさない出力に対しては確率を 0 として無視する．そして，これにより確率の総和が 1 にならない場合は，その差分を条件を満たす出力の確率に均等に割り振ることで，$\hat{p}_{tag}$の総和が 1 となるように補正する．

ケース 1：時刻t'においてタグが閉じている場合

タグ付けモデルの出力は，単語（w），開始タグ（<c>），終了タグ（</c>）のいずれかである．ただし，$w \in V, c \in C$である．ここで，時刻t'の出力が</c>である場合は，出力系列の条件 1 に反するため無視する．

ケース 1-1：$y'_{t'}$が単語wの場合

時刻t'においてタグが閉じているので，$y'_{t'}$の単語wの NE クラスは O である．したがって，$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$は式（5）の通りとなる．

$$\hat{p}_{tag}(o_t = (w, \mathrm{O})|\boldsymbol{x}', \boldsymbol{o}_{1:t-1}) = p_{tag}(y'_{t'} = w|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}) \quad (5)$$

ケース 1-2：$y'_{t'}$が開始タグ<c>の場合

時刻t'の出力が開始タグの場合，o_tの単語は$y'_{t'}$だけでは決まらない．そこで，$y'_{t'}$の出力確率に加えて，次の時刻$t'+1$における要素$y'_{t'+1}$の出力確率も用いて$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$を算出する．その際，時刻$t'+1$におけるタグ付けモデルの出力のうち，開始タグと終了タグの場合は出力系列の条件に反するため，出力が単語の場合のみを考える．$y'_{t'}$が開始タグ<c>で，$y'_{t'+1}$が単語wの場合，wの NE クラスはcとなる．したがって，$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$は式（6）の通りとなる．

$$\hat{p}_{tag}(o_t = (w, c)|\boldsymbol{x}', \boldsymbol{o}_{1:t-1}) = p_{tag}(y'_{t'} =< c >|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}) \times p_{tag}(y'_{t'+1} = w|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}, y'_{t'} =< c >) \quad (6)$$

$\hat{p}_{tag}$の補正

前述の$\hat{p}_{tag}$の算出過程では，タグ付けモデルの出力系列の条件に反する出力は無視されるため，全てのo_tに対する$\hat{p}_{tag}$の合計が 1 にならない場合がある．そこで，式（5）と式（6）で求めた$\hat{p}_{tag}$を，式（7）及び式（8）に示すように補正し，補正後の確率$\tilde{p}_{tag}$を式（2）の$\hat{p}_{tag}$として用いる．

$$\delta = \frac{1 - \sum_{o_t \in V \times \{C \cup \{\mathrm{O}\}\}} \hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})}{|V| \times |\{C \cup \{\mathrm{O}\}\}|} \quad (7)$$

$$\tilde{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1}) = \hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1}) + \delta \quad (8)$$

ケース 2：時刻t'においてタグ<c>が開いている場合（終了タグ</c>が未出力の場合）

ケース 2 においては，時刻t'のタグ付けモデルの出力のうち，開始タグまたはc以外の NE クラスの終了タグの場合は，出力系列の条件に反するため無視する．

ケース 2-1：$y'_{t'}$が単語wの場合

時刻t'において開始タグ<c>が開いているので，$y'_{t'}$の単語wの NE クラスはcである．したがって，$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$は式（9）の通りとなる．

$$\hat{p}_{tag}(o_t = (w, c)|\boldsymbol{x}', \boldsymbol{o}_{1:t-1}) = p_{tag}(y'_{t'} = w|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}) \quad (9)$$

ケース 2-2：$y'_{t'}$が終了タグ</c>の場合

o_tの単語は$y'_{t'}$だけでは決定できないため，次の時刻$t'+1$における要素$y'_{t'+1}$の出力確率も用いて$\hat{p}_{tag}(o_t|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$を算出する．$y'_{t'}$が終了タグ</$c$>の場合，時刻$t'+1$では全ての開始タグが閉じているため，時刻$t'+1$における$\hat{p}_{tag}$の算出は，前述のケース 1 と同様に考えることができる．そのため，$y'_{t'+1}$が単語の場合の$\hat{p}_{tag}$は式（10）に示すように計算され，$y'_{t'+1}$が開始タグで$y'_{t'+2}$が

単語の場合の$\hat{p}_{tag}$は式（11）に示すように計算される．

$$\hat{p}_{tag}(o_t = (w, 0)|\boldsymbol{x}', \boldsymbol{o}_{1:t-1}) = p_{tag}(y'_{t'} = </c>|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}) \times p_{tag}(y'_{t'+1} = w|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}, y'_{t'} = </c>) \quad (10)$$

$$\hat{p}_{tag}(o_t = (w, c')|\boldsymbol{x}', \boldsymbol{o}_{1:t-1}) = p_{tag}(y'_{t'} = </c>|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}) \times p_{tag}(y'_{t'+1} = < c' >|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}, y'_{t'} = </c>) \times p_{tag}(y'_{t'+2} = w|\boldsymbol{x}', \boldsymbol{y}'_{1:t'-1}, y'_{t'} = </c>, y'_{t'+1} = < c' >) \quad (11)$$

ただし，$\hat{p}_{tag}(o_t = (w, c)|\boldsymbol{x}', \boldsymbol{o}_{1:t-1})$は，式（9）と$c' = c$の場合の式（11）の 2 つから求まるため，これらの確率値の和とする．また，ケース 1 と同様，式（9）から式（11）を用いて求めた$\hat{p}_{tag}$は式（7）と式（8）のように補正し，補正後の確率$\tilde{p}_{tag}$を式（2）の$\hat{p}_{tag}$として用いる．

2.2.4 実験

2.2.4.1 実験設定

本研究では，WMT2014 の英語とドイツ語間のニュース翻訳タスク（教師データ：3,891,998 文対，開発データ：39,330 文対，評価データ：3,003 文対）と WMT2020 の英語と日本語間のニュース翻訳タスク（教師データ：7,455,119 文対，開発データ：1,774 文対，評価データ：1,000 文対）において提案モデルの有効性を評価した．実験では，提案のアンサンブルモデル（*Ens*）を，NE を活用しない通常の Transformer NMT モデル（*BS*），原言語文と目的言語文の両方の NE を活用するタグ付けモデル（*Tag*），原言語文と目的言語文の両方の NE を活用する埋め込みモデル（*Emb*）の 3 つの NMT モデルと比較した．なお，*Ens* は *Tag* と *Emb* のアンサンブルモデルである．全ての NMT モデルのベースには Fairseq[7]の Transformer *base* を使用し，ハイパーパラメータは Vaswani ら[14]の設定に従った．各 NMT モデルの学習は，検証データに対する性能が 5 エポック連続で向上しなくなったら終了させ，最後から 5 エポック分のモデルのパラメータを平均したものを最終的なモデルとした．評価指標には BLEU（%）[8]を用いた．

各データセットは Byte Pair Encoding（BPE）[11]でサブワード分割し，英語とドイツ語間の実験では語彙を共有し，語彙サイズは 40,000 とした．英語と日本語間の実験では語彙は共有せず，語彙サイズはそれぞれ 32,000 とした．BPE 適用後，教師データと開発データから，文長が 250 以上の文を含む文対及び原言語文と目的言語文の文長比が 1.5 以上の文対を削除した．文中の NE を特定するための NER 器は，英語とドイツ語間の実験では Stanza（英語文には CoNLL03 モデル，ドイツ語文には GermEval14 モデル）[9]を，英語と日本語間の実験では spacy（英語文には en_core_web_trf モデル，日本語文には ja_core_news_trf モデル）[4]を使用した．

2.2.4.2 実験結果

実験結果を表 1 に示す．表 1 より，提案モデルによって翻訳性能が向上することを確認した．具体的には，英独翻訳では最大 1.13 ポイント，独英翻訳では最大 1.62 ポイント，英日翻訳では最大 2.08 ポイント，日英翻訳では最大 1.24 ポイント BLEU が向上することを確認した．また，モデル間の性能差に対してブートストラップによる有意差検定[5]（有意差水準 5%）を行い，

表 1 翻訳性能（BLEU（%））

モデル	英独	独英	英日	日英
BS	26.65	30.84	20.95	21.46
Emb	26.49	30.99	21.94	21.60
Tag	27.01	31.24	22.43	21.88
Ens	**27.62**	**32.46**	**23.03**	**22.70**

全ての言語対において，提案モデルによる性能改善が比較対象の全モデルに対して統計的に有意であることを確認した．

2.2.5 考察

2.2.5.1 単一モデルのアンサンブルとの比較

Ens は *Tag* と *Emb* の二つの異なるモデルをアンサンブルしたモデルだが，機械翻訳の分野では，単一のモデルにおいても，シード値を変えて学習した複数モデルをアンサンブルすることで翻訳性能が向上することが報告されている[3]．そこで本節では，提案モデルの効果を明らかにするため，二つの *Emb* をアンサンブルしたモデル（*Ens-Emb*）と二つの *Tag* をアンサンブルしたモデル（*Ens-Tag*）の翻訳性能を評価し，提案モデルの翻訳性能と比較する．*Ens-Emb* は二つの *Emb* の出力確率を平均した確率に基づいて翻訳を行うモデルであり，*Ens-Tag* は二つの *Tag* の出力確率を平均した確率に基づいて翻訳を行うモデルである．結果を表 2 に示す．

表 1 と表 2 の比較より，従来の報告同様，タグ付けモデルと埋め込みモデルの各モデルにおいてもシード値を変えて学習した複数モデルをアンサンブルすることで翻訳性能が改善することを確認した．また，表 2 において，全ての言語対で *Ens* の翻訳性能が *Ens-Emb* や *Ens-Tag* の翻訳性能よりも高いことを確認した．特に，独英翻訳及び日英翻訳においては，*Ens* は *Ens-Emb* や *Ens-Tag* と比較して有意な性能改善（独英 : $p < 0.1$，日英 : $p < 0.05$）であった．これより，単一モデル同士をアンサンブルするモデルよりもタグ付けモデルと埋め込みモデルの異なるモデルをアンサンブルする提案アンサンブルモデルが有効であることを確認した．

2.2.5.1 アンサンブルモデルにおけるリランキングの活用

NMT においては，出力候補文をリランキングすることで翻訳性能が改善できる場合がある[10, 15]．しかし，NE を活用する NMT では，リランキングの活用は研究されていない．そこで本節では，提案モデルの翻訳性能をさらに向上させるため，提案モデルにリランキングを導入する．具体的には，後藤ら[15]の逆翻訳の尤度に基づくスコア（S_{BT}）と，Salazar ら[10]の Masked Language Model に基づくスコア（S_{MLM}）を用いたリランキングを導入する．S_{MLM}の算出には RoBERTa を使用した．表 3 に英語と日本語間の翻訳実験の結果を示す．

表 3 より，英日翻訳のS_{MLM}を除いて，リランキングを導入することで提案モデルの翻訳性能が向上することを確認した．また，S_{BT}とS_{MLM}の両方を用いてリランキングすることで，翻訳性能がさらに向上することを確認した．

表 2 アンサンブルモデルの翻訳性能（BLEU（%））

モデル	英独	独英	英日	日英
Ens-Emb	27.36	32.12	22.93	22.18
Ens-Tag	27.58	32.25	22.56	22.32
Ens	**27.62**	**32.46**	**23.03**	**22.70**

表 3 リランキング導入後の翻訳性能（BLEU（%））

モデル	英日	日英
Ens	23.03	22.70
Ens+S_{BT}	23.15	22.75
Ens+S_{MLM}	22.85	22.90
Ens+S_{BT} + S_{MLM}	**23.28**	**23.01**

2.2.6 おわりに

本研究では，原言語文と目的言語文の両方の NE を活用するタグ付けモデルと埋め込みモデルをアンサンブルにより統合し，2 つのモデルの生成確率の平均から次の単語とその NE クラスを予測する NMT モデルを提案した．WMT2014 の英語とドイツ語間の翻訳タスク及び WMT2020 の英語と日本語間の翻訳タスクにおいて，提案モデルにより翻訳性能が改善できることを確認した．また，提案モデルはタグ付けモデル同士のアンサンブルや埋め込みモデル同士のアンサンブルよりも翻訳性能が高いことを確認した．そして，英語と日本語間の翻訳実験において，リランキングを導入することで提案モデルの翻訳性能が向上することを確認した．今後の展望として，活用する NE や翻訳言語対を変更した場合の提案モデルの有効性の検証が考えられる．

参考文献

[1] Barrault, L., Biesialska, M., Bojar, O., Costa-juss`a, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubeˇsi´c, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). “Findings of the 2020 Conference on Machine Translation (WMT20).” In Proc. of WMT 2020, pp.1–55.

[2] Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). “Findings of the 2014 Workshop on Statistical Machine Translation.” In Proc. of WMT 2014, pp.12–58.

[3] Garmash, E. and Monz, C. (2016). Ensemble learning for multi-source neural machine translation. In Proc. of COLING 2016, pp.1409–1418.

[4] Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). “spaCy: Industrial-strength Natural Language Processing in Python.”.

[5] Koehn, P. (2004). “Statistical Significance Tests for Machine Translation Evaluation.” In Proc. of EMNLP 2004, pp.388–395.

[6] Nowakowski, A., Pa lka, G., Guttmann, K., and Pokrywka, M. (2022). “Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation.” In Proc. of WMT 2022, pp.326–334.
[7] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). “fairseq: A Fast, Extensible Toolkit for Sequence Modeling.” In Proc. of NAACL 2019 (Demonstrations), pp.48–53.
[8] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “Bleu: a Method for Automatic Evaluation of Machine Translation.” In Proc. of ACL 2002, pp.311–318.
[9] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.” In Proc. of ACL 2020 (System Demonstrations), pp.101–108.
[10] Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked Language Model Scoring. In Proc. of ACL 2020, pp.2699-2712.
[11] Sennrich, R., Haddow, B., and Birch, A. (2016). “Neural Machine Translation of Rare Words with Subword Units.” In Proc. of ACL 2016 (Volume 1: Long Papers), pp.1715–1725.
[12] Siekmeier, A., Lee W., Kwon, H., and Lee, J. -H. (2021). Tag Assisted Neural Machine Translation of Film Subtitles. In Proc. of IWSLT, 2021 pp.255-262.
[13] Ugawa, A., Tamura, A., Ninomiya, T., Takamura, H., and Okumura, M. (2018). “Neural Machine Translation Incorporating Named Entity.” In Proc. of COLING 2018, pp.3240–3250.
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). “Attention is All You Need.” Advances in Neural Information Processing Systems 30, pp.5998–6008.
[15] 後藤 功雄,田中英輝 (2018). ニューラル機械翻訳での訳抜けした内容の検出. 自然言語処理, 25(5), pp.577- 597.
[16] 南端 尚樹, 田村 晃裕, 加藤 恒夫 (2022). 目的言語文の固有表現タグ付与に基づく Transformer ニューラル機械翻訳. NLP2022 発表論文集, pp.937-941.

2.3 語義曖昧性解消に着目した
英日マルチモーダル機械翻訳の評価セット構築と分析

一橋大学大学院　ソーシャル・データサイエンス研究科
小町　守

2.3.1 はじめに

自然言語処理とコンピュータビジョンの融合が注目を集めている．マルチモーダル機械翻訳（MMT）は，その融合の一分野であり，視覚情報を利用して翻訳品質を向上させることが提案されている．機械翻訳 （MT） モデルが曖昧な文を翻訳する場合，文脈だけでは十分な情報が得られないことがある．そこで MMT はより正確な翻訳のために視覚情報を追加し入力文の文脈情報を補完する．しかし，既存の MMT システムは，翻訳品質を効果的に向上できていない [1, 2]．このような結果の原因として，モデル構造，学習データ，評価データが考えられる．本研究では，評価データの質の影響に注目する．

MMT の標準的なベンチマークは，Flickr30K データセット [3] から英語のキャプションをドイツ語 [4]，フランス語 [5]，チェコ語 [2]，日本語 [6] に翻訳することで構築されている．英語のキャプションは画像を曖昧性なく詳細に記述しているため，正確な翻訳を生成するために視覚情報で補完する必要がないものがほとんどである [7]．したがって，このようなベンチマークは MMT における語義曖昧性解消への画像の寄与を評価するのには適していない．

より正確な評価に向けて，原文に語義曖昧性を明示的に挿入した評価データセットがいくつか存在する [8, 9]．Futeral ら [9] は，フランス語に翻訳した際に異なる表層となるような複数の語義を持つ英単語をもとに曖昧性解消指向のベンチマークを提案した．これらの研究はアルファベット言語が対象であり，文化的距離の遠い非アルファベット言語はあまり注目されていない．そこで我々は，画像が曖昧性解消の手がかりとなる場合のみを含む英日 MMT 評価セットを構築した．具体的には，WordNet [10] を用いて複数の語義を持つ英単語を抽出し，画像を与えることで語義が判別できる 250 ペアを手動で選択した．また，先行研究 [9] に倣って複数の訳語候補があり得る語義曖昧性を含む文を作成し，その語義に対応する画像を ImageNet [11] から収集した（図 1）．[1]

[1] 本研究は言語処理学会第 30 回年次大会で発表したものである。
佐藤郁子，平澤寅庄，金輝燦，岡照晃，小町守．語義曖昧性解消に着目した英日マルチモーダル機械翻訳の評価セット構築と分析．言語処理学会第 30 回年次大会．（2024）

En: This is a photo of a **seal**.
Ja: これは**封**の写真である。

En: This is a photo of a **seal**.
Ja: これは**アザラシ**の写真である。

図 1: 英日翻訳における視覚的文脈による曖昧性解消の例.

また, 既存の MMT モデルを我々のデータセットで評価し, 語義曖昧性解消の能力を評価した. さらに, 出力の定性的分析を行った. その結果, MMT システムはテキストのみのシステムよりも翻訳品質が若干向上する程度であり, ほとんどの曖昧な単語を曖昧性解消して翻訳することはできなかった. この結果は, MMT システムが画像を取り込むことができないのは, 評価データによるものではなく, モデル構造や学習データによるものであることを示している.

2. 3. 2 関連研究

MMT を評価するためのデータセットを以下に示す. Caglayan ら [12] は入力文の一部をマスクすることで文脈情報を意図的に制限し, 画像の効果を分析している. この分析から, マスクされた入力文では MMT モデルが MT モデルを上回り画像が有効であることが示された. しかし, マスク入力は実用的な MMT システムには適していない. 本研究では, 入力文をマスクせず語義の曖昧性に着目することで, 翻訳候補が複数存在する設定を提案する.

Lala らは視覚的文脈やテキスト文脈が翻訳にどの程度寄与するかを調査するために, Multimodal Lexical Translation Dataset [8] を構築した. このデータセットは視覚的文脈に限定されておらず, 画像で表現できない単語も含まれているため, MMT における視覚的文脈の寄与を評価するには不向きである. そこで, 視覚的文脈のみによる曖昧性解消のための高品質な MMT 評価データセットを構築する.

CoMMuTE [9] は, 視覚的文脈によって訳語が決定される曖昧文からなる英仏データセットである. 各事例は, 曖昧な英文, 2 つの翻訳可能な対象文, 対象文に対応する 2 つの画像から構成される. 彼らは, 曖昧な英文を 談話評価データ [13] から 29 収集し, さらに 21 の文を自作し, 合計 50 の文を作成した. 英仏翻訳で発生する語義の曖昧さは, 英日翻訳でも同じ単語で発生する. しかし, 単にフランス語から日本語に翻訳するだけでは, 評価データセットが比較的小さくなってしまう. そこで, より効率的にデータサイズを拡張するために, WordNet を用いた曖昧語抽出法を提案する.

2.3.3 データセット構築

2.3.3.1 語義ペア候補の自動選定

このステップでは, WordNet から画像で表現しやすい名詞を抽出する. WordNet には動植物

など専門的な名詞も多く含まれるが，使用範囲が限定されるため学習データに含まれる可能性は低く，より出現頻度の高い一般的な単語の抽出を目指す．図 2 にデータセット構築の概要を示す．

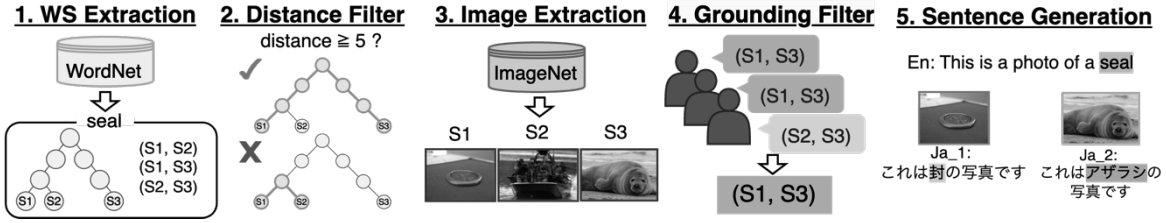


図 2: データセット構築の概要. S1, S2, S3 は語義 1, 語義 2, 語義 3 を表す.

Step 1: Word-senses extraction from WordNet
以下の条件に従って，WordNet から多義名詞と木構造語義を抽出する．(1) 名詞の長さが 10 文字以下（一般的な単語を抽出するため）．(2) 物理的実体に属する（画像で表現できる語義を抽出する）．次に，抽出した語義から語義対を作成する．

Step 2: Distance Filter
語義間の距離は 2 つの語義ノードを結ぶ辺の数として定義される．距離が 5 未満の語義ペアは除外する．フィルタリング後の単語数は 725 であり，各単語の語義対の平均数は 2.07 である．各単語について，距離の降順に語義対をソートする．

Step 3: Image Extraction from ImageNet
各語義に対応する画像を ImageNet から取得する．どちらかのノードに対応する画像がないペアは削除される．

2.3.3.2 語義ペアの自動アノテーション

自動的に抽出された語義ペアの中から，各単語に適切なペアを手動で選択する．さらに，選択されたペアに対応する画像が不適切な場合は，画像を置き換える．アノテーションは，日本語を母語とし，コンピュータサイエンスの修士課程に在籍する 3 名が担当した．すべてのアノテーターは，語義対のリストから同じ順序で語義対を選択する．

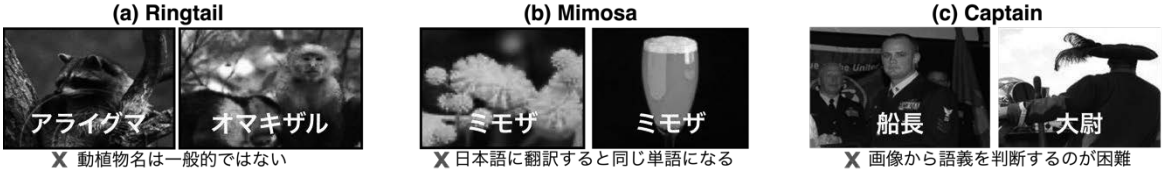


図 3: 人手によるアノテーションで除外された事例とその理由.

Step 4: Grounding Filter

単語と意味のペアをチェックし, 両方の意味が一般的で画像で表現できるペアを選択する. 適切なペアがない場合, 対象語は除外される. 図 3 は, 除外すべき不適切な語義ペアの例である.

各アノテータはそれぞれ 194, 123, 158 のペアを選択した. 選択された語義ペアの全ペア中の一致率は,Fleiss の Kappa 値で計算すると 0.256% であり, おおよそ一致している. 少なくとも 1 人に選択されたペアは合計で 197 組ある [2].

また, 対応する語義を適切に表し, 特徴抽出に十分な品質の画像を確保するため, 123 枚の不適切な (解像度が著しく低い, 教師ラベルの語義が正しくない) 画像を, Flickr から CC BY ライセンスで取得した代替画像に置き換える.

Step 5: Sentence Generation

対象単語はテンプレート文 “This is a/an/the [ ].” に挿入される. このテンプレート文は文章の文脈から語義が判断できないように曖昧さを維持した形式になっている. 各単語に 2 つの語義が使われているため, 最終的な文数は対象単語数の 2 倍となる. 構築したデータセットの統計量を表 1 に示す.

表 1: データセットの統計情報.

単語数	画像数	データサイズ	平均語義間距離
250	500	500	9.28

2.3.4 既存モデルにおける評価

2.3.4.1 実験設定

データ

評価には我々のデータセットを使用し,学習と評価の両方に Flickr30k Entities-JP を使用した. Flickr30k Entities-JP は 29,000 の学習データ, 1,014 の検証データ, 1,000 の評価データがある. 英語は Multi30K task 1 [4] に従ってトークン化し, 日本語は MeCab を使って単語分割した. (IPA 辞書) を用いて単語分割を行った. サブワード分割は BPE を用いて行う.

モ デ ル

MMT モデルと MT モデルを比較し, 画像の寄与を評価した. テキストベースの MT モデルとして Transformer-Tiny [15] を用いた. また, Transformer-based Attentive multimodal Transformer (Attentive) [16], Gated multimodal Transformer (Gated) [15], Visual Translation Language Modelling (VTLM) [17] を MMT モデルとして用いた. VTLM は Con- ceptual Captions データセットで事前学習されている. CoMMuTE の研究で提案されたモデルは, 大量のキャプションデータに対する事前学習が必要であり, 計算コストの点から本研究では使用しなかった. 画像特徴として CLIP [18], Vision Transformer [19], ResNet-50 [20] を用いた. MT と MMT モデルのアーキテクチャは, 層数を 4, 注意メカニズムのヘッド数を 4, 隠れ層の次元数

[2] データセットを拡張するために, CoMMuTE と Word-in-Context Dataset [14] から 53 の語義ペアを選択する. Step 2 の語義ペアを組み合わせることで, 最終的に 250 のペアを得る.

を 256 とした.

評価指標

sacreBLEU [21] と COMET [22] を使用した. 訳文の本質的でない摂動（文末の変更など）の影響を軽減するため, 3 つの参照文を作成し, その平均値を報告する. さらに, モデルの曖昧性解消を評価するために, [8] で提案された指標も採用し, $\frac{C}{N}$を計算する. ここで, C は出力中の対象単語が参照中の対象単語と正確に一致した回数, N はデータセットサイズである. この指標を本研究では Lexical Accuracy (LA) と呼ぶ.

2.3.4.2 結果

表 2 は, 既存の MMT モデルの自動評価における性能である. MMT モデルは MT モデルを上回り, 画像の寄与が示された. 特に, Attentive (RCNN) は顕著な改善を示し, このモデルが画像に対してより敏感であることを示唆している.

Flickr30k のスコアは, BLEU では −3.61 から 1.06 の範囲, COMET では $-3.40 \times 10{-}3$ から $1.30 \times 10{-}3$ の範囲で改善した. 一方, 我々のデータセットは,BLEU で −1.50 から 2.30, COMET で $4.26 \times 10{-}2$ から $5.24 \times 10{-}2$ の範囲で改善した. 我々のデータセットは Flickr30k よりも大幅に改善し, 画像の寄与をより敏感に評価することができた.

表 2: (M)MT モデルの結果. 太字は, MT モデルを上回っていることを示す.

Metric	Eval data	Transformer Tiny	Gated		Attentive		VTLM
			CLIP	ResNet	CLIP	R-CNN	R-CNN
BLEU	Flickr30k	43.42	**43.48**	**44.12**	**44.48**	**43.99**	39.81
	Ours	29.40	**29.68**	**30.07**	**30.43**	**31.69**	27.90
COMET	Flickr30k	0.9679	0.9672	0.9672	**0.9688**	**0.9692**	0.9645
	Ours	0.8888	**0.9314**	**0.9314**	**0.9399**	**0.9381**	**0.9412**
LA	Ours	0.1900	**0.1960**	0.1860	**0.1960**	**0.1980**	**0.2200**

2.3.4.3 分析

視覚情報による訳語変化

システムの出力を詳細に分析した. 図 4 は 2 つの出力例を示しており, MT モデルは Transformer-Tiny , MMT モデルは (a) VTLM (RCNN), (b) Attentive (RCNN) である. (a) の例では, MT モデルは両方の語義を「フード」に翻訳した. 一方, MMT モデルは, 対応する画像を参照することで, 「ボンネット」と区別することができた. しかし, MMT モデルが正しい訳語に変換できたのは 8 例だけであった. また, 目的語以外の単語が変化している例 （文末の変化, 読点の挿入など）も複数あった. これらの結果から, 自動評価スコアの向上は, 対象単語以外の変更によるトークン数の変化に大きく影響される可能性があることが示唆された.

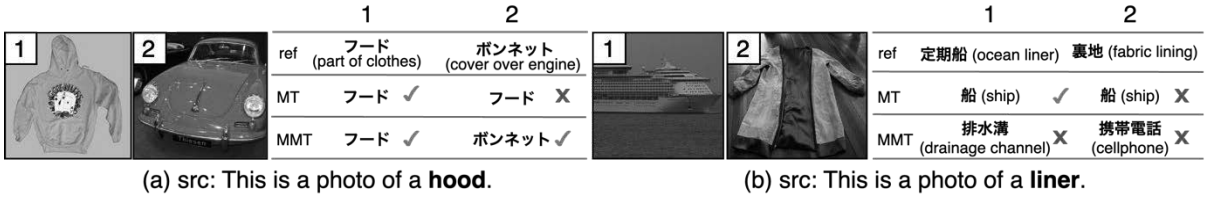


図 4: いくつかの出力例. 太字は対象単語を示す.

翻訳品質が向上したのは 8 例だけであったが, 視覚情報が出力対象語に影響を与えたと思われる例も複数あった（図 4 の liner など）. このような事例の数をモデルごとに表 3 に示す. 画像によって訳語が変化した例のうち, 正しく訳されたのは 7%だけであった. つまり, 既存のモデルは視覚情報をわずかしか利用しておらず, 改善の余地がある.

表 3: 全 MMT モデルにおける視覚情報によって翻訳が変わった事例の数.

Model	Correct	Incorrect
Gated (CLIP)	0	2
Gated (ResNet-50)	0	3
Attentive (CLIP)	1	3
Attentive (Faster R-CNN)	2	42
VTLM (Faster R-CNN)	5	56

対象単語の学習/検証データでの存在割合

本研究では英日翻訳の際に曖昧性が生じる単語（対象単語）をもとに曖昧性のある文対を作成している. 前節での分析の結果より, 翻訳品質の改善は限定的であることが示唆される. この原因として対象単語が学習データに含まれないためその単語についての学習が不十分であることが考えられる. そこで, 本節では対象単語が学習/検証データに含まれる割合を調査する.

表 4: 全対象単語のうち学習/検証データに含まれる単語の割合. 英語はソース側に対象単語の表層が含まれる割合, 日本語はターゲット側に対象単語の語義が含まれる割合, 英日は対象単語が 1 つの文についてソースとターゲットの双方に含まれる割合を示す.

	学習データ	検証データ
英語	0.916	0.208
日本語	0.690	0.218
英日	0.000	0.000

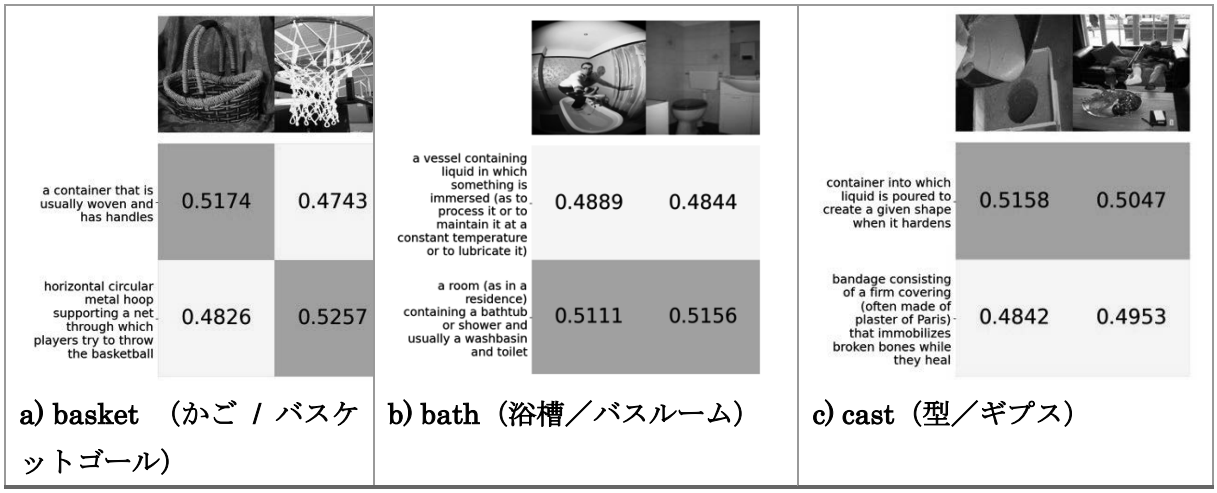
表 4 の結果より, 英日双方に含まれる対象単語の割合は 0 であることがわかる. すなわち, 対象単語が評価データと同じ英日翻訳で使用されている文は学習/検証データには存在しないにも

かかわらず，わずかな事例は正確に翻訳できているということである．一方で, 91.6% の対象単語はソース側の学習データに含まれており, 69.0% の対象単語はターゲット側の学習データに含まれている．すなわち，ソース側とターゲット側それぞれで単語が画像中の物体とマッピングされている場合，画像が仲介することでソース側の表層とターゲット側の語義を対応付けている可能性が考えられる．

画像の識別能力評価

既存 MMT モデルにおいて画像情報の効果が限定的である原因として, エンコードしているモデル自体が語義と画像を結び付けられていない可能性が挙げられる．モデルが語義と画像の対応関係を学習できているか検証するため，画像の識別能力を調査する. 検証では，対象単語の語義ペアごとに，各画像についてどちらの語義を表しているかを 2 値分類し,その正解率を報告する. 分類のモデルには CLIP [18] を使用する．

検証の結果，画像 2 値分類の正解率は 92.4% で, 500 語義のうち誤分類は 38 語義のみであった．すなわち，モデルはほとんど全ての画像と語義の対応付けを正しく行なっていると言える. 分類結果の一部を図 5 に示し，定性分析を行う．basket （かご / バスケットゴール，図 5a）の例では，画像と対応する語義が正しく分類されている．一方で, cast （型 / ギプス，図 5b）と bath （浴槽 / バスルーム，図 5c）の例では誤った分類結果となっている．cast の例は，画像がギプス以外の物体が多く画像のエンコード自体が困難なことが原因として考えられる. bath の例は，浴槽自体がバスルームに包含されるものであるという関係から対応語義を選択することが困難であったと考えられる．



a) basket （かご / バスケットゴール）　b) bath（浴槽／バスルーム）　c) cast（型／ギプス）

図 5: 画像の 2 値分類における分類結果.

2.3.5 おわりに

MMT における視覚情報の寄与を正確に評価するために，英日評価データセットを構築した. 本データセットで既存のモデルを評価した結果，画像が翻訳品質を向上させるケースはわずかであることがわかった．つまり, MMT がうまく機能しない原因は，画像を必要としない設定の評価

データではなく，モデル構造や学習データに改善の余地があることがわかった．

参考文献

[1] Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. The MeMAD submission to the WMT18 multimodal translation task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 2018.

[2] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 2018.

[3] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, Vol. 2, pp. 67–78, 2014.

[4] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In Proceedings of the 5th Workshop on Vision and Language, 2016.

[5] Desmond Elliott, Stella Frank, Lo¨ic Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In Proceedings of the Second Conference on Machine Translation, 2017.

[6] Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. A visually-grounded parallel corpus with phrase-to-region linking. In Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020.

[7] Stella Frank, Desmond Elliott, and Lucia Specia. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. Natural Language Engineering, 2018.

[8] Chiraag Lala and Lucia Specia. Multimodal lexical translation. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[9] Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023.

[10] Christiane Fellbaum, editor. WordNet: An Electronic Lexical Database. MIT Press, 1998.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), Vol. 115, No. 3, pp. 211–252, 2015.

[12] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
[13] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long
Papers), 2018.
[14] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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), 2019.
[15] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.
[16] Jindřich Libovický, Jindřich Helcl, and David Mareček. Input combination strategies for multi-source transformer decoder. In Proceedings of the Third Conference on Machine Translation: Research Papers, 2018.
[17] Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. Cross-lingual visual pre-training for multimodal machine translation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021.
[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, 2021.
[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations, 2021.
[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems 28, 2015.

[21] Matt Post. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, 2018.
[22] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

3. 国際ワークショップ開催報告

3. 国際ワークショップ開催報告：PSLT 2023

静岡大学　綱川　隆司
奈良先端科学技術大学院大学　須藤　克仁
NHK 放送技術研究所　後藤　功雄

3.1　開催概要

本研究会では 2005 年から特許翻訳に関するワークショップを主催しており、2015 年以降は特許・技術文書翻訳ワークショップ（Workshop on Patent and Scientific Literature Translation）として隔年で開催している。第 10 回となる今回も中国・マカオにて開催された機械翻訳サミット（Machine Translation Summit）との併催として 9 月 5 日（火）に開催した。オーガナイザは Co-chair として綱川、須藤、後藤の 3 名が務め、プログラム委員として研究会委員の研究者にご協力いただいた。前回 2021 年の第 9 回ワークショップは新型コロナウイルス拡大の影響で完全オンライン形式での開催となったが、今回は機械翻訳サミット本会議と同じく現地とオンラインのハイブリッド形式での開催となった。

本ワークショップでは特許庁、機械翻訳ユーザー、および機械翻訳研究者のそれぞれの立場から 3 件の招待講演をいただいた。現地にて 22～23 名程度、オンラインで 15 名程度の参加があり、質疑も活発であり盛況であったと言える。

3.2　招待講演

招待講演 1 件目は日本特許庁の名和大輔氏から、特許情報ユーザーのための特許庁の情報サービスに関するご講演をいただいた。特許庁では即時に正確な情報を提供するため多様な情報提示を行っている。特許情報プラットフォーム（J-PlatPat）は特許庁が運営する日本で最大の知的財産データ検索サービスであり、その中で海外特許および日本特許のニューラル機械翻訳等による翻訳結果を提供しており、企業等に向けた API を介したサービス提供も行っていることが紹介された。

招待講演 2 件目は中国の知的財産権出版社（IPPH；知识产权出版社）から YAN Jiang 氏による、IPPH 及びその特許機械翻訳サービスである“中知慧译”（ZhongZhi HuiYi）のご紹介があった。IPPH は 2010 年に特許向け AI・機械翻訳サービスの提供を開始し、2023 年には中国の知的財産情報を扱う最大のメディアグループとなっている。特許機械翻訳の課題である省略、長い文章、特許に特有のパターン、専門用語の一貫性等に対応する、ニューラル機械翻訳ベースの 10 言語に対応した機械翻訳サービス“中知慧译”は、1 日に 10 万件の特許を翻訳しており、中国語と英語、日本語、ドイツ語等の特許翻訳において高い性能を示している。

3 件目の招待講演は澳門大学の Derek W. Wong 氏による、Wong 氏らの研究グループにおけるニューラル機械翻訳の研究に関係したテーマのご発表であり、不確実性を考慮したカリキュラム学習、文ベクトルのノルムによるカリキュラム学習、自己適応学習、分野適応のためのメタカリキュラム学習、難易度に基づく機械翻訳評価、分散を考慮した機械翻訳評価テストセットといった

研究テーマのご紹介があった。

ニューラル機械翻訳モデルの機械学習においては、人間に対する効果的な学習方法を模したカリキュラム学習の有効性が知られている。効果的なカリキュラムとして、訓練データの難易度や不確実性を考慮した手法や、モデルが自己の学習を調整するアプローチについて、翻訳性能やその収束速度が評価、分析されている。また特定の分野に適応させるため、カリキュラム学習の方法を学ぶメタカリキュラム学習を取り入れ、過学習を軽減し、長文に対してより優れた性能を示している。また、機械翻訳システムはテストセットにおいける正解訳との類似性を求めることで性能評価されているが、テストセット中の文の翻訳の難しさに応じた評価や、テストセットを分散指標により選別することでよりバラつきの多いものにし、翻訳システムの弁別力を測る評価指標を提案している。

3.3 所感

本会議前のワークショップとしては比較的多くの方にご参加いただけたこと、また特に現地会場での質疑が活発であり、プログラム編成時に想定していた総合討論の時間をも活用して招待講演に対する質疑が続いたことはワークショップ提案者としては喜ばしい。招待講演が 1 件オンライン、1 件が録画放映であった点はワークショップの盛り上がりの面で懸念していたが、ハイブリッド開催に対する参加者の慣れもあってか、大きな障害にはならなかったと言える。

一方で、一般講演については論文投稿がなく招待講演のみの短時間での開催になったことについては残念であり、今後の企画に向けた大きな課題である。ニューラル機械翻訳の進展を上回る勢いで急速に活用が広がる大規模言語モデルは、特許や技術文書の翻訳やその活用の上でも重要な役割を果たすことがほぼ確実であり、応用事例も含めて幅広く技術的課題の解決に向けた議論の場として本ワークショップをご活用いただくべく、募集分野やワークショップ名の再検討を行った上で引き続き実施していきたいと考えている。

2023年度AAMT/Japio特許翻訳研究会報告書

発 行 日　2024 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