対訳関係にノイズのある対訳文からの新しい翻訳知識の学習

後藤功雄　衣川和尭　美野秀弥　河合吉彦　山田一郎
NHK 放送技術研究所
{goto.i-es,kinugawa.k-jg,mino.h-gq,kawai.y-lk,yamada.i-hy}@nhk.or.jp

概要

ニュースでは新しい語彙や表現が出現するため，ニュースの機械翻訳では新しい語彙や表現の翻訳知識を学習していく必要がある．英語ニュースは日本語ニュースの単純な翻訳ではないため，日英ニュース記事から抽出できる対訳文は，対訳関係にノイズがある場合が多い．対訳関係にノイズがあるデータから翻訳を学習すると，ノイズも学習してしまうという課題がある．また，新しい表現は頻度が少ない場合が多く，低頻度の表現の学習は難しいという課題もある．そこで，ノイズの学習を抑え，低頻度の表現を学習しやすくすることを目指した NMT の手法として．学習したい表現部分と後続文脈の一部のみから誤差逆伝播で学習する手法を提案する．

1　はじめに

NHK では，日本語ニュースを基にして英語ニュースを制作しており，この制作支援に日英機械翻訳システムを活用している．ニュースは新しい話題を報道するため，新しい語や名前，表現がしばしば出現する．機械翻訳システムは構成的に翻訳できない表現に関しては，学習していないものは基本的には翻訳できない．情報発信のための英語ニュース制作では，正確な表現であることが重要であることから，新しい語や表現の訳は，最初はニュースライターが正確な訳を調べて英語ニュースを制作する．過去の英語ニュースに既に出現している語や表現は，機械翻訳システムがそこから翻訳知識を自動的に学習して翻訳できるようになることが期待されている[1)]．しかし，英語ニュースの制作は日本語ニュースの単なる翻訳ではない [1] ため，同じニュースを伝える日英の記事対から抽出した対訳文対は，対訳関係にノイズが多いデータとなる．また，出稿する記事数は限りがあり，多くの話題の頻度は低い．すなわち，ノイズを含む対訳データから低頻度の新しい語や表現の翻訳知識の学習が必要となっている．

ニューラル機械翻訳（NMT）で対訳関係にノイズを含む対訳文を追加学習した場合，対訳関係のノイズを学習してしまうという課題と，低頻度の新しい語や表現は学習しても出力されにくいという課題がある．

訓練データの対訳文に含まれる対訳関係のノイズの影響を低減する NMT の手法がある [2]．この手法は単語対応の情報を利用する．単語対応の推定は難しいため推定誤りが発生し，それが NMT の学習におけるノイズの影響の低減効果に影響してしまうという課題がある．また，データ固有のアンカートークンを入力文に追加することで対訳ノイズの影響を低減する NMT の手法もある [3]．しかし，アンカートークンが低頻度の新しい語や表現の翻訳元として学習されてしまう恐れがあるため，この手法は，低頻度の新しい語や表現の学習には不向きと考えられる．

本稿では，これらの課題に向けた対策手法を提案する．対訳関係にノイズを含む対訳文をニューラル機械翻訳（NMT）で追加学習する際に，学習したい表現と後続文脈の一部のみから誤差逆伝播させる．これによって，訳抜けなどの対訳関係のノイズを学習してしまうことが低減される効果と，対象の表現を集中的に学習することで，学習されやすくなる効果が期待される．単語対応の情報を利用しないため，単語対応推定の誤りの影響を受けない．NHK のニュース文を用いた日英翻訳の実験を実施した．追加学習により翻訳品質が下がってしまう課題は，この方法では効果が見られなかったが，対象の表現が出力されやすくなる効果が確認された．

1)　高性能な日英機械翻訳システムを開発するために，これまで人手をかけて高品質な対訳データを整備して，高性能な日英機械翻訳システムを開発した [1]．しかし、現状の訓練データでは，今後出現する新しい語や表現は翻訳できないため，日英ニュースから自動的に翻訳知識を学習したい．

2 提案手法

例えば，下記に示す，日英記事（2021/4/4 NHK ニュース）中の文対から，“スエズ運河庁：SCA” の翻訳知識を学習したい場合を想定する．

- エジプトのスエズ運河では先月 23 日、愛媛県の正栄汽船が所有し、台湾の会社が運航する大型コンテナ船が座礁して運河が塞がれ、6 日後の 29 日にコンテナ船の離礁に成功し、運河の通航が再開されました。運河を管理するスエズ運河庁は、待機を余儀なくされた 422 隻の船舶すべてが 3 日、運河を通過したと発表しました。
- The SCA said on Saturday that all 422 ships stranded by the Japanese-owned, Taiwanese-operated Ever Given had passed through the canal by the end of the day.

青文字が学習したい対象部分であるが，低頻度データでは対訳部分を正確に推定して抽出することは困難である．赤文字は対応先がない部分を表しており，この文対では，日本語側に対応していない部分が多く存在する．この文対を NMT でそのまま学習すると，赤文字部分の訳抜けを学習してしまい，翻訳品質が下がってしまう．

そこで，提案手法では，NMT の訓練時に，訓練データに出現していないもしくは低頻度などで翻訳知識を学習したい目的言語表現部分と後続文脈の一部（n トークン）にのみ誤差逆伝播する対象を制限する．例えば，後続文脈の長さを 1 トークン（$n = 1$）とした場合で，“SCA” が BPE などによるサブワード分割で 3 トークン（S@@, C@@, A）[2] で表される場合，この文対の目的言語文 “The S@@ C@@ A said on Saturday that ...” のうち，誤差逆伝播させる出力トークンを 2 番目から 5 番目のトークン，すなわち，S@@，C@@，A，said の 4 トークンのみとする．このように学習したい表現部分を中心に部分的に学習することで，学習を必要としていない部分を学習することで訳抜けなどの対訳関係のノイズを学習してしまうことを抑制し，対象の表現の学習に集中することで対象の表現が出力されやすくなることが期待できる．後続文脈の一部も誤差逆伝播させる理由は，NMT のデコーダーは言語モデルの機能を持っており，新しい表現は学習していない文脈となり，その後続のトークンの推定性能が低くなることが想定され，新しい表現に続くトークンも学習するためである．

誤差逆伝播を制限する実装方法はいくつか考えられるが，本研究ではマスキングを用いる．誤差逆伝播させる直前の各出力トークンの尤度に対して、誤差逆伝播させたくないトークンの尤度に 0 を掛け，誤差逆伝播させたいトークンの尤度には 1 を掛ける．すなわち，出力トークン列に対応する 0 と 1 からなるベクトル（マスクベクトル）を出力トークンの尤度のベクトルに要素ごとに掛けて（アダマール積）から誤差逆伝播する．

2） ここでは@@はサブワード分割された分割点の左側を表している．

3 評価実験

提案手法の効果を検証するために，NHK のニュースデータを用いた実験を行った．

3.1 評価用データセットの構築

ベースとなる NMT モデルの構築用のデータには，日本語側もしくは英語側が NHK ニュース文でその対訳を人手で構築した高品質な日英対訳 1M 文対（日英ニュース 1）と開発データを用いた．この訓練データには 2020 年までのニュースを用いた．

次に，2021 年 1 月から 2022 年 10 月までの NHK 英語ニュース（英語ニュース 2）に含まれる英単語のうち，2017 年 6 月から 2020 年 12 月までの NHK 英語ニュース（英語ニュース 1）と日英ニュース 1 の英語に含まれない英単語のリスト（新英単語リスト）を取得．

英語ニュース 2 の各記事と対応する日本語ニュース記事から，対訳文を抽出し，新英単語リストの英単語を含む対訳文を選択．さらに，この対訳文の日本語側で新英単語リストの英単語に対応する日本語表現が存在する対訳文を選択．この対訳文をノイズあり対訳と呼ぶ．新英単語リストの英単語と対応する日本語表現のペア（新英単語＆日本語表現ペア）に対応するノイズあり対訳は，各ペアあたり 1 文対から最大で 3 文対である．

2021 年以降のニュースの高品質な日英対訳文対で，新英単語＆日本語表現ペアを含む対訳文対を抽出．これをテストデータ 2 ＆参照訳 2（426 文対）とする．テストデータ 2 ＆参照訳 2 に含まれる新英単語＆日本語表現ペア（25 ペア）を含むノイズあり対訳は全てノイズあり訓練データ 2（65 文対）とする．残りの新英単語＆日本語表現ペア（137 ペア）

を含むノイズあり対訳は，各ペアあたり 1 つしかノイズあり対訳がないものを除外し，各ペアあたり 2 つまたは 3 つのノイズあり対訳文対のうち，1 つの日本語文をテストデータ 1（137 文），それ以外の 1 つもしくは 2 つの対訳文をノイズあり訓練データ 1（全体で 219 文対）とした．

追加学習には，ノイズあり訓練データ 1（219 文対）とノイズあり訓練データ 2（65 文対）を合わせたデータ（284 文対）を用いた．

3.2 実験設定

日本語の単語分割には Mecab（IPA 辞書）を使い，数字は 1 トークンにまとめた．英語のトークナイザには Moses 付属のトークナイザを用いた．NMT のモデル構造には，Transformer[4] を用い，実装は独自開発で，設定値には Sockeye[5] の標準設定を用いた．その設定値は 6 層，dropout=0.1，label smoothing=0.1 などである．ただし，NHK のニュース文は文長が長いため，学習に利用する最大文長は 200 トークンに設定した．

最適化手法には adam を用いた．ベースとなる NMT モデルは，最大エポック数 100 で学習し，開発データでモデル選択を行なった．翻訳時はビーム幅 5 でビームサーチした．

追加学習は，10 エポック学習した．

3.3 実験結果

テストデータ 1 を翻訳した際に，新英単語＆日本語表現ペアの新英単語が訳出される率を調べた．結果を表 1 に示す．「追加学習なし」がベースとなる NMT の出力，「通常追加学習」がベースモデルに対して、ノイズあり対訳提案追加手法の数字 n は後続文脈として利用するトークン数である．カッコ内の数値は件数を表している．追加訓練データが 1 文のみの場合，55 件のテストデータの内，訳出できたのは「追加学習なし」および「通常追加学習」で 1.8%，提案手法は最大で，後続文脈なし (0) の 12.7%で，提案手法は改善しているが大半は訳出できていない．追加訓練データが 2 文の場合，82 件のテストデータの内，訳出できたのは「追加学習なし」が 0%および「通常追加学習」で 9.8%，提案手法は最大で，後続文脈なし (1) の 50.0%で，提案手法は通常学習に比べて改善が見られ，半数が訳出されている．なお，テストデータ 1 には高品質な参照役が存在しないため，BLEU スコアは算出していない．

表 1 テストデータ 1 における追加データ数毎の訳出率

		訳出率 (%)			
追加データ数		1		2	
追加学習なし		1.8	(1/55)	0.0	(0/82)
通常追加学習		1.8	(1/55)	9.8	(8/82)
提案追加学習	$n = 0$	12.7	(7/55)	43.9	(36/82)
	$n = 1$	7.3	(4/55)	50.0	(41/82)
	$n = 2$	7.3	(4/55)	43.9	(36/82)
	$n = 3$	7.3	(4/55)	35.4	(29/82)
	$n = 4$	5.5	(3/55)	34.1	(28/82)
	$n = 5$	3.6	(2/55)	28.0	(23/82)

テストデータ 2 を翻訳した際の BLEU スコアと，新英単語＆日本語表現ペアの新英単語が訳出される率を調べた．結果を表 2 に示す．

追加学習なしのベースモデルのスコアに比べて、追加学習したモデルはいずれも BLEU スコアが低下した．この主な原因は，追加学習で過学習してしまったためと考えられる．この結果からは，対訳関係にノイズがあるデータから学習する際にノイズの影響を抑える効果は確認できなかった．一方で，訳出率は，「追加学習なし」では 0%，「通常追加学習」では，2.1%に対して，提案手法では，最大 61.5%（後続文脈 1 トークンの場合）に向上している．

表 2 テストデータ 2 における翻訳品質

		BLEU (%)	訳出率 (%)	
追加学習なし		19.19	0.0	(0/426)
通常追加学習		15.94	2.1	(9/426)
提案追加学習	$n = 0$	15.53	58.5	(249/426)
	$n = 1$	15.14	61.5	(262/426)
	$n = 2$	15.61	35.9	(153/426)
	$n = 3$	14.87	28.6	(122/426)
	$n = 4$	15.79	18.5	(79/426)
	$n = 5$	15.70	13.6	(58/426)

4 おわりに

ニュース翻訳に向けて，ノイズのある対訳データから低頻度の新しい語や表現を学習する手法を提案した．評価実験では，低頻度の新しい語が訳出される率の向上が確認された．一方で，追加学習により全体の翻訳品質が低下した．今後，過学習を抑える対策を取り入れることで，この問題の改善に取り組んでいく予定である．

謝辞

本研究成果の一部は、国立研究開発法人情報通信研究機構の委託研究（課題 225）により得られたものです。

参考文献

[1] 後藤功雄. 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. AAMT ジャーナル, No. 77, pp. 4–10, 2022.

[2] 後藤功雄, 美野秀弥, 山田一郎. 訳抜けを含む訓練データと訳抜けのない出力とのギャップを埋めるニューラル機械翻訳. 言語処理学会第 26 回年次大会, 2020.

[3] 根石将人, 吉永直樹. ニューラル機械翻訳のためのノイズ寛容なアンカー学習. 言語処理学会第 29 回年次大会, 2023.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of NeurIPS**, pp. 5998–6008. 2017.

[5] Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. Sockeye 3: Fast neural machine translation with pytorch, 2022.