用語制約が多い翻訳に対する手法の提案

加藤 優吾 [1]　小川 泰弘 [2]　外山 勝彦 [1]
[1] 名古屋大学大学院情報学研究科　[2] 名古屋市立大学データサイエンス学部
kato.yugo.c4@s.mail.nagoya-u.ac.jp　ogawa@ds.nagoya-cu.ac.jp
toyama@is.nagoya-u.ac.jp

概要

翻訳における訳語の揺れは混乱や誤解を招く可能性があるため，決められた訳語に従った翻訳が必要となる．すなわち，用語制約のある翻訳が求められる．そのための手法として入力拡張 [1] などが提案されている．しかし，学習に使用する対訳文が用語制約に従っているとは限らず，それらの対訳文が学習においてノイズとなる可能性がある．そこで，本研究では学習データを選別する手法を導入し，入力拡張と組み合わせる手法を提案する．さらに，日本法令の対訳データセットを用いて実験し，提案手法が有効であることを示す．

1　はじめに

ニューラル機械翻訳 (NMT) モデルは，同じ意味を表す 1 つの用語（単語・熟語など）を異なる訳語で翻訳するときがあり，訳語に揺れが生じる可能性がある．製品マニュアルや契約書，法令などの翻訳では訳語の揺れが混乱や誤解を招く原因となる．そのため，1 つの用語に対して同じ訳語を使用するべきである．そこで，いくつかの用語の訳語をあらかじめ決めておき，それに準拠して翻訳するシステムが求められる．本研究において，訳語が定められた用語とその訳語のペアの集合を「用語制約」と呼ぶ．

本研究では，用語制約がある法令翻訳に取り組む．その理由は，用語制約がある翻訳が法令翻訳において必要な技術であることに加えて，対訳文と用語制約のデータが公開されていて，容易に入手できるためである．実際，対訳文として日本法令外国語訳データベースシステム[1)](JLT) が，また，用語制約として「法令用語日英標準対訳辞書[2)]」（以下「対訳辞書」という）が公開されている．対訳文と対訳辞書の例をそれぞれ表 1，表 2 に示す．JLT で公開されている訳文は対訳辞書を参考に人手で作成されているが，完全には準拠していない．例えば，対訳辞書における「労働組合」の訳語は「labor union」であるが，表 1 の例では「trade union」となっている．

表 1　JLT の対訳文の例

原文	訳文
... 労働者の過半数で組織する**労働組合**との間における ...	... with a **trade union** organized by a majority of the workers ...

表 2　対訳辞書 (ver. 15) の一部

例	用語	訳語候補
1	労働組合	labor union
2	労働者	worker | employee
3	組合	partnership

用語制約に準拠して翻訳するために入力拡張 [1] などの手法が提案されている．しかし，本研究における実験を通じて，日本法令の翻訳においてはその効果がわずかであることが判明した．その原因は，対訳辞書に準拠していない対訳文が学習データに含まれるためと考えられる．そこで，本研究では学習データを選別する手法を導入し，入力拡張と組み合わせる．また，実験によりその有効性を示す．

2　日本法令外国語訳データベースシステム

本節では JLT で公開されているデータについて，本研究と関連する事柄を記す．なお，2022/11/25 時点で公開されていたデータを実験（第 5 節）に使用したため，本節でもその時点での情報を示す．

2.1　対訳辞書

1 つの用語に対して 1 つの訳語を使用するのが理想である．しかし，用語とその訳語が一対一で対応するとは必ずしも限らず，表 2 の例 2 のように訳語が複数存在する用語が存在する．そのため，1 つの

1）https://www.japaneselawtranslation.go.jp/ja/
2）https://www.japaneselawtranslation.go.jp/ja/dicts

用語に対する訳語の集合を「訳語候補」とする．訳語候補のうちどの訳語を使用するべきかは文脈などから判断するべきである．そのため，用途や例文が記載されている語が存在する．例えば，「労働者」の訳語は一般的には「worker」であるが，民法の雇用関係に関する場合は「employee」とされている．

対訳辞書はバージョンが約 1 年ごとに更新されており，現在，ver. 15 まで公開されている．

2.2 対訳データ

JLT では日本の法令の英訳文が公開されている．訳文は対訳辞書を参考にして人手で作成されており，ネイティブや法律の専門家などによる検査を経ている．しかし，第 1 節で述べたように対訳辞書に必ずしも準拠していない．

また，対訳辞書のバージョン更新に合わせてすべての訳文を修正するのは現実的でない．ゆえに，法令によって異なるバージョンの対訳辞書を用いて翻訳されている．本稿では，それを「翻訳に使用した対訳辞書」という．

3 先行研究

用語制約に準拠して翻訳する手法は，大きく分けてハードな語彙制約手法とソフトな語彙制約手法の 2 種類が存在する．

ハードな語彙制約手法として，デコード時のビームサーチにおいて用語制約を満たすように探索する手法が提案されている [2, 3]．この手法は，与えられた用語制約をすべて満たすことが保証される．しかし，似たフレーズが繰り返し出力されるなどの問題があり，翻訳性能が低下する可能性がある．

ソフトな語彙制約手法として，入力に変更や情報を加えて学習する手法が提案されている [1, 4, 5]．これらの手法は，ハードな語彙制約手法より翻訳性能は高くなりやすいが，用語制約が満たされない可能性がある．

ソフトな語彙制約手法の 1 つに Ailem らが提案した入力拡張 [5] がある．それに対して，Wang ら [1] は 1 つの用語に複数の訳語がある場合にも対応できるよう改良した．本研究では，Wang らの手法を入力拡張と呼び，これを使用する．

入力拡張では，原文中において用語制約のある用語の後に<T>タグで囲んだ訳語を追加し，その用語と訳語の全体を<S>タグで囲む．訳語が複数あるときには<SEP>タグで区切る．例を表 3 に示す．入力拡張の目的は，<T>タグ内の訳語を出力に含めるようモデルに学習させることである．

表 3 入力拡張の例

入力拡張後
... **<S>労働者<T> worker <SEP> employee </T>** **</S>**の過半数で組織する**<S>労働組合<T> labor union </T> </S>**との間における...

4 提案手法

本研究では，用語制約のある法令翻訳において入力拡張を使用する手法を提案する．しかし，本研究における実験を通じて，その効果がわずかであることが判明した．その原因は，対訳辞書に準拠していない対訳文が学習のノイズとなるためであると考えられる．そこで，対訳文の対訳辞書に対する「準拠率」を定義し，準拠率が高い文のみで学習を行う「データ選別」を導入する．

4.1 入力拡張

第 3 節で述べた入力拡張を手法 1 とする．入力拡張は検証データとテストデータにも適用する．

ある用語と別の用語が共通文字列を持つことがある．この場合，最長一致した用語にのみタグ・訳語を追加する．例えば，表 3 の例のように「労働組合」に含まれる「組合」にはタグ・訳語を追加しない．また，各対訳文において翻訳に使用した対訳辞書を用いて入力拡張をする．

4.2 準拠率の定義

原文 s とその訳文 t のペアである対訳文の対訳辞書 $D(t)$ に対する準拠率を式 (1) で定義する．

$$\text{準拠率} = \frac{\displaystyle\sum_{(p_s,P_t)\in D(t)} \min\left(\text{count}(p_s,s), \sum_{p_t\in P_t}\text{count}(p_t,t)\right)}{\displaystyle\sum_{(p_s,P_t)\in D(t)} \text{count}(p_s,s)} \quad (1)$$

ここで，$D(t)$ とは，訳文 t の翻訳に使用した対訳辞書を表す．対訳辞書は用語 p_s と訳語候補 P_t のペアからなる集合であり，訳語候補は訳語の集合である．$\text{count}(s_1, s_2)$ は文字列 s_2 における文字列 s_1 の出現回数を示す．

対訳文の原文に用語が 1 つも含まれないとき，式 (1) の分母は 0 となるため計算できない．このと

き，その対訳文の準拠率は未定義とする．

訳語が複数あるとき，2.1 節で述べたように用途等が記載されることがある．そのため，用途に沿っているかどうかも考慮することが望ましい．しかし，それは容易でないため，訳語候補のいずれかを使用していれば辞書に準拠していることとする．

ある用語の出現回数よりその訳語の出現回数の方が多くなる可能性がある．例えば，原文で省略された主語が訳文で補われている場合などである．そこで，式 (1) において，min 関数を用いて分子の値を分母の値以下にする．すなわち，準拠率が 100%を超えないよう制限する．

4.3 データ選別

学習データにおいて「準拠率が閾値 θ 以上」，「準拠率が未定義」のいずれかの条件を満たす対訳文を抽出する．本稿では，これを「データ選別」と呼ぶ．データ選別により，準拠率が低い対訳文を学習データから除くことができる．なお，一般的に学習データは多い方が望ましいため，準拠率が未定義の文を抽出対象に含める．このデータ選別を手法 2 とする．

準拠率の計算には原文と訳文（参照訳）が必要となる．しかし，検証データ・テストデータにおいて参照訳を用いた前処理を行うのは望ましくない．ゆえに，検証データ・テストデータに対してはデータ選別を行わない．

さらに，データ選別と入力拡張を組み合わせる手法を手法 3 とする．

5 実験

評価実験を行い，提案手法の有効性を検証した．

5.1 実験設定

手法 1〜3 を用いてそれぞれ実験した．ベースラインとしてデータ選別も入力拡張も用いない実験を行った．データ選別に使用する閾値は，$\theta = 0$，50，60，70，80，90，100 の 7 種類を試した．なお，$\theta = 0$ とはデータ選別を使用しないことと同義であり，ベースラインと手法 1 が $\theta = 0$ となる．

データセット　2022/11/25 の時点で JLT に公開されていたデータを使用した．

コーパスは，XML 形式で入手した．<Paragraph-Sentence>の子要素にある<Sentence>要素を 1 文として扱った．入手したデータを法令単位で分割し，学習データ 406 法令（75,776 文），検証データ 56 法令（8,423 文），テストデータ 51 法令（11,053 文）を得た．また，データ選別の閾値によって学習データ数が変化するため，それぞれのデータ数を表 4 に示す．検証データ・テストデータに対してはデータ選別を行わないため，データ数は全実験において同一である．

表 4　閾値による学習データ数の変化

閾値 θ	データ数
0	75,776
50	70,118
60	61,716
70	47,710
80	31,359
90	14,448
100	10,080

対訳辞書は，ver. 3 から ver. 15 のデータを CSV 形式で入手した．ver. 1 と ver. 2 は PDF 形式のみで公開されているため，本実験では使用しなかった．

評価方法　評価には，BLEU [6] と平均準拠率を用いる．平均準拠率とは，4.3 節で定義した準拠率のマイクロ平均であり，式 (2) で計算される．4.3 節で述べたように準拠率が未定義の文が存在するため，マクロ平均ではなく，マイクロ平均を用いる．

$$\text{平均準拠率} = \frac{\displaystyle\sum_{(s,t)\in C}\sum_{(p_s,P_t)\in D(t)} \min\left(\text{count}(p_s,s), \sum_{p_t\in P_t}\text{count}(p_t,t)\right)}{\displaystyle\sum_{(s,t)\in C}\sum_{(p_s,P_t)\in D(t)}\text{count}(p_s,s)} \tag{2}$$

翻訳モデル　翻訳モデルとして，事前学習済み mT5 [7] のベースサイズモデル[3)]を使用した．

5.2 結果

実験結果を表 5 に示す．

まず，ベースラインと各手法を比較する．手法 1 は，BLEU と平均準拠率ともにわずかにベースラインを上回った．手法 2 と手法 3 は閾値によって BLEU と平均準拠率が向上する場合もあるが，変化しない場合や低下する場合もある．特に，$\theta = 90$，100 では BLEU が低下した．次に，手法 2 と手法 3 を同じ閾値で比較すると，後者の方が全体的に BLEU，平均準拠率ともに高い．最後に，手法 1

3）https://huggingface.co/google/mt5-base

表 5　各手法を用いた実験結果

手法		閾値 θ	BLEU	平均準拠率 (%)
入力拡張なし	ベースライン	0	36.5	63.8
	手法 2	50	37.2	65.1
		60	37.2	66.4
		70	35.6	65.3
		80	34.7	66.5
		90	31.6	63.4
		100	26.3	56.8
入力拡張あり	手法 1	0	37.0	65.9
	手法 3	50	**37.7**	66.7
		60	37.6	68.3
		70	37.4	69.7
		80	36.1	**70.5**
		90	29.5	67.2
		100	30.3	70.3
参照訳			-	70.8

と手法 3 を比較する．手法 3 の $\theta = 50 \sim 70$ において BLEU，平均準拠率がともに向上した．一方で，$\theta = 90$，100 では平均準拠率は向上しているものの，BLEU が大幅に低下した．

出力された翻訳例を付録 A に示す．

5.3　考察

すべての手法において，表 7 の例 3 のように，参照訳は用語制約を満たしていないが，モデルの出力は用語制約を満たしているケースが見られる．このとき，出力が適切な訳文であっても BLEU が低下する．すなわち，参照訳が用語制約を満たしていないために BLEU が低下したと考えられる．

手法 1 は，ベースラインをわずかに上回ったのみであり，入力拡張の効果はわずかであった．入力拡張の目的は，第 3 節で述べたように<T>タグ内の訳語を出力に含めるようモデルに学習させることである．しかし，対訳辞書に準拠していない文がノイズとなり，その学習が十分にできなかったと考えられる．

手法 3 において，閾値が高いほど BLEU や平均準拠率が向上するわけではなかった．その原因は，閾値を上げれば学習データ数が減少するためと考えられる．表 4 に示したように $\theta = 100$ における学習データ数は 10,080 文であり，元データの約 13%しかない．

また，$\theta = 100$ の学習データは，似ている原文が多い．表 6 に 2 つの正規表現と，それにマッチした文数・割合を示す．2 つとも $\theta = 70$ より $\theta = 100$ の方が学習データに占める割合が高くなっており，似た文が多いことを示している．原文が似ている場合，訳文も似ることが予想される．すなわち，$\theta = 90$，100 では学習データの数が少ないうえに，似た対訳文が多いため，翻訳性能が低下したと考えられる．

表 6　正規表現とそれぞれにマッチした文数・学習データに占める割合

正規表現	文数（割合）	
	$\theta = 70$	$\theta = 100$
この.*は、.*から施行する。$	1,826 (3.8%)	1,487 (14.8%)
この法律において.*とは.*をいう。$	480 (1.0%)	156 (1.5%)

以上により，閾値を適切に設定したとき，データ選別と入力拡張を組み合わせる手法 3 が，それぞれ単体で使用する手法 1 や手法 2 よりも有効であることを確認した．

データ選別は，1 文に用語制約が多く含まれる場合に特に有効であると考える．仮に 1 文に含まれる用語制約が高々 1 つであるとすると，各対訳文の準拠率は 0%，100%，未定義の 3 通りとなる．このとき，閾値を $\theta = 100$ 以外に設定しても，結果は変化しないので，意味をなさない．一方で，用語制約が多いときには準拠率が多くの値をとるため，閾値の変更に伴って結果も変化し，適切な閾値を探索することが可能となる．JLT の対訳データは 1 文あたり 11.4 個の用語制約が含まれる．このような用語制約が多い翻訳において，データ選別は特に有効であると考えられる．

6　まとめ

本研究では用語制約のある法令翻訳に取り組んだ．JLT の対訳文には，対訳辞書に準拠していない訳文が存在するため，入力拡張の効果はわずかであった．そこで，準拠率が高い対訳文のみを抽出するデータ選別を提案した．閾値を適切に設定した場合，データ選別と入力拡張を組み合わせた手法が入力拡張単体より有効であることを実験により示した．

しかし，閾値は高いほどよいということではないため，適切な閾値の探索が必要であり，それにはコストがかかる．そこで，$\theta = 100$ の場合でもよい結果が得られるように，学習データを拡張することは今後の課題である．

謝辞

本研究は JSPS 科研費 JP21H03772，JP22H03901 の助成を受けたものです．

参考文献

[1] Ke Wang, Shuqin Gu, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. TermMind: Alibaba's WMT21 Machine Translation Using Terminologies Task Submission. **Proceedings of the Sixth Conference on Machine Translation**, pp. 851–856, 2021.

[2] Chris Hokamp and Qun Liu. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, pp. 1535–1546, 2017.

[3] Matt Post and David Vilar. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1314–1324, 2018.

[4] Elise Michon, Josep Crego, and Jean Senellart. Integrating Domain Terminology into Neural Machine Translation. **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 3925–3937, 2020.

[5] Melissa Ailem, Jingshu Liu, and Raheel Qader. Encouraging Neural Machine Translation to Satisfy Terminology Constraints. **Findings of the Association for Computational Linguistics:ACL-IJCNP2021**, pp. 1450–1455, 2021.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[7] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 483–498, 2021.

表 7　参照訳と出力の比較

例 1	秘密　　（参照訳：対訳辞書に**準拠**、出力：対訳辞書に**非準拠**） 役員　　（参照訳：対訳辞書に**準拠**、出力：対訳辞書に**準拠**）
原文	医療事故調査・支援センターの**役員**若しくは職員又はこれらの者であつた者は、正当な理由がなく、調査等業務に関して知り得た**秘密**を漏らしてはならない。
参照訳	No officer or employee of the medical accident investigation and support center, or any former **officer** or employee thereof must divulge any **secret** obtained in connection with the investigation services, etc. without justifiable grounds.
出力	A person who is or used to be an **officer** or employee of the Medical Safety Research and Support Center must not divulge any **confidential information** learned in connection with operations for investigations, etc., without legitimate grounds.
例 2	施行の日　　（参照訳：対訳辞書に**非準拠**、出力：対訳辞書に**非準拠**）
原文	この法律は、商法等の一部を改正する法律（平成十二年法律第九十号）の**施行の日**から施行する。
参照訳	This Act comes into effect on the effective date of the Act Partially Amending the Commercial Code (Act No. 90 of 2000).
出力	This Act comes into effect as of **the date on which the Act** Partially Amending the Commercial Code (Act No.90 of 2000) **comes into effect**.
例 3	執行官　　（参照訳：対訳辞書に**非準拠**、出力：対訳辞書に**準拠**）
原文	**執行官**は、次の事務を取り扱う。
参照訳	A **court enforcement officer** handles the following affairs.
出力	A **court execution officer** handles the following functions:

表 8　表 7 の例に出現する用語制約の一部

例	用語	訳語候補
1	秘密	secret | secrecy
2	役員	officer
3	施行の日	the date on which the Act comes into effect
4	執行官	court execution officer

A　生成された翻訳の例

手法 3（データ選別+入力拡張）において，$\theta = 70$ のとき BLEU，平均準拠率ともに比較的高い．そのため，手法 3 の $\theta = 70$ における出力例を表 7 に示す．それぞれの 1 行目には注目する用語制約を記載しており，その内容を表 8 に示す．表 7 の例 1 では，「秘密」の訳について参照訳は準拠しているが，出力は準拠していない．また，「役員」の訳は参照訳・出力ともに準拠している．例 2 の出力では，「施行の日」の訳語である「the date on which the Act comes into effect」に習って出力の最後に「comes into effect」が追加されている．ただし，訳語の途中に別の単語が含まれるため，対訳辞書に準拠していないと判定される．例 3 では，「執行官」の訳が参照訳は準拠していないが，出力では準拠している．