改正後法令文翻訳のための疑似三つ組コーパスの構築

山腰貴大 [1,3]　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
[3] リーガル AI 株式会社
yamakoshi@legalai.jp　ogawa@ds.nagoya-cu.ac.jp
toyama@is.nagoya-u.ac.jp

概要

法令の改正に伴い，その訳文を修正する際には，改正箇所のみを差分的に翻訳し直す必要がある．改正前の訳文（旧訳文）の表現をコピーして訳文を生成できる 2 入力 NMT は，差分的な翻訳を可能とするが，2 入力 NMT の訓練には原文，旧訳文，訳文の三つ組が必要であり，言語資源の調達コストが大きい．そこで，本研究では，法令文の対訳コーパスから疑似的な三つ組を構築する．対訳コーパスの訳文に対してフレーズをランダムに削除又は置換した文を旧訳文とみなし，元の原文，訳文と合わせて三つ組とする．構築した三つ組コーパスを 2 入力 NMT の訓練に使用したところ，三つ組の生成元である対訳コーパスにより訓練した NMT の性能を上回った．

1　はじめに

社会の基盤をなす法令は，一度制定すれば終わりではなく，社会の変化に応じて改正される．例えば，2024 年にわが国において制定された法律は 81 件であるが，そのうち 53 件は，もっぱら既存の法律の一部を改正する一部改正法律である．また，国際間取引の円滑化のために，法令翻訳の提供が重要であるが [1]，法改正の際には，改正後法令の訳文を速やかに提供することが望ましい．わが国では，法務省が日本法令外国語訳データベースシステム JLT (Japanese Law Translation Database System) [1)] によって主要法令の対訳を提供しているが，法改正のペースに英訳の修正が追いついていない．

改正後の法令文を翻訳する際には，改正前の訳文（旧訳文）をベースとし，改正箇所のみを差分的に翻訳し直す必要がある．旧訳文を不必要に修正すると，改正の趣旨が訳文の読者に的確に伝わらないからである [2, 3]．この要件（変更極小性 (focality)）は，改正後法令文の翻訳においては，流暢性，妥当性と並んで充足すべき要件である．変更極小性は，法令に限らず，改正によりアップデートする文書，例えば，利用規約，技術文書，製品マニュアルなどの翻訳でも重要となる．

機械翻訳システムは，訳文を低コストかつ迅速に生成できる．しかし，原文を入力し訳文を出力する一般的な機械翻訳システムは，旧訳文を入力としないため，変更極小性の充足が原理上困難である．そのため，本タスクにおいては，原文に加えて旧訳文を入力し，旧訳文の表現を適宜コピーしつつ訳文を生成できるニューラル機械翻訳（2 入力 NMT）の適用が望ましい．しかし，2 入力 NMT の訓練には，原文，旧訳文，訳文からなる三つ組が必要となり，言語資源の調達コストが大きい．

そこで，本研究では，2 入力 NMT の訓練データに使える疑似的な三つ組を構築する．具体的には，法令文の対訳コーパスから訳文を取り出し，その一部を編集したものを疑似的な旧訳文とみなすことにより，原文，旧訳文，訳文の三つ組を得る．訳文の編集方法は，(1) フレーズ（文の部分木）の削除，(2) フレーズの置換の 2 種類とする．編集対象のフレーズをランダムに選択することにより，一つの対訳から複数個の三つ組を生成する．これにより，元の対訳データの件数と比べて数倍の三つ組を生成できる．なお，本手法は，法令に限らず，係り受け解析を利用できる任意の文書ドメインに適用できる．

2　背景

法令の一部改正について 2.1 節で述べ，本研究で扱うタスクとその要件を 2.2 節で整理する．

2.1　法令の一部改正

法令の一部改正を行う際は，いわば「ソフトウェアパッチ」の要領で被改正法令を書き換える [4]．

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

第百六十四条①第四項を削り、②第三項後段を削り、③同項第一号中「の父母」を「（十五歳以上のものに限る。）」に改め、④同項第二号中「前号に掲げる」を「…に対し親権を行う」に改め、⑤同項第三号を削り、⑥同項を同条第六項とし、⑦同項の次に次の一項を加える。
7　特別養子適格の…（省略）

民法等の一部を改正する法律（令和元年法律第34号）から抜粋

図 1　改め文

すなわち，文字列や条項を置換，挿入，削除する指示（改め文）を一部改正法令において定め，改め文を改正対象に対して適用することにより法令を改正する．改め文の例を図 1 に示す．改め文によって規定される指示は，次のように類型化できる [5].

1. 文の一部に対する (a) 置換，(b) 挿入，(c) 削除
2. 条，項，号等の法令構造に対する (a) 置換，(b) 挿入，(c) 削除.
3. 法令構造に付与された番号の (a) 変更，(b) 割当，(c) シフト.
4. 法令構造の番号と見出しの変更

2.2　改正後法令文の翻訳

本研究は，前節の分類 1.（文の一部に対する修正）に伴う翻訳を扱う．以降，1 節で定義した旧訳文（改正前の訳文）に加えて，改正前の原文を旧原文，改正後の原文を新原文，その訳文を新訳文と定義し，旧原文，旧訳文，新原文，新訳文からなる組を四つ組という．本研究のタスクは，既存の旧原文と旧訳文を参考に，新原文から新訳文を翻訳するタスクである．その際，新訳文に求める要件は，流暢性（訳文が自然な言い回しであること）と妥当性（原文の意味が訳文に過不足なく含まれていること）だけではない．その例を表 1 に示す．

表 1 に示す二つの新訳文#1 と#2 は，どちらも，言い回しは自然であり，新原文の意味を反映している．すなわち，どちらも流暢かつ妥当である．しかし，新訳文#2 では，旧訳文中の “a document stating” を “writing describing” に変えているが，これは改正の対象ではない．そのため，日本語の分からない読者が旧訳文と新訳文#2 を読んだとき，上述の変更も改正によるものだと誤解するおそれがある．この点で，新訳文#2 は新訳文として適切ではない．したがって，本タスクにおいては，流暢性，妥当性に加えて，「改正の影響を受けていない部分の翻訳を不必要に変更しない」という要件も必要となる．本研究では，この要件を変更極小性 (focality) と呼ぶ.

3　関連研究

機械翻訳において変更極小性を高める方法に，ベースとなる訳文から単語をコピーする方法がある．統計的機械翻訳（SMT）でこの方法を導入した手法として，Koehn らの手法 [6] や，それを改正後法令文の翻訳に特化させた小酒井らの手法 [7] がある.

ニューラル機械翻訳（NMT）にこのような考えを導入した手法として，機械翻訳文を自動で後編集するタスクで考案された手法 [8, 9] や，改正後法令文翻訳への応用を考慮した手法 [3] がある．これらの手法は，原文に加えてベース訳文を入力し，訳文を生成する 2 入力 NMT を用いている．訳文を出力する際には，訳文の各単語について，原文の訳語として生成される尤度と，ベース文中からコピーされる尤度を合計し，その尤度がもっとも高い単語を出力する．2 入力 NMT の訓練には，原文，ベース訳文，訳文（改正後法令文翻訳タスクにおいては，新原文，旧訳文，新訳文）の三つ組が必要となる.

改正後法令文翻訳のための言語資源として，小酒井らが自らの手法の評価のために構築した四つ組コーパスがある [7]．このコーパスは，JLT の法令対訳データのうち，改正に伴う複数の版が存在する法令 17 件から構築され，四つ組 158 組を含む．四つ組から新原文，旧訳文，新訳文を抽出することにより，2 入力 NMT の訓練データに転用できる．しかし，NMT の訓練データとしては四つ組の数が極めて少ない上，変更極小性が完全に保たれていない [2]．人手翻訳により変更極小性を担保した四つ組コーパス [2] も存在するが，四つ組の数は 4,630 件にとどまり，NMT の訓練には十分とはいえない.

4　疑似三つ組コーパスの構築

NMT の訓練に耐えうる量の三つ組コーパスを得るために，本研究では，対訳コーパスから三つ組を機械的に生成する．具体的には，対訳コーパス中のペア (原文 s_J，訳文 s_E) に対して，s_E 中のフレーズ（文の部分木となる単語列）を文字列操作した文 s'_E を作成し，疑似的な三つ組 (s_J, s'_E, s_E) を得る．ここで，s_J, s'_E, s_E は，それぞれ新原文，旧訳文，新訳文とみなす．なお，2 入力 NMT は，訓練の際に旧原文を要しないため，s'_E の原文は作成しない.

フレーズの操作は，(1) フレーズの削除，(2) フレーズの置換の二つを定める．操作 (1) において，訳文 s_E は，ベース訳文 s'_E にフレーズを追加した

表 1　新訳文の比較（実線の下線が改正箇所）

種類	文
旧原文	前項の申立ては、海難の事実を示して、書面でこれをしなければならない。
新原文	前項の申立ては、海難の事実及び受審人に係る職務上の故意又は過失の内容を示して、書面でこれをしなければならない。
旧訳文	The request shall be made in a document stating the facts of the marine accident.
新訳文#1	The request shall be made in a document stating the facts of the marine accident and the details of the intentional or negligent act committed in the course of duties of the examinee.
新訳文#2	The request shall be made in writing describing the facts of the marine accident and the details of the intentional or negligent act committed in the course of duties of the examinee.

文となるため，この操作による三つ組をフレーズ追加型（Phrase Addition; PA）三つ組という．また，操作 (2) により生成した三つ組は，フレーズ置換型（Phrase Substitution; PS）三つ組という．

PA 三つ組，PS 三つ組の生成手順をそれぞれ 4.1 節，4.2 節にて説明する．いずれの操作も，汎用的な係り受け解析と文字列操作のみで実現可能であり，文書ドメイン固有の性質・知識は使用しない．

4.1　フレーズ追加型三つ組の生成

PA 三つ組の生成手順を図 2 に示す．PA 三つ組は，次の手順で構築する．

1. s_E の係り受けを求める．
2. s_E から単語をランダムに選択する．
3. 当該単語を主辞とするフレーズを求める．
4. 求めたフレーズを削除し，フレーズ削除後の文を s'_E とする．

PA 三つ組の生成において，2 入力 NMT の訳文生成能力を向上させるため，以下の工夫を施す．

- 各単語について，その単語を主辞とするフレーズの長さに応じて選択確率を高める．これにより，広範囲を削除しやすくする．なお，文の主辞が選択された場合，文全体が削除されるため，原文のみから訳文を生成することになる．
- 各対訳について，処理対象の単語を複数回サンプリングする．すなわち，手順 2. から手順 4. を複数回繰り返す．これにより，一つの対訳から複数個の三つ組を生成できる．

4.2　フレーズ置換型三つ組の生成

PS 三つ組の生成手順を図 3 に示す．PS 三つ組は，次の手順で構築する．

1. 単語からその単語を主辞とするフレーズにマッピングするリスト（フレーズリスト）をあらかじめ作成する．
2. s_E の係り受けを求める．
3. s_E から単語 w をランダムに選択する．
4. w を主辞とするフレーズ p を求める．
5. フレーズリストから w が主辞のフレーズ p' をランダムに選択する．
6. s_E 中の p を p' に置換する．

フレーズ置換後の文のトピックを元の文のトピックと似たものにするため，手順 3. において選択する単語を名詞に限定する．また，PA 三つ組と同様，処理対象の単語は複数回サンプリングする．

5　実験

本稿の疑似三つ組コーパスを用いて 2 入力 NMT を訓練し，他のシステムと性能を比較する．

5.1　実験設定

比較するシステムの一覧を表 2 に示す．

表 2 の「翻訳手法」の列において，「Transformer」は，1 入力の Naive な Transformer [10]，「SMT」は，改正後法令文翻訳に特化した SMT ベースの手法 [7]，「CTT」は，改正後法令文翻訳への応用を考慮した 2 入力 NMT [3] である．

表 2 の「訓練データ」の列において，「JLT 対訳」は JLT から取得した対訳コーパス，「人手四つ組」は変更極小性を担保した四つ組コーパス [2] 中の訓練データである．「疑似三つ組」は本手法によって構築した疑似三つ組コーパスであり，「JLT 対訳」を用いて構築した．なお，PA 三つ組および PS 三つ組のサンプリング数は，予備実験（付録 A）の結果に基づき，それぞれ 3 および 1 と定めた．

ここで，三つのシステム Trm，Kozakai，CTT_{gen} の訓練データは，いずれも同じ言語資源「JLT 対訳」から構築されているため，これらの比較により，2 入力 NMT の有効性を検証できる．

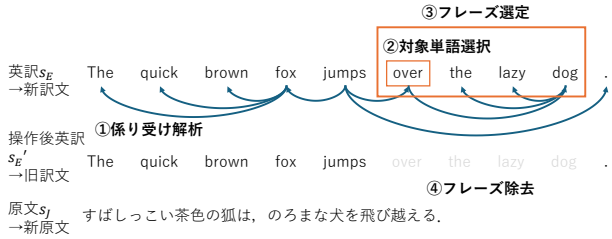

図 2　フレーズ追加型三つ組の生成

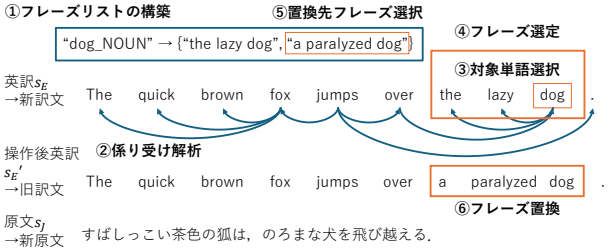

図 3　フレーズ置換型三つ組の生成

表 2　比較システム（★が本手法）

名称	翻訳手法	訓練データ（件数）
Trm	Transformer	JLT 対訳 (391,758)
Kozakai	SMT	JLT 対訳 (391,758)
CTT_{hm}	CTT	人手四つ組 (4,297)
CTT_{gen} ★	CTT	疑似三つ組 (1,562,316)

表 3　実験結果

システム	BLEU	RIBES	ISDIT
Trm	80.31	93.44	71.36
Kozakai	82.79	92.04	77.53
CTT_{hm}	74.46	93.19	77.64
CTT_{gen}	**85.39**	**96.41**	**84.94**

テストデータとして，前述の四つ組コーパス [2] のテストデータ 201 組を用いた．各手法は，BLEU [11]，RIBES [12]，ISDIT [2] により評価した．訓練時の主なハイパーパラメータを付録 B に示す．

5.2　結果

実験結果を表 3 に示す．CTT_{gen} が BLEU，RIBES，ISDIT のいずれにおいても最大スコアを記録した．CTT_{gen} の訓練に用いた「疑似三つ組」は，Trm，Kozakai の訓練に用いた「JLT 対訳」から生成したものである．すなわち，疑似三つ組の生成と 2 入力 NMT の利用によって，同じ言語資源からより高性能な機械翻訳モデルを構築できたことが示された．

出力例を表 4 に示す．本手法である CTT_{gen} は，変更極小性が高く正しい翻訳を生成した．一方，Trm および Kozakai は，旧訳文中の「審査請求人」“the requestor for review” を旧訳文とは異なる訳 “the applicant for ···” に変えており，変更極小性が保たれていない．CTT_{hm} は旧訳文をそのまま出力した．

表 4　出力例（実線の下線が改正箇所）

種類	文
旧原文	経済産業大臣は、前条の意見の聴取の期日及び場所を定め、審査請求人又は異議申立人に通知しなければならない。
新原文	経済産業大臣は、前条の意見の聴取の期日及び場所を定め、審査請求人に通知しなければならない。
旧訳文	the minister of economy , trade and industry must specify the date and place of the hearing of opinions prescribed in the preceding article , and notify the requestor for review or the petitioner for objection .
参照訳	the minister of economy , trade and industry must specify the date and place of the hearing of opinions prescribed in the preceding article , and notify the requestor for review .
Trm	the minister of economy , trade and industry must specify the date and place of the hearing of opinions prescribed in the preceding article , and notify the applicant for examination of such date and place .
Kozakai	the minister of economy , trade and industry must specify the date and place of the hearing of opinions prescribed in the preceding article , and notify applicant for examination for the .
CTT_{hm}	the minister of economy , trade and industry must specify the date and place of the hearing of opinions prescribed in the preceding article , and notify the requestor for review or the petitioner for objection .
CTT_{gen}	the minister of economy , trade and industry must specify the date and place of the hearing of opinions prescribed in the preceding article , and notify the requestor for review .

6　まとめ

本研究では，改正後法令文翻訳のための疑似三つ組コーパスを構築した．対訳コーパスの訳文からフレーズを削除又は置換し，それを旧訳文とみなすことにより，新原文および旧訳文を入力とする 2 入力 NMT の訓練データとして利用できるようにした．実験において，本コーパスにより訓練した 2 入力 NMT が高水準な翻訳を生成できることを示した．

今後は，旧訳文がない場合でも 2 入力 NMT によって訳文を生成できるかを調査し，新設法令文・改正後法令文のいずれにも対応した翻訳手法の実現を目指す．また，法令文以外のドメインで疑似三つ組コーパスを構築し，その有効性を調査したい．

謝辞

本研究は JSPS 科研費 JP23K25155 の助成を受けた.

参考文献

[1] 外山勝彦, 小川泰弘. 自然言語処理の応用に基づく法令外国語訳支援. 人工知能学会誌, Vol. 23, No. 4, pp. 521–528, 2008.

[2] Takahiro Yamakoshi, Takahiro Komamizu, Yasuhiro Ogawa, and Katsuhiko Toyama. Evaluation scheme of focal translation for Japanese partially amended statutes. In **Proceedings of the 8th Workshop on Asian Translation (WAT2021)**, pp. 124–132, 2021.

[3] Takahiro Yamakoshi, Yasuhiro Ogawa, and Katsuhiko Toyama. Differential-aware transformer for partially amended sentence translation. In Yasufumi Takama, Katsutoshi Yada, Ken Satoh, and Sachiyo Arai, editors, **New Frontiers in Artitificial Intelligence, JSAI 2022 Conference and Workshops, Revised Selected Papers, Lecture Notes in Computer Science**, Vol. 13859, pp. 5–22, 2023.

[4] 法制執務研究会（編）. 新訂ワークブック法制執務 第 2 版. ぎょうせい, 2018.

[5] Yasuhiro Ogawa, Shintaro Inagaki, and Katsuhiko Toyama. Automatic consolidation of Japanese statutes based on formalization of amendment sentences. **New Frontiers in Artificial Intelligence: JSAI 2007 Conference and Workshops, Revised Selected Papers, Lecture Notes in Computer Science**, Vol. 4914, pp. 363–376, 2008.

[6] Phillip Koehn and Jean Senellart. Convergence of translation memory and statistical machine translation. In **AMTA Workshop on MT Research and the Translation Industry**, pp. 21–31, 2010.

[7] 小酒井款雄, 小川泰弘, 大野誠寛, 中村誠, 外山勝彦. 新旧対照表の利用による法令の英訳修正. 言語処理学会第 23 回年次大会論文集（NLP2017）, pp. 859–862, 2017.

[8] Xuancheng Huang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. Learning to copy for automatic post-editing. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 6122–6132, 2019.

[9] Marcin Junczys-Dowmunt and Roman Grundkiewicz. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, pp. 822–826, 2018.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of Advances in Neural Information Processing Systems 30**, pp. 6000–6010, 2017.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 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, 2002.

[12] 平尾努, 磯崎秀樹, 須藤克仁, Kevin Duh, 塚田元, 永田昌明. 語順の相関に基づく機械翻訳の自動評価法. 自然言語処理, Vol. 21, No. 3, pp. 421–444, 2014.

A　予備実験

本節では，2 入力 NMT 用の疑似三つ組コーパスの構築に先立ち，PA 三つ組と PS 三つ組の最適なサンプリング数を調査する．

A.1　三つ組ファイルの構築

JLT から取得した対訳文 391,758 組から三つ組ファイルを生成する．PA 三つ組について，サンプル数を 0 から 4 まで 1 刻みで増やすことにより，5 件の三つ組ファイル PA_0, PA_1, PA_2, PA_3, PA_4 を生成する．PS 三つ組についても，サンプル数を 0 から 4 まで変化させた 5 件の三つ組ファイル PS_0, PS_1, PS_2, PS_3, PS_4 を生成する．ここで，PA_0 と PS_0 は，それぞれ空のファイルである．

係り受けの取得には，spaCy の ‘en_core_web_sm’ モデルを使用した [2)]．また，PS 三つ組において使用するフレーズリストは，三つ組の構築に用いた JLT の対訳文の英訳 391,758 文から作成した．

各ファイルに含まれる三つ組の数は，表 5 のとおりである．置換先が存在しないなどの理由により三つ組が生成できなかったケースがあるため，PA_n や PS_n は，391,758 の単純な n 倍にはならなかった．

A.2　翻訳モデルの訓練

前節で構築した三つ組ファイルに対し，(PA_0, PS_0) を除く 24 通りの組み合わせ (PA_0, PS_1), (PA_0, PS_2), $\cdots$, (PA_4, PS_4) を求め，各ファイルを結合した 24 件の三つ組ファイルを作成した．

各三つ組ファイルを用い，2 入力 NMT である CTT [3] を訓練した．訓練時の主要なハイパーパラメータを 付録 B に記載する．翻訳性能の評価用のデータとして，変更極小性を担保した四つ組コーパス [2] において開発データとして確保された四つ組 132 組（法改正 8 件分）を用いた．生成された訳文は，BLEU [11] によって評価した．

表 5　生成した三つ組の数

ファイル	三つ組数	ファイル	三つ組数
PA_0	0	PS_0	0
PA_1	391,562	PS_1	387,630
PA_2	783,125	PS_2	775,260
PA_3	1,174,686	PS_3	1,162,890
PA_4	1,566,248	PS_4	1,550,520

2）　https://spacy.io/models/en#en_core_web_sm

表 6　予備実験の結果

	PA_0	PA_1	PA_2	PA_3	PA_4
PS_0	—	83.84	85.55	85.61	85.81
PS_1	81.23	85.27	86.44	**87.23**	86.61
PS_2	83.38	85.42	86.15	86.48	86.07
PS_3	82.66	86.16	86.56	87.13	86.31
PS_4	82.52	85.89	85.18	85.76	86.52

A.3　結果

各三つ組ファイルにより訓練した CTT の BLEU スコアを表 6 に示す．表 6 より，(PA_3, PS_1) の組み合わせによる三つ組を用いた場合がもっとも高いスコアとなった．また，PS を増やすよりも，PA を増やす方が高いスコアとなる傾向がみられた．

B　ハイパーパラメータ

CTT の事前学習（付録 A）における主なハイパーパラメータは，次のとおりである．

- 隠れ層の数：6
- アテンションヘッド数：8
- 次元数：512
- バッチサイズ：8
- ドロップアウト率：0.1
- 最大系列長：256
- 反復数：200,000
- ビームサーチ数：4

本学習（5 節）における主なハイパーパラメータは，反復数を 2,000,000（Trm, CTT_{gen}），20,000（CTT_{hm}）にした以外は上記と同じである．