執筆・翻訳のための制限語彙の構築とその自動化の検討

杉野 峰大　　宮田 玲　　小川 浩平　　佐藤 理史
名古屋大学大学院工学研究科
sugino.hodai@c.mbox.nagoya-u.ac.jp

1 はじめに

工業製品のマニュアルなど産業文書は、大規模になることが多い上、製品のリリースに合わせて短期間で完成させることが求められる。また、制作・多言語展開のプロセスには複数の編集会社・翻訳会社が携わり、表現の一貫性を保つことが難しいという問題もある。このような課題を解決するための枠組みとして制限言語がある。制限言語により、構文や語彙を制限し、曖昧な表現を防ぐことで、テキストの一貫性や翻訳可能性の向上が期待できる [1, 2]。ここで肝心なことは、執筆者・編集者・翻訳者が各作業プロセスにおいて共通に参照できるルールや辞書を事前に整備し、その利用を技術的に支援することである。

制限語彙は、制限言語の重要な要素の１つであり、特定のドメイン向けに設計された使用可能な語のリストのことである [1, 3]。予め、使用できる語を規定しておくことによって、執筆者間の語彙使用のばらつきを抑制し、文書全体を通した表現の一貫性を向上させることができる。これまで、主に産業文書の執筆のために各種の制限語彙が構築されてきたが [4, 5]、一般に公開されているものは多くない。公開されている制限語彙の 1 つに ASD Simplified Technical English（ASD-STE 100）[6] がある。これは、もともと航空宇宙分野の整備マニュアル用に開発されたもので、現在では産業界で広く使用されている。8 つの品詞（動詞、名詞、形容詞、副詞、冠詞、前置詞、接続詞、代名詞）を対象に、承認語 (approved word) と非承認語 (unapproved word) のリストが提供されている。

我々は、ASD-STE 100 をモデル例と捉え、自動車の修理書を対象とした日英対訳の制限語彙の構築に取り組んでいる。ASD-STE 100 は有用な枠組みではあるが、他の目的、ドメイン、言語に直接移植することは容易ではない。また、どの単語を使うべきかについては規定されているが、それらをどのように使うかに関する指針は必ずしも十分ではなく、執筆や翻訳に役立てるには改良の余地がある。さらに、単言語（英語）での提供であるため、多言語に制作される文書に対しては効果が限定的といえる。

そこで、本稿では、ASD-STE 100 の枠組みを拡張し、動詞を対象とした執筆・翻訳のための日英対訳の制限語彙の形式を提案し、人手構築の現状を報告する（2 節）。提案する制限語彙は、執筆支援や翻訳支援・機械翻訳への応用が考えられる（3 節）。さらに、制限語彙の人手構築で得た知見を踏まえ、構築作業の自動化に取り組んだ結果を報告する（4 節）。

2 制限語彙の設計と人手構築

2.1 設計

本研究では、執筆や翻訳に特に関わる文の主要な構成要素であり、なおかつ格などの構文に関する情報を付与できる、動詞を対象とする。

表 1、表 2 に提案する自動車修理書を対象とした制限語彙の形式を示す。表 1 は承認語、表 2 は非承認語のエントリの例である。この形式は ASD-STE 100 の形式を基に、以下の 3 つの観点から拡張したものである。

表現文型　各動詞について取りうる格の順序と選択選好（意味カテゴリ）[7, 8] の情報を与える。これらを合わせて動詞の表現文型と呼ぶことにする。日本語では、格順序が比較的自由であり、

1. GTS を DLC3 に 接続する
2. DLC3 に GTS を 接続する

という 2 文は格の順序が入れ替わっているが、いずれも文法的には正しい。また、これらの違いは必ずしも読者の理解に大きな影響を与えるわけではない。しかし、一貫性のある文書を作成する観点から、格順序を制限し、これらの揺れを取り除くことは望ましく、テキスト検索の効率や機械翻訳出力の一貫性の向上にも貢献する。

表 1　承認語のエントリ例

見出し	交換する
品詞 (カテゴリ)	動詞 (Action ->Part)
表現文型	{[Part] を} {[Part] に} 交換する (**内容要素：動作**)
用例	センサーを新品に交換する。
英語 承認語	replace
英語 表現文型	replace {[Part] を} with {[Part] に}
英語 用例	Replace the occupant detection ECU with a new one.

表 2　非承認語のエントリ例

見出し	取り替える
品詞 (カテゴリ)	動詞 (Action ->Part)
代替承認語	交換する
用例	必ず新品に交換する。

また、選択選好を規定することで、用語集との連携を強化できる。技術文書では、パーツ名などの専門用語の使用頻度が高く、一貫性を保つために用語集で管理することが望ましい。表現文型の格によく使われる語句の意味カテゴリの情報を付与し、用語集と対応させることで用語集の検索や利用を円滑にする（定義しているカテゴリは付録 A を参照）。

文書構造／内容要素　我々は、すでに情報の配置に関するルールとなる修理書の文書構造やそれに対応した内容要素を定めている [9]（一覧は付録 B を参照）。これらを表現文型を対応させることで、構造と表現が結びつき、表現文型の利用をより円滑なものにする。

多言語対応　日本語の承認語に対し、多言語の承認語を対応させておくことで翻訳時の辞書として利用でき、目標言語での一貫性を高めることにつながる。今回は、日本語のエントリを起点に英語の承認語と表現文型を定める。

2.2　制限語彙の人手構築

制限語彙構築の基本要件は以下の通りである（構築手順の詳細は、文献 [10] を参照）。

1. 語彙の全集合により当該ドメインのあらゆる内容を表現できる。
2. ある意味内容に対する表現形が 1 つに定まる。

ある意味内容に対する表現形を集めるには、単に辞書的な類義語を取得するのみでは不十分である。例えば、「使用する」と「着用する」は一見、類義語ではないが、「この作業では、必ず手袋を使用すること」と「この作業では、必ず手袋を着用すること」では、同じ内容を指している。そのため、用例における入れ替え可能性に着目し、その語と入れ替え可能な語があれば、どちらか一方を承認語とする。

現在、トヨタ自動車株式会社から提供を受けた 10 車種 17 タイプの自動車修理書から抽出した 1,053,111 文の日本語文（以下、修理書コーパス）を用いて人手により日本語の制限語彙を構築している。複数の修理書から網羅的に文を抽出しており、当該ドメインを広くカバーしていると考えられる。

自動車修理書で使用される動詞全 852 種類を精査し、承認語 717 語と非承認語 135 語を規定した。異なりでは 15.9%、延べでは 3.1%が非承認語であった。これは、プロの執筆者、編集者によって作成された文書であっても、さらなる語彙統制の余地があることを示唆する。

また、717 の承認語に対して、900 の表現文型を規定した。全文型により、現時点で修理書コーパス中の文を延べで 85.2%カバーしているが、異なりでは 40.3%のカバレッジにとどまる。これは、数多く存在する頻度が少ない文型を十分網羅できていないことを意味する。また、格順序が規定のものと異なるバリエーションは延べで 2.7%存在した。

3　執筆・翻訳への応用

執筆・翻訳において重要な役割を果たす表現文型はよく使われる表現である基本形のみが定義されている。実際に使用する際には、文書構造の位置によって望ましい形に変形する必要がある。そこで、そのメカニズムを変形規則として実現した（表現文型の変形規則一覧は付録 B を参照）。以上を踏まえつつ、具体的な応用について述べる。

3.1　執筆支援

文を一から執筆するシナリオでは以下のプロセスからなる執筆支援方法が有効であろう。

1. 書きたい内容に応じた文書構造要素を選ぶ。
2. 文の核となる動詞を入力し、動詞と内容要素に対応した表現文型を呼び出す。
3. 表現文型のスロットを埋める。

表現文型のスロットを埋める形式で文を書くことで執筆者間で一貫性を保てる。さらに、文書構造、内容要素、表現文型がトップダウンに対応しており、文レベルではなく文書レベルでの制限オーサリングを支援することが可能になる。

他にも、構築した制限語彙は、既存テキストの診断シナリオにも活用できる。具体的には、非承認語の使用と、表現文型に違反する格の順序を検出し、それぞれ正しい表現候補を提案することで執筆者や編集者を支援できる。

3.2 翻訳支援・機械翻訳

「{[Part] を} {[Part] に}交換する」と「replace {[Part] に} with {[Part] を}」など言語間で表現文型の対応がある。格要素は対訳用語集として整備しておく。表現文型と用語集は、人手翻訳ではガイドラインとして使えるだけでなく、的確な変換規則を定めればテンプレートベース機械翻訳にも応用できる。

機械翻訳について、あらゆる入力文に対し、表現文型による翻訳を適用できるわけではない。現在、日本語の文型のカバレッジは約 85%である。表現文型の数を増やすことで、カバレッジを高めることは可能であるが、その分適切な管理が難しくなる。そこで、表現文型でカバーできない表現にはニューラル機械翻訳を用いることで、一貫性と柔軟さを兼ね備えた機械翻訳を実現できると考えている。

4 自動構築の検討

制限語彙は有用であるが、人手による構築は時間や労力がかかる。また、1000 種類以上の語を対象に制限語彙の構築を人手で行う中で、見落としも生じる。完全な自動化は難しいが制限語彙構築の機械的支援が求められる。本節では、承認語、非承認語決定の自動化の検討と実験について報告する。

4.1 自動構築手法

2.2 節で述べた制限語彙の構築方針を以下の手続きとして具体化した。

1. 注目する動詞の類義語を収集する。
2. 注目する動詞の用例を用いて、類義語の入れ替え可能性を検証する。
3. 入れ替え可能であれば、承認語を 1 つに決定する。

自動化への課題は、大きく 2 つある。1 つ目は、用例に即した入れ替え可能性をどのように自動で判定するかである。本研究では、BERT [11][1)] を利用する。BERT は同じ語であっても文脈により異なる埋め込み表現を与えるため、用例に対する入れ替え可能性を測ることに利用できると考えた。

2 つ目は、何を基準に承認語を決定するかである。基準としては、例えば、使用頻度、平易さ、体系性が挙げられる。使用頻度、平易さはそれぞれよく使われる表現、わかり易い表現に統一するという考えから重要であるといえる。また、辞書としての体系性も重要な要素と考えられる。例えば、「送る」、「受信する」が承認語、「送信する」、「受け取る」が非承認語であるのは不自然であり、承認語と非承認語の決定は対称的になっていることが望ましい。本来はこのような要素を複合的に考慮すべきであるが、今回の自動化では人手構築においても特に重視した使用頻度を基準とする。

本手法の入力は制限語彙を構築したい文書群、出力は、動詞のみを対象とした承認語のリスト及び非承認語と代替承認語のペアのリストの 2 つである。処理の全体像を図 1 に示す。この処理は対象コーパス中における頻度上位の語から順に行う。以下、各モジュールの詳細を述べる。

(1) 置換候補抽出　対象動詞の代替承認語の候補として対象コーパス中で対象動詞より出現頻度が多い動詞から以下のどちらかを満たすものを選ぶ。

- WordNet[2)] における類義語
- word2vec[3)] で対象動詞との cos 類似度が閾値 α 以上で上位 30 件以内かつ WordNet における対義語でない語

(2) 用例サンプリング　対象動詞を含むコーパス中の全用例を検証するのは実行時間の点から現実的ではないため、用例のサンプリングを行う。動詞は複数の用法で使用されることもあり、それらを広く含むようサンプリングする必要がある。本研究では、BERT のベクトルを利用した用例のクラスタリング [12] を利用する。なお、クラスタ数は未知であるため、上限クラスタ数を 5 とし、x-means 法によってクラスタリングを行う。得られたクラスタの重心に最も近い用例を取得することによって、多様な用法を含むサンプリングを実現する。

(3) 入れ替え可否判定　Panzer [13]、HaoriBricks3 [14] を利用して、(2) で得た用例（原文）の動詞を (1) で得た置換候補に置き換えた文（置換文）を生成

1) 今回は、日本語では https://github.com/cl-tohoku/BERT-japanese で公開されている BERT-base-mecabipadic-bpe-32k whole-word-mask モデル、英語では https://github.com/google-research/bert で公開されている BERT-Large Cased (Whole Word Masking) モデルを用いる。

2) http://compling.hss.ntu.edu.sg/wnja/

3) Wikipedia で学習したモデルを対象コーパスで追加学習した。

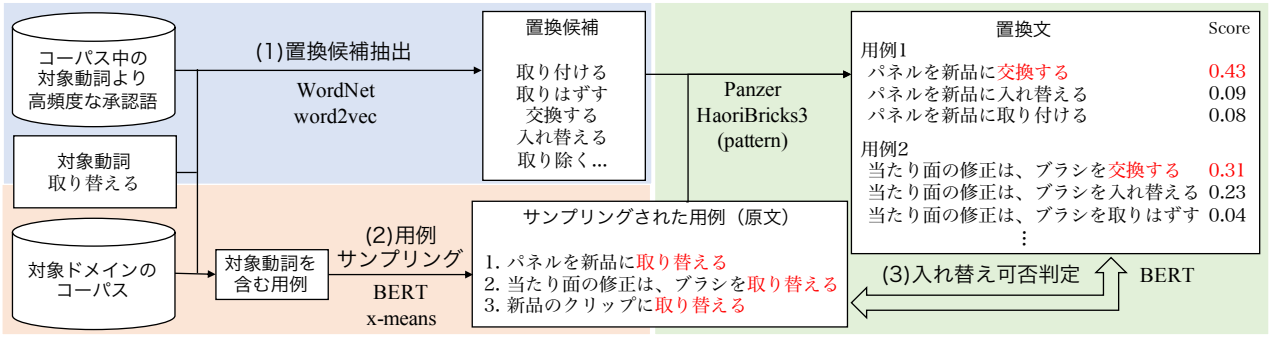

図 1　制限語彙の自動構築の流れ（動詞「取り替える」と承認語「交換する」が入れ替え可能か判定する例を示す。）

表 3　自動で取得した非承認語と代替承認語のペアに対する評価結果

日本語		英語
人手構築との一致率	精度	精度
14/99 (0.141)	63/99 (0.636)	42/61 (0.689)

する。なお、英語では、同様の処理に pattern [15] を用いた。次に、入れ替え可否判定のために、BERT ベースの平易化手法 [16] を参考に原文と置換文の動詞部分を [MASK] トークンで置き換えた文を結合して BERT に入力する。

[MASK] に置換文の動詞が出現する尤度を p、[MASK] に原文の動詞が出現する尤度 p' として、入れ替え可能性スコアは以下のように定義する[4]。

$$\text{Score} = \begin{cases} p & \text{(動詞部分の token 数が異なる場合)} \\ \frac{p}{1-p'} & \text{(動詞部分の token 数が同じ場合)} \end{cases}$$

スコアが候補中で最大かつ閾値 β 以上であれば、用例において動詞の入れ替えが可能と判定する。

(2) で得た全ての用例のについて入れ替え可能な語があれば、それらを代替承認語とし、対象動詞を非承認語とする。

4.2　検証

自動車分野の修理書コーパス[5]を用いて、日本語と英語で制限語彙の自動構築を行った。

今回は、閾値 α, β はともに 0.2 と設定した。自動で同定した非承認語と代替承認語のペアに対する評価結果を表 3 に示す。日本語において、人手構築結果（2.2 節）との一致度は 14.1%と低いが、精度は 63.6%と人手構築の判断材料として使える水準であった。制限語彙の作り方は 1 通りには定まらず、必ずしも人手で構築したものが最良とは言えないことを示唆する。特に、人手では承認語としていた動詞の内、自動構築によって新たに 41 件に対し妥当な代替承認語を獲得することができた。人手と自動構築の違いとして、人手の方が漢語から和語への言い換えを好む傾向が見られた。また、英語においても、68.9%と日本語と同程度の精度が確認できた。

エラーとして、対義語が代替承認語と判定されている事例が複数存在した。原因は、BERT や word2vec が分布仮説の考え方に従っており、同義語と対義語を十分区別できないためと考えられる。今回は候補抽出の段階で WordNet で獲得できる対義語は除外したが、それでも多くの対義語が代替承認語として出力されている。解決策としては、対義語辞書の拡充など対義語判定手法の改良が考えられる。

5　おわりに

本稿では、執筆・翻訳作業において制限語彙を有効に活用するために、ASD-STE 100 の枠組みを拡張し、新たな制限語彙の形式を提案した。本制限語彙は、執筆・翻訳作業の支援システムへの様々な応用が可能である。さらに、制限語彙の構築作業の自動化に取り組み、人手作業の支援に利用できる一定の精度が得られることを確認した。対義語への対処は今後の課題である。

今後は、表現文型を起点として、執筆・翻訳支援の枠組みを発展させる。具体的には、(1) 副詞句などの修飾要素の種類と文中での順序を整理すること、(2) 複文の構成に関して節間接続表現の列挙と規則を定めることに取り組む。また、これらの要素を執筆や翻訳に活用するためのシステムの開発を行う。

謝辞　研究データの自動車修理書はトヨタ自動車株式会社からご提供いただいた。本研究の一部は科研費（19H05660, 19K20628）の支援を受けた。

4）　BERT に同じ長さの文を結合した 2 文を入力した場合、2 文目の [MASK] に対して、1 文目と同じ token を予測する働きが強まるため、入れ替え可能な他の表現の出現尤度が低くなる。この BERT の挙動の影響を補正するため、場合分けを行う。

5）　日本語は 2.2 節と同じ 1,053,111 文、英語は 10 車種 15 タイプの自動車修理書から抽出した 1,202,738 文を利用した。

参考文献

[1] Eric Nyberg, Teruko Mitamura, and Willem-Olaf Huijsen. Controlled language for authoring and translation. In Harold Somers, editor, *Computers and Translation: A Translator's Guide*, pp. 245–281. John Benjamins, Amsterdam, 2003.

[2] Tobias Kuhn. A survey and classification of controlled natural languages. *Computational Linguistics*, Vol. 40, No. 1, pp. 121–170, 2014.

[3] Kara Warburton. Developing lexical resources for controlled authoring purposes. In *Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use*, pp. 90–103, 2014.

[4] Linda Means and Kurt Godden. The Controlled Automotive Service Language (CASL) project. In *Proceedings of the 1st International Workshop on Controlled Language Applications*, pp. 106–114, 1996.

[5] Kurt Godden. The evolution of CASL controlled authoring at General Motors. In *Proceedings of the 3rd International Workshop on Controlled Language Applications*, pp. 14–19, 2000.

[6] ASD Simplified Technical English. Specification ASD-STE100, Issue 7., （2020-12 閲覧）. http://www.asd-ste100.org.

[7] Philip Resnik. Selectional preference and sense disambiguation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pp. 52–57, 1997.

[8] Yorick Wilks. An intelligent analyzer and understander of English. *Communications of the ACM*, Vol. 18, No. 5, pp. 264–274, 1975.

[9] Hodai Sugino, Rei Miyata, and Satoshi Sato. Formalising document structure and automatically recognising document elements: A case study on automobile repair manuals. In Adam Jatowt, Akira Maeda, and Sue Yeon Syn, editors, *Digital Libraries at the Crossroads of Digital Information for the Future. ICADL 2019. Lecture Notes in Computer Science*, pp. 249–262. Springer, Cham, 2019.

[10] Rei Miyata and Hodai Sugino. Building a controlled lexicon for authoring automotive technical documents. In *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, pp. 171–180, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] 馬ブン, 田中裕隆, 曹鋭, 白静, 新納浩幸. Bert を利用した単語用例のクラスタリング. 言語資源活用ワークショップ発表論文集, 2019.

[13] 佐野正裕, 佐藤理史, 宮田玲. 文末述語における機能表現検出と文間接続関係推定への応用. 言語処理学会第 26 回年次大会発表論文集, pp. 1483–1486, 2020.

[14] 佐藤理史. HaoriBricks3: 日本語文を合成するためのドメイン特化言語. 自然言語処理, Vol. 27, No. 2, pp. 411–444, 2020.

[15] Tom De Smedt and Walter Daelemans. Pattern for Python. *The Journal of Machine Learning Research*, Vol. 13, No. 1, pp. 2063–2067, 2012.

[16] Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu. Lexical simplification with pretrained encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 8649–8656, 2020.

A　格要素の意味カテゴリ一覧

No.	カテゴリ		説明	例
1	Act		作業者、パーツの動作	車両の節電制御, タップアップ操作
2	Data		ソフトウェアによって出力されるデータ	ダイアグコード, 車両制御履歴
3	Direction		作業中の動作の方向	下方向, 時計方向
4	Feature		電流値などの特性値や設定値	抵抗値, 電圧値
5	Location		作業の対象となる場所や位置	タイヤの内側, 点が重なった場所
6	Meta		修理書内の別項目	以下の表, 手順 2
7	Part		脱着可能なパーツ	カバーパネル, ジャンクションブロック
8	Phenomenon		パーツに関連する物理現象	光, 熱, 電圧, 電流
9	Quantity		具体的な数値	10V, 2500rpm
10	Service		車両などとは直接関係のないサービス	オーナーズデスク, メール配信サービス
11	Signal		パーツ間の通信に用いられる信号	作動要求信号, 操作信号
12	Software		ソフトウェアツール、システム	ＧＴＳの選択画面, プルダウンリスト
13	State		作業者、パーツの状態	判定結果が正常になること, やけど
14	Supply		パーツに作用する消耗品	ナット, エンジンオイル
15	Tool		作業者が作業時に用いる工具	SST, レンチ
16	General	Car	車両	車両, 他車
17		Document	作業中に用いる書類	問診票
18		Gas	気体	CO2, 有害ガス
19		Group	企業や店舗	取扱店, ディーラー
20		Item	抽象的なもの	異物, 遮蔽物
21		Liquid	液体	水, 水滴
22		Person	作業者などの人間	ドライバー, お客様
23		Rule	法律や規則	道路交通法, 法規
24		Sound	音	作動音、アラーム音
25		Time	時間	作業にかかる時間

B　文書構造、内容要素、表現文型変形規則の対応一覧

内容要素は、文末動詞を含む要素のみを記載している（全要素は文献 [9] を参照のこと）。また、表現文型は、内容要素のうち、操作、判断、推測、行為、機能、状態、物理、禁止事項のいずれかと対応付けられている（2 節参照）。

No.	文書構造			内容要素	変形規則
1	attention			推奨事項	作業 + 「ことを推奨」
				禁止事項	禁止事項（変形なし）
					作業 + 「てはいけない」
					「絶対に」+ 作業 + [否定]
					機能 + 「させない」
				危険性	状態 + 「おそれがある」
					機能 + 「ことがある」
				徹底事項	「必ず」 + 作業 + 「こと」
2	info			モノ (機能、状態、物理)	（変形なし）
3	steps	step	cmd	作業（操作、判断、推測、行為）	（変形なし）
4			attention	1 と同様	
5			tutolialinfo	作業方法	作業 +「こと」
6			followup	作業（操作、判断、推測、行為）	（変形なし）
7			info	2 と同様	