文化財関連の専門用語を対象とした平易な説明生成

永井 利季† 宮田 玲† 立見 みどり‡ 佐藤 理史†
†名古屋大学大学院工学研究科 ‡立教大学異文化コミュニケーション研究科

1 はじめに

日本における在留外国人や外国人観光客が増える中、様々な情報を「やさしい日本語」で伝える動きが広がってきている [1]。これまで災害情報や自治体情報を対象とした実践が多かったが、近年は観光情報を対象とした「やさしい日本語」の活用も注目されており[1]、我々も、建築物をはじめとした日本の文化財を説明する文章（以下、文化財説明文）を対象として、テキスト平易化の研究を進めている [2]。

文化財説明文には「潜り戸」や「書院造り」といった専門用語が多く含まれており、文章読解を難しくしている。専門用語は、上位概念に置き換えることはできるものの[2]、一般に、別の平易な類義語が存在するわけではないため、単語レベルで言い換えることはできない。すなわち辞書などの外部知識を参照しながら、**説明する**必要がある[3]。

これまで、自然言語処理分野におけるテキスト平易化は、主に一般語と構文を対象とした言い換えを扱ってきた [3]。読解支援のための説明生成の研究 [4] や専門用語に関する情報抽出 [5] の研究もあるが、既存研究の多くは最終読者のための付加的な情報提供を主眼としており、執筆者が文章中の専門用語をどのような手順で平易に説明すべきか、という問題はあまり扱われていない。人間による執筆・平易化の支援の枠組みで、専門用語の平易な説明の方法を提案・実装することが求められている。

そこで本研究では、まず予備調査として、文章中の専門用語の人手書き換え事例を分析する（2 節）。その結果を踏まえ、文化財関連の専門用語の平易な説明生成の方法を提案し、必要な知識資源や説明生成プロセスの自動化方法を検討する（3 節）。さらに提案手法によって実際に生成された専門用語の説明を流暢性と正確性の観点から評価し、エラー分析を行う（4 節）。

[1]例えば、「やさしい日本語ツーリズム研究会」（https://yasashii-nihongo-tourism.jp）や訪日旅行者に向けの WEB マガジン「MATCHA」（https://matcha-jp.com/easy）を参照。

[2]例えば、「潜り戸」「書院造り」はそれぞれ「戸」「住居様式」に置き換えることが可能である。

[3]例えば、「潜り戸」であれば、「門に付けた小さい戸」のように説明することが可能である。

(3) 2 階の書院造りの客間は邸内で最も格式が高く、緑豊かな庭園も見渡せます。

書院造り	室町時代に始まり桃山時代に完成した武家住宅の様式。基本として座敷に、床の間・違い棚・付（つけ）書院・帳台構えを設備するもの。銀閣寺（慈照寺）の足利義政の書斎であった東求堂同仁斎は、ほぼその形式が整った現存最古の例。
客間	来客を応接する部屋。客室。
邸内	屋敷のうち。屋敷内。
庭園	計画的に草木・池などを配し、整えられた庭。「日本庭園」「屋上庭園」

図 1: 提示した書き換え対象文と辞書定義文の例

2 専門用語の書き換え事例の調査

専門用語の平易な説明とは、どのような操作であり、どのような課題があるかを明らかにするため、複数の作業者に専門用語を含む文化財説明文を書き換えてもらい、その結果を分析した。

2.1 調査方法

東京都内の文化財である Y 邸を紹介するパンフレットを対象に、辞書の情報を使いながら、文を平易に書き換えるタスクを設計した。作業者は、(a) パンフレットの文章の全体、(b) 書き換え対象文、(c) 文中の専門用語の辞書定義文を見ながら、文単位で (b) を書き換える（図 1 は (b)(c) の例）。作業に関して、以下の指示を与えた。

- 日本語学習者である外国人や小学校高学年生にもわかりやすいように書き換えること
- 専門用語に限らず文全体を書き換えること
- 提示された専門用語の定義文のみを使用しその他の情報検索はしないこと

同じ情報を与えた場合に、作業者によりどのような違いが現れるかを見るために、辞書の定義文はこちらで用意した[4]。

書き換え対象文は 6 文で、計 30 個の専門用語が含まれている。作業者は Y 邸のボランティアガイド 11 名（いずれも日本語母語話者）で、説明対象である Y 邸に関する知識は豊富であるが、言語学や日本語教育の知識は必ずしも有していない。書き換え作業は、手書き、キーボード入力のいずれかで行われた。

各文の書き換え作業終了時に、難しかった点や判断に悩んだ点もコメントとして残してもらった。

[4]基本的には『デジタル大辞泉』の定義文を利用し、一部『大辞林』『建築用語.net』も使用した。

書き換え操作	頻度	例
変更なし	48	現在は、**枯山水**の石組みやカヤ、カシワなどの大樹が... → 現在は**枯山水**の石組とカヤ・カシワの大きな木は...
削除	63	... 枯山水（かれさんすい）の庭に**園路**を設け、芝生を取り入れた和洋折衷の... → 庭は、水を使わず、石や砂で自然の風景を表した枯山水です。大正時代の庭の...
定義文情報利用書き換え	164	**邸内**唯一の洋間である応接間には... → **家の中**唯一の洋室である接待室には...
定義文外情報利用書き換え	55	邸内唯一の洋間である応接間には修復された**家具調度**が置かれ... → 邸内の応接間の室内は修復された**レコード蓄音機・ピアノや家具**は...

表 1: 書き換え結果分析

2.2 調査結果

1 文全体の書き換え結果の内、専門用語の処理に注目し、得られた専門用語の書き換え事例 330 件を 4 つの操作に分類した。表 1 に書き換え操作の分類を出現頻度、例とともに示す。

「変更なし」は書き換え後の文中に専門用語がそのまま使われている操作で、「削除」は専門用語を書き換えるのではなく文中から消してしまうという操作である。この 2 つの操作に関しては作業者が意図的にこのような書き換えを行った場合も考えられるが、良い書き換えが思いつかず、仕方がなくこのようにした場合も考えられる。実際、「〜〜の説明が難しいので省略している。」というコメントも見受けられた。

「定義文情報利用書き換え」は定義文中の情報を何かしら利用した書き換えであり、164 件と一番件数が多かった[5]。「定義文外情報利用書き換え」は定義文中からは読み取れない情報を利用した書き換えであり、更に「テキスト内の文脈知識を利用した書き換え」と「テキスト外の外界知識を利用した書き換え」に細分化できる。ここから、専門用語の説明には文脈知識や外界知識も用いられるが、辞書の定義文情報が比較的よく活用されていることが分かる。

個別の事例を観察すると、作業者による相違も多く見られた。例えば、「枯山水」に対して、特徴を示して「水を使用しない庭」とする書き換えもあれば、材料・製法を示して「石で作った池」とする書き換えもあった。また書き換えは行われたものの、難しい表現が残ってしまう事例も多数あった。例えば、「応接間」の書き換え事例である「接待室」「来客と接する部屋」「応接間（来客に会う部屋）」には、「接待」「来客」という日本語能力試験の旧試験 2〜4 級の範囲外の語彙が使用されていた[6]。「来客」については、定義文の「**来客**に応対する部屋。」で使われている語がそのまま使用されたものと思われる。

以上より、定義文が与えられても、適切な情報を選ぶことは容易ではないこと、難解語が残りがちであることが課題として明らかになった。

3 専門用語の説明生成

上記の調査を踏まえ、定義文情報を活用して、専門用語の平易な説明を生成する方法を提案する。大きく、以下の 2 つのプロセスからなる。

(1) 定義文から説明に必要な情報を選択する
(2) 選択した情報を結合し言語表現として出力する

それぞれ概ね、「何を言うか（what to say）」と「どう言うか（how to say）」に関わる。(1) に関しては完全な自動化は現実的ではなく、人間による用語情報選択を支援する仕組みを提案する。(2) に関しては自動化が可能と考えられるので、言語表現生成の方法についても説明する。

3.1 用語情報選択

ここで取り組むべき課題は大きく、(i) どのような形で定義文情報を整備しておくか、(ii) 情報選択の意思決定をどのように支援するかの 2 つである。

(i) については、用語ごとに、定義文に書かれた情報をなるべく小さい単位で分解し、系統的に分類する方法をとる。分類された情報を**用語情報**と呼ぶ。我々は、長尾による用語の意味の記述法 [6] を参考にしながら、『文化財用語辞典』[7] の見出し語から選んだ 100 語に対する『デジタル大辞泉』の定義文の分析を通じて、表 2 に示す 11 種類の用語情報を定義した。

「潜り戸」を例にとると、「門の扉などに設けた、くぐって出入りする小さい戸口。切り戸。くぐり。」（デジタル大辞泉）と「潜り戸（くぐりど）は主たる門扉に付属していて高さが低く頭を下げて通る門戸。」（Wikipedia）の 2 つの定義文から情報を抽出・分類することで、表 3 を構成できる。

専門用語の定義文から得られた用語情報をそのまま使用すると難しい語が残ってしまうという問題に対処するためには、用語情報の平易化が重要である。我々

[5]定義文情報と定義文外情報を併用している場合は、定義文外情報利用書き換えとしてカウントしている。
[6]「お客さんをお迎えする部屋」と平易に言い換えた例もあった。

種類	説明	例（**用語**：用語情報）
上位概念	より一般化、またはより抽象化した語	**絹本**: 書画
下位概念	より特定化、またはより具体化した語	**極書**: 箱書き, 極め札
別名	別の語で言い換えた語	**両部鳥居**: 四脚鳥居, 稚児柱鳥居, 権現鳥居, 枠指鳥居
時代	成立した時期や使用されていた時期など時代背景的な情報	**南画**: 江戸中期以降
場所	用いられる場所や存在する場所などの情報	**中板**: 茶室のまん中に設けられた
地域	用いられる地域や存在する地域などの情報	**御所人形**: 京都
用途	使われる用途や使用方法を含む情報	**注口土器**: 液体を注ぐ
外見	色や形など外見的な特徴を含む情報	**虹梁**: 虹のようにやや弓なりに曲がっている
材料	何を材料として構成されているかを表す情報	**枡**: 木製, 金属製
製法	作り方やどのように構成されているかを表す情報	**石組**: 自然石を組み合わせて配置する
その他	用語説明の上で重要となると考えられるが上記に当てはまらない情報	**鏡天井**: 禅宗様建築に多くみられる

表 2: 文化財専門用語向け用語情報一覧

種類	抽出した用語情報〈平易版〉
上位概念	戸口〈戸〉, 門戸〈門, 戸〉
別名	切り戸, くぐり
場所	門の扉などに設けた〈門にある, 門に付けた〉, 主たる門扉に付属していて〈主な門に付いている〉
用途	くぐって出入りする〈かがんで入る, 出たり入ったりする〉, 頭を下げて通る
外見	小さい, 高さが低く

表 3: 「潜り戸」の用語情報

説明観点	対応する用語情報
見た目	上位概念, 外見, 場所
用途	上位概念, 用途
つくり方	上位概念, 材料 or 製法
背景	上位概念, 時代, 地域, 場所
具体要素	上位概念, 下位概念
その他特徴	上位概念, その他

表 4: 説明観点と対応する用語情報

は日本語能力試験の旧試験用の難易度別語彙表[7]に照らして、2 級以内（できれば 3, 4 級）の語彙に収まる必要があると考えている。難解語に対して、分類語彙表 [8] や日本語 WordNet から類義語を取得し、それらの語彙難易度や Wikipedia での頻度を参照しながら、平易版の用語情報を事前に人手で構築しておく。

(ii) については、文章中で専門用語をどのように説明するかに関するより大局的な判断材料として、用語情報と対応づけた**説明観点**を用意する。現在、表 4 の 6 つを定義している。

以上をまとめると、用語情報選択プロセスとは、表 4 の説明観点を参考にしながら、表 3 のような用語情報の一覧から、説明に使いたい用語情報の組を決定することである。

[7] http://human.cc.hirosaki-u.ac.jp/kokugo/CATtwo/youziyougoziten/youziyougoziten_96_165.pdf から入手した。

3.2 言語表現生成

続いて、以下の用語情報の組を入力例として言語表現の生成手順を説明する。

{ **上位概念**：戸, **場所**：門のとびら, **外見**：高さが低く }

(1) 用語情報の並び順の決定　文化財の専門用語の定義文を分析しながら、なるべく係り受け曖昧性の問題が生じないように、用語情報を結合する順番を**下位概念** → **場所** → **用途** → **製法 or 材料** → **外見**→ **時代** → **地域** → **上位概念**と一意に定めた。例えば、以下の 2 通りの並び順は、いずれも文法的には可能である。

[**場所**：門のとびら, **外見**：高さが低く, **上位概念**：戸]

[**外見**：高さが低く, **場所**：門のとびら, **上位概念**：戸]

しかし、後者については、**外見**の修飾先が**場所**か**上位概念**か曖昧である。上記で定めた並び順に従うと、前者のみが選択される。

(2) 語の補完と活用の変換　隣り合う用語情報のペアに対して、先行する用語情報の「種類」と「末尾 1 語の品詞」、後続の用語情報の「種類」と「先頭 1 語の品詞」の計 4 つの情報を用いて、つなぎ部分の補完語および動詞・形容詞の活用形をルールにより決定する[8]。これを用語情報列の先頭から順に適用することで、言語表現生成が可能である。入力例に対しては、

[場所, 名詞, 外見, 形容詞]
→ 補完語：にある, 活用形：ϕ

[外見, 形容詞, 上位概念, 名詞]
→ 補完語：ϕ, 活用形：連体形

が順に適用され、「門のとびら**にある**高さが**低い**戸」が生成される。

現在、補完と活用に関するルール数は 106 である。次節で本言語表現生成手法の評価を行う。

[8] 品詞情報は MeCab ver.0.996 (http://taku910.github.io/mecab/) の結果を用いる。

4 言語表現生成手法の評価

4.1 実験設定

『文化財用語辞典』[7] から、これまで分析で使用していない専門用語を 100 語選択し、『デジタル大辞泉』および Wikipedia の定義文から人手で用語情報を同定し、平易化を行った。

各用語に対して、上位概念があれば必ず含め、使用する用語情報が最大 3 つになるように、ランダムに用語情報の組を決定し、入力データとした。また用語情報の平易化による言語生成への影響を確認するため、**平易化前**と**平易化後**の用語情報の組を用意した。

出力された用語説明の結果は、流暢性（1:自然, 2:ぎこちない, 3:文法的に誤り）、正確性（1:正しい, 2:曖昧性がある, 3:誤り）の観点から評価した。

4.2 実験結果とエラー分析

表 5 に評価結果を示す。入力した専門用語 100 語のうち、上位概念が存在せず説明を生成できなかったものは 4 語存在した。

実験結果において、**平易化前**では流暢性と正確性が 1 である割合がそれぞれ 8 割を超えており、**平易化後**では流暢性と正確性が 1 である割合がそれぞれ 8 割弱である。使用する用語情報を最大 3 つに制限した場合には良い精度で説明を生成できることがわかる。

また平易化後では、流暢性、正確性の両方の評価が下がっている。これは平易化することによって言語表現が変わり、用語情報間の意味的な繋がりが切れてしまったことが原因の一つに挙げられる。例えば、用語「露地」の説明生成結果は、平易化前は「茶室に**付随する**屋根など覆いのない土地」であるのに対して、平易化後は「茶室に**付いた**屋根などカバーのない土地」である。「付随する」を「付いた」に変えたことで、説明生成後に、「付いた」と「土地」の意味的な繋がりが弱くなり、不自然な表現になっている。

流暢性のエラーを分類すると、平易化前は、補完・活用のルールの誤りが 8 件で最も多かった。平易化後は、補完・活用のルールの誤りが 6 件、語句の繋がりの不自然さが 8 件存在した。正確性のエラーを分類すると、平易化前は、係り受けの曖昧さが 6 件、補完・活用のルールの誤りが 5 件あり、平易化後は、係り受けの曖昧さが 10 件、語句の繋がりによる非文が 10 件あった。

補完・活用のルールの誤りについては新しいルールを追加することで、係り受けの曖昧さに対しては新たに読点をつけるようなルールを整備することで、ある程度改善できる。語句の繋がりの悪さについては、人手の平易化段階で極力回避することや、生成後にさらに別の語に換言することが対処法として挙げられる。

	流暢性			正確性		
	1	2	3	1	2	3
平易化前	83	9	4	82	11	3
平易化後	76	14	6	72	13	11

表 5: 評価結果（生成された 96 件の分類）

5 おわりに

本稿では、人手によるテキスト平易化事例の分析に基づき、文化財関連の専門用語に対する平易な説明を生成する方法を提案した。本方法は、定義文から同定・平易化した用語情報を元に、(1) 必要な用語情報を選択し、(2) それらを結合して言語表現として生成するプロセスからなる。(2) のプロセスは自動化しているため、作業者は (1) において、整理された情報を見ながら「どの情報を伝えるか」という情報選択タスクに専念することができる点が、本提案手法の大きな特徴である。また文化財の専門用語 100 語を対象に、(2) の生成結果を流暢性、正確性の観点から評価した。評価の結果、入力の 8 割程度に対して良い説明を生成することができた一方、ルールの不足や用語情報の係り受け曖昧性の問題が明らかになった。今後は、学習ベースの生成手法も検討しながら、提案手法をユーザインタフェースを含めたシステムとして実装する予定である。

謝辞 本研究の一部は科研費（17K00466, 19K20628）の支援を受けた。

参 考 文 献

[1] 庵功雄, イ・ヨンスク, 森篤嗣 (編). 「やさしい日本語」は何を目指すか : 多文化共生社会を実現するために. ココ出版, 2013.
[2] 立見みどり. 翻訳テクノロジー論考「やさしい日本語」と翻訳テクノロジー その 1. JTF ジャーナル, No.300, pp.34–35, 2019.
[3] Saggion, H. *Automatic Text Simplification*. Morgan & Claypool, 2017.
[4] Shardlow, M. “A Survey of Automated Text Simplification”. *International Journal of Advanced Computer Science and Applications*, Vol.4, No.1, pp.58–70, 2014.
[5] 桜井裕, 佐藤理史. ワールドワイドウェブを利用した用語説明の自動生成. 情報処理学会論文誌, Vol.43, No,5, pp.1470–1480, 2002.
[6] 長尾真. 辞典形式での専門分野の知識の体系的構成法. 人工知能学会誌, Vol.7, No.2, pp.320–328, 1992.
[7] 京都府文化財保護基金 (編). 文化財用語辞典. 淡交社, 1989.
[8] 国立国語研究所 (編). 分類語彙表増補改訂版. 大日本図書, 2004.