「やさしい日本語」変換システムの試作

李 真奈見　　山本 和英
長岡技術科学大学 電気系
{moku, yamamoto}@jnlp.org

1 はじめに

現在，日本に在住する外国人は200万人を超え，その中でも日常生活に必要な日本語能力を持たない外国人は数十万人に及ぶ．しかし，一般的に日本社会で日本語以外は使用されない．よって外国人が日本で生活していくために日本語の知識は必要不可欠である．

外国人のために，必要最低限の日本語を提示する「やさしい日本語」[1]がある．「やさしい日本語」とは，日本語母語話者が日本語の文法や語彙に制限をかけて難しい日本語から「やさしい日本語」へ変換を行ったものを指す．ここでの難しいとは，最低限の文法と語彙を習得した日本語初学者の外国人が理解できないこと，やさしいとは日本語初学者でも理解できることである．

本研究の「やさしい日本語」で対象としている文書は公的文書である．公的文書とは，市役所や病院，学校等の公共施設で配られる文書を指しており，これらの文書は生活するために重要な情報を多く含んでいる．しかし，日本語初学者が学習する文に比べ理解が困難であり，特有な表現も含むため，「やさしい日本語」へ変換する必要がある．

そこで我々は公的文書における最も伝えるべき内容を抽出し，「やさしい日本語」で出力するシステムの作成を目指した．本稿では，システムの概要とシステムの工程の 1 つである「やさしい日本語」への変換について評価とともに述べる．システムは日本語初学者に公的文書に含まれる情報や指示を可能な限り端的に伝える状況を想定して作成した．システムの動きは接続詞やいくつかのキーワードに注目し，短文化を行い，その短文化したものをさらに「やさしい日本語」へと変換した．変換には「やさしい日本語」コーパスを使用している．また日本語初学者のために最も重要な部分を強調して伝える．

「やさしい日本語」への変換についての評価は2種類行った．1つは日本語母語話者による日本語についての評価，もう1つは留学生による入力と出力のどちらがやさしいかという評価である．

2 関連研究

関連研究としていくつかの「やさしい日本語」がある．美野ら[2]は国語辞典の見出しとその説明文より平易化対を取得し，日本語能力試験(JLPT)を基にした単語への級の付与により難易語と平易語を定めている．また美野ら[3]は基本語彙を使用頻度が高く，使用領域が広いものと定義し，頻度や情報量，相乗平均により放送ニュースの基本語彙を定めている．このように語彙の制限により外国人に伝わりやすい「やさしい日本語」を目指している研究が主だったものである．しかし，我々は単語単位ではなく，文単位での「やさしい日本語」にできないかを目指している．これは，日本語母語話者が日本語初学者である外国人に何かを伝える際，単語をやさしくするよりも文をやさしくするという表現が適していると考えるからである．

「やさしい日本語」のシステムとしては松田ら[4]の Plain Japanese (PJ)システムがある．これは日本での工学教育で使用するために開発されたものである．教育がたいてい日本語で行われる日本では，留学生は日常会話のための日本語だけでなく，専門のための日本語も学ばなくてはならない．その支援のためこのシステムは語彙と文法を制限する．このシステムと我々のシステムは似ているが，対象物が PJ システムは工学教育，我々は公的文書である点が異なっている．

また，我々はシステムに重要部分の抽出を利用する．重要文抽出は要約の分野などでよく使われる．これは文単位で抽出することで日本語の自然さを維持できるからである．しかし人が要約文を作成する際，複数文を基に 1 つの文を構成することが多い．そこで鈴木ら[5]は SVM(Support Vector Machine)を用いた重要文節抽出による要約を行った．これは複数文から1文を再構成する際に重要文抽出よりも有効であった．我々はこの重要文節の抽出が「日本人は日本語初学者である外国人に要件を伝える場合，要点のみを伝える」ことと類似していると考える．しかし我々は SVM のための大量の公的文書における重要部分のデータを有していない．そこで小規模の実験データからルールベースの重要部分抽出を行う．

3 使用データ

3.1 「やさしい日本語」コーパス

本研究は「やさしい日本語」コーパスを利用した．これは「やさしい日本語」のプロジェクトで作成されたものであり，日本語教師が公的文書の日本語を「やさしい日本語」に訳したものである．公的文書は日本語初学者が学習する文に比べ理解が困難であり，特有な表現も含む[6]ため，「やさしい日本語」へ変換する必要がある．

「やさしい日本語」コーパスは約 40 名の日本語教師によって作成され，42,274 文の公的文書を含む．このコーパスは原文である公的文書と共に逐語訳，意訳，要約という3段階の訳を含む．これらは一定の文法基準[1]と JLPT2 級(現試験におけるN2)レベルの語彙のみに制限されている．コーパスにおける難しい，やさしいの基準は日本語教師の主観である．

以下に「やさしい日本語」へ変換した例を示す．

例 1)
公的文書:　　予防接種
「やさしい日本語」: 予防注射，病気にならないための注射

「予防接種」は重要な情報だが，日本語の学習内容として一般的ではないため，理解できない外国人が多い．しかし「接種」を一般的な語彙である「注射」に変換することによって，意味を理解しやすくなる．また複数人で作業をしているため「予防注射」だけでなく「病気にならないための注射」にも訳すなど，1 つの公的文書に対して複数の訳がある場合があった．

本研究では「やさしい日本語」にするため，公的文書と「やさしい日本語」の変換対を用いた．また公的文書を用いて表現意図データの作成も行った．「やさしい日本語」コーパスに含まれる原文に対する各訳の例を示す．

例 2)
原　文:
　ニュース等で報道されておりますように，世界的に新型(豚)インフルエンザの流行が危惧されています．
逐語訳:
　ニュースなどにもあるように，世界中で新型インフルエンザの流行が心配されています．
意　訳:
　さて，ニュースでもありますが，世界中で新型インフルエンザが増えています．
要　約:
　さて，世界中で新型インフルエンザが増えています．

3.2 表現意図タグ

このデータは，元日本語教師 1 名が「やさしい日本語」コーパス中の公的文書 503 文に含まれる文節(672 文節)に対してそれが表現する意図をタグ付けしたものである．表現意図とは自己表出，理解要請，行動展開の 3 つの表現である[7]．その中でも理解要請と行動展開の表現意図を基に表 1 の表現意図のタグを定めた．また，ここでの文とは句点や改行で区切られたもの，文節とは文を句読点やいくつかのキーワードで区切ったものとした．キーワードの例は次の通りである．

　キーワード:「場合」,「際」,「について」,「ので」

　これらいくつかの単語と接続詞，接続助詞，そして形態素解析で動詞や助動詞で「仮定形」とされたものをキーワードとして用いる．これらのキーワードとその前後の助詞や句読点を考慮して文を自動的に区切る．
　各キーワードや文末表現などから各文節の表現意図を読み取り，その表現意図から各文節の関係を図示するシステムの構築を行った．理由のタグと指示・命令のタグの関係を基に構築の例を次に示す．

例 3)
　○○なので，　　⇒　　××してください．
【タグ:理由】　　　　【タグ:指示・命令】

　矢印等の記号や，関係を階層として表して出力するなど，図示を用いることによって文節と文節の関係を明確にする．

3.3 「やさしい日本語」変換対

　これは「やさしい日本語」のプロジェクトで作成されたものである．「やさしい日本語」コーパスに含まれる公的文書と「やさしい日本語」において対応する差異を対として構成している．例を次に示す．

例 4)
公的文書:　その他，申請に関してご不明な点がありましたら，
　　下記までお問い合わせください．
タグ対象語(原文):　ご不明な点がありましたら
逐語訳:　わからないことがありましたら
意　訳:　わからなかった
要　約:　わからなかった

表 1. 表現意図を表すタグ

タグの種類	例
忠告・助言	～したほうがいいですよ
勧告	～しませんか・しましょうよ
依頼	～してもらえますか/くれませんか
指示・命令	～してください・しなさい・お願いします
許可与え	～してもいいです
申し出	～してあげましょうか
許可求め	～してもいいですか
確認	～してもいいですね
通知・宣言	～します・させてもらいます
条件・仮定	～の場合・際，～すれば(仮定形)
理由	～ので
題目・タイトル	～について
項目	(各種項目の形式となっているもの)
既定の事実・結果	(過去形)
禁止	～いけません

　この原文，逐語訳，意訳，要約の 4 つの組み合わせから，原文－逐語訳，原文－意訳，原文－要約の 3 つの変換対を作成した．

4 「やさしい日本語」書き換えシステム

　「やさしい日本語」書き換えシステムは次の 4 つの工程で構成した．
(1) 重要部分の抽出
(2) 短文化
(3) 表現意図を用いた図示への変換
(4) 「やさしい日本語」への変換
　それぞれの仕組みとその出力の例について次に述べる．

4.1 重要部分の抽出

　現在，重要部分の抽出は係り受け解析を用いて行っている．これは係り受け関係にある文節をフレーズとし，そのフレーズの中でも動詞の数と，含まれる格助詞の種類によってそれぞれ順位付けし，上位のフレーズを重要部分とした．格助詞による順位は著者の 1 人の判断でヲ，ノ，ガ，ハ…とした．しかし格助詞は動詞に依存するため，単純に順位づけできない[8]．よって新しいデータを現在，作成中である．重要部分の抽出の例を次に示す．下線部分が重要部分と考える．

例 5)
入力:　インフルエンザにかかった人が咳やくしゃみなどをすることにより，ウイルスが空気中に広がり，それを吸い込むことによって感染します．
出力:　インフルエンザにかかった人が咳やくしゃみなどをすることにより，ウイルスが空気中に広がり，それを吸い込むことによって感染します．

4.2 短文化

　短文化は日本語初学者にとって複雑な日本語の構造が解消され，わかりやすい日本語の出力にすることができると考え

る. これは 3.2 節の表現意図タグ作成と同じ方法である. 短文化の例を次に示す. この例の場合, 5 つの文節に分けている.

例 6)
入力: また, すでにお手持ちの2回(前期・後期)の受診票につきましては, 平成20年度から一部内容が変更されますので, 平成20年4月1日以降に受診の際は, 医療機関にて新票と差し替えさせていただきますのでご了承ください
出力: 1 また,
2 すでにお手持ちの2回(前期・後期)の受診票につきましては,
3 平成20年度から一部内容が変更されますので,
4 平成20年4月1日以降に受診の際は,
5 医療機関にて新票と差し替えさせていただきますのでご了承ください

4.3 表現意図を用いた図示への変換

本システムでは表現意図タグを用いて図示化を行った. 形態素解析器 1)で品詞付けした各形態素を, 表現意図タグ作成を基に作成したルールを用いて表現意図タグを付与した.

例として例 6 の出力にタグを付けた結果を表 2 に示す.

またタグを付与した文節を用いた図示化の例を次に示す.

例 7)
入力: また, すでにお手持ちの2回(前期・後期)の受診票につきましては, 平成20年度から一部内容が変更されますので, 平成20年4月1日以降に受診の際は, 医療機関にて新票と差し替えさせていただきますのでご了承ください
出力: ＋: また,
題目・タイトル:すでにお手持ちの2回(前期・後期)の受診票につきましては,
理由: 平成20年度から一部内容が変更されますので,
↓
↓○条件・仮定: 平成20年4月1日以降に受診の際は,
↓
指示・命令: 医療機関にて新票と差し替えさせていただきますのでご了承ください

4.4 「やさしい日本語」への変換

次に公的文書の日本語を「やさしい日本語」へ変換した. 予備実験として直接的表現へ変換しようと試みた[8]. しかし, この変換対は数が少なく効果が小さいことがわかっている. また直接的表現の対は「やさしい日本語」変換対とほとんど変わらず, 「やさしい日本語」変換対の方が量は多いため, 「やさしい日本語」変換対を使用することにした. 原文－逐語訳, 原文－意訳, 原文－要約の異なり語における変換対は以下の通りとなった.

原文－逐語訳	: 5893 対
原文－意訳	: 4772 対
原文－要約	: 3944 対

それぞれの対は原文側の語の文字数が多いもの順に並べ, かつ, 出現頻度の情報も付随して, 多いものを優先的に変換することにした. これによって, できるだけ長い文字列の変換を行い, 文の意味が変わることを防ぐ. また複数の日本語教師がコーパス作成に携わっているため同じ文字列でも複数の「やさしい日本語」が存在する問題も, 出現頻度が高いものを優先的に変換することとした. 原文－逐語訳の対を用いた変換例を次に示す.

表 2. タグの付与例

No.	文節	タグ
1	また,	接続詞
2	すでにお手持ちの2回(前期・後期)の受診票につきましては,	題目・タイトル
3	平成20年度から一部内容が変更されますので,	理由
4	平成20年4月1日以降に受診の際は,	条件・仮定
5	医療機関にて新票と差し替えさせていただきますのでご了承ください	指示・命令

例 8)
入力: 65 歳以上の高齢者及び 60 歳～65 歳未満のハイリスク者の自己負担金は, 1, 000 円必要になります.
出力: 65 歳より多いお年寄り・60 歳~65 歳より小さいハイリスク者が自分で払うお金は, 1,000 円いります.

4.5 システムの出力

4.1 節から 4.4 節に示した各ステップを合わせることによって「やさしい日本語」への変換システムの出力を構築する. 図 1 に最終出力の例を示す.

このシステムはインターネット上で利用できるようにする予定である. これによって日本語初学者だけでなく, 一般の人も公的文書を入力して出力を得ることができる.

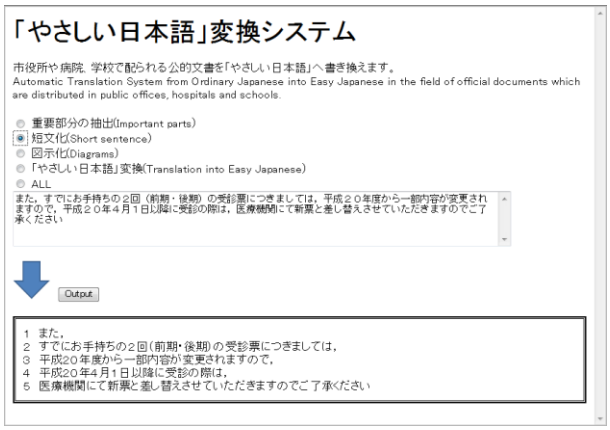

図 1. システムの出力例

5 評価

5.1 評価手法

本稿では「やさしい日本語」変換システムの中心部である, 「やさしい日本語」への変換について 2 種類の小規模な評価実験を行った.

入力文として「やさしい日本語」コーパスより無作為に 100 文

抽出した．それらを原文－逐語訳の変換対を用いて変換した．変換方法は，対における原文が公的文書に含まれていた場合，対における逐語訳へと変換する．ただし，入力文は形態素解析器 1)で分かち書きを行い，1 つまたは複数の形態素が対における原文と等しいか否かで判断する．また，名詞連続，“数字+助数詞”，“動詞+こと”は複合名詞と考える．そして，形態素の一部分を用いた変換は行わないこととした．複合名詞処理をした形態素解析結果の例を次に示す．下線部を複合名詞として結合する．

例 9)
入学/手続/が/済め/ば/，/日本語/学校/の/職員/が/代理/で/就学/ビザ/の/在留/資格/認定/証明/書/を/申請/する/こと/が/でき/ます/．

例 9 において，“在留資格”の対があった場合，“在留資格”は“在留資格認定証明”の一部であるため変換しない．また同様に，“学校の職員”の対があった場合も変換しない．しかし，“日本語学校の職員”であれば変換する．

これらの処理で出力した評価文を用いて，日本語母語話者による日本語の評価と日本語非母語話者によるやさしさの評価を行った．評価文の漢字は全てルビを振った．

5.2 日本語の正しさについての評価

まず日本語母語話者である著者の 1 名が評価文として入力と出力を見て，日本語の文法と意味の観点から日本語として適切か否かを判断した．結果，変換が行われた文が 82 文であった．82 文中の日本語の正しさの評価結果を表 3 に示す．

文法または意味のどちらかが間違っているものが 26 文であった．これの多くは文法の間違いであり，特に助詞や動詞の活用の変化の間違いであった．どちらも間違っているものが 27 文あった．間違いの多くは公的文書が名詞+接尾辞であり，名詞部分のみが変換されたものであった．これは接尾辞が単体で変換される場合を考慮して，接尾辞は複合名詞に含めないとした．しかし名詞のみを変換する場合は文としての意味が変わってくる場合が多いからと考える．

文法の間違いの原因は対の形態素情報を用いていないことである．現システムでは，対と対応する文字列があれば変換しているが，形態素解析の情報を用いた改善が必要である．意味の間違いについても，変換で用いるルールを実験の繰り返しにより増やすことによって対応していきたい．

5.3 「やさしさ」についての評価

評価者は日本語学習者 6 名で，全員がマレーシア国籍，JLPT の N2 保有者である．6.1 節において日本語が正しいと判断された 47 文のうち無作為に 15 文を抽出し，入力である公的文書と出力である「やさしい日本語」の文の評価を行った．評価は，各評価文において公的文書と「やさしい日本語」の文のどちらがやさしいかを多数決方式で決めた．評価者全員の結果と，日本での在住期間が 1 年未満の評価者の結果を表 4 に示す．

結果，全員の場合は「やさしい日本語」の方がやさしいと判断された文数の方が多かった．しかし公的文書の方がやさしいと判断された文数との差は小さい．これは日本に 1 年以上住んだ経験がある方が 3 名いたため，公的文書の文に慣れていたと考える．そこで日本での在住期間が 1 年未満の 3 名の

表 3．日本語の評価結果

日本語の正しさ	正しい	文法または意味が間違い	間違い
	47	26	27

表 4．やさしさの評価結果

	公的文書	「やさしい日本語」	同じくらい
全員	6	9	0
日本在住期間 1 年未満	3	12	0

結果をみると，「やさしい日本語」の方がやさしいと判断された文数が増加した．このことから，システムの出力は公的文書に慣れていない日本語初学者に効果があった．

7 おわりに

本研究では，「やさしい日本語」変換システムの構築を目指した．本稿で述べた「やさしい日本語」への変換の小規模実験により，形態素情報の未使用による不具合を発見した．しかし，日本語初学者に対しては有効であった．今後，システムを構成する他の工程においても小規模実験を繰り返しながら改善し，システムを完成させたい．

参考文献

[1]庵功雄．「やさしい日本語」をめぐって．多文化共生社会における日本語教育研究会 第4回研究会，pp.1-12 (2008)

[2]美野秀弥・田中英輝．国語辞典を使った放送ニュースの名詞の平易化．言語処理学会第 16 回年次大会発表論文集，pp.760-763 (2010)

[3]美野秀弥・田中英輝．放送ニュースの動詞連用形名詞の平易化．言語処理学会第 17 回年次大会発表論文集，pp.744-747(2011)

[4]松田真希子・竹元勇太・石坂達也・柴木優美・児玉茂昭．Plain Japanese システム (2009) http://twinning.nagaokaut.ac.jp/PJ/PJ.html

[5]鈴木大介・内海彰．Support Vector Machine を用いた文書の重要文節抽出―要約文生成に向けて―．人工知能学会論文誌 21 巻 4 号 B，pp.330-339 (2006)

[6]筒井千絵.試用版書き換えコーパスの作成.日本語教育学会大会 2009(平成 21)年度春季大会予稿集，pp.86-87 (2010)

[7]蒲谷宏・川口義一・坂本惠．敬語表現．大修館書店 (1998)

[8]Manami MOKU・Kazuhide YAMAMOTO・Ai MAKABI. Automatic Easy Japanese Translation for information accessibility of foreigners. Proceedings of the Workshop on Speech and Language Processing Tools in Education, COLING 2012, Mumbai, India, pp.85-90 (2012)

使用した言語資源及びツール

1) 形態素解析器 MeCab, Ver.0.993, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html,