社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

Web サイトへのアクセシビリティ向上を目的とした難語の平易化

中野智子[1]　遠藤淳[1]　菅原昌平[1]　乾健太郎[2]　藤田篤[3]

1: 株式会社 NTT データ　〒104-0033　東京都中央区新川 1-21-2

2: 奈良先端科学技術大学院大学　〒630-0101 奈良県生駒市高山町 8916-5

3: 京都大学　情報科学研究所　〒606-8501　京都市左京区吉田本町

E-mail: nakanot@nttdata.co.jp, endouat@nttdata.co.jp, sugawaras@nttdata.co.jp, inui@is.naist.jp, fujita@pine.kuee.kyoto-u.ac.jp

あらまし　Webページに含まれるテキスト情報は，利用者の知識や閲覧環境，学習経歴などによって分かりやすさが異なる．特に，高齢者や障害者などの情報弱者にとってこの問題は深刻である．本稿では，問題の一つである“難語”を取り上げ，言い換え技術を用いてWebページ上の難語の自動変換する実験について報告する．対象ドメインを限定すれば，現在の資源・技術のチューニングによって言い換えの正確性と網羅性をある程度確保できる見通しを得た．

キーワード　インターネット，アクセシビリティ，難語，言い換え

Lexical Paraphrasing for Improving Accessibility to the Web

Tomoko NAKANO[1]　Atsushi ENDO[1]　Shohei SUGAWARA[1]　Kentaro INUI[2]　Atsushi FUJITA[3]

1: NTTDATA Corporation　21-2,Shinkawa 1-chome, Chuo-ku, Tokyo, 104-0033, Japan

2: Nara Institute of Science and Technology　8916-5, Takayama-cho, Ikoma-city, Nara, 630-0101, Japan

3: Graduate School of Informatics, Kyoto University　Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan

E-mail: nakanot@nttdata.co.jp, endouat@nttdata.co.jp, sugawaras@nttdata.co.jp, inui@is.naist.jp, fujita@pine.kuee.kyoto-u.ac.jp

Abstract Textual information included in Web pages is not equally accessible to users depending, for example, on their language proficiency and browsing methods. This paper reports on the preliminary results of our experiment on the task of automatically simplifying unfamiliar words in Web documents. In the experiment, we aim at examining present techniques for automatic lexical paraphrasing to our task and are obtaining promising results indicating our approach would work as far as the target domain is carefully restricted.

Keyword　The Internet, Accessibility, Difficult/unfamiliar words, Paraphrasing

1. はじめに

近年の情報処理技術（IT）の進歩は，生活者の情報入手に大きな変化をもたらした．しかし，障害等によってこの恩恵を十分に享受できない人々は，いわゆる情報弱者として情報の流通からますます疎外される傾向にある．例えば，視覚障害者が読み上げソフトを使う場合，元テキストに読み上げを想定した配慮が施されていないと，内容が理解できない場合がある．また，聴覚障害者の中には，幼少時の日本語学習が困難であったため，文章を読むことに不自由を感じている人が少なくない．

こうした Web 上のテキスト情報のアクセシビリティを向上させるためには，どのような要因でテキストの読解が困難になるのかを分析し，その問題をひとつひとつ解決する必要がある．

本稿では，テキスト情報の分かりにくさを解消する方法として，テキスト中の難解な言語表現を平易な表現に自動変換する技術を取り上げ，Web サイトへの適用の有効性について検証し，実用化への課題について報告する．

2. 難語の平易化

2.1. 難語への言い換え要件

我々は，高齢者や視覚障害者がどのような場合に読解困難と感じるかを調査するため，ヒヤリングやユーザテストを重ねてきた．その中で，新語や馴染みのない単語の影響が大きいことが確認されている．特に，視覚障害者の場合，漢字などの視覚的な手がかりがないため，そうした難語の使用が，テキストの理解を妨げる大きな要因となっていることがわかった[1].

そこで本研究では，テキスト中の難語を平易な表現に自動的に言い換えるというタスクに取り組む.

2.2. 言い換え技術

テキストの自動言い換えについては，近年盛んに研究が行われている[2]．しかし，現状では，語義の曖昧性解消や文脈に応じた表現の選択という技術的課題が残されている．また，難語を平易化して読解を支援するという問題に対して,難語の集合をどう定義するか,名詞や動詞などのオープンクラスの語に対していかにして平易な表現を記述するかという実践的な課題もある．ただし，テキストの対象領域（ドメイン）を限定すれば，上記の実践的課題については見通しが良くなる．また，語義曖昧性の問題も生じにくくなる可能性がある.

これらをふまえ，本研究では，テキストの対象ドメインを限定したときにどの程度のコストでどれくらいの質，量の平易化が可能になるのかを検証し，実用化に向けての課題を整理する．具体的には，我々が開発中の言い換え生成システム KURA[3]を用い，名詞を対象とした言い換えの生成と評価を行う．そして，難語の平易化の際に生じる問題を整理し，その対処法について検討する.

3. 実験内容

3.1. 対象テキスト

本研究では,自治体の Web ページ中のテキストを対象とする．地域住民に有用な情報であるにもかかわらず,馴染みのない単語や言い回しが多く含まれており,アクセシビリティ向上のニーズが高いためである.

今回の実験では，例題として「住居の移動に伴う手続き」を取り上げ，複数の自治体の Web サイトからコンテンツをダウンロードし，5,493 文を抽出した．このうち 5,393 文を用いて平易化辞書を作成した.また,残りの 100 文を言い換えの生成と評価に用いた.

3.2. 平易化辞書作成

開発用サンプル文5,393文中の難語を抽出し，各々に対する平易な表現を付与して平易化辞書を作成した.概要を図 1 に示す.

まず，図1(1)に示す手順でサンプル集合から難語のリストを作成した.(1-1)では日本語形態素解析システム茶筌[4]により，名詞1,683語と固有名詞を含む未知語391語を抽出した．そして，これらの語の中から，単語親密度[5]が6.0未満，かつサンプル集合における出現頻度が2以上の739語（一般名詞328語，サ変名詞411語）を取り出し，難語のリストとした．一般に難語と考えられる語は無数にあるが，このようにドメインを固定することで，ある程度のカバレージを持つ難語リストを作成することができた.

次に図 1(2)に示す手順で，難語の各々に対して人手で平易な表現を付与した．ここでは，できる限り，一般に平易だと考えられる表現を付与することが望ましい．そこで，国語辞典の語釈文やシソーラスの同概念語を参照して平易な表現の候補を取り出した．たとえば，「支援」という語に対しては次の８つの候補表現が得られた.

援助，加勢，手助け，力添え，
助け船，与力，応援，ヘルプ

そして，これらの候補の中から，サンプル集合の文脈に適した表現を人手で１つ選択した．ここでも，ドメインを固定して具体的な例文を参照したことで，語の多義性を網羅するという問題が軽減でき，およそ50人時間という、比較的安いコストで辞書を作成できた.品詞ごとの平易化辞書中のエントリの例を表1に示す.複合名詞の平易化辞書は，難語を含む複合名詞をサンプル集合より抽出し，各々に平易な表現を人手で付与して作成したものである.

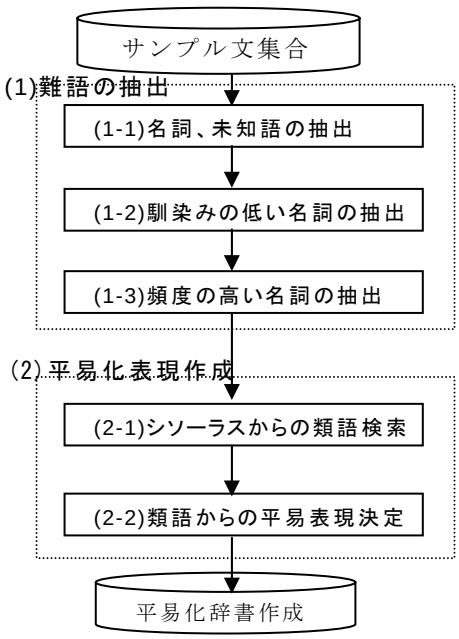

図 1.平易化辞書作成の流れ

表1．平易化辞書中のエントリの例

種類	語数	例
一般名詞	328 語	居宅：　すまい 目途：　目的 利便：　便利さ 卑属：　子孫 尊属：　祖先
サ変名詞 名詞用法	379 語	受診：　診察 検診：　診察 診療：　診察 推計：　推定の計算 進捗：　進み具合
サ変名詞 動詞用法	411 語	従事する：携わる 来訪する：訪ねてくる 干渉する：立ち入る 到達する：達する 逸脱する：脱線する
複合名詞	314 語	産業振興：産業を繁栄させる 施策展開：施策のひろがり 施設入所：施設に入る 資格喪失：資格を失うこと

3.3. 言い換えの自動生成

平易化辞書を言い換えシステム KURA 上に実装し，評価用サンプル文 100 文に対する言い換え文を自動生成した．結果，100 文中 86 文に対する言い換えが生成された．単語レベルで言い換えられた箇所の述べ数は 191 であり，高い頻度で言い換えられたと言える．これは，少なくとも「住居の移動に伴う手続き」という狭い領域に関する限り，自治体ごとに用いられる単語や言い回しに大きな違いがなかったためと考えられる．

4. 評価と考察

この節では，どの程度適切な言い換えが生成できたか，どの程度分かりやすさを向上できたかの評価結果を示す．また，明らかになった問題およびその対処法について述べる．

4.1. 評価手順

評価は，日頃業務でインターネットを利用している 20 代から 30 代の 5 名の被験者に対してアンケート方式で実施した．アンケートでは，被験者に言い換え前後の文を見比べてもらい，以下に示す設問 1 から 3 に答えてもらった．

設問 1 では，文法的な誤りがないか質問する．次のような例の場合，「損なわれる」と言い換えるべきであるため，「いいえ」と評価する．

[設問 1] 文法的な崩れはないか？
（はい，いいえ，わからない）
s. プライバシーが《侵害さ》れる恐れもあります．
t. プライバシーが《損なう》れる恐れもあります．

設問 1 において文法が適切と評価した場合（「はい」を選択した場合）のみ，設問 2 に進む．設問 2 では，意味が保存されているかどうかを質問する．次のような例の場合，資格が認められるという「認定」の意味が損なわれるため，「いいえ」と評価する．

[設問 2] 意味が保存されているか？
（はい，いいえ，わからない）
s.介護資格の《認定》が必要です．
t.介護資格の《決定》が必要です．

設問 2 において，意味が保存されていると評価された場合（「はい」を選択した場合）のみ，設問 3 を質問する．設問 3 では，言い換えによって分かりやすさが向上したかどうかを質問する.次の例のように,「代理」という単語の分かりやすさの判断は個人により異なるため，分かりにくいよ答えた場合はその理由についても確認した．

[設問 3] 分かりやすくなったか？
（はい，いいえ，わからない）
s．本人以外の方が《代理》で請求する．
t．本人以外の方が《代わり》で請求する．

4.2. 評価結果

今回の評価用サンプル文に対しては，適切な言い換えが生成されなかった文が全体の 4 割に及んだが，適切な言い換えが生成された場合は，7 割の文に効果(分かりやすくなった）が確認された（図 2 参照）．以下，各設問についての分析結果を述べる．

4.2.1. 文法的な誤り

言い換えによって多くの文法的な誤りが生じた．とくに，次の例のような誤りが多く見られた．これは，動詞として用いられているサ変名詞「受領する」を名詞用法「受領」と誤認したために生じた誤りである．実験を通じて明らかになったこれらの形態素レベルの変換の不備はひとつひとつ修正していく必要がある．

s．市民係の窓口でカードを《受領》します．
t．市民係の窓口でカードを《受け取り》します．

また，ある単語を言い換える際，同じ品詞の一語に置き換えられる場合と，複数の語から構成される句，節でしか表現できない場合がある．今回の辞書作成の過程では言い換え後の表現に関する制限を特に設けなかったため，44%の語に対して，句，節の形の言い換え表現を登録していた．複数の語からなる表現に言い換える場合に，次の例のように，文脈との整合性が損な

われる場合があった.

s. 細かな《手法》の確立を行う.
t. 細かな《物事のやり方》の確立を行う.

言い換え表現を文中に適切に埋め込むためには，対象文中の語句と言い換え表現中の語句の対応付け[6]などの処理が必要である.

4.2.2. 意味の変化

意味変化を指摘した理由のうち，次の例のような，固有名詞部分に該当するものが 7 割に及んだ.

s. 本人の《承諾》書が必要です.
t. 本人の《承知》書が必要です.

これは，申請名や届出名など，固有名詞の一部である名詞に対して言い換えが発生したためである．このため，固有名詞は言い換えを行わない制御が必要と考えられる.

対象ドメインを限定したにもかかわらず，語の多義性のために文脈にあわない言い換えが生成される場合もわずかに確認された．この問題は，言い換え後の表現の適切さを評価するモデル[7]によって解決できる可能性がある．多義語の数は限られると予想されるため，それらを中心にモデルの訓練データを収集すれば良いと考えている.

4.2.3. 分かりやすさ

適切な言い換えが行われた場合は，７割の文に言い換えの効果（分かりやすくなった）が見られた．また，分かりにくくなったと指摘された文は，理由として，「冗長である」,「一般的ではない」などのコメントが挙げられた.

s.《事前》にお問い合わせください.
t.《物事の行われる前》にお問い合わせください.

評価アンケートでは「はい」と答えた場合のみ次の設問に進むため，ある設問の回答者は言い換え文ごとに異なる．そこで，「はい」と答えた人が回答者に占める割合を言い換え文ごとに算出し，その平均を回答の一致率とした．その結果，評価文への回答一致率の平均は,「文法的に正しい」が 81%,「意味が同じ」が 86%,「分かりやすい」が 48%となった．これは，分かりにくさの尺度や，単語ごとの印象について，個人差が大きいためと考えられる．このため,「分かりにくい」理由を，利用者特性やドメインの特徴など，多角的に分析し，原因を分類していくことが必要と考える.

5. まとめ

対象ドメインを限定した結果，比較的安いコストで難語のリストおよび平易化辞書を作成することができた．言い換えの生成実験では，適切に言い換えが行われた文の約 7 割が分かりやすくなったと評価された．また，文法的な誤り，意味の変化など，適切な言い換えを生成できなかった原因を整理し，その対処法を論じた.

今後は,課題解決を進めるとともに，他のドメインにおける平易化の実現可能性についても検討したい.

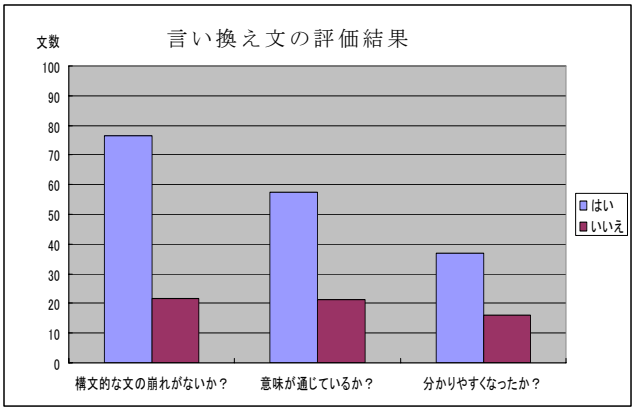

図 2. 言い換え文の評価結果
文数は 5 名の評価の平均を示す.

参 考 文 献

[1] 中野智子,堀晋久,遠藤淳,菅原昌平, ”視覚障害者の Web 閲覧を考慮した.Web 閲覧支援システム”, 2005 年電子情報通信学会総合大会, March, 2003.

[2] 乾健太郎,藤田篤,”言い換え技術に関する研究動向”,自然言語処理, Vol.11, No.5, pp.151-198, 2004.

[3] Takahashi. T.. Iwakura. T.. Iida. R.. Fujita. A. and Inui. K. KURA: A transfer-based lexico-structural paraphrasing engine. NLPRSWorkshop on Automatic Paraphrasing: Theories and Applications. pp. 37-46. 2001.

[4] 松本裕治,北内啓,山下達雄,平野善隆,松田寛,高岡一馬,浅原 正幸.,“日本語形態素解析システム『茶筌』”, 2000.

[5] 天野成昭,近藤公久, “NTT データベースシリーズ 日本語の語彙特性 1: 単語親密度”, 三省堂, 1999.

[6] 鍛治伸裕,黒橋禎夫,佐藤理史, “国語辞典に基づく平易文へのパラフレーズ”, 情報処理学会研究報告, NL-144-23, pp. 167-174, 2001.

[7] 藤田篤, 乾健太郎, 松本裕治, “自動生成された言い換え文における不適格な動詞格構造の検出”, 情報処理学会論文誌, Vol.45, No.4, pp. 1176-1187, 2004.