国語辞典を使った放送ニュースの名詞の平易化

美野　秀弥　　田中　英輝

NHK 放送技術研究所 〒157-8510 東京都世田谷区砧 1− 10− 11

E-mail:　{mino.h-gq , tanaka-h.ja}@nhk.or.jp

1. はじめに

本稿では、国語辞典を使った気象災害ニュースの語彙の平易化について報告する。近年、永住、長期滞在の外国人の数が急増していることから、いろいろな情報を「やさしい日本語」で伝える動きが広がりつつある。特に生活に必須な自治体のお知らせ[1]や、災害時の安全確保に関わる気象災害ニュースを「やさしい日本語」で伝える動きが活発である[2]。これは、永住、長期滞在の外国人は初歩的な日本語能力を持っていることが多く、これに合わせた「やさしい日本語」であれば理解してもらえる可能性が高いことによる[3]。

著者らは NHK のニュース、特に、気象災害ニュースを対象とした平易化の検討を始めた。ニュースの多くはデジタル放送、あるいはインターネットで文字によって提供されている。これらは日本語母語話者を対象としているため、日本語レベルが初級の人にはわかりにくい可能性が高い。そこで、通常のニュースに加えて、平易にした文字ニュースのサービスを想定し、問題点を検討することとした。

このサービスの実現には、まず外国人にとってのやさしい日本語の基準が必須となる。これまで、野元らの提案[4]に始まり現在もさまざまな「やさしさ」の基準が提案されているが[1]、まだ統一的な基準は作られていない。そこで、著者らは先行研究の知見を元に実験的に基準を作っていくのが現実的だと考えている。

これまでのやさしさの基準は、およそ表記、語彙、構文のレベルで検討されている。本稿では、第一歩として語彙、特に名詞の平易化について検討した。語彙についてのやさしさの基準はいくつか提案されているが[5,6,7]、ここでは佐藤ら[8]に従い、日本語能力試験(JLPT)を利用した。

名詞の平易化は、難しい語とやさしい語の対を獲得し、利用した。対の獲得には、小学生用の国語辞典[9]を利用した 2 手法を検討した。

以下、2 節で提案手法を説明し、3 節でニュース記事中の語彙の平易化の実験概要と結果を報告する。最後に 4 節でまとめを行う。

2. 提案手法

2.1. 概要

用言は、格パタンを持っていることから、その言い換えには周辺の情報が不可欠である[10]。一方、名詞などの体言は、格パタンを持たない。そこで、難しい名詞は、名詞単独に着目して言い換えることができると仮定した。そして、ニュース中の難しい名詞をやさしく言い換えるには、難しい名詞とそれに対応する平易な名詞の対を獲得しておき、それらを使う手法を考えた。

本稿では、語彙のやさしさの基準として佐藤らの基準[8]を採用し、日本語能力試験(JLPT)の 3,4 級の語彙をやさしい語（平易語）、1,2 級、および級外の語彙を難しい語（難語）とした。

ニュース中の名詞の難語の平易化の具体的な手順は以下の通りである。

i) 平易化対(難語,平易語)を獲得

ii) ニュース中の単語の難語・平易語を認定

iii) 獲得した平易化対を用い、ニュース中の難語を平易語に置換

図 1 の例では「校舎の壁にひびが入っている」というニュース文は、平易化対(校舎,建物)を用い、「建物の壁にひびが入っている」と言い換えられる。

この手順に従い、2.2.(手法 1)と 2.3.(手法 2)で、国語辞典を用いた 2 つの平易化対獲得手法を提案する。手法 1 は鍛冶らの手法[11]を参考に、手法 2 は藤田らの手法[12]を参考にしている。そして、2.4.で、ニュース中の単語に級を付与して、難語・平易語を認定する手順を説明する。

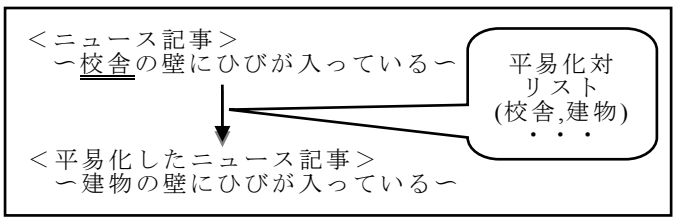

図 1 ：難語から平易語への言い換え

2.2. 手法1:国語辞典の見出し語と語釈文からの平易化対獲得

一般に、国語辞典の見出し語を平易に説明したものが語釈文だと考えられる。そこで、国語辞典の見出し語に対応する平易語が語釈文内に含まれていると仮定して、語釈文の中から言い換え可能な平易語を獲得する手法を検討した。図 2 の見出し語「過ち(1 級)」を例にして、平易化対獲得の手順を示す。

i) 主要文抽出

鍛冶らは、国語辞典の語釈文は、言い換えに使える要素と冗長表現と不要表現からなるとしている[10]。また、語釈文を係り受け解析し、パタンを使うことで、冗長表現と不要表現を取り除く手法を提案している。そこで、鍛冶らの手法を用いて言い換えに使える部分を抽出することとした。表 2 に抽出に使ったパタンの一部を示す。

本稿では、抽出した言い換えに使える部分を「主要文」と呼ぶことにする。図 2 では、語釈文中の「こと」が不要表現となり、主要文は「まちがう~~こと~~。うっかりしてやった失敗。過失。」となる。

ii) 平易化対候補

著者らは、名詞の見出し語では主要文の最終文節(係り受け木の根ノード)が、見出し語の意味を表していると考えた。そこで、主要文の最終文節中の自立語を言い換え可能な平易語の候補とした。図 2 では、見出し語「過ち(1 級)」に対して、「まちがう(2 級)」、「失敗(3 級)」、「過失(2 級)」が候補となる。

iii) 平易化対獲得

上記の候補の内、平易語(3,4 級)であるものを採用する。図 2 では、「失敗（3 級)」が平易語なので、(過ち, 失敗)が平易化対となる。

以上の手順により、国語辞典の名詞見出し語 35,083 に対して、16,172 の平易化対を獲得した。

この手法を使うと語釈文内の最終文節中の自立語が難語、見出し語が平易語となる平易化対も獲得できる。しかしこれを使ったニュースの平易化の予備実験では、不適切な言い換えが多く、採用しなかった。図 3 の例は、このような平易化対が得られる様子を示している。ここでは(鉄道,地下鉄)が平易化対となるが、鉄道を地下鉄で言い換えると不適切であった。

これは、多くの場合、見出し語が下位語に、語釈文内の最終文節中の自立語が上位語になっているためである。下位語「地下鉄」をその上位語である「鉄道」に言い換えることはできても、逆の言い換えはできない場合が多い。

種類	パタン
不要表現	「の一つ。」,「動詞＋こと。」など
冗長表現	「*など、」、「*のように、」 「連用形＋、」,「タ形＋、」など

表 2：冗長表現、不要表現の一部

見出し語：過ち(1 級)
語釈文：まちがうこと。うっかりしてやった失敗。過失。
i) 主要文：まちがう。うっかりしてやった失敗。過失。
ii) 候補：「まちがう(2 級)」、「失敗(3 級)」、「過失(2 級)」
iii) 平易化対：(過ち,失敗)

図 2：平易化対獲得の例(見出し語が難語)

見出し語：地下鉄(4 級)
語釈文：地下にトンネルをほって走るようにした鉄道。
i)主要文：語釈文に同じ
ii)候補：「鉄道(2 級)」
iii)平易化対：(鉄道,地下鉄)

図 3：平易化対獲得の例(見出し語が平易語)

2.3. 手法2:国語辞典の見出し語同士からの平易化対獲得

国語辞典の見出し語は、語釈文（主要文）間の類似度が高ければ、類似していると言える。この性質を利用して、平易化対を獲得する手法を考えた。図 4 を例にして手順を示す。

i) 主要文の抽出

前手法と同様に、語釈文から主要文を抽出する。図 4 では、「会釈」、「挨拶」の見出し語から主要文を得ている。係り受け解析の結果も併せて表示している。

ii) 主要文間の類似度の算出

次に、主要文間の類似度を式(1)により計算する。

類似度は、主要文間の共通語を探索し、共通語の重要度を加点していく。この重要度は、a)ーc)の 3 つの観点で計算されている。

a) 主要文の文長に対する補正
b) 最終文節からの距離に対する重み付け
c) 主要文における語の重要度（idf(逆文書頻度)に相当）

$$C(P_i,P_j)=\underbrace{\frac{1}{|P_i|}\frac{1}{|P_j|}}_{\text{a)}}\sum_{x\in P_i}\sum_{y\in P_j}\delta_x(y)\underbrace{\frac{1}{l(x)}\frac{1}{l(y)}}_{\text{b)}}\underbrace{\log\left(\frac{S}{s(x)}\right)}_{\text{c)}}\quad \delta_x(y)=\begin{cases}1 & (x=y)\\ 0 & (x\neq y)\end{cases} \tag{1}$$

P_i：国語辞典見出し語 i の語釈文の主要文（主要文を複数持つ見出し語の場合は別々に分ける）

$C(P_i,P_j)$：P_i,P_j 間の類似度

$|P_i|$：P_i 内の総文節数

x：P_i 内の各文節内の自立語

$l(x)$：x の文節から最終文節に係るまでの係り受けのエッジの数 ＋1（最終文節からの距離）

S：辞典内の主要文総数

$s(x)$：国語辞典内における、x の出現主要文数

図 4 では、「おじぎ」を共通語として抽出し、重要度を計算する。

iii) 平易化対の獲得

主要文間の類似度が基準値以上の場合、その見出し語同士を平易化対の候補とする。基準値は、予備実験を行って設定した。そして、見出し語同士が難語と平易語の対であれば、平易化対として獲得する。図 4 では、基準値を超えているとして、(会釈,挨拶)を平易化対として獲得する。

以上の手順により、平易化対の候補 315,978 に対して、7,261 の平易化対を獲得した。

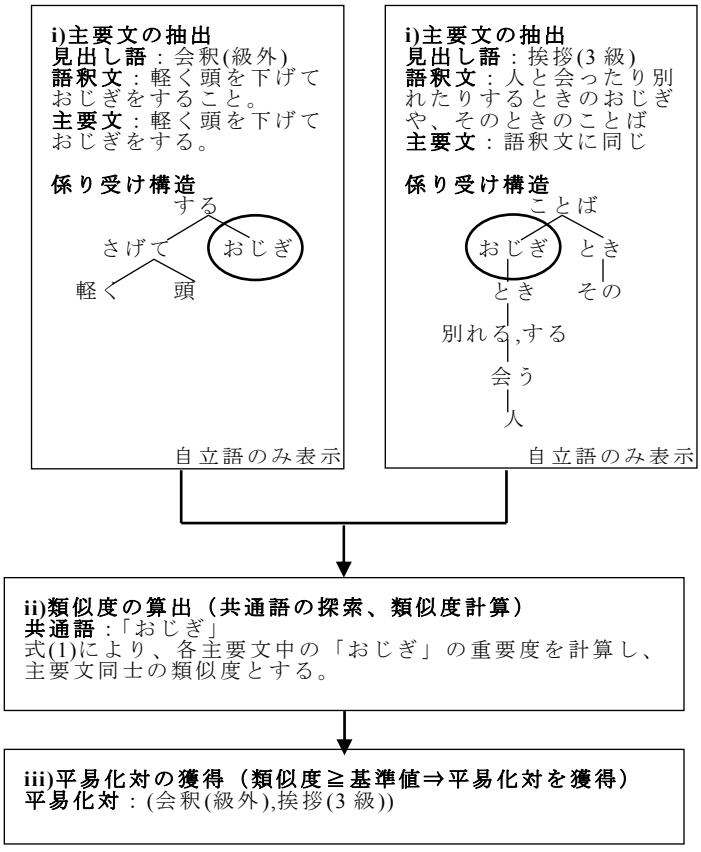

図 4：平易化対獲得の例

2.4. 級の付与

ニュース中の単語を平易語に言い換えるためには、級を付与して、難語・平易語を認定する必要がある。また、手法 1 と手法 2 を用いて平易化対を獲得するためには国語辞典の見出し語と語釈文にも同様の認定が必要になる。

ここでは、ニュース文と、国語辞典の見出し語、語釈文に級を付与する手順を説明する。

JLPT の見出し語は（校舎,こうしゃ,2 級）のように、(表記,読み,級)の 3 つの情報がある。このままでは、(表記,読み)のみを使った照合となるが、精度を上げるために品詞情報も使うこととした。JLPT の見出し語への品詞の付与には、形態素解析器(MeCab)を用いた。

・ ニュース文への級の付与

ニュース文は平文なので、形態素解析器を用い、(表記,読み,品詞情報)を付与した。

・ 国語辞典の見出し語と語釈文への級の付与

国語辞典の見出し語にはあらかじめ(表記,読み,品詞情報)が付与されているのでそのまま利用した。語釈文は平文なので、ニュース文と同様に付与した。

図 5 に、国語辞典の見出し語と語釈文に級を付与した例を示す。なお、照合できなかった単語は級外単語とした。助詞や助動詞などの機能語及び、表記のゆれなどにより照合できなかったものは照合の対象から除いた。

見出し語：過ち(あやまち,名詞,1 級)
語釈文：まちがう(動詞,2 級)こと。うっかり(副詞,2 級)してやった失敗(しっぱい,名詞,3 級)。過失(かしつ,名詞,2 級)。

図 5：国語辞典の級の付与例

3. 評価実験

3.1. 実験概要

新潟中越地震に関連した NHK の災害ニュース 11 記事（2004/10/23ー11/6）に、2 節で獲得した平易化対を適用し、結果を評価した。

1 記事は複数文から成っている。また、平易化の対象は、名詞の難語(1,2 級,級外)の内、品詞細分類(MeCab)が一般、副詞可能、形容動詞語幹のものとした。

ニュース記事と JLPT の見出し語の照合はニュース記事の形態素（MeCab）ごとに行った。このため、JLPT の見出し語が複数の形態素（MeCab）からなるときは、照合できなかった。これに対して、今回は特別な処理を行なっていない。

記事中の対象となる名詞の難語は 517 あった。

そして、手法 1 と手法 2 を使って得られた平易化結果をそれぞれ評価した。評価基準は、表 3 に示すように、言い換えることができるかどうかを評価する「言い換え可能性」と、言い換え可能かどうかに関わらず、語の説明としてふさわしいかどうかを評価する「類似性」の 2 項目を設定し、それぞれ 3 段階で評価した。

言い換え可能性	言い換え可能 2	修正が必要 1	言い換え不可 0
類似性	ほぼ同義 A	関連性有 B	全く異なる C

表 3：評価基準

3.2. 結果と分析

評価結果を表 4 に、評価例を表 5 に示す。

表 4 は手法 1 と手法 2 の Recall と Precision を示している。表 5 は評価値とニュース文の言い換え例である。評価値(2,A)は平易化前と平易化後で違和感なく置き換えることができる場合である。評価値(1,B)は言い換

えると多少違和感がある場合である。評価値(0,B)、(0,C)は、多少意味を想像できるものもあるが、置き換えると不適切になる場合である。

手法	総数 Recall	言い換え可能性			類似性		
		2	1	0	A	B	C
1	251	28	69	154	34	141	76
	49%	11%	27%	61%	13%	56%	30%
2	31	7	0	24	6	2	23
	6%	22%	0%	77%	19%	6%	74%

表 4：評価結果(難語総数：517 語)

例(評価値)	平易化前	平易化後
(2,A)	学校も含めた市内(全域)でー	学校も含めた市内(全体)でー
(1,B)	(無事)を喜びー	(元気)を喜びー
(1,B) サ変名詞	さらに詳しく(分析)する	さらに詳しく(調べる)する
(0,B) サ変名詞	道路が(寸断)されー	道路が(切る)されー
(0,C)	技術(センター)は、	技術(人)は、

表 5：評価例(ニュース文の一部)

・ Recall

平易化の対象となった 517 語の内の平易化できた語の割合を Recall とする。手法 1 は 49%、手法 2 は 6%である。特に手法 2 は全体(517)の内、31 しか平易化できていない。

手法 1 と手法 2 の差は、それぞれの手法で獲得した平易化対の数の差によるものと思われる(手法 1:16,172、手法 2:7,261)。手法 2 では見出し語の対が獲得される。この場合、候補はたくさん獲得されるものの、平易化の関係になるものが少なく平易化対の数が小さくなっている。

・ Precision

「言い換え可能性」の評価得点が 1 と 2 のものを言い換えとして有効とすると、手法 1 は 38%、手法 2 は 22%となり、手法 1 の方が高い。また、「類似性」の評価が A と B のものを、語の説明として有効とすると、手法 1 は 69%、手法 2 は 25%となり、同様に手法 1 の方が高い。

この結果から次のようなことが考えられる。手法 1 で獲得した平易化対は、上位下位関係にあるものが多い。一方、手法 2 で獲得した平易化対は、文脈が類似している。すなわち、言い換えにおいては、文脈が類似した語より、上位下位関係にある語を使うのが有効である。

最後に本実験で見つかった問題点を記す。

・ サ変名詞の平易化

(1,B)と評価した部分はサ変名詞の平易化が多かったが、本手法は十分に対応できていない。表 5 の例「さらに詳しく分析する」は、平易化対(分析,調べる)により平易化されるが、平易化対が(名詞,動詞)なのでそのまま埋め込むことはできない。平易化対を、(サ変名詞＋する,動詞)とすべきである。

・語義曖昧性

対象となる語が語義を複数持っている場合、これを解消しなくてはならないが、本手法は行っていない。表 5 の例では、平易化対(センター,人)によって「技術センター」が「技術人」に置き換えられている。これは、平易化対を、語釈文「中心になる施設」ではなく、「外野の真ん中を守る人」から獲得しているからである。平易化対を正確に適用するには、平易化対内の単語と文脈内の単語との間で、語義の照合が必要である。

4. おわりに

本稿では、国語辞典を用いた名詞の平易化手法を提案し、気象災害ニュース記事の平易化の検討を行った。

提案した 2 手法の内、手法 1(国語辞典の見出し語の平易な言い換えを語釈文から獲得する手法)の方が平易化対獲得に有効な手法であることが分かった。しかし、手法 1 でも、全ての難語を平易語で記述することはできなかった。平易化対の不足が 1 つの原因であり、単一の辞書を用いることには限界がある。今後は、複数の辞書を用いることを考えたい。また、サ変名詞の平易化や語義曖昧性の解消も今後の課題である。

文　献

[1] 庵, 岩田, 森, 「やさしい日本語」を用いた公文書の書き換え, 2009 年度日本語教育学会秋季大会, pp.135-140, 2009

[2] 蔡垂功, 外国人への災害情報提供を巡る事例と取り組み, 日本災害情報学会第 7 回研究発表大会, pp.157-162, 2005

[3] 松田, 外国人のための災害時の日本語, 月刊言語, vol.28(8), pp.42-51, 1999

[4] 野元, 川又, 義本, 簡約日本語の創成, 日本語学, vol.10(4), pp.94-105, 1991

[5] 国立国語研究所, 日本語教育のための基本語彙調査, 1984

[6] 秋本, 押尾, 新しい日本語能力試験のための語彙表・漢字表作成中間報告, 日本語学, vol.27(10), pp.36-49, 2008

[7] 橋本, 山内, 日本語教育のための語彙リストの作成, 日本語学, vol.27(10), pp.50-58, 2008

[8] 佐藤, 災害時の言語表現を考える, 日本語学, vol.23, pp.34-45, 2004

[9] 例解小学国語辞典, 第 5 版, 三省堂

[10] 鍛治, 河原, 黒橋, 佐藤, 格フレームの対応付けに基づく用言の言い換え, 自然言語処理, 10(4), pp.65-81, 2003

[11] 鍛治, 黒橋, 佐藤, 国語辞典に基づく平易文へのパラフレーズ, 自然言語処理, 2001

[12] 藤田, 乾健太郎, 乾裕子, 名詞言い換えコーパスの作成環境, 電子情報通信学会, 2000