使用者数による語彙制限を用いた日本語学習者のための文章読解支援

塩田健人，梶原智之，小町守（首都大学東京） shioda-kent@ed.tmu.ac.jp

はじめに

語彙平易化とは，難解な語や句を平易な語や句に言い換えることにより，子どもや言語学習者などの文章読解を支援する技術である

特に読者の**理解語彙**に言い換え対象の語彙を制限することで，文章の可読性を向上させることができる．そこで，本研究では複数の尺度を用いて語彙制限した言い換えを行い，文の難易度について日本語学習者の主観評価を受けた

先行研究

SemEval-2012 English Lexical Simplification Taskでは，単純頻度のみを使用したベースラインシステムが全12システム中2位の成績を示し，語彙平易化タスクにおける高頻度語への言い換えの有効性が示された（Specia et al., 2012）

国語辞典に収載されている語を「自然な日本語」として判定するタスクにおいて，Twitterに投稿されたテキストから獲得した「語の使用者数」という統計量が単純頻度よりも優れた指標であることが示された（Aramaki et al., 2013）

[1] Lucia Specia, Sujay Kumar Jauhar, Rada Mihalcea. SemEval-2012 Task 1: English Lexical Simplification. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012), pp.347-355, 2012.
[2] Eiji Aramaki, Sachiko Masukawa, Mai Miyabe, Mizuki Morita, Sachi Yasuda.Word in a Dictionary is used by Numerous Users. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp.874-877, 2013.

実験手順

語彙的換言知識の作成

名称	品詞	例	換言対
動詞含意関係DB（Ver.1.3.1）	動詞	チンする→加熱する	89,784
日本語異表記対DB（Ver.1.1）	名詞	ゴミ置き場⇆ゴミ置場	5,513,606
基本的意味関係の事例ベース（Ver.1.4）	名詞	短大⇆短期大学	78,260
PPDB Japanese（Ver.0.2.0）	語→句	光速→光の速度	33,150
内容語換言辞書	語→句	案内→連れて行く	25,504
日本語WordNet同義語DB（Ver.1.0）	名詞	故障⇆トラブル	11,753
使用した全ての言い換え対			11,355,676

言い換え

実験対象：難解語を1語だけ含む文
- 平易語：各指標の上位N語に共通の語
- 難解語：平易語に含まれない全ての語

実験対象文に含まれる難解語を言い換える
- 換言知識を再帰的に用いて，平易語になるまで言い換える

警察の車を**乗っ取る**→**奪う**→**取る**

評価

言い換え前後で意味が保持できているかを主観評価
↓
日本語能力試験N1級保持者（1名）
- 難解文と平易文を読み，理解の可否を◯×で評価
- 難解文と平易文を平易な順にランキングを付ける

言い換えをすることにより，理解できるようになった文について，日本語教育語彙表を用いて平易化された語彙の難易度を調査した

実験結果

1，使用したデータ

指標	データ	語数
頻度	Web日本語Nグラム	2,565,424語
	Twitter	48,324語
使用者数	Twitter	48,324語

2，語彙平易化（各指標上位N語への語彙制限）

N = 5,000	原文	頻度（Web）	頻度（Twitter）	使用者数
◯	85	85	86	91
◯→×	-	15	9	8
N = 7,500	**原文**	**頻度（Web）**	**頻度（Twitter）**	**使用者数**
◯	88	77	91	93
◯→×	-	21	8	7
N = 10,000	**原文**	**頻度（Web）**	**頻度（Twitter）**	**使用者数**
◯	81	86	89	85
◯→×	-	11	9	9

3，ランキング評価結果

- 全てのNにおいて言い換えた方が原文より分かりやすい
- N = 5,000，7,500の場合，使用者数 = Twitter頻度 > Web頻度
- N = 10,000の場合，Web頻度 > 使用者数 = Twitter頻度

4，言い換え後の語彙の難易度

- 言い換えをすることにより理解できるようになった文のうち，言い換え前に比べ，難易度が上がっている語はなかった
- 全ての指標において上級前半の語が言い換えられていた

考察

- Web日本語Nグラムの頻度よりもTwitterの頻度や使用者数で語彙制限する方が平易語とする語彙の数が少ない場合（Nが小さい場合）には学習者の読解支援に有効である
- 使用者数で語彙制限すると，誤って難解になりにくい
- 今後はN2級～N4級保持者に評価をしてもらう必要がある