日本語学習者の文章読解支援のための語彙制限

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

概要

本研究では，頻度および使用者数の指標を用いた語彙平易化を行い，日本語学習者の文章読解支援のための語彙制限を実現する．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で，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．本研究では，公開されている日本語の語彙的換言知識を組み合わせて言い換えを行い，文の難易度について日本語学習者の主観評価を受けた．その結果，Web日本語Nグラムから得た頻度よりも，Twitterから得た使用者数の指標で行った語彙制限が読解支援に有効であった．

先行研究

SemEval-2012 English Lexical Simplification Taskでは，単純頻度のみを用い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）

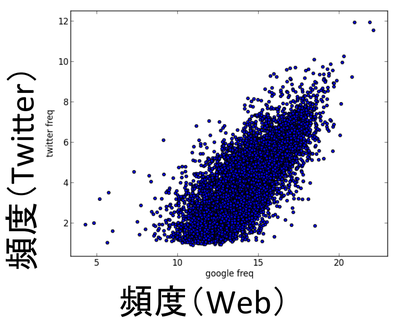
「自然な日本語」を表す「語の使用者数」という新たな統計量が提案された（Aramaki et al., 2013）．
ここで取り扱っている「自然な日本語」は，平易な日本語と関連がある可能性がある．

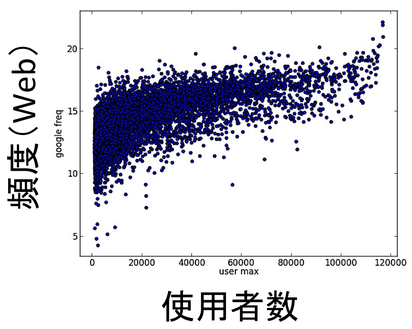
[1] Lucia Specia, Sujay Kumar Jauhar, Rada Mihalcea. SemEval-2012 Task 1: English Lexical Simplification. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012), pp.347-355, 2012.
[2] Eiji Aramaki, Sachiko Masukawa, Mai Miyabe, Mizuki Morita, Sachi Yasuda.Word in a Dictionary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

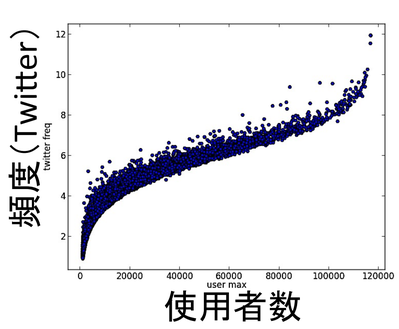
実験結果

1，各指標の相関

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（各指標の上位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 = 7,000	**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

語彙制限の手順

語彙的換言知識の作成

名称	換言対	品詞	例
PPDB: Japanese（Version0.2.0）	33,150	語→句	光速→光の速度
内容語換言辞書	25,504	語→句	案内→連れて行く
動詞含意関係DB（Version1.3.1）	89,784	動詞	チンする→加熱する
日本語異表記対DB（Version1.1）	5,513,606	名詞	ゴミ置き場↔ゴミ置場
基本的意味関係の事例ベース（Version1.4）	78,260	名詞	短大↔短期大学
日本語WordNet同義語DB（Version1.0）	11,753	名詞	故障↔トラブル
獲得した全ての言い換え対			**11,355,676**

語彙制限：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る
  ・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

・日本語能力試験N1級保持者（1名）
1. 難解文と平易文を読む
2. 理解できたら○，理解できなかったら×
3. 難解文と平易文を平易な順にランキング

考察

・Web日本語Nグラムの頻度よりもTwitterの頻度や使用者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
・N1級の学習者は難解文が読める（被験者に不相応）
・今後はN2級～N4級保持者に評価をしてもらう必要がある