均衡コーパスを用いた語彙平易化データセットの構築

首都大学東京　小平 知範　梶原 智之　小町 守

首都大日本語語彙平易化データセット

github.com/KodairaTomonori/EvaluationDataset

研究室データHP:http://cl.sd.tmu.ac.jp/research/dataset

語彙平易化

難しい文：はるかに変化**に富む**
易しい文：はるかに変化**が多い**

第二言語学習者，子供等の読解を支援する目的

評価用データセットの有用性

語彙平易化システムの自動評価の実現
人手評価のコストと再現性の課題を克服

先行研究のデータセット例

文と対象語	Kajiwara and Yamamoto (2015)
「技を出し合い，気分が**高揚する**のがたまらない」とはいえ，技量で相手を上回りたい気持ちも強い	
難易度ランキング	
1.高まる 高ぶる 2.上がる 3.**高揚する** 4.興奮する	

本研究のデータセット例

文と対象語
もっとも安上がりにサーファー**を装う**方法は，ガラムというインドネシアさんのタバコを，これ見よがしに吸うことです.
難易度ランキング
1.のふりをする 2.に見せかける 3.の真似をする のふりをする 4.を真似る 5.に成りすます 6.**を装う** 7.を偽る

問題点と解決策

新聞コーパス内の文のみで構成されている

新聞コーパスのみだと分野に偏りが出てしまうため，本研究では，BCCWJから文を選定

難易度ランキングで同順を許していない

アノテータに言い換えを並び替えてもらう際，同程度の難易度でも順位をつけるという制約をつけていたが，同程度の難易度の語は存在するため,同順をつけることを許可した

文中に難解語が残ってしまう

対象語を平易化後，文中に難解語が残ってしまうことがあるので，これをなくすため，文の選定時に難解語を一語しか含まないものとした

助詞の交替を考慮してない

用言の言い換えには助詞の交替が起こることがあるが，考慮されていないので，言い換え獲得時にこれも考慮に入れた
(統合ランキングでは，同順は含まれる)

データセットの比較

各指標での正解率

頻度，使用者数，親密度，語彙表，JLPTの5つの指標で一番平易なものを当てた場合に正解として正解率を出し，言い換えのランキングで得られた統合してない人手のランキングデータに対しても，同じく正解率を出し，人手ランキングの正解率との相関を調べた.

	本研究	K&Y	人手
頻度	41.6	35.8	41.0
使用者数	32.9	25.0	31.5
親密度	30.4	31.5	32.5
語彙表	38.2	35.7	38.7
JLPT	42.0	40.9	43.3
相関係数	0.963	0.930	N/A

統合ランキングと人手との相関

平均スコアで統合したもの(ベース)と，信頼度の低い人を除いた後平均スコアで統合したもの(除外)との比較

	ベース	除外
順位相関係数	0.541	0.580

データセット構築の流れ

文の抽出

例：はるかに変化に富む

難解語：日本語教育語彙表の上級の単語
対象語(各30語)：動詞，名詞，形容詞，形容動詞，副詞，サ変名詞，サ変動詞
文：BCCWJから難解語を1語しか含まない文を抽出
難解語1語につきランダムに10文選択

言い換えの獲得

が多い，に金持ち，が豊富

クラウドソーシングを利用し，5人のアノテータが難解語の言い換えを列挙，この際前後の助詞を含めた言い換えを許可

言い換えの評価

が多い，~~に金持ち，~~が豊富

クラウドソーシングを利用し，5人のアノテータが前行程で獲得した言い換えが正しいかどうか判定
過半数が適切な言い換えだと判断したものを採用

ランキングの獲得

1.が多い，2.に富む，3.が豊富 ×5

クラウソーシングを利用し，5人のアノテータが難解語と言い換えを平易な順にランキング．同順を4つまで許可

ランキングの統合

1.が多い，2.に富む，3.が豊富

5つのランキングを平均で統合．MLEでアノテータの信頼度(Matsui et al.)を獲得し，信頼度の低い人を除外した