語彙平易化システム評価のためのデータセットの改良

小平知範、梶原智之、小町守　　首都大学東京 kodaira-tomonori@ed.tmu.ac.jp

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**

平易化

語彙平易化は、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の例

難解語を含む文

「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。

言い換えリスト[平易な順]

高まる 高ぶる　上がる 高揚する 興奮する

(Kajiwara and Yamamoto 2015)

問題点と解決策

対象の難解語を平易にしても
文中に他の難解語が残っていた

対象の難解語平易にしたら、
文全体が平易になるようにしたい

上記の例では難解語[技量]が残る

難解語を一語のみ含む文を選定

対象語の前後の助詞を含めた
言い換えを考慮していない

用言の言い換えの際に助詞の交替が起こる

上記の例では、[の高まり]に言い換えられる

言い換え獲得時に助詞を含めた言い換えを考慮

ランキングの統合方法が単純

複数人のアノテータのランキングの
平均を取る単純なもの

最尤推定を用いた順序統合方法

同順を考慮していない

上記の例の[高まる][高ぶる]のように同程度の
難易度の語を並び替える時に同順として扱えない

並び替える時に同順位も許可

データセット構築の流れ

文の抽出

難解語：日本語教育語彙表ver.1.0の上級の単語
対象語：動詞、名詞、形容詞、形容動詞
副詞、サ変名詞、サ変動詞
BCCWJコーパスから難解語を
１語しか含まない文の抽出
文脈に依存して言い換えも難易度も変わるため、
難解語１語につきランダムに１０文選択

言い換えの獲得

クラウドソーシング（ランサーズ）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
一致率：19.3%

言い換えの評価

クラウドソーシングを利用し、５人のアノテータによる
獲得した言い換えが正しいかどうかの判定
過半数が適切な言い換えだと判断したものを採用
一致率：66.9%

言い換えのランキング

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+言い換えを平易な順にランキング
同順を４つまで許可
スピアマン相関係数：　0.559
先行研究は　0.332

ランキングの統合

BaseLine(K & Y 2015)
５人の平均をとり昇順に並べる

最尤推定を用いた順序統合方法
(Matsui et al.2014)

提案手法

最尤推定を用いた順序統合方法で求めた
アノテータのパラメータを用いて、
不真面目なアノテータを除いた方法

	Kajiwara 法	Matsui 法	提案手法
スピアマン相関係数	0.559	0.545	0.573

これから

(K & Y2015)と本研究のデータセットを
複数の語彙平易化システムを用いて比較する。

[1]Tomoyuki Kajiwara and Kazuhide Yamamoto. Evaluation Dataset and System for Japanese Lexical Simplification. ACL 2015 Student Research Workshop.
[2]Toshiko Matsui, Yukino Baba, Toshihiro Kamishima and Hisashi Kashima. Crowdordering.
In Proc. 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp.336-347, Tainan, Taiwan, 2014.