Simple PPDB: Japanese

https://github.com/tmu-nlp/simple-jppdb/

首都大学東京　梶原智之　小町守　kajiwara-tomoyuki@ed.tmu.ac.jp

Kajiwara-15a
国語辞典の見出語と定義文から自動的に獲得した言い換え辞書を用いる先行研究

Kajiwara-15b
人手で構築された複数の言い換え辞書を用いる先行研究

Glavaš-15
w2vとcos類似度で類義語を集め、頻度や言語モデルでリランキング

日本語のテキスト平易化のために大規模な2つの辞書を構築しました

1. 単語難易度辞書（57万語）
   - 【初級】夕食
   - 【中級】ディナー
   - 【上級】晩餐

2. 平易な言い換え辞書（34万語対）

晩餐	（上級）→	（初級）夕食	0.317
晩餐	（上級）→	（中級）ディナー	0.176
ディナー	（中級）→	（初級）夕食	0.217

語彙平易化の評価

$$Accuracy = |ref \cap output| / |input|$$

$$Precision = |ref \cap output| / |output|$$

$$Changed = |output| / |input|$$

語彙平易化	Accuracy	Precision	Changed
Kajiwara-15a	0.060	0.114	0.522
Kajiwara-15b	0.127	**0.236**	0.539
Glavaš-15	0.135	0.181	0.746
本研究	**0.181**	0.210	**0.861**

Kajiwara-15a	Kajiwara-15b	Glavaš-15	本研究

例１　こうして企業の【筆頭】｛トップ, 先頭, 頂点｝に立つ人間は、社内で最年長の人間ということになる。

※【難解語】↓左ほど平易 ｛平易な言い換え｝

最初	先頭	中心	トップ

例２　そしてこの調査は【疑わしい】｛疑問がある, 怪しい｝。

***	変だと思う	興味深い	怪しい

例３　なるほど、立場が上の人が、下の者にたいして、相手を尊重して【謙虚な】｛おとなしい, 控えめな｝態度で接するのはよいことだ。

***	***	誠実な	彼な

① Wikipediaの本文に5回以上出現する571,023単語に、3段階の難易度を付与

② PPDB: Japaneseのうち、Wikipediaの本文に5回以上出現する単語のみからなる512,284単語対について、単語難易度を付与して平易な言い換え辞書を構築

③ 日本語の語彙平易化タスクでの性能を評価 (Kodaira et al., 2016)

今後の課題（上のURLで辞書を更新していきます）
- 句への拡張
- 言い換え確率の改善
- 難易度推定精度の改善

3クラス分類のAccuracy	① 単語の 難易度	② 単語対の 難易度差
Baseline（頻度+閾値）	0.557	0.497
基本素性（文字数, 文字種, 頻度）	0.582	0.508
基本素性 + CBOW（100次元）	0.689	0.591
基本素性 + SGNS （100次元）	**0.708**	**0.607**

辞書の統計	難易度	収録語数	重複
JLPT: 日本語能力試験	4段階	7,759	95.6%
JEV: 日本語教育語彙表	6段階	17,207	95.6%
本研究	3段階	571,023	ー

- SVM（RBFカーネル）を使って3段階（初級・中級・上級）の単語の難易度を推定
- 教師データは日本語教育語彙表（JEV）の3段階の単語難易度
- 素性は「文字数」「文字種（平仮名/片仮名/漢字）」「対数頻度」「単語分散表現」
- 単語分散表現を使う気持ちは「難解な単語は難解な文脈で使用されやすく、平易な単語は平易な文脈で使用されやすい。」
- WikipediaとJEVの両方に出現する16,447語（MeCabとNEologdで単語分割）を使って10分割交差検証した精度（Accuracy）を評価 ①

- 日本語の最大規模の言い換え辞書である PPDB: Japanese (Mizukami et al., 2014) のうち、JEVの単語のみからなる40,309単語対を使って、単語対の難易度の差を推定
- ①と同様に各単語の難易度を求め、「言い換え先が平易」「言い換え先が難解」「言い換え元と言い換え先が同じ難易度」の3クラス分類を行ったときの精度を評価
- 同様の3値分類を実施している英語の先行研究 (Pavlick and Callison-Burch, 2016) はAccuracy=0.604、本研究でも同等の精度で単語対の難易度差を推定できた ②

梶原智之, 小町守. Simple PPDB: Japanese. 言語処理学会第23回年次大会, P8-5, pp.529-532, March 15th, 2017.