2D-01

中国語母語話者を対象とした日本語単語の難易度推定

林　妙玉*　　白井　清昭†

北陸先端科学技術大学院大学 先端科学技術研究科

1 はじめに

日本語教育において，日本語単語の難易度は重要な役割を果たす．学習者に教える単語の優先順位を決めたり，初学者に対して難しい単語の使用を避けたりするなど，単語の難易度を必要とする場面は多い．一般に，単語の難易度は日本語学習者の背景知識に依存すると考えられる．例えば，中国語を母語とする学習者は漢字を知っているので，漢字を使用しない言語を母語とする学習者よりも漢字表記の単語を学習しやすいと考えられる．ところが，日本語単語の難易度に関する先行研究 [3, 4] では，学習者の母語の違いは考慮されていなかった．

本稿では，日本語の辞書と中国語の辞書における語義の違いや漢字表記の違いに基づき，中国語母語話者を対象とした日本語単語の難易度の推定手法を提案する．

2 提案手法

2.1 日本語単語難易度の定義

本研究では，文化庁による漢語の分類 [1] を基に，日本語単語の難易度を表 1 のように定義する．難易度のクラスは，難易度が低い順に，大きく S,O,D,N の 4 つのクラスに分かれている．漢字で表記された日本語単語に対し，それと同じ単語が中国語にない場合は N とする．それ以外のとき，日本語単語と中国語単語の語義を比較し，両者の語義の類似性に応じて S,O,D と判定する．また，クラス S,O,D は，日本語単語と中国語単語の漢字表記が同じときは X-1，異なるとき (例えば「議論」と「议论」) は X-2 のように細分化する．

2.2 単語難易度の推定

日本語単語の難易度の推定手法を以下に示す．

1. 難易度を推定する日本語単語を w^J とする．ここでは w^J は漢字で表記された単語に限る．
2. 中国語辞書を検索し，w^J と同じ表記を持つ中国語の単語 w^C を検索する．もし見つからない場合，日中漢字マッピングテーブルを参照し，日本語の漢字を中国語の漢字に変換してから w^C を検索する．

表 1: 本研究における日本語単語の難易度の定義

クラス	説明	漢字表記
S-1	Same(日本語単語と中国語単語の語義が全て同じ)	同じ
S-2		異なる
O-1	Overlap(日本語単語と中国語単語の語義の一部が同じ)	同じ
O-2		異なる
D-1	Different(日本語単語と中国語単語の語義が全く異なる)	同じ
D-2		異なる
N	Nothing (日本語と同じ単語が中国語にない)	–

3. w^C が見つからない場合，w^J の難易度クラスを N と判定する．それ以外は次のステップへ．
4. 日本語の辞書から w^J の語釈文の集合 $S^J = \{s_1^J, \cdots, s_n^J\}$ を得る．s_i^J は w^J の i 番目の語義の語釈文を表す．同様に，中国語の辞書から w^C の語釈文の集合 $S^C = \{s_1^C, \cdots, s_m^C\}$ を得る．本研究では，日本語の辞書として岩波国語辞典を，中国語の辞書として白水社中国語辞典[1]と現代漢語詞典 [2] を用いる．現代漢語詞典では語釈文は中国語で書かれているが，Baidu 通用翻訳 API[2]を用いて日本語に翻訳する．
5. 日中の語義の集合 S^J と S^C に対し，語義の対応付けを行い，その結果に応じて w^J を S,O,D のいずれかに分類する．詳細は 2.3 項で述べる．
6. w^J と w^C の表記が一致しているときは S-1, O-1, D-1 のいずれかを，一致していないときは S-2, O-2, D-2 のいずれかを難易度クラスとする．

2.3 語義の対応付け

S^J と S^C から難易度クラス S,O,D のいずれかを判定するアルゴリズムを Algorithm 1 に示す．4 行目の $sim(s_i^J, s_j^C)$ は 2 つの語釈文の類似度を表す．ここでは，文 s_i^J に含まれる単語の分散表現の平均ベクトルを $\vec{v}_i^J$，同様に文 s_j^C の文のベクトルを $\vec{v}_j^C$ とし，2 つのベクトルのコサイン類似度を $sim(s_i^J, s_j^C)$ とする．日本語の語義と中国語の語義の全ての組み合わせのうち，$sim(s_i^J, s_j^C)$ が最も高くなる語義の組を (i', j') とし (4

Estimation of Difficulty of Japanese Words for Chinese Native Speaker

*Miaoyu Lin, Japan Advanced Institute of Science and Technology, †Kiyoaki Shirai, Japan Advanced Institute of Science and Technology

[1] https://cjjc.weblio.jp/category/cgkgj

[2] https://fanyi-api.baidu.com/

Algorithm 1 S, O, D の判定

```
1: procedure DIFFICULTY-CLASS(S^J, S^C)
2:     SA = ∅
3:     while S^J ≠ ∅ and S^C ≠ ∅ do
4:         (i', j') ← arg max_(i,j) sim(s_i^J, s_j^C)
5:         sim_sense ← sim(s_i'^J, s_j'^C)
6:         if sim_sense ≥ T_m then
7:             SA ← SA ∪ { (s_i'^J, s_j'^C, sim_sense) }
8:             S^J ← S^J \ {s_i'^J},  S^C ← S^C \ {s_j'^C}
9:         else
10:            break the loop
11:        end if
12:    end while
13:    if SA is ∅ then return D
14:    else if S^J ≠ ∅ or S^C ≠ ∅ then return O
15:    else return S
16:    end if
17: end procedure
```

行目)，その最大の語義間の類似度を sim_{sense} とする (5 行目)．語義 $s_{i'}^J$ と $s_{j'}^C$ は同じ意味を持つとみなして対応付け，集合 SA に追加する[3](7 行目)．対応付けられた語義を S^J と S^C から除き (8 行目)，残りの語義について同じ処理を繰り返す．ただし，sim_{sense} が閾値 T_m 以上ではないとき，つまり語義間の類似度が十分に大きくないとき，2 つの語義は同じ意味を持つとはみなさず，語義の対応付けを終了する (6,10 行目)．

語義の対応付けの終了後，13〜16 行目では，対応付けできる語義の組が 1 つも見つからないときは D と判定し，対応付けできる語義の組はあるが全ての語義について対応付けできないときは O と判定し，全ての語義について対応付けができたときは S と判定する．

Algorithm 1 は，中国語の辞書として白水社中国語辞典と現代漢語詞典を用いたときのそれぞれについて適用する．2 つの辞書による判定結果が異なる場合は，式 (1) に示す語義の対応付けのスコア (対応付けられた語義間の類似度の平均値) を算出し，大きい方の判定結果を採用する．

$$\text{sense-align-score} = \frac{1}{|SA|} \sum_{(*,*,sim_k)\in SA} sim_k \qquad (1)$$

3 評価実験

提案手法による難易度クラスの判定の性能を評価する．N とそれ以外の区別，X-1 と X-2 の区別は自明なので，ここでは S, O, D の判定結果のみを評価する．

[3] SA は対応付けされた語義の組とその語義間類似度を記録する．

表 2: 実験結果

	TEST	CV1	CV2	CV3	平均
T_m	0.196	0.34	0.23	0.23	–
正解率	0.763	0.731 (0.769)	0.785 (0.758)	0.753 (0.774)	0.756 (0.767)

資料「中国語と対応する漢語」[1] から，S,O,D に分類される漢語をランダムに選択した．この中から日本語辞書・中国語辞書に記載がないものを除いた．さらに第 1 著者によって S,O,D の分類をやり直した．最終的に 279 語の漢語からなるテストデータを用意した．難易度の内訳は，S が 136，O が 121，D が 22 となった．

実験結果を表 2 に示す．TEST はパラメタ T_m を 0.196 と設定したときの S,O,D の判定の正解率を示す．ここで T_m は以下のように決定した．テストデータにおける全ての単語の全ての語義の組み合せについて語義間の類似度を計算する．類似度の分布が正規分布にしたがうと仮定し，$[-\infty, T_m]$ の範囲の確率密度関数の確率の累積が全体の 20%になるように (全体の語義の組のうち 20%が対応付け不可となるように) 定めた．このときの正解率が 0.763 であることから，提案手法による難易度の判定はある程度妥当であると言える．

上記のパラメタ T_m は最適ではない可能性がある．そこで，3 分割交差検定を行い，3 分の 2 のデータを開発データとして T_m を最適化し，残りの 3 分の 1 のデータをテストデータとして判定の正解率を測った．表 2 の CV1〜CV3 は，交差検定の 3 回の試行について，最適化された T_m，テストデータの正解率，開発データの正解率 (括弧内の数値) を示している．正解率については 3 回の試行のマイクロ平均も示す．開発データとテストデータの正解率にやや差が見られるが，0.02〜0.04 ポイント程度の差に留まっている．また，T_m は 3 回の試行で大きな違いはない．提案手法による判定の正解率は T_m の設定に大きな影響を受けないと言える．

発表では，本論文で提案する日本語単語難易度の尺度が妥当かどうかを検証するために行った日本語学習者に対するアンケートの調査結果についても報告する．

参考文献

[1] 文化庁．中国語と対応する漢語．日本語教育研究資料, pp.85–143, 1978.

[2] Institute of Linguistics, Chinese Academy of Social Sciences. Contemporary Chinese Dictionary (現代漢語詞典), 5th edition. The Commercial Press, 2005.

[3] 劉志宇, 内田理. 日本語を学習する外国人を対象とした日本語テキスト難易度推定手法. 情報処理学会自然言語処理研究会, Vol. 2012, No. 11, pp. 1–5, 2012.

[4] Yuriko Sunakawa, Jae-ho Lee, and Mari Takahara. The construction of a database to support the compilation of Japanese learners' dictionaries. *Acta Linguistica Asiatica*, Vol. 2, No. 2, pp. 97–115, 2012.