N-024

CEFR 読解指標に基づく日本語例文分類手法の検討
On Japanese document classification method based on CEFR reading comprehension index

高田 宏輝† 宮崎 佳典‡ 谷 誠司*
Hiroki Takada Yoshinori Miyazaki Seiji Tani

1. はじめに

近年，相互理解や自律学習・生涯学習を重視する流れを受け，言語で何ができるか（Can-Do）という考えの基に記述された言語能力記述の枠組みに関心が集まっている．その中でも欧州評議会が開発した CEFR（Common European Framework of Reference for Languages，ヨーロッパ参照枠）[1]は，世界の外国語教育に導入されている．CEFR は具体的な言語能力レベルを A1 レベルから C2 レベルまでの 6 段階で設定しており，Reading，Writing，Speaking，Listening といった技能項目に対して，それぞれのレベルで言葉を使ってできることを能力記述文(Can-Do Statements,以下 CDS)で記述している．CDS は例として，「さらに詳細に読む必要があるかどうかを決定するために，広範囲にわたる専門的な話題についてのニュース，記事，レポートの内容と関連性をすばやく確認することができる．」のような抽象的な表現で記述されており，実利用とのイメージを結びつけるのが困難である場合がある．そのため，実際に言語能力レベルを調べる際には例文が用いられている．

CEFR に関連する研究としては英語教育で CEFR レベルごとに例文中に出現する文法事項の抽出[2]や例文中に出現する単語の共起性における難易度の測定[3] など，コーパスを用いて例文から CEFR のレベル決定における基準特性を抽出する研究が行われている．

一方，日本語教育分野に限ると，同様の取り組みは著者らの調べる限り多くは行われていない．これは日本語を対象とした CEFR では CEFR レベル情報を持つテキストコーパスが作成されておらず，分析に必要な資料が不足していることが一因であると考えられる．英語を対象とした CEFR には EFL 教材が多く存在し，それらを基に作成したコーパスを用いた研究が行われている．日本語教育においては CEFR に準拠したものは少ないため同様の手法を用いてコーパスを作成するのは容易ではない．そこで本研究では文章（例文）を与えることで CEFR 中の対応する CDS の項目番号を付与する分類器を作成し，テキストコーパス作成を支援することを目指す．さらに，テキストコーパスを利用することでレベル決定に寄与する要因の抽出を行う．

本発表では，例文を CDS 項目へ分類する際の足がかりとして，複数の CDS に現れる“専門性”に着目した分類の検討を行う．

†静岡大学大学院 総合科学技術研究科 情報学専攻
Department of Informatics, Graduate School of Integrated Science and Technology, Shizuoka University
‡静岡大学学術院情報学領域
College of Informatics, Shizuoka University
*常葉大学 外国語学部
Faculty of Foreign Studies, Tokoha University

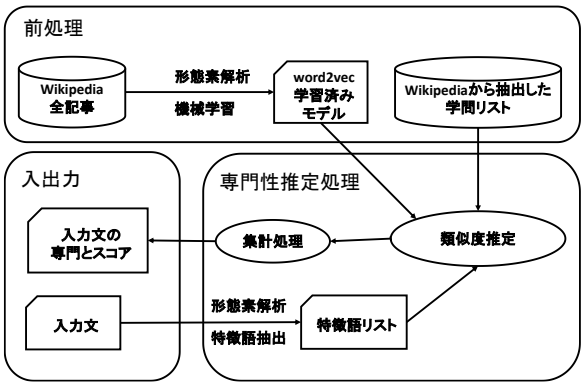

図 1 専門性推定手法の構成

2. 例文の専門性推定手法

専門性に着目した分類を行うために，例文から専門性の推定を行う手法を提案する．専門性推定処理の構成図を図 1 に示す．また，今回の専門性推定では専門分野を学問分野と仮に設定し，Wikipedia の“学問の一覧”ページより抽出した語を専門（用語）として利用する．

2.1 特徴語の抽出

例文の特徴を示す語を抽出する目的で，収集した各例文に対して形態素解析を行う．形態素解析には MeCab[4]に対して Wikipedia の見出し語を学習させたものを利用した．

同一 CDS に分類される例文を 1 つのグループとして Jaccard 係数を求め，各例文に対して上位 9 単語ずつをその例文の特徴語として利用する．ある語 w における Jaccard 係数については以下の値を計算する．

$$Jaccard = \frac{a}{F_1 + F_2 - a} \quad (1)$$

(1)式内において，aは wが当該例文内で出現した回数，F_1は当該例文が所属する CDS 内の例文数，F_2は wが当該 CDS における全例文中に出現した回数を表す．

2.2 専門性の推定

2.1 で求めた特徴語に対して Wikipedia から抽出した専門分野との類似度を比較することで，例文における専門性の推定を行う．ここで行う単語間の類似度の算出には word2vec[5]を用いる．この手法では単語の共起関係を基に単語間の意味関係を考慮したベクトル化を行う．類似した意味を持つ単語や関連する語はベクトルの角度も近接し，専門分野と特徴語の関係を推定することが可能であると考えられる．このとき，抽出した特徴語と専門の類似度を調べる際にはコサイン類似度を用いる．

各特徴語における類似度の高い専門を上位 3 つずつ抽出する．その結果，1 つの例文が 9 個の特徴語に対してそれぞれ 3 つの専門を持っている状態となる．それらの中から

専門が同一であるもの同士のコサイン類似度の値を加算し，コサイン類似度の合計が最大である専門を求め，それを当該例文の専門性として抽出する．また，このとき抽出された専門の持つコサイン類似度の合計を各例文の専門性の適合度として利用する．

3. CDS と専門性の関係性調査

例文における専門性が CDS を分類する際の指標として妥当かどうかを調査する目的で，提案手法を用いて例文の専門性と適合性を算出し，CDS と専門性の関係の調査を行った．

まず，関係性の調査を行うために実利用環境を想定して大韓民国内で日本語教育経験のある韓国人，日本人へ例文の収集・作成を依頼した．例文は 27 個の CDS に対して 1 人あたり最大 3 文を依頼しており，現在 10 人分，約 700 文が収集済みである．

収集した例文全てに対して提案手法を用いて専門性の抽出を行い，例文の適合度を求めた．この適合度を CDS 毎に平均し，CDS と専門性の関連性に関する分析を行ったものを表 1 に示す．ただし，文章の短さ等が原因で抽出した特徴語の数が 4 以下の例文は分析の対象外とした．

表 1 より，専門性との適合度が上位 20 パーセント以上であるものを抽出したところ，CDS 番号 1, 7, 8, 9, 23 が抽出された．この中で 1, 8, 23 は話題の専門性に関する CDS である．そのため，本手法で例文の専門性を提示できていると考えられる．また，本結果より抽出した CDS の適合度は 1.22 以上となった．そこで，例文の専門性に関する分類を行うための閾値を仮に 1.22 と定める．一方，CDS 番号 14 は専門性に関する CDS であるが，適合度の平均が 0.73 であり，他の CDS と比べても低い．この CDS は手紙やメールなどの通信文を対象とするものであるため，特徴語を抽出する際に手紙や通信文特有の表現を抽出しているためであると考えられる．そのため同様に手紙や通信文を対象とする CDS である 15, 16, 17, 18 の結果も同様に CDS 全体の平均である 0.99 よりも低い結果となっている．

適合度の平均値が 1.22 以上であったものの，話題の専門性に関する CDS ではないものとしては 7 番と 9 番が挙げられる．これらの CDS はレベルがそれぞれ B2, B1 であり，今回取り上げた CDS の中では比較的難易度の高いものである．その他の CDS に関しても適合度を CDS レベルごとに平均し，図 2 にグラフ化した．ただし同グラフでは，特徴語の抽出がうまく行えていないと考えられる手紙，通信文を対象とする CDS は対象から除いている．この結果より，難易度と話題の専門性には相関があると想定される．これは難易度の高い例文では専門的な話題が多く出現することを示唆していると考えられる．例文は複数の CDS に該当する可能性もあるため，今回抽出された CDS 番号 7, 9 といった適合度が高めの値となる CDS 項目は，専門性を持つ CDS として挙げられた 1, 8, 14, 23 の要素を持つものが多く含まれていると考えられる．

4. まとめ

CEFR の C1, C2 レベルを除く 27 個の項目に該当する日本語例文の収集を行い，特徴語に対して word2vec を用いた専門性推定手法を適用することにより，例文の専門性の抽出と CDS と専門性の関係の考察を行った．

表 1　CDS と専門性の適合度

CDS番号	レベル	適合度	CDS番号	レベル	適合度
1	B2	1.25	15	B1	0.73
2	B1	1.19	16	A2	0.74
3	B1	0.90	17	A2	0.68
4	A2	0.66	18	A1	0.43
5	A2	0.96	19	B1	1.05
6	A1	0.69	20	A2	1.21
7	B2	1.25	21	A1	1.16
8	B2	1.72	22	B2	1.10
9	B1	1.26	23	B1	1.22
10	B1	1.14	24	A2	1.17
11	B1	0.91	25	A2	0.90
12	A2	0.96	26	A2	1.07
13	A1	0.89	27	A1	0.81
14	B2	0.73		平均	0.99

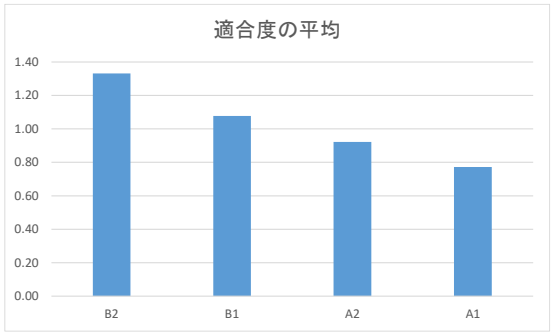

図 2 レベルと適合度の関係(対象が通信文のものを除く)

分析結果より，提案手法を用いて例文における専門性との適合度を求め，仮に設定した閾値と比較することで，手紙，通信文以外の文章に関して話題が専門性を持つかどうかの判定を行うことの妥当性が一部示唆された．

今回は手紙や通信文に対して専門性に関する特徴語をうまく抽出することができなかったため，今後は手紙や通信文等の特有の表現を排除した上で特徴語の抽出を行い，例文の専門性抽出精度の向上を目指す．また，本稿では専門性に関する分類についてのみを検討したが，今後は文の長さや文章構造など他の特徴にも注目し，例文を CDS の各項目へ分類することを目指す．

参考文献

[1] Council of Europe : “Common European Framework of Reference for Languages: Learning , Teaching , Assessment,” Cambridge University Press, 2001.
[2] 投野 由紀夫，石井 康毅 : “英語 CEFR レベルを規定する基準特性としての文法項目の抽出とその評価,” 言語処理学会 第 21 回年次大会発表論文集，pp. 884-887, 2015.
[3] 内田 諭 : “基本動詞のコロケーション難易度測定―CEFR レベルに基づくテキストコーパスからのアプローチ―,” 言語処理学会 第 21 回年次大会発表論文集，pp. 880-883, 2015.
[4] 工藤 拓 : “Yet Another Part-of-Speech and Morphological Analyzer, ” http://taku910.github.io/mecab/, 2016 年 6 月 25 日参照.
[5] Tomas Mikolov, Ilya sutskever, Kai Chen, Greg S Corrado, Jeff Dean : “Distributed Representations of Words and Phrases and their Compositionality,” Neural Information Processing Systems (NIPS), pp. 3111-3119, 2013.