様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２５　日現在

機関番号：　８２６２６
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１６０５９

研究課題名（和文）多様な読み手のための単語難易度指標指標の統計的構築手法の開発

研究課題名（英文）Developing Statistical Methods for Measures of Word Difficulty for Diverse Readers

研究代表者

江原　遥（Ehara, Yo）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員

研究者番号：６０７３８０２９

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：本研究では，第二言語学習者支援のための単語難易度指標の統計的構築方法を開発した．従来の単語難易度指標は，作成に語学教師の多大な労力を要するのにもかかわらず，専門分野の考慮や，言語間での単語難易度の比較ができない問題があった．本研究では，複数の言語資源から統計的に，ある単語がどのような特徴を持つ読み手にどの程度の確率で知られているかを計算できる手法を構築した．この手法は単語親密度や単語心像性などの予測にも役立つ．また単語難易度手法構築のためのデータセットの開発や，第二言語学習者が翻訳者として働く場合の翻訳能力を単語テストのみから求める手法の開発も行った．

研究成果の概要（英文）：In this study, we developed a statistical method for the construction of word-difficulty measures. Existing measures require intensive effort to build and do not account for the various specialties of language learners. Nor are they comparable between languages. The method we constructed can estimate the probability that a second-language learner knows a word by taking their specializations into account, using multiple language resources. This method could help estimate the psychological properties of words, such as word familiarity and word imagery. We also constructed datasets for measures of word difficulty, as well as a method for estimating the translation ability of language learners by using only vocabulary tests.

研究分野：　自然言語処理，知能情報学

キーワード：　語学学習支援　語彙学習　項目反応理論　機械学習　単語難易度　能力測定　テスト理論　e-learning

１．研究開始当初の背景

本研究では，多様な読み手のための単語難易度指標の統計的構築方法を開発した．従来の単語難易度指標は，作成に語学教師の多大な労力を要するのにもかかわらず，専門分野の考慮や，言語間での単語難易度の比較ができない問題があった．

こうした単語難易度指標は，文章の読みやすさ（リーダビリティ推定）や語彙平易化などの言語教育のタスクにおいて，有効な素性となる重要な言語資源であることが報告されていた．

しかし，これらの指標の多くは，単語ごとに被験者実験や語学教師によるアノテーションを要するため，作成コストが高い．そのため，指標が網羅する単語数を増やしたり，多言語で大規模な指標を作成することが難しい．これらの指標を他の素性から高精度に予測する事が可能になれば，リーダビリティ推定や語彙平易化などのタスクの精度向上に有用であると期待される．また，予測結果を人手で確認して半自動でアノテーションを行うなどすることにより，作成コストを低下させる事も可能と期待される．

こうした目的のため，単語難易度指標を他の素性（特徴量）から予測する研究が行われてはきた．この「他の素性」として，従来主に用いられてきたのは，次の２種類の素性である．１つは，均衡コーパスのような，様々な分野のテキストを含み言語使用の全体像を捉えたコーパスを用意し，そこでの単語頻度を基にした素性であり，古くから用いられている．もう１つは，WordNet のような，語の意味について人手の深いアノテーションを施した言語資源を用いた素性であり，直近の研究で使用されている．しかし，どちらの素性の元となる言語資源も，容易に作成できるものではない．このため，英語のような言語資源の豊富な言語で提案された予測手法を，そのまま他の言語に転用する事は難しいという問題があった．また，後者のような人手のアノテーションを要する言語資源を素性に用いてしまうと，英語のような主要言語であっても，素性の元となる言語資源に収録された単語以外の語について，高精度の予測を行うことは難しいという問題もあった．

このように，既存の単語難易度指標は，作成に大きなコストがかかる問題があった．作成コストが大きいことが，学習者の多様性を考慮することを難しくしていた．

２．研究の目的

本研究では，複数の言語資源を組み合わせて統計的に，すなわち半自動的に，単語難易度指標を構築することで，単語難易度指標の作成コストを削減しつつ，学習者の多様性を考慮することが可能な手法を開発することを目的とした．具体的には，ある単語がどのような特徴を持つ読み手にどの程度の確率で知られているかを計算できる手法を構築した．この手法は単語親密度や単語心像性などの予測にも役立つ．また単語難易度手法構築のためのデータセットの開発や，第二言語学習者が翻訳者として働く場合の能力を求める手法の開発も行った．**これら一連の研究の詳細については，本稿の最後に載せた「その他」欄に記載した研究代表者のホームページから最新情報にアクセスできるようにしているので，可能な限りこちらを合わせて参照することを推奨する**．

ここでは，これら複数の提案手法の１つとして，具体的に，生コーパスから直接，単語難易度関連指標を予測する手法について述べる．この提案手法の主要なアイデアは，様々な分野のテキストを集めたコーパスを作りそのコーパスの単語頻度を素性に使う代わりに，分野ごとの単語頻度相当の素性を作り，どの分野を重視すればいいかを，目的変数である指標に合わせて自動的に決定するということである．例えば，単語難易度関連指標のうち，単語親密度については，話し言葉の分量が書き言葉に比べて多いコーパスとの相関が強い事が分かっている．そこで，話し言葉が多いコーパスを一旦作成してそこでのコーパスの単語頻度を素性に使う代わりに，生コーパスから，話し言葉を多用する分野での単語頻度に相当する素性と，書き言葉を多用する分野での単語頻度に相当する素性をそれぞれ別に求め，単語親密度との相関が高くなるような両者の重み付けを自動的に決定すれば，指標の予測精度が改善すると考えられる．ただし，生コーパスには，通常，どのテキストがどの分野であるかといったアノテーションは与えられていない．さらに，そのようなアノテーションが着いていたとしても，分野ごとに分かれた少量のテキストからの単語頻度をそのまま使おうとするとデータ欠乏の問題が発生する．

３．研究の方法

提案手法では，この両者の問題を解決するため，Latent Dirichlet Allocation などのトピックモデルを利用する．トピックモデルは，生コーパスを，人手の教師情報なしにトピック（分野）に分解する手法である．提案手法では，各分野のテキストからの単語頻度を，各トピックからの単語の出現確率で代用することで対応する．本稿では，スペースの都合により，単語難易度関連指標のうち，特に，単語親密度についてのみ報告する．

単語親密度などの言語心理学的な指標のタグ付けデータとして，英語では，MRC Psycholinguistic Database（以後，MRC）を用いた．MRC には，単語親密度の他，具象性(Concreteness)，心象性(Imagery)，獲得年齢(Age of Acquisition)といった指標が収録されている．

LDA の実装には**gensim** ツールキットを用いた．英語においては，次の３種のコーパスを用意した．

1. Wikipedia:非均衡コーパス．約 29 億語．Wikipedia 英語版（2016 年 8 月 13 日時

点）の全体に LDA を適用した．
2. BNC:均衡コーパス．約 1 億語．British National Corpus の全体に LDA を適用した．
3. Brown:均衡コーパス．約 100 万語．Brown corpus の全体に LDA を適用した．

実験の対象とする語の集合については，次のように選んだ．まず，英語でも日本語でも，Wikipedia 上の頻度上位 100,000 語を取り出し，実験対象候補の語集合とする．次に，この候補の語集合の中で，各単語難易度関連指標と文字列が完全に一致するものを取り出し，各指標の実験対象語の集合とした．結果として，単語親密度 4,566 語が実験対象語として抽出され，既存研究[1]に従い，このうちの半分を訓練データ，半分をテストデータとした．

回帰手法としては，下記の手法を比較した．

1) Ridge 回帰(Ridge)．線形回帰にパラメタが極端な値を取りすぎないように極端なパラメタ値に罰則（正則化）をつけたもの．
2) Support Vector Regression (SVR) に線形カーネルを用いたもの．
3) SVR に Radial Basis Function (RBF)カーネルを用いたもの．

下記の素性セットを比較した．

FREQ(コーパス名)： ()内のコーパス中の単語頻度

LDA(コーパス名) ：()内のコーパスに LDA をかけて算出した各トピックの単語出現確率

Word2Vec：分散表現による素性．前述の Wikipedia コーパスに word2vec ツールを適用して分散表現を作成した．

４．研究成果

評価尺度には，先行研究と同様，目的の指標と予測値とのピアソンの相関係数(r)とスピアマンの順位相関係数ρを用いた．

	素性	ρ	r
-	FREQ(Wiki)	0.6208	0.5894
SVR-Linear	LDA(Wiki)	0.6879	0.6207
	w2v	0.6533	0.6096
	LDA(Wiki)+w2v	0.7470	0.6890
SVR-RBF	LDA(Wiki)	0.7566	0.7329
	w2v	0.7677	**0.7362**
	LDA(Wiki)+w2v	0.7801	0.7514
Ridge	LDA(Wiki)	0.7109	0.6509
	w2v	0.6869	0.6370
	LDA(Wiki)+w2v	**0.7821**	0.7324

表 1: 単語親密度予測結果（Wiki）

[1] Gustavo Paetzold and Lucia Specia. Inferring psycholinguistic properties of words. In Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 435–440.

結果を表１に示す．まず，Wikipedia は均衡コーパスではないが，多くの言語で大きい語数のコーパスを入手できる．Wikipedia のデータだけを用いて，どれだけ相関係数を向上させられるかが，本研究の主眼となる．表１より，LDA(Wiki)素性を与える事によって，元々のコーパス頻度である FREQ よりρ，r ともに大幅に向上する事が分かる．

w2v 素性を加えると，SVR-RBF の場合のみ，性能が LDA(Wiki)を与えた場合よりも向上する．SVR-RBF のみが，非線形である RBF カーネルを 用い，カーネルトリックによって組み合わせ素性を追加した場合に相当する効果が見込まれる．w2v 素性は，次元ごとに意味を見出すことは難しく，組み合わせ素性まで考慮しないと性能が向上しない事が，この実験によって示されている．一方，LDA(Wiki)素性では，各次元が，各トピックからのその単語の単語出現確率とみなせるので，線形の Ridge や SVR-Linear でも大幅な性能向上が見込めると考えられる．

最後の，どの手法を用いた場合でも，２種の素性を同時に用いた場合が最も性能が向上している．

以上の内容は，学会発表 3.の内容の概略であり，雑誌論文 2.にさらに詳細を記載してある．

この他，雑誌論文 1.の研究では，学習者の特性を考慮した語彙知識の予測モデルの研究を行った．

雑誌論文 3.の研究では，語彙テストを利用して，第二言語学習者が翻訳者として翻訳作業を請け負うクラウドソーシングという仕組み上の翻訳者の能力を計測し，能力の高い翻訳者に優先的に発注することによって翻訳コストを下げる手法を開発した．この研究は，人工知能分野のトップ査読付き国際会議である IJCAI 2016 に採録された(雑誌論文 3)．また，国内では，NLP 若手の会第 10 回シンポジウム奨励賞（学会発表 6）や，情報処理学会 DICOMO 2015 シンポジウム優秀論文賞（学会発表 8）を受賞しており，高く評価されている．

さらに，語彙テストデータのデータセットを，クラウドソーシングを用いて作成し，これについて学会発表を行った（学会発表 1)．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

1. Yo Ehara, Issei Sato, Hidekazu Oiwa, Hiroshi Nakagawa. Mining words in the minds of second language learners: learner-specific word difficulty. Journal of Information Processing. Vol. 26. March 2018.

2. Yo Ehara. Language-Independent Prediction of Psycholinguistic Properties of Words. In the Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017, short paper). Taipei, Taiwan. December 2017.
3. Yo Ehara, Yukino Baba, Masao Utiyama, Eiichiro Sumita. Assessing Translation Ability through Vocabulary Ability Assessment. In the Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16). New York, USA. July 2016.

〔学会発表〕（計 9 件）
1. クラウドソーシングを用いた語彙テスト結果データセット作成. 江原遥（産総研). 2018 年 3 月.
2. 単語親密度の分析・予測モデルの提案. 江原遥（産総研). NLP 若手の会第 12 回シンポジウム. 2017 年 9 月.
3. 生コーパスからの単語難易度関連指標の予測. 江原遥（産総研).言語処理学会第 23 回年次大会. 2017 年 3 月.
4. 単語分散表現と単語難易度. 江原遥（産総研). NLP 若手の会第 11 回シンポジウム. 2016 年 9 月.
5. 語彙知識予測問題の概説～ロジスティック回帰と項目反応理論の対応関係を中心に～. 江原遥, 石川博. 言語処理学会第 22 回年次大会. 2016 年 3 月.
6. 語彙能力と翻訳能力の関係. 江原遥（NICT）, 馬場雪乃（京大）, 内山将夫（NICT）, 隅田英一郎（NICT). NLP 若手の会第 10 回シンポジウム. 2015 年 9 月.（**奨励賞受賞**）
7. クラウドソーシング翻訳のコスト削減のための翻訳者割り当て手法の提案. 江原　遥（NICT)・馬場雪乃（京大)・内山将夫・隅田英一郎（NICT）.第 14 回情報科学技術フォーラム. 愛媛. 2015 年 9 月.
8. クラウドソーシング翻訳のコストを削減する翻訳品質の事前予測モデル. 江原遥（情報通信研究機構), 馬場 雪乃(京都大学), 内山 将夫, 隅田 英一郎(情報通信研究機構). 情報処理学会 DICOMO 2015 シンポジウム. 安比高原. 2015 年 7 月. （**優秀論文賞受賞**）
9. 単語テストを利用した翻訳品質事前予測モデル. 江原遥, 馬場雪乃, 内山将夫, 隅田英一郎. 人工知能学会第 29 回年次大会. 北海道. 2015 年 6 月

〔図書〕（計 0　件）

〔産業財産権〕
○出願状況（計 0　件）
○取得状況（計　0　件）

〔その他〕
ホームページ等
http://yoehara.com/

６．研究組織
(1)研究代表者
江原　遥　（EHARA, Yo）
産業技術総合研究所・情報・人間工学領域・研究員
研究者番号：60738029