4A2-1

単語難易度関連指標の多言語での予測

江原　遥 *1
Yo Ehara

*1産業技術総合研究所　人工知能研究センター
National Institute of Advanced Industrial Science and Technology, Artificial Intelligence Research Center

1. はじめに

言語教育分野おける、単語の難しさを表す指標を総称して、本稿では「単語難易度関連指標」と呼ぶことにする。「難易度」という語を名に含む指標もあるため、混乱を避けるための措置である。単語難易度関連指標には様々な種類があるが、コーパスからの単語頻度のように簡単に計算可能なものもある一方で、母語話者を対象に大規模な単語に対する親密度のアンケートを取って計測する必要がある単語親密度など、作成にコストを要するものも多い。コストを要する単語難易度関連指標の中には、教育上有用であるばかりでなく、自然言語処理タスクの特徴量として用いることにより性能を向上させることが近年わかってきている指標もある [Paetzold 16]。しかし、作成にコストを要する単語難易度関連指標は、入手可能な言語が限られていたり、収録されている語彙数が少ないなどの問題がある。本稿では、この作成にコストを要する単語難易度関連指標を、どの言語でも比較的入手しやすい言語資源を用いて予測することによって、この問題の解決を目指す。

単語難易度関連指標のうち、単語親密度など心理言語学的な指標については、過去に予測の研究が行われている。単語難易度関連指標では、単語に、その単語の難しさが実数値で付与されていると考えることができ、機械学習の観点からは、単語に関する特徴量からこの実数値を予測する回帰問題に属する。この特徴量として、過去には、大別して 2 種類の特徴量が用いられてきた。1 つは、複数の種類の異なるコーパスからの単語頻度を用いた特徴量 [Tanaka-Ishii 11, Paetzold 16] である。もうひとつは、語の意味に関連する特徴量であり、この双方を用いることによって高い予測性能が達成できることが報告されている [Paetzold 16]。本研究の目的である入手しやすさを考えると、後者の語の意味に関する特徴量のうち、WordNet のような人手の深いアノテーションを介した特徴量は入手にくく、また、前者のように複数の種類の異なるコーパスを用意することも難しい。

そこで、有効性が示されている二種の特徴量を、他の特徴量で代替することを考える。後者の意味を捉える特徴量については、近年、word2vec[Mikolov 13] などの単語分散表現を用いて容易に代替が可能であり、[Paetzold 16] でも用いられている。一方、前者の複数のコーパスについては、考慮が必要である。なぜ複数のコーパスを用いることが有効であるのかは、[Tanaka-Ishii 11, Paetzold 16] では明確に述べられていないものの、次の例から直感的に理解することができる。例えば、“mother” と “family” は、意味的には異なる意味を持つ単語であるが、どちらも、高い単語親密度を示す。それは、これらの語が同じトピック（分野）を共有するからであると考えられる。このように、意味的に似ているかどうかは、必ずしも単語難易度指標の予測に有効とは限らず、語の属するトピックを認識することが予測のためには重要となる。コーパスによって、含まれるトピックの分布には偏りがあるため、複数のコーパスを用いることで、初めて、語のトピックを認識し、予測性能が向上する、という仮説が立つ。

本研究では、この仮説に基づき、特にアノテーションもされていないコーパスから、語のトピックを捉える特徴量と、語の意味を捉える特徴量の 2 つを抽出し、回帰問題として指標を予測する。語の属するトピックを捉える特徴量としては、Latent Dirichlet Allocation (LDA) [Blei 03] 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替する。さらに、回帰予測手法に、各特徴量に対して重みパラメタが付き、解釈が容易な手法を用いれば、どのトピックからの単語の出現確率が予測に貢献しているかを数値的に示すことが可能となる。この分析は、実際に予測に最も有効なトピックを列挙することで予測モデルの妥当性を質的に検証するなど、様々な応用が考えられる。

本研究の貢献は，以下のとおりである．

- どの言語でも比較的容易に利用できる特徴量を用いて単語難易度指標の予測を行い、比較的高精度な予測精度を達成した．
- 近年意味を捉えたベクトルとして注目され既存研究でも利用されている word2vec [Mikolov 13] のような単語のベクトル表現より，LDA の各トピックからの単語出現確率の方が，予測精度の向上に有効な素性であった．
- 重みを分析することにより、どのようなトピックが予測に貢献しているかを数値的に示すことが可能となり、様々な応用が考えられる。

2. 関連研究

本研究では，各単語について，語の難しさの一面を表していると考えられる数値が紐付いた形式のデータを，まとめて「単語難易度関連指標」と呼ぶことにする．単語難易度関連指標は，様々な分野で作成されているが，代表的なものは大別して下記の 3 種が考えられる．そして，この前者 2 つについては，

連絡先: 江原　遥，産業技術総合研究所　臨海副都心センター別館，東京都江東区青海 2-4-7，03-3599-8609，y-ehara@aist.go.jp

コーパス中の単語頻度との相関を示した研究が存在する．最後のものについては，そもそも単語頻度表を元に人手を加えて作成している．

本稿で対象とする単語難易度関連指標は、大別して以下のようにわけられる。

心理言語学的指標 母語話者の単語に対する親密度 (familiarity) や、単語の具象性 (conreteness)、単語が意味するものを想起できるか、等の点について、母語話者へのアンケート調査を通じて求める。

教育用語彙リスト コーパスからの単語頻度の高い順に並べた語のリストを語学教育の専門家が人手で確認し、より語学教育に適切な形で修正を施したもの。

テスト理論からの指標 語学学習者に大規模な単語テストを行い、そのテストのデータから項目反応理論などを用いて統計的に求めた指標。

3. 提案手法

本稿では，簡単のため，特に断らない場合，単語頻度の対数を取った値や確率値に対数を取った値を，単に単語頻度と呼ぶ．提案手法では，単語難易度指標の予測のための素性に，コーパスからの単語頻度を用いる代わりに，まず，コーパスに LDA を適用し，トピックからの単語の出現確率の対数値を素性に使う．

以下，なぜそのようにすると性能が向上すると思われるのかについて，線形回帰を例に説明する．

今，あるコーパス C があるとする．C 中の w を単語とし，$f(w)$ を C 中のその単語の頻度，N を C の延べ単語数，$p(w) = \frac{f(w)}{N}$ を w の出現確率とする．また，m をある単語難易度指標として，$m(w)$ を w に対する単語難易度指標の値だとする．

指標 m と単語頻度 f が相関しているということは，次の単回帰式で $m(w)$ がうまく予測可能であるという事である．

$$m(w) = \beta_0 + \beta_1 \log(f(w)) + \epsilon \quad (1)$$

式 1 において，ϵ は正規分布に従う誤差とする．式 1 のように単語頻度と指標 m が相関するのであれば，単語の出現確率である $p(w)$ も，定数 $-\log(N)$ の分切片が平行移動しただけなので，指標 m と相関する．

$$\begin{aligned} m(w) &= \beta_0 + \beta_1 \log(p(w)) + \epsilon & (2)\\ &= \beta_0 + \beta_1 \log\left(\frac{f(w)}{N}\right) + \epsilon & (3)\\ &= (\beta_0 - \log(N)) + \beta_1 \log(f(w)) + \epsilon & (4) \end{aligned}$$

ここで，LDA を C にかけると，$p(w)$ を近似し，各トピック（分野）t からの単語の出現確率 $p(w|t)$ と，トピックの出現確率 $p(t)$ を用いて，次の式で合わすことが出来る．ここで，K をトピック数とする．

$$m(w) \approx \beta_0 + \beta_1 \log\left(\sum_{k=1}^{K} p(w|t_k)p(t_k)\right) + \epsilon \quad (5)$$

今，ここで，トピック（分野）の出現確率 $p(t)$ の意味合いを考えると，これは，コーパス C 中に，あるトピック（分野）t がどれ位の割合で出現しているのかを表していると考えられる．一方，指標 m では，C とは違うトピックが重視されているかもしれない．そこで，$p(t)$ をパラメタ λ_k で置き換え，この部分も指標 m に合わせて推定することにする．ただし，$\lambda_k \geq 0; \forall k \in \{1, \ldots, K\}, \sum_{k=1}^{K} \lambda_k = 1$ とする．

$$m(w) = \beta_0 + \beta_1 \log\left(\sum_{k=1}^{K} \lambda_k p(w|t_k)\right) + \epsilon \quad (6)$$

式 6 では，$\log p(w|t_k)$ を β_1 と λ_k の 2 つのパラメタで重み付けしている点が肝要である．そこで，式 6 とはモデルが異なっているものの，この点は共通している，次の単純化した問題を解くことにする．

$$m(w) = \beta_0 + \sum_{k=1}^{K} \lambda_k \log(p(w|t_k)) + \epsilon \quad (7)$$

式 7 は，$\log(p(w|t_k))$ を素性とする単純な線形回帰である．ただし，線形回帰は外れ値を含む場合，推定する β_0 や λ_k といった回帰係数が極端な値になる場合がある．これを防ぐため，正則化によって極端な回帰係数に罰則を施した Ridge 回帰を実験では用いる．

4. 評価実験

4.1 データセット

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[Coltheart 81] （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 Aquisition) といった指標が収録されている．

LDA の実装には **gensim** 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．

Wiki 非均衡コーパス．約 29 億語．Wikipedia 英語版 [*1] の全体に LDA を適用した．

BNC 均衡コーパス．約 1 億語．British National Corpus[The BNC Consortium 07] の全体に LDA を適用した．

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．

4.2 比較手法

回帰手法としては，下記の手法を比較した．

Ridge Ridge 回帰 [Tikhonov 63]．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．

SVR-Linear Support Vector Regression (SVR) [Smola 97] に線形カーネルを用いたもの．

*1 2016 年 8 月 13 日時点での enwiki-latest-pages-articles.xml.bz2

SVR-RBF SVR に Radial Basis Function (RBF) カーネルを用いたもの.

GPR-RBF Gaussian Process Regression [Williams 06] に 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，各トピックの単語出現確率

w2v(コーパス名) Word2Vec[*2] 素性．word2vec は，() 内のコーパスより作成した.

4.3 評価尺度

評価尺度には，[Paetzold 16] 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．前者は予測器が目的とする指標の数値をどれだけ正確に予測出来ているかを表し，後者は，予測器が，目的とする指標のテストセット中での順位をどれだけ正確に予測できているかを表す．例えば，テストセットが 3 単語のみから構成されており，目的とする指標の正解値がそれぞれ [1.1, 3.3, 2.2] である時に，予測器が [1.9, 3.8, 2.7] と予測した場合，ピアソンの相関係数は 0.9958 であるが，順位は正しく並べられているため，スピアマンの順位相関係数は 1.0 である.

この 2 つの評価尺度のうち，どちらの方が有用であるかは予測値の使い方による．言語教育における基本語は有限であるため，どの単語より難しいか/簡単か，だけがわかれば良い場合も多いと思われる．一方，[Paetzold 16] では，語彙簡単化タスクの素性に使うことを考慮した場合，ピアソンの相関係数のほうがスピアマンの順位相関係数と比べて重要であるとされている.

4.4 実験結果

まず，Wikipedia は，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．実験結果を表 1 に示す．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(Wiki) 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，各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.

次に，表 1 に，BNC を用いて単語親密度を予測した場合の実験結果を示す．BNC は均衡コーパスであり，人手で分野が均衡になるよう調整されているにもかかわらず，どの手法を用いた場合においても，LDA(BNC) の方が，単純な単語頻度との相関である FREQ より向上している事である．これは，前節に述べたように，各トピック（分野）をどの程度重視するか，という値を，目的の指標に合わせて推定した効果が出ているからであると考えられる．一方，BNC は Wikipedia に比べると，小さいコーパスであるにもかかわらず，Wikipedia を利用した場合と比較して，全体的に予測性能が向上している．これは，単純に語数の大きなコーパスを用意するよりも，人間が分野を調整して作成した均衡コーパスの方が，単語親密度のような単語難易度関連指標の予測に適している事を示している.

次に、表 2 に、単語親密度を LDA(Wiki) 設定で予測する際に、有効であったトピック上位 3 つを上げる。家族や小説といったドメインが捕捉できている事、そして、このようなドメインが単語親密度の予測に貢献していることがわかる。

4.5 日本語での予測結果

次に、日本語における結果を表 3 に載せる．日本語のコーパスとしては，日本語版 Wikipedia と均衡コーパスである BCCWJ[Maekawa 07] に対して，MeCab[Kudoh] を用いて形態素解析したものを用いた．日本語単語親密度 [Amano 98] を求める実験設定とし，手法は英語において全般的に高い性能を達成した GPR-RBF に限定した．word2vec は日本語版 Wikipedia に対して適用したものを用いた．その他の設定は英語と同じである.

英語の場合と同様，LDA 特徴量の方が word2vec 特徴量を用いた場合よりも良い性能を示すこと，両者を組み合わせることでより高い性能が発揮できることが示された．日本語の場合は，興味深いことに，Wikipedia の方が高い性能を達成した．この理由については，今後の課題としたい.

5. おわりに・考察

本稿では，単語難易度関連指標を、単一の大きなコーパスのみを用いて予測する手法を提案した．提案手法では，LDA の単語出現確率を特徴量に用いる事によって，分野ごとの重みを再調整し，単純な単語頻度よりも性能が向上させられる事，また，この場合の再調整にも限界があり，小規模でも人手で分野が調整されている均衡コーパスを用意して同じ手法を適用したほうが性能が高いことを示した．紙面の都合上，本稿では，数ある単語難易度指標のうち，心理言語学的指標についてのみの報告にとどめたが，例えば，JACET 等の英語の教育用語彙表や，日本語学習辞書支援グループ (2015)「日本語教育語彙表 Ver 1.0」[*3] でも同様の結果が出ている．今後の課題としては，予測精度の向上と，具体的にどのような分野が重く重み付けられているかについて，定性的な分析を行うことが挙げられる.

謝辞

本研究は，JSPS 科研費 15K16059 の助成を受けた.

参考文献

[Amano 98] Amano, S. and Kondo, T.: Estimation of mental lexicon size with word familiarity database, in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)* (1998)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022 (2003)

[Coltheart 81] Coltheart, M.: The MRC psycholinguistic database, *The Quarterly Journal of Experimental Psychology*, Vol. 33, No. 4, pp. 497–505 (1981)

*2 word2vec

*3 http://jisho.jpn.org/

Method	Feature	親密度		獲得年齢		具象性		心像性	
		ρ	r	ρ	r	ρ	r	ρ	r
FREQ(Wiki)		0.620	0.589	0.327	0.360	0.018	0.021	0.097	0.116
SVR-Linear	LDA(Wiki)	0.695	0.654	0.695	0.709	0.689	0.681	0.639	0.613
SVR-Linear	w2v(Wiki)	0.016	0.013	0.064	0.055	-0.003	-0.002	0.031	0.034
SVR-Linear	All	0.759	0.731	0.695	0.710	0.721	0.720	0.695	0.678
SVR-RBF	LDA(Wiki)	0.760	0.735	0.710	0.732	0.744	0.741	0.721	0.710
SVR-RBF	w2v(Wiki)	0.010	0.013	0.058	0.047	-0.006	-0.007	0.028	0.035
SVR-RBF	All	0.762	0.746	0.713	0.735	0.747	0.744	**0.727**	**0.716**
Ridge	LDA(Wiki)	0.738	0.695	0.700	0.716	0.716	0.709	0.685	0.665
Ridge	w2v(Wiki)	0.008	0.005	0.019	0.011	0.000	0.007	0.018	0.025
Ridge	All	0.745	0.705	0.680	0.695	0.645	0.635	0.620	0.600
GPR-RBF	LDA(Wiki)	0.758	0.725	0.712	0.730	0.747	0.745	0.707	0.691
GPR-RBF	w2v(Wiki)	0.009	0.006	0.020	0.012	0.000	0.007	0.019	0.025
GPR-RBF	All	**0.779**	**0.750**	**0.716**	**0.735**	**0.754**	**0.752**	0.723	0.708
FREQ(BNC)		0.756	0.724	0.405	0.450	0.107	0.118	0.001	0.022
SVR-Linear	LDA(BNC)	0.804	0.785	0.814	0.823	0.678	0.673	0.667	0.657
SVR-Linear	w2v(BNC)	-0.003	-0.014	0.045	0.041	0.015	0.013	0.033	0.038
SVR-Linear	All	0.823	0.803	**0.814**	**0.823**	0.687	0.683	0.684	0.675
SVR-RBF	LDA(BNC)	0.836	0.816	0.813	0.820	0.722	0.720	0.724	0.714
SVR-RBF	w2v(BNC)	-0.014	-0.017	0.044	0.043	0.011	0.009	0.031	0.033
SVR-RBF	All	0.836	**0.816**	0.813	0.820	0.722	0.720	**0.724**	**0.715**
Ridge	LDA(BNC)	0.832	0.806	0.812	0.818	0.703	0.697	0.694	0.684
Ridge	w2v(BNC)	0.017	0.014	0.039	0.036	0.009	0.011	0.023	0.024
Ridge	All	0.817	0.787	0.773	0.782	0.611	0.603	0.622	0.609
GPR-RBF	LDA(BNC)	0.837	0.813	0.810	0.817	0.725	0.720	0.717	0.707
GPR-RBF	w2v(BNC)	0.017	0.014	0.039	0.036	0.009	0.011	0.023	0.024
GPR-RBF	All	**0.837**	0.814	0.810	0.817	**0.726**	**0.720**	0.717	0.708
[Paetzold 16]		0.863	0.846	0.871	0.862	0.876	0.869	0.835	0.823

表 1: 英語における予測結果.

[Kudoh] Kudoh, T.: MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.jp/

[Maekawa 07] Maekawa, K.: Kotonoha and BCCWJ: development of a balanced corpus of contemporary written Japanese, in *Corpora and Language Research: Proceedings of the First International Conference on Korean Language, Literature, and Culture*, pp. 158–177 (2007)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[Paetzold 16] Paetzold, G. and Specia, L.: Inferring Psycholinguistic Properties of Words, in *Proc. of NAACL-HLT*, pp. 435–440, San Diego, California (2016)

[Smola 97] Smola, A. and Vapnik, V.: Support vector regression machines, Vol. 9, pp. 155–161 (1997)

[Tanaka-Ishii 11] Tanaka-Ishii, K. and Terada, H.: Word familiarity and frequency, *Studia Linguistica*, Vol. 65, No. 1, pp. 96–116 (2011)

[The BNC Consortium 07] The BNC Consortium, : The British National Corpus, version 3 (BNC XML Edition) (2007), Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: `http://www.natcorp.ox.ac.uk/` (Retrieved on October 26, 2012)

[Tikhonov 63] Tikhonov, A.: Solution of incorrectly formulated problems and the regularization method, *Soviet Math. Dokl.*, Vol. 5, pp. 1035–1038 (1963)

[Williams 06] Williams, C. K. and Rasmussen, C. E.: Gaussian processes for machine learning, *the MIT Press*, Vol. 2, No. 3, p. 4 (2006)

Weight	Top 15 words	Possible Interpretation of topic
0.327	mc, ron, roland, barker, ronald, tr, sue, hubbard, archived, panorama, mins, kaufman, gladys, partridge, domino	Novels
0.324	mother, father, plot, get, love, tells, friend, girl, man, friends, story, wife, find, young, help	Family
0.222	player, politician, writer, singer, actor, footballer, football, actress, poet, refer, surname, events, author, journalist, artist	Player

表 2: 単語親密度を LDA(Wiki) 設定で予測する際に、有効であったトピック上位３つ.

Method	Feature	親密度	
		ρ	r
GPR-RBF	LDA(Wiki)	0.4147	0.3498
GPR-RBF	w2v	0.1344	0.1191
GPR-RBF	All	**0.4205**	**0.3516**
GPR-RBF	LDA(BCCWJ)	0.3725	0.2906
GPR-RBF	w2v	0.1344	0.1191
GPR-RBF	All	**0.3821**	**0.3036**

表 3: 日本語における予測結果.