生コーパスからの単語難易度関連指標の予測

江原 遥
産業技術総合研究所 人工知能研究センター
y-ehara@aist.go.jp

1 はじめに

単語親密度などの単語難易度に関連する指標は，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る [6]．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，単語難易度関連指標を他の素性から予測する研究が行われてきた．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 [8, 5]．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 [6]．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．

本研究では，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を提案する．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，各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 [8]．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，各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．

提案手法では，この両者の問題を解決するため，Latent Dirichlet Allocation (LDA) [2] 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．

本研究の貢献は，以下のとおりである．

- 均衡コーパスや深いアノテーションを施した言語資源に頼らないことで，他言語への適用が容易な単語難易度関連指標の予測手法を提案する．
- 単語難易度関連指標の予測精度は，LDA の各トピックからの単語出現確率を基にした素性を用いることで，従来のようにコーパスからの単語頻度を素性を用いる場合と比べ，大幅に向上した．
- 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 [4] 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．

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，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．

1. 単語親密度 (Word Familiarity) や単語獲得年齢 (Age of Aquisition) のような心理言語学的な指標 [8].
2. 項目反応理論に語を項目（item）とみなした時の困難度パラメタの値 [1].
3. JACET8000 や SVL12000 のような言語教育のために人手を加えて作成された単語難易度表.

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[3] （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[9] 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．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，[6] 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．

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 [10]．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) [7] に線形カーネルを用いたもの．

SVR-RBF 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，各トピックの単語出現確率

w2v Word2Vec[2] 素性．Word2Vec は，§4.1 の **Wiki** を用いて作成した．

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

[2] word2vec

4.3 評価尺度

評価尺度には，[6] 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，[6] 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

	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）

実験結果を表 1，表 2，表 3 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，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．

	素性	ρ	r
-	FREQ(BNC)	0.7565	0.7254
SVR-Linear	LDA(BNC)	0.7880	0.7720
	w2v	0.6532	0.6096
	LDA(BNC)+w2v	0.7772	0.7664
SVR-RBF	LDA(BNC)	0.8360	0.8162
	w2v	0.7677	0.7363
	LDA(BNC)+w2v	0.8531	0.8318
Ridge	LDA(BNC)	0.8304	0.8057
	w2v	0.6869	0.6370
	LDA(BNC)+w2v	0.8337	0.8150

表 2: 単語親密度予測結果（BNC）

次に，表 2 に，均衡コーパスである 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．特徴的であるのは，BNC は均衡コーパスであり，人手で分野が調整されているにもかかわらず，LDA(BNC) の方が，どの手法を用いた場合においても，単純な単語頻度との相関 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と言えよう．一方，BNC は Wikipedia に比べると，語数がずっと小さなコーパスであるにもかかわらず，表 2 の結果は，表 1 と比較して，全体的に向上している．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が分かる．

最後に，表 3 に，Brown Corpus を素性に用いた場合の実験結果を示す．Brown Corpus は BNC よりさらに小さな均衡コーパスであり，Wikipedia と比較すると，そのサイズは非常に小さい．にも関わらず，表 3 では，LDA(Wiki)+w2v において，表 1 よりも高い予測性能を達成している．これは，やはり，Wikipedia のような語数のみ大きなコーパスを用いるよりも，小規模でも均衡コーパスを用いた方が，単語難易度関連指標を上手く予測出来ることを示す．

	素性	ρ	r
-	FREQ(Brown)	0.7224	0.6776
SVR-Linear	LDA(Brown)	0.6963	0.6332
	w2v	0.6532	0.6096
	LDA(Brown)+w2v	0.7580	0.7020
SVR-RBF	LDA(Brown)	0.7360	0.6906
	w2v	0.7677	0.7362
	LDA(Brown)+w2v	0.8157	0.7835
Ridge	LDA(Brown)	0.7089	0.6435
	w2v	0.6869	0.6370
	LDA(Brown)+w2v	0.7798	0.7238

表 3: 単語親密度予測結果（Brown）

5 おわりに・考察

本稿では，単語難易度関連指標を，均衡コーパスを使わず，生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 の単語出現確率を素性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．紙面の関係上，本稿では，単語難易度関連指標のうち，特に英語の単語親密度についてのみ報告したが，例えば単語の獲得年齢といった心理言語学的な指標や，JACET 8000 のような教育用の単語難易度関連指標，さらに，日本語の単語親密度についても同様の結果が出ている．今後の課題としては，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

[1] David Beglar. A rasch-based validation of the vocabulary size test. *Language Testing*, Vol. 27, No. 1, pp. 101–118, 2010.
[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.
[3] Max Coltheart. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, Vol. 33, No. 4, pp. 497–505, 1981.
[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.
[5] Gustavo Paetzold and Lucia Specia. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proc. of COLING*, pp. 1669–1679, Osaka, Japan, December 2016.
[6] Gustavo Paetzold and Lucia Specia. Inferring psycholinguistic properties of words. In *Proc. of NAACL-HLT*, pp. 435–440, San Diego, California, June 2016.
[7] Alex Smola and Vladimir Vapnik. Support vector regression machines. Vol. 9, pp. 155–161, 1997.
[8] Kumiko Tanaka-Ishii and Hiroshi Terada. Word familiarity and frequency. *Studia Linguistica*, Vol. 65, No. 1, pp. 96–116, 2011.
[9] 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).
[10] Andrey Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, Vol. 5, pp. 1035–1038, 1963.