

文脈化単語表現空間上の範囲の学習による 語の多義性を考慮した頻度計数法

江原 遥^{1,a)}

概要: 語彙テストの統計解析から推定される語の難易度や単語親密度など、多数の被験者に対する反応の計測値を集計して得られる語の指標の中には、単語頻度やその対数値と強く相関するものがある事が知られている。被験者の関わる計測は高コストであるため、入手が容易な単語頻度からこうした指標を精度よく予測できれば有用である。こうした単語頻度の計数においては、語の多義性が課題であった。語の多義性も考慮して単語頻度を補正すればより精度よくこうした指標を予測できると考えられるが、どの程度の粒度で語の多義性を考慮すべきかが明らかでなく、また、語の多義性を人手で付与することも高コストであるという問題があった。本研究では、語の多義性を考慮した頻度計数法を提案する。BERT など、個々の単語の出現ごとに語の埋め込み表現が獲得できる文脈化単語埋め込み表現の空間上で、被験者は限られた範囲の語に反応していると仮定する。提案手法では、被験者から収集したデータから、この空間上の範囲をパラメタとして学習しながら、頻度を計数する。実験の結果、提案手法を通じた単語計数法がより精度よく語の難易度を予測できた。

1. はじめに

自然言語処理で扱われる言語資源には、訓練されたアノテータが多くのデータに対してアノテーションを行うことによって作成するものが多い。訓練を重ねることによって、どのアノテータのアノテーションも一致し、アノテータによる差が出ないことが理想とされる設定では、この方法は有効であろう。また、アノテータによって差が出ないことを想定しているため、アノテータを複数人用意することは「差が出ない」想定を正しさを検証する意味合いが強い。同一のコストで有用なアノテーションを行うためには、できるだけ多くのケースをカバーするため、テキストや文など文脈を多分に含み一回性の高い単位で言語資源を作成することが有効な戦略となる。

一方、そもそも、言語の認知に関する個人差に興味がある設定などでは、多数の被験者^{*1}からデータを収集することが必要な場合があり、こうした設定では有効な戦略が大きく変わってくる。例えば、学習者によって背景知識に差

がある語学学習や心理言語学などでは、こうした言語資源が重要となる [6], [9]。こうした多数の被験者を扱う状況では、被験者一人ひとりに多くのデータに対して作業してもらうことにはコストがかかるため、テキストや文など一回性の高い単位ではなく、できるだけ後の研究でも再利用しやすい単位でデータを取得することが望ましい。これゆえ、基礎的な単位である「語」に対してデータを収集しておき、後の研究でも再利用することが広く行われる。語学学習支援における単語難易度 [5] や、心理言語学における単語親密度 [10] などがこうしたデータの例である。自然言語処理分野では、文や文章構造といった、語より大きな単位に興味に移りつつあるが、個人差が重要である場合には、語を単位とした言語資源も引き続き重要である。

語という基本的な単位でさえ種類数は多く、同一の被験者の多数の語についての反応を記録することは多大な労力がかかり、多数の被験者を対象とした実験を行うことはコストがかかる。被験者実験を用いずにコーパスから得られる素性から、こうした調査コストの高い指標の値を高い精度で予測することができれば、実用上有用であると考えられる。また、分析の観点からも、こうした多数の被験者を通じてはじめて得られる指標が、被験者とは関係ないテキスト自体の性質から予測可能であることは重要な知見であろう。

こうした単語頻度の計数においては、語の多義性が課題

¹ 静岡理科大学
SIST, Fukuroi, Shizuoka. 437-8555, Japan.
^{a)} ehara.yo@sist.ac.jp

^{*1} 「被験者」は実験参加者や実験協力者と呼ぶべきという提言が存在するが、自然言語処理分野においては一般的に理解されやすい用語として普及したとは言えず、また、日本語の被験者には英語の “subjects” ほどには主従関係を想起させないと思われることから、本稿では一貫して「被験者」を用いる。

であった [6], [9]. 語の多義性も考慮して単語頻度を補正すればより精度よくこうした指標を予測できると考えられるが, どの程度の粒度で語の多義性を考慮すべきかが明らかでなく, また, 語の多義性を人手で付与することも高コストであるという問題があった.

本研究では, 語の多義性考慮した頻度計数法を提案する. BERT[3] など, 個々の単語の出現ごとに語の埋め込み表現が獲得できる文脈化単語埋め込み表現の空間上で, 被験者は限られた範囲の語に反応していると仮定する. 提案手法では, 被験者から収集したデータから, この空間上の範囲をパラメータとして学習しながら, 頻度を計数する. 実験の結果, 提案手法を通じた単語計数法がより精度よく語の難易度を予測できた.

本研究の貢献は下記の通りである.

- (1) 語の多義性を考慮しつつ被験者反応データをうまく説明する語の計数法を提案した.
- (2) 提案手法は単に数を修正する手法とは異なり, 文脈化単語表現空間上で被験者反応データの説明に有効な範囲を学習・特定する事により, コーパス中で具体的にどの使用事例が被験者反応データに沿わないとしてカウントされなかったのかを出力することができる. このため, 語の個々の使用事例ベースの議論が可能となり, 質的な議論との親和性が高いという点において, 解釈性が高い.
- (3) さらに解釈性を高めるため, 被験者反応データにあわせて文脈化単語表現空間の可視化を行い, 可視化された空間上で, 被験者反応データの説明に有効な範囲を学習・特定する可視化手法も提案する.
- (4) 提案する頻度計数法によって修正された頻度を用いて, 被験者反応データの予測タスクを行ったところ, 単純な単語頻度より高い精度で予測できた.

2. 関連研究

2.1 被験者反応データの定式化

本研究では学習者, 語, そして語彙テストの結果が訓練データとして与えられ, 学習者が新規の語に対して正答し得るかを予測する. これを踏まえ, 記法を整理する. I 種類の語彙 $\{v_1, \dots, v_I\}$ と, J 人の学習者集合について考えよう. 以後, 語の添字として i を, 学習者の添字として j を一貫して用いる. 語彙テストの結果は $y_{i,j} \in \{0, 1\}$ で与えられるものとし, $y_{i,j} = 1$ のとき, 学習者 l_j が語 v_i に正答するとき $y_{i,j} = 1$, 誤答であるとき $y_{i,j} = 0$ とする.

既存研究 [4], [5], [7] では, この時の学習者の反応 $y_{i,j}$ を予測するため, 学習者 l_j の能 a_j と語 v_i の難易度 d_i から $y_{i,j}$ の反応が予測可能できるとする, テスト理論 (項目反応理論, 項目応答理論) [1] の考え方を採用している. ここで $\sigma(x) := \frac{1}{1+\exp(-x)}$ はロジスティックモイド関数である.

$$P(y_{i,j} = 1 | l_j, v_i) = \sigma(a_j - d_i) \quad (1)$$

(1) は, $a_j > d_i$ であるとき, 確率が 0.5 を超え, 被験者 l_j が語 v_i に関する問題に正答するとモデル化される. すなわち, 学習者の能力と語の難易度の差 $a_j - d_i$ という値の正負で, 反応を予測するというモデルである.

テスト理論は, いわゆる「試験」のデータを解析するためのモデルであるため, 適用範囲を広くするために, 被験者の正答・誤答データ $y_{i,j}$ を除いて, 被験者や設問に関する情報は入手できないことを通常想定しており, 各被験者・各設問ごとに能力値や難易度パラメータを用意する.

これに対して, 本研究の目的では, 各設問が何らかの「語」に対する反応を問うていることがわかっているため, d_i に直接特徴量を入れるモデル化が考えられる [5]. 例えば, 実際にある被験者反応データについて, (1) を用いて d_i を求めた後, 均衡コーパス中の単語頻度 $\text{freq}(v_i)$ の対数値を用いて d_i を回帰すると, よく相関するという報告がある [2], [12]. この結果を陽にモデルに取り込み, 各語に対して重みパラメータ w_i を用いて, 次式のモデルを考えることができる.

$$d_i = -w_i \cdot \log(\text{freq}(v_i) + 1) \quad (2)$$

(1) に (2) を代入してまとめ, 求めるパラメータが $\{a_j, w_i\}$ であることに注意すると, これは, 自然言語処理分野で多用される, 単純な 2 値のロジスティック回帰と一致することがわかる [5].

2.2 その他の関連研究

語学学習者にとって難しい語を与えるタスクは Complex Word Identification (CWI) と呼ばれる [11], [13]. これらのタスクでは, 語学学習者がどの単語を知っているかを記録したデータが評価のために必要となるが, CWI の shared task では, 文章中で単語が難しいと答えた学習者の人数の情報のみが提供されており, 個々の学習者を区別できない. この点が解決されたデータセットとしては, 100 語, 100 人について, 実際に選択式の語彙テストを受けさせたデータが公開されており [4], 本研究ではこれを用いた.

一方, 学習者の自己申告式で, 多数の語彙について反応を計測したデータもある. 語学学習については, 従来は, [5] による 12,000 語, 16 人のデータの他は入手が難しかったが, 近年 [8] にて 15,000 語, 11 人のデータが公開された. その他, 単語親密度についても, 英語・日本語で大規模なデータセットが公開されている.

3. 提案手法

既存手法では, 単純にコーパス中の単語頻度を素性としたロジスティック回帰を行うだけであった. このような単純な手法では, 単語が多義語や幅広い意味をもつ語であった場合に, あまり用いられないような例外的な使用事例ま

でカウントしてしまうことが容易に考えられる。語彙テスト結果と均衡コーパス中の単語頻度の関係を論じる応用言語学分野でも、同様の問題は議論されてきたが、コーパス中のテキストを入れ替えたり、頻度順位表をみて個別に入れ替えたりする手法が中心的に議論されていた。

提案手法では、語の出現ごとに単語表現ベクトルが生成される、文脈化単語表現ベクトルを用いて、語の頻度を修正する。今、語 v_i の出現が K_i 個あるとして、この K_i 個に対する T_1 次元空間上の文脈化単語表現ベクトルの集合を $X_i = \{x_{i,1}, \dots, x_{i,K_i}\}$ と表記する。また、 $N(\vec{c}_i, \epsilon, X_i)$ を、ベクトル集合 X_i 中のベクトルのうち、 \vec{c}_i との距離が ϵ 以下であるものの個数と定義する。ここで、距離はユークリッド距離とし、各ベクトルの次元は適切に定義されているものとする。 $\epsilon = \infty$ の時、集合 X_i の K_i 個の要素を全てカウントしていることになるため、 $N(\vec{c}_i, \infty, X_i) = K_i$ である。また、 T_1 次元空間ベクトルを $T_2 \leq T_1$ 次元空間ベクトルに射影する行列を考え、 A とする。この時、提案手法による語 v_i の難しさ d_i は、修正頻度 $\text{freq}^{\text{adj}}(v_i)$ を用いて、次のように表される。

$$d_{v_i} = -\log(\text{freq}^{\text{adj}}(v_i) + 1) \quad (3)$$

$$\text{freq}^{\text{adj}}(v_i) = N(A\vec{c}_i, \epsilon, AX_i) \quad (4)$$

$$\approx \sum_{k=1}^{K_i} \tanh(M \cdot \text{ReLU}(\epsilon - \sqrt{\|A\vec{c}_i - A\vec{x}_{i,k}\|^2})) \quad (5)$$

ただし、上記において、また、 $\text{ReLU}(x) = \max(0, x)$ とし、 M は大きい定数とする。また、 \tanh はハイパボリックタンジェント関数である。

(4) が、ReLU 関数と \tanh を用いて (5) の形で近似できることが、実装上の重要な点である。ReLU 関数は負の数を与えられれば 0 を返し、 \tanh は大きい正の数に対してほぼ 1 を返す $\tanh(0) = 0$ を満たす関数であるため、 \tanh の入力に大きな定数 M (例えば $M = 100$) をかけることによって、個数を (5) で近似することが可能となる。

これらの関数は、PyTorch などの、自動微分や高度な最適化関数を備えた標準的なニューラルネットワーク構築フレームワークに装備されているため、提案モデルもニューラルネットワークとして実装することが可能である。

(5) において、単語の難易度に関して特に最適化すべきパラメータは、 ϵ 、変換行列 A である。 A が単位行列の場合、単純に、文脈化単語表現空間上での半径 ϵ の超球を考慮することになる。一方、 A が $T_1 \times 2$ 次元行列である場合は、データから 2 次元空間への射影、すなわち、可視化そのものを学習することが可能となる。

その他の拡張として、中心点 \vec{c}_i をも学習する、半径を学習者の能力に依存させて ϵ_j とする、2 値ではなく、語を知っている度合いなどの順序付きクラスに拡張する (Ordered

表 1 BNC コーパスの 10,000 文中のドメインごとの文数

imaginative	21,946
arts	18,289
natural sciences	5,256
social science	7,777
commerce	4,378
leisure	20,300
belief and thought	3,441
world news	764
applied science	2,625
world affairs	15,224

logistic regression の拡張) などが考えられるが、どれも (1) と (5) から容易に拡張できる。

4. 実験結果

本研究では、均衡コーパスとして標準的に用いられている British National Corpus のうち、100,000 文に対して、BERT [3] の bert-base-uncased*2 を適用し、表層から最も遠い層の 768 次元ベクトルを各語の出現に対して集めることにより、100,000 文の全ての語の出現に対して文脈化単語表現ベクトルを得た。各語の \vec{c}_i には、各語の文脈化単語表現ベクトルの重心に、最も近い文脈化単語表現ベクトルを用いた。重心そのものではないので、多義性の高い語で、重心から見るとどの出現も意味的に遠い、といった問題が発生しないように配慮した。

また、 $\{y_{i,j}\}$ としては、一般公開されている語学学習者の語彙テストに対する反応データ [4] を用いた。100 語、100 人のデータを、60% を訓練データ、40% のテストデータにランダムに分割した。本稿では、訓練データにない語の難易度を予測することが主眼であるため、60% の訓練データ中のうち、15 語に対するデータを完全に削除した。

4.1 頻度修正の評価

本研究は、頻度修正が主眼であるため、不均衡なコーパスが与えられた時の頻度修正能力の評価が重要である。British National Corpus には、各文書のタイトルの後にカッコ書きで、人手で付与されたドメインが付されている ()。このうち、数の多い “arts” を用いた。

まず、(5) で $A = I$ 、すなわち、文脈化単語表現ベクトルと同じ次元数の場合で実験を行った。表 2 に結果を記す。arts ドメインのみ、という偏ったテキストを用いた頻度で、語学学習者の単語テストデータを予測する場合、全ドメインの単語頻度を用いる場合と比べて、精度が低いことがわかる。提案手法は、arts ドメインのみの単語頻度も、被験者が反応していると推定される文脈化単語表現ベクトル空間上の領域中の頻度のみを数えることで、精度が向上するような修正を行えている。この効果は、全ドメインについ

*2 <https://github.com/google-research/bert>

表 2 各ドメインの単語頻度のみを用いて計測した精度

ドメイン	修正	精度
arts	修正前	0.61
arts	修正後	0.64
全ドメイン	修正前	0.67
全ドメイン	修正後	0.72

表 3 “period” の中心点から遠い・近い事例

period pains can be severe and disruptive.
to produce a slight spread of magnetic field period .
design during this period was in the plan .
the pub designer of the period ,

でも認められた。これは、提案手法が外れ値となるような用例を除外して頻度カウントを行った効果であると思われる。(5)における修正前後での精度の向上は、arts ドメインのみの単語頻度を修正した場合も、全ドメインでの単語頻度を修正した場合も、統計的有意であった ($p < 0.01$, Wilcoxon 検定)。

4.2 可視化の学習結果

次に A を $T_1 \times 2$ 行列とし、2次元空間への可視化を学習した場合を、図 1, 図 2 に示す。 A の初期値は、主成分分析による 2次元への射影行列とした。初期値こそ主成分分析による射影ではあるが、主成分分析を単純に行っただけではなく、射影行列 A そのものが、半径 ϵ とともに、データから学習されていることに注意されたい。

図 1, 図 2 から、提案手法が、外れ値となるような語の使用事例を除いて、中心的な語義のみをカウントしていることが見て取れる。表 3 に、“period” の中心点から最遠の事例 2 つ、最近の事例 2 つを、それぞれ表示した。最遠の事例は“period pain”や“magnetic field period”といった専門用語のような使用例であるのに対し、最近の事例 2 つは、“this period”や“the period”など、期間を表す名詞として使われている事例であることがわかる。

5. 議論

自然言語処理の観点からは、語義曖昧性解消を行い、語義ごとに頻度をカウントする方法が容易に考えられる。本研究も文脈を考慮して計数する点はこの考え方に類似しているが、語義曖昧性解消とは目的が異なる。例えば、語を区別する粒度の問題が挙げられる。語義曖昧性解消における厳密な区別と、語学学習の目的における区別は当然異なると考えられる。単純に語義曖昧性解消を適用してから本稿のタスクに適用することは、目的の異なるタスクの結果を無理に利用しようとして、語義曖昧性解消という、より複雑な問題に帰着させて問題を解決しようとしていることになり、問題の解決策としての有効性が不明である。また、語義曖昧性解消タスクでは、通常、人手で語義

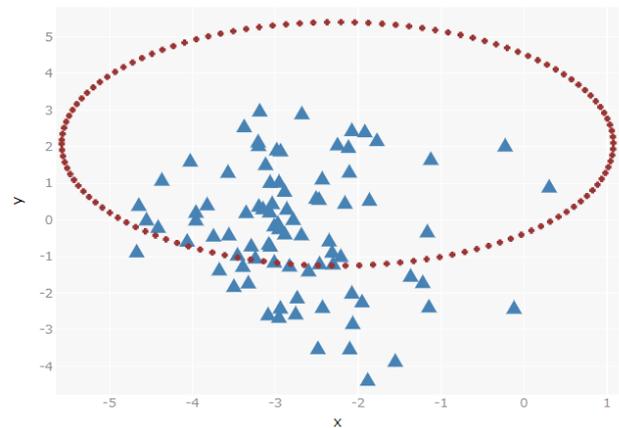


図 1 語 “period” の用例の学習済み可視化。三角形の各点が BNC 中の arts ドメインの語の出現（使用事例）を表し、赤い円中の使用事例のみをカウントするように学習された。

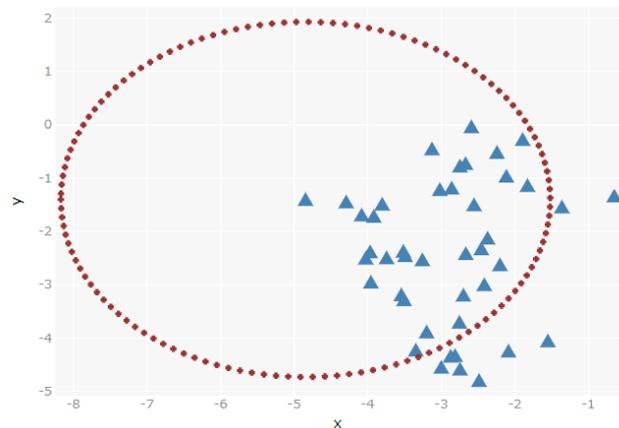


図 2 語 “figure” の学習済み可視化。同様に、赤い円中の使用事例のみをカウントするように学習された。

が付与されたデータセットが使用される。他言語で同一のデータセットは通常存在しないため、他言語への手法の拡張が困難になると想定される。

6. おわりに

本稿では、被験者反応データに対して、コーパス中の使用事例のうち被験者が反応している領域を推定し、推定領域内の頻度のみを計数することで頻度修正を行う手法を提案した。提案手法は、使用事例や単語頻度という解釈性の高い単位でモデルを理解することが可能である上、2次元空間への射影により、被験者反応データに即した可視化をも同時に学習することが可能である。実際の被験者反応データに対する実験により、提案手法は、使用事例が特定のドメインに偏っている場合でも、頻度を修正してより精度良く被験者反応を予測可能であることが示された。今後の課題としては、語学学習のためのインタラクティブな可視化や、語彙学習支援で語を覚えた時期を記録し、語を学習すべきタイミングまで組み込んだモデルの構築が挙げら

れる。

謝辞

本研究は、科学技術振興機構、ACT-I 研究費 (JP-MJPR18U8) の支援を受けた。また、産業技術総合研究所の AI 橋渡しクラウド (ABCI) を使用した。

参考文献

- [1] Baker, F. B.: *Item Response Theory : Parameter Estimation Techniques, Second Edition*, CRC Press (2004).
- [2] Beglar, D.: A Rasch-based validation of the Vocabulary Size Test, *Language Testing*, Vol. 27, No. 1, pp. 101–118 (online), available from <https://doi.org/10.1177/0265532209340194> (2010).
- [3] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. of NAACL*, Minneapolis, Minnesota, pp. 4171–4186 (2019).
- [4] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing, *Proc. of LREC* (2018).
- [5] Ehara, Y., Sato, I., Oiwa, H. and Nakagawa, H.: Mining Words in the Minds of Second Language Learners: Learner-Specific Word Difficulty, *Proceedings of COLING 2012*, Mumbai, India, The COLING 2012 Organizing Committee, pp. 799–814 (2012).
- [6] Laufer, B. and Ravenhorst-Kalovski, G. C.: Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension, *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 15–30 (online), available from <https://eric.ed.gov/?id=EJ887873> (2010).
- [7] Lee, J. and Yeung, C. Y.: Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language, *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1–4 (online), DOI: 10.1109/ICNLSP.2018.8374392 (2018).
- [8] Maddela, M. and Xu, W.: A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 3749–3760 (online), DOI: 10.18653/v1/D18-1410 (2018).
- [9] Nation, I.: How Large a Vocabulary is Needed For Reading and Listening?, *Canadian Modern Language Review*, Vol. 63, No. 1, pp. 59–82 (2006).
- [10] Paetzold, G. and Specia, L.: Inferring Psycholinguistic Properties of Words, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, Association for Computational Linguistics, pp. 435–440 (online), DOI: 10.18653/v1/N16-1050 (2016).
- [11] Paetzold, G. and Specia, L.: SemEval 2016 Task 11: Complex Word Identification, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, Association for Computational Linguistics, pp. 560–569 (online), DOI: 10.18653/v1/S16-1085 (2016).
- [12] Tamayo, J. M.: Frequency of Use as a Measure of Word Difficulty in Bilingual Vocabulary Test Construction and

Translation, *Educational and Psychological Measurement*, Vol. 47, No. 4, pp. 893–902 (online), available from <https://doi.org/10.1177/0013164487474004> (1987).

- [13] Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A. and Zampieri, M.: A Report on the Complex Word Identification Shared Task 2018, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 66–78 (online), DOI: 10.18653/v1/W18-0507 (2018).