意味の集中度に基づいた意味変化検出

永田亮 [1]　高村大也 [2] 大谷直輝 [3]　川崎義史 [4]
[1] 甲南大学 [2] 産業総合技術研究所 [3] 東京外語国語大学 [4] 東京大学
nagata-vmf @ ml.hyogo-u.ac.jp.

概要

本稿では，二つのコーパスで意味が異なる単語を検出する手法を提案する．提案手法は，単語ベクトルの方向のばらつきから算出される「意味の集中度」という新しい指標に基づく．これにより，コーパスや単語ベクトルに対する条件が緩和され，適用範囲が広がる．また，計算量も低減され，大規模なコーパスにも適用可能である．更に，意味の差異を検出するだけでなく，意味の広狭の判定および意味が異なる事例の抽出も可能となる．

1　はじめに

本稿[1)]では，「意味の集中度」という新しい指標に基づいた意味変化の検出手法を提案する．意味変化の検出とは，二つのコーパスで意味が異なる単語を検出するタスクのことである．例えば，1800年代/2000年代の文書から，そのような単語としてdynamo（元々の「発動機」という意味に加えて，近年では「エネルギッシュな人」という意味でも使われる）などを検出するタスクである．なお，本稿では，通時的変化以外にも，任意の二つのコーパスにおけるの意味，用法の違いも検出対象とする．

2節で詳述するように，従来は，文脈なし単語ベクトルを利用する手法が主流である．しかしながら，このアプローチは，比較対象のコーパスや単語ベクトルに強い仮定を必要とするという問題がある．そのため，任意のコーパスペアに必ずしも適用できるわけではない（例：母語話者／非母語話者コーパス）．また，複雑な処理を要し，時間的計算量が大きいという課題もある．

本稿では，意味の集中度に基づいて，従来の問題を解決する手法を提案する．提案手法の特徴と本稿の寄与は次の7点にまとめられる．(1) 提案手法は，シンプルで計算コストも低い；(2) また，コーパス／単語ベクトルに対する仮定を必要としない；(3) それにもかかわらず，SemEval-2020 Task1で最高検出精度（0.73）を達成する；(4) 意味の差異の有無だけでなく，意味が広がったのか狭まったのかも判定できる；(5) 更に，意味の差異を表す代表的な事例をコーパスから抽出できる；(6) 実際に，提案手法を用いて，母語話者／非母語話者英語における意味に差異がある単語を代表な事例と共に明らかにする；(7) 提案手法には，von Mises-Fisher分布という数学的裏付があることを示す．

1) 本稿は，文献 [1] の内容を拡張したものである．

2　関連研究

意味変化検出の代表的な手法に，文脈なし単語ベクトルを用いた手法 [2, 3, 4, 5, 6, 7] がある．これらの手法では，対象コーパスに対する強い仮定を必要とするという課題がある．例えば，文献 [3, 4] の手法では，二種類のコーパスから得られた単語ベクトルが線形変換により対応付け可能であると仮定する．意味変化検出は，意味的対応がつかない単語を検出するというタスクであるため，この仮定は好ましくない．より最近の手法 [5, 8, 9] では，仮定の緩和が試みられているが，対象単語以外は意味変化が起こらないなど，依然，一定の仮定を必要とする．一方，提案手法は，これらの仮定は一切必要としない．

文脈付き単語ベクトルを用いる手法 [6, 10] もある．しかしながら，これらの手法は，クラスタリングなどで対象単語の持つ意味の数を推定するなど，別の難しいタスクを解く必要がある．

両アプローチに共通する課題として計算量が多いということもある．例えば，コーパスに出現する単語タイプ全てに対してクラスタリングを行うということは計算量の観点からも好ましくない．提案手法は，計算量が少なく，大規模なコーパス中の全単語を対象にして意味変化の検出が行える．

3 手法

3.1 意味変化検出手法

提案手法では，ある単語における意味の多様性を文脈付き単語ベクトル（以降，単に単語ベクトルと表記する）の方向の多様性として捉える．すなわち，単語ベクトルが様々な方向を向いている単語ほど，幅広い意味を持つと考える．極端な場合，ある単語が，常に同一の文脈（よって，常に同じ意味）で使われる場合，得られる単語ベクトルも同一となり，その方向は一点に集中することになる．

以上を考慮して，単語ベクトルの方向の集中度合いを「意味の集中度」と定義する．幸い，ベクトルの方向の集中度合いは，von Mises-Fisher 分布 [11] で定量化が可能である．この分布は，d 次元の単位ベクトル $\mathbf{x}$ に対して，

$$f(\mathbf{x};\boldsymbol{\mu},\kappa) = z_\kappa \exp\left(\kappa \boldsymbol{\mu}^\mathsf{T}\mathbf{x}\right). \tag{1}$$

と定義される．ここで，$\boldsymbol{\mu}$ ($\|\boldsymbol{\mu}\| = 1$) と κ ($\kappa \geq 0$) は，それぞれ，平均方向と集中度と呼ばれるパラメータである．また，z_κ は正規化定数である．この分布は，ベクトル $\mathbf{x}$（の先端）は平均方向 $\boldsymbol{\mu}$ を中心に集中度 κ で超球面上に分布するとみなす．これを流用し，提案手法では，ベクトル $\mathbf{x}$ を単語ベクトルに対応させ，κ が意味の集中度を表すと考える．

より厳密な議論のために，次の記号を導入する．ある単語タイプに対応する（文脈付き）単語ベクトルを $\boldsymbol{x}$ とする（その次元を d とする）．ただし，$\|\boldsymbol{x}\| = 1$ となるように正規化した単語ベクトルを想定する．また，（ある単語タイプについて）平均した単語ベクトルを $\bar{\boldsymbol{x}}$ と表す（以降，平均単語ベクトルと呼ぶ）．また，そのノルムを l で表す．更に，比較対象の二種類のコーパスを S（ソース）と T（ターゲット）と表す．例えば，l_S はソースコーパス S から得られた平均単語ベクトルのノルムを表す．

以上の記号を用いて，意味変化のスコアを

$$c(S,T) = \log \frac{\kappa_T}{\kappa_S} \tag{2}$$

と定義する．すなわち，二つのコーパスにおける意味の集中度の比により意味変化の度合いを定量化する．対数をとるのは，スコアの符号により意味の広狭を表すためである；スコアの値が負の場合，ソースコーパス S に比べ，ターゲットコーパス T では意味が広がったことを意味する（κ が小さいほど意味が広いことに注意されたい）．

式 (2) を評価するためには，κ の値が必要となる．文献 [11] によると，κ の最尤推定の近似解は，

$$\kappa \approx \frac{l(d - l^2)}{1 - l^2} \tag{3}$$

となる．この式を，式 (2) に代入すると，

$$\log \frac{\kappa_T}{\kappa_S} = \log \frac{\frac{l_T(d-l_T^2)}{1-l_T^2}}{\frac{l_S(d-l_S^2)}{1-l_S^2}} \approx \log \frac{l_T(1-l_S^2)}{l_S(1-l_T^2)} \tag{4}$$

となる．ただし，最後の近似は，$d \gg l^2$ であることを用いた．式 (4) より，スコアは平均単語ベクトルのノルムのみで決定され，非常にシンプルであることがわかる．このことは，使用される文脈が多様であるほど，対応する単語ベクトルの方向も多様になり，その平均ベクトルが短くなると直感的には理解される；極端な場合，常に同一の文脈であれば，各単語ベクトルと平均ベクトルは同一となりノルムは最大の 1 となる．

このスコアを用いた意味変化検出は次の 5 ステップからなる：(1) コーパス S，T 中の全単語を言語モデルにより単語ベクトルへ変換；(2) 単語ベクトルをノルム 1 となるように正規化；(3) 単語タイプごとに，平均単語ベクトル $\bar{\boldsymbol{x}}_S$ と $\bar{\boldsymbol{x}}_T$ とそのノルム l_S，l_T を算出；(4) 式 (4) で定義されるスコアの降順に単語タイプをソート；(5) 結果を出力．

3.2 意味変化の代表事例の抽出手法

片方のコーパスで頻繁に使われ，もう片方のコーパスではほとんど（もしくは全く）使われない意味の事例を抽出する手法について説明する．この条件を，von Mises-Fisher 分布の枠組みで解釈すると，片方のコーパスでは確率が高く，もう片方のコーパスでは確率が低い $\boldsymbol{x}$ を見つける問題と捉えることができる．すなわち，二つのコーパスから得られる対数尤度（LLR）の絶対値が大きい単語ベクトルに対応する事例を抽出すればよい．式 (1) より，

$$\mathrm{LLR} = \log \frac{z_{\kappa_S} \exp(\kappa_S \boldsymbol{\mu}_S{}^\mathsf{T}\boldsymbol{x})}{z_{\kappa_T} \exp(\kappa_T \boldsymbol{\mu}_T{}^\mathsf{T}\boldsymbol{x})} \tag{5}$$

となる．詳細は付録 A に示すが，この比較は，

$$\left(\frac{1}{1-l_S^2}\bar{\boldsymbol{x}}_S - \frac{1}{l-l_T^2}\bar{\boldsymbol{x}}_T\right)^\mathsf{T}\boldsymbol{x} \tag{6}$$

の比較に帰着される．この値が大きい $\boldsymbol{x}$ に対応する事例を抽出することで，コーパス S のみに頻繁に表れる意味の事例が得られる．逆に，値が小さい場合は，両コーパスに共通した意味と解釈される．

このスコアを用いた意味変化の代表事例抽出は次の 4 ステップからなる：(1) 対象単語タイプの平均単語ベクトル $\bar{\boldsymbol{x}}_S$, $\bar{\boldsymbol{x}}_T$ を算出；(2) S 中の対象単語タイプの各事例に対して式 (6) を計算；(3) その値の降順に事例をソート；(4) ソート結果を出力.

4 評価実験

4.1 使用データと実験条件

歴史コーパス CCOHA [12] と英語学習者コーパス ICNALE [13] を用いた．前者は，意味変化の検出精度，意味の広狭判定に用いた．後者は，母語話者／非母語話者の英文から意味の差が大きい単語を代表事例と共に抽出する実験に用いた．いずれの場合も頻度が 10 より大きい単語タイプのみを対象とした．

単語ベクトルの取得には，BERT [14] を用いた（最終層の出力を単語ベクトルとした）．定量評価には，'bert-large-cased' を，定性評価には，'bert-large-uncased' を用いた．

4.2 意味変化検出精度

SemEval-2020 Task 1 [15] sub-task 1 の英語データを対象にして意味変化の検出精度を評価した．同タスクは，CCOHA の 1810〜1860 年代と 1960〜2010 年代の文書における 37 種類の単語に対して，意味変化の有無を判定する．提案手法では，37 の単語に対して意味変化スコアを算出し，k-means++クラスタリングを適用して，意味変化のあり／なしの 2 クラスタに分類した（値が高いほうを変化ありとした）．

表 1 に検出精度を示す．提案手法に加えて，SemEval-2020 Task 1 で最高性能を達成した手法 [16]，von Mises-Fisher 分布の平均方向を利用した手法（$1 - \cos(\bar{\boldsymbol{x}}_S, \bar{\boldsymbol{x}}_T)$ をスコアとした），平均方向と集中度を使用した手法（$1 - \cos(\bar{\boldsymbol{x}}_S, \bar{\boldsymbol{x}}_T)$ と式 (4) の平均値をスコアとした）の検出精度も示している．表 1 より，提案手法は最高性能手法 [16] と同等の検出精度を達成していることがわかる．文献 [16] の手法は，単語ベクトルの次元削減とクラスタリングという複雑な処理を要する．一方，提案手法は平均単語ベクトルのノルムを計算するだけで非常にシンプルである．それにもかかわらず同等の性能を示すことは特筆すべきことである．また，平均方向を利用した手法は，提案手法よりも検出精度が低い．このことは，平均ベクトルの方向よりも集中度のほうが意味変化の検出には重要である可能性を示唆する．

表 1 意味変化検出精度.

手法	検出精度
提案手法	0.730
Rother らの手法 [16]	0.730
平均方向のみ	0.622
平均方向と集中度の平均	0.702

4.3 意味の広狭判定性能

我々が知る限り，意味の広狭判定用の公開評価データは存在しない．そこで，DWUG [17] を利用して独自データを作成した．DWUG は，上述の 37 種類の単語に対して，人手で意味の同一性を判定したデータを収録している．更に，その結果にクラスタリングを適用し，意味ラベルを付与した事例も収録している（元コーパスは SemEval-2020 Task 1 の一部である）．この意味ラベルの相対頻度を確率とみなして算出したエントロピーを意味の多様性の指標としたとした（意味ラベルの種類が多いほど，分布が一様であるほどエントロピーは高くなる）．最終的に，二つの年代（1960〜2010 年と 1810〜1860 年）の文書から求められたエントロピーの差を意味の広狭の指標とした（正負が意味の広狭に対応する）．このエントロピーの差と提案手法の意味変化スコアとのピアソンの積率相関係数で性能を評価した．

その結果を散布図と共に図 1 に示す．図 1 から，エントロピーと意味変化スコアは正の相関を示し，意味変化スコアにより，意味の広狭の判定がある程度できることがわかる．ただし，エントロピーの差が負となる単語（すなわち意味が狭まった単語）に対して，意味変化スコアでは正の値が与えられることが多いことも見て取れる．一つの可能性として，主に現代の文書で訓練された BERT が 1800 年代の文書に対して十分に機能しないということが考えられる（文脈に応じて単語ベクトルの方向を十分に分散させられない）．この予想が正しいとすると，意味変化スコア 0 でなく，より大きい値が意味の広狭の境目となる．この点については，更なる調査が必要である．ただし，前節や以降の節のように，単語タイプ同士を比較して意味変化が大きい順に並べる場合には，影響が少ないことが予想される．なぜなら，意味変化スコアは集中度の比をとるため，二つのコーパスにおける集中度のスケールの差異は，単語タイプの比較では共通するためである．

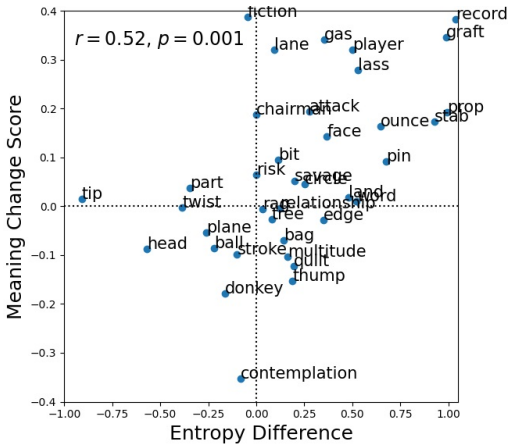

図 1　意味ラベル分布に対するエントロピーの差と意味変化スコアの関係.

4.4　定性評価：母語話者/非母語話者の比較

ICNALE の母語話者と非母語話者の文書に提案手法を適用し，意味の差異が大きい単語を代表事例と共に検出した．紙面の関係から，詳細な結果は付録に示すが結果は次のようにまとめられる．

イディオム／比喩的表現：母語話者の文書では，比喩的な用法が多く検出された．例えば，place という単語は，母語話者／非母語話者とも，字義的な「場所」という意味で頻繁に使用するが，母語話者は加えて，in place や fall into place など比喩的に使用することも多い．イディオムが逆の働きをするケースも見られた．例えば，非母語話者は，course を of course というイディオムで頻繁に使用していた．驚くべきことに，course の 87%は当該イディオムとし出現していた．一方，母語話者は，「授業」などのより広い意味で使用しており，相対的に当該イディオムの頻度が低くなっていた（63%が該当）．

品詞／構文の違い：典型例は near と knowledge である．near は，母語話者コーパスに 11 回出現したが，意味の差異がある代表事例は副詞で「ほとんど」と訳される事例であった．この用法は非母語話者コーパスにおける near の 267 事例には確認出来なかった．したがって，ICNALE の非母語話者は，この用法を習得していない可能性が高い．更に，この例は，提案手法が低頻度な単語に対してもある程度有効であることを示す（母語話者コーパス中に near は 11 回しか出現しなかったが，それでも，非母語話者コーパスに比べ意味が広い事例を捉えられている）．同様なことが，knowledge についてもいえる．knowledge は，母語話者コーパスに 16 回しか登場していない．この 16 回のうち，代表事例上位 2 件は，主語 it の内容が，that 以下で説明される構文での出現であった（knowledge は，主語 it と be 動詞で結ばれるため，knowledge の内容を that 以下が表しているとも解釈できる）．この構文は，非母語話者の 574 の knowledge の事例のうちたった 2 件しか該当しなかった（コーパスを確認したところ，knowledge that という表現は，22 件存在し，19 件が関係代名詞，2 件が当該構文，残り 1 件は誤りにより判断がつかなかった）．この例は，BERT から得られた単語ベクトルに，この構文と that-関係代名詞を区別するための情報が内在していることを示唆しており興味深い．

5　考察

前節の評価実験より，意味の集中度に基づいた提案手法の有効性を確認した．3.1 節で示したように，意味の集中度の計算は，平均単語ベクトルの算出に帰着される．このことにより，従来手法のような仮定を必要としない，計算コストも低いという利点が得られる．更に，表 1 は，意味の集中度（言い換えれば，分散の逆数）が，意味変化の検出において重要なことを示唆する．同様なことを，Aida ら [18] は，ガウス分布の平均と共分散に基づいた手法で示している（ただし，彼らの手法は，平均も使っており，共分散だけの効果は明らかでない）．従来は，平均のみで意味変化を捉える試みが多く，分散も重要であることを示したこの二つの研究は，意味変化検出の新しい方向性を示したといえる．

更に，4.3 節で示したように，提案手法は意味の広狭も判定できる．これもまた，意味の集中度がもたらす利点である．すなわち，集中度の大小で，意味の広狭を自然に表すことができる．一方で，従来研究で主流である平均（ベクトル）に基づく手法で，意味の広狭を定量化することは困難である．

6　おわりに

本稿では，「意味の集中度」という新しい指標を提案し，意味変化の検出とその代表事例の抽出に有効であることを示した．また，意味の広狭判定に対しても一定の効果があることも示した．提案手法には，(1) 従来手法より制約が少なく適用範囲が広い，(2) 計算量も少ないという優位な点がある．更に，提案手法を用いて母語話者／非母語話者英語における意味変化の具体的な事例を示した．以上の優位な点は，von Mises-Fisher 分布という数学的裏付けに起因することを示した．

謝辞

本研究の一部は JSPS 科研費 JP22K12326 および JP23K12152 により実施した.

参考文献

[1] Ryo Nagata, Hiroya Takamura, Noki Otani, and Kawasaki Yoshifumi. Variance matters: Detecting semantic differences without corpus/word alignment. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2023.

[2] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In **Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science**, pp. 61–65, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[3] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In **Proceedings of the 24th International World Wide Web Conference**, 2015.

[4] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[5] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In **Proceedings of the 11th ACM International Conference on Web Search and Data Mining**.

[6] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3960–3973, Online, July 2020. Association for Computational Linguistics.

[7] Matej Martinc, Petra Kralj Novak, and Senja Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4811–4819, Marseille, France, May 2020. European Language Resources Association.

[8] Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. Time-out: Temporal referencing for robust modeling of lexical semantic change. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 457–470, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. A comprehensive analysis of PMI-based models for measuring semantic differences. In **Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation**, pp. 21–31, Shanghai, China, 11 2021. Association for Computational Lingustics.

[10] Kazuma Kobayashi, Taichi Aida, and Mamoru Komachi. Analyzing semantic changes in Japanese words using BERT. In **Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation**, pp. 270–280, Shanghai, China, 11 2021. Association for Computational Lingustics.

[11] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. **Journal of Machine Learning Research**, Vol. 6, No. 46.

[12] Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean corpus of historical American English. In **Proc. of the 12th Language Resources and Evaluation Conference**, 2020.

[13] Shinichiro Ishikawa. **A new horizon in learner corpus studies: The aim of the ICNALE project**. University of Strathclyde Publishing, Glasgow.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In **Proceedings of the Fourteenth Workshop on Semantic Evaluation**, pp. 1–23, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[16] David Rother, Thomas Haider, and Steffen Eger. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In **Proceedings of the Fourteenth Workshop on Semantic Evaluation**, pp. 187–193, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[17] Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. DWUG: A large resource of diachronic word usage graphs in four languages. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7079–7091, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[18] Taichi Aida and Danushka Bollegala. Unsupervised semantic variation prediction using the distribution of sibling embeddings. In **Findings of the Association for Computational Linguistics: ACL 2023**, Toronto, Canada, July 2023. Association for Computational Linguistics.

付録

A　式 (6) の証明

式 (5) は，

$$\begin{aligned} \mathrm{LLR} &= \log \frac{z_{\kappa_S} \exp\left(\kappa_S \boldsymbol{\mu}_S{}^{\mathsf{T}} \boldsymbol{x}\right)}{z_{\kappa_T} \exp\left(\kappa_T \boldsymbol{\mu}_T{}^{\mathsf{T}} \boldsymbol{x}\right)} \\ &= \log \frac{z_{\kappa_S}}{z_{\kappa_T}} + (\kappa_S \boldsymbol{\mu}_S - \kappa_T \boldsymbol{\mu}_T)^{\mathsf{T}} \boldsymbol{x} \end{aligned} \tag{7}$$

と変形できる．この式の平均方向 $\boldsymbol{\mu}$ の最尤推定は $\boldsymbol{\mu} = \frac{\overline{\boldsymbol{x}}}{l}$ で与えられる [11]．$\boldsymbol{x}$ に関しては，式 (7) の二行目の第二項のみが大小関係に影響を与える．したがって，$\boldsymbol{\mu} = \frac{\overline{\boldsymbol{x}}}{l}$ と式 (3) で表される κ の最尤推定を第二項に代入すると $(\frac{d-l_S^2}{1-l_S^2}\overline{\boldsymbol{x}}_S - \frac{d-l_T^2}{1-l_T^2}\overline{\boldsymbol{x}}_T)^{\mathsf{T}}\boldsymbol{x}$ が得られる．ここで，$d \gg l^2$ に注目すると，式 (6) が得られる．

B　母語話者／非母語話者の比較

学習者コーパス ICNALE に収録されている母語話者／非母語話者コーパスに提案手法を適用し，意味，用法に差異がある単語を代表事例と共に検出した．頻度が 10 より大きい単語を対象とした．また，複数のサブワードに分割される単語，アルファベット以外の文字を含む単語，エッセイのプロンプトに含まれる単語は対象外とした．

表 2 に，母語話者コーパスでより広い意味を持つと判定された単語上位 10 件を示す．代表事例カラムの S_i は，ソースコーパス（母語話者コーパス）から抽出された代表事例の i 番目であることを示す（T_i も同様の意味である）．表 2 の結果は，4.4 で論じたように，「イディオム／比喩的表現」と「品詞／構文の違い」の二グループに大きくまとめられる．

表 3 に，非母語話者コーパスでより広い意味を持つと判定された単語上位 10 件を示す．形式は，表 2 と同様である．表 3 から，綴り誤り（form vs. from），文法誤り（hope money, a man who responsible），コロケーションの違い（great damage；母語話者はポジティブな語との共起が中心）などが見て取れる．

表 2　母語話者コーパス（*S*ource）で非母語話者コーパス（*T*arget）より幅広い意味を持つと検出された単語上位 10 件．f_S，f_T: S，T での出現頻度．

$c(S,T)$	単語タイプ	f_S	f_T	代表事例
0.54	near	11	267	S_2: it has become *near* impossible
0.46	concerned	11	113	T_1: as far as I'm *concerned*
0.45	third	11	348	T_1: Third, $\cdots$
0.39	period	13	115	S_1: Period!
0.38	first	87	1512	T_1: First, $\cdots$
0.37	course	46	489	T_1: Of *course* $\cdots$
0.37	place	67	1764	S_1: put a ban in *place* / S_6: fall into *place*
0.36	taking	34	461	S_1: *taking* away / T_1: *taking* a part time job
0.34	hold	16	111	S_1 *hold* down a job
0.34	knowledge	16	574	S_1: it is common *knowledge* that smoking and passive smoking kill people

表 3　非母語話者コーパス（*S*ource）で母語話者コーパス（*T*arget）より幅広い意味を持つと検出された単語上位 10 件．f_S，f_T: S，T での出現頻度．

$c(S,T)$	単語タイプ	f_S	f_T	代表事例
0.76	form	92	21	S_1: learn *form* the books / T_1: some *form* of
0.72	degree	70	39	S_2: in some *degree* / S_3: to some *degree*
0.64	pretty	44	28	S_5: *pretty* clothes
0.62	hope	195	24	S_3: must *hope* money from their parent
0.61	worry	94	27	S_2: a lot *worry* / S_5 their *worry*
0.55	section	85	11	S_3: important *section* of students / T_1: a smoking *section*
0.53	great	429	67	S_1: *great* harm / S_2: *great* damage
0.52	responsible	127	44	S_1: a man who *responsible* / S_3: their *responsible*
0.51	staff	45	14	S_1: a concert *staff* / S_4: a part time *staff*
0.51	yes	68	13	S_1: If we say *yes*,