深層異常検知に基づく多義語のコアミーニングを考慮した既習語予測モデルの定式化

江原 遥
静岡理工科大学
ehara.yo@sist.ac.jp

1 はじめに

外国語語彙学習における多義語の扱いについては，類似した複数の語義の背後に核となる「コアミーニング」があることを仮定し，これを核に，派生する意味を教える教育法が提案されている [15]．例えば，英語の “pool” には水泳のプールの意味の他に，“car pool” など何かをためておいたものを表す意味があり，実際に自然言語処理分野における英語の語義集合である WordNet でも別の語義としてカウントされている．最終的な目的が計算機による自然言語理解であるならば，このように細かく語義を分けることは機械可読性の観点から重要であろう．しかし，外国語の語彙学習の目的では，これらを全く別の語義として，１つ１つ教育していく方法は非効率であり,「ためておく」というコアミーニングを最初に教え，そこから派生して他の語義を覚えさせた方が，学習者にとって理解しやすいことは容易に分かる．こうした事例は，[15] において大まかに解説されている．

一方，機械学習においては，近年，文脈を考慮して単語の各出現（用例）に対して異なるベクトル表現を求める「文脈化単語埋め込み」の手法が自然言語理解の上で，大きなブレイクスルーを起こしている [4]．これにより，従来は難しかった語義ごとの単語埋め込み表現を出力する方法が研究されている [13]．語彙学習における「コアミーニング」についても，文脈化単語埋め込みの技術を活用して，実際に単語埋め込み表現を求めることはできないだろうか？これが，本稿の主たる動機である．

「コアミーニング」については，教育における概念的な内容であることから，実際にどの語のコアミーニングがどういうものであるか，などを記したデータセットが，著者の知る限り存在しない．すなわち，教師データがほとんど存在しない状況である．一方，語彙学習支援においては，単語テストのデータは広く入手でき，学習者が単語テストの各設問に正解するかどうかを予測する，既習語予測モデルについては教師データが存在する [5]．そこで，本稿では，既習語予測モデルの中間表現として文脈化単語埋め込みから計算したコアミーニングの表現を利用するニューラルモデルとして，どのようなものが考えられるのかを議論する．特に，コアミーニングは，細かい違いを捨象した表現であるので，類似した複数の事例を「正常」か「異常（外れ値）」かの２つに教師なしで分類する，深層異常検知モデルが既習語予測モデルに組み入れられないか議論する．

2 関連研究

2.1 深層異常検知

深層異常検知の近年の代表的な手法として，DAGMM[12] が挙げられる．これは，混合ガウスモデルを深層化し，高次元ベクトルを扱えるようにした手法である．語の多義性については，文脈化単語埋め込み表現をクラスタリングして，各クラスタを語義とみなしてまとめる手法が容易に考えられるが，混合ガウスモデルはこのクラスタリングの代表的な手法であるため，親和性が高いと考えられる．

この手法では，入力ベクトル $\mathbf{x}$ をオートエンコーダを用いて低次元潜在表現 $\mathbf{z}$ を通じて再構成するニューラルネットワークを考える．再構成したベクトルを $\mathbf{x}' = g(\mathbf{z}_c;\theta_d)$ とし，潜在表現を $\mathbf{z}_c = h(\mathbf{x};\theta_e)$ とする．再構成したベクトルと元の入力の近さを測る関数を $\mathbf{z}_r = f(\mathbf{x},\mathbf{x}')$ とする．ここで，この近さとして複数の尺度を利用して良い．最終的な潜在表現は潜在表現と，再構成の誤差の表現をつなげた $\mathbf{z} = [\mathbf{z}_c,\mathbf{z}_r]$ となる．

潜在表現から，先は，基本的な混合ガウスモデルの表記にしたがって，次のように定義される．ここで，

K がクラスタ数で，N はデータの数である．MLN は Multi layer network の略である．

$$\mathbf{p} = MLN(\mathbf{z}; \theta_m), \hat{\gamma} = \mathrm{softmax}(\mathbf{p}) \tag{1}$$

クラスタ k の負担率などは次で定義され，平均と分散共分散行列は下記になる．

$$\hat{\phi_k} = \sum_{i=1}^{N} \frac{\hat{\gamma}_{ik}}{N}, \tag{2}$$

$$\hat{\mu_k} = \frac{\sum_{i=1}^{N} \gamma_{ik} \mathbf{z}_i}{\sum_{i=1}^{N} \gamma_{ik}}, \hat{\mathbf{\Sigma}}_k = \frac{\sum_{i=1}^{N} \hat{\gamma_{ik}} \left(\mathbf{z}_i - \hat{\mu}_k\right) \left(\mathbf{z}_i - \hat{\mu}_k\right)^\top}{\sum_{i=1}^{N} \hat{\gamma}_{ik}} \tag{3}$$

最終的に，ある入力 $\mathbf{x}$ の潜在表現 $\mathbf{z}$ について，これが異常値である度合いは下記のエネルギー関数であらわされる．これは，直感的には，k 番目のクラスタの中心から $\mathbf{\Sigma}_k$ を用いて $\mathbf{z}$ への距離を測り，どのクラスタからも距離が大きいものを異常値とみなしていると説明できる．

$$E(\mathbf{z}) = -\log \left(\sum_{i=1}^{K} \hat{\phi}_k \frac{\exp\left(-\frac{1}{2} \left(\mathbf{z} - \hat{\mu}_k\right)^\top \mathbf{\Sigma}_k^{-1} \left(\mathbf{z} - \hat{\mu}_k\right)\right)}{\sqrt{\left|2\pi \hat{\mathbf{\Sigma}}_k\right|}} \right) \tag{4}$$

最後に，結局最適化する目的関数は下記となる．L はベクトルの再構成に関する損失関数，P は罰則項であり，λ はハイパーパラメタである．この目的関数をニューラル機械学習における標準的な最適化法によって，ニューラルネットワークの良いパラメタを求める．

$$J(\theta_e, \theta_d, \theta_m) = \frac{1}{N} \sum_{i=1}^{N} L(\mathbf{x}_i, \mathbf{x}'_i) + \frac{\lambda_1}{N} \sum_{i=1}^{N} E(\mathbf{z}_i) + \lambda_2 P(\hat{\Sigma}) \tag{5}$$

2.2 その他の関連研究

テキスト中で語学学習者にとって難しい語を探すタスクは Complex Word Identification (CWI) と呼ばれる [9, 11]．これらのタスクは，必ずしも語学学習者が語を知っているかどうかには関わらず，母語話者や語学学習者が，他の語学学習者にとって難しいと思われる語を発見するタスクであるところが異なる．

3 既習語予測問題の定式化

既習語予測問題に関しては，[14] の報告がある．この方法では，やはり，既習語予測問題をニューラルなモデルとして定式化している．

既習語予測問題では学習者，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．M 種類の語彙 $\{v_1, \ldots, v_M\}$ 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 m 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 $y_{m,j} \in \{0, 1\}$ で与えられるものとし，$y_{m,j} = 1$ のとき，学習者 l_j が語 v_m に正答するとき $y_{m,j} = 1$，誤答であるとき $y_{m,j} = 0$ とする．

既存研究 [5, 7, 6] では，この時の学習者の反応 $y_{m,j}$ を予測するため，学習者 l_j の能 a_j 力と語 v_m の難易度 d_m から $y_{m,j}$ 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] の考え方を用いている．ここで $\sigma(x) := \frac{1}{1+\exp(-x)}$ はロジスティックシグモイド関数である．

$$P(y_{m,j} = 1 | l_j, v_m) = \sigma(a_j - d_m) \tag{6}$$

この式は $a_j > d_m$ であるとき，確率が 0.5 を超え，被験者 l_j が語 v_m 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 $a_j - d_m$ 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 $y_{m,j}$ 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_m に直接特徴量を入れるモデル化が考えられる [6]．例えば，実際にある被験者反応データについて，式 6 を用いて d_m を求めた後，均衡コーパス中の単語頻度 $\mathrm{freq}(v_m)$ の対数値を用いて d_m を回帰すると，よく相関するという報告がある [10, 2]．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タ w_m を用いて，次式のモデルを考えることができる．

$$d_m = -w_m \cdot \log(\mathrm{freq}(v_m) + 1) \tag{7}$$

式 6 に式 7 を代入してまとめ，求めるパラメタが $\{a_j, w_m\}$ 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 [6]．

4 ニューラルモデルにおける中心からの範囲内の単語頻度カウント

前述の方法は，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 [8].

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_m の出現が I_m 個あるとして，この I_m 個に対する文脈化単語埋め込みベクトルの集合を $X_m = \{\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,I_m}\}$ と表記する．

前述のように，単語の難易度パラメタは単語頻度と相関があることが知られている．本稿の目的では，この単語頻度を，単純に表層的な単語の頻度ではなく，個々の出現の文脈まで考慮し,「外れ値ではない」単語の出現の回数が難易度パラメタと相関すると考えることが自然であるように思われる．前述のように，ある単語のある出現 $\mathbf{x}_{m,i}$ について，その異常値である度合いは，$E(\mathbf{z}_{m,i})$ であらわされるので，これが閾値 ϵ 以内である単語の数だけカウントする事で単語頻度を修正し $\mathrm{freq}^{\mathrm{adj}}(v_m)$ とすることが考えられる．

$$d_{v_m} = -\log(\mathrm{freq}^{\mathrm{adj}}(v_m) + 1) \quad (8)$$

$$\mathrm{freq}^{\mathrm{adj}}(v_m) \approx \sum_{i=1}^{I_m} \tanh\left(C \cdot ReLU(\epsilon - E(\mathbf{z}_{m,i}))\right) \quad (9)$$

ここで，上記において，$\mathrm{ReLU}(x) = \max(0, x)$ とし，C は大きい定数とする．

頻度を，ReLU 関数と tanh を用いて式 9 の形で近似できることが，重要な点である．ReLU 関数は負の数が与えられれば 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 C（例えば $C = 100$）をかけることによって，個数を式 9 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．

式 9 は，DAGMM を適用して $\mathbf{z}$ を求めてしまってから，あとで適用することも考えられるし，最終的にパラメタを最適化するための目的関数を DAGMM と組み合わせることによって，既習語予測モデルと同時に解くことも可能となると予想される．また，閾値 ϵ については，単語の語種によって異なることも考えられるので，語種ごとに事なる閾値を設けたり，ϵ_m に対してハイパーパラメタを定義して，閾値の値が語種によって大きくは変わらないように設定することも可能であると考えられる．

また，もともとの目的のコアミーニングに関して言えば，結局，$\mathbf{z}_{m,i}$ が単語 v_m の i 番目の出現（用例）に対するコアミーニングのベクトル表現とみなせることが予想される．

5 おわりに

本稿では，語彙学習におけるコアミーニングのベクトル表現を深層異常値検知モデル DAGMM を基に，中間表現として陽に求めながら既習語の予測ができるモデルが構築できるモデルという新しい研究の方向性を考え，具体的にそれを達成することが可能と予想されるモデルを定式化した．語彙学習においては，コアミーニングのように，概念は直感的に理解できるものの，具体的なデータを作りにくいタスクが多くあり，本稿がそのような研究の発展に役立てば幸いである．今後の課題としては，実際にこのモデルを実装して実験を行い，既習語の予測精度を評価すること，また，求められたコアミーニングの解釈を説明可能 AI などの観点から評価することが挙げられる．また，[3] などを拡張し，語彙学習を強化学習の観点からモデル化することも面白い．

謝辞

本研究は，科学技術振興機構 ACT-I 研究費 (JPMJPR18U8)，ならびに日本学術振興会科学技術研究費補助金（18K18118）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] Frank B. Baker. *Item Response Theory : Parameter Estimation Techniques, Second Edition*. CRC Press, July 2004.

[2] David Beglar. A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, Vol. 27, No. 1, pp. 101–118, January 2010.

[3] Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie. DAS3h: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, May 2019. arXiv: 1905.06873.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, pp. 4171–4186, Minneapolis, Minnesota, June 2019.

[5] Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.

[6] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining Words in the Minds of Second Language Learners: Learner-Specific Word Difficulty. In *Proceedings of COLING 2012*, pp. 799–814, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[7] John Lee and Chak Yan Yeung. Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1–4, April 2018.

[8] I. S. P. Nation and Rob Waring. *Teaching Extensive Reading in Another Language*. Routledge, November 2019. Google-Books-ID: xRu_DwAAQBAJ.

[9] Gustavo Paetzold and Lucia Specia. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 560–569, San Diego, California, June 2016. Association for Computational Linguistics.

[10] Jose M. Tamayo. 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, December 1987.

[11] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. A Report on the Complex Word Identification Shared Task 2018. *arXiv:1804.09132 [cs]*, April 2018. arXiv: 1804.09132.

[12] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

[13] 芦原和樹, 梶原智之, 荒瀬由紀, 内田諭. 多義語分散表現の文脈化. 自然言語処理, Vol. 26, No. 4, 2019.

[14] 江原遥. 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. 第 243 回自然言語処理研究発表会予稿集, 2019.

[15] 中田達也. 英単語学習の科学. 研究社, 2019.