上位語・下位語の射影関係とそのクラスタの同時学習

Jointly Learning Hypernym-Hyponym Relations and their Clusters

山根 丈亮 † 高谷 智哉 ‡ 山田 整 ‡ 三輪 誠 † 佐々木 裕 †
Josuke Yamane Tomoya Takatani Hitoshi Yamada Makoto Miwa Yutaka Sasaki
† 豊田工業大学
Toyota Technological Institute
‡ トヨタ自動車株式会社
Toyota Motor Corporation
†{sd12087, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp
‡{tomoya_takatani, hitoshi_yamada_aa}@mail.toyota.co.jp

1 はじめに

日本語の意味的階層情報（日本語版 WordNet [1] など）は，英語版と比較しても網羅率が十分とは言えない．階層情報を広く網羅するには上位概念語の自動獲得が有効であると考えられる．Wikipedia から上位概念語を自動獲得する手法 [2] では，多くの上位概念語を自動獲得することが可能であるが，Wikipedia に記述がない未知語に対しては上位概念語を獲得することができない．

単語を数値ベクトルで表す単語ベクトル（word embedding; WE）を用いた Fu らの上位概念語推定手法 [3] は高い性能を示している．この手法では，下位概念語と上位概念語の単語ベクトルの差に関してクラスタリングを行い，各クラスタで下位概念後の単語ベクトルを上位概念語のベクトルへ射影する行列を求める 2 段階の手法により上位概念語を推定している．この手法では単語ベクトルの学習データ内に単語が出現していれば，上位概念語を推定することができるので，Wikipedia の記述にない単語にも対応できる．しかしながら，クラスタリングが射影行列の学習と独立しているため，射影行列の学習が単語ベクトルの表現やクラスタリングの性能に依存している．また，学習データ内の上位・下位概念語ペアのみ学習するため，下位概念語を射影した先の近くに正解の上位概念語が位置していても，不正解の上位概念語がそれより近くにある場合，不正解の上位概念語を推定してしまう．

これらの問題に対処するため，本研究では学習データの上位・下位概念語ペアに出てこない単語の組み合わせについても負例として学習を行い，射影行列の学習中にクラスタリングを同時に行うことで，上位概念語の自動獲得の精度向上を目指す．

2 関連研究

2.1 ivLBL

WE を獲得する手法の一種である inverse vector log-bilinear language model (ivLBL) [4] では，対象の単語 w_t と，その単語から相対的に i 語離れた文脈内の単語 w_{t+i} とのスコア関数 $s(w_t, w_{t+i})$ が大きくなるように単語ベクトルを学習している．次の (1) 式に $s(w_t, w_{t+i})$ を示す．

$$s(w_t, w_{t+i}) = \mathbf{w}_{t+i} \cdot \mathbf{w}_t + b_{w_t} \tag{1}$$

ここで，$\mathbf{w}_t$ は対象の単語 w_t に対応する単語ベクトル，b_{w_t} は対象の単語 w_t におけるバイアスである．ただし i の範囲はウィンドウサイズ n によって決まり，$-n \leq i \leq n\ (i \neq 0)$ である．ivLBL では，次の (2) 式における g_t を最大化することで $\mathbf{w}_t$ と $\mathbf{w}_{t+i}$ を学習する．

$$g_t = \log[\sigma(s(w_t, w_{t+i}))] + \sum_{w'_t \sim P_n}^{k} \log[1 - \sigma(s(w_t, w'_t))] \tag{2}$$

$\sigma(x)$ はロジスティック関数，k は単語の頻度分布 P_n から選んだ文脈外の単語 w'_t の数である．

2.2 射影行列を用いた上位概念語ベクトルの表現

Fu ら [3] は，下位概念語のベクトルをその上位概念語のベクトルへ射影する行列を学習する手法を提案している．一般に上位・下位概念語間のベクトル関係は多様であり，唯一の写像行列によって全ての下位概念

語を上位概念語へと射影することは難しい．そこで上位・下位概念語間のベクトルの差分 $\mathbf{y}-\mathbf{x}$（オフセットと呼ぶ）をもとに k-means 法でクラスタリングを行い，各クラスタで別々の射影行列を学習する．(3) 式によってクラスタ k の射影行列 Φ_k を学習し Φ_k^* を得る．

$$\Phi_k^* = \arg\min_{\Phi_k} \sum_{(y,x)\in C_k} \|\Phi_k \mathbf{x} - \mathbf{y}\|^2 \qquad (3)$$

ただし $\mathbf{y}$, $\mathbf{x}$ は上位・下位概念語ペア (y,x) のそれぞれの単語に対応するベクトル，C_k はクラスタ k に含まれる上位・下位概念語ペアの集合である．このモデルでは予測した上位概念語ベクトルと正解の上位概念語ベクトルの距離の二乗が小さくなるように学習を行っている．このモデルは単語の意味階層グラフを作成するというタスクで 73.74% の F 値を出している．

しかし，単語ベクトルの学習の際には (1) 式のように内積とバイアスを用いており，単語ベクトルの学習と射影行列の学習で単語間の類似度尺度が異なっているため，モデル内での類似度尺度の一貫性が保たれていない．また，このモデルでは学習データ内の上位・下位概念語ペアのみしか学習しないため，下位概念語を射影した先の近くに正解の上位概念語が位置していても，不正解の上位概念語がそれよりも近くにある場合，不正解の上位概念語を推定してしまう．

2.3 DP-means クラスタリング

DP-means クラスタリング [5] はクラスタ数を自動推定しながらクラスタリングを行う手法であり，ディリクレ過程混合モデルを基に k-means 法を拡張したものである．k-means 法とは異なり，ハイパーパラメータとしてクラスタ数を与えるのではなく，しきい値を与える．しきい値が大きいほどクラスタ数が多くなり，しきい値が小さいほどクラスタ数が少なくなる．以下にアルゴリズムを示す．

DP-means クラスタリング

入力: 入力データ：$x_1,\ldots,x_n$，しきい値：λ
出力: クラスタ：$l_1,\ldots,l_k$，クラスタ数：k

1. 初期化　$k=1$, $l_1=\{x_1,\ldots,x_n\}$，μ_1 を全データの平均値とする．また，すべての $i\ (=1,\ldots,n)$ について $z_i=1$ とする．
2. 収束するまで以下を繰り返す．
   - 各データ x_i について
     - すべての $c\ (=1,\ldots,k)$ について
       $d_{ic}=\|x_i-\mu_c\|^2$ を計算する．
     - $\min_c d_{ic} > \lambda$ の場合，
       $k=k+1, z_i=k, \mu_k=x_i$.
     - それ以外の場合，$z_i=\arg\min_c d_{ic}$.
   - $l_j=\{x_i \mid z_i=j\}$ に基づきクラスタ $l_1,\ldots,l_k$ を生成する．
   - 各クラスタ l_j について，$\mu_j=\frac{1}{|l_j|}\sum_{x\in l_j} x$ を計算する．

3 提案手法

本研究では，2.2節で挙げた Fu らの手法 [3] の問題点を解決するために，単語ベクトルの学習と同様の類似度距離尺度，同様の負例を用いた目的関数を用いて射影行列の更新とクラスタリングを同時に行う手法を提案する．[3] では類似度尺度に二乗距離を用いていたが，単語ベクトルを学習する過程では (1) 式のように単語間の類似度として内積とバイアスを用いているため，これと同様の類似度尺度を用い，モデル内で類似度尺度の一貫性を持たせる．(4) 式に下位概念語 x と上位概念語 y のクラスタ k でのスコア関数 $\mathrm{sim}_k(x,y)$ を示す．

$$\mathrm{sim}_k(x,y) = \sigma(\Phi_k \mathbf{x}\cdot\mathbf{y} + b_k) \qquad (4)$$

ただし，b_k はクラスタ k におけるバイアスである．(4) 式は x を Φ_k で射影したベクトルと y の類似度関数であり，うまく射影できていれば値は大きくなる．(3) 式の類似度尺度を (4) 式に変更した目的関数を (5) 式に示す．

$$\Phi_k^*, b_k = \arg\max_{\Phi_k, b_k} \sum_{(\mathbf{x},\mathbf{y})\in C_k} \log(\ \mathrm{sim}(x,y)\ ) \qquad (5)$$

ここで，(5) 式の右の項は b_k を大きくするほど大きくなるため，解くことが出来ない．これを解決するために，単語ベクトルの学習手法を参考にして，次の (6) 式のように負例の項を追加する．

$$\Phi_k^*, b_k = \arg\max_{\Phi_k, b_k}[\ \log(\mathrm{sim}_k(x,y)) + \sum_{y'\sim C}^{m} \log(1-\mathrm{sim}_k(x,y'))\ ] \qquad (6)$$

y' は推定した上位概念語との類似度が最も高い単語（ただし $y'\neq y$）であり，m は負例の数である．この項の追加により，[3] では考慮できていなかった，誤った上位概念語に近づかないように制約をかけることが可能になる．

さらに，本手法ではクラスタリングを行いながら (6) 式によって最適化を行うことで射影行列の学習とクラ

スタリングを同時に行う．これにより [3] よりも単語ベクトル・クラスタリングに整合した射影行列を得ることができると考える．クラスタリングにはしきい値 λ に合わせてクラスタ数を自動推定することができる DP-means 法を応用した方法を用いる．以下にそのアルゴリズムを示す．

入力：上位下位概念語ペア：$(y_1, x_1), \ldots, (y_n, x_n)$（ただし，$i = 1, \ldots, n$），$\lambda$: しきい値
出力：クラスタ：$l_1, \ldots, l_k$，クラスタ数：k

1. 初期化　$k = 1$, $l_1 = \{(y_1, x_1), \ldots, (y_n, x_n)\}$, Φ_1 を乱数行列とする．また，すべての $i\ (= 1, \ldots, n)$ について $z_i = 1$ とする．
2. 収束するまで以下を繰り返す．
   - 各 (y_i, x_i) について
     - すべての $c\ (= 1, \ldots, k)$ について $\mathrm{sim}_c(y_i, x_i)$ を計算する．
     - $\max_c \mathrm{sim}_c(y_i, x_i) < \lambda$ の場合，$k = k + 1$, $z_i = k$ とし，Φ_k は乱数行列，$b_k = 0$ とする．さらに (6) 式を用いて Φ_k, b_k を (y_i, x_i) について更新する．
     - それ以外の場合，$z_i = \arg\max_c sim_c(y_i, x_i)$ とし，(6) 式を用いて Φ_k, b_k を (y_i, x_i) について更新する．
   - $l_j = \{(y_i, x_i) \mid z_i = j\}$ に基づきクラスタ $l_1, \ldots, l_k$ を生成する．

射影行列およびバイアスの更新においてはオンライン学習アルゴリズムである Adam [6] を用いる．Adam では目的関数の更新量に合わせて学習率を減衰させ，大きなデータに対しても省メモリで高速・高精度な学習を行うことができる．

4　実験

4.1　実験設定

評価データを用いて既存手法と提案手法を評価した．単語ベクトルの学習は ivLBL [4] で行い，その際の学習データは Yahoo! 知恵袋のデータ（約 3,800 万文）を用いた．このときの文脈の領域 n は前後 5 単語とし，出現頻度の上位 100 万単語の単語ベクトルを学習した．また，クラスタリング及び射影行列更新の際の教師データには上位語階層データ [2] を用いた．このデータは表 1のように教師データ・開発データ・評価用データに分割した．ただしそのうち上位・下位概念語の両方の単語ベクトルがモデル内に存在するペアを用いた．評価の指標には Mean Reciprocal Rank (MRR) を用いた．MRR の計算式を次の (7) 式に示す．

$$\mathrm{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathrm{rank}_i} \tag{7}$$

ただし，N は評価データの要素数，rank_i は，i 番目の評価データにおける正解データの順位である．MRR が高いほど上手く推定できていると言える．

事前実験の結果より，クラスタリングのしきい値 λ は 0.075，負例数 m を 1 個とした．Adam のハイパーパラメータは全て文献 [6] の推奨値を用いた．

比較対象のモデルとして k-means 法によりオフセット $\mathbf{y} - \mathbf{x}$ をクラスタリングした後 (3) 式によって最適化をしたモデル（二乗距離＋ k-means）と，同様にオフセットを k-means 法でクラスタリングした後 (6) 式によって最適化をしたモデル（内積距離＋ k-means）を用意し，提案手法（内積距離＋ DP-means）と比較した．

4.2　結果と考察

評価データで評価した結果を表 2に示す．表 2を見ると，二乗距離＋ k-means のモデルよりも内積距離＋ k-means のモデルの方が MRR が高い．これより，距離尺度を内積距離にすることと負例を追加することが正確な射影行列を求めることに寄与していると言える．さらに，内積距離＋ k-means のモデルよりも内積距離＋ DP-means の方が高い MRR を出している．これよ

表1　教師データの分割

データ	ペア数	モデル内に含まれるペア数
教師データ	64,345	8,868
開発データ	8,000	1,074
評価データ	20,000	2,802
合計	92,345	12,744

表2　評価データにおける各モデルの MRR

モデル	MRR
二乗距離＋ k-means	0.213
内積距離＋ k-means	0.294
内積距離＋ DP-means	0.316

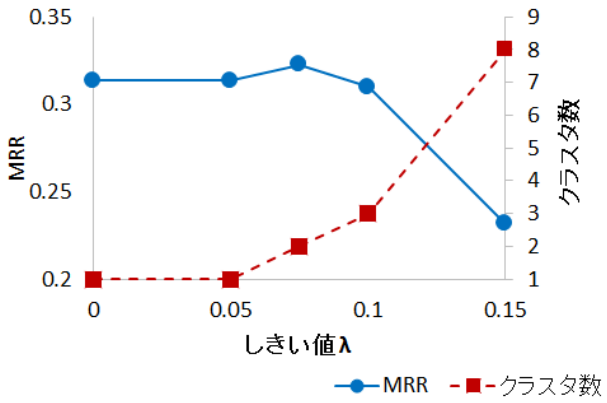

図1　しきい値 λ と MRR およびクラスタ数の関係

り，射影行列の学習と同時にクラスタリングを行う方が，オフセットでクラスタリングするよりも有効であることが分かる．

表 3に (6) 式における負例 y' の選び方と，そのモデルを開発データで評価した時の MRR の関係を示す．比較対象は全単語から無作為に選ぶ方法 (ランダム法) と，推定した上位概念語との類似度が最も高い単語を y' とする方法 (最近傍法) である．ただし，$y' \neq y$ とする．表 3より，負例の選び方は MRR に大きな影響を及ぼすことが分かる．

また図 1に提案手法を開発データで評価した時のしきい値 λ と MRR およびクラスタ数の関係を表すグラフを示す．これを見ると λ が 0.075 で MRR が最大となっており，その時の MRR は 0.323，クラスタ数は 2 であった．この結果から，提案手法ではしきい値は MRR に大きく影響するため，適切に値を決定する必要があることが分かる．今回はクラスタ数 2 が最適となったが，教師データが大きくなればクラスタ数も大きくなり，本手法の効果がより明確になるのではないかと考えられる．

5　おわりに

上位概念語の推定精度向上を目的に，単語ベクトルの学習と同様の類似度距離尺度，同様の負例を用いた目的関数を利用し，射影行列の更新とクラスタリングを同時に行う手法を提案した．結果として既存手法よりも MRR が 0.103 向上し，本手法が既存手法と比べてより正確に上位・下位概念語間の関係を学習できることを示した．今後は，自動獲得したデータ [2] を利用して，より正確に上位・下位概念語間の関係を表現する予定である．

表3　負例の選び方と開発データにおける MRR の関係

モデル	MRR
ランダム法	0.239
最近傍法	0.318

謝辞

本研究ではヤフー株式会社より提供を受けた Yahoo! 知恵袋のデータを利用させて頂いた．関係各位に感謝の意を表する．

参考文献

[1] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63, 2012.

[2] 黒田航, 李在鎬, 野澤元, 村田真樹, and 鳥澤健太郎. 鳥式改の上位語データの人手クリーニング. In 言語処理学会 *15* 回大会発表論文集, pages 76–79, 2009.

[3] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning semantic hierarchies: A continuous vector space approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):461–471, 2015.

[4] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, pages 2265–2273, 2013.

[5] Brian Kulis and Michael Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 513–520, 2012.

[6] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.