単語に対して複数の表現を使用した上位概念の発見

盛野晃平[1] Tad Gonsalves[1]
[1]上智大学　理工学研究科　理工学専攻　情報学領域
k-morino-5m6@eagle.sophia.ac.jp　t-gonsal@sophia.ac.jp

概要

単語同士の関係性を機械学習を用いて見つけることは自然言語処理の分野で重要なタスクである．その中でも本研究では上位概念の発見という，単語をより抽象度の高い単語で表すタスクに取り組む．先行研究では，１つの単語に対して１つの表現を使用して学習を行っていたが提案するモデルでは単語に対して意味を考慮した複数の表現を用いて学習することを提案する．SemEval2018 task9 [5]というコンペティションで提供されているデータで学習と評価を行った．

1 はじめに

自然言語の上位概念・下位概念の関係は言語の意味的な関係においてとても重要なものである．ある単語のより抽象的な意味を表す上位語と，ある単語をより具体的に説明する下位語の関係である．本研究で取り組むタスクは上位概念の発見である．このタスクは，単語が与えられたときに上位概念の候補の単語の中から，その単語の上位概念としてふさわしい単語にランキングをつけて出力することが目的である.本研究では教師あり学習を使用し, CRIM [4]が提案するモデルをもとに，単語に対して意味を考慮した複数の表現を出力できる Adaptive Skip-gram [3] (AdaGram)を使用して学習および精度の検証を行う．

2 関連研究

上位概念の発見にはパターンベースのアプローチと分散表現を使用したアプローチがある．
パターンベースのアプローチでは Hearst [6]が提案した文法のパターンを使用して上位語と下位語の関係を見つけ出す．簡単にできる一方で，計算時間の問題，言語依存，上位語と下位語が同じ文章内で現れる必要があることによる低い再現率といった問題が挙げられる．
２つ目の方法として分散表現を用いたアプローチは，教師あり学習または教師なし学習を使用し，単語の埋め込み表現を使用して関係性を見つけ出すことを目標にしている．Adapt [8]は上位概念の発見のために Skip-gram [9]を用いて単語の埋め込み表現を学習し，得られた単語の埋め込み表現をもとにコサイン類似度を用いて上位概念か否かの判定をする教師なし学習のモデルを提案している．より一般的なのは教師あり学習のモデルである．教師あり学習では単語の埋め込み表現が入力され，正解のペアデータを用いてモデルを学習する．CRIM は教師なし学習と教師あり学習を共に使うことで精度の向上をした．
[2]のモデルでは教師なし学習で上位語の発見のタスクに挑戦している，このモデルではコサイン類似度とコーパスにおける単語の出現の回数を基にしたランキング付けを行うメソッドを基に学習を行っている．
精度を向上させるための別のアプローチとして，単語の埋め込み表現に工夫を加えるものが挙げられる．[1]は BoxE という，エンティティをポイントして埋め込み，基本的な論理的な性質を特徴づけた超長方形のセットに関係性を埋め込む方法を提案した．[10]は box embedding を上位概念の発見で使用するモデルである HyperBox を提案した．

3 提案手法

3.1 前処理

まず，単語の表現のモデルを作成するためにラベル付けされていない文章の単語の前処理を行う．全ての単語の文字を小文字にして，複数の単語で成り立つ言葉の単語間のスペースをアンダースコアに変換し１つの単語として学習を行う．さらに，ノイズになる単語を取り除く．

3.2 単語埋め込み表現のモデル

本研究では, Skip-gram を使用して単語の埋め込み表現を得ている CRIM と異なり，それぞれの単語に対して複数の表現を用いる. そのために AdaGram を使用する．この手法は，1 つの表現で単語のすべての意味を考慮することは難しいのではないかという推測のもと提案され, word similarity のタスクで従来の Skip-gram に比べて良い精度を示した．AdaGram を用いて学習することで意味を考慮した単語埋め込み表現を得ることができる．また，AdaGram は自動的に単語の意味の数を学習することができる．それぞれの単語は意味ごとの表現に従ってインデックスが付与される.

3.3 ペアの選択

それぞれの単語に対して複数の表現を用いるためにはトレーニングのペアの選択を慎重に行う必要がある. 学習で用いるラベル付けされたデータは, 1 つの下位語と 1 つ以上の上位語で成り立っている．それぞれの単語に対してどの単語の表現を使用するかを決める必要があるので，この選択はコサイン類似度をもとに行う．下位語 q が n 個の意味を，上位語 h_i (i=1,2,...k)が m_i 個の意味を持っていた場合に，すべての下位語と上位語の意味を考慮した表現同士のコサイン類似度を計算するので,

$$P = n \times \sum_{i=1}^{k} m_i \qquad (1)$$

P 個の計算結果を比較し，その中でベストな値を示したペアをトレーニングデータとして採用する．この時に，コサイン類似度が 0 より大きいペアのみ採用する．Algorithm1 に実際の選択方法を示す.

図 1 に上記の計算を行うことで選択されるペアの例を示す．Algorithm1 に従って，それぞれの単語には事前学習済みの単語埋め込み表現モデルが出力する意味を考慮した表現の数に従ってインデックスが付与される．インデックスが付与された単語のペアを使用して学習を行う．検証，テスト用のペアは意味を考慮したインデックスが付与されていない．コサイン類似度を用いれば，これらのペアにもインデックスを付与することは可能だが正確性に欠けてしまう．推論の際に出力されるのは意味のインデックスが付与された単語であるので，検証とテストを行う際には，それらのインデックスを取り除く必要がある．学習済みのモデルは，単語ごとに上位語として予測した単語をスコアとともに出力する．このスコアは上位語としてどれだけふさわしいかを表したものである．このスコアをもとにランキングを作成し上位 n 語の単語を取得し評価を行う．評価手法に関しては 4.2 で説明する.

Algorithm 1: Selecting training pairs

Input: queries, candidate hypernyms, num_senses: dictionary having the number of meanings per word
Output: pairs: dictionary for saving pairs

```
for q of queries do
  q_sense_num = number of meanings of q
  for i =1 to q_sense_num do
    q_sense_i = index of q with sense i
    q_embed = embedding of q_sense_i
    for h of candidate hypernyms do
      h_sense_num = number of meanings of h
      for j = 1 to h_sense_num do
        h_sense_j = index of h with sense j
        h_embed = embedding of h_sense_j
        score = cosine similarity of q_embed and
        h_embed
        if score > 0:
          Save q_embed, h_embed and score
        end
      end
    end
    Add a pair that has best score to the pairs
  end
end
```

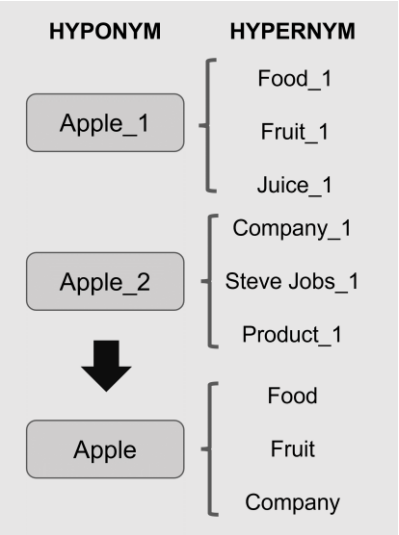

図 1　意味に応じてインデックスが付与されたペア

3.4 上位概念発見のモデル

上位概念の関係を見つけるためのモデルは CRIM が提案したモデルをもとに作成する．学習は Projection learning [11]という単語の埋め込み表現の射影ベクトルを使用した方法で行う．下位概念の単語の埋め込み表現の射影ベクトルと上位語の候補の単語の埋め込み表現のベクトルの距離を計算しスコアを出力する．上位語のランキングを作成するときには，出力されたこのスコアが高いものから順に上位語としてふさわしいと判断をした．活性化関数には sigmoid 関数を使用し，損失関数は Binary cross entropy を使用した．Negative sampling を用いて学習を行うので，出力されたスコアを正解のデータでは 1 に不正解のデータを 0 に近づけるように学習を行った．学習の際のロスは正例と負例におけるロスを合計したものを使用した．最適化には Adam [7]を使用した．図２にモデルの学習手順を示す．

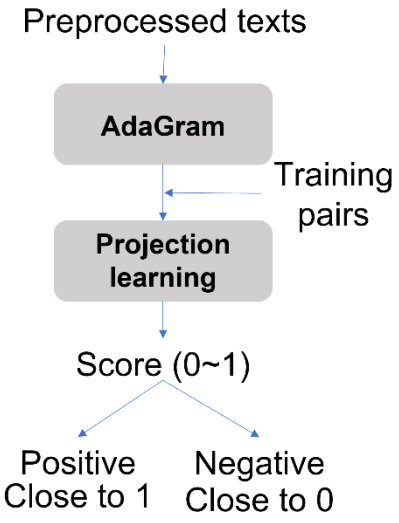

図２　トレーニング手順

4 実験

4.1 データセット

本研究では，SemEval2018 task9 で提供されている英語のデータセット（1A，2A，2B）を用いて実験を行った．それぞれのデータセットはラベル付けをされていないテキストデータと上位語・下位語のペアのデータで構成されている. 2A のデータは医療の領域，2B のデータは音楽の領域の単語のペアを持つ. 1A のデータが 1 番多く, 学習とテストデータは 1,500 個あり，2A と 2B は 500 個ある．それぞれ 1 つ以上の上位概念・下位概念のペアを持つ．1A のデータの検証用のデータ数が 50 個あるのに対して 2A と 2B の検証用のデータの数は 15 語しかないので学習の精度に影響を及ぼす可能性がある．単語の埋め込み表現のモデルはそれぞれのデータセットのテキストデータを基に作成した．

4.2 評価指標

提案するモデルは最大で上位 15 語の予測結果を３つの指標を用いて評価した．SemEval 2018 task 9 で使われている評価指標と同じものを使用した．
１つ目は，Mean Average Precision(MAP)である．MAP を提案するモデルの評価の主要な指標として用いた．Average Precision は，その順位までの正解率を特定のデータにおいて平均を求めたものである．この指標の平均が MAP にあたる．

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{q\in Q} AP(q) \qquad (2)$$

２つ目は，Mean Reciprocal Rank(MRR)である．MRR は最初に現れた正解データの順位の逆数の平均をとったものである．

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \qquad (3)$$

３つ目が，P@1 である．P@n (n は自然数)は上位 n 個における Precision である．HyperBox と精度の比較を行う際には P@5 を使用した．

4.3 実験手順

まず，3.1 で説明した前処理を行った．次に，単語をベクトル表現するためのモデル(Emb model)を AdaGram を用いて作成した．このモデルが出力する単語埋め込み表現の次元は CRIM が用いている単語埋め込み表現の次元に合わせて 300 である. 加えて, Emb model はコーパスで現れる回数が 5 回未満の単語を無視した．さらに，そのモデルにおいて単語がどれくらいの数の意味を持っているかは，AdaGram の論文で採用されている閾値 10^{-3} を用いて定めた．作成された Emb model をもとに 3.3 で説明されている手順に従ってトレーニング用の上位概念・下位概念のペアを作成した．学習時に使用する不正解のデータは，その単語の上位概念の正解データに含まれない他のペアの上位語をランダムに取得し使用した．

CRIM のモデルは結果の評価をする際にはデータの種類に関わらず, 1A のデータのみでパラメータのチューニングしたモデルを使用していたので，それに倣って 1A のデータのみでチューニングしたパラメータを 2A，2B のデータセットでのトレーニングにも用いた．トレーニング中で最も高い MAP を得ることができたモデルをテストに用いた．テストでは先ほど説明した評価指標をもとに CRIM との精度の比較を行った.

5 実験結果

5.1 英語のデータにおける実験結果

表 1 に英語のデータにおける実験の結果を示す. CRIM の精度は提案するモデルと公平に評価するために，教師あり学習でパターンを基にしたアプローチを使用していないものを用いた. CRIM では Hearst が提案した上位概念・下位概念の関係を見つけることができる文法のパターンを用いて上位語の発見を行っている. 提案するモデルは CRIM と比べると 2A データセットの MAP の結果を除き他の全ての指標で精度を上回った．特に，MRR と P@1 の指標で大幅に精度を改善することができた.
Algorithm1 で説明をしたコサイン類似度が 0 より大きいペアのみ選択する処理を行わないと 1A のデータセットにおいて MAP が 35.96%, MRR が 18.77%, P@1 が 29.73%となった.

表 1　実験結果

		CRIM	Ours
1A	MAP	19.11	**19.25**
	MRR	34.99	**37.28**
	P@1	28.8	**31.07**
2A	MAP	**28.51**	25.68
	MRR	37.63	**43.41**
	P@1	34.4	**36**
2B	MAP	39.95	**42.68**
	MRR	57.34	**63.7**
	P@1	43	**52.6**

5.2 チューニングを行った際の実験結果

CRIM との結果を比較したが，提案するモデルは 2A, 2B のデータに対してはパラメータのチューニングを行っていないので，それぞれのデータセットに対してチューニングを行えばより高い結果が得られると推測した．結果の比較の際には，2A と 2B のデータセットで CRIM に比べて高い精度を示している HyperBox を使用する．提案するモデルでは 2A に対してパラメータのチューニングを行ったときには精度の改善が見られたが，２B のデータで行った際には精度の改善ができなかったので 2B の実験結果に関しては表１と同じ結果を使用する．表２に結果の比較を示す．全ての項目で提案するモデルは HyperBox の精度を上回った．特に MRR のスコアに関しては大幅に精度を上回った．英語のデータでの実験の際にも同じように MRR の大幅な改善が見られた．このことから提案するモデルは，上位語の単語をより高い順位で出力することができたことがわかった.

表２　チューニング後の実験結果

		HyperBox	Ours
2A	MAP	27.79	**29.16**
	MRR	43.71	**53.41**
	P@5	30.22	**30.8**
2B	MAP	41.39	**42.68**
	MRR	58.15	**63.7**
	P@5	43.13	**43.6**

6 まとめ

本研究では，上位概念の発見のタスクにおいて，１つの単語に対して意味を考慮した複数の単語の表現を用いることで予測の精度を改善できることを示した．さらに，モデルを構築するためのトレーニングペアを選択する方法を説明した．提案するモデルは，意味を考慮した単語の表現を得るために人間が構築した辞書などを必要としないので，学習用の上位概念・下位概念のペアがあれば容易に適応することができる.

参照文献

1. ABBOUD, Ralph, et al. Boxe: A box embedding model for knowledge base completion. *Advances in Neural Information Processing Systems*, 2020, 33: 9649-9661.
2. ATZORI, Maurizio; BALLOCCU, Simone. Fully-unsupervised embeddings-based hypernym discovery. *Information*, 2020, 11.5: 268.

3. BARTUNOV, Sergey, et al. Breaking sticks and ambiguities with adaptive skip-gram. In: *artificial intelligence and statistics*. PMLR, 2016. p. 130-138.
4. BERNIER-COLBORNE, Gabriel; BARRIERE, Caroline. Crim at semeval-2018 task 9: A hybrid approach to hypernym discovery. *Proceedings of the 12th international workshop on semantic evaluation*. 2018. p. 725-731.
5. CAMACHO-COLLADOS, Jose, et al. SemEval-2018 task 9: Hypernym discovery. *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018); 2018 Jun 5-6; New Orleans, LA. Stroudsburg (PA): ACL; 2018. p. 712–24.* ACL (Association for Computational Linguistics), 2018.
6. HEARST, Marti A. Automatic acquisition of hyponyms from large text corpora. *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. 1992.
7. KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
8. MALDONADO, Alfredo; KLUBIČKA, Filip. Adapt at semeval-2018 task 9: Skip-gram word embeddings for unsupervised hypernym discovery in specialised corpora. *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018. p. 924-927.
9. MIKOLOV, Tomas, et al. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013, 26.
10. PARMAR, Maulik, et al. HyperBox: A Supervised Approach for Hypernym Discovery using Box Embeddings. *arXiv preprint arXiv:2204.02058*, 2022.
11. YAMANE, Josuke, et al. Distributional hypernym generation by jointly learning clusters and projections. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016. p. 1871-1879.