単語ペアと依存構造パスの共起モデリングを用いた語の意味関係の分類

鷲尾 光樹* 加藤 恒昭†
東京大学大学院総合文化研究科
kokiwashio@g.ecc.u-tokyo.ac.jp* ; kato@boz.c.u-tokyo.ac.jp†

1 はじめに

単語間の意味関係は、含意関係認識のような高度な意味処理を要するタスクにおいて重要な情報である。WordNet などの人手による知識ベースはこの種の情報を提供するが、その拡張にはコストが掛かり、カバーされているドメインも限られている。この問題を解決するために、コーパスから語の意味関係を自動的に識別し、獲得する方法が研究されてきた。

語の意味関係の識別タスクでは、対象の単語ペアが共起した文中において二語を結びつける語の系列あるいは依存構造のパスが、識別に有用であることが先行研究において示されている [2]。たとえば、*A dog is a kind of animals.* という文において、*is a kind of* という語の系列から、*dog* が *animal* の下位語であることが推測できる。近年では、このような二語を結びつける語の系列や依存構造のパスを素性として教師あり学習を行う方法が、**Path ベース**の手法として確立されている [9][8][7]。

Path ベースの意味関係識別は対象となる任意の単語ペアについてコーパス上での共起を必要とする。しかし、たとえ大規模コーパスであったとしても、意味関係を持つ語が必ずしも共起するとは限らない。そのような、文中で共起しなかった単語ペアについては、このアプローチでは手がかりが得られず、適切に意味関係を識別することができない。

この問題に対処するために、本研究ではニューラルネットワークを用いて、単語ペア w_1, w_2 とそれらを結びつける依存構造パス $path$ の共起確率 $P(path|w_1, w_2)$ のモデルを教師なし学習する手法を提案する。学習済みの $P(path|w_1, w_2)$ のニューラルネットワークモデルにより、コーパス上で共起しなかった単語ペアについても、二語を結びつける尤もらしい依存構造パスを予測することができる。本研究では、学習済みの $P(path|w_1, w_2)$ のモデルを以下の二通りの方法で用いる。

- 二語を結びつける依存構造パスを予測し、分類に用いる依存構造パスのデータベースを拡張する。
- ニューラルネットワークの隠れ層を、二語（単語ペア）の分散表現として分類に用いる。

実験により、提案手法を先行研究のニューラルネットワークを用いた Path ベースの手法に適用した場合、性能が向上することを示す。さらに、分析により、提案手法が適切に問題を緩和できていることを示す。

2 背景

2.1 Path ベースの語の意味関係識別

初期の Path ベースの手法として、Hearst は意味関係を示唆すると考えられるパターンについてコーパス上を検索することで、そのパターンで現れた単語ペアの意味関係を識別した [2]。Snow らは、コーパス上で二語を結びつけた依存構造パスを用いて単語ペアの特徴ベクトルを作り、教師あり学習を行った [9]。

Shwartz らは Snow らの手法における疎な特徴空間を避けるために、Recurrent Neural Network(RNN) を用いた手法 **LexNET** を提案した [8][7]。LexNET は、単語ペアをコーパス上で結びつける依存構造パスを系列とみなし、RNN を用いて低次元で密なベクトルに変換する。たとえば、*A dog is a mammal.* という文がに関して、$X = dog, Y = mammal$ としたとき、二語を結びつける依存構造パスは、`X/NOUN/nsubj/> be/VERB/ROOT/- Y/NOUN/attr/<`である。依存構造パスの各エッジは、レンマ、品詞、依存構造ラベル、依存方向の分散表現の結合で表現され、それが RNN の入力ベクトルとなる。単語ペアは、共起した依存構造パスのベクトルの平均として表現された後に分類される。Snow らのように、それぞれの依存構造パスを別々の素性として用いるのではなく、RNN で分散表

現に変換することで、疎な特徴空間を回避し、識別に有効な依存構造パスを一般化した形で捉えることができる。Shwartz らは実験によって、この手法が既存の Path ベースの手法を性能で上回ることを示した。さらに、依存構造パスベクトルの平均に、語の分散表現を結合して分類を行うと性能が向上することから、語の分散表現の結合を分類に用いる個別の単語の特徴に依拠した手法とは相補的な二語の関係の情報を、Path ベースの手法が捉えていることを示した。

2.2 問題点

Path ベースの手法は、コーパス上で二語を結びつける語の系列あるいは依存構造パスを通して、語の関係を捉えることができるが、対象の単語ペアについて、コーパス上での共起を必要とする。語の頻度は Zipf の法則に従うことが知られており、大抵の内容語の出現は低頻度であるため、たとえ単語ペアが何らかの意味関係を有していたとしてもコーパス上で共起するとは限らない。共起しなかった単語ペアに関しては、Path ベースの手法は適切に分類を行うことができない。この疎データ問題を、本論文では**パス欠落問題**と呼称する。パス欠落問題の解決は、Path ベースの手法において重要な課題である。

Necsulescu らはコーパスをグラフで表現することでこのパスの欠落問題の緩和を試みた [4]。語と様々な語との依存関係をグラフで表現し、それらを重ね合わせることで、コーパス上で共起しなかった二語についても、グラフ上でのパスを得られる。彼らはそのようなパスについて素性選択をし分類を行った。結果として Recall が大幅に改善されたが、依然としてパス欠落問題による誤分類があることを報告している。

本研究ではこの問題を、単語ペアと二語を結びつける依存構造パスの共起をニューラルネットワークによってモデリングすることで解決を試みる。

3 提案手法

3.1 $P(path|w_1, w_2)$ のモデリング

本節では、$P(path|w_1, w_2)$ をニューラルネットワークによってモデリングする手法について述べる。単語ペアと二語を結びつける依存構造パスのコーパス上での共起について、ニューラルネットワークを用いてモデリングを行うことで、コーパス上で同一文中に出現しなかった単語ペアについても、二語を結びつける尤もらしい依存構造パスを予測することができる。さらに、Path ベースの教師あり学習においては、対象の単語ペアが共起した依存構造パスだけ用いるため、コーパスの一部しか学習に使用できないが、本手法を適用することで、コーパス全体を学習に用いることができる。

提案手法においては、単語ペアと依存構造パスをそれぞれ分散表現として表現し、それらの内積から共起する程度を予測することで、コーパスを用いた教師なし学習を行う。単語ペア (w_1, w_2) は以下のように分散表現化する。

$$\boldsymbol{h}_{(w1,w2)} = tanh(\boldsymbol{W}_1[\boldsymbol{v}_{w_1}; \boldsymbol{v}_{w_2}] + \boldsymbol{b}_1) \quad (1)$$

$$\tilde{\boldsymbol{h}}_{(w_1,w_2)} = tanh(\boldsymbol{W}_2\boldsymbol{h}_{(w1,w2)} + \boldsymbol{b}_2) \quad (2)$$

ただし、$[\boldsymbol{v}_{w_1}; \boldsymbol{v}_{w_2}]$ は w_1 と w_2 の分散表現の結合、$\boldsymbol{W}_1$, $\boldsymbol{b}_1$, $\boldsymbol{W}_2$, $\boldsymbol{b}_2$ は線形変換の行列とバイアスベクトルである。$\tilde{\boldsymbol{h}}_{(w_1,w_2)}$ が単語ペアの分散表現となる。依存構造パス $path$ の分散表現は、$\tilde{\boldsymbol{h}}_{(w_1,w_2)}$ と同次元のベクトル $\boldsymbol{v}_{path}$ をランダムに初期化し、割り当てる。依存構造パスをベクトル化する際に RNN を用いることもできるが、今回は学習時間の短縮のために行わなかった。

目的関数には、学習の効率化のために Negative sampling 目的関数 [3] を用いた。

$$L = \sum_{(w_1,w_2,path)\in D} \log \sigma(\boldsymbol{v}_{path} \cdot \tilde{\boldsymbol{h}}_{(w_1,w_2)}) + \sum_{(w_1,w_2,path')\in D'} \log \sigma(-\boldsymbol{v}_{path'} \cdot \tilde{\boldsymbol{h}}_{(w_1,w_2)}) \quad (3)$$

ただし、σ はシグモイド関数、$(w_1, w_2, path) \in D$ はコーパスから抽出した単語ペアと依存構造パスの共起の集合であり、$(w_1, w_2, path') \in D'$ は D からランダムに生成された負例サンプルである。

3.2 語の意味関係識別への適用

依存構造パスのデータ拡張 (Add)

提案したモデルをコーパスから学習すると、$\sigma(\boldsymbol{v}_{path}, \tilde{\boldsymbol{h}}_{(w_1,w_2)})$ によって、単語ペアと依存構造パスの共起の尤もらしさのスコアを計算できる。これにより、単語ペアとの共起が尤もらしい依存構造パスを教師あり学習の際に用いることができる。本研究では、$(X = w_1, Y = w_2)$ とした時の最もスコアが高い上位 k 個の依存構造パスと、$(X = w_2, Y = w_1)$ とした時の上位 k 個のパス[1]を、w_1 と w_2 が共起した依存構造パスに追加する。k はハイパーパラメタである。この手法を **Add** と呼ぶ。

単語ペアの分散表現抽出 (Rep)

$\tilde{\boldsymbol{h}}$ は、コーパス上で共起する依存構造パスの情報を捉えた単語ペアの表現として用いることができる。以

[1] このとき、予測されたパスの X と Y を入れ替える。

データセット	事例数	パスが得られた事例数	割合
K&H+N	57509	8866	15.4%
BLESS	14558	8775	60.3%
ROOT09	8602	6582	76.5%
EVALution	3240	3199	98.7%

表 1: 訓練データごとの事例数とパスが得られた事例数の割合

下のベクトル $\boldsymbol{v}_{pair}$ を単語ペア (w_1, w_2) の分散表現として用いる。

$$\boldsymbol{v}_{pair} = [\tilde{\boldsymbol{h}}_{(w_1,w_2)}; \tilde{\boldsymbol{h}}_{(w_2,w_1)}] \quad (4)$$

この単語ペアの分散表現を、教師あり学習の際に用いる手法を **Rep** と呼ぶ。

4 実験

提案手法の評価のために、Add と Rep を LexNET に適用した場合の性能を確かめた。データセットには、K&H+N[4]、BLESS[1]、EVALution[6]、ROOT9[5] から、名詞ペアのみを抽出して用いた[2]。これらのデータセットでは単語ペアに数種類の意味関係が注釈されており、訓練データを用いて多クラス分類器を訓練し、テストデータで性能を評価する。評価指標は F1 スコアである[3]。

D を得るために、英語版 Wikipedia を spaCy[4]で依存構造解析し、名詞の単語ペアと依存構造パスの共起を抽出した。表 1 は、訓練データごとの事例数、パスが得られた事例数、その割合である。

ベースラインである LexNET は、文献 [7] に従い、以下のように実装した。依存構造パスをエンコードする RNN には、入力層 60 次元、隠れ層 60 次元の 2 層 LSTM を用いた。入力ベクトルは、50 次元の単語（レンマ）ベクトル、4 次元の品詞ベクトル、5 次元の依存構造ラベルベクトル、1 次元の依存方向ベクトルの結合である。また、同じく先行研究 [7] で用いられた LexNET_h も比較に用いた。これは、LexNET の出力層とその一つ手前の層の間に、隠れ層を一つ追加することで、依存構造パスと各単語の特徴の相互作用を捉えようとしたモデルである。LexNET_h の追加の隠れ層の次元は 60 次元とした。LSTM に用いる単語ベクトルと、出力層の一つ手前の層に結合する単語ベクトルには、訓練済みの GloVe を用いた[5]。ミニバッチ数は 100 とし、Adam(学習率 0.01) を用いて訓練を行い、入力ベクトルの各コンポーネントのドロップアウト率をハイパーパラメータチューニングした。

$P(path|w_1, w_2)$ のモデリングに関しては、$\boldsymbol{v}_{w_1}, \boldsymbol{v}_{w_2}$ に同じ GloVe を用い、各隠れ層の次元は 100 次元とした。負例サンプル数は 5 とし、学習率 0.001 の Adam を用いてモデルを 5 エポック訓練した。

Add を適用する際は $k \in \{1, 3, 5\}$ をハイパーパラメータチューニングした。Rep を適用する際は LexNET の出力層の手前の層に、$\boldsymbol{v}_{pair}$ を結合した。このとき、純粋な教師なし学習による性能向上を見るために、このコンポーネントは教師あり学習中には更新しないようにした。

[2]評価のためのデータの分割は、文献 [7] で提供されているものを用いた。

[3]先行研究 [7] にならい、F1 スコアの計算には、scikit-learn の average セッティングを用いた

[4]https://spacy.io

[5]https://nlp.stanford.edu/projects/glove/

手法	K&H+N	BLESS	ROOT09	EVALution
LexNET	0.969	0.922	0.776	0.539
LexNET_h	0.968	0.927	0.810	0.540
LexNET+Add	**0.970**	0.927	0.806	0.545
LexNET+Rep	**0.970**	**0.944**	**0.832**	0.565
LexNET+Add+Rep	0.969	0.942	0.820	**0.567**

表 2: 性能評価（F1 スコア）

結果

結果を表 2 に示す。提案手法を適用した場合、LexNET の性能が向上することがわかる。また、各データセットでの最高性能は、提案手法を用いた場合に達成されており、提案手法が有効であることを示している。

さらに、LexNET+Rep は LexNET_h よりも教師あり学習時に調整されるパラメータが少ないにも関わらず、性能が上回っている。この結果は、$\boldsymbol{v}_{pair}$ が LexNET_h における単語ベクトル間の相互作用以上に有用な特徴を提供していることを示している。これは、教師なし学習によって、$\boldsymbol{v}_{pair}$ が共起した依存構造パスの情報を適切に捉えているからだと考えられる。

Add と Rep はそれぞれ LexNET の性能を向上させるが、同時に適用した場合、必ずしも性能がさらに向上するわけではなかった。これは両方の手法が同じモデルに基づいているため、余剰な情報が最適化に悪影響を及ぼしているためであると思われる。

5 分析

本節では、Add によって予測される依存構造パスと、Rep が提供する単語ペアの分散表現がどのような性質を捉えているかを分析する。

Add に関しては、コーパスで共起が観測できなかったペアについて、提案モデルが予測したパスを検証した。結果として、予測されるパスにおいては不適切なものを含みつつも、各単語ペアの意味関係を示

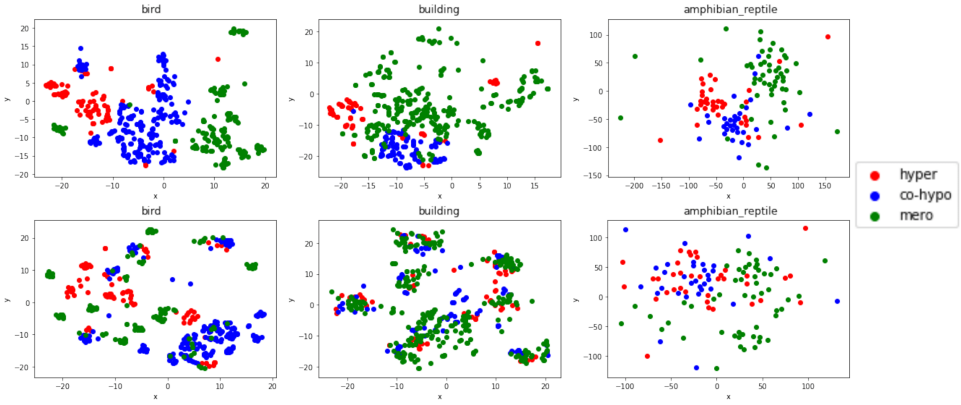

図 1: t-SNE による単語ペア分散表現の可視化 (上段は $\boldsymbol{v}_{pair}$、 下段は $[\boldsymbol{v}_{w_1}; \boldsymbol{v}_{w_2}]$)

唆する依存構造パスが予測されていた。たとえば、$X = jacket, Y = commodity$ という上位下位関係のペアにおいては、`X/NOUN/nsubj/> be/VERB/ROOT/-Y/NOUN/attr/< manufacture/VERB/acl/<`という is-a の依存構造パスが高いスコアで予測された。また、$X = owl, Y = rump$ という部分全体関係のペアにおいては、`X/NOUN/nsubj/> have/VERB/ROOT/-Y/NOUN/dobj/< be/VERB/relcl/>`という、所有の関係を示唆する依存構造パスが予測された。

次に、Rep によって得られる $\boldsymbol{v}_{pair}$ の性質を調べるために、BLESS 内の単語ペアに注釈されたドメインごとに、t-SNE によるデータ点の可視化を行った。比較のため、各単語ベクトルの結合も同時に可視化を行った。結果として、いくつかのドメインに関して、$\boldsymbol{v}_{pair}$ の空間においては、教師あり学習なしで、各意味関係のクラスタが形成されていることがわかった。図 1 に例を示す。この図においては、上位下位関係 (hypernymy)、共通の上位語を持つ関係 (co-hyponymy)、部分全体関係 (meronymy) のデータ点がプロットされている。図を見ると、単語ベクトルの結合のプロットは、各意味関係の点が散らばっているか、混ざり合っているかのどちらかであるが、$\boldsymbol{v}_{pair}$ のプロットは、各意味関係ごとにクラスタを形成している。これは、$\boldsymbol{v}_{pair}$ が語の関係の類似性 [10] を捉えており、語の意味関係の分類に望ましい性質を有していることを示している。

以上の分析は、提案手法が適切にパス欠落問題を緩和できることを示している。

6 結論

本研究では Path ベースの語の意味関係識別におけるパス欠落問題を解決するために、単語ペアと依存構造パスの共起をニューラルネットワークによってモデリングする手法を提案した。実験により、提案手法は先行研究の手法の性能を向上させることを示した。今後は、提案したモデルを語の関係の類似性が関わる様々なタスクに適用していく。

本研究は JSPS 科研費#17H01831 および#15K12873 の助成を受けた。

参考文献

[1] Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *GEMS*, 2011.

[2] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING*, 1992.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[4] Silvia Necsulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In **SEM*, 2015.

[5] Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. Nine features in a random forest to learn taxonomical semantic relations. In *LREC*, 2016.

[6] Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *LDL-2015*, 2015.

[7] Vered Shwartz and Ido Dagan. Path-based vs. distributional information in recognizing lexical semantic relations. In *CogALex-V*, 2016.

[8] Vered Shwartz, Yoav Goldberg, and Ido Dagan. Improving hypernymy detection with an integrated path-based and distributional method. In *ACL*, 2016.

[9] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*, 2004.

[10] Peter D. Turney. Similarity of semantic relations. *Computational Linguistics*, 2006.