

語義の箱埋め込み学習とその応用

小田 康平 白井 清昭

北陸先端科学技術大学院大学 先端科学技術研究科
{s2210042,kshirai}@jaist.ac.jp

概要

語義曖昧性解消の先行研究では語義の埋め込みを単一のベクトルで表現していたのに対し、本研究は語義の箱埋め込みを学習する手法を提案する。箱埋め込みの学習により、語義間の上位下位関係の推定や新語義の判定が可能となる。Prototypical Networksに基づく MetricsWSD を語義の箱埋め込みを学習するように拡張し、モデル学習に必要なエピソードを作成する2つの戦略を提案する。評価実験を行い、語義曖昧性解消、新語義の判定、新語義の上位語義の推定という3つのタスクについて、提案手法の有効性を確認した。

1 はじめに

語義曖昧性解消 (Word Sense Disambiguation; WSD) とは、文脈における単語の意味 (語義) を特定するタスクである [1]。従来の WSD の研究のほとんどは、訓練データに出現する語義の中からテストデータにおける単語の語義を選択することを想定している。しかし、単語の語義は日々変化し、新しい語義も新しく生まれている。そのため、単にあらかじめ定義された語義の中から該当する語義を選ぶのではなく、対象単語が訓練データに存在しない未知の語義 (あるいは新しい語義) であるのかを判定することが求められる。

従来の WSD に関する研究では分類モデルを学習する際に語義をベクトルで表現していたのに対し、本研究では語義の箱埋め込みを学習することを目的とする。語義の箱埋め込みとは、ベクトル空間における領域によって表された語義の抽象表現である。語義を箱埋め込みで表現することにより、箱埋め込みが表す領域の重なりによって語義間の上位下位関係を自然に表現できる。また、対象単語を含む用例が与えられたとき、その対象単語の語義が新語義なのかを判定したり、新語義の上位語義を推定するこ

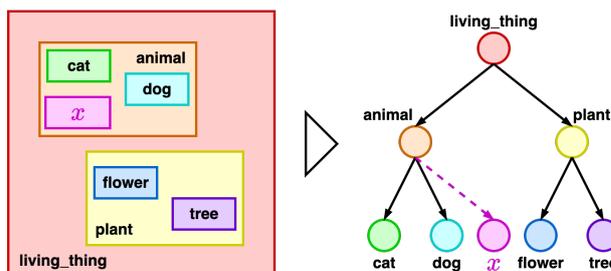


図1 語義の箱埋め込みによる上位語義の推定

とができる。図1の例では、 x の箱埋め込みが cat や dog といった他の語義と領域が重ならないためにこれらとは異なる語義であること、 $animal$ の領域に包含されていることから上位語義は $animal$ であることが推定できる。このように、語義の箱埋め込みは、新語義の自動検出や WordNet[2] などの概念辞書の自動拡張に応用できる。

本研究では既存の WSD 手法である MetricWSD[3] を語義の箱埋め込みを学習するよう拡張する。さらに、モデルの損失を計算するために用いる小さいデータセットを作成する2通りの方法を提案する。提案手法を WSD、新語義の判定、新語義の上位語義の推定の3つのタスクに適用する実験を行い、語義をベクトルで表現する従来手法と比較する。

2 関連研究

2.1 WSD

最近の WSD の研究には、語義の定義文を用いる手法 [4, 5, 6]、語義の定義文に加え上位下位関係といった意味関係を用いる手法 [7, 8] などがある。Bevilacqua らは、対象単語を含む用例から BART[9] によって語義の定義文を生成し、これと辞書の語義の定義文との類似度を測ることで語義の曖昧性を解消する手法を提案している [10]。また、この研究では新語義を判定することも目的としている。しかし、これらの手法では、訓練データにお

ける低頻度語義に対する WSD の性能が低いことが問題とされている [11]. Chen らは, Prototypical Networks[12] とよばれるメタ学習手法を用いて WSD を行う MetricWSD を提案し, 低頻度語義に対する WSD の性能を向上させた [3].

2.2 タクソノミ拡張

上位下位関係からなる知識グラフはタクソノミと呼ばれる. タクソノミ拡張 (Taxonomy Expansion) とは, 現在のタクソノミに含まれない単語と文脈が与えられたとき, その単語がタクソノミにおけるどの単語を親にもつかを予測するタスクである [13]. タクソノミ拡張の代表的な手法として, グラフニューラルネットワークに基づく手法 [14], 対象ノードと根ノードとの最短パスを学習する手法 [15] などがある. Jiang らは箱埋め込みを用いてタクソノミ拡張を行う手法を提案している [16]. 本研究でも箱埋め込みを用いるが, 単語ではなく語義の箱埋め込みを学習する点, MetricWSD を拡張して箱埋め込みを学習する点が異なる.

3 提案手法

3.1 MetricWSD

Prototypical Networks[12] は分類問題を解くモデルを学習するためのメタ学習手法である. 図 2 にその概要を示す. f_θ は訓練データをベクトル表現に変換するモデルである. 各分類クラスの訓練データをサポートセットとクエリセットに分割し, サポートセットのベクトルの平均から各分類クラスのプロトタイプ表現を獲得する. プロトタイプ表現とクエリセットのベクトル表現との距離を損失とし, それが最小になるようにモデル f_θ のパラメータを更新する.

MetricWSD[3] は Prototypical Networks を WSD に適用した手法である. 分類クラスは語義であり, プロトタイプ表現が語義のベクトルに該当する. 訓練データは語義を含む文 (用例) の集合であり, 用例をベクトル表現に変換するモデル f_θ として Bidirectional Encoder Representations from Transformers (BERT)[17] が用いられている.

3.2 語義の箱埋め込みの学習

本研究では, 先行研究 [16, 18] に倣い, 語義の箱埋め込みを $\mathbf{b} = (\mathbf{c}, \mathbf{o})$ と表現する. \mathbf{c} は中心を表すベクトル, \mathbf{o} は箱の辺の長さを表すベクトルであり,

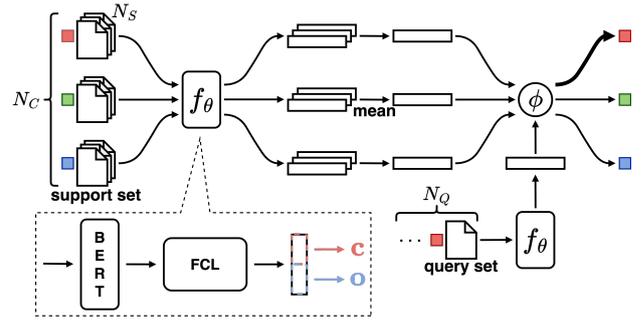


図 2 提案手法の概要

両者の次元数は同一である.

箱埋め込みを学習するために, 図 2 の点線の枠内に示すように, MetricWSD における f_θ の役割を担う BERT の最終層の後に全結合層 (Fully Connected Layer; FCL) をつなげる. この層の入力は BERT の対象単語に対する埋め込み, 出力は \mathbf{b} である. \mathbf{b} は半分に分割され, それぞれ \mathbf{c} と \mathbf{o} となる.

あるアイテム x がアイテム y を包含する確率は, アイテム x の箱埋め込み \mathbf{b}_x とアイテム y の箱埋め込み \mathbf{b}_y を用いて式 (1) のように計算される.

$$P(\mathbf{b}_y|\mathbf{b}_x) = \frac{\text{Vol}(\mathbf{b}_x \cap \mathbf{b}_y)}{\text{Vol}(\mathbf{b}_x)} \quad (1)$$

Vol は箱の体積を計算する関数である. $\text{Vol}(\mathbf{b}_x \cap \mathbf{b}_y)$ は 2 つの箱が重なっている部分の体積を意味する. ただし, Vol のハードな定義では \mathbf{b}_x と \mathbf{b}_y が重なっていない場合に値が 0 となり, 勾配がなくなるため, ソフトな定義を用いる必要がある. 本研究では, 先行研究 [16, 18] と同じく, Gumbel Box[19] によって関数 Vol を定義する. Gumbel Box では, 2 つの箱が重なっている部分の体積を Gumbel 分布に基づき計算する.

モデルを学習する際には, クエリ表現と各プロトタイプ表現のペア $(\mathbf{b}_x, \mathbf{b}_y)$ ごとに損失を計算する. 損失は, 先行研究 [18] と同じく, 式 (2) に示す binary cross-entropy とする.

$$L = -\delta \cdot \log P(\mathbf{b}_y|\mathbf{b}_x) - (1 - \delta) \cdot \log(1 - P(\mathbf{b}_y|\mathbf{b}_x)) \quad (2)$$

ここで, δ は正例の場合に 1, 負例の場合に 0 の値を取る変数である. ここでの正例とは, 図 3 に示すように, 語義 y が語義 x と同じもしくは語義 x の上位概念 (祖先) である場合, 負例とはそれ以外の場合を指す. 直観的には, ある語義の箱埋め込みは, 自身もしくは上位概念の語義との箱埋め込みが重なるように, それ以外とは重ならないように学習される.

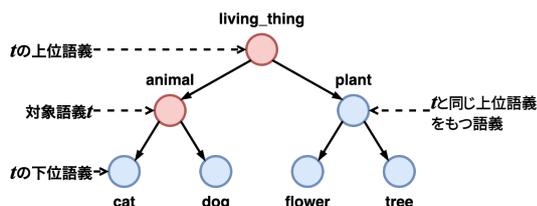


図3 損失計算における正例と負例 (赤が正例, 青が負例)

3.3 エピソードの作成

一般に損失を計算するために分割された少数のサンプルからなるデータセットはバッチ (Prototypical Networks のようなメタ学習ではエピソード) と呼ばれる。MetricWSD では, 1つの単語に対して1つのエピソードを用意し, 対象単語を含む用例によってエピソードを作成する。本研究では語義間の関係を学習するため, 1つの語義に対して1つのエピソードを用意する。語義 t に対するエピソードの作成手順を以下に示す。

1. N_C 個の語義を選択する。正例と負例を含む。
2. それぞれの語義について, その語義の用例をランダムに N_S 個選択し, サポートセットとする。
3. サポートセットに含まれない対象語義 t の用例の中からランダムに N_Q 個の用例を選択し, クエリーセットとする。

上記の手続きのステップ1において, N_C 個の語義を選択する2つの戦略を提案する。

戦略 S_r t と t の (直近の) 上位語義を正例として選び, 残りの語義をランダムにサンプリングする。 t の祖先に該当する語義が選ばれたときは正例, それ以外は負例となる。

戦略 S_n t と t の上位語義を正例, t の下位語義ならびに t と同じ上位語義をもつ語義 (兄弟関係にある語義) を負例として選ぶ。 S_r と同様に残りの語義はランダムにサンプリングする。

戦略 S_n では, ある語義の箱埋め込みが, 特に下位語義や兄弟関係にある語義の箱埋め込みと重ならないことを重視している。

4 語義の箱埋め込みの応用

4.1 WSD

ある用例に出現する対象単語 w の語義を決定する。訓練データから作成された対象単語 w の k 個の語義のプロトタイプ表現を $\mathcal{B}_w = \{\mathbf{b}_1^p, \dots, \mathbf{b}_k^p\}$ とす

る。テストデータの用例の箱埋め込みを \mathbf{b}^q とし, 式 (3) で定義する類似度 $\text{sim}(\mathbf{b}_i^p, \mathbf{b}^q)$ が最大となる語義を選択する。

$$\text{sim}(\mathbf{b}_i^p, \mathbf{b}^q) = 2 \times \frac{P(\mathbf{b}_i^p | \mathbf{b}^q) P(\mathbf{b}^q | \mathbf{b}_i^p)}{P(\mathbf{b}_i^p | \mathbf{b}^q) + P(\mathbf{b}^q | \mathbf{b}_i^p)} \quad (3)$$

4.2 新語義の判定

ある用例に出現する対象単語 w の語義が新語義かどうかを判定する。全ての語義について, \mathcal{B}_w と \mathbf{b}^q との類似度 $\text{sim}(\mathbf{b}_i^p, \mathbf{b}^q)$ が閾値 α より小さい場合は新語義, それ以外は新語義ではないと判定する。

閾値 α は対象単語毎に設定する。開発データを用意し, 対象単語の開発データにおけるクエリ表現と, 訓練データから学習された対象単語の語義のプロトタイプ表現との類似度の平均を α とする。

4.3 新語義の上位語義の推定

対象単語 w の新語義を含む用例をテストデータとし, その上位語義の候補をランキングする。訓練データにおける全ての単語の全ての語義のプロトタイプ表現を $\mathcal{B} = \{\mathbf{b}_1^p, \dots, \mathbf{b}_m^p\}$ とする。まず, \mathcal{B} の中から $P(\mathbf{b}_i^p | \mathbf{b}^q)$ が閾値 β より大きい語義を選択し, 上位語義の候補の集合を得る。次に, これらの候補を $\text{Vol}(\mathbf{b}_i^p)$ と $\text{Vol}(\mathbf{b}^q)$ の差が小さい順にソートする。

5 実験

5.1 データセット

実験には既存の WSD のデータセットを用いる。Reganato らの提案した枠組み [20] に従い, SemCor 3.0 [21, 22] を訓練データ, SemEval-2007 [23] を開発データ, Senseval-2 [24], Senseval-3 [25], SemEval-2013 [26], SemEval-2015 [27] をテストデータとする。それぞれのデータセットは正しい語義が付与された用例の集合である。

上記のデータセット群から, $\mathcal{D}_{\text{living_thing.n.01}}$, $\mathcal{D}_{\text{artifact.01}}$, $\mathcal{D}_{\text{entity.n.01}}$ という3つのデータセットを作成する。living_thing.n.01, artifact.n.01, entity.n.01 は WordNet 上における Synset (語義) であり, 各データセットはこれらの Synset の下位の語義を持つ用例から構成される。 $\mathcal{D}_{\text{entity.n.01}}$ は名詞全体, それ以外のデータセットは名詞のサブセットである。訓練データならびに4節で述べた3つのタスクのテストデータの詳細を付録 A に示す。

5.2 実験設定

単一のベクトルで語義を表現する手法をベースラインとし、箱埋め込みを学習する提案手法と比較する。具体的には、事前学習済みのBERTによる埋め込みを用いる手法 (BERT-NN) と MetricWSD をベースラインとする。

BERT-NN, MetricWSD, 提案手法のいずれも、事前学習済みBERTとしてbert-base-uncased [28] を使用した。箱埋め込みを出力するための全結合層の次元数は256 (\mathbf{c} と \mathbf{o} の次元数は128), サポートサイズ N_S は5, クエリサイズ N_Q は20, 各エピソードの語義の数 N_C は128とした。

5.3 結果と考察

本項では、3つのタスクの実験結果について述べ、それを考察する。なお、学習された語義の箱埋め込みの例を付録Bに示す。

5.3.1 WSDの実験結果

各手法のWSDの正解率を表1に示す。ProtoBox S_r と ProtoBox S_n は、それぞれ S_r, S_n をサンプリング戦略としたときの提案手法を表す。「ALL」はテストデータ全て、「 ≤ 10 」はテストデータのうち訓練データにおける出現頻度が10回以下の語義のみを対象としたときの正解率である。提案手法は、小規模なデータセットである $\mathcal{D}_{\text{living_thing.n.01}}$ と $\mathcal{D}_{\text{artifact.n.01}}$ では、おおむねベースラインを上回った。しかし、名詞全体のデータセット $\mathcal{D}_{\text{entity.n.01}}$ では、提案手法の正解率はベースラインよりも低かった。 $\mathcal{D}_{\text{entity.n.01}}$ では語義の数が他の2つのデータセットと比べて非常に多いため、語義の箱埋め込みを適切に学習できていない可能性がある。

表1 WSDの結果(正解率)

Model	$\mathcal{D}_{\text{living_thing.n.01}}$		$\mathcal{D}_{\text{artifact.n.01}}$		$\mathcal{D}_{\text{entity.n.01}}$	
	ALL	≤ 10	ALL	≤ 10	ALL	≤ 10
BERT-NN	81.6	72.7	74.4	77.5	57.9	60.2
MetricWSD	82.1	77.3	85.9	92.5	73.3	57.9
ProtoBox S_r	83.2	80.3	87.2	87.5	62.8	57.2
ProtoBox S_n	78.4	81.8	84.6	87.5	63.2	58.3

5.3.2 新語義の判定の実験結果

各手法の新語義の判定のF1スコアを表2に示す。表1に示したWSDの結果とは異なり、 $\mathcal{D}_{\text{entity.n.01}}$ では提案手法がベースラインを上回り、残りの2つの

データセットでは下回った。2つのサンプリング戦略を比較すると、全体的には S_r の方が S_n よりも高いF1スコアが得られた。

表2 新語義の判定の結果(F1スコア)

Model	$\mathcal{D}_{\text{living_thing.n.01}}$	$\mathcal{D}_{\text{artifact.n.01}}$	$\mathcal{D}_{\text{entity.n.01}}$
BERT-NN	16.4	10.2	19.5
MetricWSD	15.2	17.7	21.2
ProtoBox S_r	14.5	14.1	25.1
ProtoBox S_n	10.9	15.5	22.5

5.3.3 新語義の上位語義推定の実験結果

新語義の上位語義推定の結果を表3に示す。4.3項で述べた閾値 β は0.5, 0.7, 0.9のいずれかと設定した。ACCはランクが最上位の上位語義の正解率, MRRは正解の上位語義のランクの逆数の平均, WPはランクが最上位の上位語義と正解の語義とのWu-Palmer類似度 [29] である。ACCとWPは全てのデータセットで提案手法がベースラインを上回った。 $\mathcal{D}_{\text{entity.n.01}}$ において、MRRがベースラインを下回り、ACCやWPでもベースラインとの差が小さいのは、ランキングの対象とする上位語義の候補の選定方法の違いによるものと考えられる。4.3項で述べたように、提案手法では $P(\mathbf{b}_i^p | \mathbf{b}^q) > \beta$ という条件を満たす語義を上位語義の候補とするが、ベースラインでは全ての語義を上位語義の候補としている。より多くの語義を評価対象とする $\mathcal{D}_{\text{entity.n.01}}$ では、真の上位語義が $P(\mathbf{b}_i^p | \mathbf{b}^q) > \beta$ の条件を満たさずに候補から除外されていると考えられる。よって、上位語義の候補を絞り込む条件を洗練する必要がある。

表3 新語義の上位語義推定の結果

Model	β	$\mathcal{D}_{\text{living_thing.n.01}}$			$\mathcal{D}_{\text{artifact.n.01}}$			$\mathcal{D}_{\text{entity.n.01}}$		
		ACC	MRR	WP	ACC	MRR	WP	ACC	MRR	WP
BERT-NN	-	15.0	25.9	75.4	9.4	15.0	56.7	6.8	11.3	46.0
MetricWSD	-	10.3	22.3	76.5	3.1	15.1	47.5	7.9	13.2	50.1
ProtoBox S_r	0.5	37.4	51.5	86.3	12.5	15.6	62.4	5.9	12.5	52.6
ProtoBox S_n		53.3	59.3	86.9	9.4	10.9	62.6	6.8	12.2	50.8
ProtoBox S_r	0.7	51.4	58.9	88.7	18.8	18.8	64.9	6.1	11.9	54.2
ProtoBox S_n		53.3	55.0	87.1	3.1	4.7	59.1	6.7	11.4	50.8
ProtoBox S_r	0.9	53.3	54.7	86.6	9.4	9.4	65.4	10.0	12.7	52.5
ProtoBox S_n		34.6	34.6	76.2	3.1	3.1	60.3	4.4	6.0	47.8

6 おわりに

本研究では語義の箱埋め込みを学習する新しい手法を提案した。今後の課題として、 $\mathcal{D}_{\text{entity.n.01}}$ のような語義の数が多き場合でも語義の箱埋め込みを適切に学習できるようにエピソード作成戦略を改善すること、動詞・形容詞など名詞以外の品詞での有効性を検証すること、などが挙げられる。

参考文献

- [1] Roberto Navigli. Word sense disambiguation: A survey. **ACM Comput. Surv.**, Vol. 41, No. 2, 2009.
- [2] George A. Miller. Wordnet: A lexical database for english. **Commun. ACM**, Vol. 38, No. 11, pp. 39–41, 1995.
- [3] Howard Chen, Mengzhou Xia, and Danqi Chen. Non-parametric few-shot learning for word sense disambiguation. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1774–1781, 2021.
- [4] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3509–3514, 2019.
- [5] Terra Blevins and Luke Zettlemoyer. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1006–1017, 2020.
- [6] Edoardo Barba, Tommaso Pasini, and Roberto Navigli. ESC: Re-designing WSD with extractive sense comprehension. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4661–4672, 2021.
- [7] Michele Bevilacqua and Roberto Navigli. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2854–2864, 2020.
- [8] Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. Zero-shot word sense disambiguation using sense definition embeddings. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5670–5681, 2019.
- [9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, 2020.
- [10] Michele Bevilacqua, Marco Maru, and Roberto Navigli. Generational or “how we went beyond word sense inventories and learned to gloss”. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7207–7221, 2020.
- [11] Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. Nibbling at the hard core of Word Sense Disambiguation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4724–4737, 2022.
- [12] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.
- [13] Georgeta Bordea, Els Lefever, and Paul Buitelaar. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**, pp. 1081–1091, 2016.
- [14] Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In **Proceedings of The Web Conference 2020**, pp. 486–497, 2020.
- [15] Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. STEAM: Self-supervised taxonomy expansion with mini-paths. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, pp. 1026–1035, 2020.
- [16] Song Jiang, Qiyue Yao, Qifan Wang, and Yizhou Sun. A single vector is not enough: Taxonomy expansion via box embeddings. In **Proceedings of the ACM Web Conference 2023**, pp. 2467–2476, 2023.
- [17] 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, 2019.
- [18] Yasumasa Onoe, Michael Boratko, Andrew McCallum, and Greg Durrett. Modeling fine-grained entity types with box embeddings. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 2051–2064, 2021.
- [19] Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. Improving local identifiability in probabilistic box embeddings. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 182–192, 2020.
- [20] Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. Word sense disambiguation: A unified evaluation framework and empirical comparison. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 99–110, 2017.
- [21] George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. Using a semantic concordance for sense identification. In **Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994**, 1994.
- [22] SemCor 3.0, (2024 年 1 月 閱覽). <https://web.eecs.umich.edu/~mihalcea/downloads.html#semcor>.
- [23] Sameer Pradhan, Edward Loper, Dmitry Dligach, and Martha Palmer. SemEval-2007 task-17: English lexical sample, SRL and all words. In **Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)**, pp. 87–92, 2007.
- [24] Philip Edmonds and Scott Cotton. SENSEVAL-2: Overview. In **Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems**, pp. 1–5, 2001.
- [25] Benjamin Snyder and Martha Palmer. The English all-words task. In **Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text**, pp. 41–43, 2004.
- [26] Roberto Navigli, David Jurgens, and Daniele Vannella. SemEval-2013 task 12: Multilingual word sense disambiguation. In **Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)**, pp. 222–231, 2013.
- [27] Andrea Moro and Roberto Navigli. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In **Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)**, pp. 288–297, 2015.
- [28] BERT base model (uncased), (2023 年 12 月 閱覽). <https://huggingface.co/bert-base-uncased>.
- [29] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In **32nd Annual Meeting of the Association for Computational Linguistics**, pp. 133–138, 1994.

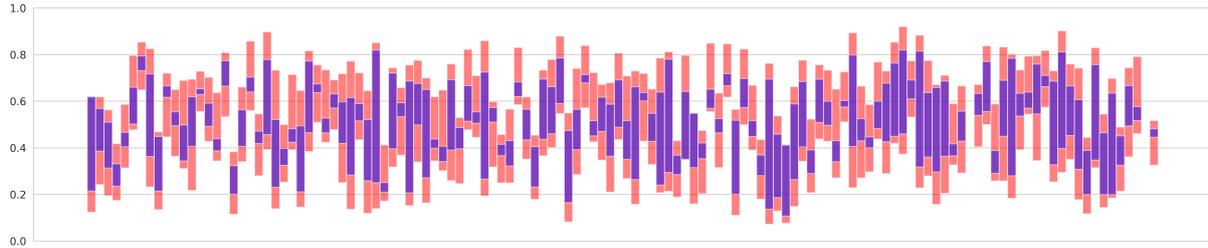


図4 赤：animal.n.01 の箱，青：dog.n.01 の箱. $P(\text{animal.n.01}|\text{dog.n.01}) = 0.98$, $P(\text{dog.n.01}|\text{animal.n.01}) = 2.04e^{-9}$.

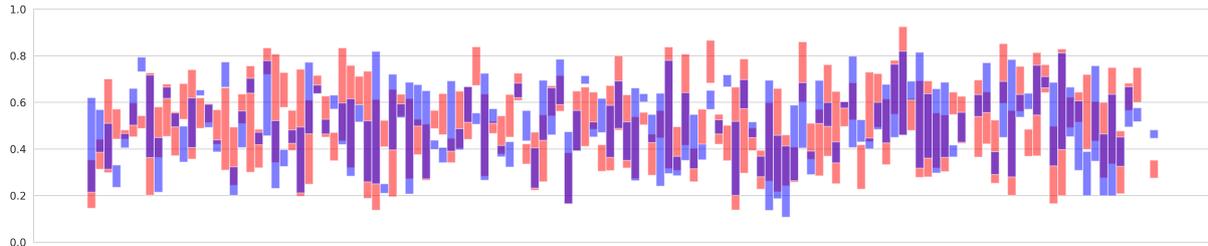


図5 赤：tree.n.01 の箱，青：dog.n.01 の箱. $P(\text{tree.n.01}|\text{dog.n.01}) = 8.44e^{-5}$, $P(\text{dog.n.01}|\text{tree.n.01}) = 3.84e^{-7}$.

A データセットの詳細

A.1 訓練データ

訓練データの統計を表4に示す.

表4 訓練データの統計

	$\mathcal{D}_{\text{living_thing.n.01}}$	$\mathcal{D}_{\text{artifact.n.01}}$	$\mathcal{D}_{\text{entity.n.01}}$
語義数	1,713	1,938	12,757
単語数	1,809	1,993	11,026
用例数	15,835	8,704	84,944

A.2 テストデータ

WSD WSDタスクのテストデータは，多義語を含む用例で，かつその正解の語義が訓練データに出現する用例の集合である．その統計を表5に示す.

表5 WSDタスクのテストデータの用例数

	$\mathcal{D}_{\text{living_thing.n.01}}$		$\mathcal{D}_{\text{artifact.n.01}}$		$\mathcal{D}_{\text{entity.n.01}}$	
	ALL	≤ 10	ALL	≤ 10	ALL	≤ 10
用例数	190	66	78	40	2,514	992

新語義の判定 新語義判定タスクのテストデータは，対象単語が訓練データに出現する用例の集合である．正解語義が訓練データに出現していなければ新語義の用例，そうでない場合は新語義ではない用例としてラベル付けする．その統計を表6に示す.

表6 新語義判定タスクのテストデータの用例数

	$\mathcal{D}_{\text{living_thing.n.01}}$	$\mathcal{D}_{\text{artifact.n.01}}$	$\mathcal{D}_{\text{entity.n.01}}$
新語義	20	12	295
新語義ではない	500	221	3,379

新語義の上位語義推定 新語義の上位語義推定タスクのテストデータは，訓練データに出現しない語義の用例の集合である．表6の「新語義」の用例とは異なり，対象単語自体が訓練データに出現しない語義の用例も含まれる．その統計を表7に示す.

表7 新語義の上位語義推定タスクのテストデータの用例数

	$\mathcal{D}_{\text{living_thing.n.01}}$	$\mathcal{D}_{\text{artifact.n.01}}$	$\mathcal{D}_{\text{entity.n.01}}$
用例数	107	32	658

B 語義の箱埋め込みの例

$\mathcal{D}_{\text{living_thing.n.01}}$ を訓練データとし，提案手法 Proto-Box S_r で学習した箱埋め込みの例を示す．図4は2つの語義 animal.n.01 と dog.n.01 のプロトタイプ表現(語義の箱埋め込み)を可視化したものである．横軸は箱の次元であり，縦軸は各次元の辺の長さである．紫は2つの語義の箱埋め込みが重なっている領域である．同図ならびに $P(\text{animal.n.01}|\text{dog.n.01})$ が0.98と高いことから，animal.n.01の箱埋め込みがdog.n.01の箱埋め込みをおおむね包含していることがわかる．また，tree.n.01とdog.n.01のプロトタイプ表現を図5に示す． $P(\text{tree.n.01}|\text{dog.n.01})$ も $P(\text{dog.n.01}|\text{tree.n.01})$ も低い値であることから，tree.n.01の箱埋め込みとdog.n.01の箱埋め込みはほとんど重なっていないことがわかる．