語彙的含意関係識別による単語意味属性の推定

長谷川 美夏　　小林 哲則　　林 良彦
早稲田大学理工学術院
mika@pcl.cs.waseda.ac.jp

1 はじめに

単語概念の意味を有限の意味属性の集合により表すことにより，人間の知覚や経験に基づく明示的な意味表現を得ることが期待される．本論文では，単語の分散表現から意味属性 (以下，属性) を推定するニューラルネットワークモデル *Word2Attr* を提案する．Word2Attr では，まず単語とそれが有する属性を整理した学習データを用いて，単語分散表現から属性ベクトルへの変換を学習する．次に，単語概念の階層性を利用した属性の伝播が行われることを期待し，語彙的含意関係の方向性識別タスクにより ファインチューニングを行う．

視覚的属性を整備した VisA dataset [1] を属性の学習データとし，語彙的含意関係を整理した HyperLex [2] を用いたファインチューニングを行った結果，意味的・視覚的な類似度・関連度を推定する評価タスクにおいて，属性推定に関する既存研究 [3, 4] による結果と同等以上の精度を得た．また，ファインチューニングによって特に視覚的類似度の精度が向上することを確認した．一方，学習データに与えられている属性がどの程度推定できるかという再現性の評価においては，学習データには存在しないが妥当と考えられる属性をある程度推定できることを確認した．

2 意味属性とその推定

2.1 属性を用いた意味表現

単語概念の意味表現として，大規模コーパスから導出する分散表現 [5, 6] が広く用いられている．これらは汎用性や頑健性を有する反面，意味の共通点や相違点を明示することには適していない．一方，単語概念の意味表現を属性の集合により表わそうとする考え方は従来より存在し，これによれば，属性集合間の関係によって単語間の関係を明示的に表すことが可能となる．ただし，どのような属性群を規定しておくか，また，ある単語・概念が持つ属性を定める手法も自明ではない．このため，人手の評定による属性データセットが整備されてきたが，自然言語処理の応用に適用するには，規模や完備性に問題がある．

属性データセット： 541 種類の英単語概念に対し多数の評定者が 2,526 種類の属性を与えた McRae semantic feature norms [7] がよく知られている．このデータセットを視覚属性の観点から整備したデータセットとして VisA dataset [1] があり，本研究ではこれを用いる．

2.2 属性を推定する研究

以上から，未知の単語に対しても，それが持つであろう属性を推定するための研究が行われている．Făgărăşan らの研究 [3] では，単語の分散表現を PLSR を用いて属性空間にマッピングする．結果として得られる属性ベクトルがスパースであること，具体単語と比べて抽象単語の属性推定精度が低いことが課題として指摘されている．Bulat らの研究 [4] では，意味属性と視覚空間の間で学習したクロスモーダルマップを用いて意味属性を推定する方法を提案しており，特に画像から抽出した視覚特徴の有効性を示している．本研究では，これらの研究との比較を行う．

3 語彙的含意関係

3.1 非対称な意味関係としての含意関係

語彙的含意関係 (Lexical Entailment) は，単語概念の間に成立する非対称な包含関係を表す意味関係であり，より具体的には，上位・下位関係 (hyponymy/hypernymy) のことを指す．例えば下位概念 dog のインスタンス集合は，上位概念 animal のインスタンス集合に包含されるが，この逆は成立しない．

3.2 語彙的含意関係と属性

論理的には hypernym の属性は hyponym に継承されるべきである．これに hyponymy 固有の属性が加わ

ることにより非対称な意味関係が定式化されるが，語彙知識において必ずしもこれは成立しない．しかし，分布包含仮説 (distributional inclusion hypothesis)[8] の考えに基づけば，hyponym の属性の多くは hypernym から継承されるべきであり，実際，本研究で用いた VisA dataset 内の単語から構成できる，上位・下位の関係にある 79 の単語ペアでは，hyponym に付与されている属性の約 70%以上が hypernym から継承されていた．よって，語彙的含意関係の方向性の識別 [9] を学習することにより，未知語に対しても適切な属性を上位関係の単語概念から伝播させることが期待できる．

4 提案手法: Word2Attr

本提案の Word2Attr の構成を図 1 に示す．

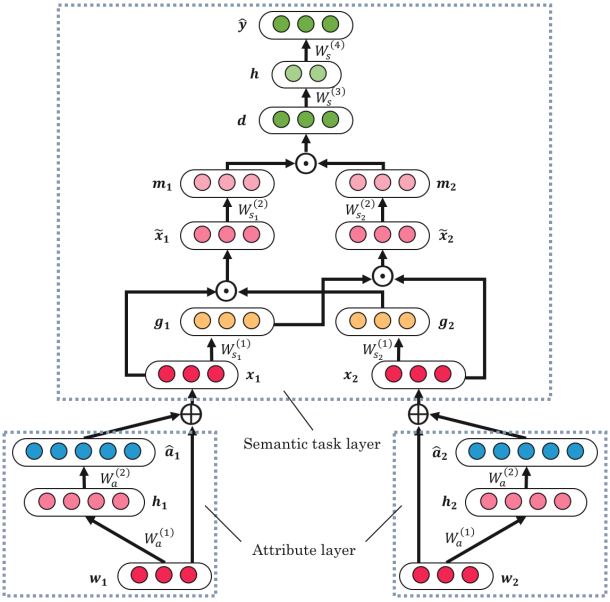

図 1: Word2Attr の構成

このネットワークは，与えられた単語対それぞれの単語分散表現[1]からそれぞれの属性ベクトルを推定する 2 つの Attribute layer と，単語対の語彙的含意関係の方向性を識別する Semantic task layer から構成される．2 つの Attribute layer はパラメータを共有しており，Siamese Neural Network [11] の構造をなす．

Word2Attr の学習は $(w_1, w_2, \texttt{type})$ の 3 項目を用いて行う．type は単語 w_1 から見た単語 w_2 がどのような関係性 (hyper:上位概念, hypo:下位概念, cohyp:兄弟概念, no-rel:無関係) にあるかを表す．これらは，学習データから与えられる．なお，単語概念がどのような属性を持つかは本来，概念の階層関係から推測されるべきであるが，これを補助するため，単語概念の粗い意味的なカテゴリ分類を利用する (4.2.2 節)．

[1]fastText [10] を利用．

4.1 学習に利用するデータセット

属性データ: VisA dataset に含まれる 510 概念から複数語義がある単語のうち片方を取り除いた 506 概念を属性データとして利用する．各単語概念は，'is_red','is_round' などの属性を各次元とし，その有無を 1/0 で表した 721 次元のベクトル (属性ベクトル) で表される．本研究に用いる 506 種類の単語概念は平均 15.14 種類の属性を持つ．事前学習では，404 件で学習を行い，残りの 102 件で評価を行う．

語彙的含意関係データ: 単語対に対して，その含意関係を注釈付けたデータセットである HyperLex[2] を用いて，含意関係の方向性識別を学習することにより，事前学習のパラメータをファインチューニングする．HyperLex には，含意関係の度合いも付与されているが，これは用いず，$(w_1, w_2, \texttt{type})$ の形で利用する．データセットの分割は，データセット内でのペアの重複だけではなく単語の重複も取り除いた lexical split (厳しい分割) と，ペアの重複はしていないが単語の重複は許す random split (緩い分割) の 2 種類を用いて比較する．lexical split の推定精度がよければ，単語対の関係性を正確に学習できていると考えられる．

4.2 事前学習

4.2.1 Attribute layer

Attribute layer のパラメータ $W_a^{(1)}$,$W_a^{(2)}$，$b_a^{(1)}$,$b_a^{(2)}$ は，1 層の隠れ層を持つ多層パーセプトロンで学習する．入力は L1-norm で正規化された，単語 w の分散表現 $\boldsymbol{w}$ であり，出力は各次元が対応する属性の度合いを表す 721 次元の属性ベクトル $\boldsymbol{a}$ である．損失関数 $\mathcal{L}$ は，推定結果の属性ベクトル $\hat{\boldsymbol{a}}$ と正解の属性ベクトルの $\boldsymbol{a}$ の最小二乗誤差とした．

4.2.2 カテゴリ分類モデルの学習

VisA dataset の単語には'animals','food','vehicles' のような粗い意味分類を表すカテゴリ情報 (16 種類) が付与されている．そこで，単語の分散表現を入力とし，カテゴリを分類する 1 層の隠れ層を持つ多重パーセプトロンを学習する．推定される結果は，Word2Attr において学習時に欠損している属性の学習データとし

て用いる．具体的には，単語 w が学習データに含まれる既知語 ($w \in \boldsymbol{S}$) の場合はその属性 a_w を利用するが，未知語 ($w \notin \boldsymbol{S}$) だった場合は，上記モデルで推定されたカテゴリ c における平均属性 a_c を用いる．

4.3 ファインチューニング

4.3.1 Attribute layer

Attribute layer では事前学習したパラメータを初期値とし，Semantic task layer における含意関係の方向性識別によりファインチューニングする．このとき，左右の Attribute layer のパラメータは共有されており，Attribute layer における損失関数 $\mathcal{L}_{attr}$ は，左右の layer の二乗誤差の和とする．

4.3.2 Semantic task layer

含意関係の方向性の識別においては，単語対の関係が hypernym/hyponym のどちらであるかを識別する基本タスク (`dir`) と，これに `cohyp` (兄弟関係) を含めた拡張タスク (`ext`) を行い，結果を比較する．`ext` タスクの設定では，兄弟関係にある単語概念間での属性の伝播が期待できる．例えば，('poodle', 'dog', `hyper`) と ('poodle', 'corgi', `cohyp`) が既知である場合，上記の情報から ('corgi', 'dog', `hyper`) という関係性を推定することが可能となる．なお，学習時には単語対が無関係 (`no-rel`) であるデータを加えて学習する．

Semantic task layer のネットワークの構造は，Supervised Directional Similarity Network (SDSN)[12] と呼ばれる，語彙的含意関係のスコアを推定するモデルを参考としているが，最終層付近ではスコア推定ではなく関係性識別を行う．

Semantic task layer では単語分散表現 $\boldsymbol{w_i}$ と Attribute layer で推定した属性ベクトル $\hat{\boldsymbol{a}}_{\boldsymbol{i}}$ を連結した特徴量を $\boldsymbol{x_i}$ と定義し入力とする．連結時の重みは，属性のみ，単語分散表現のみ，両者を単純連結，の 3 通りを実験で比較する．

次の段階は，含意関係における非対称性の獲得を目的とする．入力 x_1, x_2 から g_1, g_2 を導出し，単語 w_1 と w_2 にそれぞれ対応する g_1, g_2 と要素積を取る．この層では，単語対を分析するときにどの特徴が重要であるかを決定するため，g_1,g_2 は関係のない単語の特徴の影響を軽減するマスクの役割が期待される．

$$\boldsymbol{g_i} = \mathrm{sigmoid}(W^{(1)}_{s_i}\boldsymbol{x_i} + b^{(1)}_{s_i}) \quad (1)$$

$$\tilde{\boldsymbol{x}}_{\boldsymbol{1}} = \boldsymbol{x_1} \circ \boldsymbol{g_2} \quad (2)$$

$$\tilde{\boldsymbol{x}}_{\boldsymbol{2}} = \boldsymbol{x_2} \circ \boldsymbol{g_1} \quad (3)$$

この，$\tilde{\boldsymbol{x}}_{\boldsymbol{1}}, \tilde{\boldsymbol{x}}_{\boldsymbol{2}}$ を用いて `type` の識別を行う．

$$\boldsymbol{m_i} = \tanh(W^{(2)}_{s_i}\boldsymbol{x_i} + b^{(2)}_{s_i}) \quad (4)$$

$$\boldsymbol{d} = \boldsymbol{m_1} \circ \boldsymbol{m_2} \quad (5)$$

$$\boldsymbol{h} = \tanh(W^{(3)}_{s}\boldsymbol{d} + b^{(3)}_{s}) \quad (6)$$

$$\hat{\boldsymbol{y}} = W^{(4)}_{s}\boldsymbol{h} + b^{(4)}_{s} \quad (7)$$

Semantic task layer の損失関数 $\mathcal{L}_{sem}$ は，type 推定結果 $\hat{\boldsymbol{y}}$ と正解 `type` $\boldsymbol{y_{type}}$ の softmax cross entoropy (smxe) を用いる．

$$\mathcal{L}_{sem} = \frac{1}{N}\sum^{N} \mathrm{smxe}(\hat{\boldsymbol{y}}, \boldsymbol{y_{type}}) \quad (8)$$

全体の損失関数 $\mathcal{L}$ は Attribute layer と Semantic task layer の loss L_{attr}, L_{sem} を重み付けした和とする．

5 評価実験

意味タスクにおける有効性，および，学習データの再現性により，推定される属性ベクトルの評価を行う．

5.1 意味タスクにおける有用性

意味的・視覚的な類似度 (similarity)，および，関連度 (relatedness) の予測タスクにより有効性を評価する．評価データセットとして，表 1 に示す SemSim, VisSim[13]，MEN[14]，SimLex999[15] の 4 つを用いる．これらはいずれも人手により単語同士の類似性や関係性をスコアリングしたデータセットであり，$(word1, word2, score)$ という形式で表すことができる．評価では，$(word1, word2)$ が名詞同士のペアのみを用いた．SemSim, VisSim は VisA dataset に含まれる単語のみでペアが構成されるが，MEN と SimLex999 についてはその限りではない．

表 2 に結果を示す．単語同士の類似度/関連度の定量化には属性ベクトル同士のコサイン類似度を用い，評価指標には Spearman の順位相関係数を用いた．

表 1: 評価データセットの詳細

データセット	ペアデータ数	カバー数	評価内容
SemSim	7576	7576 (100%)	意味的類似度
VisSim	7576	7576 (100%)	視覚的類似度
MEN	2005	101 (5%)	意味的関連度
SimLex999	666	43 (4%)	意味的類似度

Word2Attr で推定した属性は，事前学習時の属性に比べ SemSim, VisSim, MEN (100%), SimLex (100%)

で相関係数が向上している．これらの結果は，2 件の先行研究の再現実験の結果と比べても同等もしくはそれ以上であり，VisSim の相関係数が特に向上が見られる．また，`dir` と `ext` の結果には大きな差はなかった．

MEN，SimLex に関しては属性ベクトルよりも fastText の表現の方が相関係数が優れており，これは収録語の種類に依存すると考えられる．SemSim，VisSim の収録語は視覚化可能な単語に限定されているが，MEN，SimLex についてはその限りではなく，抽象的な単語も含まれている．したがって，視覚化されない情報も保有している fastText の方が相関係数が高く示されると考えられる．

表 2: 意味的/視覚的・類似度/関連度の相関係数

データセット	入力	task	SemSim 100%	VisSim 100%	MEN 100%	SimLex 100%
		fastText	0.66	0.56	**0.82**	**0.49**
		VisA binary attr	0.69	0.59	NA	NA
		pre-train (baseline)	0.73	0.63	0.62	0.38
HLex rand.	a	`dir`	0.74	0.65	**0.68**	0.4
		`ext`	0.74	0.65	0.67	0.4
	f+a	`dir`	0.75	**0.66**	**0.68**	0.4
		`ext`	0.74	0.65	0.66	**0.41**
HLex lex.	a	`dir`	0.75	0.65	**0.68**	0.4
		`ext`	0.74	0.66	**0.68**	0.4
	f+a	`dir`	**0.76**	**0.66**	**0.68**	0.39
		`ext`	0.75	0.65	**0.68**	0.4
		Făgărăşan et al.[3]	0.75	0.61	**0.68**	0.4
		Bulat et al.[4]	0.74	0.61	**0.68**	0.42

5.2 属性の再現性

推定される属性ベクトルは実数値ベクトルであるが，閾値処理を行うことにより二値ベクトルへの変換が可能であり，学習データにおける属性の再現性を精度 (P)/再現率 (R)/F1 により評価できる．F1 を指標とするグリッドサーチにより閾値 θ を定めた結果を表 3 に示す．いずれの場合も安定した結果を得た．なお，VisA dataset の 1 概念あたりの平均属性数は約 15 であり，これに近い数の属性が推定された．

表 3: 閾値 θ による再現性評価

dataset	task	θ	P	R	F1	Ave. #
HyperLexr	`dir`	0.96	0.78	0.73	0.73	14.42
	`ext`	0.93	0.78	0.75	**0.75**	14.52
HyperLexl	`dir`	0.93	0.79	0.75	**0.75**	14.5
	`ext`	0.92	0.78	0.75	0.74	14.63
pre-train	-	0.7	0.93	0.81	0.85	13.01

ファインチューニングした結果は事前学習の結果より劣るが，推定された属性の中には，属性として妥当であるものも一定数含まれており，提案手法は，既存の属性データセットの不完備性を補完するために利用できる可能性がある．

6 むすび

本論文では，単語概念が有しうる意味属性を表す属性ベクトルを推定するニューラルネットワーク Word2Attr を提案し，その意味タスクにおける有効性を確認した．また，正解データの再現性の評価結果から，既存の属性データセットの不完備性を補完できる可能性を示した．今後は WordNet などの意味的語彙資源の利用などを検討する．

参考文献

[1] C. Silberer *et al.*, Models of Semantic Representation with Visual Attributes, ACL (2013) 572–582.

[2] I. Vulić *et al.*, HyperLex: A large-scale evaluation of graded lexical entailment, Computational Linguistics 43 (4) (2017) 781–835.

[3] L. Făgărăşan *et al.*, From distributional semantics to feature norms : grounding semantic models in human perceptual data as, IWCS (2015) 52–57.

[4] L. Bulat *et al.*, Vision and Feature Norms : Improving automatic feature norm learning through cross-modal maps, NAACL-HLT (2016) 579–588.

[5] M. Baroni *et al.*, Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, ACL (2014) 238–247.

[6] T. Mikolov *et al.*, Distributed Representations of Words and Phrases and Their Compositionality, NIPS 9 (2013) 3111–3119.

[7] K. McRae *et al.*, Semantic feature production norms for a large set of living and nonliving things., Behavior Research Methods 37 (4) (2005) 547–559.

[8] M. Geffet, I. Dagan, The distributional inclusion hypotheses and lexical entailment, ACL (2005) 107–114.

[9] K. A. Nguyen *et al.*, Hierarchical Embeddings for Hypernymy Detection and Directionality, EMNLP.

[10] P. Bojanowski *et al.*, Enriching Word Vectors with Subword Information, TACL 5 (2016) 135–146.

[11] G. Koch *et al.*, Siamese Neural Networks for One-shot Image Recognition, ICML.

[12] M. Rei *et al.*, Scoring Lexical Entailment with a Supervised Directional Similarity Network, ACL.

[13] C. Silberer, M. Lapata, Learning Grounded Meaning Representations with Autoencoders, ACL (2014) 721–732.

[14] E. Bruni *et al.*, Multimodal distributional semantics, Journal of Artificial Intelligence Research 49 (December) (2014) 1–47.

[15] F. Hill *et al.*, SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation, Computational Linguistics 41 (4) (2015) 665–695.