語義・概念の分散表現を利用した単語間の意味関係分類

金田 健太郎　　小林 哲則　　林 良彦
早稲田大学 理工学術院
kanada@pcl.cs.waseda.ac.jp

1 はじめに

本稿では単語間の意味関係分類を行う手法を提案する．例えば (bank, slope) という単語ペアが与えられたとき，その意味関係は下位-上位と判断するのが自然であるが，これは人間がそれぞれの単語の語義を自然に考慮して（bank を土手と解釈して）いるためだと考えられる．近年，単語の分散表現を使用した教師付き学習による単語間意味関係分類手法が提案され [6]，高い精度での分類を可能にしているが，語義や概念に関する分散表現が利用可能であれば，これを適切に利用することで更に高い精度を得ることが期待できる．

そこで本稿では,「単語ペアに特定の意味関係が成立する可能性は，その意味関係に応じて各単語と関連付けられた語義・概念の集合間の類似度によって表される」と仮定し，ターゲットとなる意味関係に対して計算される類以度群，及び，類以度を計算するのに用いた語義・概念レベルの分散表現を素性とする教師付き学習による意味関係分類手法を提案する．本稿で提案する手法を標準的なデータセットに適用したところ，従来手法を上回る結果が得られ，語義・概念レベルの分散表現を用いることの有用性が示唆された．

2 語義・概念の分散表現

単語 (word) はいくつかの語義 (sense) を持ち，同じ意味を持つ語義の集合として概念 (concept) が構成される．語義･概念レベルの分散表現を作成する手法はいくつか提案されている [7, 2] が，中でも AutoExtend[8] と呼ばれる手法では，任意の単語の分散表現に任意の語義資源を併せることで，少ない計算量で辞書資源中の語義･概念に対し，単語ベクトルと同次元の分散表現を与えることができる．そのため今回はこの手法を利用し，Google の配布する Word2Vec モデル [4]（CBOW, 300 次元）[1] と WordNet[5] 語義・概念の分散表現を作成した．

[1]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing

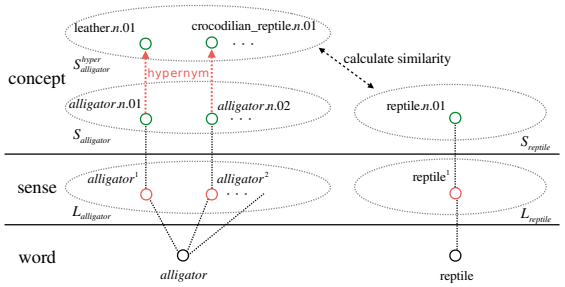

図 1: 語義・概念ベクトルを利用した類以度の計算（下位-上位関係時）

3 提案手法

本稿では教師付き学習による単語間意味関係の分類手法を提案する．その際，単語ペアに特定の意味関係が成立する可能性は，その意味関係に応じて各単語と関連付けられた語義・概念の集合間の類似度によって表されると期待する．例えば，(alligator, reptile) という単語ペアが与えられたとき，そこに下位-上位の関係 (hypernym) があるかどうかの尺度は，aligator と hypernym の関係にある概念の集合と，reptile の概念の集合間の類以度によって表されると期待する (図 1)．この期待のもとで次の手順により類以度を求める．

1. 単語と結びつける語義・概念の集合を構成する（3.1 節）
2. 単語間に成立しうる意味関係に応じて 7 通りの語義・概念集合の組み合わせを構成する（3.2 節）
3. 組み合わせた語義・概念集合それぞれに対して 3 通りの手法で類以度を計算する（3.3 節）

最終的に 1 つの単語ペアに対して，21 種類（7 通りのノード集合 × 3 通りの計算手法）の類以度が計算される．この類以度に単語ベクトル対によるコサイン類以度を加えた 22 種類の類以度と，それぞれの類以度をもたらすベクトル対を，学習の際の素性として用いる．

3.1 単語に対するノード集合の構成

WordNet 中で定義された語義・概念間の意味的ネットワークを利用し，単語 w に対して以下のように 5 種類の語義・概念の集合（ノード集合）を構成する．これらのうち一部の例（w=alligator, reptile）が図 1 に示されている．

- 語義集合 L_w: 単語 w の持つ語義の集合
- 概念集合 S_w: 単語 w が持つ語義が含まれる概念の集合
- 上位概念集合 S_w^{hyper}: S_w 中の概念と直接 hypernymy の関係（下位-上位）の関係で結びついた概念の集合
- attribute 概念集合 S_w^{attri}: S_w 中の概念と直接 attribute（主体-性質，または性質-主体）の関係で結びついた概念の集合
- meronym 概念集合 S_w^{attri}: S_w 中の概念と直接 meronymy（全体-部分）の関係で結びついた概念の集合

3.2 類以度計算に使用するノード集合対の構成

単語ペア w_1，w_2 が与えられたとき，以下 5 種類の意味関係について，それぞれが成立する可能性を 7 種類のノード集合の組み合わせを用いた類以度で表す．意味関係とノード集合対の対応は次の通りであり，またノード集合の組み合わせ方を以下括弧内のように呼ぶ．

- 類似性/関連性：L_{w_1} - L_{w_2} $(sense)$, S_{w_1} - S_{w_2} $(concept)$
- 下位-上位関係性：$S_{w_1}^{hyper}$ - S_{w_2} $(hyper)$
- 兄弟関係性：$S_{w_1}^{hyper}$ - $S_{w_2}^{hyper}$ $(coord)$
- 形容関係性：$S_{w_1}^{attri}$ - S_{w_2} $(attri_1)$, S_{w_1} - $S_{w_2}^{attri}$ $(attri_2)$
- 全体-部分関係性：$S_{w_1}^{mero}$ - S_{w_2} $(mero)$

類似性/関連性の有無について考えるときは，単語の持つ意味そのもの，すなわち単語の語義・概念について比較をするのが適切である．よって語義集合同士，また概念集合同士の類以度が類似性/関連性の有無の尺度となると仮定する．

w_2 が w_1 の上位語であるならば，w_1 の上位概念（w_1 の上位語の概念）と w_2 の概念が類似するはずである（図 1 参照）．よって $S_{w_1}^{hyper}$ と S_{w_2} の類以度が下位-上位関係性の有無の尺度になると仮定する．

w_1 と w_2 が兄弟関係にある（同じ上位語を持つ）のであれば，両者が似たような上位概念を持つはずである．よって $S_{w_1}^{hyper}$ と $S_{w_2}^{hyper}$ の類以度が兄弟関係性の有無の尺度になると仮定する．

w_1 と w_2 が形容関係にあるのであれば，w_1 の attribute 概念（w_1 を形容するような単語の概念）と w_2 の概念，または w_1 の概念と w_2 の attribute 概念（w_2 によって形容される単語の概念）が類似するはずである．よって $S_{w_1}^{attri}$ と S_{w_2}，また S_{w_1} と $S_{w_2}^{attri}$ の類以度が形容関係性の尺度になると仮定する．

w_1 と w_2 が全体-部分関係にあるのであれば，w_1 の meronym 概念（w_1 の一部分になるような単語の概念）と w_2 の概念が類似するはずである．よって $S_{w_1}^{hyper}$ と $S_{w_2}^{hyper}$ の類以度が全体-部分関係性の有無の尺度になると仮定する．

3.3 類以度計算手法

単語ペア w_1，w_2 に対し，c$\in$ $\{sense,\ concept,\ hyper,\ attri_1,\ attri_2,\ mero\}$ という組み合わせ方でノード集合の組 X_{w_1}, X_{w_2} が与えられたとき，その類以度を以下の 3 通りの手法によって計算する．

sim^c_{max} ： sim^c_{max} は，各集合中のノードを組み合わせ，それぞれの分散表現でコサイン類以度を算出し，その最大値をノード集合間の類以度とする手法であり，以下の式で与えられる．

$$sim^c_{max}(w_1, w_2) = \max_{x_1 \in X_{w_1}, x_2 \in X_{w_2}} sim(\vec{x_1}, \vec{x_2})$$

ただし，$sim(\vec{x_1}, \vec{x_2})$ は $\vec{x_1}, \vec{x_2}$ のコサイン類以度である．ノード集合はそこに含まれるノードの数だけ意味的曖昧性（多義性）を持つといえるが，その曖昧性を互いに最も近い意味同士と解釈することで解消した後に類以度を算出するような手法である．類以度最大となるときのノード対の分散表現 ($\vec{x_1}$ と $\vec{x_2}$) も素性として使用する．

sim^c_{sum} ： sim^c_{sum} は，各集合に含まれるノードについて，分散表現の総和を求めた後に，総和同士のコサイン類以度を算出する手法であり，以下の式で与えられる．

$$sim^c_{sum}(w_1, w_2) = sim(\sum_{x_1 \in X_{w_1}} \vec{x_1}, \sum_{x_2 \in X_{w_2}} \vec{x_2})$$

この手法では，それぞれのノード集合について，その全体的な意味を表すようなベクトルを作成し，それ

らの類以度を求めている．各ノード集合に含まれる分散表現の総和 ($\sum_{x_1 \in X_{w_1}} \vec{x_1}$ と $\sum_{x_2 \in X_{w_2}} \vec{x_2}$) も素性として利用する．

sim^c_{med}： sim^c_{med} は，各集合中のノードを組み合わせ，それぞれの分散表現でコサイン類以度を算出し，その中央値をノード集合間の類以度とする手法であり，以下の式で与えられる．

$$sim^c_{med}(w_1, w_2) = \underset{x_1 \in X_{w_1}, x_2 \in X_{w_2}}{\text{median}} sim(\vec{x_1}, \vec{x_2})$$

この手法においては中間的な類以度が算出されるが，この類以度に寄与するノードは各集合の中から一つづつのみである．類以度が中央値となるときのノード対の分散表現 ($\vec{x_1}$ と $\vec{x_2}$) も素性として使用する．

4 実験

4.1 実験設定

提案手法の評価のために，BLESS データセット [1] を用いて単語間意味関係分類の評価を行う．このデータセットには14400組の target concept（名詞）- relatum（単語）ペアが含まれている．target concept は 17 大まかな意味（爬虫類，家具等）によって 17 のカテゴリに分けられ，relatum は以下の 5 つの関係で名詞と結び付けられている．

- COORD (3565 組): 兄弟関係（例： alligator - lizard）
- HYPER (1337 組): 下位 - 上位関係（例：alligator - animal）
- MERO (2943 組): 全体 - 部分関係（例：alligator - mouth）
- ATTRI (2731 組): 主体 - 形容の関係（例：alligator - aquatic）
- EVENT (3824 組): 主体 - 動作の関係（例：alligator - swim）

比較評価の対称としては，単語の分散表現のみを用いて教師付き学習を行う従来手法 [6] を取り上げる．従来手法では，単語ペアが与えられたとき，それぞれの単語に対して word2vec (CBOW)[4] による単語分散表現の差分（$WECE^{offset}_{BoW}$）または連結（$WECE^{concat}_{BoW}$），及び dependency-based skip-gram[3] による単語分散表現の差分（$WECE^{offset}_{Dep}$）または連結（$WECE^{concat}_{Dep}$）のいずれかを用いて教師付き学習を行っている．

提案手法では RandomForest 分類器を用いて教師付き学習を行う．この際，22 種類の類以度に加え 22 種類のベクトル対の差分または連結を素性として用いる．差分を使用した場合を $Proposal_{offset}$，連結を使用した場合を $Proposal_{concat}$ と表記する．また教師データとテストデータの分割は従来手法に従い，以下の 2 通りで行う．また，評価尺度も従来手法に従い Precision (P) , Recall (R) , F-measure (F) を用いる．

In-domain (ID)： target concept のカテゴリに従いデータを 17 分割し，そのカテゴリ 1 つ 1 つに対して 5 分割の交差検定を行う．17 × 5 回の試行の平均スコアを算出する．

Out-of-domain (OoD)： target concept のカテゴリに従いデータを 17 分割し，そのうち 16 カテゴリ分のデータを学習，残り 1 カテゴリのデータをテストに用いる．17 回の試行の平均スコアを算出する．

提案手法においては，素性となる類以度を算出する際に WordNet の情報を利用している．そのため，単にこの情報を利用して単語間意味関係分類を行ったときよりも高いスコアが得られているかどうかを確認することが重要である．よって次のようにベースラインを設定する．

ベースライン： 単語ペア w_1，w_2 が与えられたとき，S_{w_1} 中の概念と S_{w_2} の概念のうちいずれかが，WordNet 上において直接何らかの関係によって結ばれているのであれば，単語ペアに対しその関係が存在しているとする．

辞書中に EVENT の関係は定義されておらず，また BLESS の ATTRI と Wordnet における attribute は定義が異なるため，ベースラインの手法において発見できる関係は COORD, HYPER, MERO の 3 つのみである．よってこの 3 つの関係についてのみ P, R, F を算出し比較評価を行う．

4.2 実験結果

表 1 に従来手法との比較結果を示す．ベクトルの差分を利用した場合と連結を利用した場合のどちらにおいても，P, R, F 全てのスコアにおいて提案手法が従来手法を超えるスコアを獲得している．これにより，語義・概念のベクトルを利用することの有効性が確認できた．

OoD 分割時の $Proposal_{concat}$（$Proposal^{OoD}_{concat}$）における結果の詳細と，ベースラインとの比較結果を表 2 に示す．提案手法においては，異なる品詞間にあ

	In-domain			Out-of-domain		
	P	R	F1	P	R	F1
$WECE_{BoW}^{offset}$	0.900	0.909	0.904	0.680	0.669	0.675
$WECE_{Dep}^{offset}$	0.853	0.865	0.859	0.687	0.623	0.654
$Proposal_{offset}$	0.913	0.907	0.906	0.766	0.762	0.753
$WECE_{BoW}^{concat}$	0.899	0.910	0.904	0.838	0.570	0.678
$WECE_{Dep}^{concat}$	0.859	0.870	0.865	0.782	0.638	0.703
$Proposal_{concat}$	**0.973**	**0.971**	**0.971**	**0.839**	**0.819**	**0.812**

表 1: BLESS データセットにおける従来手法との比較

	$Proposal_{concat}^{OoD}$			ベースライン		
	P	R	F1	P	R	F1
COORD	0.761	0.559	0.645	0.550	0.108	0.180
HYPER	0.767	0.654	0.706	0.746	0.199	0.314
MERO	0.625	0.809	0.705	0.934	0.034	0.065
ATTRI	0.913	0.995	0.952	-	-	-
EVENT	0.974	0.983	0.979	-	-	-

表 2: $Proposal_{concat}^{OoD}$ における詳細結果とベースラインの比較

る関係の ATTRI, EVENT に比べ，名詞-名詞の関係分類である COORD, HYPER, MERO の分類が低い結果となっているが，ベースラインと比較した際には Recall スコアについて大きく上回っている．このことから，辞書情報を単に利用するよりも広い適用範囲を持っていることがわかる．

	P	R	F1	F1 変化量
$Proposal_{concat}^{OoD}$	0.839	0.819	0.812	-
−単語の分散表現	0.845	0.827	0.819	**0.008**
$-sense$	0.833	0.815	0.806	-0.006
$-concept$	0.826	0.809	0.802	-0.010
$-coord$	0.834	0.811	0.803	-0.009
$-hyper$	0.826	0.803	0.800	-0.012
$-attri_1$	0.826	0.806	0.798	-0.014
$-attri_2$	0.842	0.820	0.814	0.002
$-mero$	0.835	0.813	0.806	-0.006

表 3: 各ノード集合対（単語の分散表現対）に由来する素性を取り除いたときの結果

表 3 では，$Proposal_{concat}^{OoD}$ において各ノード対に由来する 3 つの類以度，あるいは単語の分散表現対に由来する 1 つの類以度及びベクトル対を除いて学習を行った結果を示す．ノード対を除いた場合はほぼスコアが低下していることから，各組み合わせの有効性が確認できる．一方で単語対を除いた場合はスコアが上昇していることから，語義・概念の分散表現を適切に利用した場合，単語の分散表現はむしろノイズになることが示唆される．

表 4 では，各計算手法に由来する 7 つの類以度及びベクトル対を除いて学習を行った結果を示す．表 3 に比べて多くの素性を除いているにも関わらず，どの場合においてもさほど大きくスコアが変化しないことから，1 つの手法を他の 2 つの手法が補完しあっていることが示唆される．

	P	R	F1	F1 変化量
$Proposal_{concat}^{OoD}$	0.839	0.819	0.812	-
$-sim_{max}$	0.835	0.812	0.805	-0.007
$-sim_{sum}$	0.843	0.822	0.816	0.004
$-sim_{med}$	0.838	0.811	0.805	-0.007

表 4: 各計算手法に由来する素性を取り除いたときの結果

5 おわりに

本稿では単語間意味関係分類において，単語の分散表現と辞書資源から作成した語義・概念のベクトルを適切に利用することの有効性を確認した．しかし，この手法の適用範囲は辞書中に含まれる単語に限られる．辞書中に無い単語：「未知語」に対しても，それを辞書に結びつけるなどして適用範囲を広げていくことが今後の課題となる．

謝辞

本研究は JSPS 科研費#25280117 の助成を受けた．

参考文献

[1] Marco Baroni et al. How we BLESSed distributional semantic evaluation. *GEMS '11*, pages 1–10, 2011.

[2] Ignacio Iacobacci et al. SENSEMBED: Learning Sense Embeddings for Word and Relational Similarity. *Acl*, (1):95–105, 2015.

[3] Omer Levy et al. Dependencybased word embeddings. *Proceedings of the ACL*, 2:302–308, 2014.

[4] Tomas Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pages 1–9, 2013.

[5] George a. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[6] Silvia Necules cu et al. Reading Between the Lines : Overcoming Data Sparsity for Accurate Classification of Lexical Relationships. **SEM 2015*, pages 182–192, 2015.

[7] Arvind Neelakantan et al. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. *Emnlp*, pages 1059–1069, 2014.

[8] Sascha Rothe et al. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. *Proceedings of the ACL*, pages 1793–1803, 2015.