辞書と文脈情報を用いた対義語モデルの学習

Learning Antonym Detection Models using Thesauri and Contextual Information

小野 正貴 三輪 誠 佐々木 裕
Masataka Ono Makoto Miwa Yutaka Sasaki
豊田工業大学
Toyota Technological Institute
{sd12412, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 序論

自然言語処理の分野において，単語の意味や単語間の関係性を獲得することは非常に重要である．Word Embedding (WE) は，教師なしデータであるテキストから学習することで，単語の文法的・意味的情報を単語ごとに割り当てられた数値ベクトルに埋め込むことのできるモデルとして注目されており [3]，特に単語間の類似性を捉えることに高い性能を発揮している．一方対義性については，単語と文脈に関する分布仮説「同じ文脈に現れる単語は似た意味を持つ傾向にある」を前提としながら，対義語同士もまた同じ文脈に出やすいため，類義語と対義語は区別することができない．

対義語を判断する手法として，潜在的意味解析をベースとした，辞書知識を利用する手法があるが [6, 7]，どちらも文脈情報を十分に活用できているとは言えない．

本研究では，対義語モデルの精度向上を目的として辞書と文脈情報から対義語を判断することができるモデルを新たに提案し，人間向けの対義語選択問題への回答を通して評価する．

2 関連研究

Skip-Gram with Negative Sampling (SGNS) は 1 式を最大化することで文脈情報から WE を学習する．

$$\sum_{w\in V}\sum_{c\in V}\{\#(w,c)\log\sigma(sim(w,c)) + k\#(w)P_0(c)\log\sigma(-sim(w,c))\} \quad (1)$$

ここで，V は単語の集合，k は負例数，$\#(w,c)$ は単語 w と単語 c の共起回数，$\#(w)$ は w の出現回数，$P_0(c)$ は c の出現確率，σ はシグモイド関数である．第 1 項は共起回数に比例した単語の WE 同士の類似度の総和であり，第 2 項は負例として共起しない単語の WE との逆類似度の総和である．また，2 つの単語の WE 同士の類似度は 2 式のように計算される．

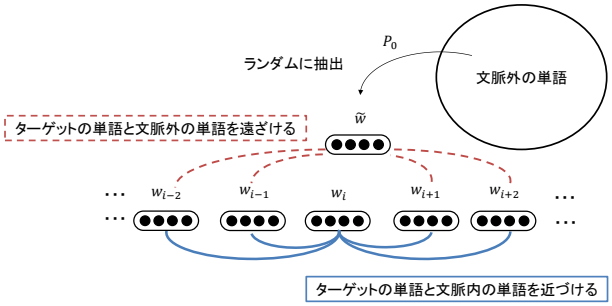

図 1: SGNS

$$sim(w_1,w_2) = v_{w_1}\cdot v_{w_2} + b_{w_1} \quad (2)$$

ここで v_w は単語 w の WE であり，b_w はバイアス項である．

3 提案手法

SGNS モデルでは，1 式に示した通り，重み付けされた単語間の類似度と逆類似度の対数の総和で，類義性のような意味関係性をうまく得ることができている．したがって，SGNS と同様の方法で，より対義性に焦点を当てたモデルを構築することで，対義性の情報を効率的に獲得することができるのではないかと推測される．

そこで我々は SGNS の単語ペアへの類似度・逆類似度による目的関数から WE を学習するのに倣い，辞書のみから学習を行うモデルと，それに加えて SGNS と同様の文脈情報を活用して学習を行うモデルの 2 つを提案する．

3.1 Word Embedding from Thesauri

まず，辞書情報のみから WE を学習するモデルとして，Word Embedding from Thesauri (WE-T) を提案する．目的関数を 3 式に示す．

$$\sum_{w\in V}\sum_{s\in S_w}\log\sigma(sim(w,s)) + \alpha\sum_{w\in V}\sum_{a\in A_w}\log\sigma(-sim(w,a)) \tag{3}$$

S_w,A_w はそれぞれ，w に対する類義語の集合，対義語の集合を表す．第 1 項はすべての類義語のペアの類似度の合計を表しており，第 2 項はすべての対義語のペアの逆類似度を表している．この目的関数を最大化することで，類義語同士の類似度は高くなり，対義語同士の逆類似度高く，すなわち類似度は低くなる．3 式を最大化するために，最適化手法の一つである AdaGrad[1] を用いる．

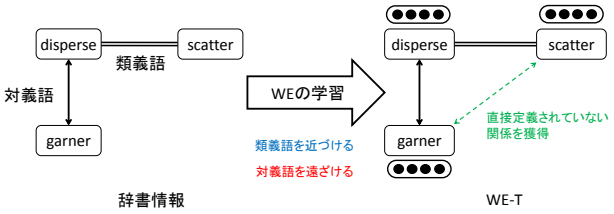

図 2: WE-T

3.2 Word Embedding from Thesauri and Distributional Information

辞書情報と文脈情報を用いて WE を学習するモデルとして，Word Embedding from Thesauri and Distributional Information (WE-TD) を提案する．1 式と 3 式の和をハイパーパラメータ β で重要度のバランスをとった 4 式を目的関数とする．

$$\begin{aligned}&\beta\{\sum_{w\in V}\sum_{s\in S_w}\log\sigma(sim(w,s)) \\ &+\alpha\sum_{w\in V}\sum_{a\in A_w}\log\sigma(-sim(w,a))\} \\ &+\sum_{w\in V}\sum_{c\in V}\{\#(w,c)\log\sigma(sim(w,c)) \\ &+k\#(w)P_0(c)\log\sigma(-sim(w,c))\}\end{aligned} \tag{4}$$

ここで，2 つの項がどちらも単語間の類似度及び逆類似度に，単語間テーブルからの重み付けをしたものの総和であることに着目すると各項を 5 式のようにまとめることができる．

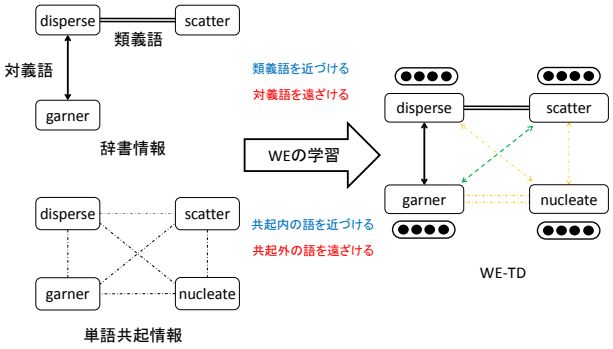

図 3: WE-TD

$$\sum_{w\in V}\sum_{c\in V}\{(T_p(w,c)\log\sigma(sim(w,c)) + T_n(w,c)\log\sigma(-sim(w,c))\} \tag{5}$$

ここで，T_p，T_n はそれぞれモデルに対する単語 w,c の類似度及び逆類似度の重みを示すテーブルである．

4 評価設定

4.1 GRE Antonym question task

対義語に関する判断が行えるかを評価するために，対義語判断の評価でよく用いられる GRE Antonym Question Task[5] を用いて評価を行う．この問題では，一つの問はターゲットの単語と 5 つの対義語候補で構成されている．以下に例を示す．

tawdry (下品な)

A. sagacious (賢明な)
B. truthful (信頼する)
C. hilarious (陽気な)
D. elegant (上品な)
E. functional (便利な)

この問の場合には，正答は D となる．

公開されているデータセットは 162 問の開発セットと 950 問の評価セットで構成されているが，この評価セット中 160 問は開発セットの問が含まれている．したがって本実験では，他の研究と比較をするために元の評価セット（評価セット A），正確な汎化性能を測るために評価セット A から開発セットの問題を取り除いた評価セット B の両方に対して性能を評価した．

表 1: GRE antonym question task における実験結果

	開発セット			評価セット A			評価セット B		
	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
直接辞書引き	1.00	0.49	0.66	0.98	0.45	0.62	0.98	0.45	0.61
WE-T	0.92	0.71	0.80	0.90	0.72	0.80	0.90	0.72	0.80
SGNS	0.09	0.08	0.09	0.08	0.07	0.07	0.07	0.07	0.07
Yih ら，2012[6]	0.88	0.87	0.87	0.81	0.80	0.81	—	—	—
Zhang ら，2014[7]	0.88	0.88	0.88	0.82	0.82	0.82	—	—	—
WE-TD	0.92	0.91	0.91	0.90	0.88	0.89	0.89	0.87	0.88

提案したモデルで学習した WE でこの問題に回答するために，まずターゲットと各候補との類似度を 2 式を用いて計算し，最も類似度の低いものを選択することとする．ただし，問題中に一つでも未知語が存在した場合には回答を行わないこととした．

4.2 学習に用いた言語リソース

モデルの学習に用いる教師ありデータとして，Zhang ら [7] が WordNet[4] 及び Roget[2] から作成した類義語と対義語のペアを用いた．このデータセットには，52,760 語に対して平均 11.7 語の類義語，21,3319 語に対して平均 6.5 語の対義語が定義されている．

一方教師なしデータとして，Wikipedia 全ページからタグ情報などを取り除いたテキストを用いた．出現単語数は約 7 千万語である．これらに対し前処理として，すべての文字を小文字化した後，出現頻度上位 100,000 位タイの単語から単語共起テーブルを作成した．

4.3 パラメータ設定

前述の GRE Antonym Question Task の開発セットを用い，モデル学習に必要なパラメータを決定した．まず，WE-T モデルについて，ベクトルの次元数は 300 次元，AdaGrad による試行回数は 20 回，学習率は 0.03，とした．また 3 式におけるハイパーパラメータ α は対義語と類義語のそれぞれのペア数の比から 3.2 とした．次に WE-TD モデルについて，まず Wikipedia からは出現する上位 100,000 語のみについて共起情報を持つテーブルを生成し，学習に用いた．4 式におけるハイパーパラメータ β は 100，SGNS における負例数 k は 5，共起をカウントする前後のウィンドウサイズ C は 5, サブサンプリングのしきい値は 10^{-8} とした．

4.4 評価指標

関連研究の結果と比較を行うために，同様に F 値を用いて評価を行った．F 値は，適合率と再現率の調和平均である．ただし，本評価においては，適合率は回答を行った問題数に対する正答数の割合，再現率は全問題数に対する正答数とした．

5 結果・考察

実験結果を表 1 に示す．まず SGNS は，ほとんど正しく答えられおらず，文脈情報のみでは対義語判断が難しいことを示している．次に WE-T では，評価セット対して行列分解をベースにした既存手法に迫る性能をあげている．これは辞書情報から類推される対義語関係を効率的に学習できたためであると考えられる．最後に WE-TD は，F 値 0.89 と最も高い性能を示した．さらに開発セットにおける F 値に対して評価セットにおける F 値の低下が既存手法より少なく，WE-TD は高い汎化性能を持つと言える．

次に，評価セット B に対する WE-TD の誤答割合について表 2 に示す．複数の候補語が対義語辞書にあるパターンであるが，誤答及び正答についてどちらもターゲット単語に対して類似度は 0.01 以下と非常に低くなっており，WE としては辞書上の対義性を反映ものが得られていると言える．しかしながら，間接的な対義語を用いて学習してしまっているがために，意味的には近いものの対義語ではない単語間においても類義性が低くなってしまったと考えられる．また，正答に対するターゲット単語の類似度に応じて分類し考察を行う．まず，類義性が十分に低い (< 0.1) のパターンについては，先ほどの学習データの辞書にある場合のものだけでなく，辞書上にはない対義語のものもあった．意味を見た場合でも，おおよその意味的方向性が似ていて極性が異なるというように捉えること

表 2: WE-TD の評価セット B における誤答の分類

誤答の分類	誤答数	全誤答中の割合	全問題数に対する割合
全誤答	88	1.00	0.12
候補が辞書にある	15	0.17	0.02
候補が複数辞書にある	4	0.05	0.01
正答語との類似度 < 0.1	28	0.32	0.04
正答語との類似度 < 0.5	53	0.40	0.05
正答語との類似度 > 0.9	13	0.15	0.02

もでき，おおよそうまく WE が学習できていると考えられる．しかしながら，これら微妙なニュアンスの違いを正確に判断して確かな対義語を選ぶためには，より高い水準での意味性獲得が必要と予想される．次に，対義語であるはずにも関わらず，正答とターゲットの単語との類似度が高く (> 0.9) 判断されているパターンについては，辞書上に間接的にも対義性を示す情報が辞書になく，かつ同様な文脈で出現する場合が多いためであると考えられる．

6 結論

本研究では，対義語の判断精度の向上を目的として，新たな 2 つのモデルを提案した．GRE Antonym Question Task を用いて対義語の判断性能の評価を行い，提案した辞書情報と文脈情報から学習するモデルで既存手法から 7%の F 値向上を達成した．今後の課題として，他の意味関係に対する学習を行うことが挙げられる．

参考文献

[1] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159, July 2011.

[2] Barbara Ann Kipfer. *Roget's 21st Century Thesaurus*. Philip Lief Group, third edition edition, 2009.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[4] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, November 1995.

[5] Saif Mohammad, Bonnie Dorr, and Graeme Hirst. Computing word-pair antonymy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 982–991, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[6] Wen-tau Yih, Geoffrey Zweig, and John Platt. Polarity inducing latent semantic analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1212–1222, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[7] Jingwei Zhang, Jeremy Salwen, Michael Glass, and Alfio Gliozzo. Word semantic representations using bayesian probabilistic tensor factorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1522–1531. Association for Computational Linguistics, 2014.