構成性に基づく関係パタンの意味計算

高瀬翔† 岡崎直観 †‡ 乾健太郎†
東北大学† 科学技術振興機構さきがけ‡
{takase, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

近年，分布仮説 [3] に基づく単語の意味ベクトル（word embeddings）の学習手法がめざましい発展を遂げている．Mikolov ら [9] が提案した Skip-gram モデルは，単語ベクトルが加法構成性（有名な例は $\boldsymbol{v}_{\text{king}} - \boldsymbol{v}_{\text{man}} + \boldsymbol{v}_{\text{woman}} \approx \boldsymbol{v}_{\text{queen}}$）を持つことで注目を浴びた．Levy ら [6, 7] は，単語ベクトルが加法構成性を示す理由と，Negative Sampling を用いた Skip-gram モデルが単語文脈行列の相互情報量（にバイアスをかけたもの）をモデル化していることを示した．Pennington ら [12] は，単語—文脈共起行列を単語ベクトルから直接的にモデル化する手法を提案し，Skip-gram モデルを上回る性能を報告した．

分布仮説による意味のモデル化は，単語に限られた話ではない．Mikolov らは Skip-gram モデルを名詞句（例えば *New York Times*）に適用するため，コロケーションに基づいて名詞句を認定し，名詞句を「単語」と見なすことで意味ベクトルを学習した [9]．関係知識抽出では，コーパスから関係パタン（例えば *X increase the risk of Y* と項（例えば *X*: *smoking*, *Y*: *cancer*）の共起行列を構築し，関係インスタンスや関係パタンの言い換え関係（例えば *X increase the risk of Y* は *X cause Y* と近い意味）を抽出する [8, 11]．

しかしながら，関係パタンをコロケーションで抽出するのは非現実的である．例えば，*X increase the risk of Y* という関係パタンを少し改変するだけで，*X increase the major risk of Y* や *X decrease the risk of Y* というパタンが得られる．関係パタンは無数に生成できるため，「単語」として認定すべき関係パタンは組み合わせ爆発的に増大する．また，複数の語から構成される関係パタンはデータ疎問題（出現頻度の低下）が発生し，意味ベクトルの質が著しく低下する．このような問題に対処するには，関係パタンの意味を構成要素から計算する（例えば *increase* と *risk* の意味から *increase the risk of* の意味を計算する）手法が不可欠である．

構成要素から計算する単純な手法として，Skip-gram モデルにより得られる単語ベクトルの加法構成性を用いる手法がある．しかしながら，これは演算に参加する全ての単語の内容的な意味を合成してしまう．例えば *X increase the risk of Y* ≈ *X risk Y* であり，ここでの *increase* は機能的な振る舞いをする表現であるが，加法構成性を用いた場合には *X increase the risk of Y* ≈ *X risk Y* ≈ *X increase Y* となってしまう．

Socher らは構成的に意味を計算する手法として Recursive Neural Network（RNN）を提案した [13]．これは，2 単語のベクトルを結合し，重み行列を掛け合わせて親ノードのベクトルを作成する，という処理を再帰的に行う手法である．この手法では単語が行列とベクトルを持ち，機能的な振る舞いを行列で，内容的な意味をベクトルで表現する事とし，行列とベクトルの学習を同時に行う．各単語が行列とベクトルを持つ場合，学習パラメータが膨大となってしまう．Muraoka らは親ノードのベクトルを計算する際の行列について，adj-noun など係り受け関係毎に構築し，パラメータを削減する手法を提案した [10]．また，Socher らの手法は教師あり学習手法であるため，ラベルなしコーパスから単語やフレーズの意味ベクトルを獲得する事はできない．Hashimoto らは行列ではなく重みベクトルを用い，ラベルなしコーパスから単語のベクトル表現と重みベクトルを同時に学習する手法を提案した [5] が，関係パタンは対象としていない．言い換えれば，*X increase the risk of Y* ≈ *X cause Y* という意味計算が可能となるようなモデルではない．上記をふまえ本研究は，Skip-gram モデルを関係パタンの意味ベクトルの学習に向けて拡張し，以下の 2 点に貢献する．

1. 関係パタンの意味を構成的に計算する学習手法を提案する．提案手法は，関係パタンを機能的な振る舞いをする動詞列（機能的な表現列）と内容的な意味を表す単語列（内容的な表現列）に分解し，内容的な表現列は意味ベクトルの平均，機能的な表現列は RNN [13] で意味の計算を行う．すなわち，機能的な振る舞いをする表現，内容的な意味を持つ表現をあらかじめ特定し，パラメータを削減した RNN により関係パタンの意味を計算する．機能的な表現列の意味を RNN でモデル化したことにより，Hashimoto らの提案した活性／不活性のような意味的極性 [4] をはじめとして，全体の意味を変性させる機能的振る舞いを扱うことができる．
2. 既存のデータセットによる評価結果から，提案手法が関係パタンの意味を構成的に計算できること，ベースライン手法（例えばベクトル和で関係パタンの意

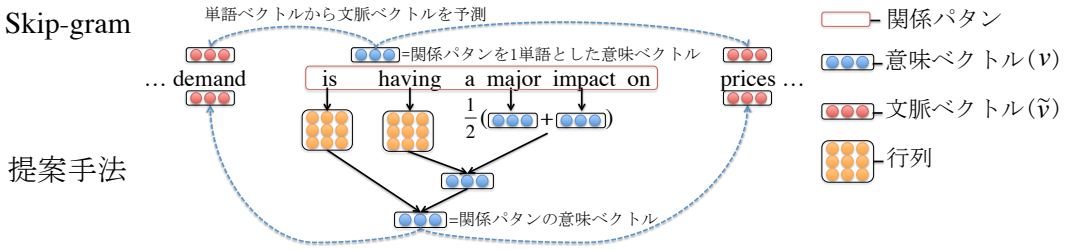

図 1: 提案手法の概要（Skip-gram モデルとの対比）

味を計算する手法）を上回る性能を示すことを実証する．重み行列の可視化結果から，提案手法が機能的な表現による意味変性を捉えていることを確認する．

2 Skip-gram モデル

まず，本研究がベースとする Skip-gram モデルを説明する．単語列からなるコーパスを $\mathcal{D} = w_1, w_2, ..., w_T$，コーパス中に含まれる単語の集合を V とする．Skip-gram モデルは，式 1 の目的関数を最小化する．

$$J = -\sum_{w \in \mathcal{D}} \sum_{c \in C_w} \log p(c|w) \tag{1}$$

ただし，C_w は単語 w の文脈単語列で，単語 w の p 個前の単語を w_{-p}，p 個後の単語を w_{+p} と表すことにすると，$C_w = \{w_{-h}, ..., w_{-1}, w_{+1}, ..., w_{+h}\}$（$h$ は文脈の広さを調整するパラメータ）である．$P(c|w)$ は単語 w から文脈単語 c を予測する確率で，log-bilinear モデルで定式化される．

$$p(c|w) = \frac{\exp(\boldsymbol{v}_w \cdot \tilde{\boldsymbol{v}}_c)}{\sum_{c' \in V} \exp(\boldsymbol{v}_w \cdot \tilde{\boldsymbol{v}}_{c'})} \tag{2}$$

ここで，$\boldsymbol{v}_w$ は単語 w のベクトル（次元数 d），$\tilde{\boldsymbol{v}}_c$ は文脈 c のベクトル（次元数 d）である [1]．式 2 の分母はコーパスに含まれる全ての単語に関する内積の和を求めるものであり，計算コストが高い．そこで，Mikolov らは Noise Contrastive Estimation [2] に基づく Negative Sampling を提案している．Negative Sampling では，観測された文脈語 c と人工的にサンプリングした k 個のノイズ（擬似負例単語 z）を識別できるようなロジスティック回帰モデルを学習する．

$$p(c|w) \approx \log \sigma(\boldsymbol{v}_w \cdot \tilde{\boldsymbol{v}}_c) + k \mathop{\mathbb{E}}_{z \sim P_n} \left[\log \sigma(-\boldsymbol{v}_w \cdot \tilde{\boldsymbol{v}}_z)\right] \tag{3}$$

ここで，P_n はノイズをサンプリングするための確率分布である．本研究では Mikolov らと同様に，コーパス中での各単語の出現確率を 0.75 乗したものを用いる．

3 関係パタンの構成的な意味計算

次に，Skip-gram モデルを関係パタンの語構成を考慮したモデルに拡張する．まず，関係パタン P は機能的な表現列 $p_1, ..., p_n$ に続いて内容的な表現列 $p_{n+1}, ..., p_m$ から構成されると考える．例えば，*is having a major impact on* という関係パタンは，図 1 に示すように，*is having* という機能的な表現列と *a major impact on* という名詞句+前置詞からなる．ここで，*a major impact on* に含まれる内容的な表現は *major* と *impact* である．したがって，関係パタン *is having a major impact on* は (p_1, p_2, p_3, p_4) =(*is*, *having*, *major*, *impact*) であり，機能的な表現列と内容的な表現列の境界は $n = 2$ で，総単語数 $m = 4$ である．なお，関係パタン，機能的な表現，内容的な表現の定義については 4.1 節で説明する．

本研究では，内容的な表現を組み合わせて関係パタンの意味を計算する際，ベクトル空間上で加法構成性が成り立つと仮定する．すなわち，内容的な表現列 $p_{n+1}, ..., p_m$ の意味は，各内容的な表現の意味ベクトルの平均で計算する．

$$\frac{\boldsymbol{v}_{p(n+1)} + \boldsymbol{v}_{p(n+2)} + ...\boldsymbol{v}_{p_m}}{m - n} \tag{4}$$

これに対し，各機能的な表現 p_i は写像 $f_{p_i} : \mathbb{R}^d \to \mathbb{R}^d$ で意味ベクトルを変換すると仮定する．したがって，関係パタン P の意味ベクトル $\boldsymbol{v}_P$ は式 5 で計算する．

$$\boldsymbol{v}_P = f_{p_1}\left(f_{p_2}\left(...f_{p_n}\left(\frac{\boldsymbol{v}_{p(n+1)} + \boldsymbol{v}_{p(n+2)} + ...\boldsymbol{v}_{p_m}}{m - n}\right)\right)\right) \tag{5}$$

写像 f_{p_i} は RNN に基づいて設計する．すなわち，機能的な表現 p_i の意味変換写像は，行列 W_i（$d \times d$ 次元）と活性化関数でモデル化する．

$$f_{p_i}(\boldsymbol{v}) = \tanh(W_{p_i}\boldsymbol{v}) \tag{6}$$

まとめると，本手法は関係パタンの内容的な表現列をベクトルの平均で合成し，機能的な表現の意味を RNN で変換するモデルである．例として，*is having a major impact on* という関係パタンの意味ベクトルは，図 1 に示したように，*major* と *impact* の意味ベクトルの平均を *having* および *is* の行列を用いて変換する事で得られる．こうして得られた関係パタンの意味ベクトルから文脈単語 *demand* や *prices* を予測できるよう学習を行う．

単語と関係パタンの意味ベクトルは，Negative Sampling に基づく Skip-gram モデルと同様の手順で学習する．ただし，本研究では以下の拡張を行う．

1. 学習するのは，単語ベクトル $\boldsymbol{v}, \tilde{\boldsymbol{v}}$ に加えて，機能的な表現の意味行列 W である．単語ベクトル $\boldsymbol{v}, \tilde{\boldsymbol{v}}$ は，

[1] 学習後に 2 種類のベクトル $\boldsymbol{v}$ と $\tilde{\boldsymbol{v}}$ が獲得されるが，$\boldsymbol{v}$ のみを単語ベクトルとして採用し，$\tilde{\boldsymbol{v}}$ は利用しない．

B* N* V A* P
B ::= *be* verb
N ::= 'not'
V ::= verb
A ::= noun | adj | adv | pronoun | det
P ::= preposition | particle

図 2: 本研究で関係パタンと認定する品詞列

関係パタンを考慮しない通常の Skip-gram モデルでの学習結果で初期化する．機能的な表現の意味行列 W の要素は標準正規分布から得た値で初期化する．

2. 関係パタンの意味ベクトルは式 5 で計算する．例えば，関係パタンから周辺の文脈語を予測する場合は，式 3 の $\boldsymbol{v}_w$ の代わりに，式 5 の $\boldsymbol{v}_P$ を用いる．関係パタンを構成する内容的な表現のベクトルと機能的な表現の行列は，誤差逆伝搬法に基づいて更新する．なお，関係パタンを周辺の単語から予測する際には 5 式を使わず，各関係パタン毎に割り当てたベクトル $\tilde{\boldsymbol{v}}_P$ を用いる．
3. 本研究では機能的な表現の意味ベクトルの計算の部分で活性化関数 tanh を適用しているため，関係パタンの意味ベクトルの各次元は [-1, 1] の範囲に収まる．一方，元々の Skip-gram モデルでは活性化関数を用いていないため，単語の意味ベクトルの値域に制限がない．本研究では，関係パタンと単語の意味ベクトルの値域を合わせるため，単語ベクトル $\boldsymbol{v}$ にも活性化関数 tanh を適用する．

4 実験

4.1 実験設定

単語の意味ベクトルや機能的な表現の行列を学習するためのコーパスとして，ukWaC[2] を利用した．このコーパスは，uk ドメインから収集した Web ページのテキストを収録しており，Tree Tagger[3] で品詞とレンマが付与されている．実験では，小文字に変換したレンマを単語とし，動詞の過去分詞形だけは表層形をそのまま用いた [4]．さらに，*a*, *b* のような 1 文字からなるトークン，*the* などの冠詞や *what* のような疑問詞などをストップワードとして除去した．

教師なし関係抽出器 Reverb [1] を参考に，図 2 の品詞パタンに合致する単語列を関係パタンとして認定した．機能的に振る舞う動詞のリストを以下の手順でマイニングした．まず，図 2 の品詞列に合致し，かつ名詞か形容詞を含む単語列の中で，20 回以上コーパス中で出現するものを抽出した．抽出された関係パタンにおいて，10 種類以上の関係パタンに含まれている動詞 961 件を機能的な表現とした．なお，否定表現が意味ベクトルの写像としてどのように捉えられるかを調べるため，*not* も機能的な表現として扱う事にした．関係パタンに含まれる単語のうち，上記で認定された 962 件を機能的な表現，それ以外の冠詞，前置詞以外を内容的な表現として扱う．

[2] http://wacky.sslmit.unibo.it
[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[4] 能動態と受動態の区別を付けるため

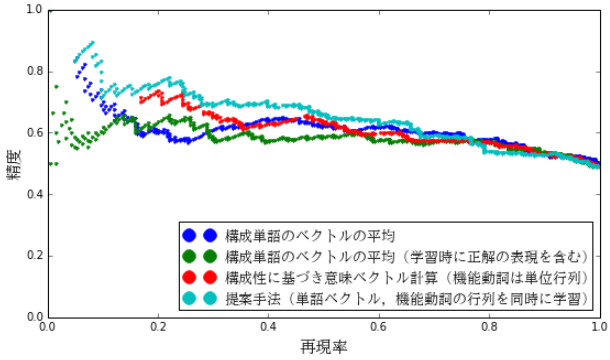

図 3: 各手法での精度と再現率

ukWaC コーパス中に出現する単語や関係パタンのうち，10 回以上出現するものを学習対象とした．なお，機能的な表現として認定された動詞（例えば *cause*）が関係パタン以外で出現した場合は，単語が出現したと見なし，通常の Skip-gram と同様に単語ベクトルを更新する．ベクトルの次元 $d = 50$，文脈の広さ $h = 5$，Negative Sampling の数 $k = 5$ とし，10^{-5} サブサンプリングで Skip-gram を学習した．

4.2 評価データ

評価データとして，Zeichner らが述部のペアに対し含意か否かを付与したデータセット [14] を用いる．このデータセットは，*prevent* と *reduce the risk of* のようなペアに同一の項を付与して関係インスタンス化（例えば *Cephalexin prevent the bacteria* と *Cephalexin reduce the risk of the bacteria*）し，関係インスタンスペアが含意関係にあるかどうかを人手で判定したものである．図 2 のルールに合致する関係パタンのペア，もしくは関係パタンと動詞のペアを抽出し，426 件のペア（内 209 ペアが含意関係）を評価対象とした．評価データに含まれる関係パタンのペアに対し，提案手法で得た意味ベクトル間のコサイン類似度を計算し，コサイン類似度がペアの含意関係をどのくらい推定できるか，適合率—再現率曲線を描く．なお，特に断りがある場合を除き，評価データに収録されている関係パタンは学習の際の語彙から除去した．すなわち，評価データに出現する関係パタンについては，単語のみを個別に学習する事とし，評価では，提案手法が内容的な表現と機能的な表現の未知の組み合わせである関係パタンの意味を合成できるか検証した．

4.3 結果

提案手法と比較手法の適合率—再現率曲線を図 3 に示した．提案手法（水色）はほとんどの領域において，他の手法よりも高い性能を達成している．関係パタンの意味ベクトルの計算に，Skip-gram モデルで学習された単

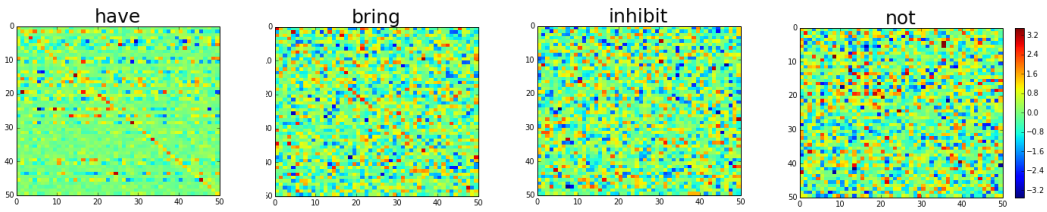

図 4: 行列の学習結果の例

語ベクトルの平均[5]を用いる手法（青色）よりも，提案手法の方が高い性能を示している．

また，機能的な表現の写像に対応する行列を単位行列に固定した場合（赤色）でも，単語ベクトルの平均を用いる手法を上回る性能を示している．今回の評価データでは，機能的な表現の意味を関係パタンに反映させず，内容語の意味ベクトルだけで関係パタンの意味を計算した方が良いことを示唆している．ただ，行列を学習する提案手法の方が高い性能を示していることから，機能的な表現を無視するのではなく，機能的な表現による意味の変性を RNN でモデル化することが有効であることが分かった．単語ベクトルの平均手法と提案手法とで類似度上位のペアを比較した際に，提案手法でのみ獲得できているペアとしては *inhibit* と *prevent the growth of* や *prevent* と *reduce the incidence of* などがあった．提案手法では，機能的な表現の行列により *growth* や *incidence* の意味を変性させる事に成功している事が分かる．

評価データに収録されている関係パタンも学習対象とし，学習済みである関係パタンについてはその関係パタン自体のベクトルを用いる手法（緑色）はほとんどの領域で最も悪い結果となった．この事から，関係パタンの意味を構成要素から計算する事が有効と言える．

4.4 行列の学習結果の可視化

図 4 に，提案手法で学習された *have*, *bring*, *inhibit*, *not* の重み行列を可視化したものを示した．*have* の行列は対角成分の値が高く，それ以外の要素は 0 に近い値になることが多く，*have* の意味変換は単位行列に近いものとなった．これは，*have* を含む関係パタン（*have acceess to* や *have an impact on*）において，その意味を内容語の部分に預けていることが多いからである．同様の傾向は *make* や *take* などの動詞にも見られた．

bring も対角成分の値が高いが，それ以外の要素にも値が割り当てられている．これは，*bring* という動詞が *bring an end to* や *bring a wealth of* などの関係パタンにおいて，機能的な側面に加えて「もたらす」という意味を追加しているためだと考えられる．これに対し，*inhibit* や *not* は *have* や *bring* とは大きく異なる行列が学習されている．これは，例えば *inhibit* は *inhibit the development of* のように，内容的な表現部分の意味を打ち消すような働きを持っているためで，重み行列が *development* の意味から *prevent* に類似した意味に写像するような役割を担っているためだと考えられる．

[5]提案手法での計算と同様，関係パタンに含まれる冠詞と前置詞以外の単語ベクトルの平均を計算した

5 おわりに

本研究では，機能的な表現を行列とし，RNN を利用する事で，関係パタンの意味を構成的に計算する手法を提案した．Skip-gram モデルの拡張として，関係パタンの意味を構成的に計算可能な行列および単語ベクトルを同時に学習する手法を提案し，実験の結果，提案手法は未知の関係パタンについても意味を構成的に計算可能である事，機能的な表現による意味の変性を捉えられる事を明らかにした．本研究では関係パタンや機能的な表現の判定は非常に簡単なヒューリスティクを用いて行っている．今後は，関係を表す単語列の特定とその意味計算を同時に行う手法を考えたい．

謝辞 本研究は，文部科学省科研費課題 26・5820 および課題 23240018 の一環として行われた．また JST 戦略的創造研究推進事業「さきがけ」から部分的な支援を受けて行われた．

参考文献

[1] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proc. of EMNLP 2011*, pages 1535–1545.
[2] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*.
[3] Z. Harris. Distributional structure. *Word*, 10(23):146–162.
[4] C. Hashimoto, K. Torisawa, S. De Saeger, J.-H. Oh, and J. Kazama. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proc. of EMNLP-CoNLL 2012*.
[5] K. Hashimoto, P. Stenetorp, M. Miwa, and Y. Tsuruoka. Jointly learning word representations and composition functions using predicate-argument structures. In *Proc. of EMNLP 2014*.
[6] O. Levy and Y. Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proc. of CoNLL 2014*.
[7] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *NIPS 27*, pages 2177–2185.
[8] D. Lin and P. Pantel. DIRT - discovery of inference rules from text. In *Proc. of SIGKDD 2001*.
[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 26*, pages 3111–3119.
[10] M. Muraoka, S. Simaoka, K. Yamamoto, Y. Watanabe, N. Okazaki, and K. Inui. Finding the best model among representative compositional models. In *Proc. of PACLIC 28*.
[11] N. Nakashole, G. Weikum, and F. M. Suchanek. PATTY: A taxonomy of relational patterns with semantic types. In *Proc. of EMNLP-CoNLL 2012*, pages 1135–1145.
[12] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proc. of EMNLP 2014*, pages 1532–1543.
[13] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proc. of EMNLP-CoNLL 2012*.
[14] N. Zeichner, J. Berant, and I. Dagan. Crowdsourcing inference-rule evaluation. In *Proc. of ACL 2012 (short papers)*.