構成性と非構成性を同時に考慮した動詞句の表現学習

橋本和真　鶴岡慶雅
東京大学 工学系研究科
{hassy,tsuruoka}@logos.t.u-tokyo.ac.jp

1　はじめに

自然言語処理の分野では, word2vec [7] に代表されるように, 単語を実数値ベクトルとして学習する手法が盛んに研究されている. 大規模コーパス上での共起統計により学習された単語ベクトルは, 様々なタスクに応用されている. 自然言語を表現するうえでは単語だけでなく句も同様に扱うことができると有用である. そこで最近では, 句に関する共起統計を利用した句ベクトルの学習手法も注目されている [3, 9]. これらの研究は, 単語の共起により単語のベクトルが学習できるのであれば, 句に関しても共起情報によって同様に学習ができるという考え方に基づいている.

句をベクトルとして表現するアプローチとしては, 構文情報に基づいた行列演算によるもの [3] や, 単語ベクトルの加算によるもの [9] が存在する. いずれの手法も, 句の表現は構成要素の単語の表現から計算することができるという仮定をおいている. そのような表現を**構成的表現**と呼ぶ. しかし, 全ての句を構成的表現で扱ってよいわけではない. 例えば, 動詞句 “use computer” の意味は各単語の意味から容易に推定できる. その一方で, “take part” などのように, 各単語の意味から全体を考えるのでなく, イディオムとして使われることが多い句も存在する. イディオムや定型表現を全て統一して構成的表現として扱うことは, 構成的表現の学習の精度の低下につながると考えられる. 定型表現などは, それ自体をひとまとまりとして扱ってベクトル化する**非構成的表現**のほうが適している.

そこで本研究では, 構成的表現と非構成的表現の間のバランスをとり, それらを同時学習する手法を提案する. 各句ごとに構成的表現と非構成的表現に対する重みを計算し, それによる両者の重み付き和によって句ベクトルを計算する. その重みの計算モデルは句ベクトルと同時に学習されるため, 辞書などの事前知識を要しない. 実験では, 動詞句の表現学習に提案手法を適用した結果, 他動詞の語義曖昧性解消タスクにおいて最高精度を達成したモデルの改善に成功した. さらに, 自動で学習した構成的表現に対する重みを, 人手の評価と比較したところ, 先行研究と同等以上の相関係数を示すことがわかった.

2　構成的・非構成的表現の同時学習

2.1　構成的表現の学習

単語ベクトルを学習する研究に加え, 句ベクトルを学習する研究も注目されている [3, 7, 9]. 行列演算を介するもの [3] や, 単語ベクトルの加算を用いるもの [7, 9] など様々な手法が提案されている. それらの多くは

> 句の意味はその構成要素の単語の意味の組み合わせによって決まる

という句の意味構成の仮定に基づいている.

意味構成に関して単語ベクトルと句ベクトルを同時学習する手法 [3, 9] では, ある長さ L の句 $phr = (w_1, w_2, \cdots, w_L)$ の構成的表現を n 次元ベクトル $\boldsymbol{c}(phr) \in \mathbb{R}^{d\times 1}$ として

$$\boldsymbol{c}(phr) = f(\boldsymbol{v}(w_1), \boldsymbol{v}(w_2), \cdots, \boldsymbol{v}(w_L)) \qquad (1)$$

のように計算する. ここで, $\boldsymbol{v}(w_i) \in \mathbb{R}^{d\times 1}$ は単語 w_i の単語ベクトルであり, f はそれらを用いて句ベクトル $\boldsymbol{c}(phr)$ を計算する意味構成関数である. 単語ベクトルや関数 f を構成する全パラメータは, ある目的関数を最小化することによって学習される. 例として, 大規模コーパス中で句と単語の共起統計に関する目的関数を設計して学習する手法がある [3, 9].

2.2　非構成的表現の学習

非構成的表現の学習に関しては, 特定の句を一まとまりとして考えてベクトル化する手法が研究されている [7]. つまり, 式 (1) と対比して, 句のベクトルが固有の d 次元ベクトル $\boldsymbol{p}(phr) \in \mathbb{R}^{d\times 1}$ としてパラメータ化され, 句の構成要素の個々の単語の情報は考慮されない. 例えば, “New York” などを一単語として扱ってベクトル化することができる. これにより, “New” と “York” それぞれの単語の意味から「ニューヨーク」という意味を構成する必要がなくなる.

Mikolov ら [7] は固有名詞に類する名詞句に主に焦点を当てて非構成的表現を学習したが, 動詞句などにも同じ考えが適用できると考える. その際, 意味構成関数を用いるべき場合とそうでない場合を区別する必要がある. しかし, 先行研究では構成的表現と非構成的表現の比較をすることしか行われていない.

2.3 同時学習

本研究では, 2.1, 2.2 節をふまえて, 構成的表現と非構成的表現のバランスをとったベクトル表現の学習手法を提案する. 具体的には, ある句の構成的表現 $\boldsymbol{c}(phr)$ と非構成的表現 $\boldsymbol{p}(phr)$ を用いて, その句のベクトル表現 $\boldsymbol{v}(phr) \in \mathbb{R}^{d\times 1}$ を

$$\boldsymbol{v}(phr) = \alpha(phr)\boldsymbol{c}(phr) + (1-\alpha(phr))\boldsymbol{p}(phr) \quad (2)$$

として計算する. $\boldsymbol{c}(phr)$ と $\boldsymbol{p}(phr)$ のノルムのバランスをとるため, それぞれをノルム 1 に正規化した後に式 (2) を適用する手法も試みる. ここで, $\alpha(phr) \in [0,\ 1]$ は各句ごとに計算される実数値であり, 意味構成性の度合いを表す値である. つまり,

- $\alpha(phr)$ が 1 に近いときは構成的表現を重視し,
- $\alpha(phr)$ が 0 に近いときは非構成的表現を重視する

ということになる. この $\alpha(phr)$ の計算に関してもパラメータ化し, 全て対象のタスクに応じて同時に学習を行う. これにより, 各句の意味構成性の度合いを自動で学習することができる.

本稿では, 非構成的表現で扱う句の候補はあらかじめ与えられているとする. その候補に入っていない句に関しては $\alpha(phr) = 1$, つまり $\boldsymbol{v}(phr) = \boldsymbol{c}(phr)$ として計算する. 初見の句に関して個々の単語から全体の意味を解釈しようとすることは自然である.

$\alpha(phr)$ の計算方法に関しては,

1. $\alpha(phr) = 0.5$ (**Half**)
2. $\alpha(phr) = \sigma(\boldsymbol{W}\cdot\boldsymbol{\phi}(phr) + b)$ (**Feature**)
3. $\alpha(phr) = \sigma(\boldsymbol{c}(phr)\cdot\boldsymbol{p}(phr))$ (**Dot**)
4. $\alpha(phr) = \sigma(\boldsymbol{W}\cdot\boldsymbol{\phi}(phr) + b + \boldsymbol{c}(phr)\cdot\boldsymbol{p}(phr))$ (**Feature-Dot**)

の 4 種を試みる. ここで, σ はロジスティック関数であり, $\alpha(phr)$ を $[0,\ 1]$ の範囲にする際に用いる. また, $\boldsymbol{\phi}(phr)$ は phr に関する素性ベクトルであり, $\boldsymbol{W}$ はそれに関する重みベクトル, b はバイアス項である. Half では $\alpha(phr)$ の値は 0.5 で常に一定とする. Feature では $\alpha(phr)$ は phr 自身やその構成要素の単語に関して設計した素性を基に計算される. Dot では $\alpha(phr)$ は $\boldsymbol{c}(phr)$ と $\boldsymbol{p}(phr)$ の類似度スコア (内積) によって計算される. Dot の解釈としては, $\boldsymbol{c}(phr)$ と $\boldsymbol{p}(phr)$ の類似度が高ければ高いほど, 全体を一まとまりとした表現と, 個々の単語を考慮した表現が近い, つまり意味構成性の度合いが高いということである. Feature-Dot は両者の組み合わせである.

ここで新たに導入される $\boldsymbol{W}$ や b もモデル全体の目的関数を最小化する際に誤差逆伝播法によって同時に学習される. 学習における正則化に関しては, $\lambda_W(\|\boldsymbol{W}\|^2 + b^2)$ と $-\lambda_\alpha \log\alpha(phr)$ の二つを導入した. 前者は重みベクトルの L2 ノルム正則化であり, 後者は構成的表現を重視する正則化である. これは, 多くの句が構成的であるという仮定に基づく. また, λ_W と λ_α は正則化の強さを決めるハイパーパラメータである.

3 他動詞句の意味構成性の自動検出

本稿では, 提案手法の有効性を確認するための試金石として他動詞句に着目した実験を行う. 我々は, コーパス中における主語-動詞-目的語の共起統計と, 他動詞句-前置詞-名詞の共起統計に基づいて, 動詞句のベクトル表現を学習する手法を提案した [3]. 目的関数は, その 2 種の組み合わせに関して, コーパス中に出現するものに関してはスコアが高くなるように, 出現しないものに関してはスコアが低くなるように設計されている. つまり, コーパス中で出現するものとしないものを識別するタスクになっている.

我々の手法 [3] では, 主語 S, 動詞 V, 目的語 O が与えられたとき, SVO の動詞句ベクトルが

$$\boldsymbol{v}(SVO) = \boldsymbol{v}(S)\odot\boldsymbol{v}(VO) \quad (3)$$
$$\boldsymbol{v}(VO) = \boldsymbol{M}(V)\boldsymbol{v}(O) \quad (4)$$

と計算され, 上記のタスクで学習される. ここで, 他動詞 V は行列 $\boldsymbol{M}(V) \in \mathbb{R}^{d\times d}$ であり, 主語 S と目的語 O は d 次元ベクトル $\boldsymbol{v}(S), \boldsymbol{v}(O) \in \mathbb{R}^{d\times 1}$ である. また, $\odot$ はベクトルの要素積を表す.

本稿では式 (4) に提案手法を適用 ($phr = VO$) し,

$$\boldsymbol{v}(VO) = \alpha(VO)\boldsymbol{c}(VO) + (1-\alpha(VO))\boldsymbol{p}(VO) \quad (5)$$

として計算する. ここで, $\boldsymbol{c}(VO) = \boldsymbol{M}(V)\boldsymbol{v}(O)$ であり, $\alpha(VO) = 1$ とすると元の手法と同じになる.

3.1 実験設定

学習には, British National Corpus (BNC) を構文解析器 Enju[1] によって解析した結果から抽出した 138 万

[1] http://kmcs.nii.ac.jp/enju.

事例の主語-動詞-目的語, 93 万事例の他動詞句-前置詞-名詞の組み合わせを用いた [2]. 目的関数の最適化や, 開発データを用いたハイパーパラメータの調整は我々の以前の実験設定 [3] と同様にした. 非構成的表現で扱う句の候補は, 学習コーパス中に出現する動詞-目的語のうち出現頻度上位 50,000 句とした.

本稿では単語ベクトルの次元 d は 25 で固定とし, 正規乱数によって初期化を行った. また, Feature, Feature-Dot における $\alpha(VO)$ の計算のためのパラメータ $\boldsymbol{W}$, b は 0 で初期化した. そのなかで, $\alpha(VO)$ の計算方法と正規化の有無の設定を変えて実験を行った. AdaGrad [1] の学習率の候補は $\{0.05, 0.06, 0.07, 0.08, 0.09, 0.1\}$ とし, 正則化係数 λ_W, λ_α の候補は $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 0\}$ とした. その結果, 実験設定の数は合計 1,272 となった.

$\alpha(VO)$ の計算の Feature, Feature-Dot における素性は, McCarthy ら [6] が用いている動詞-目的語の出現頻度, 動詞-目的語の自己相互情報量に加え, 動詞, 目的語, 動詞-目的語の番号 (0/1 の素性) を用いた.

4 動詞句の意味構成性の検出に関する評価

4.1 評価用データセット

McCarthy ら [6] が提供しているデータセットでは, 638 組の動詞-目的語に関して, 二人の英語話者によって意味構成性の度合いが 1 から 6 の 6 段階でスコア付けがなされている. 例えば, “use computer” には最高スコアの 6 が付けられており, “take part” には 1 と 2 のスコアが付けられている.

このデータセットにつけられているスコアと, 提案手法で算出した $\alpha(VO)$ の相関を調べるために, スピアマンのランク相関係数を用いて評価を行った. つまり, その相関係数が高ければ高いほど, $\alpha(VO)$ の値の傾向が人間の感覚と合っているということになる.

4.2 結果と考察

表 1 に結果を示す. 提案手法の結果に加えて, McCarthy ら [6] の結果も示した. 出現頻度と自己相互情報量によるスコアは, それらによって降順に並べた場合の順位を用いてデータセットとの相関を測った結果である. これを見ると, 提案手法の Feature-Dot (正規化有り) の場合に先行研究と同等以上のスコアを達成したことがわかる. DSPROTO は, 自動構築したシソーラスに基づく単語の類似度情報 [5], 係り受け解析

[2]学習データは `http://www.logos.t.u-tokyo.ac.jp/~hassy/publications/cvsc2015/` で公開されているもの [3] を用いた. 学習データと, パラメータ調整用の開発データの分け方は `https://github.com/hassyGo/SVOembedding` に公開されているコードに従った.

	正規化	有り	無し
提案手法	Feature	0.350	0.088
	Dot	0.116	0.105
	Feature-Dot	**0.414**	0.064
出現頻度 [6]		0.141	
自己相互情報量 [6]		0.274	
DSPROTO [6]		0.398	

表 1: 動詞句の意味構成性の検出タスクのスコア.

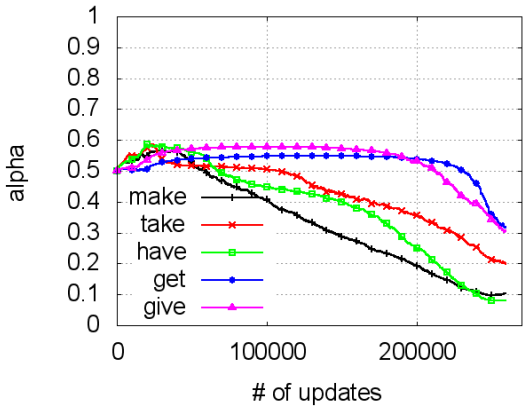

図 1: 軽動詞の意味構成性への寄与分の推移.

結果 (動詞-目的語, 名詞-名詞, 形容詞-名詞の関係) などを利用した手法である. 本研究と DSPROTO で用いられている情報は類似しているが, 我々の手法は意味構成性の度合いの指標の計算と共に, 動詞句のベクトルも同時に学習することが出来ている点が異なる. また, 名詞句など他種の句ベクトルの学習, 他のタスクを利用した学習など, 拡張性も高い.

提案手法での結果を見ると, $\boldsymbol{c}(VO)$ と $\boldsymbol{p}(VO)$ の正規化が効果的であることがわかる. 予備実験において, $\boldsymbol{c}(VO)$ と $\boldsymbol{p}(VO)$ の大きさに非常に大きな差が生じていることを確認したが, そのような場合に $\alpha(VO)$ を重要度として用いてバランスをとることは困難である. そこで正規化をすることでベクトルの大きさの極度な不均衡を避けることが出来るため, 学習が効果的に行われたと考えられる. Dot を用いた場合のスコアは低いが, Feature と合わせる (Feature-Dot) ことでスコアが向上した. つまり, 今回用いた素性では捉えられない有用な情報が Dot によって付加されたと考えられる. また, Feature-Dot (正規化有り) において, 構成性を重視するための正則化係数 λ_α は 0 でなく 10^{-3} が選択されており, ある程度の有効性がうかがえる.

次に, 学習中の $\alpha(VO)$ の変化の様子を確認した.

	正規化	有り	無し
提案手法	Half	0.462	0.588
	Feature	0.540	0.554
	Dot	0.573	0.544
	Feature-Dot	0.531	**0.593**
$\alpha(VO) = 1$ [3]		0.574	
Milajevs ら [8]		0.456	

表 2: 他動詞の語義曖昧性解消タスクのスコア.

Feature-Dot (正規化有り) で学習した後, $\alpha(VO)$ を計算する際に, 動詞のインデックスの素性のみを用いた場合の $\alpha(VO)$ の変化の様子を図 1 に示す. 横軸は学習中のパラメータの更新回数である. これは, 各動詞が意味構成性の度合いの計算にどのように寄与するか, ということを示していると解釈できる. ここで選択された動詞は軽動詞と呼ばれるものであり, イディオム表現を作りやすい動詞である. 軽動詞に関しては $\alpha(VO)$ を小さくする傾向が見て取れるため, 期待通りの挙動をしている. さらに, 全ての素性などを用いて $\alpha(VO)$ を計算した結果, α(“take part”) ≈ 0, α(“make payment”) $= 0.25$, α(“make money”) $= 0.56$, α(“use computer”) $= 0.96$ となり, 様々な動詞句によってその意味構成性の度合いが異なる結果が確認できた.

5 動詞の語義曖昧性解消タスクによる評価

5.1 評価用データセット

Grefenstette と Sadrzadeh [2] が提供しているデータセットは他動詞の語義曖昧性解消に関する評価を行うものである. 具体的には, 他動詞の組に対して共通の主語と目的語を与えたときに, その意味的な類似度がどれだけ高いか, ということを人手でスコア付けしたものである. 例えば, 他動詞 “write” と “spell” が共通の主語 “student” と目的語 “name” を伴った場合には, その意味が近いということで高いスコアが与えられている. このようなデータが 200 事例あり, 人手のスコアとシステムによって出力したスコアのスピアマンのランク相関係数を用いて評価を行った. 本研究では式 (3), (5) を用いて計算した主語-動詞-目的語を表すベクトル間のコサイン類似度をスコアとして用い, 人手のスコアとの相関を測った.

5.2 結果と考察

表 2 に結果を示す. 提案手法による結果に加え, 我々が以前に報告した結果 [3] ($\alpha(VO) = 1$ に相当) と Milajevs ら [8] の結果を掲載している. この結果からわかる通り, Feature-Dot (正則化無し) によって最高スコアを達成した. また, Half (正則化無し) の場合にも $\alpha(VO) = 1$ の場合を上回っており, 構成的表現と非構成的表現を併用することの有用性が示唆された.

結果の傾向として, $\boldsymbol{c}(VO)$ と $\boldsymbol{p}(VO)$ の正則化をしないほうがスコアが高いことが多いといえる. この理由としては, 式 (3), (4) で SVO の句ベクトルを計算する意味構成関数 [4] が, ベクトルの正規化などを仮定していないということが考えられる. 理想的には, 基礎とする意味構成関数との親和性なども考慮し, 表 1, 2 の傾向が一致することが望ましい.

6 おわりに

本稿では, 動詞句の表現学習に焦点を当てて, 句の構成的表現と非構成的表現を併用した同時学習手法を提案した. 実験では, 意味構成性の度合いの指標 $\alpha(phr)$ が, 人手の評価と大きな相関を示すことがわかった. また, 学習した動詞句ベクトルを用いることで, 他動詞の語義曖昧性解消のタスクにおいても最高スコアを達成した. 今後は, さらに有力な $\alpha(phr)$ の計算方法の提案や, 対象の句の種類を増やすことなどを試みる.

謝辞

本研究は、JST、CREST の支援を受けたものである。

参考文献

[1] J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR*, 12:2121–2159, 2011.

[2] E. Grefenstette and M. Sadrzadeh. Experimental Support for a Categorical Compositional Distributional Model of Meaning. In *EMNLP*, 2011.

[3] K. Hashimoto and Y. Tsuruoka. Learning Embeddings for Transitive Verb Disambiguation by Implicit Tensor Factorization. In *CVSC*, 2015.

[4] D. Kartsaklis, M. Sadrzadeh, and S. Pulman. A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments. In *COLING*, 2012.

[5] D. Lin. Automatic Retrieval and Clustering of Similar Words. In *ACL/COLING*, 1998.

[6] D. McCarthy, S. Venkatapathy, and A. Joshi. Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *EMNLP/CoNLL*, 2007.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 2013.

[8] D. Milajevs, D. Kartsaklis, M. Sadrzadeh, and M. Purver. Evaluating Neural Word Representations in Tensor-Based Compositional Settings. In *EMNLP*, 2014.

[9] N. T. Pham, G. Kruszewski, A. Lazaridou, and M. Baroni. Jointly optimizing word representations for lexical and sentential tasks with the C-PHRASE model. In *ACL/IJCNLP*, 2015.