4K1-2

意味と構造の統一的なカーネル埋め込みによる非線形類似度学習

Non-linear Similarity Learning with Kernel Embeddings
for Compositional Semantics and Structure Representations

椿真史 *1 Masashi Tsubaki　Kevin Duh*1 Kevin Duh　新保仁 *1 Masashi Shimbo　松本裕治 *1 Yuji Matsumoto

*1奈良先端科学技術大学院大学
Nara Institute of Science and Technology (NAIST)

NLP applications rely on existence of a good similarity over their text data. Semantic vector spaces with distributed model provide a good similarity between words. However, such spaces fail to capture composed phrasal and sentential similarities. In this work, we propose a new method of non-linear similarity learning for compositionality. Ours can learn new word representations through similarity learning with kernels taking into account the non-linearity in compositional semantics. We evaluated our method on the prediction task of similarity between two sentences, and achieved the state-of-the-art without feature engineering and deep recursive neural networks.

1. はじめに

自然言語処理において，何らかの二つの対象 (単語や句，文や文書など) の類似度を計算することは重要である．この類似度に基づき，情報の検索や抽出，分類やクラスタリングを行う．従来手法では，対象の表層的な特徴を用いることが多いが，これには言語が持つ意味的な情報を無視しているという大きな問題がある．しかし近年では意味研究が盛んに行われ，単語ベクトル空間モデルはその基礎を提供する．

単語ベクトル空間モデルは，単語の意味的な情報をベクトル空間において表現する一般的な枠組みである．古くは潜在意味解析，近年ではニューラルネットワークを用いた手法が提案されている．しかし単語ベクトル空間のみでは，句や文といったより複雑な意味をどのように表現するのかという問題が残る．この問題を解決するために，非線形関数を階層的に適用して句や文の意味表現を構成する再帰的ニューラルネットワーク [Socher 12, Socher 14] が，特に近年注目されている．

一方で機械学習では，距離学習と呼ばれる研究分野が存在する．この研究は，すでに得られているデータ間の距離を，タスクに合わせて処理しやすい距離へ変換することを目的とする．例えば，同じラベルを持つデータ間の距離は近く，異なるラベルを持つデータ間の距離は遠くなるように，ベクトル自体を変換する [Xing 02]．また，類似度学習においては同様のモチベーションで，主に内積を学習する [Chechik 09]．さらに，データをカーネルを用いて高次元空間へ写像した後，その空間のユークリッド距離を学習する手法も提案されている [Kedem 12]．適切な距離ないしは類似度を持つベクトル空間の学習は，機械学習の多くの問題において最も重要であり，カーネルを用いたその非線形拡張は特に近年注目されている．

本稿では，前述した単語ベクトル空間における意味構成と類似度学習という 2 つの研究を踏まえ，以下の問題に焦点を当てる．

- 文をその意味と構造を含めてどのように統一的に表現するか？そして文の類似度をどのように学習するか？

この問題を解決するために我々は，以下の仮説を立てる．

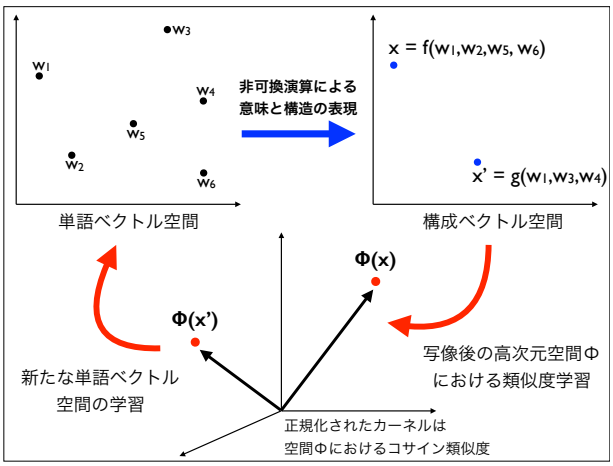

図 1: 我々は，個々の単語ベクトルから構成された 2 つの文ベクトル $\mathbf{x}$ と $\mathbf{x}'$ を，カーネルによって高次元空間 ϕ に写像した後に，それらの類似度を学習する．この空間 ϕ は，単語から文の構成に伴って生じる，より複雑な意味を表現する高次元空間と捉えることができる．最終的に我々は，文の意味を適切に構成するような新たな単語ベクトル空間を得る．

- ベクトル空間において，意味の情報はベクトルで，構造の情報はベクトル間の演算によって表現されうる．そして文は，単語とは異なる高次元空間において表現されうる．

ここで我々は，文の意味と構造を非可換演算によってベクトル空間で表現した上で，カーネルを用いた非線形類似度学習を適用する手法を提案する．従来手法では，依存構造や木構造を詳細に考慮した上で，複雑に文の表現を計算している [Socher 12, Socher 14]．さらにこの時，文は単語と同一次元の固定された空間で表現される *1．一方で我々の提案法ではまず，文の意味と構造の情報を，非可換演算の性質を利用してシンプルに

連絡先: 椿真史，奈良先端科学技術大学院大学，
masashi-t@is.naist.jp

*1　これは主に，文のデータ構造と再帰的ニューラルネットワークの持つ計算上の制約による．

表現する．次に，その文表現をカーネルを用いて高次元空間 ϕ へ写像し，その空間において類似度学習を適用する．そして最終的に，詳細な文ベクトルを陽に計算することなく，適切な文の意味構成のための新たな単語表現を獲得する．図 1 に提案法の全体像を示す．

本稿の貢献は以下の通りである．

1. 非可換演算の性質と，表現力の高い高次元空間の性質の双方を活かし，カーネルを用いた意味と構造の非線形類似度学習法を提案した．
2. 提案法はシンプルかつ実装も容易ながら，文の意味的類似度評価データセットにおいて，世界最高性能に迫る結果を出すことに成功した．

2. 提案法

訓練データは，$\{(S_i, S'_i), y_i\}_{i=1}^n$ の形式で与えられる（3.1.1 節）．S と S' は文，$y \in [-1, +1]$ はその類似度を表す．目標は，新たな文の類似度を予測することである．我々はまず，文における意味と構造の情報の，非可換演算による表現法を幾つか提案する (2.1 節)．次に，文の類似度計算をカーネルを用いて非線形に拡張する (2.2 節)．最後に，本稿で設計するカーネルと，非線形類似度学習について述べる (2.3 節)．

2.1 文の意味と構造のベクトル表現

まず最もシンプルな文のベクトル表現として，文 S のベクトル $\mathbf{x}$ を，

$$\mathbf{x} = f_{ADD}(S) = \sum_{w \in S} \mathbf{d}(w) \tag{1}$$

と計算する．ここで，$\mathbf{d}(w)$ は単語 w の n 次元ベクトル表現とする．この計算法は，文内に現れるすべての単語の共起情報を考慮することができる反面，N-gram や係り受け関係などの系列や構造の情報は一切考慮できない欠点がある．そこで次に，文ベクトル $\mathbf{x}$ を，D_S を文 S 内の係り受け関係にある単語ペアの集合とした上で，

$$\mathbf{x} = f_{SUBT}(D_S) = \sum_{(w_i, w_j) \in D_S} (\mathbf{d}(w_i) - \mathbf{d}(w_j)) \tag{2}$$

と計算する．ここで，w_i と w_j は係り受け関係にある単語ペアである．これは，最も基本的な非可換演算である減算を用いることで，文内の係り受け関係にある単語間 w_i と w_j の順序情報をエンコードする．ただし減算の場合，$(a-b)+(c-d)$ と $(a-d)+(c-b)$ が等価となるため，係り受け関係の情報を厳密に保持することはできない．そこで 3 つ目の文の表現法として $\mathbf{x}$ を，

$$\mathbf{x} = f_{DIV}(D_S) = \sum_{(w_i, w_j) \in D_S} \frac{\mathbf{d}(w_i)}{\mathbf{d}(w_j)} \tag{3}$$

と計算する [*2]．除算も減算と同様の非可換演算であるが，前述した減算のような問題が生じることはなく，より構造の情報を保持したベクトル表現が得られると期待できる．そして最後の表現法として，ベクトル間の連結演算，

$$\mathbf{x} = f_{CONC}(D_S) = \sum_{(w_i, w_j) \in D_S} [\mathbf{d}(w_i); \mathbf{d}(w_j)] \tag{4}$$

を用いて文ベクトルを計算する．本稿では，これら 4 つの非可換演算を用いて意味と構造の情報を低次元空間で表現した上で，後述するカーネルを用いた非線形類似度学習法を適用し，高次元空間においてより詳細に意味と構造を最適化する．

2.2 カーネルによる類似度計算

まず我々は，意味的類似度計算のためのカーネル関数 K に，自然言語処理において幅広く用いられる，線形カーネルのコサイン類似度 K_{cos} を用いる．

$$K_{\cos}(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}'}{\sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}}\sqrt{\mathbf{x}'^{\mathrm{T}}\mathbf{x}'}} \tag{5}$$

以降，本稿で述べるすべてのカーネルは正規化されているものとし，以下のように表現する．

$$K_{\cos}(\phi(\mathbf{x}), \phi(\mathbf{x}')) = \frac{K(\mathbf{x}, \mathbf{x}')}{\sqrt{K(\mathbf{x}, \mathbf{x})}\sqrt{K(\mathbf{x}', \mathbf{x}')}} \tag{6}$$

ここで ϕ は，次に述べる非線形カーネルによって写像される高次元空間である．つまり我々は，正規化されたカーネルを用いることで，高次元空間 ϕ においても適切な意味的類似度であるコサイン類似度を考えることができる．

本稿で用いる非線形カーネルは，以下の多項式カーネル K_{poly} と RBF カーネル K_{rbf} [*3] の 2 つである．

$$\begin{aligned} K_{poly}(\mathbf{x}, \mathbf{x}') &= (c + K_{\cos}(\mathbf{x}, \mathbf{x}'))^p \\ s.t &\quad c \geq 0, \quad p \in \mathbb{N} \end{aligned} \tag{7}$$

$$\begin{aligned} K_{rbf}(\mathbf{x}, \mathbf{x}') &= \exp\left(-\frac{1 - K_{\cos}(\mathbf{x}, \mathbf{x}')}{\sigma^2}\right) \\ s.t &\quad \sigma \geq 0 \end{aligned} \tag{8}$$

2.3 非線形類似度学習

提案する非線形類似度学習に用いるカーネルは，本稿では以下の 4 つとする．

$$K(S, S') = K(f_{ADD}(S), f_{ADD}(S')) \tag{9}$$

$$\begin{aligned} K(S, S') &= K(f_{ADD}(S), f_{ADD}(S')) \\ &\times K(f_{SUBT}(D_S), f_{SUBT}(D_{S'})) \end{aligned} \tag{10}$$

$$\begin{aligned} K(S, S') &= K(f_{ADD}(S), f_{ADD}(S')) \\ &\times K(f_{DIV}(D_S), f_{DIV}(D_{S'})) \end{aligned} \tag{11}$$

$$\begin{aligned} K(S, S') &= K(f_{ADD}(S), f_{ADD}(S')) \\ &\times K(f_{CONC}(D_S), f_{CONC}(D_{S'})) \end{aligned} \tag{12}$$

そして，ロス関数は以下の通りである．

$$L(\Theta) = \sum_{i=1}^{n} \frac{1}{2}\{y_i - K(S_i, S'_i)\}^2 + \frac{\lambda}{2}||\Theta||^2 \tag{13}$$

*2 ここでの除算は，ベクトルの各要素に対して適用する element-wise な演算とする

*3 一般的に用いられる RBF カーネルはユークリッド距離を用いた $K_{rbf}(\mathbf{x}, \mathbf{x}') = \exp\left(-||\mathbf{x} - \mathbf{x}'||^2/2\sigma^2\right)$ である．これはまず $||\mathbf{x} - \mathbf{x}'||^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x} + \mathbf{x}'^{\mathrm{T}}\mathbf{x}' - 2\mathbf{x}^{\mathrm{T}}\mathbf{x}'$ と内積を用いて書き下すことができる．次に，これらの内積を任意のカーネルに置き換えることが可能であるため，すべてをコサイン類似度に置き換え $||\mathbf{x} - \mathbf{x}'||^2 = K_{\cos}(\mathbf{x}, \mathbf{x}) + K_{\cos}(\mathbf{x}, \mathbf{x}) - 2K_{\cos}(\mathbf{x}, \mathbf{x}')$ を得る．そして最終的に，$||\mathbf{x} - \mathbf{x}'||^2 = 2 - 2K_{\cos}(\mathbf{x}, \mathbf{x}')$ が得られるため，本稿での RBF カーネルは式 (8) となることに注意されたい．

$L(\Theta)$ は，データセットで与えられた文の類似度 y とカーネルとの正則化付き二乗誤差である．Θ は学習するパラメータ集合であり，単語ベクトルとカーネル内パラメータ (多項式カーネルの場合は c，RBF カーネルの場合は σ) である．

我々はカーネル関数を用いることで，2.1 節で述べた意味と構造の情報を非可換演算によってエンコードしたベクトルを，より表現力の高い高次元空間 ϕ へを写像することができる．この空間 ϕ においては，低次元ベクトルに表現された意味と構造の情報が，より適切に学習されることが期待できる．最終的に，高次元空間に写像されたベクトル $\phi(\mathbf{x})$，つまり詳細な文ベクトルを陽に計算することなく，カーネルを通してその類似度のみを学習することで，新たな単語ベクトル表現のみを陽に獲得する (図 1).

3. 実験結果と考察

3.1 実験

3.1.1 データセット

提案法は，SemEval 2014 の Sentences Involving Compositional Knowledge (SICK) [Marelli 14] のデータセット [*4] を用いて評価した．このデータセットは，2 つの文の意味的な類似度を人手でスコアリングしたものであり，訓練データとテストデータは各々約 5000 文対から成る．評価には，提案法によって計算された二つの文ベクトルの類似度と人手の類似度スコアとの，ピアソンの相関係数 r とスピアマンの相関係数 ρ，そして平均二乗誤差 (MSE) を用いる [*5]．我々の目標は，単語ベクトル表現を用いた意味構成モデルによって，新たに与えられた文の意味的類似度を正確に予測することである．

3.1.2 比較する既存研究

既存研究では主に，2 つの文に含まれる単語や N-gram のマッチングやオーバーラップ，品詞や木構造のアライメント，さらには WordNet などの外部知識などを用いて様々な素性を考え，それらを用いてサポートベクター回帰で学習するものが一般的である．SemEval2014 の SICK に関しても，同様の素性エンジニアリングの手法に基づいたアプローチが多数を占めている（Illinois-LH_run1 [Lai 14] UNAL-NLP_run1 [Jimenez 14] Meaning_Factory_run1 [Bjerva 14] ECNU_run1 [Zhao 14]）．特に，単語ベクトル空間と構成モデルのみを用いた意味的なアプローチのみでは，相関係数が 0.7 程度に留まるという報告がある [Marelli 14]．一方で，Deep Neural Network を用いた手法も幾つか提案されている [Socher 14, Tai 15]．これらは主に，Recursive Neural Network あるいは Recurrent Neural Network を用いて，単語ベクトルから文ベクトルを直接計算するモデルである．文の依存構造や木構造を考慮した上で多数の重み行列や非線形関数を階層的に適用し，低次元の文表現を学習する手法となっている．

3.1.3 実装の詳細

提案法で再学習する単語ベクトル表現については，初期値として 300 次元の Global vector [Pennington 14] [*6] を用いた．また構文解析には Enju [*7] を用いた．

最終的なコスト関数は式 (13) の $L(\Theta)$ であり，これを最小化する．最適化には AdaGrad [Duchi 11] を用いた．単語ベクトルの学習率は $\alpha = 10^{-1}$，カーネル内パラメータの学習率は $\beta = 10^{-3}$，正則化項については $\lambda = 10^{-6}$ とした．データセットに対してはイテレーション数を上限 1000 に統一し実験を行い，比較検証した．

カーネル	r	ρ	MSE
ADD			
コサイン	0.7674	0.7460	0.4678
多項式 (p=2)	0.8304	0.7818	0.3195
多項式 (p=3)	0.8384	0.7839	0.3120
多項式 (p=4)	0.8385	0.7836	0.3130
RBF	0.8356	0.7817	0.3166
ADD × SUBT			
コサイン	0.7521	0.7445	0.5342
多項式 (p=2)	**0.8414**	**0.7886**	**0.3040**
多項式 (p=3)	0.8410	0.7870	0.3073
多項式 (p=4)	0.8387	0.7845	0.3136
RBF	0.8373	0.7860	0.3127
ADD × DIV			
コサイン	0.589	0.567	0.759
多項式 (p=2)	0.691	0.667	0.611
RBF	0.685	0.659	0.598
ADD × CONC			
コサイン	0.7785	0.7209	0.4370
多項式 (p=2)	0.7862	0.7303	0.4072
多項式 (p=3)	0.7991	0.7456	0.3763
多項式 (p=4)	0.8034	0.7501	0.3551
RBF	0.8089	0.7513	0.3456

表 1: 様々なカーネルを用いて学習した際のピアソンとスピアマンの相関係数，そして平均二乗誤差 (MSE) を示す．

3.2 結果と考察

3.2.1 線形 vs. 非線形

線形カーネルであるコサイン類似度よりも，多項式カーネルと RBF カーネルを用いた非線形類似度学習によって，ピアソンの相関係数が最大で 0.1 ポイント程度上昇する結果となった．これは，単語ベクトル空間とは異なる高次元空間，つまり単語より表現力の高い空間において文の類似度を学習することが，非常に有効であることを示している．

3.2.2 ADD vs. SUBT vs. DIV vs. CONC

単語ベクトル間の演算において，可換の ADD と非可換の SUBT，DIV，CONC とを比較した．ADD では，文の系列や構造の情報はすべて失われてしまうが，文に出現する単語の共起情報をすべて考慮できるため，全体的に相関係数が高い結果となっている．また ADD と，非可換演算である SUBT とを組み合わせたモデルでは，非線形カーネル，特に多項式カーネルを用いた場合に相関係数の上昇が見られた一方で，線形カーネルでは逆に相関係数が下がる結果となった．これは，SUBT による文の意味と構造の表現が，単語と同一次元の空間においては適切なエンコードではないが，異なる高次元空間においてその表現がより適切に学習されていることを示している．しかし DIV では，全体的に相関係数が低くなる結果となり，また CONC では，ADD や SUBT と比較すると相関係数の上昇は低かった．これは除算や連結演算が，意味と構造の情報を表現し学習する演算としては適さないことを示唆している．

3.2.3 提案法 vs. 既存研究

提案法は，ピアソンとスピアマンの相関係数，平均二乗誤差 (MSE) のすべてにおいて，素性エンジニアリングをベースとした SemEval 2014 の上位のチームと，[Socher 14] らの提案した RNN を上回る結果となった．一方で，Constituency

*4 http://alt.qcri.org/semeval2014/task1/
*5 SemEval 2014 のオフィシャルのランキングにおいては，ピアソンの相関係数 r が用いられている．
*6 http://nlp.stanford.edu/projects/glove/
*7 http://www.nactem.ac.uk/enju/index.ja.html

手法	r	ρ	MSE
Illinois-LH_run1 [Lai 14]	0.7993	0.7538	0.3692
UNAL-NLP_run1 [Jimenez 14]	0.8043	0.7458	0.3593
Meaning_Factory_run1 [Bjerva 14]	0.8268	0.7722	0.3224
ECNU_run1 [Zhao 14]	0.8280	0.7689	0.3250
Dependency Tree-RNN [Socher 14]	0.7863	0.7305	0.3983
Semantic Dependency Tree-RNN [Socher 14]	0.7886	0.7280	0.3859
Constituency Tree LSTM [Tai 15]	0.8491 (2)	0.7873 (3)	0.2852 (2)
Dependency Tree LSTM [Tai 15]	**0.8627 (1)**	**0.8032 (1)**	**0.2635 (1)**
提案法 (ADD×SUBT，多項式カーネル (p=2))	**0.8414 (3)**	**0.7886 (2)**	**0.3040 (3)**
提案法 (ADD×SUBT，多項式カーネル (p=3))	0.8410	0.7870	0.3073
提案法 (ADD×SUBT，多項式カーネル (p=4))	0.8387	0.7845	0.3136

表 2: 提案法と様々な既存研究との比較．グループ分けは上から順に，素性エンジニアリングベースの手法 (これは SemEval 2014 の上位 4 チームの結果である)，依存構造 (Dependency Tree) や木構造 (Constituency Tree) を考慮した再帰的ニューラルネットワーク (Recursive Neural Network(RNN)) と Long Sort-Term Memory(LSTM)，そして我々の提案法である．括弧内は上位 3 位の手法を示している．我々の提案法は，ピアソンの相関係数と MSE で 3 位，スピアマンの相関係数で 2 位にランクしている．

Tree LSTM と Dependency Tree LSTM [Tai 15] については，提案法が同程度かあるいは若干下回る結果となった．しかし，LSTM の手法と比較すると，提案法はよりシンプルかつ実装も容易であり，今後様々な拡張を考えることができる．特に我々の結果は，単語ベクトル空間における意味と構造の非可換演算と，カーネルによる非線形類似度学習のみによって達成されており，その点において遥かに優位性があると考えられる．

4. 結論と今後の課題

本稿で我々は，意味と構造の表現学習のための非線形類似度学習法を提案した．提案法では，意味と構造の情報をベクトルとその非可換演算によって表現した上で，カーネルを用いてより表現力の高い高次元空間へ写像し類似度学習を適用する．最終的に我々は，詳細な文ベクトルを陽に計算することなく，意味と構造の双方を踏まえた適切な文表現を構成するための，新たな単語表現を獲得することができる．

今後の課題は以下の通りである．

1. 文が持つ階層構造を適切に表現する非可換演算を考え，非線形類似度学習法を適用する．
2. 深層カーネル (Deep Kernel) を用いて，非線形類似度学習法をより拡張させる．

参考文献

[Bjerva 14] Bjerva, J., Bos, J., Goot, van der R., and Nissim, M.: The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity, in *SemEval* (2014)

[Chechik 09] Chechik, G., Shalit, U., Sharma, V., and Bengio, S.: An online algorithm for large scale image similarity learning, in *NIPS* (2009)

[Duchi 11] Duchi, J., Hazan, E., and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *JMLR* (2011)

[Jimenez 14] Jimenez, S., Duenas, G., Baquero, J., Gelbukh, A., Bátiz, A. J. D., and Mendizábal, A.: UNAL-NLP: Combining soft cardinality features for semantic textual similarity, relatedness and entailment, in *SemEval* (2014)

[Kedem 12] Kedem, D., Tyree, S., Sha, F., Lanckriet, G. R., and Weinberger, K. Q.: Non-linear metric learning, in *NIPS* (2012)

[Lai 14] Lai, A. and Hockenmaier, J.: Illinois-lh: A denotational and distributional approach to semantics, in *SemEval* (2014)

[Marelli 14] Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R.: SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment, in *SemEval* (2014)

[Pennington 14] Pennington, J., Socher, R., and Manning, C. D.: Glove: Global vectors for word representation, in *EMNLP* (2014)

[Socher 12] Socher, R., Huval, B., Manning, C. D., and Ng, A. Y.: Semantic Compositionality through Recursive Matrix-Vector Spaces, in *EMNLP-CoNLL* (2012)

[Socher 14] Socher, R., Le, Q. V., Manning, C. D., and Ng, A. Y.: Grounded Compositional Semantics for Finding and Describing Images with Sentences, *TACL* (2014)

[Tai 15] Tai, K. S., Socher, R., and Manning, C. D.: Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, *arXiv preprint arXiv:1503.00075* (2015)

[Xing 02] Xing, E. P., Jordan, M. I., Russell, S., and Ng, A. Y.: Distance metric learning with application to clustering with side-information, in *NIPS* (2002)

[Zhao 14] Zhao, J., Zhu, T. T., and Lan, M.: ECNU: One Stone Two Birds: Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment, in *SemEval* (2014)