単語埋め込みの二種類の加法構成性

Geewook Kim[†,††]　横井 祥[‡,††]　下平 英寿[†,††]
† 京都大学大学院　‡ 東北大学大学院　†† 理化学研究所 革新知能統合研究センター
geewook@sys.i.kyoto-u.ac.jp, yokoi@ecei.tohoku.ac.jp, shimo@i.kyoto-u.ac.jp

1　はじめに

大規模コーパス中の単語の共起情報を用いて学習された**単語の分散表現**によって自然言語処理は長足の進歩を遂げた．単語の分散表現には，単語の意味的な情報，構文的な情報，あるいは単語間の関係を計算するための情報が格納されることが知られている．とりわけ驚くべきは，分散表現の和によって意味の合成が近似的に実現できる**加法構成性**（Additive Compositionality）[6] と呼ばれる性質で，“king”≈“man”＋“royal” などの例が知られている．

こうした単語の分散表現の性質については，理論的な解析も少しずつ進んでいるものの，未解明の点も数多く残されている．例えば，単語は一般に複数の意味を有するが，この性質は学習済みの単語の分散表現の中でどのように表現されているのだろうか？ あるいは，学習された単語の分散表現は一般に正規化してから活用されるが（コサイン類似度），分散表現のノルム（長さ）にはどのような情報が内包されているのだろうか？ 本稿ではこれらの問いに対して，加法構成性を足がかりに統一的な解釈を与える．

本稿ではまずはじめに，「複数の語義の OR の意味を持つ単語」と「複数の語義の AND の意味を持つ単語」でそれぞれ別の加法構成性が成立することを示す．そしてこれに基づいて，単語の分散表現のノルムには各単語の意味の狭さ・広さの情報が格納されていることを示す．実データを用いた実験では，単語の分散表現に想定された AND と OR の二種類の加法構成性が確かに成り立つこと，またノルムの情報だけである程度の含意関係推論ができることを確認した．

2　背景：単語埋め込み

本章では，Noise Contrastive Estimation（NCE）[4] に基盤する単語埋め込み手法に関して説明し，その具体例として Skip-Gram with Negative Sampling（SGNS）を紹介する [6]．

NCE による単語埋め込みでは学習コーパスで共起した単語ペアとそうでないペアを区別する分類タスクを考える．そして以下のような条件付き確率をモデリングの対象としている [2]．

$$p(d=1|(w_i,w_j)) = \frac{p(w_j|w_i)}{p(w_j|w_i)+k\times q(w_j)} \quad (1)$$

$$p(d=0|(w_i,w_j)) = 1-p(d=1|(w_i,w_j)) \quad (2)$$

ここで d は単語ペアがコーパスで共起したかを表す 0-1 変数である．$q(w)$ は単語をサンプリングするための分布であり，ネガティブサンプリング分布と呼ばれる．単語の分散表現を含むモデルのパラメータ $\boldsymbol{\theta}$ はコーパスで共起した単語ペアの集合 $\mathcal{D}$ に対して以下の目的関数を最大化することで学習される．

$$\sum_{(w_i,w_j)\in\mathcal{D}} \Big\{ \log p_{\boldsymbol{\theta}}(d=1|(w_i,w_j)) + \sum_{\hat{w}\sim q}^{k} \log p_{\boldsymbol{\theta}}(d=0|(w_i,\hat{w})) \Big\}$$

NCE の具体例として，SGNS [6] は以下のモデルを用いる．

$$p_{\boldsymbol{\theta}}(d=1|(w_i,w_j)) = \sigma(s_{\boldsymbol{\theta}}(w_i,w_j)) \quad (3)$$

ここで，$\sigma(x) = (1+\exp(-x))^{-1}$，$s_{\boldsymbol{\theta}}(w_i,w_j) = \langle \boldsymbol{u}_{w_i}, \boldsymbol{v}_{w_j} \rangle$ である．$\boldsymbol{\theta} = \{\boldsymbol{u}_w, \boldsymbol{v}_w\}_{w\in\mathcal{V}}$ は単語の分散表現である（$\mathcal{V}$ は辞書）．モデル (3) を (1) へ代入して式を整理すると，

$$\log \frac{p(w_i,w_j)}{p(w_i)q(w_j)} - \log k = \langle \boldsymbol{u}_{w_i}, \boldsymbol{v}_{w_j} \rangle \quad (4)$$

が導かれる．左辺で $q(w_j)=p(w_j)$ と仮定すれば自己相互情報量 (PMI) になることから，SGNS は自己相互情報量行列を分解すると解釈できる [5]．また，(4) において $b_{w_j} = \log(k\,q(w_j))$ とおくと，

$$p(w_j|w_i) = \exp(\langle \boldsymbol{u}_{w_i}, \boldsymbol{v}_{w_j} \rangle + b_{w_j}) \quad (5)$$

となるが，これは SGNS が簡素化した言語モデルとして解釈できることを示す．この解釈は本来の Skip-Gram が言語モデルに基づいているため自然だが，SGNS は

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

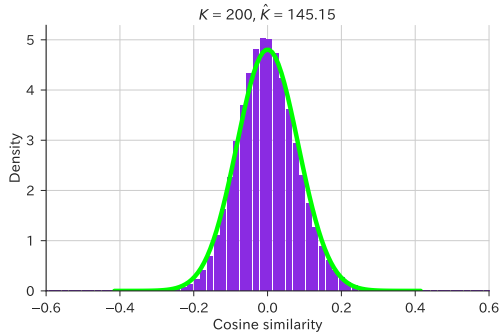

図 1: SGNS で事前学習済みの分散表現からランダムにサンプリングした $\boldsymbol{v}_w$ と $\boldsymbol{u}_{w'}$ のコサイン類似度のヒストグラム．曲線は正規分布の密度関数．分散表現はランダムではないが補題 1 の傾向が観測できる．K は埋め込みの次元，$\hat{K}$ は標本分散の逆数で，埋め込みの有効次元と解釈できる．分散表現の成分が必ずしも独立ではなく，また分散も一定でないとき，$\hat{K} < K$ である．

Skip-Gram に NCE のアイデアを導入することでモデリングの対象が異なる点に注意されたい．

3　二種類の加法構成性

単語には，**多義語**（複数の語義を持つが，具体的な使用においては特定の語義のみが現れる単語）や，**複数の語義を常に持つ単語**（たとえば語義 “man” と語義 “royal” の意味が両方とも常に生じる “king”）が存在する．このセクションでは，簡単な仮定をおくことで，「OR の単語」「AND の単語」の分散表現が，構成要素となる語義の分散表現から再構成できることを示す．

あとの実験セクションで，実際のコーパスから学習された SGNS で我々の考える加法構成が経験的によく成り立つことを示す．

3.1　OR の加法構成性

■OR の共起の確率モデル　多義語は複数個の語義を持つ単語で，実際の使用においてはいずれかの意味で用いられる．例えば “crane” は「鶴」や「クレーン（重機）」の意味を持つ多義語である．

多義語 w が s 個の語義を持ち，単語 $\{w_i\}_{i=1}^s$ で表されるとする．各語義は同時には生じないと仮定すると，確率の和の法則から次の関係が成り立つ．

$$p(w|w') = \sum_{i=1}^{s} p(w_i|w'), \forall w'. \tag{6}$$

つまり，ある単語 w' が与えられた時に周りで単語 w が出現する確率は，w' の周りで単語 w_i が出現する確率の和であると考えられる．

■OR の加法構成性　以上の簡単な仮定のもとで，多義語 w の分散表現が w_i たちの分散表現の**重み付き平均**となることを定理 1 で示す．その準備として，次の補題は中心極限定理より容易に示すことができる．

補題 1. K 次元確率ベクトル $\boldsymbol{X} = (X_1, \cdots, X_K)$, $\boldsymbol{Y} = (Y_1, \cdots, Y_K)$ の各成分は独立で期待値 0，分散は一定値とする．このとき，$\boldsymbol{X}, \boldsymbol{Y}$ 間のコサイン類似度は $K \to \infty$ の極限で期待値 0，分散 $1/K$ の漸近正規分布に従う．

図 1 をみると実際にコサイン類似度は 0 に近い．ほとんどの単語ペア w, w' に対して，分散表現はほぼ直交して $|\langle \boldsymbol{v}_w, \boldsymbol{u}_{w'} \rangle| \ll 1$ と仮定すると，

$$\exp(\langle \boldsymbol{v}_w, \boldsymbol{u}_{w'} \rangle) \approx 1 + \langle \boldsymbol{v}_w, \boldsymbol{u}_{w'} \rangle \tag{7}$$

と近似でき，これから次の定理が得られる．

定理 1. 単語 w, $\{w_i\}_{i=1}^s$ が，モデル (5) と，関係 (6) を満たすとき，その分散表現は次式で表される．

$$\boldsymbol{v}_w \approx \sum_{i=1}^{s} \frac{\exp(b_{w_i})}{\exp(b_w)} \boldsymbol{v}_{w_i} \tag{8}$$

$$b_w \approx \log\Big(\sum_{i=1}^{s} \exp(b_{w_i})\Big) \tag{9}$$

導出には近似 (7) を 2 回使う．

$$\begin{aligned}
&p(w|w') \\
=&\exp(b_w)\exp(\langle \boldsymbol{u}_{w'}, \boldsymbol{v}_w \rangle) \\
\approx&\exp(b_w)(1 + \langle \boldsymbol{u}_{w'}, \boldsymbol{v}_w \rangle) \\
\approx&\exp(b_w)\left\{1 + \langle \boldsymbol{u}_{w'}, \sum_{i=1}^{s} \frac{\exp(b_{w_i})}{\exp(b_w)} \boldsymbol{v}_{w_i} \rangle\right\} \\
\approx&\sum_{i=1}^{s} \exp(b_{w_i})(1 + \langle \boldsymbol{u}_{w'}, \boldsymbol{v}_{w_i} \rangle) \\
\approx&\sum_{i=1}^{s} \exp(b_{w_i})\exp(\langle \boldsymbol{u}_{w'}, \boldsymbol{v}_{w_i} \rangle) \\
=&\sum_{i=1}^{s} p(w_i|w')
\end{aligned}$$

重みの和は 1 になる（$\sum_{i=1}^{s} \exp(b_{w_i})/\exp(b_w) = 1$）．モデル (5) で $q(w_j) = p(w_j)$ を仮定すれば，定理 1 は Arora et al. [1] の定理 2 と同様に，多義語の分散表現は構成要素となる語義の分散表現のそれぞれの頻度に比例した重み付き平均であることを示唆している．

3.2　AND の加法構成性

■AND の共起の確率モデル　ある単語 w とその語義を常に構成する複数個の単語の集合 $\{w_i\}_{i=1}^s$ には以下の関係が成り立つと考える．

$$p(w|w') = \prod_{i=1}^{s} p(w_i|w'), \forall w' \tag{10}$$

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

つまり，ある単語 w' が与えられた時に，その周りに単語 w が出現する確率は，単語 w を構成する複数の語義に対応するそれぞれの単語が出現する確率の掛け算であると考える．

■AND の加法構成性　以上の簡単な仮定のもとで，ある単語 w の分散表現はその語義を常に構成する各単語 w_i たちの分散表現の和であることが示せる．

定理 2. 単語 $w, \{w_i\}_{i=1}^s$ が，モデル (5) と，関係 (10) を満たすとき，その分散表現は次式で表される．

$$\boldsymbol{v}_w = \sum_{i=1}^{s} \boldsymbol{v}_{w_i} \tag{11}$$

$$b_w = \sum_{i=1}^{s} b_{w_i} \tag{12}$$

この結果は次のように導出される．

$$\begin{aligned} & p(w|w') \\ =& \exp(b_w) \exp(\langle \boldsymbol{u}_{w'}, \boldsymbol{v}_w \rangle \\ =& \exp(\sum_{i=1}^{s} b_{w_i}) \exp(\langle \boldsymbol{u}_{w'}, \sum_{i=1}^{s} \boldsymbol{v}_{w_i} \rangle) \\ =& \exp(\sum_{i=1}^{s} (b_{w_i} + \langle \boldsymbol{u}_{w'}, \boldsymbol{v}_{w_i} \rangle)) \\ =& \prod_i p(w_i|w') \end{aligned}$$

これは “king”≈“man”＋“royal” などの関係性が分散表現の足し算に現れる現象を説明する．

3.3　分散表現のノルムへの影響

二種類の加法構成性は分散表現の**重み付き和**という点で共通しており，OR の加法構成性は重み付き平均，AND の加法構成性は単なる足し算に相当する．このことから，単語の分散表現のノルムは単語の意味の広さ・狭さと関連すると考察することができる．

はじめに，複数の意味を持ち広く用いられる単語（OR の単語）の分散表現は，様々な方向を向いた語義の分散表現の重み付き**平均**であり，結果として原点に近くノルムの小さい分散表現になると予想される．

逆に，特殊な意味を持つ単語（例えば専門用語）は関連する（似た方向を向いた）複数の語義の共通部分（AND の単語）として解釈できる．その分散表現は語義の分散表現の**和**であり，ノルムは大きいと予想される．極端な例として，同一の語義を重ねて “very and very” のような意味を持つ新しい単語を作る場合，その分散表現のノルムは元の単語の分散表現のノルムよりも大きくなる（$\mathrm{AND}(\boldsymbol{v}_w, \boldsymbol{v}_w) = 2\boldsymbol{v}_w$）．

単語 1	単語 2	作成した多義語
digit	spacing	digit_OR_spacing
health	aircraft	health_OR_aircraft

表 1: 作成した多義語の例.

以上の考察から，単語の分散表現のノルムは単語の意味の広さ・狭さと関連すると考えられる．あとの実験セクションで，実際のコーパスから学習された SGNS でこのような性質が経験的によく成り立つことを示す．

4　実験

本章では SGNS を用いた実験により二種類の加法構成性の存在を確認し，また，分散表現のノルムが単語の意味の広さ・狭さに関する情報を格納することを示す．本稿では Mu and Viswanath [7] に従い単語の分散表現を中心化して用いた．

4.1　OR の加法構成性の確認

OR の加法構成性（8）を確認するため，以下のように架空の多義語を作成してその性質を調べた．まず Wikipedia コーパス[*1]からランダムに 600 個の単語を選び 300 個の架空の多義語を作成した．表 1 にその例を示す．次に，単語を架空の多義語へ置き換えたコーパスと元のコーパスを結合することで多義語と元の単語が共存する学習コーパスを作成し，SGNS によって単語の分散表現を学習した．多義語 w について，学習によって得られた分散表現 $\boldsymbol{v}_w$ と，(8) によって $\{\boldsymbol{v}_{w_i}\}_{i=1}^s$ から計算された分散表現がどれほど近いかをコサイン類似度を用いて評価した．

結果 : OR の加法構成性 (8) で計算された分散表現と多義語の分散表現との平均コサイン類似度は **0.827** だった．一方で，AND の加法構成性（11）を用いた結果は **0.777** であり，多義語に対しては OR の加法構成性の方がより近い分散表現が計算できることが確認できた．

4.2　AND の加法構成性の確認

実験設定のほとんどは 4.1 節と同じである．AND の加法構成性（11）を確認するためには，“king”≈“man”＋“royal” のような例が必要である．本実験では Farahmand et al. [3] のデータセットを用いた．このデータは 2 つの単語から構成される複合語に対して，複合語と元の 2 つの単語の間にどれほど意味的な構成性が成り立つかを人手で評価したものである．

[*1] 200MB の英語版 Wikipedia コーパスを用いた．

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

複合語	構成性スコア
card_game	1.0
convenience_store	0.75
zip_code	0.0

表 2: データの例．スコアが大きいほど構成性がある．

単語 1	単語 2	含意性スコア
snake	animal	8.75
material	rubber	1.28

表 3: データの例．スコアが大きいほど単語 2 は単語 1 を包含する．

表 2 にデータの例を示す．複合語 w について，SGNS の学習によって得られた分散表現 $\boldsymbol{v}_w$ と，(11) によって $\{\boldsymbol{v}_{w_i}\}_{i=1}^{s}$ から計算された分散表現がどれほど近いかをコサイン類似度を用いて評価した[*2]．さらに，AND の加法構成性より計算されたコサイン類似度が人手で評価された構成性スコアとどれほど傾向が一致するかもスピアマン順位相関係数より評価した．

結果：AND の加法構成性 (11) で計算した分散表現と複合語の分散表現との平均コサイン類似度は **0.237** だった．一方で，OR の加法構成性 (8) を用いた結果は **0.222** だった．さらに，AND の加法構成性より計算されたコサイン類似度と構成性スコアとの相関係数は **0.197** だった．これらの結果から，構成性を持つ複合語の分散表現は AND の加法構成性によってより近い分散表現が計算できることが確認できる．

4.3 分散表現のノルムと含意関係推論への応用

単語の分散表現のノルムに単語の意味の広さ・狭さに関する情報が格納されていることを確認するため，Vulić et al. [9] のデータセットを用いて以下の実験を行なった．このデータセットには，2 つの単語ペア (w_1, w_2) の単語の意味の広さ・狭さから成り立つ，含意関係が人手で評価されている．表 3 にデータの例を示す．本実験では単語ペアの分散表現のノルムの比 $||\boldsymbol{v}_{w_1}||/||\boldsymbol{v}_{w_2}||$ と含意スコアがどれほど同じ傾向を持つかを，Nickel and Kiela [8] と同様にスピアマン順位相関係数より評価した．

結果：計算された相関係数は **0.334** だった．なお，一般に公開されている単語の分散表現[*3]を用いた実験でも相関係数は **0.282** だった．これは単語の含意関係に関する教師データを与えていないことを考慮すると，とても高い値であることがわかる [8]．

5 まとめと今後の課題

本稿では，単語埋め込みの加法構成性には「または」(OR，論理和) の意味をなすものと「かつ」(AND，論理積) の意味をなすものが存在することを示し，実際のコーパスから SGNS で学習した分散表現において二種類の加法構成性を確認した．また，単語の分散表現のノルムには意味の狭さ・広さの情報が格納されていることを示し，分散表現のノルム情報だけで含意関係推論がある程度できることを確認した．今後の課題として，近年注目されている多層ニューラルネットワーク言語モデルへ解析の範囲を広めて行きたい．また，今回の実験では素朴にベクトルの長さの情報のみを用いて単語の含意関係を推論したが，向きの情報も併用できるため，その手法の開発についても考えていきたい．

参考文献

[1] S. Arora et al. “Linear Algebraic Structure of Word Senses, with Applications to Polysemy”. In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 483–495.

[2] C. Dyer. “Notes on Noise Contrastive Estimation and Negative Sampling”. In: *arXiv:1410.8251* (2014).

[3] M. Farahmand et al. “A Multiword Expression Dataset: Annotating Non-Compositionality and Conventionalization for English Noun Compounds”. In: *Workshop on Multiword Expressions*. 2015, pp. 29–33.

[4] M. Gutmann and A. Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *International Conference on Artificial Intelligence and Statistics*. 2010, pp. 297–304.

[5] O. Levy and Y. Goldberg. “Neural Word Embedding as Implicit Matrix Factorization”. In: *Advances in Neural Information Processing Systems 27*. 2014, pp. 2177–2185.

[6] T. Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26*. 2013, pp. 3111–3119.

[7] J. Mu and P. Viswanath. “All-but-the-Top: Simple and Effective Postprocessing for Word Representations”. In: *International Conference on Learning Representations*. 2018.

[8] M. Nickel and D. Kiela. “Poincaré Embeddings for Learning Hierarchical Representations”. In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 6338–6347.

[9] I. Vulić et al. “HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment”. In: *Computational Linguistics* 43.4 (2017), pp. 781–835.

[*2] 構成性スコアが 0.5 以上の複合語を用いた．

[*3] `https://code.google.com/archive/p/word2vec`

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.