加法構成性を活用した最適輸送による文書類似度の定量化

赤松 朋哉 [1] 中川 慧 [2]
[1] 大阪大学大学院理学研究科 [2] 野村アセットマネジメント株式会社
u149852g@ecs.osaka-u.ac.jp, kei.nak.0315@gmail.com

概要

本研究は，自然言語処理の基本タスクである Senmantic text similarity (STS) の精度向上を目的とし，単語ベースと文章ベースの STS 手法の差異に着目する．特に，金融経済分野をはじめとする専門分野のフレーズ表現に焦点を当て，加法構成性を持つ単語埋め込みによって得られている単語の分散表現に対し，加法構成性を活用した最適輸送問題に基づく新しい文書類似度の定量化手法を提案する．

1 はじめに

Senmantic text similarity (STS) は，二つのテキスト間の類似度を計算し，それらの意味的な等価性を測定するタスクである [1]．STS は自然言語処理において最も基本的なタスクの一つであり [2]，情報検索やテキスト要約などの様々なタスクへ応用されている．そのため，正確な STS の計測が可能になると，様々な自然言語処理タスクの精度向上が期待できる [3].

STS の手法は単語アライメントベースのものと文ベースのものに大別される [4]．ここで，前者はアライメントが得られるため直感的かつ解釈可能である一方で，性能自体は文ベースの方が良い [5,6]．この理由として単語のアライメントでは，複数の単語の加法で構成される “フレーズ表現”をうまくとらえることができていない一方，文ベースの手法ではこれをうまくとらえているためであると考えている．特に，金融経済といった専門領域の文章には独自のフレーズ表現に対応するための専門の辞書が用意されており [7]，こういった表現が多く含まれると考えられる．ゆえに金融や経済分野といった専門領域ほどフレーズ表現を捉えることは特に重要である．例えば，「卵を一つのかごに盛るな（Don’t put all your eggs in one basket）」という表現と「一つの投資先に依存するな（Don’t rely on a single investment）」という表現はともに，分散投資の重要性を指摘するフレーズ表現である．これに対して，単語ベースの分散表現である Word2Vec と文章ベースの分散表現である BERT の埋込空間を用いて類似度を計測すると，前者は 0.66，後者は 0.90 と文章ベースはフレーズ表現に対応できていることがわかる．また同様に，「人が売るときに買い，人が買うときには売れ（Buy when others sell; Sell when others buy）」という表現と「 大衆はいつも間違える（The public is always wrong）」という群集心理に逆らう重要性を指摘するフレーズ表現でも，前者は 0.31，後者は 0.85 と文章ベースはフレーズ表現に対応できている．

以上を踏まえて本研究では，加法構成性を持つ単語埋め込みによって得られている単語の分散表現に対し，加法構成性を活用した最適輸送問題を構成することで，Word mover’s distance [8] による STS の精度を向上させる手法を提案する．（図1）

2 準備

以下では，正の整数 n に対し，$[n] \coloneqq \{1, \dots, n\}$ と定義する．また，集合 X に対し，$\#X$ で X の濃度（要素の数）を表す．

2.1 問題設定

まず単語埋め込みの設定について，以下のように定義する．

- 単語全体の集合を $\mathbb{W}$ とおく．
- 文全体の集合を $\mathbb{S}$ とおく．
- 各文 s を形態素解析によって $\#s$ 個の単語からなる順序対とみなし，$s = (s_i)_{i \in [\#s]}$ と記す．
- 各単語をある次元 D のユークリッド空間 $\mathbb{R}^D$ 内の点と対応付ける操作

$$\mathbf{vec} : \mathbb{W} \to \mathbb{R}^D$$

を**単語埋め込み**と呼ぶ．単語 w に対し，$\mathbf{vec}(w)$ $(\in \mathbb{R}^D)$ を w の**単語ベクトル**という．

単語埋め込みの最も単純な例として，ある成分のみ 1 でその他の成分は全て 0 であるベクトルによって

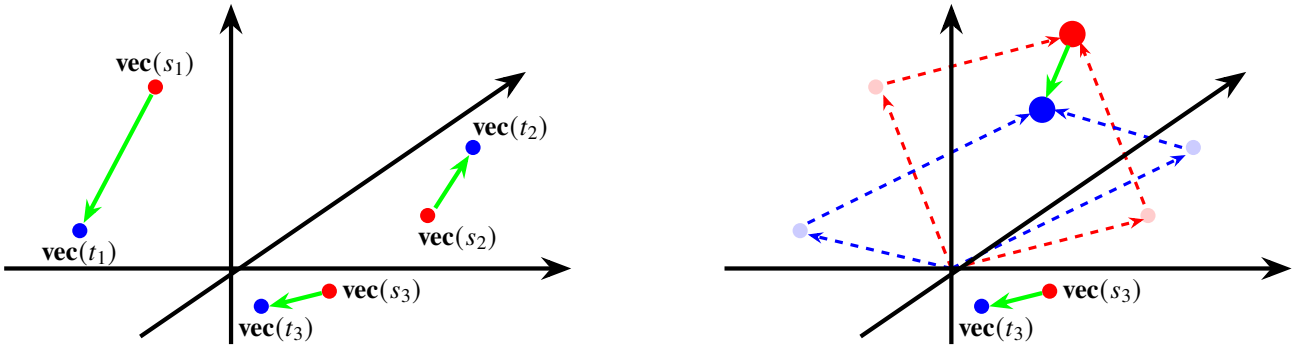

図 1: 文 $s = (s_1, s_2, s_3), t = (t_1, t_2, t_3)$ に対し，それぞれの単語埋め込みが左図のように与えられたとする．このとき，WMD では緑の矢印のような直接的な輸送により文 s, t の類似度を測る．本紙では，例えば単語 s_1, s_2 及び t_1, t_2 の組合せが 1 つのフレーズとして検知できた場合に，これらをまとめて 1 つの “単語”として扱う輸送問題を提案する（右図）．ただしフレーズは必ずしも 1 つの単語として表せないため，単語埋め込みには加法構成性を仮定し，この性質による近似を使用する．また，フレーズとして認識されなかった単語は WMD による通常の輸送を行う．

各単語を表現する One-hot 表現がある．この手法では単語の総数分の次元が必要，すなわち埋め込み先の空間の次元が $D = \#\mathbb{W}$ となるため，扱う単語数が大きい場合は各操作の計算時間が大きくなり実用的とはいえなくなる．このため埋め込み先の空間の次元を $\#\mathbb{W}$ に比べて可能な限り削減するというのは単語埋め込みにおける重要な問題意識である．この問題を上手く克服した埋め込み手法の 1 つに Mikolov ら [9] によって提案された Word2vec がある．Word2vec による単語埋め込みには，特徴を表す単語ベクトルを複数足し合わせることでそれらの特徴を兼ね備えた単語の単語ベクトルを（$\mathbb{R}^D$ 上の点として）近似できるという**加法構成性**がある [10]．加法構成性によって例えば次のよく知られた近似が可能となる：

$$\mathbf{vec}(\text{royal}) + \mathbf{vec}(\text{man}) \approx \mathbf{vec}(\text{king}). \tag{1}$$

Word2vec 以外にも Pennington らによる [11] も加法構成性を持つことが知られている．本研究では，このような加法構成性を持つ単語埋め込みを扱う．

2.2 最適輸送

最初に，最適輸送による記述を行うための用語をまとめる．

- **確率測度に関して．** 確率測度 μ の**サポート** $\operatorname{supp}\mu$ を $\operatorname{supp}\mu := \{x \in X \mid \mu(x) > 0\}$ と定義する．断りのない限り，本紙で扱う確率測度は全てサポートが有限とする．また，集合 X に対し，X 上の有限サポートを持つ確率測度全体の集合を $\mathscr{P}(X)$ と書く．すなわち $\mu \in \mathscr{P}(\mathbb{R}^D)$ は有限個の点 $x \in \mathbb{R}^D$ に重み $\mu(x) > 0$ が乗った確率分布である．$\pi \in \mathscr{P}(X \times X)$ が確率測度 $\mu, \nu \in \mathscr{P}(X)$ の**カップリング**であるとは，任意の部分集合 $S \subset X$ に対して $\pi(S \times X) = \mu(S)$ かつ $\pi(X \times S) = \nu(S)$ であることをいう．μ, ν のカップリング全体の集合を $\Pi(\mu, \nu)$ $(\subset \mathscr{P}(X \times X))$ と書く．
- **輸送距離に関して．** $c : X \times X \to \mathbb{R}$ を集合 X 上の**コスト関数**とする．このとき，$\mathscr{P}(X) \times \mathscr{P}(X)$ 上の関数 $\mathsf{W}_1(\cdot, \cdot; c)$ を

$$\mathsf{W}_1(\mu, \nu; c) := \min_{\pi \in \Pi(\mu,\nu)} \sum_{x,y \in X} c(x, y)\pi(x, y)$$

と定義する．$c(x, y)$ は $x \in X$ から $y \in X$ への単位質量あたりの輸送コストを表していることから，右辺の目的関数は μ から ν への輸送コストの総量である．従って $\mathsf{W}_1(\mu, \nu; c)$ とは，確率測度 μ, ν 間のコスト関数 c のもとでの最小輸送費用である．扱う確率測度は全てサポートが有限であるため，右辺の最小値が存在することも分かる．ここで，コスト関数 c によっては $\mathsf{W}_1(\cdot, \cdot; c)$ は距離関数とならない（例えば三角不等式を満たさない）場合があることに注意する．本紙ではこのような場合においても $\mathsf{W}_1(\cdot, \cdot; c)$ を（輸送）距離と呼ぶことがある．X が距離関数 $d : X \times X \to \mathbb{R}_{\geq 0}$ を備えた距離空間であるとき，$\mathsf{W}_1(\cdot, \cdot; d)$ は $\mathscr{P}(X)$ 上の **Earth mover's distance** あるいは **L^1-Wasserstein 距離**と呼ばれる距離尺度である[1)]．

1) 正確には (X, d) が完備可分距離空間である必要がある

次に, [8] で提案された Word mover's distance (WMD) を導入する．WMD では 2 つの文の間の類似度を測る距離尺度であるが，そのためには (1) 文を確率測度に変換する，(2) それらの間をユークリッド距離をコストとした W_1 で測る，という 2 つのステップがある．これらを式で書き下すと以下のようになる：

(1)まず，各文 s を $\mathbb{R}^D$ 上の確率測度に変換する写像[2] $STUM : \mathbb{S} \to \mathscr{P}(\mathbb{R}^D)$ を，各 $i \in [\#s]$ に対し

$$STUM\big(\mathbf{vec}(s_i)\big) := \frac{1}{\#s} \tag{2}$$

であるものと定義する．すなわち $STUM$ とは，与えられた文 $s = (s_i)_{i\in[\#s]}$ に対し，各単語 s_i の単語ベクトル $\mathbf{vec}(s_i)$ の上に合計 1 の重みを均等に乗せる操作である．

(2)そして $\mathbb{S}$ 上の距離 **Word mover's distance** WMD : $\mathbb{S}\times\mathbb{S} \to \mathbb{R}_{\geq 0}$ を

$$\mathsf{WMD}(s,t) := \mathsf{W}_1\Big(STUM(s), STUM(t); d_{\mathbb{R}^D}\Big)$$

と定義する．ここで，$d_{\mathbb{R}^D}$ は単語の埋め込み先である空間 $\mathbb{R}^D$ のユークリッド距離である．

3 提案手法

3.1 背景

1 章で述べたように専門領域においては特有の用語・フレーズが豊富である[3]．特定の単語が集まった場合に普段使われているのと異なる意味として対処する必要があるため，文脈を考慮しない "静的"な埋め込みである Word2vec などではその特定のタスクにフォーカスした訓練が必要となる．また，文ベースの STS，文脈をくみ取ってその文の中での単語の意味を表現する Embeddings from Language Models による埋め込み [14] などが有効な手法となる．ただしこれらの手法には加法構成性は期待できないことが多い．

一方で文の距離尺度である輸送距離に関しては，WMD では個々の単語ベースの輸送で文の類似度を測るため，特殊な表現を含む文への適用に十分な妥当性を持たないことがしばしば見受けられる．WMD はフレーズを単語単位に分解した状態で 2 文の距離を測るため，特にフレーズに対しては埋め込みの段階で得られている各単語の分散表現の精度が低下してしまうことが原因だと考えられる．

3.2 加法構成性を活用した最適輸送問題

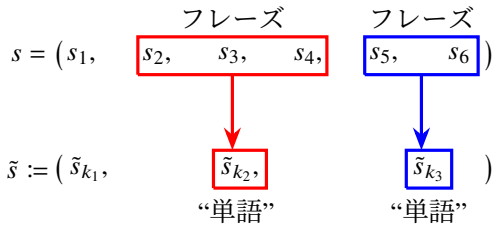

図 2: 文 s から $\tilde{s}$ への変換（式 (4)）．単語の組 (s_2, s_3, s_4), (s_5, s_6) がフレーズとして判定された場合，これらをまとめて "単語"$\tilde{s}_{k_2}, \tilde{s}_{k_3}$ として扱う．

本研究では，加法構成性を持つ単語埋め込みによって得られている単語の分散表現に対し，加法構成性を活用した最適輸送問題を構成することで WMD の精度を向上させる手法を提案する．すなわち，単語埋め込みが与えられた（単語が $\mathbb{R}^D$ に埋め込まれた）状況において WMD を修正する．

以下では**フレーズ**とは連続する 2 語以上からなる単語の組とする．与えられた文 $s = (s_i)_{i\in[\#s]}$ に対し，以下のように定義する．

- 文 s に対し，s 内で検出できたフレーズ全体の集合を $P(s)$ と表す．フレーズは最低 2 語からなるものであり，また文そのものがフレーズであることもあるため $0 \leq \#P(s) \leq \#s/2$ である．
- 検出した各フレーズ $p \in P(s)$ について，以下の流れで p を構成する s の単語たちをまとめて 1 つの "単語"として同一視する．フレーズ p を構成する s の単語の数を $\#p$ とし，$p = (p_j)_{j\in[\#p]}$ と表す．このとき，単語埋め込み **vec** の記号を濫用し，

$$\mathbf{vec}(p) = \mathbf{vec}\Big((p_j)_{j\in[\#p]}\Big) := \sum_{j\in[\#p]} \mathbf{vec}(p_j) \tag{3}$$

と定める．任意の $j \in [\#p]$ に対して $\mathbf{vec}(p_j) \in \mathbb{R}^D$ であるため $\mathbf{vec}(p) \in \mathbb{R}^D$ である．ここで $p \notin \mathbb{W}$ であり，また $\mathbf{vec}(w) = \mathbf{vec}(p)$ なる単語 $w \in \mathbb{W}$ は必ずしも存在しないことに注意する．
- このようにして新たに構成された "単語"の順序対を $\tilde{s}$ とおき，改めて番号付けしておく（図2）：

$$\tilde{s} = \Big(\tilde{s}_{k_1}, \ldots, \tilde{s}_{k_{\#[\tilde{s}]}}\Big). \tag{4}$$

（詳しくは [12, 13] などを参照）．任意の次元 D に対してユークリッド空間 $\mathbb{R}^D$ は完備可分距離空間である．

2) Sentence To Uniform (Probability) Measure.

3) 実際に本研究と同様の問題意識で専門用語については termextracthttp://gensen.dl.itc.u-tokyo.ac.jp/pytermextract/のように専門用語を抽出する技術が存在する．

いま，$\tilde{s}$ の定義より次が成り立つ：

$$\#\tilde{s} = \#s - \sum_{p \in P(s)} \big(\#p - 1\big) \ \big(\leq \#s \big).$$

2 つのフレーズ $p = (p_i)_{i \in [\#p]}, q = (q_j)_{j \in [\#q]}$ があり，これらは互いに似た意味を持っているとする．このとき，単語埋め込み **vec** に加法構成性を仮定しているため，$\mathbb{R}^D$ での (1) のような近似

$$\mathbf{vec}(p) = \sum_{i \in [\#p]} \mathbf{vec}(p_i) \approx \sum_{j \in [\#q]} \mathbf{vec}(q_j) = \mathbf{vec}(q) \quad (5)$$

が期待できる．

以上の考察をもとに，WMD を修正する．まず，与えられた文 s から確率測度への変換[4] $STWM : \mathbb{S} \to \mathscr{P}(\mathbb{R}^D)$ を考える．本設定では重みは s ではなく $\tilde{s}$ を構成する単語の単語ベクトルに乗せる．ここで，$p \notin \mathbb{W}$ であることと同様に $\tilde{s} \notin \mathbb{S}$ であるが，s から $\tilde{s}$ への変換と合成することで $\mathbb{S} \to \mathscr{P}(\mathbb{R}^D)$ の写像を構成できることに注意する．また，検出したフレーズに対して重みを割り当てることができるため，式 (2) のような一様な重みではなく，横井ら [4] に導入された Word rotator's distance (WRD) と同様の方法でフレーズの重要度を反映させた重みを考える．実際，フレーズは複数個の単語の組合せであるため，本研究のような定義を採用する場合，その埋め込み先（=フレーズを構成する単語の単語ベクトルの和，定義式 (3) 参照）のノルム（重要度）は特に意味を持つ．つまり，各 $\ell \in [\#\tilde{s}]$ に対し，

$$STWM\Big(\mathbf{vec}\big(\tilde{s}_{k_\ell}\big)\Big) := \frac{\big\|\mathbf{vec}\big(\tilde{s}_{k_\ell}\big)\big\|}{\sum_{\ell \in [\#\tilde{s}]} \big\|\mathbf{vec}\big(\tilde{s}_{k_\ell}\big)\big\|}$$

と定める．ただし $\|\cdot\| : \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ は $\mathbb{R}^D$ のユークリッドノルムである．続いて輸送距離を決める．式 (5) の近似を活用するため，ここでは純粋に $\mathbb{R}^D$ のユークリッド距離を採用する．したがって本研究で導入した輸送距離[5] $\mathrm{MWMD} : \mathbb{S} \times \mathbb{S} \to \mathbb{R}_{\geq 0}$ を

$$\mathrm{MWMD}(s,t) := \mathrm{W}_1\Big(STWM(s), STWM(t); d_{\mathbb{R}^D}\Big)$$

と定義する（図1も参照）．

4 数値例

株式相場の格言を 47 個取得[6]し，それらのペアについて (i) word2vec (ii) BERT それぞれの手法による類似度の数値例を示す．

4) Sentence To Weighted (Probability) Measure.
5) Modified Word Mover's Distance.
6) https://www.jsda.or.jp/jikan/proverb/proverb26.html

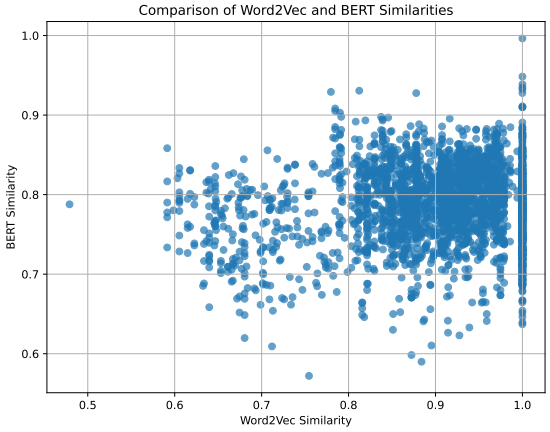

図 3: (i) word2vec (ii) BERT の類似度の散布図

まず，(i) と (ii) の相関を計算すると 0.101 となり，相関が低いことがわかる (散布図3)．なお，(i) word2vec の類似度において単語分割ができないことにより類似度が 1 となっているペアが存在するが，これを除外すると相関は-0.405 と逆相関となる．

また，(i) と (ii) の類似度の差の統計量は平均:-0.114，中央値:-0.121，標準偏差:0.099，最小:-0.362，最大: 0.308 である．定性的にも，「山高ければ谷深し」，「株価の里帰り」というほぼ同じ意味を持つ格言に対して，(i) 0.61, (ii) 0.79 という類似度が計算されている傾向がみられた．

「山高ければ谷深し」，「株価の里帰り」．(i) 0.61, (ii) 0.79.

(i) < (ii) となるのは BERT が word2vec に比べて埋め込みがより文レベルであることによる．一方，これらは同等の意味を持つため，これらのフレーズが検知できた場合は MWMD の値は十分小さくなる（すなわち類似度をより高く評価できる）ことが期待できる．

5 まとめ

本研究では，加法構成性を持つ単語埋め込みのもとでの WMD の修正案として，検知したフレーズを形式上の単語として扱った状態で類似度を求める最適輸送問題を構成した．今回の輸送尺度はフレーズを単に単語とみなしているが，このような “単語”を通常の単語と区別するべきかどうかなどの問題はまだ改良の余地があると考えられる．

参考文献

[1] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 385–393, 2012.

[2] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, Vol. 54, No. 2, pp. 1–37, 2021.

[3] John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. Beyond bleu: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4344–4355, 2019.

[4] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word rotator's distance. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

[6] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, Vol. 34, pp. 27263–27277, 2021.

[7] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, Vol. 66, No. 1, pp. 35–65, 2011.

[8] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

[9] T Mikolov, K Chen, G Corrado, and J Dean. Efficient estimation of word representations in vector space. in y. bengio & y. lecun (eds.), 1st international conference on learning representations, iclr 2013, scottsdale, az, may 2–4, 2013, workshop track proceedings. 2013.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, Vol. 26, , 2013.

[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[12] Cédric Villani, et al. *Optimal transport: old and new*, Vol. 338. Springer, 2009.

[13] Cédric Villani. *Topics in optimal transportation*, Vol. 58. American Mathematical Soc., 2021.

[14] E Matthew. Peters, mark neumann, mohit iyyer, matt gardner, christopher clark, kenton lee, luke zettlemoyer. deep contextualized word representations. In *Proc. of NAACL*, Vol. 5, 2018.

謝辞

第一著者は JST 次世代研究者挑戦的研究プログラム JPMJSP2138 の支援を受けている．