W-2-3

『リグ・ヴェーダ』讃歌の特徴を示す分散表現から得られる文書間類似度

塚越柚季

要旨

サンスクリット文献『リグ・ヴェーダ』(RV) の各巻は、複数の詩節から成る讃歌の集合である。Doc2Vec とは、文書の ID と単語列をもとにその列に続く単語を推測するアルゴリズムと、文書の ID をもとに文書内の単語列を推測するアルゴリズムとを統合したものである。単語の意味を表現するベクトルは、共起する前後の数単語とその頻度から学習されるように、その単語の用法に基づいたモデルである。単語ベクトルが結合された文書ベクトルにより、文書間の類似度の計算や、新規文書が既存文書群のいずれの類に属するかの推測などができる。そこで本発表では、サンスクリット文献『リグ・ヴェーダ』を、Doc2Vec と呼ばれる機械学習手法を用いて分析する。Doc2Vec を用いて計算した RV の巻、讃歌、詩節ごとの類似度を、グラフによって視覚化すると、類似度の高い讃歌/詩節群あるいは低い群を明確に示せる。サンスクリットという言語特有の問題への対処を行いつつ、Doc2Vec による文書の類似度計算によって得られる結果の分析を行う。

1　はじめに

近年、自然言語処理技術は目覚ましい発展を遂げており、とりわけ機械学習や中でも深層学習は隆盛を極めている。ところで、初期のサンスクリットが用いられる『リグ・ヴェーダ』は現在、種々の電子テクスト、特に形態情報を付与したテクストを使えるようになっている。つまり膨大な文献のあるサンスクリットを対象に、自然言語処理技術を用いることが可能な状況にある。はじめに、分散表現、すなわち、多次元のベクトルによって単語ならびに文書を表現する技術の概略を示す。ついで、『リグ・ヴェーダ』の構造、特徴を確かめたのち、Doc2Vec という機械学習手法を用いて分析方法を提示する。最後に分析結果の提示およびそこから得られる考察を行う。

2　文書の特徴を示すベクトル表現

2.1　Word2Vec

次節の Doc2Vec を見る前にそのもととなった Word2Vec を概観する。

Word2Vec とは、CBOW (continuous bag-of-words) というモデルと skip-gram というモデルとを含むものである。CBOW とは、文中におけるある単語の前後の (数) 単語から、中央にあるその単語を推測するニューラルネットワークモデルである。skip-gram は CBOW とは逆に、ある単語から、その前後の (数) 単語を推測するモデルである。ここではアルゴリズムの詳細に踏み入れないが、その周辺に位置する単語から中央の単語の「意味」を推測するという、単語の実際の用法に基づいた学習モデルである。

Word2Vec によって、単語がベクトルによって表現されることで、単語間の類似度が計算できる。例えば、『リグ・ヴェーダ』を学習のテクストとして用いると、神格「アグニ」(agní) という語に最も類似する語は「ジャータヴェーダス」(jātávēdas 「生き物の知識を持つ」) であることが示される。この語がアグニを指す名称であることは知られている (Grassmann 1873:483, Mayrhofer 1992:583)。この 2 語の共起頻度が最も高いわけではないことに注目すべきである。『リグ・ヴェーダ』中で *agní* という語との共起頻度が高い語は、

代名詞類 (人称代名詞、指示代名詞、関係代名詞)、*dēvá*「神」や他の神名、動詞類 (*as*「である、いる」、*dʰā*「置き定める」など)、接語や動詞接頭辞などで、*jātávēdas-* との共起頻度は上から 36 番目にある。このように、Word2Vec による単語間の類似度計算は、単純に計算される共起頻度よりも単語の「意味」をより正確にとらえているように見える。

2.2 Doc2Vec

Doc2Vec とは、PV-DM (Paragraph Vector: A distributed memory model) と PV-DBoW (Paragraph Vector: Distributed bag of words) とを含むものである。

PV-DM は Word2Vec の CBOW に相当するモデルで、文書の ID と単語列をもとにそれに続く単語を推測するモデルである。学習手順を簡潔に述べると、文書の ID とその文書内の一部の単語のベクトルを用意し、それらを結合 (平均または連結) して、その単語列の次に来る単語を推測する。

PV-DBoW は Word2Vec の skip-gram に相当するモデルで、文書の ID をもとにその文書内の単語列を推測するモデルである。ただし、これは文書中の語順を無視する。

Word2Vec が単語 (word) の特徴を表すベクトルを作る手法である一方で、Doc2Vec は文書 (document) の特徴を表すベクトルを作る手法である。Doc2Vec 内のアルゴリズムは、語順を無視したり文単位でしか表現できなかったかつての手法の短所を補いながら、任意の長さの文書 (= 文レベル、段落レベル、書類レベルなど) を入力に入れ、その文書のベクトル表現を得ることができる。

3 『リグ・ヴェーダ』

『リグ・ヴェーダ』(RV) は、サンスクリットの初期段階の言語が用いられるヴェーダ文献の中でも最古の文献である。全部で 10 の巻 (maṇḍala) から成る。各巻は、それぞれ特定の神格への讃歌 (sūkta) から構成される。さらに讃歌はいくつかの詩節 (ṛc) から成り、各詩節は一定の音節数とリズムを備えた韻律詩である。巻・讃歌・詩節が並ぶ順序は、後代の編集者らが定めた規則に従ったものであり、必ずしも詩人らの年代順に並べられたものではない。讃歌/詩節はそれぞれ特定の神格を対象としている。例えば RV 第 1 巻の 1 番目にアグニ (agní) 讃歌が配置されている。続く 2 番目の讃歌は、1-3 詩節がヴァーユ (vāyú)、4-6 詩節がインドラ (índra) とヴァーユ、7-9 詩節がミトラ (mitrá) とヴァルナ (váruṇa) と同一讃歌内でも異なる神格の詩節がまとめられる (Van Nooten & Holland 1994: 1)。

2-7 巻は、単一の詩人家系による詩節で構成されており家集と呼ばれる。8 巻は 2 つの詩人家系の詩節から成り、1, 10 巻は様々な詩人家系の詩節を含む。9 巻は「ソーマ・パヴァマーナ」(sṓma pávamāna 「自らを清めるソーマ」) を神格とする讃歌のみで構成される。これらの讃歌は本来他の巻に属するものと考えられているが、「ソーマ・パヴァマーナ」を神格とする讃歌が抜き出されて 1 つの巻にまとめられた。他の神格は、それぞれの巻で共通する。

4 『リグ・ヴェーダ』の詩節間の類似度

4.1 手法

『リグ・ヴェーダ』のテクストは、一般的にサンヒター・パータと呼ばれるテクストが使われる。このテクストは単語間の連声によって、隣り合う単語が融合することや、語末音、語頭音が変化することがある。例

えば日本語のテクストを扱う場合は、基本的に分かち書きを施してから処理するように、『リグ・ヴェーダ』のこのようなテクストをそのまま使うことは適切でなく何かしらの処理が必要である。ところで、サンヒター・パータに並んでよく用いられるテクストとしてパダ・パータというテクストがある。パダ・パータは、もとの詩を創作した詩人らの時代よりも後の時代の編集者が作ったもので必ずしも正確とは言えないが、連声を解除した語形を並べたテクストである。

パダ・パータを用いる他、『リグ・ヴェーダ』の電子テクストとして、VedaWeb より取得したデータから辞書形を取り出したテクストも用いる。名詞や動詞などの屈折語尾を考慮に入れないことによる精度の低下はありうるが、同一の語幹を持つ語形を 1 つのグループに入れることができる。

パダ・パータ、辞書形テクストのいずれも、繰り返し行はそのまま残す。同一の句、節が何度も繰り返されると学習に影響を与えると考えられる。しかし、繰り返し行を特定の 1 詩節にのみ残し他の出現箇所を削除することは、当該の文書 (= 讃歌、詩節) の特徴を消すことになると想定し、そのような削除は行わない。

Doc2Vec による計算は gensim (Řehůřek & Sojka 2010) を使う。学習アルゴリズムは PV-DM および PV-DBOW、ベクトルの次元は 300、窓幅は 15 (15-gram)、単語の最低出現回数は 1 とする。1 つの文書を巻・讃歌・詩節のそれぞれの単位で計算する。『リグ・ヴェーダ』は韻律詩という性質から、音節の軽重のリズムを整えるために、語の並びが複雑である。主語の名詞を修飾する形容詞/動形容詞が、述語動詞や目的語の名詞 (句) を越え離れて位置する場合がよくある。韻律が関わらない箇所において優勢な語順があること (Gunkel & Ryan 2015) や、語順によって意味が異なることはあるが、結果に大きな影響は与えないと考え、語順を考慮に入れない PV-DBOW も採用する。

巻の類似度に関しては、それを 1 つの文書としてみなして計算する方法の他に、また別の方法でも計算する。ある巻を構成する讃歌がどれだけ類似しているかを考えるために、讃歌ごとの類似度計算から得られる数値を用いてその平均を巻ごとの類似度とみなす。すなわち、類似する讃歌を多く含む巻ほどより類似した巻であると言える。

4.2 結果

はじめに巻ごとの類似度を見る。グラフにおいて黒に近い方は類似度が高く、白に近い方は類似度が低い。図 1 から 9 巻が他の巻よりも類似度の高い讃歌を含んでいることが分かる。2 つモデル、2 つのテクストで分析を行っているため本来 4 つのグラフを提示するべきではあるが紙幅の関係上、モデルは PV-DBOW、テクストはパダ・パータを用いた場合の分析結果を示す (図 2)。

次に讃歌ごとの類似度を見る。2 つのモデルの比較および 2 つのテクストの比較のため 4 つの図 (3, 4, 5, 6) を提示する。

4.3 考察

1 つのテーマつまり 1 神格のみを取り扱う 9 巻が類似度の高い讃歌を多く含むことが示されていることから、Doc2Vec が有効に働いていることが分かる。

詩節ごとの類似度を示すグラフ (図 3, 4, 5, 6) から分かるように、PV-DM を用いた方は全体的に類似度が高くなる一方で、PV-DBOW を用いた方は類似度が高い詩節がより顕著に示される。

アグニ讃歌が配置されている各巻頭部分もそれぞれ類似度が高いことが示されている。さらに、前述のよ

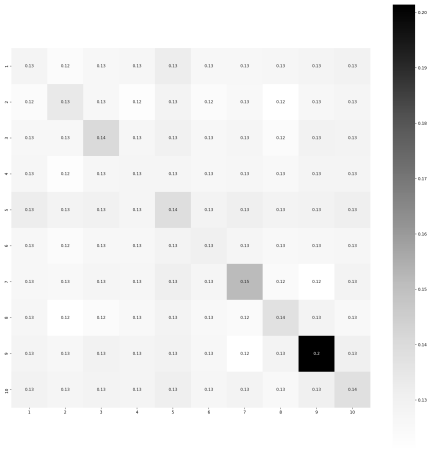

図1　巻ごとの類似度 (讃歌の平均)

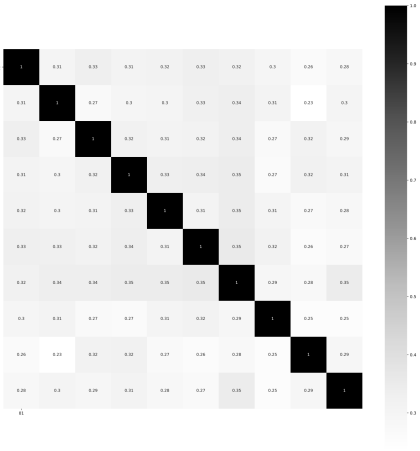

図2　巻ごとの類似度 (PV-DBOW, パダ・パータ)

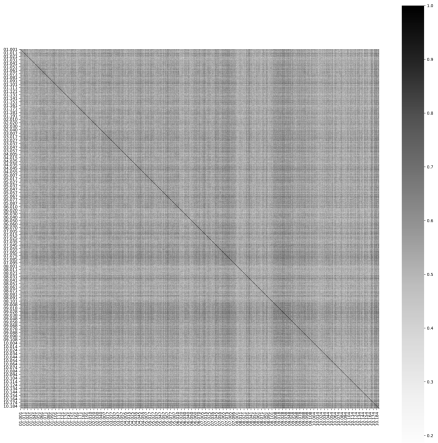

図3　讃歌ごとの類似度 (PV-DM, パダ・パータ)

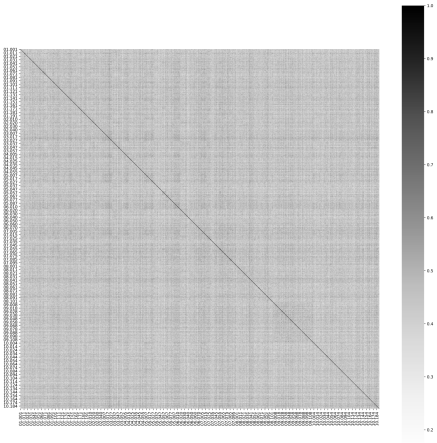

図4　讃歌ごとの類似度 (PV-DM, 辞書形テクスト)

うに 9 巻に当たる箇所は類似度が高いほか、ヴァシシュタ家系による 7 巻に当たる箇所も類似度が高いことが示されている。また、図 3 から見えるように 8 巻に当たる部分が薄い、すなわち他の讃歌のいずれとも類似度が低い。3 節で簡単に触れたとおり、8 巻は前半部の 1 家系と後半部の 1 家系との 2 家系の讃歌集で、創作された年代も 1 巻に近いことが知られている (Witzel 1995)。学習精度の問題も十分にありうるが、これらは文献研究の課題に残す。

予備実験中に得られた結果を提示するのは適切ではないだろうが、特定のパラメータでの分析の結果、一部の讃歌は他のどの讃歌とも類似度が低い結果を示した。RV 1.112 をはじめとするアシュヴィン双神讃歌

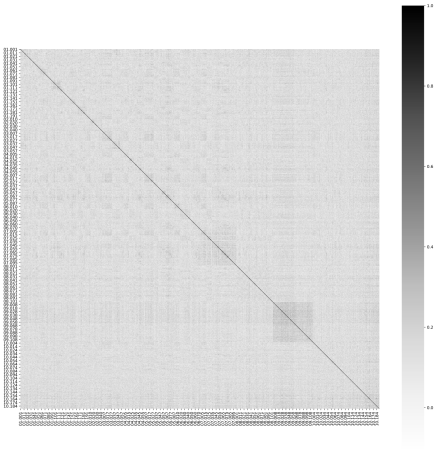

図5　讃歌ごとの類似度 (PV-DBOW, パダ・パータ)

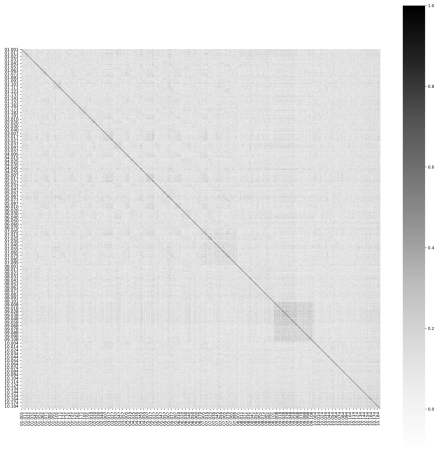

図6　讃歌ごとの類似度 (PV-DBOW, 辞書形テクスト)

がそのような讃歌である。確かな結論のためには抽出された讃歌を全て精査する必要はあるが、文書の内容の他に双数形が特徴的であることが示されている可能性がある。

PV-DBOW において、パダ・パータという連声を解除したテクストと辞書形に変換したテクストとで結果に大差がないのは注目すべき点である。『リグ・ヴェーダ』以外の文献で POS タグ付きのテクストがなくとも有意義な結果が期待される。サンスクリット語は、前述した連声もあり形態情報の自動タグ付けはおろか、自動で連声を解除して単語ごとに切り分けることすら困難である。Adinarayanan et al. (2019) のような研究はあるが、ヴェーダのサンスクリットへの応用は未だ研究途上にある。それゆえ連声を解除することさえできれば、ヴェーダ文献を対象として、Doc2Vec を用いた分析が可能である。

詩節ごとの類似度はヴェーダ文献研究への活用が見いだせる。天野 (2021)、京極 (2021) のマントラ共起関係の研究のように、本実験では行わなかった詩節の 1 行ごとの分析を行えば類似度の高い行を『リグ・ヴェーダ』内外から見るけることが容易となる。

5　課題

本実験で辞書形に変換したテクストを用いたが、そこでは本来の語形が有する情報を大幅に失っている。サンスクリットは印欧語の中でも、名詞、形容詞の変化や動詞の変化が多く、アプラウトという母音交替によって派生もする。形態や統語研究に活かす可能性を拓くため、これらの情報を含めた形の分散表現を用いて文書の類似度を計算することが必要である。

『リグ・ヴェーダ』は、音節の軽重の一定のリズムを厳格に守るため、しばしば語が離れて位置する。語順訂正の技術 (Nishida & Nakayama 2017) などを活かして『リグ・ヴェーダ』を散文の語順に直すことができるならば、語順を考慮した分析の可能性がある。

参考文献

Adinarayanan, Sharada, J. Naren, P. Sriranjanie & Ganesan Vithya (2019) Rule based POS Tagger for Sanskrit. In: A. J. Anderson (ed.) *International Journal of Psychosocial Rehabilitation.* Vol 23-1, 336-345. London: Hampstead Psychological Associates.

天野恭子 (2021) 「マントラ共起関係の可視化から読み解くヴェーダ学派間の関係性」古代文献の言語分析から読み解く社会背景のダイナミズム口頭発表。2021 年 2 月 12 日。

Grassmann, Hermann (1873) *Wörterbuch zum Rig-Veda.* Leipzig: F.A. Brockhaus.

Gunkel, Dieter & Kevin Ryan (2015) *Investigating Rigvedic word order in metrically neutral contexts.* University of Vienna Colloquium.

京極祐樹 (2021) 「ベクトル空間モデルによる『タイッティリーヤ・サンヒター』の章間類似度比較」古代文献の言語分析から読み解く社会背景のダイナミズム口頭発表。2021 年 2 月 12 日。

Le, Quac & Tomas Mikolov (2014) Distributed representations of sentences and documents. In: Eric P. Xing & Tony Jebara (eds.) *Proceedings of the 31st International Conference on Machine Learning.* Vol II. 1188-1196. Beijing: JMLR.

Mayrhofer, Manfred (1992) *Etymologisches Wörterbuch des Altindoarischen I Band.* Heidelberg: Universitätsverlag Carl Winter.

Nishida, Noriki & Hideki Nakayama (2017) Word ordering as unsupervised learning towards syntactically plausible word representations. In: Greg Kondrak & Taro Watanabe (eds.) *Proceedings of the The 8th International Joint Conference on Natural Language Processing.* 70-79. Taipei: Asian Federation of Natural Language Processing.

Řehůřek, Radim & Petr Sojka (2010) Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* 45-50 Valletta, Malta: ELRA.

Van Nooten, Barend A. & Gary B. Holland (1994) *Rig Veda: a metrically restored text with an introduction and notes.* Cambridge, Massachusetts: Department of Sanskrit and Indian. Studies, Harvard University.

VedaWeb, online reserach platform for Old Indic texts. `https://vedaweb.uni-koeln.de/`

Witzel, Michael (1995) Ṛgvedic history: poets, chieftains and politics. In: Gavin Flood (ed.) *The Blackwell companion to Hinduism*, 68-101. Oxford; Malden, Mass.: Blackwell.