

大規模言語モデルを用いたサンスクリット辞書の横断的意味探索と意味提示

塚越柚季¹ 大向一輝¹

¹ 東京大学大学院人文社会系研究科
{yuzuki, i2k}@l.u-tokyo.ac.jp

概要

本研究は、サンスクリット辞書の意味記述を横断的に探索する新たなアプローチを提示する。Cologne Digital Sanskrit Dictionaries が公開する 21 種類のサンスクリット辞書を用い、大規模言語モデルに追加学習を行うことで、個々の辞書の限界を超えた柔軟な意味解釈を可能にするモデルを構築した。提案モデルは、既存辞書の記述を統合しつつ、辞書に記載されていない単語についても意味記述を生成する能力を示した。また、妥当性がありながら、いずれの辞書でもなされていない意味記述を生成することも可能であることを示した。

1 はじめに

人文研究では、しばしば膨大なテキストから関連する情報を検索・抽出する必要がある。このような課題に対して、自然言語処理技術の応用によって、表面的な文字列検索から文脈や意味に基づく情報検索が可能となっている。本研究では、既に整備されている膨大な量のサンスクリット辞書を用いて、その横断的な意味記述を大規模言語モデルによって 1 つのモデルに集約することで、辞書検索をこえた意味記述の探索の枠組みを提案する。サンスクリット原文そのものを対象としないながらも、辞書や翻訳はそれを作成した研究者の解釈の結果であるため、そのような 2 次文献であっても関連する記述を横断的に探索することで、文献解釈の一助となることが期待される。人文学の他分野にも応用可能な上、辞書と類似する構造 – 見出し語と意味記述の組 – をもつ文献に対しても適用可能である。

本論文では、まず、背景となるサンスクリット辞書の概要 (2.1 節) と、サンスクリット辞書の横断的検索について述べる (2.2 節)。その後、具体的なモデル構築の手法 (3 節) と、モデルの意味記述の出力

結果を示し (4 節)、最後に考察 (5 節) を行う。

2 サンスクリット

2.1 サンスクリット辞書

サンスクリットは、古代インドから用いられてきた言語である。この言語は、古くは紀元前 1000 年前後に成立したヴェーダ文献群に用いられ、叙事詩や哲学文献、宗教文献など幅広いジャンルの文献で使用されている。そのようなサンスクリット文献の読解において、辞書は不可欠な工具書の 1 つである。

サンスクリット辞書は、19 世紀半ばから数多く編纂されてきた。それぞれの辞書は、言語・歴史・文化など様々な観点に基づいて編纂され、独自の目的と特徴を持っている。その中でも、用例を豊富に提示する Böthlingk & Roth の Sanskrit-Wörterbuch [1] や、それをもとにした Monier-Williams の Sanskrit-English Dictionary [2]、Apte の Practical Sanskrit-English Dictionary [3] などが広く用いられている。また、Grassmann の Wörterbuch zum Rig-Veda [4] のように『リグ・ヴェーダ』という特定の文献に特化した辞書も存在する。

このような辞書群は、編纂者による文献解釈の違いが反映されている場合がある。そのため、理想的には、複数の辞書を照らし合わせることで、より正確な意味解釈を得ることができる。しかしながら、実際にデジタル化されているものだけで数十ある辞書を実際に照合することは、時間の点においても労力の点においても、非常に困難である。それゆえ、一般的に先に挙げた有力な辞書で十分であると考えられていることも多いため、常に複数の辞書を照らし合わせることは行われていない。

19 世紀から 20 世紀初めにかけて編纂されたサンスクリット辞書は、デジタル化されているものが

多い。また、一部のウェブサイトで検索システムとともに利用できる [5, 6, 7]。特に、Cologne Digital Sanskrit Dictionaries ([5] 以下, CDS D) は、約 50 の辞書をデジタル化し公開している。さらに、CDS D は、タグをつけることで構造化した辞書のテキストデータを作成し公開している。これらのテキストは XML で構造化されている。このようなデジタル化の進展は、サンスクリット文献解釈の効率化に寄与しており、さらなる応用が期待される。

2.2 横断的検索

CDS D は、このようなデジタル辞書の構造化テキストの作成・公開を行うとともに、各辞書に対しての検索機能も提供している。しかしながら、横断的な検索を行うことはできず、個別の辞書ごとのページにおいて検索することしかできない。そのような状況において、複数の辞書データを接続する試み [8] がなされているものの、データの公開および全辞書への適用は未だ実現していない。

また、ある語形の見出し語をどのように定めるかは、まさに語の解釈そのものであるため、辞書ごとに大きな違いがある。特に同音異義語か多義語か判断がつきにくい場合、項目を別立てにするか否かの方針が辞書ごとに異なることがある。たとえば見出し語が同じであっても、母音階梯という音韻システムの中でどの階梯に基づいて見出し語を設定するかについても統一されていない。例えば、動詞 $\sqrt{star/sty}$ 「撒く」は、Böhlingk & Roth では \sqrt{star} (標準階梯) が見出し語として挙げられている一方、Monier-Williams では \sqrt{str} (ゼロ階梯) が見出し語として挙げられており、Grassmann では \sqrt{star} および \sqrt{stir} (標準階梯) が同一の動詞として扱われている。このような差異は、複数辞書を単純に横断検索するシステムの構築を極めて困難にしている。

本来、複数の辞書を照らし合わせることで、より正確な意味解釈を得ることが望まれる一方で、数十に及ぶ辞書を、紙媒体やウェブ上で並べて 1 つずつ検索することは、相当の労力を要する。そこで、いわゆる横断検索ではないにせよ、複数の辞書の記述を一つの意味空間にまとめ上げることができれば、それぞれの辞書の記述を考慮した、より豊かで深い意味記述を得ることができると考えられる。

辞書とは、一般的に見出し語と意味記述の組で構成される。サンスクリット辞書の場合、見出し語はサンスクリット単語であり、意味記述は、その単語

の意味や用法に関する記述が英語やドイツ語、フランス語などで記されている。

本研究では、そのような辞書群を用いて言語モデルを追加学習し、あるサンスクリット単語の入力に対して、その意味記述を出力するモデルを作成することを考える。

3 手法

本研究では、サンスクリット単語が与えられたときに意味記述が生成できるモデルを構築するために、CDS D が公開しているサンスクリット辞書データを活用する。使用する辞書は、サンスクリット英語辞書、サンスクリットフランス語辞書、サンスクリットドイツ語辞書、サンスクリットラテン語辞書、サンスクリットサンスクリット辞書の計 21 個を用いる (表 1)。

ただし、使用する辞書において、以下の条件を設定する。まず、Monier-Williams による Sanskrit-English Dictionary のうち、古い辞書 [26] ではなく、より新しい辞書 [2] を用いる。また、Böhlingk & Roth の Sanskrit-Wörterbuch については、簡易版の Sanskrit-Wörterbuch in kürzerer Fassung ではなく、用例も含む Grosses Petersburger Wörterbuch [1] を採用する。さらに、英語からサンスクリット単語を引く辞書や名前辞典を初めとする索引的性質を持つ辞書は除外する。

CDS D が提供するサンスクリットデジタル辞書は、XML によって構造化されている。それぞれの辞書がもつ見出し語と意味記述の組は、 $\langle H1 \rangle$ タグ内に含まれる。見出し語形は 2 種類あり、一つはアクセントのない $\langle key1 \rangle$ タグに、もう一つはアクセントや形態素境界が明示された $\langle key2 \rangle$ タグに格納される。これらはどちらも SLP1 (Sanskrit Library Phonological Text Encoding Scheme 1 (basic)) という 1 音素 1 文字の転写方式で記述されている。意味記述は、 $\langle body \rangle$ タグ内に含まれ、場合によって出典情報や文法情報などがさらにマークアップされている。CDS D のデジタル辞書はこのような構造をもつため、それぞれの辞書に含まれる見出し語形と意味記述の組を取得するのは容易である。本研究では、2 種ある見出し語のうち、より単純な $\langle key1 \rangle$ 、つまりアクセントも形態情報もない形式を用いることとする。ただし、1 つの見出し語に対して複数の意味記述がある場合は、それぞれの記述がタグによって区切られていないため、複数の意味記述全体を 1 つ

表 1 サンسكريット辞書の一覧

辞書名	記述言語
Wilson Sanskrit-English Dictionary [9]	英語
Yates Sanskrit-English Dictionary [10]	英語
Goldstücker Sanskrit-English Dictionary [11]	英語
Benfey Sanskrit-English Dictionary [12]	英語
Lanman's Sanskrit Reader Vocabulary [13]	英語
Vaidya Sanskrit-English Dictionary [14]	英語
Apte Practical Sanskrit-English Dictionary [3]	英語
Cappeller Sanskrit-English Dictionary [15]	英語
Macdonell Sanskrit-English Dictionary [16]	英語
Monier-Williams Sanskrit-English Dictionary [2]	英語
Shabda-Sagara Sanskrit-English Dictionary [17]	英語
Burnouf Dictionnaire Sanscrit-Français [18]	フランス語
Stchoupak Dictionnaire Sanscrit-Français [19]	フランス語
Böhtlingk and Roth Grosses Petersburger Wörterbuch [1]	ドイツ語
Grassmann Wörterbuch zum Rig Veda [4]	ドイツ語
Cappeller Sanskrit Wörterbuch [20]	ドイツ語
Schmidt Nachträge zum Sanskrit-Wörterbuch [21]	ドイツ語
Bopp Glossarium Sanscritum [22]	ラテン語
Abhidhānaratnamālā of Halāyudha [23]	サンسكريット
Vacaspatyam [24]	サンسكريット
Śabda-kalpadruma [25]	サンسكريット

の意味記述として扱うこととする。

学習に用いる大規模言語モデルとして Llama-2-13b-chat-hf を採用する。ここでは、辞書の見出し語と意味記述の組を対話とみなす。すなわち、見出し語を質問として、意味記述を回答とみなす形式で学習データを構築し、見出し語の意味を問う質問を入力として、その単語の辞書の意味記述を出力するようにファインチューニングする。¹⁾

4 結果

3 節の手順で構築したモデルとの対話の数例を表 2 に挙げる。ただし、本来サンسكريット部分には SLPI 方式の転写を採用しているが、ここでは可読性を確保するため、ISO 15919 に従うローマ字表記で例示する。同一の問いに対して複数回の回答を生成し、最も意味記述の長いものを選択した。

また、今回用いた辞書のいずれにも載せられていない単語の意味を正しく推定している例を表 3 に示す。

1) コードは https://github.com/Yuzki/cdsL_llama に公開されている。

5 考察

本研究で構築した新たなモデルは、サンسكريット単語に対して比較的正確な意味記述を生成できる能力を示した。特に、既存の辞書群に掲載されていない単語であっても、適切であると判断できる意味記述を生成する点は注目に値する。追加学習を行う前のモデルの学習テキストには、有名な人物や神などの一部の固有名詞が入っていると推測されるが、ベースモデルだけではほとんどのサンسكريット単語の確かな意味記述は可能でない。これは、ウェブ上の多くのテキストはデーヴァナーガリー文字表記や一般的ではない ASCII のみによるラテン文字表記などが見受けられることも一因である。学術的需要を鑑みると、昔から使われてきた IAST (International Alphabet of Sanskrit Transliteration) や現在使われつつある ISO 15919 による表記が望まれる。よって、本研究の成果は、学術的な文献読解において有用であると考えられる。また、入出力の翻字方法は本質的ではなく、このモデルへの入出力の前処理として、デーヴァナーガリー文字表記や、IAST, ISO 15919,

表2 対話例

問	What does <i>agni</i> mean?
返答	fire, L.
問	What is the meaning of <i>prthivī</i> ?
返答	the earth, L.
問	What is the meaning of <i>sakhi</i> ?
返答	A friend, companion. A trusty friend. A fellow-soldier. A wife. A messmate. A witness. A friend, companion. A wife.
問	What does <i>varuṇa</i> mean?
返答	n. the god of prosperity, prosperity, wealth, riches, RV. ; AV. ; VS. ; SBr. &c.
問	What does <i>vṛṣabha</i> mean?
返答	mfn. Belonging to a bull. m. A bull..E. a bull, and aff.
問	What does <i>ūrdhva</i> mean?
返答	mfn. 1 Upwards, aloft, elevated, ascended..2 Northwards, northern..3 Upwards, up..4 Rising, springing, growing, advancing..5 Thick, strong, powerful..6 A particular form of metre, (in rhetoric.) f. 1 The north, the northern quarter..2 The point of the compass..3 A plant.
問	What is the meaning of <i>barhis</i> ?
返答	mf. any tree or shrub bearing a fruit, grain, or pulse, grain, RV. ; AV.
問	What is the meaning of <i>vrata</i> ?
返答	m. 1. Sacrificing, offering, oblation. 2. A religious ceremony, a vow, religious observance, a religious rite. 3. An oath.4. A vow, a promise. 5. A religious order or observance, a kind of religion, mode of worship. 6. A vow, a promise, a religious obligation. E. to vow, aff.

表3 対話例(未知語)

問	What is the meaning of <i>candrayāna</i> ?
返答	n. a lunar journey, MBh. i, 6674.

SLP1 など他の転写方式を用いるようにすれば十分である。

全般的な意味記述はベースモデルよりは正確でありながら、元の辞書の記述をすべて網羅しているわけではない。その点において、複数辞書の記述を照らし合わせるという目的はまだ十分に達成されていない。

一方で、*vrata*「誓戒」の意味記述にも見られるように、既存の辞書群には存在しなくとも示唆に富む意味記述が含まれていることは注目すべき点である。²⁾ 実際に利用したサンスクリット英語辞書には“oath”という語が含まれておらず、他言語の辞書

2) *vrata*「誓戒」についての一部の辞書の記述は、付録 A に記載している。

でも相当する語を用いて記述されていない。さらにベースモデルによる生成回答にも見られない。このような独自の意味記述は、従来の文献読解に新たな示唆を与える可能性があると考えられる。また、このケースから示唆されるのは、モデルが独自の知識統合を行い、新しい解釈を生成する能力を持っているという点である。

学習データの整備および評価にはいくつかの課題が残されている。まず、今回用いている辞書の意味記述部分は、タグ付けされた電子テキストからタグを除去しただけのテキストである。このテキストは、省略記号や意味記述内の他のサンスクリット単語、出典などの情報が適切に処理されていない。それゆえ、表 2、表 3 に見られるように、不必要であり時には誤っている情報が付与されることがある。

また、意味記述には、サンスクリット原文の用例や出典情報が含まれており、また辞書によっては名詞や動詞の屈折形が詳細に載せられている。これらの情報を除外することなく学習データに取り込んだ結果、特に正確ではない出典情報が生成される問題が確認された。出典情報の正確性を確保することはまた別の課題であり、これを切り離して扱うべきである。同様に、名詞や形容詞の文法性に関する形式的な情報も、モデルの主目的に不要な場合には学習データから事前に削除することが望ましい。しかし一方で、出典や文法情報を簡単に取得する手段は現状では限られており、通常の辞書を用いた手作業に頼らざるを得ない。

6 おわりに

本研究では、CDS D に公開されているサンスクリット辞書を用いて言語モデルを追加学習し、サンスクリット単語の意味記述を生成するモデル構築した。その結果、モデルはサンスクリット単語の意味記述をある程度正確に生成できることが示された。特に、既存の辞書には記載されていないような新しい視点を提示できる点は、文献研究における重要な示唆となる。一方で、複数の辞書記述を完全に統合することや、生成される意味記述の精度向上には課題が残っている。特に、学習データのタグ付けや構造情報の活用が、モデルの性能をさらに高める鍵となると考えられる。

謝辞

本研究は JSPS 科研費 23K18646 および東京大学次世代知能科学研究センターの助成を受けたものです。

参考文献

- [1] Otto Böhtlingk and Rudolph Roth. **Sanskrit Wörterbuch, Herausgegeben von der Kaiserlichen Akademie der Wissenschaften**. Eggers, St. Petersburg, 1855-1875.
- [2] M. Monier-Williams. **A Sanskrit-English dictionary: Etymologically and philologically arranged with special reference to Cognate indo-european languages**. The Clarendon Press, Oxford, 1899.
- [3] V. S. Apte. **The practical Sanskrit-English dictionary: containing appendices on Sanskrit prosody and important literary & geographical names in the ancient history of India, for the use of schools and colleges**. Shiralkar, Poona, 1890.
- [4] Herman G. Grassmann. **Wörterbuch zum Rig-Veda**. F.A. Brockhaus, Leipzig, 1873.
- [5] Cologne digital sanskrit dictionaries, version 2.4.79.
- [6] Sanskrit dictionary.
- [7] GRETIL - göttingen register of electronic texts in indian languages and related indological materials from central and southeast asia.
- [8] Francisco Mondaca and Felix Rau. Transforming the cologne digital Sanskrit dictionaries into OntoLex-lemon. In **Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)**, pp. 11-14, Marseille, France, May 2020. European Language Resources Association.
- [9] H. Wilson. **A dictionary in Sanskrit and English: Translated, amended, and enlarged from an original compilation**. The Education Press, Calcutta, 1832.
- [10] W. Yates. **A dictionary in Sanscrit and English: designed for the use of private students and of Indian colleges and schools**. Baptist Mission Press, Calcutta, 1846.
- [11] T. Goldstücker and H. H. Wilson. **A Dictionary, Sanskrit and English: extended and improved from the second edition of the dictionary of Prof. H. H. Wilson; together with a supplement, grammatical appendices and an index, serving as an English-Sanskrit vocabulary**. Asher, Berlin, 1856.
- [12] T. Benfey. **A Sanskrit-English dictionary: with references to the best editions of Sanskrit authors and etymologies and comparisons of cognate words chiefly in Greek, Latin, Gothic, and Anglo-Saxon**. Longmans, Green, and Co., London, 1866.
- [13] C. R. Lanman. **A Sanskrit Reader: with vocabulary and notes**. Ginn and Co, Boston, 1888.
- [14] L. R. Vaidya. **The Standard Sanskrit-English Dictionary: containing appendices on Sanskrit prosody and names of noted mythological persons &c. (For the use of schools and colleges)**. Mrs. Radhabai Atmaram Sagoon, Bombay, 1889.
- [15] C. Cappeler. **A Sanskrit-English dictionary: based upon the St. Petersburg lexicons**. Trübner, Strassburg, 1891.
- [16] A. A. Macdonell. **A Sanskrit-English dictionary: being a practical handbook with transliteration, accentuation, and etymological analysis throughout**. Longmans, Green, London, 1893.
- [17] J. V. Bhaṭṭācārya. **A comprehensive Sanskrit-English lexicon: chiefly based on Professor Horace Hayman Wilson's Sanskrit-English Dictionary and compiled from various recent authorities for the use of schools and colleges**. Calcutta Press, Calcutta, 1900.
- [18] Eugène Burnouf and Léon Leupol. **Dictionnaire classique**

sanskrit-français où sont coordonnés, révisés et complétés les travaux de Wilson, Bopp, Westergaard, Johnson, etc. et contenant le dévanāgarī, sa transcription européenne, l'interprétation, les racines et de nombreux rapprochements philologiques. Maissonneuve, Paris, 1866.

- [19] N. Stchoupak. **Dictionnaire Sanskrit-Français**. Librairie d'Amérique et d'Orient, Adrien Maissonneuve, 1959.
- [20] Carl Cappeller. **Sanskrit Wörterbuch, nach den Petersburger Wörterbüchern**. K.J. Trübner, Strassburg, 1887.
- [21] R. Schmidt and Otto Böhtlingk. **Nachträge zum Sanskrit-Wörterbuch in kürzerer Fassung von Otto Böhtlingk**. O. Harrassowitz, Leipzig, 1928.
- [22] Franz Bopp. **Glossarium Sanscritum in quo omnes radices et vocabula usitatissima explicantur et cum vocabulis Graecis, Latinis, Germanicis, Lithuanicis, Slavicis, Celticis comparantur**. Dümmler, Berolini, 1847.
- [23] A.V. Rao and H.S. Ayyangar. **Abhidhānaratnamālā of Halāyudha: with the Kannada tīke of Nāgavarma**. Madras University Kannada series. University of Madras, 1940.
- [24] T. V. Bhaṭṭācārya. **Vācaspatyam: Bṛhat Saṃskṛtābhidhānam**, Vol. 5. Caukhambā Saṃskṛta Sīrīja Āphisa, Vārāṇasī, 1873.
- [25] Rādhākāntadeva, Barada Prasad Vasu, and Hari Charan Vasu. **Śabdakalpadruma**. Barada Prasad Basu and Hari Charan Basu, Kalikātā, 1886.
- [26] M. Monier-Williams. **A Sanskrit-English dictionary etymologically and philologically arranged with special reference to Greek, Latin, Gothic, German, Anglo-Saxon, and other cognate Indo-European languages**. The Clarendon Press, Oxford, 1872.

A 辞書記述の例

本節では、本研究で使用した辞書の記述例を示す。例として用いる語は、5節で示した *vrata* 「誓戒」である。紙幅の都合上、一部の辞書の記述例の一部を抜粋して示す。本研究で用いた辞書は全て、<https://www.sanskrit-lexicon.uni-koeln.de/>にて検索可能である。

A.1 Monier-Williams Sanskrit-English Dictionary

vrata n. (ifc. f(ā). ; fr. √ 2. vr) will, command, law, ordinance, rule, RV. obedience, service, ib.; AV.; ĀśvGr. dominion, realm, RV. sphere of action, function, mode or manner of life (e.g. śuci-vr°, ‘pure manner of life’, Śāk.), conduct, manner, usage, custom, RV. &c. &c. (以下略)

A.2 Böhtlingk and Roth Grosses Petersburger Wörterbuch

1. vratá (von 2. var)

1) n. (nach Ak. 2,7,37 und Trik. 3,5,11 auch masc., welches wir nur durch M. 2,3 zu belegen vermögen). am Ende eines adj. comp. f. ā .

a) Wille, Gebot; Gesetz, vorgeschriebene Ordnung (用例略)

— b) Botmässigkeit, Gehorsam; Dienst (用例略)

— c) Gebiet: yāḥ pāṛthivāso yā apāmapī vrāte (用例略)

— d) Ordnung so v. a. geordnete Reihe, Reich: trīṇī vratā vīdathē antareṣām (用例略) (以下略)

A.3 Apte Practical Sanskrit-English Dictionary

vrataḥ -taḥ [vraj-gha jasya taḥ]

1 A religious act of devotion or austerity, vowed observance, a vow in general; (用例略); (there are several vratas enjoined in the different Purāṇas; but their number cannot be said to be fixed, as new ones, e. g. satyanārāyaṇavrata, are being added every day.)

2 A vow, promise, resolve; (用例略)

3 Object of devotion or faith, devotion; as in (用例略) (以下略)

A.4 Stchoupak Dictionnaire Sanscrit-Français

vrata- nt. règle, observance ; conduite, manière d’être ; vœu (qu’on prononce), devoir religieux ; pratique religieuse, chasteté (am car- observer un vœu, not. le vœu de chasteté); m. fils d’un Manu ; -āt en vertu d’un vœu ; vant a. qui observe les pratiques religieuses, qui observe

un vœu, not. celui de la chasteté. °grahaṇa- nt. fait de prononcer un vœu, d’entrer en religion. °caryā- f. observance religieuse, fait d’observer un vœu. (以下略)

A.5 Bopp Glossarium Sanscritum

vrata m. n. (fortasse a r. vṛ i. e. var eligere, transposito var in vra, suff. ta) votum, devotio, pietas. SA. 4.3. 6. BH. 6.14. N. 5.21. 13.69. IN. 4.7. Saepe in fine compos. BAH. e. c. pativratā erga conjugem pietatem, devotionem habens, (以下略)

A.6 Vacaspatyam

vrata na0 vraja — gha jasya taḥ . 1 bhakṣaṇabhede uṇādi0 2 puṇya- sādhanē 3 upavāsādinīyamabhede ca vrataṅca samyaksāṅkalpajānitānuṣṭheyakriyācīseyarūpaṃ tacca pravṛttinivṛttyubhayarūpaṃ tatra dūṣyaviśeṣabhojanapūjādikaṃ pra- vṛttirūpaṃ upavāsādikaṅca nivṛttirūpaṃ tacca nityaṃ naimi- ttikaṃ kāmyaṅca . nityamekādaśyādivratam naimittikaṃ cāndrāyaṇādivratam kāmyaṃ tattattithyupavāsādirūpaṃ “sattvake sasādhanam karma kartavyamadhikāriṇā . niṣkāmeṇa sa- vīra! kāmyaṃ kāmānvitena ca” bhaviṣyapu0 . samyak pravama (以下略)