独立成分分析による事前学習済み多言語モデルの層を横断した単語埋め込み表現の分析

北野雄士　西田悠人　坂上温紀　上垣外英剛　渡辺太郎
奈良先端科学技術大学院大学
{kitano.yuji.la6,sakajo.haruki.sd9}@naist.ac.jp
{nishida.yuto.nu8,kamigaito.h,taro}@is.naist.jp

概要

多言語モデルの有用性は埋め込みによって支えられている。しかし、多言語で事前学習された言語モデルの埋め込みに内在すると考えられる単語の意味表現の分析は行われていない。本研究では、多言語翻訳モデルを対象に独立成分分析を用いて埋め込みを分析・可視化することによって、モデルの言語理解や層を横断した言語処理の流れについて調査した。分析の結果、入力に近い層では表層形によって軸が分離し、出力に近い層では意味によって分離する傾向がみられ、層を経るごとに文脈や意味情報を獲得することが示唆された。

1　はじめに

多言語で事前学習された言語モデル (以下、多言語モデル) は機械翻訳、感情分析、質問応答、文書分類などのさまざまなタスクにおいて、単一のモデルで複数の言語を効果的に処理することを可能にしている。多言語モデルがこのような成功を収める鍵の一つは、異なる言語間で同じ意味や文脈を共有して扱える能力にある。

これを実現するためには、多言語モデルの埋め込み表現が異なる言語の情報や意味や文脈に基づいた情報を統合している必要がある [1]。そのため、多言語モデルが異なる言語の間で意味や文脈をどのように扱っているのか、具体的にどの部位でその処理が行われているのかを明らかにすることは、モデルの理解と応用の上で極めて重要である。

しかし、このような内部構造の解析を行う際には、言語の違いによる影響を排除しなければ正確にその部位を特定することが難しい。このような課題を克服することは、多言語モデルのさらなる内部表現分析に向けて重要な一歩となる。

本研究では、この問題に着目し、多言語モデル内部の言語処理メカニズムを明らかにするために、埋め込み表現を独立成分分析 (ICA: Independent Component Analysis) によって分析・可視化する。ICA はデータを統計的に独立した成分に分解する手法であり、埋め込み空間に内在する意味的特徴を個別に捉える能力を有し、いくつかの研究で用いられている。多言語モデルの文脈を考慮した埋め込み表現を ICA によって分析・可視化することによって、モデルが学習していると考えられる言語的特徴を層ごとに単語単位で示す。

分析の結果、入力に近い層は主に表層形によって軸が分離し、出力に近い層は言語に関係なく意味によって分類している可能性があることが明らかとなり、モデルが層を経るごとに言語の意味情報や文脈を獲得していることが示唆された。

2　関連研究

2.1　独立成分分析による埋め込み分析

従来、単語埋め込み表現を解釈可能にするための分析手法として、主成分分析 (PCA) [2] や特異値分解 (SVD) [3], 特異ベクトル正準相関分析 (SVCCA) [4] などが用いられてきた。しかし、これらの手法による分析は、埋め込み表現の意味的特徴の解釈性において ICA と比較して不明確であることが示され、ICA を用いた埋め込み表現分析が近年行われている [5, 6, 7]。ICA は PCA とは異なり、次元を圧縮することなく、各次元の非ガウス性を最大化するようにベクトル表現を分解することができる。また、中心極限定理により非ガウス性が大きい次元は、線形混合されたデータから抽出された成分が統計的に独立であることを示唆する。これらを応用して単語の埋め込み表現を解釈可能な特徴の組み合わせと見な

すことで、その特徴を表す軸に分離することを目標に ICA が埋め込み表現分析に用いられてきた。

Tomáš ら [5] は自動単語侵入者テストを用いて、ICA によって変換された単言語静的埋め込みが主成分分析によって変換された埋め込み表現よりも意味的特徴を持つ軸の解釈性が高く、ある軸において高い値をとる単語の一貫性がより大きいことをいくつかの軸で示した。Hoagy ら [6] は残差ストリームで学習した埋め込み空間におけるベクトルの方向は、自動解釈スコアにおいて ICA が PCA よりも高い値を示し、解釈性が高いことを示した。Yamagiwa ら [7] は、ICA により変換された単言語埋め込み表現は異方的な構造や加法構成性を持ち、解釈可能性が比較的高いことが示された。単言語静的埋め込みが言語間で一貫性を持つことが示唆された。

ICA を用いたこれらの研究は、文脈に依存せず同じ単語は常に同一のベクトル表現をもつ静的な埋め込みや、単言語モデルの埋め込みを分析対象としている。しかし、ICA を用いた多言語モデルの各層における文脈を考慮した動的埋め込み表現の意味構造を明らかにする取り組みは行われてこなかった。

本研究では、埋め込み表現をモデルが獲得している特徴と見なし ICA を用いて変換することで、多言語モデルが学習していると考えられる言語的特徴について単語の意味レベルで分析・考察する。

2.2 多言語埋め込み表現分析

多言語埋め込みは、言語の意味だけでなく言語の種類も同一の埋め込み空間内に符号化されている。Rochelle ら [8] は多言語モデルの文埋め込みで WALS を用いた言語類型論的分析を行った。Ethan ら [9] は構造プローブによって、mBERT [10] が各言語で文構造を学習していること、文法を表す部分空間が言語間で共有されていることを示した。しかし、これらの研究は構文的な特徴や言語の種類などの分析に焦点を当てており、語の文脈上での意味単位で分析している研究はほとんどない。本研究では、ICA により文脈で条件付けられた単語埋め込みが持つ意味構造を明らかにする。

3 実験手法

3.1 対象とする多言語埋め込み

本研究では、100 言語の翻訳データで学習された多言語翻訳モデルである M2M100 [11] の 418M モデル [1] を用いる。埋め込みを取得するために、対訳コーパスである flores200 [12] のうち、M2M100 モデルがカバーする言語と共通する 5 言語（英語、フランス語、ドイツ語、ロシア語、日本語）のテキストを用いた。各言語 1024 個の文を入力し、モデルのエンコーダ 12 層の次元 $d = 1024$ の埋め込み表現を取得する。本実験では、埋め込み表現の意味構造を明らかにするために、トークン表現を単語単位で平均した表現を対象として分析を行った。日本語の単語の分割には MeCab [13] による分かち書きを用い、その他の言語は空白区切りを 1 単語とみなす。

3.2 分析手法

ICA による埋め込み表現の分析は、Yamagiwa ら [7] に倣い、以下の手順で行った。

1. 英語、フランス語、ドイツ語、ロシア語、日本語の 5 言語のコーパスそれぞれのトークン埋め込み表現を M2M100 モデルの各層で取得する。
2. 取得したトークン埋め込みから記号や未知語を取り除いた後、単語内のトークン埋め込みを各単語ごとに平均することにより単語埋め込みを計算する。その後、計算した単語埋め込みを言語を横断して各層で連結する。つまり、5 言語における総単語数を n とすると、連結後の各層の単語埋め込みは $\mathbf{X} \in \mathbb{R}^{n \times d}$ となる。
3. 各隠れ層において、得られた全ての埋め込み表現 $\mathbf{X}$ を FastICA [2] によって独立成分に変換する。ここで、FastICA の計算は $\mathbf{X} = \mathbf{AS}$ と表され、独立成分 $\mathbf{S} \in \mathbb{R}^{n \times d}$ を取得する。ただし、$\mathbf{A}$ は独立成分を合成信号 $\mathbf{X}$ に変換する行列である。
4. 変換された埋め込み表現 $\mathbf{S}$ の各軸の歪度の値 $\mathbf{E}(X^3)$ を計算し、その絶対値の降順に全ての単語の軸を並び替える。歪度は分布の非ガウス性を測るための指標の 1 つであり、歪度の絶対値が大きいほどデータ固有の情報を反映していると考えられる。ここでは、単語埋め込みの意味的な構造を代表する軸の選定に利用した。
5. 各軸における全ての単語の値を MinMaxScaler によって正規化する。その後、上位軸の歪度の符号が正であれば降順、負であれば昇順で上位 5 単語を重複がないように選択し、ヒートマップによって可視化する。これは、歪度が正であ

1) https://huggingface.co/facebook/m2m100_418M
2) https://scikit-learn.org/1.5/modules/generated/sklearn.decomposition.FastICA.html

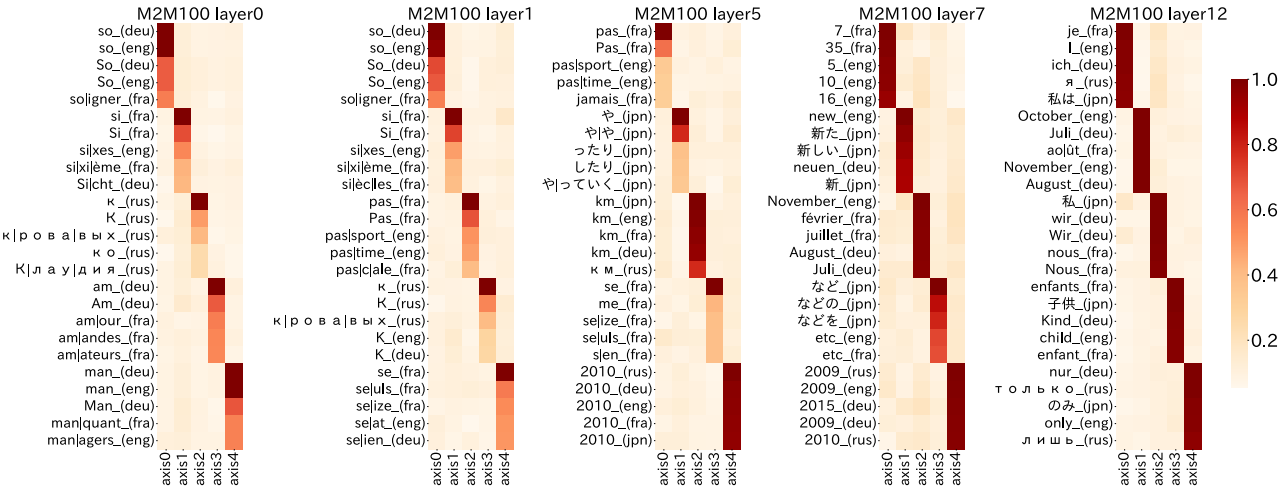

図 1　M2M100 モデルの ICA 後の各軸での上位単語。ただし、単語内の'|' はトークンの句切れを表し、単語の末尾にはそれぞれの単語がどの言語のテキストのものかを表す言語コードを付与している。

ればその軸における値が大きい単語、歪度が負であればその軸における値が小さい単語が、他の単語とは比較的異なる値となるため、その軸が表現している意味を反映している単語だと考えられるためである。

4　結果と考察

4.1　層ごとの分析

M2M100 モデルの一部の層における埋め込みの分析結果を図 1 に示す。すべての層の結果と考察は付録 A を参照。この図は、歪度の絶対値の上位 5 軸において、各軸の値の上位 5 単語ずつを並べて可視化したものである。また、第 0 層における埋め込みは、M2M100 モデルの埋め込み層の出力である。

0 層目から 1 層目はモデルの層を経ているが、分離された特徴に大きな変化はなく、どちらも言語の表層形で軸が分離されていることが見てとれる。この特徴は言語に関係なく、同じラテン文字を使う英語・ドイツ語・フランス語間でも近い値を持つことがわかる。しかし、注目したいのは 1 層目の axis3 で、キリル文字を用いるロシア語とラテン文字を使う英語・ドイツ語が同じ軸の上位単語として現れていることである。これは、文字の種類は異なるが似たアルファベットとしてモデルが認識していると考えられる。5 層目では、歪度の絶対値が上位である 5 軸の中に、上位単語がほぼ同じ値を持つ軸が現れている。これらは表層形は同じであるが、分析対象の 5 言語全てが上位単語に含まれている。7 層目では表層系ではなく意味で軸が分離し始めていることが見てとれる。例えば axis1,2,3 では、それぞれ “new” を意味する単語、月名を表す単語、「など」を意味する単語が複数言語で近い値を取っていることがわかる。このことから、M2M100 モデルは 7 層目付近で言語の意味を獲得し始めていることが考えられる。M2M100 モデルの最終層である 12 層目では、上位軸における上位単語がその軸においてほぼ同じ値をとり、複数言語間で同一もしくは似た意味で軸が分離されていることが見てとれる。また、出力に近い層ほど上位軸における上位単語の値の差が小さいことも分かる。

4.2　言語特徴分析

層を経ることによる単語の変化に着目し、モデルが学習していると考えられる言語特徴について分析を行う。入力に近い隠れ層では、各軸上位 5 単語が似た表層形を持っており、言語の表層形で軸が分離しているように見える。この結果は Andera ら [14] による研究の mT5-XL [15] の初期トークン埋め込みの分析結果と類似している。この研究では、トークン埋め込みの分布はモデル構造によって異なることを示している。実際、M2M100 モデルと mT5-XL はどちらも Transformer [16] ベースのエンコーダ・デコーダモデルであり、似たアーキテクチャを採用しているので、M2M100 モデルの入力に近い隠れ層の結果が似た結果になったと考える。

また、中間の層から徐々に言語を横断した意味で軸が分離されていき、出力に近い層では軸が言語に依らず意味で分離している傾向にあることが分かった。このことから、M2M100 モデルは出力に近い層

ほど単語の意味を学習していると考えられる。言い換えると、M2M100 モデルが層を通過するごとに入力されたテキストに対して意味や文脈情報を付与していることが示唆された。

4.3 単語単位での分析

ここでは、より具体的な単語特徴について考察する。7 層目や 12 層目における歪度の絶対値が上位の軸の上位単語は、主に機能語であることが伺える。広義的には、同義語が少ない単語が上位に多いと考えられる。これは、言語内で意味の似た語が少ない単語は、他の単語に置き換えることが難しい特徴的な単語であるので、意味による独立成分への分離がより明確であることに起因していると考えられる。加えて、これらの語の意味を表すと考えられる軸の上位単語は、比較的多くの種類の言語が混在していることが分かる。

これに関連して、層を経るごとに意味で軸が分離している様に見られるが、同義語が少ない単語が優先的に分離されていることも見てとれる。例えば、7 層目 2 軸目と 4 軸目の上位 5 単語はそれぞれ “km”、 “2010” を表す単語であるが、どちらも代替が難しい単位や数詞である。この現象は、独特な意味を表す軸における単語の分布が、ガウス分布からの歪みが大きいために、歪度の絶対値が大きくなるからである。

出力に近い層の軸に現れる上位単語の特徴として、基本語が多いことが分かる。これは、学習データに高頻度で現れやすい基本語は他の語に比べて適切に埋め込みを表現しやすいことに起因していると考えることができる。

4.4 上位軸における値の推移

ここでは、上位軸における上位単語の値に注目して分析を行う。入力に近い層より出力に近い層の方が、上位軸の上位単語の値が近いことが分かる。これは、出力に近い層の埋め込みがより独立成分に分解しやすい言語特徴を学習していることを示している。このことから、多言語埋め込み空間において表層形よりも言語の意味の方がより統計的な独立に分離可能な特徴であると考えることができる。このことから、M2M100 モデルは散らばって分布した単語を区別し適切に処理する可能性があると考察する。

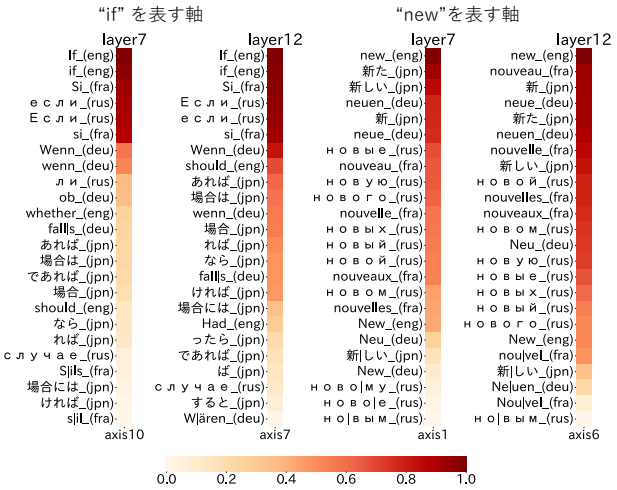

図 2 “if” と “new” を表す軸の 7,12 層における上位単語

4.5 同じ意味を表す軸の推移

同じ意味を表すと考えられる軸の値が層を経ることでどのように推移するかに着目する。図 2 は “if” と “new” を表す軸の 7 層と 12 層における上位単語の値を示している。軸が表す意味はほとんど同じにもかかわらず、7 層から 12 層になると、どちらの意味を表す軸においても単語がより近い値をとるようになっている。また、層を経ても上位単語は大きく変化していないことが分かる。これらの結果から、単語の意味ごとに独立した構造が層を経るにつれてより強く形成されているといえる。

5 おわりに

本研究では、ICA を用いて事前学習済み多言語モデルの埋め込み表現の分析・可視化を行った。ICA によって、M2M100 モデルの層が深くなるにつれて言語の意味で軸が分離できることを示した。このことから、M2M100 モデルは層を経るごとに言語の意味や文脈情報を学習していると考えられる。

本研究では 5 言語のみを扱ったが、今後は Daoyang ら [17] のように比較的学習データが少ない言語に対しても実験を行い、結果の分布の違いからモデルが学習している言語特徴を言語の種類もしくは意味の分類で定量的に示すことを目指す。

また、本研究より、ICA による多言語モデルの内部表現分析は、モデルの言語理解を理解するのに有効な手段である可能性が示唆された。今後は他の多言語モデルに適用し比較することによって、モデル間の言語理解の差を示したいと考える。

参考文献

[1] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics.

[2] Tomáš Musil. Examining structure of word embeddings with pca. In **Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings**, p. 211–223, Berlin, Heidelberg, 2019. Springer-Verlag.

[3] Tyler Chang, Zhuowen Tu, and Benjamin Bergen. The geometry of multilingual language model representations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 119–136, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[4] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems 30**, pp. 6076–6085. Curran Associates, Inc., 2017.

[5] Tomáš Musil and David Mareček. Exploring interpretability of independent components of word embeddings with automated word intruder test. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 6922–6928, Torino, Italia, May 2024. ELRA and ICCL.

[6] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[7] Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Discovering universal geometry in embeddings with ICA. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 4647–4675, Singapore, December 2023. Association for Computational Linguistics.

[8] Rochelle Choenni and Ekaterina Shutova. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties.

[9] Ethan A. Chi, John Hewitt, and Christopher D. Manning. Finding universal grammatical relations in multilingual BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5564–5577, Online, July 2020. Association for Computational Linguistics.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. **J. Mach. Learn. Res.**, Vol. 22, No. 1, January 2021.

[12] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

[13] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] Andrea Wen-Yi and David Mimno. Hyperpolyglot llms: Cross-lingual interpretability in token embeddings. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, p. 1124–1131. Association for Computational Linguistics, 2023.

[15] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 483–498, Online, June 2021. Association for Computational Linguistics.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, December 2017. Curran Associates Inc.

[17] Daoyang Li, Mingyu Jin, Qingcheng Zeng, Haiyan Zhao, and Mengnan Du. Exploring multilingual probing in large language models: A cross-language analysis, 2024.

A　各層における上位軸の上位単語の可視化

M2M100 モデルの全ての層における埋め込みの分析結果を図 3 に示す。すべての層における上位軸の上位単語を観察する場合でも、層を経るごとに単語の意味で軸が分離されていることが確認できる。

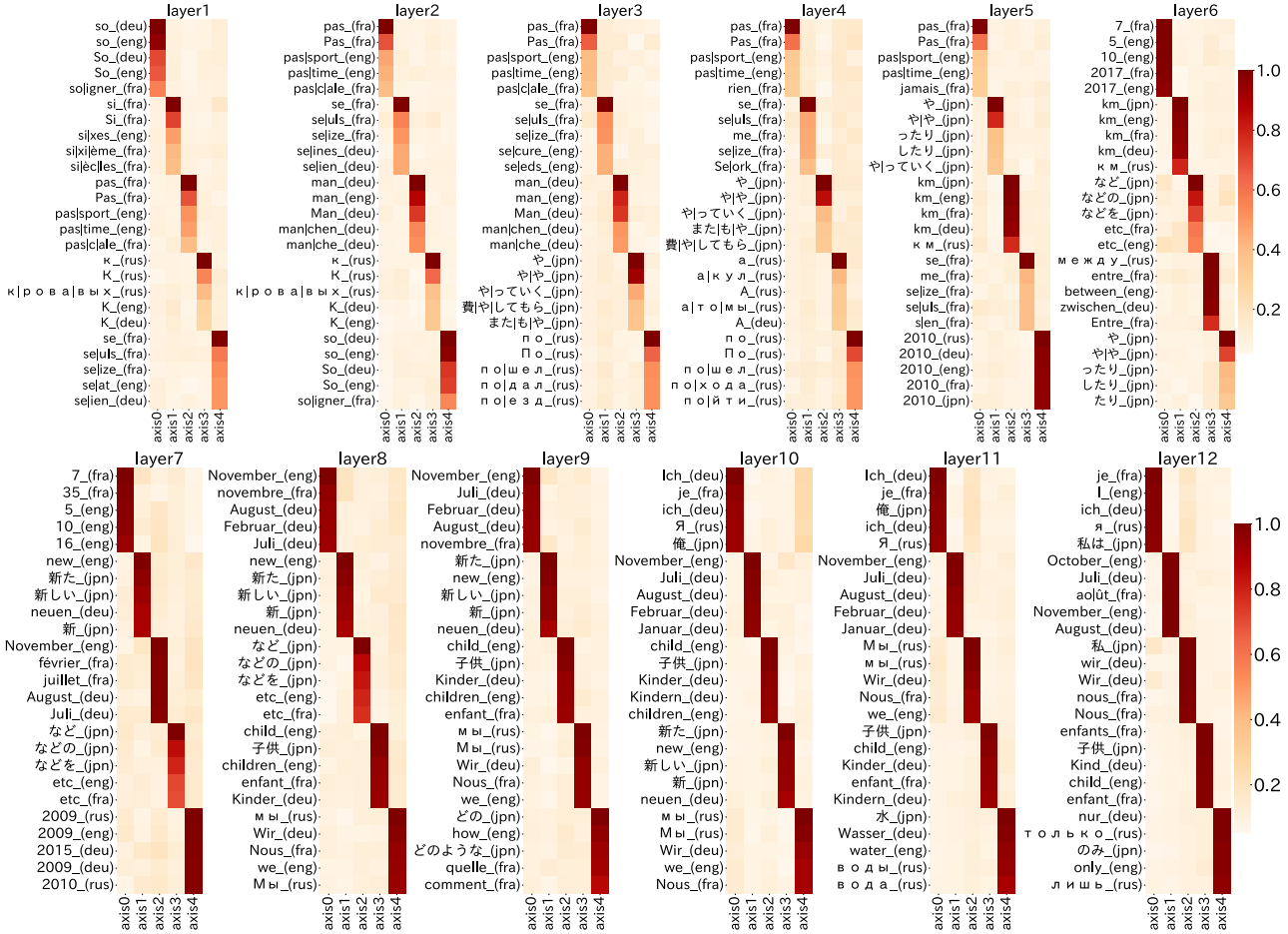

図 3　M2M100 モデルの全ての層における ICA 後の埋め込み