スパースオートエンコーダーを用いた 大規模言語モデルのチェックポイント横断分析

稲葉 達郎 [1,2] 乾 健太郎 [3,4,5] 宮尾 祐介 [6,2] 大関 洋平 [6]
Benjamin Heinzerling[5,4*] 高木 優 [2*]

[1] 京都大学 [2] 国立情報学研究所 大規模言語モデル研究開発センター
[3]MBZUAI [4] 東北大学 [5] 理化学研究所 [6] 東京大学
inaba@sap.ist.i.kyoto-u.ac.jp kentaro.inui@mbzuai.ac.ae
yusuke@is.s.u-tokyo.ac.jp oseki@g.ecc.u-tokyo.ac.jp
benjamin.heinzerling@riken.jp yu-takagi@nii.ac.jp

概要

大規模言語モデルは優れた多言語能力と広範な知識を有するが，これらの能力が内部的に形成される過程は定かでない．本研究ではこの疑問を解明するため，モデルの内部表現に含まれる情報が学習経過に伴いどう変化していくかを分析する．具体的には，モデルのチェックポイント別にスパースオートエンコーダーを学習し，その解釈結果をチェックポイント横断で比較する．実験の結果，大規模言語モデルは言語を個別に学習した後に言語間の対応関係を学習すること，トークンレベルの知識を学習した後に抽象度の高い概念レベルの学習をすることが明らかとなった．

1 はじめに

大規模言語モデル（Large Language Models; LLMs）が驚異的な性能を発揮するにつれ，モデルが知識や思考を内部的にどのように表現し，処理しているかへの関心が高まっている [1]．大規模言語モデル内の表現が次元数以上の情報を持ちうること（Superposition）が，高い能力を発揮する一つの要因になっている一方で，内部表現の分析を困難にする原因にもなっている [2]．

このような多義的な内部表現を理解する手法の一つとして，内部表現を少数の単一意味表現（Feature）の重ね合わせに分解するスパースオートエンコーダー（Sparse Auto Encoder; SAE）がある [3, 4]．どのようなトークン系列を LLMs に入力した時に，各意味表現（Feature）が再構成に使用されるかを分析することで，内部表現が捉えている情報を解釈することができる．近年ではこれを応用し，内部表現の層間 [5, 6]・モデル間 [7, 8]・ファインチューニング前後 [8, 9] での比較分析が行われ，大規模言語モデルの内部機序に対する知見が蓄積されている．

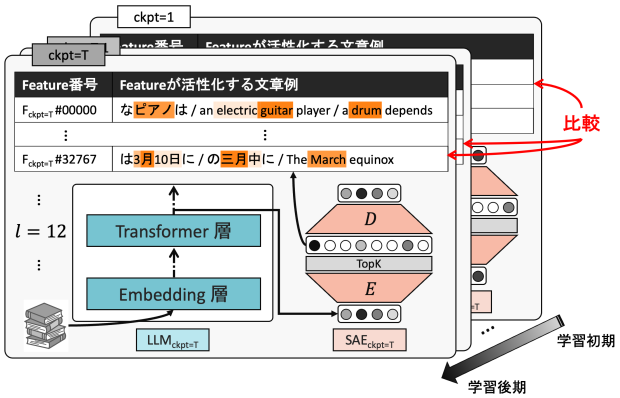

図 1 SAE の Feature の活性化傾向（≒ LLM の内部表現が含む情報）が学習経過でどう変化していくか．

しかし，学習過程において，大規模言語モデル内部がどう変化しているかは未だ明らかでない部分が多い．大規模言語モデルの驚異的な言語汎化能力が形成される過程を深く理解するためには，モデルの入出力だけでなく，モデル内部がどう進化しているかも理解することが重要である．

本研究では，SAE を大規模言語モデルの複数チェックポイントでそれぞれ学習し解釈結果の比較を行う．各 SAE が捉えている意味表現の傾向が学習経過でどう変化するかを分析することで，大規模言語モデルが内部的に捉える情報がどう進化していくかを可視化する．実験の結果，大規模言語モデル

* 共同最終著者．

は言語を個々に学習した後に，言語間の共通意味を習得する（4.3 章）こと，トークンレベルの情報を学習した後に，概念レベルの抽象度の高い知識を習得する（4.4 章）ことを定量的・定性的に確認した．

2 スパースオートエンコーダー

スパースオートエンコーダー（SAE）は，中間層が疎になるように制約をかけて学習を行うオートエンコーダーである．本研究では，TopK を中間層に適用する TopK-SAE を採用する [10]．ReLU を中間層に適用する ReLU-SAE [3, 4] に比べ，TopK-SAE は学習が容易でスパース性を保ちながらより高い再構成性能を発揮できる [11]．

入力ベクトルを $x \in \mathbb{R}^d$，中間層の次元数を n とすると，そのエンコーダーとデコーダーはそれぞれ以下のように定義される：

$$z = \mathrm{TopK}\left(W_{\mathrm{enc}}\left(x - b_{\mathrm{pre}}\right)\right) \tag{1}$$

$$\hat{x} = W_{\mathrm{dec}} z + b_{\mathrm{pre}} \tag{2}$$

ここで，$W_{\mathrm{enc}} \in \mathbb{R}^{n \times d}$ と $W_{\mathrm{dec}} \in \mathbb{R}^{d \times n}$ はそれぞれ入出力の学習可能な線形変換層であり，$b_{\mathrm{pre}} \in \mathbb{R}^d$ は学習可能なバイアスパラメータを表す．W_{dec} は $W_{\mathrm{enc}}^{\top}$ で初期化され，b_{pre} は入力データの幾何中央値（Geometric median）で初期化される．

次に，学習時の損失は次式で定義される：

$$L = ||x - \hat{x}||_2^2 \tag{3}$$

ここで，$||x - \hat{x}||_2^2$ は平均二乗誤差であり，入力ベクトルを復元するように学習が行われる．

TopK-SAE の調整可能パラメータは，中間層の次元を決める拡大係数（n/d）とスパース性を制御する K の二つである．W_{dec} を n 個の d 次元ベクトルとみなすと，Top-K SAE は n 個の中から K 個のベクトル表現を選び，その重み付きベクトル和として入力ベクトルを再構成するネットワークとみなせる．本研究では，エンコーダーの出力（$z \in \mathbb{R}^n$）の各要素を Feature と呼び，ある Feature に対応するベクトル表現が再構成に使用される時（TopK に選ばれた時），その Feature が活性化しているとみなす．

3 予備実験

チェックポイント横断で実験を行う前に，学習済みモデルの内部表現で SAE のパラメータ調整と，解釈可能な分析結果が得られるかの確認をする．

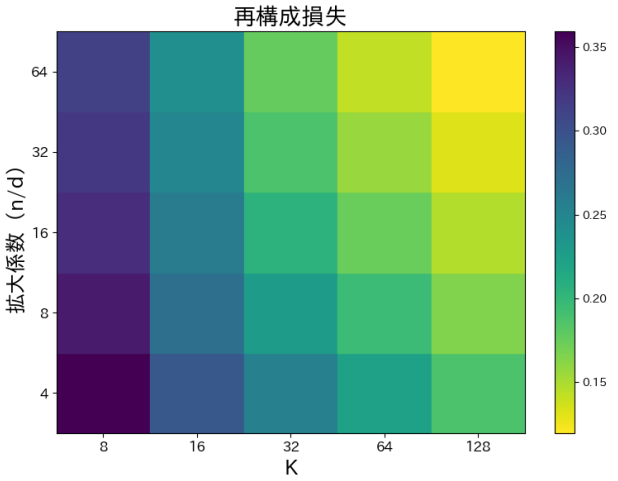

図 2 拡大係数と K に対する再構成損失．拡大係数と K が大きくなるほど再構成精度が向上する．学習率はグリッドサーチを行い最適なものを選択した．

3.1 学習設定

事前学習済みの 24 層からなる LLM-jp-3-1.8B [1] の 12 層目の出力（$d = 2048$）で TopK-SAE を学習する．データは LLM-jp Corpus v3[2] に含まれる日本語 Wikipedia と英語 Wikipedia を一対一で使用する．文書の最初 65 トークンを LLM へ入力し，[BOS] トークンの表現を除いた 64 トークン分の中間表現に対し L2 正規化を行なったものを SAE への入力とする．日英混合計 165M トークン分の表現のうち 10 %を検証用，10 %をテスト用，残りの 80 %を訓練用に割り振る．バッチサイズは 32,768 に固定し，ウォームアップステップは 1,000 とする．

3.2 拡大係数（n/d）・K と再構成損失

図 2 に拡大係数・K と再構成損失の関係を示す．拡大係数を大きく・K を大きくするほど，再構成損失が減少することが確認された．ただし，拡大係数と K の値は，入力表現の復元に使用できるベクトルの種類および数に密接に関係している（2 章参照）．そのため，拡大係数や K が大きすぎる場合，一つの概念が複数の特徴（Feature）に分割される可能性がある．逆に，小さすぎる場合には，複数の概念が一つの特徴に内包される可能性がある．再構成精度を保ちながら人間にとって解釈しやすい Feature を得るためのパラメータ探索は今後の研究課題の一つである [12, 13]．なお，特記がない限り，本章以降では

1) https://huggingface.co/llm-jp/llm-jp-3-1.8b
2) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

	Feature番号	Featureが活性化する文章例	言語傾向（4.3章）	意味粒度（4.4章）
(a)	$F_{ckpt=100}$ #00002	・）は、「日本の貴婦人 ・または HMG-CoA レダクターゼ ・ called radiological pollution, is	日英混合	無関連
	$F_{ckpt=100}$ #00004	・から20世紀前半にかけて ・は日本の防衛官僚。 ・ investigations are performed by geotechnical	日英混合	無関連
(b)	$F_{ckpt=10000}$ #00004	・ dorsalis), also known as the scrub ・ regnans, known variously as ・ nerve) also known as the fourth	英語	トークンレベル:「 known」
	$F_{ckpt=10000}$ #00009	・石油生産設備から ・冷暖房設備、冷凍冷蔵設備、動力設備又は ・のプラント設備を	日本語	トークンレベル:「設備」
(c)	$F_{ckpt=988240}$ #00009	・ここで言う「都市」には ・, where fluency is defined as linguistic ・ ."Arbitrary" here means that the	日英混合	概念レベル（同義）:「定義」
	$F_{ckpt=988240}$ #00016	・特有の臭気のある白色個体で、 ・物で、白色の粉末である ・. It is a colorless liquid with a smell reminiscent	日英混合	概念レベル（意味的共通性）:「物質の特性」

図 3　(a) 学習初期（ckpt=100），(b) 学習中期（ckpt=10000），(c) 学習済みモデル（ckpt=988240）の内部表現でそれぞれ SAE を学習し，得られた Feature の例．(a) 学習初期は無関係な文章に活性化しているが，学習が進むにつれ，(b) 言語ごとやトークンレベルの意味を捉え，(c) 言語間に共通した意味や抽象的な概念を捉えるようになる．その他の例は図 6 参照．

拡大係数を 16，K を 32 として実験を進める．

3.3　Feature の活性化パターン

図 3 (c) に Feature の活性化パターン例を示す．文字の背景色の濃淡が活性化値の大小を表している．9 番目の Feature（$F_{ckpt=988240}$#00009）は言葉の定義を行う部分で強く活性化し，16 番目の Feature（$F_{ckpt=988240}$#00016）は物質の匂いや色，状態について述べる部分で強く活性化している．その他の活性化パターン（図 6 (c)）からも十分に解釈性の高い Feature が獲得できていることが確認できた．

4　実験

学習チェックポイントごとに SAE を学習し，Feature の活性化パターンの変化を分析する．

4.1　学習設定

LLM-jp-3-1.8B の学習チェックポイント 10, 100, 1000, 10000, 100000, 988240（学習済みモデル）の 6 つを分析対象のモデルとする．SAE の拡大係数は 16，K は 32 とし，その他の学習条件は 3.1 章と同一である．

4.2　活性化パターンの評価法

Feature を活性化させる文章をそれぞれ最大 50 個用意し，その活性化パターンの言語傾向と意味粒度を分類する．

言語傾向　Feature を活性化させる文章・トークンの言語をルールベースで判定し，日本語が 90 %以上の場合は「日本語」，英語が 90 %以上の場合は「英語」，その他の場合は「日英混合」とする．各チェックポイントごとに全ての Feature の言語傾向を集計する．

意味粒度　各 Feature について，その意味的な粒度を以下の 4 つのカテゴリーに分類する：

- **トークンレベル**：同一のトークンに活性化が観察される場合（例：「猫」と「猫」）
- **概念レベル（同義）**：同一の意味を表すトークンまたは文に対して活性化が観察される場合（例：「猫」と「ネコ」）
- **概念レベル（意味的共通性）**：意味的共通性を持つトークンまたは文に対して活性化が観察される場合（例：「猫」と「犬」）
- **無関連**：活性化するトークンや文章に解釈可能な関係性が見られない場合

各チェックポイントごとに 100 個の Feature に対し，それぞれの意味粒度を著者が手動で分類する．

4.3　言語傾向の変化

図 4 に結果を示す．学習初期には日英混合 Feature が大半を占め，その後学習が進むにつれてその割合は一度減少し，再び増加する傾向を示した．学習初期の日英混合 Feature が活性化する文章（図 3 (a) と図 6 (a) 参照）では，意味的関連性を持たない無作為

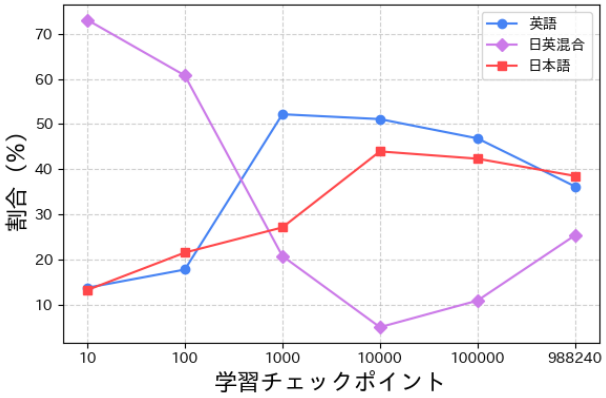

図 4　学習チェックポイント別の各言語傾向の割合

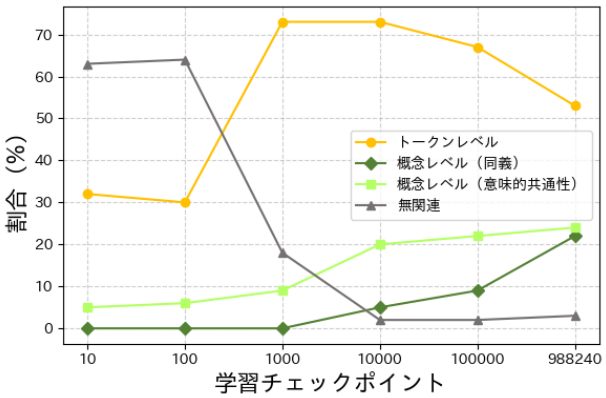

図 5　学習チェックポイント別の各意味粒度の割合

なトークンに活性化する傾向が観察された．一方，学習後期（図 3 (c) と図 6 (c) 参照）では，日英間で同一の意味を持つトークンや文章に対して活性化することが確認された．

対照的に，英語 Feature と日本語 Feature は学習中期に高い割合を示した．学習中期の日英混合 Feature（図 3 (b) と図 6 (b) 参照）は，同一言語内における類似した意味を持つトークンに対して活性化する傾向が確認された．

これらの分析結果から，大規模言語モデルの学習は以下の 2 段階で進行していることが示唆される：

1. 学習初期から中期にかけて，トークンや文章の意味を言語別に習得
2. 学習中期から後期にかけて，トークンや文章の言語間での対応関係を習得

4.4　意味粒度の変化

図 5 に結果を示す．学習初期から中期にかけてトークンレベルの意味表現を持つ Feature が増え，学習中期から後期にかけて概念レベル（同義・意味的共通性）の意味粒度を持つ Feature の割合が増加した．逆に，活性化パターンが無関連な Feature の割合は学習が進むにつれて減少した．

この分析結果は，大規模言語モデルが学習初期から中期にかけてトークンレベルの知識を獲得し，その後，中期から後期にかけて概念レベルの知識体系を構築していくことを示唆している．

5　関連研究

SAE を用いた往来研究の多くは，単一の SAE の分析に焦点を当ててきたが，近年では異なる状態で訓練された SAE 間での比較分析が進められつつある．具体的には，層間比較 [5, 6]，異なるモデルアーキテクチャ間の比較 [7, 8]，ファインチューニング前後の比較 [8, 9] などがある．また，同時研究として各学習段階の大規模言語モデルで順に SAE を継続学習し，Feature が形成されていく過程を追跡した研究もある [14]．ただし，[14] は Feature 形成過程の定性的分析に留まっており，定量的評価が十分に行われていない．本研究の独自性は，SAE を各チェックポイントごとに独立で学習し，Feature の定性的分析に加え，定量的評価も実施している点にある．

大規模言語モデルの学習過程に関して注目されている現象の一つにグロッキング（grokking）がある [15]．グロッキングとは，モデルが過学習状態に陥った後，ある時点を境に突然未知のデータに対しても高い汎化性能を示すようになる現象を指す．この現象のメカニズムを解明するために，簡易なモデルとタスクによりグロッキング発生過程をリバースエンジニアリング的に解析した研究がある [16]．モデルの内部状態とグロッキングの関連性を解明することは今後の重要な研究課題の一つである．

6　おわりに

本研究では，スパースオートエンコーダーを用いて大規模言語モデルの内部表現をチェックポイント横断で分析した．その結果，大規模言語モデルは個別言語でのトークンや文章の意味を習得し，その後言語間の対応関係を理解する（4.3 章）ことと，トークンレベルの知識を習得した後に概念レベルの知識体系を構築している（4.4 章）ことが明らかとなった．今後は，（1）意味粒度判別の自動化，（2）他の評価軸の検討（3）言語モデルの様々な能力，（例：多言語能力，多段推論能力等）が発現するタイミングに焦点を当てた分析，（4）異なるモデルサイズ，多言語モデルでの実験，を行っていきたい．

謝辞

本研究成果は，データ活用社会創成プラットフォーム mdx を利用して得られたものです [17].

参考文献

[1] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety - a review. **Transactions on Machine Learning Research**, Aug 2024.

[2] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. **Transformer Circuits Thread**, 2022.

[3] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. **Transformer Circuits Thread**, 2023.

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

[5] Daniel Balcells, Benjamin Lerner, Michael Oesterle, Ediz Ucar, and Stefan Heimersheim. Evolution of sae features across layers in llms. **arXiv preprint arXiv:2410.08869**, 2024.

[6] Nikita Balagansky, Ian Maksimov, and Daniil Gavrilov. Mechanistic permutability: Match features across layers. **arXiv preprint arXiv:2410.07656**, 2024.

[7] Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models. **arXiv preprint arXiv:2410.06981**, 2024.

[8] Jonathan Marcus Thomas Conerly Joshua Batson Christopher Olah Jack Lindsey, Adly Templeton. Sparse cross-coders for cross-layer features and model diffing. **Transformer Circuits Thread**, 2024.

[9] Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. **arXiv preprint arXiv:2410.06672**, 2024.

[10] Alireza Makhzani and Brendan Frey. k-sparse autoencoders. **arXiv preprint arXiv:1312.5663**, 2014.

[11] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. **arXiv preprint arXiv:2406.04093**, 2024.

[12] Abhinav Menon, Manish Shrivastava, Ekdeep Singh Lubana, and David Krueger. Analyzing (in)abilities of SAEs via formal languages. In **MINT: Foundation Model Interventions**, 2024.

[13] Anonymous. Sparse autoencoders do not find canonical units of analysis. In **Submitted to The Thirteenth International Conference on Learning Representations**, 2024. under review.

[14] Yang Xu, Yi Wang, and Hao Wang. Tracking the feature dynamics in llm training: A mechanistic study. **arXiv preprint arXiv:2412.17626**, 2024.

[15] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. **arXiv preprint arXiv:2201.02177**, 2022.

[16] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In **The Eleventh International Conference on Learning Representations**, 2023.

[17] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

	Feature番号	Featureが活性化する文章例	言語傾向（4.3章）	意味粒度（4.4章）
(a)	$F_{ckpt=100}$ #00000	・大学国際教養学部教授。 ・が販売するクロスオーバーSUVである。 ・sphere (variations are known as spherical	日英混合	無関連
	$F_{ckpt=100}$ #00005	・そつぎょうけんてい）とは、 ・うめによる日本のゲーム業界漫画 ・Railroad's class N2sa comprised rebuilds	日英混合	無関連
(b)	$F_{ckpt=10000}$ #00000	・イギリスの女流小説家。 ・女流棋士初の女流タイトルグランドスラム ・の女流王将戦である。	日本語	トークンレベル:「女流」
	$F_{ckpt=10000}$ #00005	・guy-wired aerial masts for ・Aerial reconnaissance spotted a ・flapping-winged aerial robot, and	英語	トークンレベル:「erial」
	$F_{ckpt=10000}$ #00008	・In late mornings and during ・a Saturday morning animated series ・licensed to Morningside, Maryland	英語	トークンレベル:「morning」
	$F_{ckpt=10000}$ #00010	・live flagship daytime show. It ・both daytime and primetime television. ・, and only 1 watt nighttime	英語	トークンレベル:「time」
	$F_{ckpt=10000}$ #00011	・ある。類語には鶏鳴の助や ・ジャーゴン（俗語、隠語）である。 ・微妙」の略語。開経	日本語	トークンレベル:「語」
	$F_{ckpt=10000}$ #00048	・皇女として生まれ、のちにボイオーティアの ・生まれや生い立ちは不明だが時宗の ・裕福な家庭で育つが、父親から「	日本語	概念レベル（意味的共通性）:「出生や生い立ち」
	$F_{ckpt=10000}$ #00058	・近接型の工業団地・ニュータウン ・音楽・工芸（クラフトとフォークアート） ・工場制機械工業（こうじょうせい	日本語	概念レベル（同義）:「工業」
(c)	$F_{ckpt=988240}$ #00006	・about a mile (1.6 km) east of the ・36.6 square miles (94.8 km), of ・Located 4 miles north from Wasilla	英語	トークンレベル:「mile(s)」
	$F_{ckpt=988240}$ #00007	・旧表記（数え年）にて表記。 ・0から数え始め、1 ・一つに数えられることがある。	日本語	トークンレベル:「数え」
	$F_{ckpt=988240}$ #00017	・津海道（しんかい-どう）は ・かけての津藩の藩士である。 ・は岡山県御津郡にあった村。	日本語	トークンレベル:「津」
	$F_{ckpt=988240}$ #00021	・for cyclists (e.g. cyclist-only paths ・itself, for example on signage. ・languages spoken, such as Belgium	英語	概念レベル（同義）:「例示」
	$F_{ckpt=988240}$ #00026	・よりはダーク・ファンタジー ・Darkened Skye is a ・ baryonic dark matter is hypothetical dark matter	日英混合	概念レベル（同義）:「ダーク」
	$F_{ckpt=988240}$ #00039	・に上海美術映画作成所より制作された ・The game was developed by Beam Software ・It was part of Mutual Film Corporation's	日英混合	概念レベル（意味的共通性）:「社名」
	$F_{ckpt=988240}$ #00041	・is located nine kilometers south-west of ・airport located 13 km northwest of ・airport located seventeen miles (	英語	概念レベル（意味的共通性）:「距離数値」

図 6 (a) 学習初期（ckpt=100），(b) 学習中期（ckpt=10000），(c) 学習済みモデル（ckpt=988240）の内部表現でそれぞれ SAE を学習し，得られた Feature の例.