# TrendScape 1.0: 言語モデルの潜在空間上の概念探索

本田純也 [1,5]　坂本航太郎 [2]　高木志郎 [3]　林祐輔 [4]
小川修平 [5]　松尾豊 [2]
[1] 豊橋技術科学大学　[2] 東京大学
[3] 独立研究者　[4] 一般社団法人 AI アライメントネットワーク　[5] 株式会社エモスタ
{kotaro.sakamoto,matsuo}@weblab.t.u-tokyo.ac.jp　{jhonda,sogawa}@emosta.com
hayashi@aialign.net

## 概要

本稿では，言語モデル内の潜在的な概念空間を探索・可視化するための手法である TrendScape 1.0 を紹介する．本手法では，自然言語を潜在空間にマッピングして近傍グラフを構築する．グラフ上の経路探索を通じて概念間の関係を調べる．手法の検証として，文学作品間の概念経路を可視化し，得られたネットワークを分析することで，言語モデルの概念理解に関する洞察を提供する．

## 1　はじめに

概念と表象に関する研究は長年哲学の分野で行われてきた [1]．プラトンの「洞窟の寓話」にあるように，人間の知覚する世界は実在の影に過ぎないという考え方がある．これは，我々の理解する概念が，実際の世界とは別に存在するという見方です．近年，AI や深層学習モデルの発展に伴い，この古典的な問いが新たな形で再考されています．AI モデル，特に大規模言語モデルにおける概念の表現が，プラトンのこの考えに類似していると指摘した [2]．これを「プラトン的表象仮説 (The Platonic Representation Hypothesis)」と名付けた．一方，言語学の分野では，ソシュールが概念は他の概念との「差異」によって意味づけられると論じた．現代の計算言語学では，これらの考えを踏まえ，単語や文をベクトル空間上に表現する手法が発展している．例えば，skipgram モデルは単語の分散表現を学習し，King - Man + Woman = Queen のような類推を可能にする．

さらに，大規模言語モデル内部の概念表現を可視化・解析する試みも進んでいる．これらの研究は，モデルが獲得した概念同士の関係性を明らかにし，AI システムの解釈可能性を高めることを目指している．本研究では，大規模言語モデルにおける概念の潜在空間を分析し，モデルがどのように概念を認識・処理しているかを探る．これにより，AI システムの概念理解メカニズムの解明に貢献し，より洗練された自然言語処理技術の開発につながることが期待される．

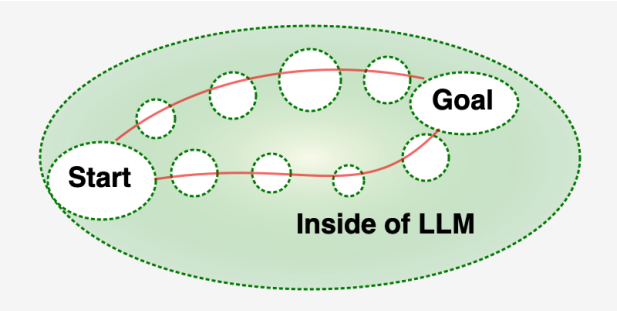



図 1: 提案法 TrendScape 1.0 の概念図

## 2　提案法

本研究では，言語モデルにおける潜在的な概念空間を分析するために，TrendScape 1.0 法を提案する．この方法は，入力を潜在空間にマッピングし，概念間の関係の探索と可視化を可能にする．提案するアプローチは 5 つの主要なステップから構成される．まず，入力は潜在空間における埋め込み表現と整列され，格納される．次に，探索空間のサンプリングが行われ，例えば概念 A から概念 B へのベクトル空間の関係が抽出され，それがネットワークに構造化される．続いて，探索空間を再サンプリングしながら，パスの周辺を探索する．最後に，空間内の関係が再構築され，可視化される．

この方法は，Word2Vec [3] のものと Sparse Autoencoder (SAE) [?, ?] の 2 つの異なるアプローチを採用している．これらの視点を組み合わせることで，より包括的な概念の分析が可能となる．Word2Vec

バージョンは “word2vec-google-news-300” モデルの単語埋め込み空間を利用する．このアプローチでは，「word2vec-google-news-300」がモデルとして機能し，その言語空間が単語埋め込み空間として扱われる．探索のための概念は，テキストを特徴付ける単語をトピックモデル化することによって決定される．

対照的に，SAE バージョンは Gemma SAE モデルを採用し，その中間層を言語空間として扱う．この手法の特筆すべき点は，文脈を含む単語を扱い，約 16,000 次元の特徴量を管理することである．

提案手法の有効性を検証するために，2 つのアプローチを採用した．第一に，Word2Vec を用いて単語の差分間の関係を構築し，概念の探索が可能かどうかを検証する．第二に，SAE により文脈化されたトークンを用いた概念探索の可能性を検証する．これらの手法により，言語モデル内の概念表現を可視化・分析し，モデルがどのように概念を認識・処理するかを探る．さらに，Word2Vec アプローチと SAE アプローチを比較することで，包括的な概念理解の背後にあるメカニズムを解明することを目指す．これらの知見は，自然言語処理の進歩に貢献し，AI システムの解釈可能性を向上させることが期待される．

# 3 実験設定

## 3.1 データセット

物語作品としては，グリム童話の代表的な 3 作品を選択した：

- 『ヘンゼルとグレーテル』
- 『ラプンツェル』
- 『白雪姫』

これらの作品は，テーマが異なり，多様な概念空間を提供すると考えられる．グリム童話のテキストは Project Gutenberg からダウンロードし．得られた HTML ファイルに対して BeautifulSoup を使用して HTML ファイルからテキストを抽出し，章ごとに分割した．

論文のデータとしては，“Learning representations by back-propagating errors“ と “Parallel visual computation“ の 2 本の論文の abstract を使用した．

## 3.2 トピックモデル

テキスト分析は 2 段階で行われた．まず，テキストの分割と前処理を行い，次にトピックモデリングを適用した．

1. テキスト分割：SudachiPy を用いて各作品のテキストを分かち書きした．この過程では，カスタムクラス SudachiText を実装し，より細かな制御を可能にした．具体的には以下の処理を行った：

- 品詞フィルタリング：名詞，動詞，形容詞のみを抽出し，文の主要な意味を担う語彙に焦点を当てた．
- ストップワードの除去：一般的な語を除外し，作品特有の語彙を浮き彫りにした．
- 数字や特殊文字の除去：テキストのノイズを低減し，純粋な言語的特徴に注目した．

これらの処理により，各作品の本質的な語彙構造を抽出することができた．トピックモデリングによって文書集合のトピックをモデル化した．具体的には，gensim ライブラリを用いて LDA（Latent Dirichlet Allocation）モデルを各作品に適用した．LDA は文書集合に潜在するトピックを確率的に推定する手法であり，本研究では以下の手順で実施した：

- 文書-単語行列の作成：前処理済みのテキストから，各単語の出現頻度を記録した行列を生成.
- トピック数の設定：予備実験の結果を踏まえ，各作品のトピック数を 5 に設定.
- LDA モデルの学習：ギブスサンプリングを用いてモデルを最適化.
- トピックの可視化：pyLDAvis を用いて，得られたトピックを視覚的に表現.
- 特徴的単語の抽出：各トピックにおいて最も重要度の高い単語を抽出し，作品の特徴を把握.

この過程により，各作品の主要なテーマや特徴的な概念を客観的に抽出することができた．

### 3.2.1 Word Embeddings

テキスト分析で抽出された単語群を，計算可能な形式に変換するため，事前学習済みの単語埋め込みモデル “word2vec-google-news-300”（version 1.3-mc90）を使用した．このモデルは，Google News データセットを用いて訓練されたモデルである．このモデルでは約 300 万の単語と語句を 300 次元のベ

クトル空間で表現していることです．

gensim ライブラリを用いて “word2vec-google-news-300” モデルをロードし，抽出された各単語を対応する 300 次元ベクトルに変換した．

このモデルは，大規模な日本語コーパスで学習されており，日本語の意味的・構文的特徴を 300 次元のベクトル空間に効果的に符号化している．

#### 3.2.2 Trendscape 法の適用

Trendscape 法は，単語埋め込みを基に作品間の概念的な「道筋」を可視化する本研究の核心的手法である．その適用は以下の 4 つの主要ステップで構成される：

1. 探索空間の構築：二つの作品（source と target）間の概念的橋渡しとなる単語群を特定するため，“word2vec-google-news-300” の’closer_than’ メソッドを用いてサンプリングを行った．このメソッドは，source と target の両方に「近い」単語を効率的に抽出する．さらに，計算効率と多様性のバランスを取るため，サンプリングした単語群からランダムに 500 語を選択し，初期探索空間とした．

2. ネットワーク構造化：抽出された単語群の関係性を構造化するため，scikit-learn ライブラリの AgglomerativeClustering を用いて階層的クラスタリングを実行した．クラスタリングの結果に基づいて単語間にエッジを作成し，その重みをコサイン類似度の 2 乗として設定した．この手法により，意味的に近い単語同士が強く結びついたネットワーク構造が形成された．

3. 経路探索：構築されたネットワーク上で，source から target への最短経路を探索した．この過程では，networkx ライブラリの最短経路アルゴリズムを基本としつつ，意味的な関連性をより重視するため，独自のヒューリスティック関数を実装した．具体的には，各ステップで top-k（k=3）の候補を考慮する確率的な探索を行い，局所的な最適解に陥るリスクを軽減した．

4. 結果の統合：探索の安定性と網羅性を高めるため，上記のプロセスを複数回繰り返し，得られた結果を統合した．最終的なネットワークは，pyvis ライブラリを用いて可視化し，概念間の関係性を直感的に理解できるよう工夫した．

これらのステップを通じて，作品間の概念的な架け橋となる単語の系列を特定し，その関係性を視覚化することに成功した．

2

### 3.3 追加実験 1：SAE を用いた拡張

Trendscape 手法の汎用性と拡張性を検証するため，より高度な言語モデルを用いた追加実験を行った．具体的には，Self-Explaining Autoencoder (SAE) を利用し，文レベルの埋め込みに基づく概念ネットワークの構築を試みた．

この追加実験は以下の手順で実施された：

1. データ準備：Project Gutenberg からグリム童話の英語テキストを取得し，BeautifulSoup を用いて章ごとに分割した．これにより，より大規模かつ多様なテキストデータセットでの検証が可能となった．

2. 文埋め込みの生成：Google 社の大規模言語モデル Gemma-2b の中間層（layer 2）の出力を利用し，各文を 16,384 次元の特徴ベクトルとして表現した．この過程で，各文を最大 512 トークンに切り詰め，モデルに入力した．

3. Trendscape 法の適用：生成された高次元ベクトルに対し，コサイン類似度に基づく距離行列を計算した．その後，階層的クラスタリングを適用してネットワーク構造を構築した．効率的な近傍探索のため，FAISS ライブラリを活用した．

4. 評価：生成されたネットワークを plotly ライブラリを用いて可視化し，文レベルでの意味的関連性を分析した．さらに，物語構造の観点から生成されたパスの解釈を試みた．

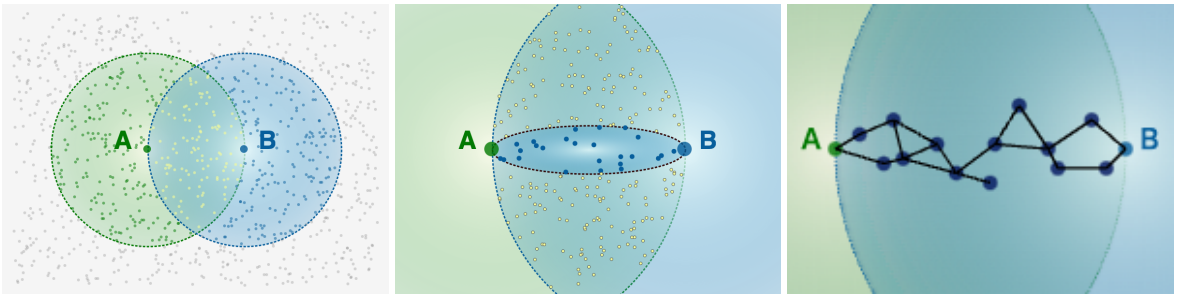
この追加実験により，Trendscape 手法が単語レベルだけでなく，より高次の意味表現に対しても適用可能であることを示すとともに，異なる言語や文学形式における概念ネットワークの構築可能性を探った．

以上の実験設計と手順により，Trendscape 手法の有効性と汎用性を多角的に検証した．次節では，これらの実験から得られた具体的な結果とその分析について詳述する．

## 4 実験結果

The results of the TrendScape visualisation of the concept space are shown in Figure 3. Only SNOWDROP with HANSEL AND GRETEL results are shown due to space constraints, for other results see Appendix **??**.

先述したように，ナイーブに LLM の概念空間上を遷移させると，例えば Iceland から Ireland に遷移する際に Iceland, Ieeland, ..., Ipeland, Iqueland といっ




(a) Exploration Space Sampling (b) Exploration Space Re-Sampling (c) Concept Network Visualization

図 2: The TrendScape method involves several interconnected steps to analyse and visualize conceptual relationships in latent spaces. It begins by mapping and storing the input to its corresponding embedding representation in the latent space. Next, it samples the exploratory space by extracting the vector space between two concepts (A and B) and constructing a network based on the relationships in this vector representation (a). The method then finds the conceptual path from A to B within this network. To refine the results, it performs a neighbourhood exploration around the identified path, resampling the exploratory space in the process (b). Finally, the method reconstructs and displays the relationships within the explored space, providing a visual representation of the conceptual connections (c).

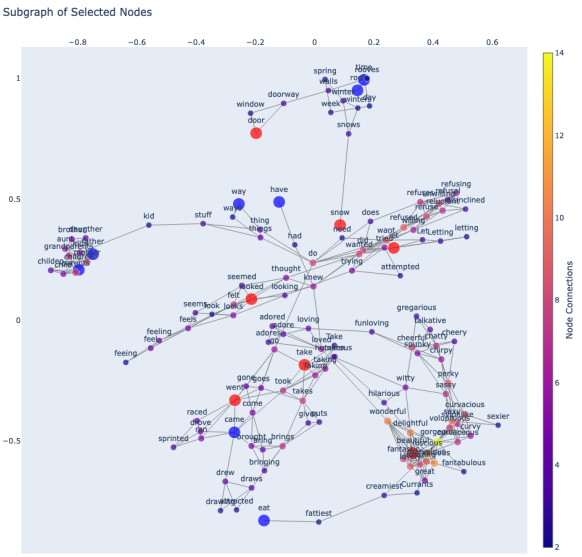



図 3: TrendScape between “SNOWDROP” and “HANSEL AND GRETEL”

たように，表層的な類似性でつながるモーフィングのようになってしまうかもしれず必ずしも有意味な潜在空間上を遷移するとは限りません．これに対して，図 **??** を見ると，例えば “came” という概念と “gone” という意味的に近い二つの概念がグラフ上の近い距離でパスで結ばれていることがわかります．これは，TrendScape が LLM の世界モデルや LLM が物語を意味的にどのように捉えているかを，より適切に可視化できている可能性を示唆します．

例えばこの図を見ると "winter", "snow", "snows" といった単語が一つのクラスターを形成していますが，これは SNOWDROP に特異な単語と考えられ，一方 "children" や "brother" といった表現は HANSEL AND GRETEL を代表するようなクラスターであることがわかります．一方で “do” や “knew” のような単語はどちらの物語にも出てくるような一般的な単語で，多くのクラスターの結節点となっています．これは 2 つの異なる物語の関係をどのように把握しているかについての考える上での示唆を与えるような結果と言えるかもしれません．

### 4.1 Additional Experiment: Extension Using SAE

Word2Vec の場合は先に単語の集団を作ってそれを埋め込んでいますが， SAE の場合は文章を入れてそのトークン毎の特徴量空間を作成してそれを埋め込んでいます．また，SAE の場合は文脈付きの単語の埋め込みをしています．また，Word2Vec は 300 次元全てで特徴を表現するのでコサイン類似度を使用したのに対して，SAE の時は特徴量がそれぞれ別の独立した特徴量として捉えられるため，cos 類似度による比較が適さないと考え，ワッサースタイン距離を用いて距離を測った．また，SAE は互いの同じ特徴量が高く活性化した時にのみ，同じ特徴を持っていると判断してエッジを結ぶようにした．

# 参考文献


[1] Jun Otsuka. What machine learning tells us about the mathematical structure of concepts. **arXiv preprint arXiv:2408.15507**, 2024.

[2] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In **Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024**. OpenReview.net, 2024.

[3] Tomas Mikolov. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.