文脈化された単語ベクトルからの単語領域表現

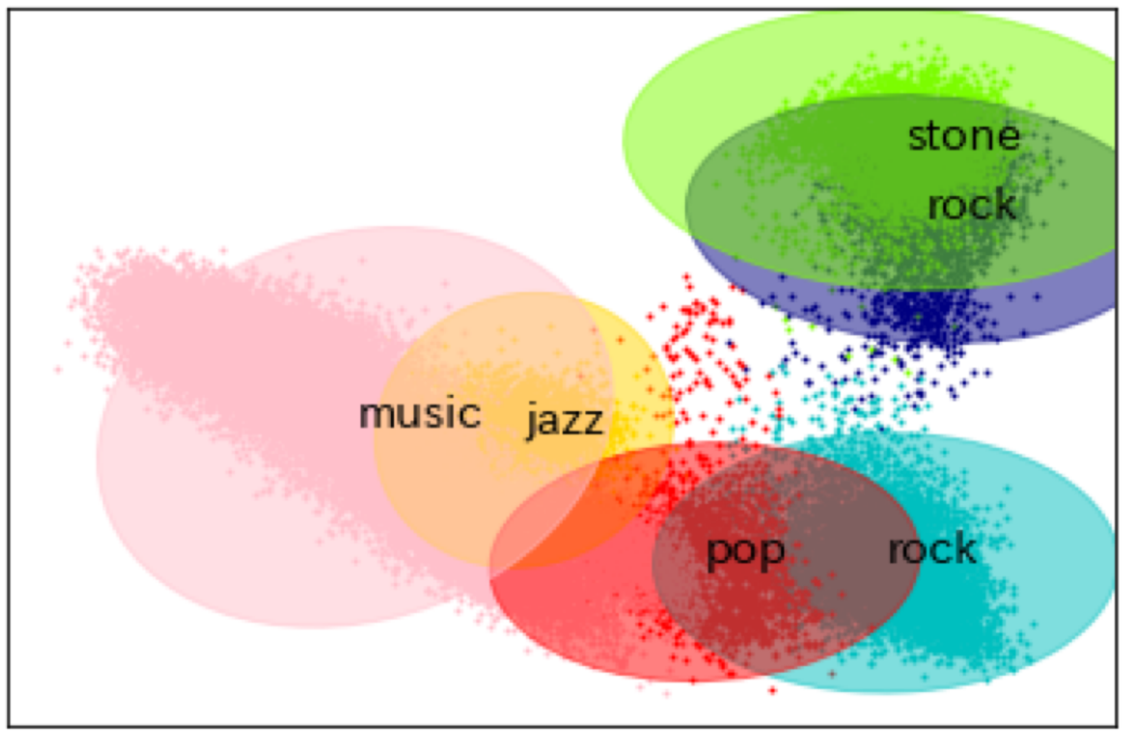
山内崇史*　梶原智之**　荒瀬由紀***

*大阪大学電子情報工学科　**大阪大学データビリティフロンティア機構　***大阪大学大学院情報科学研究科

OSAKA UNIVERSITY Live Locally, Grow Globally

研究背景

- **包含関係、上位下位関係は単語の重要な意味的関係である**
  - 検索や応答生成などの分野に応用される
- **静的な単語分散表現（Word2Vec, GloVe, etc.）[1, 2]**
  - １単語に対して１ベクトルを割り当てる
  - × 多義性を表現できない
  - × 意味の包含関係や上位下位関係の表現が困難
- **動的な単語分散表現（ELMo, BERT）[3, 4]**
  - 文脈ごとに分散表現を生成する
  - 多義性に対応
  - × 意味の包含関係や上位下位関係の表現が困難
- **既存の領域表現 [5, 6]**
  - 混合ガウス分布により多義性に対応
  - × モデルの訓練時に規定の窓幅の文脈しか考慮できない
  - × 語義数を動的に決定しない



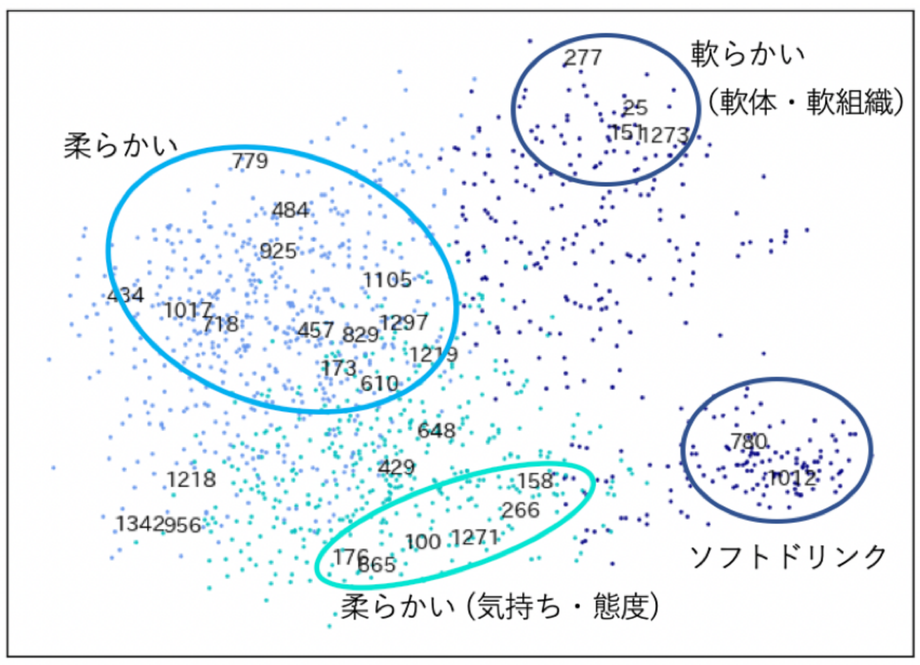
提案手法

- **BERTで得られる文脈化された単語ベクトルを混合ガウスモデルに当てはめる**
  - 意味的に近いベクトルが同じ領域に含まれることを期待
  - 文全体を考慮した単語表現を得られる
- **各ガウス分布が単語の意味に対応**
  - 多義性に対応
- **ガウス分布の分散によって領域の広がりを表現**
  - 単語間の意味の包含関係や上位下位関係を表現

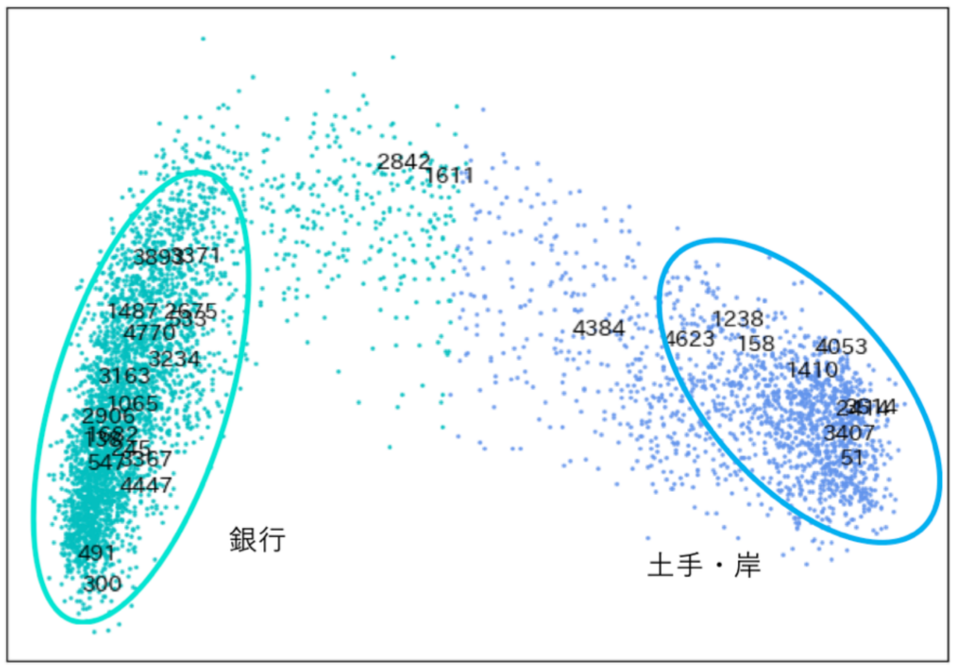
実験

- **意味的に近いものが同じ領域に属するかを可視化**
  - English Wikipedia 内の10万記事から得た‘soft’、‘bank’の単語ベクトルをプロット、混合ガウス分布を当てはめ

soft



bank



文番号	原文（一部抜粋）
610	has two coats - a harsh, wiry topcoat and a **soft** warm undercoat.
718	In contrast to southern Dutch cuisine, which tends to be **soft** and moist,
665	Todd claimed Coleman was **soft** on crime and
1012	appeared in advertising campaigns for batteries, a **soft** drink ,
1273	Eunicella cavolini is a much-branched **soft** coral growing to a height of about

文番号	原文（一部抜粋）
3407	, and lies on The southern **bank** of The River Clyde.
1410	While elements moved south to protect the north **bank** of the Loire River,
2842	several small creek running in different Directions **bank** generally high and dry
3163	It was at the time the largest **bank** failure in the history of the country,
3893	by debit or credit cards or online **bank** transfers.

今後の課題

- 語義数を自動で決定
- 領域表現を活かした新しい類似度尺度の提案：Stanford Contextual Word Similarity データセットで評価
- 包含関係、上位下位関係の表現： Lexical Entailment のタスクにより評価

References：[1] Mikolov et al. Efficient Estimation of Word Representations in Vector Space. In Proc. of ICLR, pp.1-12, 2013. [2] Pennington et al. GloVe: Global Vectors for Word Representation. In Proc. of EMNLP, pp.1532-1543, 2014.
[3] Peters et al. Deep Contextualized Word Representations. In Proc. of NAACL, pp.2227-2237, 2018. [4] Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
[5] Athiwaratkun and Wilson. Multimodal Word Distributions. In Proc. of ACL, pp.1645-1656, 2017. [6] Athiwaratkun et al. Probabilistic FastText for Multi-Sense Word Embeddings. In Proc. of ACL, pp.1-11, 2018.

大阪大学 OSAKA UNIVERSITY