文脈を考慮した単語ベクトル集合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

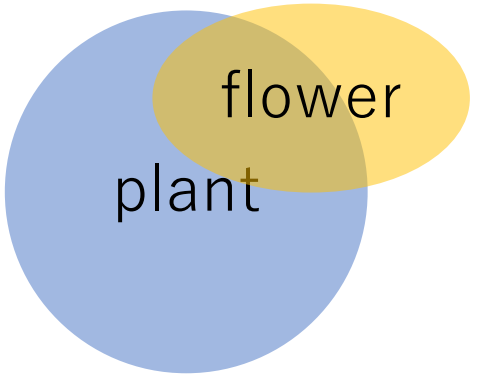
はじめに

研究背景

- 単語分散表現

  単語の意味を考慮したベクトル空間への埋め込み

- 一般的に各単語の意味を1つのベクトルで表現
  - 多義語における複数の意味が表現できない
  - 単語間の上位下位関係の表現が困難



研究目的

多義性や**上位下位関係**が表現可能な単語表現の獲得

単語分散表現

- 文脈を考慮しない単語分散表現（Word2Vec, GloVe）[1, 2]
  1単語を1ベクトルで表現
  多義性を表現できない

plant ⟶ [0.22, 0.91, 0.45,・・・]

- 文脈を考慮した単語分散表現（ELMo, BERT）[3, 4]
  各文脈ごとに分散表現を生成
  ⭕ 多義性を表現可能

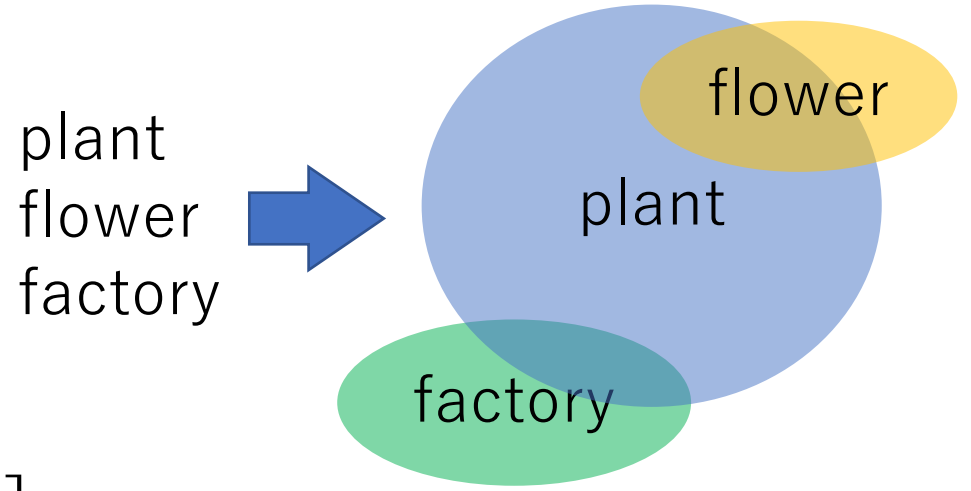
… is the **plant** manager in …
↳ [-0.12, 0.87, 0.05,・・・]
… a tropical **plant** called …
↳ [0.32, 0.13, -0.24,・・・]

単語間の上位下位関係の表現が困難

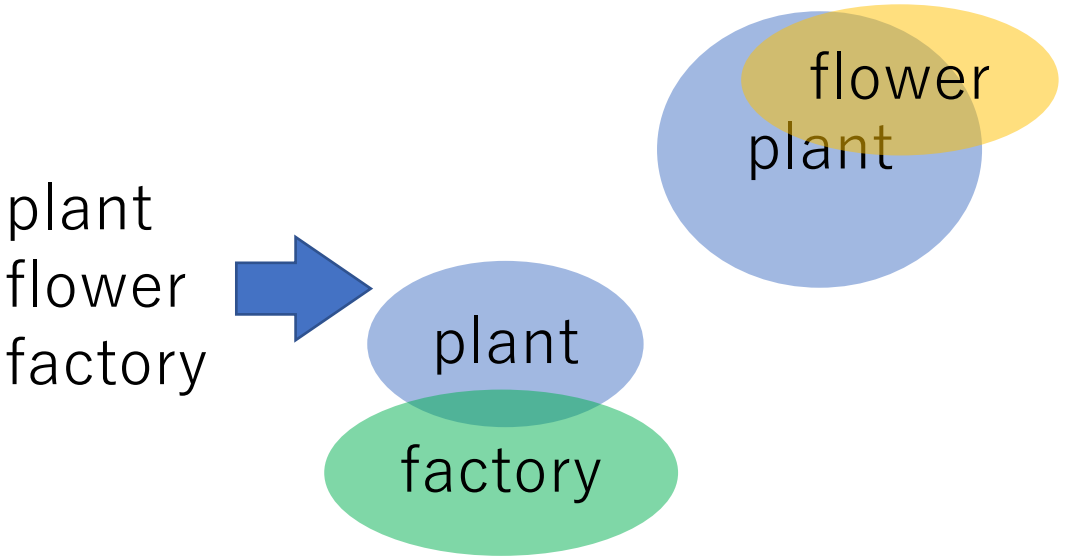
[1]Mikolov et al. Efficient Estimation of Word Representations in Vector Space.(ICLR 2013) [2]Pennington et al. GloVe: Global Vectors for Word Representation.(EMNLP 2014)
[3]Peters et al. Deep Contextualized Word Representations.(NAACL 2018) [4]Devlin et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding.(NAACL 2019)

単語領域表現

- word2gauss（w2g）[5]
  - 各単語をガウス分布で表現
  - 多義性を表現できない



- word2gaussian mixture（w2gm）[6]
  - 各単語を混合ガウス分布で表現
  - ⭕ 多義性を表現可能
  - 学習時に文全体を考慮しない
  - 全ての単語に同じ語義数



[5] Vilnis and McCallum. Word Representations via Gaussian Embedding. (ICLR 2015)
[6] Athiwaratkun and Wilson. Multimodal Word Distributions. (ACL 2017)

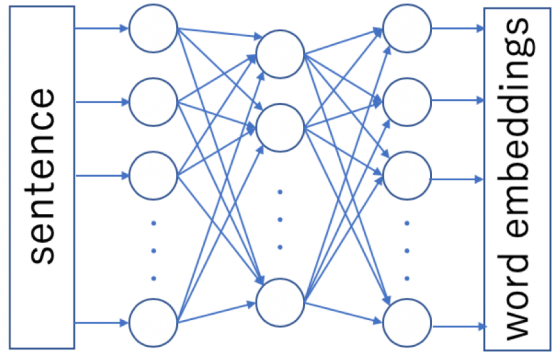
関連研究まとめ

	文脈	語義数	多義性	上位下位
Word2Vec	×	１単語１ベクトル	×	×
ELMo, BERT	文全体	１文脈１ベクトル	○	×
word2gauss	規定の窓幅 (W=5, 10)	１単語１分布	×	○
word2gm	規定の窓幅 (W=5, 10)	ハイパーパラメータとして設定 (K=2)	△	○
提案手法	**文全体**	**動的に決定**	○	○

提案手法

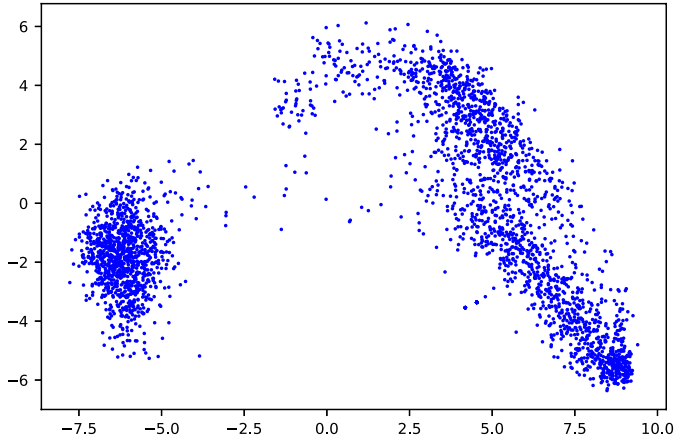
コーパス

ELMo, BERTの
訓練済みモデル



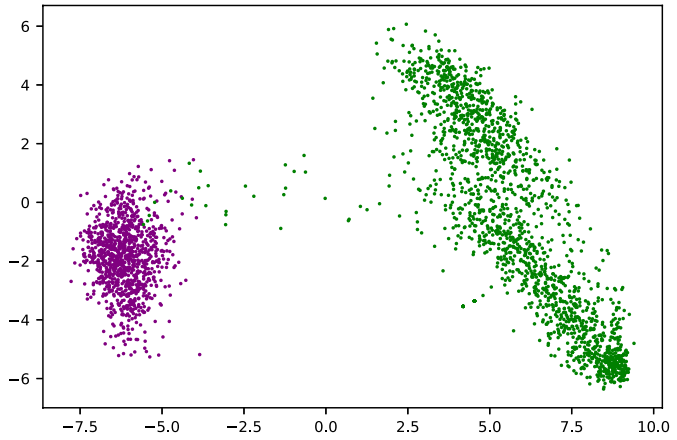
文脈ごとの単語分散表現

plant

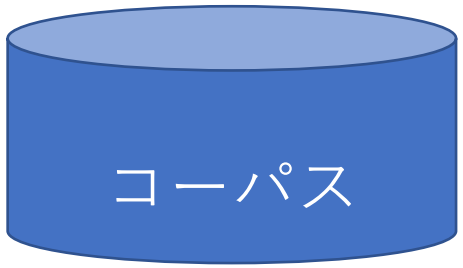


単語分散表現を
クラスタリング

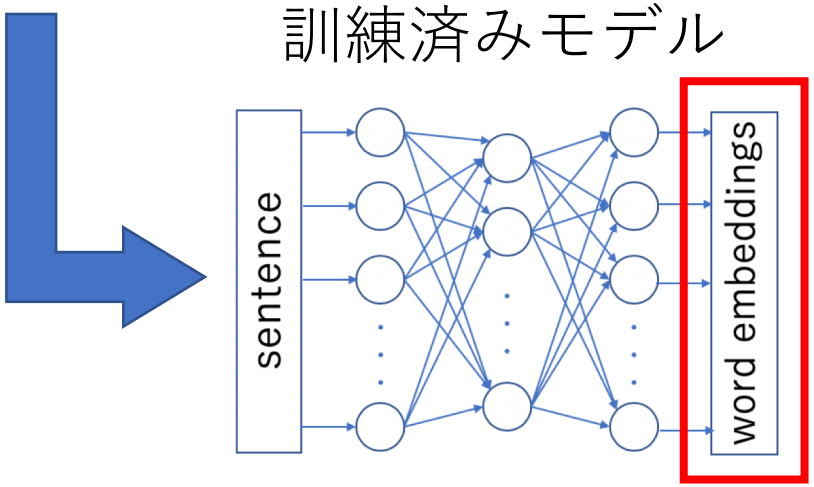
plant



提案手法

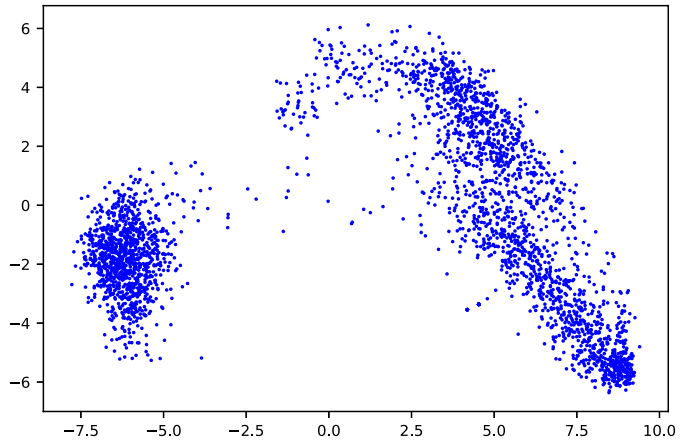


ELMo, BERTの
訓練済みモデル



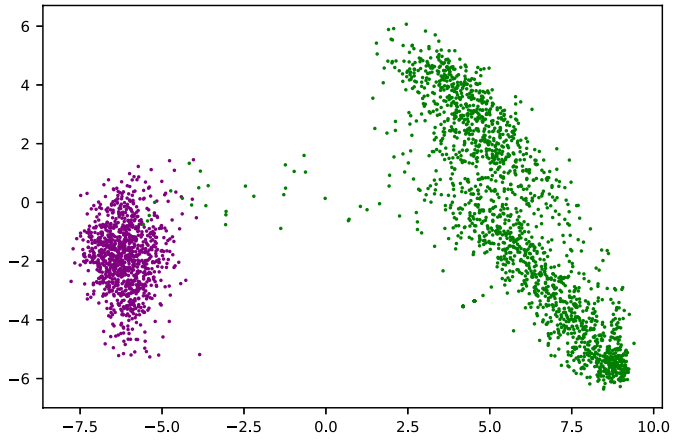
文脈ごとの単語分散表現

plant



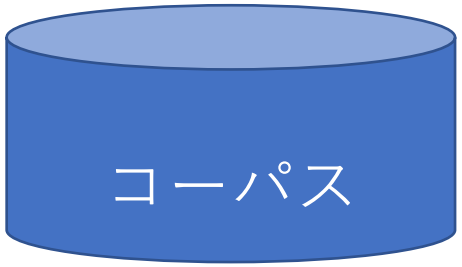
単語分散表現を
クラスタリング

plant

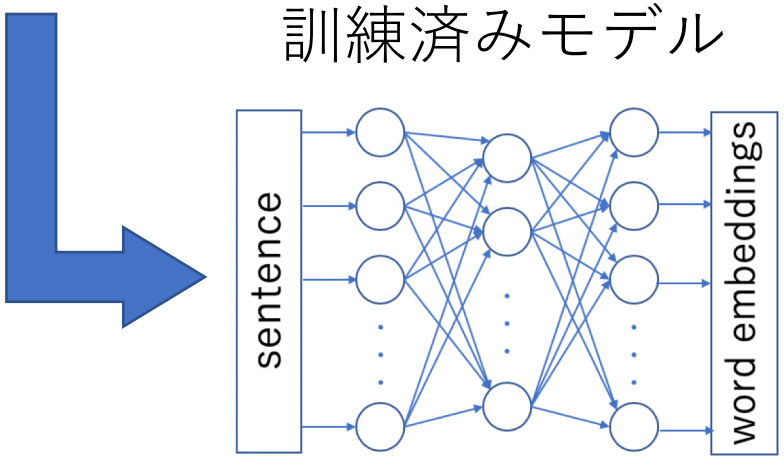


文脈を考慮した単語分散表現を得る

提案手法

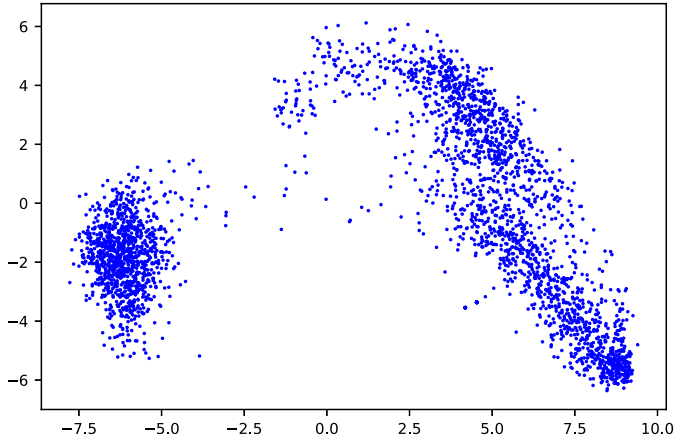


ELMo, BERTの
訓練済みモデル



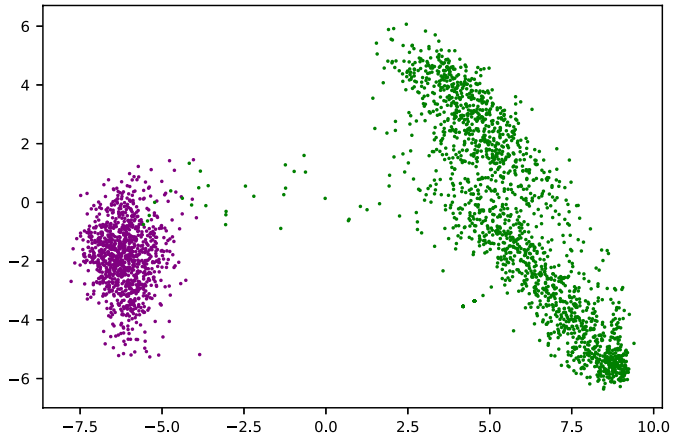
文脈ごとの単語分散表現

plant



単語分散表現を
クラスタリング

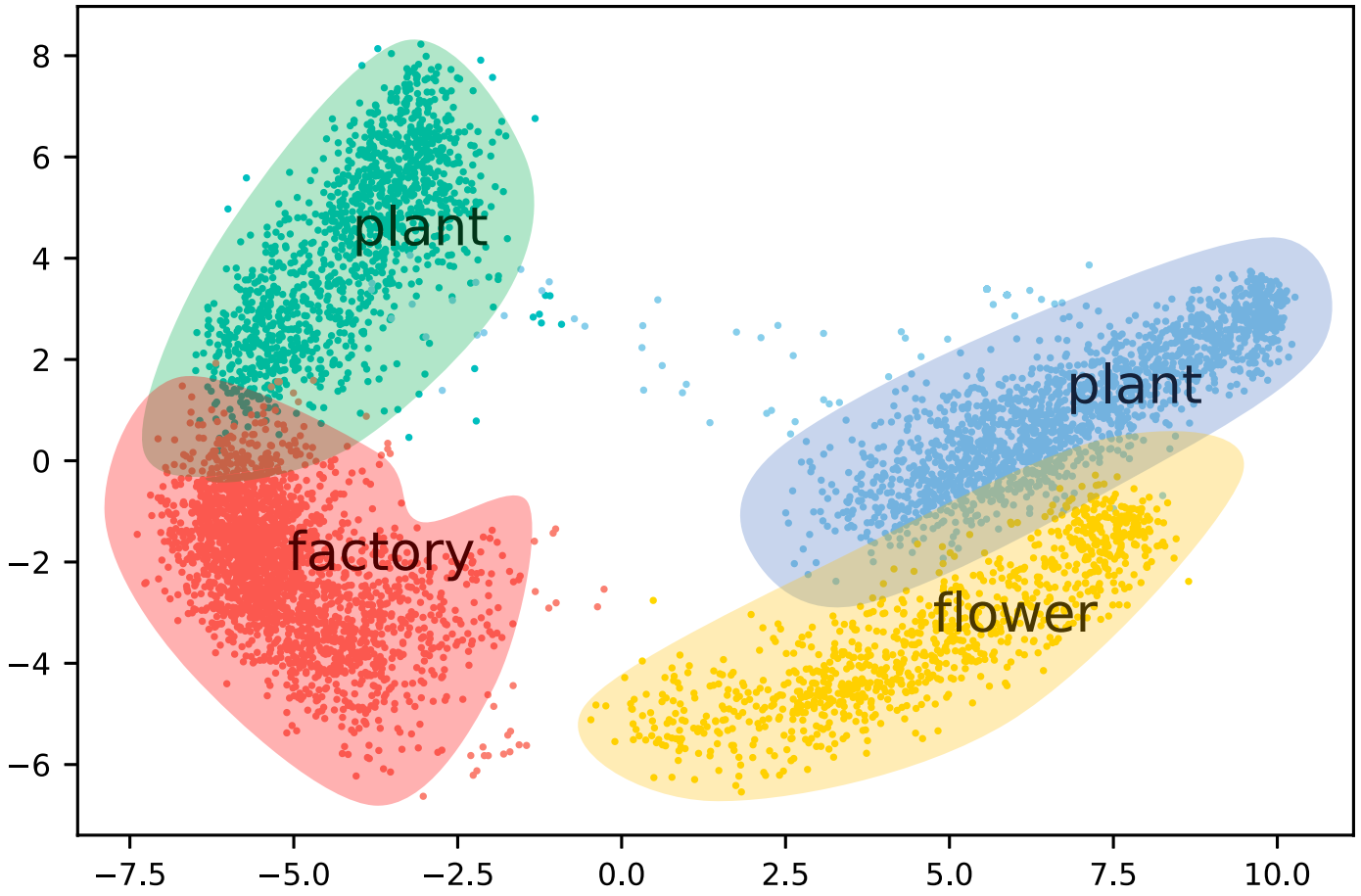
plant



得られた分散表現をDBSCAN[7]を用いて
クラスタリングし、**領域表現**を獲得

[7] Ester et al. A density-based algorithm for discovering clusters in large spatial databases with noise. (AAAI 1996)

領域表現の可視化



plant（工場、植物）が**factory**（工場）および**flower**（植物）と近い領域を持つ

→ 多義性を表現できている

評価実験

- 文脈を考慮しないタスク
  - 単語間の意味的類似度推定
  - 単語間の意味的関係推定

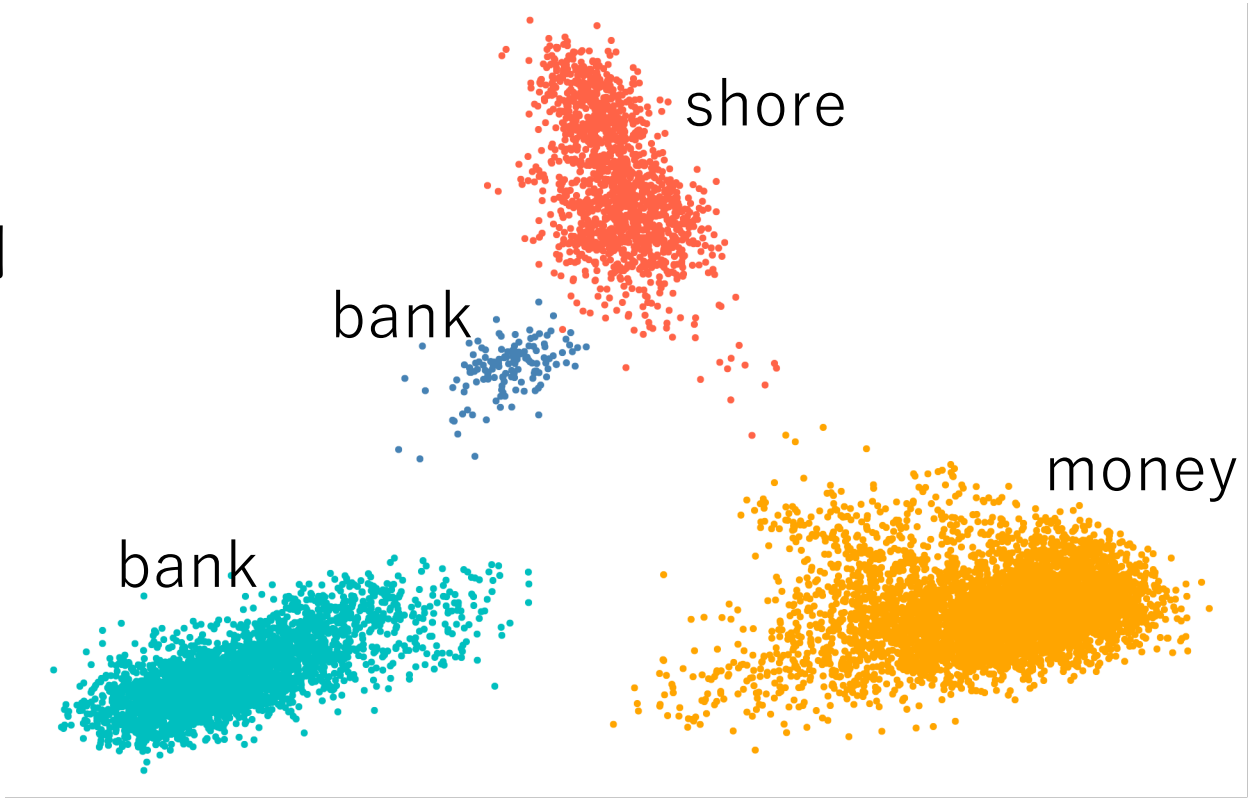
- 文脈を考慮するタスク
  - Stanford Contextual Word Similarity (SCWS)

単語間の意味的類似度推定

- データセット
  ９つの意味的類似度データセット（WS-353は開発セットとして用いる）
- 評価指標
  人手で付与されたスコアとのスピアマンの順位相関係数
- 比較手法
  w2g [6]、w2gm [6]
- 提案手法の推定スコア
  各単語の要素数が最大の領域の重心間
  コサイン類似度

データ例（ SimLex-999 ）

word1：bad	word1：bad
word2：terrible	word2：great
score ：7.78	score：0.35



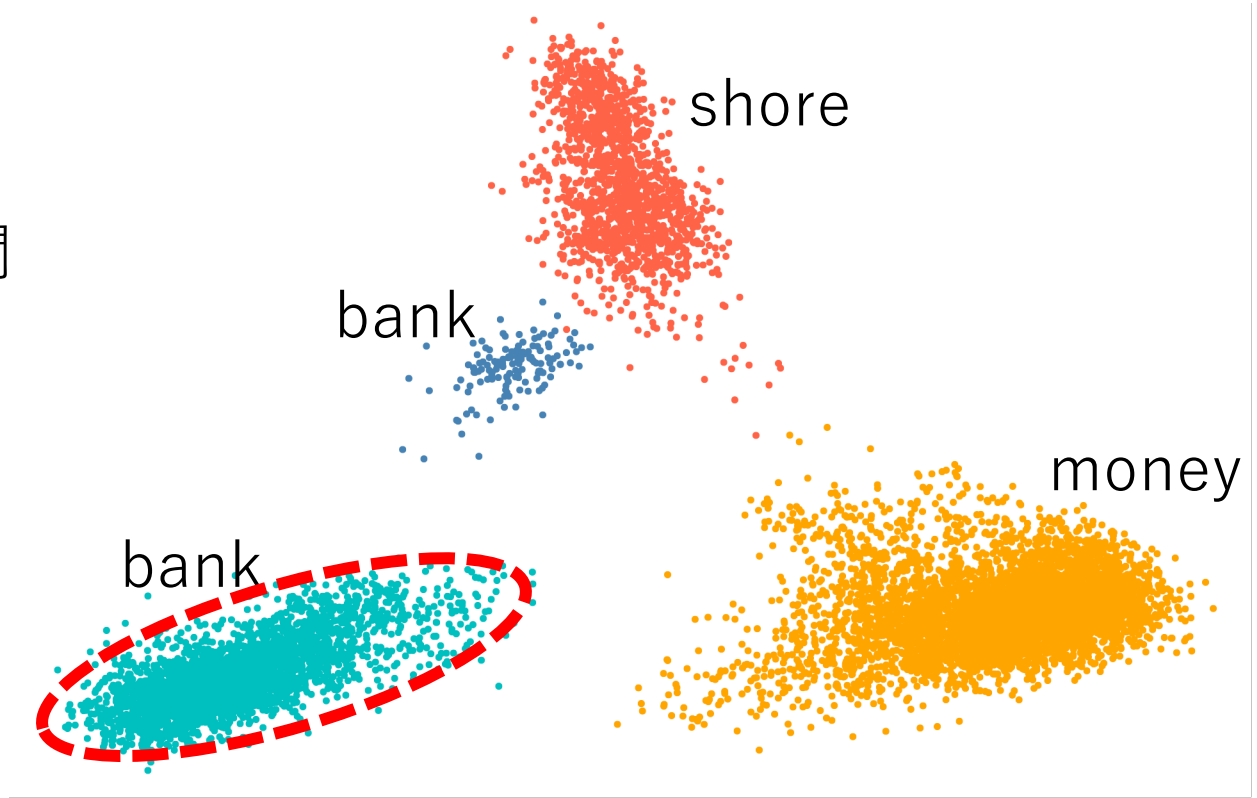
[6] Athiwaratkun and Wilson. Multimodal Word Distributions. (ACL 2017)

単語間の意味的類似度推定

- データセット
  ９つの意味的類似度データセット（WS-353は開発セットとして用いる）
- 評価指標
  人手で付与されたスコアとのスピアマンの順位相関係数
- 比較手法
  w2g [6]、w2gm [6]
- 提案手法の推定スコア
  各単語の要素数が最大の領域の重心間
  コサイン類似度

データ例（ SimLex-999 ）

word1：bad	word1：bad
word2：terrible	word2：great
score ：7.78	score：0.35



[6] Athiwaratkun and Wilson. Multimodal Word Distributions. (ACL 2017)

単語間の意味的類似度推定

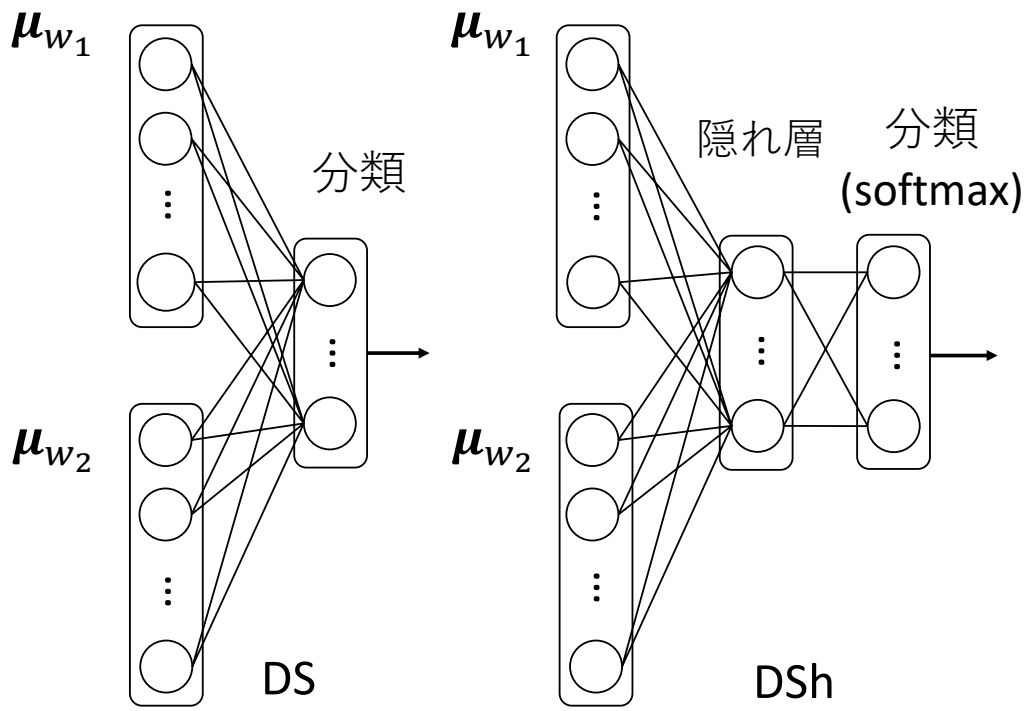
単語間の意味的類似度データセットにおけるスピアマンの順位相関係数

	SL	MEN	MC	RG	YP	micro平均
w2g [6]	29.4	72.6	76.5	73.3	42.0	52.4
w2gm [6]	29.3	**73.6**	79.1	74.5	45.1	52.9
ELMo+GMM	45.5	60.7	61.6	64.1	38.7	55.1
ELMo+DBSCAN	45.5	61.6	63.8	64.4	39.2	56.4
BERT+GMM	42.5	65.6	65.4	64.0	50.8	58.3
BERT+DBSCAN	**47.1**	71.0	**85.5**	**78.6**	**59.3**	**64.1**

- 多くのデータセットにおいてBERT+DBSCANが既存の領域表現よりも高い性能
- クラスタリング手法として、GMMよりもDBSCANが高い性能
- micro平均において提案手法が既存手法よりも高い性能

単語間の意味的関係推定

- データセット
  BLESS、ROOT09、EVALution
- 評価指標
  F値
- 比較手法
  DS[8]、DSh[8]
- 提案手法
  重心間の距離が最も近い領域の
  重心を使用



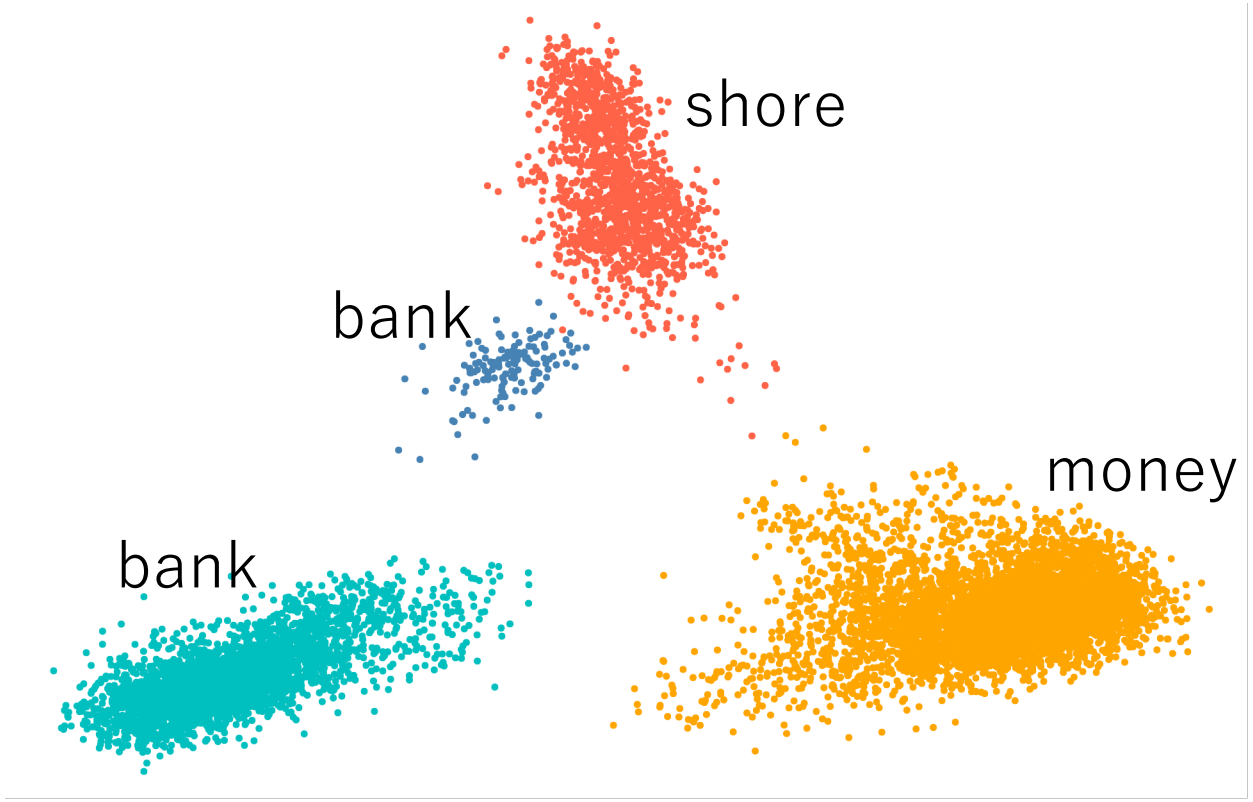
データ例（ROOT09）

word1 : tiger	word1 : apple	word1 : apple
word2 : animal	word2 : pear	word2 : animal
relation : HYPER	relation : COORD	relation : RANDOM

[8] Shwartz and Dagan. Path-based vs. Distributional Information in Recognizing Lexical Semantic Relations. (COLING 2016)

単語間の意味的関係推定

- データセット
  BLESS、ROOT09、EVALution
- 評価指標
  F値
- 比較手法
  DS[8]、DSh[8]
- 提案手法
  重心間の距離が最も近い領域の
  重心を使用

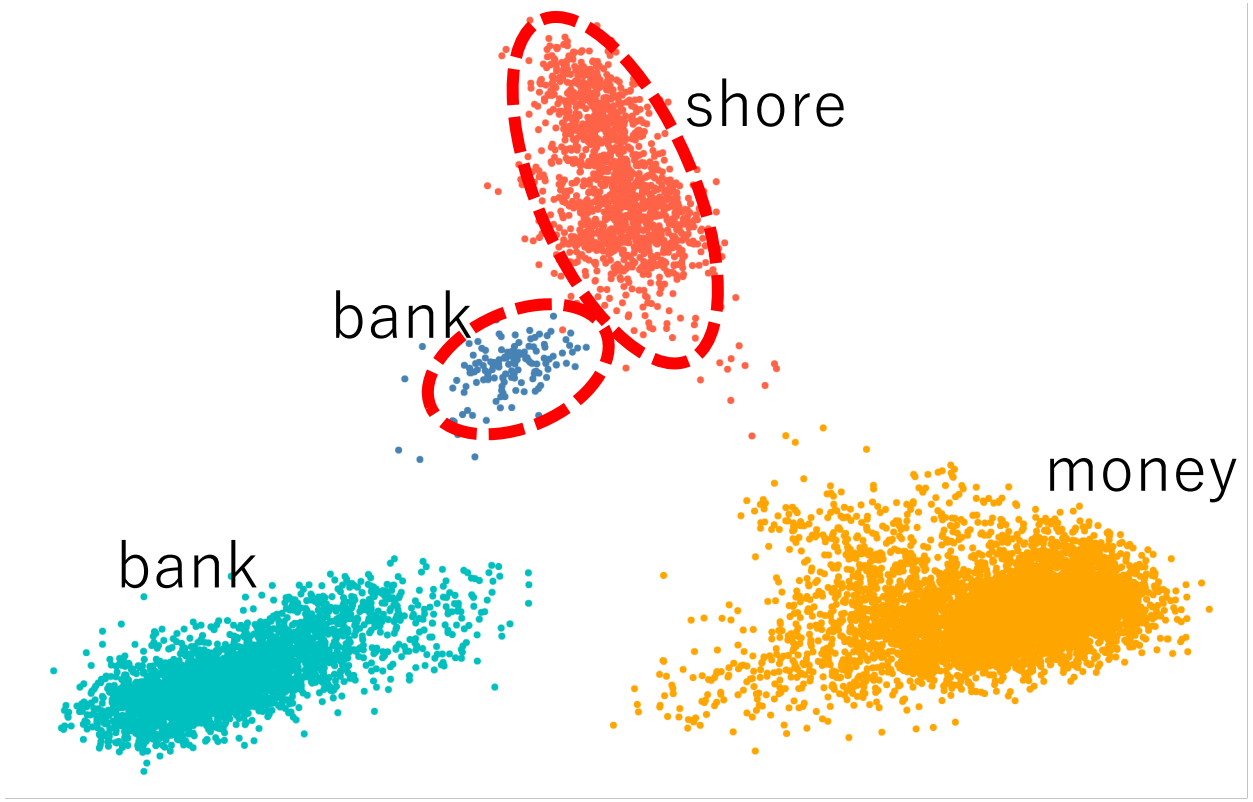


データ例（ROOT09）

word1 : tiger	word1 : apple	word1 : apple
word2 : animal	word2 : pear	word2 : animal
relation : HYPER	relation : COORD	relation : RANDOM

単語間の意味的関係推定

- データセット
  BLESS、ROOT09、EVALution
- 評価指標
  F値
- 比較手法
  DS[8]、DSh[8]
- 提案手法
  重心間の距離が最も近い領域の
  重心を使用



データ例（ROOT09）

word1 : tiger	word1 : apple	word1 : apple
word2 : animal	word2 : pear	word2 : animal
relation : HYPER	relation : COORD	relation : RANDOM

単語間の意味的関係推定

単語間の意味的関係データセットにおけるスコア（F値）

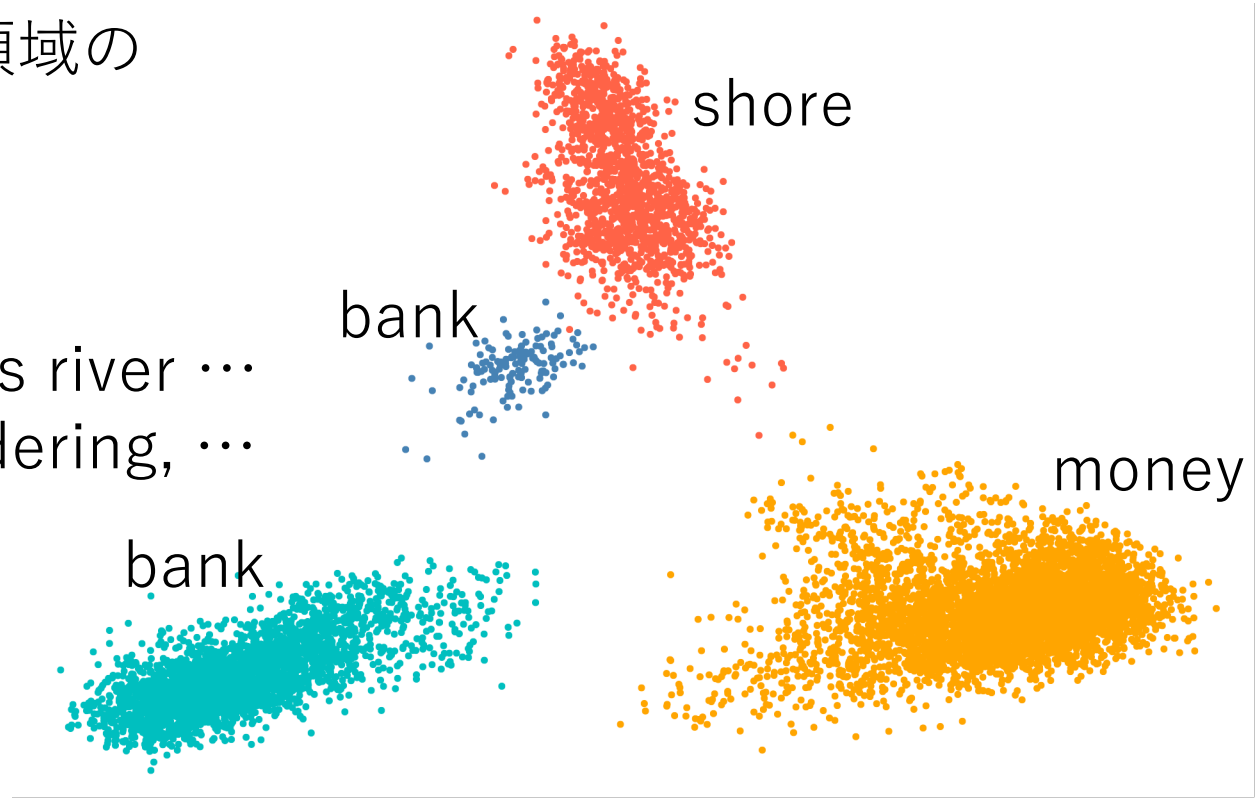
	BLESS	ROOT09	EVALution
DS[8]	0.811	0.646	0.525
DS_h [8]	**0.889**	0.716	0.571
ELMo+GMM	0.852	0.734	0.575
ELMo+DBSCAN	0.854	**0.743**	**0.588**
BERT+GMM	0.834	0.655	0.587
BERT+DBSCAN	0.822	0.671	0.568

- ROOT09とEVALutionにおいて既存手法よりも高い性能
- 既存手法と同等の性能で単語間の関係推定が可能

Stanford Contextual Word Similarity (SCWS)

- 評価指標
  人手で付与されたスコアとのスピアマンの順位相関係数
- 比較手法
  w2g [6]、w2gm [6]、ELMo、BERT
- 提案手法の推定スコア
  文脈から得られる分散表現が属する領域の
  重心間のコサイン類似度

データ例 (SCWS)
文脈１：… the east **bank** of Des Moines river …
文脈２：… the basis of all **money** laundering, …
スコア：2.5

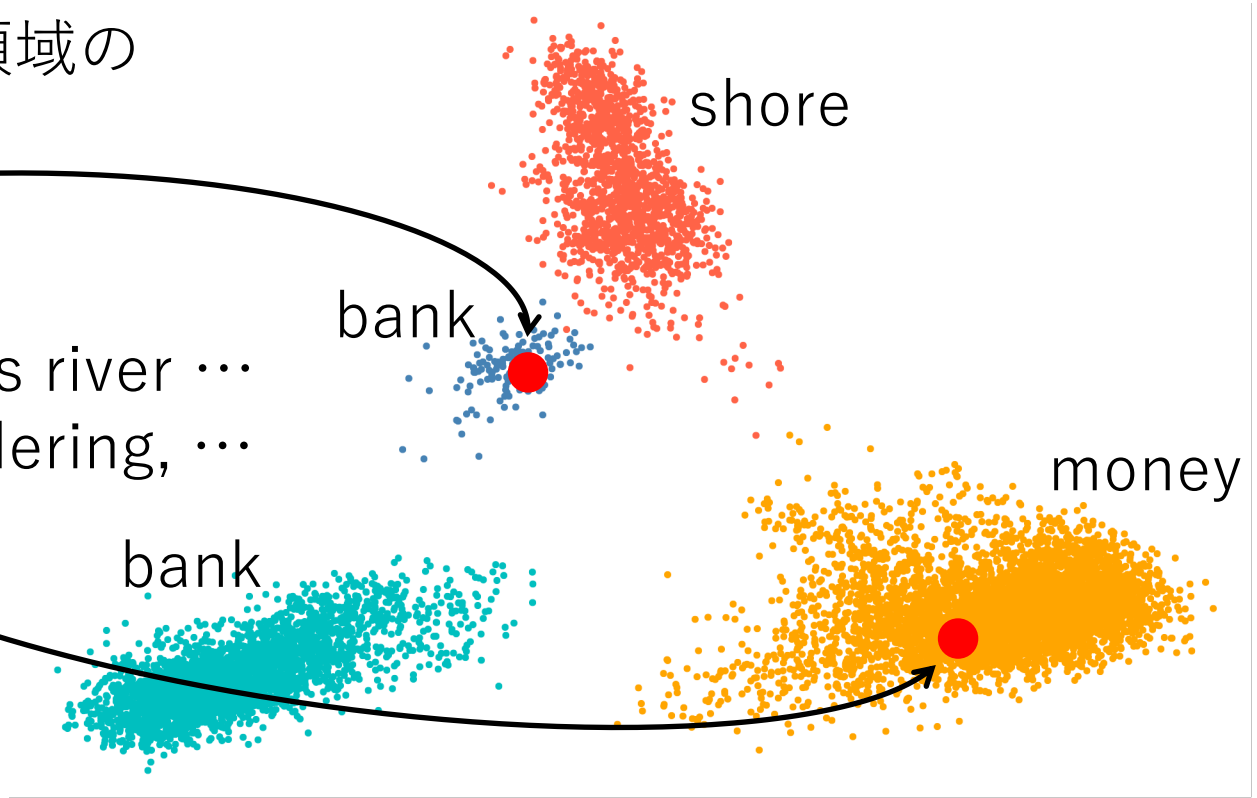


[6] Athiwaratkun and Wilson. Multimodal Word Distributions. (ACL 2017)

Stanford Contextual Word Similarity (SCWS)

- 評価指標
  人手で付与されたスコアとのスピアマンの順位相関係数
- 比較手法
  w2g [6]、w2gm [6]、ELMo、BERT
- 提案手法の推定スコア
  文脈から得られる分散表現が属する領域の
  重心間のコサイン類似度

データ例 (SCWS)
文脈１：… the east **bank** of Des Moines river …
文脈２：… the basis of all **money** laundering, …
スコア：2.5

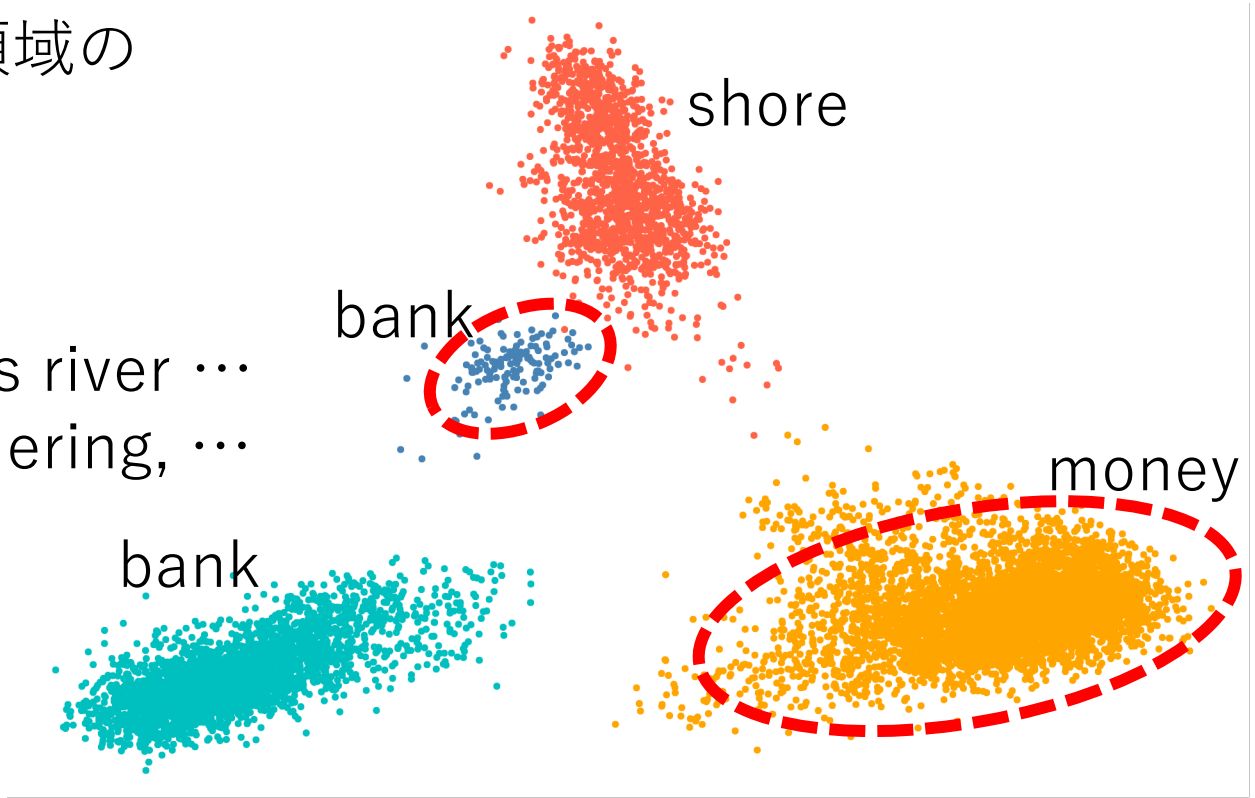


[6] Athiwaratkun and Wilson. Multimodal Word Distributions. (ACL 2017)

Stanford Contextual Word Similarity (SCWS)

- 評価指標
  人手で付与されたスコアとのスピアマンの順位相関係数
- 比較手法
  w2g [6]、w2gm [6]、ELMo、BERT
- 提案手法の推定スコア
  文脈から得られる分散表現が属する領域の
  重心間のコサイン類似度

データ例 (SCWS)
文脈１：… the east **bank** of Des Moines river …
文脈２：… the basis of all **money** laundering, …
スコア：2.5



[6] Athiwaratkun and Wilson. Multimodal Word Distributions. (ACL 2017)

SCWS

SCWSデータセットにおける
スピアマンの順位相関係数

Model	$\rho \times 100$
w2g [6]	66.2
w2gm [6]	65.5
ELMo	67.6
ELMo+GMM	66.0
ELMo+DBSCAN	**68.0**
BERT	61.7
BERT+GMM	64.5
BERT+DBSCAN	65.0

- ELMo+DBSCANが既存の領域表現よりも高い性能
  →文脈を考慮した単語分散表現から領域を獲得することが有効
- ELMoやBERTで得られる分散表現をそのまま用いるよりも高い性能

SCWS

SCWSデータセットにおける
スピアマンの順位相関係数

Model	$\rho \times 100$
w2g [6]	66.2
w2gm [6]	65.5
ELMo	67.6
ELMo+GMM	66.0
ELMo+DBSCAN	**68.0**
BERT	61.7
BERT+GMM	64.5
BERT+DBSCAN	65.0

- ELMo+DBSCANが既存の領域表現よりも高い性能
  →文脈を考慮した単語分散表現から領域を獲得することが有効
- ELMoやBERTで得られる分散表現をそのまま用いるよりも高い性能

まとめ

目的

多義性や上位下位関係が表現可能な単語表現の獲得

提案手法

文脈を考慮した単語ベクトル集合から単語領域表現を獲得

結果

文脈を考慮しないタスクでは既存手法と同等以上の性能

文脈を考慮するタスクでは既存手法よりも高い性能

今後の課題

クラスタそのものの性質を利用した距離尺度