単語埋め込みの決定的縮約

仲村 祐希[1]　鈴木 潤[1,2]　高橋 諒[1,2]　乾 健太郎[1,2]

東北大学[1]

理化学研究所[2]

背景：単語埋め込みは自然言語処理で必須要素

- 自然言語処理では単語埋め込みが必要不可欠
  - 離散的な記号である単語と連続的な数値の橋渡し

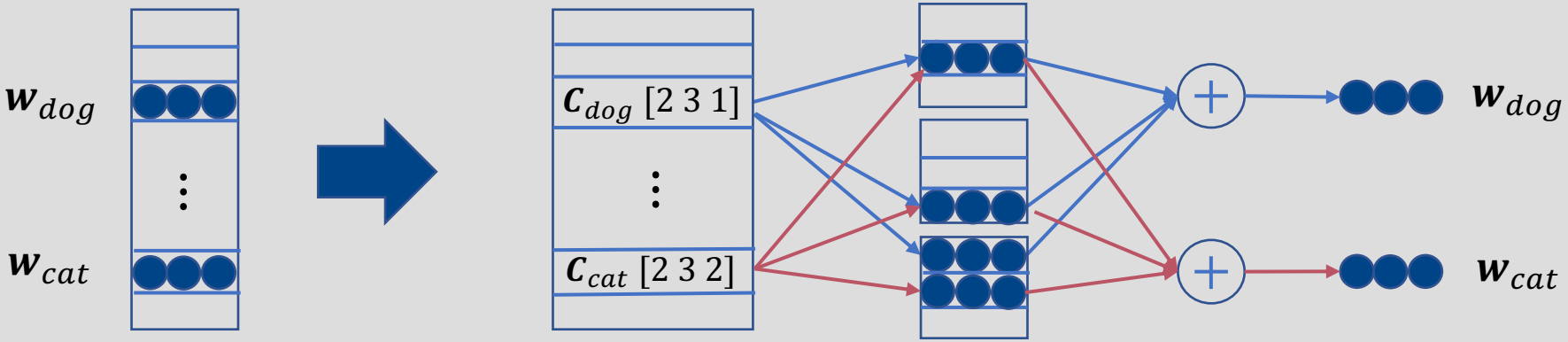
- 単語埋め込みのパラメータ数は膨大
  - 単語埋め込みは語彙数 × 次元数の行列
  - GloVeの例：語彙数400000 × 次元数300 = 1億2000万パラメータ

既存研究：代表的な単語埋め込みの圧縮手法

単語埋め込みを基底番号のリスト(離散符号)と基底ベクトル(コードブック)で表現

- 似た単語は似た単語埋め込みをしている

➤似た単語同士で部分的に同じパラメータを共有することで圧縮



既存研究：深層学習による圧縮手法は乱数のシード値によって異なる

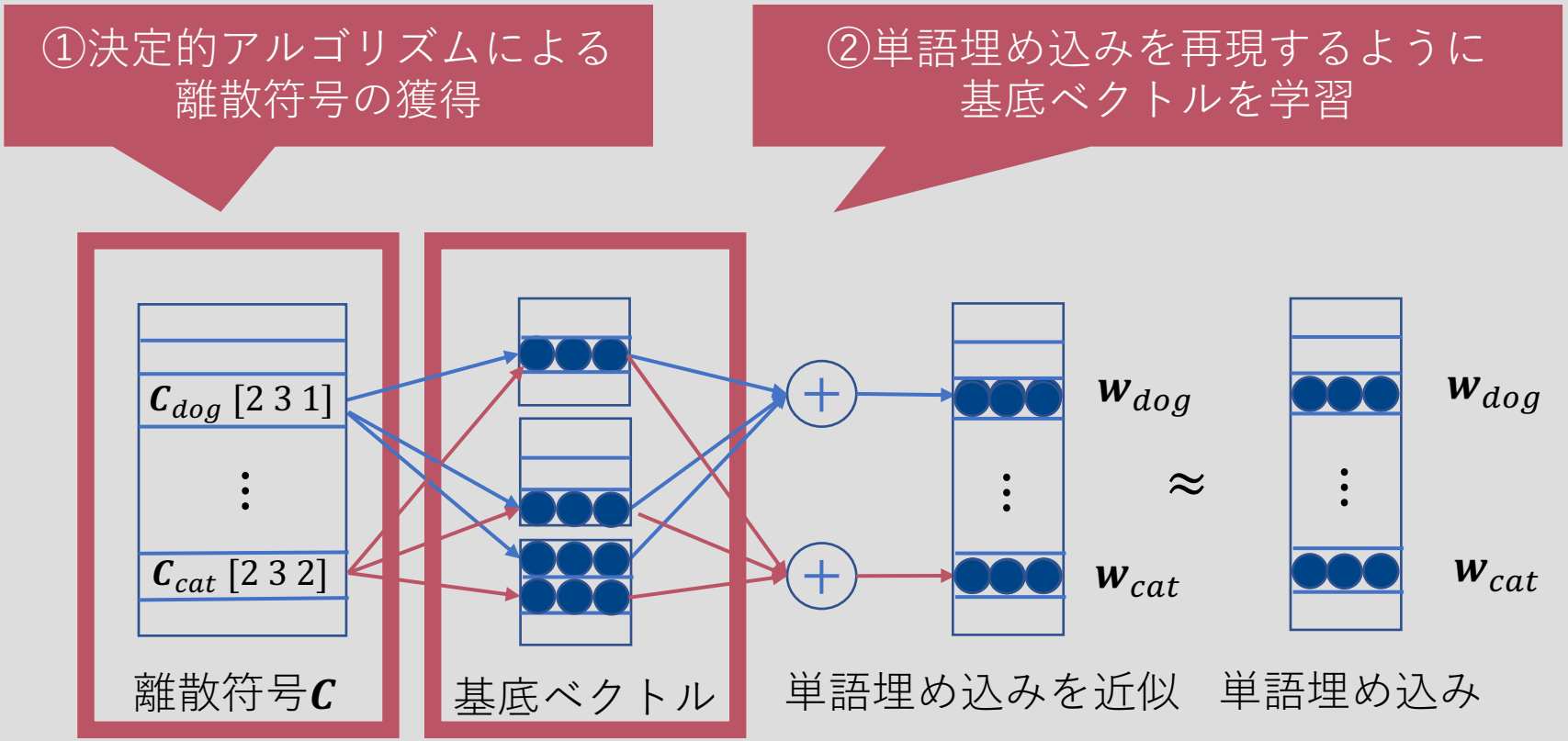
- 深層ニューラルネットワーク(DNN)による手法
  - Compressing Word Embeddings via Deep Compositional Code Learning [Shu+,ICLR'18]
- DNNによる手法は**ランダム性があり**，乱数のシードによって離散符号が**異なる**
  - ➤ **ランダム性がない離散符号の獲得手法を考案**

実行結果が**決定的**であることの**メリット**

- **再現性**や**信頼性**の観点で優れる
- 実応用上，コストの低下につながる

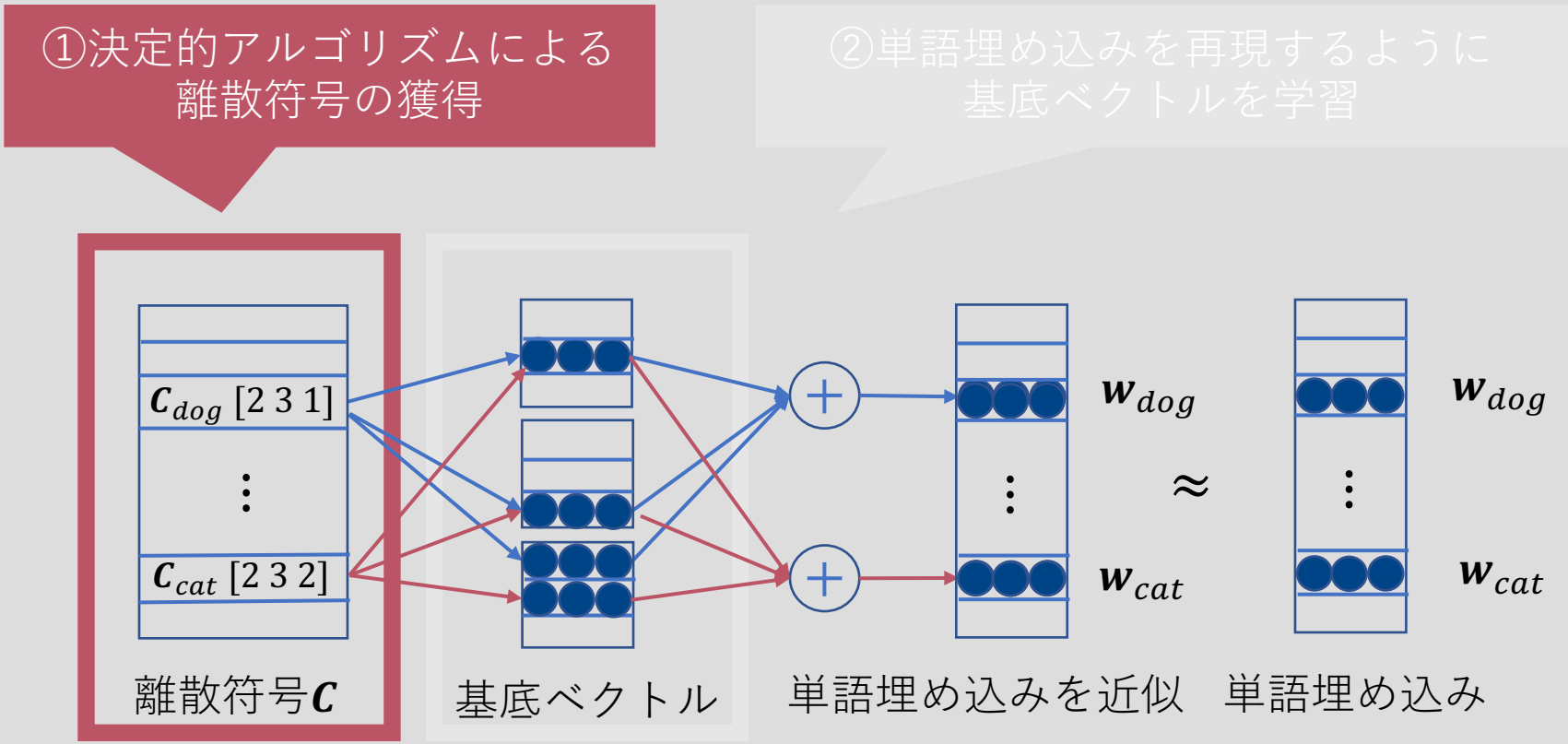
提案手法の圧縮方法の概要

- 提案手法は「離散符号の獲得」と「基底ベクトルの学習」から構成される



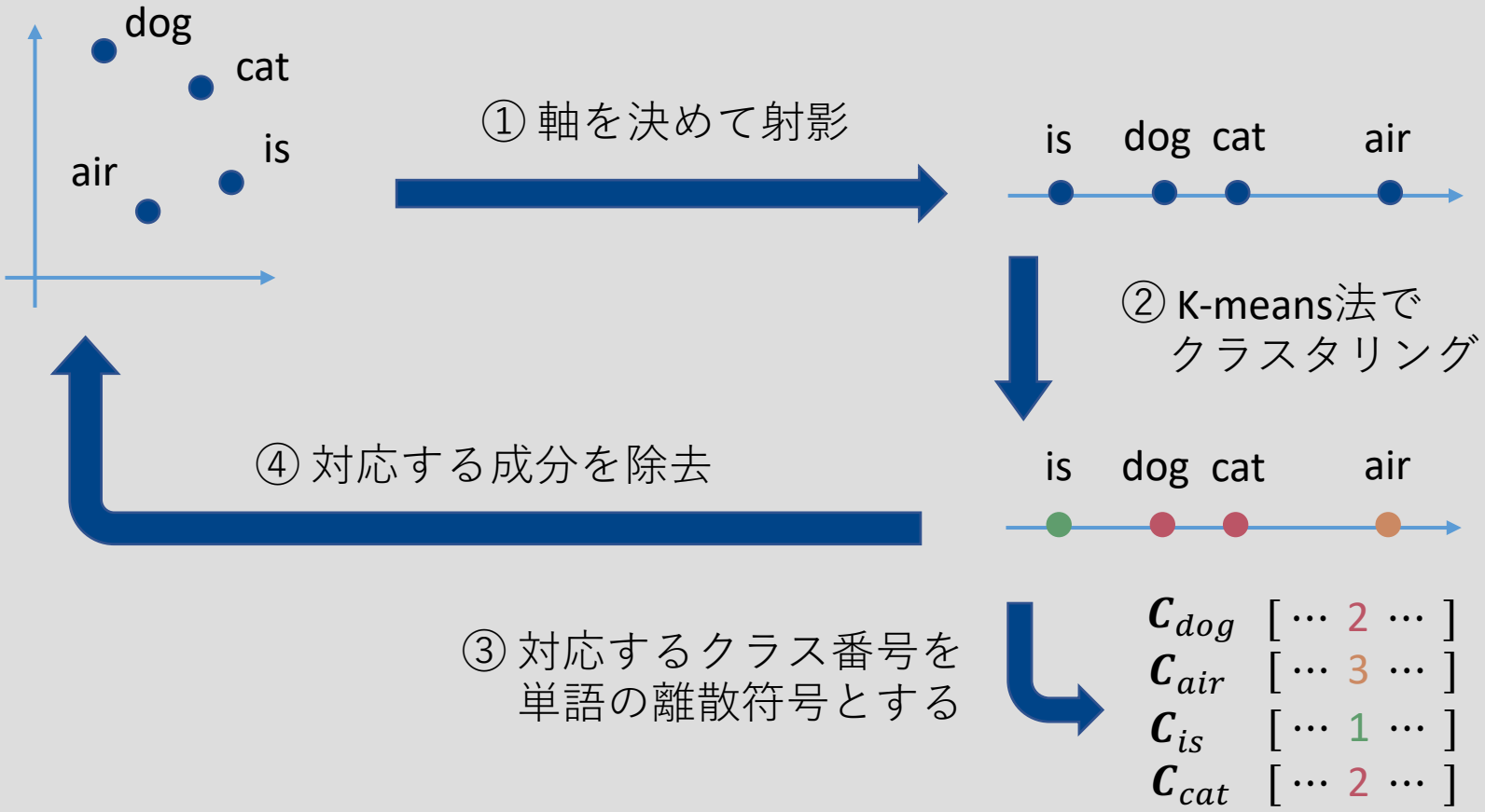
提案手法の圧縮方法の概要

- 提案手法は「離散符号の獲得」と「基底ベクトルの学習」から構成される



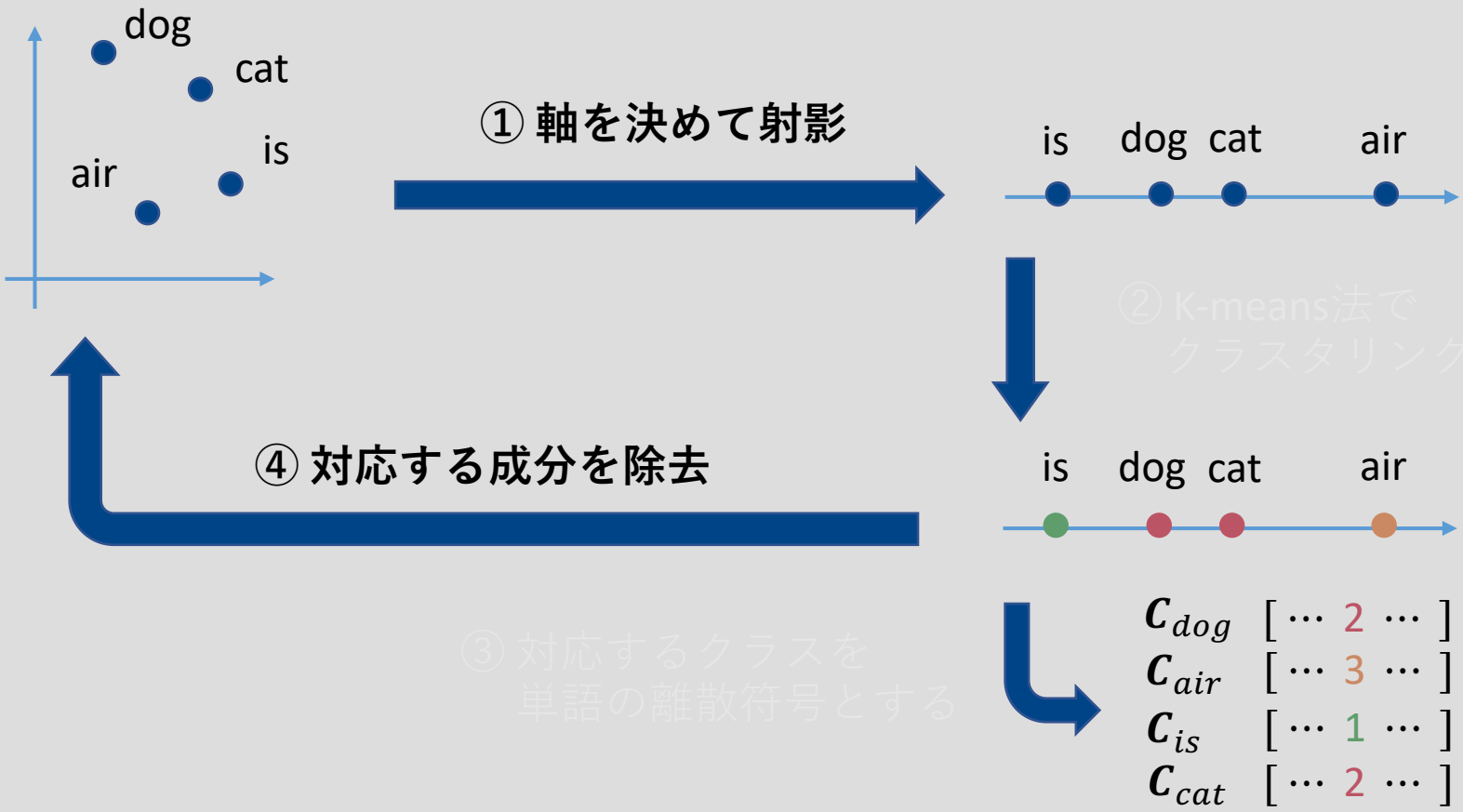
提案手法：決定的アルゴリズムによる離散符号の獲得手法の概要

- 1次元のK-means法は最適解が多項式時間で求まり**決定的**



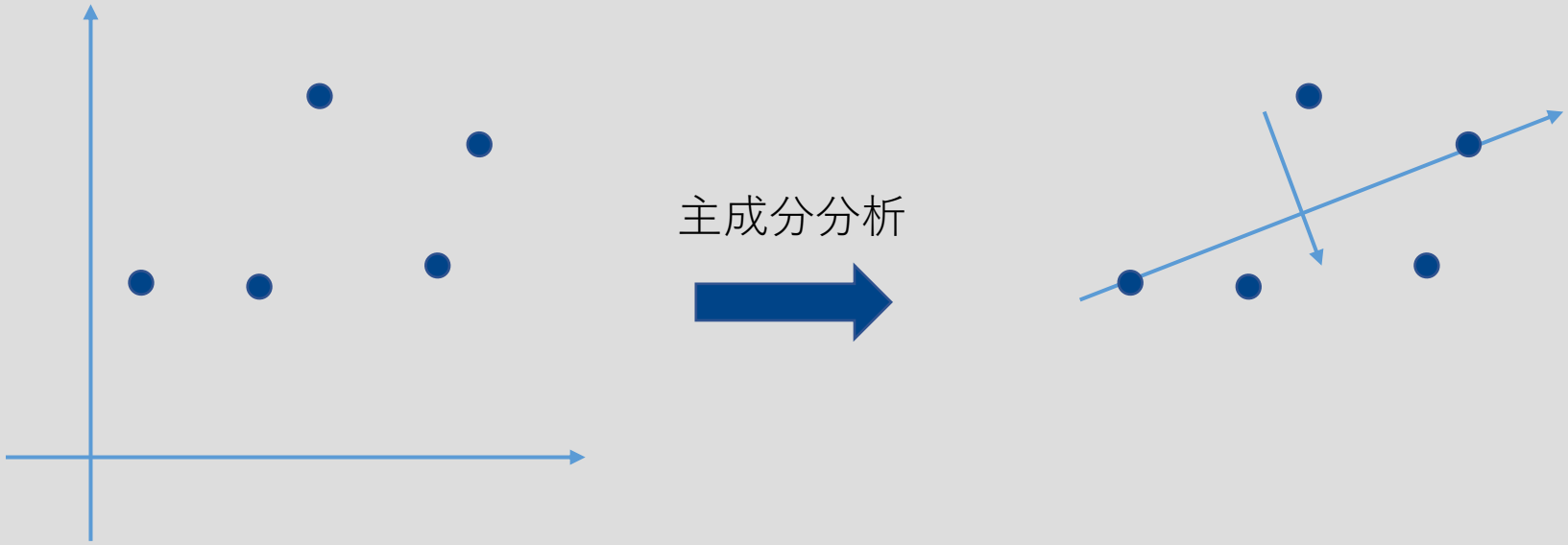
提案手法：決定的アルゴリズムによる離散符号の獲得手法の概要

- 1次元のK-means法は最適解が多項式時間で求まり決定的



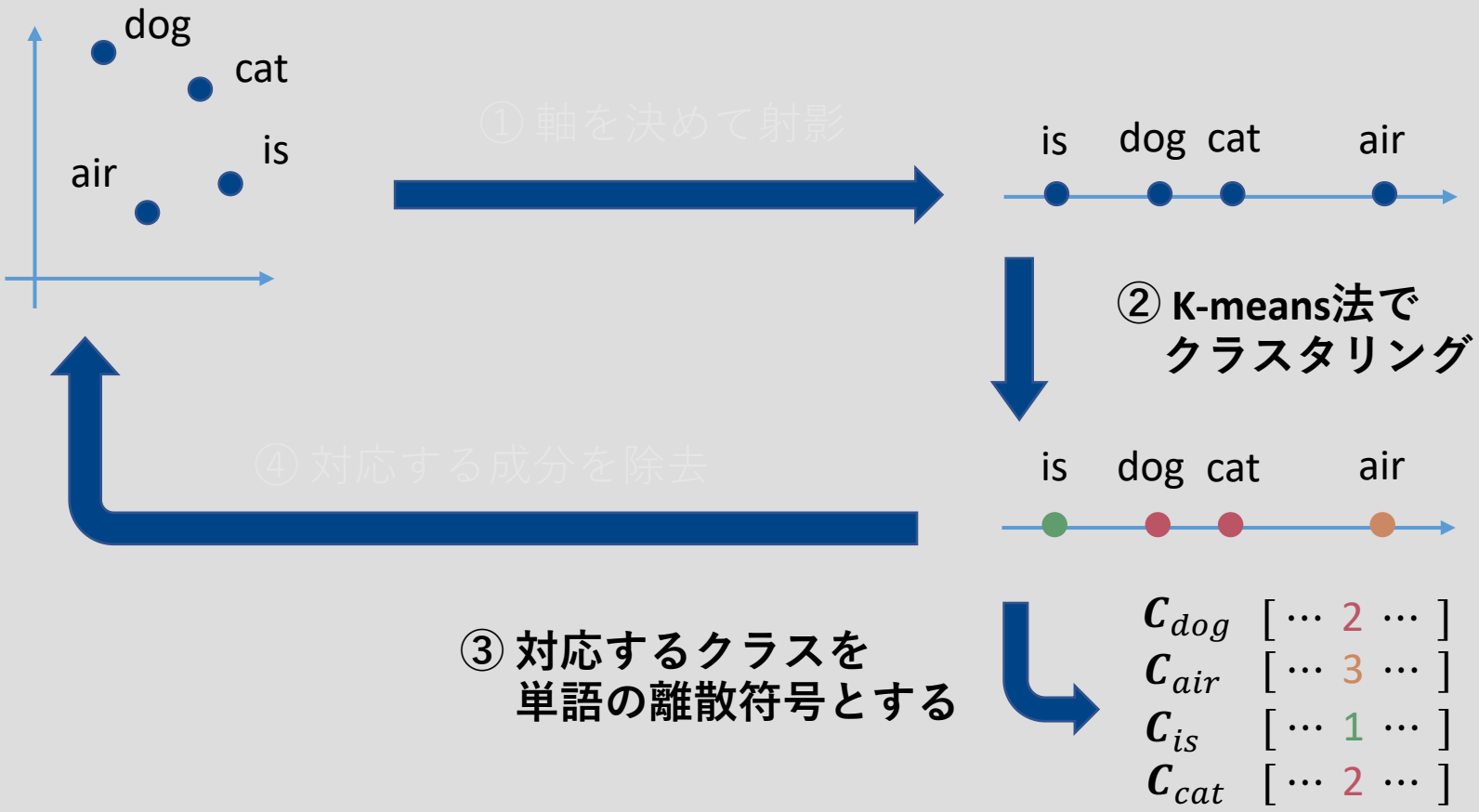
提案手法：主成分分析を用いて1次元の軸へ射影

- 単語埋め込みを1次元の軸上に射影するために**主成分分析**を使用
  - 主成分分析は**決定的**
  - 第M主成分までを用いることで，M個の離散符号を獲得



提案手法：決定的アルゴリズムによる離散符号の獲得手法の概要

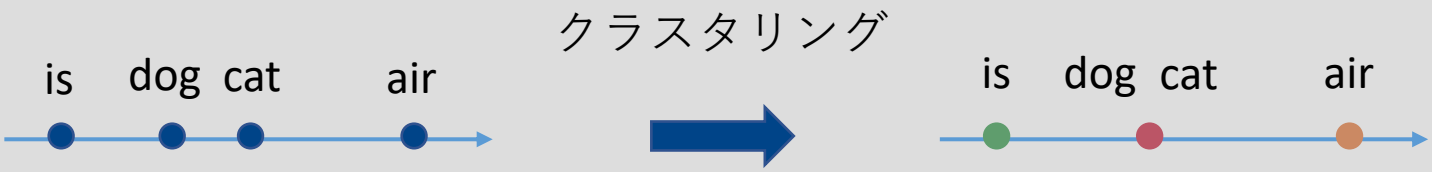
- 1次元のK-means法は最適解が多項式時間で求まり決定的



提案手法：1次元K-means法は最適解が多項式時間で計算可能

- 1次元K-means法は最適解を$O(V \log V + VK)$で計算可能
  - V: 語彙数　K: クラス数

- 似た単語には似たクラス番号が割り振られることが期待
  - クラスタリングしたクラス番号(離散符号)は順序関係を保持

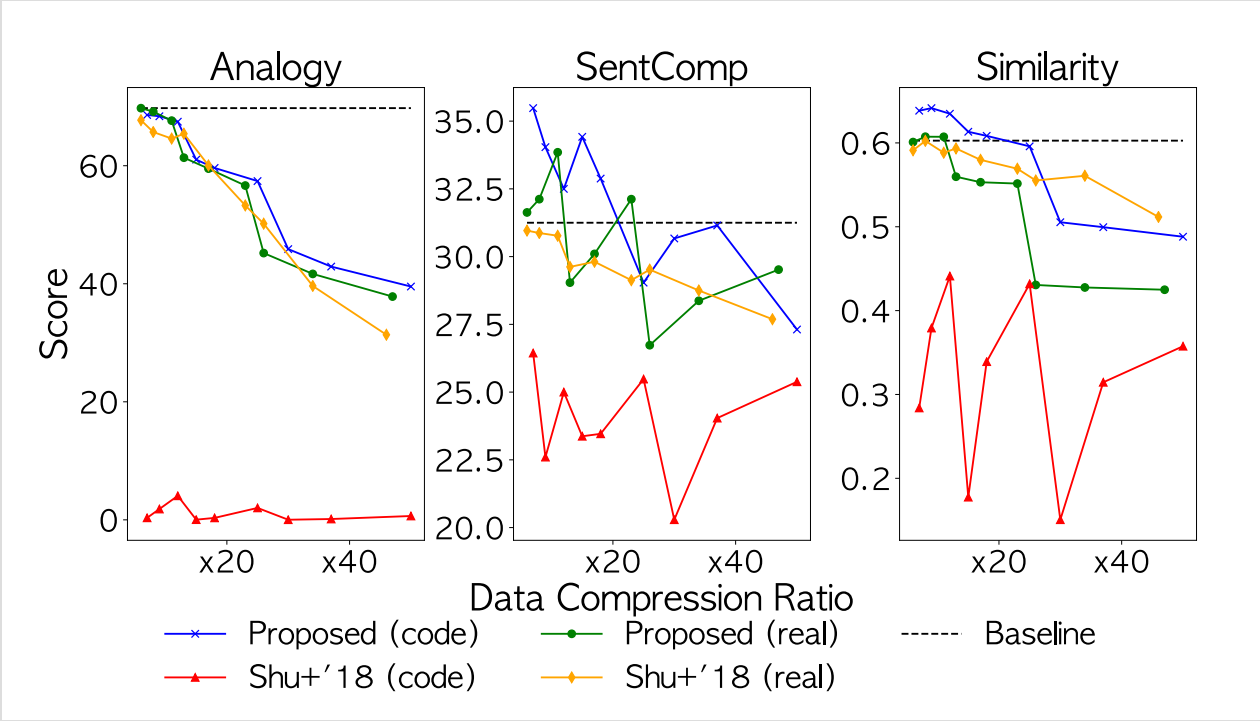


実験設定：単語埋め込みの内的評価タスク

提案手法が単語埋め込みの情報を保持しているか調べるために,
単語埋め込みを圧縮したときのスコアの変化を測定

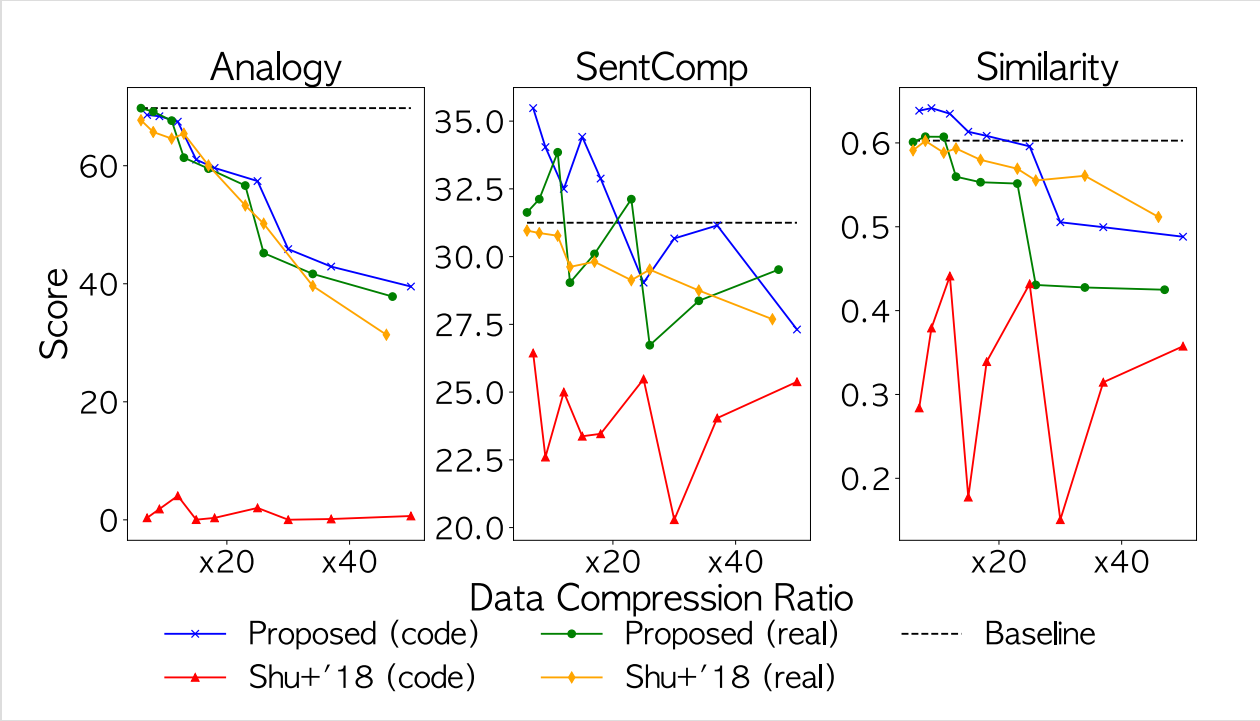
- 実験タスク：
  - 単語類推タスク(Analogy)
    - 評価指標：正答率 (acc)
  - 文穴埋めタスク(SentComp)
    - 評価指標：正答率 (acc)
  - 単語類似性判定タスク(Similarity)
    - 評価指標：スピアマンの順位相関係数 (ρ)
- 単語埋め込み：
  - GloVe.6B.300d (語彙数400000,　次元数300)

実験結果：単語埋め込みの内的評価タスクの実験結果



Code: 離散符号　Real: 離散符号に対応する基底ベクトルの和

実験結果：単語埋め込みの内的評価タスクの実験結果



Code: 離散符号　Real: 離散符号に対応する基底ベクトルの和

- 提案手法は離散符号は軸上の単語埋め込みの順序関係を保持
  - ➤ 離散符号の数値の計算が**意味がある**

実験結果：提案手法は乱数によらず再現性に優れている

DNNの方法で乱数のシード値を変えて10回測定

	最小値	最大値	差
Analogy (acc)	63.79	65.05	1.29
SentComp (acc)	29.81	31.63	1.82
Similarity (ρ)	0.58	0.60	0.2

✗乱数によってスコアが**大きく変化**

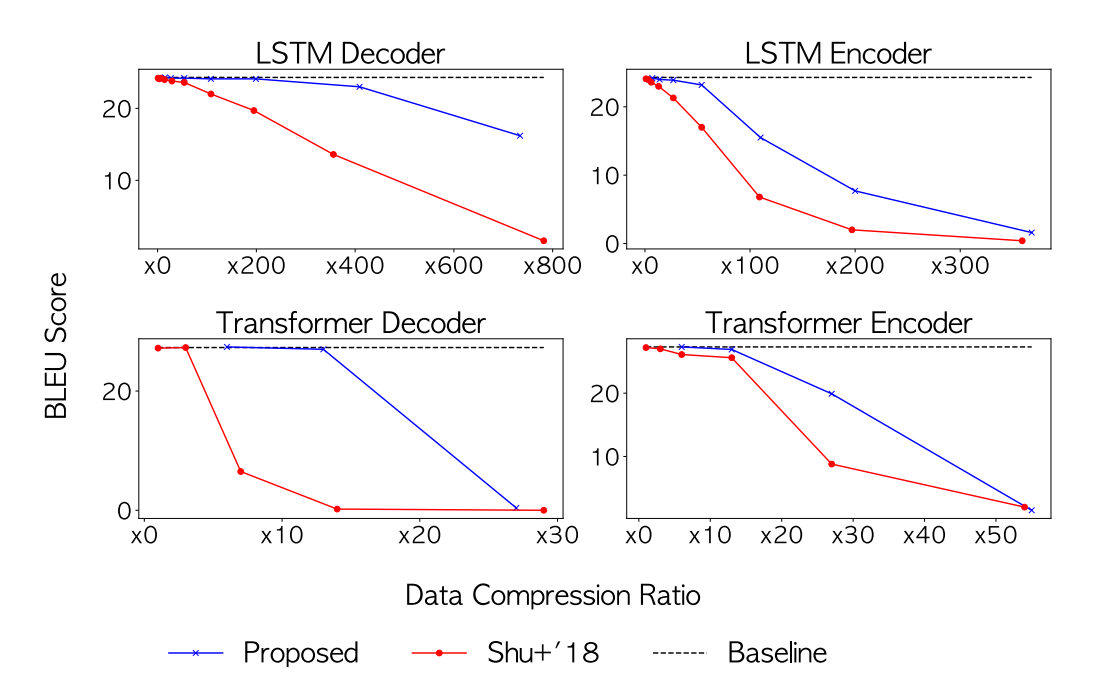
✓一方，提案手法の離散符号の結果は乱数によらないため**再現性**に**優れる**

実験設定：機械翻訳タスク

深層学習モデルの単語埋め込み層を圧縮したときのBLEUと圧縮比を測定

- 実験タスク：
  - WMT 2016 英独翻訳タスク
- モデル：
  - 双方向LSTM，Transformer
- 単語埋め込み：
  - 学習後の埋め込み層

実験結果：提案手法の方が圧縮比に対するBLEUの値が高い



- ✓ 提案手法の方が圧縮比に対するBLEUが**高い**
- ✓ LSTMとTransformerではLSTMの方が圧縮比に対するスコアの低下が小さい

まとめ

- 単語埋め込みから乱数によらない**再現性に優れた**離散符号の獲得手法を考案

- 提案手法により獲得した離散符号は単語埋め込みの情報を十分に保持していることを実験で確認

Appendix

Appendix：手法によるサイズの違い

- 既存手法 [Shu+,ICLR'18]

$$\underbrace{VM\log_2 K}_{\text{離散符号}} + \underbrace{4MKH}_{\text{コードブック}}\ [\text{Byte}]$$

M:離散符号の数
K:離散符号の種類数
H:単語埋め込みの次元数
V:語彙数

- 提案手法

$$\underbrace{VM\log_2 K}_{\text{離散符号}} + \underbrace{4MH + 4MK}_{\text{コードブック}}\ [\text{Byte}]$$