分散表現を利用したタグ集合の階層化

尾崎研究室　鈴木宏明

目次

- 研究背景と目的
- 関連研究
- 提案手法
- 評価実験
- 実験結果
- まとめ、今後の課題

研究背景と目的

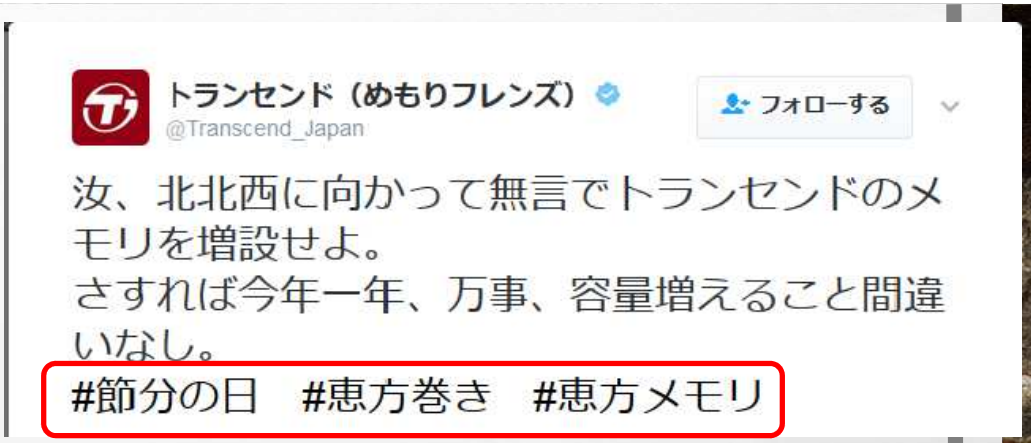
SNSなどに投稿されたもの

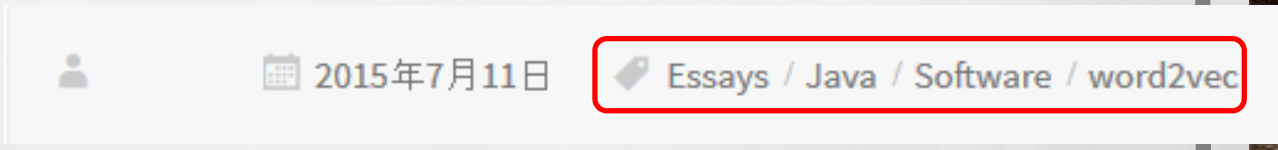
→タグによって特徴付け

しかし…

コンテンツ同士の関係性

→明示的でない





研究背景と目的

コンテンツ同士に関係が有るのか無いのかを明示的に

コンテンツに付与されたタグの集合を構造化する

関連研究：タグの階層化

- 『動画投稿サイトで付与された動画タグの階層化』：村上直至,伊藤栄典

  →出現頻度と共起確率で決定するISR手法

- 『ニコニコ動画における共起関係を用いたタグの階層化』：

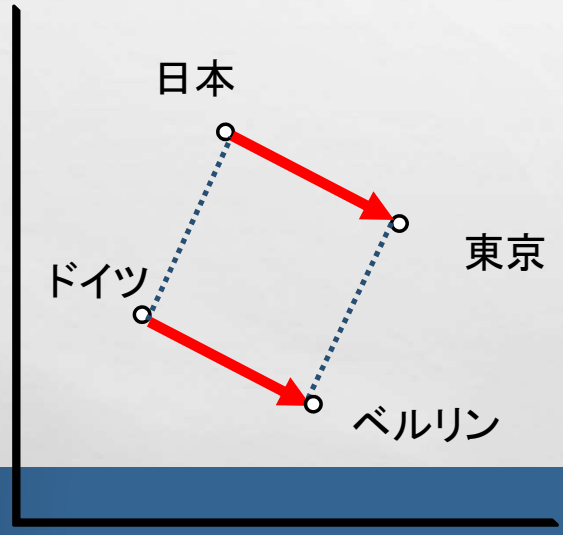
  高橋文彦,山本雅人,古川正志

  →共起関係を用いた手法

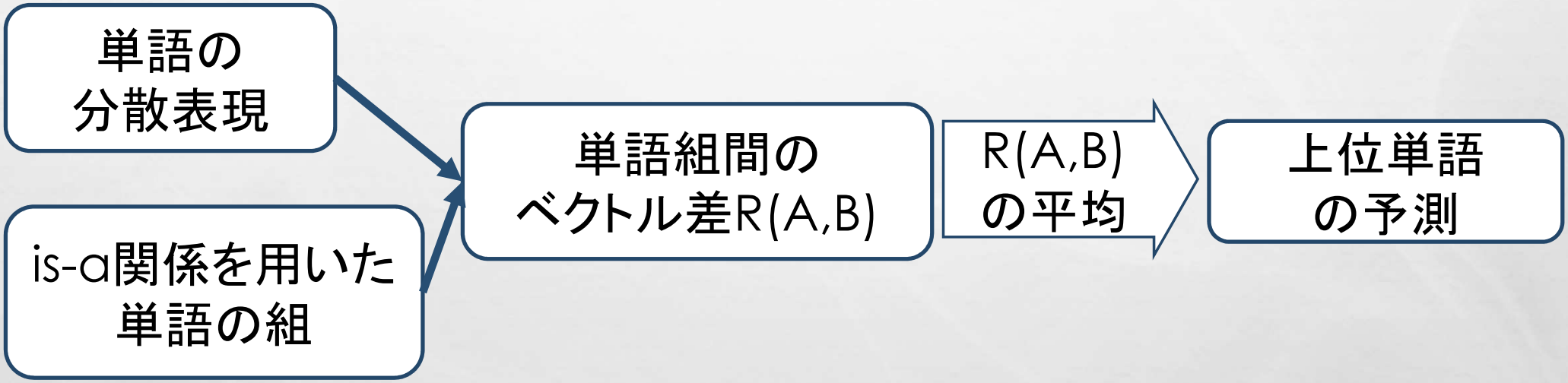
関連研究：分散表現

- 『分散表現とオントロジーの関係』：市瀬龍太郎,荒川直哉

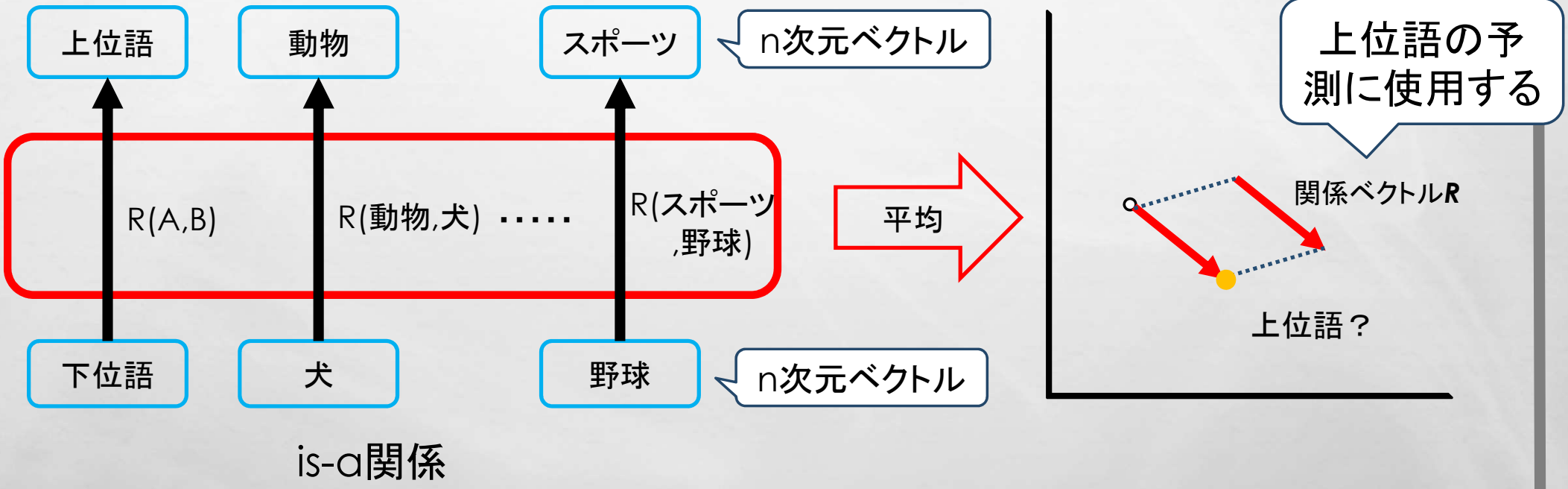
  →語のis-a関係と単語ベクトルの加法構成性を用いたオントロジーの構築



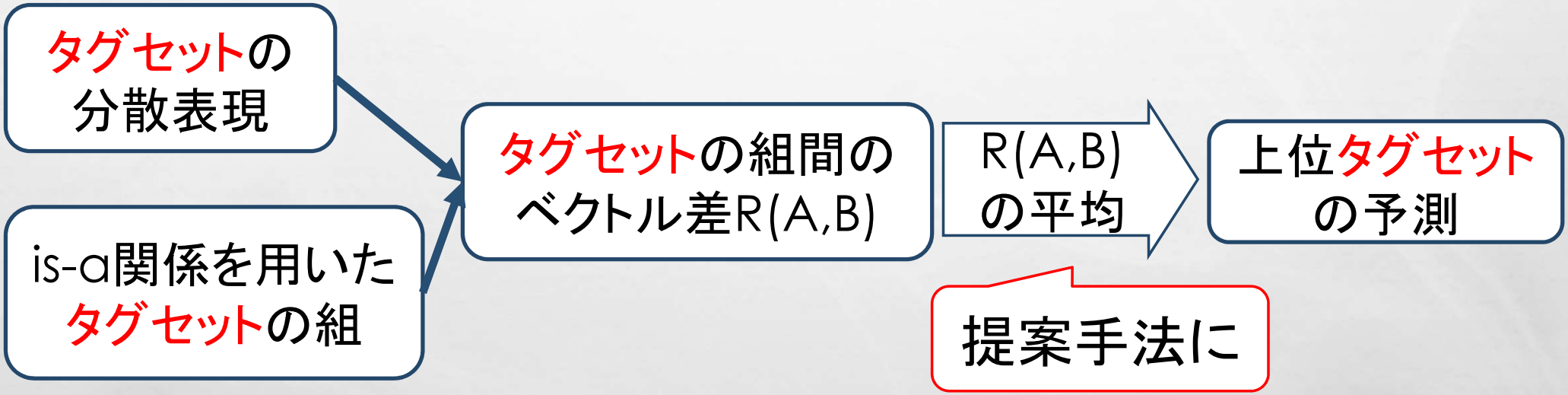
関連研究：分散表現



関連研究：分散表現


上位語
動物
スポーツ
n次元ベクトル
R(A,B)
R(動物,犬) ・・・・・
R(スポーツ,野球)
平均
上位語の予測に使用する
関係ベクトルR
上位語？
下位語
犬
野球
n次元ベクトル
is-a関係

手法の概要


タグセットの
分散表現
is-a関係を用いた
タグセットの組
タグセットの組間の
ベクトル差R(A,B)
R(A,B)
の平均
上位タグセット
の予測
提案手法に

提案手法1：クラスタリング

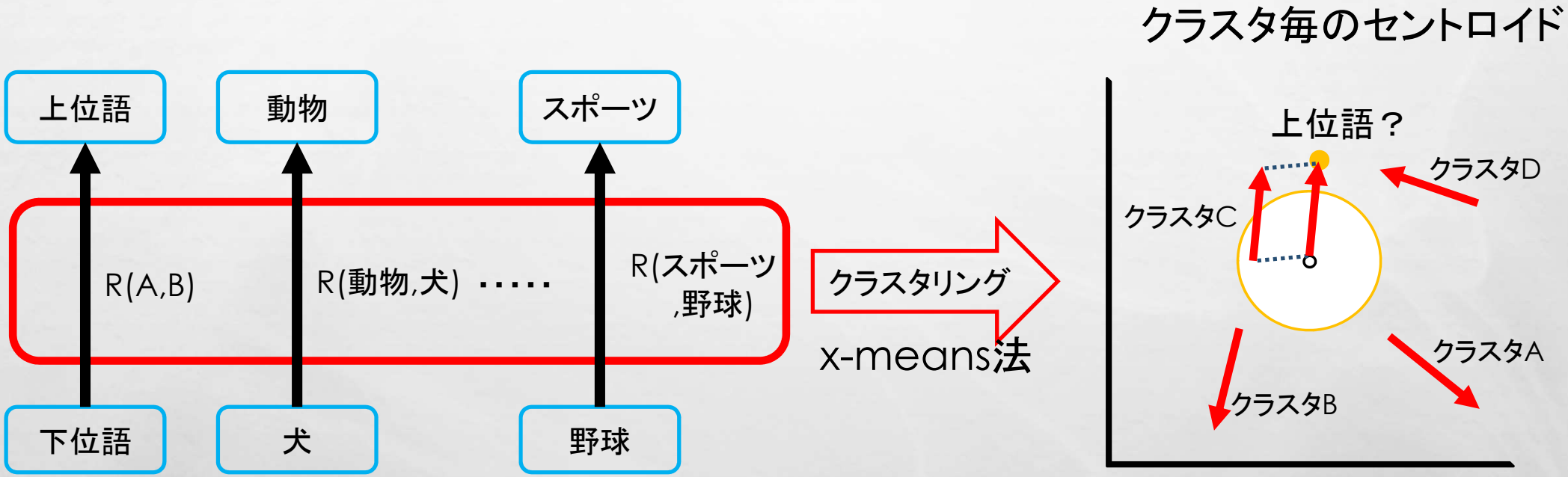
- ベクトル差R(A,B)

  →様々な向き

  →正反対の向きのベクトルの平均は0

  →同じ様な向き（=似たような関係）のベクトルをまとめる

提案手法1:クラスタリング



提案手法2：近傍のR(A,B)の平均

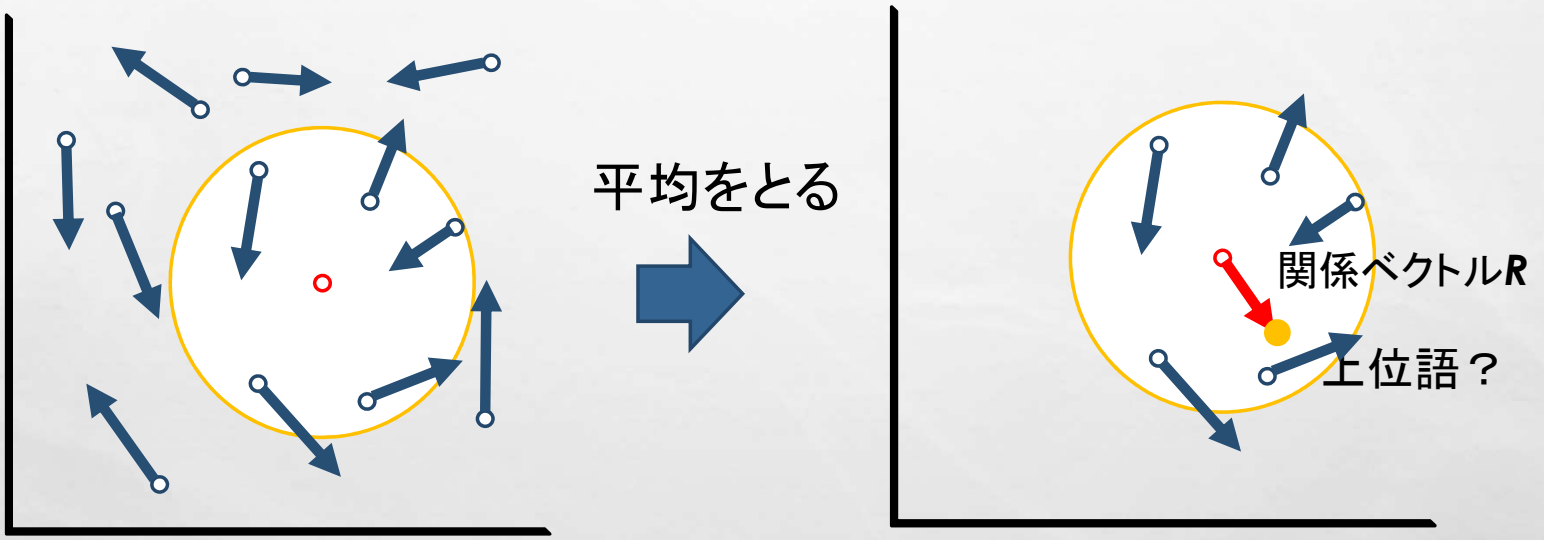
- 分散表現

  →意味的に近い語は近いベクトルを持つ

  →関係ベクトルも似たようなものになるはず

  →近くの語のベクトル差を用いる

提案手法2：近傍のR(A,B)の平均

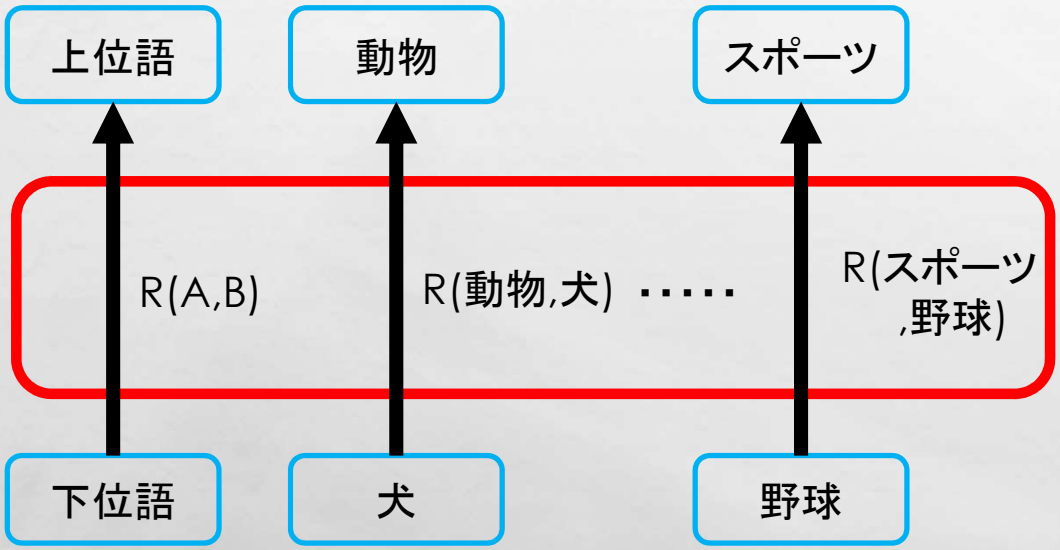


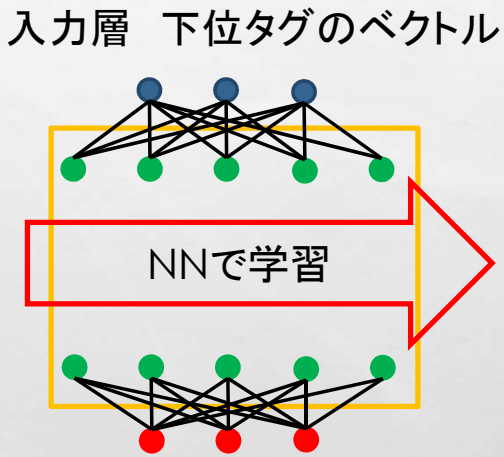
k個の近傍の
下位タグを取り出す

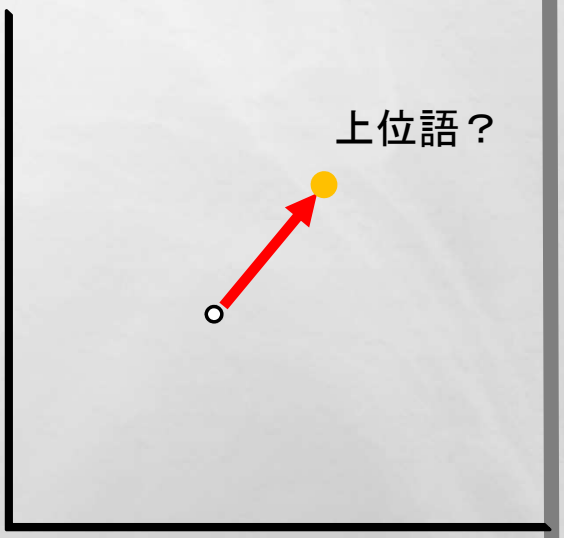
提案手法3: R(A,B)をNNで学習

- 下位語を入れたら上位語が出てくれば良い

→教師付き学習







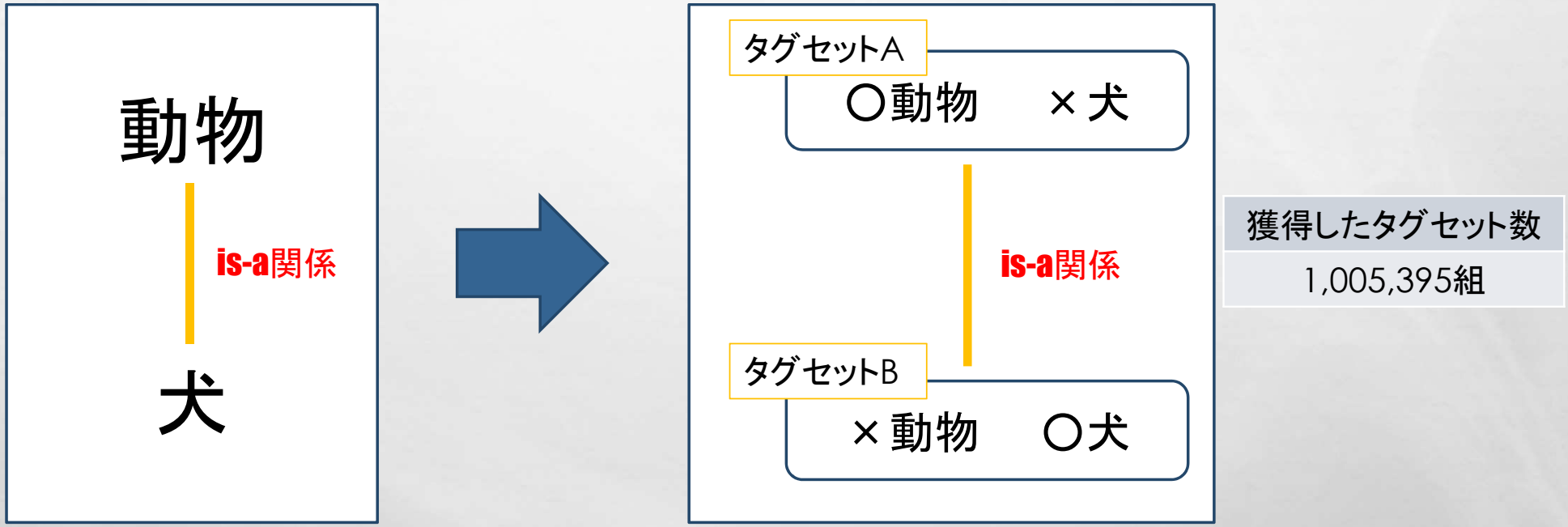
タグセットのペア

- 正解データが存在しない

  →人手で作るのは難しい

  →単語のis-a関係をタグセットに拡張する

タグセットのペア

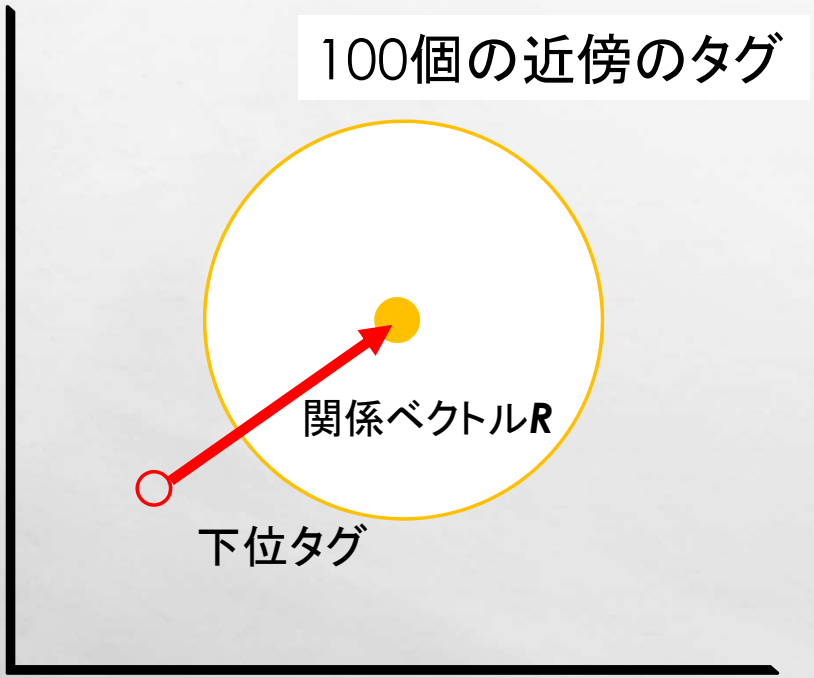


獲得したタグセット数
1,005,395組

評価実験1

- タグのis-a関係とベクトルを用いる
  - 3つの提案手法と既存手法の結果を比較する
  - タグのベクトル：Word2Vec,300次元
  - 正しい上位タグを予測できるかを評価する

評価方法



- 正しい上位タグが含まれていれば正解
- 約20,000タグから100個

実験結果

NN以外は10fold交差検証を行った平均値
NNは学習7:テスト3にデータを分けて学習

手法	データ数	近傍	クラスタ数	正解数	正解率	平均順位
既存	4980			197.47	4.08	36.89
クラスタリング	4980	1	17.4	534.60	10.99	31.40
近傍	4980	1		206.89	4.23	18.95
近傍	4980	5		300.83	6.16	24.75
近傍	4980	10		263.17	5.38	29.47
NN	29044			102	0.35	

評価実験2

- タグセットのペアとベクトルを用いる
  - 3つの提案手法の結果を比較する
  - タグセットのベクトル：Doc2Vec,300次元
  - 正しい上位タグセットを予測できるかを評価する
  - 評価方法は実験1と同じ

実験結果

NN以外は10fold交差検証を行った平均値
データ量が多いので1/10ずつ実験

上位100個/約320,000動画

手法	データ数	近傍	クラスタ数	正解数	正解率	平均順位
クラスタリング	10053	1	20	102.3	0.96	50.86
近傍	10053	1		139.8	1.36	48.01
近傍	10053	5		132.1	1.27	48.49
近傍	10053	10		121.2	1.16	48.23
NN	60324			106	0.18	

まとめ

- タグに対する実験

  →既存手法より精度が向上。

  →絶対的な精度が低い。獲得したis-a関係に問題？

- タグセットに対する実験

  →十分な精度は得られなかった。近傍がわずかに良い。

  →タグセットのペアの作り方に問題？

- NNを用いた学習

  →タグ、タグセットともに十分な精度は得られなかった。

今後の課題

- is-a関係

  →抽出方法の改善、日本語Wikipediaオントロジーを用いる

- タグセットのペア

  →異なるペアの作成方法の考案

- NNの学習方法の改善
- ベクトル化手法の検討