



文脈語間の対訳関係を用いた単語ベクトルの翻訳

石渡祥之佑[†], 鍛冶伸裕[♣], 吉永直樹^{◇♡}, 豊田正史[◇], 喜連川優^{◇♠}

{ishiwatari, kaji, ynaga, toyoda, kitsure} @ tkl.iis.u-tokyo.ac.jp

[†] 東京大学, [♣] ヤフー株式会社, [◇] 東京大学生産技術研究所, [♡] 情報通信研究機構, [♠] 国立情報学研究所

1. 概要

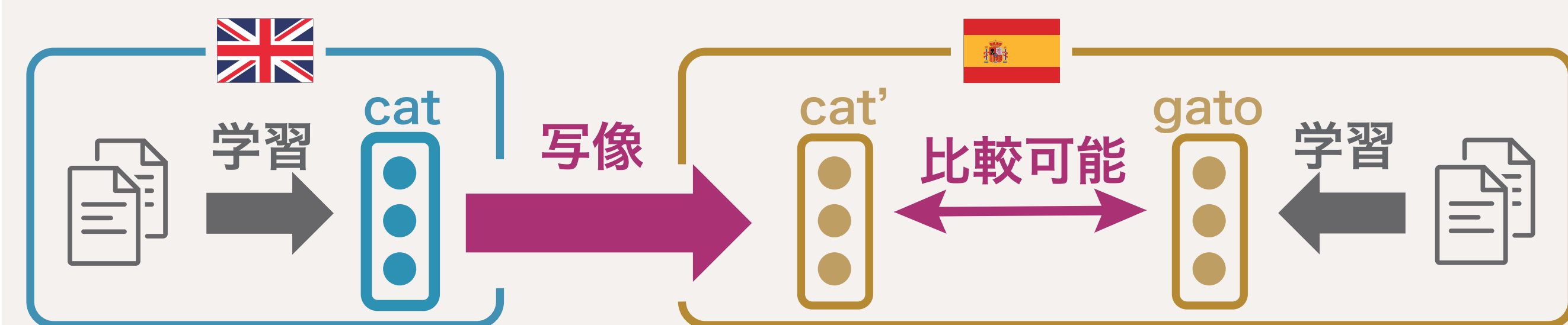
課題

異なる言語の単語ベクトルは言語によって異なるコーパスから学習されるため、比較不能



アプローチ

単語ベクトルを他の言語の空間に写像することにより比較可能にする [Fung 98][Mikolov+ 13]



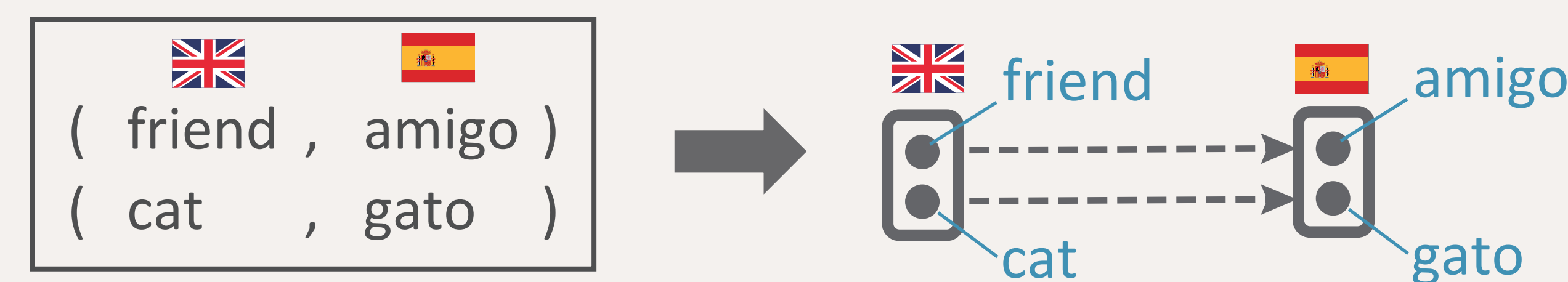
提案手法

2 種類の先行研究 [Fung+, 98][Mikolov+, 13] の手法を組み合わせ、ベクトル翻訳の精度を向上

2. 先行研究

辞書によるベクトル翻訳 [Fung+, 98]

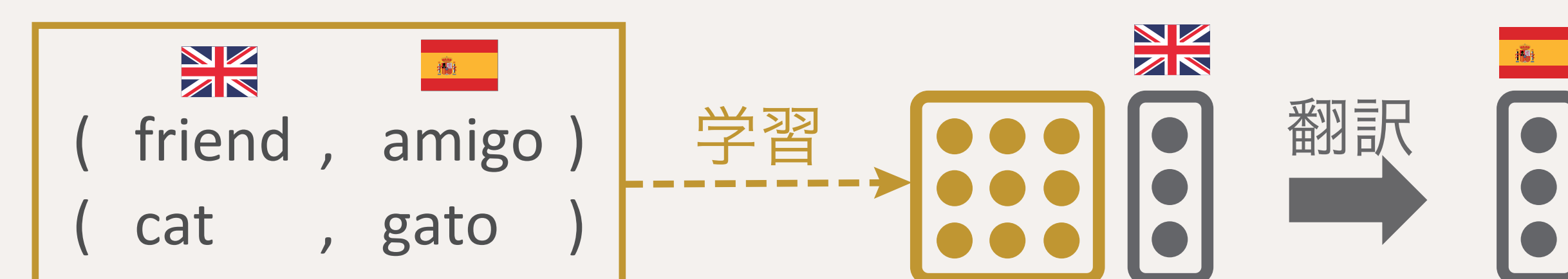
- ・ 共起頻度ベクトルを用いる
- ・ 次元ごとに辞書を用いて直接写像する



- ☺ 文脈語の対訳関係を直接利用できる
- ☹ 次元間は 1:1 にしか写像することができない

学習によるベクトル翻訳 [Mikolov+, 13]

- ・ 分散表現によるベクトル (word2vec) を用いる
- ・ ベクトル間の線形変換 (i.e., 行列) を学習する



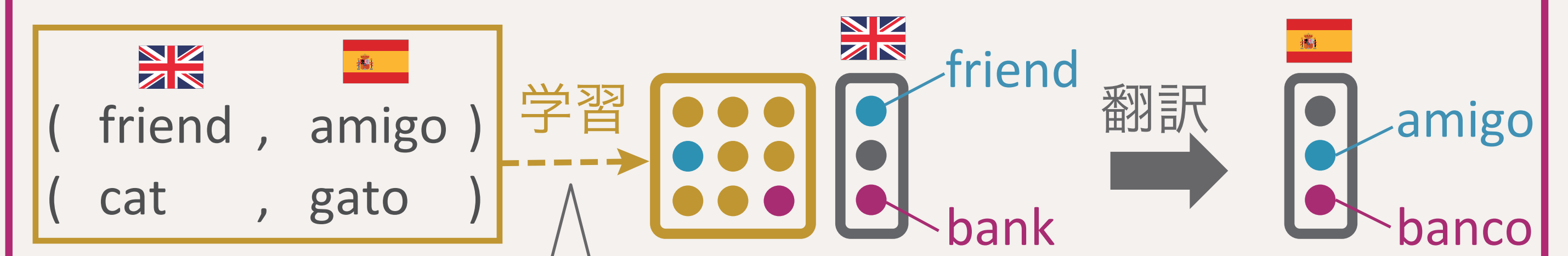
- ☺ 次元間の多: 多の写像を自動的に見つけれられる
- ☹ 文脈語の情報は利用できない

3. 提案手法

学習 + 辞書によるベクトル翻訳

- ・ 文脈語の対訳関係を活用するため、共起頻度ベクトルを用いる
- ・ 文脈語の対訳関係は下記 2 つの手がかりから自動的に獲得する

1. 学習データに含まれる対訳ペア
2. 文脈語間のつづりの類似度



目的関数:

$$\min_W \sum_{i=1}^n \|Wx_i - z_i\|^2 - \beta_{train} \sum_{(j,k) \in \mathcal{D}_{train}} w_{jk} - \beta_{sim} \sum_{(j,k) \in \mathcal{D}_{sim}} w_{jk}$$

[Mikolov+, 13]

x : 原言語のベクトル, z : 目的言語のベクトル, W : 行列

- ☺ 次元間の多: 多の写像を自動的に見つけれられる
- ☺ 文脈語の対訳関係を直接利用できる

4. 実験

データ

- ・ 単語ベクトル: Wikipedia の dump データ
- ・ 学習 / テスト: Open Multilingual Wordnet

評価手順

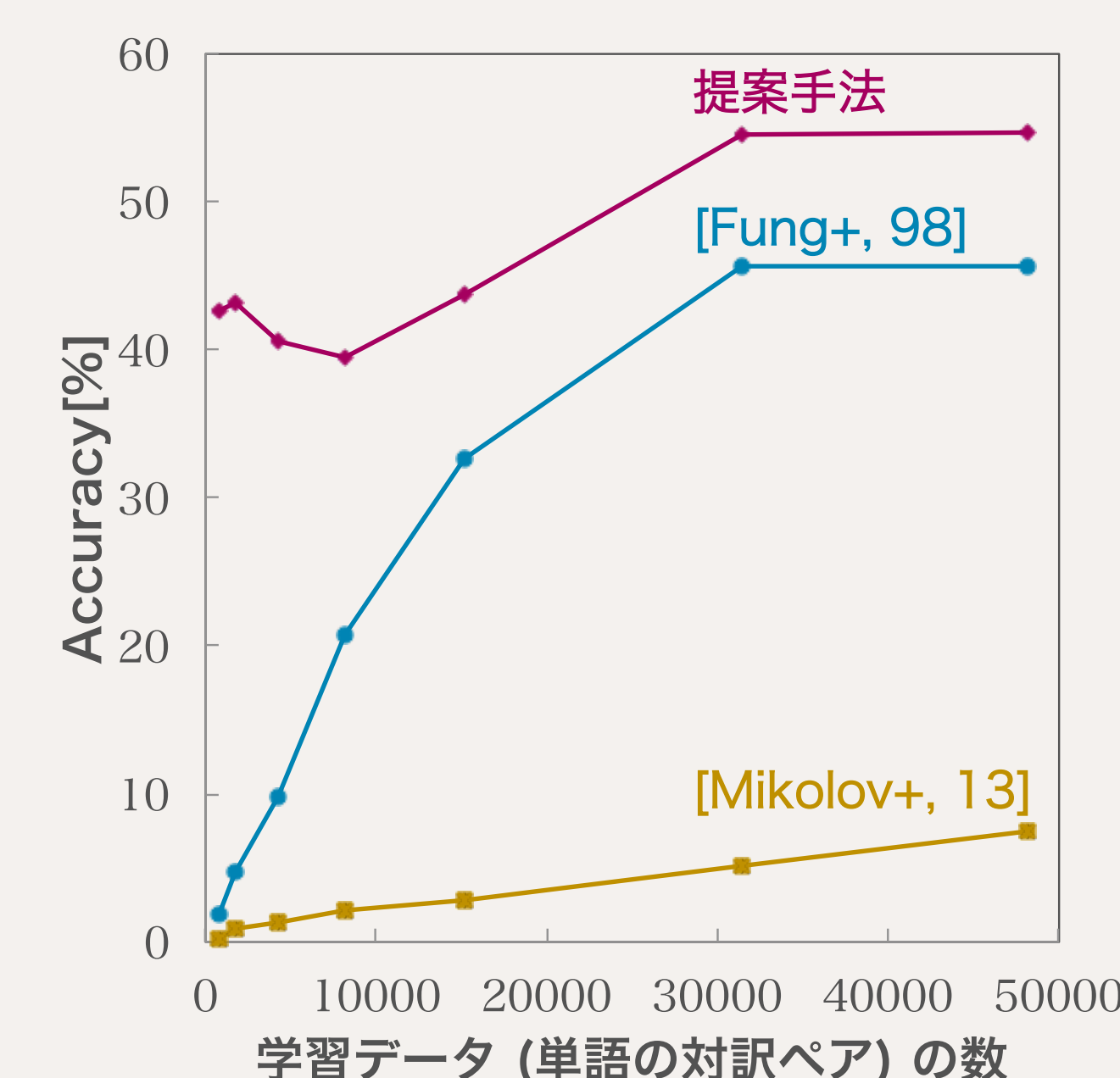
1. 単語ベクトルを原言語→目的言語に翻訳
2. 目的言語の単語ベクトルとの類似度を測り、類似度順に目的言語の単語をソート
3. 上位 n 語以内に正しい訳語があるかを確認

“ゴーレム” の翻訳候補 (日→英)

[Mikolov+, 13]	[Fung+, 98]	提案手法
1. monster	1. hand	1. creature
2. creature	2. turn	2. golem
3. beast	3. creature	3. robot
4. evil	4. robot	4. monster
⋮	⋮	
6. robot	8. golem	
⋮		
26. golem		

正解

学習データの規模が翻訳精度に与える影響 (英→西)



5. 今後の予定

翻訳精度向上に向けて

- ・ 対義語や上位語, 下位語への誤翻訳を防ぐには?
- ・ コーパスのドメイン変化による影響は?

機械翻訳システムへの導入

案 1: デコーダの対数線形モデルの特徴量として類似度を追加

案 2: フレーズテーブルの自動拡張