Improving Word Translation via Two-Stage Contrastive Learning

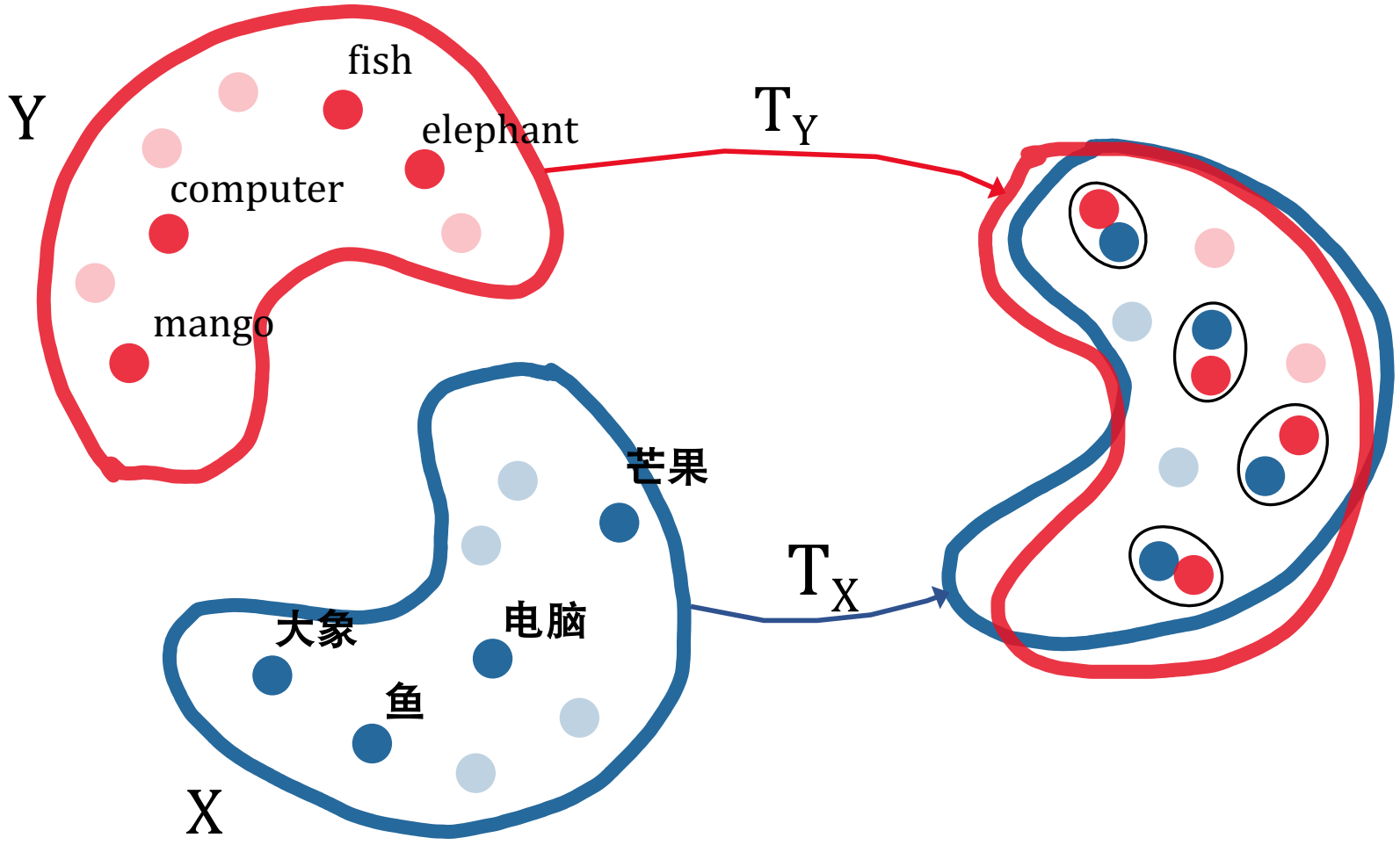
Language Technology Lab, University of Cambridge, UK

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, Ivan Vulić

UNIVERSITY OF CAMBRIDGE

Language Technology Lab

The Task: Bilingual Lexicon Induction



The Task: Bilingual Lexicon Induction

- An important and long-standing task
- Applications of BLI
    - Language learning & language acquisition
    - Machine Translation
    - Cross-lingual Transfer Learning
    - Low-resource NLP

    ……

The Task: Bilingual Lexicon Induction

- Supervised
  - 5K pairs
- Semi-supervised
  - 1K pairs
- Unsupervised

The Task: Bilingual Lexicon Induction

- Supervised
  - 5K pairs
- Semi-supervised
  - 1K pairs
- Unsupervised

Previous Approaches

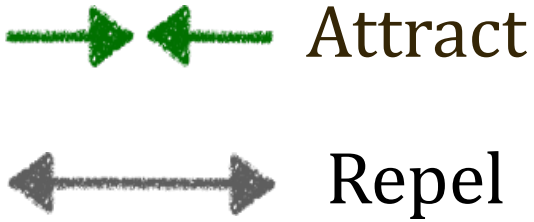
- Mapping-based
  - Orthogonal (Xing et al., 2015)
  - Linear (Joulin et al., 2018)
  - Non-linear (Mohiuddin et al., 2020)

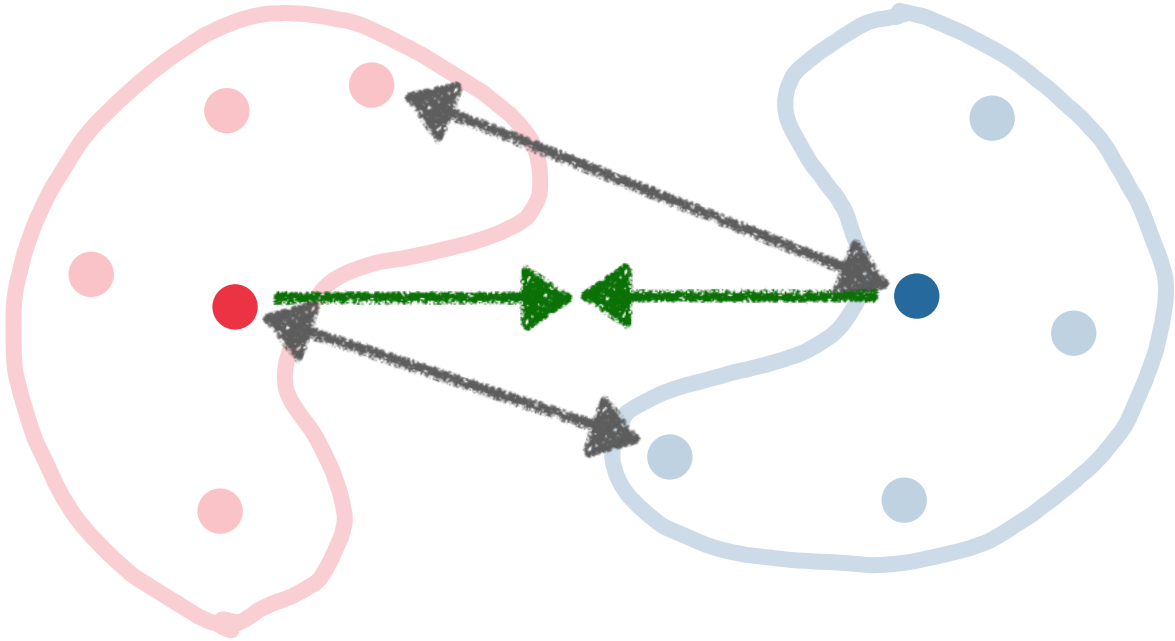
Previous Approaches

- Mapping-based
  - Orthogonal (Xing et al., 2015)
  - Linear (Joulin et al., 2018)
  - Non-linear (Mohiuddin et al., 2020)
- Static WEs ≻ pretrained LMs
  - fastText ≻ monolingual BERT (Vulić et al., 2020)
  - fastText ≻ multilingual LMs (e.g. mBERT) (Gonen et al., 2020)
  - Why?

Motivation

- Contrastive Learning (CL)
  - Success in sentence encoders (Gao et al., 2021; Liu et al., 2021)
  - BLI ?





Motivation

- We propose a novel two-stage CL approach.
  - Stage C1 can be evaluated independently
  - Stage C2, further improvement

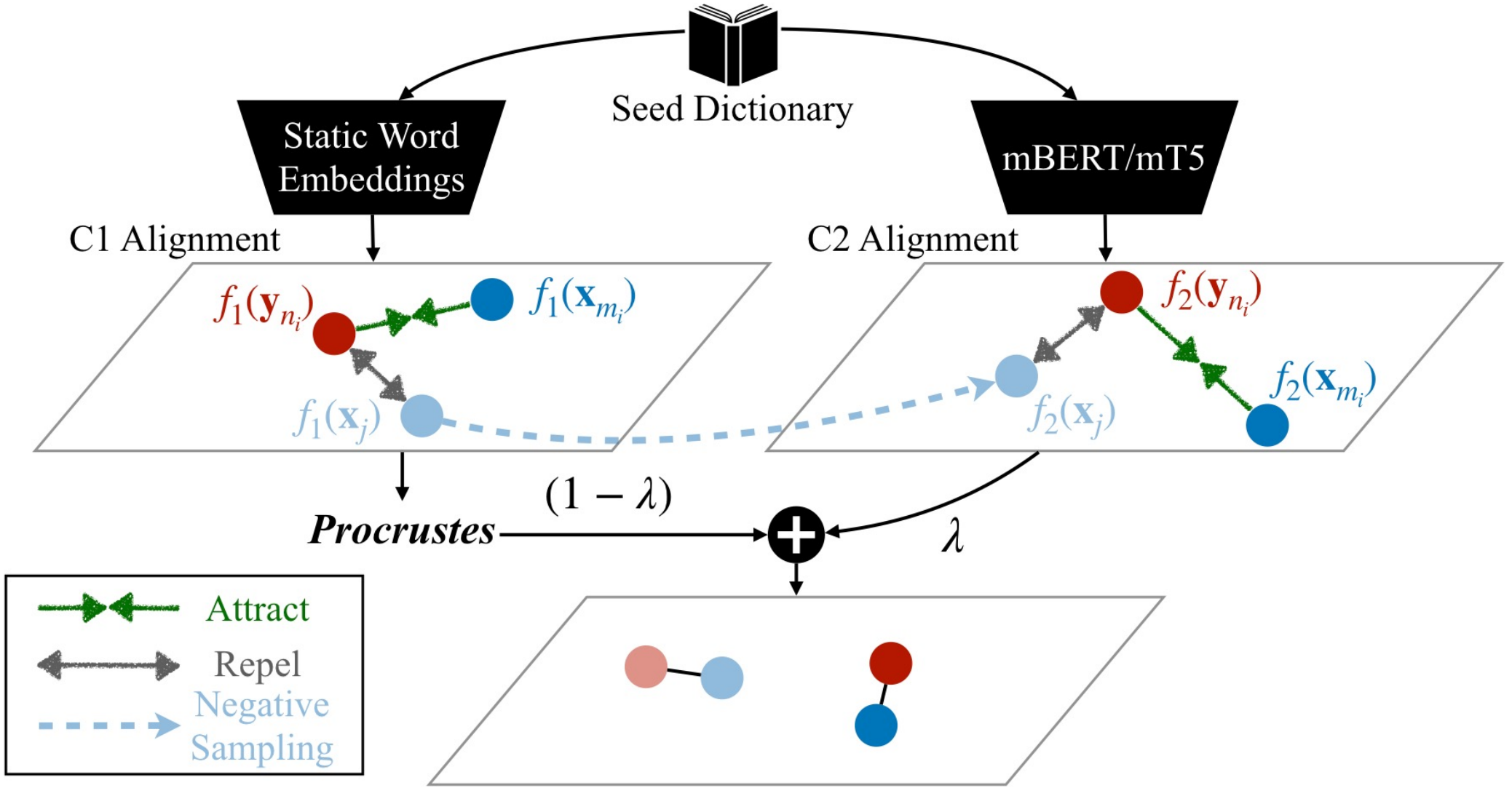
Motivation

- We propose a novel two-stage CL approach.
  - Stage C1 can be evaluated independently
  - Stage C2, further improvement
- Stage C1
  - Static WEs + CL = SotA (RQ 1)

Motivation

- We propose a novel two-stage CL approach.
  - Stage C1 can be evaluated independently
  - Stage C2, further improvement
- Stage C1
  - Static WEs + CL = SotA (RQ 1)
- Stage C2
  - LMs + CL = Lexical Encoders (RQ 2)
  - CL exposes translation knowledge (RQ 2)
  - Combine LMs with static WEs (RQ 3)

Method: Two-Stage CL



Method: CL in Stage C1 (Static WEs)

Linear Map on L_X WE

Linear Map on L_Y WE

$$s_{i,j} = \exp(\cos(\boldsymbol{x}_i \boldsymbol{W}_x, \boldsymbol{y}_i \boldsymbol{W}_y)/\tau)$$

A Positive Pair

$$p_i = \frac{s_{m_i,n_i}}{\sum_{w_j^y \in \{w_{n_i}^y\} \cup \bar{w}_{n_i}^y} s_{m_i,j} + \sum_{w_j^x \in \bar{w}_{m_i}^x} s_{j,n_i}}$$

Hard Negative Pairs

$$\min_{\boldsymbol{W}_x, \boldsymbol{W}_y} -\mathbb{E}_{\left(w_{m_i}^x, w_{n_i}^y\right) \in \mathcal{D}_{CL}} \log(p_i)$$

Method: CL in Stage C2 (multilingual LMs)

Encoder L_X Word with LM

Encode L_Y Word with LM

$$s'_{i,j} = \exp(\cos(f_\theta(w_i^x), f_\theta(w_j^y))/\tau)$$

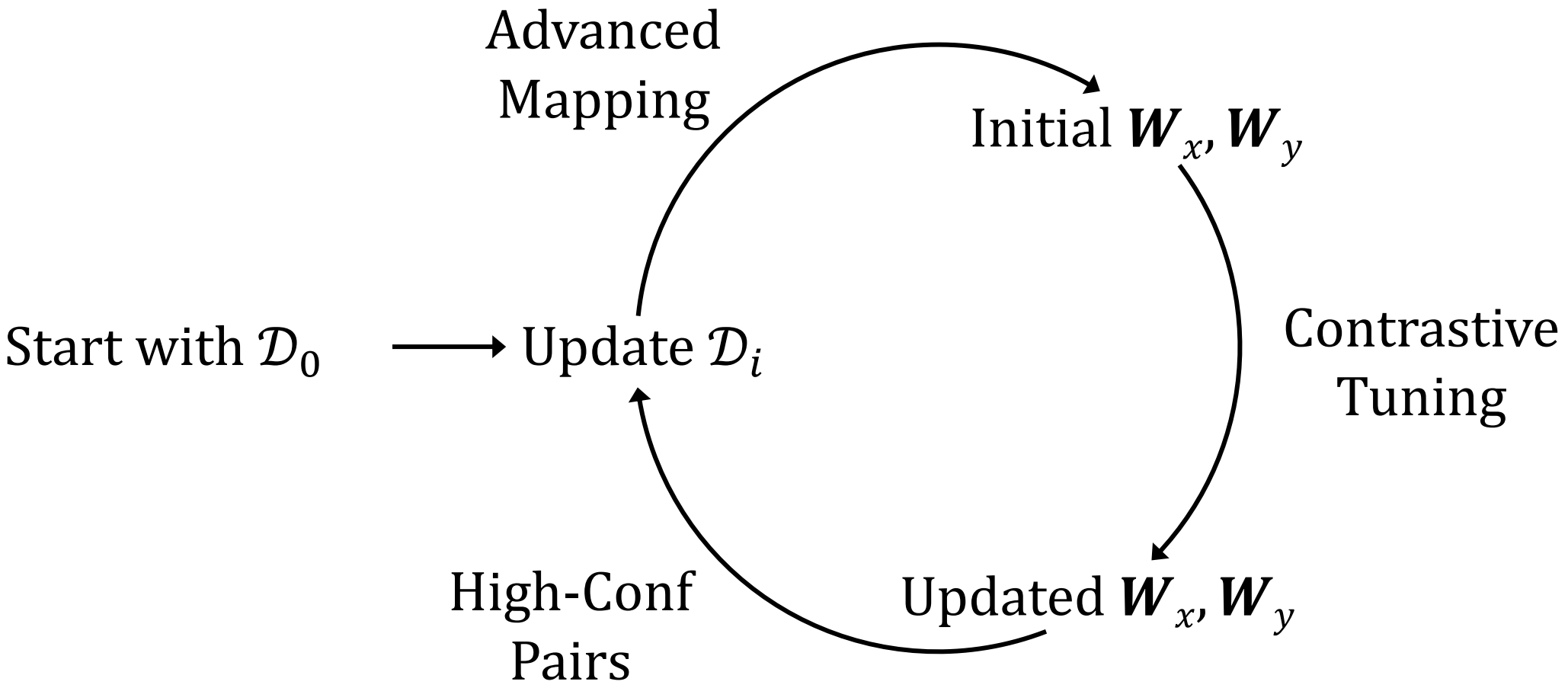
A Positive Pair

$$p'_i = \frac{s'_{m_i,n_i}}{\sum_{w_j^y \in \{w_{n_i}^y\} \cup \bar{w}_{n_i}^y} s'_{m_i,j} + \sum_{w_j^x \in \bar{w}_{m_i}^x} s'_{j,n_i}}$$

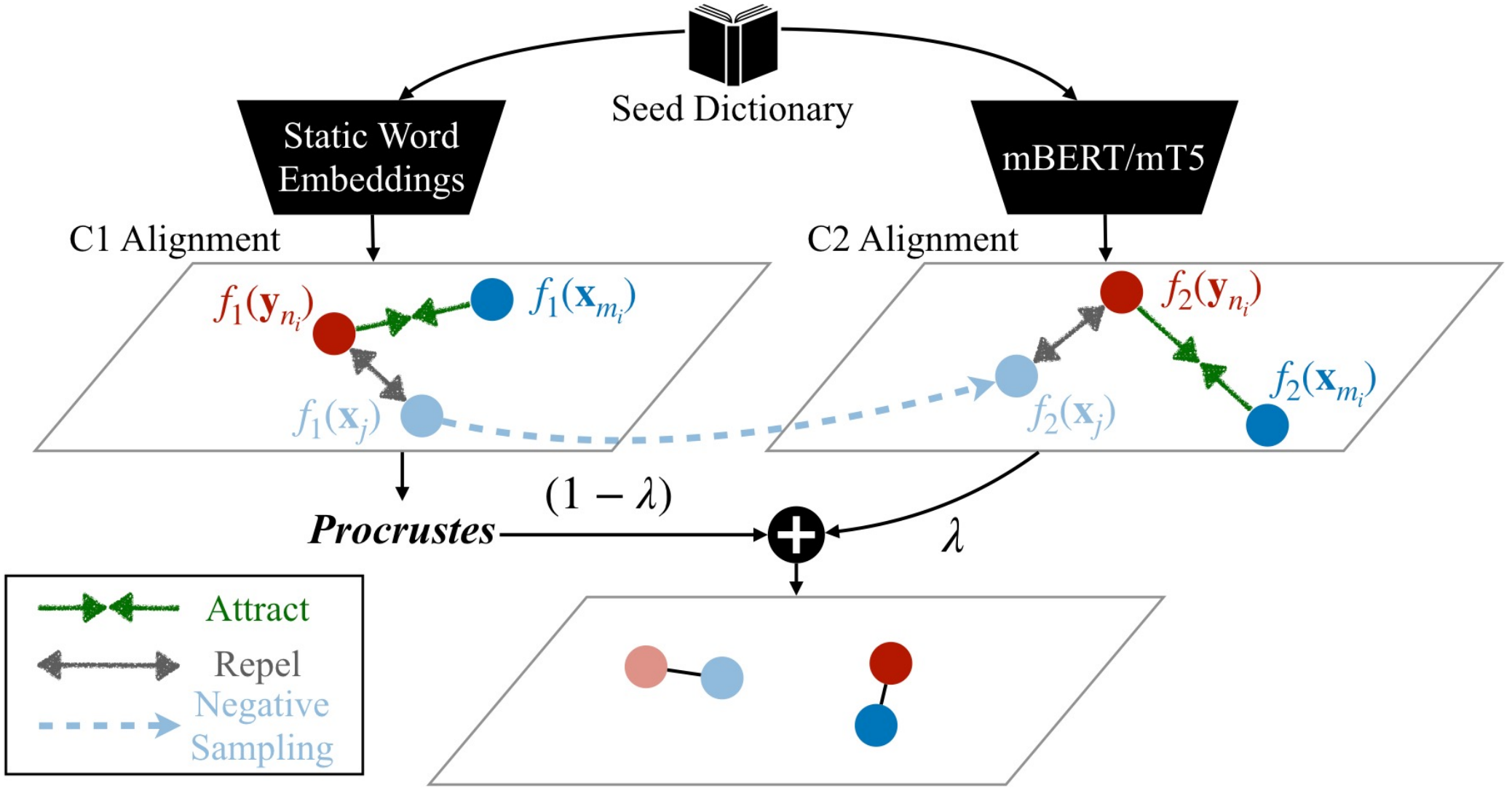
$$\min_\theta - \mathbb{E}_{(w_{m_i}^x, w_{n_i}^y) \in \mathcal{D}_{CL}} \log(p'_i)$$

Hard Negative Pairs

Method: Stage C1



Method: Stage C2



Evaluation

- Datasets
  - XLING (112 setups) (Glavaš et al., 2019)
  - PanLex-BLI (lower-resource languages) (Vulić et al., 2019)
  - Vocabulary size: 200K

Evaluation

- Datasets
  - XLING (112 setups) (Glavaš et al., 2019)
  - PanLex-BLI (lower-resource languages) (Vulić et al., 2019)
  - Vocabulary size: 200K
- Baselines
  - RCSLS (w/o SL) (Joulin et al., 2018)
  - VecMap (w/ SL) (Artetxe et al., 2018)
  - LNMap (w/ SL) (Mohiuddin et al., 2020)
  - FIPP (w/ SL) (Sachidananda et al., 2021)

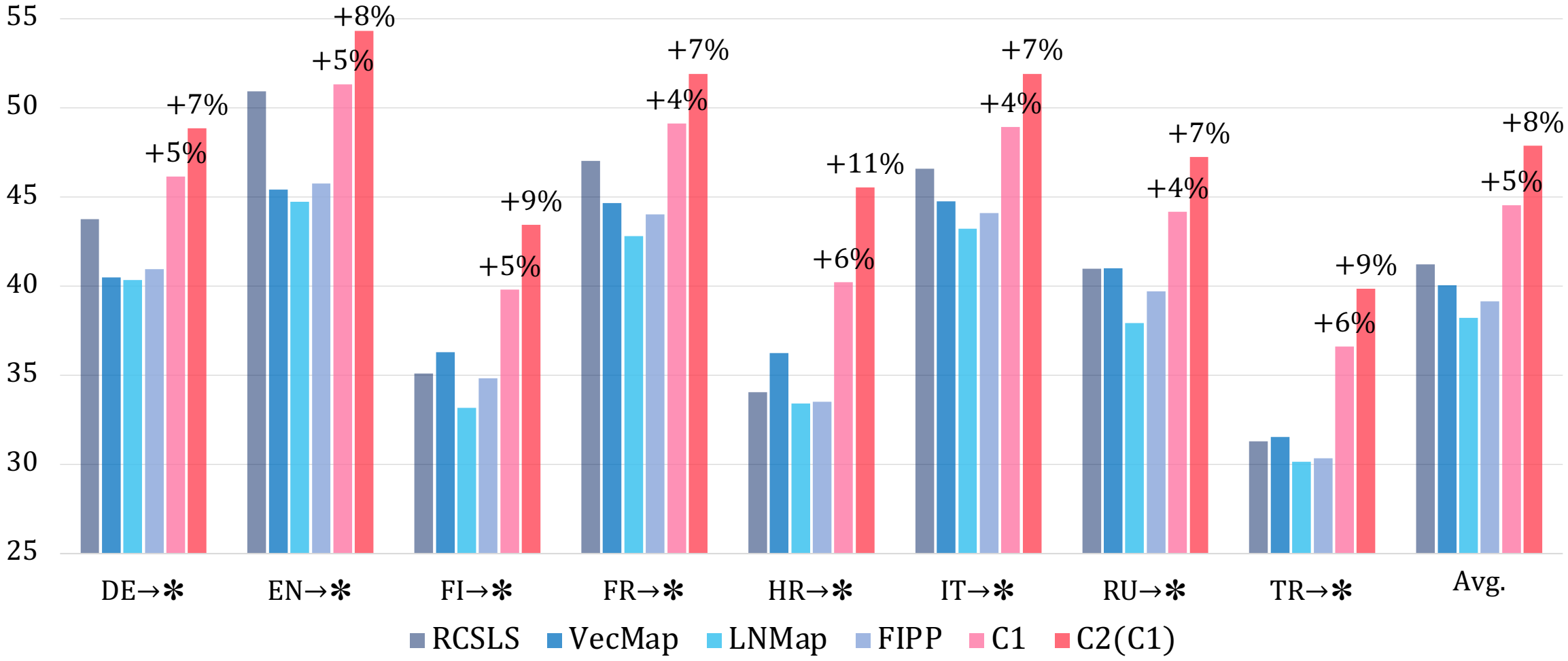
Evaluation

- Statics WEs & Multilingual LMs
    - fastText
    - mBERT (main); XLM, mT5 (comparison)

Evaluation

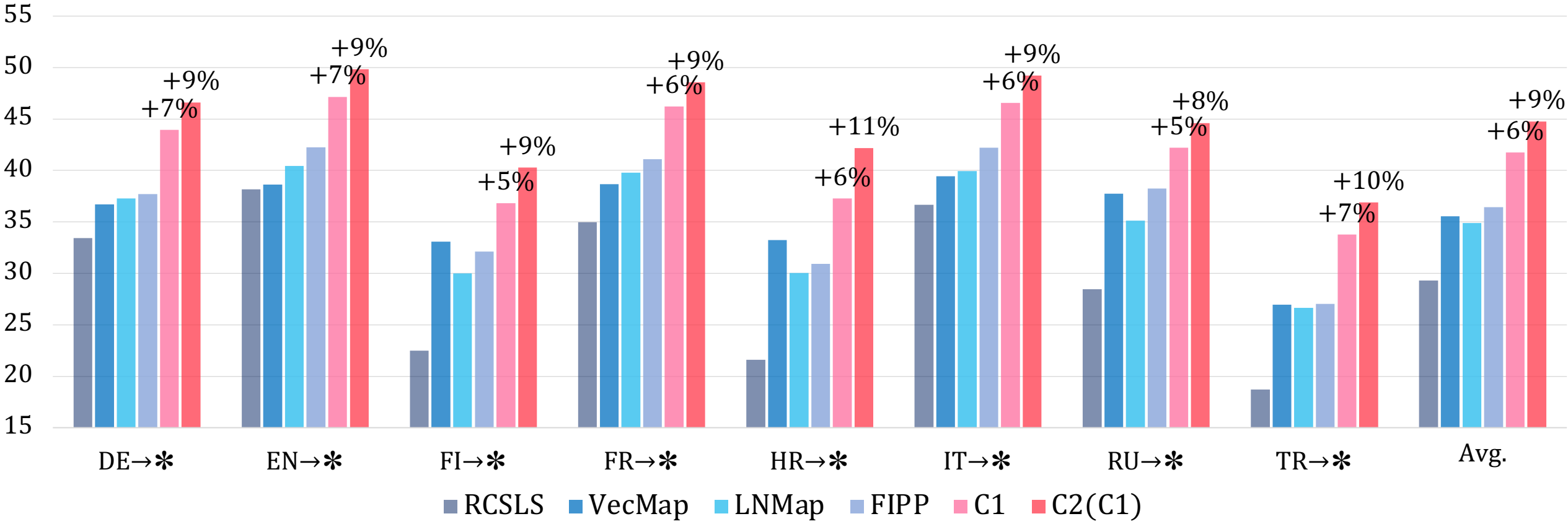
- Statics WEs & Multilingual LMs
  - fastText
  - mBERT (main); XLM, mT5 (comparison)
- Hyperparameters
  - Tuned on EN→TR
  - λ=0.2 (fixed)

Evaluation: XLING 5K



- C1: ≈ 5% higher than SotA (Avg. of 4 Baselines)
- C2: ≈ 8% higher than SotA (Avg. of 4 Baselines)

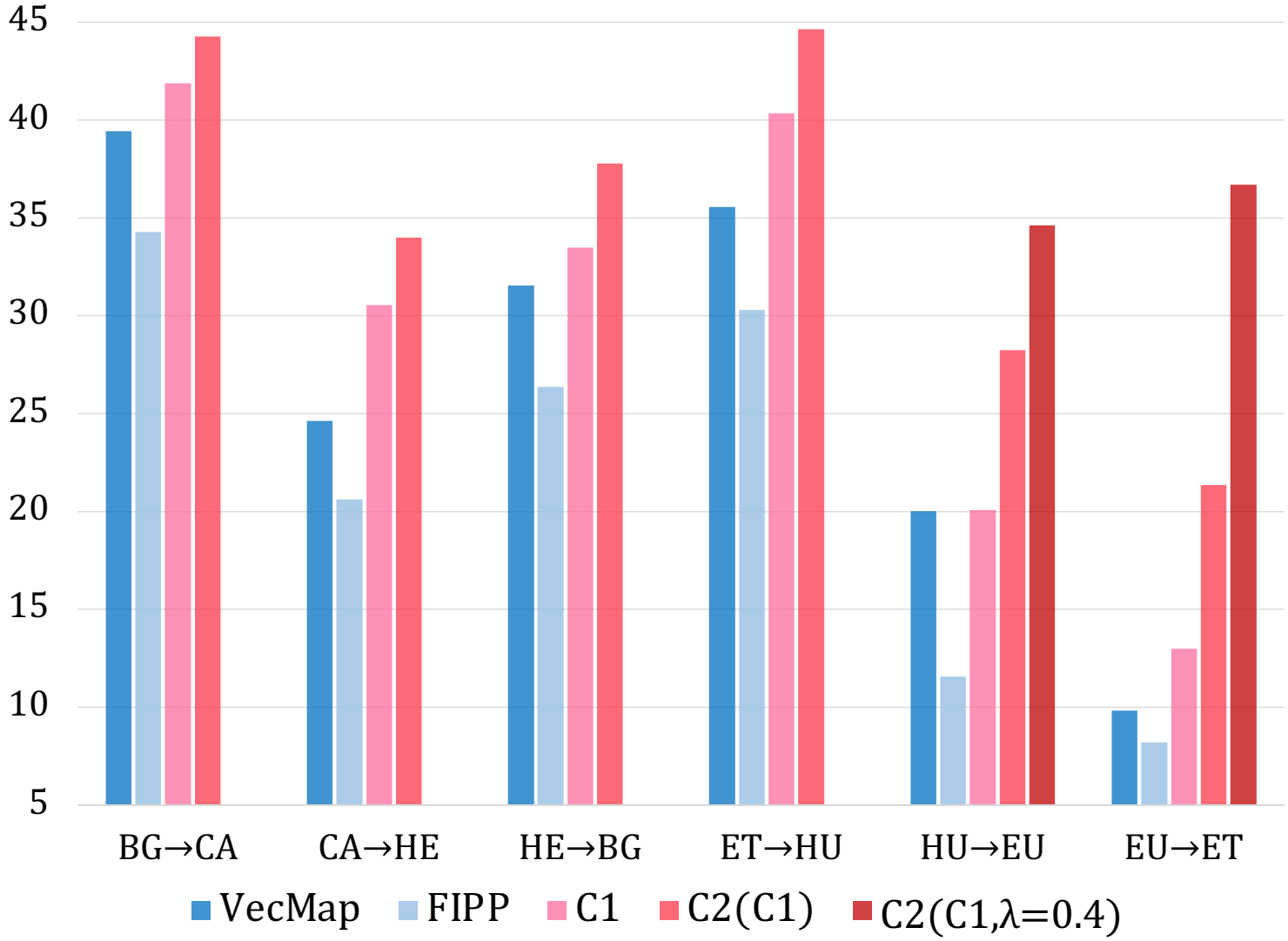
Evaluation: XLING 1K



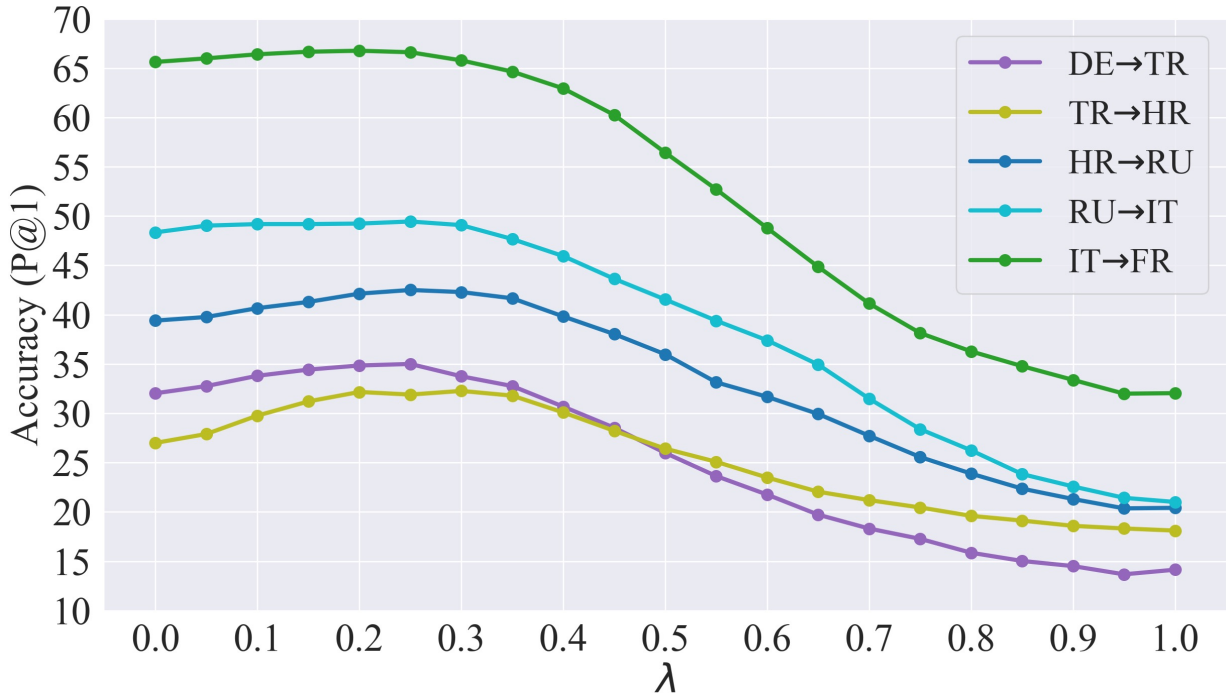
- C1: ≈ 6% higher than SotA (Avg. of 3 stronger baselines)
- C2: ≈ 9% higher than SotA (Avg. of 3 stronger baselines)

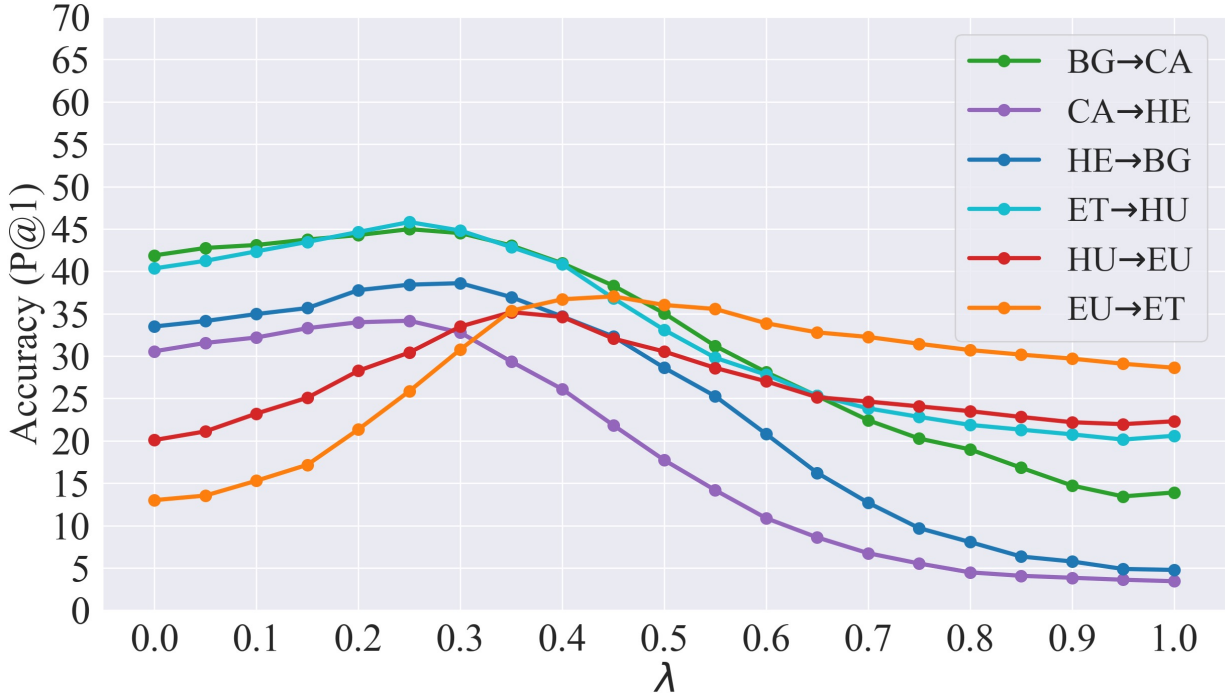
Evaluation: PanLex-BLI 1K

- Same trends
- PLMs more useful?
  - Lower-resource

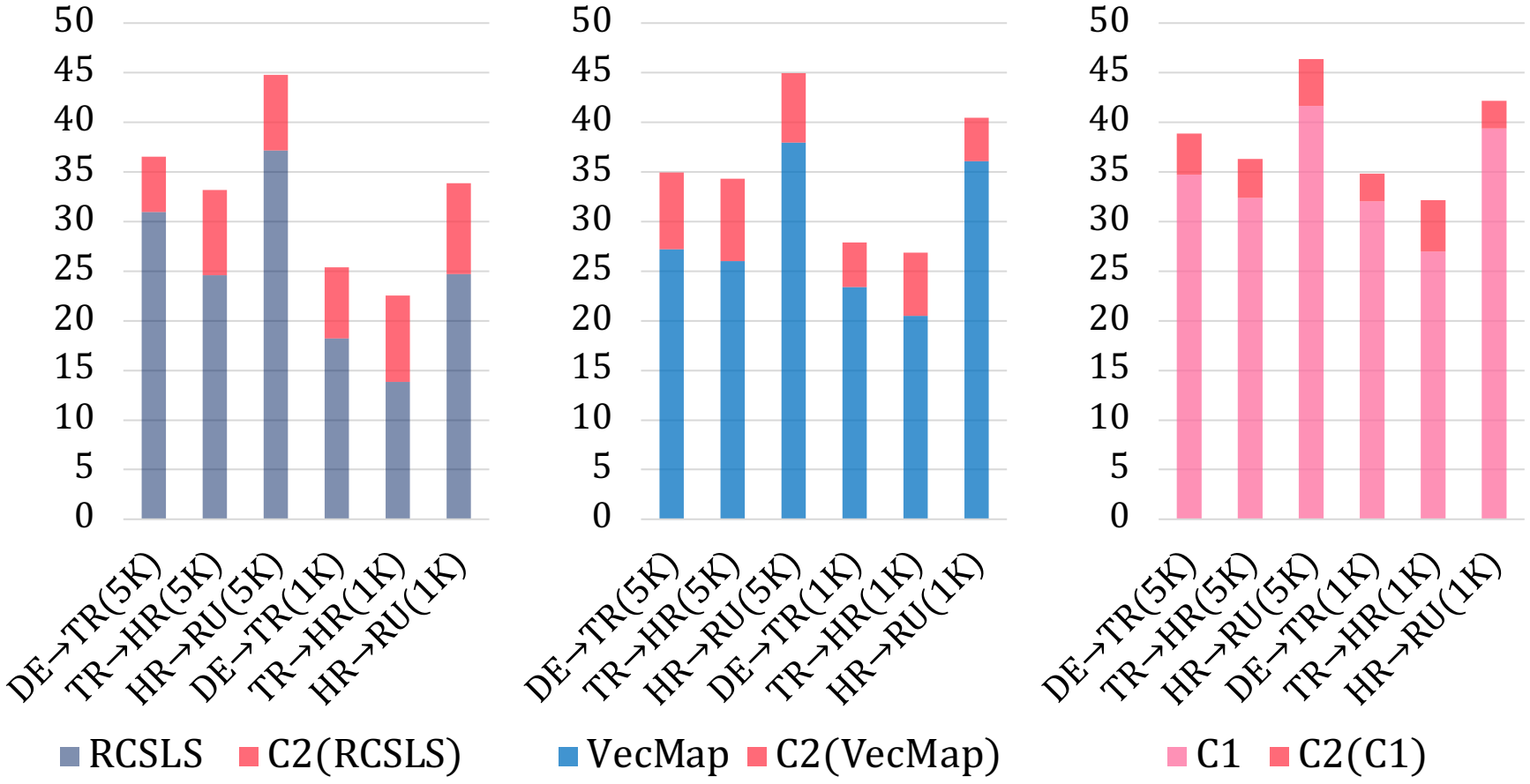


Analysis: λ Values



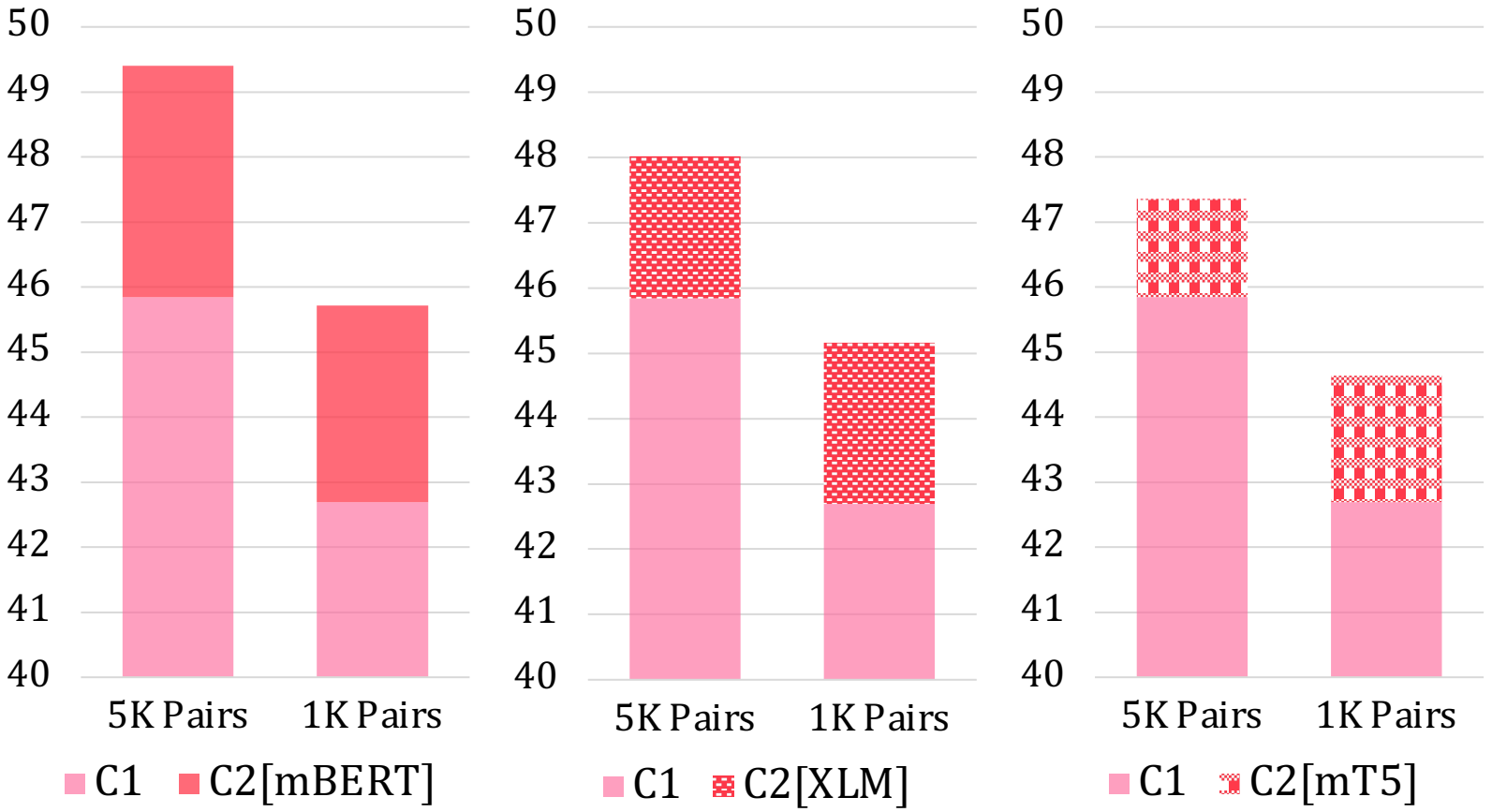


Analysis: Different ‘Support’ Methods



- C2(mod)
  - C2(RCSLS) ✓
  - C2(VecMap) ✓
  - C2(C1) ✓

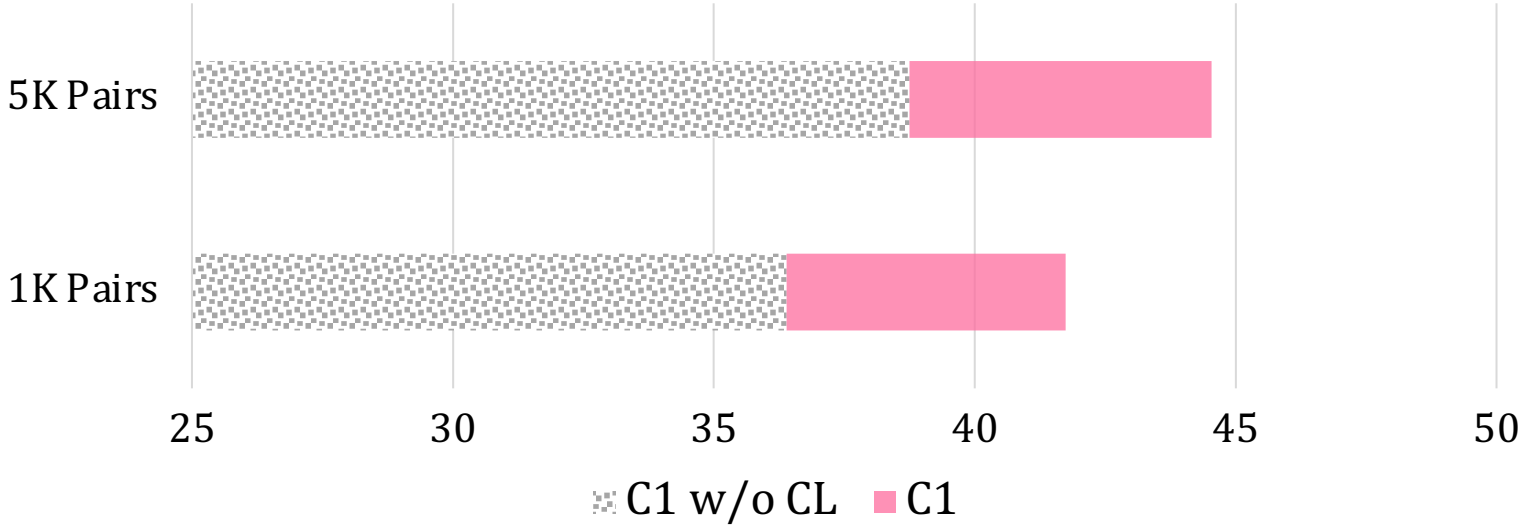
Analysis: Different Multilingual LMs



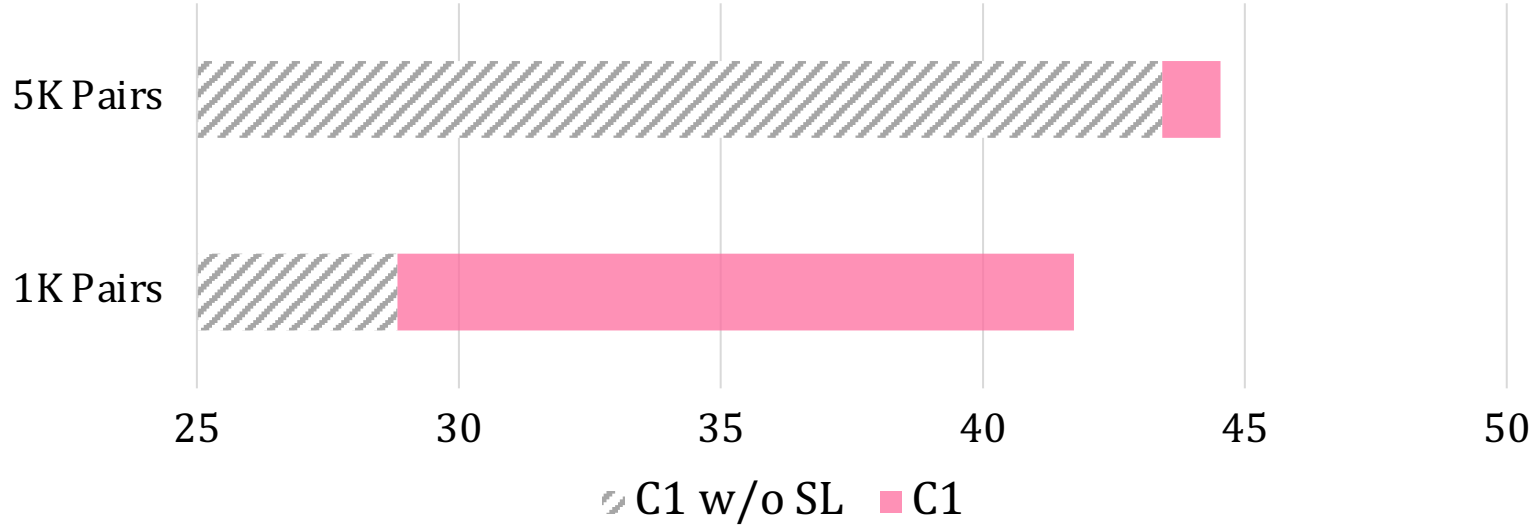
- C2[LM]
  - C2[mBERT] ✓
  - C2[XLM] ✓
  - C2[mT5] ✓

Analysis: Ablation Study — Stage C1

- Contrastive Learning
  - 5K ✓
  - 1K ✓

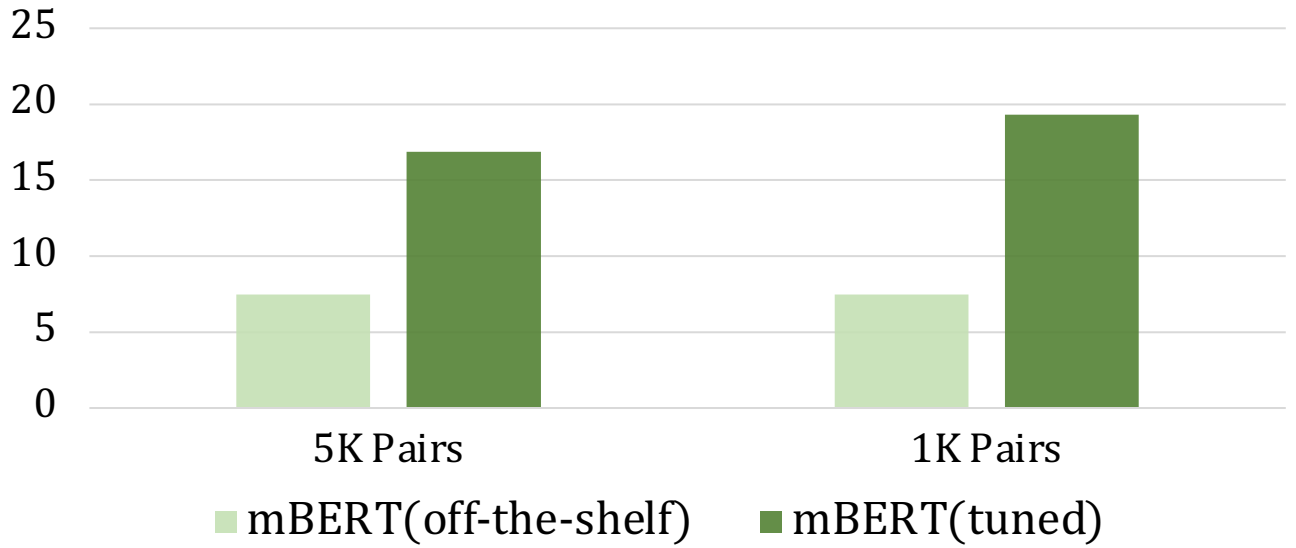


- Self-Learning
  - 5K (slight gains)
  - 1K ✓

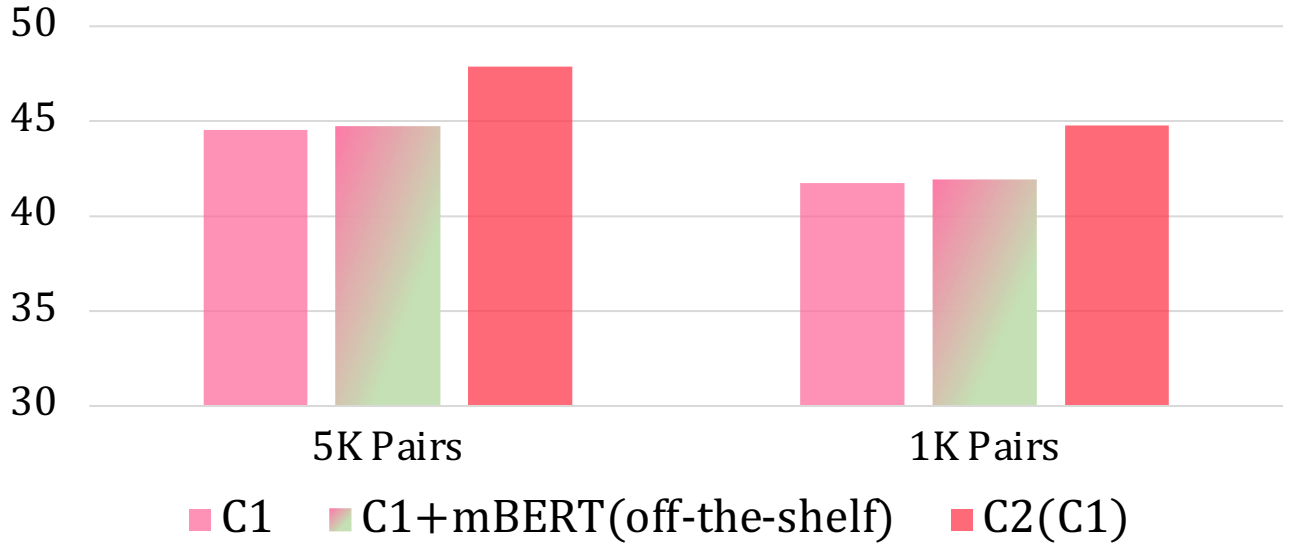


Analysis: Ablation Study — Stage C2

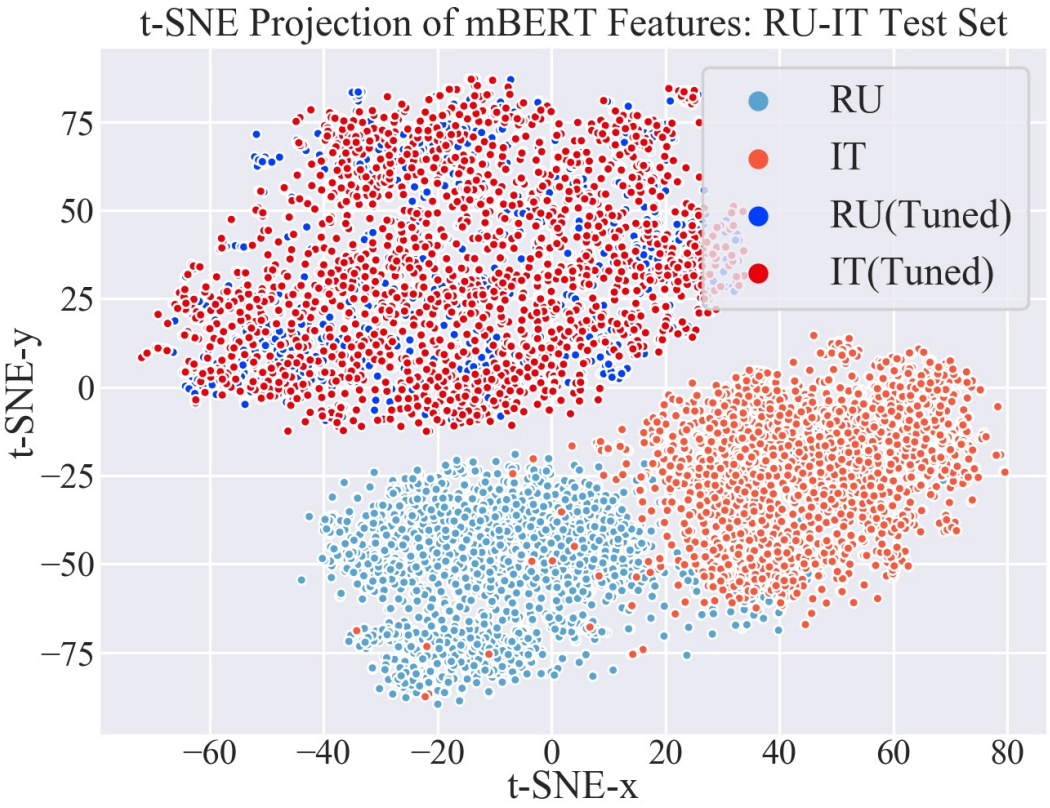
- Contrastive Tuning
  - Tuned ≻ Off-the-shelf
  - fastText (C1) ≻ mBERT

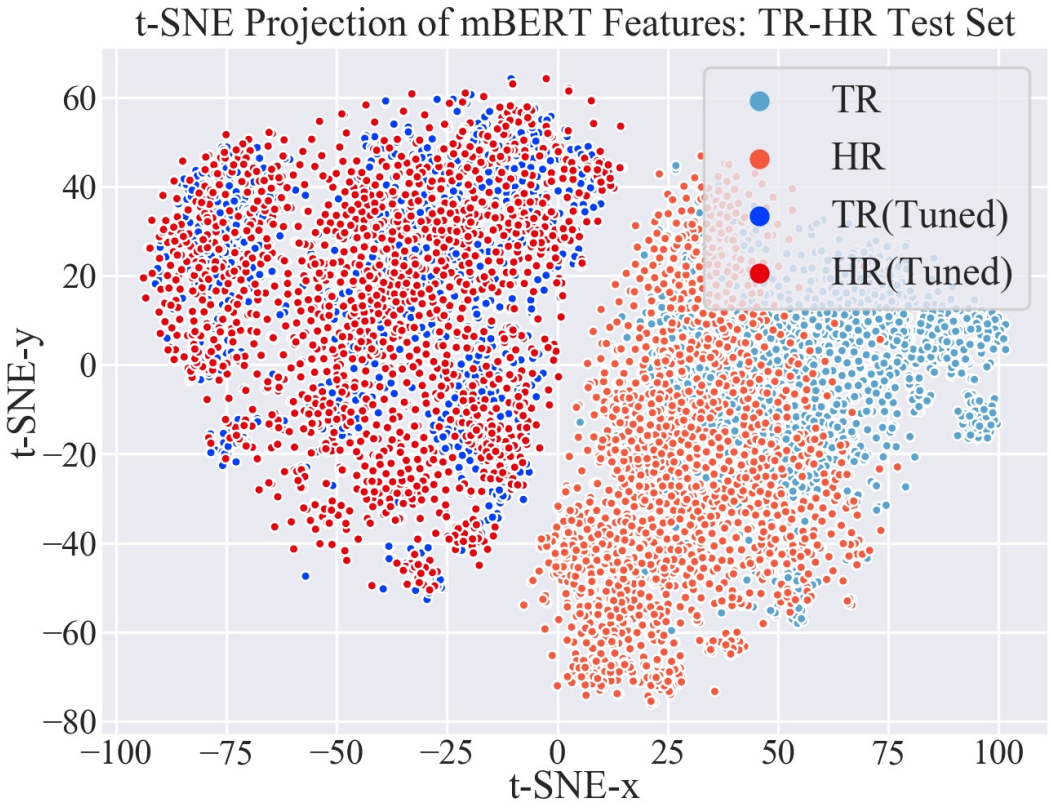


- Combing C1 & mBERT
  - CL exposes word translation knowledge in mBERT

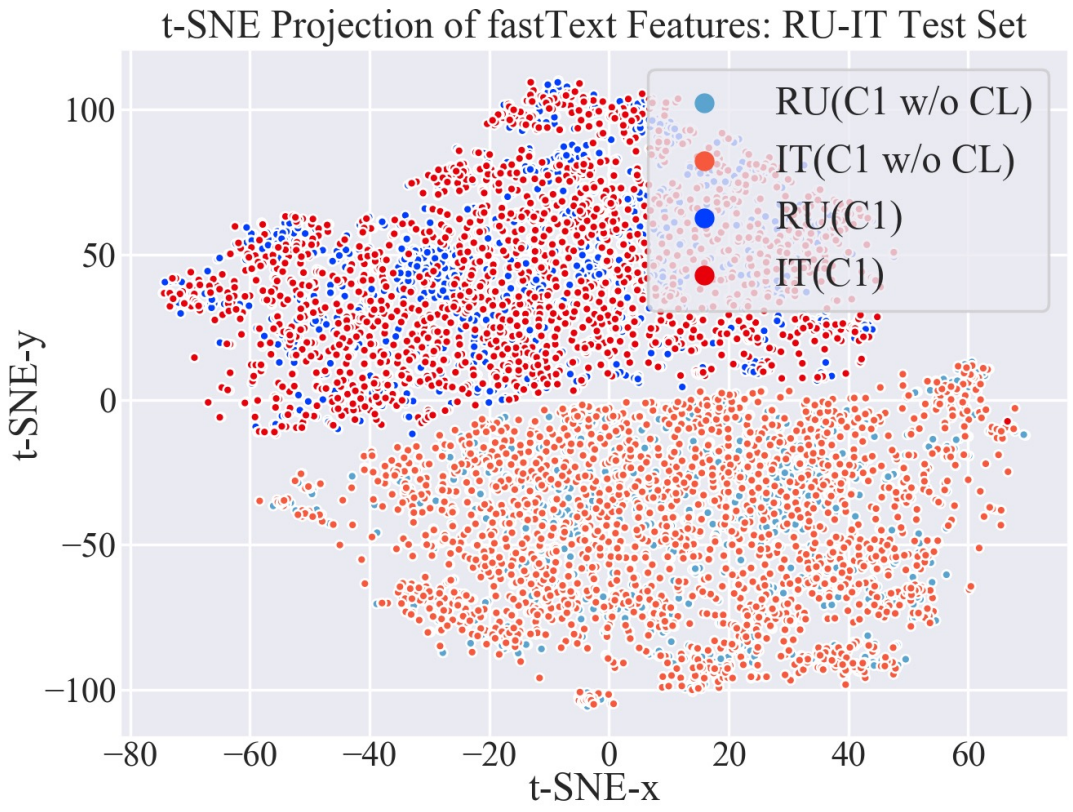


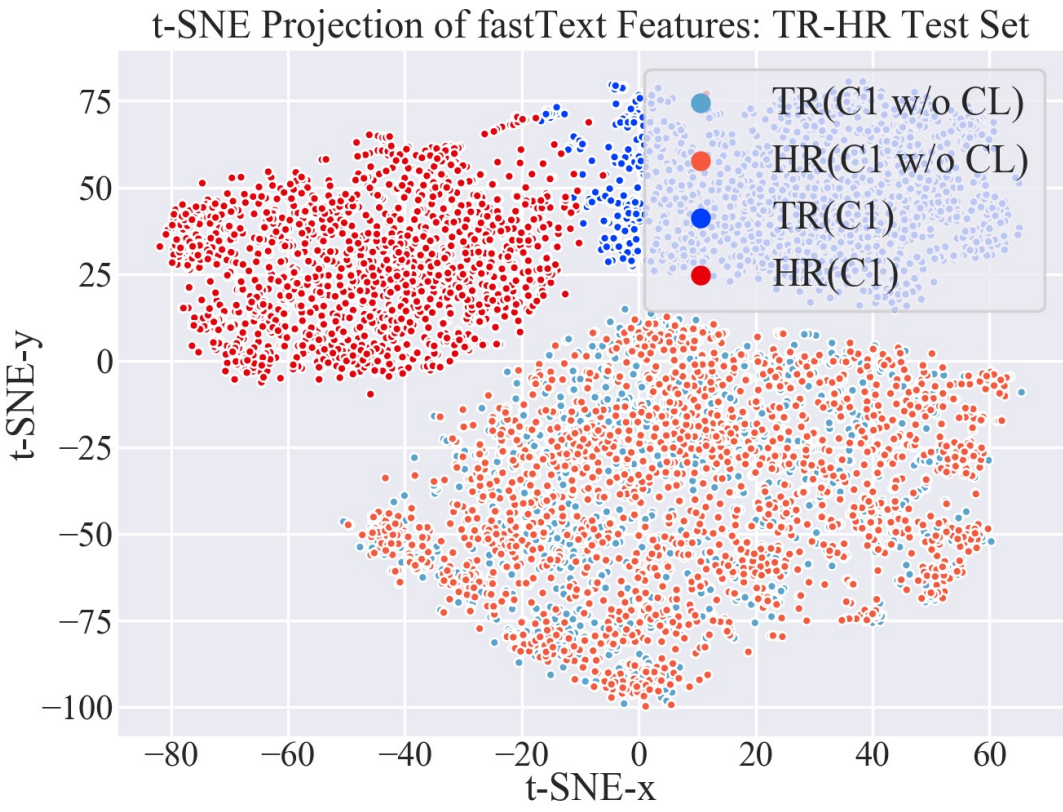
Analysis: Effectiveness of CL (Stage C2)





Analysis: Effectiveness of CL (Stage C1)





Summary

- A two-stage framework for BLI
    - Simple, effective and robust

Summary

- A two-stage framework for BLI
  - Simple, effective and robust
- Stage C1
  - Static WEs only, e.g., fastText
  - SotA ← evaluated independently
- Stage C2
  - CLWEs + Multilingual LMs, e.g., mBERT
  - Extra gains

Summary

- A two-stage framework for BLI
  - Simple, effective and robust
- Stage C1
  - Static WEs only, e.g., fastText
  - SotA ← evaluated independently
- Stage C2
  - CLWEs + Multilingual LMs, e.g., mBERT
  - Extra gains
- A series of analyses & ablation studies

Summary (Technically)

- Contrastive Learning

Static WEs (C1)

Multilingual LMs (C2)

Summary (Technically)

- Contrastive Learning

Static WEs (C1) Multilingual LMs (C2)

- Self-Learning
  - Semi-supervised
- Multilingual LMs
  - Word translation knowledge

- Might be useful for future BLI research: CL, SL, and multilingual (also monolingual) LMs.

Thanks for Watching!

Code: https://github.com/cambridgeltl/ContrastiveBLI

Contact: yl711@cam.ac.uk