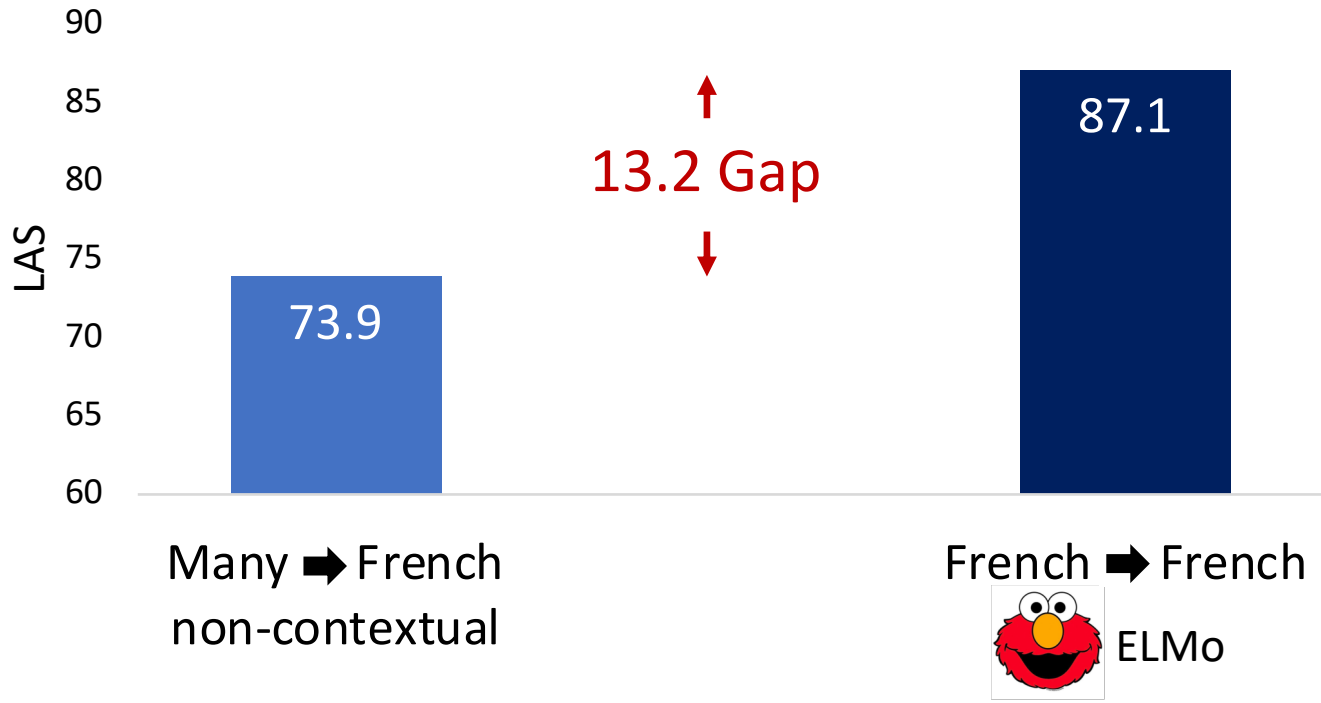
Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing

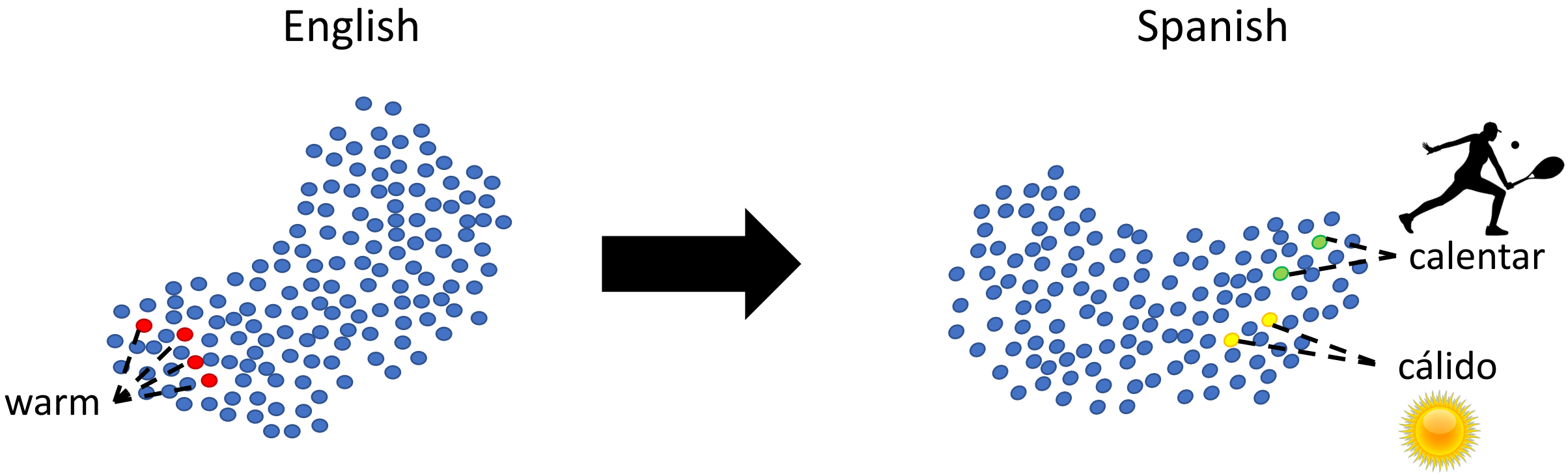
Tal Schuster*, Ori Ram*, Regina Barzilay, Amir Globerson

Task: Cross-lingual Zero-shot Dependency Parsing

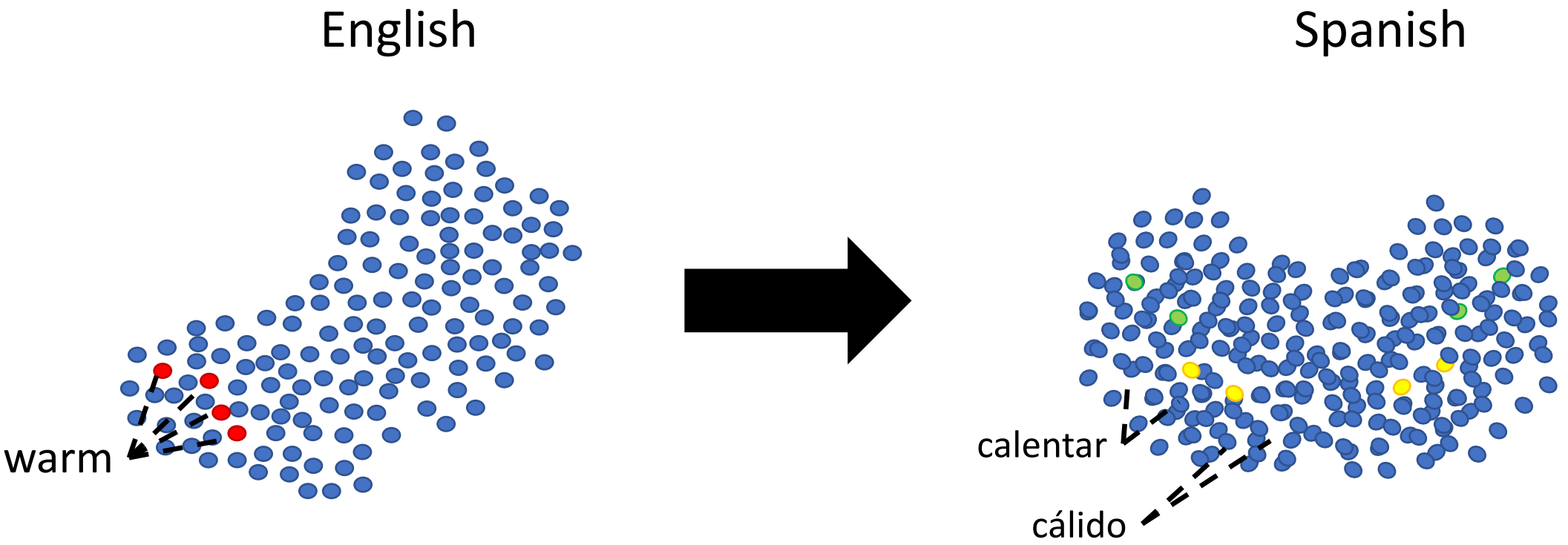


Goal: Utilize universal space of contextual embeddings

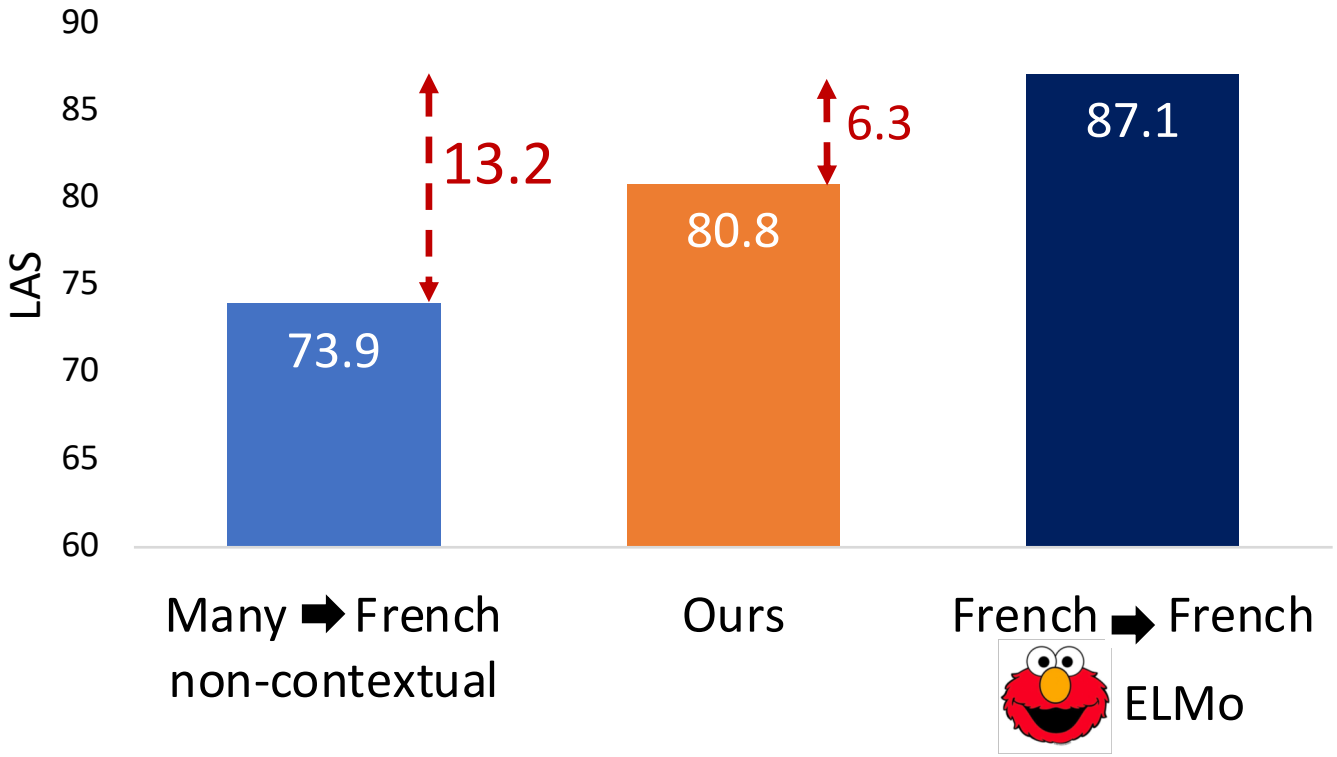
Idea: Align Contextual Word Embeddings



Idea: Align Contextual Word Embeddings



Our Results – zero-shot



By aligning ELMo contextual embeddings

Problem Definition

<u>English</u>

- WIKIPEDIA

ELMo embeddings

- POS tags

<u>Dictionary</u>

bear	oso
warm	cálido
…	…

<u>Spanish</u>

- WIKIPEDIA

ELMo embeddings

- POS tags

<u>Goal</u>: Learn a linear alignment (W)

Problem Definition - Extensions

English

- WIKIPEDIA

ELMo embeddings

- ~~POS tags~~

Spanish

- WIKIPEDIA

ELMo embeddings

- ~~POS tags~~

Goal: Learn a linear alignment (W)

Problem Definition - Extensions

<u>English</u>

- WIKIPEDIA

ELMo embeddings

- ~~POS tags~~

<u>Dictionary</u>

bear	oso
warm	cálido
…	…

<u>Spanish</u>

Small

- WIKIPEDIA

Deficient ELMo embeddings

- ~~POS tags~~

<u>Goal</u>: Alignment (W) and improve the embeddings

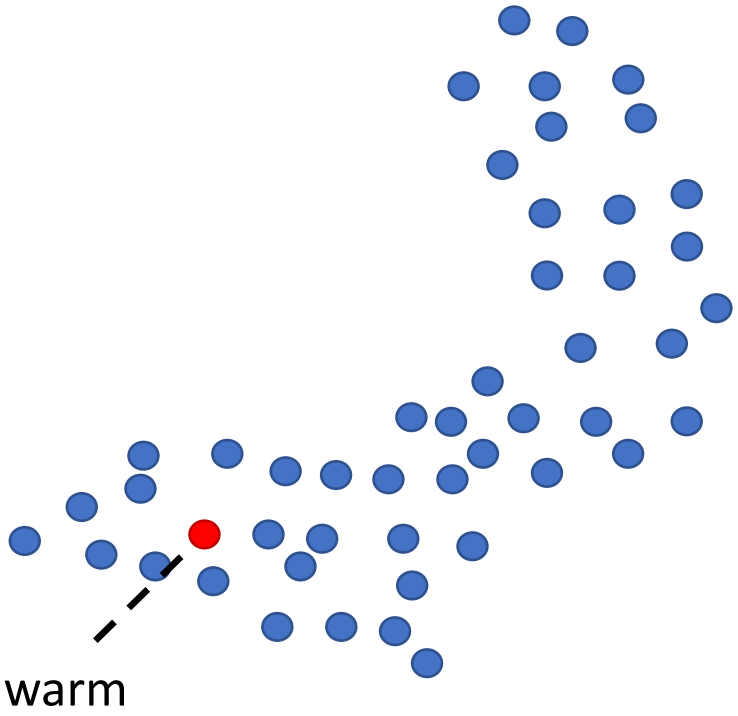
Aligning Embeddings - Static Case

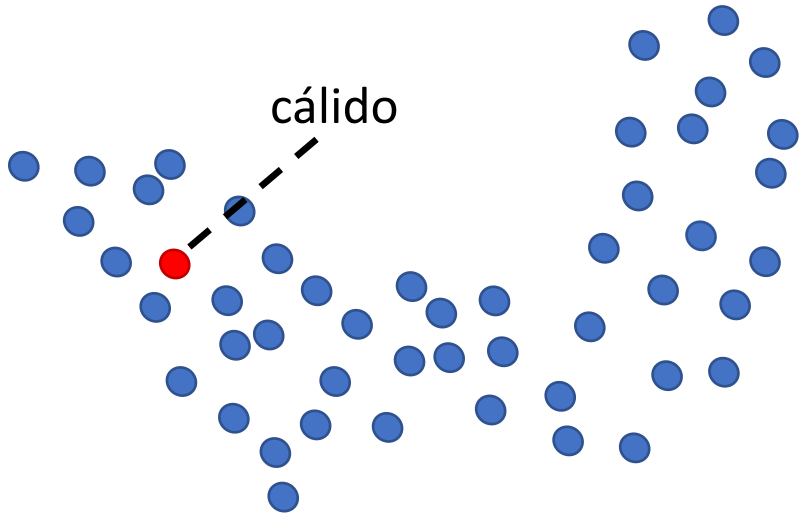
English

Spanish

$$e_i^{EN} = W e_i^{ES}$$

$$W = \underset{W \in O_d}{\operatorname{argmin}} \sum \left\| e_i^{EN} - W e_i^{ES} \right\|^2$$

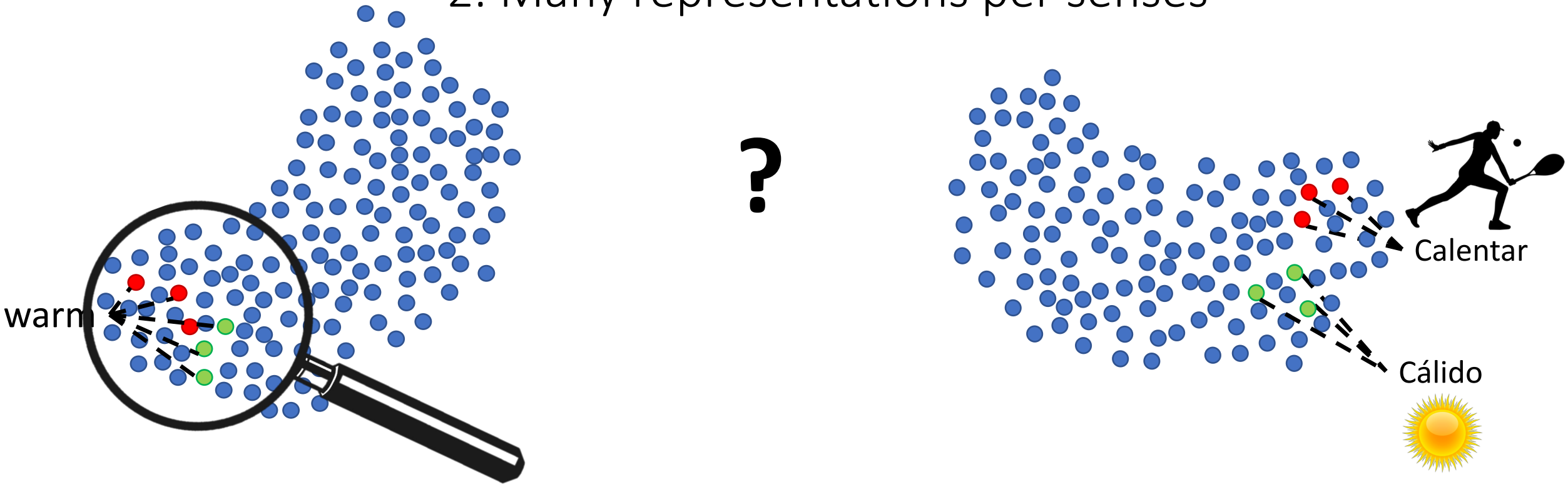




(Mikolov et al., 2013)

Aligning Embeddings - Contextual Case

Challenges: 1. Multiple senses per token
2. Many representations per senses

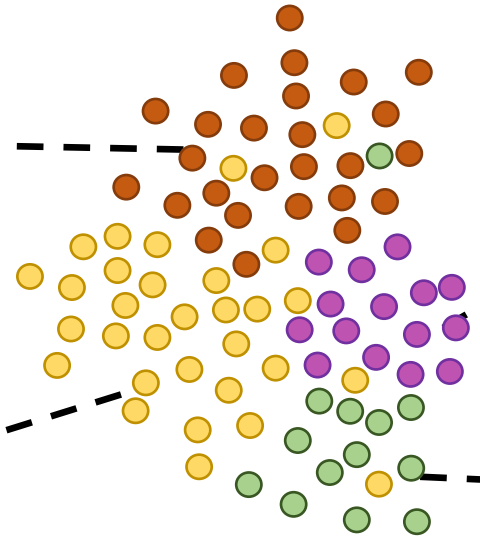


The Contextual Component

- Contextual embeddings of the word “*warm*”:

*He was a **warm** friend of Pope St. Gregory.*

*Sunday was a glorious day , clear and **warm**.*

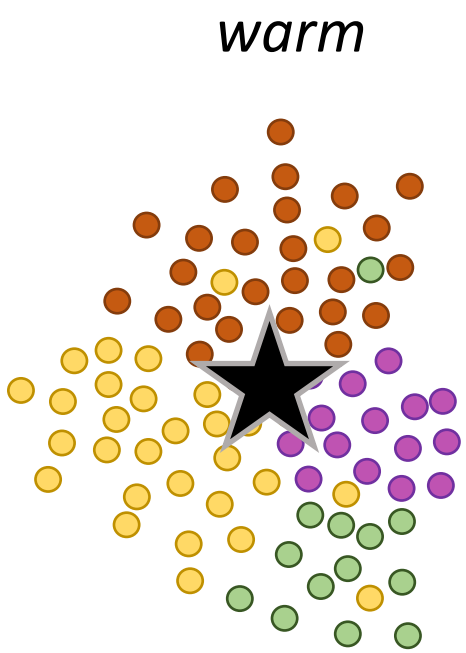


*Fuzz (electric guitar) , distortion effects to create "**warm**" and "dirty" sounds.*

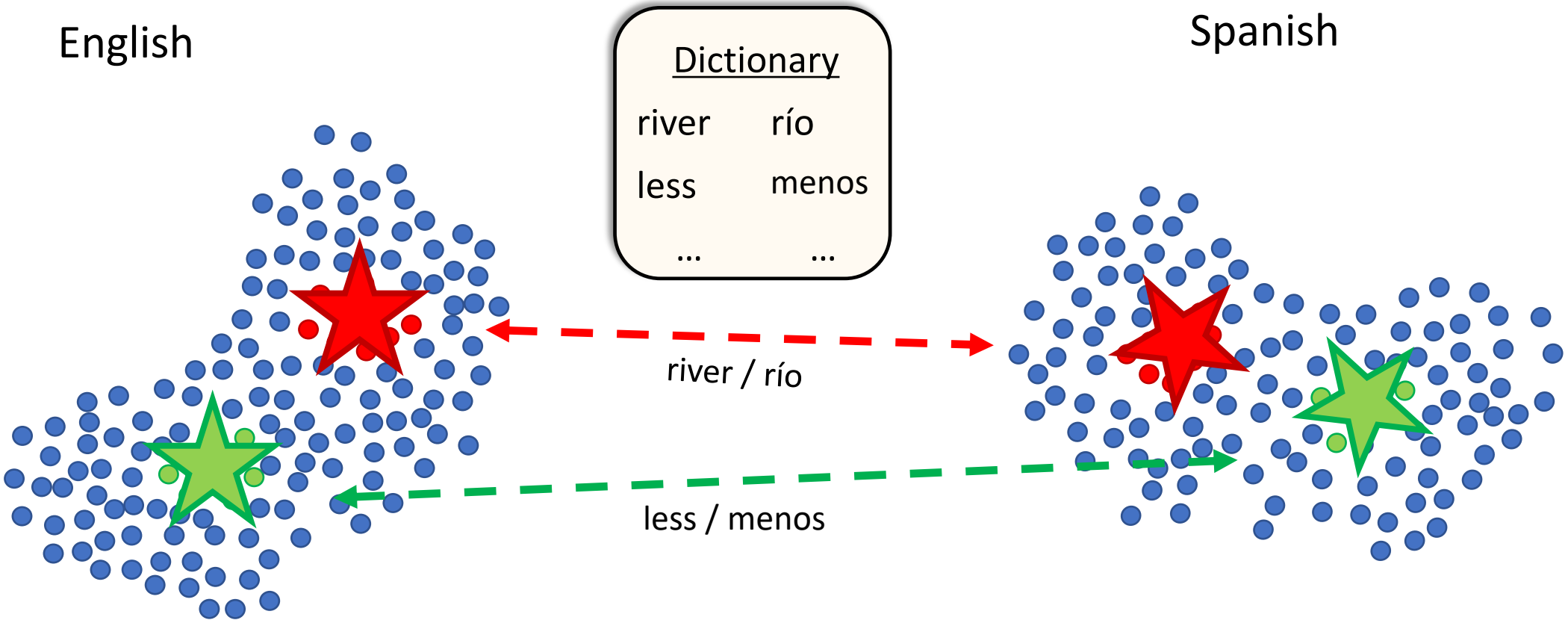
*winning just four matches in her Wimbledon **warm** up tournaments*

Per Token Anchor

$$\bar{\boldsymbol{e}}_i = \mathbb{E}_c[e_{i,c}]$$

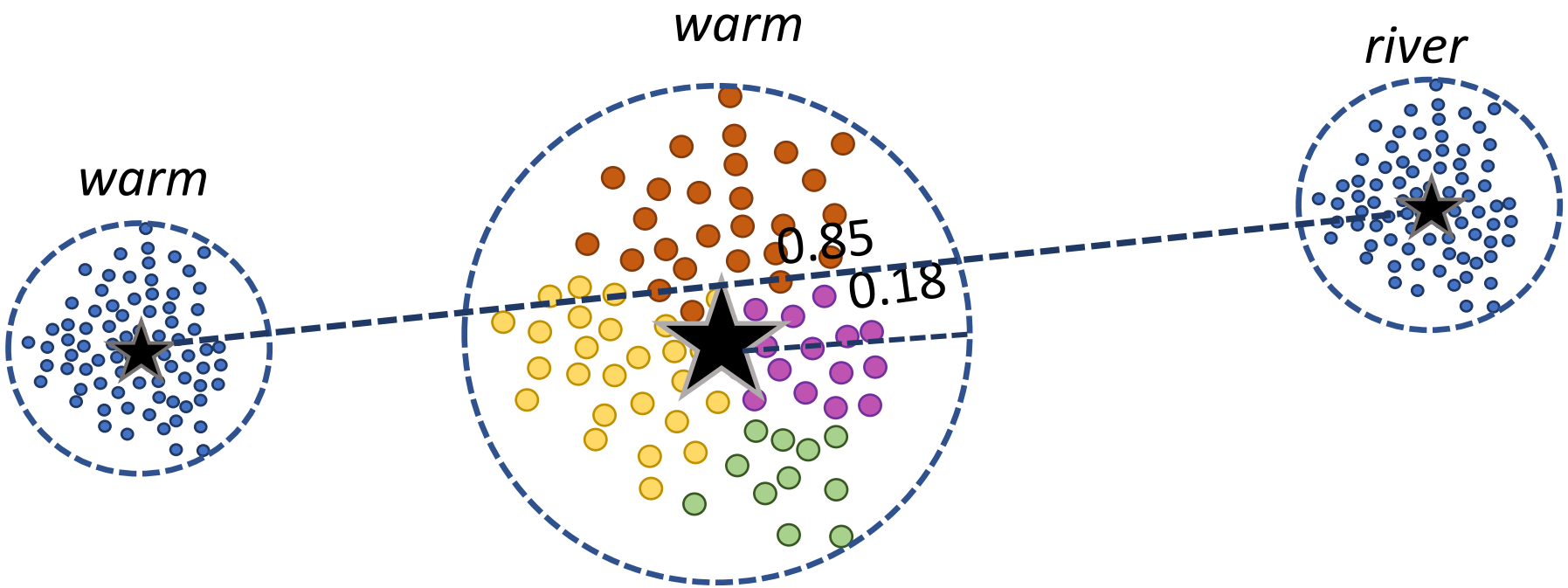


Utilizing Lexical Anchors for Alignment



Geometry of the Contextual Space

- Contextual representation of the same token are clustered together
- The average distance between tokens is larger than within each token

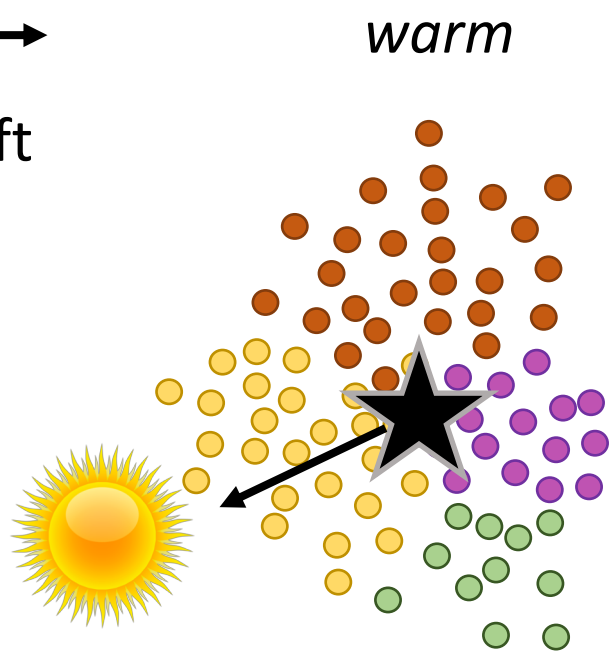


Factorizing the Contextual Embedding

$$\boldsymbol{e}_{i,c} = \bar{\boldsymbol{e}}_i + \hat{\boldsymbol{e}}_{i,c}$$

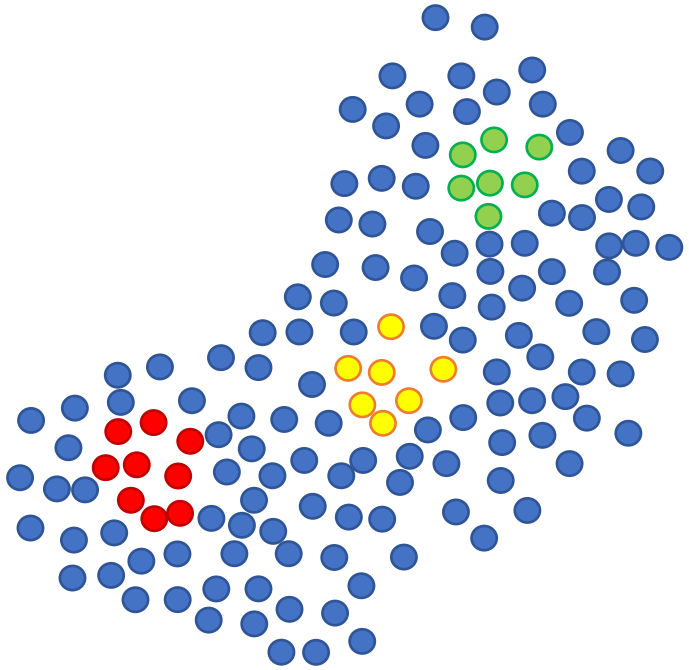
Anchor + Shift

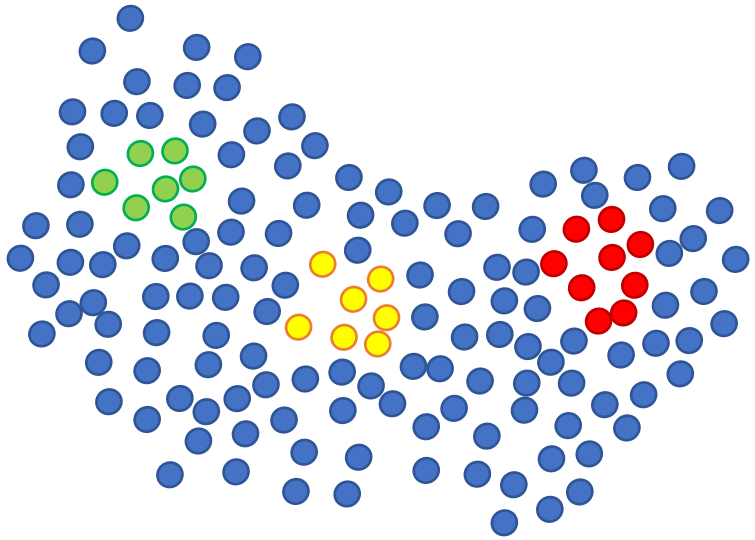
$$\bar{\boldsymbol{e}}_i = \mathbb{E}_c[e_{i,c}]$$



Anchor Based Alignment

A. Train ELMo model per language

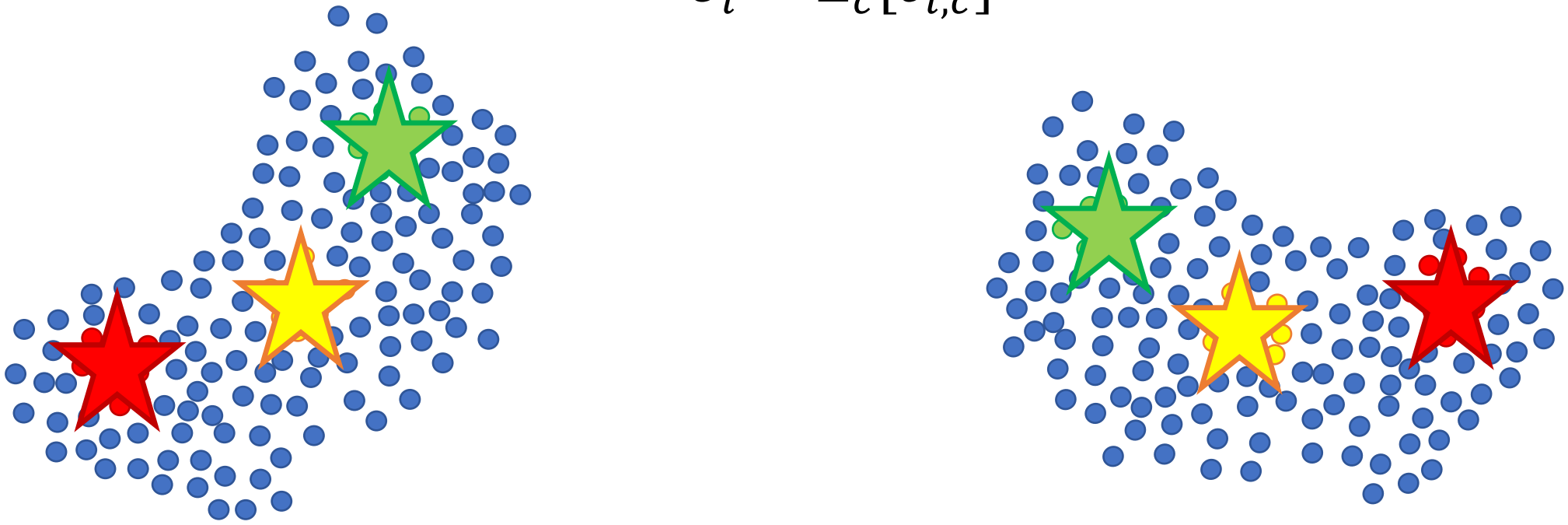




Anchor Based Alignment

B. Extract anchors

$$\bar{\boldsymbol{e}}_i = \mathbb{E}_c[e_{i,c}]$$

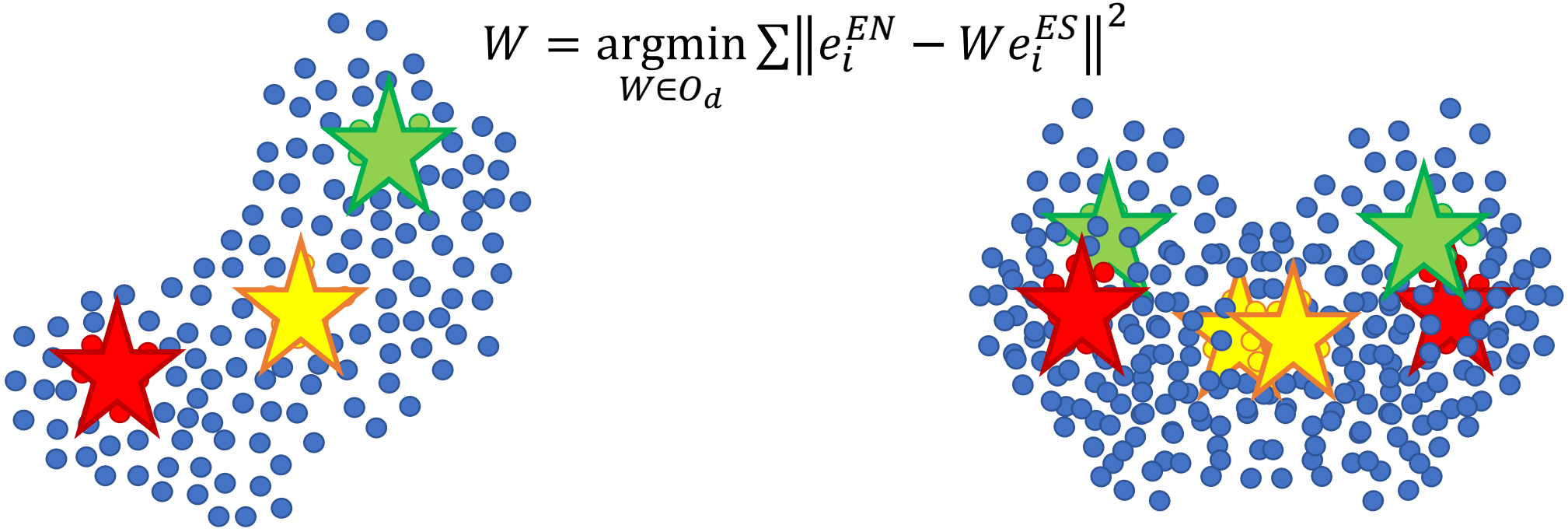


Anchor Based Alignment

Dictionary	
river	río
less	menos
…	…

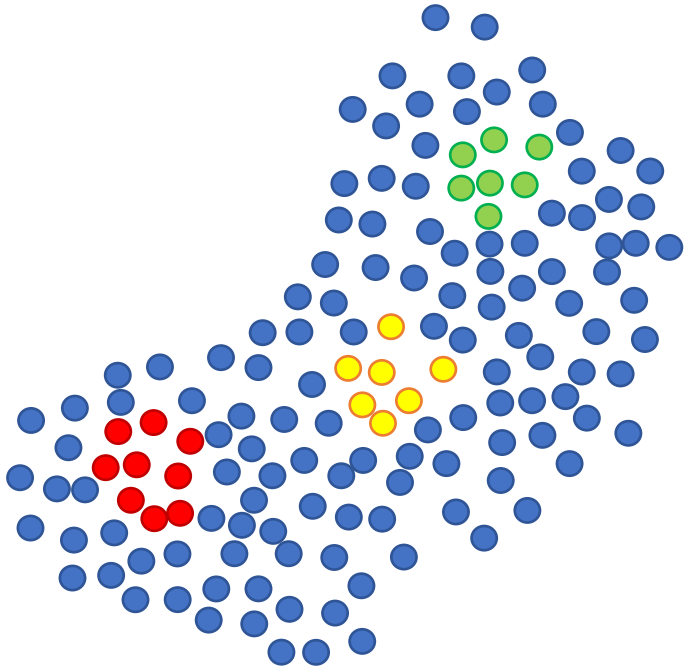
C. Compute alignment by anchors

$$W = \underset{W \in O_d}{\operatorname{argmin}} \sum \left\| e_i^{EN} - W e_i^{ES} \right\|^2$$



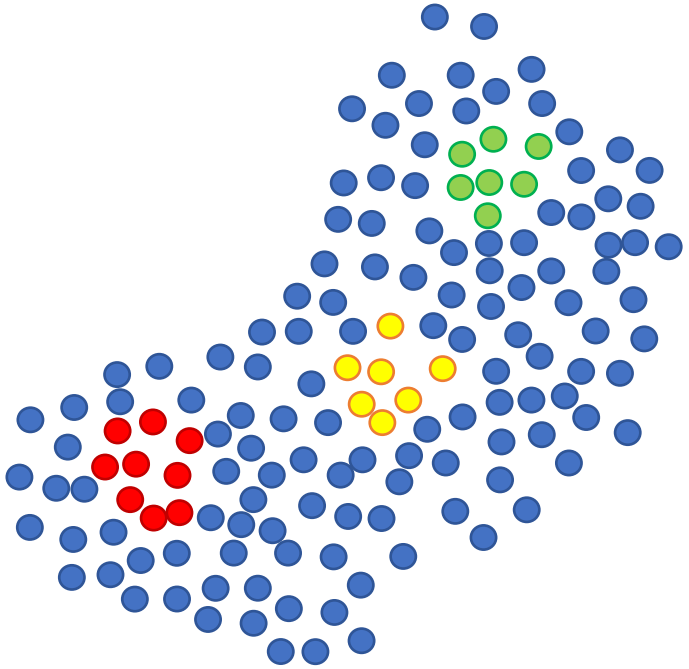
Anchor Based Alignment

D. Apply alignment on contextual space



$$\boldsymbol{e}_{i,c}^{EN} = W\boldsymbol{e}_{i,c}^{ES}$$

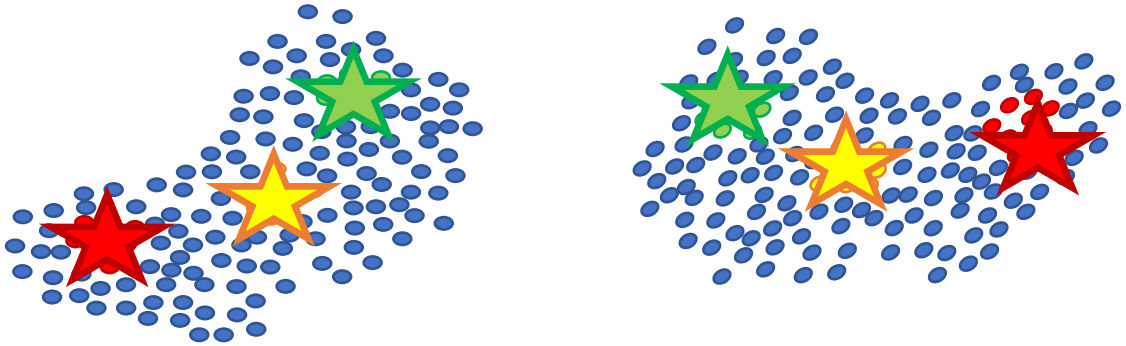
$$= W(\bar{\boldsymbol{e}}_i + \hat{\boldsymbol{e}}_{i,c})$$



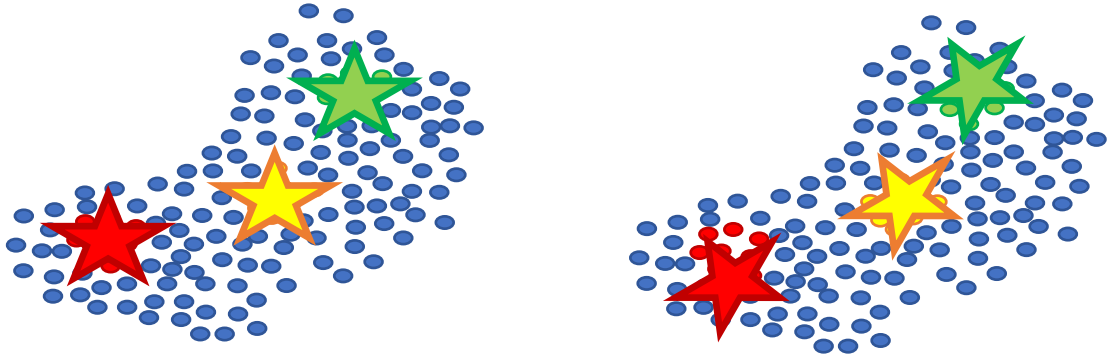
Anchor Based Alignment

A. Train ELMo model per language

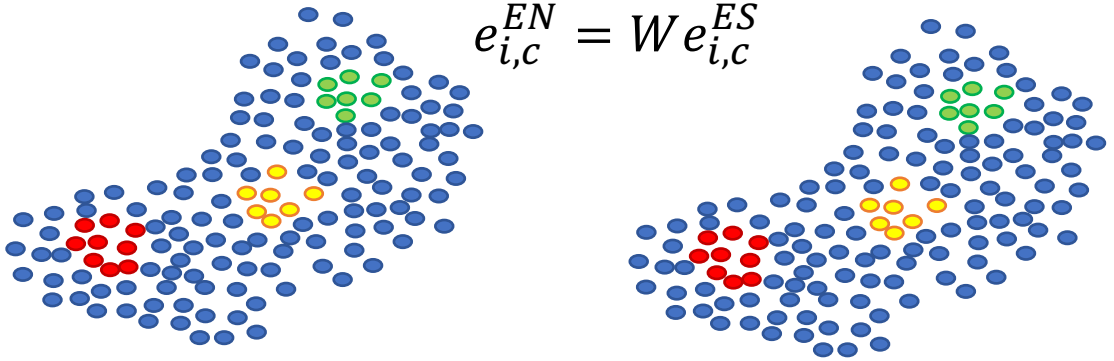
B. Extract anchors $\bar{\boldsymbol{e}}_i = \mathbb{E}_c[e_{i,c}]$



C. Align by anchors $W = \underset{W \in O_d}{\operatorname{argmin}} \sum \left\| e_i^{EN} - W e_i^{ES} \right\|^2$



D. Apply alignment on contextual space $e_{i,c}^{EN} = W e_{i,c}^{ES}$

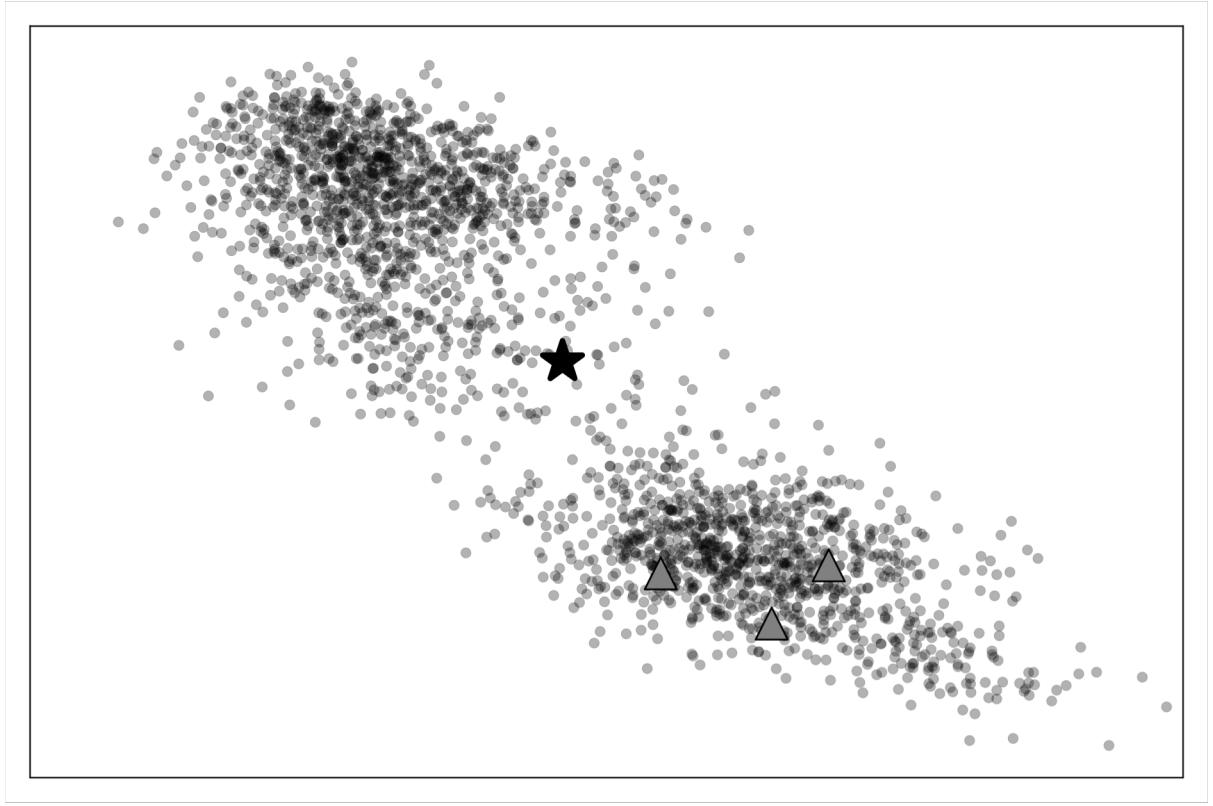


Potential Problem: Multi-sense Words

- Contextual embeddings of the word "*bear*":

bear *her name*

bear *the pain*



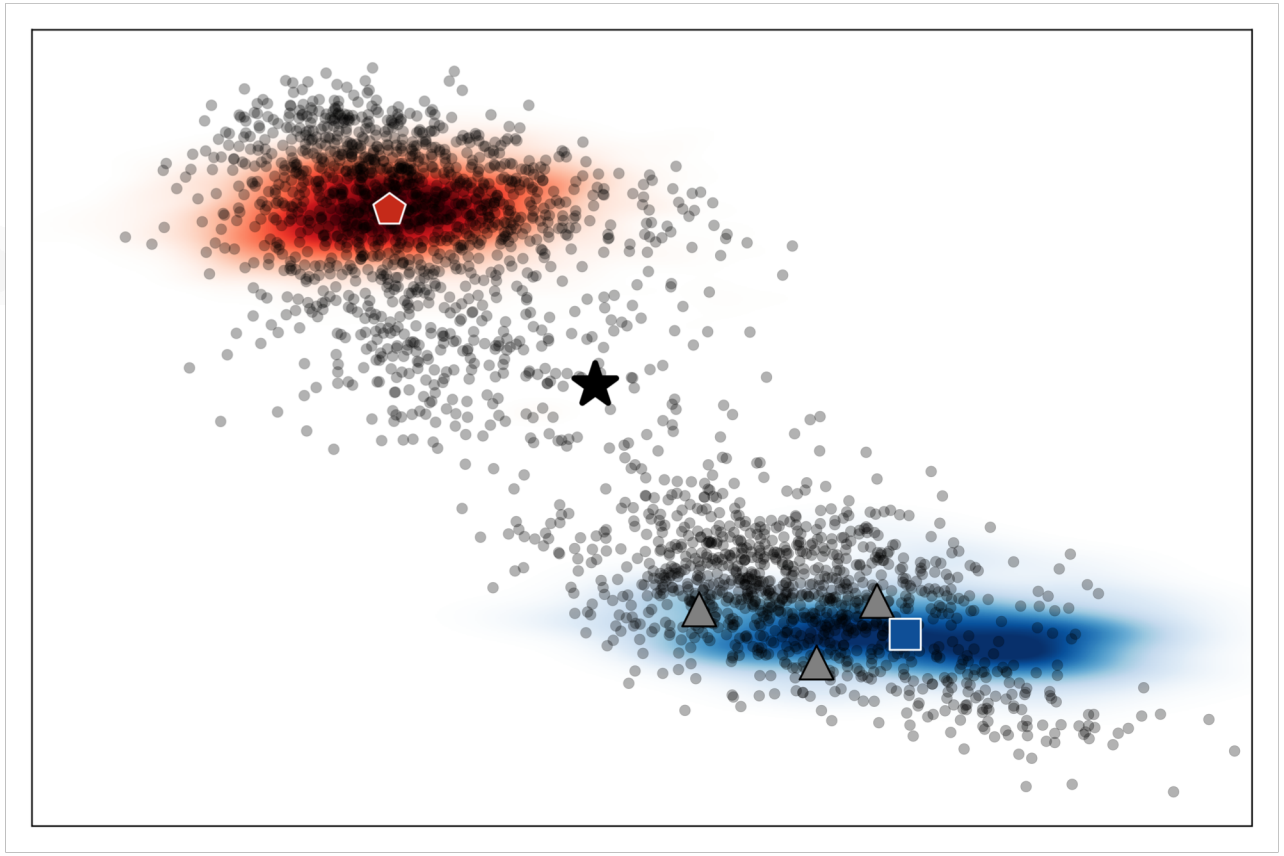
polar ***bear*** *cub*

teddy ***bear***

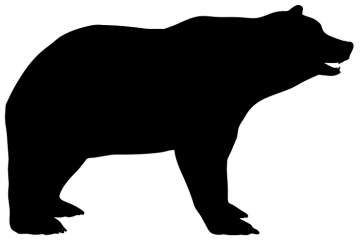
The Alignment Works for Multi-sense Words

tener

bear



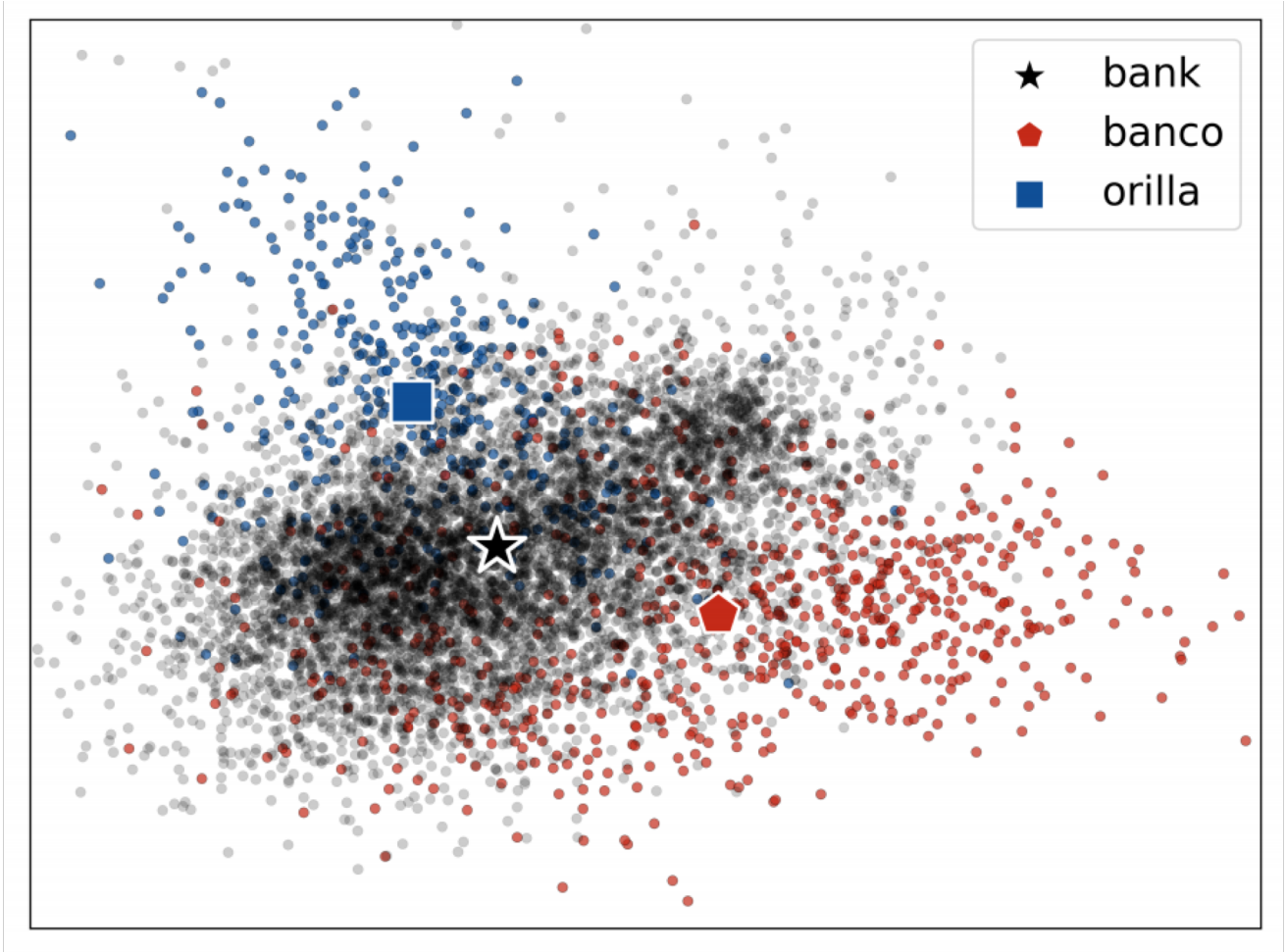
oso



The Alignment Works for Multi-sense Words

bank *of the river*

eastern ***bank*** *of …*



soil seed ***bank***

battery ***bank***

clue ***bank***

No Dictionary

English

- WIKIPEDIA

ELMo embeddings

- POS tags

Spanish

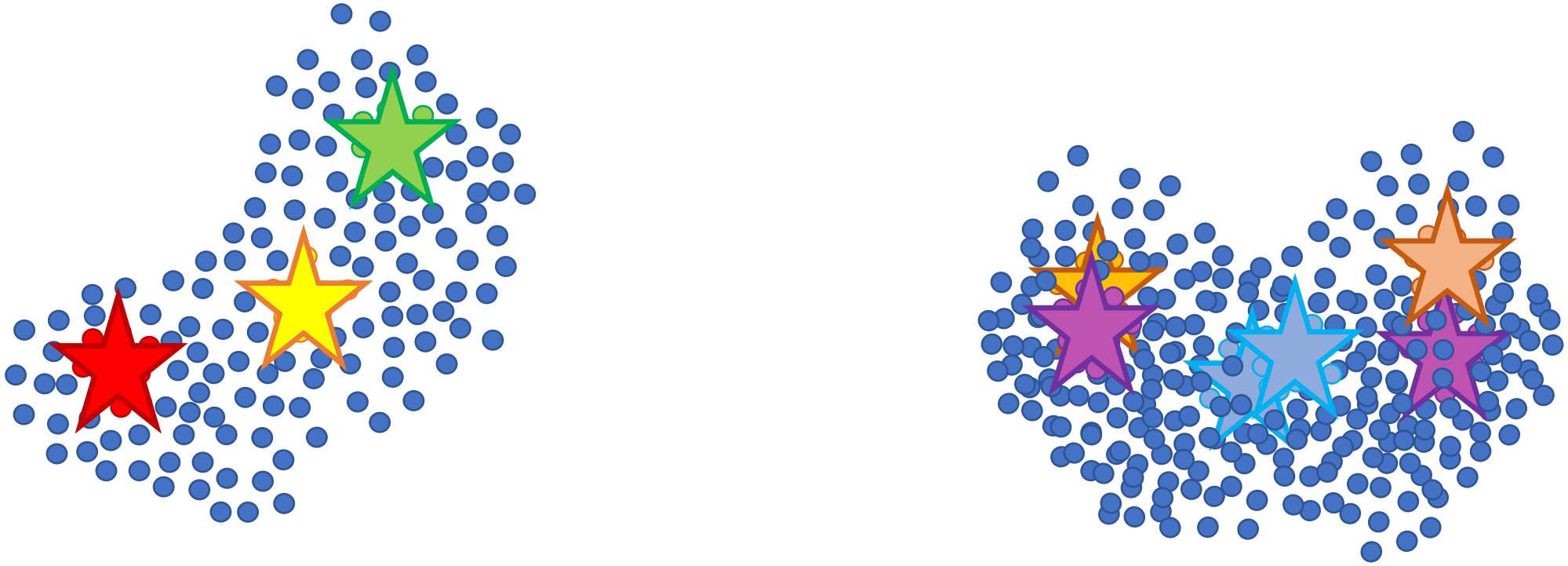
- WIKIPEDIA

ELMo embeddings

- POS tags

Anchor Based Alignment - Unsupervised

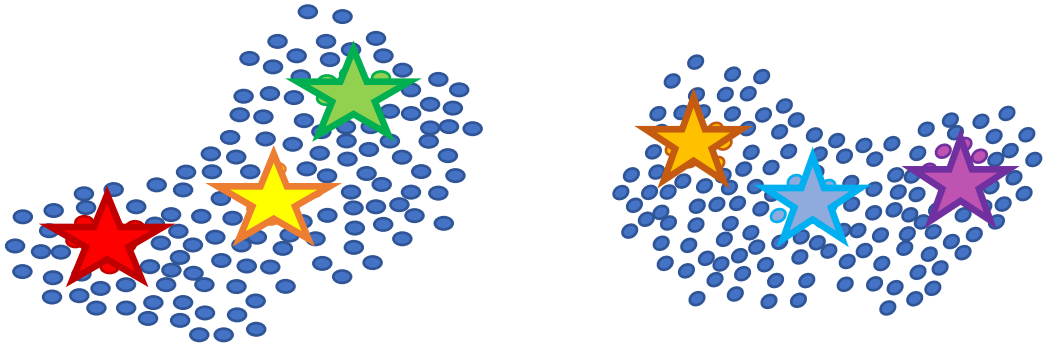
Compute alignment by anchors via **adversarial training**



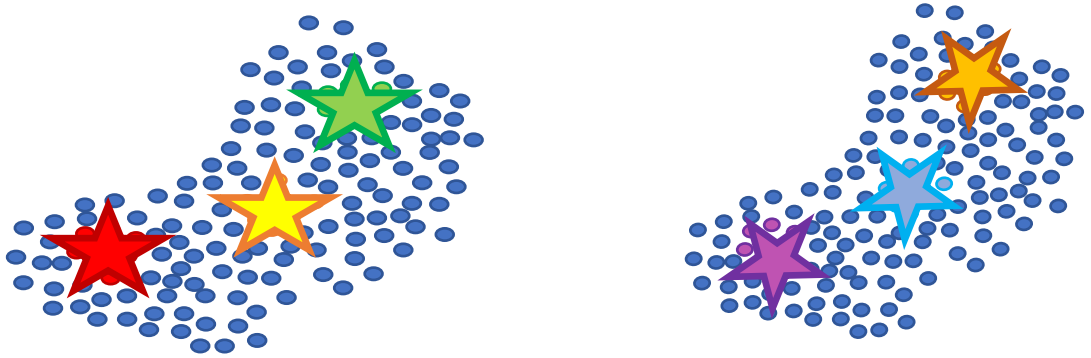
Anchor Based Alignment - Unsupervised

A. Train ELMo model per language

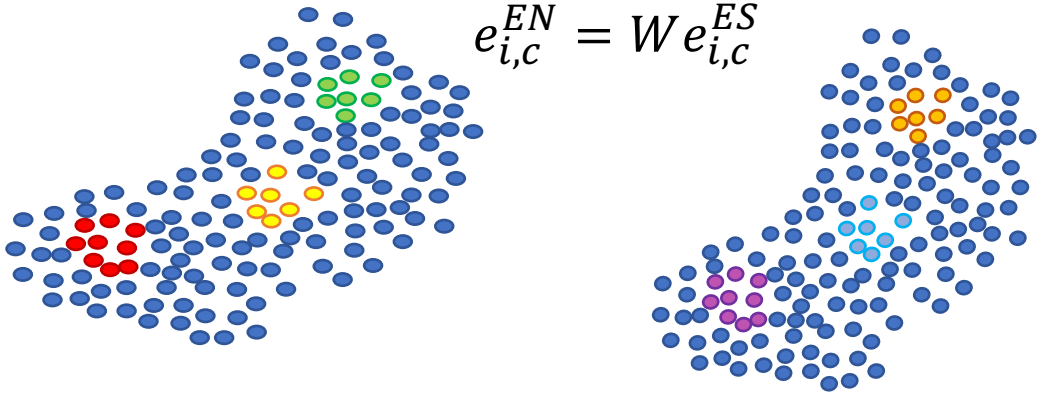
B. Extract anchors $\bar{\boldsymbol{e}}_i = \mathbb{E}_c[e_{i,c}]$



C. Align by anchors – adversarial training



D. Apply alignment on contextual space $e_{i,c}^{EN} = W e_{i,c}^{ES}$



Low Resource Languages

WIKIPEDIA articles per language

English	2.5M
German	800k
Spanish	400k
Turkish	100k
Kazakh	3k

Low Resource Languages

- WIKIPEDIA

ELMo embeddings

- ~~POS tags~~

Dictionary

bear	oso
warm	cálido
…	…

Small

- WIKIPEDIA

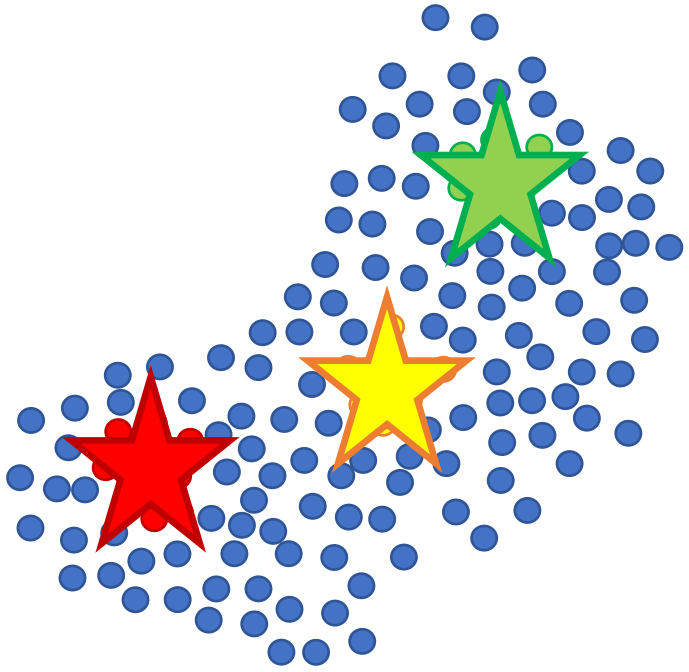
Deficient ELMo embeddings

- ~~POS tags~~

Goal: Alignment (W) and improve the embeddings

Anchored Language Model

A. Extract anchors from English model

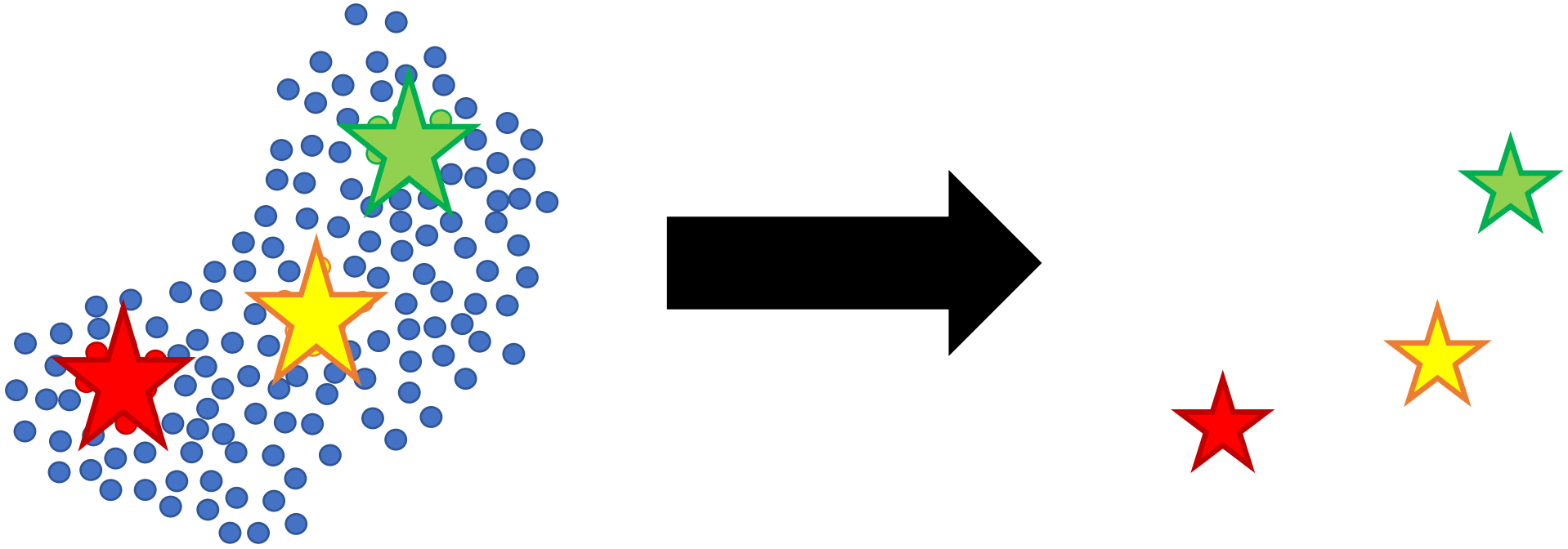


$$\bar{\boldsymbol{e}}_i = \mathbb{E}_c[e_{i,c}]$$

Anchored Language Model

B. Use anchors as seeds for the low resource language

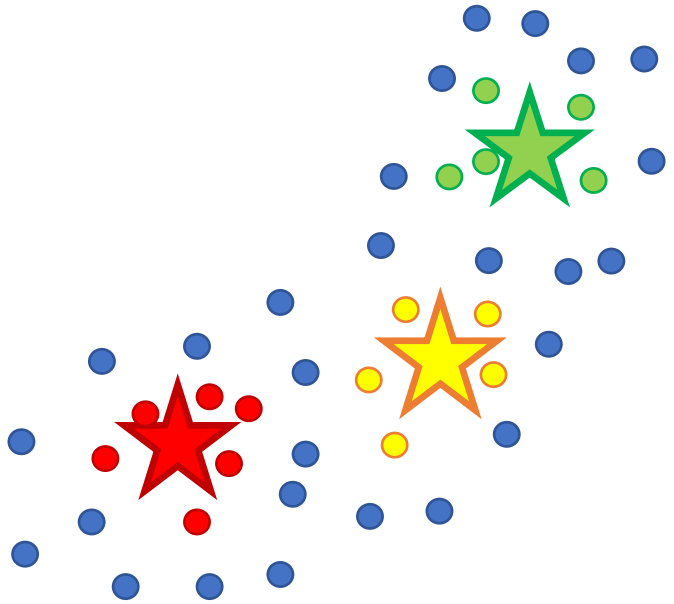
Dictionary	
river	río
less	menos
…	…



Anchored Language Model

C. Learn language model for low resource language

$$\left\| e_{i,c} - \bar{e}_i \right\|^2$$

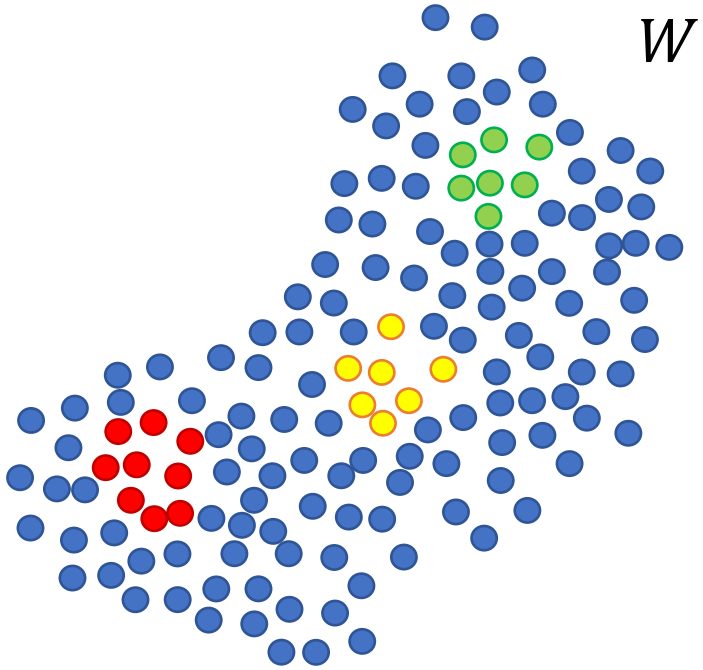


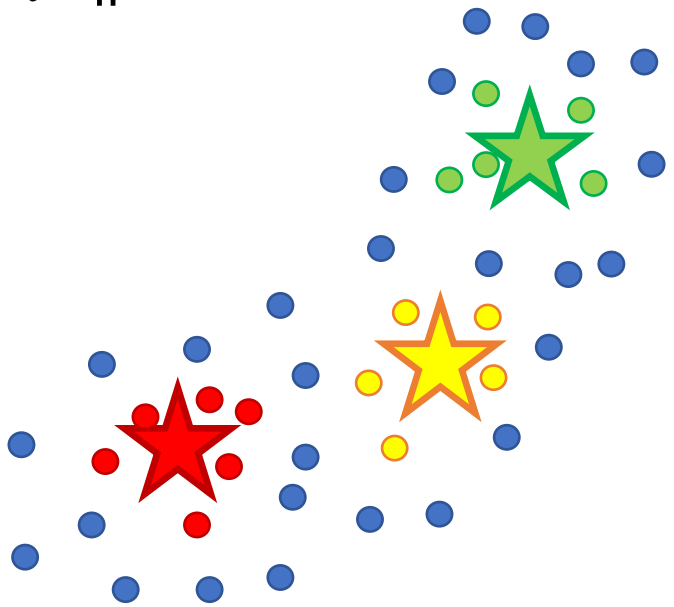
Anchored Language Model

Dictionary	
river	río
less	menos
…	…

D. Learn and apply finer alignment

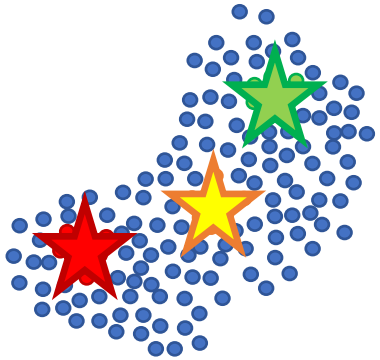
$$W = \underset{W \in O_d}{\operatorname{argmin}} \sum \left\| e_i^{EN} - W e_i^{ES} \right\|^2$$



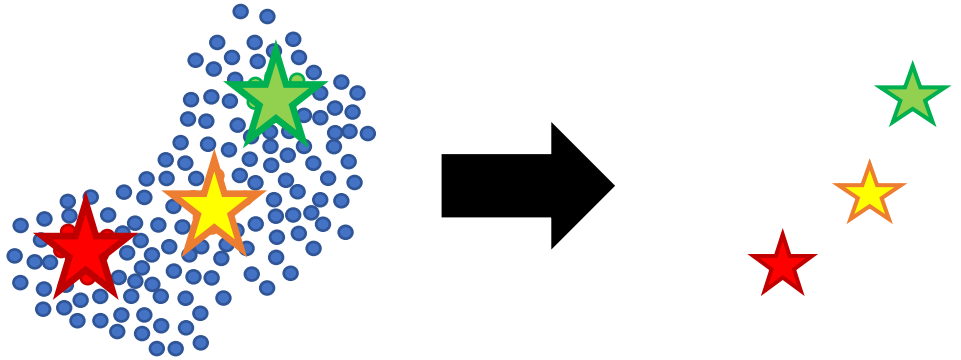


Anchored Language Model

A. Extract anchors from English model

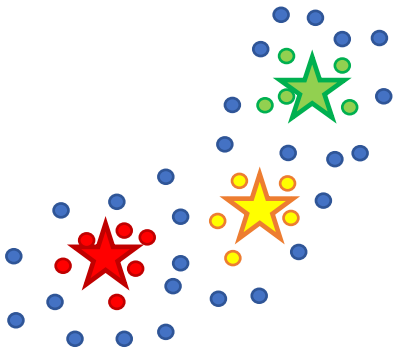


B. Learn language model for low resource language



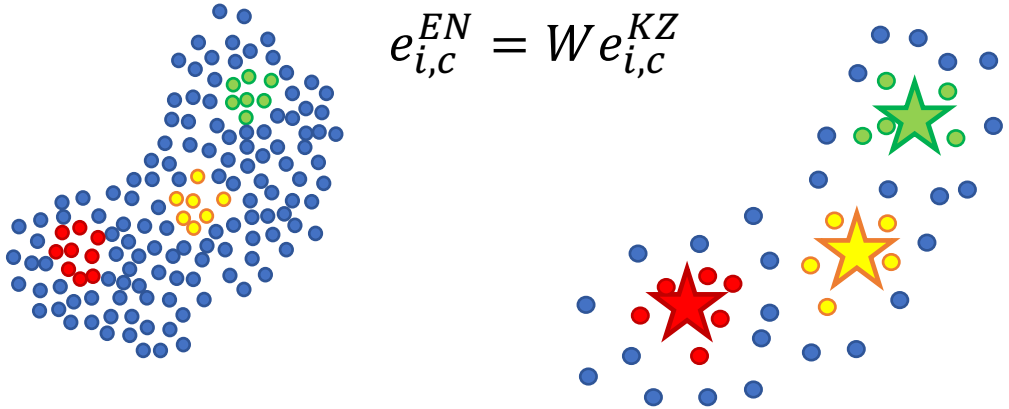
C. Learn language model for low resource language

$$\left\| e_{i,c} - \bar{e}_i \right\|^2$$



D. Learn and apply finer alignment

$$e_{i,c}^{EN} = W e_{i,c}^{KZ}$$



Related Work

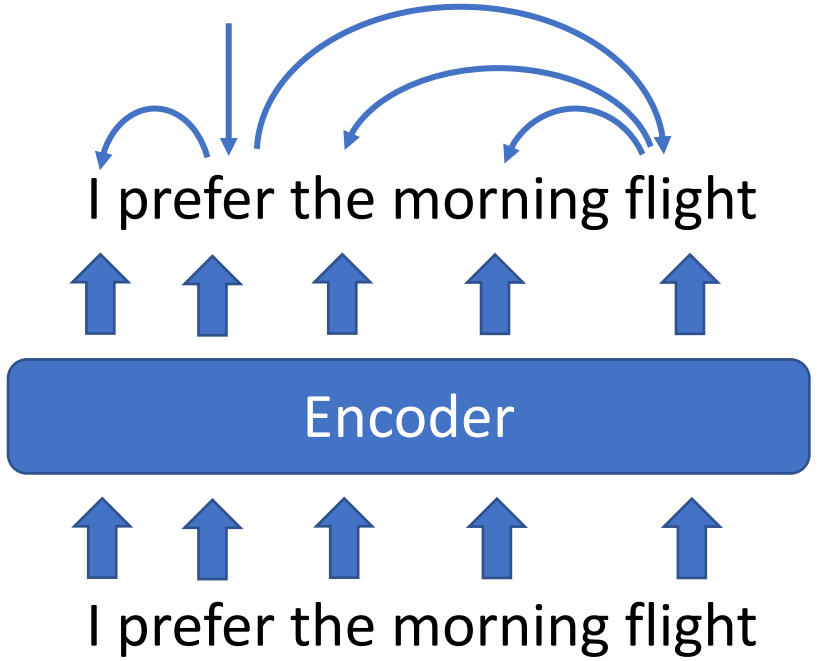
- Contextual embeddings *(Peters et al., 2018; McCann et al., 2017; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2018)*

- Cross-lingual alignment *(Mikolov et al., 2013; Smith et al., 2017; Artetxe et al., 2017; Conneau et al., 2018)*

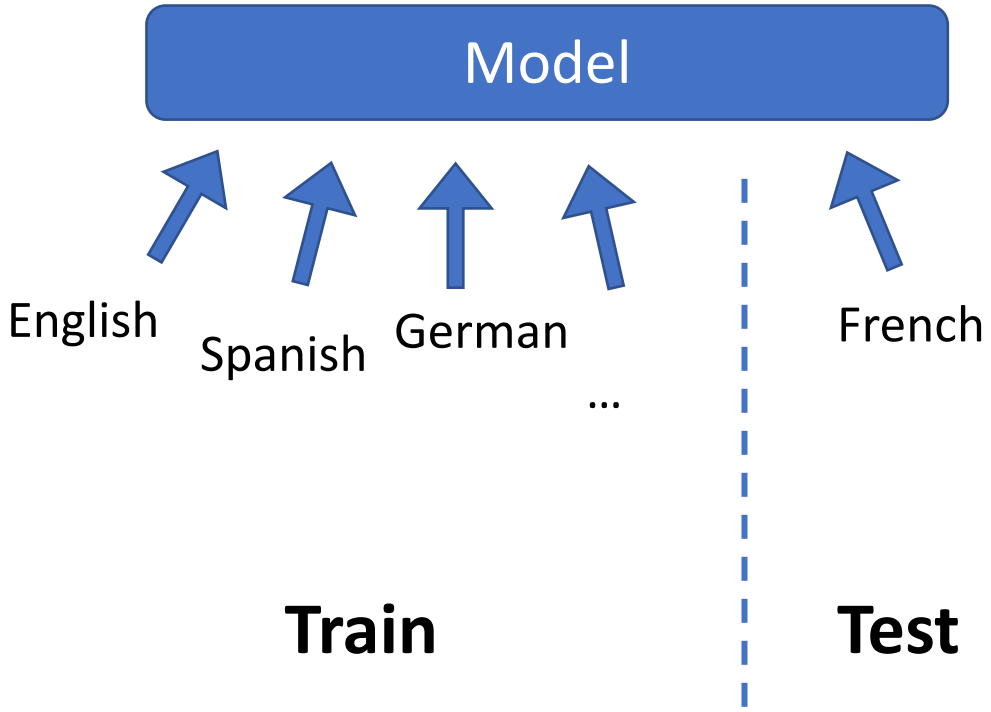
- Multilingual parsing *(Duong et al., 2015; Guo et al., 2016; Ammar et al., 2016; de Lhoneux et al., 2018; Che et al., 2018; Wang et al., 2018; Clark et al., 2018)*

Dependency Parsing

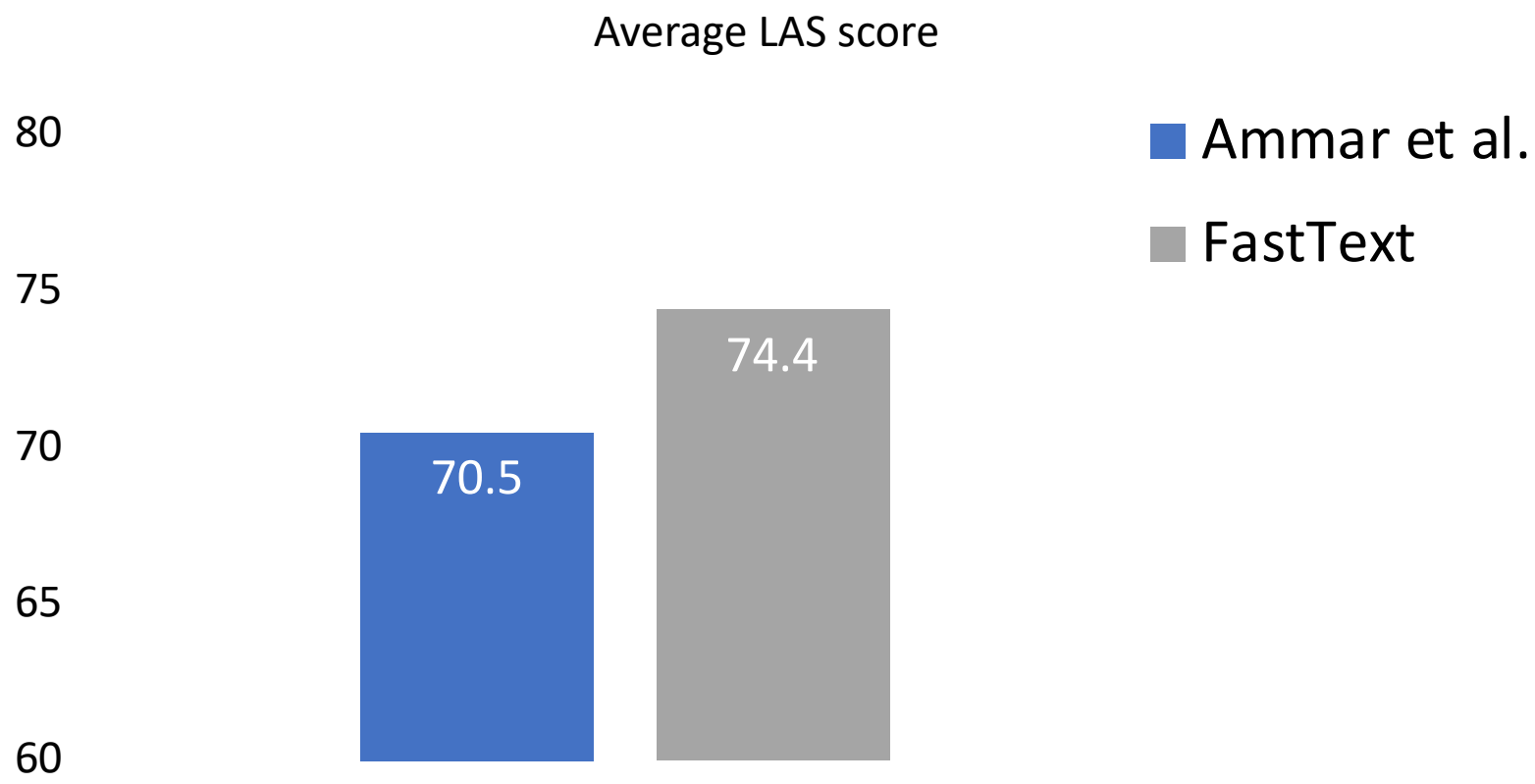
- first-order graph-based model (Dozat and Manning, 2017)



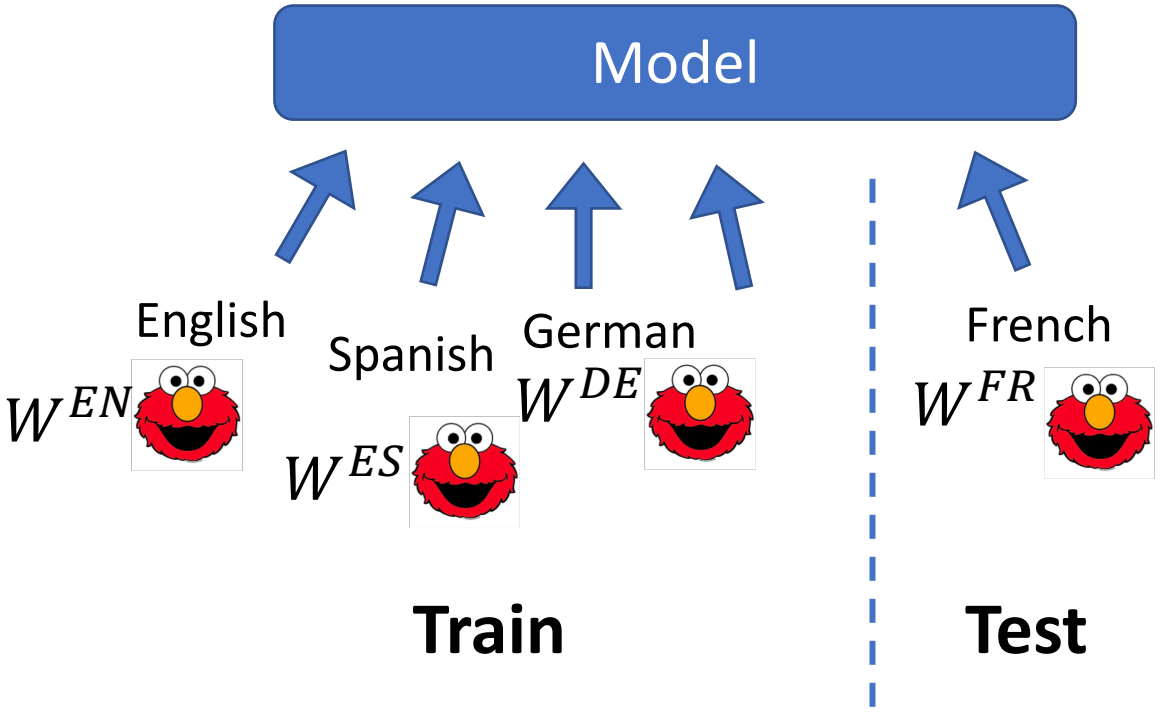
Cross-lingual Zero-shot



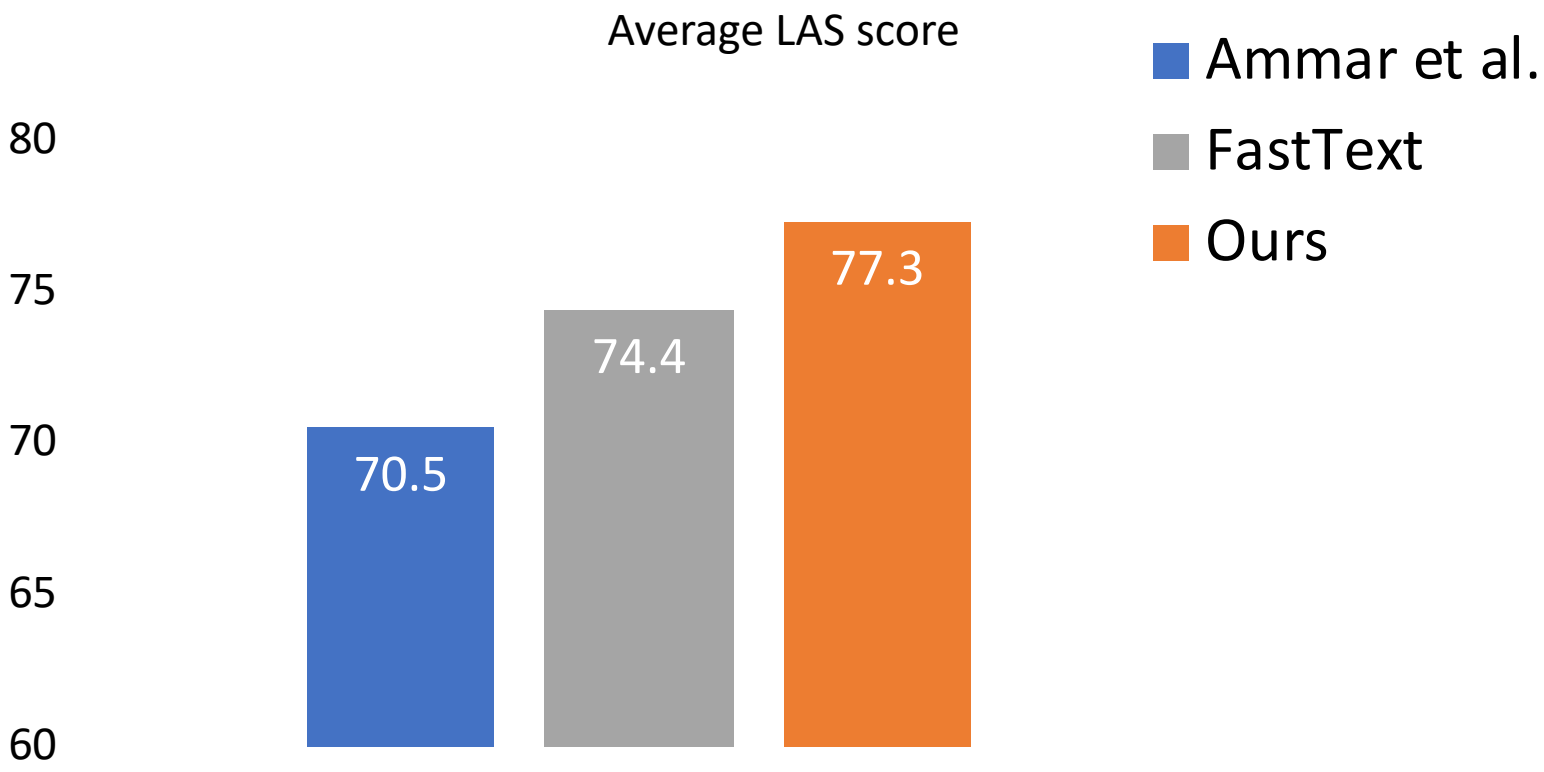
Cross-lingual Transfer for Dependency Parsing



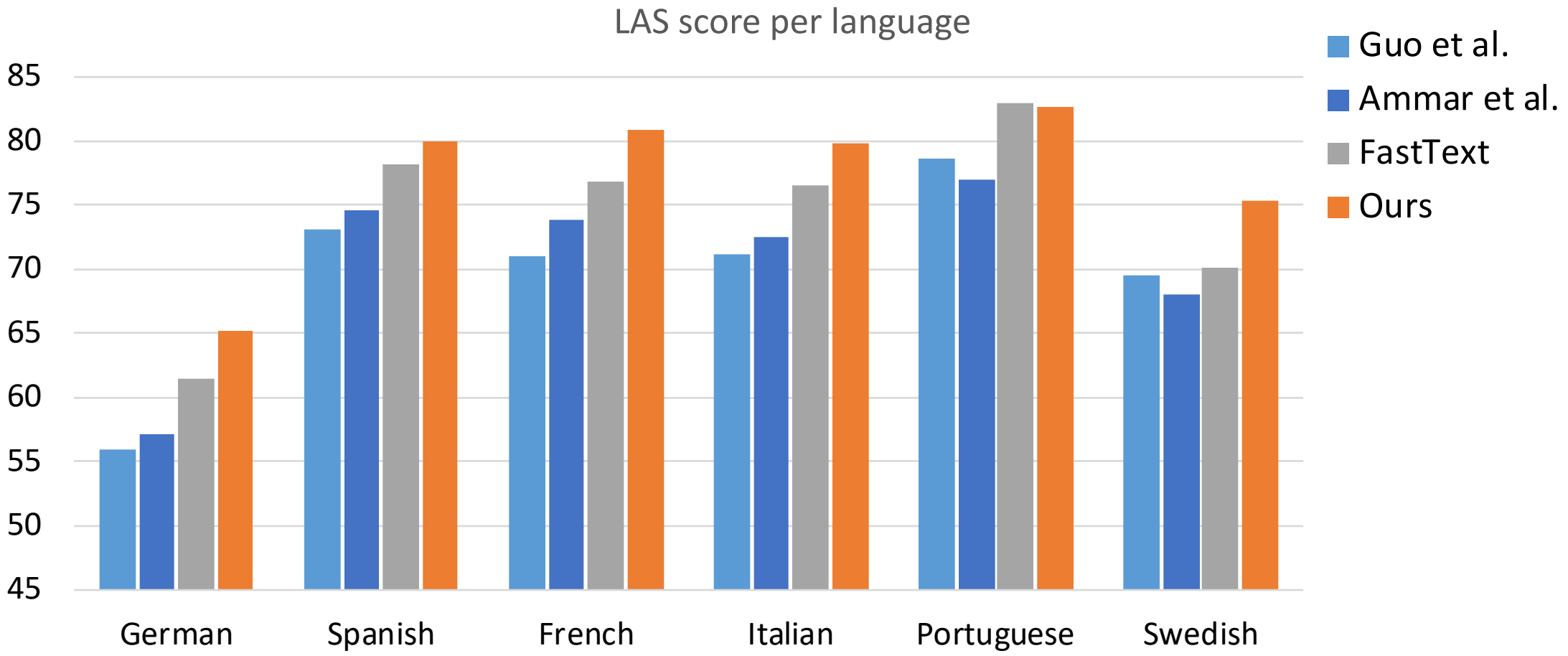
Cross-lingual Zero-shot



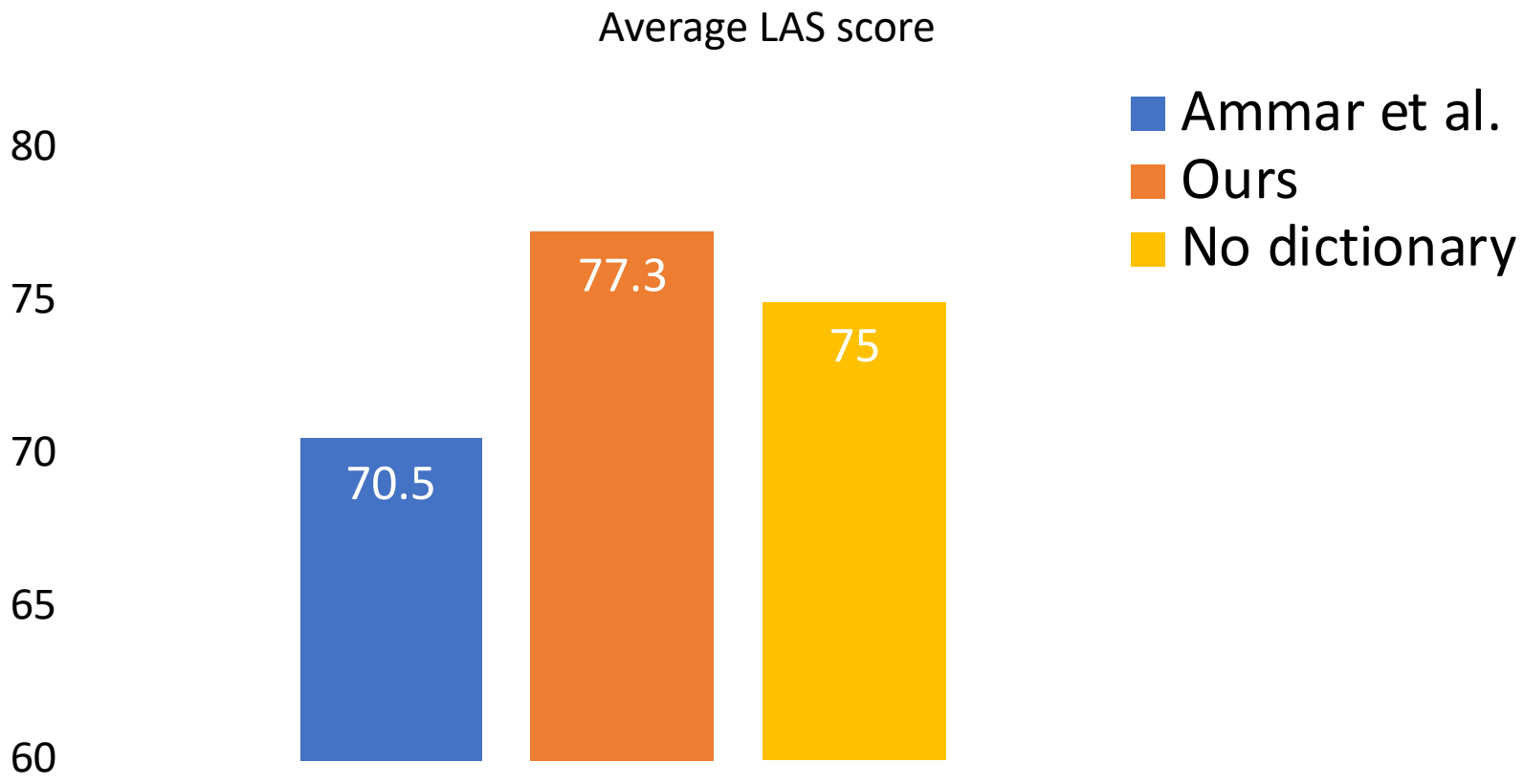
Cross-lingual Transfer for Dependency Parsing



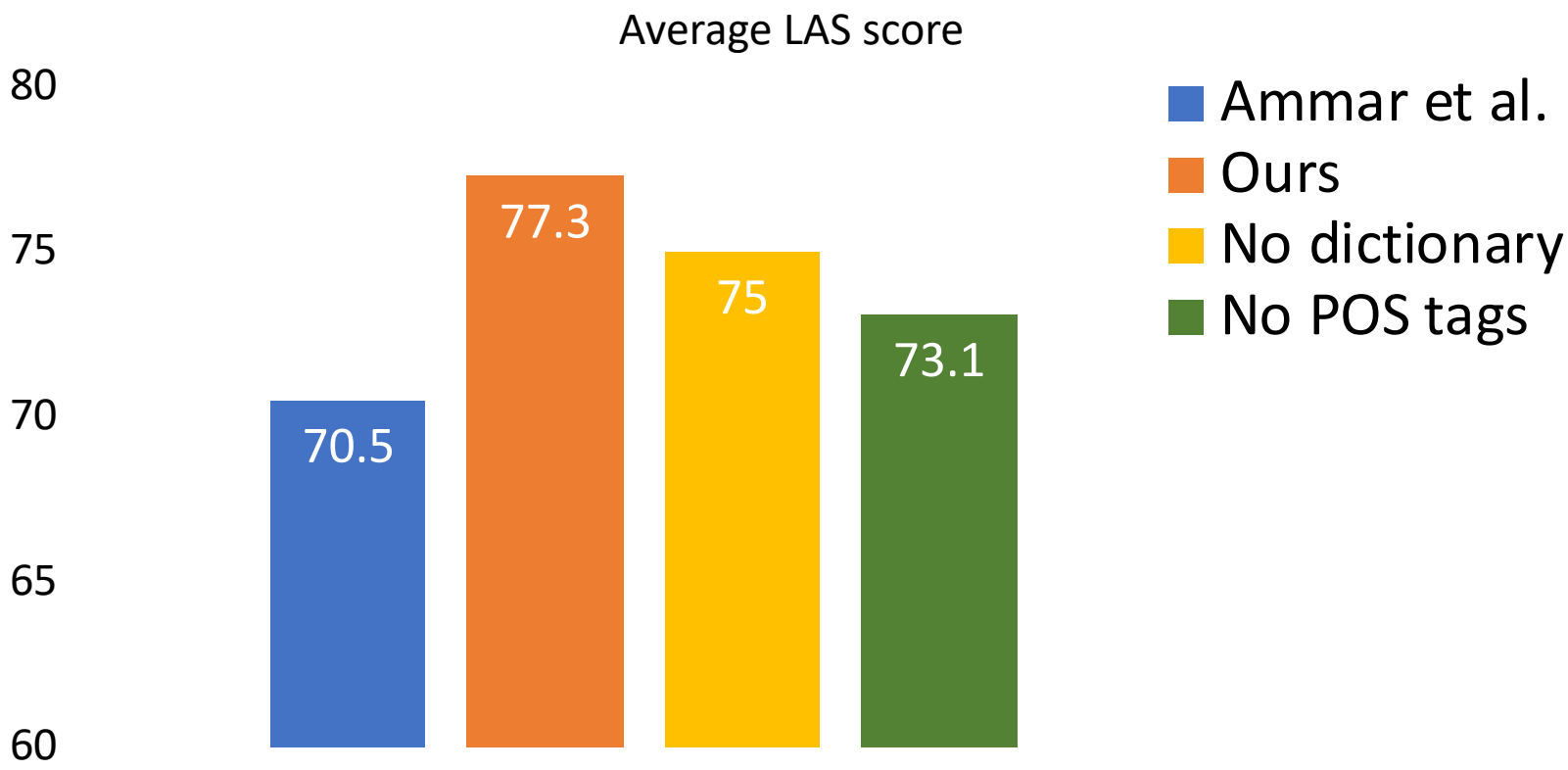
Cross-lingual Transfer for Dependency Parsing



Cross-lingual Transfer for Dependency Parsing



Cross-lingual Transfer for Dependency Parsing



Low Resource Language

English

- WIKIPEDIA

ELMo embeddings

- ~~POS tags~~

Dictionary

bear	oso
warm	cálido
…	…

10k sentences (vs. 28M)

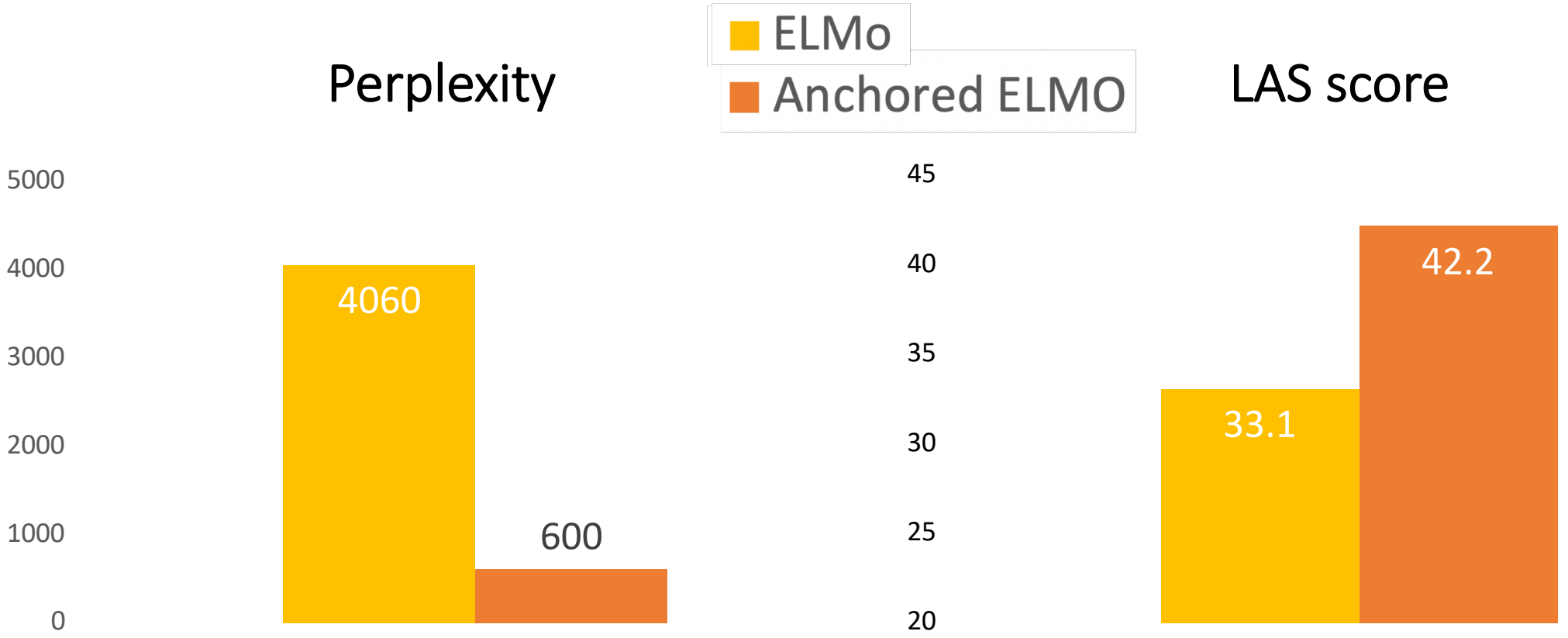
Small

- WIKIPEDIA

Deficient ELMo embeddings

- ~~POS tags~~

Low Resource Language (10k sentences)



Conclusions

★ ELMo embeddings are clustered around their anchor

- Anchor based alignment preserves the contextual component

- Effective for cross-lingual transfer learning (not task-specific)

Code available at:
https://github.com/TalSchuster/CrossLingualELMo
https://github.com/TalSchuster/allennlp-MultiLang
(soon part of the AllenNLP repo)