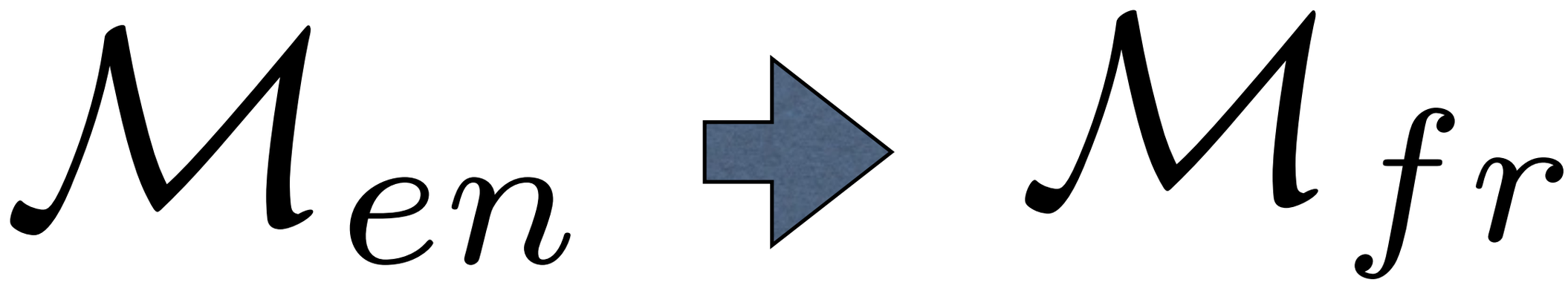
Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing

Tal Schuster[*,1], **Ori Ram**[*,2], **Regina Barzilay**[1], **Amir Globerson**[2]

[1]Computer Science and Artificial Intelligence Lab, MIT
[2]Tel Aviv University

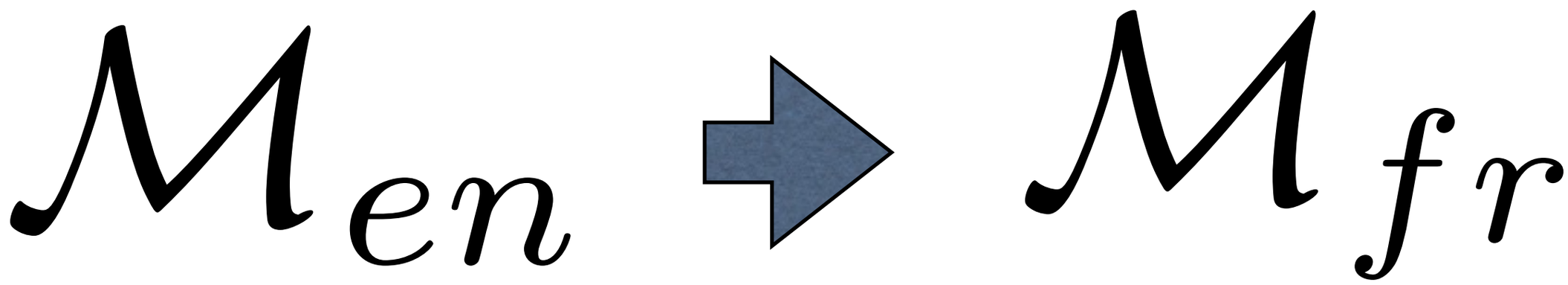
{tals, regina}@csail.mit.edu, {ori.ram, gamir}@cs.tau.ac.il

Task: Transfer Parsing



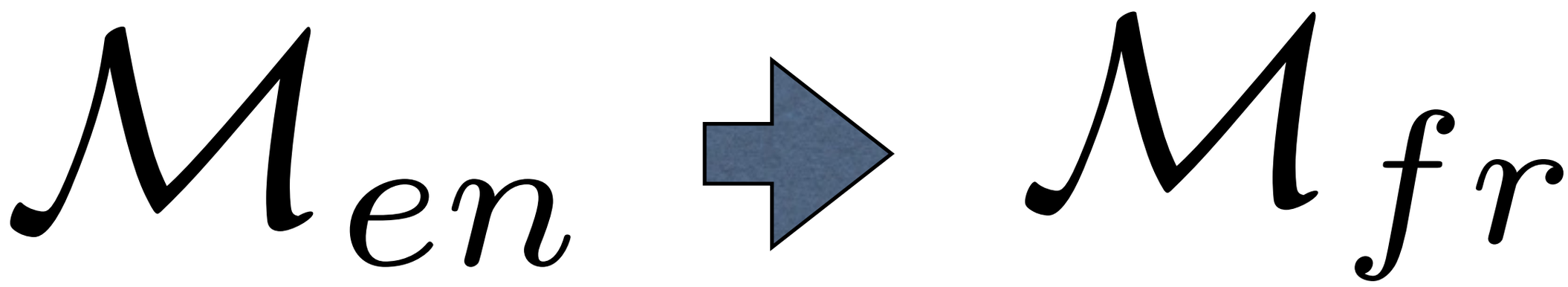
- Word order
- Lexical information

Task: Transfer Parsing



- Word order
- Lexical information

Lexicalized Transfer Parsing



- Monolingual word embedding
- Align to the same vector space
- => multilingual word embedding

Lexicalized Transfer Parsing

$$\mathcal{M}_{en} \Rightarrow \mathcal{M}_{fr}$$

- Multilingual word embedding?
- Current success in ELMo and BERT tells us that contextual word embedding is much better.
- But how ?

Aligning Contextual Word Em..

- Our goal:
  - point clouds are well separated.

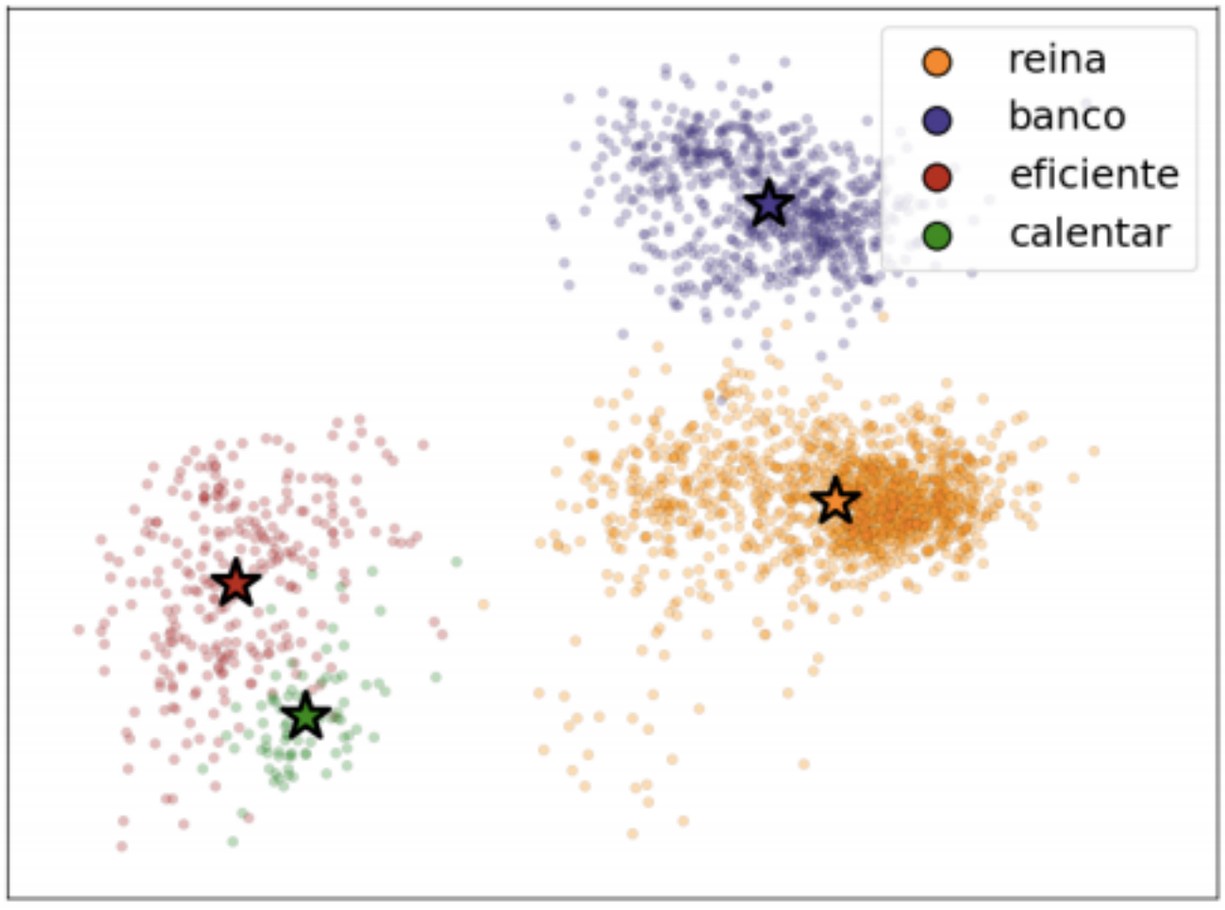


Figure 1: A two dimensional PCA showing examples of contextual representations for four Spanish words. Their corresponding anchors are presented as a star in the same color. (best viewed in color)

Aligning Contextual Word Em..

- From contextual vectors to anchor word vector

contextual words

$$\mathbf{e}_{i,c}$$

center word

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

- Embedding alignment

$$\boldsymbol{e}_{i,c}^{s\to t} = W^{s\to t}\boldsymbol{e}_{i,c}^{s}$$

How about Word Sense?

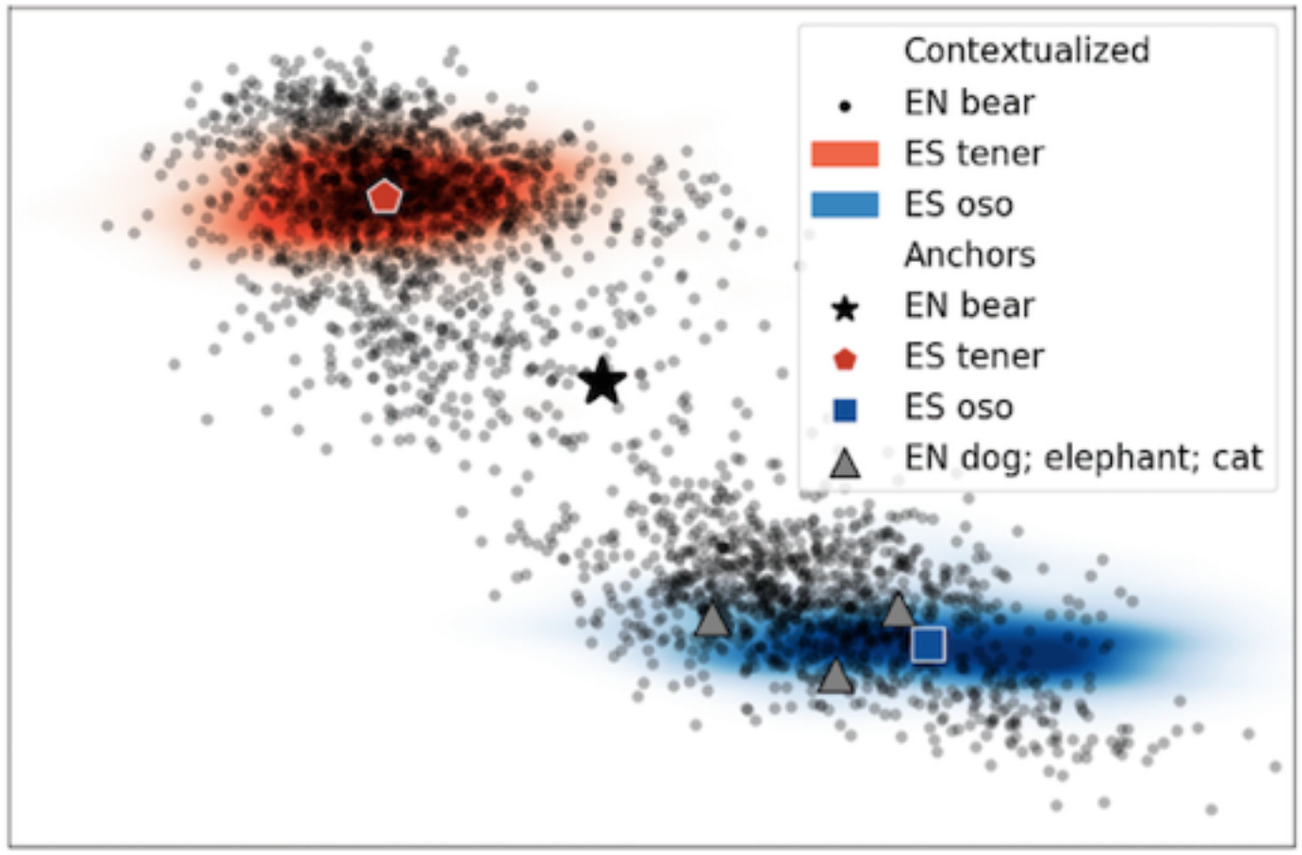


Figure 2: Contextual embeddings for the English word "bear" and its two possible translations in Spanish — "oso" (animal) in blue and "tener" (to have) in red. The figure shows a two dimensional PCA for the aligned space of the two languages. The symbols are the anchors, the clouds represent the distribution of the contextualized Spanish words, and the black dots are for contextualized embeddings of "bear". The gray colored triangles show the anchors of the English words "dog", "elephant", "cat", from left to right respectively.

How to Learn Parameters?

- For matrix W,

$$W^{s \to t} = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| W \boldsymbol{e}_i^s - \boldsymbol{e}_i^t \right\|^2$$

- For contextual vectors from language models,

ELMo loss + $\lambda_{\text{anchor}} \cdot \sum_i \| \boldsymbol{v}_i^s - \boldsymbol{v}_{D(i)}^t \|_2^2$

Experiments

- Base parser: Biaffine

- ELMo vectors are obtained by,

$$\boldsymbol{e}_{i,s}^{\ell \to J} = W^{\ell \to J} \boldsymbol{e}_{i,s}$$

- Transfer from english language to six languages.

Experiments

MODEL	DE	ES	FR	IT	PT	SV	AVERAGE
Zhang and Barzilay (2015)	54.1	68.3	68.8	69.4	72.5	62.5	65.9
Guo et al. (2016)	55.9	73.1	71.0	71.2	78.6	69.5	69.9
Ammar et al. (2016)	57.1	74.6	73.9	72.5	77.0	68.1	70.5
ALIGNED FASTTEXT	61.5	78.2	76.9	76.5	**83.0**	70.1	74.4
ALIGNED $\bar{e}$	58.0	76.7	76.7	76.1	79.2	71.9	73.1
OURS	**65.2**	**80.0**	**80.8**	**79.8**	82.7	**75.4**	**77.3**
OURS, NO DICTIONARY	64.1	77.8	79.8	79.7	79.1	69.6	75.0
OURS, NO POS	61.4	77.5	77.0	77.6	73.9	71.0	73.1
OURS, NO DICTIONARY, NO POS	61.7	76.6	76.3	77.1	69.1	54.2	69.2

Table 3: Zero-shot cross lingual LAS scores compared to previous methods, for German (DE), Spanish (ES), French (FR), Italian (IT), Portuguese (PT) and Swedish (SV). Aligned FASTTEXT and $\bar{e}$ context-independent embeddings are also presented as baselines. The bottom three rows are models that don't use POS tags at all and/or use an unsupervised anchored alignment. Corresponding UAS results are provided in App. B.

Experiments

Alignment Method	DE	ES	FR	IT	PT	SV	Average
Supervised Anchored	78	85	86	82	74	68	79
Unsupervised Anchored	63	61	70	58	35	22	52
+ Refine	72	74	**81**	**77**	53	**33**	**65**
Unsupervised Context-based	57	68	59	57	53	*	49
+ Refine	**73**	**82**	77	73	**66**	*	62

Table 2: Word translation to English precision @5 using CSLS (Conneau et al., 2018a) with a dictionary (supervised) and without (unsupervised) for German (DE), Spanish (ES), French (FR), Italian (IT), Portuguese (PT) and Swedish (SV). Each of the unsupervised results is followed by a line with the results post the anchor-based refinement steps. * stands for 'Failed to converge'.

# Sentences	Language Model	UAS / LAS Dev	UAS / LAS Test	Perplexity Train	Perplexity Dev	Align
28M	ELMo	72.3 / 62.8	72.5 / 61.3	22	44	85
10K	ELMo	52.9 / 38.3	50.1 / 33.1	4	4060	4
	Anchored ELMo	**59.2 / 47.3**	**57.2 / 42.2**	92	600	12

Table 4: Zero-shot, single-source results for the Spanish limited unlabeled data experiments. The parsing results are UAS/LAS scores, the perplexity is of the ELMo model, and the alignment scores are precision@5 on the held-out set, based on CSLS. All embeddings were aligned to English using supervised anchored alignment.