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

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

MIT

TEL AVIV UNIVERSITY
Task: Cross-lingual Zero-shot Dependency Parsing

Goal: Utilize universal space of contextual embeddings
Idea: Align Contextual Word Embeddings
Idea: Align Contextual Word Embeddings

English

warm

Spanish

calentar
cálido
Our Results – zero-shot

By aligning ELMo contextual embeddings
Problem Definition

**English**

- **WIKIPEDIA**
  - ELMo embeddings
  - POS tags

**Spanish**

- **WIKIPEDIA**
  - ELMo embeddings
  - POS tags

**Dictionary**

- bear
- oso
- warm
- cálido

... ... ...

**Goal:** Learn a linear alignment ($W$)
Problem Definition - Extensions

**Goal:** Learn a linear alignment ($W$)

- **English**
  - Wikipedia
  - ELMo embeddings
  - **•** POS tags

- **Spanish**
  - Wikipedia
  - ELMo embeddings
  - **•** POS tags

**Extensions**

**Dictionary**

- bear
- oso
- warm
- caliente

... ...
Problem Definition - Extensions

**English**

- Wikipedia
- ELMo embeddings
- **POS tags**

**Spanish**

- Wikipedia
- Deficient ELMo embeddings
- **POS tags**

**Dictionary**

- bear
- oso
- warm
- cálido
- ...
- ...

**Goal:** Alignment ($W$) and improve the embeddings
Aligning Embeddings - Static Case

\[ e_i^{EN} = W e_i^{ES} \]

\[ W = \arg\min_{W \in O_d} \sum \| e_i^{EN} - W e_i^{ES} \|^2 \]

(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

\[ \bar{e}_i = \mathbb{E}_c[e_{i,c}] \]
Utilizing Lexical Anchors for Alignment

English

Spanish

Cross-Lingual Alignment of Contextual Word Embeddings
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

\[ e_{i,c} = \bar{e}_i + \hat{e}_{i,c} \]

Anchor + Shift

\[ \bar{e}_i = \mathbb{E}_c[e_{i,c}] \]
Anchor Based Alignment

A. Train ELMo model per language
Anchor Based Alignment

B. Extract anchors

\[ \bar{e}_i = \mathbb{E}_c[e_{i,c}] \]
Anchor Based Alignment

C. Compute alignment by anchors

\[
W = \arg\min_{\sum_{W \in O_d}} \sum \| e_i^{EN} - W e_i^{ES} \|^2
\]
Anchor Based Alignment

D. Apply alignment on contextual space

\[ e_{i,c}^{EN} = W e_{i,c}^{ES} \]

\[ = W (\bar{e}_i + \hat{e}_{i,c}) \]
Anchor Based Alignment

A. Train ELMo model per language

B. Extract anchors

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

C. Align by anchors

\[ W = \arg\min_{W \in O_d} \sum_{i} \| e_{i,N}^{EN} - We_{i,S}^{ES} \|^2 \]

D. Apply alignment on contextual space

\[ e_{i,c}^{EN} = We_{i,c}^{ES} \]
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
The Alignment Works for Multi-sense Words

bank of the river
eastern bank of ...

Cross-Lingual Alignment of Contextual Word Embeddings
No Dictionary

**English**

- **WIKIPEDIA**
  - ELMo embeddings
- POS tags

**Spanish**

- **WIKIPEDIA**
  - ELMo embeddings
- POS tags

Cross-Lingual Alignment of Contextual Word Embeddings
Anchor Based Alignment - Unsupervised

Compute alignment by anchors via adversarial training
Anchor Based Alignment - Unsupervised

A. Train ELMo model per language

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

C. Align by anchors – adversarial training

D. Apply alignment on contextual space
   \[ e^{EN}_{i,c} = W e^{ES}_{i,c} \]
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**

- Small

  Dictionary
  - bear
  - oso
  - warm
  - cálido

  ... ... ...

- **WIKIPEDIA**
  - Deficient ELMo embeddings
  - **POS tags**

**Goal**: Alignment ($W$) and improve the embeddings
Anchored Language Model

A. Extract anchors from English model

\[ \bar{e}_i = E_c[e_{i,c}] \]
Anchored Language Model

B. Use anchors as seeds for the low resource language
Anchored Language Model

C. Learn language model for low resource language

$$\| e_{i,c} - \bar{e}_i \|^2$$
Anchored Language Model

D. Learn and apply finer alignment

\[ W = \arg\min_{W \in O_d} \sum_{i} \| e_i^{EN} - W e_i^{ES} \|^2 \]
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

D. Learn and apply finer alignment

\[ \| e_{i,c} - \bar{e}_i \|^2 \]

\[ e_{i,c}^{EN} = W e_{i,c}^{KZ} \]
Related Work


• Cross-lingual alignment \cite{mikolov-etal-2013-effective, smith-etal-2017-crosslingual, artetxe-etal-2017-cross, conneau-etal-2018-advanced}

Dependency Parsing

- first-order graph-based model (Dozat and Manning, 2017)
Cross-lingual Zero-shot
Cross-lingual Transfer for Dependency Parsing

Cross-Lingual Alignment of Contextual Word Embeddings
Cross-lingual Zero-shot
Cross-lingual Transfer for Dependency Parsing

Average LAS score

- Ammar et al.: 70.5
- FastText: 74.4
- Ours: 77.3
Cross-lingual Transfer for Dependency Parsing

LAS score per language

Guo et al.  
Ammar et al.  
FastText  
Ours

Cross-Lingual Alignment of Contextual Word Embeddings
Cross-lingual Transfer for Dependency Parsing

Average LAS score

- Ammar et al.
- Ours
- No dictionary

Cross-Lingual Alignment of Contextual Word Embeddings
Cross-lingual Transfer for Dependency Parsing

Average LAS score

- Ammar et al.
- Ours
- No dictionary
- No POS tags

Cross-Lingual Alignment of Contextual Word Embeddings
Low Resource Language

**English**

- **WIKIPEDIA**
  - ELMo embeddings
  - **POS tags**

**10k sentences (vs. 28M)**

- **WIKIPEDIA**
  - Small
  - Deficient ELMo embeddings
  - **POS tags**
Low Resource Language (10k sentences)

Cross-Lingual Alignment of Contextual Word Embeddings
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)