Learning bilingual word embeddings with (almost) no bilingual data

Mikel Artetxe, Gorka Labaka, Eneko Agirre

IXA NLP group – University of the Basque Country (UPV/EHU)
Who cares?
Who cares?

word embeddings are useful!
Who cares?

word embeddings are useful!
Who cares?

word embeddings are useful!
Who cares?

bilingual word embeddings are useful!

even more
Who cares?

- inherently crosslingual tasks

bilingual word embeddings are useful!

- inherently crosslingual tasks
Who cares?

Bilingual word embeddings are useful!

- inherently crosslingual tasks
- crosslingual transfer learning
Who cares?

bilingual word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning

bilingual signal for training
Who cares?

bilingual signal for training

bilingual word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning

even more
Who cares?

Previous work

bilingual signal for training

bilingual word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning
Who cares?

Previous work
- parallel corpora

bilingual signal for training

bilingual word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning

even more
Who cares?

bilingual word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning

Previous work
- parallel corpora
- comparable corpora
Who cares?

- inherently crosslingual tasks
- crosslingual transfer learning

Previous work
- parallel corpora
- comparable corpora
- (big) dictionaries

bilingual word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning

bilingual signal for training
Who cares?

bilingual signal for training

- parallel corpora
- comparable corpora
- (big) dictionaries

Even more

bilingual word embeddings are useful!

- inherently crosslingual tasks
- crosslingual transfer learning
Who cares?

- inherently crosslingual tasks
- crosslingual transfer learning

**bilingual**

- word embeddings are useful!
  - inherently crosslingual tasks
  - crosslingual transfer learning

**bilingual signal for training**

**Previous work**
- parallel corpora
- comparable corpora
- (big) dictionaries

**This talk**
Who cares?

- inherently crosslingual tasks
- crosslingual transfer learning

word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning

Previous work
- parallel corpora
- comparable corpora
- (big) dictionaries

This talk
- 25 word dictionary
Who cares?

bilingual word embeddings are useful!
- inherently crosslingual tasks
- crosslingual transfer learning

bilingual signal for training

Previous work
- parallel corpora
- comparable corpora
- (big) dictionaries

This talk
- 25 word dictionary
- numerals (1, 2, 3...)

- parallell corpora
- comparable corpora
- (big) dictionaries
Bilingual embedding mappings
Bilingual embedding mappings

Basque
- Zaurka
- Katu
- Txakur
- Egutegi
- Behi
- Banana
- Sagar
- Udare
- Etxe
- Miau
- Murr}

English
- X
- Z
- House
- Calendar
- Pear
- Apple
- Banana
- Cow
- Dog
- cat
- Moo
- Bark
- Meow
Bilingual embedding mappings

Seed dictionary

Basque
- Zaunka
- Katu
- Txakur
- Egutegi
- Miau
- Marru
- Behi
- Sagar
- Udare
- Etxe

English
- House
- Calendar
- Pear
- Apple
- Cow
- Dog
- Moo
- Bark
- Meow
- Banana
- Cat
- Meow
Bilingual embedding mappings

Basque

Seed dictionary

English

X

Z

Txakur
Miau
Marru
Zaunka
Katu
Behi
Etxe
Egutegi
Sagar
Udare

Dog
Apple
House
Calendar
Cow
Cat
Meow
Bark
Moo

Txakur
Sagar
::
Egutegi

Dog
Apple
::
Calendar
Bilingual embedding mappings

Example mappings:

- Basque: Txakur, Sagar, Egutegi
- English: Dog, Apple, Calendar

Diagram:

- Basque words mapped to English words using a mapping matrix $W$.
- Words are projected into a 2D space for visualization.
Bilingual embedding mappings

Seed dictionary

Basque

<table>
<thead>
<tr>
<th>Txakur</th>
<th>Sagar</th>
<th>Egutegi</th>
</tr>
</thead>
</table>

English

<table>
<thead>
<tr>
<th>Dog</th>
<th>Apple</th>
<th>Calendar</th>
</tr>
</thead>
</table>

$$W$$
Bilingual embedding mappings

\[ X_1,*, X_2,*, \cdots, X_n,* \]

\[ W \approx Z_1,*, Z_2,*, \cdots, Z_n,* \]

Basque

Seed dictionary

English

Txakur
Sagar
Egutegi

Dog
Apple
Calendar

Miau
Marru
Banana
Udare

Katu
Behi
Sagar

House
Etxe
Calendar

Egutegi

Moo
Marru
Bark

Meow
Meow
Bilingual embedding mappings

\[
\begin{align*}
\begin{bmatrix}
X_1, * \\
X_2, * \\
\vdots \\
X_n, *
\end{bmatrix}
\end{align*}
\approx
\begin{align*}
\begin{bmatrix}
Z_1, * \\
Z_2, * \\
\vdots \\
Z_n, *
\end{bmatrix}
\end{align*}
\]

Txakur
Sagar
\vdots
Egutegi

Dog
Apple
Calendar
Bilingual embedding mappings

\[
X_1^*, X_2^*, \ldots, X_n^* \approx Z_1^*, Z_2^*, \ldots, Z_n^*
\]

\[
\text{arg min}_{W \in O(n)} \sum_i \|X_i^* W - Z_j^*\|^2
\]

\[
\begin{bmatrix}
X_1^* \\
X_2^* \\
\vdots \\
X_n^*
\end{bmatrix}
\approx
\begin{bmatrix}
Z_1^* \\
Z_2^* \\
\vdots \\
Z_n^*
\end{bmatrix}
\]

Basque

Egutegi
Miau
Marru
Katu
Txakur

Sagar
Udare
Etxe

English

Dog
Apple
Calendar

Moo
Marru
Bark
Miau
Meow

Katu
Zaunka
Txakur
cat
Bilingual embedding mappings

\[ W \approx \arg\min_{W \in \mathcal{O}(n)} \sum_i \|X_i^* W - Z_j^*\|^2 \]

\[ \begin{bmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_n^* \end{bmatrix} [W] \approx \begin{bmatrix} Z_1^* \\ Z_2^* \\ \vdots \\ Z_n^* \end{bmatrix} \begin{bmatrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{bmatrix} \]
Bilingual embedding mappings

\[ \begin{align*}
X_1^*, X_2^*, \ldots, X_n^* \approx Z_1^*, Z_2^*, \ldots, Z_n^* \end{align*} \]

\[
\text{arg min}_{W \in O(n)} \sum_i \| X_i^* W - Z_j^* \|^2
\]
Bilingual embedding mappings

\[
\begin{align*}
\arg \min_{W \in \mathbb{O}(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2
\end{align*}
\]

\[
\begin{bmatrix}
X_{1,*} \\
X_{2,*} \\
\vdots \\
X_{n,*}
\end{bmatrix}
\approx
\begin{bmatrix}
Z_{1,*} \\
Z_{2,*} \\
\vdots \\
Z_{n,*}
\end{bmatrix}
\]

Basque

\[X\]

English

\[Z, XW\]
Bilingual embedding mappings
Bilingual embedding mappings
Bilingual embedding mappings

Monolingual embeddings

Dictionary
Bilingual embedding mappings

Diagram:

- Dictionary
- Monolingual embeddings
Bilingual embedding mappings

Diagram:
- Dictionary
- Monolingual embeddings
- Mapping
Bilingual embedding mappings

- Dictionary
- Monolingual embeddings
- Mapping
Bilingual embedding mappings

Monolingual embeddings

Dictionary

Mapping

Dictionary
Bilingual embedding mappings

Dictionary → Monolingual embeddings → Mapping → Dictionary

better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary → Mapping → Dictionary

better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary ➔ Mapping ➔ Dictionary

Mapping ➔ Mapping

better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary → Mapping → Dictionary

Mapping → Mapping

better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary → Mapping → Dictionary

Mapping → Dictionary

better!
Bilingual embedding mappings

Monolingual embeddings


better!
even better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary

Mapping

Dictionary

better!

Mapping

Dictionary

even better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary → Mapping → Dictionary — better!

Dictionary → Mapping → Dictionary — even better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary → Mapping → Dictionary

better!

Dictionary → Mapping → Dictionary

even better!

Dictionary → Mapping → Dictionary

Dictionary → Mapping

Dictionary
Bilingual embedding mappings

Monolingual embeddings

Dictionary -> Mapping -> Dictionary

Mapping -> Dictionary

Mapping -> Dictionary

Mapping -> Dictionary

better!

even better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary → Mapping → Dictionary

better!

Dictionary → Mapping → Dictionary

even better!

Dictionary → Mapping → Dictionary

even better!
Bilingual embedding mappings

Monolingual embeddings

Dictionary → Mapping → Dictionary

Dictionary
Bilingual embedding mappings

proposed self-learning method
Bilingual embedding mappings

proposed self-learning method
formalization and implementation details in the paper based on the mapping method of Artetxe et al. (2016)
Bilingual embedding mappings

**proposed self-learning method**
formalization and implementation details in the paper based on the mapping method of Artetxe et al. (2016)

Too good to be true?
Bilingual embedding mappings

proposed self-learning method
formalization and implementation details in the paper based on the mapping method of Artetxe et al. (2016)

Too good to be true?

Nope, it works!
Experiments
Experiments

- Dataset by Dinu et al. (2015)
Experiments

• Dataset by Dinu et al. (2015)
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish

<table>
<thead>
<tr>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
</table>
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*

| English-Italian | English-German | English-Finnish |
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs

<table>
<thead>
<tr>
<th></th>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Pairs</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>25</td>
<td>5,000</td>
</tr>
<tr>
<td></td>
<td>num.</td>
<td>25</td>
</tr>
<tr>
<td>5,000</td>
<td>25</td>
<td>5,000</td>
</tr>
<tr>
<td>num.</td>
<td></td>
<td>25</td>
</tr>
</tbody>
</table>
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>num.</td>
<td>num.</td>
<td>num.</td>
</tr>
</tbody>
</table>
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  - Monolingual embeddings (CBOW + negative sampling)
  - Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  - Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
<tr>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
<tr>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
</tbody>
</table>

word translation induction
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
<tr>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
<tr>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
</tbody>
</table>

Mikolov et al. (2013a)
Xing et al. (2015)
Zhang et al. (2016)
Artetxe et al. (2016)

word translation induction
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>num.</td>
<td>num.</td>
<td>num.</td>
</tr>
</tbody>
</table>

Mikolov et al. (2013a)
Xing et al. (2015)
Zhang et al. (2016)
Artetxe et al. (2016)
Our method

word translation induction
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th></th>
<th>English-Italian</th>
<th></th>
<th>English-German</th>
<th></th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
<td>5,000</td>
<td>25</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>34.93%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>35.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87%</td>
<td>0.00%</td>
<td>0.13%</td>
<td>41.27%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>36.73%</td>
<td>0.07%</td>
<td>0.27%</td>
<td>40.80%</td>
<td>0.13%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>39.27%</td>
<td>0.07%</td>
<td>0.40%</td>
<td><strong>41.87%</strong></td>
<td>0.13%</td>
</tr>
<tr>
<td>Our method</td>
<td><strong>39.67%</strong></td>
<td><strong>37.27%</strong></td>
<td><strong>39.40%</strong></td>
<td>40.87%</td>
<td><strong>39.60%</strong></td>
</tr>
</tbody>
</table>

word translation induction
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  - Monolingual embeddings (CBOW + negative sampling)
  - Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  - Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th></th>
<th>English-Italian</th>
<th></th>
<th></th>
<th>English-German</th>
<th></th>
<th></th>
<th>English-Finnish</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
<td>5,000</td>
<td>25</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>34.93%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>35.00%</td>
<td>0.00%</td>
<td>0.07%</td>
<td>25.91%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87%</td>
<td>0.00%</td>
<td>0.13%</td>
<td>41.27%</td>
<td>0.07%</td>
<td>0.53%</td>
<td>28.23%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>36.73%</td>
<td>0.07%</td>
<td>0.27%</td>
<td>40.80%</td>
<td>0.13%</td>
<td>0.87%</td>
<td>28.16%</td>
<td>0.14%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>39.27%</td>
<td>0.07%</td>
<td>0.40%</td>
<td><strong>41.87%</strong></td>
<td>0.13%</td>
<td>0.73%</td>
<td><strong>30.62%</strong></td>
<td>0.21%</td>
</tr>
<tr>
<td>Our method</td>
<td><strong>39.67%</strong></td>
<td><strong>37.27%</strong></td>
<td><strong>39.40%</strong></td>
<td>40.87%</td>
<td><strong>39.60%</strong></td>
<td>40.27%</td>
<td>28.72%</td>
<td>28.16%</td>
</tr>
</tbody>
</table>

word translation induction
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  - Monolingual embeddings (CBOW + negative sampling)
  - Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  - Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th></th>
<th>English-Italian</th>
<th>English-German</th>
<th>English-Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>34.93%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87%</td>
<td>0.00%</td>
<td>0.13%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>36.73%</td>
<td>0.07%</td>
<td>0.27%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>39.27%</td>
<td>0.07%</td>
<td>0.40%</td>
</tr>
<tr>
<td>Our method</td>
<td><strong>39.67%</strong></td>
<td><strong>37.27%</strong></td>
<td><strong>39.40%</strong></td>
</tr>
</tbody>
</table>

word translation induction
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  - Monolingual embeddings (CBOW + negative sampling)
  - Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  - Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th></th>
<th>English-Italian</th>
<th></th>
<th></th>
<th>English-German</th>
<th></th>
<th></th>
<th>English-Finnish</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>34.93%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>35.00%</td>
<td>0.00%</td>
<td>0.07%</td>
<td>25.91%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87%</td>
<td>0.00%</td>
<td>0.13%</td>
<td>41.27%</td>
<td>0.07%</td>
<td>0.53%</td>
<td>28.23%</td>
<td>0.07%</td>
<td>0.56%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>36.73%</td>
<td>0.07%</td>
<td>0.27%</td>
<td>40.80%</td>
<td>0.13%</td>
<td>0.87%</td>
<td>28.16%</td>
<td>0.14%</td>
<td>0.42%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>39.27%</td>
<td>0.07%</td>
<td>0.40%</td>
<td>41.87%</td>
<td>0.13%</td>
<td>0.73%</td>
<td>30.62%</td>
<td>0.21%</td>
<td>0.77%</td>
</tr>
<tr>
<td>Our method</td>
<td><strong>39.67%</strong></td>
<td><strong>37.27%</strong></td>
<td><strong>39.40%</strong></td>
<td><strong>40.87%</strong></td>
<td><strong>39.60%</strong></td>
<td><strong>40.27%</strong></td>
<td><strong>28.72%</strong></td>
<td><strong>28.16%</strong></td>
<td><strong>26.47%</strong></td>
</tr>
</tbody>
</table>

word translation induction
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  - Monolingual embeddings (CBOW + negative sampling)
  - Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  - Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th></th>
<th>English-Italian</th>
<th></th>
<th>English-German</th>
<th></th>
<th>English-Finnish</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
<td>5,000</td>
<td>25</td>
<td>num.</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>34.93%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>35.00%</td>
<td>0.00%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87%</td>
<td>0.00%</td>
<td>0.13%</td>
<td>41.27%</td>
<td>0.07%</td>
<td>0.53%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>36.73%</td>
<td>0.07%</td>
<td>0.27%</td>
<td>40.80%</td>
<td>0.13%</td>
<td>0.87%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>39.27%</td>
<td>0.07%</td>
<td>0.40%</td>
<td>41.87%</td>
<td>0.13%</td>
<td>0.73%</td>
</tr>
<tr>
<td>Our method</td>
<td>39.67%</td>
<td>37.27%</td>
<td>39.40%</td>
<td>40.87%</td>
<td>39.60%</td>
<td>40.27%</td>
</tr>
</tbody>
</table>

word translation induction
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th></th>
<th>English-Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,000</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>34.93%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>36.73%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>39.27%</td>
</tr>
<tr>
<td>Our method</td>
<td>39.67%</td>
</tr>
</tbody>
</table>

word translation induction
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  \[\Rightarrow\] Monolingual embeddings (CBOW + negative sampling)
  \[\Rightarrow\] Seed dictionary: 5,000 word pairs / 25 word pairs / numerals
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

crosslingual word similarity
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th>EN-IT</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>RG</td>
</tr>
</tbody>
</table>

crosslingual word similarity
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th>Bi. data</th>
<th>EN-IT</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>RG</td>
<td>WS</td>
</tr>
</tbody>
</table>

crosslingual word similarity
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th>Bi. data</th>
<th>EN-IT</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luong et al. (2015)</td>
<td>WS</td>
<td>RG</td>
</tr>
<tr>
<td>Europarl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

crosslingual word similarity
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th></th>
<th>EN-IT</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi. data</td>
<td>WS</td>
<td>RG</td>
</tr>
<tr>
<td>Luong et al. (2015)</td>
<td>Europarl</td>
<td></td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>5k dict</td>
<td></td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>5k dict</td>
<td></td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>5k dict</td>
<td></td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>5k dict</td>
<td></td>
</tr>
</tbody>
</table>

crosslingual word similarity
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / numerals*

<table>
<thead>
<tr>
<th>Bi. data</th>
<th>EN-IT</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Luong et al. (2015)</em></td>
<td>Europarl</td>
<td></td>
</tr>
<tr>
<td><em>Mikolov et al. (2013a)</em></td>
<td>5k dict</td>
<td></td>
</tr>
<tr>
<td><em>Xing et al. (2015)</em></td>
<td>5k dict</td>
<td></td>
</tr>
<tr>
<td><em>Zhang et al. (2016)</em></td>
<td>5k dict</td>
<td></td>
</tr>
<tr>
<td><em>Artetxe et al. (2016)</em></td>
<td>5k dict</td>
<td></td>
</tr>
<tr>
<td><em>Our method</em></td>
<td>5k dict</td>
<td>25 dict num.</td>
</tr>
</tbody>
</table>

crosslingual word similarity
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th>Method</th>
<th>Bi. data</th>
<th>EN-IT</th>
<th>RG</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luong et al. (2015)</td>
<td>Europarl</td>
<td>33.1%</td>
<td>33.5%</td>
<td>35.6%</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>5k dict</td>
<td>62.7%</td>
<td>64.3%</td>
<td>52.8%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>5k dict</td>
<td>61.4%</td>
<td>70.0%</td>
<td>59.5%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>5k dict</td>
<td>61.6%</td>
<td>70.4%</td>
<td>59.6%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>5k dict</td>
<td>61.7%</td>
<td>71.6%</td>
<td>59.7%</td>
</tr>
<tr>
<td>Our method</td>
<td>5k dict</td>
<td>62.4%</td>
<td>74.2%</td>
<td>61.6%</td>
</tr>
<tr>
<td></td>
<td>25 dict</td>
<td>62.6%</td>
<td>74.9%</td>
<td>61.2%</td>
</tr>
<tr>
<td></td>
<td>num.</td>
<td>62.8%</td>
<td>73.9%</td>
<td>60.4%</td>
</tr>
</tbody>
</table>

**crosslingual word similarity**
Experiments

• Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th></th>
<th>EN-IT</th>
<th></th>
<th>EN-DE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bi. data</td>
<td>WS</td>
<td>RG</td>
<td>WS</td>
</tr>
<tr>
<td>Luong et al. (2015)</td>
<td>Europarl</td>
<td>33.1%</td>
<td>33.5%</td>
<td>35.6%</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>5k dict</td>
<td>62.7%</td>
<td>64.3%</td>
<td>52.8%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>5k dict</td>
<td>61.4%</td>
<td>70.0%</td>
<td>59.5%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>5k dict</td>
<td>61.6%</td>
<td>70.4%</td>
<td>59.6%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>5k dict</td>
<td>61.7%</td>
<td>71.6%</td>
<td>59.7%</td>
</tr>
<tr>
<td>Our method</td>
<td>5k dict</td>
<td>62.4%</td>
<td>74.2%</td>
<td>61.6%</td>
</tr>
<tr>
<td></td>
<td>25 dict</td>
<td>62.6%</td>
<td>74.9%</td>
<td>61.2%</td>
</tr>
<tr>
<td></td>
<td>num.</td>
<td>62.8%</td>
<td>73.9%</td>
<td>60.4%</td>
</tr>
</tbody>
</table>

crosslingual word similarity
Experiments

- Dataset by Dinu et al. (2015) extended to German and Finnish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / numerals

<table>
<thead>
<tr>
<th></th>
<th>EN-IT</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bi. data</td>
<td>WS</td>
</tr>
<tr>
<td>Luong et al. (2015)</td>
<td>Europarl</td>
<td>33.1%</td>
</tr>
<tr>
<td>Mikolov et al. (2013a)</td>
<td>5k dict</td>
<td>62.7%</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>5k dict</td>
<td>61.4%</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>5k dict</td>
<td>61.6%</td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>5k dict</td>
<td>61.7%</td>
</tr>
<tr>
<td>Our method</td>
<td>5k dict</td>
<td>62.4%</td>
</tr>
<tr>
<td></td>
<td>25 dict</td>
<td>62.6%</td>
</tr>
<tr>
<td></td>
<td>num.</td>
<td><strong>62.8%</strong></td>
</tr>
</tbody>
</table>

**crosslingual word similarity**
Why does it work?
Why does it work?

Monolingual embeddings

Dictionary → Mapping → Dictionary
Why does it work?

small

Dictionary -> Mapping -> Dictionary

Monolingual embeddings
Why does it work?

Monolingual embeddings

Dictionary

Mapping

Dictionary

small

large
Why does it work?

Monolingual embeddings

Dictionary

Mapping

Dictionary

small

no error

large
Why does it work?

- Small, no error
- Large, errors

Monolingual embeddings

Dictionary ➔ Mapping ➔ Dictionary
Why does it work?

Monolingual embeddings

Dictionary -> Mapping -> Dictionary

small
no error

large
errors
Why does it work?
Why does it work?

Monolingual embeddings

Dictionary

Mapping

Dictionary

small

no error

Mapping

Dictionary

large

errors

better?

worse?
Why does it work?

Monolingual embeddings

Dictionary

Mapping

Dictionary

small
no error

Dictionary

Mapping

Dictionary

large
errors

better?
worse?
Why does it work?

- Monolingual embeddings
- Dictionary
- Mapping
- Dictionary
- Mapping
- Dictionary
- Mapping
- Dictionary
- Mapping
- Dictionary
- Mapping
- Dictionary

small
no error

large
errors

better?
worse?
even better?
Why does it work?

- Monolingual embeddings
- Dictionary
- Mapping
- Dictionary

small
no error

large
errors

better?
worse?

even better?
even worse?
Why does it work?
Why does it work?

Implicit objective: $$W^* = \arg \max_W \sum_i \max(X_i W) \cdot Z_j$$

s.t. $$WW^T = W^T W = I$$
Why does it work?

Implicit objective: \[ W^* = \arg \max_W \sum_i \max(X_i, W) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^TW = I \]

Independent from seed dictionary!
Why does it work?

Implicit objective:  
\[ W^* = \arg \max_W \sum_i \max_j (X_i^* W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

Implicit objective: \( W^* = \arg \max_W \sum_i \max_j (X_i^*W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^TW = I \)
Implicit objective: 

\[ W^* = \arg \max_w \sum_i \max(X_i^*W) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

Implicit objective: \[ W^* = \arg \max_W \sum_i \max(X_i, W) \cdot Z_j \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

Implicit objective: \[ W^* = \arg \max_w \sum_i \max(X_i^* W) \cdot Z_{j^*} \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

Implicit objective: \( W^* = \arg \max_W \sum_i \max(X_iW) \cdot Z_j \)  

s.t. \( WW^T = W^TW = I \)
Implicit objective: $W^* = \arg \max_W \sum_i \max (X_i^* W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^T W = I$
Why does it work?

Implicit objective: \[ W^* = \arg \max_W \sum_i \max(X_i*W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

Implicit objective: $W^* = \arg \max_W \sum_i \max(X_i \cdot W) \cdot Z_j^*$ s.t. $WW^T = W^T W = I$
Why does it work?

Implicit objective: \( W^* = \arg \max \sum_i \max(X_i W) \cdot Z_j \) \quad \text{s.t.} \quad WW^T = W^T W = I \
Why does it work?

Implicit objective: \( W^* = \arg \max_W \sum_i \max(X_iW) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^TW = I \)
Why does it work?

Implicit objective: \( W^* = \arg\max_W \sum_i \max(X_i^*W) \cdot Z_{j^*} \) \quad \text{s.t.} \quad WW^T = W^TW = I
Implicit objective: 

\[ W^* = \arg \max_W \sum_i \max(X_i^*W) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^TW = I \]
Why does it work?

Implicit objective: \[ W^* = \arg \max_W \sum_i \max(X_i^*W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^TW = I \]
Why does it work?

Implicit objective:  
\[ W^* = \arg \max_W \sum_i \max(X_i W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

Implicit objective: \( W^* = \arg \max_w \sum_i \max(X_i \cdot W) \cdot Z_j \) \( \text{s.t. } WW^T = W^T W = I \)
Why does it work?

**Implicit objective:**

$$W^* = \arg \max_W \sum_i \max(X_{i*}W) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^TW = I$$
Why does it work?

Implicit objective: \( W^* = \arg \max_W \sum_i \max(X_i^* W) \cdot Z_{j^*} \) \quad \text{s.t.} \quad WW^T = W^TW = I

Independent from seed dictionary!
Why does it work?

Implicit objective:  
\[ W^* = \arg \max_W \sum_i \max(X_iW) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^TW = I \]

Independent from seed dictionary!

So why do we need a seed dictionary?
Why does it work?

Implicit objective: \( W^* = \arg \max_W \sum_i \max_j (X_i^* W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^T W = I \)

Independent from seed dictionary!

So why do we need a seed dictionary?

Avoid poor local optima!
Implicit objective: 

\[ W^* = \arg\max_w \sum_i \max_j (X_i^* W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^T W = I \]
Conclusions
Conclusions

• Simple self-learning method to train bilingual embedding mappings
Conclusions

• Simple self-learning method to train bilingual embedding mappings

• High quality results with almost no supervision (25 words, numerals)
Conclusions

• Simple self-learning method to train bilingual embedding mappings

• High quality results with almost no supervision (25 words, numerals)

• Implicit optimization objective independent from seed dictionary
Conclusions

• Simple self-learning method to train bilingual embedding mappings

• High quality results with almost no supervision (25 words, numerals)

• Implicit optimization objective independent from seed dictionary

• Seed dictionary necessary to avoid poor local optima
Conclusions

- Simple self-learning method to train bilingual embedding mappings
- High quality results with almost no supervision (25 words, numerals)
- Implicit optimization objective independent from seed dictionary
- Seed dictionary necessary to avoid poor local optima
- Future work: fully unsupervised training
One more thing...
One more thing...

> git clone https://github.com/artetxem/vecmap.git
One more thing...

> git clone https://github.com/artetxem/vecmap.git
>
One more thing...

> git clone https://github.com/artetxem/vecmap.git
> python3 vecmap/map_embeddings.py
One more thing...

> git clone https://github.com/artetxem/vecmap.git
> python3 vecmap/map_embeddings.py --self_learning --numerals
One more thing...

> git clone https://github.com/artetxem/vecmap.git
> python3 vecmap/map_embeddings.py --self_learning --numerals
SRC_INPUT.EMB TRG_INPUT.EMB
One more thing...

> git clone https://github.com/artetxem/vecmap.git
> python3 vecmap/map_embeddings.py --self_learning --numerals
SRC_INPUT.EMB TRG_INPUT.EMB SRC_OUTPUT.EMB TRG_OUTPUT.EMB
One more thing...

> git clone https://github.com/artetxem/vecmap.git
> python3 vecmap/map_embeddings.py --self_learning --numerals
  SRC_INPUT.EMB TRG_INPUT.EMB SRC_OUTPUT.EMB TRG_OUTPUT.EMB
>
One more thing...

> git clone https://github.com/artetxem/vecmap.git
> python3 vecmap/map_embeddings.py --self_learning --numerals
> SRC_INPUT.EMB TRG_INPUT.EMB SRC_OUTPUT.EMB TRG_OUTPUT.EMB
> vecmap/reproduce_acl2017.sh
One more thing...

```bash
> git clone https://github.com/artetxem/vecmap.git
> python3 vecmap/map_embeddings.py --self_learning --numerals
SRC_INPUT.EMB TRG_INPUT.EMB SRC_OUTPUT.EMB TRG_OUTPUT.EMB
> vecmap/reproduce_acl2017.sh
```

<table>
<thead>
<tr>
<th>5,000 WORD DICTIONARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Mikolov et al. (2013a)</td>
</tr>
<tr>
<td>- Xing et al. (2015)</td>
</tr>
<tr>
<td>- Zhang et al. (2016)</td>
</tr>
<tr>
<td>- Artetxe et al. (2016)</td>
</tr>
<tr>
<td>- Proposed method</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>25 WORD DICTIONARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Mikolov et al. (2013a)</td>
</tr>
<tr>
<td>- Xing et al. (2015)</td>
</tr>
<tr>
<td>- Zhang et al. (2016)</td>
</tr>
<tr>
<td>- Artetxe et al. (2016)</td>
</tr>
<tr>
<td>- Proposed method</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NUMERAL DICTIONARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Mikolov et al. (2013a)</td>
</tr>
<tr>
<td>- Xing et al. (2015)</td>
</tr>
<tr>
<td>- Zhang et al. (2016)</td>
</tr>
<tr>
<td>- Artetxe et al. (2016)</td>
</tr>
<tr>
<td>- Proposed method</td>
</tr>
</tbody>
</table>
Thank you!

https://github.com/artetxem/vecmap