Cross-linguality and machine translation without bilingual data

Eneko Agirre
@eagirre

Joint work with: Mikel Artetxe, Gorka Labaka

IXA NLP group – University of the Basque Country (UPV/EHU)
Motivation

Previous work on **cross-lingual word representations**:

- **Word embeddings** key for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations
Motivation

Previous work on cross-lingual word representations:

- **Word embeddings** key for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Our focus: **extend mappings to any pair of languages**

- Most language pairs have **very few bilingual resources**
- Key research area for **wide adoption** of NLP tools
Motivation

Previous work on cross-lingual word representations:
• Word embeddings key for Natural Language Processing
• Mapped embeddings represent languages in a single space
  • Depend on seed bilingual dictionaries
• Exciting results in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Our focus: extend mappings to any pair of languages
• Most language pairs have very few bilingual resources
• Key research area for wide adoption of NLP tools

In particular: no bilingual resources at all
• Unsupervised embedding mappings
• Unsupervised neural machine translation
Overview

Arabic monolingual corpora

Chinese monolingual corpora
Overview

Arabic monolingual corpora

Chinese monolingual corpora

Arabic embeddings

Chinese embeddings

Bilingual embeddings
Overview

Arabic monolingual corpora

Chinese monolingual corpora

Arabic embeddings

Chinese embeddings

Bilingual embeddings

Bilingual dictionaries

Crosslingual & multilingual applications

Machine translation
Overview

Arabic monolingual corpora

No bilingual resource

Chinese monolingual corpora

Arabic embeddings

Chinese embeddings

Bilingual embeddings

Bilingual dictionaries

Crosslingual & multilingual applications

Machine translation
Outline

• Bilingual embedding mappings
  • Introduction to vector space models (embeddings)
  • Introduction to bilingual embedding mappings
  • Reduced supervision
    • Self-learning, semi-supervised (ACL17)
    • Self-learning, fully unsupervised (ACL18)
  • Conclusions

• Unsupervised neural machine translation
  • Introduction to NMT
  • From bilingual embeddings to uNMT (ICLR18)
  • Conclusions
Outline

• Bilingual embedding mappings
  • *Introduction to vector space models (embeddings)*
  • *Introduction to bilingual embedding mappings*
  • *Reduced supervision*
    • Self-learning, semi-supervised (ACL17)
    • Self-learning, fully unsupervised (ACL18)
  • *Conclusions*

• Unsupervised neural machine translation
  • *Introduction to NMT*
  • *From bilingual embeddings to uNMT (ICLR18)*
  • *Conclusions*
Introduction to vector space models
Introduction to vector space models
Introduction to vector space models

Geographical space
Introduction to vector space models

Geographical space
- Cities
Introduction to vector space models

Geographical space
- Cities
- Meaningful distances
Introduction to vector space models

Geographical space
- Cities
- Meaningful distances
Introduction to vector space models

Geographical space
- Cities
- Meaningful distances
Introduction to vector space models

- Geographical space
  - Cities
  - Meaningful distances
  - Meaningful relations
Introduction to vector space models

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
Introduction to vector space models

**Geographical space**
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
Introduction to vector space models

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Geographical space
Introduction to vector space models

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space
Introduction to vector space models

Semantic space
- Words

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

**Semantic space**
- Words

**Geographical space**
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

Semantic space
- Words
- Meaningful distances

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

Semantic space
- Words
- Meaningful distances

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

**Semantic space**
- Words
- Meaningful distances

**Geographical space**
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space
- Words
- Meaningful distances
- Meaningful relations
Introduction to vector space models

Semantic space
- Words
- Meaningful distances
- Meaningful relations

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

Semantic space
- Words
- Meaningful distances
- Meaningful relations

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

**Semantic space**
- Words
- Meaningful distances
- Meaningful relations

**Geographical space**
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

Semantic space
- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to vector space models

Semantic space
- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Neural networks / linear algebra from co-occurrence counts

Geographical space
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world
Introduction to embedding mappings
Introduction to embedding mappings
Introduction to embedding mappings
Introduction to embedding mappings

Bilbo
Baiona
Iruña

X

Bilbao
Bayona
Pamplona

Z
Introduction to embedding mappings

Bilbo
Baiona
Iruñea

Bilbao
Bayona
Pamplona
Introduction to embedding mappings

Bilbo
Baiona
Iruña

Bilbao
Bayona
Pamplona
Introduction to embedding mappings
Introduction to embedding mappings

Bilbo
Baiona
Iruñea

Bilbao
Bayona
Pamplona
Introduction to embedding mappings
Introduction to embedding mappings
Introduction to embedding mappings

Seed dictionary

Basque
- Zaunka
- Mjau
- Marru
- Katu
- Txakur
- Egutesi
- Etxe

Basque
- X
- Marru
- Łagăr
- Udare

English
- IHouse
- Calendar
- Pear
- Apple
- Banana
- Moo
- Bark
- Meow
- Cow
- Dog
- Cat
Introduction to embedding mappings

Seed dictionary

Basque

X

Z

English

Txakur
Sagar
Egutegi

Dog
Apple
Calendar

Mjau
Marru
Banana
Udare

Katu
Behi
Sagar

Egutegi

Etxe

Pear
Apple

Calendar

House

Cow
Dog

Moo

Bark

Meow

cat
Introduction to embedding mappings

Seed dictionary

Basque
- Txakur
- Sagar
- Egutegi

English
- Dog
- Apple
- Calendar

$X$ $\Rightarrow$ $Z$

$W$
Introduction to embedding mappings

The diagram illustrates the concept of embedding mappings between two languages: Basque (left) and English (right). The seed dictionary includes words like "Txakur" (Basque), "Dog" (English), "Sagar" (Basque), "Apple" (English), and "Egutegi" (Basque). The mapping is represented by the matrix $W$, transforming the Basque seed dictionary into the English one. Notable words in the Basque seed dictionary include "Mjau" (cat), "Marru" (cow), "Etxe" (house), "Calendar", and "Egutegi".
Introduction to embedding mappings

Seed dictionary

Basque

Miu
Marru
Yanaka
Katu
Behi
Sagar
Udare
Egutegi
Etxe

English

Moo
Marru
Bark
Miu
Yanaka
Meow

X

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

W

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

Dog
Apple
Calendar
Introduction to embedding mappings

\[
\begin{align*}
\text{Basque} & \quad \text{Seed dictionary} \\
\text{English} & \quad X \xrightarrow{W} Z \cdot XW
\end{align*}
\]

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

- Txakur
- Sagar
- Egutegi
- Mjau
- Marru
- Banana
- Shag
- Udare
- Cat
- Dog
- Cow
- Calendar
- House
- Etxe
- Bark
- Meow
- moo
- Marru
- Katu
- Txakur

\[
\text{Introduction to embedding mappings}
\]
Introduction to embedding mappings

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

Mikolov et al. (2013b)

- Txakur
- Sagar
  \[
  \begin{bmatrix}
  X_{1,*} \\
  X_{2,*} \\
  \vdots \\
  X_{n,*}
  \end{bmatrix}
  \]
- Egutegi

\[
\begin{bmatrix}
Z_{1,*} \\
Z_{2,*} \\
\vdots \\
Z_{n,*}
\end{bmatrix}
\approx
\begin{bmatrix}
\text{Dog} \\
\text{Apple} \\
\vdots \\
\text{Calendar}
\end{bmatrix}
\]
Introduction to embedding mappings

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

Mikolov et al. (2013b)

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

Basque

- Txakur
- Sagar
- Egutegi

English

- Apple
- Calendar
- Dog
- Etxe
- House
- Moo
Introduction to embedding mappings

\[
\underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min} \sum_i \| \mathbf{x}_i \mathbf{w} - \mathbf{z}_j \|^2
\]

Mikolov et al. (2013b)

\[
\begin{bmatrix}
\mathbf{X}_{1,*} \\
\mathbf{X}_{2,*} \\
\vdots \\
\mathbf{X}_{n,*}
\end{bmatrix} [\mathbf{W}] \approx
\begin{bmatrix}
\mathbf{Z}_{1,*} \\
\mathbf{Z}_{2,*} \\
\vdots \\
\mathbf{Z}_{n,*}
\end{bmatrix}
\]

Dog

Apple

Calendar
Introduction to embedding mappings

Mikolov et al. (2013b)

$$\arg\min_{w \in \mathbb{R}^d} \sum_i \|X_i W - Z_j\|^2$$

$$\begin{bmatrix}
X_{1,*} \\
X_{2,*} \\
\vdots \\
X_{n,*}
\end{bmatrix} W \approx \begin{bmatrix}
Z_{1,*} \\
Z_{2,*} \\
\vdots \\
Z_{n,*}
\end{bmatrix}$$
State-of-the-art in supervised mappings

Artetxe et al. AAAI 2018

• Use 5000 sized seed bilingual dictionary
• Framework subsuming previous work, a sequence of (optional) linear mappings:
  (opt.) Pre-process: Normalize length, mean centering
  1. (opt.) Whitening
  2. Orthogonal mapping, solved with SVD (Procrustes)
  3. (opt.) Re-weighting
  4. (opt.) De-whitening
• Optional steps, properly combined, bring up to 5 points improvement
Why does it work?
Why does it work?
Why does it work?
Why does it work?
Why does it work?
Why does it work?
Why does it work?

Languages are largely isometric in embedding space (!)
Outline

• Bilingual embedding mappings
  • Introduction to vector space models (embeddings)
  • Introduction to bilingual embedding mappings
• Reduced supervision
  • Self-learning, semi-supervised (ACL17)
  • Self-learning, fully unsupervised (ACL18)
• Conclusions

• Unsupervised neural machine translation
  • Introduction to NMT
  • From bilingual embeddings to uNMT (ICLR18)
• Conclusions
Reducing supervision
Reducing supervision
Reducing supervision

Previous work

bilingual signal for training
Reducing supervision

Previous work

- parallel corpora
- comparable corpora
- (big) dictionaries

Bilingual signal for training
Reducing supervision

Previous work

- parallel corpora
- comparable corpora
- (big) dictionaries

bilingual signal for training


Removing supervision

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

Our work:
- 25 word dictionary
- numerals (1, 2, 3...)
- nothing
Self-learning
Self-learning

Monolingual embeddings
Self-learning

Monolingual embeddings

Dictionary
Self-learning

Dictionary ➔ Monolingual embeddings
Self-learning

Dictionary ➔ Monolingual embeddings ➔ Mapping
Self-learning

Diagram:
- Dictionary
- Monolingual embeddings
- Mapping
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary

better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary

better!
Self-learning

Monolingual embeddings

Dictionary -> Mapping -> Dictionary

Mapping

better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary → Mapping

better!
Self-learning

Monolingual embeddings


better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary
better!

Dictionary → Mapping → Dictionary
even better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary → even better!

Dictionary → Mapping → Dictionary → better!
Self-learning

Monolingual embeddings

Dictionary ➔ Mapping ➔ Dictionary ➔ even better!

Mapping ➔ Dictionary ➔ even better!

Mapping ➔
Self-learning

Monolingual embeddings

Dictionary ➔ Mapping ➔ Dictionary ➔ even better!

Dictionary ➔ Mapping ➔ Dictionary ➔ better!
Self-learning

Monolingual embeddings


better!

even better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary

even better!

Mapping → Dictionary

even better!

Mapping → Dictionary

even better!

Mapping → Dictionary

better!
Self-learning
Self-learning

proposed self-learning method

Too good to be true?
Semi-supervised experiments (ACL17)
Semi-supervised experiments (ACL17)

• Given monolingual embeddings plus seed bilingual dictionary (*train* dictionary):
  • 25 word pairs
  • Pairs of numerals
Semi-supervised experiments (ACL17)

- Given monolingual embeddings plus seed bilingual dictionary (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals
- Induce bilingual dictionary using self-learning for full vocabulary
Semi-supervised experiments (ACL17)

• Given monolingual embeddings plus seed bilingual dictionary \((\textit{train} \text{ dictionary})\):
  • 25 word pairs
  • Pairs of numerals
• Induce bilingual dictionary using self-learning for full vocabulary
• Evaluation
  • Compare translations to existing bilingual dictionary \((\textit{test} \text{ dictionary})\)
  • Accuracy
Semi-supervised experiments (ACL17)

![Graph showing accuracy (%) over seed dictionary size for different methods.](image)

**English-Italian**

- **Method**
  - Our method
  - Artetxe et al. (2016)
  - Xing et al. (2015)
  - Zhang et al. (2016)
  - Mikolov et al. (2013a)
Why does it work?
Why does it work?
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: \( W^* = \arg\max_W \sum_i \max(X_i^*W) \cdot Z_j^* \quad \text{s.t.} \quad WW^T = W^TW = I \)
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: \[ W^* = \arg \max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \] \text{s.t. } 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: \[ W^* = \arg \max_W \sum_i \max(X_{i*}, W) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

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

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

**Implicit objective:**

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

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

Implicit objective: $W^* = \arg \max_W \sum_i \max(Z_{i*}, W) \cdot Z_{j*}$ s.t. $WW^T = W^T W = I$
Why does it work?

Implicit objective: 

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

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

Implicit objective: $W^* = \arg \max_w \sum \max(X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$
Why does it work?

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

Independent from seed dictionary!
Why does it work?

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

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!
Why does it work?

Implicit objective: 

\[ \mathbf{W}^* = \arg \max_{\mathbf{W}} \sum_{i} \max(X_{i*}, W) \cdot Z_{j*} \quad \text{s.t.} \quad \mathbf{W}^T \mathbf{W} = \mathbf{I} \]
Next steps

Is there a way we can avoid the seed dictionary?

Would an initial noisy initialization suffice?
Unsupervised experiments (ACL18)
Unsupervised experiments (ACL18)

Initial dictionary:
- Compute intra-language similarity
- Words which are translations of each other
  would have analogous similarity histograms (isometry hyp.)
Unsupervised experiments (ACL18)

Initial dictionary:
- Compute intra-language similarity
- Words which are translations of each other
  would have analogous similarity histograms (isometry hyp.)
Unsupervised experiments (ACL18)

Initial dictionary:
- Compute intra-language similarity
- Words which are translations of each other
  would have analogous similarity histograms (isometry hyp.)
Unsupervised experiments (ACL18)

Initial dictionary:
- Compute intra-language similarity
- Words which are translations of each other
  would have analogous similarity histograms (isometry hyp.)
Unsupervised experiments (ACL18)

Initial dictionary:
- Compute intra-language similarity
- Words which are translations of each other
  would have analogous similarity histograms (isometry hyp.)

- It works, but very weak: Accuracy 0.52%
Unsupervised experiments (ACL18)

Initial dictionary:
- Compute intra-language similarity
- Words which are translations of each other
  would have analogous similarity histograms (isometry hyp.)

- It works, but very weak: Accuracy 0.52%

- For self-learning to work we added:
  1) Stochastic dictionary induction
  2) Frequency-based vocabulary cut-off
  3) Instead of inducing dictionary with nearest-neighbour
     use CSLS (Lample et al. 2018), due to hubness problem
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015)
Unsupervised experiments (ACL18)

• Dataset by Dinu et al. (2015)

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015), extended German, Finnish, Spanish

<table>
<thead>
<tr>
<th>Supervision Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ Monolingual embeddings (CBOW + negative sampling)

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN-DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN-FI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN-ES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5k dict.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25 dict.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / none
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 dict.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

• Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / none
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td>Mikolov et al. (2013)</td>
<td>34.93⁺</td>
<td>35.00⁺</td>
<td>25.91⁺</td>
<td>27.73⁺</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27</td>
<td>41.87⁺</td>
<td>30.62⁺</td>
<td>31.40⁺</td>
</tr>
<tr>
<td></td>
<td>Smith et al. (2017)</td>
<td>43.1</td>
<td>43.33⁺</td>
<td>29.42⁺</td>
<td>35.13⁺</td>
</tr>
<tr>
<td></td>
<td>Our method (AAAI18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 dict.</td>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - Monolingual embeddings (CBOW + negative sampling)
  - Seed dictionary: 5,000 word pairs / 25 word pairs / none
  - Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td>Mikolov et al. (2013)</td>
<td>34.93†</td>
<td>35.00†</td>
<td>25.91†</td>
<td>27.73†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27</td>
<td>41.87*</td>
<td>30.62*</td>
<td>31.40*</td>
</tr>
<tr>
<td></td>
<td>Smith et al. (2017)</td>
<td>43.1</td>
<td>43.33†</td>
<td>29.42†</td>
<td>35.13†</td>
</tr>
<tr>
<td></td>
<td>Our method (AAAI18)</td>
<td><strong>45.27</strong></td>
<td><strong>44.13</strong></td>
<td><strong>32.94</strong></td>
<td><strong>36.60</strong></td>
</tr>
<tr>
<td>25 dict.</td>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / none
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td>Mikolov et al. (2013)</td>
<td>34.93†</td>
<td>35.00†</td>
<td>25.91†</td>
<td>27.73†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27</td>
<td>41.87*</td>
<td>30.62*</td>
<td>31.40*</td>
</tr>
<tr>
<td>25 dict.</td>
<td>Our method (ACL17)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / none
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td>Mikolov et al. (2013)</td>
<td>34.93†</td>
<td>35.00†</td>
<td>25.91†</td>
<td>27.73†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27</td>
<td>41.87*</td>
<td>30.62*</td>
<td>31.40*</td>
</tr>
<tr>
<td>25 dict.</td>
<td>Our method (ACL17)</td>
<td>37.27</td>
<td>39.60</td>
<td>28.16</td>
<td>-</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / none
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td>Mikolov et al. (2013)</td>
<td>34.93†</td>
<td>35.00†</td>
<td>25.91†</td>
<td>27.73†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27</td>
<td>41.87*</td>
<td>30.62*</td>
<td>31.40*</td>
</tr>
<tr>
<td></td>
<td>Smith et al. (2017)</td>
<td>43.1</td>
<td>43.33†</td>
<td>29.42†</td>
<td>35.13†</td>
</tr>
<tr>
<td></td>
<td>Our method (AAAI18)</td>
<td>45.27</td>
<td>44.13</td>
<td>32.94</td>
<td>36.60</td>
</tr>
<tr>
<td>25 dict.</td>
<td>Our method (ACL17)</td>
<td>37.27</td>
<td>39.60</td>
<td>28.16</td>
<td>-</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  ⇒ *Test dictionary: 1,500 word pairs*

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td>Mikolov et al. (2013)</td>
<td>34.93†</td>
<td>35.00†</td>
<td>25.91†</td>
<td>27.73†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27</td>
<td>41.87*</td>
<td>30.62*</td>
<td>31.40*</td>
</tr>
<tr>
<td></td>
<td>Smith et al. (2017)</td>
<td>43.1</td>
<td>43.33†</td>
<td>29.42†</td>
<td>35.13†</td>
</tr>
<tr>
<td></td>
<td>Our method (AAAI18)</td>
<td>45.27</td>
<td>44.13</td>
<td>32.94</td>
<td>36.60</td>
</tr>
<tr>
<td>25 dict.</td>
<td>Our method (ACL17)</td>
<td>37.27</td>
<td>39.60</td>
<td>28.16</td>
<td>-</td>
</tr>
<tr>
<td>None</td>
<td>Zhang et al. (2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conneau et al. (2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Our method (ACL18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ Monolingual embeddings (CBOW + negative sampling)
  ⇒ Seed dictionary: 5,000 word pairs / 25 word pairs / none
  ⇒ Test dictionary: 1,500 word pairs

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k dict.</td>
<td>Mikolov et al. (2013)</td>
<td>34.93†</td>
<td>35.00†</td>
<td>25.91†</td>
<td>27.73†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27</td>
<td>41.87*</td>
<td>30.62*</td>
<td>31.40*</td>
</tr>
<tr>
<td></td>
<td>Smith et al. (2017)</td>
<td>43.1</td>
<td>43.33†</td>
<td>29.42†</td>
<td>35.13†</td>
</tr>
<tr>
<td></td>
<td>Our method (AAAI18)</td>
<td>45.27</td>
<td>44.13</td>
<td>32.94</td>
<td>36.60</td>
</tr>
<tr>
<td>25 dict.</td>
<td>Our method (ACL17)</td>
<td>37.27</td>
<td>39.60</td>
<td>28.16</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Zhang et al. (2017)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Conneau et al. (2018)</td>
<td>13.55</td>
<td>42.15</td>
<td>0.38</td>
<td>21.23</td>
</tr>
<tr>
<td></td>
<td>Our method (ACL18)</td>
<td>48.13</td>
<td>48.19</td>
<td>32.63</td>
<td>37.33</td>
</tr>
</tbody>
</table>
Conclusions
Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Matches results of supervised methods with no supervision
- Implicit optimization objective independent from seed dictionary
Conclusions

• Simple self-learning method to train bilingual embedding mappings
• Matches results of supervised methods with no supervision
• Implicit optimization objective independent from seed dictionary
• High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
Conclusions

• Simple self-learning method to train bilingual embedding mappings
• Matches results of supervised methods with no supervision
• Implicit optimization objective independent from seed dictionary
• High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
• Full reproducibility (including datasets):
  https://github.com/artetxem/vecmap
Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Matches results of supervised methods with no supervision
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries: Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
- Full reproducibility (including datasets): https://github.com/artetxem/vecmap
- Shows that languages share “semantic” structure to a large degree
Conclusions

• Simple self-learning method to train bilingual embedding mappings
• Matches results of supervised methods with no supervision
• Implicit optimization objective independent from seed dictionary
• High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
• Full reproducibility (including datasets):
  https://github.com/artetxem/vecmap

• Shows that languages share “semantic” structure to a large degree
• No need of magic
Conclusions


Outline

• Bilingual embedding mappings
  • Introduction to vector space models (embeddings)
  • Introduction to bilingual embedding mappings
  • Reduced supervision
    • Self-learning, semi-supervised (ACL17)
    • Self-learning, fully unsupervised (ACL18)
  • Conclusions

• Unsupervised neural machine translation
  • Introduction to NMT
  • From bilingual embeddings to uNMT (ICLR18)
  • Conclusions
Introduction to (supervised) NMT
Introduction to (supervised) NMT

• Given pairs of sentences with known translation \((x_1...x_n, y_1...y_m)\)

  This is my dearest dog \(</s>\)
  Este es mi perro preferido \(</s>\)
Introduction to (supervised) NMT

• Given pairs of sentences with known translation \((x_1...x_n, y_1...y_m)\)
  
  This is my dearest dog </s>
  Este es mi perro preferido </s>

• Train an encoder based on Recurrent Neural Nets
  return all hidden states, encoding input \(x_1...x_n\)
Introduction to (supervised) NMT

• Given pairs of sentences with known translation \((x_1\ldots x_n, y_1\ldots y_m)\)
  
  This is my dearest dog </s>
  
  Este es mi perro preferido </s>

• Train an **encoder** based on Recurrent Neural Nets
  
  return all hidden states, encoding input \(x_1\ldots x_n\)

• Train a **decoder** based on Recurrent Neural Nets
  
  - based on hidden states and last word in translation \(y_{i-1}\)
  
  - plus an **attention** mechanism
  
  - classifier guesses next word \(y_i\)
Introduction to (supervised) NMT

• Given pairs of sentences with known translation \((x_1...x_n, y_1...y_m)\)
  
  This is my dearest dog \(<s>\)
  Este es mi perro preferido \(<s>\)

• Train an **encoder** based on Recurrent Neural Nets
  return all hidden states, encoding input \(x_1...x_n\)

• Train a **decoder** based on Recurrent Neural Nets
  - based on hidden states and last word in translation \(y_{i-1}\)
  - plus an **attention** mechanism
  - classifier guesses next word \(y_i\)

End-to-end training
Introduction to (supervised) NMT

Source: Wu et al. 2016 (~ 30 authors – Also known as Google NMT)
Introduction to (supervised) NMT

Encoder for L1

L1 embeddings

L2 decoder

softmax

attention

...
Unsupervised neural machine translation

• Now that we can **represent words in two languages in the same embeddings space** without bilingual dictionaries...
  ... what can we do?
Unsupervised neural machine translation

- Now that we can represent words in two languages in the same embeddings space without bilingual dictionaries...
  ... what can we do?

- We change the architecture of the NMT system:
  - Handle both directions together (L1 -> L2, L2 -> L1)
  - Shared encoder for the two languages (E)
  - Two decoders for each language (D1, D2)
  - Fixed embeddings
Unsupervised neural machine translation
Unsupervised neural machine translation
Unsupervised neural machine translation
Unsupervised neural machine translation

• We change the **architecture of the NMT system**:  
  - Handle both directions together (L1 -> L2, L2 -> L1)  
  - Shared encoder for the two languages (E)  
  - Two decoders for each language (D1, D2)  
  - Fixed embeddings

• We change the **training regime, mixing mini-batches**:  
  - Denoising autoencoder: noisy input in L1, output in the same language (E+D1)  
  - Denoising autoencoder: noisy input in L2, output in the same language (E+D2)  
  - Backtranslation: input in L1, translate E+D2, translate E+D1, output in L1  
  - Backtranslation: input in L2, translate E+D1, translate E+D2, output in L2
Unsupervised neural machine translation

Training
Unsupervised neural machine translation

Training

Une fusillade a eu lieu à l’aéroport international de Los Angeles.
Unsupervised neural machine translation

Training
- Supervised

Une fusillade a eu lieu à l’aéroport international de Los Angeles.
Unsupervised neural machine translation

Training

- Supervised

There was a shooting in Los Angeles International Airport.

Une fusillade a eu lieu à l’aéroport international de Los Angeles.
Unsupervised neural machine translation

Training
- Supervised

Une fusillade a eu lieu à l’aéroport international de Los Angeles.

There was a shooting in Los Angeles International Airport.
Unsupervised neural machine translation

Training
- Supervised

Une fusillade a eu lieu à l’aéroport international de Los Angeles.

There was a shooting in Los Angeles International Airport.
Unsupervised neural machine translation

Training

- Supervised

Une fusillade a eu lieu à l’aéroport international de Los Angeles.
Unsupervised neural machine translation

Training
- Supervised
- Autoencoder

Une fusillade a eu lieu à l’aéroport international de Los Angeles.
Unsupervised neural machine translation

Training

- **Supervised**
- **Autoencoder**

A fusillade a eu lieu à l'aéroport international de Los Angeles.
Unsupervised neural machine translation

Training
- Supervised
- Denoising Autoencoder

Une fusillade a eu lieu à l’aéroport international de Los Angeles.
Unsupervised neural machine translation

Training

- **Supervised**
- **Denoising Autoencoder**

![Diagram of neural machine translation process](image)

*Une fusillade a eu lieu à l’aéroport international de Los Angeles.*

*Une lieu fusillade a eu à l’aéroport de Los Angeles.*
Unsupervised neural machine translation

Training
- **Supervised**
- **Denoising Autoencoder**

There was a shooting in Los Angeles International Airport.
Unsupervised neural machine translation

Training

- Supervised
- Denoising Autoencoder

There was a shooting in Los Angeles International Airport.
Unsupervised neural machine translation

Training
- Supervised
- Denoising
- Backtranslation

Une fusillade a eu lieu à l’aéroport international de Los Angeles.
Unsupervised neural machine translation

Training
- **Supervised**
- **Denoising**
- **Backtranslation**

*Une fusillade a eu lieu à l’aéroport international de Los Angeles.*
Unsupervised neural machine translation

**Training**
- **Supervised**
- **Denoising**
- **Backtranslation**

Une fusillade a eu lieu à l’aéroport international de Los Angeles.

A shooting has had place in airport international of Los Angeles.
Unsupervised neural machine translation

**Training**
- **Supervised**
- **Denoising**
- **Backtranslation**

---

*Une lieu fusillade a eu à l’aéroport de Los international Angeles.*

*A shooting has had place in airport international of Los Angeles.*
Unsupervised neural machine translation

Training
- Supervised
- Denoising
- Backtranslation

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

A shooting has had place in airport international of Los Angeles.
Unsupervised neural machine translation

- We change the **architecture of the NMT system**:  
  - Handle both directions together (L1 -> L2, L2 -> L1)  
  - Shared encoder for the two languages (E)  
  - Two decoders for each language (D1, D2)  
  - Fixed embeddings

- We change the **training regime, mixing mini-batches**:  
  - Denoising autoencoder: noisy input in L1, output in the same language (E+D1)  
  - Denoising autoencoder: noisy input in L2, output in the same language (E+D2)  
  - Backtranslation: input in L1, translate E+D2, translate E+D1, output in L1  
  - Backtranslation: input in L2, translate E+D1, translate E+D2, output in L2
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
</table>


Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
</tbody>
</table>

Unsupervised
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised</th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
</tbody>
</table>

It works!
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
</tbody>
</table>

Semi-supervised
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
<tr>
<td>Semi-supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed (full) + 10k parallel</td>
<td>18.57</td>
<td>17.34</td>
<td>11.47</td>
<td>7.86</td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
<tr>
<td>Semi-supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed (full) + 10k parallel</td>
<td>18.57</td>
<td>17.34</td>
<td>11.47</td>
<td>7.86</td>
</tr>
<tr>
<td>Proposed (full) + 100k parallel</td>
<td>21.81</td>
<td>21.74</td>
<td>15.24</td>
<td>10.95</td>
</tr>
</tbody>
</table>
## Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Method</th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unsupervised</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
<tr>
<td><strong>Semi-supervised</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed (full) + 10k parallel</td>
<td>18.57</td>
<td>17.34</td>
<td>11.47</td>
<td>7.86</td>
</tr>
<tr>
<td>Proposed (full) + 100k parallel</td>
<td>21.81</td>
<td>21.74</td>
<td>15.24</td>
<td>10.95</td>
</tr>
<tr>
<td><strong>Supervised</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(upperbound)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
<tr>
<td>Semi-supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervised (upperbound)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparable NMT</td>
<td>20.48</td>
<td>19.89</td>
<td>15.04</td>
<td>11.05</td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (emb. nearest neighbor)</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td>Proposed (denoising)</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td>Proposed (+backtranslation)</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
<tr>
<td>Semi-supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervised (upperbound)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparable NMT</td>
<td>20.48</td>
<td>19.89</td>
<td>15.04</td>
<td>11.05</td>
</tr>
<tr>
<td>GNMT (Wu et al., 2016)</td>
<td>-</td>
<td>38.95</td>
<td>-</td>
<td>24.61</td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
</table>

210
Unsupervised neural machine translation

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Une fusillade a eu lieu à l’aéroport international de Los Angeles.</td>
<td>A shooting occurred at Los Angeles International Airport.</td>
</tr>
</tbody>
</table>
### Unsupervised neural machine translation

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Une fusillade a eu lieu à l'aéroport international de Los Angeles.</td>
<td>A shooting occurred at Los Angeles International Airport.</td>
</tr>
<tr>
<td>Cette controverse croissante autour de l’agence a provoqué beaucoup de speculations selon lesquelles l’incident de ce soir était le résultat d’une cyber-operation ciblée.</td>
<td>This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.</td>
</tr>
</tbody>
</table>
### Source

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Une fusillade a eu lieu à l’aéroport international de Los Angeles.</td>
<td>A shooting occurred at Los Angeles International Airport.</td>
</tr>
<tr>
<td>Cette controverse croissante autour de l’agence a provoqué beaucoup de speculations selon lesquelles l’incident de ce soir était le résultat d’une cyber-operation ciblée.</td>
<td>This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.</td>
</tr>
<tr>
<td>Le nombre total de morts en octobre est le plus élevé depuis avril 2008, quand 1 073 personnes avaient été tuées.</td>
<td>The total number of deaths in May is the highest since April 2008, when 1 064 people had been killed.</td>
</tr>
<tr>
<td>Source</td>
<td>Translation</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Une fusillade a eu lieu à l’aéroport international de Los Angeles.</td>
<td>A shooting occurred at Los Angeles International Airport.</td>
</tr>
<tr>
<td>Cette controverse croissante autour de l’agence a provoqué beaucoup de speculations selon lesquelles l’incident de ce soir était le résultat d’une cyber-operation ciblée.</td>
<td>This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.</td>
</tr>
<tr>
<td>Le nombre total de morts en octobre est le plus élevé depuis avril 2008, quand 1 073 personnes avaient été tuées.</td>
<td>The total number of deaths in May is the highest since April 2008, when 1 064 people had been killed.</td>
</tr>
<tr>
<td>À l’exception de l’opéra, la province reste le parent pauvre de la culture en France.</td>
<td>At an exception, opera remains of the state remains the poorest parent culture.</td>
</tr>
</tbody>
</table>
Why does it work?
Why does it work?

Early to say... but intuition:
Why does it work?

Early to say... but intuition:

• Mapped embedding space provides information for k-best possible translations

• Encoder-decoder figures out how to best “combine” them
Why does it work?

Early to say... but intuition:

• Mapped embedding space provides information for k-best possible translations

• Encoder-decoder figures out how to best “combine” them

• No need of magic
Conclusions

• New research area – unsupervised Machine Translation

The main Machine Translation competition (WMT18) has now an unsupervised track

• New papers are coming out, reporting 25 BLEU

• Code for replicability
  https://github.com/artetxem/undreamt

Final words

• **Word embeddings key** for Natural Language Processing
• Mappings represent *languages in common space*
  • Most of language pairs have **very few resources**
• New research area: **only monolingual resources**
Final words

- **Word embeddings key** for Natural Language Processing
- Mappings represent **languages in common space**
  - Most of language pairs have **very few resources**
  - New research area: **only monolingual resources**
- **Cross-lingual unsupervised mappings** enabled breakthroughs in
  - Bilingual dictionary induction
  - Unsupervised machine translation
  - Also (Conneau et al. 2018; Lample et al. 2018)
Final words

• **Word embeddings key** for Natural Language Processing
• Mappings represent **languages in common space**
  • Most of language pairs have **very few resources**
  • New research area: **only monolingual resources**
• **Cross-lingual unsupervised mappings** enabled breakthroughs in
  • Bilingual dictionary induction
  • Unsupervised machine translation
  • Also (Conneau et al. 2018; Lample et al. 2018)
• Unexplored area in its **infancy**
  • Potential for **MT in low resource languages and domains**
  • Potential for **transforming the NLP landscape**
    • From monolingual NLP (e.g. English) to multilingual tools
    • Universal sentence representations
Thank you!

@eagirre
http://ixa2.si.ehu.eus/eneko
https://github.com/artetxem/vecmap
https://github.com/artetxem/undreamt