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)
http://ixa.eus
Motivation

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

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

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

Arabic embeddings

Chinese embeddings
Overview

Arabic monolingual corpora

Chinese monolingual corpora
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)
  ▫ Bilingual embedding mappings (AAAI18)
  ▫ 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)
  ▫ Unsupervised statistical MT (EMNLP18)
  ▪ Conclusions
Outline

• Bilingual embedding mappings
  • *Introduction to vector space models (embeddings)*
  • *Bilingual embedding mappings (AAAI18)*
  • *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)*
  • *Unsupervised statistical MT (EMNLP18)*
  • *Conclusions*
Introduction to vector space models
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

Seed dictionary

Basque

English

X

Z

Miau
Marru
Banana
Udare

Katu
Behi
Sagar
Etxe

Zaunka
Txakur
Egutegi

Apple
Banana
Moo
Bark
Meow

Pear
Calendar
Cow
Dog
cat

House
Introduction to embedding mappings

Basque | Seed dictionary | English

- Txakur
- Sagar
- Egutegi

- Dog
- Apple
- Calendar
Introduction to embedding mappings

\[ W \]

Basque

Seed dictionary

English

- Txakur
- Sagar
- Egutegi

- Dog
- Apple
- Calendar
Introduction to embedding mappings

Basque

Seed dictionary

English

Txakur

Sagar

Egutegi

Dog

Apple

Calendar
Introduction to embedding mappings

Basque

Seed dictionary

English

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

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

- Txakur
- Sagar
- Egutegi

- Dog
- Apple
- Calendar
Introduction to embedding mappings

\[ \mathbf{X} \rightarrow \mathbf{W} \rightarrow \mathbf{Z} \]

Basque

Seed dictionary

English

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

Txakur
Sagar
Egutegi

Dog
Apple
Calendar
Introduction to embedding mappings

Mikolov et al. (2013b)

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

Txakur
Sagar
Egutegi

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

Dog
Apple
Calendar
Introduction to embedding mappings

\[ \text{arg min}_{W \in O(n)} \sum_{i} \|X_{i,*}W - Z_{i,*}\|^2 \]

Mikolov et al. (2013b)

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

Txakur, Sagar, Egutegi, ...
Introduction to embedding mappings

\[
\arg\min_{W \in \mathbb{R}^{d \times d}} \sum_{i} \|X_{i,*}W - Z_{j,*}\|_2^2
\]

Mikolov et al. (2013b)

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

Txakur, Sagar, Egutegi, ...
Introduction to embedding mappings

Mikolov et al. (2013b)

\[
\arg \min_{W \in \mathbb{R}^{n \times m}} \sum_{i} \| X_{i,*} W - Z_{j,*} \|^2
\]

Txakur, Sagar, ..., Egutegi

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

Basque

English

\begin{align*}
&\text{Miau} \quad \text{Marru} \\
&\text{Zaunka} \\
&\text{Katu} \quad \text{Behi} \quad \text{Sagar} \quad \text{Udare} \\
&\text{Txakur} \quad \text{Etxe} \\
&\text{Egutegi} \\
&\text{Banana} \\
\end{align*}

\begin{align*}
&\text{Udare} \quad \text{Sagar} \\
&\text{Pear} \quad \text{Apple} \\
&\text{Calendar} \\
&\text{House} \quad \text{Etxe} \\
&\text{Moo} \quad \text{Marru} \\
&\text{Bark} \quad \text{Miau} \quad \text{Zaunka} \\
&\text{Katu} \quad \text{Bark} \quad \text{Meow} \\
&\text{Meow} \quad \text{Egutegi} \\
&\text{Dog} \quad \text{Txakur} \quad \text{cat}
\end{align*}
State-of-the-art in supervised mappings
Artetxe et al. AAAI 2018

• Use 5000 sized seed bilingual dictionary
• Framework subsuming previous work, learns two mappings $W_X, W_Z$ as sequences of (optional) linear mappings:
  
  (opt.) Pre-process
  1. (opt.) Whitening
  2. **Orthogonal mapping**
  3. (opt.) Re-weighting
  4. (opt.) De-whitening
• The optional steps, properly combined, bring up to 5 points improvement
State-of-the-art in supervised mappings

S0 (opt.) Pre-processing: length normalization, mean centering
State-of-the-art in *supervised* mappings

Two sequences of (optional) linear transformations:

\[ W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)} \]

S0 (opt.) Pre-processing: length normalization, mean centering
State-of-the-art in supervised mappings

Two sequences of (optional) linear transformations:

\[ W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)} \]

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening: turn covariance matrices into the identity matrix

\[ W_{X(1)} = (X^T X)^{-0.5} \]
\[ W_{Z(1)} = (Z^T Z)^{-0.5} \]
State-of-the-art in **supervised** mappings

Two sequences of (optional) linear transformations:

\[
W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}
\]

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening: turn covariance matrices into the identity matrix

\[
W_{X(1)} = (X^T X)^{-0.5} \\
W_{Z(1)} = (Z^T Z)^{-0.5}
\]

S2 Orthogonal mapping: map into a shared space (Procrustes)

\[
W_{X(2)} = U \\
W_{Z(2)} = V \quad USV^T = X_{(1)}^T Z_{(1)}
\]
State-of-the-art in supervised mappings

Two sequences of (optional) linear transformations:

$$W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}$$

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening: turn covariance matrices into the identity matrix

$$W_{X(1)} = (X^T X)^{-0.5}$$
$$W_{Z(1)} = (Z^T Z)^{-0.5}$$

S2 Orthogonal mapping: map into a shared space (Procrustes)

$$W_{X(2)} = U \quad USV^T = X_{(1)}^T Z_{(1)}$$
$$W_{Z(2)} = V$$

S3 (opt.) Re-weight each component according to its cross-correlation

$$W_{X(3)} = S, \quad W_{Z(3)} = I$$
$$W_{X(3)} = I, \quad W_{Z(3)} = S$$
State-of-the-art in supervised mappings

Two sequences of (optional) linear transformations:

\[
W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}
\]

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening : turn covariance matrices into the identity matrix

\[
W_{X(1)} = (X^T X)^{-0.5} \\
W_{Z(1)} = (Z^T Z)^{-0.5}
\]

S2 Orthogonal mapping: map into a shared space (Procrustes)

\[
W_{X(2)} = U \\
W_{Z(2)} = V \quad USV^T = X_{(1)}^T Z_{(1)}
\]

S3 (opt.) Re-weight each component according to its cross-correlation

\[
W_{X(3)} = S, \quad W_{Z(3)} = I \\
W_{X(3)} = I, \quad W_{Z(3)} = S
\]

S4 (opt.) De-whitening: restore original variance in every direction

\[
W_{A(4)} = W_{B(4)}^T W_{B(1)}^{-1} W_{B(2)}
\]
State-of-the-art in supervised mappings

Two sequences of (optional) linear transformations:

\[ W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)} \]

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening: turn covariance matrices into the identity matrix

\[ W_{X(1)} = (X^T X)^{-0.5} \]
\[ W_{Z(1)} = (Z^T Z)^{-0.5} \]

S2 Orthogonal mapping: map into a shared space (Procrustes)

\[ W_{X(2)} = U \quad W_{Z(2)} = V \quad USV^T = X_{(1)}^T Z_{(1)} \]

S3 (opt.) Re-weight each component according to its cross-correlation

\[ W_{X(3)} = S, \quad W_{Z(3)} = I \]
\[ W_{X(3)} = I, \quad W_{Z(3)} = S \]

S4 (opt.) De-whitening: restore original variance in every direction

\[ W_{A(4)} = W_{B(2)}^T W_{B(1)}^{-1} W_{B(2)} \]

S5 (opt) Dimensionality reduction: keep the first \( n \) components only

\[ W_{X(5)} = W_{Z(5)} = (I_n \ 0)^T \]
State-of-the-art in *supervised* mappings

<table>
<thead>
<tr>
<th></th>
<th>S0 (l)</th>
<th>S0 (m)</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4 (src)</th>
<th>S4 (trg)</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OLS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mikolov et al. (2013)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>src</td>
<td>trg</td>
<td>trg</td>
<td></td>
</tr>
<tr>
<td>Shigeto et al. (2015)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>trg</td>
<td>src</td>
<td>src</td>
<td></td>
</tr>
<tr>
<td><strong>CCA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faruqui and Dyer (2014)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td><strong>Orth.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith et al. (2017)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
State-of-the-art in supervised mappings

<table>
<thead>
<tr>
<th>Method</th>
<th>S0 (l)</th>
<th>S0 (m)</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4 (src)</th>
<th>S4 (trg)</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>src</td>
<td>trg</td>
<td>trg</td>
<td></td>
</tr>
<tr>
<td>Mikolov et al. (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shigeto et al. (2015)</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>trg</td>
<td>src</td>
<td>src</td>
<td></td>
</tr>
<tr>
<td>CCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>src</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Faruqui and Dyer (2014)</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orth.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artetxe et al. (2016)</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith et al. (2017)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our method (AAAI18)</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>trg</td>
<td>trg</td>
</tr>
</tbody>
</table>
Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish
Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish
⇒ Monolingual embeddings (CBOW + negative sampling)
Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish
⇒ Monolingual embeddings (CBOW + negative sampling)
⇒ Seed dictionary: 5,000 word pairs
Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish
⇒ Monolingual embeddings (CBOW + negative sampling)
⇒ Seed dictionary: 5,000 pairs
⇒ Test dictionary: 1,500 pairs (Nearest neighbor, P@1)
Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish
⇒ Monolingual embeddings (CBOW + negative sampling)
⇒ Seed dictionary: 5,000 pairs
⇒ Test dictionary: 1,500 pairs (Nearest neighbor, P@1)

<table>
<thead>
<tr>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
</table>

76
Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish
⇒ Monolingual embeddings (CBOW + negative sampling)
⇒ Seed dictionary: 5,000 pairs
⇒ Test dictionary: 1,500 pairs (Nearest neighbor, P@1)

<table>
<thead>
<tr>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<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>Faruqui and Dyer (2014)</td>
<td>38.40*</td>
<td>37.13*</td>
<td>27.60*</td>
<td>26.80*</td>
</tr>
<tr>
<td>Shigeto et al. (2015)</td>
<td>41.53†</td>
<td>43.07†</td>
<td>31.04†</td>
<td>33.73†</td>
</tr>
<tr>
<td>Dinu et al. (2015)</td>
<td>37.7</td>
<td>38.93*</td>
<td>29.14*</td>
<td>30.40*</td>
</tr>
<tr>
<td>Lazaridou et al. (2015)</td>
<td>40.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87†</td>
<td>41.27†</td>
<td>28.23†</td>
<td>31.20†</td>
</tr>
<tr>
<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>Zhang et al. (2016)</td>
<td>36.73†</td>
<td>40.80†</td>
<td>28.16†</td>
<td>31.07†</td>
</tr>
<tr>
<td>Smith et al. (2017)</td>
<td>43.1</td>
<td>43.33†</td>
<td>29.42†</td>
<td>35.13†</td>
</tr>
</tbody>
</table>

† our publicly available reimplementations
Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish
⇒ Monolingual embeddings (CBOW + negative sampling)
⇒ Seed dictionary: 5,000 pairs
⇒ Test dictionary: 1,500 pairs (Nearest neighbor, P@1)

<table>
<thead>
<tr>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<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>Faruqui and Dyer (2014)</td>
<td>38.40*</td>
<td>37.13*</td>
<td>27.60*</td>
<td>26.80*</td>
</tr>
<tr>
<td>Shigeto et al. (2015)</td>
<td>41.53†</td>
<td>43.07†</td>
<td>31.04†</td>
<td>33.73†</td>
</tr>
<tr>
<td>Dinu et al. (2015)</td>
<td>37.7</td>
<td>38.93*</td>
<td>29.14*</td>
<td>30.40*</td>
</tr>
<tr>
<td>Lazaridou et al. (2015)</td>
<td>40.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Xing et al. (2015)</td>
<td>36.87†</td>
<td>41.27†</td>
<td>28.23†</td>
<td>31.20†</td>
</tr>
<tr>
<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>Zhang et al. (2016)</td>
<td>36.73†</td>
<td>40.80†</td>
<td>28.16†</td>
<td>31.07†</td>
</tr>
<tr>
<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><strong>Our method (AAAI18)</strong></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>
</tbody>
</table>
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 (to a large extent) isometric in word embedding space (♥)
Outline

• Bilingual embedding mappings
  - Introduction to vector space models (embeddings)
  - Bilingual embedding mappings (AAAI18)
• 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)
  - Unsupervised statistical MT (EMNLP18)
  - 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 for training bilingual signal

- parallel corpora
- comparable corpora
- (big) dictionaries
Reducing supervision

bilingual signal for training

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

Monolingual embeddings

Dictionary → 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 → Dictionary

better!
Self-learning

Monolingual embeddings

Dictionary -> Mapping -> Dictionary

better!

Mapping -> Dictionary
Self-learning

Monolingual embeddings


better!
eveng better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary → even better!

Dictionary → Mapping → Dictionary → better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary → better!

Mapping → Dictionary → even better!

Mapping → Dictionary
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary

Mapping

Dictionary

Mapping

Dictionary

Mapping

better!

even better!
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary

better!

Mapping → Dictionary

even better!

Mapping → Dictionary
Self-learning

Monolingual embeddings

Dictionary → Mapping → Dictionary → better!

Dictionary → Mapping → Dictionary → even better!

Dictionary → Mapping → Dictionary → even better!
Self-learning

Diagram:

- Monolingual embeddings
- Dictionary
- Mapping
- Dictionary

Arrow indicating flow from Dictionary to Monolingual embeddings, then to Mapping, and back to Dictionary.
Self-learning

Monolingual embeddings

Dictionary ➔ Mapping ➔ Dictionary

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 (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals
- Induce bilingual dictionary using self-learning for full vocabulary
- Evaluation
  - Compare translations to existing bilingual dictionary (*test* dictionary)
  - Accuracy
Semi-supervised experiments (ACL17)

Accuracy (%)

Seed dictionary size

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?

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

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

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

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

Implicit objective:

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

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

[Diagram showing data points before and after transformation]
Why does it work?

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

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

Implicit objective: 

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

\[ W^* = \arg \max_w \sum_l \max_j (x_{i*} \cdot w) \cdot z_{j*} \quad \text{s.t.} \quad WW^T = W^T W = I \]
Why does it work?

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

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

Implicit objective: \( W^* = \arg \max_W \sum_l \max_{j} (X_{i*}W) \cdot Z_{j*} \) \hspace{1cm} s.t. \( WW^T = W^TW = I \)
Why does it work?

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

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

Implicit objective:

\[ W^* = \arg \max_W \sum_i \max_j (X_i^* W) \cdot Z_j^* \hspace{1cm} \text{s.t.} \hspace{0.5cm} WW^T = W^T W = I \]
Why does it work?

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

Implicit objective: 

\[ W^* = \arg \max_w \sum_j \max_i (X_{i*}W) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^TW = I \]

Independent from seed dictionary!
Why does it work?

Implicit objective: 

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

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

Independent from seed dictionary!

So why do we need a seed dictionary?

Avoid poor local optima!
Why does it work?

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

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

Initial dictionary:

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

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

Initial dictionary:

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

Initial dictionary:

1. Compute intra-language similarity
2. 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:

1. Compute intra-language similarity
2. 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 had to add:

1. Stochastic dictionary induction
2. Frequency-based vocabulary cut-off
3. Hubness problem: Instead of inducing dictionary with nearest-neighbour use CSLS (Lample et al. 2018)

\[ 2\cos(x, y) - mnn_T(x) - mnn_S(y) \]

\[ mnn_T(x) = \frac{1}{K} \sum_{i=1}^{K} \cos(x, nn_i) \]
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>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

171
Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  \[ \Rightarrow \text{Monolingual embeddings (CBOW + negative sampling)} \]

<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)
  - 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>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 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>
</tr>
<tr>
<td>25 dict.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>45.27</td>
<td>44.13</td>
<td>32.94</td>
<td>36.60</td>
</tr>
<tr>
<td>25 dict.</td>
<td></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
  \[\Rightarrow\text{ Monolingual embeddings (CBOW + negative sampling)}\]
  \[\Rightarrow\text{ Seed dictionary: 5,000 word pairs / 25 word pairs / none}\]
  \[\Rightarrow\text{ Test dictionary: 1,500 word pairs}\]

<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>

178
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>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>Conneau et al. (2018)</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

Previous work convergence problems!
Also observed by Sogard et al. (2018)

<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></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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></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>
</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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5k dict.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></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>Zhang et al. (2017)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>None</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>
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
- Unsupervised matches results of supervised methods!
- Implicit optimization objective independent from seed dictionary
Conclusions

• Simple self-learning method to train bilingual embedding mappings
• Unsupervised matches results of supervised methods!
• 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
• Unsupervised matches results of supervised methods!
• 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
• Unsupervised matches results of supervised methods!
• 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
References: cross-lingual mappings


Outline

• Bilingual embedding mappings
  ▫ Introduction to vector space models (embeddings)
  ▫ Bilingual embedding mappings (AAAI18)
  ▫ 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)
  ▪ Unsupervised statistical MT (EMNLP18)
  ▪ 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 \((x_1...x_n, y_1...y_m)\)
  
  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 \(y_i\)
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

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

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.
Unsupervised neural machine translation

Training
- Supervised
- Autoencoder

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

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**
- **Denoising Autoencoder**

*Une 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**

There was a shooting in Los Angeles International Airport.

There a shooting was in Airport Los Angeles International.
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

Only 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 NMT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised neural machine translation

Only WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised NMT</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 neural machine translation

Only WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised NMT</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

Only WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised NMT</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

Only 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>Unsupervised NMT</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 NMT</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>

It can be easily combined with training data (interesting for low resource MT)
Unsupervised neural machine translation

Only WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised NMT</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>
<tr>
<td>Lample et al. 2018* (Same conference!)</td>
<td>14.31</td>
<td>15.06</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Experimental conditions are not exactly the same

State-of-the-art (not anymore...)

237
Unsupervised neural machine translation

Only WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised NMT</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><strong>15.56</strong></td>
<td><strong>15.13</strong></td>
<td><strong>10.21</strong></td>
<td><strong>6.55</strong></td>
</tr>
<tr>
<td>Lample et al. 2018</td>
<td>14.31</td>
<td>15.06</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lample et al. 2018b</td>
<td>15</td>
<td>13</td>
<td>10</td>
<td>6.55</td>
</tr>
</tbody>
</table>

Lample et al. 2018b (EMNLP)

- No embedding mappings
- BPE jointly over monolingual corpora. Fails for less related languages (Russian).
- Shared decoder for both languages
- Transformer (instead of LSTM)
Unsupervised neural machine translation

Only WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised NMT</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>
<tr>
<td>Lample et al. 2018</td>
<td>14.31</td>
<td>15.06</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lample et al. 2018b</td>
<td>24.2</td>
<td>25.1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Lample et al. 2018b (EMNLP)

- No embedding mappings
- BPE jointly over monolingual corpora. Fails for less related languages (Russian).
- Shared decoder for both languages
- Transformer (instead of LSTM)
Unsupervised neural machine translation

Only WMT released data (test and monolingual corpora)

<table>
<thead>
<tr>
<th>Unsupervised NMT</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>
<tr>
<td>Lample et al. 2018</td>
<td>14.31</td>
<td>15.06</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lample et al. 2018b</td>
<td>24.2</td>
<td>25.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lample and Conneau 2019</td>
<td><strong>33.3</strong></td>
<td><strong>33.4</strong></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Lample and Conneau 2019 (Arxiv)

- Similar to 2018b
- Pre-train encoder and decoder on (masked) language modeling
- And ... larger batch sizes (+6 Bleu!)
Unsupervised statistical machine translation
Unsupervised statistical machine translation

Artetxe et al. 2018b (EMNLP)

• Estimate PBMT parameters
  • Learn monolingual embeddings for bigrams and trigrams
  • Initialize phrase table using prob. estimates from cross-lingual mappings
  • Unsupervised tuning based on back-translation
• Use backtranslation and train reverse PBMT from scratch. Iterate.
Unsupervised statistical machine translation

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

<table>
<thead>
<tr>
<th>Method</th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised NMT</td>
<td>Artetxe et al. 2018</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lample and C. 2019</td>
<td>33.3</td>
<td>33.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsupervised PBMT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Artetxe et al. 2018b (EMNLP)

• Estimate PBMT parameters
  • Learn monolingual embeddings for bigrams and trigrams
  • Initialize phrase table using prob. estimates from cross-lingual mappings
  • Unsupervised tuning based on back-translation
• Use backtranslation and train reverse PBMT from scratch. Iterate.
Unsupervised statistical machine translation

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

<table>
<thead>
<tr>
<th>Method</th>
<th>Artetxe et al. 2018</th>
<th>Artetxe et al. 2018b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised NMT</td>
<td>15.56 15.13 10.21 6.55</td>
<td>25 87 26 22 17 43 14 08 23 05 18 23</td>
</tr>
<tr>
<td>Unsupervised PBMT</td>
<td>33.3 33.4 34.3 26.2</td>
<td>25.87 26.22 17.43 14.08 23.05 18.23</td>
</tr>
</tbody>
</table>

Artetxe et al. 2018b (EMNLP)

• Estimate PBMT parameters
  • Learn monolingual embeddings for bigrams and trigrams
  • Initialize phrase table using prob. estimates from cross-lingual mappings
  • Unsupervised tuning based on back-translation
• Use backtranslation and train reverse PBMT from scratch. Iterate.
Unsupervised statistical machine translation

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised NMT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artetxe et al. 2018</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lample and C. 2019</td>
<td>33.3</td>
<td>33.4</td>
<td></td>
<td></td>
<td>34.3</td>
<td>26.2</td>
</tr>
<tr>
<td>Unsupervised PBMT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artetxe et al. 2018b</td>
<td>25.87</td>
<td>26.22</td>
<td>17.43</td>
<td>14.08</td>
<td>23.05</td>
<td>18.23</td>
</tr>
<tr>
<td>Lample et al. 2018b</td>
<td>27.16</td>
<td>28.11</td>
<td></td>
<td></td>
<td>22.68</td>
<td>17.77</td>
</tr>
</tbody>
</table>

Artetxe et al. 2018b (EMNLP)

• Estimate PBMT parameters
  • Learn monolingual embeddings for bigrams and trigrams
  • Initialize phrase table using prob. estimates from cross-lingual mappings
  • Unsupervised tuning based on back-translation
• Use backtranslation and train reverse PBMT from scratch. Iterate.
Unsupervised statistical machine translation

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
<th>DE-EN</th>
<th>EN-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsupervised NMT</td>
<td>Artetxe et al. 2018</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lample and C. 2019</td>
<td>33.3</td>
<td>33.4</td>
<td>34.3</td>
<td>26.2</td>
<td></td>
</tr>
<tr>
<td>Unsupervised PBMT</td>
<td>Artetxe et al. 2018b</td>
<td>25.87</td>
<td>26.22</td>
<td>17.43</td>
<td>14.08</td>
<td>23.05</td>
</tr>
<tr>
<td></td>
<td>Lample et al. 2018b</td>
<td>27.16</td>
<td>28.11</td>
<td>22.68</td>
<td>17.77</td>
<td></td>
</tr>
<tr>
<td>NMT +</td>
<td>Lample et al. 2018b</td>
<td>27.7</td>
<td>27.6</td>
<td>25.2</td>
<td>20.2</td>
<td></td>
</tr>
<tr>
<td>PBMT</td>
<td>Artetxe et al. SUBM</td>
<td>33.5</td>
<td>36.2</td>
<td>27.0</td>
<td>22.5</td>
<td>34.4</td>
</tr>
</tbody>
</table>

**Combinations** improve further

UMT is at the level MT was at 2014
Unsupervised machine translation

Getting closer to supervised machine translation!

Figure 2: Comparison between supervised and unsupervised approaches on WMT’14 En-Fr, as we vary the number of parallel sentences for the supervised methods.

Source: (Lample et al. 2018)
Unsupervised machine translation

Getting closer to supervised machine translation!

Figure 2: Comparison between supervised and unsupervised approaches on WMT’14 En-Fr, as we vary the number of parallel sentences for the supervised methods.

Source: (Lample et al. 2018)
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
• NMT / PBMT figures out how to best “combine” them
Conclusions

• New research area – unsupervised Machine Translation

The main Machine Translation competition (WMT) has now an **unsupervised subtrack**

• Performance up, 33 BLEU En-Fr

• Plenty of margin for improvement

• Code for replicability
  
  https://github.com/artetxem/undreamt
  https://github.com/artetxem/monoses (soon)
References: unsupervised MT


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
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
- 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
https://github.com/artetxem/monoses