Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings

Yuan Zhang, David Gaddy, Regina Barzilay, Tommi Jaakkola

MIT, CSAIL
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

98.2%

Supervised
700k tokens
(Brants, 2000)
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

- Multilingual Transfer
  2m parallel sentences (Das and Petrov, 2011)
  82.8

- Supervised
  700k tokens (Brants, 2000)
  98.2
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

Prototype-driven
14 prototypes
(Haghighi et al., 2006)

25.5

Multilingual Transfer
2m parallel sentences
(Das and Petrov, 2011)

82.8

Supervised
700k tokens
(Brants, 2000)

98.2
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Tagging Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype-driven (14 prototypes)</td>
<td>25.5</td>
</tr>
<tr>
<td>Multilingual Transfer (10 Translation Pairs, No parallel sentences)</td>
<td>?</td>
</tr>
<tr>
<td>Multilingual Transfer (2m parallel sentences)</td>
<td>82.8</td>
</tr>
<tr>
<td>Supervised (700k tokens)</td>
<td>98.2</td>
</tr>
</tbody>
</table>

(Haghighi et al., 2006)
(Das and Petrov, 2011)
(Brants, 2000)
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Accuracy</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype-driven 14 prototypes</td>
<td>25.5</td>
<td>(Haghighi et al., 2006)</td>
</tr>
<tr>
<td>Multilingual Transfer Ten Translation Pairs</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Multilingual Transfer 2m parallel sentences</td>
<td>82.8</td>
<td>(Das and Petrov, 2011)</td>
</tr>
<tr>
<td>Supervised 700k tokens</td>
<td>98.2</td>
<td>(Brants, 2000)</td>
</tr>
</tbody>
</table>

How little parallel data is necessary to enable multilingual transfer?
Our Work

• Task: multilingual transfer of part-of-speech (POS) tagging
• Data:

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>
| Unlabeled| ✓      | ✓      | *(non-parallel data)*
Our Work

• Task: multilingual transfer of part-of-speech (POS) tagging
• Data:

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

(non-parallel data)

Ten Translation Pairs

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td></td>
</tr>
<tr>
<td>,</td>
<td></td>
</tr>
<tr>
<td>der</td>
<td></td>
</tr>
<tr>
<td>die</td>
<td></td>
</tr>
<tr>
<td>in</td>
<td></td>
</tr>
</tbody>
</table>
Our Work

- Task: multilingual transfer of part-of-speech (POS) tagging
- Data:

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Ten Translation Pairs**

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>,</td>
<td>,</td>
</tr>
<tr>
<td>der</td>
<td>der</td>
</tr>
<tr>
<td>die</td>
<td>die</td>
</tr>
<tr>
<td>in</td>
<td>in</td>
</tr>
<tr>
<td>und</td>
<td>und</td>
</tr>
<tr>
<td>dem</td>
<td>dem</td>
</tr>
<tr>
<td>von</td>
<td>von</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>zu</td>
<td>zu</td>
</tr>
</tbody>
</table>

**POS Accuracy on German**

- Prototype: 25.5%
- Ours: 68.7%

(Haghighi et al., 2006)
1. Learn **coarse mapping** between embeddings via ten translation pairs

2. Refine embedding transformations and model parameters via **unsupervised learning** on the target language
Coarse Mapping between Embeddings

- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs

**Monolingual Embedding**

*Source (English)*
- red
- cat
- is

*Target (German)*
- Katze (cat)
- ist (is)
- Hund (dog)
- rot (red)
Coarse Mapping between Embeddings

- **Goal:** find a linear transformation from target to source embedding space
- **Objective:** minimize the distance between translation pairs

**Monolingual Embedding**

<table>
<thead>
<tr>
<th>Source (English)</th>
<th>Target (German)</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>Hund</td>
</tr>
<tr>
<td>cat</td>
<td>Katze</td>
</tr>
<tr>
<td>is</td>
<td>ist</td>
</tr>
</tbody>
</table>

**Translation Pairs**

- dog || Hund
- cat || Katze
- red || rot
Coarse Mapping between Embeddings

- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs

**Monolingual Embedding**

**Source (English)**
- red
- cat
- is

**Target (German)**
- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot

Too many degrees of freedom

- dimension: 20
- # pairs: 10
- degree of freedom: 10
Coarse Mapping between Embeddings

- **Goal:** find a linear transformation from target to source embedding space
- **Objective:** minimize the distance between translation pairs

---

**Monolingual Embedding**

**Source (English)**
- dog
- cat
- is

**Target (German)**
- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot

---

**Too many degrees of freedom**
- dimension: 20
- # pairs: 10
- degree of freedom: 10

**Solutions need to be constrained!**
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

Monolingual Embedding

Source (English)

- red
- cat
- is

Target (German)

- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

Isometric Constraints

$P^T P = I$

Translation Pairs

dog || Hund
| cat || Katze
| red || rot
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations
  \[
  \cos \langle \text{cat, dog} \rangle \approx \cos \langle \text{Katze, Hund} \rangle, \quad \cos \langle \text{dog, red} \rangle \approx \cos \langle \text{Hund, rot} \rangle
  \]

**Monolingual Embedding**

**Source(English)**
- dog
- cat
- is
- red

**Target(German)**
- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

**Isometric Constraints**

\[
P^T P = I
\]

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

**Monolingual Embedding**

**Source (English)**
- dog
- red
- is

**Target (German)**
- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

**Isometric Solution**

**Isometric Constraints**

$P^T P = I$

**Translation Pairs**

- dog $||$ Hund
- cat $||$ Katze
- red $||$ rot
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

**Monolingual Embedding**

**Source (English)**
- dog
- cat
- is
- red

**Target (German)**
- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

**Isometric Solution**

### Isometric Constraints

$P^T P = I$

### Translation Pairs

- dog || Hund
- cat || Katze
- red || rot
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix.
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations.
- Use the steepest descent algorithm (Abrudan et al., 2008).

**Monolingual Embedding**

**Source (English)**
- dog
- red
- is

**Target (German)**
- Hund (dog)
- Katze (cat)
- rot (red)
- ist (is)

**Isometric Solution**

- Hund (dog)
- Katze (cat)
- red (red)
- ist (is)

**Isometric Constraints**

$$P^T P = I$$

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot
Validation of Isometric Constraints

- Validation for $\cos\langle\text{cat, dog}\rangle \approx \cos\langle\text{Katze, Hund}\rangle$
- Verify whether nearest neighbors are preserved after translations

For 50% of word pairs, $k \leq 2$
Validation of Isometric Constraints

• Validation for $\cos\langle\text{cat, dog}\rangle \approx \cos\langle\text{Katze, Hund}\rangle$
• Verify whether nearest neighbors are preserved after translations

**English: nearest neighbor**

For 50% of word pairs, $k \leq 2$

**German: $k$-th ($k\leq 2$) nearest neighbor?**

**English: nearest neighbor**

For 90% of word pairs, $k \leq 10$

**German: $k$-th ($k\leq 10$) nearest neighbor?**
Direct Transfer Model

- Supervised source language HMM
  - Feature-based HMM (Berg-Kirkpatrick et al., 2010)
  - Word embeddings as emission features
Direct Transfer Model

- Supervised source language HMM
  - Feature-based HMM (Berg-Kirkpatrick et al., 2010)
  - Word embeddings as emission features

Source

\[ p(x|y) \propto \exp\{v_x^T \mu_y\} \]

Target

\[ p^{dt}(x|y) \propto \exp\{v_x^T P \mu_y\} \]

Coarse mapping is not accurate
Our Two-phase Method

1. Learn **coarse mapping** between embeddings via ten translation pairs

2. Refine embedding transformations and model parameters via **unsupervised learning** on the target language
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$

\[ p(x|y) \propto \exp\{v_x^T PM \mu_y + \theta_{x,y}\} \]
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$.

$$
p(x|y) \propto \exp\{v^T_x PM \mu_y + \theta_{x,y}\}
$$
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$.

$$p(x|y) \propto \exp\{\nu_x^T PM \mu_y + \theta_{x,y}\}$$
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$.

$$p(x|y) \propto \exp\{\nu^T_x P M \mu_y + \theta_{x,y}\}$$
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$

$$p(x|y) \propto \exp\{\nu_x^T PM \mu_y + \theta_{x,y}\}$$
Learning

• Parameters: $\mu_y, \theta_{y,y'}, M, \theta_{x,y}$

• Optimization method: standard Expectation-Maximization (EM)
  ✦ E-step: forward-backward
  ✦ M-step: gradient ascent using L-BFGS
Experimental Setup

• Datasets: Universal Dependencies Treebanks v1.2
  ♦ Source: English
  ♦ Target (Indo-European): Danish (da), German (de), Spanish (es)
  ♦ Target (non-Indo-European): Finnish (fi), Hungarian (hu), Indonesian (id)

• Universal tagset: 14 tags (noun, verb, adjective etc.)

• Word embeddings: 20-dimension vectors trained on Wiki dumps using word2vec
Indo-European Results

Averaged Accuracy on Indo-European Languages

- Prototype (Haghighi et al., 2006): 31.8
- Direct Transfer: 60.9
- Transfer+EM (Ours full): 72.9
Non-Indo-European Results

Averaged Accuracy on non-Indo-European Languages

- Prototype (Haghighi et al., 2006): 27.6%
- Direct Transfer: 57.7%
- Transfer+EM (Ours full): 62.1%
Prediction of Linguistic Typology

- Task: predict whether a language is verb-object or object-verb (five typological properties)
- Features: bigrams and trigrams of POS tags
Impact of Amount of Supervision

- Transfer+EM with 10 pairs = 150 prototypes
Impact of Amount of Supervision

- Transfer+EM with 10 pairs = 150 prototypes
- Prototype improves with large amount of annotations
Impact of Amount of Supervision

- Transfer+EM with 10 pairs = 150 prototypes
- Prototype improves with large amount of annotations
- Transfer+EM consistently improves over Direct Transfer

![Accuracy on German](image)
Conclusion

• Ten translation pairs are sufficient to enable multilingual transfer of POS tagging

• Our model significantly outperforms the direct transfer and the prototype-driven method

Source code available at: https://github.com/yuanzh/transfer_pos
Thank You!
Impact of Embedding Dimensions and Window Size

![Impact of Embedding Dimensions and Window Size](image_url)

- Accuracy: 55, 60, 65, 70
- Dimensions: 10, 20, 50, 100, 200
- Windows: window=1, window=5