Learning Structural Correspondences Using Synchronous Neural Language Models

Stephan Gouws\textsuperscript{1}, GJ van Rooyen\textsuperscript{1} and Yoshua Bengio\textsuperscript{2}

\textsuperscript{1}Stellenbosch University
and
\textsuperscript{2}University of Montreal
The oncogenic mutated forms of the ras proteins are constitutively active and interfere with normal signal transduction.

The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.
The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.

The oncogenic mutated forms of the ras proteins are constitutively active and interfere with normal signal transduction.
Domain adaptation

- How can we train a system on a source distribution to perform well on a target distribution?

**Diagram:**
- **Source** (Lots of labels)
- **Target** (Little/no labels)
- **Classifier**: POS, NER..
Domain adaptation

• How can we train a system on a source distribution to perform well on a target distribution?
Domain adaptation

- How can we train a system on a source distribution to perform well on a target distribution?
The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.
The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.
### English Language (WSJ)

<table>
<thead>
<tr>
<th>DT</th>
<th>NN</th>
<th>VBZ</th>
<th>DT</th>
<th>NN</th>
<th>IN</th>
<th>DT</th>
<th>JJ</th>
<th>NN</th>
<th>CC</th>
</tr>
</thead>
</table>
| The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.

### Different Language (Afrikaans)

<table>
<thead>
<tr>
<th>??</th>
<th>??</th>
<th>??</th>
<th>??</th>
<th>??</th>
<th>??</th>
<th>??</th>
<th>??</th>
<th>??</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die botsing is ’n teken van ’n nuwe taaiheid en verdeeldheid in Japan se eens knus finansiële kringe.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Part-of-Speech Tagging**

**English Language (WSJ)**

<table>
<thead>
<tr>
<th>DT</th>
<th>NN</th>
<th>VBZ</th>
<th>DT</th>
<th>NN</th>
<th>IN</th>
<th>DT</th>
<th>JJ</th>
<th>NN</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Different Language (Afrikaans)**

<table>
<thead>
<tr>
<th>DT</th>
<th>??</th>
<th>VBZ</th>
<th>DT</th>
<th>??</th>
<th>IN</th>
<th>DT</th>
<th>??</th>
<th>??</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die botsing is ‘n teken van ‘n nuwe taaiheid en verdeeldheid in Japan se eens knus finansiële kringe.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SOURCE**

**TARGET**
The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.
The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.

Die botsing is ’n teken van ’n nuwe taaiheid en verdeeldheid in Japan se eens knus finansiële kringe.
Structural Correspondence Learning [Blitzer et al. 2006]

- $\Phi$ should make the domains look as similar as possible
- But should also allow us to classify well
- $\Phi$ defines a \textit{linear} mapping between feature spaces of both domains
This Work: Deep Structural Correspondence Learning

We use Neural Language Models [Bengio et al., 2003]
This Work: Deep Structural Correspondence Learning

- Embedding Space $\mathbf{R}_1$
- Embedding Space $\mathbf{R}_2$
- Model $f_1(\mathbf{R}_1, \Theta_1)$
- Model $f_2(\mathbf{R}_2, \Theta_2)$
- Data from domain 1
- Data from domain 2

B: Correspondence Learning
A: Dimensionality Reduction
**This Work: Deep Structural Correspondence Learning**

- Embedding Space $\mathbb{R}_1$
  - Model $f_1(\mathbb{R}_1, \Theta_1)$
  - Data from domain 1
- Embedding Space $\mathbb{R}_2$
  - Model $f_2(\mathbb{R}_2, \Theta_2)$
  - Data from domain 2

Constrain models to be as similar as possible

Constrain representations of **pivot pairs** to be as similar as possible
This Work: *Deep Structural Correspondence Learning*

**Hypothesis**: Aligning a *well-chosen subset* of the vocabularies ("pivot pairs") w.r.t. each other, will iteratively align the rest of the words in the 2 vocabularies in a meaningful way w.r.t. each other in the shared feature space.
The Log Bilinear NLM

[Mcih & Hinton 2007]

\[ \hat{r} = \sum_i C_i r_{w_i} \]

- Multinomial over output vocabulary
- Predicted next word’s embedding vector
- Previous 4 words’ low-dim embedding vectors
- Previous 4 words’ high-dim representation
- \( P(w_1) \) to \( P(w_{|V|}) \)

4\(|V|\)-length Embedding layer

|\(|V|\)-length Softmax layer

```
000000..1..0000
```

```
| "cat" |
| "sits" |
| "on" |
| "the" |
```

```
```

```
```

```
```

```
```

```
```
Augmented cost function

\[ J_{TOT} = J_{1}^{nll} + J_{2}^{nll} \]

Fit the data well
Augmented cost function

\[ J^{TOT} = J_1^{nll} + J_2^{nll} + \alpha J^R(\theta_1, \theta_2) + \beta J^f(\theta_1, \theta_1) \]

Fit the data well

Learn structural similarities
Augmented cost function

$$J_{TOT} = J_{1}^{nll} + J_{2}^{nll} + \alpha J_{R}^{R}(\theta_1, \theta_2) + \beta J_{f}^{f}(\theta_1, \theta_1)$$

- Fit the data well
- Learn structural similarities
Constraining the Embeddings

$$J^R(\theta_1, \theta_2) = \sum_{r_i, r_j \in \text{Pivots}} \lambda_{ij} \| r_i - r_j \|^2$$

Push known pivot pairs \((r_i, r_j) \in V_1 \times V_2\) closer together by minimizing the weighted sum of squared distances between them.
Augmented cost function

\[ J_{TOT} = J_{1}^{ll} + J_{2}^{ll} + \alpha J^{R}(\theta_{1}, \theta_{2}) + \beta J^{f}(\theta_{1}, \theta_{1}) \]

Fit the data well

Learn structural similarities
Constraining the learned Functions

• Intuitively: In the cases where it is appropriate, we constrain the learned NLM functions $f$ to be as similar as possible while modelling the data as accurately as possible
  – Reduces degrees of freedom

• [Erhan et al. 2010]
Experiments I: Synthetic Data

- Sampled two datasets each **without** replacement from LA Times, encoded each in a **different vocabulary**
  - “the_1 president_1 of_1 the_1 united_1 ...”
  - “between_2 the_2 hours_2 of_2 midnight_2...”
- Give the networks $k$ % of pivot pairs (chose most frequent words)
  - (president_1, president_2), ...

**EVALUATE**: We measure the similarity of vectors for all translation pairs
Experiments I: Synthetic Data

![Graph showing the average distance between all pivots over training time for different pivot percentages. The x-axis represents training time in units of 10^3, ranging from 0 to 5. The y-axis represents the average distance between pivots, ranging from 0.75 to 1.15. There are four lines: green for 1% pivots, blue for no pivots, red for 5% pivots, and pink for 10% pivots. The graph illustrates how the average distance between pivots decreases as training time increases.]
Experiments II: English-French

- Sampled two non-parallel datasets from English and French newswire text
- Provided models with list of $k\%$ translation pairs

**EVALUATE**: We measure the similarity of vectors for all translation pairs
Experiments II: English-French

- No pivots
- 1% Pivots
- 10% Pivots, function constraint
- 10% Pivots, no function constraint

Graph showing the average distance between all pivots over training (minibatches) for different pivot percentages and conditions.
Experiments II: English-French

NLL of individual models on validation: 4.56 (EN) and 4.09 (FR)
NLL of coupled models on validation: 4.57 (EN) and 4.10 (FR)
Conclusion

• We investigated the hypothesis that NLMs can learn similar features for similar words in two domains, given a well-chosen prior subset of pivot words in the two domains.
• Our results indicate that even 5% pivot pairs are sufficient to start convergence.
• Previous work has shown these features to be useful for POS-tagging and NER in the single-domain setting [Turian et al, 2010]
• Future work will evaluate these learned features in multi-domain sequence-tagging tasks.
Thank you!