DATNet: Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition

DATNet: Background

Named entity recognition, also known as NER, classifies named entities that are present in a text into pre-defined categories like person, organization, location, dates, etc.

NER is challenging and detects not only the type of named entity, but also the entity boundaries, which requires deep understanding of contextual semantics to disambiguate the different entity types of same tokens.
**DATNet: Background**

### Traditional Method for NER
- Conditional Random Field (CRF), Support Vector Machine (SVM), Perceptron, etc.
- Hand-craft features by expertise.

- **Drawbacks:** require a lot of domain-knowledge to design features.

### Deep Learning for NER
- Deep Neural Nets (DNN), Convolutional Nets (CNN), Recurrent Nets (RNN), etc.
- Requires little feature engineering and domain knowledge.

- **Limitations:** mass of data is required for better generalization ability.

### Transfer Learning for Low-resource NER
- When annotated corpora is small, NN-based methods degrade significantly, since hidden features cannot be learned adequately.
- *Transfer learning* is a way to overcome such obstacle by borrowing knowledge from other resources.
Although the existing transfer-based methods show promising performance in low-resource settings. There are two issues deserved to be further investigated on:

1. **Representation Difference**: They did not consider the representation difference across source and target in different scenarios (Cross-languages/domains).
2. **Resource Data Imbalance**: the training size of high-resource is usually much larger than that of low-resource.

Most existing methods ignore the above two issues in their models, thus resulting in poor generalization.

*The Dual Adversarial Transfer Nets (DATNet) is proposed to solve these two issues.*
Representation Difference

**Partially Share (DATNet-P) and Fully Share (DATNet-F)**

DATNet-P decomposes the BiLSTM units into the **shared** component and the **private** one.

In DATNet-F, the BiLSTM units are **fully shared** by both resources while word embeddings for different resources are disparate.

In the experiment, we will investigate the performance of two different shared representation architectures on different tasks and give their corresponding recommendation.
**DATNet: Experiments**

In this experiment, CoNLL-2003 English NER is source data, CoNLL-2002 and WNUT are target data.

**Cross-language transfer:** CoNLL-2003 $\rightarrow$ CoNLL-2002

**Cross-domain transfer:** CoNLL-2003 $\rightarrow$ WNUT

1. **DATNet-P** model advocates Cross-language.
2. **DATNet-F** model advocates Cross-domain.

---

**Comparison with State-of-the-art Results in CoNLL and WNUT datasets (F1-score).**

<table>
<thead>
<tr>
<th>Mode</th>
<th>Methods</th>
<th>Additional Features</th>
<th>CoNLL Datasets</th>
<th>WNUT Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>POS Gazetteers Orthographic</td>
<td>Spanish</td>
<td>Dutch</td>
</tr>
<tr>
<td>Mono-language /domain</td>
<td>Gillick et al. [74]</td>
<td>$\times$ $\times$ $\times$</td>
<td>82.59</td>
<td>82.84</td>
</tr>
<tr>
<td></td>
<td>Lample et al. [4]</td>
<td>$\times$ $\sqrt{}$ $\times$</td>
<td>85.75</td>
<td>81.74</td>
</tr>
<tr>
<td></td>
<td>Partalas et al. [67]</td>
<td>$\sqrt{}$ $\sqrt{}$ $\sqrt{}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Limspopatham et al. [68]</td>
<td>$\times$ $\times$ $\sqrt{}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Lin et al. [75]</td>
<td>$\sqrt{}$ $\sqrt{}$ $\times$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>Our Base Model</strong></td>
<td>Best Mean &amp; Std</td>
<td>85.53</td>
<td>85.55</td>
</tr>
</tbody>
</table>

**Cross-language /domain**

|                 | Yang et al. [13]   | $\times$ $\sqrt{}$ $\times$ | 85.77 | 85.19 | 47.19* | 40.83* |
|                 | Ying et al. [35]  | $\times$ $\sqrt{}$ $\times$ | 85.88 | 86.55 | 46.53* | 40.79* |
|                 | Feng et al. [21]   | $\sqrt{}$ $\times$ $\times$ | 86.42 | **88.39** |
|                 | Von et al. [76]    | $\times$ $\sqrt{}$ $\times$ | - | - | - | - |
|                 | Aguilar et al. [33] | $\sqrt{}$ $\times$ $\sqrt{}$ | - | - | 40.78 | - |
| **DATNet-P**    | Best Mean & Std    | $\times$ $\times$ $\times$ | **88.16** | 88.32 | 50.85 | 41.12 |
| **DATNet-F**    | Best Mean & Std    | $\times$ $\times$ $\times$ | 87.89±0.18 | 88.09±0.13 | 50.41±0.32 | 40.52±0.38 |

- The scores with “*” denote produced results by the corresponding official tools/codes.
Transfer Learning Performance

- The transfer learning component in the DATNet consistently improves over the results of the base model and the improvement margin is more distinct when the target data ratio is lower.

- DATNet-F outperforms DATNet-P on cross-language transfer when the target resource is extremely low, however, this results are reversed when the target dataset size is large enough (i.e., more than 100 sentences);

- DATNet-F is generally superior to DATNet-P on cross-domain transfer.
**Resource Data Imbalance**

**Generalized Resource-Adversarial Discriminator (GRAD)**

GRAD takes self-attention output and computes the resource label. Its loss is defined as

\[
\ell_{GRAD} = - \sum_i \{I_i \in D \} \alpha (1 - r_i)^\gamma \log r_i + \sum_i \{I_i \in D \} (1 - \alpha) r_i^\gamma \log (1 - r_i)
\]

\(\alpha\) is a weighting factor to balance the loss contribution from high and low resource.

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>0.1</th>
<th>0.15</th>
<th>0.2</th>
<th>0.25</th>
<th>0.3</th>
<th>0.35</th>
<th>0.4</th>
<th>0.45</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio CoNLL-2002 Spanish NER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\rho = 0.1)</td>
<td>78.37</td>
<td>78.63</td>
<td><strong>78.70</strong></td>
<td>78.32</td>
<td>77.96</td>
<td>77.92</td>
<td>77.88</td>
<td>77.78</td>
<td>77.85</td>
<td>77.90</td>
<td>77.65</td>
<td>77.57</td>
<td>77.38</td>
<td>77.49</td>
<td>77.29</td>
</tr>
<tr>
<td>(\rho = 0.2)</td>
<td>80.99</td>
<td>81.71</td>
<td><strong>82.18</strong></td>
<td>81.57</td>
<td>81.53</td>
<td>81.55</td>
<td>81.44</td>
<td>81.25</td>
<td>81.32</td>
<td>81.16</td>
<td>81.02</td>
<td>81.16</td>
<td>80.63</td>
<td>80.79</td>
<td>80.54</td>
</tr>
<tr>
<td>(\rho = 0.4)</td>
<td>83.76</td>
<td>83.73</td>
<td>84.18</td>
<td><strong>84.48</strong></td>
<td>84.26</td>
<td>84.12</td>
<td>83.54</td>
<td>83.40</td>
<td>83.52</td>
<td>84.18</td>
<td>83.42</td>
<td>83.47</td>
<td>83.28</td>
<td>83.33</td>
<td>83.19</td>
</tr>
<tr>
<td>(\rho = 0.6)</td>
<td>85.18</td>
<td>85.24</td>
<td>85.85</td>
<td>85.68</td>
<td>85.84</td>
<td><strong>86.10</strong></td>
<td>85.71</td>
<td>85.74</td>
<td>85.42</td>
<td>85.60</td>
<td>85.20</td>
<td>85.40</td>
<td>85.26</td>
<td>85.24</td>
<td>84.98</td>
</tr>
</tbody>
</table>

Table 5: Analysis of Discriminator Weight \(\alpha\) in GRAD with Varying Data Ratio \(\rho\) (F1-score).

[1] Focal Loss for Dense Object Detection, ICCV 2017
Resource Data Imbalance

Generalized Resource-Adversarial Discriminator (GRAD)

\[
\ell_{GRAD} = - \sum_i \{ I_i \in D_S \alpha (1 - r_i)^\gamma \log r_i \\
+ I_i \in D_T (1 - \alpha) r_i^\gamma \log (1 - r_i) \}
\]

\((1 - r_i)^\gamma \) (or \(r_i^\gamma\)) controls the loss contribution from individual samples by measuring the discrepancy between prediction and true label (easy samples have smaller contribution).

- For the sample from the high resource \(D_s\), its corresponding loss term is \(I_i \in D_S \alpha (1 - r_i)^\gamma \log r_i\), where the controlling factor \((1 - r_i)^\gamma\) is inverse proportion to \(r_i\). In other words, \(r_i \to 1\), this well-classified sample is down-weighted due to \((1 - r_i)^\gamma\) goes to 0. As \(\gamma\) increases, the approaching speed increases. In this case, for sample from high resource data, a large \(\gamma\) is preferred.
- On the contrary, for the sample from low resource data, a small \(\gamma\) is preferred.
Ablation Study of GRAD

This experiment reports the quantitative performance comparison between models with different components.

1. GRAD shows the stable superiority over the normal AD regardless of other components.

The recommendation of $\gamma = 2$ for GRAD in practical use.
DATNet: Feature Visualization

The visualization of extracted features from shared bidirectional-LSTM layer. The left, middle, and right figures show the results when no Adversarial Discriminator (AD), AD, and GRAD is performed, respectively. Red points correspond to the source CoNLL-2003 English examples, and blue points correspond to the target CoNLL-2002 Spanish examples.

GRAD in DATNet makes the distribution of extracted features from the source and target datasets much more similar by considering the data imbalance, which indicates that the outputs of BiLSTM are resource-invariant.
**DATNet: Adversarial Training (AT)**

Adversarial samples are widely incorporated into training to **improve the generalization and robustness of the model**, which is called adversarial training. It emerges as a powerful regularization tool to **stabilize training and enable the model to escape from the local minimum**.

An adversarial sample is built by adding the original sample with a perturbation bounded by a small norm $\varepsilon$ to maximize the loss function as

$$\eta_x = \arg \max_{\eta: \|\eta\|_2 \leq \varepsilon} \ell(\Theta; x + \eta)$$

where $\Theta$ is the current model parameters set. $\eta$ is estimated by

$$\eta_x = \varepsilon \frac{g}{\|g\|_2}, \text{ where } g = \nabla \ell(\Theta; x)$$
DATNet: Experiments  Ablation Study

The aforementioned results show AT helps to enhance the overall performance by adding perturbations into inputs with the limit of $\epsilon = 5$.

This experiment indicates that less training data require a larger $\epsilon$ to prevent over-fitting, which further validates the necessity of AT in the case of low resource data.

<table>
<thead>
<tr>
<th>$\epsilon_wT$</th>
<th>1.0</th>
<th>3.0</th>
<th>5.0</th>
<th>7.0</th>
<th>9.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>CoNLL-2002</td>
<td>Spanish NER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.1$</td>
<td>75.90</td>
<td>76.23</td>
<td>77.38</td>
<td>77.77</td>
<td><strong>78.13</strong></td>
</tr>
<tr>
<td>$\rho = 0.2$</td>
<td>81.54</td>
<td>81.65</td>
<td>81.32</td>
<td><strong>81.81</strong></td>
<td>81.68</td>
</tr>
<tr>
<td>$\rho = 0.4$</td>
<td>83.62</td>
<td>83.83</td>
<td>83.43</td>
<td><strong>83.99</strong></td>
<td>83.40</td>
</tr>
<tr>
<td>$\rho = 0.6$</td>
<td>84.44</td>
<td>84.47</td>
<td><strong>84.72</strong></td>
<td>84.04</td>
<td>84.05</td>
</tr>
</tbody>
</table>
Thank you