Auxiliary Objectives for Neural Error Detection Models

Marek Rei & Helen Yannakoudakis
I want to thank you for preparing such a nice evening.

1. **Independent learning**
   Providing feedback to the student.

2. **Scoring and assessment.**
   Helping teachers and speeding up language testing.

3. **Downstream applications.**
   Using as features in automated essay scoring and error correction
I want to thank you for preparing such a nice evening.

I know how to cook some things like potatoes.

I’m looking forward to seeing you and good luck to your project.

We can invite also people who are not members.

The main material that have been used is dark green glass.
Error Types in Learner Writing
Neural Sequence Labelling

\[ P(y_t = k|d_t) = \frac{e^{W_o,k}d_t}{\sum_{k \in K} e^{W_o,k}d_t} \]

\[ d_t = \tanh(W_d h_t) \]

\[ h_t = \begin{bmatrix} \overrightarrow{h_t} \\ \overleftarrow{h_t} \end{bmatrix} \]

\[ \overleftarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t+1}}) \]

\[ \overrightarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t-1}}) \]

char-based word embeddings
Neural Sequence Labelling

\[ E = - \sum_{t=1}^{T} \log(P(y_t|d_t)) \]

Rei and Yannakoudakis (2016, ACL); Rei et al. (2016, COLING)
Auxiliary Loss Functions

• Learning all possible errors from training data is not possible.
• Let’s encourage the model to learn generic patterns of grammar, syntax and composition, which can then be exploited for error detection.

• Introducing additional objectives in the same model.
• Helps regularise the model and learn better weights for the word embeddings and LSTMs.
• The auxiliary objectives are only needed during training.
Auxiliary Loss Functions

\[ d_t^{(n)} = W_f^{(n)} h_t^{(f)} + W_b^{(n)} h_t^{(b)} \]

\[ E = -\sum_t \sum_n \alpha_n \cdot \log(y_t^{(n)}) \]
Auxiliary Loss Functions

1. Frequency
Discretized token frequency, following Plank et al. (2016)

\[ \text{int}(\log(\text{freq}_{\text{train}}(w))) \]

5 3 8 4 8 5 7 9 5 8 0 10
My husband was following a course all the week in Berne.
2. Native language
The distribution of writing errors depends on the first language (L1) of the learner. We can give the L1 as an additional objective.

My husband was following a course all the week in Berne.
3. Error type
The data contains fine-grained annotations for 75 different error types.

My husband was following a course all the week in Berne.
4. Part-of-speech
We use the RASP (Briscoe et al., 2006) parser to automatically generate POS labels for the training data.

APP$  NN1  VBDZ  VVG  AT1  NN1  DB  AT  NNT1  II  NP1  .
My  husband  was  following  a  course  all  the  week  in  Berne  .
5. Grammatical Relations
The Grammatical Relation (GR) in which the current token is a dependent, based on the RASP parser, in order to incentivise the model to learn more about semantic composition.

My husband was following a course all the week in Berne.
Evaluation: FCE

First Certificate in English dataset (Yannakoudakis et al, 2011)
28,731 sentences for training, 2,720 sentences for testing,
Evaluation: CoNLL-14

CoNLL 2014 shared task dataset (Ng et al., 2014)
Alternative Training Strategies

Two settings:
1. Pre-train the model on a different dataset, then fine-tune for error detection.
2. Train on both datasets at the same time, randomly choosing the task for each iteration.

Three datasets:
2. NER dataset with 8 labels (CoNLL 2003).
3. Part-of-speech tagging dataset with 48 labels (Penn Treebank).
### Alternative Training Strategies

#### Pre-training

<table>
<thead>
<tr>
<th>Aux dataset</th>
<th>FCE</th>
<th>CoNLL-14 TEST1</th>
<th>CoNLL-14 TEST2</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>43.4</td>
<td>14.3</td>
<td>21.9</td>
</tr>
<tr>
<td>CoNLL-00</td>
<td>42.5</td>
<td>15.4</td>
<td>22.3</td>
</tr>
<tr>
<td>CoNLL-03</td>
<td>39.4</td>
<td>12.5</td>
<td>20.0</td>
</tr>
<tr>
<td>PTB-POS</td>
<td><strong>44.4</strong></td>
<td>14.1</td>
<td>20.7</td>
</tr>
</tbody>
</table>

#### Switching

<table>
<thead>
<tr>
<th>Aux dataset</th>
<th>FCE</th>
<th>CoNLL-14 TEST1</th>
<th>CoNLL-14 TEST2</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>43.4</td>
<td>14.3</td>
<td>21.9</td>
</tr>
<tr>
<td>CoNLL-00</td>
<td>30.3</td>
<td>13.0</td>
<td>17.6</td>
</tr>
<tr>
<td>CoNLL-03</td>
<td>31.0</td>
<td>13.1</td>
<td>18.2</td>
</tr>
<tr>
<td>PTB-POS</td>
<td>31.9</td>
<td>11.5</td>
<td>14.9</td>
</tr>
</tbody>
</table>
Training on a larger corpus (17.8M tokens):
  • Cambridge Learner Corpus (Nicholls, 2003)
  • NUS Corpus of Learner English (Dahlmeier et al., 2013)
  • Lang-8 (Mizumoto et al., 2011)

<table>
<thead>
<tr>
<th>Task</th>
<th>$F_{0.5}$</th>
<th>R&amp;Y (2016)</th>
<th>$F_{0.5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCE DEV</td>
<td>60.7</td>
<td>61.2</td>
<td></td>
</tr>
<tr>
<td>FCE TEST</td>
<td>64.3</td>
<td>64.1</td>
<td></td>
</tr>
<tr>
<td>CoNLL-14 TEST1</td>
<td>34.3</td>
<td>36.1</td>
<td></td>
</tr>
<tr>
<td>CoNLL-14 TEST2</td>
<td>44.0</td>
<td>45.1</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

• We performed a **systematic comparison** of possible auxiliary tasks for error detection, which are either available in existing annotations or can be generated automatically.

• **POS tags, grammatical relations and error types** gave the largest improvement.

• The **combination** of several auxiliary objectives improved the results further.

• Using **multiple labels** on the same data was better than using out-of-domain datasets.

• Multi-task learning also helped with **large training sets**, getting the best results on the CoNLL-14 dataset.
Thank you!