Auxiliary Objectives for Neural Error Detection Models

Marek Rei & Helen Yannakoudakis

UNIVERSITY OF CAMBRIDGE

Error Detection in Learner Writing

```
I want to thak 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

Error Detection in Learner Writing

Spelling error (8.6%)

I want to thak you for preparing such a nice evening .

Missing punctuation (7.4%)

I know how to cook some things like potatoes .

Incorrect preposition (6.3%)

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

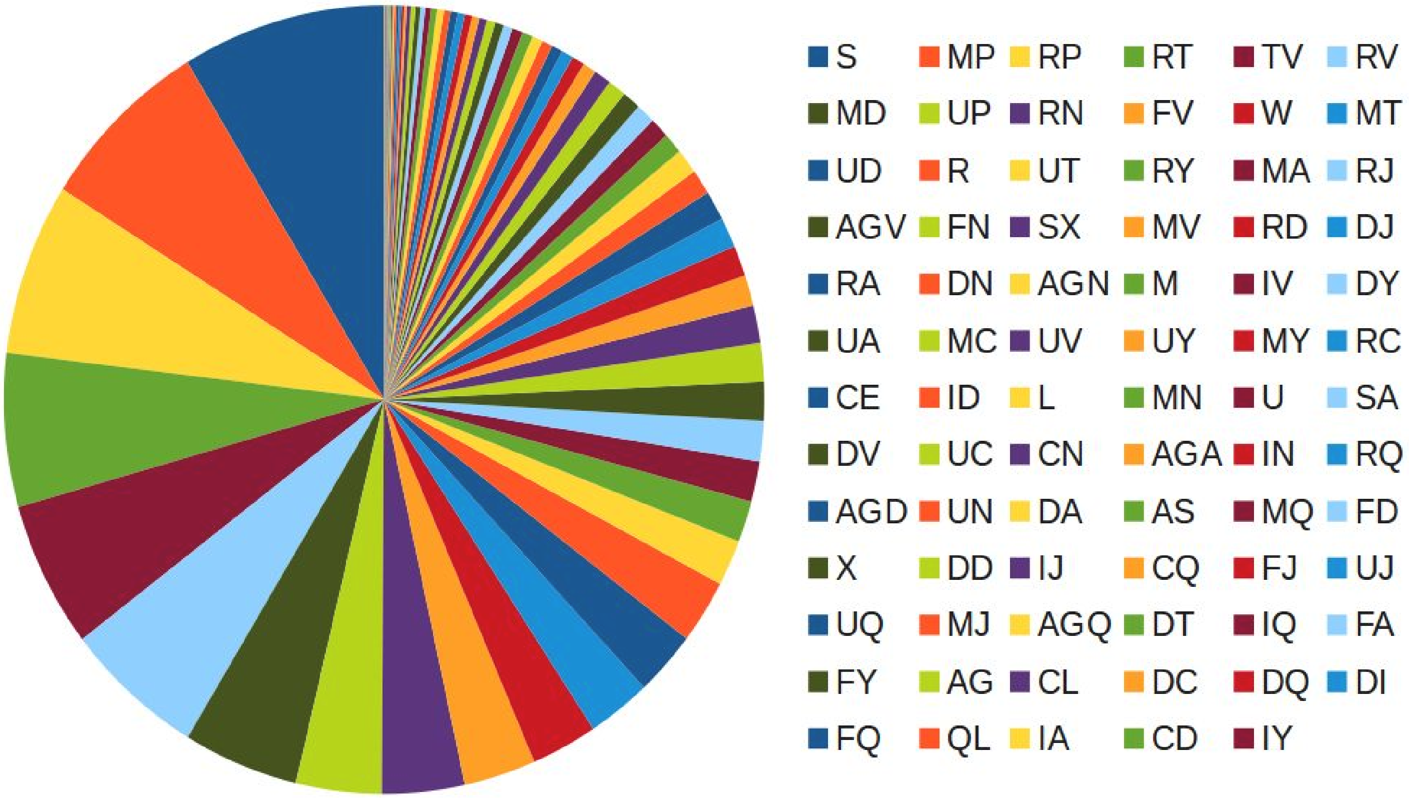
Word order error (2.8%)

We can invite also people who are not members .

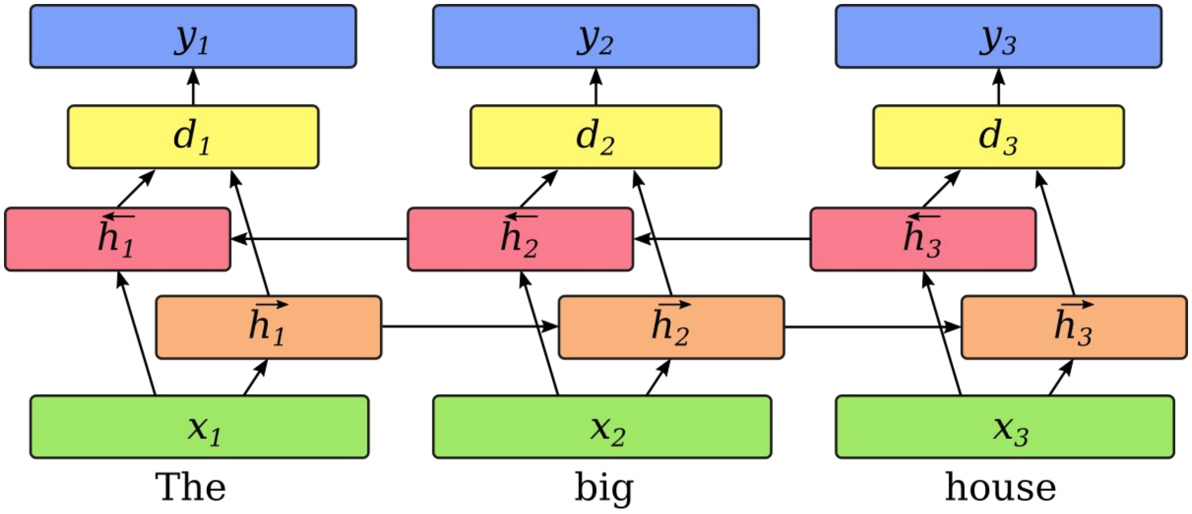
Verb agreement error (1.6%)

The main material that have been used is dark green glass .

Error Types in Learner Writing



Neural Sequence Labelling



$$P(y_t = k|d_t) = \frac{e^{W_{o,k} d_t}}{\sum_{\tilde{k} \in K} e^{W_{o,\tilde{k}} d_t}}$$

$$d_t = tanh(W_d h_t)$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

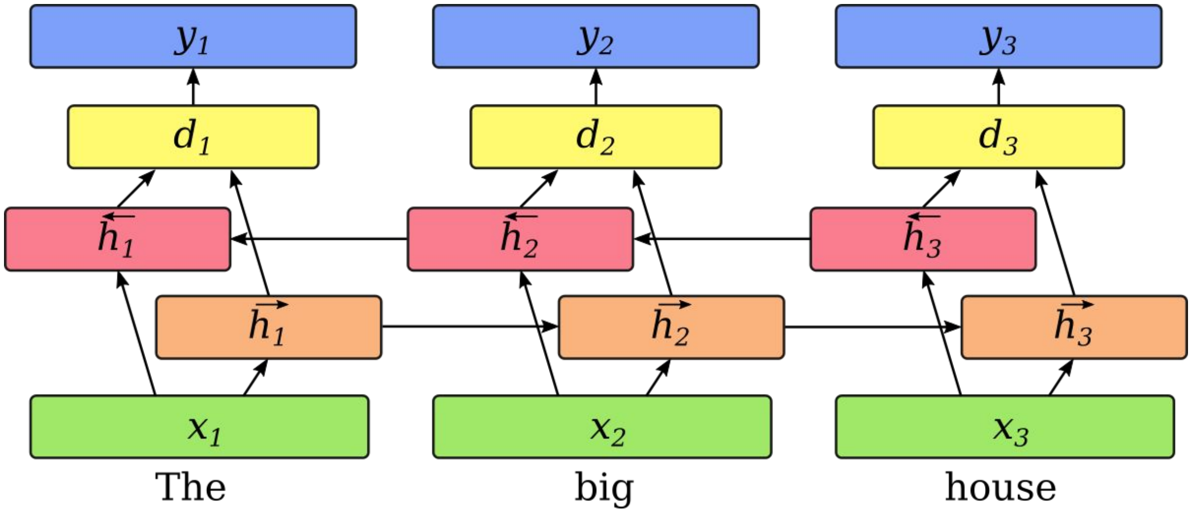
$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}})$$

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}})$$

char-based word embeddings

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

Neural Sequence Labelling



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

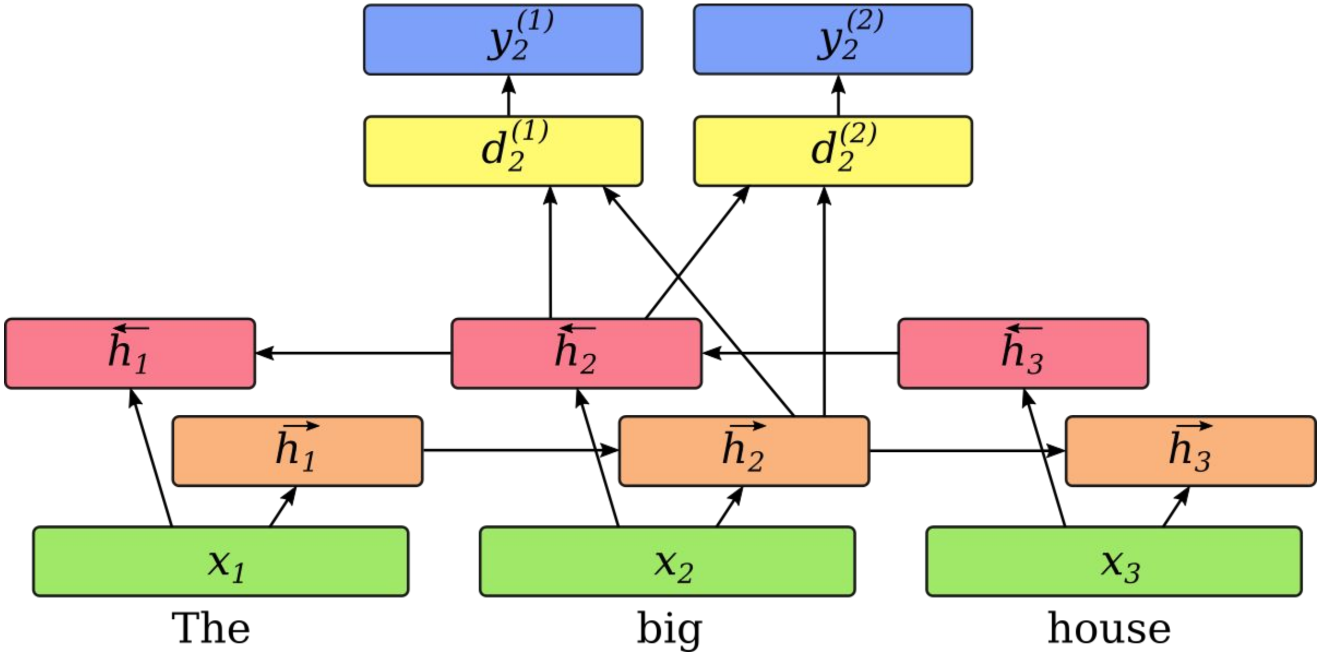
Rei and Yannakoudakis (2016, ACL); Rei et al. (2016, COLING)

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



$$d_t^{(n)} = W_f^{(n)} h_t^{(f)} + W_b^{(n)} h_t^{(b)} \qquad E = -\sum_t \sum_n \alpha_n \cdot \log(y_t^{(n)})$$

Auxiliary Loss Functions

1. Frequency

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

$$\mathrm{int}(\log(\mathrm{freq}_{\mathrm{train}}(w)))$$

5	3	8	4	8	5	7	9	5	8	0	10
My	husband	was	following	a	course	all	the	week	in	Berne	.

Auxiliary Loss Functions

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.

fr	fr	fr	fr	fr	fr	fr	fr	fr	fr	fr	fr
My	husband	was	following	a	course	all	the	week	in	Berne	.

Auxiliary Loss Functions

3. Error type
The data contains fine-grained annotations for 75 different error types.

```
_  _        _   RV          _ _      _   UD  _    _  _      _
My husband was following a course all the week in Berne .
```

Auxiliary Loss Functions

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 .
```

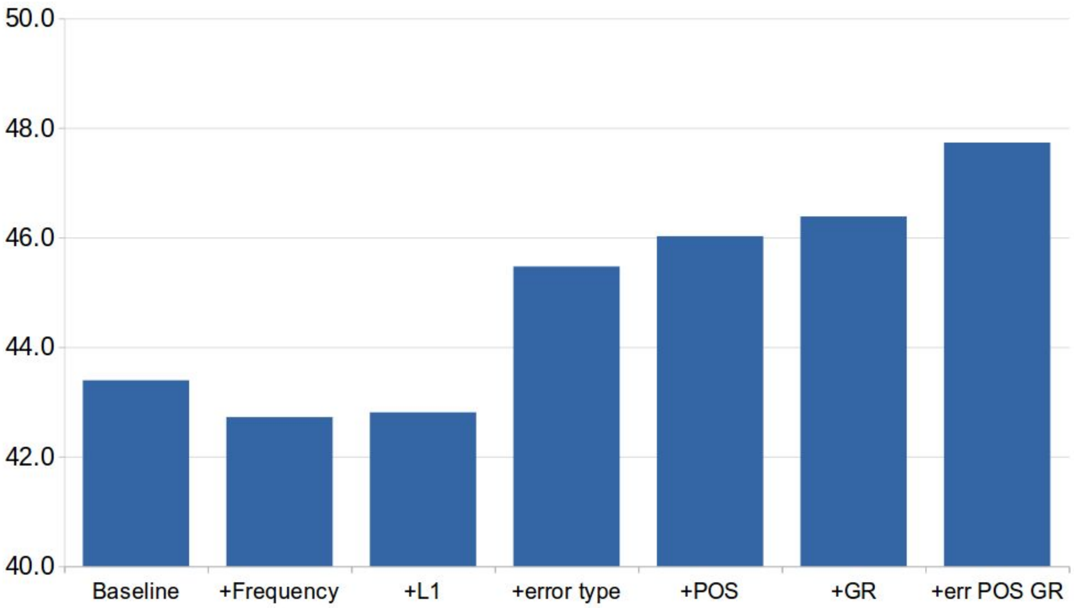
Auxiliary Loss Functions

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.

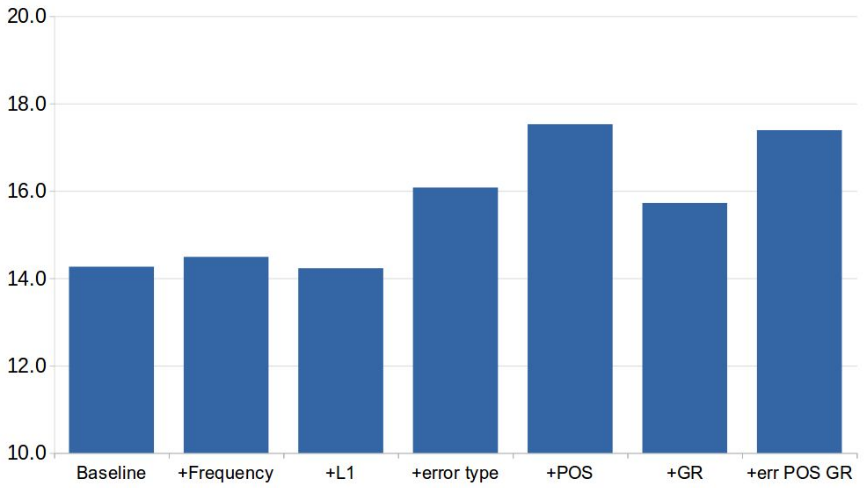
```
det ncsubj     aux    null          det  dobj     ncmod  det   ncmod  ncmod  dobj    null
My  husband  was  following  a   course  all   the  week  in    Berne  .
```

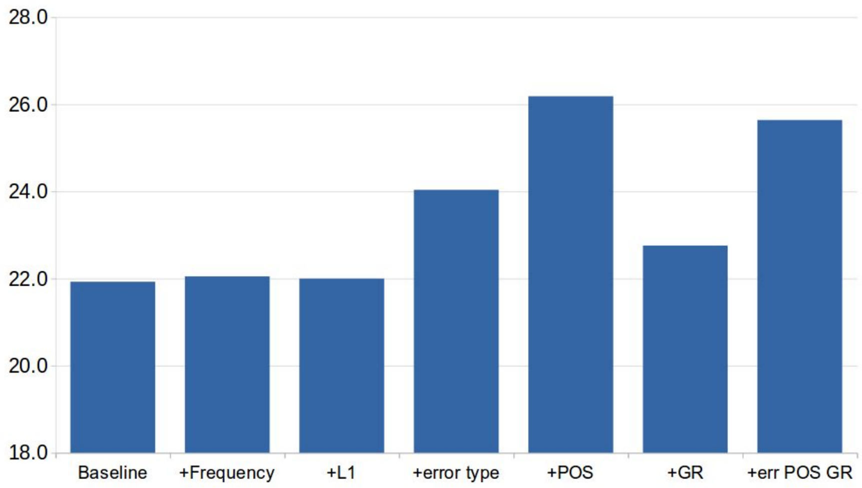
Evaluation: FCE

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:

1. Chunking dataset with 22 labels (CoNLL 2000).

2. NER dataset with 8 labels (CoNLL 2003).

3. Part-of-speech tagging dataset with 48 labels (Penn Treebank).

Alternative Training Strategies

Pre-training

Aux dataset	FCE	CoNLL-14 TEST1	CoNLL-14 TEST2
None	43.4	14.3	21.9
CoNLL-00	42.5	**15.4**	**22.3**
CoNLL-03	39.4	12.5	20.0
PTB-POS	**44.4**	14.1	20.7

Switching

Aux dataset	FCE	CoNLL-14 TEST1	CoNLL-14 TEST2
None	**43.4**	**14.3**	**21.9**
CoNLL-00	30.3	13.0	17.6
CoNLL-03	31.0	13.1	18.2
PTB-POS	31.9	11.5	14.9

Additional Training Data

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)

Task	$F_{0.5}$ R&Y (2016)	$F_{0.5}$
FCE DEV	60.7	**61.2**
FCE TEST	**64.3**	64.1
CoNLL-14 TEST1	34.3	**36.1**
CoNLL-14 TEST2	44.0	**45.1**

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!