INTRODUCTION
The number of non-native speakers of English is growing every year, and automated learner error detection and correction has recently become a popular application area for machine learning (ML) algorithms in natural language processing. Most previous research focuses on function words and casts the task as a multi-class classification problem. In our research, we look at error detection and correction for more challenging errors in content words and investigate how ML algorithms can be applied.

OBJECTIVES
The focus and objectives of this research:
1. We automatically detect and correct learner errors in written English.
2. We investigate errors in the choice of content words: adjectives, nouns and verbs.
3. We take the meaning into account → use compositional distributional semantics.
4. We use machine learning (ML) algorithms to detect and correct errors.

DATA & METHODS
◆ Data: extracted from the Cambridge Learner Corpus (CLC), and contains texts written by non-native English speakers with the examples of the correctly as well as incorrectly chosen words.
◆ The task is to automatically distinguish between the two classes.
◆ Previous research has cast the task as multi-class classification, but focused on predefined set of classes (= number of potential corrections).
◆ Challenges for content words:
  - How many classes (e.g., as many as there are adjectives in English)?
  - Corrections depend on the original word: “big history” vs “long history”
  - Confusions are caused by different reasons: “big anger” vs “great anger” (meaning)
  - “classic dance” vs “classical dance” (form)
◆ Method: treat as binary classification (correct vs. incorrect); encode semantics in the features

ML FOR ERROR DETECTION
Features encode properties of semantic vectors.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Accuracy</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN_context</td>
<td>0.8113</td>
<td>0.7889</td>
<td>0.8650</td>
</tr>
<tr>
<td>AN_front</td>
<td>0.6535</td>
<td>0.5084</td>
<td>0.7467</td>
</tr>
<tr>
<td>VN_context</td>
<td>0.6577</td>
<td>0.5557</td>
<td>0.8217</td>
</tr>
<tr>
<td>VN_front</td>
<td>0.6491</td>
<td>0.6086</td>
<td>0.8467</td>
</tr>
</tbody>
</table>

Table 1: Results

- LB = lower bound, majority class distribution
- UB = upper bound, inter-annotator agreement

CONCLUSION
◆ We have showed that our algorithm detects errors with high accuracy (close to UB).
◆ There is still some room for improvement.
◆ The features derived using semantics and capturing word meaning are useful.
◆ The algorithm shows high precision → it is reliable in practice.
◆ Major source of misclassification – cases where confusion occurs due to similarity in meaning: “small speech” vs “short speech”

There is an increasing need in error detection and correction algorithms for non-native speakers and writers. We plan to extend current research investigating error types other than those currently addressed, wider use of context (e.g., via topic modelling), feature engineering and other feature types (e.g., neural network language models currently applied), and other machine learning algorithms. The next step is to apply an error correction algorithm to the errors identified.

CONTACT INFORMATION
Web: www.cl.cam.ac.uk/~ek358/
Email: Ekaterina.Kochmar@cl.cam.ac.uk
Data: ilexir.co.uk/media/an-dataset.xml

REFERENCES