MuLVE
A Multi-Language Vocabulary Evaluation Data Set

Language Resources and Evaluation Conference 2022

Anik Jacobsen $^1$, Salar Mohtaj $^{1,2}$, Sebastian Möller $^{1,2}$

$^1$ Technische Universität Berlin, Berlin, Germany
$^2$ German Research Centre for Artificial Intelligence (DFKI), Labor Berlin, Germany
Outline

● Motivation
● Approach
● Format
● Analysis
● Availability
● Experiments & Results
● Conclusion
Motivation

- Vocabulary learning is an essential part of foreign language learning.
- Repetition and appropriate feedback are crucial to achieve long-term memory of words and their meaning.
- **phase6**\(^1\) is a digital vocabulary trainer focused on pupils.
- Users can import publisher content they are learning in class.

\(^1\)https://www.phase-6.de

MuLVE, A Multi-Language Vocabulary Evaluation Data Set, LREC 2022
Motivation

- Language learning systems operate on simple rules that compare the user’s answer to an existing answer.

→ User frustration: semantically correct solutions are not accepted

<table>
<thead>
<tr>
<th>question</th>
<th>answer</th>
<th>user answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wir sind aus Berlin.</td>
<td>We are from Berlin.</td>
<td>We are from Berlin.</td>
</tr>
<tr>
<td>Wir sind aus Berlin.</td>
<td>We are from Berlin.</td>
<td>We are Berlin.</td>
</tr>
<tr>
<td>Wir sind aus Berlin.</td>
<td>We are from Berlin.</td>
<td>We’re from Berlin.</td>
</tr>
<tr>
<td>Wir sind aus Berlin.</td>
<td>We are from Berlin.</td>
<td>We come from Berlin.</td>
</tr>
</tbody>
</table>
Approach

User learning data since 2015
~ 456 M data points

per target language: **EN, FR, ES**

- top 1,250 most learned publisher vocabulary cards
  (question language: **DE**)

MuLVE, A Multi-Language Vocabulary Evaluation Data Set, LREC 2022
**Approach**

- "**I was right**" option in the app let users give a feedback to system shows their answer was correct.
- The initial approach used the "I was right" and Wrong classes as labels for the data points.
- We found different user behavior for "I was right" option based on manual checking of the data.
**Challenge**

- Different user behavior for “I was right” option

---

**Labeling Process**

1. Generating correct answer variations from A: A*
2. Comparing A’ with A*
3. Computing Edit Distance between A’ and A
4. High Edit Distance: hand labeling A’

---

A = Answer  
A’ = User Answer  
A* = Answer Variations

---

MuLVE, A Multi-Language Vocabulary Evaluation Data Set, LREC 2022
Format

original
A'₁: We are from Berlin. <br/>{d14e32fr43f32f2}
A'₂: We’re from Berlin. <br/>{d14e32fr43f32f2}
A'₃: We’re from Berlin. <br/>{d14e32fr43f32f2}

A'₁: we are from berlin
A'₂: we are from berlin
A'₃: we are from berlin

Preprocessing: Removing HTML Tags, Audio IDs, etc.
Normalization: Lowercase, removing punctuation, long form

Deduplication: Removing duplicate user answers

MuLVE, A Multi-Language Vocabulary Evaluation Data Set, LREC 2022
Analysis

- **Duplicate** data sets
  - are sampled to 1M data points
  - have original True / False balance.

- **Deduplicated** datasets are a lot smaller because of
  - small number of possible correct answers
  - undersampling

MuLVE, A Multi-Language Vocabulary Evaluation Data Set, LREC 2022
Availability

The data set is available on European Language Grid.
https://live.european-language-grid.eu/catalogue/corpus/9487
Experiments

- Fine-tune a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model using the described data set as a downstream task.

- Fine tune pre-trained models for each language to ensure compatibility and a multilingual BERT with the concatenated datasets of all languages.

- Hyperparameters:
  - 4 epochs
  - batch size 32 (16 for English)
  - learning rate 3e-5 for the English and Spanish model and 2e-5 for the French and multilingual model

- Duplicate datasets were downsampled to 1 million data points.

MuLVE, A Multi-Language Vocabulary Evaluation Data Set, LREC 2022
Experiments

- Each data set reaches >92 accuracy, and for each language, there also exists a model with >95.5 accuracy, showing the models are able to learn from the available data.

- F2-Score performance is comparable to accuracy, showing the models are able to balance precision and recall while highlighting recall.

MuLVE, A Multi-Language Vocabulary Evaluation Data Set, LREC 2022
We introduce Multi-Language Vocabulary Evaluation Data Set (MuLVE): a data set containing different variations of vocabulary cards and real-life user answers with a binary label indicating whether the answer is correct or not.

We provide a first experiment and validation of a transformer model trained and tested on the available data set variations.

We make the data set variations available to the research community. It can, for example, be used to train and evaluate vocabulary and language evaluation systems.
MuLVE
A Multi-Language Vocabulary Evaluation Data Set

Language Resources and Evaluation Conference 2022

Anik Jacobsen $^1$, Salar Mohtaj $^{1,2}$, Sebastian Möller $^{1,2}$

$^1$ Technische Universität Berlin, Berlin, Germany
$^2$ German Research Centre for Artificial Intelligence (DFKI), Labor Berlin, Germany

a.jacobsen@campus.tu-berlin.de, {salar.mohtaj, sebastian.moeller}@tu-berlin.de