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.

- phase 6 [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

💥 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

question	answer	user answer	
Wir sind aus Berlin.	We are from Berlin.	We are from Berlin.	🟢
Wir sind aus Berlin.	We are from Berlin.	We are Berlin.	🔴
Wir sind aus Berlin.	We are from Berlin.	We're from Berlin.	🟢
Wir sind aus Berlin.	We are from Berlin.	We come from Berlin.	🔴

Approach

User learning data since 2015

~ 456 M data points

phase 6

per target language:
EN, FR, ES

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

top 1,000
Training Set

1,001 - 1,250
Test Set

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

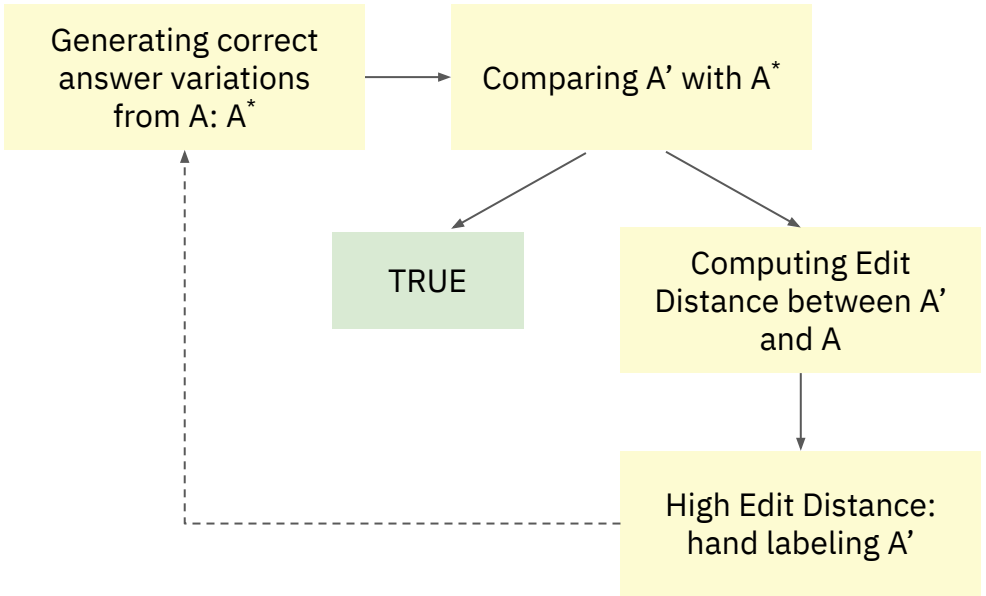
Approach

Challenge

- Different user behavior for “I was right” option

Labeling Process

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



Format

original
A'$_1$:
 We are from Berlin.
{d14e32fr43f32f2}
A'$_2$:
 We're from Berlin.
{d14e32fr43f32f2}
A'$_3$:
 We're from Berlin.
{d14e32fr43f32f2}

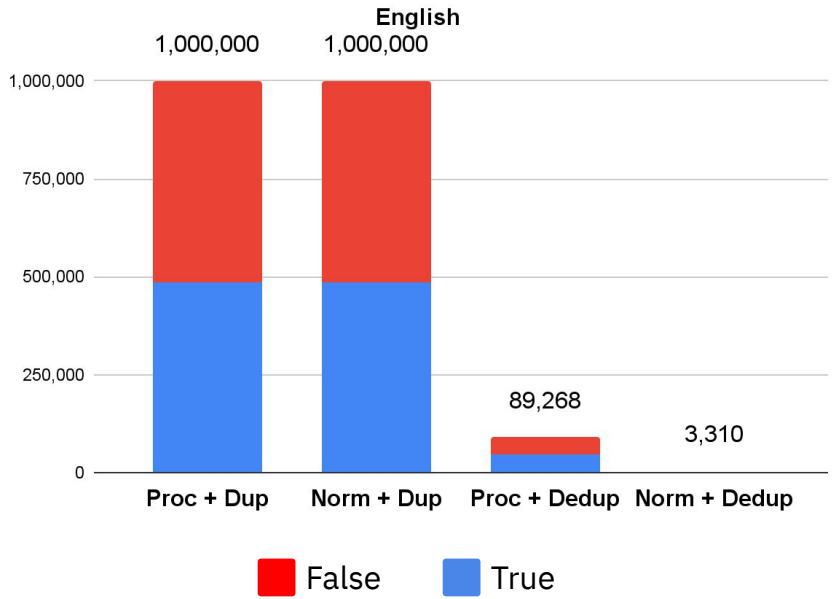
Preprocessing: Removing HTML Tags, Audio IDs, etc.
Normalization: Lowercase, removing punctuation, long form

	Pre-processed Data	Normalized Data
Including Duplicates	A'$_1$: We are from Berlin. A'$_2$: We're from Berlin. A'$_3$: We're from Berlin.	A'$_1$: we are from berlin A'$_2$: we are from berlin A'$_3$: we are from berlin
Deduplicated	A'$_1$: We are from Berlin. A'$_2$: We're from Berlin.	A'$_1$: we are from berlin

Deduplication: Removing duplicate user answers

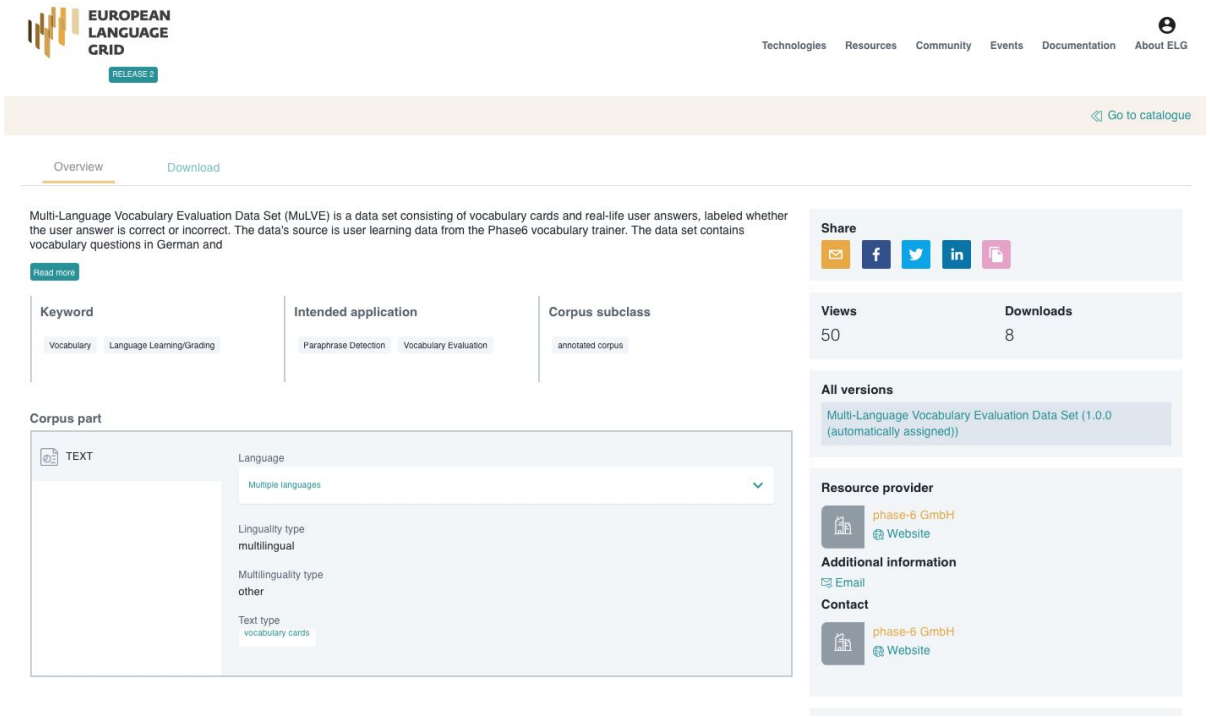
📊 Analysis

- **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



Availability

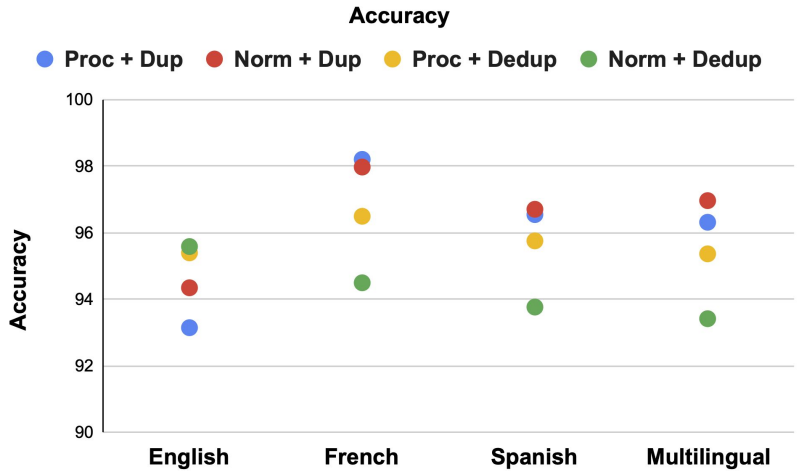
The data set is available on European Language Grid.
https://live.european-language-grid.eu/catalogue/corpus/9487

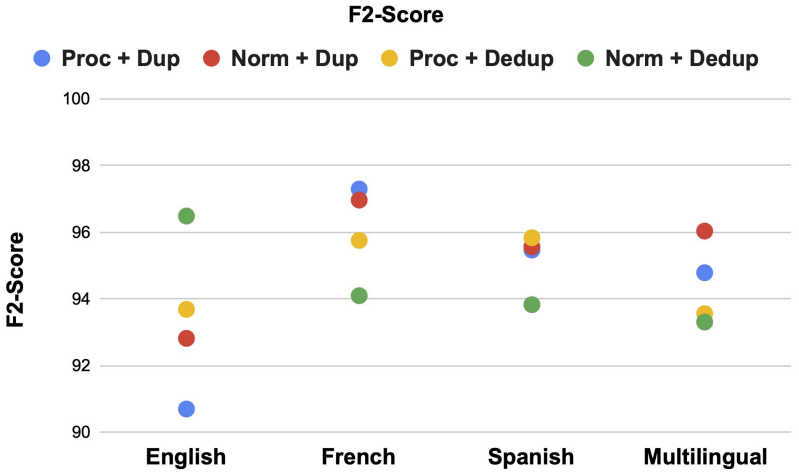


Experiments

- 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.

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.

💫 Conclusion

🗂️ 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