

MuLVE

A Multi-Language Vocabulary Evaluation Data Set

Language Resources and Evaluation Conference 2022

Anik Jacobsen¹, Salar Mohtaj^{1,2}, Sebastian Möller^{1,2}

¹Technische Universität Berlin, Berlin, Germany

²German 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>

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



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



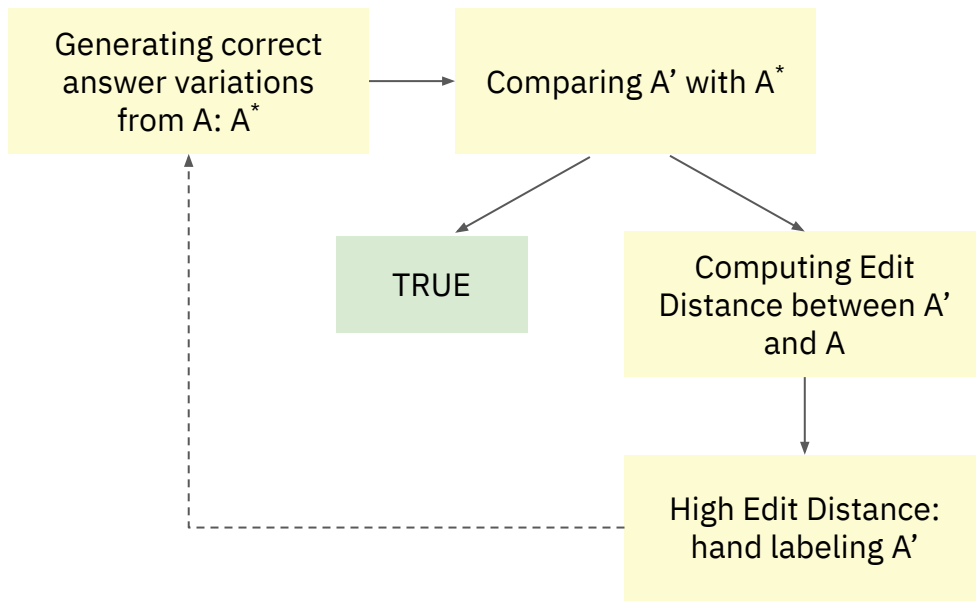
Approach

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}

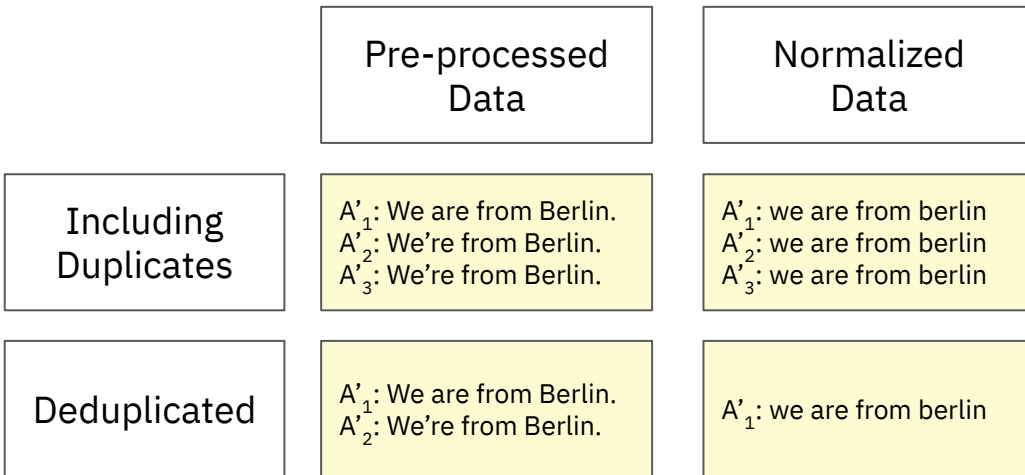
A'₂:
 We're from Berlin.
{d14e32fr43f32f2}

A'₃:
 We're from Berlin.
{d14e32fr43f32f2}

Preprocessing: Removing HTML Tags, Audio IDs, etc.

Normalization: Lowercase, removing punctuation, long form

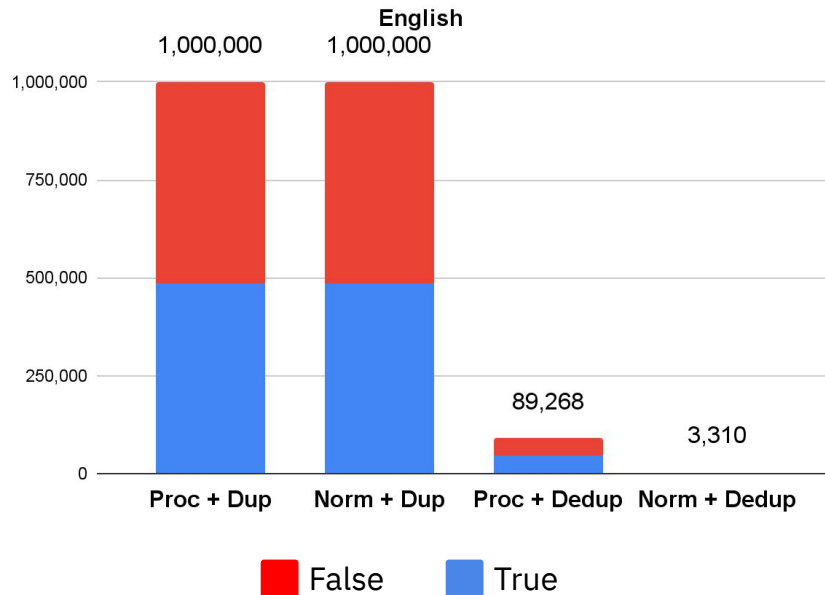
Deduplication: Removing duplicate user answers





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>

The screenshot shows the European Language Grid website interface. At the top left is the logo for the European Language Grid, labeled 'RELEASE 2'. To the right is a navigation menu with links for Technologies, Resources, Community, Events, Documentation, and About ELG. Below the navigation is a 'Go to catalogue' link. The main content area has two tabs: 'Overview' (selected) and 'Download'. The overview text describes the Multi-Language Vocabulary Evaluation Data Set (MuLVE) as a data set of vocabulary cards and real-life user answers, labeled for correctness. It mentions the data source is user learning data from the Phase6 vocabulary trainer. Below this text is a 'Read more' link. The 'Keyword' section lists 'Vocabulary' and 'Language Learning/Grading'. The 'Intended application' section lists 'Paraphrase Detection' and 'Vocabulary Evaluation'. The 'Corpus subclass' section lists 'annotated corpus'. The 'Corpus part' section shows a 'TEXT' document icon and a 'Language' dropdown menu set to 'Multiple languages'. Other metadata includes 'Linguality type: multilingual', 'Multilinguality type: other', and 'Text type: vocabulary cards'. On the right side, there is a 'Share' section with icons for email, Facebook, Twitter, LinkedIn, and Print. Below that is a 'Views' and 'Downloads' section showing 50 views and 8 downloads. The 'All versions' section lists 'Multi-Language Vocabulary Evaluation Data Set (1.0.0 (automatically assigned))'. The 'Resource provider' section lists 'phase-6 GmbH' with a 'Website' link. The 'Additional information' section includes an 'Email' link. The 'Contact' section also lists 'phase-6 GmbH' with a 'Website' link.

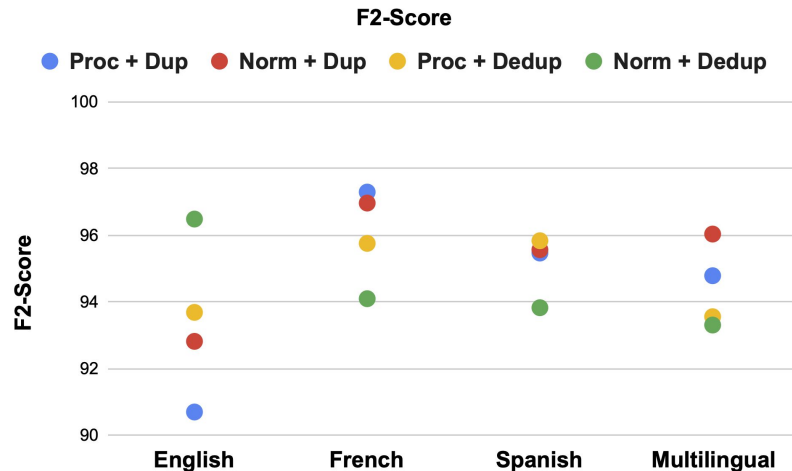
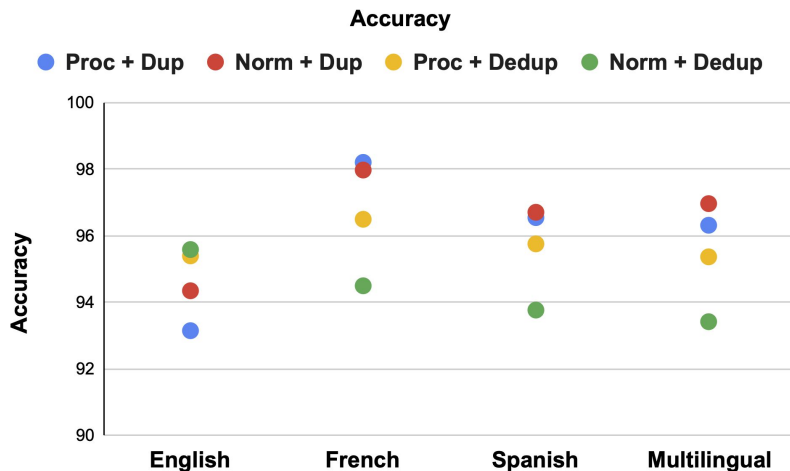


Experiments

- 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