Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 673–679
Marseille, 20-25 June 2022

MuLVE, A Multi-Language Vocabulary Evaluation Data Set

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

Abstract

Vocabulary learning is vital to foreign language learning. Correct and adequate feedback is essential to successful and satisfying vocabulary training. However, many vocabulary and language evaluation systems perform on simple rules and do not account for real-life user learning data. This work introduces Multi-Language Vocabulary Evaluation Data Set (MuLVE), a data set consisting of vocabulary cards and real-life user answers, labeled indicating whether the user answer is correct or incorrect. The data source is user learning data from the Phase6 vocabulary trainer. The data set contains vocabulary questions in German and English, Spanish, and French as target language and is available in four different variations regarding pre-processing and deduplication. We experiment to fine-tune pre-trained BERT language models on the downstream task of vocabulary evaluation with the proposed MuLVE data set. The results provide outstanding results of > 95.5 accuracy and F2-score. The data set is available on the European Language Grid.

Keywords: Data Sets, Vocabulary Evaluation, Paraphrase Detection

1. Introduction

Vocabulary learning is an essential part of foreign language learning. Building an extensive vocabulary is necessary to master a foreign language and communicate successfully (Alqahtani, 2015). To achieve long-term memory of words and their meaning, repetition and appropriate feedback are crucial (Metcalfe and Kornell, 2007). We introduce a Multi-Language Vocabulary Evaluation Data Set (MuLVE), including real-life user vocabulary learning data, aiming to improve vocabulary evaluation.

phase-6 GmbH[1] (hereinafter referred to as Phase6) offers a digital education tool in the domain of language learning. Their service is a vocabulary trainer, available for various digital devices, optimizing vocabulary training for long-term memory. Students can study vocabulary independently and aligned with the content of their school lessons. Phase6 operates in the German market and focuses on pupils.

The area of vocabulary evaluation and training has not been addressed in Natural Language Processing (NLP). Most existing language learning systems operate on simple rules that compare the user's answer to an existing answer, neglecting potential correct answers such as synonyms or various ways of formatting. These inflexible systems lead to user frustration and a limited learning experience, resulting in users losing interest in language learning completely.

We aim to establish a robust and significant multilingual data set for vocabulary evaluation from available user learning data provided by Phase6 to allow for the development and training of more flexible systems for the task of vocabulary evaluation. Our contributions are as follows:

- 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.
- 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[2]. It can, for example, be used to train and evaluate vocabulary and language evaluation systems.

To the best of our knowledge, it is the first data set focusing on vocabulary evaluation.

Next, we discuss related work and relevant corpora in the related domain of paraphrase detection (Section 2). Section 3 describes the creation of the MuLVE data set, focusing on the retrieval of the data and the annotation process. In Section 4, we present the format of the data set and how to access it. Section 5 analyses the data set distribution. We experiment with the data set and validate the results in Section 6. Finally, we conclude the paper and provide future work.

2. Related Work

The research problem is closely related to tasks in the area of NLP. Semantic similarity, usually of sentences or documents, is discussed within paraphrase detection. Paraphrases are semantically identical sentences that convey the same meaning but use different wording. The research in this field goes beyond the sentence level and further considers documents composed

[1] https://www.phase-6.de/

[2] https://live.european-language-grid.eu/catalogue/corpus/9487

of multiple sentences. The task of paraphrase detection has many applications, such as plagiarism detection (Wahle et al., 2021), Q&A systems (Bogdanova et al., 2015), and text summarization (Agarwal et al., 2018).

The benchmark corpus in the field of paraphrase detection is the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005). It consists of 5,801 sentence pairs, where each pair is binary labeled, indicating whether it is a paraphrase or not by a human annotator. The sentences have been collected from newswire articles over two years. The corpus linked with the task of paraphrase detection is included in the GLUE benchmark (Wang et al., 2018), a collection of nine natural language understanding tasks.

Finkelstein et al. (2001) developed WordSimilarity-353 as one of the benchmark corpora for the task of word similarity detection. The data consists of 353 pairs of words that are annotated in range of 0 to 10 by 13 experts. SimLex-999 is another relevant corpus which is proposed by Hill et al. (2015). It is a gold standard resource for evaluating distributional semantic models which quantified similarity between pairs of entities. The corpus covers adjective, noun, and verb, and 500 annotators participated in the annotation process, in which each pair is annotated by almost 50 subjects.

Another relevant corpus, including cross-lingual sentence pairs, is PAWS-X (Yang et al., 2019). PAWS-X is derived from PAWS (Paraphrase Adversaries from Word Scrambling (Zhang et al., 2019)), containing challenging English sentence pairs from Wikipedia and Quora. The noisy paraphrase detection labeled sentence pairs highlight the importance of modeling structure, context, and word order information in the domain of paraphrase identification. PAWS-X contains 23,659 human translated and 296,406 machine-translated PAWS evaluation pairs in six typologically distinct languages: French, Spanish, German, Chinese, Japanese, and Korean. It contains only examples from PAWS-Wiki. In their paper, Yang et al. show the effectiveness of deep, multilingual pre-training on PAWS-X.

Xu et al. (2015) propose a data set consisting of short and noisy texts retrieved from Twitter. It contains 17,790 sentence pairs in the training and development set and 972 sentence pairs in the test set from 500+ trending topics on Twitter (the collection period was between April 24th and May 3rd, 2013). The sentence pairs have a label indicating whether they are paraphrases, not paraphrases, or debatable.

Quora Question Pairs (QQP)[3] is another data set used to train and evaluate paraphrase detection approaches. It consists of over 400,000 question pairs. Each question pair is annotated with a binary value indicating whether the two questions are paraphrases of each other.

[3] https://www.kaggle.com/c/quora-question-pairs

In contrast to the aforementioned data sets, MuLVE focuses on the new task of vocabulary evaluation and contains vocabulary cards and their respective user answers. MuLVE is constructed from real user learning data and aims to improve language evaluation systems.

3. Approach

This section explains in detail the creation of the MuLVE data set. Section 3.1 provides information on the retrieval of the data set using user learning data from Phase6. Section 3.2 describes the data annotation process.

3.1. Data Retrieval

The source of the data is the Phase6 user input database. Phase6 has collected anonymous user inputs, resulting in more than 450 M available data points. Currently, Phase6 offers support for more than 20 languages. The company focuses on the German-speaking market; thus, most vocabulary questions are in German. The data set will focus on the three most popular target languages: English, French, and Spanish.

When a user's answer is flagged as incorrect by the Phase6 system, the user has the option to select **"accept as correct"** in which case the vocabulary is marked as correct. These data points will be referred to as **I was right (IWR)**. A screenshot of the user interface can be seen in Figure 2.

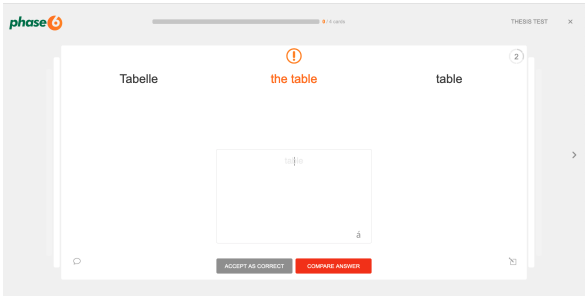

Figure 2: Screenshot of the vocabulary learning environment. If an answer is flagged as incorrect, the user can select **accept as correct**.

All user answers for the top 1,000 most learned vocabulary cards are included to build a representative and not too extensive data set. Also, the user answers are limited to the IWR and Wrong user answers, as these include misclassifications which we aim to include in the data set.

3.2. Data Annotation

The initial approach used the IWR and Wrong classes as labels for the data points. During data exploration, we discovered, however, that user behavior is quite variant; entailing that some users accept the system's decision, while others use the option to select "Accept as correct" (even when their answer might certainly

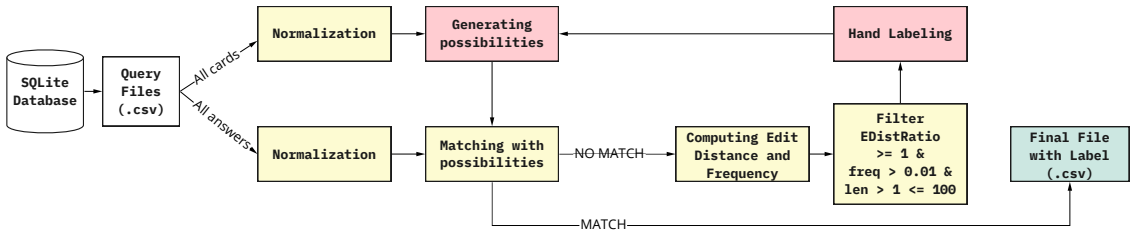

Figure 1: Relabeling process.

be incorrect). Thus, user answers containing the exact same text are present in both the IWR and Wrong class; this results in a low data quality.

A semi-automatized process for relabeling is shown in Figure 1. It focuses on generating possibilities of additional correct answers, combined with an extra loop to identify synonyms and other correct solutions which are not yet included. The user answers are then compared with these lists of possible correct answers and labeled **True** if there is a match and **False** if there is no match. The individual steps will be discussed in detail.

3.2.1. Normalization

First, the data is normalized to make the format as consistent as possible and then compare the user answers with the possibilities. The text is converted to lower case, and everything in parentheses in the answer is ignored, as it mainly contains irrelevant additional information. Punctuation is removed because it is not essential on vocabulary level. Further, spacing is adapted to single spacing, and line breaks are removed. Since children primarily use the vocabulary trainer, wrong versions of the apostrophe character (“ ’ ”) are replaced; many children are not yet used to the spelling of foreign languages, and apostrophes are rarely used in German. Thus, accents and other characters that appear similar to an apostrophe are corrected.

3.2.2. Generating Possibilities

The generation of possibilities encompasses two different aspects. Firstly, it aims to include additional semantically correct solutions such as synonyms. Moreover, it aims to include additional differently formatted correct solutions.

Synonyms are words that have the same semantic meaning and, for example, in English, often occur when comparing British and American English. Another example is the use of different words depending on the desired formality.

The formatting is not unique for all vocabulary cards. There are several such formatting cases for each language, and we explain a few as examples. Despite these differences in formatting, the system should accept a correct user answer, as we aim for the user to learn the vocabulary and not its format.

An example in the English language is the use of *“to”* in the case of the *to-infinitive*. There are several ways to answer an English verb vocabulary question correctly.

answer	**answer possibilities**
till twelve o’clock	till twelve o’clock, until twelve o’clock
(to) ask	to ask, ask
Thank you.	Thank you., Thanks.
neighbour (BE)	neighbour, neighbor
... ¿no?	¿no, ¿verdad
escuchar algo	escuchar algo, escuchar, escuchar (algo)
todo/-a	todo/-a, todo, toda, todo toda, toda todo, todoa
soy	soy, yo soy, soy estoy, estoy soy, yo soy estoy, yo estoy soy
la télé / la télévision	la télé, la télévision, la télé la télévision, la télévision la télé
l’effaceur (m.)	l’effaceur, l’effaceur m
toi	toi, tu
gratuit/gratuite	gratuit, gratuite, gratuit gratuite, gratuite gratuit

Table 1: Examples of the list of correct possibilities, which can be compared with user answers to find possible matches.

In Romance languages, in the case of the available languages French and Spanish, there often exist different versions of nouns and adjectives for the male and female forms. In this case, the male and female, as well as any combination of the two, are correct.

Another common formatting is the use of indefinite pronouns, like: something = *sth*, somebody = *sb*. If the vocabulary card asks for the answer *“to find sth/sb”*, the indefinite pronouns are not vital for the meaning of the verb when translating it from German. Thus, correct answers are any combination of these pronouns.

These format variations can be generated automatically by defining rules for the possible formatting options. The result is a list of possibilities for each vocabulary card, which can be compared with the user’s answers to find possible matches. Examples can be found in Table 1.

3.2.3. Edit Score to Identify Synonyms

The drawback of using such possibility lists as described above is missing synonyms and other correct solutions in the user data and mislabeling them as Wrong. The edit distance is used to filter for such

possible cases and then inspect and label this data by hand. The edit distance is a valuable tool, as most incorrect solutions include typos and wrong letters and are thus, by edit distance, very close to the correct set of solutions. We choose the Levenshtein distance (Levenshtein, 1966) as edit distance.

By computing the edit distance and filtering words with a high edit ratio, it is possible to identify significantly different words to the correct solution. These words are likely to be synonyms. A list of possible synonyms and correct solutions could be identified and then labeled by hand. The hand-labeling assignment was carried out by one of the authors. Examples can be found in Table 2.

Question	Answer	Synonym / Correct Solution
um halb acht	at 7:30	at half past seven
richtig; korrekt	right	correct
die Pop-Musik	el pop	la música pop
Wie geht´s?	¿qué tal?	¿cómo estás?
braun	brun/brune	marron marron
ein Stadtviertel	un quartier	un arrondissement

Table 2: Examples of synonyms and additional correct answers detected with edit distance.

4. Data Set Format

There are different preprocessed variations of the data set. The data is either preprocessed (remove HTML tags and sound IDs present in the export) or normalized as described in Section 3.2.1. We present preprocessed data because it contains the format of possible user answers and generalizes more. In contrast, when using normalized data, the data is maximally "clean". Another aspect of the data set format is the inclusion of duplicates. Keeping duplicates might highlight the importance of common mistakes, while it could also lead to a system only adapting to these common mistakes and not generalizing enough. We decided to generate four different variations of the data set to determine the best resulting machine learning model experimentally. An example of these variations is visualized in Figure 3.

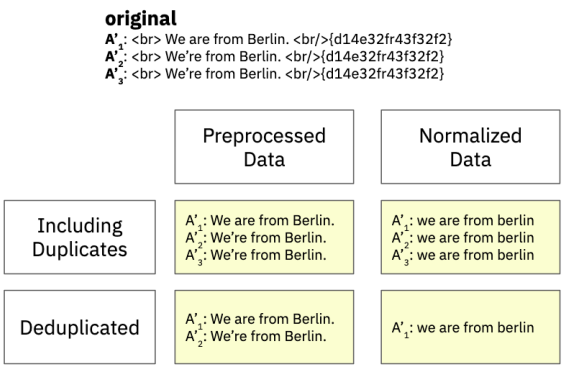

Figure 3: Data set variations with example.

4.1. Balancing Data Set

The data set variations including duplicates are balanced sufficiently (see Section 5). On the other hand, the deduplicated data set variations are extremely imbalanced. This distribution is to be expected; given that, for one vocabulary question, there are many more possible incorrect answers than correct solutions. Thus, a far higher percentage of the deduplicated user answers is **False**, and there are only a few correct variations for the **True** class. The deduplicated data sets are undersampled to achieve fully balanced data sets.

4.2. Disjoint Test Set

The most learned 1,001 to 1,250 vocabulary cards and their user answers are used to create an additional disjoint test set. The data is processed and labeled the same way as the training and validation data set. We again create the same four variations (preprocessed vs. normalized and duplicated vs. deduplicated). Furthermore, the deduplicated variations are undersampled. Models are to be evaluated on these test sets to gain more insights into their performance on new vocabulary cards closer to the real-life scenario.

4.3. Availability

The data set is available on European Language Grid[4]. It is split up in four tab separated files, one for each variation, per train and test set. The files include the following columns:

- **cardId** - numeric card ID
- **question** - vocabulary card question
- **answer** - vocabulary card answer
- **userAnswer** - user answer input
- **label** - True if user answer is correct, False if it is not
- **language** - target language (English, French or Spanish)

The processed data set variations include the following additional columns:

- **question_norm** - question normalized
- **answer_norm** - answer normalized
- **userAnswer_norm** - user answer normalized

The deduplicated processed data sets do not include the **userAnswer**, since we deduplicate on basis of the **userAnswer_norm** column.

	Preprocessed + Duplicates		Preprocessed + No Duplicates		Normalized + Duplicates		Normalized + No Duplicates	
	True	**False**	**True**	**False**	**True**	**False**	**True**	**False**
English	12,718,244		89,268		12,718,244		3,310	
	6,186,558 49%	6,531,686 51%	44,634 50%	44,634 50%	6,186,558 49%	6,531,686 51%	1,655 50%	1,655 50%
French	8,027,831		59,538		8,027,831		4,078	
	2,972,047 37%	5,055,784 63%	29,769 50%	29,769 50%	2,972,047 37%	5,055,784 63%	2,039 50%	2,039 50%
Spanish	2,248,457		31,838		2,248,457		3,858	
	811,918 36%	1,436,539 64%	15,919 50%	15,919 50%	811,918 36%	1,436,539 64%	1,929 50%	1,929 50%

Table 3: Distribution of data points per language and data set variation for the training set.

	Preprocessed + Duplicates		Preprocessed + No Duplicates		Normalized + Duplicates		Normalized + No Duplicates	
	True	**False**	**True**	**False**	**True**	**False**	**True**	**False**
English	2,329,762		15,260		2,329,762		882	
	1,155,401 50%	1,146,584 50%	7,630 50%	7,630 50%	1,155,401 50%	1,146,584 50%	441 50%	441 50%
French	1,243,814		9,478		1,243,814		944	
	484,651 39%	759,163 61%	4,739 50%	4,739 50%	484,651 39%	759,163 61%	472 50%	472 50%
Spanish	388,923		6308		388,923		1042	
	388,923 41%	228,326 59%	3,154 50%	3,154 50%	388,923 41%	228,326 59%	521 50%	521 50%

Table 4: Distribution of data points per language and data set variation for the test set.

5. Data Analysis

Table 3 shows the number of user answers for the top 1,000 most learned vocabulary cards per language and variation in the final data set. The total number of data points and the number of user answers per class are shown.

The data set variations, including duplicates, contain the most data points, as to be expected. English is the largest of the three languages as English is the most popular foreign language in the German school system, followed by French and Spanish. The duplicated data set variations are relatively balanced. Deduplicated data set variations were undersampled to achieve balanced data sets. The data sets are significantly smaller due to the undersampling of the larger possible set of False answers. Normalizing the user answers decreases the size of the data set even further since the set of possible correct answers is smaller.

The same insights can be deducted for the test set (Table 4). The test is smaller as it contains only the user answers to 250 vocabulary cards.

6. Experiments and Validation

We conduct experiments to explore the task of vocabulary evaluation and establish the usability of the data set. Eventually, we fine-tune a pre-trained BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) model using the described data set as a downstream task. We compare the results for the data set variations and individual languages.

[4] https://live.european-language-grid.eu/catalogue/corpus/9487

6.1. Parameters

We fine-tune BERT models pre-trained for each language to ensure compatibility with each language. For English, we use Vanilla BERT$_{BASE}$ (Devlin et al., 2019), for French CamemBERT (Martin et al., 2020), and for Spanish BETO$_{BASE}$ (Canete et al., 2020). In addition, we also fine-tune a multilingual BERT (mBERT) (Devlin et al., 2019) model with the concatenation of all languages.

In terms of parameters for fine-tuning the pre-trained models, we experimented with the hyperparameter space suggested by Devlin et al. (2019). Eventually, the models were trained for 4 epochs, using a batch size of 32 (16 for the English model). We used a learning rate of $3e-5$ for the English and Spanish model and $2e-5$ for the French and multilingual model.

The data sets including duplicate user answers include up to more than 12 million data points (see Table 3). Training a model with this amount of data leads to very long training times and the overfitting of the model to the training data. We downsampled the data for training and validation to 1 million data points to overcome this challenge while keeping distribution and structure in place.

6.2. Results

The results from BERT fine-tuning for each language and data set variation are visualized in Figure 4 and 5. We determined accuracy and F2-score to be the most relevant metrics to measure the models' performance. Accuracy indicates the overall quality of the model. F2-score, a variation of the F1-score, emphasizes the completeness (recall) of a system, which is important in vocabulary evaluation. F2-score is thus better suited

than the widely used F1-score, as well as, precision and recall themselves.

Overall, we can conclude that the fine-tuning results indicate excellent performance. For each data set variation in each language, we were able to fine-tune a model that reaches an accuracy of > 92. Further, a model with > 95.5 accuracy also exists for each language. These results show that the model learns to classify most vocabulary cards correctly. The high F2-scores confirm this finding. It shows that the model can learn from the available training data set and further generalize to classify new vocabulary cards correctly, demonstrated by the disjoint test set.

Most models could learn best from the data sets that include duplicates. Only Vanilla BERT seems to generalize better from deduplicated data. This can be explained by the very large dataset leading to overfitting. There is no clear performance distinction between the preprocessed and normalized data, which indicates that BERT can abstract from the textual input to perform the classification task.

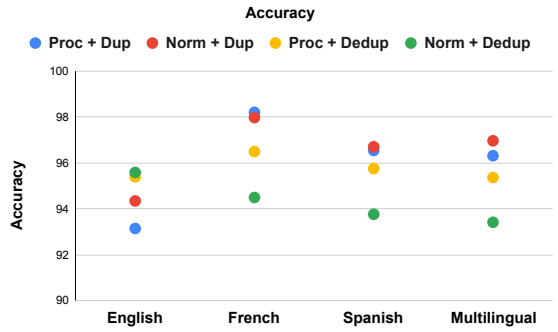

Figure 4: Accuracy results: each data set reaches > 92 accuracy, and for each language, there also exists a model with > 95.5 accuracy, showing the models are able to learn from the available data.

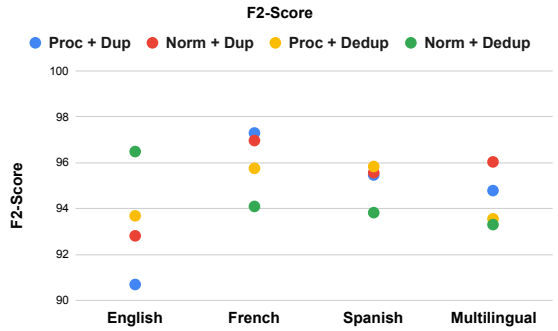

Figure 5: F2-Score results: performance is comparable to accuracy, showing the models are able to balance precision and recall while highlighting recall.

7. Conclusion

In this paper, we present a data set for the task of vocabulary evaluation called MuLVE. By using real-life pupil vocabulary training data, we are able to provide a data set for English, French, and Spanish in four variations. The primarily automated re-labeling process allows generating improved labels compared to existing language learning evaluation systems. A first experiment, fine-tuning pre-trained BERT models to the downstream task of vocabulary evaluation, shows excellent results.

We aim to extend the data set to more languages in future work. Further, we aim to incorporate qualitative vocabulary evaluations to provide fine-grained feedback to language learners.

8. Bibliographical References

Agarwal, B., Ramampiaro, H., Langseth, H., and Ruocco, M. (2018). A deep network model for paraphrase detection in short text messages. *Information Processing & Management*, 54(6):922–937.

Alqahtani, M. (2015). The importance of vocabulary in language learning and how to be taught. *International journal of teaching and education*, 3(3):21–34.

Bogdanova, D., dos Santos, C., Barbosa, L., and Zadrozny, B. (2015). Detecting semantically equivalent questions in online user forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 123–131.

Canete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Metcalfe, J. and Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, 14(2):225–229.

Wahle, J. P., Ruas, T., Meuschke, N., and Gipp, B. (2021). Are neural language models good plagiarists? a benchmark for neural paraphrase detection. *arXiv preprint arXiv:2103.12450*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Xu, W., Callison-Burch, C., and Dolan, W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019). Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692.

Zhang, Y., Baldridge, J., and He, L. (2019). Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.