Lexical Substitution is not Synonym Substitution: On the Importance of Producing Contextually Relevant Word Substitutes

Juraj Vladika[a], Stephen Meisenbacher[b] and Florian Matthes[c]
Department of Computer Science, School of Computation, Information and Technology,
Technical University of Munich, Germany
{juraj.vladika, stephen.meisenbacher, matthes}@tum.de

Keywords: Natural Language Processing, Lexical Substitution, Lexical Semantics, Language Models.

Abstract: Lexical Substitution is the task of replacing a single word in a sentence with a similar one. This should ideally be one that is not necessarily only synonymous, but also fits well into the surrounding context of the target word, while preserving the sentence's grammatical structure. Recent advances in Lexical Substitution have leveraged the masked token prediction task of Pre-trained Language Models to generate replacements for a given word in a sentence. With this technique, we introduce CONCAT, a simple augmented approach which utilizes the original sentence to bolster contextual information sent to the model. Compared to existing approaches, it proves to be very effective in guiding the model to make contextually relevant predictions for the target word. Our study includes a quantitative evaluation, measured via sentence similarity and task performance. In addition, we conduct a qualitative human analysis to validate that users prefer the substitutions proposed by our method, as opposed to previous methods. Finally, we test our approach on the prevailing benchmark for Lexical Substitution, CoInCo, revealing potential pitfalls of the benchmark. These insights serve as the foundation for a critical discussion on the way in which Lexical Substitution is evaluated.

1 INTRODUCTION

Lexical substitution (LS) can be described as the task of replacing a word in a sentence with the most appropriate different word. It is one of the essential linguistic tasks in Natural Language Processing (NLP) and is an integral component of more complex NLP tasks that deal with text rewriting. Some generative tasks where LS is an important part include paraphrasing (Fu et al., 2019), machine translation (Agrawal and Carpuat, 2019), style transfer (Helbig et al., 2020), defense against adversarial attacks (Zhou et al., 2021), text simplification (Štajner et al., 2022), or private text rewriting (Meisenbacher et al., 2024).

More so than just replacing a word in text with the one of the most similar meaning from a thesaurus, we argue that true LS also ideally regards the surrounding context of the target word and tries to produce candidates that best fit the semantic flow of the whole sentence. In tasks dealing with text rewriting, the goal is to preserve the semantic meaning of the sentence, where simply choosing synonyms may be inadequate.

While early LS methods relied on rule-based systems, pre-trained language models (PLMs) like BERT (Devlin et al., 2019) have been predominantly used in recent years. The training objective of PLMs with a masked token prediction (masked language modeling, MLM) provides a natural approach to LS. While this mechanism has been utilized in current LS methods, they are prone to overfitting to the target word by predicting pure synonyms. For this purpose, we introduce CONCAT, a method for English lexical substitution that in a simple way of concatenating the masked sentence with the target sentences provides an improved trade-off between semantics and context.

To test the performance of CONCAT, we deploy both quantitative and qualitative analysis. We first use three standard benchmarks for lexical substitution, namely LS07 (McCarthy and Navigli, 2007), CoInCo (Kremer et al., 2014), and Swords (Lee et al., 2021). Even though our model shows satisfying results on these datasets, upon manual inspection of the gold substitutes provided by the annotators, we noticed a number of problems and inconsistencies. This leads to a critical discussion of the way LS is currently evaluated in the NLP community. To address this gap, we perform a qualitative survey in form of a ques-

[a] https://orcid.org/0000-0002-4941-9166
[b] https://orcid.org/0000-0001-9230-5001
[c] https://orcid.org/0000-0002-6667-5452

Vladika, J., Meisenbacher, S. and Matthes, F.
Lexical Substitution is not Synonym Substitution: On the Importance of Producing Contextually Relevant Word Substitutes.
DOI: 10.5220/0013389000003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1481-1488
ISBN: 978-989-758-737-5; ISSN: 2184-433X

tionnaire, where users were invited to choose their preferred lexical substitutes in four different settings. The analysis of the results shows that users highly favor substitutes generated by our approach and deem them the most appropriate in given context. While this type of human evaluation is common in generative NLP tasks, to the best of our knowledge there has not been such an evaluation for LS.

Finally, we test to what extent LS preserves the semantic usefulness of text. We achieve this by lexically substituting words in the input sentences of a text classification dataset and observing the effect on a trained classification model. The results reveal that the model performance stays very close to the original text, especially in the case of ConCat substitutions.

Our contributions are as follows:

- We introduce ConCat, a simple lexical substitution method for English, capable of producing highly contextually fitting replacement words.
- We evaluate the method on three standard LS benchmarks: LS07, CoInCo, and Swords. We provide a critical discussion of the benchmarks and point out their shortcomings.
- We conduct a qualitative survey and show that users highly favor substitutes generated by ConCat and deem them most appropriate contextually, even when compared to gold substitutes.
- We examine how well ConCat preserves the semantic meaning of text, showing that ConCat is in most cases the best at preserving performance.
- We make the code for ConCat publicly available at https://github.com/sebischair/ConCat.

Original sentence: The quick brown fox jumped over the lazy dog.
Target word: jumped
ConCat: The quick brown fox jumped over the lazy dog. [SEP]
The quick brown fox [MASK] over the lazy dog.

Figure 1: ConCat: a simple and intuitive method for contextually relevant Lexical Substitution.

2 RELATED WORK

Lexical substitution was formally defined by McCarthy (2002). Early approaches used rule-based heuristics and synonym lookup in word thesauri (Hassan et al., 2007). With the advent of word embedding methods, LS approaches began representing the target word and its context with dense vectors and ranking the candidates with vector-similarity metrics (Roller and Erk, 2016). Most recent approaches utilize pre-trained language models (PLMs).

Word representations learned by PLMs are highly contextual and finding substitutes that fit the word's surrounding context was significantly improved. Since the prediction of substitutes in this manner can be highly biased towards the target word, Zhou et al. (2019) apply a dropout mechanism by resetting some dimensions of the target word's embedding and then leverage the model to predict substitutes using this perturbed embedding. Later methods by Michalopoulos et al. (2022) and Seneviratne et al. (2022) combine the PLM word embeddings with a gloss value of target word's synonyms returned by WordNet to guide the vector-space exploration towards similar words. Qiang et al. (2023) generate substitutes with paraphrase modeling and improved decoding.

Unlike the approaches that use WordNet for embedding perturbation, our approach includes querying WordNet for a rule-based filtering of unsuitable words. To the best of our knowledge, we are the first to utilize the idea of sentence concatenation for improved LS and the first to do a qualitative analysis of generated substitutes with human evaluation.

3 METHODOLOGY

3.1 The ConCat Approach

Language models are powerful predictors of words that contextually fit a sentence, due to their MLM learning objective of predicting a missing word in a sentence, performed over massive training corpora. An intuitive approach to LS is simply to replace the target word with a [MASK] token and let the PLM predict top candidates. Another approach would be to keep the original text intact to predict substitutes. From our observation, the first approach tends to produce contextually relevant words that can be semantically distant from the original word, while the second approach overfits to the target word and predicts solely its different inflectional forms or synonyms, thus lacking creativity. To bridge this gap, our method combines these two approaches – the masked sentence (with the target word masked) is concatenated with the original sentence, separated by a separator token. This was inspired by a similar concatenation method for lexical simplification (Qiang et al., 2020).

ConCat, our proposed simple and intuitive LS method, is demonstrated in Figure 1. This approach combines the best of both worlds – it increases creativity by forcing the model to predict an empty [MASK] token but also makes the model aware of the original word by including it in the next sentence. After experimenting with multiple base models such as BERT (Devlin et al., 2019) and XL-Net (Yang et al., 2019), we opted for RoBERTa (Liu et al., 2019), which seemed to produce the most fitting substitutes. Additionally, its tokenizer uses full words as tokens, facilitating direct word substitution. We use the ROBERTA-BASE model from HuggingFace since it is less resource-intensive than the large variant, without a significant performance drop.

Despite the impressive performance, there was still a considerable number of instances where our approach fell short, resorting to predicting words which are antonyms of the target word (which could fit the context but not the semantics of the sentence) or grammatical variations of the target word (which breaks the grammatical correctness of the sentence). To overcome this obstacle, we additionally deploy checks based on WordNet and filter out inadequate words, creating a hybrid approach. In particular, we use Wordnet to obtain lists of synonyms, antonyms, hypernyms, hyponyms, meronyms, holonyms, and then manually filter out these from the top *k* generated candidates of the masked target word.

3.2 Evaluation

To evaluate our new method, we conduct a multi-headed evaluation consisting of three parts: (1) evaluation on the LS07, CoInCo, and Swords benchmarks, (2) evaluation and analysis of LS on a classification task, and (3) a qualitative evaluation led by surveys.

LS Benchmarking. We begin our evaluation with measuring our method's performance on popular LS bechmarks: LS07 (McCarthy and Navigli, 2007), CoInCo (Kremer et al., 2014), and Swords (Lee et al., 2021). In all of them the goal is to provide substitutes for given target words in the provided sentences. Note that in the case of Swords, we employ two versions of the provided gold labels: (1) *Swords 1*, which use all gold labels where at least one annotator voted it to be suitable, and (2) *Swords 5*, where at least 50% of annotators agreed upon a substitute.

As metrics, we employ the four LS metrics from SemEval 2007. We briefly introduce them, but refer the reader to the original task report for more details.

- *best*: evaluates the quality of the best prediction, by scoring the existence of the gold top substitute, weighted by the order.
- *best-mode*: evaluates if the system's best prediction appears in the *mode* of the gold labels.
- *oot*: evaluates the coverage of the gold substitutes in the system predictions, i.e., the percentage of gold substitutes appearing in the system's top-10 predictions.
- *oot-mode*: evaluates the % of times the mode of the gold labels appears in the system's top-10 predictions.

In addition to these metrics, we also employ P@1 and P@3. It should be noted that for the entire benchmark, we limit the system predictions to 10 responses, so as to not skew the *best* scoring.

LS Task Performance. The second step of our evaluation includes measuring performance on a NLP task after performing LS on the original dataset. For this, we select the AG News dataset (Zhang et al., 2015), which presents a Multi-Class Topic Classification task. We take a random sample of 10,000 rows for our evaluation. For the task, we select four settings $Subst.\% \in \{0.25, 0.5, 0.75, 1.0\}$, where the percentage represents the randomly selected percentage of tokens in the dataset that are replaced via LS. We use our proposed method (CONCAT), as well as the method proposed by Zhou et al. (2019).

To measure performance (accuracy) in each of these settings, we use each resulting dataset to train a LSTM model (Hochreiter and Schmidhuber, 1997) using keras. As an input layer, we use GloVe embeddings (Pennington et al., 2014). For training, we set the batch size to 64 and train for a maximum of 30 epochs with early stopping. Accuracy is captured from the best resulting model, and the training process is run five separate times to obtain average accuracy.

Perplexity Test. *Perplexity* is a common method for the evaluation of generative LMs (Chen et al., 1998). Essentially, perplexity measures how "uncertain" a LM is in making a next token prediction. A lower perplexity, therefore, implies that an LM is less "surprised" by the context given to the model. Perplexity can be used to assess text complexity of a sentence (Vladika et al., 2022). We aim to leverage this metric to evaluate how well a word substitute fits into its context sentence. In particular, we envision two settings:

- **Top-10**: all top-10 generated substitutes are replaced into the sentence, and the average perplexity of these sentences is taken.
- **Top-Match**: only to the top *k* substitutes are measured, where *k* denotes the number of gold substitutes.

These results are compared against the perplexity of the original context sentences (*baseline*), and the average perplexity of the sentences replaced with the gold substitutes (*gold*). For perplexity, we use the GPT2 model (Radford et al., 2019).

Table 1: Benchmark results for both the LS07 and CoInCo tasks. Metrics were calculated using the approach described by (McCarthy and Navigli, 2007). Highest scores are **bolded**.

Dataset:	**LS07**						**CoInCo**					
Model:	best	best-mode	oot	oot-mode	P@1	P@3	best	best-mode	oot	oot-mode	P@1	P@3
Dropout	20.3	34.2	55.4	68.4	51.1	–	14.5	33.9	45.9	69.9	56.3	–
Dropout*	2.37	12.81	23.73	32.02	16.67	26.67	3.44	10.57	18.91	23.59	23.54	40.82
ConCat	**3.52**	**16.26**	**35.11**	**46.31**	**21.67**	**40.33**	**4.81**	**15.63**	**26.07**	**34.47**	**34.66**	**53.81**

Table 2: Benchmark results for the Swords task. Metrics were calculated and are presented as in Table 1.

Dataset:	**Swords 1**						**Swords 5**					
Model:	best	best-mode	oot	oot-mode	P@1	P@3	best	best-mode	oot	oot-mode	P@1	P@3
Dropout*	1.34	**3.39**	**13.44**	**18.64**	22.43	20.27	2.19	**8.33**	**21.92**	**27.27**	40.54	32.70
ConCat	**2.43**	**3.39**	13.17	13.56	**35.95**	**31.08**	**4.02**	**8.33**	20.76	25.00	**48.38**	**42.43**

Table 3: Comparing CoInCo gold substitutions to ours. Top-1 took the top candidate from each substitution set and replaced the target. Random-1 picked a word from the substitution set at random (same index for gold and ours).

Task:	**CoInCo (Cosine Similarity)**					
Substitution:	Top-1			Random-1		
Model:	Mini	DR	MPN	Mini	DR	MPN
CoInCo Gold	76.15	69.67	72.08	76.26	69.88	72.27
Avg:		72.63			**72.80**	
ConCat	78.00	70.24	73.10	76.53	69.44	71.81
Avg:		**73.77**			72.59	

Qualitative Analysis. In the final stage of our evaluation, we administer a survey via Google Forms, with the goal of evaluating user preference of different LS methods' predictions. To accomplish this, we divide the survey into four tasks:

1. **Single Word Replacement (SWR)**: a sentence is presented with the target word bolded, and the user is asked to select from a list of single substitutes.
2. **Single Word Replacement, Masked (SWR-M)**: a sentence is shown with a placeholder instead of the target word to be replaced, and the user is asked to select, from a list a substitutes, which replacement is most suitable.
3. **Set Replacement (SR)**: setup similar to SWR, but the list of options includes sets of three replacements, and the user must select which *set* is most suitable.
4. **Set Replacement, Masked (SR-M)**: like SR, but the target word is once again masked.

For the survey questions, we select 60 random entries from CoInCo, 15 for each task. As answer options, we present the top 1 or 3 gold substitutes from CoInCo, the top 1 or 3 using the method of Zhou et al. (2019), and the top 1 or 3 from ConCat.

4 EXPERIMENT RESULTS

Benchmark. The results for the evaluation on the three benchmark datasets are in Tables 1 and 2. For all tasks, we include 6 metrics, outlined in the previous section. These metrics are placed in juxtaposition to Zhou et al. (2019) for comparison. The original scores of Zhou et al. (2019) could not be replicated due to unavailability of the original code. Therefore, we reimplemented their method and include our replicated score as well, marked by an asterisk (*).

As an added point of comparison, we compute cosine similarity scores for our method's replacements, and the gold CoInCo substitutes. To compute these similarity scores, we utilize three Sentence Transformer models (Reimers and Gurevych, 2019): ALL-MINILM-L12-V2 (Mini), ALL-DISTILROBERTA-V1 (DR), and ALL-MPNET-BASE-V2 (MPN). In addition, we compute the scores in two settings: (1) **Top-1**, where the target word is replaced with the top annotator response / system prediction, and (2) **Random-1**, where the target is replaced with a randomly chosen substitute. These results are in Table 3.

Table 4: Accuracy scores for AG News. % Subst. denotes the percentage of tokens in the dataset per that were substituted. Scores represent an average of five evaluated models.

Task:	**AG News** (baseline = 88.41 ± 0.40)			
% Subst.	25%	50%	75%	100%
Dropout	**87.52** ± 0.13	85.91 ± 0.54	83.37 ± 0.36	81.37 ± 0.37
ConCat	87.41 ± 0.48	**86.18** ± 0.44	**84.48** ± 0.50	**83.04** ± 0.30

AG News. In Table 4, we present the results of our task-based evaluation. Accuracy scores for each subtitution setting (percentage of tokens replaced) are given, once again compared against Zhou et al. (2019). These results are also plotted in Figure 2.

Similarly to benchmark analysis, we employ co-

Table 5: Average cosine similarity scores between original dataset and datasets with a percentage of token replacements. SentenceTransformers models used: ALL-MINILM-L12-V2 (Mini), ALL-DISTILROBERTA-V1 (DR), and ALL-MPNET-BASE-V2 (MPN). Highest average score per % Subst. is **bolded.**

Task:	**AG News (Cosine Similarity)**											
% Subst.		25%			50%			75%			100%	
Model:	Mini	DR	MPN	Mini	DR	MPN	Mini	DR	MPN	Mini	DR	MPN
Dropout	91.23	91.00	91.04	80.73	81.19	80.59	69.77	71.36	69.82	55.51	59.37	55.76
Avg:		91.09			80.84			70.32			56.88	
ConCat	91.03	91.38	91.34	81.05	82.07	81.20	70.45	73.22	71.00	57.76	62.91	58.72
Avg:		**91.34**			**81.44**			**71.56**			**59.80**	

Table 6: Survey responses from 21 respondents. For each of the four tasks, we include the number of responses that preferred a given method's replacement, as well as the corresponding percentage for this task (of 21*15=315 responses per task). Note that in some cases, the total percentage may exceed 100. This occurs when more than one method outputs the same prediction – both methods would then be counted if selected by a respondent.

	Survey Responses (21 respondents, 60 questions)									
	SWR		SWR-M		SR		SR-M		**TOTAL**	
Gold	80	25.40%	**129**	**40.95%**	119	37.78%	88	27.93%	416	33.02%
Dropout	141	44.76%	114	36.19%	89	28.25%	95	30.16%	439	34.84%
ConCat	**195**	**61.90%**	120	38.40%	**125**	**39.68%**	**112**	**35.56%**	**552**	**43.81%**

sine similarity to illustrate the effect of LS methods on the overall semantics of the underlying AG News dataset. The scores are calculated for each of the substitution settings, and scores from three embedding models are averaged. The results are given in Table 5. The results of the perplexity test are found in Table 7.

Survey. The aggregate results from our survey are illustrated in Table 6. We received 21 respondents, all who were close colleagues with fully proficient or native levels of English. We separate the results by task, as well as provide total counts.

5 DISCUSSION

We now discuss our results in detail, and extract interesting insights from our analysis.

LS vs. Utility. The task-based evaluation revealed that performing LS on datasets to be fed to downstream tasks does indeed have an effect on model performance. As illustrated by Figure 2, a clear degradation in utility occurs as a higher percentage of the AG News dataset is replaced by LS. Interestingly, our method "slows" this degradation rate down. Not covered is *less-than-25%* replacement, where one could study the intersection of LS and model robustness.

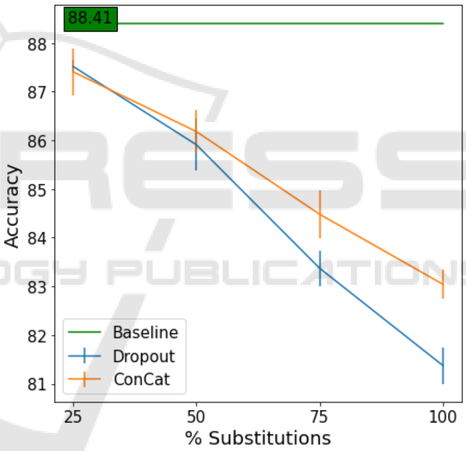

Figure 2: Accuracy scores for LS task performance. As token replacements increase, task performance decreases, but at different rates. Baseline performance shown in green.

Context Improves Metrics. The results of applying our method to LS07 and CoInCo clearly show the effects of including contextual information in the LS process. In comparison to Zhou et al. (2019), whose method does not include such context, our method shows improvement on all metrics. Particularly in the metrics that evaluate *coverage* of the gold substitutes in the system predictions, our method displays considerable improvements. We pose that especially in these cases where coverage is tested, the benefits of contextual information are evident, namely in producing lexically *and* semantically suitable replacements.

Preserving Semantics and Reducing Perplexity. In evaluating the preservation of semantic similarity, we empirically show that LS in general leads to a loss in information, measured by cosine similarity. While this is certainly plausible, the extent to which semantic similarity is affected is shown in Table 3, as well in additional task-based experiments.

Table 7: Perplexity Results. Note that for Swords, we use all provided gold labels (*Swords 1*). T-M: Top-Match.

Dataset:	**LS07**		**CoInCo**		**Swords**	
	Top-10	T-M	Top-10	T-M	Top-10	T-M
Baseline	228.62		258.19		66.78	
Gold	292.14		232.05		83.06	
CONCAT	256.36	253.80	206.73	203.58	74.69	73.75

At the same time, the results of the perplexity measurements lead to important insights about the nature of LS. In 3 out of 4 cases, using gold substitutes in sentences leads to *increased* perplexity. This would imply that the substitutes are not as suitable to maintain natural, flowing sentences. Increased perplexity can also be observed from CONCAT replacements, yet the effect of our substitutes is not as severe as those from the gold substitutes. In fact, for CoInCo our replacements lead to *lower* perplexity than baseline.

We hypothesize that our replacements might be "preferred" by LMs in perplexity measurements, as CONCAT aims to produce *contextually* meaningful replacements, rather than substitutes that might be synonymous, but not so fitting in a particular context.

A Critique of CoInCo. For evaluation of CONCAT, CoInCo was chosen as this dataset serves as the current *de facto* standard benchmark for LS tasks (Seneviratne et al., 2022). A closer study of its entries, however, reveals intriguing findings that speak to possible considerations to be made in the future.

Firstly, the inclusion of single-word and double-word sentences (e.g., "*She **said**.*"), although a challenging task, greatly biases towards dictionary-based substitutions or similar methods, as context is lacking. Along similar lines, some target words do not consist of full words, but rather word pieces, such as *don* (don't). Similarly, some gold substitute are double-word phrases (e.g., *very much* for *enough*), which is impossible to predict using the single-word prediction of PLMs. Finally, some sentences have not been fully cleaned, which may lead to issues on the system end.

More importantly, some top annotator responses contain errors, e.g., *day* → *@card@ hour period*, where *@card@* does not make sense. Moreover, some annotator suggestions are questionable, such as:

Orig. Tasha is not the **whole** of what happened on Vega IV.

Gold bulk; complete; consummate; entirety;
sum; total

While the suggested replacements represent true synonyms of the target, some of these are not contextually suitable, such as *bulk* or *consummate*.

From these qualitative insights, we decided to conduct a manual inspection of our CONCAT method on the provided benchmarks. Following a similar inspection performed by (Seneviratne et al., 2022), we analyze our method's performance on the benchmarks with two metrics: (1) *Top-3 coverage* (**T3C**): how often each of the top-3 model predictions are in *any* of the gold labels (in %), and (2) *mismatch percentage* (**MMP**): how often *none* of the top-3 predictions are in the gold set (in %). The results of this inspection for CONCAT and Dropout* are included in Table 8.

Table 8: **T3C** ↑ and **MMP** ↓ scores for the three benchmark datasets.

Dataset:	LS07		CoInCo		Swords 1		Swords 5	
	T3C	MMP	T3C	MMP	T3C	MMP	T3C	MMP
Dropout*	11.5	70.3	18.6	17.8	15.8	57.0	20.6	64.9
ConCat	18.6	51.3	27.7	11.9	25.3	47.0	31.4	53.7

As seen in Table 8, both methods "miss" all of the gold labels very often, particularly in LS07 and Swords. Both methods, however, improve in the **T3C** metric from LS07 to CoInCo to Swords 5, implying that the benchmarks are in fact improvements to each other with regard to annotator agreement on the top-choice gold substitutes. The poor performance of both methods on the **MMP** metric also leads to interesting insights. While the task-based (Table 4) and similarity-based results (Table 3) demonstrate that contextual substitutions preserve semantic coherence and overall utility, the **MMP** scores imply that the model-predicted substitutes are not suitable at all.

Table 9 presents five randomly selected samples from the set of **MMP** results, i.e., CONCAT output sets where none of the top-3 substitutes were found in the gold labels. Note that these cases imply the *worst case* predictions of CONCAT. As one can see, however, these "poor" CONCAT substitutes often do contain either semantically and/or contextually relevant replacements, whereas these are simply not reflected in the gold labels. For example, *complain* presents a good *contextual* replacement, which is not covered in the benchmark annotation scheme.

These insights call into question whether any of the benchmark gold labels extend beyond synonym replacement to *contextually relevant* replacements. Our analysis would imply this is not the case. Thus, we view that a closer investigation of the suggested alternatives be made for the LS field going forward.

User Preferences. In the analysis of our survey results, one can see that although our method's substitutes were chosen the most, there is no clear majority. This leads us to hypothesize that preference for LS

Table 9: Examples from the LS07, CoInCo, and Swords benchmarks. Presented are the original sentences with target words **bolded**, as well as the annotator gold substitutes and the top-3 predictions from CONCAT. Text in red denotes unsuitable replacements, whereas CONCAT substitutes in green denote good ones.

Sentence	Gold Substitutes	ConCat Top-3
Finally, this new rule will also have the effect of encouraging existing corporations to produce safer products	lastly, in	note, thus, together
Treatment of physical problems , particularly chronic ones , is possible as **well** as psychological therapy .	in, along, including	much, also, far
Tara **fumed**. Of all the impertinence!	anger, be steam, be upset, blow up, boil, bristle, burn, digest, flare, fret, froth, glower, grumble, howl, madden, plan, rage, rant, rave, ruffle, scowl, seethe, smolder, steam, stew, whine, yell	raged, complained, argued
I have **sent** Patti a list. For payment, we have to forecast the money two days out.	convey, deliver, dispatch, forward, mail, relay, remit, report	given, written, shipped
The factory is highly automated and designed to shift flexibly to produce many different kinds of **chips** to suit demand.	computer chip, fragment	cells, processors, screens

is highly subjective. Nevertheless, our results show that the contextually suitable replacements proposed by our model tend to be preferred. Taking a few examples of survey responses:

E1 Depending upon how many warrants and options are **exercised** prior to completion of the transaction...

Gold use
Ours execute

E2 It **hissed** thoughtfully.

Gold buzz, hoot, say
Ours whisper, say, mutter

In example E1, **90.9%** of respondents preferred our substitutes over the gold label; likewise, **86.4%** preferred ours over the gold choices in E2. Ultimately, such findings suggest the need for a more in-depth study of the human perspective in LS, namely how LS methods reflect our way of thinking, in terms of synonymous versus contextual substitutions.

6 CONCLUSION

In this paper, we introduced CONCAT, a simple and intuitive approach for English lexical substitution. The approach generates highly contextually fitting word substitutes while preserving the semantics of the surrounding sentence. We test our approach using established metrics on three standard LS benchmarks. Deeper analysis of the benchmark structure revealed certain weak points, which provided terrain for a critical discussion of current testing approaches in the LS community. For better insight into our approach's usefulness, we conduct a qualitative survey where we assessed the user preferences. The survey revealed users preferred CONCAT's substitutes more so than the competing approach and gold substitutes. Finally, we provided an analysis of how the performance of a text classification model changes when input instances have some of their words replaced with CONCAT substitutes, showing that our approach is better at preserving the semantics needed for the classification performance than the competing approach. We hope CONCAT will prove useful as a component of generative NLP tasks dealing with text simplification, stylistic transfer, or author anonymization.

LIMITATIONS

Despite showing impressive performance, our approach still falls short in some instances. There are cases when it generates incomplete words. Additionally, there are rare cases where filtering based on WordNet is too strict and removes word that could have been appropriate substitutes. Furthermore, our approach is only optimized for English The model also struggles with uncommon and rare words.

While we wished to provide a comprehensive comparison of our approach against competing approaches, we found it difficult to recreate them for a fair comparison. Even when public code repositories were provided, certain crucial files were missing or code dependencies were ill-defined. This includes the current state-of-the-art approaches ParaLS (Qiang et al., 2023) and CILex (Seneviratne et al., 2022), and their predecessor LexSubCon (Michalopoulos et al., 2022). Therefore, it is difficult to position our approach in the current research landscape. To account for this, we provided both a qualitative survey and NLP task performance experiments, which give more insight into the performance of CONCAT.

REFERENCES

Agrawal, S. and Carpuat, M. (2019). Controlling text complexity in neural machine translation. 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)*, Hong Kong, China. Association for Computational Linguistics.

Chen, S. F., Beeferman, D., and Rosenfeld, R. (1998). Evaluation metrics for language models.

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)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Fu, Y., Feng, Y., and Cunningham, J. P. (2019). Paraphrase generation with latent bag of words. *Advances in Neural Information Processing Systems*, 32.

Hassan, S., Csomai, A., Banea, C., Sinha, R., and Mihalcea, R. (2007). UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic. Association for Computational Linguistics.

Helbig, D., Troiano, E., and Klinger, R. (2020). Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 41–50, Online. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden. Association for Computational Linguistics.

Lee, M., Donahue, C., Jia, R., Iyabor, A., and Liang, P. (2021). Swords: A benchmark for lexical substitution with improved data coverage and quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

McCarthy, D. (2002). Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, pages 89–115.

McCarthy, D. and Navigli, R. (2007). SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic. Association for Computational Linguistics.

Meisenbacher, S., Chevli, M., Vladika, J., and Matthes, F. (2024). DP-MLM: Differentially private text rewriting using masked language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9314–9328, Bangkok, Thailand. Association for Computational Linguistics.

Michalopoulos, G., McKillop, I., Wong, A., and Chen, H. (2022). LexSubCon: Integrating knowledge from lexical resources into contextual embeddings for lexical substitution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1226–1236, Dublin, Ireland. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics.

Qiang, J., Li, Y., Yi, Z., Yuan, Y., and Wu, X. (2020). Lexical simplification with pretrained encoders. *Thirty-Fourth AAAI Conference on Artificial Intelligence*, page 8649–8656.

Qiang, J., Liu, K., Li, Y., Yuan, Y., and Zhu, Y. (2023). ParaLS: Lexical substitution via pretrained paraphraser. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3731–3746, Toronto, Canada. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Roller, S. and Erk, K. (2016). Pic a different word: A simple model for lexical substitution in context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Seneviratne, S., Daskalaki, E., Lenskiy, A., and Suominen, H. (2022). CILex: An investigation of context information for lexical substitution methods. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Štajner, S., Ferrés, D., Shardlow, M., North, K., Zampieri, M., and Saggion, H. (2022). Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in Artificial Intelligence*, 5:991242.

Vladika, J., Meisenbacher, S., and Matthes, F. (2022). TUM sebis at GermEval 2022: A hybrid model leveraging Gaussian processes and fine-tuned XLM-RoBERTa for German text complexity analysis. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 51–56, Potsdam, Germany. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhang, X., Zhao, J. J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *NIPS*.

Zhou, W., Ge, T., Xu, K., Wei, F., and Zhou, M. (2019). BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

Zhou, Y., Zheng, X., Hsieh, C.-J., Chang, K.-W., and Huang, X. (2021). Defense against synonym substitution-based adversarial attacks via dirichlet neighborhood ensemble. In *Annual Meeting of the Association for Computational Linguistics*.