CILex: An Investigation of Context Information for Lexical Substitution Methods

Sandaru Seneviratne[1], **Elena Daskalaki**[1], **Artem Lenskiy**[1], **Hanna Suominen**[1,2]
[1]The Australian National University (ANU) / Canberra, ACT, Australia
[2]University of Turku / Turku, Finland
{sandaru.seneviratne, eleni.daskalaki,
artem.lenskiy, hanna.suominen}@anu.edu.au

Abstract

Lexical substitution, which aims to generate substitutes for a target word given a context, is an important natural language processing task useful in many applications. Due to the paucity of annotated data, existing methods for lexical substitution tend to rely on manually curated lexical resources and contextual word embedding models. Methods based on lexical resources are likely to miss relevant substitutes whereas relying only on contextual word embedding models fails to provide adequate information on the impact of a substitute in the entire context and the overall meaning of the input. We proposed CILex, which uses contextual sentence embeddings along with methods that capture additional *C*ontext *I*nformation complimenting contextual word embeddings for *Lex*ical substitution. This ensured the semantic consistency of a substitute with the target word while maintaining the overall meaning of the sentence. Our experimental comparisons with previously proposed methods indicated that our solution is now the state-of-the-art on both the widely used LS07 and CoInCo datasets with $P@1$ scores of 55.96% and 57.25% for lexical substitution. The implementation of the proposed approach is available at https://github.com/sandaruSen/CILex under the MIT license.

1 Introduction

Lexical substitution is an important Natural Language Processing (NLP) task, which aims to generate and rank suitable candidate words to replace a given target word, while maintaining the meaning of the given sentence. Lexical substitution is used in a wide range of NLP tasks like data augmentation, paraphrase generation, word sense induction, or text simplification (Shardlow, 2014; Amrami and Goldberg, 2018).

Through the years, different approaches have been introduced for lexical substitution but, due to the paucity of annotated data, most of the lexical substitution systems rely on unsupervised methods based on lexical resources or pre-trained language models (Lacerra et al., 2021). Earlier, methods typically relied entirely on manually curated lexical resources like WordNet (Miller, 1995). The synonyms obtained from such resources were then ranked based on their suitability evaluated by a similarity metric and predefined rules. Some approaches used vector-based modelling and distributional vectors based on syntactic context to obtain the most suitable synonyms (Melamud et al., 2015b). Recent advances in contextual language models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), Embeddings from Language Models (ELMo) (Peters et al., 2018), and XLNet (Yang et al., 2019) have resulted in major breakthroughs in NLP. Because these models carry contextual information and have the ability of context-sensitive modelling of word probabilities, they have achieved the state-of-the-art (SOTA) results in lexical substitution as well. Some recent research efforts have improved lexical substitution by modifying the architecture of contextual embedding models (Zhou et al., 2019) whereas others integrated lexical resources to contextual embeddings to obtain the most suitable set of substitutes (Michalopoulos et al., 2022).

Methods based on lexical resources may fail to obtain the most relevant substitutes given that they predominantly focus on synonyms, hypernyms, and hyponyms. Moreover, they fail to consider the influence of the substitute on the global context of the given sentence (Zhou et al., 2019). Even though the contextual word embedding models consider the given context, they are unable to provide sufficient knowledge about the effect of the substitute on the overall meaning of a sentence.

To address these issues, the aim of this paper was to investigate the effect of introducing contextual sentence embeddings alongside contextual word

Proceedings of the 29th International Conference on Computational Linguistics, pages 4124–4135
October 12–17, 2022.

embeddings in the lexical substitution task. We first analysed the impact of the addition of contextual sentence information and then, investigated other methods to improve lexical substitution (Zhou et al., 2019; Michalopoulos et al., 2022). The proposed solution achieved the SOTA results on the LS07 and CoInCo datasets.

The main contributions of the paper were as follows:

- Analysis of the impact of adding sentence context for lexical substitution.
- Analysis of methods, which incorporate lexical resources and additional context information to improve lexical substitution.
- A lexical substitution solution, which outperformed previous SOTA methods, and its release at `https://github.com/sandaruSen/CILex` under the MIT license.

2 Related Work

Researchers have identified different subtasks under lexical substitution, namely substitution generation, substitution selection, and substitution ranking (Shardlow, 2014). Out of these, substitution generation and substitution ranking are considered as the two main subtasks, where the former focuses on generating possible substitutes for a target word given the context, and the latter aims to rank the substitutes (Giuliano et al., 2007; Martinez et al., 2007). Ranking of the substitutes may include ranking of the generated substitutes by the lexical substitution method or a much simpler ranking problem with ranking of the set of substitutes obtained from the human-annotated data given in the dataset (Erk and Padó, 2010; Thater et al., 2011).

Early efforts on lexical substitution relied mainly on manually curated lexical resources like WordNet (Miller, 1995) which evolved to the use of unsupervised methods and models based on distributional similarity. Word embeddings, such as word2vec (Mikolov et al., 2013) were used to obtain substitutes by selecting words with embeddings residing near the target word. The embedding similarity obtained from these models was used to rank the substitutes (Melamud et al., 2015b). The model context2vec, introduced by Melamud et al. (2016), produced the contextual embeddings for a given target word by combining the output of two bidirectional Long Short-Term Memory Networks (LSTMs) using a feedforward neural network. This model was successfully applied for the ranking of given substitutes in the lexical substitution task. ELMo used a similar approach with bidirectional LSTMs where the embedding of a given word was created based on the meaning of the context it appeared (Peters et al., 2018). ELMo was used in the lexical substitution task to rank the candidates by calculating the cosine similarity between the contextual embeddings from the ELMo for the target word and all the substitutes for the target word (Garí Soler et al., 2019).

The introduction of transformers resulted in major advances in a wide range of NLP tasks (Vaswani et al., 2017). Transformer-based language models trained on extra large corpora like BERT (Devlin et al., 2019) and a robustly optimised BERT (RoBERTa) (Liu et al., 2019) used a masked language modelling objective where tokens were replaced by a special token [MASK] in the training process. Further improving on the BERT-based language models, XLNet was introduced; it used an autoregressive pre-training method with a permutation-based language modelling objective without corrupting the input with masks (Yang et al., 2019). These contextual embedding models were extensively used for lexical substitution.

The authors in Zhou et al. (2019) relied on contextual word embeddings for lexical substitution. They modified the BERT architecture with a dropout embedding policy where the target word was partially masked with the aim of providing some information of the target word in the prediction. To evaluate the fitness of possible candidates, the authors introduced a validation score which was computed using representations in the top four layer's of BERT. The proposed method achieved the SOTA results for lexical substitution. Arefyev et al. (2020a) presented an extensive analysis on different contextual embedding models for lexical substitution. The authors, in addition to the model probability predictions for the target word, computed the word embedding similarity of the target with all the words in the model's vocabulary for final predictions. Their experimental comparisons indicated that XLNet had superior performance compared to other contextual word embedding models like ELMo, BERT, and RoBERTa at providing substitutes given no changes in the basic architecture of the models.

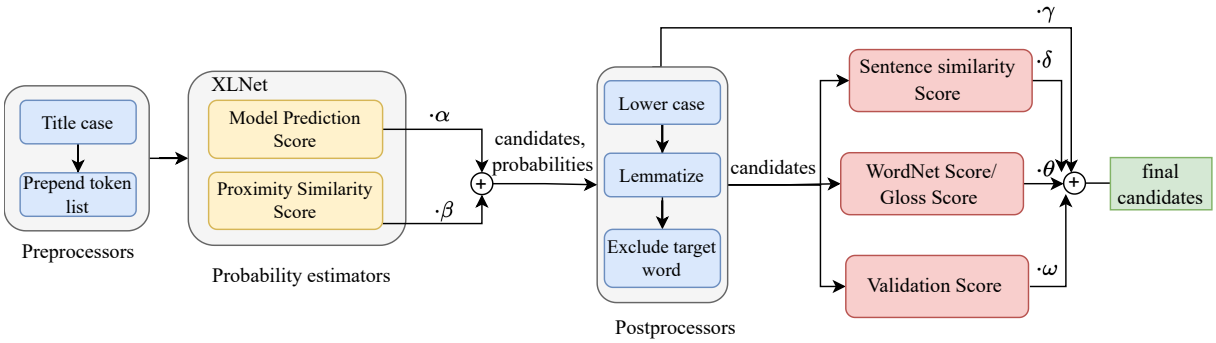

Figure 1: A flowchart of the proposed solution.

Michalopoulos et al. (2022) presented a framework, which integrated external knowledge from WordNet to BERT for lexical substitution. The authors computed a proposal score based on BERT and WordNet, a gloss sentence similarity score based on WordNet definitions, a sentence similarity score using a contextual sentence embedding model, and the validation score introduced by Zhou et al. (2019) to obtain the final set of substitutes. Compared to Zhou et al. (2019) and Arefyev et al. (2020a), additionally their approach integrated both WordNet and contextual sentence embeddings for lexical substitution. These authors' most recent methods on lexical substitution relied mostly on contextual word embedding models (Zhou et al., 2019; Arefyev et al., 2020a) and the use of a variety of methods that provide contextual information (Michalopoulos et al., 2022).

In our study, we specifically focused on the added value of sentence context for lexical substitution based on contextual sentence embeddings. We based our experiments on Arefyev et al. (2020a) which gave evidence of XLNet outperforming other contextual embedding models for lexical substitution, and analysed the impact of adding sentence context information. Additionally, we introduced a WordNet-based metric and investigated the methods proposed by Zhou et al. (2019); Michalopoulos et al. (2022) for lexical substitution.

3 The CILex Solution

In this study, we investigated the impact of sentence context and the methods that capture context information, and proposed a lexical substitution solution called CILex (Figure 1). We followed Arefyev et al. (2020a) and Michalopoulos et al. (2022) as the basis of our work.

3.1 Preprocessing Methods

To address the performance degradation of XLNet model for short contexts (Arefyev et al., 2020a), we explored two main preprocessing steps, namely, converting to title case and prepending strings to the input. Following Arefyev et al. (2020a), we tested out prepending two types of strings (a random set of strings followed by a meaningful string) to increase the input length and assessed the impact in the performance of our models. We observed a slight improvement in the results when only a meaningful string was prepended. When prepending with XLNet, to ensure the separation of the input and the string and to define the beginning of a sentence, use of a special end-of-document (<*eod*>) token (Arefyev et al., 2020b) and conversion of the first word in a sentence to title case were performed.

3.2 Contextual Word Embedding-based Scores

Model Prediction Score. Given a target word x and its context c, we obtained the probability provided by the XLNet model $P(w|c)$ as the model prediction (w is any word from the XLNet vocabulary). We used the XLNet model following Arefyev et al. (2020a) which gave evidence of XLNet outperforming other contextual word embedding models like BERT, ELMo, and RoBERTa without fine-tuning.

Proximity Similarity Score. In addition to the model prediction, we obtained the probability of possible substitutes based on their proximity to the target word $P(w|x)$ through embedding similarity which was computed using the inner product of the embedding of the target word and the embedding of the respective word $(embedding_x \cdot embedding_w^\top)$ (Arefyev et al., 2020a).

These probability scores were linearly combined for each word in the vocabulary to obtain $S_{\texttt{XLNet}}$ which is a representation of the model prediction and the embedding similarity (Eq. (1)).

$$S_{\texttt{XLNet}} = \alpha P(w|c) + \beta P(w|x) \quad (1)$$

where α is the weight for model prediction score and β is the weight for embedding similarity score. The values for parameters α and β can be fine-tuned.

Based on $S_{\texttt{XLNet}}$ score, the words were ranked to obtain the top 20 possible substitutes.

3.3 Contextual Sentence Embedding-based Scores

To evaluate the suitability of the possible candidates and their influence in the global context of the given sentence, we used contextual sentence embeddings with the assumption that contextual sentence embeddings are capable of ensuring that the possible substitutes do not change the overall meaning of the sentence.

Given a sentence s with a target word x_i, we obtained an updated sentence (s') by replacing the target word with a possible substitute. An updated sentence can be denoted as $s' = (x_1, \ldots, x'_i, \ldots)$. For each possible substitute, a sentence similarity score was then calculated using cosine similarity using the sentence embeddings for the original sentence s and the updated sentence s':

$$S_{\texttt{sent}} = \texttt{cos}(s, s'). \quad (2)$$

To obtain sentence embeddings, we experimented with a general sentence embedding model based on RoBERTa (*stsb-roberta-large*) (Reimers et al., 2019), unlike Michalopoulos et al. (2022) who used a fine-tuned RoBERTa sentence embedding model.

To investigate the added value of sentence context, the scores from the XLNet model and the sentence similarity model were linearly combined to obtain the candidate score S for the possible substitutes with γ and δ as the weights for $S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$ scores respectively. The model which relied only on $S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$ was defined as *CILex1*.

$$S = \gamma S_{\texttt{XLNet}} + \delta S_{\texttt{sent}}. \quad (3)$$

3.4 Additional Context Information-based Scores

Gloss Sentence Similarity Score. As introduced by Michalopoulos et al. (2022), we computed a gloss sentence similarity score $S_{\texttt{gloss}}$ based on WordNet and BERT (*bert-large-uncased*) which captured additional context information of the target word. For target words and possible substitutes, lists of potential definitions were obtained from WordNet. By computing the similarity score between the given sentence and the definitions, the most suitable definitions for each target word and substitute was obtained. For each substitute, a gloss similarity score was obtained by computing the cosine similarity between the best definition embedding of the target word d_t and best definition embedding of the substitute d_w.

$$S_{\texttt{gloss}} = \texttt{cos}(d_t, d_w). \quad (4)$$

WordNet Similarity Score. Similar to the $S_{\texttt{gloss}}$ score, we introduced a new score $S_{\texttt{wordnet}}$ based on WordNet and BERT (*bert-large-uncased*). Unlike $S_{\texttt{gloss}}$ score, we obtained lists of potential definitions only for the target words, from which the most suitable definition for the target word was obtained computing cosine similarity score between the given sentence and the definitions. By replacing the target word in the given sentence by possible substitutes, a list of updated sentences were obtained. For each substitute, wordnet based similarity score was obtained by computing the cosine similarity between the best definition of the target word d_t and the updated sentence s'.

$$S_{\texttt{wordnet}} = \texttt{cos}(d_t, s'). \quad (5)$$

Validation Score. We also used the validation score $S_{\texttt{val}}$ in Zhou et al. (2019) by computing the cosine similarities between the BERT-based contextual embeddings (*bert-large-uncased*) of the top four layers of every token in the original sentence and the modified sentence.

For each word filtered based on $S_{\texttt{XLNet}}$ score, $S_{\texttt{sent}}$, $S_{\texttt{gloss}}$, $S_{\texttt{wordnet}}$, and $S_{\texttt{val}}$ scores were calculated. The scores were then linearly interpolated to obtain the candidate score S for the possible substitutes with γ, δ, θ, and ω as the weights for $S_{\texttt{XLNet}}$, $S_{\texttt{sent}}$, $S_{\texttt{wordnet}}$, and $S_{\texttt{val}}$ scores respectively for *CILex2* (Eq. (6)). For *CILex3*, $S_{\texttt{wordnet}}$ score was replaced using $S_{\texttt{gloss}}$ (Eq. (7)) [1].

$$S = \gamma S_{\texttt{XLNet}} + \delta S_{\texttt{sent}} + \theta S_{\texttt{wordnet}} + \omega S_{\texttt{val}}. \quad (6)$$

$$S = \gamma S_{\texttt{XLNet}} + \delta S_{\texttt{sent}} + \theta S_{\texttt{gloss}} + \omega S_{\texttt{val}}. \quad (7)$$

[1] Both $S_{\texttt{wordnet}}$ and $S_{\texttt{gloss}}$ are based on WordNet and BERT models, and therefore not considered together.

Sentence	Candidate Substitutes with Weights
If we **take** the factual context in which the term is used into consideration ...	consider 2; accept 1; include 1; think about 1
It shouldn't **take** that long .	last 2; be 1; engage for 1
If you don't **take** the risk of dying by driving to the store ...	tolerate 1; run 1; undergo 1; accept 1; risk 1

Table 1: Three example instances for the target word $take$ from LS07 dataset.

4 Experiments

This section gives an overview of the datasets used in the experiments, evaluation metrics, the experimental setup, and the results from performance evaluations.

4.1 Datasets in Experiments

We evaluated CILex1, CILex2, and CILex3 on the two most widely used English datasets for lexical substitution task: the SemEval 2007 task dataset (LS07) (McCarthy and Navigli, 2007) and the Concepts in Context (CoInCo) (Kremer et al., 2014) dataset.

LS07 dataset is from the English Internet Corpus and consists of $2,010$ sentences for 201 target words with 10 sentences per target word. The annotators were asked to provide up to 3 candidate words for each target word.

CoInCo dataset is from the "Manually Annotated Sub-Corpus" and it consists of $15,415$ sentences with $3,874$ target words. For each target word, 6 candidate substitutes were provided by the annotators.

Each candidate substitute in the datasets was assigned a weight, which corresponds to the frequency it was chosen by the annotators. Table 1 provides an example of three instances of the target word $take$ with the candidate substitutes provided by the annotators in LS07 dataset.

4.2 Experimental Setup

We evaluated CILex on the two subtasks of lexical substitution: substitution generation and substitution ranking.

In the substitution generation task, possible substitutes for the target word were obtained from our proposed approach. We based our evaluation on the metrics proposed in the SemEval 2007 task (McCarthy and Navigli, 2007), in particular, we used *best* and *best-m* to evaluate the quality of the best predictions of the system and out of ten (*oot*) and *oot-m* to assess the coverage of the gold substitutes in the top ten best predictions respectively.[2] We also used $precision@1$ ($P@1$) and $precision@3$ ($P@3$) as evaluation metrics to have a thorough comparison with Zhou et al. (2019); Arefyev et al. (2020a); Michalopoulos et al. (2022). We computed the $P@k$, $k = \{1, 3\}$ as follows:

$$P@k = \frac{\text{acceptable substitutes in the system top-}k}{\text{substitutes in the system top-}k}.$$

To evaluate the statistical significance of the $P@1$ score of CILex methods, we used Wilcoxon Signed-Rank Test (Wilcoxon, 1992) and Pearson correlation (Benesty et al., 2009).

The substitution ranking task was performed based on the substitutes provided in the dataset. Following the previous works, we pooled all the candidate substitutes for the target word in the given instance across the dataset based on the target lemma and Part Of Speech (POS) tag and removed multi-words from the list (Melamud et al., 2015b; Arefyev et al., 2020a; Michalopoulos et al., 2022). The filtered out list was the input to the system as candidates to be ranked. As the gold standard, we used the given candidate substitutes. The proposed approach was then used to rank the possible candidates. CILex was evaluated for the candidate ranking task using the Generalised Average Precision (GAP) score (Kishida, 2005) where candidates are ranked based on their weights; candidates with higher weights should be ranked higher.

We experimented with weights when combining the scores together. Scores from the XLNet model (Eq. (1)) were computed by setting the α parameter to 1 and β parameter to 10 following (Arefyev et al., 2020a). When integrating the XLNet score $S_{\texttt{XLNet}}$ with sentence similarity score $S_{\texttt{sent}}$, empirically, we changed the γ parameter to 1, 0.5, and 0.05 keeping δ at 1. We obtained the best results when γ was 0.05. For θ and ω, we used 0.05 and

[2]For brevity, details of all the evaluation metrics are not described. More information can be found at McCarthy and Navigli (2007).

Method	*best*	*best-m*	*oot*	*oot-m*	*P*@1	*P*@3
			LS07 dataset			
Substitute Vector (Melamud et al., 2015a)	12.7	21.7	36.37	52.03	-	-
PIC (Roller and Erk, 2016)	-	-	-	-	19.7	14.8
Transfer Learning (Hintz and Biemann, 2016)	17.2	-	48.8	-	40.8	-
BERT-based substitution (Zhou et al., 2019)	20.3	34.2	55.4	68.4	51.1	-
BERT-based substitution*	12.8	22.1	43.9	59.7	31.7	-
XLNet+embs (Arefyev et al., 2020a)	21.32	37.80	55.04	73.90	50.56	36.29
LexSubCon (Michalopoulos et al., 2022)	21.1	35.5	51.3	68.6	51.7	-
CILex1	22.15	39.02	54.98	74.15	53.38	37.58
CILex2	23.17	40.98	55.51	73.90	55.43	38.15
CILex3	**23.31**	**40.98**	**56.32**	**74.88**	**55.96**	**38.5**
			CoInCo dataset			
Substitute Vector	8.1	17.4	26.7	46.2	-	-
BERT-based substitution	14.5	33.9	45.9	69.9	56.3	-
BERT-based substitution*	11.8	24.2	36.0	56.8	43.5	-
XLNet+embs	15.09	33.02	45.06	71.85	52.57	39.67
LexSubCon	14.0	29.7	38.0	59.2	50.5	-
CILex1	15.96	35.04	45.84	72.12	55.73	41.34
CILex2	16.30	35.73	46.55	72.84	56.77	42.3
CILex3	**16.39**	**35.80**	**46.87**	**72.98**	**57.25**	**42.49**

Table 2: Results of the best implementations of our approach and previous state-of-the-art models for LS07 and CoInCo datasets (Higher the value, better the performance). We reproduced the results of Arefyev et al. (2020a) and included reproduced results of the BERT-based substitution method (Zhou et al., 2019) by Michalopoulos et al. (2022) which is shown in *. Best values are bolded. (Results for the entire dataset can be found at Appendix A.)

0.5. The linear model parameters for LS07 dataset were fine-tuned against CoInCo dataset and vice versa (Arefyev et al., 2020a). We conducted our experiments on a RTX 3090 graphics card with 24 GB memory and CUDA 11.4.

4.3 Experimental Results from Performance Evaluations

Substitution Generation. Our proposed approaches outperformed all the previous SOTA lexical substitution methods for both datasets (Table 2). We compared our best performing approaches (Eq. (3), Eq. (6), Eq. (7)) with the previous best results from substitute vector-based method (Melamud et al., 2015a), PIC (Roller and Erk, 2016), transfer learning-based method (Hintz and Biemann, 2016), BERT for lexical substitution (Zhou et al., 2019), XLNet+embs method (Arefyev et al., 2020a), and LexSubCon method (Michalopoulos et al., 2022). *CILex3* outperformed the most recent method by Michalopoulos et al. (2022) on both datasets by a ∼4% improvement on LS07 dataset and ∼6.75% improvement on CoInCo dataset. *CILex3* also showed an improvement of ∼5% and ∼4.5% on LS07 and CoInCo respectively compared to (Arefyev et al., 2020a).

Method	**LS07**	**CoInCo**
Transfer Learning	51.9	-
Vector Space Modelling	52.5	47.8
PIC	52.4	48.3
Supervised Learning	55.0	-
Substitute Vector	55.1	50.2
context2vec	56.0	47.9
CILex1	56.81	51.68
CILex3	57.83	53.57
CILex2	58.25	53.92
BERT-based	58.6	55.2
XLNet+embs	60.5	55.64
LexSubCon	60.6	58.0

Table 3: Comparison of GAP scores (%) for the candidate ranking task. The results from the transfer learning (Hintz and Biemann, 2016), vector space modelling (Kremer et al., 2014), PIC (Roller and Erk, 2016), supervised learning (Szarvas et al., 2013), substitute vector (Melamud et al., 2015a), context2vec (Melamud et al., 2016), XLNet+embs (Arefyev et al., 2020a), BERT-based lexical substitution (Zhou et al., 2019), and LexSubCon (Michalopoulos et al., 2022) are presented.

Method	*best*	*best-m*	*oot*	*oot-m*	$P@1$	$P@3$	$R@10$	*Runtime*
				LS07 dataset				
$S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$	**22.15**	**39.02**	54.98	**74.15**	**53.38**	37.58	48.67	32 min 27 sec
$S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$*	21.76	38.70	**55.27**	73.90	52.38	**37.6**	**48.79**	32 min 29 sec
$S_{\texttt{XLNet}}$ and $S_{\texttt{wordnet}}$	21.53	38.37	54.59	72.76	50.85	35.39	48.22	25 min 57 sec
$S_{\texttt{XLNet}}$ and $S_{\texttt{gloss}}$	20.97	36.59	50.72	69.35	50.5	30.4	44.0	45 min 11 sec
$S_{\texttt{XLNet}}$ and $S_{\texttt{val}}$	21.98	38.46	54.39	72.93	52.73	36.78	48.15	59 min 37 sec
				CoInCo dataset				
$S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$	**15.96**	**35.04**	45.84	72.12	**55.73**	41.34	37.21	4 hrs 54 min
$S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$*	15.71	34.39	**46.07**	**72.80**	54.63	**41.88**	**37.31**	4 hrs 59 min
$S_{\texttt{XLNet}}$ and $S_{\texttt{wordnet}}$	15.23	33.39	44.48	70.84	53.07	38.14	35.77	3 hrs 18 min
$S_{\texttt{XLNet}}$ and $S_{\texttt{gloss}}$	14.95	32.35	41.46	66.22	52.37	34.54	33.27	6 hrs 28 min
$S_{\texttt{XLNet}}$ and $S_{\texttt{val}}$	15.63	34.39	44.76	70.57	54.64	39.97	36.26	8 hrs 47 min

Table 4: Ablation study of the proposed approach with $S_{\texttt{XLNet}}$ as the basis and different methods to obtain additional context information. In *, we used the fine-tuned RoBERTa model to compute sentence similarity scores.

All the proposed CILex solutions gave statistically significant improvement of $P@1$ score compared to (Arefyev et al., 2020a) for both datasets (P<0.05). Both CILex2 and CILex3 were statistically significantly better than CILex1. However, based on Pearson's correlation results for CILex2 and CILex3, we could observe a high level of correlation between the two methods.

Candidate Ranking. Our proposed approaches provided competitive results on both LS07 and CoInCo datasets for candidate ranking task (Table 3). We could observe that CILex approaches outperformed the transfer learning (Hintz and Biemann, 2016), vector space modelling (Kremer et al., 2014), PIC (Roller and Erk, 2016), supervised learning (Szarvas et al., 2013), substitute vector (Melamud et al., 2015a), and context2vec (Melamud et al., 2016) methods. However, BERT-based lexical substitution (Zhou et al., 2019), XLNet+embs (Arefyev et al., 2020a), and LexSubCon (Michalopoulos et al., 2022) reported better results than the proposed approach for candidate ranking.

Ablation Study on Substitution Generation. We conducted an ablation study to evaluate the effect of contextual sentence embeddings and different methods that capture context information introduced in (Zhou et al., 2019; Michalopoulos et al., 2022) (Table 4). We have presented results for the recall of the top 10 predictions ($R@10$) and run time for each experiment as additional metrics.

The results from our analysis of the contribution of fine-tuned contextual sentence embeddings and general contextual embeddings indicated that fine-tuned sentence embedding model on the dataset does not necessarily perform well for lexical substitution. We used $S_{\texttt{XLNet}}$ as the basis and experimented with two contextual sentence embedding models based on RoBERTa and the fine-tuned sentence embedding model based on RoBERTa (Michalopoulos et al., 2022). Based on our experiments, we observed that both models gave similar results for LS07 and CoInCo datasets.

We further performed experiments to analyse the relative contribution of additional context information obtained by $S_{\texttt{wordnet}}$, $S_{\texttt{gloss}}$, and $S_{\texttt{val}}$ with respect to the output from $S_{\texttt{XLNet}}$. Our results implied that the model achieves the worst performance when gloss sentence similarity score was used as additional context information. Based on our results, we also observed an increase in the final results when sentence similarity score was used to obtain additional context information.

To identify the efficient methods of integrating contextual information for lexical substitution, we reported the runtimes for our experiments (Table 4). The runtimes indicated that use of $S_{\texttt{wordnet}}$ is comparatively efficient. However, considering the computational vs performance trade-off, desired scores can be used for lexical substitution.

	Number of successful predictions			
Dataset	**0**	**1**	**2**	**3**
LS07	23.19	43.27	28.36	5.16
CoInCo	22.00	37.98	30.53	9.47

Table 5: The percentage of no. of samples in each dataset based on the number of successful predictions in the top three predictions.

Sentence	Gold Substitutes	Top Three Predictions
Nevertheless she gave me what i can only **describe** as as appraising glance	call, see, recount, imagine, detail, assess as	description, explain, define
Just another **wild** and crazy guy.	uninhibited, turbulent, rowdy, restless, peculiar, intense, insane, impassioned, fierce, adventurous	crazy, reckless, random
The federal complaint offers many **details** of the alleged conspiracy ...	specific, point, fact, tidbit, snippet, item, issue, facet, count, account	information, description, outline
He **said**.	state, remark, declare, comment, cite, ask, answer	speak, tell, reply

Table 6: Four example instances from CoInCo dataset whose top three model predictions were not in the gold substitutes and the extracted possible substitutes from WordNet. The target words are bolded.

4.4 Analysis of the Substitution Generation Results

We conducted experiments to further analyse the performance of the CILex architecture. For each dataset, we calculated the percentage of successful predictions as the samples for which the model returned at least one prediction that was included in the gold substitutes, considering the top three predictions (Table 5). We observed that for 76.8% of the LS07 dataset and 77.99% of the CoInCo dataset CILex provided at least one successful prediction. We further analysed the 23.19% and 22% of LS07 and CoInCo datasets for which our method did not yield a successful prediction.

For the target words in the samples which did not yield at least one successful prediction, we extracted synonyms, hypernyms, and hyponyms from WordNet as possible substitutes and checked if at least one of the model predictions was in the extracted WordNet substitutes. Results from CILex indicated that 51.64% of LS07 and 25.69% of CoInCo which did not yield a successful prediction, contained predictions that were included in the extracted WordNet substitutes. This implied that, even though for certain samples there were no successful predictions based on the gold substitutes, they included predictions with a certain relevance to the target word. We manually checked the remaining samples, for which our method did not yield a successful prediction and which were not included in the WordNet substitutes, as illustrated in Table 6. In the analysed instances, we observed predictions which could be considered as possible substitutes for the given target word.

5 Discussion

In this paper, we analysed the impact of introducing contextual sentence embeddings and methods which provide additional context information for lexical substitution, thereby ensuring that the substitutes are semantically consistent while preserving the overall meaning of the sentences. The results from the proposed CILex solution outperformed previous SOTA methods for lexical substitution on LS07 and CoInCo datasets. Our results indicated that accounting for sentence context information has improved the performance on the substitution generation task. This is demonstrated by our approach (i.e., CILex) outperforming (Arefyev et al., 2020a) SOTA contextual word embedding-based method on two datasets. However, interestingly based on our results, the performance did not improve in the candidate ranking task, which requires further investigation.

The results from our ablation study on the methods, which provide additional context information, indicated that they improve the lexical substitution task. Analogously with Michalopoulos et al. (2022), our results implied that the model achieves the worst performance when gloss sentence similarity score was used as additional context information. This could be mainly due to WordNet being manually curated and definitions for words obtained from WordNet might not reflect the meaning of the words in the given context. Furthermore, for certain words, definitions may not be available on WordNet. Based on our results, we also observed an increase in the final results when sentence similarity score was used to obtain additional context information as opposed to wordnet score, gloss sen-

tence similarity score, and validation score. This is likely because the sentence similarity can appropriately identify if a substitute fits the context and ensures that the overall meaning of the sentence is unchanged. However, when compared with Michalopoulos et al. (2022), their approach yielded an improvement in results when the validation score introduced by (Zhou et al., 2019) was used. These observations hindered us to come to a conclusion as to which of the components contributes most to lexical substitution and illustrated that different components contributed differently.

Our experiments gave evidence of the importance of the initial model/method used to obtain the first set of substitutes. The proposed CILex solution, which relied on an XLNet-based method to obtain the initial set of substitutes, outperformed Lexsubcon (Michalopoulos et al., 2022) which used a BERT-based method to obtain the initial set of substitutes. CILex showed an improvement of $\sim$4% on LS07 dataset and $\sim$6.75% on CoInCo dataset. These insights guided us to conclude that the initial model used to obtain the possible candidates also had a direct impact to the lexical substitution which requires a thorough investigation.

The proposed CILex solutions are based on pre-trained language models and therefore are generalisable. For domain specific applications, the proposed approach can be easily transferable by replacing the pre-trained language models with respective domain specific models.

Further analysis of the substitution generation results indicated that samples which did not have successful predictions may contain potential substitutes based on WordNet. This illustrated the impact of the annotation subjectivity in the interpretation of the performance of the lexical substitution task.

As future work, we intend to extend our experiments and analyse methods that can provide context information to improve lexical substitution. Moreover, the impact on the candidate ranking task will be explored further as future work. We also plan to look into the applicability of the proposed approach for other tasks like word sense induction (Amrami and Goldberg, 2018) and word sense disambiguation.

6 Conclusion

We have presented and released a solution for lexical substitution investigating the impact of sentence context obtained from contextual sentence embeddings. We have introduced and further integrated methods which capture additional context information as proposed by Zhou et al. (2019); Michalopoulos et al. (2022). The unified solution has achieved the SOTA results on two benchmark datasets; LS07 and CoInCo.

We have also analysed and evaluated effects of different methods that provide contextual information and their contribution for lexical substitution. The results have demonstrated the importance of sentence context information obtained using contextual sentence embeddings in lexical substitution.

7 Ethical Considerations

We proposed a solution for lexical substitution and analysed the impact of adding sentence context using contextual sentence embeddings. Additionally, we also incorporated scores proposed in (Zhou et al., 2019; Michalopoulos et al., 2022) for lexical substitution to evaluate the significance of each of them.

The proposed approach was tested and validated on two benchmark datasets: LS07 dataset and CoInCo dataset. According to the National Statement on Ethical Conduct in Human Research (2007) — Updated 2018 (National Health and Medical Research Council, 2018), a new ethics approval was not required for our experiments and, to the best of our knowledge, both datasets were created ethically. No new data or annotations were collected as part of our study.

Acknowledgement

This research was funded by and has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians and healthcare providers. We gratefully acknowledge the funding from the ANU School of Computing for the first author's PhD studies. We wish to thank the reviewers for their valuable comments. We also thank the authors of Arefyev et al. (2020a); Michalopoulos et al. (2022) for releasing their code and clarifying the questions we came across.

References

Asaf Amrami and Yoav Goldberg. 2018. Word sense induction with neural biLM and symmetric patterns.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels, Belgium. Association for Computational Linguistics.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020a. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020b. A comparative study of lexical substitution approaches based on neural language models. *arXiv preprint arXiv:2006.00031*.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala, Sweden. Association for Computational Linguistics.

Aina Garí Soler, Anne Cocos, Marianna Apidianaki, and Chris Callison-Burch. 2019. A comparison of context-sensitive models for lexical substitution. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 271–282, Gothenburg, Sweden. Association for Computational Linguistics.

Claudio Giuliano, Alfio Gliozzo, and Carlo Strapparava. 2007. FBK-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 145–148, Prague, Czech Republic. Association for Computational Linguistics.

Gerold Hintz and Chris Biemann. 2016. Language transfer learning for supervised lexical substitution. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 118–129, Berlin, Germany. Association for Computational Linguistics.

Kazuaki Kishida. 2005. *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 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*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Caterina Lacerra, Tommaso Pasini, Rocco Tripodi, and Roberto Navigli. 2021. Alasca: an automated approach for large-scale lexical substitution. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3836–3842.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David Martinez, Su Nam Kim, and Timothy Baldwin. 2007. MELB-MKB: Lexical substitution system based on relatives in context. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 237–240, Prague, Czech Republic. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 48–53.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015a. Modeling word meaning in context with substitute vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 472–482, Denver, Colorado. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Oren Melamud, Omer Levy, and Ido Dagan. 2015b. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, Denver, Colorado. Association for Computational Linguistics.

George Michalopoulos, Ian McKillop, Alexander Wong, and Helen Chen. 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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

National Health and Medical Research Council. 2018. National Statement on Ethical Conduct in Human Research (2007). https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018. [Online; accessed 06-January-2022].

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 671–688. Association for Computational Linguistics.

Stephen Roller and Katrin Erk. 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*, pages 1121–1126, San Diego, California. Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

György Szarvas, Róbert Busa-Fekete, and Eyke Hüllermeier. 2013. Learning to rank lexical substitutions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1926–1932, Seattle, Washington, USA. Association for Computational Linguistics.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 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.

A Results of the Entire Dataset

The CILex solutions are based on pre-trained contextual embedding models and therefore no training is performed. Hence, we have also provided the results of the proposed approaches for the entire dataset (Tables 7, 8, and 9).

Method	*best*	*best-m*	*oot*	*oot-m*	$P@1$	$P@3$
			LS07 dataset			
XLNet+embs (Arefyev et al., 2020a)	20.88	36.92	53.60	71.74	49.53	34.9
CILex1	21.50	37.68	53.53	72.02	51.92	36.25
CILex2	22.47	39.43	53.96	71.81	53.77	36.76
CILex3	**22.59**	**39.22**	**54.65**	**72.37**	**54.22**	**37.33**
			CoInCo dataset			
XLNet+embs	14.11	31.70	44.03	71.62	51.5	39.5
CILex1	15.37	34.71	45.09	72.25	56.12	42.24
CILex2	15.67	35.20	45.73	72.88	57.09	43.12
CILex3	**15.67**	**34.97**	**46.03**	**73.07**	**57.35**	**43.27**

Table 7: Results of the best implementations of our approach for the whole dataset (trial and test) and the reproduced results of XLNet+embs method (Arefyev et al., 2020a).

Method	*best*	*best-m*	*oot*	*oot-m*	$P@1$	$P@3$	$R@10$	*Runtime*
				LS07 dataset				
$S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$	**21.50**	**37.68**	**53.53**	**72.02**	**51.92**	**36.25**	**47.6**	37 min 5 sec
$S_{\texttt{XLNet}}$ and $S_{\texttt{wordnet}}$	20.83	36.99	53.08	70.41	49.23	34.18	47.1	27 min 55 sec
$S_{\texttt{XLNet}}$ and $S_{\texttt{gloss}}$	20.26	35.24	49.27	66.99	48.63	29.21	43.04	55 min 46 sec
$S_{\texttt{XLNet}}$ and $S_{\texttt{val}}$	21.29	37.19	52.78	70.41	50.97	35.1	46.95	1 hr 7 min
				CoInCo dataset				
$S_{\texttt{XLNet}}$ and $S_{\texttt{sent}}$	**15.37**	**34.71**	**45.09**	**72.25**	**56.12**	**42.24**	**36.33**	7 hr 58 min
$S_{\texttt{XLNet}}$ and $S_{\texttt{wordnet}}$	14.53	32.78	43.40	70.72	52.77	38.58	34.62	4 hr 52 min
$S_{\texttt{XLNet}}$ and $S_{\texttt{gloss}}$	14.32	31.82	40.74	66.50	52.41	35.41	32.42	10 hr 6 min
$S_{\texttt{XLNet}}$ and $S_{\texttt{val}}$	15.02	33.91	43.85	70.51	54.59	40.61	35.29	13 hr 58 min

Table 8: Ablation study of the proposed approach for the complete dataset (trial and test).

Method	**LS07**	**CoInCo**
CILex3	56.91	53.55
CILex2	57.42	53.93
BERT-based	57.9	55.5
XLNet+embs	59.61	55.64
CILex1	59.9	55.6
LexSubCon	60.3	58.0

Table 9: Comparison of GAP scores (%) for the candidate ranking task on the entire dataset (trial and test). The results of CILex methods, reproduced results of XLNet+embs (Arefyev et al., 2020a), reported results of BERT-based lexical substitution (Zhou et al., 2019), and LexSubCon by Michalopoulos et al. (2022) are presented.