Grounding the Lexical Substitution Task in Entailment

Talgat Omarov and **Grzegorz Kondrak**
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{omarov,gkondrak}@ualberta.ca

Abstract

Existing definitions of lexical substitutes are often vague or inconsistent with the gold annotations. We propose a new definition which is grounded in the relation of entailment; namely, that the sentence that results from the substitution should be in the relation of mutual entailment with the original sentence. We argue that the new definition is well-founded and supported by previous work on lexical entailment. We empirically validate our definition by verifying that it covers the majority of gold substitutes in existing datasets. Based on this definition, we create a new dataset from existing semantic resources. Finally, we propose a novel context augmentation method motivated by the definition, which relates the substitutes to the sense of the target word by incorporating glosses and synonyms directly into the context. Experimental results demonstrate that our augmentation approach improves the performance of lexical substitution systems on the existing benchmarks.

1 Introduction

Lexical substitution is the task of finding appropriate replacements for a target word in a given context sentence. This task was first introduced as an application-oriented alternative to word sense disambiguation (WSD) that does not depend on a predefined sense inventory (McCarthy, 2002). Lexical substitution has been applied in various tasks, such as word sense induction (Amrami and Goldberg, 2018), lexical relation extraction (Schick and Schütze, 2020), and text simplification (Al-Thanyyan and Azmi, 2021).

Lexical substitution continues to be an important area of research in NLP. For instance, it can be used to probe the ability of NLP models to capture contextual meaning, as substitutes can vary depending on the sense of the word. Furthermore, professional writers often need good substitutes in a specific context, which cannot be found by simply looking them up in a thesaurus.

Many definitions used in the literature to describe lexical substitution are either vague or inconsistent with the evaluation datasets. For example, Hassan et al. (2007) and Roller and Erk (2016) leave the criteria for lexical substitution to the discretion of human annotators. Studies such as Sinha and Mihalcea (2009, 2014) and Hintz and Biemann (2016) require substitutes to be synonyms, which creates a discrepancy with established lexical substitution benchmarks that allow annotators to provide slightly more general terms (hypernyms) (McCarthy, 2002; Kremer et al., 2014). For example, while the two words are not synonyms, *vehicle* can be considered as a valid substitute for *car* if the context clearly refers to a car. Most prior work requires substitutes to preserve the meaning of the original sentence (McCarthy and Navigli, 2007; Giuliano et al., 2007; Szarvas et al., 2013a,b; Kremer et al., 2014; Melamud et al., 2015; Garí Soler et al., 2019; Zhou et al., 2019; Lacerra et al., 2021; Michalopoulos et al., 2022; Seneviratne et al., 2022; Wada et al., 2022). However, as we show in this work, not all gold substitutes necessarily preserve the meaning of the sentence taken in isolation.

We propose a definition of lexical substitution that is more precise and well-founded. Our aim is not only to address the inconsistency in the literature but also to align the task definition with established evaluation datasets. We draw on insights from natural language inference (NLI), which provides a framework for understanding the semantic relationship between sentences and words. According to our definition, the sentence that results from a lexical substitution must be in the relation of *mutual entailment* with the original sentence. For example, *position* is a suitable substitute for *post* in the sentence "I occupied a *post* in the treasury" because the two sentences entail each other. The entailment criterion takes into account the implicit

Findings of the Association for Computational Linguistics: ACL 2023, pages 2854–2869

background knowledge (Dagan et al., 2005), which allows lexical substitution to generalize over simple synonym replacement, encompassing a wider range of semantic relations, such as hypernymy and meronymy (Geffet and Dagan, 2005).

The classification of the entailment relation between two sentences requires the identification of the target word's sense. For example, *position* is a proper substitute for *post* only if it is used in the sense corresponding to "job in an organization". Based on this observation, we develop an augmentation method that helps to ground the substitutes by incorporating glosses and synonyms of the target word's sense directly into the context. Since the word sense is latent, the method leverages a WSD system to account for the probabilities of each candidate sense.

We show the effectiveness of the proposed definition and our augmentation method through experiments on existing lexical substitution datasets. Our analysis indicates that the proposed definition encompasses gold substitutes that could not previously be explained by existing definitions. Furthermore, our empirical evaluation shows that our augmentation method improves the performance on the lexical substitution benchmarks by up to 4.9 F1 points, surpassing the previous state-of-the-art models in certain settings.

The main contributions of this paper are as follows.

1. We propose a task formulation for lexical substitution that is grounded in entailment and show its suitability for existing datasets.

2. We construct a new dataset for lexical substitution, which demonstrates the applicability of our theoretical definition.

3. By facilitating the identification of the latent word senses, our method improves results on existing lexical substitution benchmarks.

2 Related Work on Lexical Substitution

In this section, we review the available datasets and provide a brief overview of the prior work.

2.1 Datasets

The first English lexical substitution dataset was created by McCarthy and Navigli (2007) for SemEval-2007 Task 10. The dataset, which we refer to as SE07, consists of 2003 context sentences with one target word per sentence. The authors instructed the annotators to provide substitutes that preserve the original meaning of the sentence.

Biemann (2012) constructed Turk Bootstrap Word Sense Inventory (TWSI), which encompasses a sense inventory induced by lexical substitutes for 1,012 common English nouns. It was created by annotating 25,851 sentences with lexical substitutes using Amazon Mechanical Turk.

Kremer et al. (2014) introduced CoInCo, an "all-word" lexical substitution dataset, in which all content words in a corpus are annotated with substitutions. According to the authors, the all-word setting provides a more realistic distribution of target words and their senses. It is important to note that both McCarthy and Navigli (2007) and Kremer et al. (2014) explicitly allowed annotators to provide phrases or more general words when they could not think of a good substitute.

The SWORDS dataset (Lee et al., 2021) is based on the CoInCo dataset but uses a slightly different annotation approach. Instead of relying on annotators to come up with substitutes from their memory, they were provided with a list of candidate substitutes from a thesaurus and CoInCo for a given target word. The dataset contains 1,250 context sentences, each with a single target word.

The task of lexical substitution is not limited to the English language, and datasets have also been created for other languages, including Italian (Toral, 2009), and German (Cholakov et al., 2014); the latter dataset includes sense annotations (Miller et al., 2016). In addition, a cross-lingual dataset from SemEval-2010 Task 2 (Mihalcea et al., 2010) combines English target words and sentences with Spanish gold substitutes. While multilingual and cross-lingual tasks are beyond the scope of this paper, our proposed grounding of lexical substitution in entailment is also applicable in those settings.

2.2 Methods

Numerous methods have been proposed for lexical substitution. Early methods retrieve candidate substitutes from lexical resources such as WordNet (Miller, 1995). Approaches that rank candidate substitutes are based on web queries (Zhao et al., 2007; Martinez et al., 2007; Hassan et al., 2007), ngram models (Giuliano et al., 2007; Yuret, 2007; Dahl et al., 2007; Hawker, 2007; Hassan et al., 2007), latent semantic analysis (Giuliano et al., 2007; Hassan et al., 2007), delexicalized features (Szarvas

et al., 2013a), and word embeddings (Melamud et al., 2015, 2016; Roller and Erk, 2016).

Pre-trained neural language models (NLMs) and their contextualized embedding representations have greatly advanced the state of the art in lexical substitution. Garí Soler et al. (2019) use contextual embeddings from ELMo (Peters et al., 2018) to calculate the similarity between the target and candidate substitutes. To fix the bias toward the target word, Zhou et al. (2019) apply a dropout embedding policy that partially masks the target word's BERT embedding. Arefyev et al. (2020) propose combining a masked language model probability score with a contextual embedding-based proximity score. Lacerra et al. (2021) propose training a supervised sequence-to-sequence model that takes a context sentence containing a target word as input, and outputs a comma-separated list of substitutes. Wada et al. (2022) employ contextualized and decontextualized embeddings (the average contextual representation of a word in multiple contexts). Yang et al. (2022) inject information about the target word into context and use BERT to generate initial candidates. Furthermore, they train RoBERTa on the Multi-Genre Natural Language Inference corpus (Williams et al., 2018) to further refine the ranking by semantic similarity scores.

Similar to our method, two recent proposals leverage knowledge from WordNet to improve the quality of substitutes retrieved from pretrained neural language models. Michalopoulos et al. (2022) inject synonyms by linearly interpolating their contextual embeddings, while we insert synonyms and glosses directly into the context. Seneviratne et al. (2022) and the other approach of Michalopoulos et al. (2022) use knowledge from WordNet only at the ranking stage after candidates had been generated from an NLM. In contrast, our approach injects WordNet information into the NLM's input from the beginning, which may produce more relevant candidates initially.

3 Entailment-Based Lexical Substitution

In this section, we provide background information about entailment, present the theoretical formulation of the proposed definition, and demonstrate its suitability through empirical validation.

3.1 Entailment

A premise (P) *entails* a hypothesis (H) if a human reader of P would infer that H is most likely true (Dagan et al., 2005). Entailment is denoted as $P \models H$. For example, the premise "the water is boiling" entails the hypothesis "the water is hot". This definition of entailment assumes a common human understanding of language, as well as common background knowledge. Entailment is a directional relation, which means that $P \models H$ does not imply $H \models P$. For example, "I own a car" entails "I own a vehicle" but not the other way around. However, if $P \models H$ and $H \models P$ then H and P are semantically equivalent: $P \equiv H$ (MacCartney, 2009).

Lexical entailment is a subset of textual entailment that specifically examines the relationship between a premise and a hypothesis where the two differ by a single word or phrase (Kroeger, 2018). It has previously been established that words in context often entail their synonyms, hypernyms, and, in some cases, holonyms (Geffet and Dagan, 2005).

3.2 Lexical Substitution Definition

We anchor our definition of lexical substitution in textual entailment. Let C_t be a context sentence that contains a target word t, and let C_w be the same context sentence where t is replaced with a word or phrase w. We define w as a lexical substitute for t in C_t if and only if C_t and C_w entail each other:

$$\text{LexSub}(C_t, w) \Leftrightarrow C_t \models C_w \wedge C_w \models C_t$$

This binary definition can be adapted to the task of substitute generation by considering a finite set of all words and short phrases. Specifically, the output of the generation task would consist of all candidate substitutions that satisfy the above condition.

While entailment is recognized as an important substitutability criterion within the NLI community (Geffet and Dagan, 2004; Zhitomirsky-Geffet and Dagan, 2009), it has been largely overlooked in lexical substitution. A notable exception is Giuliano et al. (2007), who recognize the significance of the relationship between lexical substitution and entailment. Although their mutual textual entailment criterion is similar to ours, we disagree with their conclusion that the mutual equivalence requirement restricts substitutes to synonyms only. Next, we show that this criterion not only extends beyond word synonymy, but also naturally allows for the integration of common-sense reasoning and knowledge about the world.

3.3 Semantic Equivalence

In this section, we explicitly spell out our assumptions about the relationship between lexical substitution and the criterion of meaning preservation.

The first proposition states that all contextual synonyms are good substitutes.

Proposition 1. *If t and w express the same concept in C then w is a lexical substitute for t in C.*

Proof. When we replace a target word with another word that expresses the same concept in a given context, the truth conditions of the sentence do not change. This is because the truth conditions are determined by the relationships between concepts that are expressed in the sentence. Therefore, the mutual entailment between C_w and C_t must hold, which by our definition implies that w is a lexical substitute for t in the context C. □

If words express the same concept in some context, they must belong to the same wordnet synset (Hauer and Kondrak, 2020). A wordnet is a lexical ontology in which words are grouped into sets of synonyms (synsets), each representing a distinct concept (Miller, 1995). The suitability of contextual synonyms with lexical substitution provides a theoretical basis for the use of wordnets to generate substitutes (McCarthy and Navigli, 2007).

The implication in Proposition 1 is unidirectional; that is, not all substitutes must be synonyms.

Proposition 2. *If w is a lexical substitute for t in C then t and w do not necessarily represent the same concept in C .*

As evidence that the reverse implication does not hold, we provide a counter-example. Consider the following sentence from the SWORDS dataset: *"Those hospitals were not for us. They were for an expected invasion of Japan."* where the word *planned* is among the gold substitutes for the target word *expected*. While the verbs *expect* and *plan* are not synonyms, this particular substitution is correct considering the broader historical context of World War II, which has been provided in previous sentences. From the point of view of the US military, the invasion was both planned and expected. Thus, although the two words do not express the same concept, the corresponding sentences entail each other.

Taken together, these two propositions imply that synonymy within a narrow context is a sufficient but not a necessary condition for mutual entailment between the sentences. Thus, mutual entailment provides a more flexible criterion for substitution than contextual synonymy. The mutual entailment criterion captures the nuances of lexical substitution better than the definitions based on strict meaning preservation because it takes into account both context *and* background knowledge. This is essential to identify a wider range of substitutions in scenarios such as the ones described above. Furthermore, this definition may facilitate the job of annotators by breaking down lexical substitution into two concrete entailment conditions, which are easier to reason about.

		Strict meaning preservation	
		Covered	Not covered
Ours	Covered	35	6
	Not covered	0	9

Table 1: The number of substitutes in a random sample which are captured by our entailment-based definition vs. the existing definition of meaning preservation.

3.4 Empirical Validation

To validate our proposed definition, we perform a manual analysis of a random sample of 50 gold substitutes from the SWORDS dataset which are labeled "acceptable" (i.e., high quality). Our objective is to assess whether these substitutes are adequately covered by our definition. We provide a detailed description of our manual analysis procedure and examples in Appendix A.

The summary of our manual analysis is presented in Table 1. It shows that our definition successfully covers 41 (82%) of gold substitutes. All 9 substitutes that are not covered by our definition are also not covered by the existing definition of meaning preservation. This finding matches our Proposition 1, which implies that a word that is not a lexical substitute (i.e., mutual entailment does not hold), cannot express the same concept (i.e. there is a difference in meaning). We conclude that those 9 instances represent annotation errors (rows 1-9 in Table 5).

We also observe that the 6 substitutes that are not covered by the existing definition of meaning preservation are covered by our definition (rows 10-15 in Table 5). For example, consider the context *"Energy Secretary Bill Richardson went to Baghdad in 1995 while a representative for New Mexico,"* where *elected official* is a gold substitute for *representative*. The new sentence induced by the substitution does not preserve the original mean-

ing because not every elected official is a congress representative. However, the sentence provides enough historical context to validate the substitution. This observation matches our Proposition 2, which states that lexical substitutes need not represent the same concept.

3.5 Dataset Induced by Entailment

Based on Proposition 1, we use synonyms from existing semantic resources to construct a new lexical substitution dataset, which we refer to as WNSub.[1] This is because replacing target words with synonyms is guaranteed to generate sentences that satisfy the mutual entailment criterion.

To generate the WNSub dataset, we use SemCor (Miller et al., 1994), the largest corpus manually annotated with WordNet senses. The sense annotations are crucial for our dataset, as contextual synonyms are defined in relation to word senses rather than word lemmas. For example, for the sentence *"can your insurance company aid you in reducing administrative costs?"* we retrieve substitutes *help* and *assist* from the WordNet synset that corresponds to the annotated sense of the target word *aid*. In total, we obtain 146,303 sentences with 376,486 substitutes.

Although contextual synonyms do not necessarily capture all aspects of lexical substitution, WNSub can be used for pre-training supervised systems, in combination with other datasets. We verify this claim experimentally in Section 5.3.

4 Sense-based Augmentation Method

In this section, we describe our sense-based augmentation method for lexical substitution. Our approach is based on the observation that knowing the sense of the target word is key to deciding whether a substitution induces an entailment relation between the two sentences. For example, *position* is a proper substitute for *post* in some context only if the latter is used in the sense corresponding to "job in an organization". We posit that inserting sense glosses directly into the context will help lexical substitution systems identify substitutes that are mutually entailed by the original context. Our hypothesis is supported by prior findings that this technique works well for semantic tasks such as WSD (Huang et al., 2019) and idiomaticity detection (Hauer et al., 2022).

[1]Dataset and code available at https://github.com/talgatomarov/wnsub

Our method is based on two stand-alone modules: a WSD system and a lexical substitution generation system. The method is sufficiently flexible to incorporate new systems as the state of the art on those two tasks continues to improve. The only requirement is that these systems output probabilities for each candidate sense or substitute.

The formula below is used to combine the probabilities from the two systems. Figure 1 shows an example of soft constraint augmentation. Let C_t be a context sentence containing the target word t, w be a candidate substitute, and $s \in senses(t)$ be a candidate sense for t in C_t. Under the assumption that the substitutes depend on the sense of the target word, the conditional probability $P(w|C_t)$ can be derived by marginalizing the senses out:

$$P(w|C_t) = \sum_{s \in senses(t)} P(w|C_t, s) \times P(s|C_t)$$

In the equation above, we model $P(s|C_t)$ using a WSD system, and obtain $P(w|C_t, s)$ from a lexical substitution system that operates on the context augmented with sense information.

Motivated by the work of Luan et al. (2020), we experiment with two types of constraint: hard and soft. In the hard-constraint approach, a WSD system is used to identify the most likely sense of the target word, which is effectively assigned the probability of 1.0. Next, the glosses and synonyms corresponding to this sense are retrieved from a lexical resource and inserted in parentheses after the target word. This augmented context is then passed to a lexical substitution system, which generates substitutes along with their substitute probabilities. In the soft-constraint approach, for each possible sense of the target word, a WSD system first computes its probability, the context is augmented with glosses and synonyms of that sense, and finally a lexical substitution system generates and assigns final probabilities to candidate substitutes using the formula above.

Soft constraint allows grounding of lexical substitutes in the target word senses, while taking into account the probability of each candidate sense. We posit that considering all candidate senses and their probabilities should work better than committing to a single most likely sense, by improving robustness against WSD errors. In addition, in some cases, the context itself may not provide enough information to reliably disambiguate the sense of the target word. We verify this hypothesis experimentally in the next section.

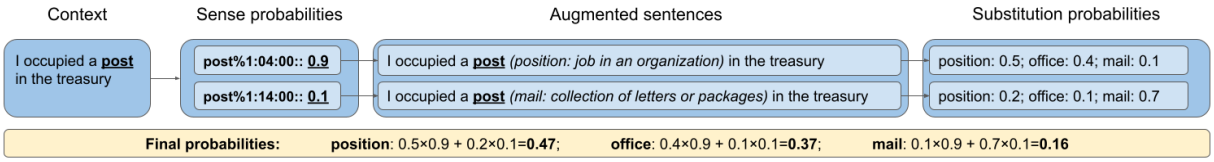

Figure 1: An example of augmenting a context with target word definitions, and calculating substitute scores. For brevity, not all candidate senses and substitutes are shown.

5 Experiments

In this section, we investigate the effectiveness of our dataset and augmentation method in improving the performance of lexical substitution systems. The experiments were conducted on a machine with two NVIDIA GeForce RTX 3090 video cards.

5.1 Evaluation Datasets and Metrics

We evaluate our methods using test splits from two benchmarks: the SemEval 2007 Task 10 (SE07) (McCarthy and Navigli, 2007) and SWORDS (Lee et al., 2021). Each benchmark has its own set of evaluation metrics, which we outline here.

The SE07 benchmark uses *best* and *oot* metrics, which measure the quality of the system's top-1 and top-10 predictions, respectively. These metrics assign weights to gold substitutes based on how frequently annotators selected them. The benchmarks also use *mode* variations of *best* and *oot*, which evaluate performance against a single gold substitute chosen by the majority of annotators, provided that such a majority exists. We consider the *mode* metrics theoretically problematic because they disregard instances without an annotation majority, and because many instances could involve multiple equally valid substitutes,

The SWORDS benchmark uses F^{10} scores, the harmonic mean of precision and recall, calculated with respect to the system's *top 10 predictions* and *acceptable* (F^{10}_a) or *conceivable* (F^{10}_c) gold substitutes. A candidate is labeled as *conceivable* if it was selected by at least one annotator and *acceptable* if selected by at least half of the annotators. Furthermore, the benchmark includes two evaluation settings: lenient and strict. In the lenient setting, any system-generated substitutes that are not in SWORDS are removed. In the strict setting, all system-generated substitutions are considered. The lenient settings were originally proposed to compare against "oracle" baselines whose predictions are guaranteed to be in SWORDS. We posit that the lenient setting provides an unreliable basis for measuring lexical substitution performance in real-world scenarios because systems are not provided with a predefined vocabulary of possible words that can occur during testing.

All existing evaluation metrics require a ranking mechanism to select top-k system predictions, which is problematic for two reasons. First, there is a lack of clarity on objective criteria for ranking substitute words. For example, in the sentence *"the FBI said that explicit conversations about the scheme had been recorded"*, it is debatable whether *disclosed* is a better substitute for *said* than *declared*. Second, the existing metrics reward systems for generating a specific number of candidates, regardless of how many substitutes actually exist. This may result in an inaccurate evaluation of the system's ability to generate correct substitutes.

Despite these limitations, our method builds upon existing systems that have been optimized using these metrics, and therefore we use them for the evaluation. However, we posit that it would be beneficial for future lexical substitution systems to consider metrics that do not depend on substitution ranking, such as the standard F1 score calculated with respect to all predicted substitutes.

5.2 Comparison Systems

On the SE07 dataset, we compare against KU (Yuret, 2007), supervised learning (Szarvas et al., 2013a), BERT for lexical substitution (Zhou et al., 2019), GeneSis (Lacerra et al., 2021), LexSubCon (Michalopoulos et al., 2022), and CILex (Seneviratne et al., 2022). The reported results are from the last two papers.

On the SWORDS dataset, we compare against GPT-3 with "in-context" learning (Brown et al., 2020), a commercial lexical substitution system Word-Tune[2], and a BERT baseline which produces substitutes according to the masked language modeling head (Devlin et al., 2019). The results of these models are reported by Lee et al. (2021). We also include the results of Yang et al. (2022).

[2] https://www.wordtune.com

5.3 WNSub Experiments

The objective of the experiments with WNSub (Section 3.5) is to determine whether the dataset could enhance the performance of supervised sequence-to-sequence lexical substitution models when used as a pre-training dataset.

The first model is our own implementation of a simple supervised sequence-to-sequence (seq2seq) model. It takes a context where the target word is tagged with two brace tokens, and generates a substitute word or phrase as a prediction. We use beam search to generate multiple likely substitutes. Our underlying seq2seq model is *bart-large* (Lewis et al., 2020). We utilize the same set of hyper-parameters for both pre-training and fine-tuning. Specifically, we train our model for 19,000 steps with a batch size of 64 and a learning rate of 4e-5.

The second model is GeneSis (Lacerra et al., 2021), also a sequence-to-sequence model. Unlike our model, GeneSis filters out words that are not in WordNet, and it incorporates a fallback strategy in the *oot* setting. When the model generates fewer than 10 substitutes, additional words are retrieved from WordNet, and ranked using NLM embeddings. To assess the model's performance based solely on annotated data, we disable both lexicon filtering and fallback strategy. We use their default settings for both pre-training and fine-tuning.

In order to evaluate the contribution of the WNSub dataset, we compare a baseline approach with a WNSub pre-training approach. In the baseline approach, we train the systems on existing datasets, specifically the CoInCo and TWSI datasets, following the methodology of Lacerra et al. (2021). In the pre-training approach (+ WNSub), we first pre-train the systems on WNSub, and then fine-tune on the union of the CoInCo and TWSI datasets. Our evaluation is on the SE07 test set only, as SWORDS includes instances from CoInCo.

The results in Table 2 indicate that pre-training on the WNSub dataset improves the results of both supervised models. The only exception is GeneSis in the *oot* setting, in which there is no penalty for attempting to fill all 10 candidate substitutes, even if some of them are incorrect. However, when evaluated using the standard F1 score that considers all predictions, pre-training does improve GeneSis' performance from 26.8 to 27.7 points. This suggests that the F1 metric may better reflect the quality of the systems when they are not forced to produce a fixed number of substitutes.

Models	best	oot
Yuret (2007)	12.9	46.2
Szarvas et al. (2013a)	15.9	48.8
Zhou et al. (2019)	20.3	55.4
GeneSis (2021)	21.6	52.4
Michalopoulos et al. (2022)	21.1	51.3
Seneviratne et al. (2022)	**23.3**	56.3
WNSub experiments		
seq2seq baseline	9.7	44.0
+ WNSub	**10.7**	**44.8**
GeneSis*	19.2	**34.3**
+ WNSub	**19.6**	34.1
Augmentation experiments		
LexSubGen (2020)	21.7	55.1
+ soft constraint	**21.9**	**57.9**
Wada et al. (2022)	21.8	58.0
+ soft constraint	**22.0**	**58.4**

Table 2: Results on the SE07 test set. *With disabled vocabulary filtering and fallback strategy.

5.4 Augmentation Experiments

We evaluate the effectiveness of our sense-based augmentation method (Section 4) on both SE07 and SWORDS test sets, using two different lexical substitution systems. We retrieve synonyms and glosses for the target word from WordNet 3.0 via NLTK (Bird et al., 2009).

As our base WSD system, we use ConSec[3] (Barba et al., 2021). The model jointly encodes the context containing the target word and all possible sense definitions, and extracts the span of the definition that best fits the target word. ConSec also leverages the senses assigned to nearby words to improve performance. Since the original implementation outputs only predicted senses, we changed the source code to capture the probability scores for all candidate senses.

As our primary base lexical substitution system, we use LexSubGen[4] (Arefyev et al., 2020). Their best-performing model injects the target word information by combining the substitute probability from XLNet (Yang et al., 2019) with the contextual embedding similarity of the substitute to the target word.

To test the generalizability of our approach, we also apply our augmentation method to the model of Wada et al. (2022). Their model is based on the similarity of contextualized and decontextualized

[3] https://github.com/SapienzaNLP/consec
[4] https://github.com/Samsung/LexSubGen

Models	F_a^{10}	F_c^{10}
GPT-3	22.7	36.3
WordTune	22.8	33.6
BERT	19.2	30.3
Yang et al. (2022)	18.3	28.7
LexSubGen (2020)	19.4	29.9
+ soft constraint	**21.5**	**34.8**
Wada et al. (2022)	24.5	39.9
+ soft constraint	**24.7**	**42.5**

Table 3: Results on the SWORDS test set.

embeddings, which represent the average contextual representation of a word in multiple contexts.

The results on SE07 in Table 2 show that our approach leads to improvements over both base models. In the *oot* setting, the result of 57.9 represents a 5% relative gain, while the result of 58.4 is higher than any reported in prior work.

Similarly, the results on SWORDS in Table 3 demonstrate consistent improvements over both base systems in the strict evaluation settings. The results in the last row represent the new state of the art on the SWORDS dataset.

5.5 Ablation and Analysis

Table 4 presents the results of an ablation study on the SWORDS dataset, which we conducted to assess the impact of various components of our augmentation method. Removal of both synonyms and glosses simultaneously is equivalent to the LexSubGen baseline shown in the first row. Our principal model, soft constraint, is in the row 3. The results in rows 2 and 3 show that hard constraint is less effective than soft constraint. This is because the former relies on a single most likely sense, which makes it less robust to WSD errors. The results in rows 4 and 5 indicate that glosses provide more information than synonyms. Overall, the ablation study provides further evidence that augmentation improves lexical substitution systems.

We also performed a manual error analysis on a randomly selected sample of 20 instances from SWORDS. We did not find any instances where the augmentation results in missed substitutes, as compared to the base model. On the other hand, we found one instance where the augmentation helps to identify two gold substitutes, *overlook* and *neglect*, as substitutes for *miss*. We note that these three verbs share a WordNet synset which is glossed as "leave undone or leave out."

Models	F_a^{10}	F_c^{10}
LexSubGen	19.4	29.9
+ hard constraint	21.2	34.2
+ soft constraint	**21.5**	**34.8**
- gloss	20.6	32.7
- synonyms	21.1	33.6

Table 4: Ablation study on the SWORDS test set.

6 Conclusion

We consider the new entailment-based definition and formalization of lexical substitution as the principal contribution of this paper. The new WNSub dataset and the context augmentation method are inspired by our theoretical analysis. The experiments demonstrate that both innovations lead to performance improvements on the standard lexical substitution benchmarks, which we interpret as empirical validation of the theoretical approach. In the future, we plan to explore the generalizability of our approach to other languages, as well as cross-lingual lexical substitution.

7 Limitations

Our augmentation approach is model-agnostic, meaning that it can be applied to any lexical substitution model. However, this also means that it inherits any limitations of the underlying model. For example, in the case of LexSubGen, it can only produce single-token words as substitutes which might prevent it from generating valid longer words or phrases as substitutes that are present in the gold annotations. Additionally, the substitutes are also limited by the vocabulary of the pre-trained language model that LexSubGen uses.

Another limitation of our method is that it relies on the presence of target words in a lexical resource, such as WordNet, together with their synonyms and glosses. If this sense-specific information is missing from the lexical resource, it cannot be used to improve the performance of a lexical substitution system.

Our entailment criterion for lexical substitution is defined for the binary classification task, rather than for generation or ranking tasks. However, if a probabilistic model is used to determine the probability of mutual entailment between sentences, this score can be utilized to rank substitutes if necessary. As explained in Section 3.2, the binary definition can also be adapted to the generation task by iterating over candidate substitutes.

8 Ethics Statement

It is important to acknowledge that our approach utilizes a large language model trained on data from the internet, which may contain inherent biases. Therefore, it is crucial to exercise caution when applying this model in applications such as writing assistance, where it may have a direct impact on individuals or groups.

We also have considered ethical considerations in the construction and use of our evaluation dataset. The dataset we used was automatically constructed from publicly available datasets and lexical resources. To the best of our knowledge, the original datasets do not contain offensive content. The names included in the datasets are from texts that are already publicly available. We did not use the help of third-party annotators to produce any additional data. The datasets we used did not include any license agreements or terms of use. The only requirement was to cite the dataset papers, which we have done in Section 3.5. Additionally, we intend to release our dataset publicly to encourage further research and development in the field of lexical substitution.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

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. 2020. 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.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris Biemann. 2012. Turk bootstrap word sense inventory 2.0: A large-scale resource for lexical substitution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4038–4042, Istanbul, Turkey. European Language Resources Association (ELRA).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Lexical substitution dataset for German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1406–1411, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

George Dahl, Anne-Marie Frassica, and Richard Wicentowski. 2007. SW-AG: Local context matching for English lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 304–307, Prague, Czech Republic. Association for Computational Linguistics.

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.

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.

Maayan Geffet and Ido Dagan. 2004. Feature vector quality and distributional similarity. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 247–253, Geneva, Switzerland. COLING.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114, Ann Arbor, Michigan. 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.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. 2007. UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413, Prague, Czech Republic. Association for Computational Linguistics.

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, and Grzegorz Kondrak. 2022. UAlberta at SemEval 2022 task 2: Leveraging glosses and translations for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 145–150, Seattle, United States. Association for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2020. Synonymy = translational equivalence. *arXiv preprint arXiv:2004.13886*.

Tobias Hawker. 2007. USYD: WSD and lexical substitution using the Web1T corpus. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 446–453, 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.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.

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.

Paul Kroeger. 2018. *Analyzing meaning: An introduction to semantics and pragmatics*. Language Science Press.

Caterina Lacerra, Rocco Tripodi, and Roberto Navigli. 2021. GeneSis: A Generative Approach to Substitutes in Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10810–10823, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mina Lee, Chris Donahue, Robin Jia, Alexander Iyabor, and Percy Liang. 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*, pages 4362–4379, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, Online. Association for Computational Linguistics.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University.

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. 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 089–115. 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, Prague, Czech Republic. 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. 2015. 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.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

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

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Tristan Miller, Mohamed Khemakhem, Richard Eckart de Castilho, and Iryna Gurevych. 2016. Sense-annotating a lexical substitution data set with ubyline. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 828–835, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

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.

Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8766–8774.

Sandaru Seneviratne, Elena Daskalaki, Artem Lenskiy, and Hanna Suominen. 2022. CILex: An investigation of context information for lexical substitution methods. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4124–4135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of the International Conference RANLP-2009*, pages 404–410, Borovets, Bulgaria. Association for Computational Linguistics.

Ravi Sinha and Rada Mihalcea. 2014. Explorations in lexical sample and all-words lexical substitution. *Natural Language Engineering*, 20(1):99–129.

György Szarvas, Chris Biemann, and Iryna Gurevych. 2013a. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1131–1141, Atlanta, Georgia. Association for Computational Linguistics.

György Szarvas, Róbert Busa-Fekete, and Eyke Hüllermeier. 2013b. 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.

Antonio Toral. 2009. The lexical substitution task at evalita 2009. In *Proceedings of EVALITA Workshop, 11th Congress of Italian Association for Artificial Intelligence, Reggio Emilia, Italy*.

Takashi Wada, Timothy Baldwin, Yuji Matsumoto, and Jey Han Lau. 2022. Unsupervised lexical substitution with decontextualised embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4172–4185, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing text provenance via context-aware lexical substitution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11613–11621.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Deniz Yuret. 2007. KU: Word sense disambiguation by substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214, Prague, Czech Republic. Association for Computational Linguistics.

Shiqi Zhao, Lin Zhao, Yu Zhang, Ting Liu, and Sheng Li. 2007. HIT: Web based scoring method for English lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 173–176, Prague, Czech Republic. Association for Computational Linguistics.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461.

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 Manual Dataset Analysis

In this section, we describe our manual analysis procedure. It consists of the following steps.

1. We randomly select 50 gold substitutes along with their corresponding contexts and target words.

2. For each sampled gold substitute, we generate a new sentence by replacing the original target with the gold substitute.

3. For each generated sentence pair, we check the following criteria:

   (a) Whether the original sentence entails the new sentence.

   (b) Whether the new sentence entails the original sentence.

   (c) Whether the new sentence fully preserves the meaning of the original sentence.

To identify textual entailment, we follow the definition outlined in Section 3.1. We verify the meaning preservation criterion by assessing whether the target word and its substitute candidate represent the same concept within the given context.

This analysis, which is summarized in Section 3.4, allows us to compare our definition, which is based on mutual entailment, with the existing definition of meaning preservation. The results of our analysis are presented in Table 5.

Context C_t	Substitute w	$Cw \models Ct$	$Ct \models Cw$	Meaning preserved
I am glad to be out of the favor-trading scene for half a **minute**	moment	No	No	No
It didn't seem like we had a lot of holes to fill. It's good, it **gives** us something we didn't have and we didn't lose much.	award	No	Yes	No
Walking out of the church , a little gust of **cold** air caught me by surprise.	icy	No	Yes	No
"It's a long way to anywhere worth going," he **said**.	declare	No	Yes	No
Taste , hearing and touch became a single blur , and I do not know if my eyes were **open**.	uncovered	No	No	No
My favorite thing about her is her **straightforward** honesty and that her favorite food is butter.	uncomplicated	No	No	No
I had almost forgotten the body lying with **broken** neck on the cathedral's hard tiles	damage	Yes	No	No
A black hallway opened into a space like a **cathedral**. The vault rose into obscurity above me, and a massive window stood ahead of me.	large church	Yes	No	No
A black hallway opened into a space like a **cathedral**. The vault rose into obscurity above me, and a massive window stood ahead of me.	house of god	Yes	No	No
"Excuse me," I **said**, ignoring Nepthys' warning look,	mention	Yes	Yes	No
Please, **walk** this way.	proceed	Yes	Yes	No
They were for (an **expected** invasion of Japan)	planned	Yes*	Yes	No
Energy Secretary Bill Richardson went to Baghdad in 1995 while a **representative** for New Mexico.	elected official	Yes	Yes*	No
Then I felt a tug on the back of my shirt and **noticed** that Amy was following me.	see	Yes	Yes	No
This story might be interesting. Does it have anything to do with why your **head** is shaved?	scalp	Yes	Yes	No
I **swear**. They all thought I was Steve Martin .	vow	Yes	Yes	Yes
...many clinical psychologists already receive **inadequate** training	insufficient	Yes	Yes	Yes
Now, will you tell me how you **know** my family?	have knowledge of	Yes	Yes	Yes
It's **okay**, you can trust him.	alright	Yes	Yes	Yes
...you **know** some way to locate the undead, don't you ?	have	Yes	Yes	Yes
But in some **areas**, the seabass are being overfished.	location	Yes	Yes	Yes
The Persian Gulf War **destroyed** much of the country's medical infrastructure	devastate	Yes	Yes	Yes
That was **very** kind of her.	exceedingly	Yes	Yes	Yes
...considers prescriptive authority a **logical** extension of psychologists' role as health-care providers	rational	Yes	Yes	Yes
...we simply want to **discover** whether this individual is in fact, a vampire.	find	Yes	Yes	Yes
But they **liked** the way (Jose) has played and they're giving him a chance.	enjoy	Yes	Yes	Yes
Karnes had his own Jeep, and **went** to the beach	head	Yes	Yes	Yes
Ochoa has played in the majors for five different teams **starting** in 1995	commence	Yes	Yes	Yes
The new plant is part of IBM 's push to gain a **strong** lead in chip-making.	formidable	Yes	Yes	Yes
He ran down a hallway and slipped behind one of the **doors**	doorway	Yes	Yes	Yes
"What would convince you to part with it?" She **considered** this , looking him over.	think over	Yes	Yes	Yes
One expert, whose job is so politically sensitive that he spoke on condition that he wouldn't be **named** or quoted, said ...	cite	Yes	Yes	Yes
We've had genies , indentured sorcerers , even golems and the **occasional** elf.	intermittent	Yes	Yes	Yes
RxP opponents charge the APA with pushing its prescription-privileges agenda without **adequately** assessing support for it in the field.	sufficiently	Yes	Yes	Yes
Comey **said** Tokhtakhounov had three residences in Italy	state	Yes	Yes	Yes
It pulled back around his fingertips, which bore **things** that might have been nails or claws.	object	Yes	Yes	Yes
Hall is to **return** to Washington on April 22	arrive back	Yes	Yes	Yes
Moreover , he said , technology now exists for stealing corporate secrets.	in addition	Yes	Yes	Yes

35 **thin** fingers waved lazily like seaweed.	narrow	Yes	Yes	Yes
The door **took** us to the bottom of a flight of wooden stairs.	bring	Yes	Yes	Yes
It's exhausting to talk to those **people**.	folk	Yes	Yes	Yes
I bet my friend can tell you everything you **need** to know.	feel the necessity for	Yes	Yes	Yes
That's a question you learn not to ask **here**.	in this place	Yes	Yes	Yes
If he got your girl, she's **probably** dead!	most likely	Yes	Yes	Yes
Rep. Tony Hall, D-Ohio, urges the UN to allow a freer **flow** of food and medicine into Iraq.	transmission	Yes	Yes	Yes
"I have made it a policy of mine never to serve seabass," said Hahn. "I refuse to **sell** it."	market	Yes	Yes	Yes
"You **idiots**! You woke it up?"	blockhead	Yes	Yes	Yes
She will have reunions of sorts with her famous kitchen in the **next** few weeks.	forthcoming	Yes	Yes	Yes
Still unresolved is Sony's effort to **hire** producers J. Peters and P. Guber to run the studio.	give job to	Yes	Yes	Yes
Ochoa will **join** the club today in Anaheim before tonight's game against the Yankees.	enter	Yes	Yes	Yes

Table 5: The table contains a random sample of 50 substitutes from the SWORDS dataset. The target words are in bold. * denotes that the specified entailment holds if we assume relevant background knowledge.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
We used ChatGPT for the assistance purely with the language of the paper (paraphrasing and polishing our original ideas). We thoroughly checked that the generated output does not contain any new ideas.

B ☑ Did you use or create scientific artifacts?

Section 3.5 describes the dataset we constructed. Section 5 describes the datasets and models we used.

☑ B1. Did you cite the creators of artifacts you used?
Sections 3.5 and 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 8

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 8

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 8

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.1

C ☑ Did you run computational experiments?

Section 5

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 5.3 and 5.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 5.3 and 5.4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5.3 and 5.4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.