

SWORDS ✂: A Benchmark for Lexical Substitution with Improved Data Coverage and Quality

Mina Lee* Chris Donahue* Robin Jia Alexander Iyabor Percy Liang

Stanford University

{minalee, cdonahue, robinjia, aiyabor, pliang}@cs.stanford.edu

Abstract

We release a new benchmark for lexical substitution, the task of finding appropriate substitutes for a target word in a context. For writing, lexical substitution systems can assist humans by suggesting words that humans cannot easily think of. However, existing benchmarks depend on human recall as the only source of data, and therefore lack coverage of the substitutes that would be most helpful to humans. Furthermore, annotators often provide substitutes of low quality, which are not actually appropriate in the given context. We collect higher-coverage and higher-quality data by framing lexical substitution as a classification problem, guided by the intuition that it is easier for humans to judge the appropriateness of candidate substitutes than conjure them from memory. To this end, we use a context-free thesaurus to produce candidates and rely on human judgement to determine contextual appropriateness. Compared to the previous largest benchmark, our SWORDS benchmark has 3x as many substitutes per target word for the same level of quality, and its substitutes are 1.4x more appropriate (based on human judgement) for the same number of substitutes.

1 Introduction

Imagine you are writing the message “I read an amazing paper today” to a colleague, but you want to choose a more descriptive adjective to replace “amazing.” At first you might think of substitutes like “awesome” and “great,” but feel that these are also unsatisfactory. You turn to a thesaurus for inspiration, but among reasonable alternatives like “incredible” and “fascinating” are words like “prodigious” which do not quite fit in your context. Ultimately, you choose to go with “fascinating,” but reaching this decision required a non-trivial amount of time and effort.

*Equal contribution.

Research on *lexical substitution* (McCarthy, 2002; McCarthy and Navigli, 2007; Erk and Padó, 2008; Szarvas et al., 2013; Kremer et al., 2014; Melamud et al., 2015; Hintz and Biemann, 2016; Zhou et al., 2019; Arefyev et al., 2020) considers the task of replacing a target word in context with appropriate substitutes. There are two widely-used English benchmarks for this task: SEMEVAL (McCarthy and Navigli, 2007) and COINCO (Kremer et al., 2014). For both benchmarks, data was collected by asking human annotators to think of substitutes from memory. Because lexical substitution was originally proposed as a means for evaluating word sense disambiguation systems (McCarthy, 2002), this data collection strategy was designed to avoid a bias towards any particular word sense inventory.

In this work, we consider a different use case for lexical substitution: writing assistance. For this use case, we are interested in evaluating a system’s ability to produce appropriate substitutes that are likely to be difficult for humans to think of. We show that the data collection strategy used in past benchmarks yields low *coverage* of such uncommon substitutes—for our previous example, they might contain words like “awesome” and “great,” but miss words like “incredible” and “fascinating.” Furthermore, we observe that these benchmarks have low *quality*, containing words like “fun,” which are easy to think of, but not quite appropriate in context.

We present SWORDS—the Stanford Word Substitution Benchmark—an English lexical substitution benchmark that raises the bar for both coverage and quality (Table 1). We collect SWORDS by asking human annotators to judge whether a given candidate word is an appropriate substitute for a target word in context, following the intuition that judging a given substitute is easier than producing that same substitute from memory. To bootstrap a set of candidates for humans to annotate, we

Context	My favorite thing about her is her <u>straightforward</u> honesty.
COINCO	candid (3), artless (1), blunt (1), complete (1), direct (1), forthright (1), frank (1), outspoken (1), plainspoken (1), truthful (1), unreserved (1)
Thesaurus	honest, simple, easy, candid, forthright, genuine, sincere, truthful, unequivocal, unambiguous, clear-cut, clear, uncomplicated, aboveboard, barefaced, direct, frank, guileless, honorable, just, laid on the line, ... 33 more
SWORDS	sincere (80%), genuine (80%), frank (70%), candid (70%), direct (70%), forthright (70%), uncomplicated (60%), up front (60%), clear-cut (60%), clear (60%), plainspoken (60%), complete (50%), straight-arrow (50%), honest (50%), open (50%), blunt (50%), outspoken (50%), truthful (50%), plain-dealing (40%), undisguised (40%), unvarnished (40%), unreserved (40%), barefaced (40%), unequivocal (30%), upright (30%), simple (30%), veracious (30%), unconcealed (30%), like it is (30%), square-shooting (20%), upstanding (20%), undissembled (20%), manifest (20%), unambiguous (20%), pretenseless (20%), level (10%), laid on the line (10%), honorable (10%), guileless (10%), ... 20 more with 0%

Table 1: We consider lexical substitution, the task of finding appropriate substitutes for a target word in context. In COINCO (the previous largest benchmark), humans are asked to think of substitutes from memory and result in low coverage (the number of annotators who produced each substitute is shown in parentheses; out of six annotators). On the other hand, looking up the target word in a thesaurus has higher coverage, but low quality, because it does not consider the context. In SWORDS, we combine the best of both worlds and provide a list of substitutes that has high coverage and high quality, along with fine-grained scores for each substitute (shown in parentheses). Substitutes with scores greater than 50% from SWORDS are **bolded**.

look up target words in an existing context-free thesaurus.¹ Because a thesaurus might miss substitutes that would not typically be synonymous with the target word outside of the provided context (e.g. “thought-provoking” for “amazing”), we also include human-proposed candidates from the previous COINCO benchmark.

Determining whether a substitute is appropriate is intrinsically subjective. To address this, we collect binary labels from up to ten annotators for each substitute, inducing a score for each substitute. In COINCO, analogous scores are derived from the number of independent annotators who thought of a substitute—hence, as we will show in Section 4, these scores tend to correspond more to ease-of-recollection than appropriateness. In contrast, scores from SWORDS correspond to appropriateness, and also allow us to explicitly trade off coverage and quality, permitting more nuanced evaluation. Our analysis shows that compared to COINCO, SWORDS has 3x more substitutes per target word for the same level of quality, and its substitutes are 1.4x more appropriate based on scores for the same number of substitutes.

We demonstrate that SWORDS is a challenging benchmark by evaluating state-of-the-art lexical substitution systems and large-scale, pre-trained language models including systems based on BERT (Devlin et al., 2019; Zhou et al., 2019) and GPT-3 (Brown et al., 2020). In our evaluation, we find

that humans substantially outperform all existing systems, suggesting that lexical substitution can be used as a downstream language understanding task for pre-trained models. We release SWORDS publicly as a benchmark for lexical substitution, coupled with a Python library that includes previous benchmarks in a common format, standardized evaluation scripts for prescribed metrics, and reproducible re-implementations of several baselines.²

2 Background

We describe lexical substitution and briefly introduce two widely-used benchmarks: SEMEVAL (McCarthy and Navigli, 2007), the first benchmark, and COINCO (Kremer et al., 2014), the largest existing benchmark. For a survey of other benchmarks, we refer readers to Kremer et al. (2014), Hintz and Biemann (2016), and Miller (2016).

Lexical substitution. Lexical substitution is the task of generating a list of substitutes w' that can replace a given target word w in a given context c (McCarthy, 2002):

$$(\text{context } c, \text{target } w) \rightarrow [\text{substitute } w'].$$

The context c is one or more sentences where the target word w is situated. The target word w is one word in the context, which is either manually chosen by humans (McCarthy and Navigli, 2007) or

¹Note that our use of a thesaurus makes SWORDS less appropriate for the original use case for lexical substitution: evaluating word sense disambiguation systems.

²SWORDS: github.com/p-lambda/swords
All experiments reproducible on the CodaLab platform:
worksheets.codalab.org/worksheets/0xc924392d555f4b4fbee47be92e3daa0b

automatically selected based on the part-of-speech of the target word (Kremer et al., 2014). The substitute w' can be a word or phrase. Note that the task of lexical substitution does not consider inflection and does not involve grammar correction; all benchmarks contain lemmas as substitutes (e.g. “run” instead of “ran”).

SEMEVAL. The first lexical substitution benchmark, SEMEVAL-2007 Task 10 (McCarthy and Navigli, 2007), contains 201 manually chosen target words. For each target word, 10 sentences were chosen as contexts (mostly at random, but in part by hand) from the English Internet Corpus (Sharoff, 2006) and presented to five human annotators. The five annotators were instructed to produce up to three substitutes from memory as a replacement for the target word in context that “preserves the meaning of the original word.” This resulted in 12,300 labels in total with four substitutes per target word on average.

COINCO. The previous largest lexical substitution benchmark, COINCO (Kremer et al., 2014), was constructed by first choosing 2474 contexts from the Manually Annotated Sub-Corpus (Ide et al., 2008, 2010). Then, all content words (nouns, verbs, adjective, and adverbs) in the sentences were selected to be target words in order to reflect a realistic frequency distribution of target words and their senses. Each target word was presented to six human annotators, who were asked to provide up to five substitutions or mark it as unsubstutable. All the annotators were instructed to provide (preferably single-word) substitutes for the target that “would not change the meaning.” This resulted in 167,446 labels in total and 7.2 substitutions per target word on average.³ For the rest of the paper, we focus on COINCO (but not SEMEVAL) as our benchmark is built on COINCO and it is the largest existing benchmark.

3 Our benchmark

SWORDS is composed of context, target word, and substitute triples (c, w, w') , each of which has a score that indicates the appropriateness of the substitute. We consider a substitute to be *acceptable* if its score is greater than 50% (e.g. bolded words in Table 1) and *unacceptable* if the score is less than

³The reported number in Kremer et al. (2014) is 167,336 and 10.71, respectively. The latter differs as they counted the same substitute multiple times when suggested by multiple humans, whereas we report the number of unique substitutes.

or equal to 50%. Similarly, a substitute with a score greater than 0% is considered *conceivable*, and otherwise *inconceivable*. Note that these terms are operational definitions for convenience, and different thresholds can be chosen for desired applications.

3.1 Addressing limitations of past work

Improving quality. In prior work, annotators were prompted to consider whether a substitute “preserves the meaning” (McCarthy and Navigli, 2007) or “would not change the meaning” (Kremer et al., 2014) of the target word. Instead, we ask annotators whether they “would actually consider using this substitute as the author of the original sentence.” We believe this wording encourages a higher standard. In Section 4.1, we provide evidence that substitutes from SWORDS have higher quality than those from COINCO on average.

Improving coverage. For prior benchmarks, annotators were asked to generate a list of substitutes from memory. Psycholinguistic studies have shown that when humans are asked to predict the next word of a sentence, they deviate systematically from the true corpus probabilities (Smith and Levy, 2011; Eisape et al., 2020). Thus, we may reasonably expect that asking humans to generate substitutes would similarly lead to systematic omissions of some appropriate substitutes.

We observe that prior benchmarks exclude many appropriate substitutes that are difficult for humans to think of (Section 4.2). To address this limitation, we first obtain a set of candidate substitutes and then ask annotators to *judge* whether they would consider using a given candidate to replace the target word in the context. That is, given a context c , target word w , and candidate substitute w' , we ask humans to judge whether w' is a good replacement for the target word:

$$(\text{context } c, \text{target } w, \text{substitute } w') \rightarrow \{0, 1\},$$

where a *positive* label 1 corresponds to “I would actually consider using this substitute as the author of the original sentence,” and a *negative* label 0 as the opposite. As described in Section 3.2, we annotate a large pool of candidate substitutes to ensure high coverage of all possible substitutes. We confirm that this increases coverage compared to COINCO in Section 4.2.

Redefining scores to reflect appropriateness. In past work, each substitute w' has an associated

Benchmark	Contexts	Targets	Substitutes	Labels	Substitutes per target		
					Inconceivable (score = 0%)	Conceivable (score > 0%)	Acceptable (score > 50%)
SEMEVAL	2010	201	8025	12,300	-	4.0	N/A
COINCO	2474	15,629	112,742	167,446	-	7.2	2.0*
SWORDS	1250	1250	71,813	395,175	37.4	20.1	3.9
COINCO (dev)	1577	10,179	67,814	98,950	-	6.7	2.0*
SWORDS (dev)	417	417	24,095	127,735	38.9	18.9	3.3
COINCO (test)	897	5450	44,928	68,496	-	8.2	2.0*
SWORDS (test)	833	833	47,718	267,440	36.6	20.7	3.3

Table 2: Benchmark statistics of SEMEVAL, COINCO, and SWORDS. Our benchmark contains 2x more acceptable substitutes (3.9) and 3x more conceivable substitutes (20.1) on average, compared to the previous largest benchmark COINCO. SWORDS has additional inconceivable substitutes which received a score of 0 from appropriateness judgement, which are useful for evaluation. Substitutes in the past benchmarks are listed as conceivable substitutes, and the number of acceptable substitutes for COINCO (numbers with *) is estimated based on the COINCO’s substitutes scored under SWORDS.

score defined as the number of annotators who produced w' given the associated context c and target word w . Instead, we define the score as the fraction of annotators who judged the w' to be an appropriate replacement of w . We argue that the previous definition of score reflects ease-of-recollection, but not necessarily appropriateness. In Section 4.3, we show that our definition of score better represents the appropriateness of each substitute.

3.2 Data collection

We collect substitutes and scores for a context and target word pair (c, w) via the following three steps.

Step 1: Select contexts, targets, and substitutes.

We use the subset of contexts and target words from COINCO. Concretely, we start with the (c, w) pairs in COINCO and randomly select one w per c to annotate. Here, the context c consists of three sentences, where the middle sentence has the target word w . Next, we choose a set of candidate substitutes w' to annotate for each (c, w) pair, as framing annotation as binary classification requires determining the set of candidate substitutes a priori. We use human-generated substitutes from COINCO, then add substitutes suggested by a thesaurus (see Appendix A.2 for details). In principle, candidate substitutes can be retrieved from any lexical resource or even sampled from a generative model, which we leave as future work. By combining candidates from COINCO and the thesaurus, we increase the coverage of acceptable substitutes.

Step 2: Reduce the pool of substitutes. Given a list of candidate substitutes from the previous step, we collect three binary labels on each (c, w, w')

triple (see Section 3.3 for details). Then, we pass any substitute with at least one positive label to Step 3 and further collect fine-grained scores. We show that the probability that an acceptable substitute gets incorrectly filtered out as an inconceivable substitute (three negative labels) is very low (0.8%) in Section 4.4.

Step 3: Collect fine-grained scores. In the final step, we collect seven more binary labels on the substitutes which received at least one positive label from Step 2. This yields a total of 10 binary labels for the substitutes.

3.3 Crowdsourcing

We used Amazon Mechanical Turk (AMT) to crowdsource labels on substitutes. Each Human Intelligence Task (HIT) contained a target word highlighted in the context and at most 10 candidate substitutes for the target word. Each candidate substitute had three radio buttons for positive, negative, and abstain. Annotators were asked to choose positive if they would actually consider using the substitute to replace the target word as the author of the context, negative if they would not consider using the substitute, and abstain if they do not know the meaning of the substitute. We treated all abstain labels (1.24% of total labels) as negative labels, thereby making it binary. The benchmark includes abstain labels to maintain the option for them to be handled separately (e.g. excluded) in the future. The interface, instructions, qualification conditions, and filtering criteria used for crowdsourcing can be found in Appendix B.

4 Data analysis

Table 2 shows overall statistics of our benchmark. SWORDS comprises a total of 1250 context and target word pairs (494 nouns, 448 verbs, 189 adjectives, 119 adverbs) and 71,813 total substitutes that have been labeled (including both acceptable and unacceptable substitutes). For brevity, we defer an analysis of annotator agreement to Appendix C.1.

4.1 High quality

With our notion of acceptability, we first observe that 75.4% of the substitutes from COINCO⁴ are considered unacceptable, and 28.6% of the substitutes are even inconceivable (receiving scores less than 50% and 0% from our human annotators). Table 3 shows examples of substitutes that received relatively high scores under COINCO, yet were considered unacceptable under SWORDS. With the same size as COINCO (by taking the subset of our benchmark with the highest scoring substitutes per target), the average score of the substitutes is 4.9 for SWORDS and 3.4 for COINCO, resulting in 1.4x higher quality. Furthermore, SWORDS minimizes the potential noise by having fine-grained scores to account for appropriateness (Section 4.3) as well as explicit inconceivable substitutes, which is useful for evaluation (Section 5.2).

4.2 High coverage

We show that SWORDS achieves high coverage. Among the conceivable substitutes in SWORDS, 14.4% are only in COINCO (COINCO-only), 14.6% are common to both COINCO and the thesaurus ($\text{COINCO} \cap \text{Thesaurus}$), and 71.1% are only from thesaurus (Thesaurus-only). Among the acceptable substitutes, 24% are from COINCO-only, 37.1% are from $\text{COINCO} \cap \text{Thesaurus}$, and 38.9% are from Thesaurus-only. This suggests that a substantial number of substitutes are not present in COINCO. Overall, SWORDS contains 3.9 acceptable and 20.1 conceivable substitutes per target word on average, increasing those numbers by nearly 2x and 3x over COINCO, respectively.

In addition, we find that substitutes from COINCO-only are more likely to be common words whereas substitutes from Thesaurus-only are more likely to be rare words. We compute the Zipf frequency (Speer et al., 2018) of each substitute based on the Google n -gram corpus (Brants and Franz,

⁴For this analysis, we consider COINCO’s substitutes that are used and labeled under SWORDS.

2006) and threshold conceivable substitutes into three groups: uncommon (≤ 3.5), neutral, common (> 4.5). We observe that substitutes from COINCO-only are more likely to be common words (52.7%) than those from Thesaurus-only (38%). On the other hand, the substitutes from Thesaurus-only tend to be more uncommon words (29%) than those from COINCO-only (17.5%).

4.3 Reflection of appropriateness in scores

We show that scores in SWORDS better reflect the appropriateness of each substitute compared to COINCO both quantitatively and qualitatively. We find that if a substitute has a high score under COINCO (score > 1), it is likely to be acceptable under SWORDS (score $> 50\%$) almost all the time (99.6%). However, the converse does not hold: the acceptable substitutes under SWORDS have low scores (score ≤ 1) under COINCO half of the time (49.4%). Intuitively, this is because COINCO’s scores reflect the ease of producing the substitute from memory, whereas SWORDS’s scores reflect the appropriateness of the substitute. Table 3 shows examples of context, target word, and substitute triples which received a low score from COINCO but a high score from SWORDS.

4.4 Validation with additional data

We show that the probability of an acceptable substitute falsely filtered out in Step 2 is very low. To this end, we collected 10 additional labels on 100 context-target word pairs randomly selected from the test set, without reducing the pool of substitutes as in Step 2. By comparing the first three labels to the entire 10 labels, we find that 35.5% of substitutes without any positive labels in Step 2 could have received one or more positive labels if they were kept in Step 3. However, we find that 99.2% of these substitutes were eventually considered unacceptable (judged by 10 labels), indicating that the probability of an acceptable substitute incorrectly filtered out in Step 2 is very low (0.8%).

4.5 Score distribution

Figure 1 shows the score distribution of substitutes in SWORDS along with the source of substitutes: COINCO-only, $\text{COINCO} \cap \text{Thesaurus}$, or Thesaurus-only. Across scores, neither COINCO nor thesaurus completely dominates substitutes, and the overlap between COINCO and thesaurus is

⁵1935 substitutes from COINCO-only, 972 from both, and 43,806 from thesaurus-only received a score of 0%.

Context with <u>target word</u>	Substitute	COINCO’s score (max: 6)	SWORDS’s score (max: 100%)
I don’t wish to be a spokesman for any <u>campaign</u> .	effort	3	0%
She was <u>heading</u> for a drink and slipped <u>out</u> of the crowd.	look	2	10%
“ <u>Name</u> me,” she said.	nickname	2	0%
The e-commerce free <u>zone</u> is situated in north Dubai.	district	1	90%
She will have reunions in the <u>next</u> few weeks.	forthcoming	1	60%
It’s <u>very</u> reassuring that I’ll not <u>only</u> be an outsider but a curiosity.	extraordinarily	0	70%

Table 3: Controversial examples of contexts, target words, and substitutes which have low scores under either COINCO or SWORDS, but high scores under the other benchmark (score > 1 for COINCO and score > 50% for SWORDS). The contexts are simplified for readability.

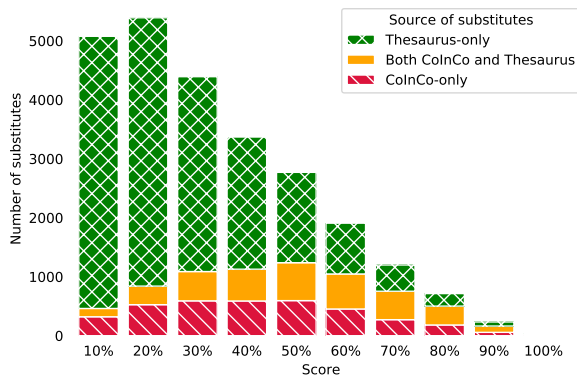


Figure 1: Score distribution of SWORDS’s substitutes with the source of substitutes. We find that neither COINCO nor the thesaurus completely dominates substitutions across scores, indicating the necessity of both human-generated substitutes as well as substitutes from the thesaurus. Substitutes with score 0% are not shown to make the bars visually distinguishable.⁵

quite small, thus indicating the necessity of both human-recalled substitutes as well as substitutes from a thesaurus. We also find that SWORDS adds more substitutes for all the scores, although substitutes from the thesaurus tend to have a lower range of scores compared to those from COINCO. Lastly, we observe that substitutes from COINCO roughly form a normal distribution, which suggests that even the substitutes provided by human annotators are controvertible, and that it is important to account for the intrinsically gradable nature of appropriateness with fine-grained scores.

5 Model evaluation

In this section, we evaluate several methods on SWORDS. The goals of this evaluation are three-fold: (1) to prescribe our recommended evaluation practice for SWORDS, (2) to measure performance of existing large-scale pre-trained models and state-of-the-art lexical substitution systems, and (3) to measure human performance for the purpose of

comparing current and future systems.

5.1 Evaluation settings

There are two primary evaluation settings in lexical substitution research: the *generative* setting (McCarthy and Navigli, 2007) and the *ranking* setting (Thater et al., 2010). In the generative setting, systems output a ranked list of candidate substitutes. There are no restrictions on the number of candidates that a system may output. In the ranking setting, systems are given all candidate substitutes from the benchmark (including those marked as unacceptable) and tasked with ranking them by appropriateness. Here we primarily focus on the generative setting, as it is more relevant to writing assistance. We defer our experiments on the ranking setting to Appendix D.

5.2 Evaluation metrics

In a writing assistance context, we envision that lexical substitution systems would be used to suggest a limited number of substitutes to users (e.g. 10 substitutes as opposed to 100). Hence, we consider evaluation metrics that examine the quality and coverage of the top-ranked substitutes from a system with respect to the substitutes that humans judged as *acceptable* (score > 50%). Specifically, we compute precision (P^k) and recall (R^k) at k ⁶:

$$P^k = \frac{\# \text{ acceptable substitutes in system top-}k}{\# \text{ substitutes in system top-}k}$$

$$R^k = \frac{\# \text{ acceptable substitutes in system top-}k}{\min(k, \# \text{ acceptable substitutes})}$$

Because we care about both quality (precision) and coverage (recall) when comparing systems, we report F^k , the harmonic mean of P^k and R^k . Likewise, we evaluate against the list of substitutes

⁶Note that our definition of recall at k is non-standard; the min compensates for the fact that there are often fewer than k acceptable substitutes.

which humans judged as *conceivable* (score $> 0\%$). P_c^k and R_c^k constitute precision and recall of systems against this larger candidate list, and F_c^k their harmonic mean. Motivated by past work (McCarthy and Navigli, 2007), we primarily examine performance for $k = 10$ and lemmatize system and reference substitutes during comparison.

We note that these metrics represent a departure from standard lexical substitution methodology, established by McCarthy and Navigli (2007). Like P^k and R^k , the previously-used BEST and OOT metrics are also measures of precision and recall, but do not take advantage of the negative labels from our binary data collection protocol as no such labels existed in the earlier benchmarks. Nevertheless, we report performance of all systems on these metrics in Appendix E as reference.

5.3 Baselines

We evaluate both state-of-the-art lexical substitution systems and large-scale pre-trained models as baselines on SWORDS. We reimplement the BERT-based lexical substitution system (BERT-LS) from Zhou et al. (2019), which achieves state-of-the-art results on past benchmarks. As another lexical substitution system, we examine WORDTUNE (AI21, 2020), a commercial system which offers lexical substitution capabilities.⁷

We also examine two large-scale pre-trained models adapted to the task of lexical substitution: BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020). To generate and rank candidates with BERT, we feed in the context with target word either masked (BERT-M) or kept intact (BERT-K), and output the top 50 most likely words according to the masked language modeling head. Because the target word is removed, BERT-M is expected to perform poorly—its main purpose is to assess the relative importance of the presence of the target word compared to the context. Note that both of these strategies for using BERT to generate candidates differ from that of BERT-LS, which applies dropout to the target word embedding to partially obscure it. To generate candidates with GPT-3, we formulate lexical substitution as natural language generation (see Appendix D.5 for details).

5.4 Human and oracle systems

Here we consider human and oracle “systems” to help contextualize the performance of automatic

lexical substitution systems evaluated on SWORDS. We evaluate the performance of HUMANS using labels from a separate pool of annotators as described in Section 4.4. Because this task is inherently subjective, this system represents the agreement of two independent sets of humans on this task, which should be thought of as the realistic upper bound for all metrics. We consider the substitutes that have score $> 0\%$ from the separate pool of annotators as HUMANS’s substitutes in the generative setting.

We also consider both of the candidate sources, COINCO and THESAURUS, as oracle systems. Each source contains a list of substitutes for every target word, and therefore can be viewed as a lexical substitution system and evaluated on SWORDS. COINCO provides substitutes for a target word that were provided by (six) human annotators. This can be thought of as a proxy for how humans perform on lexical substitution when recalling words off the top of their heads (as opposed to making binary judgements as in HUMANS). THESAURUS provides context-free substitutes for a target word (regardless of their word senses) with the default ranking retrieved from the thesaurus. This represents the context-insensitive ordering that a user of the same thesaurus would encounter.

Because these oracle systems only produce candidates which are guaranteed to be in SWORDS, they have an inherent advantage on the evaluation metrics over other systems. Hence, to be more equitable to other systems, we additionally compute F^{10} and F_c^{10} in a “lenient” fashion—filtering out model generated substitutes which are not in SWORDS (we refer to the setup without filtering as “strict”). It is our intention that future systems should *not* use COINCO or THESAURUS in any way, as they leak information about the SWORDS benchmark.

5.5 Evaluation results

Table 4 shows that the performance of all methods falls short of that of humans on all metrics. We interpret this as evidence that SWORDS is a challenging benchmark, since strong (albeit unsupervised) baselines like BERT and GPT-3 do not reach parity with humans. We also observe that two models (WORDTUNE and GPT-3) achieve higher F^{10} than COINCO. In other words, while all models perform worse than humans who are judging the appropriateness of substitutes (HUMANS), some models appear to slightly outperform humans who

⁷WORDTUNE is not optimized for lexical substitution.

Model	Lenient		Strict	
	F^{10}	F_c^{10}	F^{10}	F_c^{10}
HUMANS*	51.6	76.4	—	—
COINCO	34.6	63.3	—	—
THESAURUS [†]	24.6	59.3	—	—
THESAURUS	17.6	50.2	—	—
WORDTUNE [†]	35.2	43.7	23.4	33.2
GPT-3	35.0	47.0	22.2	34.3
WORDTUNE	34.9	43.5	22.8	32.1
GPT-3 [†]	34.7	47.2	21.8	32.8
BERT-K [†]	32.2	53.8	18.5	29.1
BERT-LS	31.6	53.3	16.8	26.1
BERT-K	31.5	53.2	15.2	23.7
BERT-M [†]	30.8	47.2	15.7	24.4
BERT-M	30.8	47.0	10.4	16.1

Table 4: Evaluation of models on SWORDS in the generative setting. Here, systems must both generate and rank a set of substitutes. We observe that the performance of all baselines on all metrics falls short of human performance. In the “lenient” columns, we filter out system candidates which do not appear in the reference list. *Computed on a subset of the test data. [†]Reranked by our best ranking model (BERT).

are thinking of substitutes off the top of their head (COINCO). This implies that some lexical substitution models may already be helpful to humans for writing assistance, with room for improvement.

Overall, we find that there is no single system which emerges as the best on all metrics. We note that, despite BERT-LS representing the state-of-the-art for past lexical substitution benchmarks, its performance falls short of that of commercial systems like GPT-3 and WORDTUNE on most criteria. Also, the BERT-based methods output around 5x as many candidates as the other models on average, thus having an inherent advantage in recall with the lenient criteria (see Table 7 in Appendix E).

In Table 4, we additionally report the performance of generative models by re-ranking their lists of substitutes using the best ranker from our candidate ranking evaluation, BERT (see Appendix D for details). This procedure unilaterally improves performance for all systems on all metrics except for GPT-3.⁸ Hence, we speculate that improved performance on the ranking setting will be complementary to improved performance on the generative setting.

From a qualitative perspective, many of the systems we evaluate already produce helpful substi-

⁸We speculate that this is because GPT-3 produces many substitutes containing multiple word pieces, and mean-pooling several word pieces may result in lower-quality scores.

tutes (Table 5). In examining errors, we find that BERT-based models and WORDTUNE tend to produce words that differ semantically from the target (e.g. “league” for zone). Substitutes generated by GPT-3 are often repetitive (e.g. for zone, GPT-3 produced 64 substitutes, out of which only 13 were unique)—we filter out duplicates before evaluating. Finally, we observe that some systems produce appropriate substitutes which are not present in SWORDS (e.g. GPT-3 produces “precinct” for zone), indicating that SWORDS still has gaps in coverage. However, the higher coverage and quality in SWORDS compared to past benchmarks still improves the reliability of our proposed evaluation.

6 Related work

As we already discussed previous lexical substitution benchmarks in Section 2 and models in Section 5, we use this section to draw connections to other related literature.

Word sense disambiguation. The task of word sense disambiguation consists of selecting the intended meaning (i.e. sense) from the pre-defined set of senses for that word in a sense inventory. The task of lexical substitution is closely related to word sense disambiguation, as many words are *sense synonyms*—some of their senses are synonymous, but others are not (Murphy, 2010). In fact, McCarthy (2002) proposed lexical substitution as an application-oriented word sense disambiguation task that avoids some of the drawbacks of standard word sense disambiguation, such as biases created by the choice of sense inventory (Kilgarriff, 1997).

Near-synonym lexical choice. Words are often *near-synonyms*—they can substitute for each other in some contexts, but not every context (DiMarco et al., 1993; Murphy, 2010). SWORDS can be viewed as a collection of human judgments on when certain near-synonyms are substitutable in a given context. The task of near-synonym lexical choice consists of selecting the original target word from a set of candidate words given a context where the target word is masked out (Edmonds and Hirst, 2002). The candidate words are composed of the target word and its near-synonyms which are often retrieved from a lexical resource such as Hayakawa (1994). In this task, systems are tested whether they can reason about near-synonyms and choose the best substitute that fits in the context, without knowing any direct semantic information about the

Context	The e-commerce free zone is situated in north Dubai, near the industrial free <u>zone</u> in Hebel Ali
Substitutes in SWORDS	sector (90%), district (90%), area (90%), region (70%), section (70%), range (60%), strip (60%), ground (50%), segment (50%), territory (50%), sphere (40%), realm (40%), place (30%), tract (30%), city (30%), belt (20%), circuit (20%), band (0%)
Reference for F^k (7)	sector, district, area, region, section, range, strip
Reference for F_c^k (17)	sector, district, area, region, section, range, strip , ground, segment, territory
COINCO (9)	area, region, district, section , city, place, range, strip , territory
THESAURUS [†] (14)	district, area , belt, territory, region , realm, sector, section , circuit, segment
WORDTUNE [†] (11)	district, area , city, region , site, league, center, system, place, zona
GPT-3 [†] (13)	district, area , territory, region , realm, sector , locality, section , quarter, precinct
BERT-LS (50)	belt, district , port, area , zones, city, park, center, strip, sector
BERT-K [†] (50)	zones, district, area , city, belt, region , park, ville, site, sector
BERT-M [†] (50)	zones, district, area , city, belt, territory, region , haven, park, site

Table 5: Qualitative comparison of top 10 candidates generated by best systems. From top to bottom, table sections show (1) a context, target word, substitutes, and scores from SWORDS (dev), (2) reference lists used to compute F^k and F_c^k (applying thresholds of $> 50\%$ and $> 0\%$ to scores), (3) candidates from oracle systems (data sources of SWORDS), and (4) candidates from best systems. For each system, we include the number of candidates produced by the system in parentheses (after removing duplicates), although we only consider the top 10 candidates for evaluation. Substitutes with scores greater than 50% from are **bolded**.

target word and without having to explicitly judge the appropriateness of other candidates.

Lexical and phrasal resources. Lexical resources such as thesauri are often used to identify possible word substitutes. WordNet (Fellbaum, 1998) is a widely used lexical resource for English that includes synonymy, antonymy, hypernymy, and other relations between words. PPDB (Pavlick et al., 2015) includes both word-level and phrase-level paraphrase rules ranked by paraphrase quality. These resources relate words and phrases in the absence of context, whereas lexical substitution requires suggesting appropriate words in context.

Paraphrase generation. Work on sentence-level paraphrase generation considers a wide range of meaning-preserving sentence transformations, including phrase-level substitutions and large syntactic changes (Madnani and Dorr, 2010; Wieting and Gimpel, 2018; Iyyer et al., 2018; Hu et al., 2019). Our work could be extended to phrases given appropriate methods for identifying target phrases and proposing candidate substitute phrases. One benefit of focusing on word substitutions is that we can cover a large fraction of all appropriate substitutes, and thus estimate recall of generative systems. Some word-level substitutions, such as function word variation and substitutions that rely on external knowledge, are also outside the scope of our work but occur in standard paraphrase datasets (Bhagat and Hovy, 2013).

Self-supervised pre-trained models. The task of suggesting words given surrounding context bears strong resemblance to masked language modeling, which is commonly used for pretraining (Devlin et al., 2019). However, for lexical substitution, appropriate substitutes must not only fit in context but also preserve the meaning of the target word; thus, additional work is required to make BERT perform lexical substitution (Zhou et al., 2019; Arefyev et al., 2020).

Modeling human disagreement. In SWORDS, we find considerable subjectivity between annotators on the appropriateness of substitutes. For the task of natural language inference, recent work argues that inherent disagreement between human annotators captures important uncertainty in human language processing that current NLP systems model poorly (Pavlick and Kwiatkowski, 2019; Nie et al., 2020). We hope that the fine-grained scores in SWORDS encourage the development of systems that more accurately capture the graded nature of lexical substitution.

Acknowledgments

We sincerely thank Frieda Rong, Nelson Liu, Stephen Musmann, Kyle Mahowald, Daniel Jiang, and all reviewers for their help and feedback throughout this project. We also thank OpenAI and Wordtune for allowing us to evaluate their systems. This work was funded by DARPA CwC under ARO prime contract no. W911NF-15-1-0462.

References

- AI21. 2020. Wordtune (accessed 2020 Oct 30). <https://www.wordtune.com/>.
- Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. A comparative study of lexical substitution approaches based on neural language models. *arXiv*.
- Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39.
- Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *arXiv preprint arXiv:2005.14165*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics (ACL)*, pages 4171–4186.
- Chrysanne DiMarco, Graeme Hirst, and Manfred Stede. 1993. The semantic and stylistic differentiation of synonyms and near-synonyms. In *AAAI Spring Symposium on Building Lexicons for Machine Translation*.
- Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*.
- Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation improves psychometric predictive power. In *Computational Natural Language Learning (CoNLL)*.
- Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 897–906.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Samuel Ichiye Hayakawa. 1994. *Choose the Right Word: A Contemporary Guide to Selecting the Precise Word for Every Situation*. Collins Reference.
- Gerold Hintz and Chris Biemann. 2016. Language transfer learning for supervised lexical substitution. In *Association for Computational Linguistics (ACL)*.
- J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
- Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: the manually annotated sub-corpus of American English. In *Language Resources and Evaluation (LREC)*.
- Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Association for Computational Linguistics (ACL)*.
- Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *North American Association for Computational Linguistics (NAACL)*.
- A. Kilgarriff. 1997. I don’t believe in word senses. *Computers and the Humanities*.
- Barbara Ann Kipfer. 2013. *Roget’s 21st Century Thesaurus, Third Edition*. Random House Publishing Group.
- Kazuaki Kishida. 2005. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical report, National Institute of Informatics.
- Gerhard Kremer, Katrin Erk, Sebastian Pado, and Stefan Thater. 2014. What substitutes tell us – analysis of an “all-words lexical substitution corpus. In *Association for Computational Linguistics (ACL)*.
- Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.
- Diana McCarthy. 2002. Lexical substitution as a task for WSD evaluation. In *SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Direction*.
- Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Workshop on Semantic Evaluations*.
- Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015. Modeling word meaning in context with substitute vectors. In *North American Association for Computational Linguistics (NAACL)*.
- Tristan Miller. 2016. *Adjusting Sense Representations for Word Sense Disambiguation and Automatic Pun Interpretation*. Ph.D. thesis, Technische Universität Darmstadt.

- M. Lynne Murphy. 2010. *Lexical Meaning*. Cambridge University Press.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 7.
- Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Association for Computational Linguistics (ACL)*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- S. Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11:435–462.
- Nathaniel Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2. <https://doi.org/10.5281/zenodo.1443582>.
- György Szarvas, R’obert Busa-Fekete, and Eyke Hüllermeier. 2013. Learning to rank lexical substitutions. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Stefan Thater, Hagen Fürstenaу, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Association for Computational Linguistics (ACL)*.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*.
- John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Association for Computational Linguistics (ACL)*.
- Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Association for Computational Linguistics (ACL)*.

A Data collection

A.1 Deduplicating context

SWORDS uses the same contexts as COINCO, but with slight modifications to avoid duplication issues and incomplete contexts found in COINCO. COINCO uses a subset of contexts from the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008, 2010), in which some sentences are erroneously repeated multiple times due to multiple IDs assigned to a single sentence. Consequently, COINCO contains duplicate sentences in some contexts, as shown below:

" –was kindly received," "But an artist who would stay first among his fellows can tell when he begins to fail." "But an artist who would stay first among his fellows can tell when he begins to fail."

Furthermore, we found that some parts of the document context are missing in COINCO because an ID was not assigned to the parts in MASC (e.g. “he said.” is missing from the above passage after the word “received”).

To address this issue, we re-extracted full contexts from MASC. Given a sentence containing a target word in COINCO, we located the sentence in MASC and used three non-overlapping adjacent MASC regions as our context. As a result, our context contains additional text that was erroneously omitted in COINCO (including newlines), thereby reducing annotator confusion. The context of the above example in our benchmark is as follows:

" –was kindly received," **he said.** "But an artist who would stay first among his fellows can tell when he begins to fail." "Oh?"

A.2 Retrieving substitutes from THESAURUS

We use `thesaurus.com`, which is based on Roget’s Thesaurus (Kipfer, 2013), as a primary source of context-free substitutes for target words in SWORDS. This resource contains substitutes for 133K words, with an average of four senses per word (median of one—a fraction of words have dozens of senses) and 25 substitutes per senses.

To select substitutes for a particular target word, we gather all substitutes from all senses that have the same part of speech as the original target, in order to disentangle lexical substitution from the task of word sense disambiguation as well as to include

challenging distractors for evaluating models.⁹ Because lemmas typically contain more substitutes than their associated word forms (e.g. “jump” has more substitutes than “jumping”), we lemmatize target words before querying the thesaurus.

B Crowdsourcing

B.1 Instructions and interface

Figures 2 and 3 show the instructions and interface we used for Amazon Mechanical Turk (AMT) to crowdsource labels on substitutes. Following the practice of COINCO, we showed a highlighted target word in the context, which consisted of three sentences to provide sufficient context. We instructed annotators to provide a negative label if the target word is a proper noun or part of a fixed expression or phrase.

Since our Human Intelligence Task (HIT) concerns acceptability judgement as opposed to substitute generation, we made the following modifications to the COINCO’s setup. First, we asked annotators whether they “would actually consider using this substitute” rather than whether the substitute “would not change the meaning” of the target word (Section 3.1). Second, we allowed annotators to abstain if they do not know the definition of the substitute, while asking them to return the HIT if they do not know the definition of the target word or more than three substitutes. Third, we asked annotators to accept a substitute which is “good but not quite grammatically correct.” Lastly, we asked annotators to accept the substitute identical to the target word, in attempt to filtering out spammed HITs (Section B.3).

B.2 Setting on Amazon Mechanical Turk

Each HIT contained at most 10 candidate substitutes for a context-target word pair. When there were more than 10 candidate substitutes, we generated multiple HITs by partitioning the candidate substitutes into multiple subsets with potentially different length, using `numpy.array_split`. We randomized the ordering of substitutes so that each HIT is likely to contain substitutes from both COINCO and the thesaurus. The following qualification conditions were used to allow experienced annotators to participate in our task:

⁹In COINCO, contexts, target words, and their part-of-speech tags all come from the Manually Annotated Sub-Corpus (Ide et al., 2008, 2010)

Instructions

Note: Even if you have done this task before, please carefully read these updated instructions!

You will be shown a sentence with a **highlighted** word and a list of possible substitutes for the highlighted word. For each substitute, please select one of the following:

- ✓ I would **actually consider using this substitute** as the author of the original sentence.
- ✗ I would **not consider using this substitute** as the author of the original sentence.
- ? I **do not know the definition** of the substitute.

Please use your best judgement when labeling each substitute. We are trying to capture what makes one word a good, natural substitute for another. Detailed criteria are listed below:

- Mark ✓ if the substitute is **good but not quite grammatically correct** (see examples below).
- Ignoring grammar, please keep a high bar for marking a substitute as ✓.
- You may end up marking many substitutes as ✗. This is fine and expected.
- Mark ✓ if the **highlighted** word itself appears as a substitute. **Not doing so will make your HIT rejected.**
- Mark ✗ if the **highlighted** word is a proper noun or part of a fixed expression / phrase (see examples below).
- Return the HIT if you don't know the definition of the **highlighted** word or more than three substitutes.

[Click to show examples](#)

Sentence	Substitute	Label	Reason
The sky was gray .	overcast	✓	Good substitute
The cow jumped over the moon.	leap	✓	Ungrammatical ("leap" vs "lept") but good otherwise
They munched on the snacks.	ate	✓	Ungrammatical ("ate on") but good otherwise
She was fascinated by science .	sports	✗	Different meaning
He said he was hungry.	pronounced	✗	Not natural
We drank coffee .	liquid	✗	Too general
He hit the sack.	punched	✗	Highlighted word is part of an expression
She drank green tea.	emerald	✗	Highlighted word is part of a phrase
The book was excellent .	gazint	?	Unknown definition

Figure 2: Our instructions ask annotators to accept a substitute if they would actually consider using the substitute as the author of the context, reject if not, or abstain if they do not know the definition of the substitutes. Examples are provided for reference, when annotators click the “click to show examples” buttons.

- HIT Approval Rate (%) for all Requesters’ HITs is greater than 98.
- Location is the United States.
- Number of HITs Approved is greater than 10,000.

Our target hourly wage for annotators was \$15. Based on our in-person pilot study with five native English speakers, we approximated the time per assignment (labeling at most twelve substitutes) to be 25 seconds. Then, we assumed that it may take 1-2x longer for crowd workers to complete the assignments and decided on the compensation of \$0.10 to fall into the range of \$7.25 (US federal minimum wage) and \$15 per hour, which corre-

sponds to 50 seconds and 24 seconds per assignment, respectively.

It may be surprising that our assignments only take 25 seconds on average, though there are several reasons why this is the case: (1) In general, making binary judgements about substitute words takes very little time for native speakers. (2) Annotators only have to read the target sentence once to provide judgements for all substitutes in an assignment. (3) Annotators usually do not need to read the two additional context sentences to make judgements. (4) Annotators can almost instantly judge two control substitutes (Section B.3), and are therefore only realistically evaluating at most ten candidates per assignment.

Model	GAP
HUMANS*	67.7
BERT	57.1
BERT-LS	53.8
BERT-LS w/o s_p	53.0
BERT-LS w/o s_v	51.9
GLOVE	49.8
RANDOM	32.6

Table 6: Evaluation of models on SWORDS in the ranking setting. Here, systems provide a score for every candidate in the benchmark. *Computed on a subset of the test data.

derived from the data to that of SWORDS and simulate the effect of having k annotators by sampling k labels per substitute without replacement a total of 100 times.

Figure 4 shows the correlation between SWORDS and k additional human annotators. We observe quite low RBO for $k < 3$ and diminishing returns as k grows. Based on this observation, we argue that there is wide variation in opinions and it is necessary to use sufficiently large k to capture the distribution.

D Model evaluation

D.1 Ranking setting

As opposed to the generative setting where candidates must generate and rank substitutes, in the (easier) ranking setting, systems are given all candidate substitutes from the benchmark (including those marked as unacceptable) and tasked with ranking them by their appropriateness.

D.2 Evaluation metrics

To evaluate ranking models, we adopt standard practice and report generalized average precision (GAP) (Kishida, 2005). GAP is similar to mean average precision, but assigns more credit for systems which produce substitutes that have higher scores in the reference list. Considering that our data collection procedure results in reference scores which correspond more to substitute appropriateness than ease-of-recollection, GAP is aligned with our high-level goals.

D.3 Baselines

We evaluate contextual embeddings from BERT and word embeddings from GLOVE (Pennington et al.,

2014), using cosine similarity of the target and substitute embeddings as the score. To compute the contextual embedding of a target or substitute with BERT, we mean pool contextual embeddings of its constituent word pieces. Because GLOVE discards contextual information, we expect it to perform worse than BERT, and is mainly used to assist interpretation of GAP scores. In the ranking setting, we are unable to evaluate GPT-3 and WORDTUNE, as we interface with these systems via an API which provides limited access to the underlying models. We report GAP scores in Table 6.

D.4 Results

We posit that contextual word embedding models should be invariant to contextual synonymy—they should embed acceptable substitutes nearby to one another. Hence, the SWORDS ranking setting may offer a useful perspective for evaluating this aspect of such models. In the ranking setting, our best contextual embedding model (BERT) achieves a GAP score of 57.1. While BERT outperforms a simple context-free baseline (GLOVE), it falls short of the 67.7 GAP score achieved by HUMANS. We interpret this as evidence that contextual embedding models have room to improve before attaining the aforementioned invariance.

D.5 Lexical substitution as natural language generation

GPT-3 is a language model which generates text in a left-to-right order, and is not designed specifically for the task of lexical substitution. To use GPT-3 to perform lexical substitution, we formulate the task in terms of natural language generation, and use in-context learning as described in (Brown et al., 2020). Specifically, we draw examples at random from the SWORDS development set to construct triplets of text consisting of (context with target word indicated using asterisks, natural language query, comma-separated list of all substitutes with score $> 0\%$ in descending score order) as follows:

Phone calls were monitored. An undercover force of Manhattan Project security agents ****infiltrated**** the base and bars in the little town of Wendover (population 103) to spy on airmen. Karnes knew the 509th was preparing for a special bombing mission, but he had no idea what kind of bombs were involved.

Q: What are appropriate substitutes for **infiltrated** in the above text?

A: penetrate, swarm, break into, infest, overtake, encompass, raid, breach

We construct as many of these priming triplets as can fit in GPT-3’s 2048-token context (roughly 12 examples on average), leaving enough room for a test example formatted the same way except without the list of answers. Then, we query the 175B-parameter `davinci` configuration of GPT-3 to generate a text continuation with up to 128 tokens. Finally, we parse the generated text from GPT-3, using its natural language ordering as the ordering for evaluation.

In an initial pilot study on a random split of our development set, we selected the sampling hyperparameters for GPT-3 as `temperature 0`, `presence_penalty 0.5`, and `frequency_penalty 0`, among possible candidates of $\{0, 1\}$ and $\{0, 0.5, 1.0\}$, and $\{0, 0.5, 1.0\}$, respectively. We used a grid search (18 runs) to select values based on highest F_c^{10} .

E Additional evaluation results

We include additional results from our evaluation. In Table 7, we break down F^{10} from Table 4 into P^{10} and R^{10} . In Table 8, we report performance of all generative baselines on traditional metrics for lexical substitution.

Model	Lenient				Strict			
	P_c^{10}	R_c^{10}	P^{10}	R^{10}	P_c^{10}	R_c^{10}	P^{10}	R^{10}
HUMANS*	75.3	77.5	44.9	60.6	–	–	–	–
COINCO	70.5	57.5	24.4	59.3	–	–	–	–
THESAURUS [†]	58.3	60.4	16.3	50.2	–	–	–	–
THESAURUS	49.4	51.1	11.6	35.9	–	–	–	–
WORDTUNE [†]	75.6	30.7	32.1	38.8	39.4	28.7	17.1	37.2
GPT-3	75.0	34.3	30.3	41.3	36.9	32.1	15.4	39.9
WORDTUNE	75.4	30.6	31.9	38.5	38.1	27.7	16.6	36.1
GPT-3 [†]	75.2	34.3	30.1	41.0	35.3	30.7	15.1	39.1
BERT-K [†]	67.4	44.8	24.2	48.0	28.2	30.1	12.1	38.6
BERT-LS	67.2	44.2	23.9	46.8	25.3	27.0	11.0	35.0
BERT-K	66.6	44.3	23.7	46.9	23.0	24.5	10.0	31.8
BERT-M [†]	69.2	35.8	25.4	39.1	23.6	25.2	10.3	32.7
BERT-M	69.0	35.7	25.4	39.1	15.6	16.6	6.8	21.7

Table 7: Expansion on the results from Table 4, breaking down F-measures by precision and recall. *Computed on a subset of the test data. [†]Reranked by our best ranking model (BERT).

Model	BEST	BEST-M	OOT	OOT-M	P^1
ORACLE	7.8	98.1	73.6	99.4	100.0
HUMANS*	3.9	29.8	56.1	93.0	91.0
CoINCo	6.3	28.6	41.6	71.0	87.9
THESAURUS [†]	2.3	18.2	37.7	51.8	79.6
THESAURUS	2.3	7.3	30.3	37.9	59.1
WORDTUNE [†]	2.8	22.8	20.0	42.2	88.1
GPT-3	2.6	24.1	22.9	45.6	86.7
WORDTUNE	2.8	17.6	19.4	41.3	83.9
GPT-3 [†]	2.6	19.8	22.1	44.3	86.2
BERT-K [†]	0.7	18.3	22.9	47.1	84.8
BERT-LS	0.7	17.9	20.6	41.7	81.2
BERT-K	0.7	14.9	18.8	40.4	77.7
BERT-M [†]	0.6	18.9	19.2	38.5	83.0
BERT-M	0.6	8.0	13.0	27.5	76.0

Table 8: Evaluation of models on SWORD in the generative setting using traditional evaluation metrics. We also include numbers for an ORACLE, as (unlike for F^{10} and GAP), the oracle does not achieve a score of 100. *Computed on a subset of the test data. [†]Reranked by our best ranking model (BERT).