

Measuring Context-Word Biases in Lexical Semantic Datasets

Anonymous ACL submission

Abstract

State-of-the-art pretrained contextualized models (PCM) eg. BERT use tasks such as WiC and WSD to evaluate their *word-in-context* representations. This inherently assumes that performance in these tasks reflect how well a model represents the coupled word and context semantics. We question this assumption by presenting the first quantitative analysis on the context-word interaction being tested in major contextual lexical semantic tasks. To achieve this, we run probing baselines on masked input, and propose measures to calculate and visualize the degree of context or word biases in existing datasets. The analysis was performed on both models and humans. Our findings demonstrate that models are usually not being tested for word-in-context semantics in the same way as humans are in these tasks, which helps us better understand the model-human gap. Specifically, to PCMs, most existing datasets fall into the extreme ends (the retrieval-based tasks exhibit strong target word bias while WiC-style tasks and WSD show strong context bias); In comparison, humans are less biased and achieve much better performance when both word and context are available than with masked input. We recommend our framework for understanding and controlling these biases for model interpretation and future task design.

1 Introduction

Meaning contextualization (i.e., identifying the correct meaning of a target word in linguistic context) is essential for understanding natural language, and has been the focus in many lexical semantic tasks. Pretrained contextualized models (PCMs) have brought large improvements in these tasks including WSD (Hadiwinoto et al., 2019; Loureiro and Jorge, 2019; Huang et al., 2019; Blevins and Zettlemoyer, 2020), WiC (Pilehvar and Camacho-Collados, 2019; Garí Soler et al., 2019) and entity

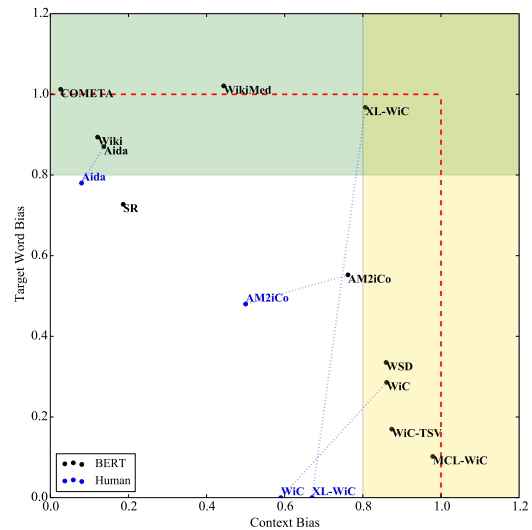


Figure 1: Plotting context and target word biases from BERT (black) and humans (blue) across popular context-aware lexical semantic datasets. The green shade and the yellow shade roughly indicate the areas for high target word bias and high context bias (>0.8). We would ideally want a dataset to lie towards the bottom left corner which is bias-free. The dashed red lines indicate 1.0 context (right) and 1.0 target word bias (top), implying a dataset is in effect dealt with by relying on target words alone or context alone.

linking (EL) (Wu et al., 2020; Broscheit, 2019).

These superior performances have been taken as proof that PCMs can successfully model *word-in-context* semantics. Many studies have investigated the process of lexical contextualization in these PCMs. Specifically, Vulić et al. (2020); Aina et al. (2019) found language models ‘contextualize’ words in higher layers while the type-level information is better kept in lower layers¹. Voita et al. (2019) point out different learning objectives affect the contextualization process, and Garí Soler

¹In this study, we do not perform layer-wise analysis as we fine-tune the PCMs to achieve the optimal performance in each task and we expect the relevant information is already surfaced to the last layer (Peters et al., 2019)

054 and Apidianaki (2021); Pimentel et al. (2020) show 101
055 PCMs can capture words’ ambiguity levels. 102

056 While most these studies have focused on probing 103
057 the innerworkings of the PCM feature space, 104
058 there is no systematic study to quantify the word- 105
059 context interaction (either learned by PCMs or in- 106
060 trinsic) across different lexical semantic tasks. On 107
061 one hand, these datasets often vary in their em- 108
062 phasis on context vs target words. For example, 109
063 we could expect tasks such as WSD and WiC to 110
064 rely more on context by design as the target words 111
065 are either given or the same in each input pair 112
066 ². On the other hand, models may find shortcuts 113
067 from datasets to avoid learning the complex word- 114
068 context interaction. **What is missing in the cur- 115
069 rent literature is an accurate quantification of 116
070 this word-context interplay being tested in each 117
071 task so that we can fully understand task goals 118
072 and model performance.** In particular, we need 119
073 to flag the situation where a model can solve a task 120
074 by relying solely on context or the target words. 121
075 Such heavy word or context reliance hinders a sci- 122
076 entific assessment of the models’ meaning contex- 123
077 tualization abilities as it essentially bypasses the 124
078 key word-context interaction challenge in human 125
079 understanding of lexical semantics. Therefore, we 126
080 refer to such heavy reliance on target words or 127
081 context in a contextual lexical semantic dataset as 128
082 target word biases or context biases ³.

083 This study presents an analysis framework to 129
084 quantify this context-word interaction by measur- 130
085 ing context and target word biases across lexical 131
086 semantic tasks. We first run controlled probing 132
087 baselines by masking the input to show the context 133
088 or the target word alone. Based on model’s perfor- 134
089 mance on these probing baselines, we calculate two 135
090 ratios that reflect how much of the model perfor- 136
091 mance in this dataset can be achieved from simply 137
092 relying on context alone or the target word alone, 138
093 i.e. the degree of context or target word biases (See 139
094 Figure 1 which will be discussed fully in Section 3). 140
095 The design of the probing baselines follows pre- 141
096 vious studies that applied input perturbation tech- 142
097 niques for model and task analysis in GLUE (Pham 143
098 et al., 2020), NLI (Poliak et al., 2018; Wang et al., 144
099 2018; Talman et al., 2021) and relation extraction 145
100 (Peng et al., 2020). While previous probing studies 146

²Notice that the exact amount of context/target word re-
liance in these tasks is to be tested as humans naturally use
both to make prediction.

³This is also in line with Gardner et al. (2021)’s claim that
all simple feature correlations from partial input are spurious.

usually assume no meaningful information from 101
corrupted input with no human verification, we pro- 102
vide fairer comparison with model performance by 103
collecting human judgment on the same masked in- 104
put in four tasks. Such comparison reveals whether 105
the biases are learned by models from the datasets 106
or are inherent in the tasks. 107

Our key findings are (1) the tasks can be clearly 108
divided into target-word-biased (the retrieval-based 109
tasks), and context-biases (WiC-style tasks and 110
WSD). Among the retrieval-based tasks, domain 111
affects ambiguity level and thus the target word 112
bias: models even achieve the best performance us- 113
ing target words alone in the medical domain. (2) 114
AM²ICO and Sense Retrieval show less extreme 115
model biases and challenge a model more to rep- 116
resent both the context and target words; and (3) 117
a similar trend of biases exists in humans but is 118
much less extreme, as humans find semantic judg- 119
ment more difficult on masked input and require 120
both word and context to do well in each task. This 121
analysis helps us better understand the nuanced 122
differences between models and humans in exist- 123
ing tasks, and we recommend the framework to 124
be applied when designing new datasets to check 125
whether word and context are required and whether 126
the models rely on the coupled word and context 127
semantics in a similar way to humans. 128

2 The Analysis Framework 129

2.1 Task Selection 130

We examine the following contextual lexical se- 131
mantic tasks, and for illustration, we list example 132
data for each task in Appendix A in the appendix. 133

Word Sense Disambiguation (WSD). WSD (Nav- 134
igli, 2009; Raganato et al., 2017) requires a model 135
to assign sense labels to target words in context 136
from a set of possible candidates for the target 137
words. Following the standard practice, we use 138
SemCor (sense-annotated texts created from Brown 139
Corpus) as the train set, Semeval2007 as dev, and 140
report accuracy on the concatenated ALL testset. 141

**The WiC-style Tasks (WiC, WiC-TSV, MCL- 142
WiC and XL-WiC).** To alleviate WSD’s require- 143
ment for a sense inventory, WiC (Pilehvar and 144
Camacho-Collados, 2019) presents a pairwise clas- 145
sification task where each pair consists of two word- 146
in-context instances. The model needs to judge 147
whether the target words in a pair have the same 148
contextual meanings. WiC-TSV (Breit et al., 2021) 149
extends the WiC framework to multiple domains 150

and settings. This study adopts the combined sets used for training. For both AIDA and Wiki, the search space is the full Wikipedia entity list. instance paired with a definition and a hypernym, WikiMed and COMETA. To test domain effects, we evaluate on two medical EL tasks. We use the WikiMed corpus (Vashishth et al., 2020), an automatically extracted medical subset from Wikipedia, for medical wikification. Each mention is mapped to a Wikipedia page linked to a concept in UMLS (Bodenreider, 2004), a massive medical concept KB. We define the search space as the Wikipedia entities covered in UMLS. With the same Wikipedia ontology but a different domain subset, WikiMed can be directly compared with Wiki for assessing domain influence. We also test on COMETA (Basaldella et al., 2020), a medical EL task in social media. COMETA consists of 20k English biomedical entity mentions from online posts in Reddit. The expert-annotated labels are linked to SNOMED CT (Donnelly et al., 2006), another widely-used medical KB. We report accuracy for WSD and all the WiC style tasks, and accuracy@1 for retrieval-based tasks including Wiki, AIDA, etc.

2.2 Probing Baselines

Sense Retrieval (SR) With the same train and test data as WSD, SR (Loureiro and Jorge, 2019) requires a model to retrieve a correct entry from the full sense inventory of WordNet (Miller, 1998). AIDA and Wiki cation. An important application scenario for testing meaning contextualization is Entity Linking (EL). EL maps a mention (an entity in its context) to a knowledge base (KB) which is usually Wikipedia in the general domain. The target word and its context help solve name variations and lexical ambiguity, which are the main challenges in EL (Shen et al., 2014). In addition, the context itself can help learn better representations for rare or new entities (Schick and Schütze, 2019; Ji et al., 2017). We test on two popular Wikipedia-based EL benchmarks: AIDA (Hoffart et al., 2011) and Wikification (Wiki) (Ratinov et al., 2011; Bunescu and Paşca, 2006). AIDA provides manual annotations of entities with Wikipedia and YAGO2 labels for 946, 216 and 231 articles as train, dev and test sets respectively. The Wiki Dataset is based on the hyperlinks from Wikipedia. We randomly sampled 50k sentences from Wikipedia as the test and another 50k as the dev set. The rest

Context vs. Word: For the main experiment, we design the WORD baseline where we input only the target word to the model, and the CONTEXT baseline where the target word is replaced with a [MASK] token in the input. The model is then trained and tested on the perturbed input. A high performance in CONTEXT or WORD will indicate strong context or target word bias. Example baseline input is shown in Table 1.

Lower Bound: Apart from a RANDOM baseline, we also set up a LABEL baseline where all the input is masked and the learning is only from the label distribution in the task. Notice that training the LABEL baseline is preferable to simply counting label occurrences in the data as the former can work with both continuous and categorical label space. All the probing baselines are compared with model performance on the full input (FULL). We refer to model M's performance in WORD, CONTEXT, LABEL and FULL as M_W , M_C , M_L and M_{Full} respectively.

Human Evaluation: To measure the inherent task biases, we collect human judgment (HUM) for a subset of WiC, XL-WiC, AM²iCo and AIDA) as being representative of the tasks described in Section 2.1

⁴We performed the analysis on other datasets AM²iCo and found the trend is similar

⁵In the surveyed tasks, a target word can show different surface variations of number, case and etc. E.g. red breeds

and feasible given resources for annotation. WiC target word biases in a dataset $Bias^{M_C}$ is the ratio of M_C to M_{Full} with the LABEL performance M_L in different languages; AIDA is chosen as a representative retrieval-based task. We follow the quality control procedures in [Pilehvar and Camacho-Collados \(2019\)](#); [Liu et al. \(2021\)](#) to recruit two different annotators for each baseline input from the CONTEXT, WORD and for FULL input in each task. The annotators are recruited from Prolog. They have graduate degrees and are fluent or native in the language of the dataset. In each setup, an annotator is assigned a randomly sampled 100 examples from the test set of each task. 50 example overlap between the two annotators for a dataset can be achieved from context alone or agreement calculation. The annotators are asked to perform meaning judgment in WiC, XL-WiC and AM²ICO, and to find the corresponding Wikipedia pages for entities for AIDA. For CONTEXT input where the target words are masked, annotators are encouraged to first guess what the target word could be⁶. As to the WORD input, annotators are asked to think of the most representative meaning of the out-of-context words when performing the target word and since this gap is small with a high tasks. As the pairs of input are always the same word by design in WiC and XL-WiC, we assume humans will give true judgment for all the examples and therefore will score 0.5 for WiC and XL-WiC. As to human LABEL baseline performance, while humans are not given any prior indication of how the task labels will be distributed, it is reasonable to expect that an annotator will give a random choice between the available labels or stick with one label when there is no input. Therefore, we approximate the human baseline as being 0.5 for WiC, XL-WiC and AM²ICO, and 0 for AIDA.

2.3 Calculating the Bias Measures

Based on a model's performance on the full input and on the baseline input, we propose $Bias^{M_C}$ and $Bias^{M_W}$ (as calculated in Equation (1) and Equation (2)) to measure the model's context and

$$Bias^{M_C} = \frac{(M_C - M_L)}{(M_{Full} - M_L)} \quad (1)$$

$$Bias^{M_W} = \frac{(M_W - M_L)}{(M_{Full} - M_L)} \quad (2)$$

Like models, humans can also be biased as they can use their prior knowledge (eg. humans can guess the typical meaning of a word without knowing the context) to make predictions based on partial input ([Gardner et al., 2021](#)). To measure how much humans can perform on the baseline input will help us understand the biases inherent in a task. We therefore calculate the context and target word bias scores for humans in the same way.

2.4 Experiment setup

The underlying model for our main experiments is BERT ([Devlin et al., 2019](#)), the most popular PCM that offers dynamic contextual word representations as bidirectional hidden layers from a transformer architecture. To ensure the general trend of our findings are consistent across different models, we also performed the analysis using RoBERTa ([Liu et al., 2019](#)), which improves upon BERT by optimized design decisions during training.

⁶Notice the annotator profiles in this study may be different from the original annotation scheme which was not always clearly specified. Therefore results on FULL in this study may be different from figures originally reported.

⁷We cannot use the test set for WiC and XL-WiC as the test labels are undisclosed. As the dev set comes from the same distribution of the test, we use dev to estimate human performance in these two tasks.

⁸We provide an example annotation guideline in appendix G. Human has lower but still reasonable agreement in the probing baselines (where there is naturally less information) than with FULL input (appendix F).

Input	Sentence1	Sentence2	BERT	HUMAN
FULL	Google represents a new [breed] of entrepreneurs .	The [breed] of tulip .	F	F
CONTEXT	Google represents a new [MASK] of entrepreneurs .	The [MASK] of tulip .	F	T
WORD	breed	breed	T	-
GUESSEWORD	Google represents a new [type] of entrepreneurs .	The [type] of tulip .	F	T

Table 1: Example input of FULL, CONTEXT and WORD in WiC. Target words are in brackets and the original WiC label for the FULL example is F. GUESSEWORD shows human-elicited target words based on CONTEXT. Comparing CONTEXT and GUESSEWORD also shows BERT’s contextual bias in WiC as BERT is not sensitive to the target word change.

We adopt standard model netuning setups in results (black dots) in Figure 1 (We also report each task. We use the base uncased variant of BERT biases in Appendix E and found a similar trend). For comparison, we plot human baseline performance and biases alongside the model performance in each figure.

that learns a sentence-gloss pair classification model based on BERT. For the WiC-style tasks, we follow the SuperGlue (Wang et al., 2019) practices to concatenate BERT’s last layer of [CLS] and the target words’ token representations for each input pair, followed by a linear classifier. For the retrieval-based tasks including SR and EL, we adopt a bi-encoder architecture to model query and target candidates with BERT (Wu et al., 2020). For the query, we insert [t and] to mark the start and end positions of the target word in context. Each target candidate is reformatted as [CLS]Name || Description[SEP] “. Name is an entity title (EL) or synset lemmas from WordNet (SR). Description is the first sentence in an entity’s Wikipedia page (Wiki & WikiMed), a gloss (SR), or n/a (COMETA). The model learns to draw closer the true query-target pairs’ representations using triplet loss with triplet miners during netuning (Liu et al., 2020). For each experiment, we perform grid search for the learning rate in [1e -5; 2e -5; 3e -5] and select models with early stopping on the dev set. We also run all the models with three random seeds and select the models with the best performance on the dev set. The performance across random seeds are stable as shown in small standard deviations which can be referred to in Table 7 in the appendix.

3 Main Results and Discussion

We report BERT’s baseline performance in Figure 2, based on which we calculate Bias^{BERT c} and Bias^{BERT w} for each dataset and plot the

⁹All PCM configurations are listed in Appendix D. We also conducted experiments with BERT (Liu et al., 2019) and reported the results in Appendix E

Figure 2: BERT and human performance on probing baselines across popular context-aware lexical semantic tasks. For the retrieval-based tasks, we report @1 accuracy, and the LABEL and RANDOM baselines are not visible as they are close to 0.

3.1 Model biases

Models can learn extreme context or target word biases from the datasets. One obvious observation from Figure 1 is that, probed with BERT, most of the datasets lie close to the dashed red lines: tasks such as WiC and MCL-WiC lie towards the right and are close to the vertical red line which indicates 1.0 context bias; the retrieval-based tasks such as WikiMed and Wiki lie towards the top and are close to or even surpass the horizontal red line which indicates 1.0 target word bias. This pattern indicates that BERT can score highly on these datasets by relying only on the target words or only on the context. In other words, context or target words can be much ignored when the model learns to solve the tasks. It is therefore questionable how much word-context interaction, which requires the modeling of both word and context representations,

is actually learned by BERT when applied to these tasks.

Moreover, the datasets tend to concentrate in two corners. That is, models usually learn strong bias from either context or the target word: the retrieval-based datasets (eg. Wiki) lie in the top left corner, showing large target word bias and low context bias; the WiC style datasets and WSD lie in the bottom right corner with large context bias and low target word bias. XL-WiC is an exception as it contains both strong context and target word bias. We will come back to this later in Section 3.2 where we compare model and human performance.

AM²ICo and SR are closest to testing word-context interaction from models. There are few existing datasets that in effect require the modeling of the context-word interaction, which should result in both low context and target word biases. SR and AM²ICo can be seen as two such datasets which, in Figure 1, can be found further inside of the red lines towards the bias-free left bottom corner. This is because these two tasks are designed to require balanced attention over context and target words. In SR, a system needs to model the target words in order to retrieve all the possible senses associated with the word, and because there is plenty of ambiguity in the dataset, context is also crucial to identify the correct sense. AM²ICo was specifically designed to include adversarial examples to penalize models that rely only on the context, and therefore elicits the lowest context bias from models among the WiC-style tasks. As such, SR and AM²ICo are the closest tasks that we have to test word-context interaction.

Domains affect lexical ambiguity and the target word bias.

	SemCor	Wiki	AIDA	WikiMed	COMETA
Sense Entropy	0.2102	0.060	0.0438	0.026	0.0004
Bias _{BERT} ^w	0.7274	0.8939	0.8705	1.0208	1.0124
Bias _{RoBERTa} ^w	0.7315	0.8994	0.8319	0.9957	1.1798

Table 2: Target Word Bias and Sense Entropy across retrieval-based tasks

across domains is arguably caused by the different degrees of lexical ambiguity in these tasks. In particular, domain could reduce ambiguity (Magnini et al., 2002; Koeling et al., 2005), and therefore affect the importance of the context and therefore the target word bias. As a quantitative measure for lexical ambiguity, we calculate average sense entropy across all words in each task’s training data, see Table 2. Confirming our hypothesis, sense entropy (lexical ambiguity) in a task does roughly correlate with the model’s target word bias: the medical domain tasks (WikiMed and COMETA) contain the lowest lexical ambiguity as reflected by the lowest sense entropy, and therefore missing context in these two tasks will not bring so much negative impact on the model performance, resulting in the highest target word biases; whereas higher sense entropy and thus higher lexical ambiguity (eg. Wiki and then SR) will necessarily require context alongside the target word, which leads to lower target word biases.

Context can harm model performance in Medical EL. We notice that the model’s target word bias in COMETA and WikiMed can go beyond 1.0, indicating that the model learning is dominated entirely by the target words with the context being useless or even harmful. This comes as a surprise as medical EL has been treated as a contextual lexical semantic task where the context is usually provided in the hope for higher modeling accuracy. We examined the errors from FULL as compared with WORD, and we found that the model tends to get distracted by related context words. Table 3 shows an example where the retrieval model selects the entry that is closer to a context word (“Miltonia”) than to the target word (“Miltoniopsis”), but in fact knowing the target word alone in this case is sufficient to retrieve the correct label. This indicates that the model has not learned a good strategy to incorporate word and context representations from the datasets (i.e. not knowing when to focus on the context and when to focus on the target words).

3.2 Human vs Model

The retrieval-based tasks in this study offer comparison between two domains, general vs medical, by comparing Wiki/AIDA and WikiMed. The target word bias is increased in the medical domain, where relying on the target words alone gives the best performance (i.e. COMETA and WikiMed both have > 1.0 target word bias). Such divergence is that humans show a similar trend of biases in the tasks in comparison to model biases (except for XL-WiC). This is evident from Figure 1 where, with the human bias indicators, WiC still lies near the bottom right corner with relatively high context bias; AIDA lies near the top left corner with high

	Input	Retrieved concept entry	Result
FULL	Formerly many more species were attributed to "Miltonia", ... including [Miltoniopsis] and Oncidium ...	miltonia: miltonia is an orchid genus comprising twelve epiphyte species and eight natural hybrids.	Wrong
WORD	Miltoniopsis	miltoniopsis: miltoniopsis is a genus of orchids native to costa rica and etc.	Correct

Table 3: Error analysis on FULL and WORD BERT predictions on WikiMed.

target word bias and AM²iCo remains in the middle. This confirms that there are some degrees of biases inherent in the task design so that humans can also rely on either target words or context alone to perform the task to some extent.

Humans are less biased than models.

Figure 3: The minimum gap between FULL and CONTEXT or WORD, i.e. $\min(\text{FULL-CONTEXT}, \text{FULL-WORD})$ with BERT and human performance. A small gap will indicate strong bias.

That being said, the second finding and the more important one is that humans exhibit overall much weaker biases in comparison with models in all the four tasks. If we compare human performance with model performance in Figure 2, we can see the CONTEXT and WORD baseline scores are lower in comparison to FULL from human performance. For clearer comparison, we calculate and plot the minimum gap between FULL to either of the two baselines in Figure 3, and we can see substantial difference between humans and models where humans exhibit much larger gaps across the four tasks. The much larger gaps from humans also result in all the four tasks moving further towards the left-bottom "bias-free" corner as shown Figure 1. In other words, humans are more likely than models to rely on both word and context as the absence of either part will lead to much more negative impact

for humans when performing these tasks.

The most dramatic difference is in XL-WiC where the model's strong target word bias disappears in humans. The task of XL-WiC by nature should not leak any information from the target word alone (hence 0 target word bias for humans) as the input pair will always contain the same target word. The high target word bias from models comes from the fact the dataset does not contain sufficient ambiguous cases where the same word pair can have both true and false labels dependent on the contexts. We confirm this by calculating the per-word average label entropy of the training data for all WiC-style tasks¹⁰ in Table 4, and we found XL-WiC has the lowest label entropy as 0.2084, and on average a word pair has the same label for 87% of the examples it appears in the dataset. Therefore, the model learns correlation between the word itself and the label without needing context for disambiguation.

	WiC	AM ² iCo	XL-WiC	MCL-WiC	WiC-TSV
LE	0.4022	0.5383	0.2084	0.2976	0.4212

Table 4: Label Entropy (LE) across WiC style tasks

Finally, the fact that models cannot achieve a similarly large jump of performance from masked input to FULL like humans could indicate the word-context interaction is particularly challenging for models and this might eventually explain the model-human gap. We take AM²iCo as an example that explicitly requires word-context interaction (Figure 5). While BERT achieves comparable results with humans in CONTEXT and WORD, a significantly larger human-model gap is found in FULL, indicating the word-context interaction is what the model lacks the most to achieve human-like performance.

	CONTEXT	WORD	FULL
Human	69	68.5	87.9
BERT	66	61	71
Human - BERT	3	7.5	16.9

Table 5: Model-human gap in CONTEXT, WORD and FULL in AM²iCo

Target words are important in WiC for humans. The much lower context bias from humans in tasks such as WiC suggests that the absence of the target words drastically decreases performance. In fact,

¹⁰We disregard words that only occur once

humanCONTEXT baseline (0.61) is even worse thus low bias measures. When interpreting the re-
 thanBERT (0.65) as shown in Figure 2. This may sults, apart from evaluating model performance on
 also come as a surprise, considering that targettheFULL input, we should also ensure the model
 words are always the same and only the context isshows a human-like gap in performance (between
 different in each pair of input. We examined hu-FULL and the baseline(s)) on the same data.
 man response iCONTEXT and found that humans
 can guess another valid target word based on the

5 Conclusion and Limitations

As qualitative analysis on the human-model dis-
 crepancy onCONTEXT, we examined 20 cases
 where annotators did not predict WiC labels (from
 the correspondingFULL input) whileBERT did.
 In 11 cases, humans guessed other valid target
 words to justify their predictions. We then perform
 preliminary analysis to testBERT on all the 11
 GUESSEWORD cases where the human-elicited
 target words change the labels (We show more ex-
 amples in Table 8), and found that for 7 out of 11,
 BERT is insensitive to the changed target words
 and maintains its original prediction. This suggests
 BERT does not appreciate the same word-context
 interaction as humans, and is making prediction
 mainly based on contexts rather than modeling con-
 textual lexical semantics in WiC.

4 Implications for future dataset design

We recommend this analysis framework in future
 dataset design and result interpretation for contex-
 tual word representation evaluation. In particular,
 we recommend (1) creating probing baselines by
 masking the context and word (if relevant), and
 (2) providing a sample to humans (details in sec-
 tion 2.2, (3) and comparing human and model per-
 formance of full input vs the masked baseline/s,
 and then calculate bias indicators. In terms of task
 design, we would ideally want both models and
 humans to show low baseline performance and

This study presented an analysis framework to dis-
 entangle and quantify context-word interplay in
 application of popular contextual lexical semantic
 benchmarks. With our proposed bias measures, we
 plot datasets on a continuum, and we found that, to
 models, most existing datasets lie on the two ends
 with excessive biases (WiC-style tasks and WSD
 are heavily context-biased while retrieval-based
 tasks are heavily target-word-biased) that essen-
 tially bypass the key challenges in word-context
 interaction. SR andAM²ICO have been identi-
 ed as two tasks that have less extreme biases
 and therefore can better test the representation of
 both word and context, and we call for more tasks
 that challenge models to do so. In addition, we
 identify that the degree of lexical ambiguity as a
 byproduct of domain affects target word bias (med-
 ical>general) in retrieval-based tasks. Most impor-
 tantly, we differentiate biases learned by models
 and task-inherent biases by collecting human re-
 sponses on the same baseline input. We found that
 models' heavy context and target word biases are
 not attested to the same extent in humans who usu-
 ally need both context and target words to perform
 well in the tasks. This suggests that models are
 relying on different cues instead of modeling con-
 textual lexical semantics as intended by the tasks.
 Our paper highlights the importance of understand-
 ing these biases in existing datasets and encourages
 future dataset and model design to control for these
 biases and to focus more on testing the challeng-
 ing word-context interaction in context-sensitive
 lexical semantics.

One limitation of this study is that we do not
 have large-scaled quantitative evidence to pinpoint
 the cues the models rely on from partial input.
 Possible future directions will be to design such ab-
 lation studies to identify any spurious correlations
 the models have learned and introduce adversarial
 examples that penalize sole reliance on context or
 target words in both task design and model training.

¹¹It could be that models like BERT (trained with masked
 language modelling) are genuinely better than humans in ex-
 ploiting partial input

647	References		
648	Laura Aina, Kristina Gulordava, and Gemma Boleda.	Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data . Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	703 704 705 706 707 708 709 710
649	2019. Putting words in context: LSTM language models and lexical ambiguity . Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3342–3348, Florence, Italy. Association for Computational Linguistics.		
650			
651	Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media . Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3122–3137, Online. Association for Computational Linguistics.	Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses . Transactions of the Association for Computational Linguistics 9:825–844.	711 712 713 714 715
652			
653			
654			
655			
656			
657			
658			
659			
660			
661	Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders . Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1006–1017, Online. Association for Computational Linguistics.	Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2019. Word usage similarity estimation with sentence representations and automatic substitutes . In Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019), pages 9–21, Minneapolis, Minnesota. Association for Computational Linguistics.	716 717 718 719 720 721 722
662			
663			
664			
665			
666			
667	Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. <i>Nucleic acids research</i> 32(suppl_1):D267–D270.	Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing .	723 724 725 726 727
668			
669			
670			
671	Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context . Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1635–1645, Online. Association for Computational Linguistics.	Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations . 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 5297–5306, Hong Kong, China. Association for Computational Linguistics.	728 729 730 731 732 733 734 735 736
672			
673			
674			
675			
676			
677			
678			
679	Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking . In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 677–685, Hong Kong, China. Association for Computational Linguistics.	Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenu, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text . In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.	737 738 739 740 741 742 743 744
680			
681			
682			
683			
684			
685	Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation . In 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy. Association for Computational Linguistics.	Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge . 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.	745 746 747 748 749 750 751 752 753
686			
687			
688			
689			
690			
691	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.	Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. Dynamic entity representations in neural language models . Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1830–1839, Copenhagen, Denmark. Association for Computational Linguistics.	754 755 756 757 758 759 760
692			
693			
694			
695			
696			
697			
698			
699			
700	Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for health studies in health technology and informatics. <i>Studies</i> 21:279.		
701			
702			

761	Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition . In <i>Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing</i> , pages 419–426, Vancouver, British Columbia, Canada. Association for Computational Linguistics.	817
762		818
763		819
764		820
765		821
766		822
767		823
768	Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pre-training for biomedical entity representations. arXiv preprint arXiv:2010.11784	824
769		825
770		826
771		827
772	Qianchu Liu, Edoardo M. Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. Am2ico: Evaluating word meaning in context across low-resource languages with adversarial examples .	828
773		829
774		830
775		831
776	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692	832
777		833
778		834
779		835
780		836
781	Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 5682–5691, Florence, Italy. Association for Computational Linguistics.	837
782		838
783		839
784		840
785		841
786		842
787		843
788	Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and A lo Gliozzo. 2002. The role of domain information in word sense disambiguation. <i>Natural Language Engineering</i> 8(4):359–373.	844
789		845
790		846
791		847
792	Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In <i>Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)</i>	848
793		849
794		850
795		851
796		852
797		853
798	George A Miller. 1998. <i>WordNet: An electronic lexical database</i> MIT press.	854
799		855
800		856
801	Roberto Navigli. 2009. Word sense disambiguation: A survey . <i>ACM Computing Surveys</i> 41(2):1–69.	857
802		858
803	Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 3661–3672, Online. Association for Computational Linguistics.	859
804		860
805		861
806		862
807		863
808		864
809		865
810		866
811	Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pre-trained representations to diverse tasks . In <i>Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)</i> , pages 7–14, Florence, Italy. Association for Computational Linguistics.	867
812		868
813		869
814		870
815		871
816		872

873 data sanity check: Assessing the effect of data
 874 corruption on model performance. *arXiv preprint*
 875 *arXiv:2104.04751*.

876 Shikhar Vashishth, Denis Newman-Griffis, Rishabh
 877 Joshi, Ritam Dutt, and Carolyn Rose. 2020. Im-
 878 proving broad-coverage medical entity linking with
 879 semantic type prediction and large-scale datasets.
 880 *arXiv preprint arXiv:2005.00460*.

881 Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The
 882 bottom-up evolution of representations in the trans-
 883 former: A study with machine translation and lan-
 884 guage modeling objectives. In *Proceedings of the*
 885 *2019 Conference on Empirical Methods in Natural*
 886 *Language Processing and the 9th International*
 887 *Joint Conference on Natural Language Processing*
 888 *(EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong,
 889 China. Association for Computational Linguistics.

890 Ivan Vulić, Edoardo Maria Ponti, Robert Litschko,
 891 Goran Glavaš, and Anna Korhonen. 2020. *Probing*
 892 *pretrained language models for lexical semantics*. In
 893 *Proceedings of the 2020 Conference on Empirical*
 894 *Methods in Natural Language Processing (EMNLP)*,
 895 pages 7222–7240, Online. Association for Computa-
 896 tional Linguistics.

897 Alex Wang, Yada Pruksachatkun, Nikita Nangia,
 898 Amanpreet Singh, Julian Michael, Felix Hill, Omer
 899 Levy, and Samuel R Bowman. 2019. Superglue: A
 900 stickier benchmark for general-purpose language un-
 901 derstanding systems. *Advances in Neural Informa-*
 902 *tion Processing Systems*, 32.

903 Alex Wang, Amanpreet Singh, Julian Michael, Fe-
 904elix Hill, Omer Levy, and Samuel Bowman. 2018.
 905 *GLUE: A multi-task benchmark and analysis plat-*
 906 *form for natural language understanding*. In *Pro-*
 907 *ceedings of the 2018 EMNLP Workshop Black-*
 908 *boxNLP: Analyzing and Interpreting Neural Net-*
 909 *works for NLP*, pages 353–355, Brussels, Belgium.
 910 Association for Computational Linguistics.

911 Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian
 912 Riedel, and Luke Zettlemoyer. 2020. *Scalable zero-*
 913 *shot entity linking with dense entity retrieval*. In
 914 *Proceedings of the 2020 Conference on Empirical*
 915 *Methods in Natural Language Processing (EMNLP)*,
 916 pages 6397–6407, Online. Association for Computa-
 917 tional Linguistics.

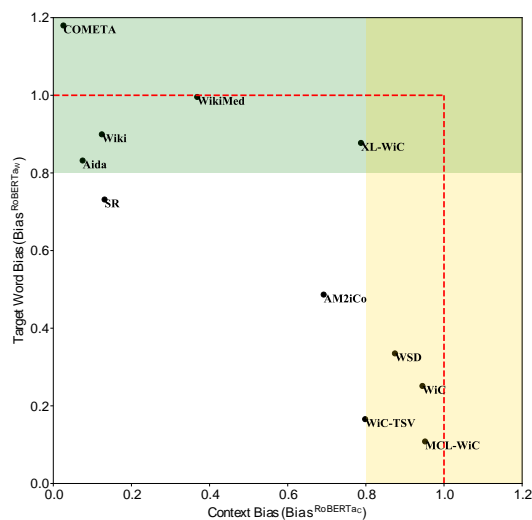


Figure 4: Plotting context and target word biases when applying ROBERTA across popular context-aware lexical semantic datasets. The green shade and the yellow shade roughly indicate the areas for high target word bias and high context bias (0.8). The dashed red lines indicate 1.0 context (right) and 1.0 target word bias (top), implying the model only requires the target words alone or context alone in this dataset.

A Task examples 918

Table 6 lists example input and labels for tasks 919 surveyed in this study. 920

B Dev performance 921

Table 7 shows BERT biases calculated over three 922 runs on the dev set with standard deviation re- 923 ported. 924

C Examples of the context bias in WiC 925

See Table 8 for two examples where the model 926 relies solely on the context to make the prediction. 927

D Model configurations 928

ALL PCMs are from <https://huggingface.co/>. 929 Model configurations are listed in Table 9. 930

931

Task	Input	Label	Label Space	Metrics
WiC	Room and [board]. He nailed [boards] across the windows.	F	T or F	Acc
WiC-TSV	I spent my [spring] holidays in Morocco. the season of growth; season, time of the year	T	T or F	Acc
MCL-WiC	Bolivia holds a key [play] in any process... A musical [play] on the same subject...	F	T or F	Acc
XL-WiC	Herr [Starke] wollte uns kein Interview geben. Das kann ich dir aber sagen: Wenn die Frau [Starke] kommt...	T	T or F	Acc
AM ² ICo	...航天员训练及[阿波罗]中飞船... ...the six [Apollo] Moon landings...	T	T or F	Acc
WSD	The [art] of change-ringing is peculiar to the English...	art : a superior skill that you can learn by study and practice and observation	art : the creation of beautiful or significant things art : the products of human creativity; works of art collectively ...(all possible meanings of <i>art</i>)	F1
SR	The [art] of change-ringing is peculiar to the English...	art : a superior skill that you can learn by study and practice and observation	art : a superior skill that you can learn by study and practice and observation door : a swinging or sliding barrier that will close the entrance... ... PLUS all other entries in WordNet	Acc
Wiki	an additional [Hash] literal syntax using colons for symbol keys...	hash table : in computing , a hash table (hash map) is a data structure...	hash table : in computing , a hash table (hash map) is a data structure ... united kingdom : the United Kingdom of Great Britain and Northern Ireland... ... (all entries in Wikipedia)	Acc@1
WikiMed	The flowers produce pollen, but no nectar. Various bees and flies visit the flowers looking in vain for nectar, for instance [sweat bees] in the genera "Lasiglossum" and "Halictus"...	halictidae : the Halictidae is the second largest family of Apoidea bees.	halictidae : the Halictidae is the second largest family of Apoidea bees. eomecon : eomecon is a monotypic genus of flowering plants in the poppy family... ... (all entries in the medical section of Wikipedia)	Acc@1
COMETA	I am [spacey] because I am thinking and daydreaming about my obsession.	dizziness (finding)	dizziness (finding) large intestine ...PLUS all other entries in SNOMED CT	Acc@1

Table 6: Examples for a selection of context-sensitive lexical semantic tasks surveyed in this thesis. Acc: accuracy; : Spearman’s correlation; r : Pearson’s correlation; P&R: precision and recall.

	WiC	WiC-TSV	WSD	MCL-WiC	XL-WiC	AM ² iCo SR	AIDA	Wiki	MedWiki	COMETA	
$Bias^{BERT_w}$	0.473 (0.016)	0.266 (0.043)	0.346 (0.015)	0.122 (0.007)	0.903 (0.002)	0.665 (0.008)	0.648 (0.012)	0.910 (0.007)	0.946 (0.002)	1.024 (0.022)	1.017 (0.034)
$Bias^{BERT_c}$	1.055 (0.017)	0.890 (0.028)	0.874 (0.020)	0.864 (0.043)	0.844 (0.002)	0.768 (0.016)	0.237 (0.011)	0.241 (0.015)	0.308 (0.003)	0.447 (0.010)	0.028 (0.010)

Table 7: Average context and target word biases over three runs with three different random seeds on the dev set in each dataset. Standard deviation is reported in the parenthesis.

Input	Sentence1	Sentence2	BERT	HUM
FULL	[Misdirect] the letter .	The pedestrian [misdirected] the out - of - town driver .	F	F
CONTEXT	[MASK] the letter .	The pedestrian [MASK] the out - of - town driver .	F	T
GUESSEWORD	[Ignore] the letter .	The pedestrian [ignored] the out - of - town driver .	F	T
FULL	[Kill] the engine .	He [kills] the ball .	F	F
CONTEXT	[MASK] the engine	He [MASK] the ball .	F	T
GUESSEWORD	[Hit] the engine .	He [hits] the ball .	F	T
FULL	[Kill] the engine .	He [kills] the ball .	F	F
CONTEXT	[MASK] the engine	He [MASK] the ball .	F	T
GUESSEWORD	[Hit] the engine .	He [hits] the ball .	F	T
FULL	His [treatment] of the race question is badly biased .	His [treatment] of space borrows from Italian architecture .	F	F
CONTEXT	His [MASK] of the race question is badly biased .	His [MASK] of space borrows from Italian architecture .	F	T
GUESSEWORD	His [understanding] of the race question is badly biased .	His [understanding] of space borrows from Italian architecture .	T	F
FULL	I could see it in the [distance] .	The [distance] from New York to Chicago .	F	F
CONTEXT	I could see it in the [MASK] .	The [MASK] from New York to Chicago .	F	T
GUESSEWORD	I could see it in the [train] .	The [train] from New York to Chicago .	T	F

Table 8: Example input of WORD, CONTEXT and FULL in WiC. The original WiC label for these examples is F. GUESSEWORD contains human-elicited target words that flip the label. Comparing CONTEXT and GUESSEWORD also shows BERT’s contextual bias in WiC as BERT is not sensitive to the target word change.

E ROBERTA Performance (Figure 4)

F Agreement in WiC-style tasks (Table 10)

G Annotation Guideline

Figure 5 shows an example annotation guideline for the CONTEXT experiment in WiC.

Model	Variant name in Huggingface	Parameters	Pretraining corpus
BERT	bert-base-uncased	12-layer, 768-hidden, 12-heads, 110M parameters	Lowercased Wikipedia + BookCorpus
PUBMEDBERT	microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	12-layer, 768-hidden, 12-heads, 110M parameters	Lowercased abstracts from PubMed and full-text articles from PubMedCentral
DEBERTA	microsoft/deberta-large	24-layer, 1024-hidden, 16-heads, 400M parameters	Wikipedia + BookCorpus + OPENWEBTEXT (public Reddit content) + STORIES

Table 9: Model details in our experiments

Figure 5: An annotation guideline for conducting the CONTEXT baseline of humans in WiC.

	AM ² iCo	XL-WiC	WiC
CONTEXT	94.0	88	76
FULL	87.9	66	64

Table 10: Human agreement in CONTEXT and FULL in WiC-style tasks