Measuring Context-Word Biases in Lexical Semantic Datasets

Anonymous ACL submission

Abstract

State-of-the-art contextualized models 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 required and being tested in major contextual lexical semantic tasks, taking into account that tasks can be inherently biased and models can learn spurious correlations from datasets. To achieve this, we run probing baselines on masked input, based on which we then propose measures to calculate the degree of context or word biases in a dataset, and plot existing datasets on a continuum. The analysis were performed on both models and humans to decouple biases inherent to the tasks and biases learned from the datasets. We found that, (1) to models, most existing datasets fall into the extreme ends of the continuum: the retrieval-based tasks and especially the ones in the medical domain (eg. COMETA) exhibit strong target word bias while WiC-style tasks and WSD show strong context bias; (2) AM^{2}ICO and Sense Retrieval show less extreme model biases and challenge a model more to represent both the context and target words. (3) A similar trend of biases exists in humans but humans are much less biased compared with models as humans found semantic judgments more difficult with the masked input, indicating models are learning spurious correlations. This study demonstrates that with heavy context or target word biases, models are usually not being tested for word-in-context representations as such in these tasks and results are therefore open to misinterpretation. We recommend our framework as a sanity check for context and target word biases in future task design and model interpretation in lexical semantics.

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 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. However, on one hand, the evaluation benchmarks often vary in their emphasis on context vs target words. For example, we could expect tasks such as WSD and WiC to rely more on context by design as the target words are either given or the same in each input pair. Notice that the exact amount of context/target word reliance in these tasks is to be tested as humans naturally use both to make prediction. On the other hand, models may find shortcuts from datasets to avoid learning the complex word-context interaction. **What is missing in the current literature is an accurate quantification of this word-context interplay required and being tested in each task so that we can fully understand task goals and model performance.** In particular, we need to flag heavy word and context reliance where a model can solve a task by relying solely on context or the target words. Such heavy word or context reliance hinders a scientific understanding of the models' meaning contextualization abilities as it essentially bypasses the key word-context interaction challenge in meaning contextualization, which requires the modeling of both target words and their contexts (Words are frequently ambiguous, but so are contexts. In "*I like XX* .", *XX* could have a number of meanings).

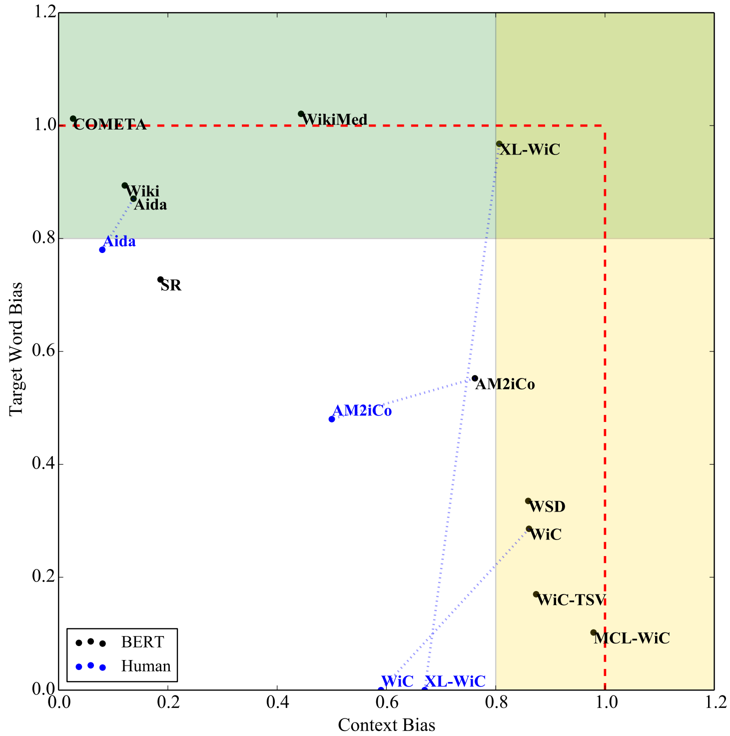

Figure 1: Plotting context and target word biases from BERT (blue) and humans (black) across popular context-aware lexical semantic datasets. The green shade and the yellow shade roughly indicate the areas for high target word bias and high context bias (>0.8). We would ideally want a dataset to lie 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.

Therefore, we refer to such heavy reliance on target words or context in a contextual lexical semantic dataset as target word biases or context biases. This is also in line with Gardner et al. (2021)'s claim that all simple feature correlations based on partial input are spurious.

This study presents an analysis framework to quantify this context-word interaction by measuring context and target word biases. We first run controlled probing baselines by masking the input to show the context or the target word alone. Based on model's performance on these probing baselines, we calculate two ratios that reflect how much of the model performance in this dataset can be achieved from simply relying on context alone or the target word alone, i.e. the degree of context or target word biases (See Figure 1 which will be discussed fully in Section 3). The design of the probing baselines follows previous studies that applied input permutation techniques for model and task analysis in GLUE (Pham et al., 2020), NLI (Poliak et al., 2018; Wang et al., 2018; Talman et al., 2021) and relation extraction (Peng et al., 2020). While previous probing studies usually assume no meaningful information from corrupted input with no human verification, we provide fairer comparison with model performance by collecting human judgment on the same masked input in four tasks. Such comparison reveals whether the biases are learned spuriously by models from the datasets or are inherent in the tasks.

2 The Analysis Framework

2.1 Task Selection

We examine a number of popular context-aware lexical semantic tasks. For illustration, we list example data for each task in Table 4 in the appendix.

Word Sense Disambiguation (WSD). WSD (Navigli, 2009; Raganato et al., 2017) requires a model to assign a sense label to a target word in context from a set of possible candidates for the target word. Following the standard practice, we use SemCor as the train set, Semeval2007 as dev, and report accuracy results on the concatenated ALL testset.

The WiC-style Tasks (WiC, WiC-TSV, MCL-WiC and XL-WiC). To alleviate WSD's requirement for a sense inventory, WiC (Pilehvar and Camacho-Collados, 2019) presents a pairwise classification task where each pair consists of two word-in-context instances. The model needs to judge whether the target words in a pair have the same contextual meanings. WiC-TSV (Breit et al., 2021) extends the WiC framework to multiple domains and settings. This study adopts the combined setting where each input consists of a word in context instance paired with a definition and a hypernym, and the task is to judge whether the sense intended by the target word in context matches the one described by the definition and is the hyponym of the hypernym. The WiC-style tasks have also been extended to the multilingual and crosslingual settings in MCL-WiC (Martelli et al., 2021), XL-WiC (Raganato et al., 2020) and more recently in AM^{2}ICO (Liu et al., 2021). MCL-WiC provides test sets for five languages with full gold annotation scores. However, MCL-WiC only covers training data in English. To ensure the analysis will be testing the same data distribution during both training and testing, we will only use the English dataset of MCL-WiC. XL-WiC extends WiC to 12 languages. While most languages in this task do not have training data, we perform analysis on its German dataset which does contain both train (50k) and test data (20k). AM^{2}ICO covers 14 datasets, each of which pairs English word-in-context instances with word-in-context instances in a target language. In this study, we perform analysis on the English-Chinese dataset which contain 13k train and 1k test data [1].

Sense Retrieval (SR). Based on WSD with the same train and test data, SR (Loureiro and Jorge, 2019) requires a model to retrieve a correct entry from the full sense inventory of all words from WordNet (Miller, 1998).

AIDA and Wikification. 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 is used for training. For both AIDA and Wiki, the search space is the full Wikipedia entity list.

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

Context vs. Word: For the main experiment, we design the WORD baseline where we input only the target word [2] 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:**

[1] We performed the analysis on other datasets of AM^{2}ICO and found the trend is similar

[2] In the surveyed tasks, a target word can show different surface variations of number, case and etc. Eg., *breed*, *breeds*.

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 (WiC, XL-WiC, AM^{2}ICO and AIDA) as being representative of the tasks described in Section 2.1 and feasible given resources for annotation. WiC, XL-WiC and AM^{2}ICO cover WiC-style datasets 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 CONTEXT,WORD and for FULL input in each task. The annotators are recruited from Prolific. 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[3] and there is a 50 example overlap between the two annotators for agreement calculation. The annotators are asked to perform meaning judgment in WiC, XL-WiC and AM^{2}ICO, 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 words could be. As to the WORD input, annotators are asked to think of the most representative meaning of the out-of-context words when performing the 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 on WORD input in WiC and XL-WiC. As to human's 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 LABEL human baseline as being 0.5 for WiC, XL-WiC and AM^{2}ICO, and 0 for AIDA.

[3]We 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.

2.3 Calculating the Bias Measures

Based on a model M'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 target word biases in a dataset. $Bias^{M_C}$ is the ratio of M_C to M_{Full} with the LABEL performance M_L deducted from both M_C and M_{Full}. M_L has to be deducted as it is unrelated to the input. Otherwise, the ratio will give an inflated bias measurement. $Bias^{M_W}$ is calculated in the same way as $Bias^{M_C}$ except that we replace M_C with M_W in the equation. The two measures can also be seen as M_C and M_W under min-max normalization where the min value is M_L and the max value is M_{Full}, and therefore the normalized values can be fairly compared across datasets. $Bias^{M_C}$ and $Bias^{M_W}$ reflect how much of what a model has learned from the input in a dataset can be achieved from context alone or target word alone, which will give us indicators of the degree of context and target word biases in the dataset. These bias indicators will in turn tell us how important the masked part of the input is. For example, we can interpret a $Bias^{M_C}$ of 0.9 as 90% of what the model has learned from the full input can be achieved from the context alone. The 10% gap can be gained from adding the masked target word and since this gap is small with a high context bias, we can conclude that the model can do pretty well just from the context alone and it is not learning much from the target word.

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

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

Like models, humans can also be biased as they can also use their prior knowledge or biases (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.

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	-
GUESSEDWORD	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. GUESSEDWORD shows human-elicited target words based on CONTEXT. Comparing CONTEXT and GUESSEDWORD also shows BERT's contextual bias in WiC as BERT is not sensitive to the target word change.

2.4 Experiment setup

The underlying model for our main experiments is BERT (Devlin et al., 2019), one of the most successful PCMs that offer 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.

We adopt standard model finetuning setups in each task. We use the base uncased variant of BERT[4] for general domain experiments and PUBMEDBERT (Gu et al., 2020) for the medical tasks. For WSD, we use GLOSSBERT (Huang et al., 2019) 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 `[` 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 finetuning (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 by small standard deviations which can be referred to in Table 5 in the appendix.

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

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 results (black dots) in Figure 1 (We also report ROBERTA 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.

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 learn a tremendous amount from 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

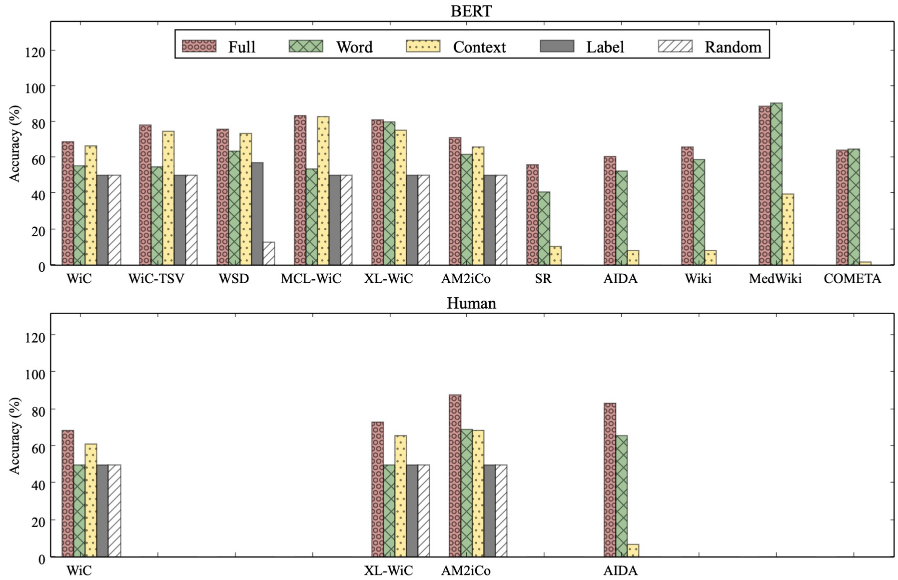

Figure 2: BERT and human performance on probing baselines across popular context-aware lexical semantic tasks. For the retrieval-based tasks, we report @1 accuracy, and the LABEL and RANDOM baselines are not visible as they are close to 0.

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 biases. We will come back to this later in Section 3.2 where we compare model and human performance.

AM^{2}iCo 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 AM^{2}iCo 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. AM^{2}iCo 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 AM^{2}iCo are the closest tasks that we have to test word-context interaction.

Domains affect lexical ambiguity and the target word bias.

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 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 along-

	SemCor	Wikification	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

side 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

There are inherent task biases. Our first finding 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 bottom right corner with relatively high context bias; AIDA lies near top left corner with high target word bias and AM^{2}ICO 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.

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

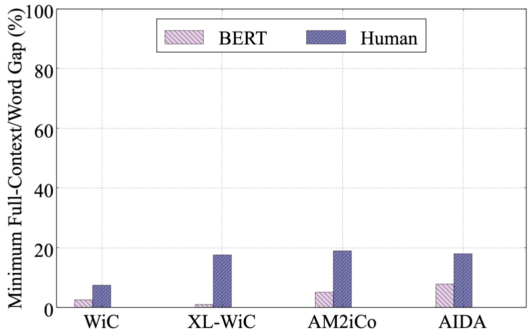

Figure 3: The minimum gap between FULL and CONTEXT or WORD, i.e. min(FULL-CONTEXT,FULL-WORD) with BERT and human performance. A small gap will indicate strong bias.

CONTEXT and WORD baseline scores are lower in comparison toFULL 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 as 0.09, and on average a word pair has the same label for 94% of the context examples it appears

Baseline	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.

in the dataset. Therefore, the model learns correlation between the word itself and the label without needing context for disambiguation.

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, human CONTEXT baseline (0.61) is even worse than BERT (0.65) as shown in Figure 2. This may also come as a surprise, considering that target words are always the same and only the context is different in each pair of input. We examined human response in CONTEXT and found that humans can guess another valid target word based on the context, which gives a different prediction. Table 1 shows such an example. While the original WiC label of the input is F, our annotator gave T for the CONTEXT input, guessing the target word is *type*. This is a reasonable prediction as *type* fits the contexts and does hold its meaning across the two sentences. We refer to this new example with human-elicited target words as GUESSEDWORD input. The same annotator was able to give the WiC label F when we reveal the original target word (*breed*) which has the specific meaning of *species* in sentence1 and *personality* in sentence2 (see the FULL input in Table 1). BERT however still predicts F regardless of the target word change in this GUESSEDWORD example.

As qualitative analysis on the human-model discrepancy on CONTEXT, we examined 20 cases where annotators did not predict WiC labels (from the corresponding FULL input) while BERT did. In 11 cases, humans guessed other valid target words to justify their predictions. We then perform preliminary analysis to test BERT on all the 11 GUESSEDWORD cases where the human-elicited target words change the labels (We show more examples in Table 6), 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 contextual lexical semantics in WiC.

4 Conclusion

This study presented an analysis framework to disentangle 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 essentially bypass the key challenges in word-context interaction. SR and AM^{2}ICO have been identified 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 (medical>general) in retrieval-based tasks. Most importantly, we differentiate biases spuriously learned by models and task-inherent biases by collecting human responses 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 usually need both context and target words to perform well in the tasks. This suggests that models are learning spurious correlations instead of modeling contextual lexical semantics as intended by the tasks. Our paper highlights the importance of understanding these biases in existing datasets and encourages future dataset and model design to control for these biases and to focus more on testing the challenging word-context interaction in context-sensitive lexical semantics. Possible future directions will be to include adversarial examples that penalize sole reliance on context or target words in both task design and model training.

References

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

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. In *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.

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.

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.

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.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

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. In *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.

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.

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.

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.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, 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.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Copenhagen, Denmark. Association for Computational Linguistics.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 419–426, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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*.

Qianchu Liu, Edoardo M. Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. Am2ico: Evaluating word meaning in context across low-resourcelanguages with adversarial examples.

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*.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

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 *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT 2019*, pages 1267–1273.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Alessandro Raganato, José Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of EACL 2017*, pages 99–110.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2019. Attentive mimicking: Better word embeddings by attending to informative contexts. 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 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. 2021. Nli data sanity check: Assessing the effect of data corruption on model performance. *arXiv preprint arXiv:2104.04751*.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

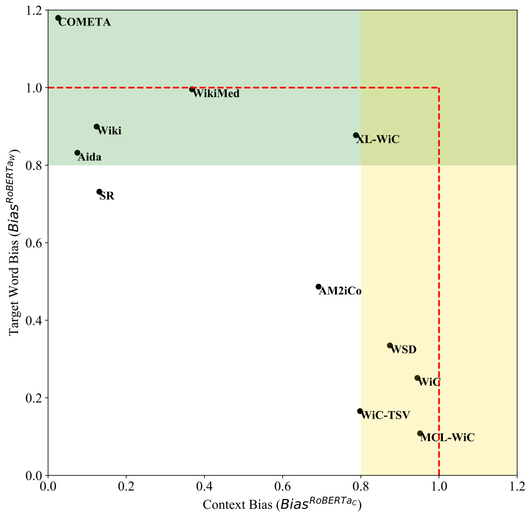

Figure 4: Plotting context and target word biases when applying RoBERTa across popular context-aware lexical semantic datasets. The green shade and the yellow shade roughly indicate the areas for high target word bias and high context bias (0.8). The dashed red lines indicate 1.0 context (right) and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

A Task examples

Table 4 lists example input and labels for tasks surveyed in this study.

B Dev performance

Table 5 shows BERT biases calculated over three runs on the dev set with standard deviation reported.

C Examples of the context bias in WiC

See Table 6 for two examples where the model relies solely on the context to make the prediction.

D Model configurations

ALL PCMs are from https://huggingface.co/. Model configurations are listed in Table 7.

E RoBERTa Performance (Figure 4)

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 *art*)	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 "Lasioglossum" 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 fl owering 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 4: Examples for a selection of context-sensitive lexical semantic tasks surveyed in this thesis. Acc: accuracy; ρ: Spearman's correlation; r: Pearson's correlation; P&R: precision and recall.

	WiC	WiC-TSV	WSD	MCL-WiC	XL-WiC	AM^2ICo	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 5: 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
GUESSEDWORD	[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
GUESSEDWORD	[Hit] the engine .	He [hits] the ball .	F	T

Table 6: Example input of WORD, CONTEXT and FULL in WiC. The original WiC label for these examples is F. GUESSEDWORD contains human-elicited target words that flip the label. Comparing CONTEXT and GUESSEDWORD also shows BERT's contextual bias in WiC as BERT is not sensitive to the target word change.

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 7: Model details in our experiments