Lost in Context? On the Sense-Wise Variance of Contextualized Word Embeddings

Yile Wang and Yue Zhang, Member, IEEE

Abstract—Contextualized word embeddings in language models have given much advance to NLP. Intuitively, sentential information is integrated into the representation of words, which can help model polysemy. However, context sensitivity also leads to the variance of representations, which may break the semantic consistency for synonyms. Previous works that investigate contextualized sensitivity focus on the token level representations, while we are taking a deeper dive into exploring representations at the fine-grained sense level. In particular, we quantifiy how much the contextualized embeddings of each word sense vary across contexts in typical pre-trained models, the results show that contextualized embeddings can be highly consistent across contexts, even for two different words with the same sense. In addition, part-of-speech, number of word senses, and sentence length have an influence on the variance of sense representations. Interestingly, we find that word representations are position-biased, where the first words in different contexts tend to be more similar. We analyze such a phenomenon and also propose a prompt-augmentation method to alleviate such bias in distance-based word sense disambiguation settings. Finally, we investigate the influence of sense-level pre-training on the performance of different downstream tasks, results show that such external tasks can improve the sense- and syntactic-related tasks, while not necessarily benefitting general language understanding tasks.

Index Terms—Language models, contextualized word embeddings, sense-wise variance, position bias.

I. INTRODUCTION

CONTEXTUALIZED word embeddings (CWE) from pre-trained language models [1], [2], [3], [4], [5], [6], [7] have achieved state-of-the-art performance in various downstream tasks. Different from static word embeddings (e.g., word2vec [8] and GloVe [9]), CWE represents words using a deep neural network, which can effectively model polysemy [1], capture syntactic [10], [11], semantic [12], [13], and commonsense knowledge [14], [15] from large corpora.

However, research shows that context sensitivity can also bring unnecessary variation for CWE. For example, Shi et al. [16] show that the ELMo [1] embeddings of a word change drastically when the context is paraphrased, which makes the downstream model not robust to paraphrasing and other linguistic variations. Such a phenomenon is beyond what we can expect from the perspective of semantic polysemy, which is not only undesirable linguistically but also has implications on NLP tasks such as word alignment [17], [18] and lexical semantics [19], [20]. In the machine learning aspect, it has been shown that stable representations are relevant to reducing over-fitting, improving generalization and relieving data hunger [21], [22].

Despite discussion on the consequences, there has been relatively little research characterizing the context-sensitivity in CWEs systematically. The only exception is Ethayarajh [23], which provides overall statistics for ELMo, BERT, and GPT on the token-level. In this paper, we aim to quantify the variance in the sense-level representations across contexts for different CWE models. Our goal is to quantitatively answer the following research questions. First, how much does CWE differ for the same word sense, and is the variation highly model-dependent? Second, what types of words and contexts are more prone to CWE variations? Third, how does word position affect CWE variations? Knowledge of the above issues can be useful for better understanding and improving CWE, and also for guiding the choice of models toward a specific problem.

Empirically comparing seven pre-trained models including ELMo, BERT, SenseBERT, RoBERTa, DeBERTa, XLM-Net, and GPT2, we find that:

1) Word sense representation in CWE is generally consistent across contexts, while different models have varying degrees of sense-wise consistency. Such representations are generally encoded on the higher layers of models, showing again that these features are highly used for the specific training objectives during pre-training (Section III, IV).

2) Sense-level representations vary differently according to word types and contexts, where they are more consistent for nouns or in short sentences. We also find that such representation is position-sensitive for the sentence beginning words, where the first words in two sentences tend to be very close. We call this position bias in CWE (Section V).

3) We find that adding some simple prompts can largely change the distribution of CWEs, especially for the first words. These findings can be used for calibrating representation via our proposed prompt augmentation method, which is shown effective for a distance-based word sense disambiguation scenario (Section VI).

Manuscript received 19 August 2022; revised 2 July 2023 and 20 November 2023; accepted 24 November 2023. Date of publication 29 November 2023; date of current version 8 December 2023. This work was supported by the National Natural Science Foundation of China under Grant 61976180. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson. (Corresponding author: Yue Zhang.)

Yile Wang is with the Institute for AI Industry Research, Tsinghua University, Beijing 100190, China (e-mail: wangyile@air.tsinghua.edu.cn).

Yue Zhang is with the School of Engineering, Westlake University, Hangzhou 310058, China (e-mail: yue.zhang@wias.org.cn).

Digital Object Identifier 10.1109/TASLP.2023.3337643

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
4) Through finetuning SenseBERT and BERT, we investigate the influence of sense-level pre-training on the performance of 12 downstream tasks of four different types, results show that such external objectives can improve the sense- and syntactic-related downstream tasks, while not necessarily benefiting general language understanding tasks (Section VII).

To our knowledge, we are the first to investigate sense-level representation variances of CWE, and the first to report the influence of both token, context, and position factors.

II. RELATED WORK

Our work is in line with existing work on BERTology [24], which discusses the characteristics of pre-trained models, aiming to provide evidence for explaining their effectiveness as well as their limitations.

Knowledge in CWE: A line of research finds that CWE can encode transferable and task-agnostic knowledge and successfully improve the performance of downstream tasks [25]. Peters et al. [1] propose ELMo and find it encodes syntactic and semantic features at different layers. Goldberg [26] finds BERT performs remarkably well on English syntactic phenomena. Jawahar et al. [12] reveal BERT composes a rich hierarchy of linguistic information, starting with surface features at the bottom, and syntactic features in the middle followed by semantic features at the top. Our work is related to discussing knowledge of CWE, with a focus on word sense and context.

Word Sense in CWE: Recent work shows that CWE is capable of encoding word sense information. Schuster et al. [27] find the ELMo embeddings can be separated into multiple distinct groups, each with a certain meaning. Reif et al. [28] show different senses of a word are typically spatially separated using BERT representation. Garí Soler and Apidianaki [29] demonstrate that BERT representations reflect word polysemy level and their partition ability into senses. Such capability makes the success of leveraging CWE into sub-tasks such as word sense disambiguation [30], [31], [32]. We also study sense-level knowledge inside CWE but differ in that our experiments are more on the variance of embeddings for the same sense, analyzing the influence from different contexts.

Contextualization of CWE: There have been works analyzing the variations of CWE according to different contexts. Etyahayaj [23] first compared ELMo, GPT2 and BERT, analyzing how the word representations change according to different models and layers. They find that the upper layers produce more context-specific representations, where the self-similarity of two identical words in different contexts becomes much lower. Our work is similar in the evaluation of cross-sentence scenarios. However, we consider sense-level comparison rather than word-level comparison. Different words in the same sense are compared. In addition, we report not only the overall statistics for different models, but also discuss breakdown factors to influence context sensitivity, both internally and externally for each sense type. Shi et al. [16] find that ELMo cannot capture semantic equivalence, where the word representations are very different in paraphrased contexts. To minimize the variance of CWE, they use paraphrases to retrofit word representation. Similarly, Elazar et al. [33] claim that the invariance under meaning-preserving alternations in its input is a desirable property for CWE, however, they find the consistency is poor for factual knowledge, leading to the unstable knowledge representation. The above work does not consider sense-level representation and the comparison is restricted to paraphrased contexts. In contrast, we use natural sentences with open context formation. In addition, we consider a wider range of models.

III. METHODS

A. Datasets and Models

We use the XL-WSD dataset [34], an extra-large Word Sense Disambiguation benchmark, to investigate the CWE of words in the corpus with labeled sense information. Specifically, the SEMCOR and WNGBT corpora are used, where the most widely used WSD dataset are gathered, including Senseval 2&3 [35], [36] and SemEval 07&13&15 [37], [38], [39]. All sense taggings have been merged from various resources using BabelNet1, which is a superset of Princeton WordNet, the standard sense inventory for English. Statistics of the dataset are shown in Table I.

We choose seven representative pre-trained models:

ELMo [1] uses a two-layer bi-directional LSTM [40] for CWE.

BERT [3] makes use of bi-directional Transformer [41] and pre-trained with the mask language modeling and next sentence prediction.

SenseBERT [42] is based on BERT, but is pre-trained to predict both masked words and their WordNet supersenses.

RoBERTa [4] is a variation of BERT, which applies more training corpora, dynamic masking, and removes the next sentence prediction (NSP) objectives.

DeBERTa [7] is similar to RoBERTa, while using relative position embedding, disentangled attention mechanism, and enhanced mask decoder.

XLNet [5] extends Transformer-XL [43] and is pre-trained using autoregressive permutation language modeling.

GPT2 [2] uses a Transformer decoder structure for generative language model pre-training.

| Types                  | Values   |
|------------------------|----------|
| #Total tokens          | 1,240,828|
| #Labeled tokens        | 273,815  |
| #Vocabulary            | 22,979   |
| #Sentences             | 83,277   |
| #Sentence length in average | 14.9    |
| #Senses inventory      | 44,046   |
| #Senses for each token in average | 1.65  |
| #Maximum senses for one token | 58     |

1[Online]. Available: https://babelnet.org
Specifically, we take the model checkpoints ELMo-SMALL, BERT-BASE-CASED, RoBERTa-BASE, DeBERTa-BASE, XLNet-BASE-CASED and GPT2-BASE from AllenNLP [44] and Huggingface [45]. For SenseBERT, we use the SENSEBERT-BASE-UNCASED released by Levine et al. [42]. Most of the model outputs are in the same 768-dimensional vector space (except for 1024 dimensions for ELMo) so they can be fairly compared.

B. Cosine Similarity

For each word \( t_i \) labeled with a specific sense \( s_i \) and the corresponding context \( t_1, \ldots, t_N \), we calculate the CWE \( h_i^l \) using the above models:

\[
h_i^1, \ldots, h_i^l, \ldots, h_i^{N_t} = \text{CWE}(t_1, \ldots, t_N, \ldots, t_{N_t})
\]

where \( l \) is the model layer and \( N_t \) is the sentence length. Special tokens such as [CLS] and [SEP] may be added according to different models while we discard them in the final. We use its first sub-token vector for representation for those words that are split into multiple tokens [46, 47].

Following Ethayarajh [23], we consider the normalized cosine similarity \( \cos(t_i, t_j) = \frac{\langle h_i, h_j \rangle}{||h_i|| \cdot ||h_j||} \) between two words \( t_i \) and \( t_j \) with hidden CWE \( h_i \) and \( h_j \), respectively. Note that \( t_i \) and \( t_j \) may come from different sentences. For aggregating evaluation of embedding similarities in a corpus, we find all pairs of hidden representations from \( t_i, t_j \), computing:

\[
\text{Sim} = \frac{\sum_{i,j} \cos(t_i, t_j)}{\text{Count}(t_i, t_j)}
\]

We consider sense-wise variance for the following targets:

**The Same Word in the Same Sense:** We compare the representation of the same words in the same word sense. For each word pair, the representations should be similar intuitively. For example, the representation of word “levels” should be similar when they share the same word sense BN:00041239N (A relative position or degree of value in a graded group) in the following sentences:

- There are three **levels** on which to treat the subject.
- The same command is repeated as many times as there are **levels** in rank from general to corporal.

**Different Words in the Same Sense:** Note that different words may have similar senses, showing that they share some common features in the semantic spaces. We also compare the CWE of different words but in the same word senses in certain contexts. For example, the words “levels”, “layers” and “strata” have the same word sense of BN:00050303N (An abstract place usually conceived as having depth) in the sentences below:

- a good actor communicates on several **levels**.
- a simile has at least two **layers** of meaning.
- the mind functions on many **strata** simultaneously.

Given the above observations, we categorize the measurement of (2) into two situations, namely the same sense expressed in the same word (\( \text{Sim}_{ss} \)) and the same sense expressed in different words (\( \text{Sim}_{ds} \)), using the equations below:

\[
\text{Sim}_{ss} = \frac{\sum_{i,j} \cos(t_i, t_j)}{\text{Count}(t_i, t_j)}, \quad t_i = t_j, s_i = s_j
\]

\[
\text{Sim}_{ds} = \frac{\sum_{i,j} \cos(t_i, t_j)}{\text{Count}(t_i, t_j)}, \quad t_i \neq t_j, s_i = s_j
\]

**Masked Words in the Same Sense:** For BERT and its variants (SenseBERT, RoBERTa, and DeBERTa), they use masked language modeling to predict the masked word during pre-training. Intuitively, the representation of token [MASK] can also reflect the sense-level information by its context. Similar to (3), we considered the comparison of such contextualized embeddings without the cues of the given token:

\[
\text{Sim}_{ss}^{\text{mask}} = \frac{\sum_{i,j} \cos(t_i, \tilde{t}_j)}{\text{Count}(t_i, \tilde{t}_j)}, \quad t_i = t_j, s_i = s_j,
\]

\[
\text{Sim}_{ds}^{\text{mask}} = \frac{\sum_{i,j} \cos(t_i, \tilde{t}_j)}{\text{Count}(t_i, \tilde{t}_j)}, \quad t_i \neq t_j, s_i = s_j
\]

where \( \tilde{t}_i \) and \( \tilde{t}_j \) are [MASK] symbols that replace the origin \( t_i \) and \( t_j \).

**Random Baseline:** The above scores reflect the context sensitivity of the same word sense across models in absolute cosine similarities. However, for some CWE models, the general cosine similarity between arbitrary vectors can be large. In order to understand both the absolute distance and the relative difference, we also randomly sample \( N = 10000 \) words and compute the average similarity among them as our baseline, and we define \( \Delta_{ss, \text{rand}} \) as the difference between \( \text{Sim}_{ss} \) and \( \text{Sim}_{\text{rand}} \):

\[
\text{Sim}_{\text{rand}} = \frac{2}{N(N-1)} \sum_{i,j} \cos(t_i, t_j)
\]

\[
\Delta_{ss, \text{rand}} = \text{Sim}_{ss} - \text{Sim}_{\text{rand}}
\]

C. Probing Method

Cosine similarity has been widely used for measuring representation similarity [8, 48]; however, it only offers one global perspective of the embeddings. As a result, we also consider adding a simple linear or MLP layer upon CWE for probing [25, 49] to show the local or transformed features. In particular, we build a sense equivalent judging task by constructing 20,000 token samples with sense annotations for training and 2,000 held-out samples for evaluation. We optimized the model by only finetuning the linear layer or MLP parameters while keeping the CWE fixed:

\[
P(y|t_i, t_j) = \text{softmax}(W_0([h_i; h_j])), \text{ or}
\]

\[
P(y|t_i, t_j) = \text{softmax}(W_1, g(W_2([h_i; h_j]))),
\]

where \( W_0, W_1 \) and \( W_2 \) are parameters, \( g \) is the ReLU function and \( [h_i; h_j] \) is the concatenation operation. \( y \) is the binary distribution indicating the sense labels \( s_i \) and \( s_j \) are equal or not.
CWE models encode word sense knowledge: ELMo and BERT show a relative large difference between $\text{Sim}_{\text{as}}$ and $\text{Sim}_{\text{rand}}$, reflecting the distinct spatial distribution according to different word senses [28]. Our findings align with Garí Soler and Apidianaki [29], who find a similar gap in BERT representation and show BERT representation can reflect the polysemous level. SenseBERT is trained with external word sense prediction and gives the largest difference, showing the usefulness of sense-level supervision signals during pre-training.

Embedding spaces of auto-regressive models are highly anisotropic: For XLNet and GPT2, the difference between $\text{Sim}_{\text{as}}$ and $\text{Sim}_{\text{rand}}$ ($\Delta_{\text{as}, \text{rand}}$) are only 0.01 and 0.02, respectively. $\text{Sim}_{\text{as}}$ is even slightly lower than $\text{Sim}_{\text{rand}}$ in XLNet. These can result from the highly anisotropic distribution of word vectors in the final layers, as has been observed by Ethayarajh [23], especially in GPT2. In this case, the word vectors assemble in a narrow cone rather than being uniform in all directions. We find a similar phenomenon in XLNet representations and hypothesize that it is due to the auto-regressive language modeling objective.

Both token and context are non-negligible for CWE: Table III shows the results of the masked token. Compared with the unmasked representation, the similarities drop consistently, showing the importance of the token feature for the final CWE. However, the difference is not significant, ranging from $-0.05$ to $-0.02$. The results of $\text{Sim}_{\text{mask}}^{\text{as}}$ are still larger than the random baseline, showing that the sentential information is also utilized in models.

Probing results are consistent with cosine similarity: Table IV shows the results of the probing task. We find that the performance of using a linear classifier is around 60% accuracy, which shows that a simple linear transformation can not fully extract the sense-level features. The results are improved much when using MLP for ELMo, BERT, and its variants, reaching 80% or more, showing that such features are indeed highly encoded. As for XLNet and GPT2, both results for linear and MLP transformation are extremely low, where CWE distribution is extremely non-uniform as in the aforementioned discussion. Overall, the performance is highly related to the cosine similarity based $\Delta_{\text{as}, \text{rand}}$ in Table II, showing that the similarity gap can reflect the capability of encoding sense level features. For brevity, we use the cosine similarity metric for further analysis in the remainder of this paper.

### Experimental Results

Tables II and III show the results of cosine similarity using the CWEs from the last layer. Overall, both $\text{Sim}_{\text{as}}$ and $\text{Sim}_{\text{ds}}$ are relatively large compared with $\text{Sim}_{\text{rand}}$. Showing that CWE keeps a high level of sense-wise similarity across contexts, which can be a useful characteristic. The values of $\text{Sim}_{\text{as}}$ are higher than those of $\text{Sim}_{\text{ds}}$, which reflects the fact the same word is more likely semantically equivalent compared with different words.

Table IV shows the results of probing task, showing that CWE can also capture semantic features through simple linear or MLP transformation.

### Comparison Between CWE Models

Different models show different similarities: In Table II, the models show large variations in vector similarities. ELMo gives the lowest $\text{Sim}_{\text{as}}$, $\text{Sim}_{\text{ds}}$, and $\text{Sim}_{\text{rand}}$ of 0.45, 0.29 and 0.17, respectively. In contrast, Transformer-based models give higher scores. The reasons can be that: 1) The dimension of embedding space (1024 vs. 768) is different; 2) Transformer-based models use an attention mechanism rather than recurrent state transition for information exchange, making the token interaction direct and the output embeddings more similar.

### Influence of Model Layers

Besides the final layer that is considered to encode more semantic-related features, we draw the layer-wise results in Figs. 1 and 2, respectively.

Features are encoded hierarchically in models: In both figures, sense-wise similarity varies according to different layers. As seen in Fig. 1, $\text{Sim}_{\text{as}}$ in the bottom layers are relatively large compared with that in the middle layer, while the values are increasing in the higher layers (except for a slight drop in the final layers of BERT, SenseBERT, and XLNet). These are related to recent findings about knowledge learned in CWE. For example, Belinkov et al. [50] find that morphology is learned at the lower layer of CWE, the large $\text{Sim}_{\text{as}}$ come from the same word
spellings. For the middle layers, BERT learns more non-local linguistic knowledge such as syntax [26], making the sense-wise similarity lower. For higher layers, the word representation is highly contextualized [23] and encodes more semantic knowledge [12], thus the sense-level similarity increases.

**Deep layers are required to extract contextual meaning:** As shown in Fig. 2, for the different words in the same sense, almost all the models have relatively low similarity scores (0.3 ∼ 0.7) in the bottom layers, where the embeddings are not highly contextualized. The $Sim_{ss}$ values generally increase as the layer moves higher, and reaches high values (0.6 ∼ 0.9) in the top layer representations, demonstrating that CWE encodes context information and recognizes the shared word sense.

**Different models show different trends,** which may be because of the different training settings such as training corpus and objective. The results are relatively stable for RoBERTa, which could due to 1) RoBERTa uses the dynamic masking strategy instead of the fixed words masking, which can make the training signals more uniform towards different words; 2) RoBERTa does not use the next sentence prediction objective, which focused more on token-level representations. On the other hand, XLNet permutes the word orders during pre-training and is optimized through the permutation language modeling objectives, which is highly different from the causal language modeling and masked language modeling, thus the curves change drastically from bottom to top layers. As for the other models, the differences between the layers are relatively smooth.

**C. Case Study**

We select BERT and show the word/senses (appearing at least five times) that contribute the most and least to $Sim_{ss}$ in Table V. **First**, three words “Microscopically”, “Petitioner”, and “Second” give the highest similarity. They have a relatively stable meaning. In contrast, the words “pictorial”, “bear”, and “neurotic” are the lowest-ranked, which are more polysemous. As in the dataset we used, the number of senses of words “Microscopically”, “Petitioner”, and “Second” are 1, 1, 2, 2, 12 and 3, respectively. **Second**, contexts for the latter three words are generally longer and more complex, such contextualization may have an effect on the word representations. **Third**, interestingly, the three words with the largest $Sim_{ss}$ appear in the sentence start position, yet the three with the lowest $Sim_{ss}$ appear in different and irregular positions of contexts.

The above case suggests that both intrinsic and extrinsic characteristics have an influence on the cross-context stability of word embeddings. We analyze the influence of different types of words (Section V-A) and contexts (Section V-B) accordingly. We also investigate if there is a position bias for high similarity for the first word, or if it just happens by chance (Section V-C).

**V. Analysis and Discussion**

For word attribution, we investigate the influence of part-of-speech and the number of senses, given that (1) word senses are
related to the PoS by definition; (2) the number of senses decides the polysemous degree of a word. With regard to contextual-level influences, we are interested in the sentence length and the word position in a sentence, for both a semantic variation is undesired.

A. Influence of Word Type

The Part-of-Speech: Similar to (3), we compute the cosine similarity according to different PoS categories:

$$\text{Sim}_{\text{ss}}[\text{PoS}] = \frac{\sum_{i,j} \cos(t_i, t_j)}{\text{Count}(t_i, t_j)}$$

where both $s_i$ and $s_j$ belong to one PoS among the noun, the verb, the adjective, or the adverb.

The results are shown in Fig. 3(a). For GPT2 and XLNet, the values are small and the difference between PoS categories is small. The differences for the other five models are larger according to different PoS. In general, the similarity of nouns is the largest across all models, which shows that the CWE of nouns is more consistent.

The Number of Senses: We also compute the cosine similarity according to different numbers of word senses:

$$\text{Sim}_{\text{ss}}[\#\text{Senses}] = \frac{\sum_{i,j} \cos(t_i, t_j)}{\text{Count}(t_i, t_j)}$$

where the sense number of token $t$ ($\#\text{Senses}$) varies from 1 to 10 or more.

Fig. 3(b) shows the $\Delta_{\text{ss.rand}}$ according to different numbers of senses. For all models, the values are higher for words with fewer senses (1 and 2 ~ 5) and decrease sharply as the number of senses increases (6 ~ 9 and 10+). The similarity value is also related to the granularity of defined word senses. For those
common words with 10+ senses, such as the word “take” labeled with 58 senses in the dataset, some of the senses can be easily confused, leading to the difficulty for sense understanding via CWE.

Each sense belongs to one PoS by definition, and we calculate the average number of senses of a word against each PoS. The results for nouns, verbs, adjectives, and adverbs are 1.35, 1.83, 1.46, and 1.24, respectively. The number of senses is relatively small for nouns, which can explain its high consistency. Adverb has the lowest number of senses but also has the lowest word frequency in corpora [51], [52], [53]. This makes adverbs less pre-trained generally, thus the results are similar to verbs and adjectives.

### B. Influence of Context

#### The Sentence Length:
We compute the similarity according to different sentence length and position distances:

$$\text{Sim}_{\text{ss}}[\text{SentLen}] = \frac{\sum_{i,j} \text{Cos}(t_i, t_j)}{\text{Count}(t_i, t_j)}$$

where $t_i \in t_1..t_{N_i}$ and $t_j \in t_1..t_{N_j}$. $N_i, N_j$ are in the same length range.

Fig. 4(a) shows the values of $\Delta_{\text{ss,rand}}$ regarding different sentence lengths. The largest similarity arises between words appearing in relatively short sentences. As the sentence becomes longer, the similarity decreases and gradually stabilized. This shows that the CWE of words is influenced by the sentential content, the sense-level knowledge is easier recognized and more distinguishable in short sentences. However, the CWE becomes more changeable when being more contextualized, and the need for long-term position embeddings could also be one of the reasons that influence the final representations (see discussions in the following subsections), thus the sense-wise consistency is prone to get lost.

#### The Relative Distance:
The different positions may also influence the sense-wise variances, we compute the similarity according to the distance between two words:

$$\text{Sim}_{\text{ss}}[\text{Dist}] = \frac{\sum_{i,j} \text{Cos}(t_i, t_j)}{\text{Count}(t_i, t_j)}$$

where the relative distance between $t_i$ and $t_j$ (i.e., $\text{Dist} = |i - j|$) is in the same distance range.

The results for word pairs in different position distances are shown in Fig. 4(b). Overall, for ELMo, word pairs in the same position (i.e., the relative distance is zero) give larger similarities. As the distance becomes longer, the similarity decreases. This may be because words with similar positions tend to share more similar contexts, thus the semantic relationship tends to be close. In contrast, a larger distance means different contextualized information, making the variation of CWE. BERT, SenseBERT, and RoBERTa use absolute position embeddings [41], [54], where word embeddings in the same positions are concatenated with the same position embeddings as inputs, making the representations more similar. However, DeBERTa uses relative position embeddings during the attention interaction, showing little similarity variance against relative distance.

### C. Position Bias for the First Word

According to the findings in Table V and Fig. 4(b), we find that there exists a position bias for the first word representation, where the similarity is irregularly higher than other positions\(^2\). The mentioned position bias in this work refers to the task-agnostic representation of the first word in a sentence.

#### Part-of-Speech vs. Position:

We first check the distribution relationship between part-of-speech and position. Although the similarity for nouns is larger than verbs, as shown in Table VI, the total numbers are similar (36.0% vs. 36.7%) and there is

\(^2\)Ma et al. [55] and Amor et al. [56] also mentioned the position bias in neural models, they focus on the specific tasks like sentiment classification and tagging tasks, respectively. The mentioned position bias in this work refers to the task-agnostic representation of the first word in a sentence.
no obvious correlation between part-of-speech and positions, the distribution for the first words are normal as other words. This shows that the position bias is not due to the imbalance distribution of part-of-speech in each position.

Similarity vs. Position: We set the relative position distance as zero and measure the sense similarity at different position indices. The results are shown in Fig. 5. Intuitively, the values of word similarity should be position-insensitive. However, we find that, except for XLNet and GPT2, the first few positions give the largest similarity. The values go down as the position moves forward, reaching a relatively stable value soon. The reason can include: First, the CWE of the sentence beginning word from ELMo gives the highest similarity because they share the same information before the first token (conveyed by hidden states $h_{<s>}$ of the sentence beginning symbol $<s>$), which serves half hidden states (together with the context information in the backward direction from the second word $h_{<2>}$) for the generation of the final vector. Second, for BERT-related models, the first token shows higher similarity, which may be because the attention mechanism inside Transformers can encode certain patterns. For example, the words tend to focus on the [CLS] token directly or the previous tokens in certain attention heads [57]. Further, words can have strong local dependencies [58], which makes the CWE be position biased, especially for the sentence beginning words that share the [CLS] token nearby.

Verification by Prompt Augmentation: To further explore the influence of position, we verify it by adding three designed prompts below for each sentence [X]. In this way, the word attribution (e.g., sense and part-of-speech) in the original sentence keeps steady, but the position for each word will shift, and the local context for the sentence beginning word is enriched, relying less on the [CLS] token. These prompts are:
- Prompt1, “[X].”
- Prompt2, She said : [X].
- Prompt3, Document : [X].

Table VII shows the results. By using prompt augmentation inputs, the similarity scores drop for the first word in the original sentence. For the other positions, the scores remain relatively stable with variations ranging from 0.01 $\sim$ 0.03. This shows that the representations of the sentence beginning word are biased and context-sensitive, leading to the higher scores in Fig. 5 compared with other positions. We also illustrate an example of such an interesting phenomenon in Fig. 6.

VI. POSITION BIAS MITIGATION FOR DISTANCE-BASED WORD SENSE DISAMBIGUATION

In the above discussion, we show that the word representation of CWE is position biased, where the CWE of the first words in different contexts tends to be much closer than in other positions. In this section, we show that such bias may influence the performance of a certain downstream task and propose a simple way to mitigate it.

A. Task Definition and Datasets

In particular, we choose the 2-way word sense disambiguation task using the Word-in-Context (WiC) dataset [59], which is also a sub-task included in the SuperGLUE benchmark [60]. Given two sentences and each containing a target word, the task is to determine whether these two target words have the same contextualized meaning.

B. Methods

Since pre-trained models such as BERT have shown the capability for encoding semantics, we proposed a distance-based method by using BERT as our baseline. Formally, given two target words $t^1_i, t^2_j$ and two contexts $s^1 = t^1_1, \ldots, t^1_i, \ldots, t^1_M, s^2 = t^2_1, \ldots, t^2_j, \ldots, t^2_N$, we assign the label to 1 (has the same
meaning) or 0 (has a different meaning) according to their cosine
distance between two CWEs:

$$
\text{Label}(t^1_i, t^2_j) = 1[\text{Cos}(t^1_i, t^2_j) > T]
$$

(11)

where $1[\cdot]$ is the indicator function, $T$ is the best pre-specified threshold according to datasets.

The position bias shows that the $\text{Cos}(t^1_i, t^2_j)$ could be higher when $i = j = 1$ for all word types $t$, thus the threshold $T$ could be higher than words in other positions. According to the experiments in the previous section, we can add some prompt (prefix or suffix) to the original text without losing too much semantics, and make the positions for each token shift for a distance. We illustrate our method in Fig. 6.

C. Results

The overall results are shown in Table VIII**The Best Results are in Bold.. Across all models, the accuracy increases from lower layers to higher layers, showing again that higher layers encode more semantic-related information. The method using original text gives 66.9 accuracy in the last layer. Adding some prefixes (and suffixes) can improve the overall accuracy by 0.8 ∼ 1.4.

We also check the fine-grained results for different positions. As Table IX shows, the accuracy for the first words are quite low (with only 53.5 by using the original text), showing that the CWEs for the first words have a low degree of distinction. As a comparison, the accuracy for the other positions is relatively high (about 67.9). By adding the prefix, the accuracy for the original first words gets a large improvement (+5.1 ∼ +8.4), which shows the effectiveness of our method.

TABLE VIII
RESULTS OF USING BERT FOR DISTANCE-BASED WSD BY USING PROMPT-AUGMENTED INPUTS

| Inputs         | Layer 1 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 | Layer 7 | Layer 8 | Layer 9 | Layer 10 | Layer 11 | Layer 12 |
|----------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|-----------|-----------|-----------|
| original text  | 63.5    | 64.5    | 65.3    | 65.7    | 65.9    | 65.8    | 65.5    | 65.4    | 65.6    | 65.3      | 64.5      | 66.9      |
| w/ prefix = "", suffix = "" | 63.0    | 63.9    | 64.7    | 65.2    | 65.4    | 65.3    | 65.1    | 65.2    | 65.6    | 65.7      | 64.9      | 67.7 (+0.8) |
| w/ prefix = She said : | 63.2    | 64.2    | 65.1    | 65.7    | 66.0    | 66.9    | 65.8    | 65.7    | 66.1    | 66.2      | 65.4      | 68.1 (+1.2) |
| w/ prefix = Document : | 62.9    | 63.8    | 64.7    | 65.1    | 65.2    | 65.1    | 64.8    | 64.8    | 65.0    | 65.2      | 64.8      | 68.3 (+1.4) |

The best results are in bold.

TABLE IX
RESULTS OF USING BERT (LAST LAYER) FOR DISTANCE-BASED WSD BETWEEN $t^1_i$ AND $t^2_j$ BY USING PROMPT-AUGMENTED INPUTS

| Inputs         | Positions $i = j = 1$ | others         |
|----------------|-----------------------|----------------|
| original text  | 53.5                  | 67.9           |
| w/ prefix = "", suffix = "" | 61.9 (+8.4) | 68.1 (+0.2) |
| w/ prefix = She said : | 59.2 (+5.7) | 68.7 (+0.8) |
| w/ prefix = Document : | 58.6 (+5.1) | 68.9 (+1.0) |

VII. THE INFLUENCE TO DOWNSTREAM TASKS

In the aforementioned study, we discuss how sense-level similarity varies across different models, layers, contexts, words, and positions by using the pre-trained models, which reflect the main characteristics of pre-trained CWE. In this section, we further investigate how does the sense-level representations influence the downstream tasks by fine-tuning models. We mainly compare SenseBERT and BERT since they have similar model architecture and pre-training settings (e.g., corpus and training consumption). During fine-tuning, we set the same training parameters for both models.

A. Tasks and Datasets

We use four types of tasks, which have been widely studied in the NLP domain.

Sense-Level Tasks: SemEval WSD [61] is used for word sense disambiguation and WiC [59] is used for Word-in-Context.
Syntactic Tasks: CoLA [62] is used for testing linguistic acceptability. The Wall Street Journal portion of the Penn Treebank [63] is used for constituency tagging.

Sense-Level Semantic Tasks: We use the GLUE [64] benchmark for general semantic understanding (except CoLA), including QNLI, MNLI, RTE for question pairs matching, SST-2 for sentiment analysis, STS-B for textual similarity.

Entity Tagging: CoNLL2003 dataset [65] is used for named entity recognition (NER).

B. Results

Fig. 7 shows the results for each task. We can find that SenseBERT shows large improvements over BERT on sense-level tasks, which is not surprising because sense-level multi-task is used during pre-training, thus SensBERT can learn more fine-grained sense-level knowledge. The performance gap remains significant for syntactic tasks such as CoLA and parsing, which is because the sense-level information also contains some syntactic knowledge (e.g., part-of-speech) and is useful for judging linguistic acceptability and constituency tagging. As for sentence-level semantic tasks and entity tagging, the performance is close, and BERT can outperform SenseBERT on tasks including RTE, QQP, SST-2, and STS-B. These results show that the external token-level sense prediction objective does not bring benefits for these general semantic tasks.

The above findings show that SenseBERT (or sense-based pre-training) is more suitable for downstream sense-level or syntactic-related tasks. As for more general language understanding tasks, traditional language modeling based pre-trained models (e.g., BERT) can already achieve satisfactory results.

VIII. CONCLUSION

In this paper, we investigated the sense-wise representations of contextualized word embeddings from seven pre-trained models. Results show that models can capture sense consistency well. However, influence from both intrinsic and extrinsic factors such as position bias also exists. Based on these findings, we further analyze such sense-level representation beyond the pre-trained models, including the influence of position bias for unsupervised word sense disambiguation and the general performance by fine-tuning models across different downstream tasks. Our work reveals some characteristics as well as limitations of pre-trained language models and we hope it can help people better understand or improve language models from a sense-level representation perspective.

ACKNOWLEDGMENT

The authors thank all the anonymous reviewers for their detailed and constructive suggestions.

REFERENCES

[1] M. E. Peters et al., “Deep contextualized word representations,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202
[2] A. Radford et al., “Language models are unsupervised multitask learners,” OpenAI blog, vol. 1, no. 8, p. 9, 2019.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
[4] Y. Liu et al., “RoBERTa: A robustly optimized bert pretraining approach,” 2019, arXiv:1907.11692.
[5] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding,” in Proc. Adv. Neural Inf. Process. Syst., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/dcfa7e7655d75840eb66733e9b6e7ccc69-Paper.pdf
[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations,” in Proc. Int. Conf. Learn. Representations, 2020.
[7] P. He, X. Liu, J. Gao, and W. Chen, “Deberta: Decoding-enhanced bert with disentangled attention,” in Proc. Int. Conf. Learn. Representations, 2021.
[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. 1st Int. Conf. Learn. Representations, 2013.
[9] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162
[10] J. Hewitt and C. D. Manning, “A structural probe for finding syntax in word representations,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2019, pp. 4129–4138. [Online]. Available: https://aclanthology.org/N19-1419
[11] Z. Wu, Y. Chen, B. Kao, and Q. Liu, “Perturbed masking: Parameter-free probing for analyzing and interpreting BERT,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4166–4176. [Online]. Available: https://aclanthology.org/2020.acl-main.383
[51] R. H. Shih, J. Y. Chiang, and F. Tien, “Part-of-speech sequences and distribution in a learner corpus of English,” in Proc. 13th Res. Comput. Linguistics Conf., 2000, pp. 171–177. [Online]. Available: https://aclanthology.org/O00-1009

[52] C. Lu et al., “Examining scientific writing styles from the perspective of linguistic complexity,” J. Assoc. Inf. Sci. Technol., vol. 70, no. 5, pp. 462–475, 2019.

[53] A. Peti-Stantic et al., “The Croatian psycholinguistic database: Estimates for 6000 nouns, verbs, adjectives and adverbs,” Behav. Res. Methods, vol. 53, 2021.

[54] P. Dufter, M. Schmitt, and H. Schütze, “Position information in transformers: An overview,” Comput. Linguistics, vol. 48, no. 3, pp. 733–763, Sep. 2022. [Online]. Available: https://aclanthology.org/2022.cl-3.7

[55] F. Ma, C. Zhang, and D. Song, “Exploiting position bias for robust aspect sentiment classification,” in Proc. Findings Assoc. Comput. Linguistics, 2021, pp. 1352–1358. [Online]. Available: https://aclanthology.org/2021.findings-acl.116

[56] M. B. Amor, M. Granitzer, and J. Mitrović, “Technical report on Token Position Bias in Transformers,” 2023, arXiv:2304.13567.

[57] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does BERT look at? an analysis of BERT’s attention,” in Proc. ACL Workshop BlackboxNLP: Anal. Interpreting Neural Netw., 2019, pp. 276–286. [Online]. Available: https://aclanthology.org/W19-4828

[58] L. Gong, D. He, Z. Li, T. Qin, L. Wang, and T. Liu, “Efficient training of BERT by progressively stacking,” in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 2337–2346. [Online]. Available: https://proceedings.mlr.press/v97/gong19a.html

[59] M. T. Pilehvar and J. Camacho-Collados, “WiC: The word-in-context dataset for evaluating context-sensitive meaning representations,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2019, pp. 1267–1273. [Online]. Available: https://aclanthology.org/N19-1128

[60] A. Wang et al., “Superglue: A stickier benchmark for general-purpose language understanding systems,” in Proc. Adv. Neural Inf. Process. Syst., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf2f24afe7fabf6046bf4923da1d6-Paper.pdf

[61] A. Raganato, J. Camacho-Collados, and R. Navigli, “Word sense disambiguation: A unified evaluation framework and empirical comparison,” in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics: Volume 1, Long Papers, 2017, pp. 99–110. [Online]. Available: https://aclanthology.org/E17-1010

[62] A. Warstadt, A. Singh, and S. R. Bowman, “Neural network acceptability judgments,” Trans. Assoc. Comput. Linguistics, vol. 7, pp. 625–641, 2019. [Online]. Available: https://aclanthology.org/Q19-1040

[63] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a large annotated corpus of English: The Penn treebank,” Comput. Linguistics, vol. 19, no. 2, pp. 313–330, 1993. [Online]. Available: https://aclanthology.org/93-2004

[64] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, “GLUE: A multi-task benchmark and analysis platform for natural language understanding,” in Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw., 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446

[65] E. F. T. Kim Sang and F. D. Meulder, “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,” in Proc. 7th Conf. Natural Lang. Learn. IHLT-AAACL, 2003, pp. 142–147. [Online]. Available: https://aclanthology.org/W03-0419