Coming to its senses: Lessons learned from Approximating Retrofitted BERT representations for Word Sense information

Anonymous ACL submission

Abstract

Retrofitting static vector space word representations using external knowledge bases has yielded substantial improvements in their lexical-semantic capacities but is non-trivial to apply to contextual word embeddings (CWE). In this paper, we propose MAKESENSE, a method that 'approximates' retrofitting in CWEs to better infer word sense knowledge from word contexts. We specifically analyze BERT and MAKESENSE-transformed BERT representations over a diverse set of experiments encompassing sense-sensitive similarities, alignment with human-elicited similarity judgments, and probing tasks focusing on sense distinctions and hypernymy. Our findings indicate that MAKESENSE imparts substantial improvements in word sense information over vanilla CWEs but largely preserves more complex usage of sense and directionally sensitive information such as hypernymy.

1 Introduction

Word sense disambiguation (WSD) is a fundamental component for language understanding (Navigli, 2009). Humans readily show this capacity by inferring word meanings from their linguistic contexts (Klein and Murphy, 2001). Recently proposed pre-trained language models (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019) represent words as a function of their sentence/paragraph contexts, producing contextualized word embeddings (CWE) that overcome the *'meaning conflation deficiency'* (Camacho-Collados and Pilehvar, 2018) of static vector space models such as word2vec (Mikolov et al., 2013). Perhaps unsurprisingly, CWEs have a clear edge in empirical performance on a range of sense-disambiguation tasks (Raganato et al., 2017; Pilehvar and Camacho-Collados, 2019; Reif et al., 2019), highlighting their relative potential as models of polysemy (Nair et al., 2020).

Incorporating external knowledge sources (Loureiro and Jorge, 2019) has further enhanced the WSD capacities of CWEs, opening up new avenues to engage in combining statistical and symbolic paradigms. An alternate route of incorporating knowledge into distributional representations of words is *retrofitting*. This paradigm operates on the enhancement of the distributional vector geometry by injecting linguistic constraints (Faruqui et al., 2015; Mrkšić et al., 2016; Lengerich et al., 2018), improving alignment with word-relatedness measures (Faruqui et al., 2015) as well as downstream tasks (Mrkšić et al., 2016). While extensively applied to static word representations, retrofitting has been rather under-explored in the context of CWEs. We speculate that this is largely due to CWEs of words being sensitive to the contexts they appear in, making the formulation of the geometrical transformations intractable due to the vastness of the range of possible contexts in which a word can occur. The one approach that does retrofit CWEs explicitly for sense-information (Bihani and Rayz, 2021) does it on a static inventory of contexts, and as such cannot be applied to instances of words in context disjoint from its training data, making it non-trivial for researchers to test its effectiveness.

In this paper, we revisit retrofitting by proposing MAKESENSE, a method that 'approximates' CWEs specialized for word sense information, and is applicable to any polysemous or homonymous word in context, thereby generalizing sense retrofitting to unseen instances. As a case study, we apply this approach to $\text{BERT}_{\text{base}}$ (Devlin et al., 2019).[1] We then take steps to clarify the sense-sensitive properties our method imparts on the BERT representational space by testing it on sense-similarity measures from discrete and graded human-elicited judgments (Erk et al., 2013). We then turn to *probing* literature (Ettinger et al., 2016a; Adi et al., 2017) and establish the extent to which MAKESENSE makes information about word-senses more readily acces-

[1] Our methods and analyses can be applied to any CWE model. We make our code available at `url-anonymized`.

sible during supervised classification. Finally, we investigate patterns of sense-sensitive *hypernymy* in MAKESENSE representations. Our experiments explicitly compare MAKESENSE against $\text{BERT}_{\text{base}}$, and are conducted in a layerwise fashion, allowing us to shed light on how-much sense-information is already present as a result of pre-training, how it evolves within the model, and whether our approximation approach enhances it.

2 Related Work

Our contributions build upon two different strands of research that focus on computational lexical semantics. The first strand of research investigates manifestation of word-sense representations within CWEs. Most studies carry out such investigations by using standard WSD benchmarks and quantify knowledge of senses based on a 1-NN (nearest neighbor) classifier built on top of popular CWEs such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), resulting in state-of-the-art performance at the time. This indicates favorable competence of CWEs in retaining sense-information as a result of pre-training, and suggests that they form representations that are largely similar for words carrying similar meaning in context. Layerwise investigations by Reif et al. (2019); Loureiro et al. (2021) suggest that deeper layers align better with sense-disambiguation information while shallow layers are closer to words' static representations and perform worse on WSD. Interestingly, smaller models ($\text{BERT}_{\text{base}}$) tend to out-perform larger ones ($\text{BERT}_{\text{large}}$) (Pilehvar and Camacho-Collados, 2019). Our analysis methods stray away from WSD benchmarks due to complete data overlap with their standard splits (see §3), we instead focus on a diverse set of tasks requiring crucial access to the sense-disambiguation signal within the representations — e.g. differentiating between same and different senses of a word, and predicting whether pairs of words in contexts have a hypernymy relation. While the latter has been recently analyzed by Ravichander et al. (2020), they only consider words with single-senses.

A large body of work focuses on augmenting BERT's pre-existing WSD capacities by incorporating external knowledge by altering its training objective (Peters et al., 2019), defining an auxiliary task (Bevilacqua and Navigli, 2020; Levine et al., 2020), leveraging gloss knowledge (Loureiro and Jorge, 2019; Blevins and Zettlemoyer, 2020; Huang et al., 2019) or diversifying contexts using knowledge-enhanced corpora (Scarlini et al., 2020a,b). We complement these findings using a different mechanism of knowledge incorporation in CWEs, which we describe next.

The second strand of research focuses on retrofitting approaches. Retrofitting was first proposed by Faruqui et al. (2015) as a graph based post-processing technique that could specialize any word embedding space, acting as an alternative to model training-dependent semantic specialization (Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014). Recent works have extended this approach to include a variety of linguistic entities such as paraphrases (Wieting et al., 2015) and word senses (Jauhar et al., 2015; Ettinger et al., 2016b), as well as lexical relations such as antonymy (Mrkšić et al., 2016), lexical entailment (Vulić and Mrkšić, 2018) and other functional relations (Lengerich et al., 2018). Joint retrofitting models have also been proposed to learn semantic specialization from cross lingual resources and are beneficial for low-resource language representation learning (Mrkšić et al., 2017). Since retrofitting methods are limited to entities seen in corpora, recent works on *post-specialization* have focused on extending the specialization learnt during retrofitting to unseen lexical instances (Glavaš and Vulić, 2018; Vulić et al., 2018), which we build upon here, for CWEs.

2.1 Retrofitting CWEs using LASeR

LASeR (Bihani and Rayz, 2021) is a sense retrofitting method that aims to encode sense information into CWEs. LASeR utilizes sense annotated corpora to modify any given vector space by injecting sense information within word vectors, while minimizing anisotropy, the tendency for vector spaces to occupy a narrow cone, resulting in inflated vector similarities (Ethayarajh, 2019). LASeR performs anisotropy reduction by removing the top common direction(s) within the vector space, making it uniformly distributed. It further extends the retrofitting update developed by Faruqui et al. over word senses, such that vector representations of same word senses are shifted closer together while retaining the distributional properties learnt during pretraining. LASeR is trained on multi-sense nouns, verbs, and adjectives from five sense-annotated resources from various SemEval and SensEval tasks, concatenated under a unified WSD framework by Raganato et al.

(2017). Although LASeR-enhanced CWEs empirically show greater sensitivity to sense-information, their generation critically depends on the existence of ground-truth sense information, which is unrealistic when encountering words embedded in sentence contexts that have not been seen during the retrofitting step. This facet of the method restricts its testing to only intrinsic analyses (see Bihani and Rayz, 2021) and prevents testing on standard WSD benchmarks due to complete data-overlap, or supervised sense-sensitive tasks thereby casting doubts about its effectiveness in NLP applications.

3 Method: Approximating LASeR

To circumvent the aforementioned issues, we propose to instead "approximate" sense-enriched CWEs from vanilla CWEs in a supervised-learning setup. Specifically, given d-dimensional CWE representations $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, and their corresponding sense-enriched LASeR representations $\mathbf{X}_s = \{\mathbf{x}_{s,1}, \dots, \mathbf{x}_{s,n}\}$, we propose to learn an approximation function $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, that maps each $\mathbf{x}_i$ to $\mathbf{x}_{s,i}$ by minimizing a regression-based loss. Approximating LASeR embeddings allows researchers to better test the benefits of inducing sense information through retrofitting — i.e., one can simply use the learned function f on word representations that are disjoint from the vocabulary that LASeR was trained for and then probe the resultant vectors for sense-information. Figure 1 illustrates our entire approximation method.

Model Investigated We perform our experiments on 768-dimensional embeddings extracted from $\text{BERT}_{\text{base}}$ (Devlin et al., 2019). We use BERT as our CWE model due to precedence in earlier research investigating word-sense information in CWEs produced by pre-trained LMs (see §2). Furthermore, this lets us narrow in on deeper analyses — e.g., investigating layerwise effects. However, our methods are agnostic to any model that encodes words in context and therefore can be extended to any distributional CWE models.

Data We first expand the coverage of our sense-enriched representations by combining the original LASeR corpus with a subset of SemCor (Miller et al., 1993) consisting only of single-word nouns, verbs, and adjectives. This is a considerable update as it results in 181,768 total instances, comprising of 16,528 unique words and 16,751 unique senses, embedded in 36,360 unique sentences. By

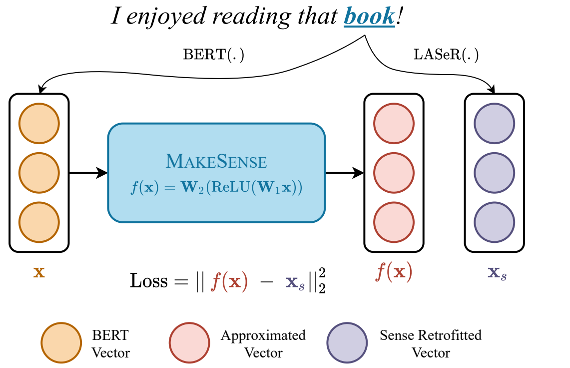


Figure 1: Illustration of the MAKESENSE approximation method. In practice, BERT(.) can be replaced by any CWE, provided one has access to the LASeR embeddings corresponding to the desired CWE.

contrast, Bihani and Rayz have 2,416 instances, 426 unique senses, 918 unique words, and 966 unique sentences. We apply LASeR[2] on the $\text{BERT}_{\text{base}}$ representations ($\mathbf{x}_i$) of target-words extracted from our augmented sense-annotated data, yielding a sense-enriched vector space $\mathbf{X}_s^l$ for each layer (l) in the $\text{BERT}_{\text{base}}$ model, amounting to 13 distinct $\mathbf{X}_s$ spaces.[3] For each layer, we lexically split the resultant set of tuples $D^l = \{(\mathbf{x}_1^l, \mathbf{x}_{s,1}^l), \dots, (\mathbf{x}_n^l, \mathbf{x}_{s,n}^l)\}$ into our experimental datasets: D^l_{train} (80%), D^l_{dev} (10%), and D^l_{test} (10%), such that their vocabularies are disjoint from one another. Our lexical-split strategy allows for a more robust model training procedure to generalize LASeR approximation as opposed to simply memorize it due to word-identity information leaks (Levy et al., 2015).

Approximation function construction Following Vulić et al. (2018), who propose *post-specialization* of retrofitted static word embeddings, we assume non-linear mappings to be a better hypothesis of how retrofitted sense information can be estimated from CWEs — owing to the fact that retrofitting injects several constraints to the vector space, making it limiting for a linear map to successfully approximate it. Therefore, we formulate our approximation function as a multi-layer perceptron, i.e., $f(\mathbf{x}_i) = \text{MLP}(\mathbf{x}_i)$. We use the standard L_2 loss between the sense-enriched embedding $\mathbf{x}_{s,i}$ and the approximated embedding $f(\mathbf{x}_i)$:

$$\mathcal{L}_m(\mathbf{x}, \mathbf{x}_s) = ||f_m(\mathbf{x}) - \mathbf{x}_s||_2^2 \quad (1)$$

[2] we use the publicly released code: https://github.com/bihani-g/LASeR

[3] 12 transformer layers and one '0-th' layer that serves as input to the first transformer layer.

We experiment with composing $h \in \{1, \ldots, 5\}$ different hidden layers, with sizes $d_h \in \{512, 1024, 2048\}$. Each layer is passed through a ReLU activation and a dropout function ($p = 0.5$). We find the best hyperparameter configuration by training multiple models on the training set (D_{train}), and choose our final model as the one that achieves the minimum average loss on the development set (D_{dev}). Henceforth, we refer our best model as MAKESENSE.

Training Details We use the Adam optimizer (Kingma and Ba, 2015) with regularization (with a weight decay of 0.001) to train all of our approximation functions. For each training regimen, the best initial learning rate for the optimizer is chosen from the space: $\{0.001, 0.0001, 0.0003\}$. Our models are trained for a maximum of 40 epochs, with a batch size of 128. For each run, we halt the training process if the loss on the development set does not reach a new minimum for five consecutive epochs. With our various parameter configurations, we train 585 different approximation functions (3 learning-rate values $\times$ 3 hidden layer sizes $\times$ 5 hidden layers $\times$ 13 distinct BERT layers). Interestingly, all of our final 13 MAKESENSE models converge to the exact same configuration: two hidden layers of size 2048 each, and an initial learning rate of 0.0001. Representations from MAKESENSE show substantial improvements over $\text{BERT}_{\text{base}}$ representations in vector space isotropy (see appendix A).

4 Does MAKESENSE make sense?

We now conduct a range of tests targeting various sense-sensitive properties that our proposed MAKESENSE method imparts to the original CWE ($\text{BERT}_{\text{base}}$). Our analyses crucially require access to sense information and serve as a holistic benchmark environment where success of a model is quantified by various metrics that allow for robust comparison and conclusions regarding the representational quality produced by performing MAKESENSE. Data used in each analysis are disjoint from those used in our approximation experiments, contributing further to the robustness of our tests.

4.1 Investigating word sense information through representation similarity

Recent work in CWE-based WSD (see §2) suggests that computational models/agents that are sensitive to word sense information should likely produce representations that are similar for surface-forms of words with same as opposed to different senses. We gauge word-sense sensitivity in MAKESENSE and the original $\text{BERT}_{\text{base}}$ embeddings by comparing their representational similarities for words used in similar versus different senses. Let, $S = \{(s_1^1, s_1^2), \ldots, (s_n^1, s_n^2)\}$ be contextual embedding space of pairs of words with the same sense and $D = \{(d_1^1, d_1^2), \ldots, (d_m^1, d_m^2)\}$ be the embedding space of pairs of words with different senses.[4] To assess sensitivity to word sense information, we utilize the Δ metric, calculated as the difference of average cosine similarity between same-sense word instances and that of different-sense word instances:

$$\Delta = \frac{1}{n}\sum_{i=1}^{n}\cos(s_i^1, s_i^2) - \frac{1}{m}\sum_{j=1}^{m}\cos(d_j^1, d_j^2). \quad (2)$$

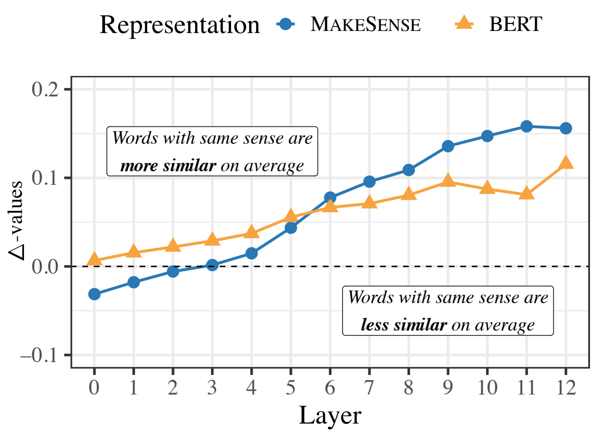


Figure 2: Δ values computed per-layer for $\text{BERT}_{\text{base}}$ and MAKESENSE representations, on the WIC corpus.

Thus, for a given word, if representations produced by MAKESENSE are on average more similar for surface-forms of the same sense and farther apart for its different senses, relative to the representations from the original model ($\Delta_{\text{MS}} > \Delta_{\text{BERT}}$), then we take this as evidence in favor of MAKESENSE in terms of the improvements it lends to the $\text{BERT}_{\text{base}}$ representations.

We rely on the WIC dataset (Pilehvar and Camacho-Collados, 2019) for this experiment. WIC consists of pairs of contexts with marked target words (e.g., row 1 of table 1), annotated for a discrete judgment of whether the surface-forms of the words carry the same sense. We use the concatenation of the training and development splits made publicly available by the authors.[5]

[4] Note that the surface-form of s_i^1 is the same as that of s_i^2, and that of d_j^1 is the same as that of d_j^2.

[5] The ground-truth data for the test split is part of an on-

Experiment	Stimulus Example	Outcome	Evaluation Metric(s)
WIC (§4.1 and §4.3)	(1a) *He designed a new* ***piece*** *of equipment.* (1b) *She bought a lovely* ***piece*** *of china.*	Same sense	Δ (similarity); Accuracy (probing)
	(1c) *Life has lost its* ***point****.* (1d) *He broke the* ***point*** *of his pencil.*	Different sense	
USIM (§4.2)	(2a) *No, we are not talking about the fortunes of a* ***rich*** *and powerful democracy.* (2b) ***Rich*** *people manage their money well.*	Avg. Human Similarity: 4.75	Spearman's ρ with human judgments
	(2c) *What are the important variables that create a* ***rich*** *online learning experience, ...* cont. (2d) ***Rich*** *people manage their money well.*	Avg. Human Similarity: 1.63	
WHIC (§4.4)	(3a) *Magnus Carlsen is the world* ***chess*** *champion.* (3b) *The championship* ***game*** *was played yesterday.*	Hypernymy	Weighted F_1 (overall); Directional-accuracy
	(3c) *He refused to give* ***titles*** *to his paintings.* (3d) *He had the* ***status*** *of a minor.*	No Hypernymy	

Table 1: Example of stimuli used in our analyses. **Note:** The outcome column represents the ground-truth label or value of the corresponding stimulus example. Dataset statistics and source URLs can be found in Appendix B.

Results and Analysis Figure 2 shows Δ-values for representations extracted at each layer of the $\text{BERT}_{\text{base}}$ model and their corresponding MAKESENSE representations. In general, we see greater Δ-values in deeper layers, suggesting that overall sensitivity to word-sense information largely increases as we move closer to the output of the BERT model. MAKESENSE substantially enhances this sensitivity in deeper layers with greater Δ-values compared to $\text{BERT}_{\text{base}}$. However, we see the opposite behavior in layers prior to layer 6, where the average similarity of surface-forms with the same sense is in fact not very different or even lower (starting at layer 3) than that of surface-forms with different senses. Since embeddings in layers closer to the input to BERT are more likely to retain information about word identity (Devlin et al., 2019), we speculate that this property makes earlier layers less susceptible to making distinctions between different usage of words in context, thereby producing low Δ-values. From this preliminary analysis, we predict that benefits of using MAKESENSE are more likely to be observed in deeper as opposed to shallow layers.

Takeaways MAKESENSE representations show greater sensitivity to sense-information compared to the original $\text{BERT}_{\text{base}}$ embeddings. However, this behavior is only local to deeper layers (layer 6 and above) and is reversed in shallow layers, suggesting that deeper layers may be more susceptible to improvements by MAKESENSE.

going competition and only allows limited access to 10 tries, which is insufficient for our experiments.

4.2 Correspondence with Graded Word Sense Similarity Judgments

Next, we turn to a setting that sheds a more nuanced light on inferring word meaning from context. This setting draws on theories of cognition advocating for 'fuzzy' concept boundaries (Zadeh, 1999; Rosch, 1973; Hampton, 2007), and casts relatedness in contextual word meaning as a graded measure (Kintsch, 2007). In table 1 for instance, *rich* in (2a) is more closely related to that in (2b) than it is to *rich* in (2c). We test the extent to which our representations are able to make word relatedness predictions consistent with this intuition. To this end we rely on the USIM dataset (Erk et al., 2013). USIM contains word meaning similarity annotations on pairs of instances of the same word appearing in different contexts. Each instance in the dataset presents a word lemma w in two contexts, where annotators judge graded similarity between their perceived word meanings on a scale of 1 (completely different) to 5 (same meaning). We compare MAKESENSE and $\text{BERT}_{\text{base}}$ based on their correspondence (measured using Spearman's ρ) with two measures: (1) USIM, the raw human-elicited similarity judgements reported by Erk et al.; and (2) UMID (McCarthy et al., 2016), the propor-

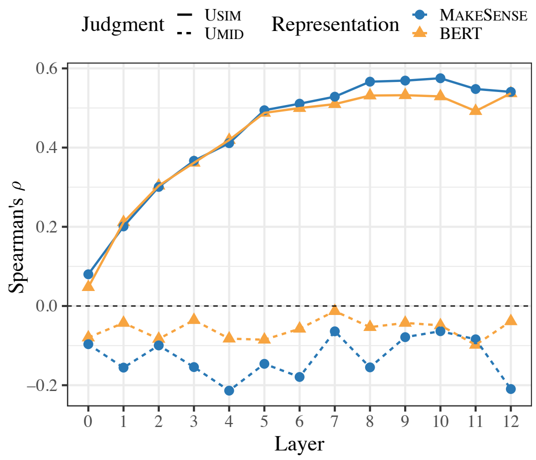


Figure 3: Spearman's ρ computed between representations' cosine relatedness and gold-standard metrics of graded sense-similarities: USIM and UMID.

tion of *mid-range* similarity judgments (between 2 and 4) on a word lemma, extracted from the USIM dataset. Correspondence with USIM denotes the alignment of the representational space with human intuitions about sense-similarity, while that with UMID reflects the uncertainty/disagreement regarding the perceived word meaning across different contexts. For a given word lemma, we expect models with enhanced sense-information to show greater positive correlation with USIM, suggesting better alignment with humans, and more negative correlation with UMID, indicating less uncertainty about word-sense similarity judgments.

Results and Analysis The correlation comparisons are plotted in Figure 3. We observe that MAKESENSE embeddings show greater correlation with USIM in deeper layers, as compared to $\text{BERT}_{\text{base}}$ embeddings, suggesting greater correspondence with overall human intuitions of sense-similarities. MAKESENSE representations also show greater negative correlation with UMID scores, especially in the middle layers. This suggests that MAKESENSE representations are better equipped to capture fine-grained gradedness in word sense similarity, i.e. they are more susceptible to distinguishing between moderately vs. highly similar instances relative to $\text{BERT}_{\text{base}}$ representations, which show more uncertainty in their similarity judgments. These findings agree with our prior observation (see §4.1) that MAKESENSE improves performance in the deeper model layers. We additionally observe that gradedness in sense similarities are better captured by MAKESENSE representations, especially in the middle layers and the final layer.

Takeaways In comparison to $\text{BERT}_{\text{base}}$, MAKESENSE representations not only encode more sense information, but also create vector spaces that show greater correspondence with gradedness in word sense similarity.

4.3 Probing for Binary Sense Judgments

We now turn to the body of work popularly known as *probing* (Ettinger et al., 2016a; Adi et al., 2017; Conneau et al., 2018) to further characterize the differences between MAKESENSE and $\text{BERT}_{\text{base}}$ in terms of word-sense information. The probing paradigm lets us explore the extent to which representations extracted from black-box models make a certain feature or property (linguistic or non-linguistic) readily accessible in a supervised setting. We hypothesize that representations that better encode sense-level information are also more conducive to successfully determining whether a given surface-form of a word carries the same meaning in a pair of minimally-overlapping sentence or phrasal contexts. Using our example from the first row of table 1, representations with better sense-level capacities should support the classification of *piece* in (1a) and (1b) as the same sense, while that of *point* in (1c) and (1d) as different.

We again rely on WIC as our experimental dataset, but instead cast our investigation as a binary classification setting, leveraging the annotated labels of "same-sense" and "different-sense" as our target labels. We follow Adi et al. (2017) and Hewitt and Liang (2019) and use a simple one-hidden-layer MLP as our probing classifier with 256 hidden-units, ReLU activation, and a sigmoid layer to generate the probability of the "same-sense" label. For each layer, we train our probe on 90% of the training split—we reserve 10% for validation—and test generalization performance using the final model's accuracy on the development set. A finer-grained description of our training details can be found in Appendix C.

Results and Analysis Figure 4 shows classification accuracies of the probe on the development set of the WIC dataset. Since WIC is balanced for its two class labels, chance performance on this task is 50%. We see that MAKESENSE elevates the probing accuracy of $\text{BERT}_{\text{base}}$ on this task across a majority of layers (all except layer 3), suggesting that the MAKESENSE method makes sense-information more accessible to the probe relative to the vanilla

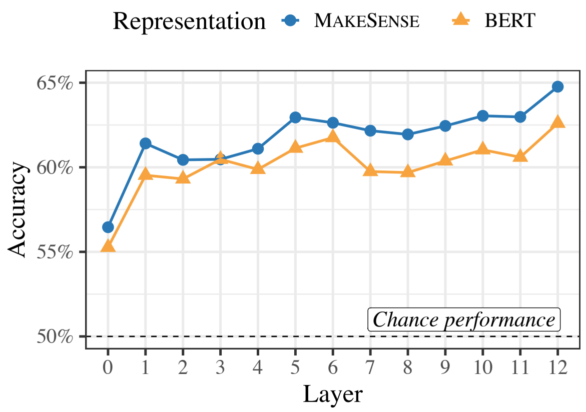


Figure 4: Probing accuracies on the WIC dataset.

$\text{BERT}_{\text{base}}$ model. However, it should be noted that the increase in the representational-capacity to binary classification is modest — with the maximum difference in performance being 2.41 percentage-points in layer 7. Nonetheless, all layers show above chance level performance, with the best accuracy being 65% for MAKESENSE at layer 12. Revisiting our hypothesis from §4.1, MAKESENSE shows its maximum benefit in deeper layers.

Takeaways $\text{BERT}_{\text{base}}$ representations transformed using the MAKESENSE method show enhanced capacities to distinguish between same and different meanings of surface-forms of words in context, in a supervised setting. These capacities generally increase as we go deeper into the model's layers, matching evidence from previous work (Reif et al., 2019).

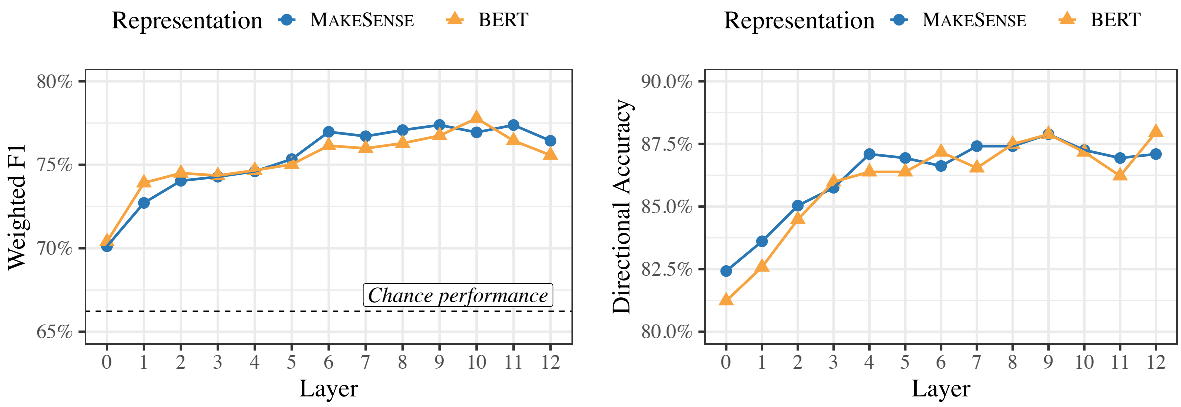
4.4 Probing for sense-sensitive hypernymy

Our final experiment deals with perhaps one of the most fundamental and well-studied lexical relation: the *hypernymy* or IS-A relation (Pustejovsky, 1995). Most linguists argue that hypernymy is a relation between word senses as opposed to surface-forms (see Murphy, 2003, and references therein). That is, *chess* in (3a) is a hyponym of *game* in (3b) but not a hyponym of *game* in *"The poachers looked to hunt the big game,"* where it corresponds to "animal hunted for food" as per WordNet. We explore in this section the extent to which MAKESENSE and $\text{BERT}_{\text{base}}$ encode this sense-sensitive relation, where the pair (*chess*, *game*) in (3a) and (3b) is classified as a case of hypernymy, while the pair (*titles*, *status*) in (3c) and (3d) is not. While MAKESENSE does not include any hierarchical component in its learning mechanism, it should at the very least preserve the hypernymy information that is already contained in $\text{BERT}_{\text{base}}$ for it to be competitive in this experiment, especially since it focuses on manipulating representations for a different—albeit related—task. Arguably, this is a non-trivial task that involves not only discerning the sense of a word from its context, but also predicting the existence as well as direction of the relation — hypernymy is asymmetric, i.e., *chess* is a hyponym of *game* (provided their senses are correctly disambiguated) but the reverse is not true.

For this experiment, we rely on the Word Hypernyms in Context (WHIC) dataset (Vyas and Carpuat, 2017). WHIC consists of pairs of sentence contexts with marked words that are annotated for whether the first word's sense is the hyponym of the second word's sense, thereby making this dataset sensitive to both the senses of words in context and the direction of the relationship. An example stimuli is shown in row 3 of table 1. The dataset comes in standard splits (70% - train, 5% - dev, and 25% - test) that have disjoint vocabulary in terms of the marked words, thereby eliminating issues related to lexical-overlap (Levy et al., 2015). Note that WHIC is an imbalanced dataset, with more negative than positive instances — the negative instances include both directionally opposite versions of the positive instances, as well as multiple cases where the senses of the two words do not have a hypernymy relation.

We again use the probing paradigm to test the extent to which MAKESENSE and $\text{BERT}_{\text{base}}$ representations make sense-sensitive hypernymy relation accessible in a manner that is directionally sensitive. To this end, we conduct tests on two versions of WHIC: (1) WHIC-FULL, which consists of the entire dataset; and (2) WHIC-DIRECTIONAL, which consists of a balanced version of WHIC with positive instances and their directionally reversed counterparts as negative instances. We use the same architecture as the probing experiments on WIC for our WHIC-probing experiments and perform layerwise probing experiments.

Results on WHIC-FULL This test focuses on the overall encoding of hypernymy information in the representations that we test. Due to the imbalanced nature of this dataset, we use the weighted-F1 score as our performance measure, following Vyas and Carpuat (2017). Figure 5a shows our results. Overall, we find that representations from all layers show above-chance performance, suggesting non-trivial access to sense-sensitive hypernmy



(a) WHIC-FULL. (b) WHIC-DIRECTIONAL

Figure 5: layerwise performance from our sense-sensitive hypernymy tests: (a) Weighted F1 scores on WHIC-FULL; and (b) Directional accuracies on WHIC-DIRECTIONAL. **Note:** Y-axes are different in (a) and (b).

information during classification. On comparing MAKESENSE and $\text{BERT}_{\text{base}}$, we see little to no difference in overall performance, suggesting that our approximation experiments show no particular benefits in inferring taxonomic relations from context. At the same time, overall high F1 scores on WHIC-FULL from $\text{BERT}_{\text{base}}$ suggests that this information is considerably imparted during pre-training.

Results on WHIC-DIRECTIONAL This test focuses on specifically shedding light on the extent to which the representations we test are sensitive to the asymmetrical nature of the hypernymy relation. We quantify this sensitivity by evaluating the 'directional accuracy' of the probe trained on the WHIC-DIRECTIONAL subset of WHIC. This metric represents the proportion of pairwise instances where directionally correct instances (*chess* in (3a) is a hyponym of *game* in (3b)) *and* their flipped counterparts (*game* in (3b) is a hypernym of *chess* in (3a)) are assigned the correct label. We observe that both MAKESENSE and $\text{BERT}_{\text{base}}$ show high directional accuracies across all layers, ranging from 81-88%, with performance roughly increasing with layer. Again, we observe that MAKESENSE shows no particular benefit in making the asymmetrical property of hypernymy more accessible during supervision, instead it largely preserves it despite numerically altering the $\text{BERT}_{\text{base}}$ representations.

Takeaways Both MAKESENSE and $\text{BERT}_{\text{base}}$ are equally conducive to making sense and directional sensitive hypernymy information readily accessible from linguistic context. Pre-training imparts a non-trivial amount of context-sensitive hypernymy information to BERT representations and MAKESENSE largely preserves this information.

5 Conclusion and Future work

We present MAKESENSE, a post-processing approach that incorporates word sense information in CWEs. MAKESENSE generalizes the retrofitting paradigm by learning a transformation to push words with similar senses closer together in vector space, while also making the space more isotropic. This way, sense information can be induced for any homonymous or polysemous word by simply passing its contextual representation through MAKESENSE. Through our analyses, we observe MAKESENSE to better impart sense-sensitive information in deeper layers of the original model, resulting in sense-similarity predictions that align better with human intuitions about word senses. Our probing studies show improvements in making sense-disambiguation information more readily accessible. However. we see that MAKESENSE largely preserves hierarchical knowledge about inferred word senses through our investigation for sense-sensitive hypernymy, opening up avenues to incorporate structured lexical semantic knowledge into CWEs in future work.

There remains substantial work to be done in capturing the nuances of lexical ambiguity in context. Our work presents a step towards building generalizable models of lexical specialization, not only at the word token level, but also word sense level. In the future, we aim to experiment with a variety of different approximation methods, as well as incorporate more diverse knowledge sources into the approximation pipeline. It would be informative to also interact MAKESENSE with more context-aware embeddings to better infer word meaning patterns from context.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'14, page 132–148, Berlin, Heidelberg. Springer-Verlag.

Geetanjali Bihani and Julia Taylor Rayz. 2021. Low anisotropy sense retrofitting (laser) : Towards isotropic and sense enriched representations. In *Proceedings of Deep Learning Inside Out (DeeLIO): The Second Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 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.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. 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.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016a. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Allyson Ettinger, Philip Resnik, and Marine Carpuat. 2016b. Retrofitting sense-specific word vectors using parallel text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1378–1383, San Diego, California. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45, Melbourne, Australia. Association for Computational Linguistics.

James A Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive Science*, 31(3):355–384.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. 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.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 683–693, Denver, Colorado. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Walter Kintsch. 2007. Meaning in context. *Handbook of latent semantic analysis*, pages 89–105.

Devorah E Klein and Gregory L Murphy. 2001. The representation of polysemous words. *Journal of Memory and Language*, 45(2):259–282.

Ben Lengerich, Andrew Maas, and Christopher Potts. 2018. Retrofitting distributional embeddings to knowledge graphs with functional relations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2423–2436, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

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.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, pages 1–55.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

M Lynne Murphy. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT press.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Eleanor Rosch. 1973. On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language*, pages 111–144. Elsevier.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 516–527, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana. Association for Computational Linguistics.

Yogarshi Vyas and Marine Carpuat. 2017. Detecting asymmetric semantic relations in context: A case-study on hypernymy detection. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 33–43, Vancouver, Canada. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 1219–1228, New York, NY, USA. Association for Computing Machinery.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550, Baltimore, Maryland. Association for Computational Linguistics.

Lotfi A Zadeh. 1999. Fuzzy logic = computing with words. In *Computing with Words in Information/Intelligent Systems 1*, pages 3–23. Springer.

A Isotropy Improvements by MakeSense

Anisotropy in contextual word embeddings (CWEs) has been shown to hinder the semantic capabilities of models (Gao et al., 2019). Moreover, the existence of anisotropy in a vector space renders vector geometry based sense similarity judgements inconsequential (Ethayarajh, 2019). To address this problem and improve lexical-semantic capabilities of CWEs, recent works have proposed methods to boost the isotropy of the underlying vector space (Gao et al., 2019; Su et al., 2021; Bihani and Rayz, 2021). In this regard, MakeSense-transformed vector spaces show significant improvements in isotropy, especially in the deeper layers of models. We plot the average similarity between 1,000 randomly sampled words (multi-sense nouns, verbs and adjectives) extracted from the sense annotated corpora, for MakeSense and $\text{BERT}_{\text{base}}$ word representations across model layers, as shown in Figure 6. It can be observed that unlike $\text{BERT}_{\text{base}}$ embeddings, where average similarity between random words increases across the model layers, MakeSense embeddings create a vector space such that random words have almost no similarity. Thus, MakeSense-transformed BERT embeddings successfully create uniformly distributed vector spaces, while retaining and even enhancing the lexical-semantic information present.

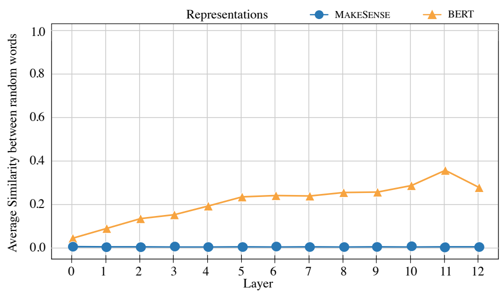


Figure 6: Average similarity between representations of randomly sampled words across model layers

B Dataset statistics

All our data are in the English language. Experimental statistics of the WiC, WHiC-full, and WHiC-directional datasets that we use in our analyses are shown in Table 2.

We collect our experimental data from the following sources:

WiC		
	Same Sense	Different Sense
Train	2,714	2,714
Dev	319	319
WHiC-full		
	Hypernymy	No Hypernymy
Train	3,693	12,023
Dev	283	1,421
Test	1,263	4,098
WHiC-directional		
	Hypernymy	No Hypernymy
Train	3,693	3,693
Dev	283	283
Test	1,263	1,263

Table 2: Statistics of experimental splits of the WiC, WHiC-full, and WHiC-directional datasets used in our probing experiments.

- WiC: https://pilehvar.github.io/wic/package/WiC_dataset.zip
- Usim: https://www.dianamccarthy.co.uk/downloads/WordMeaningAnno2012/cl-meaningincontext.tgz
- WHiC: https://github.com/yogarshi/WHiC

C Training Details for Probing Classifiers

We use probing classifiers for our analyses in §4.3 and §4.4. As described, both our probes are multilayer perceptrons (MLP) with a single hidden layer with 256 units and a final sigmoid layer for classification. Our probes takes as input concatenated representations of the marked words, and classify for same vs. different sense in the case of WiC, and whether the first marked is a hyponym of the second marked word in the context of WHiC. In both cases, we optimize for the binary cross-entropy using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and perform regularization with a weight-decay of 1e-5. Following Hewitt and Liang (2019), we halve the learning rate if after every epoch the optimizer is unable to find a new minimum loss, and stop training if we encounter 5 such epochs consecutively.

D Implementation Details

We use Pytorch (Paszke et al., 2019) and scikit-learn (Pedregosa et al., 2011) for our probing experiments and analyses. The BERT model was accessed using the `transformers` library by HuggingFace (Wolf et al., 2020). Our experiments were run on a NVIDIA V100 GPU with a 32GB RAM.