Changing the Basis of Contextual Representations with Explicit Semantics

Tamás Ficsor
Institute of Informatics,
University of Szeged, Hungary
ficsort@inf.u-szeged.hu

Gábor Berend
Institute of Informatics,
University of Szeged, Hungary
berendg@inf.u-szeged.hu

Abstract

The application of transformer-based contextual representations has became a de facto solution for solving complex NLP tasks. Despite their successes, such representations are arguably opaque as their latent dimensions are not directly interpretable. To alleviate this limitation of contextual representations, we devise such an algorithm where the output representation expresses human-interpretable information of each dimension. We achieve this by constructing a transformation matrix based on the semantic content of the embedding space and predefined semantic categories using Hellinger distance. We evaluate our inferred representations on supersense prediction task. Our experiments reveal that the interpretable nature of transformed contextual representations makes it possible to accurately predict the supersense category of a word by simply looking for its transformed coordinate with the largest coefficient. We quantify the effects of our proposed transformation when applied over traditional dense contextual embeddings. We additionally investigate and report consistent improvements for the integration of sparse contextual word representations into our proposed algorithm.

1 Introduction

In recent years, contextual word representations – such as BERT (Devlin et al., 2019) or GPT-3 (Brown et al., 2020) – have dominated the NLP landscape on leaderboards such as SuperGLUE (Wang et al., 2019) as well as on real word applications (Lee et al., 2019; Alloatti et al., 2019). These models gain their semantics-related capabilities during the pre-training process, which can be then fine-tuned towards downstream tasks, including question answering (Raffel et al., 2019; Garg et al., 2019) or text summarization (Savelieva et al., 2020; Yan et al., 2020).

Representations obtained by transformer-based language models carry context-sensitive semantic information. Although the semantic information is present in the embedding space, the interpretation and exact information it carries is convoluted. Hence understanding and drawing conclusions from them are a cumbersome process for humans. Here we devise such a transformation where we explicitly express the semantic information in the basis of the embedding space. In particular, we express the captured semantic information as finite sets of linguistic properties, which are called semantic categories. A semantic category can represent any arbitrary concept. In this paper, we define them according to WordNet (Miller, 1995) LexNames (sometimes also referred as supersenses).

Even though we present our work on supersense prediction task, our proposed methodology can also be naturally extended to settings that exploit a different inventory of semantic categories. Our results also provide insights into the inner workings of the original embedding space, since we infer the semantic information from embedding spaces in a transparent manner. Therefore, amplified information can be assigned to the basis of the original embedding space.

Sparse representations convey the encoded semantic information in a more explicit manner, which facilitates the interpretability of such representations (Murphy et al., 2012; Balogh et al., 2020). Feature norming studies also illustrated the sparse nature of human feature descriptions, i.e. humans tend to describe objects and concepts with only a handful of properties (Garrard et al., 2001; McRae et al., 2005). Hence, we also conduct experiments utilizing sparse representations obtained from dense contextualized embeddings.

The transformation that we propose in this paper was inspired by Şenel et al. (2018), but it has been extended in various important aspects, as we

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 235–247

- also utilize sparse representations to amplify semantic information,
- analyze several contextual embedding spaces
- apply whitening transformation on the embedding space to decorrelate semantic features, which also servers as the standardization step,
- evaluate the strength of the transformation in a different manner on supersense prediction task.

We also publish our source code on Github: https://github.com/ficstamas/word_embedding_interpretability.

2 Related Work

Contextual word representations provide a solution for context-aware word vector generation. These deep neural language models – such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) or GPT-3 (Brown et al., 2020) – are pre-trained on unsupervised language modelling tasks, and later fine-tuned for downstream NLP tasks. Several variants were proposed to address one or more issue corresponding to the BERT model. Some of which we exploited in this paper. Liu et al. (2019) proposed a better pre-training process, Sanh et al. (2019) reduced the number of parameters, Conneau et al. (2020) presented a multilingual model. These models form the base of our approach, since we produce interpretable representations by measuring the semantic content of existing representations.

One way to measure the morphological and semantic contents of contextual word embeddings is via the application of probing approaches. The premise of this approach is that, if the probed information can be identified by a linear classifier, then the information is encoded in the embedding space (Adi et al., 2016; Ettinger et al., 2016; Klafka and Ettinger, 2020). Others explored the capacity of language models, where they examined the output probabilities of the model in given contexts (Linzen et al., 2016; Wilcox et al., 2018; Marvin and Linzen, 2018; Goldberg, 2019). We slightly reflect the premise of these methodologies by introducing a logistic regression baseline model.

Another approach is to incorporate external knowledge into Language Models. Levine et al. (2020) devised SenseBERT by integrating supersense information into the training of BERT. K M et al. (2018) showed a method where an arbitrary knowledge graph can be incorporated into their LSTM based model. External knowledge incorporation is getting a popular approach to improve already existing state-of-the-art solutions in a domain or task specific environment (Munkhdalai et al., 2015; Weber et al., 2019; Baral et al., 2020; Mondal, 2020; Wise et al., 2020; Murayama et al., 2020). Since we deemed to investigate the effect of incorporated knowledge towards the semantic content of embedding space, SenseBERT serves a good basis for that.

Ethayarajh (2019) investigated the importance of anisotropic property of the contextual embeddings, which is a different kind of investigation than we aim to do. It still gives a good insight into the inner workings of the layers. Şenel et al. (2018) showed a method where they measured the interpretability of Glove embeddings, and later showed a method to manipulate and improve the interpretability of a given static word representation (Şenel et al., 2020). Our approach resembles Şenel et al. (2018), however, we apply different pre- and post-processing steps and more importantly, we replaced the usage of the Bhattacharyya distance with the Hellinger distance, which is closely related to it but operates in a bounded and continuous manner. Our approach also differs from Şenel et al. (2018) in that we deal with contextualited language models instead of static word embeddings and we also rely on sparse contextualized word vectors.

The intuition behind sparse vectors is related to the way humans describe concepts, which has been extensively studied in various feature norming studies (Garrard et al., 2001; McRae et al., 2005). Additionally, generating sparse features (Kazama and Tsujii, 2003; Friedman et al., 2008; Mairal et al., 2009) has proved to be useful in several areas, including POS tagging (Ganchev et al., 2010), text classification (Yogatama and Smith, 2014) and dependency parsing (Martins et al., 2011). Therefore, several sparse static representations were presented, such as Murphy et al. (2012) proposed Non-Negative Sparse Embeddings to represent interpretable sparse word vectors. Park et al. (2017) showed a rotation-based method and Subramanian et al. (2017) suggested an approach using a denoising k-sparse auto-encoder to generate sparse word vectors. Berend (2017) showed that sparse representations can outperform their dense counterparts in certain NLP tasks, such as NER, or POS tagging. Additionally, Berend (2020) illustrated

how applying sparse representations can boost the performance of contextual embeddings for Word Sense Disambiguation, which we also desire to exploit.

3 Our Approach

We first define necessary notations. We denote the embedding space with $\mathcal{E} \in \mathbb{R}^{v \times d}$ with the superscript indicating whether it is obtained from the training set t or evaluation set e. We denote the number of input words and their dimensionality by v and d, respectively. Furthermore, we denote the transformation matrix with $\mathcal{W} \in \mathbb{R}^{d \times s}$ – where s represents the number of semantic categories – and the final interpretable representation with $\mathcal{I} \in \mathbb{R}^{v \times s}$, which always denotes the interpretable representation of $\mathcal{E}^{(e)}$. Additionally, we denote the semantic categories with $\mathcal{S}$.

3.1 Interpretable Representation

Our goal is to produce such embedding spaces where we can identify semantic features by their basis. In order to obtain such an embedding space, we are constructing a transformation matrix $\mathcal{W}^{(t)}$, which amplifies the semantic information of an input representation and can be formulated as: $\mathcal{I} = \mathcal{E}_w^{(e)} \times \mathcal{W}^{(t)}$. $\mathcal{E}_w$ represents the whitened embedding space, which is the output of a pre-processing step (Section 3.2), and $\mathcal{W}$ being our transformation matrix (Section 3.3).

3.2 Pre-processing

Pre-processing consists of two steps: first we generate sparse representations of dense embedding spaces (this step is omitted when we report about dense embedding spaces), then we whiten the embedding space.

3.2.1 Sparse Representation

For obtaining sparse contextualized representations, we follow the methodology proposed in (Berend, 2020). That is, we solve the following sparse coding (Mairal et al., 2009) optimization problem:

$$\min_{\alpha^{(t)}, D} \frac{1}{2} \left\| \mathcal{E}^{(t)} - \alpha^{(t)} D \right\|_F^2 + \lambda \left\| \alpha^{(t)} \right\|_1,$$

where $D \in \mathbb{R}^{k \times d}$ is the dictionary matrix, and $\alpha \in \mathbb{R}^{v \times k}$ contains the sparse contextualized representations. The two hyperparameters of the dictionary learning approach are the number of basis vectors to employ (k) and the strength of the regularization (λ).

We obtained the sparse contextual representations for the words in the evaluation set by fixing the dictionary matrix D that we learned on the train set and optimized solely for the sparse coefficients $\alpha^{(e)}$. We also report experimental results obtained for different values of basis vectors k and regularization coefficients λ.

The output of this step is also represented with $\mathcal{E}$ instead of α since this step is optional. Among our results we mark whether we applied (*Sparse*) or skipped (*Dense*) this step.

3.2.2 Whitening

Since we handle dimensions independently, we first apply whitening transformation on the embedding space. Several whitening transformations are known – like Cholesky or PCA (e.g. Friedman (1987)) – but we decided to rely on ZCA whitening (or Mahalanobis whitening) (Bell and Sejnowski, 1997). One benefit of employing ZCA whitening is that it ensures higher correlation between the original and whitened features (Kessy et al., 2018). As a consequence, it is a widely utilized approach for obtaining whitened data in NLP (Heyman et al., 2019; Glavaš et al., 2019).

We determine the whitening transformation matrix from the training set ($\mathcal{E}^{(t)}$), which is then applied on the representation of our training ($\mathcal{E}^{(t)}$) and evaluation sets ($\mathcal{E}^{(e)}$). We denote the whitened representations for the training and evaluation sets by $\mathcal{E}_w^{(t)}$ and $\mathcal{E}_w^{(e)}$, respectively.

3.3 Transformation

In this section, we discuss the way we measure the semantic information of the embedding space and express the linear transformation matrix ($\mathcal{W}$).

3.3.1 Semantic Distribution

The coefficients of the contextual embeddings of words that belong to the same (super)sense category are expected to originate from the same distribution. Hence, it is reasonable to quantify the extent to which some semantic category is encoded along some dimension by investigating the distribution of the coefficients of the word vectors along that dimension. For every semantic category, we can partition the words whether they pertain to that category. When a dimension encodes a semantic category to a large extent, the distribution of the

coefficients of those words belonging to that category is expected to differ substantially from that of those words not pertaining to the same category.

We can formulate the distributions of our interest by function $L : x \rightarrow \mathcal{S}$, which maps each token (x) to its context-sensitive semantic category (LexName) and a function $f : x \rightarrow \mathcal{E}$, which returns the context-sensitive representation of x. Thus the devised distributions can be defined as:

$$P_{ij} = \left\{ f(x)^{(i)} \mid f(x) \in \mathcal{E}_w^{(t)},\, L(x) \in \mathcal{S}^{(j)} \right\}$$

and

$$Q_{ij} = \left\{ f(x)^{(i)} \mid f(x) \in \mathcal{E}_w^{(t)},\, L(x) \notin \mathcal{S}^{(j)} \right\},$$

where i represents a dimension and j denotes a semantic category. In other words, P_{ij} represents the distribution along the ith dimension of those words that belong to the jth semantic category, whereas Q_{ij} represents the distribution of the coefficients along the same dimension (i) of those words that do not belong to the jth semantic category.

3.3.2 Semantic Information and Transformation Matrix

For every dimension (i) and semantic category (j) pair, we can express the presence of the semantic information by defining a distance between the distributions P_{ij} and Q_{ij}. Following from the construction of the distributions P_{ij} and Q_{ij}, the larger the distance between a pair of distributions (P_{ij}, Q_{ij}), the more likely that dimension i encodes semantic information j.

Based on that observation, we define a transformation matrix $\mathcal{W}_D$ as

$$\mathcal{W}_D(i,j) = D(P_{ij}, Q_{ij}),$$

where D is the distance function. We specify the distance function as the Hellinger distance, which can be formulated as

$$\sqrt{1 - \sqrt{\frac{2\sigma_{p_{ij}}\sigma_{q_{ij}}}{\sigma_{p_{ij}}^2 + \sigma_{q_{ij}}^2}} e^{-\frac{1}{4} \cdot \frac{(\mu_{p_{ij}} - \mu_{q_{ij}})^2}{\sigma_{p_{ij}}^2 + \sigma_{q_{ij}}^2}}},$$

where we assume that $P_{ij} \sim \mathcal{N}(\mu_{p_{ij}}, \sigma_{p_{ij}})$ and $Q_{ij} \sim \mathcal{N}(, \mu_{q_{ij}}, \sigma_{q_{ij}})$, i.e. they are samples from normal distributions with expected value μ and standard deviation σ.

We decided to rely on Hellinger distance due to its continuous, symmetric and bounded nature. In contrast to out approach, Şenel et al. (2018) proposed the usage of Bhattacharyya distance – which is closely related to Hellinger distance – but it would overestimate the certainty of the semantic information of a dimension in the case of distant distributions. Another concern is that the Bhattacharyya distance is discontinuous. We discussed this topic in a earlier work (Ficsor and Berend, 2020) in relation to static word embeddings.

Bias Reduction. So far, our transformation matrix is biased due to the imbalanced semantic categories. It can be reduced by ℓ_1 normalizing $\mathcal{W}_D$ in such a manner that vectors representing semantic categories sum up to 1, which we denote as $\mathcal{W}_{ND}$ (Normalized Distance Matrix).

Directional Encoding. As semantic information can be encoded in both positive and negative directions, we modify the entries of $\mathcal{W}_{ND}$ as

$$\mathcal{W}_{NSD}(i,j) = sign(\mu_{p_{ij}} - \mu_{q_{ij}}) \cdot \mathcal{W}_{ND}(i,j),$$

where $sign(\cdot)$ is the signum function. This modification ensures that each semantic category is represented with the highest coefficients in their corresponding base of the interpretable representation.

3.4 Post-processing

The representations transformed in the above manner are still skewed in the sense that they do not reflect the likelihood of each semantic category. In order to alleviate that problem, we measure and normalize the frequency ($\mathbf{f_N} = \mathbf{f}/\|\mathbf{f}\|_2, \mathbf{f} \in \mathbb{N}^s$) of each occurrence of a supersense category in the training set and accumulate that information into the embedding space in the following manner: $\mathcal{I}_\mathrm{f} = \mathcal{I} + \mathcal{I} \odot \mathbf{1}\mathbf{f}_N^\intercal$, where $\odot$ represents the element-wise multiplication, and $\mathbf{1}$ represents a vector consisting of all ones. Finally, $\mathcal{I}_\mathrm{f}$ represents our final interpretable representations adjusted with supersense frequencies.

3.5 Accuracy Calculation

Representations generated by our approach let us determine the presumed semantic category by the highest coefficient in the word vector. In other words, a word vector should have its highest coefficient in the base, which represents the same semantic category as the annotation represents in the evaluation set. Our overall accuracy is the fraction of the correct predictions and the total number of annotated data in the evaluation set.

4 Evaluation

4.1 Experimental setting.

During our experiments, we relied on the SemCor dataset for training and the unified word sense disambiguation framework introduced in (Raganato et al., 2017a) for evaluation, which consists of 5 sense annotated corpora: *SensEval2* (Edmonds and Cotton, 2001), *SensEval3* (Mihalcea et al., 2004), *SemEval 2007* Task 17 (Pradhan et al., 2007), *SemEval 2013* Task 12 (Navigli et al., 2013), *SemEval 2015* Task 13 (Moro and Navigli, 2015) and their concatenation. We refer to the combined dataset as *ALL* througout the paper. The individual datasets contain 2282, 1850, 455, 1644 and 1022 sense annotations, respectively. These datasets contain fine-grained sense annotation for a subset of the words from which the supersense information can be conveniently inferred. We reduced the scope of fine-grained sense annotations to lexname level, in order to maintain well-defined semantic categories with high sample sizes. We used the *SemEval 2007* data as our development set in accordance with prior work (Raganato et al., 2017b; Kumar et al., 2019; Blevins and Zettlemoyer, 2020; Pasini et al., 2021).

We conducted our experiments on several contextual embedding spaces, where each model represent a different purpose. We can consider BERT (Devlin et al., 2019) as the baseline of the following contextual models. SenseBERT (Levine et al., 2020) incorporated word sense information into its latent representation. DistilBERT (Sanh et al., 2019) obtained through knowledge distillation and operates with less parameters. RoBERTa (Liu et al., 2019) introduced a better pre-training procedure. Finally, XLM-RoBERTa (Conneau et al., 2020) is a multilingual model with the RoBERTa's pre-training procedure. When available, we also conducted experiments using both `cased` and `uncased` vocabularies.

Following (Loureiro and Jorge, 2019), we also averaged the representations from the last 4 layers of the transformer models to obtain our final contextual embeddings. Furthermore, to determine the hyperparameters for sparse vector generation, we used the accuracy of BERT `Base` model with different regularizations (λ) and number of employed basis (k) on the *SemEval2007* dataset, the results of which can be seen in Table 1.

		λ		
		0.05	0.1	0.2
k	1500	63.51	64.83	57.80
	3000	65.71	**66.59**	64.61

Table 1: Results of our experiments when relying on sparse representations created by using various hyperparameter combinations. The BERT `Base` model was used on the SemEval2007 validation set. k represents the number of employed basis and λ denotes the regularization parameter.

4.2 Baselines

We next introduce those baselines we compared our approach with. Most of these approaches rely on the intact contextual representations $\mathcal{E}$, for which the dimensions are not intended to directly encode human interpretable supersense information about the words they describe.

Logistic Regression Classifier We conducted the experiments by setting the random state to 0, maximum iterations to 25,000 and turned off the utilization of a bias term. In this case the vectors that were used for making the predictions about the supersenses of words were of much higher dimensions and not directly interpretable at all, unlike our representations.

Dimension Reduction (PCA+LogReg) We also experimented with representations, which inherit the same number of dimensions as many we utilize (45). So we applied principal component analysis (PCA) based dimension reduction on the original $\mathcal{E}$ embedding space. Additionally, we applied Logistic Regression Classifier on the reduced representations with the same parametirazition to the previously described baseline.

Sparsity Makes Sense (SMS) An approach proposed by Berend (2020) yields human-interpretable embeddings like ours, since human-interpretable features are bound to the basis of the output representation. Berend (2020) originally presented the devised algorithm on fine-grained word sense disambiguation, which we altered to work similarly to our approach and predict supersense information instead. We utilized normalized positive pointwise mutual information to construct the transformation matrix because it showed the most prominent scores in the paper.

Representation		Interpretable						Latent			
Method		Our Approach				SMS		PCA+LogReg		LogReg	
Input Embedding Type		Dense		Sparse		Sparse		Dense		Dense	
Vocabulary (Cased/Uncased)		C	U	C	U	C	U	C	U	C	U
ALL-dev											
BERT	Base	65.04	62.44	69.53	68.43	65.24	63.00	57.45	54.70	73.96	72.64
	Large	63.68	62.51	68.41	64.82	62.00	57.03	55.60	51.05	73.25	71.69
SenseBERT	Base	–	66.13	–	74.59	–	74.21	–	68.57	–	79.47
	Large	–	64.62	–	74.55	–	73.75	–	71.44	–	78.99
DistilBERT	Base	62.94	64.44	70.78	72.68	66.31	68.03	59.34	61.51	74.86	74.46
RoBERTa	Base	59.47	–	65.40	–	61.91	–	52.25	–	69.44	–
	Large	64.43	–	70.27	–	65.85	–	52.91	–	75.16	–
XLM-RoBERTa	Base	63.31	–	70.10	–	67.84	–	58.43	–	76.02	–
	Large	62.10	–	67.74	–	64.63	–	57.89	–	75.54	–

Table 2: Accuracy of each model on the supersense prediction task using dense and sparse embedding spaces. *ALL-dev* denotes the evaluation on the *ALL* dataset excluding the development set. All of the sparse representations were generated using $\lambda = 0.1$ for the regularization coefficient and $k = 3000$ basis based on the experiments reported in Table 1. Our approach and SMS are interpretable representations, PCA+LogReg just represents the information in the same number of basis but there are no connection, which can be drawn to the previous two, and Logistic Regression operates on the original embedding spaces. We also include a more detailed table in the Appendix, which breaks down performances for each sub-corpora.

4.3 Results

We list the results of our experiments using different contextual encoders on the task of supersense prediction in Table 2. We calculated the accuracy as the fraction of correct predictions and the total number of annotated samples. We selected $\lambda = 0.1$ regularization and $k = 3000$ basis for sparse vector generation in accordance with the results that we obtained over the development set for different choices of the hyperparameters (see Table 1).

4.3.1 Model Performances

We consider a model's semantic capacity as the Logistic Regression model's performance, and its interpretability as the best performing interpretable representation. We do not expect to exceed the original model, since we limited its capabilities drastically by reducing the number of utilized dimensions to 45.

By looking at the performance, as expected the original latent representation expresses the most semantic information measure by Logistic Regression. Among all of them, SenseBERT dominates which is due to the additional supersense information signal it relies on during its pretraining. The incorporated supersense information helps SenseBERT to represent that information more explicitly, which becomes more obvious when we amplify it by sparse representations. So including further objectives during training just further separates the information in the basis.

4.3.2 Dense and Sparse Representations

We can see from Table 2 that relying on sparse representations further amplifies the semantic content of the latent representations. Based on the results of our approach, we can conclude that the semantic information can be more easily identified in the case of sparse representations (as indicated by the higher scores in the majority of the cases). SMS follows a similar trend to ours. Also the relatively small decrease in performance suggests that the majority of the removed signals correspond to noise.

4.3.3 Impact of Base and Large Models

In several cases, the `Large` models underperformed their `Base` counterparts (except RoBERTa). It can indicate that the `Large` version might be under-trained, which was also hypothesised in (Liu et al., 2019). Overall, choosing the `Base` pre-trained models seems to be a sufficient and often better option for performing supersense prediction.

		Mean (Std)	
		Cased	Uncased
BERT	Base	0.35 (±0.21)	0.32 (±0.21)
	Large	0.29 (±0.22)	0.28 (±0.22)
SenseBERT	Base	–	0.59 (±0.25)
	Large	–	0.55 (±0.29)
DistilBERT	Base	0.34 (±0.21)	0.33 (±0.20)
RoBERTa	Base	0.34 (±0.22)	–
	Large	0.31 (±0.21)	–
XLM-RoBERTa	Base	0.34 (±0.22)	–
	Large	0.32 (±0.22)	–

Table 3: Average Spearman Rank Correlation between the basis of our interpretable embedding space and the one obtain by the SMS approach.

4.3.4 Case-sensitivity of the Vocabulary

As the choice whether using a cased or an uncased model is more beneficial can vary from task to task, we made experiments in that respect. To this end, we compared the performance of BERT and DistilBERT, which are available in both case sensitive and case insensitive versions. Usually, the choice highly depends on the task (cased versions being recommended for POS, NER, WSD) and the language (cased can be beneficial for certain languages such as German). Overall, we can observe some advantage of using the cased vocabularies. Interestingly, the behavior of DistilBERT and BERT differs radically in that respect for all but the LogReg approach.

4.3.5 Considering Dimensionality

Other than the Logistic Regression model, every approach relies on some kind of condensed representation for supersense prediction. Even though all of the representations were condensed – into 45 dimensions from 768, 1024 dimensions for dense and 3000 dimensions for sparse representations – the performance did not decreased by a large margin. PCA-based dimension reduction approach performed the worst among the 3 approaches, whereas ours performed the best. Note that these interpretable approaches (ours and SMS) not only perform better over a standard dimension reduction, but they also associate human-understandable knowledge to the basis of the embedding space. So it can be utilized as an explicit semantic compression technique.

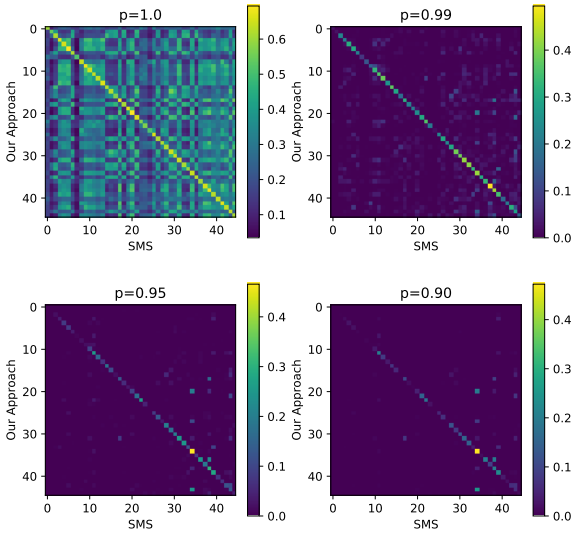

Figure 1: Rank-biased Overlap scores between the basis of our approach and SMS on sparse representations of SenseBERT Base models. Here the p value indicates the steep of decline in weights (smaller the p the more top-weighted the metric is).

4.3.6 Comparing Interpretable Representations

Both our and SMS approach are similar in the sense that we can assign human-interpretable features to the basis of output embeddings. We hence analysed the similarity of the semantic information of the two embedding spaces. We measured the Spearman rank correlation of the coefficients in each pair of basis generated by our approach and the SMS approach. We included these values in Table 3, which showcases the mean of absolute (ignoring the direction of correlation) correlation coefficients. Except for SenseBERT, we can see weak correlation scores. Higher correlation between the coefficients of these interpretable models, along the same dimension would suggest that they can represent the same semantic information to a different level and/or manner. According to the Spearman correlation between our and the SMS approach captures a different aspect of the encoded semantic content, but we futher experimented with SenseBERT.

Since the two embeddings expressed from SenseBERT – with our and SMS approach – seem to share the most semantic content, we investigated them further. During our evaluation, we rely on the maximum value of each word token, so each dimension represents the semantic information among its highest coefficients. Hence, higher value ranks a word more likely to carry the corresponding semantic information. Therefore, we calculated Rank-

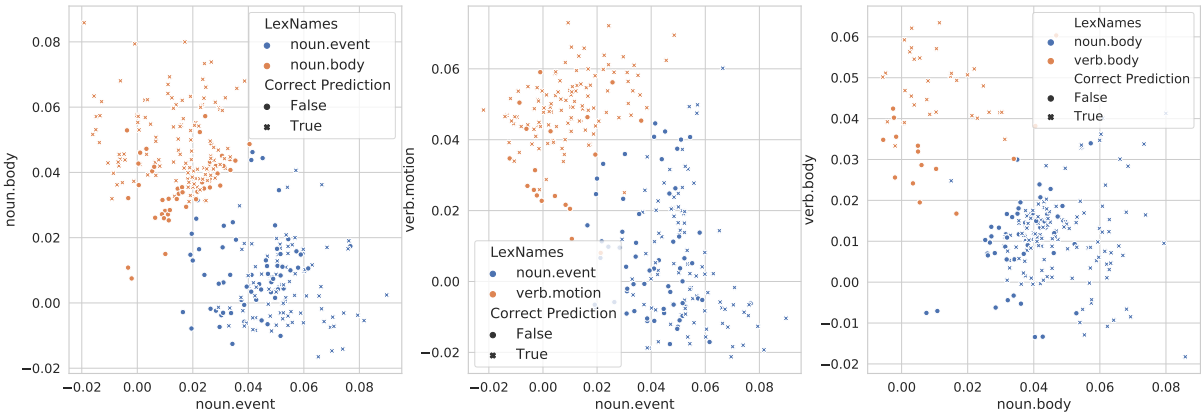

Figure 2: Representation of the coefficients of several semantic categories where the color represents the assigned label according to the corpus, whether the prediction according to the maximum is correct (True) or not (False), and both axis represent its value in their corresponding basis in our representation (SenseBERT, $k = 3000, \lambda = 0.1$).

biased Overlap (RBO) scores (Webber et al., 2010) between the sorted basis, which can be seen in Figure 1. RBO quantifies a weighted, non-conjoint similarity measure, which does not rely on correlation. RBO utilizes a p parameter, which controls the emphasis we have on top ranked items (lower p indicates more emphasis on the top ranked items). The $p = 1$ case differs from the $p < 1$ case, in that it returns the un-bounded set-intersection overlap calculated according to the proposition from Fagin et al. (2003). On the other hand, $p < 1$ prioritizes the head of the lists. Higher score indicates higher similarity between two ranked lists, which in our case means that the two models behave more similarly.

Both models perform comparable in general with slightly better scores on sparse models for our approach. We measured the statistical significance of the improvements by Berg-Kirkpatrick et al. (2012), which states the following H_0 hypothesis: *if* $p(\delta(X) > \delta(x)|H_0) < 0.05$ *then we accept the improvement of the first model and unlikely to be cause of random factors*, where $\delta(\cdot)$ represents the improvement of the first model. Furthermore, we used $b = 10^6$ bootstraps, which was sufficient according to the original paper. Between sparse models we obtained $p = 0.0016$ value, which suggests that the significance of improvement is unlikely to be caused by random factors.

4.3.7 Qualitative Assessments

Clustering We demonstrate the semantic decomposition of 3 pairs of semantic categories in Figure 2. Each marker corresponds to a concrete word occurrence with their color reflecting their expected supersense. The markers also indicate whether the prediction made according to the highest coordinate is correct (True) or not (False). Furthermore, both axis represents its actual value in its corresponding base. We can notice in these figures how well data points are separated with respect to their semantic properties.

Shared Space of Multilingual Domain The availability of multilingual encoders allows us to use our supersense classifier on languages other than English as well. In order to test the applicaility of XLM-RoBERTa in such a scenario, we tested it on some sentences in multiple languages, the outcome of which is included in Table 4.

To this experiment, we constructed $\mathcal{W}_D$ in the usual manner from Sparse XLM-RoBERTa transformer on the SemCor dataset (which is in English). After that, we generated the context aware word vectors for the sentences. We then obtained the sparse representations from them by employing the already optimized dictionary matrix from SemCor. We finally utilized the previously constructed distance matrix to obtain the interpretable representation. In Table 4, we marked the expected label above the text with blue, and the top 3 predictions with red below the text.

We included 3 typologically diverse languages German (DE), Hungarian (HU) and Japanese (JP). Overall, the expected label was within the top 3 predictions irrespective of the language, which suggests that the overlap in semantic distribution is high between languages, but further quantitative experiments are also needed to support that statement.

	adj.all	noun.cognition	verb.stative		adj.all	noun.act
Dein	bester	Lehrer	ist	dein	letzter	Fehler.
	adj.all	noun.person	verb.stative		adj.all	noun.act
	noun.event	noun.act	verb.social		noun.shape	noun.attribute
	verb.competition	noun.cognition	verb.change		verb.competition	noun.feeling

DE) Translation: Your best teacher is your last mistake. – Ralph Nader

adv.all	noun.attribute	verb.stative	adj.all	noun.attribute
Együtt	erő	vagyunk,	szerteszét	gyöngeség.
adv.all	noun.attribute	verb.stative	verb.body	noun.feeling
adj.all	noun.feeling	verb.weather	adj.all	noun.state
verb.social	noun.phenomenon	verb.consumption	verb.competition	noun.attribute

HU) Translation: We are strong together, and weak as scattered. – Albert Wass

noun.location	verb.motion	noun.time		noun.person	adv.all	verb.body
千代田町	に 着いた	時	には、	禎子	は すでに	生まれていた のです。
noun.Tops	verb.motion	noun.event		noun.Tops	adv.all	verb.change
noun.location	verb.change	noun.time		noun.person	noun.object	verb.body
noun.object	verb.contact	noun.shape		noun.animal	noun.food	adj.all

JP) Translation: Upon arriving to Chiyoda, Sadako was already born. – Eleanor Coerr

Table 4: A few example of shared knowledge between languages in XLM-RoBERTa. We used the transformation matrix learned on the English SemCor dataset with Sparse XLM-RoBERTa `Base` model. Above the text with blue we mark the expected label, and below the text with red the top 3 predictions.

5 Conclusion

In this paper, we demonstrated our approach to obtain interpretable representations from contextual representations, which represents semantic information in the basis with high coefficients. We demonstrated its capabilities by applying it on supersense prediction task. However, it can be utilized on other problems as well such as term expansion and knowledge base completion.

We additionally explored the application of sparse representations, which successfully amplified the examined semantic information. We also considered the effect of incorporated prior knowledge in the form of applying SenseBERT embeddings, which showed that its additional objective during pre-training can amplify those features. Furthermore, explored the space of condensed (DistilBERT) and multilingual (XLM-RoBERTa) spaces. We examined the improvements come by RoBERTa from a semantic standpoint. Note that our classification decision is currently made by simply finding the coordinate with the largest magnitude.

In conclusion, our experiments showed that it is possible to extract and succinctly represent human-interpretable information about words in transformed spaces with much lower dimensions than their original representations. Additionally, it allows us to make decisions about word vectors in a more transparent manner, where some kind of explanation is already assigned to the basis of a representation, which can lead us to more transparent machine learning models.

Acknowledgements

This research was supported by the European Union and co-funded by the European Social Fund through the project "Integrated program for training new generation of scientists in the fields of computer science" (EFOP-3.6.3-VEKOP-16-2017-0002) and the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program and the Artificial Intelligence National Excellence Program (2018-1.2.1-NKP-2018-00008).

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks.

Francesca Alloatti, Luigi Di Caro, and Gianpiero Sportelli. 2019. Real life application of a question answering system using BERT language model. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 250–253, Stockholm, Sweden. Association for Computational Linguistics.

Vanda Balogh, Gábor Berend, Dimitrios I. Diochnos, and György Turán. 2020. Understanding the semantic content of sparse word embeddings using a commonsense knowledge base. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7399–7406.

Chitta Baral, Pratyay Banerjee, Kuntal Kumar Pal, and Arindam Mitra. 2020. Natural language QA approaches using reasoning with external knowledge.

Anthony J. Bell and Terrence J. Sejnowski. 1997. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

Gábor Berend. 2017. Sparse coding of neural word embeddings for multilingual sequence labeling. *Transactions of the Association for Computational Linguistics*, 5:247–261.

Gábor Berend. 2020. Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8498–8508, Online. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. 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 *ACL*, pages 1006–1017. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451. 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.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SENSEVAL '01, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. 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.

Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing top k lists. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, page 28–36, USA. Society for Industrial and Applied Mathematics.

Tamás Ficsor and Gábor Berend. 2020. Interpreting word embeddings using a distribution agnostic approach employing hellinger distance. In *Text, Speech, and Dialogue*, pages 197–205, Cham. Springer International Publishing.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9:432–41.

Jerome H. Friedman. 1987. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397):249–266.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, 11:2001–2049.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2019. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection.

Peter Garrard, Matthew Ralph, and Karalyn Patterson. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive neuropsychology*, 18:125–74.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *CoRR*, abs/1902.00508.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. 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 1890–1902, Minneapolis, Minnesota. Association for Computational Linguistics.

Annervaz K M, Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. 2018. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. 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 313–322, New Orleans, Louisiana. Association for Computational Linguistics.

Jun'ichi Kazama and Jun'ichi Tsujii. 2003. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 137–144, Morristown, NJ, USA. Association for Computational Linguistics.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. 2018. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

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.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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.

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.

Julien Mairal, Francis R. Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online dictionary learning for sparse coding. In *ICML*, volume 382 of *ACM International Conference Proceeding Series*, pages 689–696. ACM.

André F. T. Martins, Noah A. Smith, Mário A. T. Figueiredo, and Pedro M. Q. Aguiar. 2011. Structured sparsity in structured prediction. In *EMNLP*, pages 1500–1511. ACL.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Ken McRae, George Cree, Mark Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37:547–59.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 english lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Ishani Mondal. 2020. Bertchem-ddi: Improved drug-drug interaction prediction from text using chemical structure information. *arXiv preprint arXiv:2012.11599*.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation*

(SemEval 2015), pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Tsendsuren Munkhdalai, Meijing Li, Khuyagbaatar Batsuren, Hyeon Park, Nak Choi, and Keun Ho Ryu. 2015. Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. *J. Cheminformatics*, 7(S-1):S9.

Yuri Murayama, Lis Kanashiro Pereira, and Ichiro Kobayashi. 2020. Dialogue over context and structured knowledge using a neural network model with external memories. In *Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP*, pages 11–20, Suzhou, China. Association for Computational Linguistics.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING 2012*, pages 1933–1950, Mumbai, India. The COLING 2012 Organizing Committee.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Sungjoon Park, JinYeong Bak, and Alice Oh. 2017. Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 401–411, Copenhagen, Denmark. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proc. of AAAI*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017a. 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.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017b. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Alexandra Savelieva, Bryan Au-Yeung, and Vasanth Ramani. 2020. Abstractive summarization of spoken and written instructions with BERT. In *Proceedings of the KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption co-located with the 26TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2020), Virtual Workshop, August 24, 2020*.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2017. Spine: Sparse interpretable neural embeddings.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4).

Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. 2019. NLProlog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6151–6161, Florence, Italy. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Colby Wise, Vassilis N. Ioannidis, Miguel Romero Calvo, Xiang Song, George Price, Ninad Kulkarni, Ryan Brand, Parminder Bhatia, and George Karypis.

2020. Covid-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature. *CoRR*, abs/2007.12731.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training.

Dani Yogatama and Noah A. Smith. 2014. Linguistic structured sparsity in text categorization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–796, Baltimore, Maryland. Association for Computational Linguistics.

L. K. Şenel, İ. Utlu, V. Yücesoy, A. Koç, and T. Çukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779.

Lütfi Kerem Şenel, İhsan Utlu, Furkan Şahinuç, Haldun M. Ozaktas, and Aykut Koç. 2020. Imparting interpretability to word embeddings while preserving semantic structure. *Natural Language Engineering*, page 1–26.