Tracking the Traces of Passivization and Negation in Contextualized Representations

Hande Celikkanat **Sami Virpioja** **Jörg Tiedemann** **Marianna Apidianaki**

Department of Digital Humanities
University of Helsinki
Helsinki, Finland
firstname.lastname@helsinki.fi

Abstract

Contextualized word representations encode rich information about syntax and semantics, alongside specificities of each context of use. While contextual variation does not always reflect actual meaning shifts, it can still reduce the similarity of embeddings for word instances having the same meaning. We explore the imprint of two specific linguistic alternations, namely passivization and negation, on the representations generated by neural models trained with two different objectives: masked language modeling and translation. Our exploration methodology is inspired by an approach previously proposed for removing societal biases from word vectors. We show that passivization and negation leave their traces on the representations, and that neutralizing this information leads to more similar embeddings for words that should preserve their meaning in the transformation. We also find clear differences in how the respective features generalize across datasets.

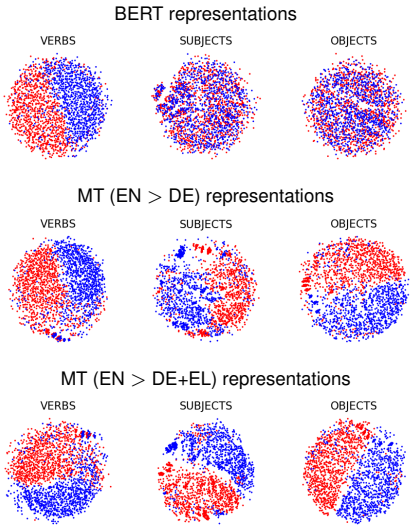

Figure 1: Multidimensional (MDS) visualization of representations obtained for verbs and nouns from active (red) and corresponding passive (blue) sentences. Data points are BERT representations (top) and the encodings from machine translation models involving one (middle) or two language pairs (bottom).

1 Introduction

Contextualized representations extracted from pretrained language models reflect the syntactic and semantic properties of words (Linzen et al., 2016; Hewitt and Manning, 2019; Rogers et al., 2020; Tenney et al., 2019) as well as variation in their context of use. We propose to explore the impact of context variation on word representations. We specifically address representations generated by the BERT model (Devlin et al., 2019), trained using a language modeling objective, and translation models involving one or more language pairs (Artetxe and Schwenk, 2019; Vázquez et al., 2020).

We run a series of controlled experiments using sentences illustrating both meaning preserving and meaning altering transformations from the SICK dataset (Marelli et al., 2014b), and examples automatically generated using a template-based method (Prasad et al., 2019). We explore the impact of specific alternations on the representations, namely passivization and negation. Examples in our datasets consist of sentences that only differ in terms of the specific alternation addressed. In order to detect the imprint of these transformations on the representations, we employ methodology inspired by work on linguistic bias detection in embedding representations (Bolukbasi et al., 2016; Lauscher et al., 2019; Ravfogel et al., 2020).

Furthermore, we investigate the impact of removing the encoding of such alternations on word similarity. Intuitively, we would expect the representations of words present in sentences that have undergone passivization (PAS) to be highly similar despite the differences in syntactic structure. Consider, for example, the words *mafia*, *millionaire*

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 136–148
Online, November 20, 2020.

and *kidnapped* in the examples ① and ②.

① *The mafia kidnapped the millionaire.*
② *The millionaire was kidnapped by the mafia.*

PAS changes the words' syntactic roles but their thematic roles remain the same. The meaning shift that results from this operation is mainly discursive,[1] shifting the focus from the theme to the agent, but the content words in the two sentences still refer to the same event and entities.[2] Their representations should thus be highly similar.

We also address a meaning altering transformation which involves inserting (or removing) the negation particle to produce contradictions, as in ③ and ④.

③ *The boy is playing the piano.*
④ *The boy is not playing the piano.*

The effect of negation (NEG) at the sentence level is obvious. However, the meaning of specific words (*boy*, *playing*, *piano*) should remain the same despite of the whole sentence having the opposite meaning. Below, we explore the extent to which this type of context variation affects the similarity of the representations of word instances in the two sentences.

We show that passivization and negation[3] have a significant imprint on the representations, and that their removal can improve word similarity estimation. Our results also highlight that this type of context variation is differently marked in representations generated by models trained with different objectives. Specifically, we find that variation in the embeddings produced by models trained with a translation objective generalize better than those derived from models trained with a masked language modeling objective, across datasets, in the sense that they seem to be encoded in features that are independent of the specific dataset.

[1]Note, however, that the impact of the alternation on the framing of the sentence can be significant. Passive avoids identifying a causal agent and therefore conceals the responsibility for an event (Greene and Resnik, 2009).

[2]In sentence ①, the *mafia* is the agent and is in subject position, while the *millionaire* is the theme in direct object position. In ②, the semantic relationship of the *mafia* and the *millionaire* to the kidnapping event is the same but their syntactic roles have changed.

[3]These two transformations were preferred on the basis that they do not change the words in the sentence, as opposed to other possible translations, which involve reformulations, eg. "a sewing machine" vs. "a machine made for sewing".

2 Related Work

The analysis and interpretation of the linguistic knowledge present in contextualized representations has recently been the focus of a large amount of work (Clark et al., 2019; Voita et al., 2019b; Tenney et al., 2019; Talmor et al., 2019). The bulk of this interpretation work relies on probing tasks which serve to predict linguistic properties from the representations generated by the models (Linzen, 2018; Rogers et al., 2020). These might involve structural aspects of language, such as syntax, word order, or number agreement (Linzen et al., 2016; Hewitt and Manning, 2019; Hewitt and Liang, 2019), or semantic phenomena such as semantic role labeling and coreference (Tenney et al., 2019; Kovaleva et al., 2019). In our work, we shift the focus from interpreting the knowledge about language encoded in the representations, to exploring the imprint of two specific transformations, passivization and negation, on word representations.

The majority of the above mentioned works address representations generated by models trained with a language modeling objective, such as LSTM RNNs (Linzen et al., 2016), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Voita et al. (2019a) propose to study the representations obtained from models trained with a different objective. We take the same stance and investigate the impact of context on representations generated by BERT, and by the encoder of neural machine translation (NMT) models involving one or more language pairs.

In order to detect the information related to the two studied transformations that is encoded in the representations, we employ methodology initially proposed for identifying and removing linguistic and other kinds of biases from representations. Such methods fall in two main paradigms: projection and adversarial methods. Projection methods identify specific directions in word embedding space that correspond to the protected attribute, and remove them. Bolukbasi et al. (2016) identify a gender subspace by exploring gendered word lists. Zhao et al. (2018) propose to train debiased word embeddings from scratch by altering the loss of the GloVe model (Pennington et al., 2014) to concentrate specific information (e.g., about gender) in a dedicated coordinate of each vector. Dev and Phillips (2019) propose a simple linear projection method to reduce the bias in word embed-

dings. Lauscher et al. (2019) develop a variation of this method that introduces more flexibility in the formation of the debiasing vector used in the projection. Adversial methods extend the main task objective with a component that competes with the encoder trying to extract the protected information from its representation (Goodfellow et al., 2014; Xie et al., 2017; Zhang et al., 2018). These models cannot, however, completely remove the protected information, and their training is difficult (Elazar and Goldberg, 2018).

Xu et al. (2017) propose a null-space cleaning operator as a privacy mechanism to minimize the exposure of confidential information in a dataset. Given a model pre-trained for a given task, they remove from the input a subspace that contains the null-space, hence removing information that is not used for the main task. Ravfogel et al. (2020) propose a similar method, Iterative Null-space Projection (INLP), for removing information regarding a certain property from representations. It is based on the mathematical notion of linear projection and is data-driven in the directions it removes, like adversarial methods. In our experiments, we repurpose the INLP method for identifying and removing traces of the passivization and negation transformations from contextualized representations.

3 Experimental Setup

In our experiments, we use contextualized representations generated by the BERT language model and two Transformer-based machine translation models (Section 3.1). We generate representations for words in two datasets with sentence pairs illustrating passivization and negation (Section 3.2). We focus on the main verb, and the nouns found in subject and object positions in the sentence pairs. We study the effect of the transformations on the representations using binary classification and iterative nullspace projection (Section 3.3).[4]

3.1 Contextualized Representations

We obtain BERT representations using `bert-base-uncased` (Devlin et al., 2019), a pre-trained language model that consists of 12 layers with 768 dimensions on each layer. We also extract representations from machine translation models involving one or more language pairs. We use a bilingual English-to-German model (which we call **MT: EN > DE**) and a model with two languages, German and Greek, on the target side (**MT: EN > DE+EL**). The latter is trained using language flag tokens in the spirit of Johnson et al. (2017). We, however, feed the flags to the decoder instead of encoder. This way, we avoid the risk that the encoder is influenced by the target language and force the model to create more generic abstractions. For the two MT models, we use Transformer architectures trained on a multiparallel subset of the Europarl dataset (Koehn, 2005), spanning ≈ 400,000 aligned sentences (Mareček et al., 2020), with the following parameters: 6 layers in the encoder and in the decoder, 16 attention heads, 512 as the dimension of the encodings, and 4,096 as the feed-forward network inner dimension.

3.2 Data

We explore the traces that the PAS transformation leaves on word representations using a dataset automatically created with the templates proposed by Prasad et al. (2019).[5] The PAS sentence pairs generated by Prasad et al. (2019) in their original study, contain relative clauses and are often syntactically very complex (e.g., *the obnoxious manager that was astonished by the interesting jobs trusted the modest receptionists last month*).[6] To reduce complexity and focus on the phenomenon of interest, we modify the templates to generate PAS sentence pairs without relative clauses (e.g., *the obnoxious manager was astonished by the interesting jobs*). 1000 PAS sentence pairs are generated in this manner. We call this dataset TEMPL-PAS.

We also use sentence pairs from the SICK (Sentences Involving Compositional Knowledge) dataset (Marelli et al., 2014b).[7] The SICK dataset has been obtained through crowdsourcing and illustrates lexical, syntactic and semantic phenomena that compositional distributional semantic models are expected to account for. PAS is one of the meaning preserving alternations in SICK, where a sentence S2 results from the passivization of an active sentence S1. We use all the 276 sentence

[4] Our code and data are available at `https://github.com/Helsinki-NLP/Syntactic_Debiasing`

[5] The code is available at `https://github.com/grushaprasad/RNN-Priming`.

[6] The complexity of the sentences also resulted in numerous syntactic analysis errors when we tried to parse them using Stanza (Qi et al., 2020).

[7] The dataset was used in SemEval 2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment (Marelli et al., 2014a).

pairs (i.e., total of 552 sentences) in SICK that illustrate the PAS transformation, and call this dataset SICK-PAS.

For exploring negation, we again generate 1,000 sentence pairs with the Prasad et al. (2019) templates, inserting negation to produce contradictions. We call this dataset TEMPL-NEG. We also use the 400 sentence pairs illustrating negation in the SICK dataset, which we call as SICK-NEG.

We distinguish nouns in subject and object positions, and call the main verb of the sentence with the label VERB. In the passivization examples, we compare nouns in subject position of active sentences with the corresponding noun in agent position of the passive sentence and label them as A-SUBJ/P-AG. Furthermore, we compare nouns in subject position of the passive examples with the nouns in object position of the corresponding active sentence, and label them as A-OBJ/P-SUBJ. In the negation examples we compare nouns in the same position and label them as SUBJECT or OBJECT.

We parse both datasets with the Stanza parser (Qi et al., 2020) to obtain the dependency trees, from which we extract the elements for our comparison.

3.3 Method

A straightforward approach for measuring the effect of the studied transformations on the contextualized word representations is to train a binary classifier to detect in which sentence variants (active/passive, affirmative/negated sentence) the word occurred. For this purpose, we form training and test sets (%70 and %30 of the SICK-PAS, SICK-NEG, TEMPL-PAS and TEMPL-NEG datasets) by grouping the noun and verb instances occurring in corresponding sentence pairs into two contrasting classes (e.g., active vs. passive). For a fair evaluation of the classifier performance, we make sure to preserve a lexical split between the training and test portions of the datasets, by grouping all instances of a specific word in one set (either train or test).

A successful classification on the test set shows that the representations encode informative features describing each variant (active vs. passive or affirmative vs. negative). The debiasing methods discussed in Section 2 are suitable for neutralizing such features. Here, we utilize Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020). Given a set of vectors $x_i \in \mathbb{R}^d$ and corresponding discrete attributes Z, $z_i \in \{1, ...,\text{k}\}$, the goal is to learn a transformation $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$, such that z_i cannot be predicted from $g(x_i)$. The method is based on iteratively (1) training a linear classifier to predict z_i from x_i, followed by (2) projecting x_i on the null-space of the classifier, using a projection matrix $P_{N(W)}$ such that $W(P_{N(W)}x) = 0$ $\forall x$, where W is the weight matrix of the classifier, and $N(W)$ is its null-space. Through the projection step in each iteration, the information detected by the trained linear classifier is removed from the representation. The procedure continues until the attempt to train a linear classifier on the projected data becomes unsuccessful. As a result of the procedure, one also obtains a projection matrix, $P = P_{N(W_m)}PN(W_{m-1})...P_{N(W_0)}$, which is the multiplication of all the null-space projections applied in all steps. This projection matrix P can then potentially be applied to uncleaned data in a single step to reproduce the effect of the whole operation.

The features used by the classifiers may be very low-level, based on specific words or their role in the sentence. Such features are not very interesting as they are easily overfitted to the particular types of sentences in the training data. By testing the same features on a second dataset, we can measure if they are abstract enough to be generalizable. Specifically, we apply the trained INLP projection to the second dataset, then train a new classifier on it. If the new classifier is able to predict the sentence variant, this means that the projection is specific to the first dataset, and is thus not useful for removing information relevant for this distinction from the second dataset.

4 Results

In this section, we present various analyses of the original data and the effects of the transformations on contextualized word representations. First, we provide a visualization of embeddings before and after null-space projection. Next, we study the classification results which demonstrate the success of INLP and, finally, we investigate the impact of the neutralization procedure on word similarity. We also provide evidence regarding the generalization capability of the algorithm and the projections it discovers. In all results, with the exception of visualizations, we report the average of 20 runs.

4.1 Visualization

One of our main goals is to explore the extent to which grammatical variation is encoded in contextualized representations. Visualization is a useful

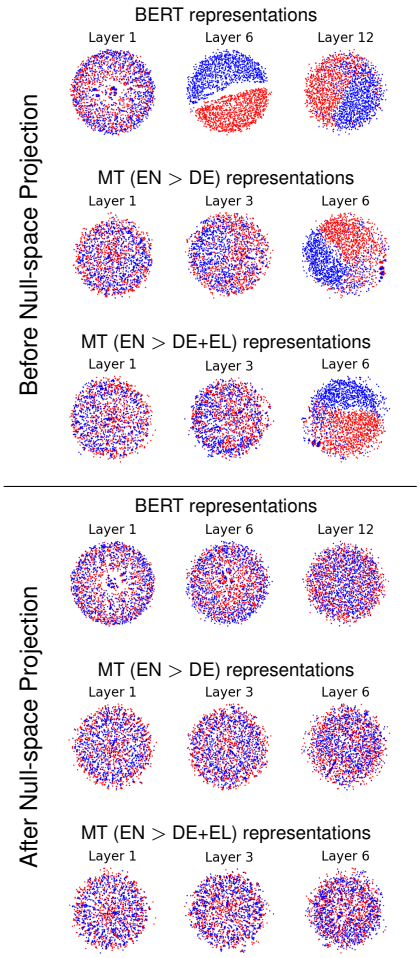

Figure 2: Multidimensional scaling (MDS) visualization of verbs in TEMPL-PAS. We show the word representations before (top part of the figure) and after INLP cleaning (lower part). The columns from left to right refer to the bottom, middle, and top layers of the encoder.

tool for demonstrating the division of the representational space into different regions in controlled examples. We use multidimensional scaling (MDS) to show the impact of the variation on the encodings. MDS reveals the level of similarity of individual points in a dataset in terms of their pairwise distance. Our data points are the contextualized representations of words in the sentences. Figure 2 reflects the distinction between active and passive verb instances present in the TEMPL-PAS dataset.

The top part of Figure 2 shows how the original representations are distributed. The separation between instances of the two classes seems almost linear, especially in the top layer of the models. For BERT, this is also the case for the middle layer (layer 6). The lower part of the figure shows that after the INLP procedure, the active and passive instances are no longer visually separable.

For nouns in corresponding thematic roles in the active and passive sentences, the situation is similar except for the BERT-based representations. Figure 1 includes the plots for the top layer of each model, and the nouns reflecting the agent and theme in corresponding sentences. The separation between active and passive examples is clear in MT models but quite fuzzy when using BERT.[8] However, the following section on classification-based results reveals that, even in this case, the distinction is still clearly present and can effectively be detected and removed by INLP.

4.2 Classification

We also explore how easy it is to correctly assign different instances in the two classes using a logistic regression classifier with inverse L2 regularization strength of 0.001.[9] We conduct this experiment on the original data using two iterations of the INLP procedure. This shows the amount of information relevant to this distinction in the original dataset that is still present after null-space projection.

Table 1 shows a successful classification of the TEMPL dataset before INLP for both transformations and all used grammatical categories, with the accuracy dropping to ≈ 0.5 by Iteration 2. This demonstrates that all representations explicitly encode the features that are altered by the PAS and NEG transformations, and that INLP can effectively remove them from the representations. This is especially informative for the BERT-based representations for nouns, a distinction that was not apparent from the visualization experiment discussed previously. The results for the SICK dataset are similar and available in the Appendix.

4.3 Similarity Estimation

We explore the similarity of individual word instances and how it is affected by the INLP neutralization procedure we apply. We study this effect on each of the encoder layers, and provide a comparison of four different measures to illustrate the impact of INLP on the embeddings. The first two metrics measure the distance between the classes C_1 and $C_2 \in C$ corresponding to our transformation variants, and we expect them to go down due to the neutralization procedure. Two additional

[8] The full picture is available in the Appendix including MDS plots for SICK-PAS and NEG transformations.

[9] Selected from among options of {0.1, 0.01, 0.001, 0.0001} to optimize the generalization of the classifier.

		Active-Passive						Positive-Negative					
		VERB		A-SUBJ/P-AG		A-OBJ/P-SUBJ		VERB		SUBJECT		OBJECT	
		It-0	It-2	It-0	It-2	It-0	It-2	It-0	It-2	It-0	It-2	It-0	It-2
BERT	L-1	0.99	0.50	1.00	0.50	0.99	0.50	0.99	0.49	0.86	0.50	0.77	0.50
	L-6	1.00	0.49	1.00	0.50	1.00	0.50	1.00	0.50	0.98	0.50	0.88	0.50
	L-12	0.99	0.50	0.99	0.50	0.95	0.50	1.00	0.50	0.92	0.50	0.90	0.50
MT (EN > DE)	L-1	0.86	0.49	0.98	0.47	0.91	0.50	0.94	0.49	0.57	0.50	0.76	0.51
	L-3	0.87	0.49	1.00	0.49	0.96	0.50	0.94	0.51	0.66	0.50	0.77	0.50
	L-6	0.90	0.49	1.00	0.53	0.97	0.50	0.96	0.47	0.77	0.50	0.81	0.49
MT (EN > DE+EL)	L-1	0.86	0.48	0.98	0.48	0.92	0.50	0.93	0.52	0.64	0.50	0.80	0.50
	L-3	0.86	0.49	0.98	0.49	0.96	0.50	0.94	0.49	0.69	0.50	0.83	0.50
	L-6	0.91	0.49	0.99	0.49	0.98	0.51	0.97	0.47	0.78	0.50	0.85	0.50

Table 1: Classification accuracy obtained on the TEMPL-PAS and TEMPL-NEG datasets before (Iteration 0, 'It-0') and after (Iteration 2, 'It-2') application of the INLP procedure.

metrics measure the distance of instances within the same class in order to verify that INLP does not produce any unwanted side effects when modifying the representations.

The first metric computes the average **pairwise inter-class distance** and is defined as:

$$\underset{i \in S}{\text{avg}} ||x_i^A - x_i^B||, \tag{1}$$

where S is the set of sentence pairs and x_i^A and x_i^B are the embeddings of the target word w_i in sentence variants A and B (e.g., active and passive). We expect this to be high prior to neutralization, and to drop significantly afterwards. We also measure the **global inter-class distance:**

$$\underset{i \in S, C_1 \in \{A,B\}}{\text{avg}} ||{x_i}^{C_1} - \underset{j \in S, C_2 \in \{A,B\}: C_2 \neq C_1}{\text{avg}} x_j^{C_2}||, \tag{2}$$

which measures the average distance of the embedding $x_i^{C_1}$ of variant C_1 to the centroid of the corresponding word embeddings of the other variant C_2, $x_j^{C_2}$. We expect this value to also decrease after the projection, but less than the previous one since it includes distances between all data points rather than only the paired sentences.

Neutralization should not significantly affect similarities between embeddings of the same word w_i in different contexts within the same sentence variant C_k. We measure this using the **same-word intra-class distance** for instances of the same word, expecting this to stay approximately same:

$$\underset{i \in S, C_k \in \{A,B\}}{\text{avg}} ||x_i^{C_k} - \underset{j \in S: w_j = w_i, j \neq i}{\text{avg}} x_j^{C_k}|| \tag{3}$$

Finally, analogous to the global inter-class distance, we also measure the **global intra-class distance:**

$$\underset{i \in S, C_k \in \{A,B\}}{\text{avg}} ||x_i^{C_k} - \underset{j \in S}{\text{avg}}\, x_j^{C_k}||, \tag{4}$$

which computes the average distance of the embeddings $x_i^{C_k}$ to the centroid of the word embeddings of variant C_k. Again, we expect this to not decrease.

Figure 3 shows the results for the verbs and nouns in the TEMPL-PAS dataset before and after INLP.[10] In all plots, especially the MT ones, we see a significant drop in pairwise inter-class distance after INLP application, which shows the effectiveness of the procedure. As expected, global inter-class distance also shows a smaller degree of drop. On the contrary, and also as expected, we do not observe drops in same-word intra-class distance or global intra-class distance, which implies that the projection does not cause major damage to the information that needs to be preserved.

4.4 Null-space Projection Transfer

Finally, we investigate the possibility to transfer null-space projections across data sets and word classes, in order to understand how generic the features representing the targeted transformation are.

4.4.1 Transfer across Datasets

We learn a projection on the TEMPL-PAS and TEMPL-NEG datasets, and use it to clean SICK-PAS and SICK-NEG respectively. We then evaluate how well the transfer works by using the cleaned dataset to train and test a classifier. If the transfer succeeds and the projection learned on the first dataset efficiently cleans the second dataset, the classification attempt will fail because all relevant information that would be useful to the classifier would have been removed. On the contrary, if a classifier can still be successfully trained on the cleaned version of the SICK datasets, then we as-

[10] TEMPL-NEG results are available in the Appendix.

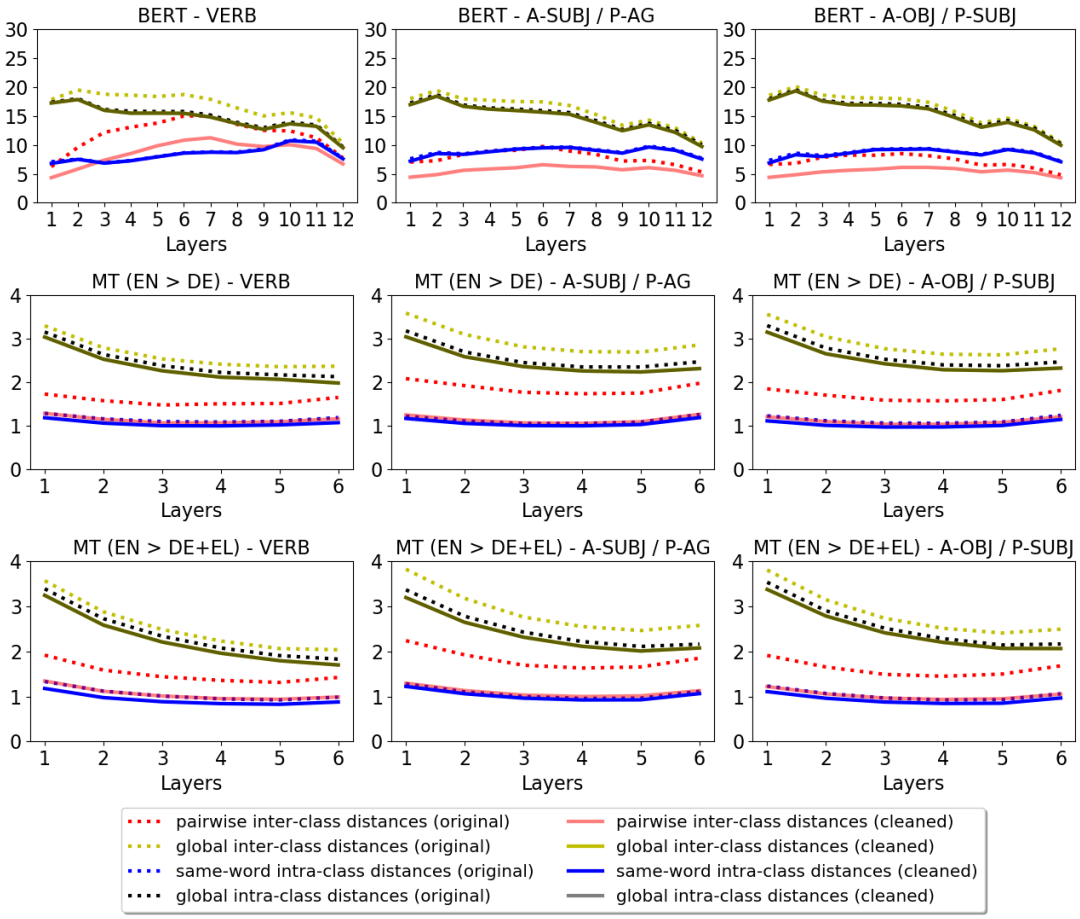

Figure 3: Average Euclidean distance for instances of nouns and verbs in the TEMPL-PAS dataset. Dashed lines show distances in the original dataset, and solid lines reflect distances after applying INLP. Distances are given for representations generated by each layer of the models.

sume that the transfer failed since information relevant to the distinction still persists.

In Figure 4, we compare (a) the classification accuracy on the original SICK-PAS and SICK-NEG datasets (dotted lines) to (b) the accuracy obtained on these datasets cleaned by using the null-space projection learned on TEMPL-PAS and TEMPL-NEG, respectively (solid lines). We report results for nouns and verbs obtained using representations generated by BERT and the MT encoders.

The transfer from TEMPL to SICK does not seem to work well with BERT representations, since a classifier trained on the cleaned SICK datasets still obtains fairly high accuracy. An exceptional to this is seen the in final layers of BERT, and subjects in the SICK-NEG dataset, for which the cleaned dataset shows slightly lower (70–90%) accurracy. For the MT representations, on the other hand, we observe low accuracies for the post-transfer classification, which suggests a successful transfer of information between the datasets. Especially for TEMPL-PAS VERB and A-SUBJ/P-AG, representations obtained with the MT model that involves two language pairs respond better to the transfer, as shown by significantly lower post-cleaning accuracies (i.e., less remaining information) than the ones obtained by the MT model with one target language. Notably, this trend is not seen for TEMPL-NEG.

4.4.2 Transfer across Grammatical Categories

We also tried to transfer null-space projection between different grammatical categories, specifically by learning the projection for verbs, subjects or objects, and then trying to apply it to one of the other two. An example of such a transfer is shown in Figure 5. Here, we apply the projection learned on verbs in the negation dataset to neutralize the same information from the noun in subject position. This seems to work surprisingly well for the MT-based representations. For BERT-based representations

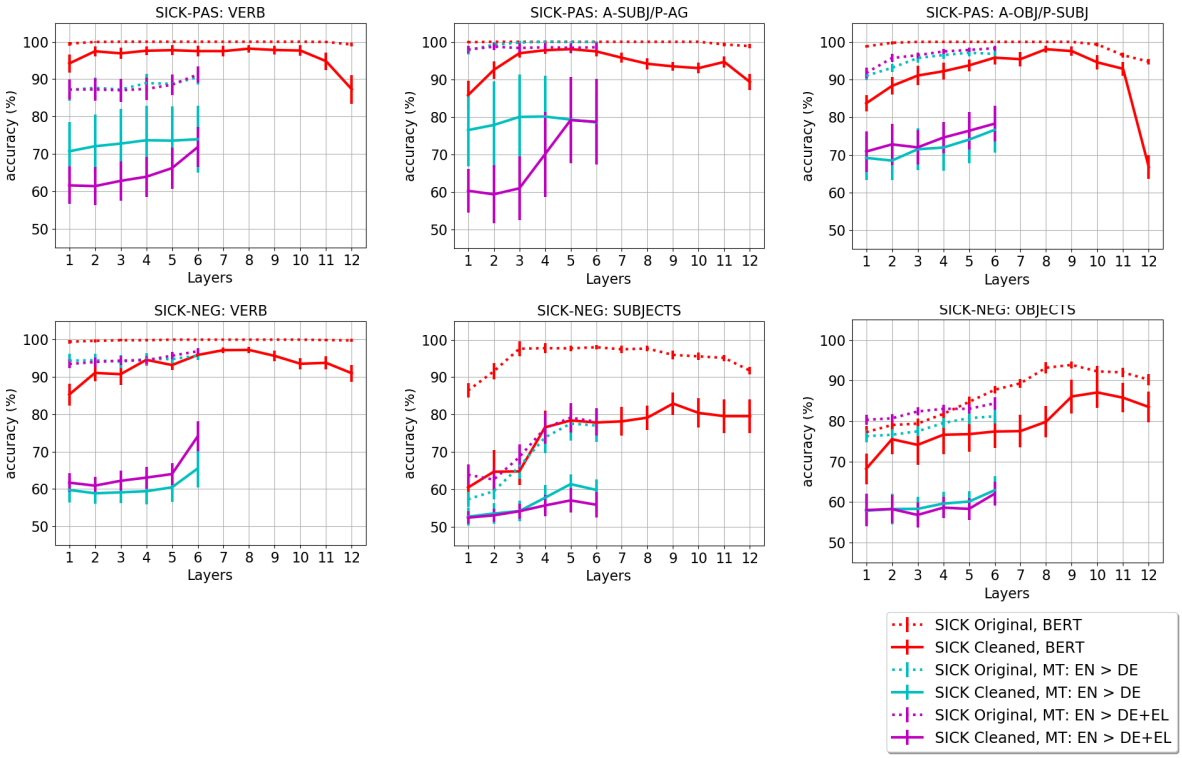

Figure 4: Classification accuracies for the SICK-PAS and SICK-NEG datasets on (1) the original version of the dataset (dotted lines) vs. (2) the cleaned version of the dataset using information from the learned INLP projection on TEMPL-PAS and TEMPL-NEG. The larger the difference between the original and cleaned versions, the more useful the transferred projection is for cleaning. Error bars indicate standard deviation of 20 experiments.

and for the passivization data set, on the other hand, the transfer across categories is not very successful with classification accuracies typically remaining above 80%. Results highlight that the information is highly specific to words of a certain grammatical category and that the projection cannot be applied as a universal neutralization procedure.

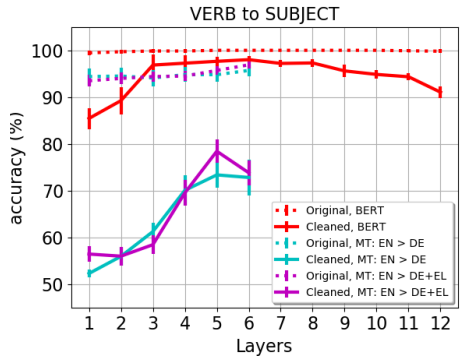

Figure 5: Classification accuracies for subjects in TEMPL-NEG on (1) the original dataset vs. (2) the dataset cleaned using learned INLP projection on verbs of TEMPL-NEG. The larger the difference between the original and cleaned versions, the more useful is the transferred projection for cleaning. Error bars indicate standard deviation of 20 runs.

5 Conclusion

We have shown that transformations such as passivization and negation leave a strong imprint on contextualized representations. We demonstrate that leveraging this information, it is possible to build classifiers that successfully identify word instances falling in either category. The traces of these transformations also affect the similarity of word instances that refer to the same entities and events. Repurposing a method initially proposed for identifying and removing societal biases from representations, we show that it is possible to neutralize the trace of such transformations from contextualized representations, and preserve the similarity of word instances having the same reference. Interestingly, the features that predict the transformation variant seem to be more generalizable in the embeddings generated by an MT encoder than in the BERT embeddings, implying that the BERT embeddings contain more surface-level information specific to each dataset.

Acknowledgements

This work has been supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 771113). We thank the reviewers for their thoughtful comments and valuable suggestions.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. volume 7, pages 597–610. MIT Press.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bertlook at? an analysis of bert's attention. In *2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. *CoRR*, abs/1901.07656.

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*, pages 4171–4186.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 503–511.

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.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. 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 4365–4374.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2019. A general framework for implicit and explicit debiasing of distributional word vector spaces.

Tal Linzen. 2018. What can linguistics and deep learning contribute to each other? *CoRR*, abs/1809.04179.

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.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 1–8.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 216–223.

David Mareček, Hande Celikkanat, Miikka Silfverberg, Vinit Ravishankar, and Jörg Tiedemann. 2020. Are multilingual neural machine translation models better at capturing linguistic features? *The Prague Bulletin of Mathematical Linguistics (in press)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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*, pages 2227–2237.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *arXiv preprint:2002.12327v1*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics – On what Language Model Pre-training Captures. arXiv preprint arXiv:1912.13283v1.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives. 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 4396–4406.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 5797–5808.

Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann. 2020. A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. *Computational Linguistics*, 46(2):387–424.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems 30*, pages 585–596.

Ke Xu, Tongyi Cao, Swair Shah, Crystal Maung, and Haim Schweitzer. 2017. Cleaning the Null Space: A Privacy Mechanism for Predictors. In *AAAI Conference on Artificial Intelligence*, pages 2789–2795.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 335–340.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

A Appendices

A.1 Visualization

Figures 6 and 7 provide the complete MDS visualizations for TEMPL-PAS and TEMPL-NEG. For TEMPL-PAS, we see a significant imprint for the nouns also. For TEMPL-NEG, the imprint is mostly visible for the verbs, however note that this does not mean the nouns are unclassifiable, since the INLP classifier is able to find a good classification for them as well (Table 1).

A.2 Classification

Table 2 shows the classification accuracies for the SICK-PAS and SICK-NEG datasets, before and after INLP. Similar to TEMPL-PAS and TEMPL-NEG results, these also show a good classification accuracy before, and a chance-level one after, demonstrating both a significant initial imprint, and the effectiveness of the INLP procedure.

A.3 Similarity Estimation

Figure 8 depicts the changes in the similarities of individual words of TEMPL-NEG using the four distance measures discussed in Section 4.3.

		Active-Passive						Positive-Negative					
		VERB		A-SUBJ/P-AG		A-OBJ/P-SUBJ		VERB		SUBJECT		OBJECT	
		It-0	It-2	It-0	It-2	It-0	It-2	It-0	It-2	It-0	It-2	It-0	It-2
BERT	L-1	0.98	0.50	0.99	0.51	0.99	0.50	0.83	0.51	0.69	0.50	0.70	0.50
	L-6	0.98	0.49	0.99	0.50	1.00	0.51	0.97	0.50	0.82	0.50	0.88	0.50
	L-12	0.98	0.50	0.96	0.50	0.82	0.50	0.92	0.50	0.80	0.50	0.89	0.50
MT (EN > DE)	L-1	0.78	0.51	0.94	0.50	0.92	0.50	0.74	0.50	0.54	0.50	0.66	0.50
	L-3	0.81	0.51	0.98	0.51	0.96	0.52	0.74	0.49	0.56	0.50	0.69	0.50
	L-6	0.82	0.54	0.98	0.52	0.97	0.53	0.84	0.51	0.64	0.50	0.72	0.50
MT (EN > DE+EL)	L-1	0.87	0.50	0.91	0.50	0.91	0.51	0.71	0.48	0.52	0.50	0.67	0.50
	L-3	0.89	0.51	0.96	0.50	0.98	0.53	0.74	0.49	0.53	0.50	0.70	0.50
	L-6	0.88	0.50	0.97	0.52	0.98	0.58	0.85	0.50	0.53	0.50	0.68	0.50

Table 2: Classification accuracy obtained on the SICK-PAS and SICK-NEG datasets before (Iteration 0, 'It-0') and after (Iteration 2, 'It-2') application of the INLP procedure.

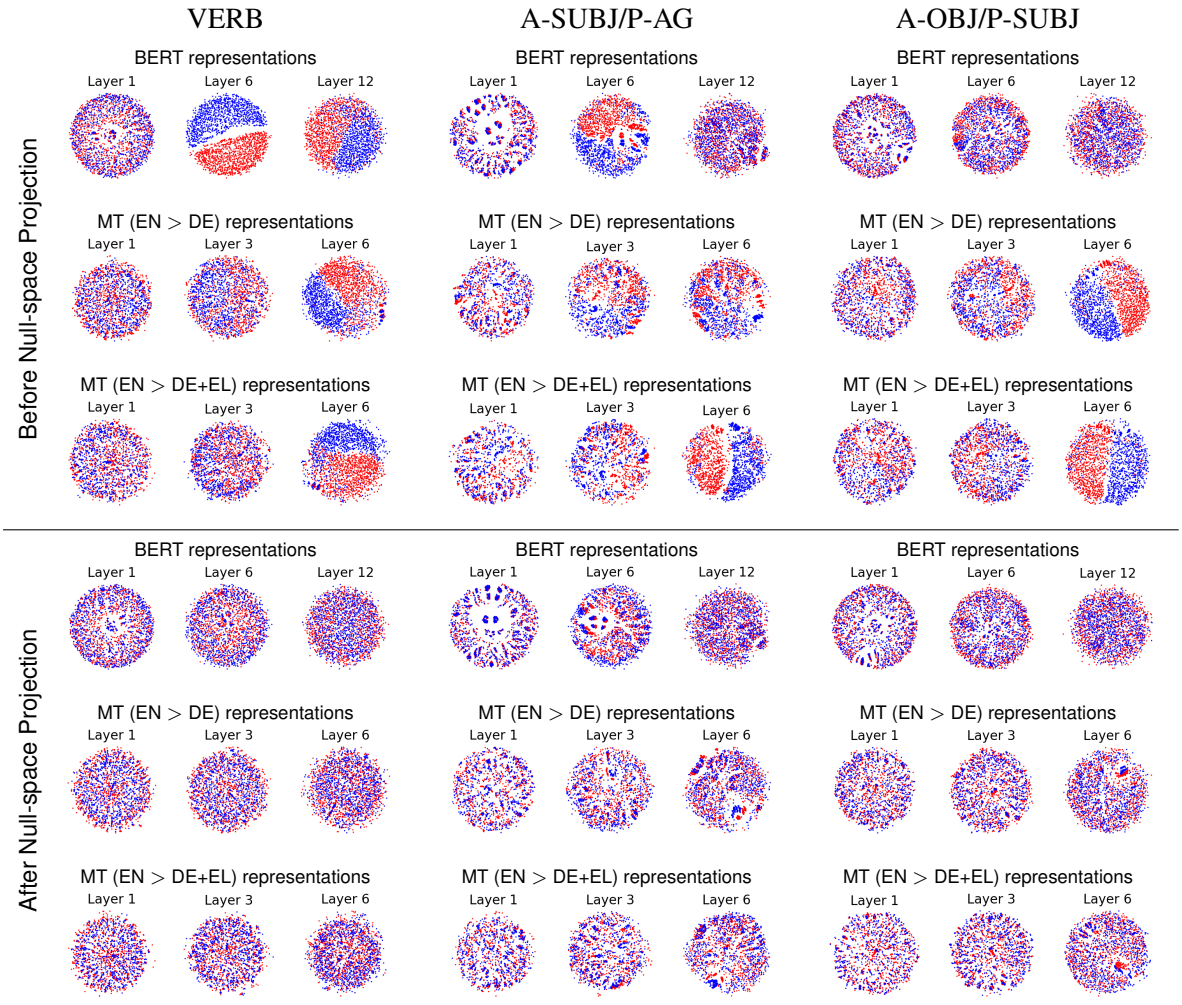

Figure 6: Multidimensional scaling (MDS) visualisation for three word instance sets in the TEMPL-PAS dataset: Verbs (Left), A-SUBJ/P-AG nouns (Middle), A-OBJ/P-SUBJ nouns (Right). The top part of the figure depicts their representations before cleaning, while the bottom part shows the same word representations after the cleaning procedure. Red and blue points indicate instances in the Active and Passive sentences, respectively.

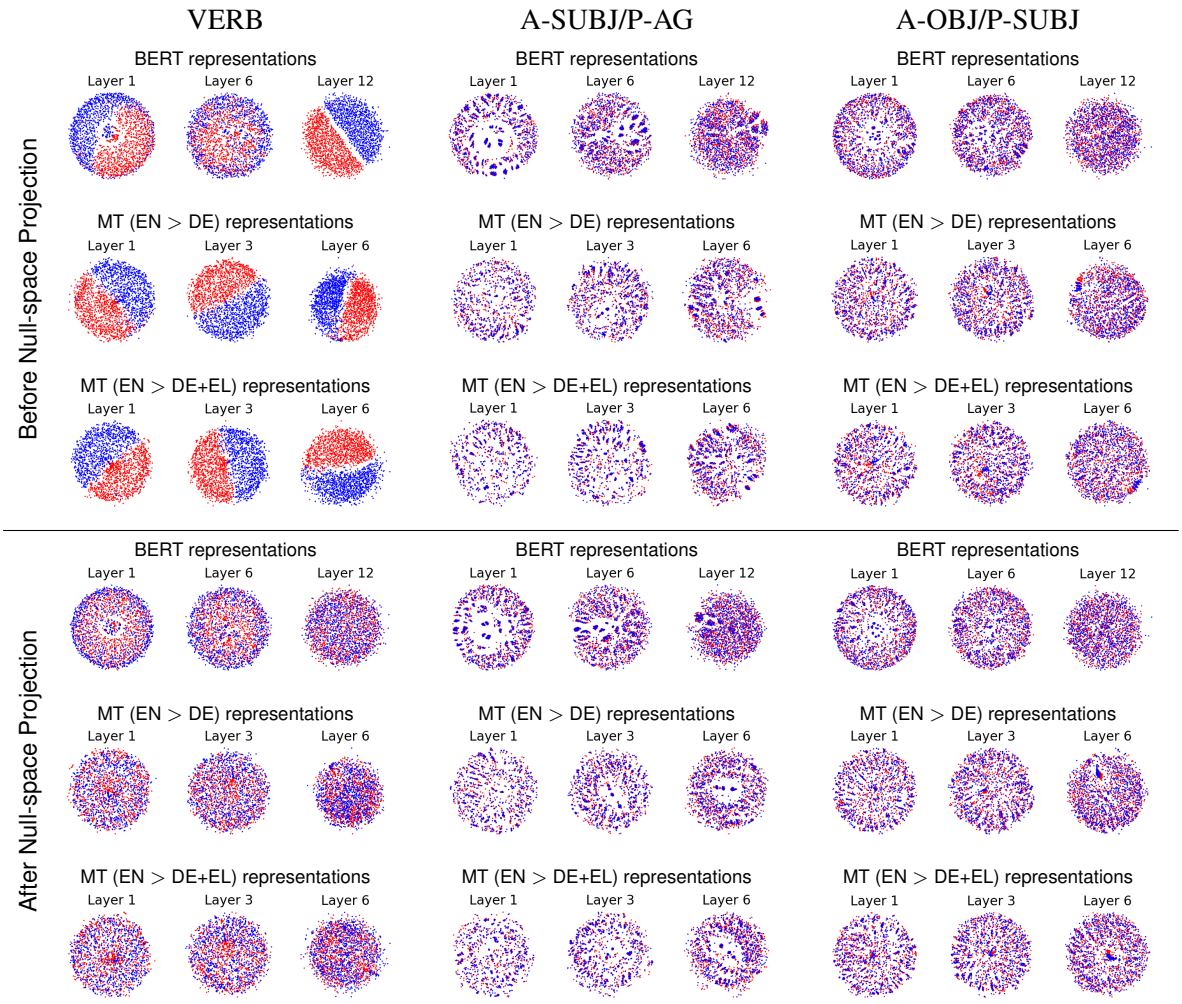

Figure 7: Multidimensional scaling (MDS) visualisation for three word instance sets in the TEMPL-NEG dataset: Verbs (Left), A-SUBJ/P-AG nouns (Middle), A-OBJ/P-SUBJ nouns (Right). The top part of the figure depicts their representations before cleaning, while the bottom part shows the same word representations after the cleaning procedure. Red and blue points indicate instances in the Active and Passive sentences, respectively.

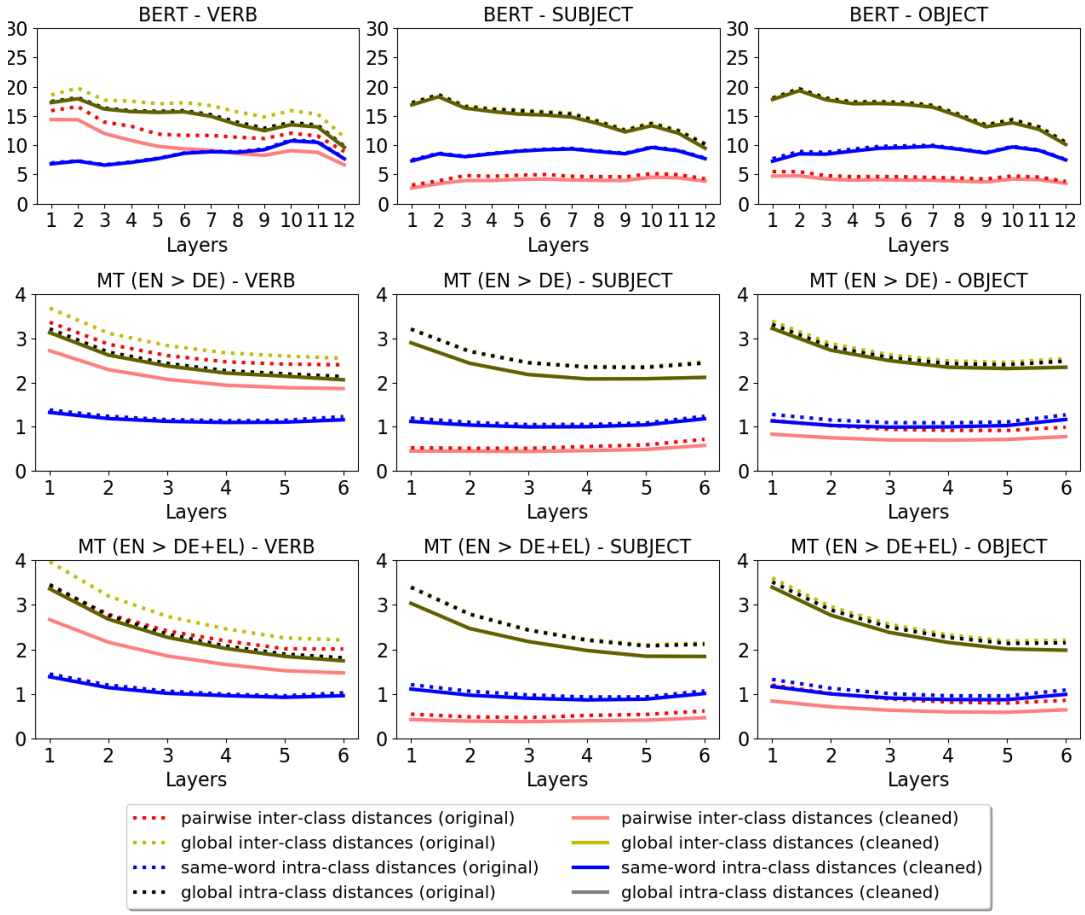

Figure 8: Average Euclidean distance for instances of nouns and verbs in the TEMPL-NEG dataset. Dashed lines show distances in the original dataset, and solid lines reflect distances after applying INLP. Distances are given for representations generated by each layer of the models.