The geometry of sentence embedding spaces is not indicative of their performance: A study of three variations of sentence representation

Anonymous ACL submission

Abstract

Transformer models learn to encode and decode an input text, and produce contextual token embeddings as a side-effect. The mapping from language into the embedding space maps words expressing similar concepts onto points that are close in the space. In practice, the reverse implication is also assumed: words corresponding to points that are close in this space are similar or related.

Does this closeness in the embedding space extend to shared properties for sentence embeddings? We compute sentence embeddings in three ways: as the averaged token embeddings, as the embedding of the special [CLS] token, and as the embedding of a random token from the sentence. We explore whether sentence embedding variations that are close in this space also have similar performance on morphology, syntax, semantic, discourse, and reasoning tasks, or whether their relative position does not offer useful clues about their relative performance and the type of linguistic information they encode.

The results show that each of the four transformer models tested – BERT, RoBERTa, DeBERTa, Electra – have their own embeddings profile, but shallow differences or commonalities between the three types of embeddings are not predictive of their performance on specific tasks. In an extreme case, Electra's [CLS] sentence embeddings and averaged token embeddings are superficially almost orthogonal, but both of them encode information about sentence chunk structure in the same way. RoBERTa's very similar sentence embeddings have very different performance on linguistic tasks. The embedding of a random token in a sentence works surprisingly well as a proxy for the sentence embedding.

1 Introduction

Projecting words and larger pieces of text into an n-dimensional space allows us to map linguistic objects into a well-defined mathematical space, with specific metrics and operations. Building this projection relies on equating word similarity in language with closeness between their corresponding vectors in the embedding space, that is, the embedding space is *smooth* (Bengio et al., 2013). The smoothness of the embedding space is assumed to work both ways: similar or related words or sentences will be projected to points that are close in the space, and words or sentences corresponding to points that are close in the space are similar or related.

Understanding the embedding space, or rather, what it means from a linguistic point of view, is difficult. On the one hand, the embedding space was shown to be anisotropic, with most words appearing in a narrow cone in this space, thus making the cosine similarity often used to estimate word similarity or relatedness seemingly less informative (Timkey and van Schijndel, 2021; Cai et al., 2021). On the other hand, the relative position of words in the embedding space was shown to encode sentence structure (Manning et al., 2020). The dimensions of sentence embeddings – as the embedding of the special [CLS] token or averaged token embeddings – were shown to have a few highly correlated groups, that mostly encode shallow information about the sentences, such as length or extreme word frequencies within the sentence (Nikolaev and Padó, 2023b).

This paper adds a few pieces to the embeddings puzzle, by studying properties of three different representations for sentences: the averaged token embeddings (S_{AVG}), the embedding of the special [CLS] token (S_{CLS}), and a random token embedding ($S_{T_{rand}}$). We establish first the geometry of the sentence embedding space – specifically their relative positions – and then investigate their linguistic properties by answering these questions:

- how different are the three representations ob-

tained from a transformer-based pretrained model?

- how do these representations change with changes to a model's training regimen, optimization process and other internal changes?
- what kind of information does each type of representation encode?
- are the relative positions of the embedding variations informative of their performance on specific tasks? In particular, do embeddings that are close in the embedding space lead to similar performance on linguistic tasks, and viceversa, do embeddings that are very distant lead to very different performance on the same tasks?

To investigate these issues we use four pretrained models from the BERT family: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021) and Electra (Clark et al., 2020) and a dataset of sentences. To establish the geometry of the sentence embedding space we perform an analysis based on the cosine similarity between the corresponding embedding vectors. For testing on linguistic tasks we use the FlashHolmes benchmark (Waldis et al., 2024), which has subsets for different types of linguistic and reasoning tasks. This provides an insight into the degree to which each type of representation encodes morphological, syntactic, semantic, discourse and reasoning information. Finally, we test whether we can reconstruct a sentence's chunk structure from each of these sentence representations.

The results show that contextual token embeddings include much contextual information (Sections 3.2, 3.3), to the degree that the embedding of a random token from a given sentence is useful even for semantic, discourse and reasoning tasks (Section 4). The relative positions of the three types of sentence representations change with the model (Section 3.3). RoBERTa shows the highest (superficial) consistency, which is not reflected in the FlashHolmes tasks, as the S_{CLS} representation has a much lower performance than S_{AVG}. For Electra and DeBERTa S_{CLS} is almost orthogonal to S_{AVG}, however they both display close performance on the FlashHolmes tasks, and also on deeper probing for syntactic structure (Sections 4 and 5). These seemingly contradictory results suggest the hypothesis that information encoded in the three types of embeddings consists of superposed layers, some of which are consistently encoded across the different representation types. We confirm this in experiments that show that all three variations of the sentence representations encode information about a sentence's chunk structure in the same way (Section 5).

2 Word and text representations in the embedding space

The evolution of the embedding space Procedurally and scale-wise, we have come a long way from the first distributional models of language inspired by Harris (1954) and Firth (1957), with tens of thousands of symbolic dimensions computed over a small (relative to what is used today) corpus (Schütze, 1992). Different methods to compute occurrence scores have been used – binary, absolute numbers, normalized scores, tf-idf. To account for similarity of dimensions, they have been clustered (Pantel and Lin, 2002; Blei et al., 2003), or reduced using singular value decomposition (Furnas et al., 1988) to perform latent semantic analysis (Landauer and Dumais, 1997), principal component analysis (Jolliffe, 2002), latent Dirichlet allocation (Blei et al., 2003).

Landauer and Dumais (1997) proposed a neural network view of their process, with a 3 layer network. Layers 1 and 2 encode the $word \times dimensions$ matrix, and layers 2 and 3 encode the $dimension \times text$ matrix. This was a theoretical exercise. Practically, Bengio et al. (2003) used a neural network to encode the probability function of word sequences in terms of the feature vectors of the words in the sequence. The vector representations of words are learned together with the parameters of the probability function. The word representations were only "method-internal". Pretrained word embeddings, as these representations have come to be known, have become the norm starting with the representations obtained through the skip-gram and continuous bag-of-words techniques proposed by Mikolov et al. (2013b,a). These word embeddings have been shown to encode several types of syntactic and semantic information, which manifest as regularities in the relative position of words in the low-dimensional vector space: plurals, derivations, analogies, and so on (Ethayarajh et al., 2019).

The latest variation are contextual embeddings obtained with transformer-based models. Models

from the BERT family (Devlin et al., 2019) work at the token level, and produce not only token embeddings, but also sentence representations as the embedding of a special [CLS] token.

The geometry of the embedding space The picture of the embedding space is complex. Mimno and Thompson (2017); Timkey and van Schijndel (2021); Cai et al. (2021) show that the embedding space is apparently anisotropic, with most points falling within a narrow cone. This is considered problematic, because it means that the space is not used properly. This influences cosine similarity measures, often used both in training the models and in the fine-tuning or task learning steps, which in this case overestimate the similarity between their corresponding words. Timkey and van Schijndel (2021) explain the phenomenon in terms of the existence of a few dominant dimensions, that can skew the similarity profile of the space. Cai et al. (2021) show that despite this shallow anisotropy, the embedding space actually contains isotropic clusters and lower-dimensional manifolds that reflect word frequency properties.

Nikolaev and Padó (2023b); Manning et al. (2020); Tenney et al. (2019) use the embeddings to uncover properties of the embedding dimensions, and information they may encode. Token embeddings were shown to encode sentence-level information (Tenney et al., 2019) (with better results when representations from multiple layers is mixed), including syntactic structure – reflected as relative positions in the embedding space that parallel a syntactic tree (Hewitt and Manning, 2019), even in multilingual models (Chi et al., 2020). Nikolaev and Padó (2023b) find that the [CLS] token embeddings have a few highly correlated groups of dimensions, that mostly encode shallow information about the sentences (sentence length, hapaxlegomena in the sentence).

Deeper exploration through probing showed that predicate embeddings contain information about their semantic roles structure (Conia and Navigli, 2022), embeddings of nouns encode subjecthood and objecthood (Papadimitriou et al., 2021), and that syntactic and semantic information can be teased apart (Mercatali and Freitas, 2021; Bao et al., 2019; Chen et al., 2019) and so can semantic roles (Silva De Carvalho et al., 2023). Probing can have issues: learning a classifier for a task does not guarantee that the model uses the targeted information (Hewitt and Liang, 2019; Belinkov, 2022; Lenci, 2023). To address this issue, Michael et al. (2020) introduce latent subclass learning, where a binary classification task has a pre-classification multiclass logistic regression step that helps probe for emergent information.

3 Sentence representation comparisons in the embedding space

As the work in exploring and probing embeddings shows, the interplay among embedding dimensions is complex, as each can contribute to various linguistic features in different measures (Bengio et al., 2013; Elhage et al., 2022). We investigate this interplay for three variations of sentence representations: averaged token embeddings, the embedding of the special [CLS] token, the embedding of a random token from the sentence.

3.1 Sentence representations

Averaged token embeddings: $\mathbf{S}_{AVG}$ The representation obtained by averaging a sentence's tokens (without the special [CLS] and [SEP] tokens) is the most frequently used representation of a sentence (Nikolaev and Padó, 2023a). Representing sentence embeddings as averages over token embeddings is justifiable as the learning signal for transformer models is stronger at the token level, with a much weaker objective at the sentence level – e.g. next sentence prediction (Devlin et al., 2019; Liu et al., 2019), or sentence order prediction (Lan et al., 2019).

The embedding of the special [CLS] token This type of representation is most commonly used after fine-tuning for specific tasks such as story continuation (Ippolito et al., 2020), sentence similarity (Reimers and Gurevych, 2019), alignment to semantic features (Opitz and Frank, 2022). Electra (Clark et al., 2020) relies on replaced token detection, which uses the sentence context to determine whether a (number of) token(s) in the given sentence were replaced by a generator sample. This training regime leads to [CLS] embeddings that perform well on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) and Stanford Question Answering (SQuAD) dataset (Rajpurkar et al., 2016), or detecting verb classes (Yi et al., 2022).

The embedding of a random token It may seem counterintuitive to use the embedding of a random token of the sentence, not even a word, as a sentence representation. This choice, however, reveals

how much contextual information each token embedding contains.

We investigate these three variations of representing sentences in four pretrained transformer models: BERT[1], RoBERTa[2], DeBERTa[3], and Electra[4]. BERT is the baseline transformer model. RoBERTa is a variation of BERT with optimized training, BPE tokenization, dynamic masking and without a next sentence objective (Liu et al., 2019). DeBERTA is another variation that introduces disentangled attention and an optimized mask decoder training (He et al., 2021). Unlike BERT, RoBERTa and DeBERTa, Electra is not a masked language model, rather implements a replaced token recognition model, predicting whether a token in the input was produced by a generator model (Clark et al., 2020). As we show in Section 3.3, these differences in the training regime and architecture of the models are reflected in the relative position of the embeddings in the embedding space.

The investigations start from shallow analyses, and move towards deeper probing of information encoded in sentence representations.

We quantify the amount of **contextual information in token embeddings** through an analysis of the pairwise (cosine) similarity between all tokens in a sentence (Section 3.2).

We study the **relative positions of the three sentence representation variations** to quantify how close they are in the embedding space (Section 3.3).

We measure **performance on linguistic tasks** to investigate the type of linguistic information each sentence representation contains, and to verify whether the similarities quantified in the previous step are reflected as similar performances (Section 4).

We probe the **encoding of phrase structure** to determine whether the three variations of sentence representation contain the same information on the chunk structure of a sentence, and if they do, whether it is encoded in the same way (Section 5).

[1] https://huggingface.co/google-bert/bert-base-multilingual-cased

[2] https://huggingface.co/FacebookAI/xlm-roberta-base

[3] https://huggingface.co/microsoft/deberta-v3-base

[4] google/electra-base-discriminator

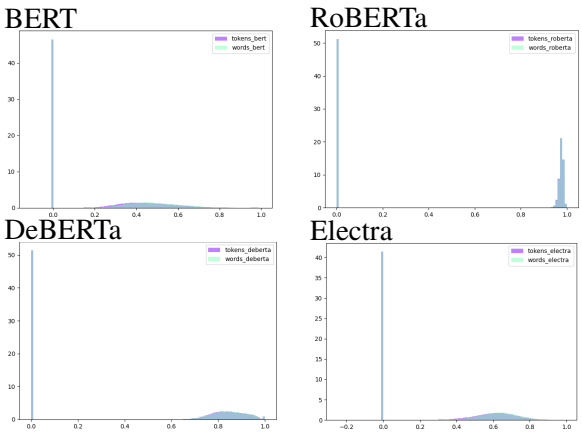

Figure 1: Histograms of cosine distances computed for words and tokens in 1000 English sentences (results for French, German, Italian, Romanian and Spanish in appendix).

3.2 Contextual information in token and word embeddings

According to the assumption that the embedding space is smooth, if two tokens in the same sentence encode much contextual information, they should be close in the embedding space. We quantify this prediction through cosine similarity distributions, which reflect pairwise comparisons of the tokens and words within each sentence. The density histogram plots are shown in Figure 1. For BERT, the similarities among the token representations have a wider distribution, while they become tighter and centered on a higher mean for the optimized BERT variations, RoBERTa and DeBERTa, and for Electra. The word embeddings – as averages of their token representations – follow similar trends. For both tokens and words we note a large out-of-distribution peak close to 0. These come (mostly) from pairings between tokens/words and punctuation marks.

3.3 Relative positions of sentence representations in the embedding space

The next step is an analysis of the distance between the three types of embeddings —token ($S_{T_{rand}}$), averaged token (S_{AVG}), sentence (S_{CLS})— for several models from the BERT family.

The dataset consists of 1000 sentences in six languages (English, French, German, Italian, Romanian, Spanish) extracted from the parallel ParaCrawl corpus (Bañón et al., 2020) (the datasets are not parallel)[5]. For each input sentence s, we obtain the output of the model, and extract the

[5] The data will be made available upon publication.

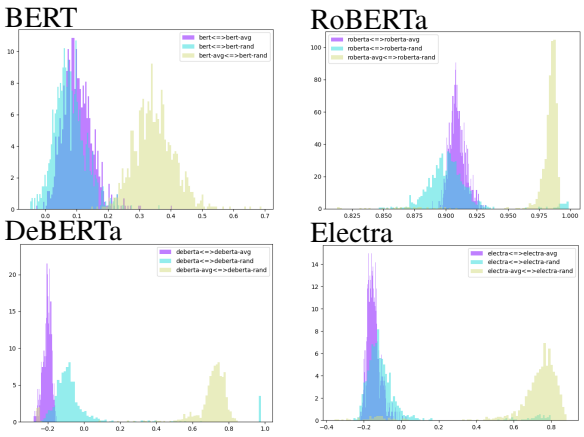

Figure 2: Histograms of cosine distances computed for 1000 English sentences (results for French, German, Italian, Romanian and Spanish in appendix). In yellow are the distances between S_{AVG} and $S_{T_{rand}}$, in blue the distances between S_{CLS} and $S_{T_{rand}}$, and in purple the distances between S_{CLS} and S_{AVG}.

embedding of the [CLS] token as s_{CLS}, the embedding of a random token t_{rand}, and we compute the averaged token embedding s_{avg}, as the average of the embeddings of all tokens of the original sentence (without special tokens, like [CLS] or [SEP]). We then compute the cosine similarities between all three pairs of embeddings, for each sentence in the datasets.

Figure 2 shows the histograms of these comparisons for the four pretrained models we consider. Different shades of the same colour indicate the type of sentence representation. This kind of analysis also shows how the sentence representations change with different training regimes and set-ups for the considered models.

For BERT, the averaged sentence embeddings (S-avg) are very similar to the embeddings of randomly picked tokens ($S_{T_{rand}}$) from the corresponding sentence. This indicates that token embeddings encode much contextual information. The holistic sentence embeddings (S_{CLS}) are quite dissimilar from both the averaged sentence embeddings, and the randomly chosen token embeddings, but slightly closer to the former than the latter. The optimized training of RoBERTa has the effect of bringing all variations of the sentence embeddings closer together. Still, S_{AVG} are closest to $S_{T_{rand}}$ with a mean very close to 1, while S_{CLS} is brought closer to S_{AVG} and $S_{T_{rand}}$. DeBERTa-s approach leads to a stronger separation of $S_{T_{rand}}$ and S_{AVG}, lowering their cosine similarity mean to around 0.6. This is also reflected in their similarities with S_{CLS}, which are more strongly separated. It is interesting to note that S_{CLS} are almost orthogonal to the S_{AVG} and $S_{T_{rand}}$. Following the assumption of the smoothness of the embedding space, this may indicate that the holistic S_{CLS} embeddings encode different types of information that the contextual embeddings. For Electra, S_{AVG} and $S_{T_{rand}}$ are very similar again, with a mean around 0.8, and their similarities with S_{CLS} are close to orthogonal.

4 Task-level comparisons

The previous analysis has shown that token embeddings encode much contextual information, and they, and the averaged token embeddings, are dissimilar from the embeddings of the special [CLS] token. We use the FlashHolmes benchmark (Waldis et al., 2024) to test the three embedding types. This benchmark consists of 216 tasks in morphology, syntax, semantics, discourse and reasoning. The results on these tasks will help determine what kind of information the three types of embedding encode, and whether the differences noted in the embedding space analysis are reflected in their relative performances.

Figure 3 presents a summary of the performance of the different sentence representation methods for each task, and on the task averages.[6]

The analysis shows that there isn't a single sentence representation method that leads to best results on all tasks. For morphology and syntax the methods using the averaged token embeddings as the sentence representation work best for most tasks – both when comparing the method for one transformer, and when considering the overall best. However, for semantic, discourse and reasoning task, this is no longer the case. For reasoning tasks in particular, it is unexpected that random token embeddings lead to higher performance on most tasks for all the transformers (although there isn't much variation among the models, indicating that they are all close to a baseline). Even for some semantics and discourse tasks, the random token embeddings have best results. The more holistic S_{CLS} embeddings have high performance particularly on semantics and discourse tasks.

In terms of performance, S_{AVG} and S_{CLS} embeddings have high and close performance for most task types, and for discourse they show a slight

[6]Detailed (task-level) results are presented in figures 9-10 and tables 1–5 in the appendix.

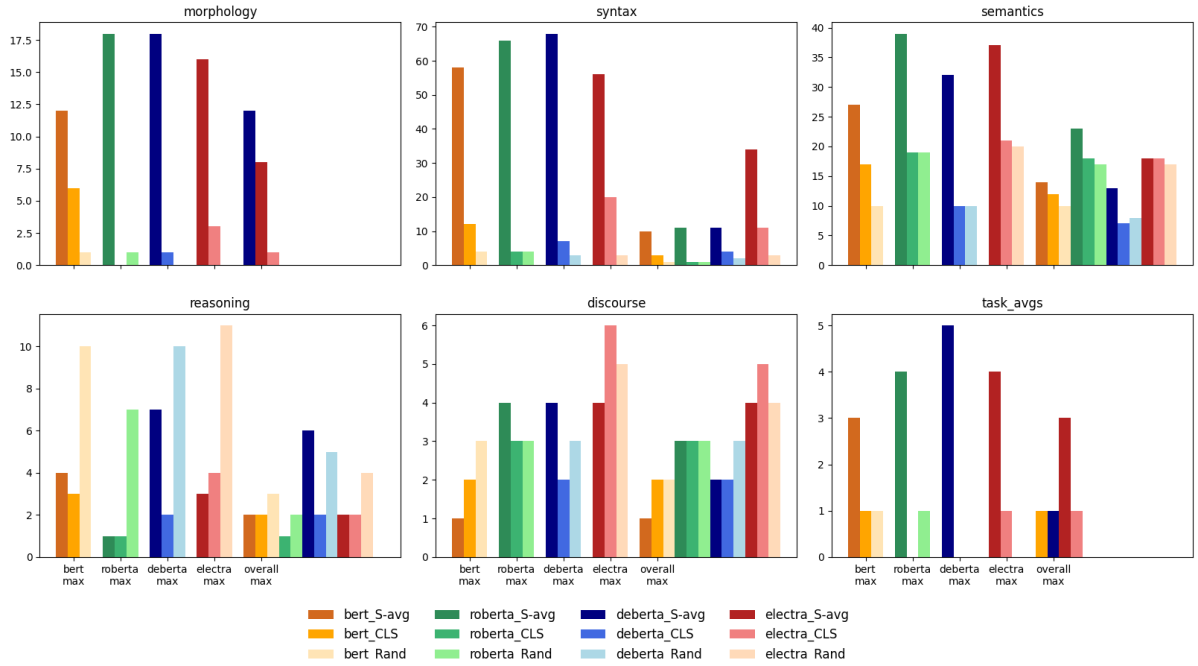

Figure 3: Statistics on the best sentence representation for each transformer, and overall for each task. The y-axis is the count of tasks for which the plotted method performs best. In case of ties we count all methods with the same score, only if not all have the same score. For each transformer, we count the methods that performed best among the transformer's variations. If all variations have the same score, we count them only if they match the highest overall scores for the task.

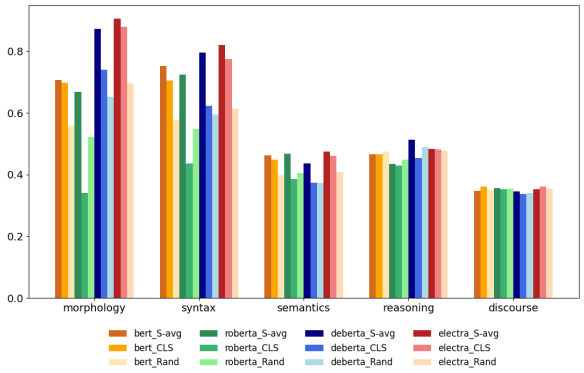

Figure 4: Comparison of embedding variations through average performance on the FlashHolmes benchmark

advantage. Compared with the analysis of the embeddings as vectors in the embedding space, this result is unexpected, as for Electra and DeBERTa in particular, the S_{CLS} and the S_{AVG} embeddings are almost orthogonal. Not only these embeddings have similar performance, but even for variations of the same task[7] S_{AVG} gives best results for one task, and S_{CLS} gives best results for the other. For RoBERTa, where the cosine similarity between these two variations is very high, their relative performance is very different.

5 Probing for structure

The previous experiments on a variety of morphological, syntactic, semantic, discourse and reasoning tasks within the FlashHolmes benchmark show very close performance on the S_{AVG} and S_{CLS} variations. In light of the analysis of the relative position of embeddings in the embedding space, these results are surprising: for Electra and DeBERTa in particular, the two representations seem to be almost orthogonal (see Figure 2). An explanation could be that the same information is encoded in a similar manner, only possibly compounded with other information which superficially obfuscates it. To investigate whether this is the case, we perform an analysis on detecting syntactic-semantic sentence structure. Nastase and Merlo (2024) have shown that some types of structural information – noun, prepositional, or verb phrase (chunks) structure – is recoverable from sentence representations. We use their code and data for the

[7] e.g. blimp_determiner_noun_agreement_with_adj_irregular_(1 and 2), blimp_irregular_plural_subject_verb_agreement_(1 and 2), blimp_principle_A_case_(1 and 2), blimp_principle_A_domain_(1 and 2)

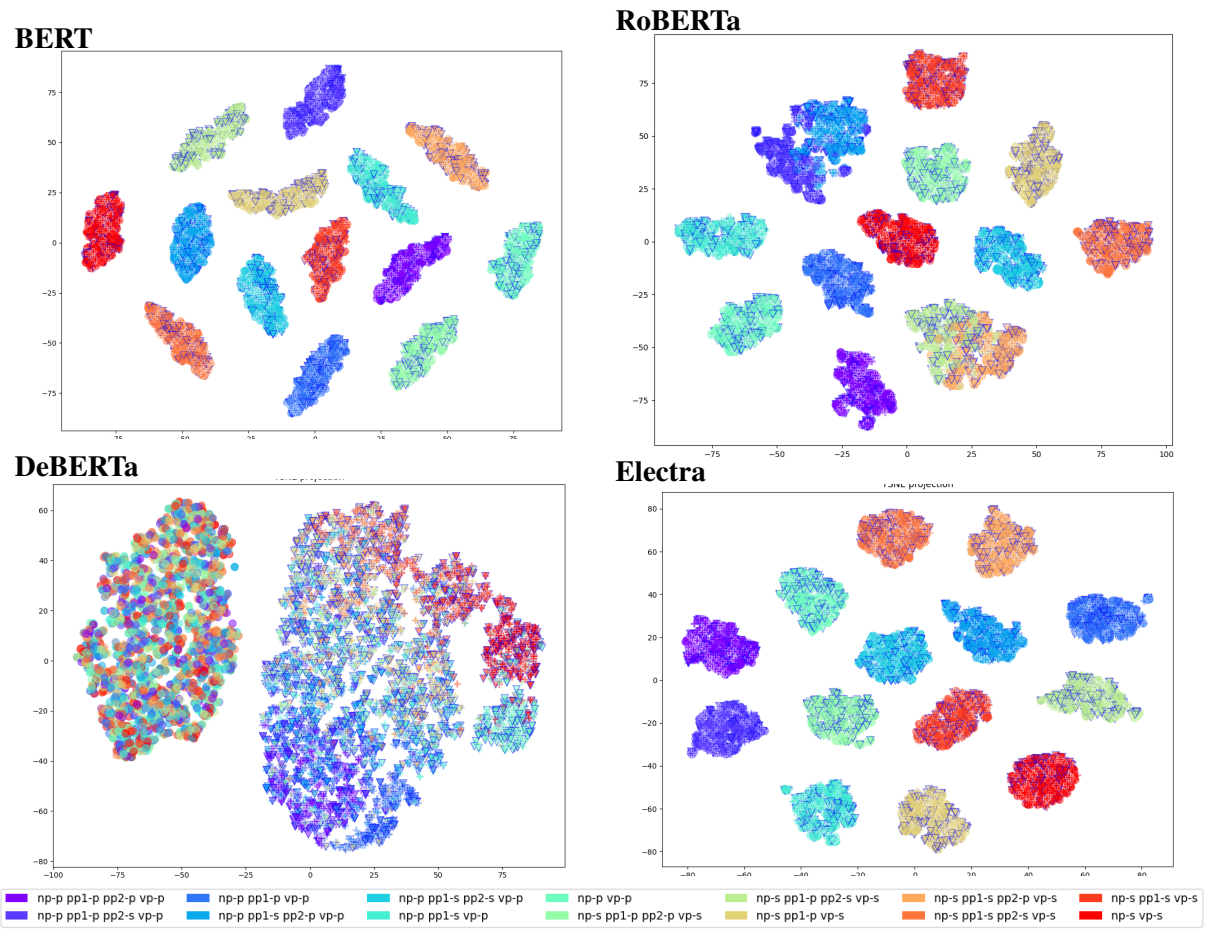

Figure 5: Comparison between models using S_{AVG} (◯), S_{CLS} (▽) and $S_{T_{rand}}$ (+) in detecting the sentence chunk structure. tSNE plots of the latent layer vectors of the training data represented using S_{AVG} , S_{CLS} and $S_{T_{rand}}$, obtained from a model trained on the S_{AVG} representation. The latent layer vectors are expected to encode the targeted information, i.e. the chunk structure. We note very sharp clusters for BERT and Electra

reported experiments[8].

The data consists of English sentences with the syntactic pattern np (pp$_1$ (pp$_2$)) vp[9], where each np, pp$_1$, pp$_2$, vp can be in the singular or plural form, and the subject (np) always agrees with the verb (vp). There are 4004 instances, evenly split across the chunk patterns.

The experimental set-up is a variational encoder-decoder, where an input sentence is decoded into a different sentence, but with the same syntactic/semantic structure. The encoder consists of a CNN layer that splits the input sentence embedding into layers of information, which it then compresses using a linear layer into a small latent representation. The decoder is a mirror image of the encoder. To encourage the sentence chunk structure to be encoded in the latent layer, each input is paired with a correct output – a different sentence but with the same chunk structure – and several (7) sentences with different structure than the input. The system does not receive information about a sentence's structure. The loss function combines the KL divergence on the latent layer, and a max-margin loss that pushes the system towards rewarding output that matches the sentence with the same structure as the input, and is maximally different from sentences with a different structure.

We apply this approach to the provided sentence data, and contrast the results when using the S_{CLS}, S_{AVG} and $S_{T_{rand}}$ sentence representations. The results tell a mixed story, shown in Figure 5. Despite high results on the syntactic and semantic Holmes tasks, detecting the chunk structure is not successful on the DeBERTa embeddings. This finding may be because of DeBERTa's optimized training, with disentangled attention matrices and token embeddings with separate position and content sections,

[8] https://github.com/CLCL-Geneva/BLM-SNFDisentangling

[9] We use BNF notation: pp$_1$ and pp$_2$ may be included or not, pp$_2$ may be included only if pp1 is included

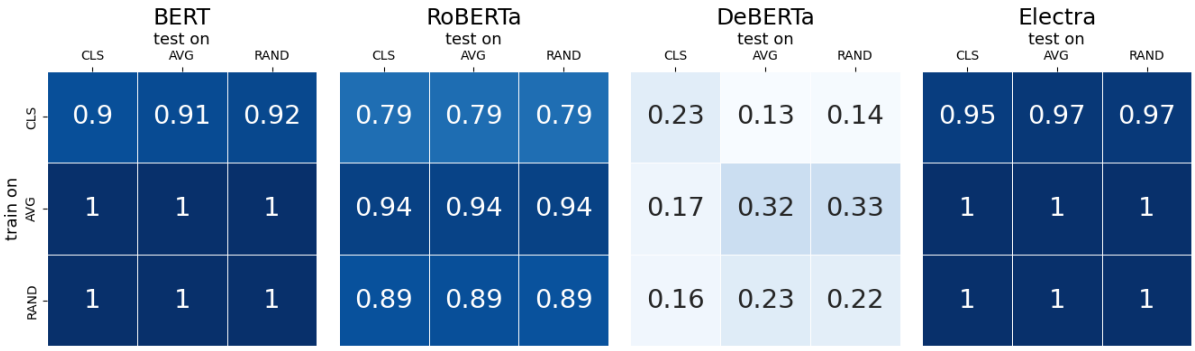

Figure 6: Comparison between models using S_{AVG}, S_{CLS} and $S_{T_{rand}}$ in detecting the sentence chunk structure in terms of average F1 scores over three runs. Detailed results in table 6 in the appendix.

which leads to a differently organized sentence embedding. BERT and Electra in particular show very high results, with results on $S_{T_{rand}}$ even higher than S_{CLS}.

For the purpose of determining whether the variations in sentence representation encode the same information in the same manner, we look at the cross-testing results – training on one representation, and testing on the others. Results are reported in Figure 6. Despite the differences revealed by the cosine similarity analysis, where for Electra the S_{CLS} representations are almost orthogonal to S_{AVG} and $S_{T_{rand}}$, these experiments show that all three representations encode information about the chunk pattern in a sentence, and moreover, this information is encoded in the same manner.

Nastase et al. (2024) have shown, through experiments on several languages, that sentence embeddings do not encode chunk structure, but rather linguistic clues – such as phrase boundaries and number information – that can be assembled into the chunk structure. Considering this, and the results in Figure 6 and the plots in Figure 5, this indicates that the S_{AVG} and S_{CLS} encode the information about phrase boundaries and number in the same manner and in the same location for BERT and Electra in particular.

6 Discussion and Conclusions

The output of pretrained language models provide embeddings for individual tokens, and a holistic sentence embedding as the embedding of a special token. A sentence is often represented through the averaged embeddings of its tokens, or through this special token embedding. In the extreme, we could even use the embedding of a random token to represent the sentence. In this work, we explored how different, or similar, these three types of representations are, and what kind of information they encode. What we found is a complex picture. Shallow analysis through cosine similarity measures shows how distinct these three representations are, and how they change relative to each other from a baseline system (BERT) with various optimizations (RoBERTa), internal organization changes (DeBERTa) or changes in the training regimen (Electra) of the system. These shallow differences or similarities are not reflected in benchmarks on five types of NLP tasks, where seemingly orthogonal representations lead to very similar results on many tasks. In a surprising twist, using the embedding of a random token as a sentence representation leads to best results on several discourse and reasoning tasks.

The close performance of the seemingly very distinct sentence representations raises another question: do they encode similar information in a similar manner, or the results come from exploiting different cues? Experiments in detecting a sentence's chunk structure – the sequence of NP/VP/PP phrases and their grammatical number attributes – showed that in fact information relevant for reconstructing this structure is encoded in the same manner, as a system trained on one sentence representation has a very similar performance when tested on the other.

The experiments presented in this paper add to the complex picture of what kind of information the embeddings induced by pretrained transformer models encode, and how. The results show that embeddings combine various layers of information, some of which is shared among all tokens in a sentence, and within the holistic sentence embedding.

7 Limitations

Synthetic data with 14 structure patterns To study the deeper question of whether the different sentence embedding variations encode sentence structure the same way, we have used a synthetic dataset, with limited variation in sentence structure,

expressed as a sequence of chunks, or phrases. In future work we plan to investigate what level of structure complexity can be recovered from these embeddings, and whether at some complexity level, differences among the embedding variations becomes apparent.

Raw output of transformer models We have focused on four pretrained models from the BERT family, and analyzed their sentence embedding space through cosine similarity, solving tasks and detecting sentence structure. We have excluded from the related work and analysis sentence transformers, which fine-tune sentence embeddings for similarity. Our aim was to study the raw output of the transformer models, and understand the properties of the different types of embeddings they induce.

Cosine similarity We reported analyses in terms of cosine similarity which is the most commonly used in the training objective. The analysis in terms of euclidean distance did not provide additional insights, so it was not included.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *J. Machine Learning Research*, 3:1137–1155.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2453–2464, Minneapolis, Minnesota. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. 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 *Proc. 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.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards Understanding Linear Word Analogies. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

J.R. Firth. 1957. *Studies in Linguistic Analysis*. Wiley-Blackwell.

George W. Furnas, Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proc. 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465–480.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced BERT with disentangled attention.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch. 2020. Toward better storylines with sentence-level language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7472–7478, Online. Association for Computational Linguistics.

Ian T. Jolliffe. 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Alessandro Lenci. 2023. Understanding natural language understanding systems. *Sistemi intelligenti, Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale*, (2/2023):277–302.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117:30046 – 30054.

Giangiacomo Mercatali and André Freitas. 2021. Disentangling generative factors in natural language with discrete variational autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3547–3556, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.

Vivi Nastase, Chunyang Jiang, Giuseppe Samo, and Paola Merlo. 2024. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In *Tenth Italian Conference on Computational Linguistics*.

Vivi Nastase and Paola Merlo. 2024. Are there identifiable structural parts in the sentence embedding whole? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 23–42, Miami, Florida, US. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023a. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 142–154, Singapore. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023b. The universe of utterances according to BERT. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 99–105, Nancy, France. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proc. 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* Edmonton, Alberta, Canada, 23-26 July 2002, pages 613–619.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hinrich Schütze. 1992. Dimensions of meaning . In *SC Conference*, pages 787–796, Los Alamitos, CA, USA. IEEE Computer Society.

Danilo Silva De Carvalho, Giangiacomo Mercatali, Yingji Zhang, and André Freitas. 2023. Learning disentangled representations for natural language definitions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1371–1384, Dubrovnik, Croatia. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *The Seventh International Conference on Learning Representations (ICLR)*, pages 235–249.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

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

David Yi, James Bruno, Jiayu Han, Peter Zukerman, and Shane Steinert-Threlkeld. 2022. Probing for understanding of English verb classes and alternations in large pre-trained language models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 142–152, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Words vs. token embedding similarities distribution

Figure 7 shows a comparison between the distribution of token and word similarities within the same sentence. A tighter distribution – as displayed by RoBERTa embeddings – indicates that all contextual embeddings are closer to each other, and thus encode more contextual information. BERT and Electra embeddings display distributions with larger standard deviation, indicating that there is more variation in the information encoded in the individual tokens and words. Electra token/word distances have a higher mean, indicating that these embeddings encode more contextual information than BERT ones. All distributions have a high spike close to 0. These pairs include punctuation and "suffix" tokens.

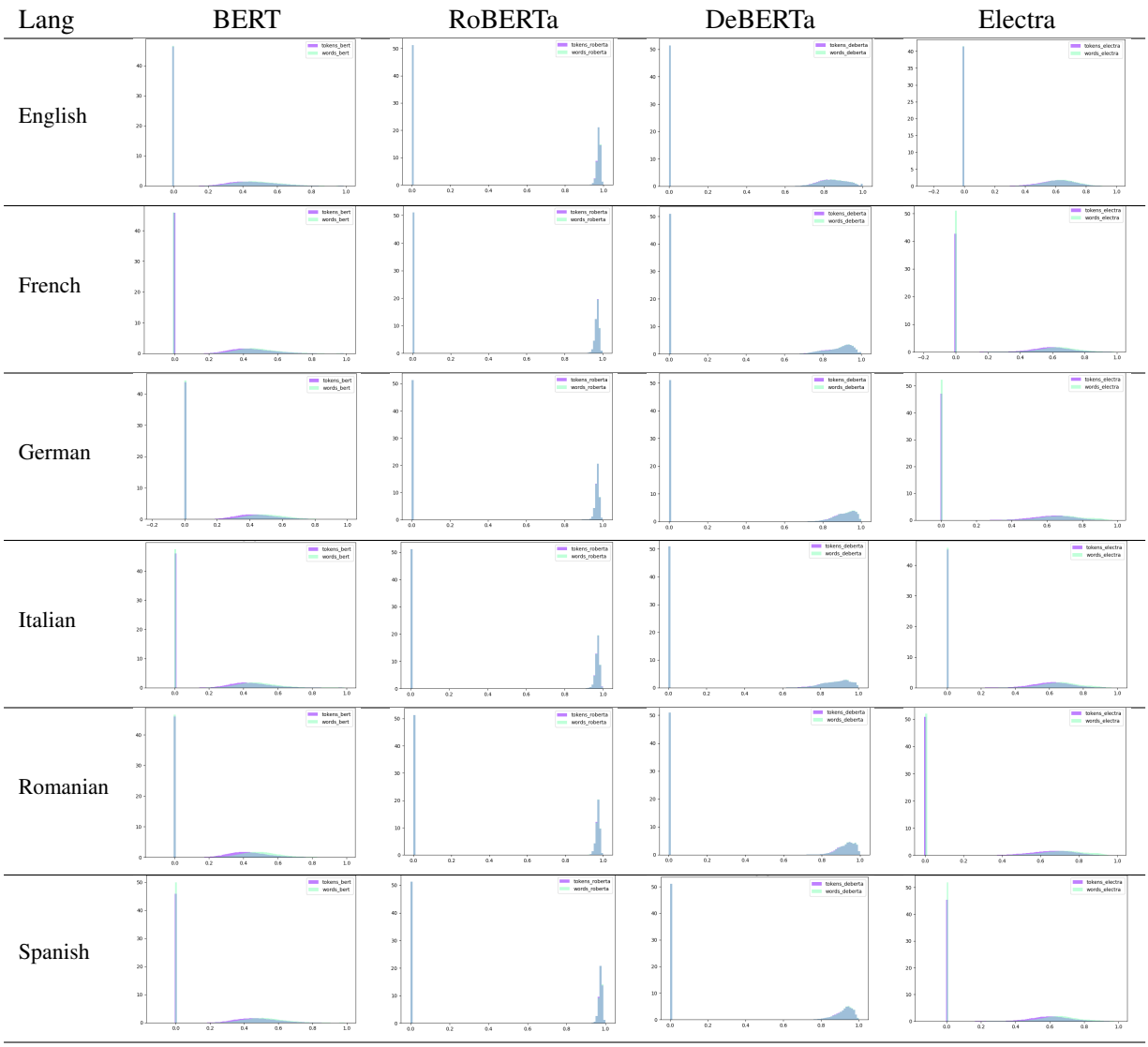

Figure 7: Cosine distances histograms computed for words and tokens from 1000 English/French/German/Italian/Romanian/Spanish.

B Multilingual embeddings comparison

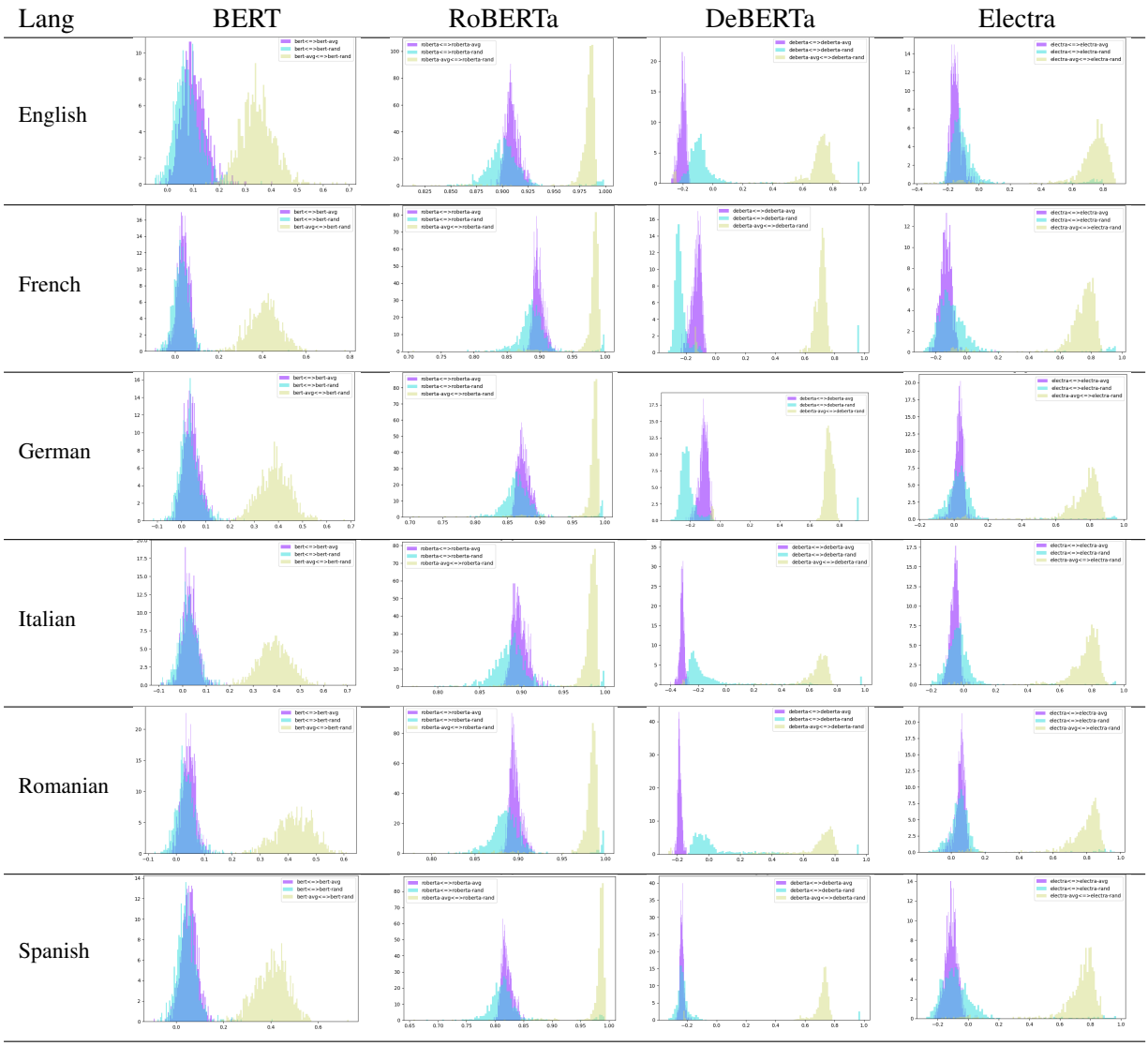

Figure 8: Cosine distances histograms computed for 1000 English/French/German/Italian/Romanian/Spanish sentences. In yellow are the distances between S_{AVG} and $S_{T_{rand}}$, in blue the distances between S_{CLS} and $S_{T_{rand}}$, and in purple the distances between S_{CLS} and S_{AVG}.

C Task results

Figure 9: Detailed results on the FlashHolmes benchmark, on morphology, syntax and semantic tasks

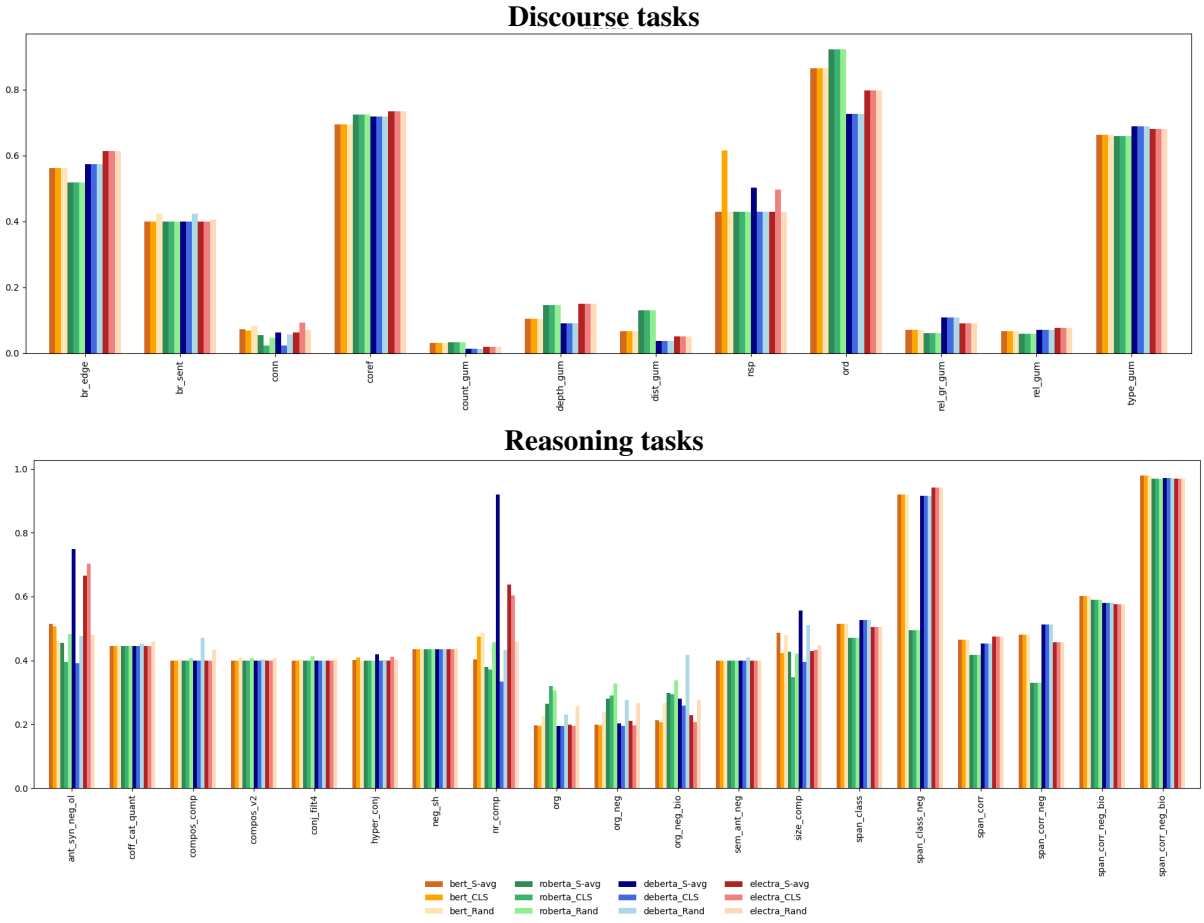

Figure 10: Detailed results on the FlashHolmes benchmark, on discourse and reasoning tasks

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-anaphor_gender_agreement	0.782	0.776	0.525	0.696	0.336	0.564	**0.939**	0.818	0.676	0.931	0.875	0.682
blimp-anaphor_number_agreement	0.782	0.727	0.586	0.825	0.345	0.582	0.954	0.835	0.717	**0.977**	0.937	0.712
blimp-determiner_noun_agreement_1	0.755	0.752	0.549	0.664	0.336	0.514	**0.974**	0.898	0.678	0.973	0.918	0.809
blimp-determiner_noun_agreement_2	0.736	0.75	0.594	0.644	0.336	0.529	**0.963**	0.922	0.677	**0.963**	0.924	0.873
blimp-determiner_noun_agreement_irregular_1	0.694	0.66	0.505	0.572	0.336	0.49	**0.842**	0.721	0.675	0.83	0.829	0.668
blimp-determiner_noun_agreement_irregular_2	0.626	0.66	0.565	0.542	0.334	0.476	**0.922**	0.857	0.768	0.884	0.885	0.758
blimp-determiner_noun_agreement_with_adj_2	0.759	0.735	0.601	0.673	0.333	0.517	0.938	0.773	0.611	**0.939**	0.873	0.620
blimp-determiner_noun_agreement_with_adj_irregular_1	0.765	0.725	0.57	0.713	0.334	0.546	**0.927**	0.82	0.656	0.909	0.879	0.613
blimp-determiner_noun_agreement_with_adj_irregular_2	0.725	0.698	0.568	0.573	0.333	0.481	**0.889**	0.775	0.670	0.857	0.864	0.627
blimp-determiner_noun_agreement_with_adjective_1	0.767	0.73	0.487	0.697	0.333	0.528	**0.929**	0.82	0.736	0.925	0.89	0.692
blimp-distractor_agreement_relational_noun	0.537	0.567	0.485	0.573	0.334	0.526	**0.868**	0.702	0.659	0.864	0.819	0.582
blimp-distractor_agreement_relative_clause	0.6	0.628	0.501	0.565	0.354	0.518	**0.746**	0.39	0.501	0.723	0.696	0.524
blimp-irregular_past_participle_adjectives	0.742	0.685	0.569	0.757	0.338	0.451	**0.808**	0.712	0.616	0.806	0.796	0.633
blimp-irregular_past_participle_verbs	0.932	0.931	0.718	0.754	0.372	0.58	0.92	0.695	0.633	**0.936**	0.884	0.736
blimp-irregular_plural_subject_verb_agreement_1	0.619	0.615	0.523	0.727	0.334	0.525	0.861	0.647	0.626	**0.875**	0.866	0.624
blimp-irregular_plural_subject_verb_agreement_2	0.651	0.716	0.577	0.764	0.349	0.608	0.892	0.591	0.667	0.908	**0.922**	0.720
blimp-regular_plural_subject_verb_agreement_1	0.779	0.785	0.588	0.739	0.341	0.544	0.885	0.785	0.711	**0.93**	0.909	0.672
blimp-regular_plural_subject_verb_agreement_2	0.75	0.746	0.637	0.804	0.354	0.49	**0.957**	0.671	0.670	**0.957**	0.938	0.747
zorro-agreement_determiner_noun-between_neighbors	0.407	0.378	0.466	0.395	0.315	0.435	0.343	0.633	0.442	**0.996**	0.983	0.937
average	0.706	0.698	0.559	0.667	0.339	0.521	0.871	0.740	0.652	**0.904**	0.878	0.696

Table 1: FlashHolmes morphology tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-adjunct_island	0.735	0.655	0.487	0.746	0.384	0.539	0.818	0.581	0.592	**0.925**	0.861	0.646
blimp-animate_subject_passive	0.743	0.663	0.531	0.665	0.346	0.531	0.709	0.577	0.527	**0.761**	0.685	0.573
blimp-animate_subject_trans	0.827	0.824	0.678	**0.828**	0.351	0.546	0.759	0.673	0.479	0.787	0.755	0.628
blimp-causative	0.714	0.661	0.544	0.605	0.36	0.501	**0.788**	0.71	0.456	0.782	0.758	0.605

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-complex_NP_island	0.652	0.634	0.478	0.65	0.341	0.467	0.791	0.566	0.589	**0.863**	0.794	0.581
blimp-coord_str_constr_complex_left_branch	0.902	0.865	0.58	0.79	0.367	0.58	0.87	0.739	0.665	**0.929**	0.831	0.724
blimp-coord_str_constr_object_extraction	0.888	0.801	0.637	0.874	0.498	0.617	0.875	0.759	0.739	**0.899**	0.85	0.476
blimp-drop_argument	0.775	0.766	0.621	0.729	0.4	0.569	0.8	0.516	0.516	0.817	**0.827**	0.599
blimp-ellipsis_n_bar_1	0.656	0.645	0.492	0.508	0.352	0.503	0.792	0.67	0.522	**0.794**	0.767	0.548
blimp-ellipsis_n_bar_2	0.723	0.72	0.543	0.664	0.34	0.45	0.813	0.589	0.46	**0.818**	0.763	0.547
blimp-existential_there_object_raising	**0.757**	0.671	0.5	0.494	0.364	0.469	0.67	0.375	0.478	0.669	0.688	0.497
blimp-existential_there_subject_raising	0.791	0.718	0.558	0.74	0.336	0.54	0.785	0.774	0.587	**0.797**	0.772	0.624
blimp-expletive_it_object_raising	0.729	0.659	0.544	0.5	0.337	0.48	0.695	0.409	0.499	0.729	**0.739**	0.532
blimp-inchoative	0.806	0.801	0.698	0.736	0.39	0.552	0.793	0.565	0.542	**0.825**	0.815	0.604
blimp-intransitive	**0.841**	0.837	0.685	0.76	0.379	0.606	0.826	0.485	0.629	0.838	0.829	0.615
blimp-left_branch_island_echo_question	0.978	0.984	0.692	0.983	0.418	0.683	0.97	0.735	0.722	**0.985**	0.95	0.835
blimp-left_branch_island_simple_question	0.921	0.821	0.63	0.904	0.405	0.62	0.937	0.771	0.777	**0.964**	0.937	0.828
blimp-only_npi_scope	0.807	0.796	0.55	0.658	0.413	0.494	0.822	0.435	0.498	**0.853**	0.795	0.559
blimp-passive_1	0.738	0.689	0.541	0.625	0.355	0.434	**0.791**	0.648	0.624	0.787	**0.791**	0.576
blimp-passive_2	0.72	0.696	0.534	0.611	0.335	0.492	0.867	0.793	0.681	**0.876**	0.859	0.691
blimp-principle_A_c_command	0.583	0.586	0.5	0.574	0.365	0.52	0.657	0.448	0.525	**0.702**	0.69	0.541
blimp-principle_A_case_1	0.944	0.86	0.654	0.826	0.336	0.577	0.949	0.913	0.657	0.939	**0.956**	0.563
blimp-principle_A_case_2	0.882	0.836	0.657	0.771	0.338	0.532	0.894	0.744	0.662	**0.921**	0.886	0.657
blimp-principle_A_domain_1	0.788	0.708	0.544	0.828	0.356	0.536	**0.862**	0.588	0.507	0.719	0.816	0.567
blimp-principle_A_domain_2	0.575	0.499	0.478	0.505	0.339	0.468	0.762	0.434	0.479	**0.877**	0.775	0.49
blimp-principle_A_domain_3	0.64	0.649	0.495	0.524	0.335	0.493	0.839	0.533	0.515	**0.877**	0.82	0.548
blimp-principle_A_reconstruction	0.948	0.88	0.58	0.752	0.333	0.557	0.808	0.467	0.584	**0.949**	0.599	0.585
blimp-sentential_negation_npi_scope	0.719	0.781	0.477	0.867	0.634	0.556	0.876	0.484	0.534	**0.931**	0.783	0.561
blimp-sentential_subject_island	0.669	0.645	0.49	0.519	0.354	0.493	0.62	0.595	0.527	**0.69**	0.645	0.534
blimp-tough_vs_raising_1	0.907	0.86	0.592	**0.908**	0.404	0.663	0.893	0.513	0.651	0.899	0.891	0.671
blimp-tough_vs_raising_2	**0.964**	0.904	0.661	0.855	0.353	0.669	0.898	0.792	0.655	0.882	0.839	0.526
blimp-transitive	0.732	0.693	0.553	0.65	0.354	0.516	0.796	0.746	0.581	**0.847**	0.774	0.641
blimp-wh_island	0.886	0.742	0.563	0.901	0.342	0.583	0.836	0.679	0.609	**0.931**	0.766	0.452
blimp-wh_questions_object_gap	0.867	0.774	0.593	0.901	0.361	0.645	0.888	0.653	0.653	**0.902**	0.779	0.607

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-wh_questions_subj_gap	0.876	0.789	0.6	**0.957**	0.402	0.635	0.9	0.771	0.657	0.919	0.844	0.63
blimp-wh_questions_subj_gap_long_dist	**0.812**	0.772	0.498	0.802	0.344	0.555	0.78	0.651	0.548	0.757	0.714	0.625
blimp-wh_vs_that_no_gap	0.885	0.813	0.649	**0.959**	0.371	0.641	0.936	0.689	0.685	0.952	0.827	0.615
blimp-wh_vs_that_no_gap_long_dist	**0.805**	0.796	0.634	0.769	0.341	0.576	0.797	0.609	0.612	0.801	0.708	0.565
blimp-wh_vs_that_with_gap	0.865	0.802	0.645	0.851	0.395	0.651	0.878	0.66	0.696	**0.908**	0.822	0.613
blimp-wh_vs_that_with_gap_long_dist	0.8	0.771	0.511	0.784	0.394	0.486	0.762	0.574	0.55	**0.849**	0.783	0.449
const	0.163	0.163	0.164	0.128	0.128	0.128	**0.186**	**0.186**	**0.186**	0.18	0.18	0.18
const_max_depth		0.815	0.668		0.78	0.658		0.812	0.539		**0.829**	0.635
const_node_length		0.879	0.721		0.89	0.73		**0.903**	0.588		0.871	0.676
context-object_number	0.705	0.691	0.609	**0.767**	0.361	0.581	0.74	0.383	0.586	0.676	0.763	0.548
context-subj_number	**0.878**	0.72	0.707	0.837	0.578	0.582	0.795	0.582	0.588	0.849	0.71	0.619
context-verb_causative	**0.842**	0.791	0.687	0.596	0.332	0.511	0.789	0.333	0.517	0.775	0.777	0.498
flesch		0.171	0.019		0.605	0.082		**0.846**	0.002		0.48	-0.074
pos	0.547	0.551	0.548	0.598	0.597	0.598	0.594	0.595	0.594	**0.646**	0.645	0.645
senteval-bigram_shift	0.841	0.828	0.786	0.85	0.804	0.706	**0.936**	0.878	0.806	0.935	0.922	0.88
senteval-obj_number	0.791	0.737	0.66	**0.799**	0.735	0.678	0.769	0.645	0.592	0.763	0.698	0.599
senteval-sentence_length	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**
senteval-subj_number	0.808	0.774	0.707	**0.813**	0.771	0.71	0.766	0.62	0.593	0.752	0.685	0.608
senteval-top_constituents	**0.398**	0.391	0.298	0.334	0.131	0.191	0.349	0.067	0.177	0.365	0.307	0.209
senteval-tree_depth	0.2	0.181	0.174	0.203	0.108	0.173	0.2	0.091	0.145	**0.209**	0.196	0.156
upos	0.415	0.415	0.415	0.434	0.434	0.434	0.388	0.388	0.388	**0.548**	**0.548**	**0.548**
xpos	0.761	0.761	0.761	0.68	0.68	0.68	0.736	0.735	0.736	**0.775**	**0.775**	**0.775**
zorro-agr_subj_v-across_prepositional_phrase	0.643	0.608	0.556	0.757	0.371	0.567	0.868	0.614	0.694	**0.971**	0.943	0.714
zorro-agr_subj_v-across_relative_clause	0.75	0.738	0.6	0.84	0.348	0.647	0.904	0.76	0.758	**0.99**	0.855	0.813
zorro-agr_subj_v-in_question_with_aux	0.429	0.383	0.447	0.417	0.333	0.423	**0.992**	0.659	0.894	0.978	0.975	0.847
zorro-agr_subj_v-in_simple_question	0.441	0.48	0.457	0.806	0.447	0.656	0.998	0.648	0.836	**1.0**	0.991	0.96
zorro-anaphor_agr-pronoun_gender	0.477	0.454	0.494	0.551	0.49	0.516	**0.991**	0.764	0.74	0.989	0.921	0.691
zorro-arg_str-dropped_argument	**0.896**	0.838	0.68	0.728	0.393	0.615	0.845	0.665	0.71	0.869	0.875	0.719
zorro-arg_str-swapped_arguments	0.974	**0.994**	0.874	0.936	0.512	0.63	0.983	0.795	0.865	0.978	0.95	0.906
zorro-arg_str-transitive	0.63	0.405	0.59	0.409	0.333	0.455	0.527	0.392	0.484	0.596	**0.754**	0.63

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
zorro-binding-principle_a	0.96	0.923	0.827	0.939	0.383	0.588	0.958	0.762	0.813	0.968	**0.97**	0.804
zorro-case-subjective_pronoun	0.91	0.793	0.669	0.895	0.493	0.643	**0.967**	0.731	0.805	0.819	0.841	0.787
zorro-ellipsis-n_bar	0.751	0.744	0.608	**0.874**	0.453	0.545	0.792	0.799	0.696	0.829	0.767	0.663
zorro-filler-gap-wh_question_object	0.959	0.848	0.769	**0.997**	0.785	0.774	0.909	0.781	0.743	0.967	0.915	0.756
zorro-filler-gap-wh_question_subject	0.99	0.904	0.827	**1.0**	0.902	0.673	0.942	0.797	0.812	0.912	0.933	0.786
zorro-island-effects-adjunct_island	0.846	0.738	0.659	0.921	0.607	0.654	0.857	0.641	0.679	0.914	**0.926**	0.799
zorro-island-effects-coord_str_constr	0.899	0.855	0.746	0.904	0.482	0.646	**0.946**	0.882	0.795	0.936	0.623	0.771
zorro-local_attractor-in_question_with_aux	0.669	0.585	0.525	0.807	0.399	0.642	0.878	0.645	0.729	**0.969**	0.908	0.807
zorro-npi_licensing-matrix_question	0.981	**0.999**	0.767	0.914	0.859	0.668	0.996	0.706	0.723	0.994	0.951	0.851
zorro-npi_licensing-only_npi_licensor	0.944	0.892	0.704	0.98	0.725	0.614	**1.0**	0.972	0.744	0.998	0.992	0.909
average	0.752	0.705	0.577	0.723	0.435	0.547	0.795	0.622	0.596	**0.819**	0.774	0.615

Table 2: FlashHolmes syntactic tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
arg-is-abstract	0.125	0.125	0.125	**0.212**	**0.212**	**0.212**	0.167	0.167	0.165	0.2	0.2	0.2
arg-is-kind	0.098	0.098	0.098	0.174	0.175	0.176	**0.209**	0.208	**0.209**	0.168	0.168	0.168
arg-is-particular	0.29	0.29	0.29	0.286	0.285	0.285	0.277	0.277	0.278	**0.359**	**0.359**	**0.359**
blimp-existential_there_quantifiers_1	0.861	0.802	0.556	**0.868**	0.402	0.634	0.859	0.741	0.627	0.841	0.775	0.512
blimp-existential_there_quantifiers_2	0.936	0.863	0.578	**0.993**	0.378	0.677	0.934	0.894	0.659	0.939	0.898	0.639
blimp-matrix_quest_npi_licensor_pres	0.995	**1.0**	0.807	0.986	**1.0**	0.862	0.973	0.704	0.733	0.989	0.948	0.691
blimp-npi_present_1	0.956	0.946	0.66	0.957	0.927	0.706	**0.973**	0.915	0.746	0.93	0.91	0.711
blimp-npi_present_2	0.953	0.954	0.702	**0.965**	0.91	0.664	0.939	0.737	0.727	0.95	0.858	0.622
blimp-only_npi_licensor_present	0.959	0.941	0.625	0.99	0.962	0.698	0.982	0.934	0.726	**0.995**	0.966	0.8
blimp-sent_neg_npi_licensor_present	0.972	0.942	0.712	0.956	0.943	0.711	**0.987**	0.877	0.699	0.985	0.966	0.652
blimp-superlative_quantifiers_1	0.932	0.773	0.626	**0.99**	0.398	0.661	0.96	0.621	0.657	0.986	0.922	0.687
blimp-superlative_quantifiers_2	0.976	0.939	0.645	**0.98**	0.378	0.612	0.888	0.738	0.643	0.863	0.723	0.606
context-object_animacy	0.765	0.894	0.8	0.753	0.339	0.736	0.918	0.339	0.71	0.955	**0.999**	0.83
context-object_gender	0.566	**0.638**	0.543	0.57	0.518	0.548	0.444	0.326	0.421	0.547	0.613	0.504
context-subject_animacy	0.942	0.756	0.752	0.814	0.346	0.679	**0.969**	0.367	0.765	0.966	0.945	0.806
context-subject_gender	**0.835**	0.77	0.671	0.75	0.744	0.594	0.533	0.367	0.501	0.496	0.566	0.484
context-verb_dynamic	0.813	**0.852**	0.691	0.724	0.334	0.542	0.742	0.43	0.54	0.743	0.623	0.519
context-verb_tense	**0.981**	0.969	0.913	0.816	0.324	0.61	0.898	0.324	0.702	0.762	0.723	0.525
cwi	0.49	0.49	0.49	0.51	0.51	0.51	0.488	0.488	0.488	**0.568**	**0.568**	**0.568**
event_structure-distributive	0.678	0.678	0.678	0.675	0.675	0.675	0.691	0.691	0.691	**0.694**	**0.694**	**0.694**
event_structure-event	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**
event_structure-has-natural-parts	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**
event_structure-has-similar-parts	0.302	0.302	0.302	0.284	0.284	0.284	0.337	0.337	0.337	**0.423**	**0.423**	**0.423**
event_structure-is-dynamic	0.364	0.364	0.364	0.363	0.363	0.363	**0.395**	**0.395**	**0.395**	0.376	0.376	0.376
event_structure-is-telic	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**
factuality	**0.466**	**0.466**	**0.466**	0.278	0.278	0.278	0.278	0.278	0.278	0.303	0.303	0.303
metaphor-lcc	0.796	0.796	0.796	0.803	0.803	0.803	**0.806**	**0.806**	**0.806**	0.798	0.798	0.798
metaphor-trofi	0.601	0.601	0.601	0.617	0.617	0.617	**0.642**	**0.642**	**0.642**	0.638	0.638	0.638
metaphor-vua_pos	**0.721**	**0.721**	**0.721**	0.719	0.719	0.719	0.712	0.712	0.712	0.708	0.708	0.708
metaphor-vua_verb	0.665	0.665	0.665	0.657	0.657	0.657	**0.678**	**0.678**	**0.678**	0.672	0.672	0.672

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
ner	**0.457**	0.456	**0.457**	0.449	0.449	0.449	0.393	0.393	0.393	0.401	0.401	0.401
passive	0.46	**0.48**	0.418	0.333	0.333	0.333	0.399	0.333	0.333	0.333	0.333	0.334
pred-is-dynamic	0.078	0.078	0.075	0.105	0.105	0.105	0.046	0.046	0.046	**0.111**	**0.111**	**0.111**
pred-is-hypothetical	0.256	0.257	0.256	0.264	**0.271**	**0.271**	0.193	0.191	0.195	0.216	0.199	0.216
pred-is-particular	0.117	0.116	**0.118**	0.076	0.086	0.086	0.074	0.071	0.071	0.102	0.104	0.103
protoroles-awareness	0.651	0.651	0.651	0.714	0.714	0.714	0.599	0.599	0.599	**0.72**	**0.72**	**0.72**
protoroles-change_of_location	0.115	0.115	0.115	**0.235**	**0.235**	**0.235**	0.141	0.141	0.141	0.197	0.197	0.197
protoroles-change_of_possession	0.238	0.238	0.238	**0.298**	**0.298**	**0.298**	0.139	0.139	0.139	0.25	0.25	0.25
protoroles-change_of_state	0.082	0.082	0.082	0.132	0.132	0.132	0.071	0.071	0.071	**0.144**	**0.144**	**0.144**
protoroles-change_of_state_continuous	0.138	0.138	0.138	**0.197**	**0.197**	**0.197**	0.107	0.107	0.107	0.179	0.179	0.179
protoroles-changes_possession	-0.203	-0.203	-0.203	-0.024	-0.024	-0.024	0.085	0.085	0.085	**0.185**	**0.185**	**0.185**
protoroles-existed_after	0.122	0.122	0.122	0.157	0.157	0.157	0.126	0.126	0.126	**0.173**	**0.173**	**0.173**
protoroles-existed_before	0.283	0.283	0.283	**0.377**	**0.377**	**0.377**	0.244	0.244	0.244	0.359	0.359	0.359
protoroles-existed_during	-0.011	-0.011	-0.011	**0.194**	**0.194**	**0.194**	0.074	0.074	0.074	0.146	0.146	0.146
protoroles-exists_as_physical	0.086	0.086	0.086	**0.304**	**0.304**	**0.304**	-0.045	-0.045	-0.045	0.079	0.079	0.079
protoroles-instigation	0.343	0.343	0.343	**0.425**	**0.425**	**0.425**	0.301	0.301	0.301	0.406	0.406	0.406
protoroles-location_of_event	-0.059	-0.059	-0.059	-0.485	-0.485	-0.485	-0.292	-0.292	-0.292	**0.184**	**0.184**	**0.184**
protoroles-makes_physical_contact	-0.035	-0.035	-0.035	**0.323**	**0.323**	**0.323**	0.149	0.149	0.149	0.146	0.146	0.146
protoroles-partitive	0.079	0.079	0.079	0.059	0.059	0.059	0.048	0.048	0.048	**0.121**	**0.121**	**0.121**
protoroles-predicate_changed_argument	**0.342**	**0.342**	**0.342**	0.099	0.099	0.099	-0.078	-0.078	-0.078	-0.175	-0.175	-0.175
protoroles-sentient	0.678	0.678	0.678	**0.748**	**0.748**	**0.748**	0.614	0.614	0.614	0.724	0.724	0.724
protoroles-stationary	-0.1	-0.1	-0.1	**0.04**	**0.04**	**0.04**	-0.154	-0.154	-0.154	-0.0	-0.0	-0.0
protoroles-volition	0.606	0.606	0.606	**0.685**	**0.685**	**0.685**	0.56	0.56	0.56	0.677	0.677	0.677
protoroles-was_for_benefit	0.327	0.327	0.327	0.36	0.36	0.36	0.259	0.259	0.259	**0.369**	**0.369**	**0.369**
protoroles-was_used	0.058	0.058	0.058	0.031	0.031	0.031	0.079	0.079	0.079	**0.105**	**0.105**	**0.105**
relation-classification	**0.362**	**0.362**	**0.362**	0.281	0.281	0.281	0.244	0.244	0.244	0.27	0.27	0.271
senteval-coordination_inversion	0.623	0.649	0.555	0.628	0.614	0.554	0.715	0.639	0.583	**0.721**	0.698	0.614
senteval-odd_man_out	0.622	0.612	0.575	0.63	0.584	0.552	**0.736**	0.7	0.619	0.73	0.714	0.661
senteval-past_present	**0.885**	0.883	0.844	0.873	0.86	0.818	0.87	0.725	0.716	0.862	0.843	0.716
senteval-word_content	**0.084**	0.04	0.02	0.02	0.0	0.014	0.006	0.0	0.002	0.006	0.004	0.005

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
sentiment-sentence	**0.425**	0.237	0.133	0.234	0.082	0.113	0.163	0.184	0.07	0.269	0.061	0.084
srl	0.114	0.114	0.114	0.105	0.105	0.105	0.118	0.118	0.118	0.125	**0.126**	**0.126**
synonym-antonym-hard	0.608	0.588	0.5	0.559	0.514	0.458	**0.63**	0.591	0.494	0.606	0.59	0.499
time	0.105	0.105	0.105	**0.106**	**0.106**	**0.106**	0.099	0.099	0.099	0.098	0.098	0.098
wordsense	**0.104**	**0.104**	**0.104**	0.084	0.084	0.084	0.071	0.071	0.071	0.086	0.086	0.086
zorro-quantifiers-existential_there	0.773	0.678	0.636	**0.905**	0.438	0.621	0.638	0.74	0.581	0.803	0.789	0.716
zorro-quantifiers-superlative	0.981	0.76	0.616	**0.997**	0.425	0.634	0.973	0.652	0.673	0.995	0.93	0.722
average	0.463	0.449	0.398	0.468	0.386	0.405	0.436	0.373	0.374	**0.474**	0.46	0.409

Table 3: FlashHolmes semantic tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
bridging-edge	0.563	0.563	0.563	0.519	0.519	0.519	0.574	0.574	0.574	**0.614**	**0.614**	**0.614**
bridging-sentence	0.4	0.4	**0.424**	0.4	0.4	0.4	0.4	0.4	**0.424**	0.4	0.4	0.405
coref	0.695	0.695	0.695	0.725	0.725	0.725	0.718	0.718	0.718	**0.735**	**0.735**	**0.735**
discourse-connective	0.073	0.068	0.082	0.055	0.023	0.047	0.064	0.023	0.056	0.064	**0.093**	0.071
gum-rst-edu-count	0.031	0.031	0.031	**0.034**	**0.034**	**0.034**	0.014	0.014	0.014	0.02	0.02	0.02
gum-rst-edu-depth	0.105	0.105	0.105	0.146	0.146	0.146	0.091	0.091	0.091	**0.151**	**0.151**	**0.151**
gum-rst-edu-distance	0.067	0.067	0.067	**0.13**	**0.13**	**0.13**	0.038	0.038	0.038	0.052	0.052	0.052
gum-rst-edu-relation	0.067	0.067	0.067	0.059	0.059	0.059	0.072	0.072	0.072	**0.076**	**0.076**	**0.076**
gum-rst-edu-relation-group	0.071	0.071	0.071	0.061	0.061	0.061	**0.109**	**0.109**	**0.109**	0.092	0.092	0.092
gum-rst-edu-successively	**0.486**	**0.486**	**0.486**	0.483	0.483	0.483	0.483	0.483	0.483	0.479	0.479	0.479
gum-rst-edu-type	0.664	0.664	0.664	0.658	0.658	0.658	**0.689**	**0.689**	**0.689**	0.68	0.68	0.68
next-sentence-prediction	0.429	**0.615**	0.429	0.429	0.429	0.429	0.502	0.429	0.43	0.429	0.497	0.429
ordering	0.865	0.865	0.865	**0.923**	**0.923**	**0.923**	0.727	0.727	0.727	0.798	0.798	0.798
averages	0.347	**0.361**	0.350	0.356	0.353	0.355	0.345	0.336	0.340	0.353	**0.361**	0.354

Table 4: FlashHolmes discourse tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
SemAntoNeg	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	**0.408**	0.4	0.4	0.4
bioscope-negation-span-classify	**0.979**	**0.979**	**0.979**	0.969	0.969	0.969	0.971	0.971	0.971	0.97	0.97	0.97
bioscope-negation-span-classify	**0.601**	**0.601**	**0.601**	0.589	0.589	0.589	0.58	0.58	0.58	0.577	0.577	0.577
bioscope-org-negation	0.212	0.207	0.266	0.299	0.294	0.338	0.279	0.259	**0.417**	0.228	0.207	0.276
fuse-negation-span-classify	0.919	0.919	0.919	0.495	0.495	0.495	0.917	0.917	0.917	**0.942**	**0.942**	**0.942**
fuse-negation-span-correspondence	0.48	0.48	0.48	0.33	0.33	0.33	**0.512**	**0.512**	**0.512**	0.458	0.458	0.458
fuse-org-negation	0.2	0.196	0.239	0.28	0.29	**0.329**	0.203	0.196	0.275	0.211	0.196	0.267
olmpics-antonym_synonym_negation	0.515	0.507	0.46	0.455	0.396	0.483	**0.749**	0.392	0.477	0.665	0.703	0.48
olmpics-coffee_cats_quantifiers	0.444	0.444	0.444	0.444	0.444	0.444	0.444	0.444	0.453	0.444	0.444	**0.46**
olmpics-composition_v2	0.4	0.4	**0.409**	0.4	0.4	0.407	0.4	0.4	0.404	0.4	0.4	0.407
olmpics-compositional_comparison	0.4	0.4	0.404	0.4	0.4	0.407	0.4	0.4	**0.47**	0.4	0.4	0.434
olmpics-conjunction_filt4	0.4	0.4	0.403	0.4	0.4	**0.412**	0.4	0.4	0.4	0.4	0.4	0.403
olmpics-hypernym_conjunction	0.401	0.408	0.4	0.4	0.4	0.4	**0.42**	0.4	0.402	0.4	0.41	0.404
olmpics-number_comparison_age_compare_masked	0.404	0.476	0.487	0.38	0.372	0.457	**0.919**	0.333	0.433	0.638	0.603	0.459
olmpics-size_comparison	0.486	0.423	0.479	0.427	0.347	0.421	**0.556**	0.394	0.51	0.429	0.434	0.447
sherlock-negation	0.434	0.434	0.436	0.434	0.434	0.434	0.435	0.434	0.435	0.434	0.434	**0.437**
speculation-org	0.196	0.195	0.224	0.265	**0.319**	0.306	0.195	0.195	0.23	0.198	0.195	0.256
speculation-span-classify	0.514	0.514	0.514	0.47	0.47	0.47	**0.526**	**0.526**	**0.526**	0.505	0.505	0.505
speculation-span-correspondence	0.465	0.465	0.465	0.417	0.417	0.417	0.453	0.453	0.453	**0.475**	**0.475**	**0.475**
average	0.466	0.466	0.474	0.434	0.430	0.448	**0.514**	0.453	0.488	0.483	0.482	0.477

Table 5: FlashHolmes reasoning tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

D Probing for structure

train on	test on	BERT	RoBERTa	DeBERTa	Electra
CLS	CLS	0.896 (0.088)	0.789 (0.027)	0.227 (0.058)	0.955 (0.006)
	AVG	0.910 (0.078)	0.793 (0.026)	0.130 (0.025)	0.971 (0.003)
	RAND	0.919 (0.070)	0.792 (0.023)	0.139 (0.019)	0.966 (0.002)
AVG	CLS	1.000 (0.000)	0.943 (0.013)	0.174 (0.020)	0.999 (0.001)
	AVG	0.999 (0.001)	0.936 (0.017)	0.325 (0.087)	0.997 (0.001)
	RAND	1.000 (0.000)	0.939 (0.018)	0.327 (0.096)	0.999 (0.001)
RAND	CLS	0.998 (0.001)	0.888 (0.009)	0.163 (0.023)	0.999 (0.001)
	AVG	0.998 (0.002)	0.895 (0.004)	0.233 (0.048)	0.998 (0.001)
	RAND	0.997 (0.003)	0.886 (0.005)	0.221 (0.048)	0.997 (0.003)

Table 6: Detailed results on detecting chunk structure in sentence embeddings. Averaged F1 scores (standard deviation) over three runs.