Sentence Ambiguity, Grammaticality and Complexity Probes

Sunit Bhattacharya[=] and **Vilém Zouhar**[=] and **Ondřej Bojar**
Charles University, Faculty Of Mathematics and Physics
Insititute of Formal and Applied Linguistics
{bhattacharya,zouhar,bojar}@ufal.mff.cuni.cz

Abstract

It is unclear whether, how and where large pre-trained language models capture subtle linguistic traits like ambiguity, grammaticality and sentence complexity. We present results of automatic classification of these traits and compare their viability and patterns across representation types. We demonstrate that template-based datasets with surface-level artifacts should not be used for probing, careful comparisons with baselines should be done and that t-SNE plots should not be used to determine the presence of a feature among dense vectors representations. We also show how features might be highly localized in the layers for these models and get lost in the upper layers.

1 Introduction

Pre-trained language models, such as BERT, M-BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019), while being very efficient at solving NLP problems, are also notoriously difficult to interpret and their analysis and interpretation is an active area of research (Belinkov and Glass, 2019). One such technique of analysis is based on probing classifiers (Belinkov, 2021), which primarily consists of training and evaluating a shallow network multi-layer perceptron (MLP) as a classifier on top of the vector representations. Probing classifiers are now fairly established in NLP (Adi et al., 2016; Tenney et al., 2019; Ma et al., 2019).

In this work, we build sentence representations from layer-wise contextual embeddings obtained from three different pre-trained language models and probe them for three linguistic traits: sentence ambiguity, grammaticality, and complexity using some well-established datasets.

In the process, we show why having a reasonable baseline is a necessity for performance interpretation. We also demonstrate why simply visually checking the clustering of embeddings on datasets using t-SNE, a popular dimension-reduction technique in probing, can lead to incorrect conclusions.

Motivation. The study of these traits is important for example in machine translation where disambiguation is necessary and grammaticality correction and simplification sometimes happen implicitly without any control. For the tasks of text simplification and grammar correction, it is crucial to be aware of whether and how general-purpose models encode these traits or whether they abstract the meaning from them. Specifically, ambiguity detection has been investigated very little in contrast to other features. All of these three traits are orthogonal in their definitions, although their mutual relationships are unknown. For example, it may be that ambiguous sentences tend to be more complex and prone to lower grammaticality. We assimilate the definition of these traits from the respective datasets but nevertheless include examples in Table 1.

Contribution. We carry out text classification tasks of ambiguity, grammaticality and complexity and demonstrate empirically that:

- having a reasonable baseline is a necessity for performance interpretation;
- sentence ambiguity is represented much less than sentence complexity in the models;
- the template-based BLiMP dataset is not suitable for probing grammaticality because of surface-level artefacts;
- t-SNE is not always an adequate tool to see whether a feature is represented in vectors.

[=] Co-first authors.

Code for the experiments in this paper is open-source: github.com/ufal/ambiguity-grammaticality-complexity

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 40 - 50

Dataset	Class	Sentence
Ambiguous COCO	Ambiguous	A metal artwork displays a clock in the middle of a floor.
MS COCO	Unambiguous	A couple sitting under an umbrella on a park bench.
HCR English	Complex	For the year, net income tumbled 61% to $ 86 million, or $ 1.55 a share.
HCR English	Simple	In part, the trust cited the need to retain cash for possible acquisitions.
CoLA	Acceptable	The sailors rode the breeze clear of the rocks.
CoLA	Unacceptable	The problem perceives easily.
BLiMP-Morphology	Acceptable	The sketch of those trucks hasn't hurt Alan.
BLiMP-Morphology	Unacceptable	The sketch of those trucks haven't hurt Alan.
BLiMP-Syntax	Acceptable	Aaron breaks the glass.
BLiMP-Syntax	Unacceptable	Aaron appeared the glass.
BLiMP-Syn_Semantics	Acceptable	Mary can declare there to be some ladders falling.
BLiMP-Syn_Semantics	Unacceptable	Mary can entreat there to be some ladders falling.
BLiMP-Semantics	Acceptable	There was a rug disappearing.
BLiMP-Semantics	Unacceptable	There was every rug disappearing.

Table 1: Sentence examples from used datasets.

2 Related Work

Ambiguity. Word-sense disambiguation has been extensively studied and is a closely related task (Navigli, 2009). This has also been the focus of work done with recent NLP tools, which has mostly concentrated on the determination of ambiguity at the lexical level and not at the sentence level. Yaghoobzadeh et al. (2019); Şahin et al. (2020); Meyer and Lewis (2020) classify ambiguous words. Chen et al. (2020) explore the geometry of BERT and ELMo (Peters et al., 2018) using a structural probe to study the representational geometry of ambiguous sentences. Bordes et al. (2019) use a combination of visual and text data to ground the textual representations and make notes on disambiguation. Ambiguity modelling has also been a focus of the MT community because translation often requires disambiguation. This applies on many levels: lexical (Higinbotham, 1991; Zou and Zou, 2017; Do et al., 2020; Campolungo et al., 2022), syntactic (Pericliev, 1984) and semantic (Baker et al., 1994; Stahlberg and Kumar, 2022). Psycholinguists have also studied the effect of ambiguity resolution on cognitive load (Altmann, 1985; Trueswell, 1996; Papadopoulou, 2005), often motivated by issues in MT (Sammer et al., 2006; Scott, 2018). Bhattacharya et al. (2022) explore ambiguity by the task of translation by human annotators.

Grammaticality. This trait has been studied historically from the perspective of human sentence processing and acceptability (Nagata, 1992; Braze, 2002; Mirault and Grainger, 2020). Many real-world applications utilize tools for automatic grammaticality prediction (Heilman et al., 2014; Warstadt et al., 2019), such as automatic essay assessment (Foltz et al., 1999; Landauer, 2003; Dong et al., 2017) or machine translation (Riezler and Maxwell III, 2006). For MT, output acceptability, or fluency, is a standard evaluation direction for which many automated metrics exist (Hamon and Rajman, 2006; Lavie and Denkowski, 2009; Stymne and Ahrenberg, 2010). In contrast to our supervised classifier approach, perplexity-based approach has been used to measure acceptability (Meister et al., 2021).

Related more closely to our setup, Hewitt and Manning (2019) use a linear probe and identify syntax in contextual embeddings. Lu et al. (2020); Li et al. (2021) examine gramaticality in BERT layers. Hanna and Bojar (2021) assess BERTScore effectiveness in spotting grammatical errors.

Complexity. Similarly to other traits, complexity was first studied in the human processing of language (Richek, 1976; Just et al., 1996; Heinz and Idsardi, 2011). Brunato et al. (2018) perform a crowd-sourcing campaign for English along with an in-depth analysis of the annotator agreement and complexity perception. Automatic complexity estimation is vital, especially in the educational setting for predicting readability (McNamara et al., 2002; Weller et al., 2020). Ambati et al. (2016) estimate sentence complexity using a parser while Štajner et al. (2017) do so using n-grams. Sarti

(2020); Sarti et al. (2021) juxtapose the effect of complexity on language models and human assessment thereof. Martinc et al. (2021) survey multiple neural approaches to complexity estimation, including using pre-trained LM representation. In contrast to our work, they report only the final results and do not investigate the issue from the perspective of probing (e.g. what representation to extract and from which layer).

Probing. Earlier probing studies have shown that the early layers of BERT capture phrase-level information and the later layers tend to capture long-distance dependencies (Jawahar et al., 2019). The syntax is also captured more in the early layers of BERT and higher layers are better at representing semantic information (Tenney et al., 2019). It is not clear if and how pre-trained models achieve compositionality (Kalchbrenner and Blunsom, 2013; Nefdt, 2020; Kassner et al., 2020) and how linguistic knowledge is represented in sentence embeddings. Liu et al. (2019) use probing on a set of tasks including token labelling, segmentation and pairwise relation extraction to test the abilities of contextual embeddings. Mutual information can be used as a viable alternative to traditional probes that require optimization (Pimentel et al., 2020). A conceptual follow-up is $\mathcal{V}$-information (Hewitt et al., 2021) which is better suited for probing. In many cases, t-SNE is the prevalent method of visualization of class clusters in high-dimensional vector space (Jawahar et al., 2019; Jin et al., 2019; Wu and Xiong, 2020; Hoyt and Owen, 2021).

3 Data

For each trait, we use a different dataset. Their overall sizes are listed in Table 2 and example sentences in Table 1. We repurpose the datasets and derive binary labels (positive/negative) from each: ambiguous/unambiguous, complex/simple and grammatical/ungrammatical.

Ambiguity. We use sentences from the MS COCO (Lin et al., 2014) dataset, for our list of ambiguous and unambiguous sentences. The MS COCO dataset comprises of a set of captions describing an image. Captions containing ambiguous verbs corresponding to 461 images (Ambiguous COCO; Elliott et al., 2016) constitute the ambiguous sentences for our experiment. 461 captions that were randomly sampled from MS COCO constituted the unambiguous sentences for the experiment.

	Dataset	Sentences
Ambiguity	COCO	0.9k
Complexity	HCR English	1.2k
	PACCSS-IT	1.1k
Grammaticality	CoLA	5k
	BLiMP	67×2k

Table 2: Number of sentences for each dataset corresponding to each trait.

Complexity. Corpus of Sentences rated with Human Complexity Judgments[1] (Iavarone et al., 2021) and PACCSS-IT (Brunato et al., 2016) contain 20 human ratings on the scale from 1 (not complex) to 7 (very complex) about sentences. We binarize these ratings and consider sentences below the average to be simple sentences and others to be complex sentences. The resulting dataset is class-balanced (complex/simple) in terms of examples (592 sentences of each class for English and 551 sentences for Italian). The average sentence length for complex and simple examples is 24.84 and 13.95, respectively for English sentences. For Italian sentences, the average sentence length for complex and simple examples is 21.61 and 12.26, respectively. The complexity could therefore be encoded solely in the sentence length.

Grammaticality. For experiments under this category, we use the Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020) and the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2019) datasets. BLiMP contains sentence pairs, one of which contains a mistake in syntax, morphology, or semantics while the other is correct. The dataset covers 67 different conditions, grouped into 12 phenomena. These phenomena are further categorized as 'syntax','morphology','syntax-semantics' and 'semantics'. The CoLA dataset is not contrastive but contains human annotations of acceptable grammaticality.

[1]English sentences were taken from the Wall Street Journal section of the Penn Treebank. Italian sentences were taken from the newspaper section of the Italian Universal Dependency Treebank.

4 Experiments

4.1 Task definition

In the following experiments, we are solving three classification tasks in parallel. The input is always the whole sentence and the output one of the two classes (ambiguous/unambiguous, complex/simple, acceptable/unacceptable), as shown in Table 1, applies to the whole sentence. The whole pipeline is also depicted in Figure 1. When using the TF-IDF feature extractor, it replaces the *pre-trained LM* block.

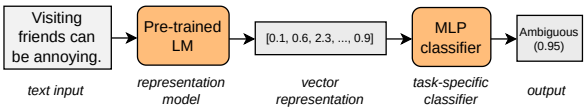

Figure 1: Example of the experiment pipeline for ambiguity classification. Ambiguous sentence from Stanley and Gendler Szabó (2000).

4.2 Setup

We use a simple MLP classifier to identify three linguistic traits from BERT (bert-base or multilingual bert-base) and GPT-2. The resulting vectors are 768-dimensional.[2] Both of these models are Transformer based models and contain 12 layers, which makes comparison convenient. We perform probing on each model separately.

- **CLS**: single vector at the `[CLS]` token.
- **Pooling:** single vector from the pooling layer.
- **Tokens:** vector representations of tokens aggregated with mean or (Hadamard) product to get a single 768-dimensional vector.

We obtain the layer-wise pre-trained model representations using Huggingface (Wolf et al., 2019) and use them to train a classifier that identifies if a sentence belongs to the positive class (e.g. ambiguous) or not. We perform a 10-fold cross-validation each with 10 runs of MLP.

Baseline. The most common class classifier (50% accuracy) is a poor baseline because it may be that the ambiguous and non-ambiguous sentences are distributed differently w.r.t. topic. In an attempt to alleviate this issue, we, therefore include as the baseline a TF-IDF-based vectorizer (with a varying number of maximum features). Probe performance of e.g. 65% would be considered at the first glance a positive result compared to 50%. However, in reality, it would be a false positive finding if a simple lexical feature extractor such as TF-IDF could yield 70%.

MLP Configuration. For probing we use `MLPClassifier` from scikit-learn 1.1.0 (Pedregosa et al., 2011) with most defaults preserved, as shown in Table 3.

Architecture	Single hidden layer (100)
Activation	ReLU
Optimizer	Adam
Learning rate	10^{-3}
Epochs	Early stopping, patience 1

Table 3: MLP classifier configuration.

4.3 Ambiguity & Complexity

Because the dataset is in Italian, we make use of multilingual BERT for both Complexity datasets. The probe performance for M-BERT is shown in Figure 2. At the first glance, it appears that the model does represent ambiguity internally since the ambiguity probe is systematically higher than 50%. However, because TF-IDF performs similarly and only uses surface-level features, the probe is very weak. This is supported by the fact that the most negative tokens from the classification (extracted from logistic regression coefficients) contained words such as *man* or *woman*, which disambiguate, based on gender, some unclear cases with an unclear referent.

In contrast, the complexity probe is systematically higher than the TF-IDF baseline. With minor exceptions, the accuracy remains high regardless of the layer. The performance for Italian (sentences taken from PACCSS-IT corpus) is identical to that for English using M-BERT (not shown). The CLS representation at layer 0 is 50% in both instances because it does not contain any information from the sentence (before the self-attention block).

4.4 Grammaticality

For the morphological task of determiner-noun agreement, Figure 3 shows a sudden drop in accuracy for the CLS representation at the 5th layer. In all the tasks concerning "Determiner-Noun Agreement", the sentence minimal pairs focus on the number agreement between the demonstrative determiners (like this/these) and an associated noun.

[2]The CLS and pooling representations apply only to BERT.

Acceptable Sentence	Unacceptable Sentence
Raymond is selling this sketch. Carmen hadn't shocked these customers.	Raymond is selling this sketches. Carmen hadn't shocked these customer.
Carl cures those horses. Sally thinks about that story.	Carl cures that horses. Sally thinks about those story.
Laurie hasn't lifted those cacti. The waitresses haven't cleaned this thesis.	Laurie hasn't lifted those cactus. The waitresses haven't cleaned this theses.
The teachers are running around this concealed oasis. Randolf buys those gray fungi.	The teachers are running around these concealed oasis. Randolf buys that gray fungi.
Cynthia scans these hard books. Jerry appreciates this lost report.	Cynthia scans this hard books. Jerry appreciates these lost report.

Table 4: Example minimal sentence pairs from the *determiner-noun* agreement task of BLiMP.

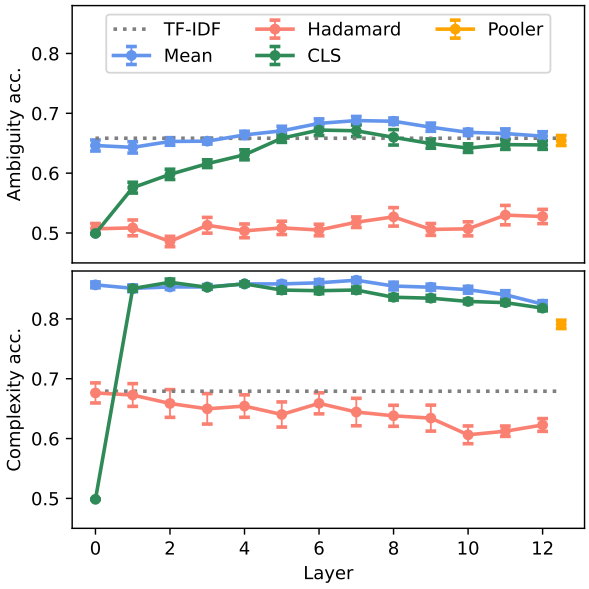

Figure 2: MLP dev accuracy for *ambiguity* and *complexity* BERT representation across layers.

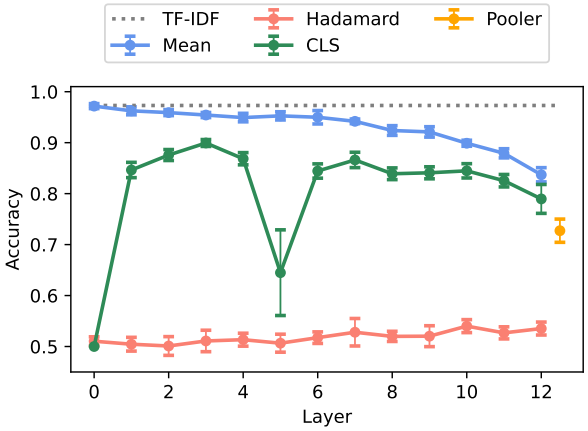

Figure 3: MLP dev accuracy for *determiner noun agreement irregular 1* task of BLiMP benchmark for BERT representation across layers. Each point is represented with a mean across 10 runs with a 95% confidence interval.

Examples of minimal pairs from the different tasks of this kind are shown in Table 4.

While the cause is unclear, it corresponds to the average norm of the representation being very low at that particular layer, making it harder for the classifier optimization.

As Figure 4 shows, many tasks can be "solved" with a simplistic TF-IDF featurizer, making them inadequate for determining the usefulness of large model representations. More adequate datasets need to be developed for probing stronger models. Systematically for all cases in morphology where the TF-IDF failed to work accurately, the performance of CLS representations was worse than the mean representations. Even in most semantics tasks, TF-IDF probes had near-perfect accuracy. For the 7 out of 26 syntactic tasks where the TF-IDF classifier was not accurate, the BERT models show a steep rise in accuracy from the 2nd/3rd layer for the mean and CLS representations, respectively. In comparison, GPT-2 does not exhibit this pattern.

5 Discussion

The experiments with ambiguity reveal that the representations of the pre-trained models do not encode the ambiguity trait well. The description detailing how the Ambiguous COCO was created (Elliott et al., 2017) states that the dataset was created with the intention of testing the capabilities of multimodal translation systems. We posit that ambiguity as a trait is not encoded in an accessi-

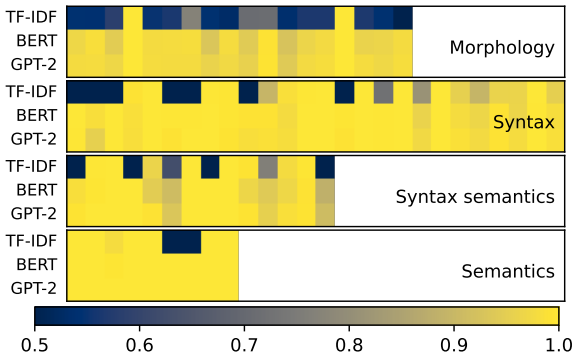

Figure 4: Accuracy on various BLiMP tasks with a max of BERT and GPT-2 representations and TF-IDF baseline. Each task+model is represented as one square. The lighter squares correspond to greater accuracy and are hence better.

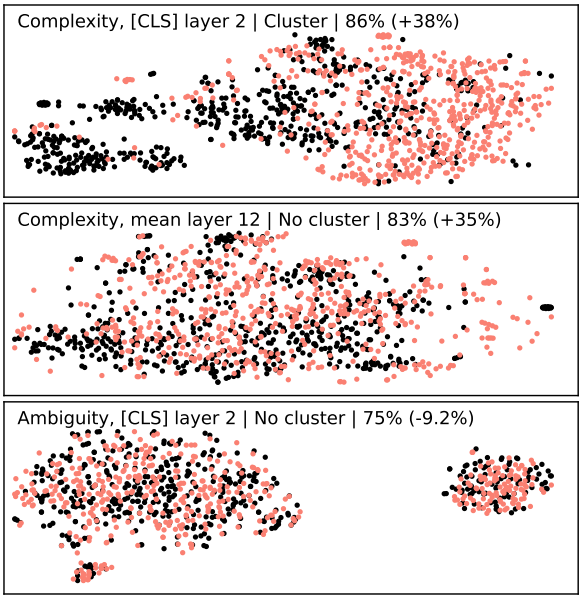

Figure 5: t-SNE projections from BERT-based embeddings. The first and the second row show high accuracy. The second and third rows show a lack of visual clusters. Red/black represent either complex/simple or ambiguous/unambiguous sentences. Percentages include classifier accuracy with the difference to the TF-IDF baseline in parentheses.

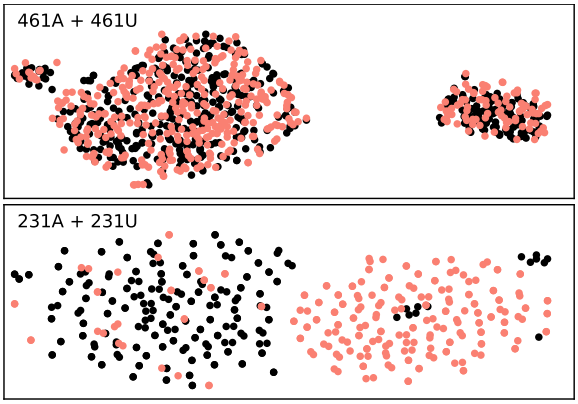

Figure 6: t-SNE projections from BERT-based embeddings (layer 1 of CLS) on ambiguous/unambiguous sentences (58% MLP and 66% TF-IDF accuracy). The first row is all the vectors and the second is half of them subsampled by Algorithm 1. Red/black represent ambiguous/unambiguous sentences.

ble way in the layer representations of pre-trained models.

For BLiMP tasks related to morphology and syntax-semantics, the accuracy goes down in the upper layers, presumably because of increasing abstraction for both models (not shown in graphs). Although we perform experiments without fine-tuning, the findings are in line with the experimental results of Mosbach et al. (2020) where finetuning on 3 tasks from the GLUE benchmark (Wang et al., 2018) showed changes in probing performance mostly in the higher layers. Fine-tuning however led to modest gains. The present setup which probes sentence representations from pre-trained models shows that the middle layers fare far better in our probing tasks than the upper layers. This leads us to posit that the features of interest are highly localized and are lost in the upper layers (even with fine-tuning).

Although both BERT and GPT-2 employ the Transformer (Vaswani et al., 2017) architecture, they have very different ways and locations for storing knowledge in their internal representations (Rogers et al., 2020; Vulić et al., 2020; Lin et al., 2019; Kuznetsov and Gurevych, 2020; de Vries and Nissim, 2021; Liu et al., 2021). The CLS representations outperform the mean representations in only a few cases. This is expected since without fine-tuning the CLS token in BERT is trained to be used for the next sentence classification tasks.

6 t-SNE Inadequacy

Given appropriate optimization and classifier, if two or more classes in a vector space form clusters, they are linearly separable and therefore the classifier performs well. Furthermore, if a classifier probe performs well and is not affected by surface-level phenomena, it means that the features are represented in the vectors. Both these statements are one-way implications:

- clear clustering → high classifier accuracy
- high classifier acc. → feature present

Because t-SNE projects vectors from high dimensional space to lower dimensions in a manner that tries to preserve distances, it may be that visual clusters are created where there were none before and vice versa. The following scenarios are possible:

- clear clusters and high classifier accuracy
- no clusters and high classifier accuracy
- no clusters and low classifier accuracy

The last combination, "clear clusters and low classifier accuracy" is impossible with proper optimization. The three scenarios on probes from the previous experiments are shown in Figure 5. The conclusion is that probes should always precede visual clustering checks using t-SNE because it may be that the data does not form clear clusters in t-SNE but the classes are still linearly separable, meaning that the feature is encoded. The last image shows two clusters but not those that separate the two classes.

A plethora of work uses t-SNE to show clusters of vectors grouped by features (Chi et al., 2020; Nigam et al., 2020; Wu et al., 2020; Zhang et al., 2021; Subakti et al., 2022), though some follow-up with reporting classifier performance. Because t-SNE visual separation is not easily quantifiable, the negative results are often underreported (Fanelli, 2012; Mlinarić et al., 2017). This issue can be resolved by using other methods, such as probes.

Algorithm 1 Forcing t-SNE clusters

▷ Vectors of sentences in the two classes
Load D_A, D_B
▷ Cluster size, e.g. $|D_A|/2$
Input c', $c \leftarrow c'/2$
▷ Two seeds from classes, most distant
$s_A, s_B \leftarrow \arg\max_{v_A \in D_A, v_B \in D_B} ||v_A - v_B||$

▷ Closest points to own seeds
$C'_A \leftarrow \text{top-}c_{\,v \in D_A} \quad -||s_A - v||$
$C'_B \leftarrow \text{top-}c_{\,v \in D_B} \quad -||s_B - v||$
▷ Furthest points to opposing seeds
$C''_A \leftarrow \text{top-}c_{\,v \in D_A} \quad ||s_B - v||$
$C''_B \leftarrow \text{top-}c_{\,v \in D_B} \quad ||s_A - v||$

$C_A \leftarrow C'_A \cup C''_A$
$C_B \leftarrow C'_B \cup C''_B$
t-SNE$(C_A \cup C_B)$

6.1 Forcing t-SNE Clusters.

It is possible to start with sentence vectors that result in a t-SNE graph that does not show any visual clusters and select half of them such that running t-SNE will show clusters between the two classes. The algorithm is described in Algorithm 1. It is based on first finding two most distant "seeds" from the two classes and then selecting vectors of the same class which are closest to the seed or most distant to the other seed.

An example is shown in Figure 6. While the original does not show any clusters between the classes, the application of the algorithm selects such vectors that t-SNE shows visual clusters. Simplu randomly subsampling the vectors would not work but this shows that using t-SNE to visually determine the presence of a feature is not robust.

7 Conclusion

In this work, we showed how large pre-trained language models represent sentence ambiguity in a much less extractable way than sentence complexity and stress the importance of using reasonable baselines. We document that template-based datasets, such as BLiMP used for sentence acceptability, are not suitable for probing because of surface-level artefacts and more datasets should be developed for probing more performant models. Finally, we discuss why using t-SNE visually for determining whether some representations contain a specific feature is not always a suitable approach.

Future work

Because both t-SNE clustering and classification (inability to establish a rigid threshold for accuracy) can fail for determining whether a specific feature is represented in the model, more robust methods for this task should be devised. These probes should also be replicated in models used for machine translation, which is the primary motivation for studying these traits.

8 Acknowledgements

This work has been funded from the 19-26934X (NEUREM3) grant of the Czech Science Foundation. The work has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ.

References

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

Gerry Altmann. 1985. The resolution of local syntactic ambiguity by the human sentence processing mechanism. In *Second Conference of the European Chapter of the Association for Computational Linguistics*.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *HLT-NAACL*, pages 1051–1057.

Kathryn Baker, Alexander Franz, Pamela Jordan, Teruko Mitamura, and Eric Nyberg. 1994. Coping with ambiguity in a large-scale machine translation system. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Sunit Bhattacharya, Věra Kloudová, Vilém Zouhar, and Ondřej Bojar. 2022. EMMT: A simultaneous eye-tracking, 4-electrode eeg and audio corpus for multi-modal reading and translation scenarios. *arXiv preprint arXiv:2204.02905*.

Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. Incorporating visual semantics into sentence representations within a grounded space. 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 696–707.

Forrest David Braze. 2002. *Grammaticality, acceptability and sentence processing: A psycholinguistic study*. University of Connecticut.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. 2020. Probing bert in hyperbolic spaces. In *International Conference on Learning Representations*.

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.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle english GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Quang-Minh Do, Kungan Zeng, and Incheon Paik. 2020. Resolving lexical ambiguity in english-japanese neural machine translation. In *2020 3rd Artificial Intelligence and Cloud Computing Conference*, pages 46–51.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233.

Daniele Fanelli. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90:891–904.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. Automated essay scoring: Applications to educational technology. In *Edmedia+ innovate learning*, pages 939–944. Association for the Advancement of Computing in Education (AACE).

Olivier Hamon and Martin Rajman. 2006. X-score: Automatic evaluation of machine translation grammaticality. In *LREC*, pages 155–160.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel R Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *ACL (2)*.

Jeffrey Heinz and William Idsardi. 2011. Sentence and word complexity. *Science*, 333(6040):295–297.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639.

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.

Dan W Higinbotham. 1991. The resolution of lexical ambiguity in machine translation. In *Deseret Language and Linguistic Society Symposium*, volume 17, page 7.

Christopher R Hoyt and Art B Owen. 2021. Probing neural networks with t-sne, class-specific projections and a guided tour. *arXiv preprint arXiv:2107.12547*.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 186–199.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89.

Marcel Adam Just, Patricia A Carpenter, Timothy A Keller, William F Eddy, and Keith R Thulborn. 1996. Brain activation modulated by sentence comprehension. *Science*, 274(5284):114–116.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182.

Thomas K Landauer. 2003. Automatic essay assessment. *Assessment in education: Principles, policy & practice*, 10(3):295–308.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is BERT surprised? Layerwise detection of linguistic anomalies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4215–4228.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual 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 1073–1094.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385*.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: augmenting bert with graph embedding for text classification. In *European Conference on Information Retrieval*, pages 369–382. Springer.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from BERT: An empirical study. *arXiv preprint arXiv:1910.07973*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Danielle S McNamara, Max M Louwerse, and Arthur C Graesser. 2002. Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Technical report, Technical report, Institute for Intelligent Systems, University of Memphis ....

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980.

Francois Meyer and Martha Lewis. 2020. Modelling lexical ambiguity with density matrices. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 276–290.

Jonathan Mirault and Jonathan Grainger. 2020. On the time it takes to judge grammaticality. *Quarterly Journal of Experimental Psychology*, 73(9):1460–1465.

Ana Mlinarić, Martina Horvat, and Vesna Šupak Smolčić. 2017. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia medica*, 27(3):447–452.

Marius Mosbach, Anna Khokhlova, Michael A Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82.

Hiroshi Nagata. 1992. Anchoring effects in judging grammaticality of sentences. *Perceptual and Motor Skills*, 75(1):159–164.

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

Ryan M Nefdt. 2020. A puzzle concerning compositionality in machines. *Minds and Machines*, 30(1):47–75.

Amber Nigam, Shikha Tyagi, Kuldeep Tyagi, and Arpan Saxena. 2020. Skillbert:"skilling" the bert to classify skills!

Despina Papadopoulou. 2005. Reading-time studies of second language ambiguity resolution. *Second Language Research*, 21(2):98–120.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vladimir Pericliev. 1984. Handling syntactical ambiguity in machine translation. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 521–524.

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

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Margaret A Richek. 1976. Effect of sentence complexity on the reading comprehension of syntactic structures. *Journal of Educational Psychology*, 68(6):800.

Stefan Riezler and John T Maxwell III. 2006. Grammatical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 248–255.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. Linspector: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Marcus Sammer, Kobi Reiter, Stephen Soderland, Katrin Kirchhoff, and Oren Etzioni. 2006. Ambiguity reduction for machine translation: Human-computer collaboration. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 193–202.

Gabriele Sarti. 2020. Interpreting neural language models for linguistic complexity assessment. Master's thesis, University of Trieste, dec.

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–60.

Bernard Scott. 2018. *Translation, brains and the computer: A neurolinguistic solution to ambiguity and complexity in machine translation*, volume 2. Springer.

Felix Stahlberg and Shankar Kumar. 2022. Jam or cream first? modeling ambiguity in neural machine translation with SCONES. *arXiv preprint arXiv:2205.00704*.

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, volume 17, pages 4096–4102.

Jason Stanley and Zoltan Gendler Szabó. 2000. On quantifier domain restriction. *Mind & Language*, 15(2-3):219–261.

Sara Stymne and Lars Ahrenberg. 2010. Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Alvin Subakti, Hendri Murfi, and Nora Hariadi. 2022. The performance of bert as data representation of text clustering. *Journal of big Data*, 9(1):1–21.

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.

John C Trueswell. 1996. The role of lexical frequency in syntactic ambiguity resolution. *Journal of memory and language*, 35(4):566–585.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018*, page 353.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Orion Weller, Jordan Hildebrandt, Ilya Reznik, Christopher Challis, E Shannon Tass, Quinn Snell, and Kevin Seppi. 2020. You don't have time to read this: An exploration of document reading time prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1789–1794.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chien-Sheng Wu, Steven CH Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Chien-Sheng Wu and Caiming Xiong. 2020. Probing task-oriented dialogue representation from language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5036–5051.

Yadollah Yaghoobzadeh, Katharina Kann, Timothy J Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753.

Xiao-Chen Zhang, Cheng-Kun Wu, Zhi-Jiang Yang, Zhen-Xing Wu, Jia-Cai Yi, Chang-Yu Hsieh, Ting-Jun Hou, and Dong-Sheng Cao. 2021. Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings in bioinformatics*, 22(6):bbab152.

Shunpeng Zou and Xiaohui Zou. 2017. Understanding: how to resolve ambiguity. In *International Conference on Intelligence Science*, pages 333–343. Springer.