Representation of ambiguity in pretrained models and the problem of domain specificity

Anonymous ACL submission

Abstract

Recent developments in pretrained language models have led to many advances in NLP. These models have excelled at learning powerful contextual representations from very large corpora. Fine-tuning these models for downstream tasks has been one of the most used (and successful) approaches to solving a plethora of NLP problems. But how capable are these models in capturing subtle linguistic traits like ambiguity in their representations? We present results from a probing task designed to test the capability of the models to identify ambiguous sentences under different experimental settings. The results show how different pretrained models fare against each other in the same task. We also explore how domain specificity limits the representational capabilities of the probes.

1 Introduction

Over the past few years, contextual embeddings have proven their worth over static embeddings (Liu et al., 2020). Pretrained models like BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) have made use of the Transformer (Vaswani et al., 2017) architecture and huge training datasets to serve as excellent base models that can be fine-tuned for multiple down-stream tasks (Qiu et al., 2020). However it is often unclear why these models work so well or what features these models learn that make them so effective. To address these questions, probing classifiers (Belinkov, 2021) are often used to analyze and interpret these models. 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. Hewitt and Manning (2019) use a linear probe to identify syntax in contextual embeddings. (Jawahar et al., 2019) show that while the earlier layers of BERT capture more phrase-level information, the later layers capture long-distance dependency information. Furthermore, (Tenney et al., 2019) show that information syntax is captured more on the earlier layers of BERT and that higher layers are better at representing semantic information. Training a shallow network multilayer perceptron (MLP) as a probe is an established technique in MLP (Adi et al., 2016; Tenney et al., 2019). Accordingly, we utilize a shallow MLP for our experiments in this work.

In this paper, we focus on the task of classifying ambiguous sentences as a probing task. We use three existing datasets to serve as our corpus of ambiguous sentences. We use the “out-of-domain (MSCOCO) test set” used in WMT17[1] (Bojar et al., 2017) that contains captions with ambiguous verbs corresponding to images. We also use the challenge test-set of the Hindi-Visual-Genome(Parida et al., 2019) (HVG) that contains sentences with ambiguous words. Sentences from HVG and COCO both contain some form of *lexical ambiguity*. Finally, we use the sentences from LAVA corpus (Berzak et al., 2015) that contain sentences with *syntactic ambiguity*. For the unambiguous sentences to feed our probe, we use the “unambiguous” sentences from the HVG and COCO corpora. For the purpose of this work, we disregard the images associated with the sentences in these corpora and just focus on the sentences itself. During the course of working with the sentences from the datasets, it was found out that many sentences contained grammatical errors. Hence, we selected a set of ambiguous and unambiguous sentences from a combination of the LAVA, COCO and HVG corpora and corrected them. An in-house annotator (a native English speaker) vetted the sentences. Then, a team of in-house annotators ranked the sentences on the basis of how “ambiguous” they seemed. We selected the top 100 sentences each of ambiguous and unambiguous types from the ranked list for experiments described later. We call this dataset “curated dataset” for the reminder of the paper.

[1] https://www.statmt.org/wmt17/multimodal-task.html

Our contributions in the paper are the following:

- Demonstrate how layer representations from different pretrained models differ for the same task, using the same data.
- Demonstrate how different "sentence representations" of models affect the performance of the probing classifiers.
- Demonstrate how probing classifiers are domain specific.

2 Related Work

Work done on this area has mostly concentrated on the determination of ambiguity at the lexical level and not at the sentence level. Yaghoobzadeh et al. (2019) use a probing task to classify ambiguous words. Şahin et al. (2020) also use probing tasks for token-level and type-level identification of ambiguities. 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. (Meyer and Lewis, 2020) use density matrices to model-word level ambiguity. Bordes et al. (2019) have used a combination of visual and text data to 'ground' the textual representations. But their work targeted visual ambiguity. Thus, quite some works in the recent past have looked at the representation of ambiguity in neural models using probing techniques. We extend that line of investigation in this paper. The following sections describe the experiments and observations.

3 Experiments

In this work, we investigate the capability of a shallow MLP classifier probe to identify ambiguous sentences from pretrained model representations. We investigate three kinds of sentence representations: *mean*, *sum* and *product*. For each sentence S_i, we first obtain the contextual representation for each word in the sentence. We then take the mean of the representations of the words to get the *mean* sentence representation. Similarly, we add the word representations to get the *sum* representation. Finally, we obtain the Hadamard product of the word representations to get the *product* sentence representation. We obtain such sentence representations of BERT and GPT-2 layers. We use the pretrained models provided by Huggingface (Wolf et al., 2019) for obtaining the representations.

The sentence representations are then used to train a probing classifier that identifies if a sentence is ambiguous. We perform the probing task in two settings:

- **In-Domain**: The training and test data came from the same source.
- **Cross-Domain**: The training and test data came from different sources.

3.1 In-Domain probing

For the experiments in this case, we considered data from COCO and HVG. As mentioned earlier, sentences in both corpora contain lexical ambiguities. For ambiguous sentences, we used the MSCOCO ambiguous test-set and the HVG challenge test-set respectively for the two experiments. The unambiguous sentences were drawn from the original (unambiguous) MSCOCO (Lin et al., 2014) and HVG datasets. As Fig. 1 and Fig. 2 show, the classifier was quite accurate across all the layers of both models for the two datasets. Across models, their performance is more or less the same. The mean sentence representation for GPT-2 seems to be the worst performer among all the other representations.

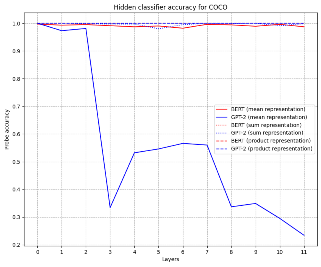

Figure 1: Probing classifier accuracy across layers for in-domain probing with COCO

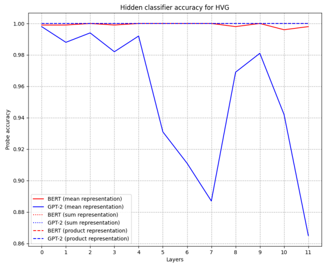

Figure 2: Probing classifier accuracy across layers for in-domain probing with HVG

As the last experiment in this category, we use the curated dataset that we created (described in Section 1) for the probing task. The results are shown in Fig. 3. It is observed that the product sentence representation performs the best and GPT-2 outperforms BERT.

Thus, even when trained on a mix of data from different datasets, the sentence representations of the models manage to encode features that help the classifier obtain reasonable scores on the task.

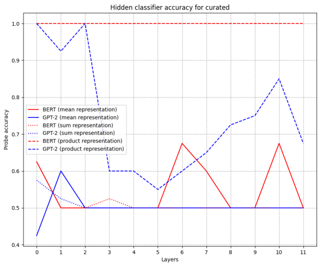

Figure 3: Probing classifier accuracy across layers for in-domain probing with curated dataset

3.2 Cross-Domain probing

In the cross-domain probing experiments, we wanted to investigate how the classifier would perform on sentences with similar ambiguity type (lexical) even if they were drawn from different datasets. In other words, does the classifier learn some unique features about the ambiguity in sentences (irrespective of what data it is being trained on)? First, we trained the classifier on the HVG corpus while we used the COCO corpus as the test-set. The results of the probe are shown in Fig. 4. We see that the classifier performs poorly across all layers for both the models. Thus, it does not seem that the classifier learns any general traits that helps it identify the lexical ambiguity.

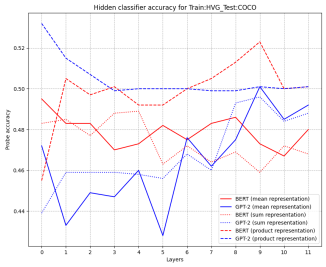

Figure 4: Probing classifier accuracy across layers for cross-domain probing with HVG and COCO

Next, we investigate the performance of the classifier when the *type* of ambiguity is changed. Hence, we replace the ambiguous sentences from COCO with those from LAVA corpus in the test-data. In other words, we train the classifier to detect lexical ambiguities and test it on syntactic ambiguities. Fig. 5 shows similar performance (if not worse) as the last experiment. It should be noted however, in terms of relative performance, the product sentence representation still performs better and GPT performing slightly better than BERT. But it doesn't seem that the classifier learns some universal features for ambiguity detection.

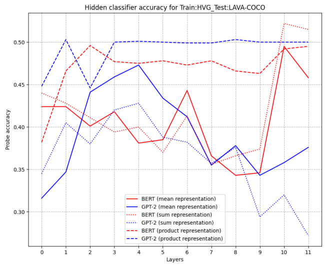

Figure 5: Probing classifier accuracy across layers for cross-domain probing with HVG and LAVA as ambiguity

We also concatenated the layer representations together and fed the concatenated representation to the probe. The rationale was to see if the probe would identify some useful feature from across all the layers. The results are shown in Table 1 and Table 2

Representation	BERT	GPT-2
mean	0.444	0.426
sum	0.449	0.475
product	0.508	0.514

Table 1: Classifier accuracy for concatenated layers (cross domain probing with HVG and COCO)

Representation	BERT	GPT-2
mean	0.378	0.310
sum	0.390	0.359
product	0.555	0.506

Table 2: Classifier accuracy for concatenated layers (cross domain probing with HVG and LAVA)

As the tables show, no significant performance gain was observed.

As the final experiment, we use HVG as the training data and the curated dataset as the test data. The results are shown in Fig. 6. Consistent with the observations before, the classifier fails to identify ambiguous sentences with accuracy.

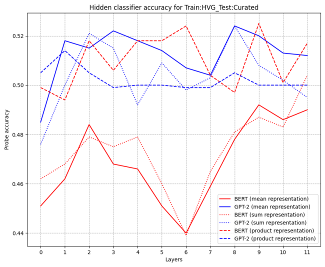

Figure 6: Probing classifier accuracy across layers for cross-domain probing with HVG and curated data

4 Discussion

As described in the Section 3.2, a number of experiments involving combinations of datasets were used to train the probes and observe how different models and different sentence representations fare against each other. The mean performance of the probes across the tasks corresponding to the three sentence representations for BERT and GPT-2 is presented in Fig. 7 and Fig. 8 respectively. It appears that across both the models, the product representation works better compared to other sentence representations. As discussed in Section 2, probing mechanisms for ambiguity detection has been explored at the word-level. Getting probes to work at the sentence level requires choosing the most appropriate form of sentence representations. And from the experiments described here, it seems that Hadamard product representations seem to work better for the purpose. Although there have been some criticisms regarding averaged word representations (Conneau et al., 2017) and proposed solutions (Rücklé et al., 2018), the objective in this paper was not to find the best way of obtaining sentence representations. The objective was to merely show how different sentence representations perform in the probing task.

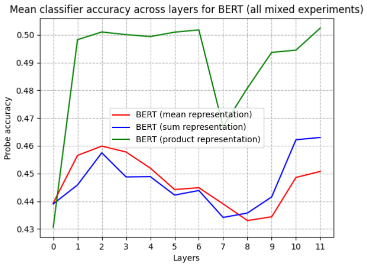

Figure 7: Mean Probing classifier accuracy for BERT across layers for cross-domain probing tasks

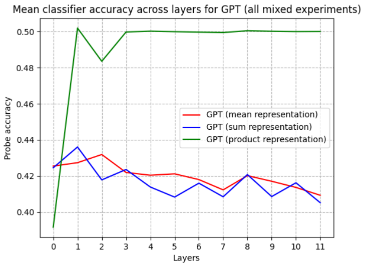

Figure 8: Mean Probing classifier accuracy for GPT-2 across layers for cross-domain probing tasks

Also, it is apparent that the sentence representations generated by the pretrained models do not seem to explicitly encode general features to identify ambiguity. The reasonable performance of the classifier in the in-domain task (including the one with the curated data) and the sub-par performance in the cross-domain task shows how probing tasks are domain-dependent.

However, it should be noted that all the ambiguity datasets used in this work (MSCOCO, LAVA, HVG) were designed to be used along with their corresponding images. And hence, it would be interesting to extend this line of analysis to a multimodal scenario to investigate if the inclusion of modalities impact the performance of the probes.

5 Conclusion

In this paper we explore how probing methods can be used to ascertain how and if pretrained models like BERT and GPT-2 identify ambiguity in sentences fed to them. We make use of three ways (mean,sum,product) to obtain sentence representations from the individual word representations to be fed to the probe. The experiments indicate that the Hadamard product representation for sentences works better than the others. We also observe how the sentence representations from both models perform remarkably well when the probing task involves a test-set drawn from the same domain as the training data and that cross-domain probing yields a bad performance.

References

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

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

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2015. Do you see what I mean? visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1487, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer. 2017. Proceedings of the second conference on machine translation. In *Proceedings of the Second Conference on Machine Translation*.

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.

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*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

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.

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.

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*.

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.

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.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings.

Francois Meyer and Martha Lewis. 2020. Modelling lexical ambiguity with density matrices. *arXiv preprint arXiv:2010.05670*.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation. *arXiv preprint arXiv:1907.08948*.

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.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

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

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

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

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.

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.

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*.

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.