Measuring Context-Dependent Syntactic Information Across Layers

Anonymous ACL submission

Abstract

Probing studies have extensively explored where in neural language models linguistic information is located. While probing classifiers are a common instrument to approach such questions, it is less clear what evaluation metrics to choose, how to compare probes, and which baselines to use. We identify angles from which the question how linguistic information is structured within a model can be approached and propose two new setups that fill the gap of explicitly modelling local information gain compared to the previous layer. We apply the new setups, along with two from the literature, to probe models for a syntactic property that explicitly needs context to be retrieved: part-of-speech tags that are not the most common for a specific token. We test the hypothesis that more information is retrieved in deeper layers than for the most common tags, and find that while this is often true, the manifestation varies among metrics and models in different languages.

1 Introduction

Probing neural language representations aims at finding hierarchical structure or other linguistic information in their parameters by testing hypotheses about features in the representations empirically (Hupkes et al., 2018; Alain and Bengio, 2017). This is often done by training a probing classifier with the model's hidden representations at different layers as the input representations and comparing the performance on diagnostic tasks across layers. While probes are conceptually simple, developing convincing evaluation setups for them is an ongoing challenge.

We contribute to this line of research by systematically investigating methods to measure usable information across layers on seven models in different languages. We group the metrics along two dimensions. The first one (as proposed by Hewitt et al. (2021)) inspects the relation of the baseline and the representation: how much more information can we extract from the representation than from the baseline (*baselined* probing), or how much information is extractable from the representation that does not overlap with information from the baseline (*conditional* probing). The second dimension, proposed by us, is the type of information intended to be measured: overall information, relative to a non-contextualized baseline (a *global* baseline), or information gain relative to the previous layer (a *local* baseline). The local setting challenges the view that a linguistic property's place in the model is the layer where most usable information for it can be extracted. Instead, we consider the layers where most usable information is *gained* relative to the previous layer to reflect the linguistic property's place within the model's hierarchy. We fill the identified gap of local metrics by developing and testing the local correspondents of baselined and conditional probing.

As a case study, we investigate which layers encode information for part-of-speech (POS) tags that are not the most frequent for a word form (¬MFTs). Recent probing work has hypothesized that ¬MFTs could be found in deeper layers of a model: The layer's contribution beyond the information already present in the uncontextualized layer of the popular BERT model (Devlin et al., 2019) is more significant in deeper layers, and therefore POS information could be found later in the model than previously assumed (Hewitt et al., 2021). While this behavior is arguably expected because lower layers share more information with layer 0 due to the closer connection in the neural architecture, such findings emphasize the relevance of the question of how we define and measure where in the model a linguistic feature is best found.

2 Related Work

How linguistic information is distributed across the layers of a neural model is one of the central ques-

tions in the probing literature. For ELMo, Peters et al. (2018) find that parts-of-speech are better predicted from the first hidden layer and word senses from the second. Tenney et al. (2019) probe BERT for a range of different NLP tasks and find that the layers that are the most predictive for each task are ordered like a classical pipeline: from parts-of-speech over syntactic dependencies, named entities and semantic roles to coreference.

How probing experiments should be designed and evaluated is an ongoing discussion. Some authors argue for simple classifiers (Alain and Bengio, 2017; Hewitt and Liang, 2019), others for more expressive models (Pimentel et al., 2020). While probes are traditionally evaluated using accuracy, recent work has proposed the use of alternative metrics that measure the effort of learning (Voita and Titov, 2020) or emphasize the performance early in the training (Talmor et al., 2020). Kunz and Kuhlmann (2021) propose to probe in an extrapolation setting, evaluating, among other setups, on the ¬MFTs in diagnostic POS tagging experiments.

3 Probing Setups

We consider a standard setup where we train a probe on a diagnostic task (part-of-speech tagging in our case) and evaluate it in terms of accuracy. More specifically, we use datasets $\mathcal{D} = \{(x_i, y_i)\}_i$ where each x_i is the representation of a neural language model (BERT in our case) at some specific layer, and y_i is the gold-standard label. By computing probe accuracy for different layers of the same model, we can compare layers in terms of how predictive they are with respect to the diagnostic task. For this we employ four different metrics:

Global Baselined Probing (GBP) In this common setup we measure the difference between the probe accuracy on a given layer l_i and the baseline layer l_0 – in BERT, this is the uncontextualized embedding layer. Thus we compute

$$\text{Acc}(l_i) - \text{Acc}(l_0) \tag{1}$$

As Hewitt et al. (2021) show, this can be interpreted as a difference between two quantities of $\mathcal{V}$-*information* (Xu et al., 2020), a theory of usable information under computational constraints. More specifically, (1) estimates the difference in $\mathcal{V}$-information between predicting the linguistic property under consideration from l_i and predicting it from layer l_0.

Global Conditional Probing (GCP) This setup has been proposed by Hewitt et al. (2021) with the intent to explicitly measure what information a layer l_i contributes *beyond* the information present in the baseline l_0. Practically, it entails computing the difference between the probe accuracy on the concatenation of l_i to l_0 and the baseline layer:

$$\text{Acc}([l_i; l_0]) - \text{Acc}(l_0) \tag{2}$$

In the framework of Hewitt et al. (2021), this measure is related to a conditional version of $\mathcal{V}$-information. More specifically, it estimates the conditional $\mathcal{V}$-information conditioned on prior information contained in the baseline.

Local Baselined Probing (LBP) Analogously to global baselined probing, we may consider a local setup where the baseline is the previous layer l_{i-1}:

$$\text{Acc}(l_i) - \text{Acc}(l_{i-1}) \tag{3}$$

This quantity provides an estimate of how much $\mathcal{V}$-information is gained when taking the step from l_{i-1} to l_i. We posit that layers with high values of (3) can be considered as layers where useful new information *emerges*.

Local Conditional Probing (LCP) To complete the picture, we propose to apply conditional probing to the local setting:

$$\text{Acc}([l_i; l_{i-1}]) - \text{Acc}(l_{i-1}) \tag{4}$$

The intention behind this metric is also to measure information gain with respect to l_{i-1}, but we account for exclusive information of l_{i-1} that is absent in l_i. Similar to Hewitt et al. (2021), we concatenate two layers and compare to scores on one of them. Our approach differs in that we do not compare to one static baseline layer (l_0) but dynamically to l_{i-1} to track the information gained across layers.

4 Experiments

As our probe, we use a simple feed-forward network with 64 hidden units and ReLU activation, and train it using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. Our implementation uses PyTorch (Paszke et al., 2019). We calculate the results for all metrics based on the mean accuracy over 10 random seeds.

Apart from English BERT (Devlin et al., 2019), we train probes on BERT models in Czech

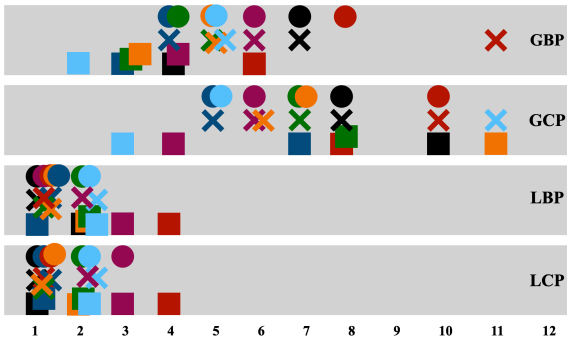
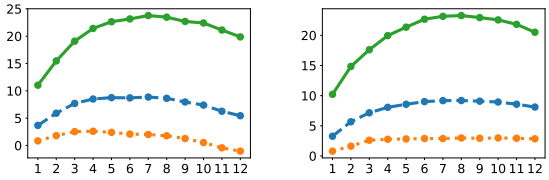


Figure 1: Layer that maximizes the score for a setup. Circle: All tags, Square: MFTs, Cross: ¬MFTs. Magenta: *cs*, black: *en*, red: *de*, dark blue: *fi*, green: *he*, orange: *sv*, light blue: *tr*. Color-free alternative: table 1.



(a) GBP English (b) GCP English

Figure 2: Global Metrics on *en* BERT. Solid green line: ¬MFTs, dotted orange: MFTs, dashed blue: all tags.

(Sido et al., 2021), Finnish (Virtanen et al., 2019), German (Chan et al., 2020), Hebrew (Seker et al., 2021), Swedish (Malmsten et al., 2020) and Turkish (Schweter, 2020). The languages are chosen to represent diverse families: Indo-European/Germanic (*de*, *en*, *sv*), Indo-European/Slavic (*cs*), Uralic (*fi*), Turkic (*tr*), and Afro-Asiatic/Semitic (*he*) to test if the syntactic information is localized in similar regions across models and languages. All models are *base* models with 12 layers, and accessed via the Huggingface Transformers library (Wolf et al., 2020).[1]

We predict UPOS tags from the Universal Dependencies treebank (Zeman et al., 2021) as the probing task[2], train on 1000 random sentences, and evaluate on three sets: the full test set (*all*), the most frequent tags for a word form (*MFT*) and tags are not the most frequent for a word form (*¬MFT*).

5 Results and Discussion

For each probing setup and language, we report that layer which maximizes the respective metric in Figure 1. Figures 2 and 3 show plots for selected setup/language pairs with interesting properties; plots for the remaining combinations can be found in Appendix A.

Global metrics. Our results for the global setups confirm the finding of Hewitt et al. (2021) that the layers that maximize conditional probing accuracy (GCP) are generally deeper than those that maximize baselined accuracy (GBP). This is the case in 5 out of the 7 models we investigated.

Zooming in on the distinction between most frequent and non-most frequent tags, however, exhibits an unexpected behavior: Hewitt et al. (2021) suggest that for the non-MFTs, GCP should be higher than GBP in deeper layers, while for MFTs it should be the other way round. Here we find that in 4 out of 7 models, the layer with the highest GCP value on non-MFTs precedes the layer with the highest value for MFTs. The highest scores of GBP on non-MFTs are consistently in deeper layers than those for MFTs. The exact layer in which the maximum scores are however varies greatly between models: for the MFTs, it ranges between 2 (*tr*) and 6 (*de*) and for the non-MFTs between 4 (*he*) and 11 (*de*).

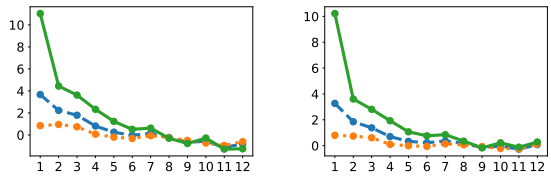
The plots for *en* BERT in Figure 2 are in line with the general trend: GCP peaks in deeper layers than GBP, but this is not explained by the non-MFTs, as their curve drops steeper with increasing layer index compared to that of the MFTs. This observation holds for most BERT models we used, except for *cs* and *tr* where the scores on MFTs drop more in deeper layers than those for non-MFTs.

Local metrics. Looking at the highest local information gain in Figure 1, both LBP and LCP show the highest gains for ¬MFTs in layer 1 for all models except *he* and *tr*, plus *cs* for LCP, where it is layer 2 to 3 (average LBP 1.3, LCP 1.6). The differences of the empirical results in LBP and LCP are small, leaving the choice between them to theoretical or practical preferences. Specifically accounting for information in l_i that is absent in l_{i-1} does not result in a different pattern.

Example curves for English BERT in Figure 3 show a typical pattern for the drop across layers. It appears to be the layer where contextual information is added first which is the layer of highest information gain. For MFTs, where the accuracy starts off much higher, and information gains are

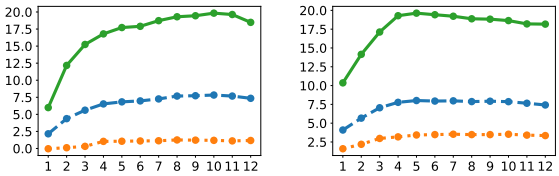
[1] All code is provided in the supplementary material and will be available on Github after the anonymity period.

[2] The treebanks for each language are: *cs*: PDT, *de*: GSD, *en*: EWT, *fi*: TDT, *he*: HTB, *sv*: Talbanken, *tr*: Kenet. Lic: CC BY-SA 4.0 (*de*, *en*, *fi*, *sv*, *tr*) / CC-BY-NC-SA 3.0 (*cs*, *he*).



(a) LBP English (b) LCP English

Figure 3: Local Metrics on *en* BERT.



(a) GCP German (b) GCP Finnish

Figure 4: Distinct patterns for *de* and *fi* BERT.

generally smaller, the highest-scoring layer ranges between 1 and 4 (average GBP 2.3, GCP 2.1) in both setups.

Model differences. As a general observation, we note that the relation between layer depth and score for each setup varies greatly across models. For example, looking at the GCP plots Figure 4, the curve for *de* shows a prolonged increase and the strongest information increase in deep layers (peak at layer 10), while for *fi*, the plot has a steep but short ascent (up to layer 5) followed by a slow drop as layers get deeper. Such differences in the distribution of information across models raise concerns about the generalizability of earlier probing results, which were limited to few popular, mostly English-language models. Future probing studies should include more diverse models and languages.

Summary. The different results on different metrics show how it depends on the perspective where a linguistic property is located. POS-relevant information, both for MFTs and ¬MFTs, are gained in early layers (LBP and LCP). Still, middle (and in some models late) layers contribute additional information that increases the amount of usable information over previous layers (GBP and GCP). The deeper the layer, the more different is the information from that in l_0 (GCP compared to GBP). ¬MFTs s are more sensitive to the layers than MFTs, with steep increases in the beginning but also a more distinct drop deeper in the model. As expected, the ¬MFTs are generally found later in the layers than MFTs. Contrary to previous assumptions, the difference is less pronounced in GCP. However, conditional probing makes the strong assumption that the probe can make optimal use of the features present in the representation. As this is probably not the case in the ¬MFTs evaluation, where the training data differs more from the test data than from the other two evaluation sets, GBP may be more meaningful here.

¬MFTs show both steeper gains in the beginning and more pronounced losses in the later layers, indicating that it is more specialized contextual information that the ¬MFTs require.

Limitations of our method. All metrics we use are based on probe accuracy. However, our setups can be easily adapted to other metrics which have been shown to be more robust towards design choices regarding the classifier, such as minimum description length (Voita and Titov, 2020), or metrics that reward fast learning (Yogatama et al., 2019; Talmor et al., 2020).

We do not consider the possibility to relate the different distribution of information across models to linguistic properties of the languages. We believe that this is impossible with the relatively small set of non-parallel models we analyze. Apart from the language, they differ in several variables: most importantly, the data they are trained on, but some also in training details.

6 Conclusion

We have collected and suggested metrics that model the information distribution in a model's layers from different perspectives: globally and locally, and with or without conditioning on the baseline. We used them to test whether information for POS tags that are not the most frequent for a word is found in deeper layers than general POS information and found that while this is the case for overall information measured by global metrics, local metrics highlight that the most significant gains consistently happen in the very first layers.

Probing experiments on seven monolingual BERT models in different languages show that the metrics' behavior varies between models. While it is currently not feasible to relate the differences to specific properties of the models such as the language or the domain of the training data, a controlled training of parallel models where the additional variables are controlled for may enable such a comparison and is an insightful direction for future work.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jenny Kunz and Marco Kuhlmann. 2021. Test harder than you train: Probing with extrapolation splits. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 15–25, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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, Online. Association for Computational Linguistics.

Stefan Schweter. 2020. Berturk - bert models for turkish. *Zenodo*.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – czech bert-like model for language representation. *arXiv preprint arXiv:2103.13031*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

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, Florence, Italy. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 183–196, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, et al. 2021. Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A All results

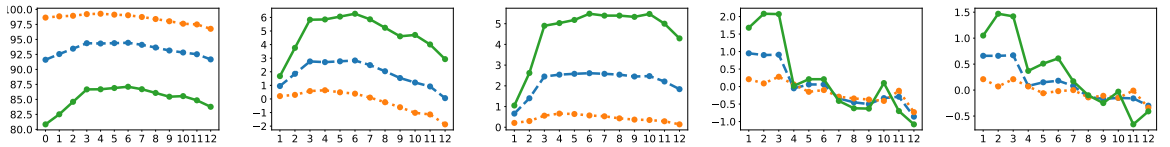
For completeness, we present all plots across models and metrics as supplementary material in figure 5. Table 1 reports the layers of the maximum score across all metrics and models that were visualized in figure 1 numerically, including average scores. A brief summary of the material is provided in the following paragraphs.

Accuracy and GBP. The accuracy shows the same highs and lows as the GBP setup, where the static l_0 baseline is subtracted from the accuracy. Its peak is in the early middle layers, with the average over models being 5.6 for all, 3.6 for MFTs and 6.1 for ¬MFTs (see table 1). Generally, the ¬MFTs peak a few layers later, indicating the need for more contextual information. Across models we see a large variation, most extremely visible in *de*, where the scores increase until layer 11 for the MFTs, and tr, where the drop for the MFTs is more distinct than for other models. *fi* and *he* have a distinct peak for the ¬MFTs in layer 4, then a decrease, and then stabilize.

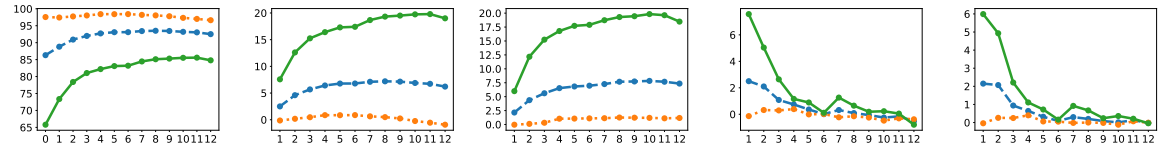
GCP. Compared to GBP, the results for ¬MFTs in GCP are shifted to later layers. Table 1 shows that the peak is on average in layer 6.9 instead of layer 5.6 for all tags. For MFTs, the difference is most pronounced with the average layer being 7.3 compared to 3.6 in GBP, while for ¬MFTs, it is closer, with 7.6 compared to 6.1.

LBP and LCP. The metrics that measure the local information gain have the most consistent pattern for the ¬MFTs, with most information generally added in the very first layer. The pattern of the curves appears to asymptotically approximate 0. There are however two exceptions: the *cs*, but most distinctly the *tr* model that gains relatively little in the first layer and makes its biggest jump in the second layer. We also observe in the accuracy curve of these two models that the increase in the beginning is less steep.

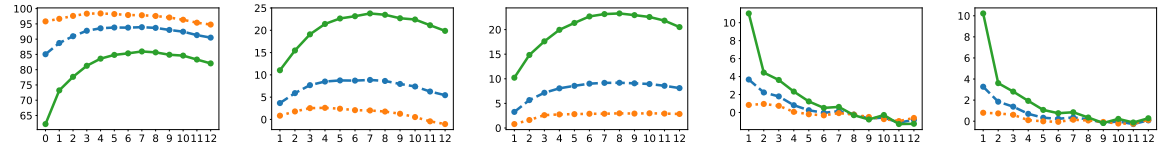
In all of the models we observe little difference in the empirical results and patterns of LBP and LCP, confirming our observations in section 5 that the choice between them can be arbitrary or based on theoretical preferences.



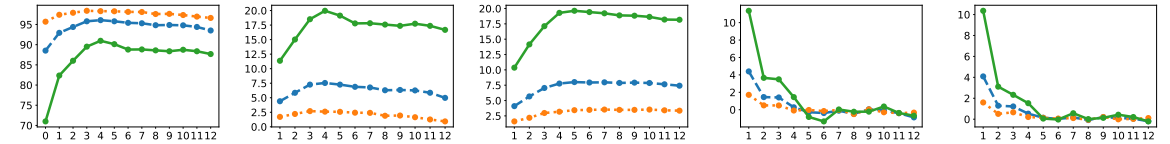
(1) Accuracy Czech (2) GBP Czech (3) GCP Czech (4) LBP Czech (5) LCP Czech



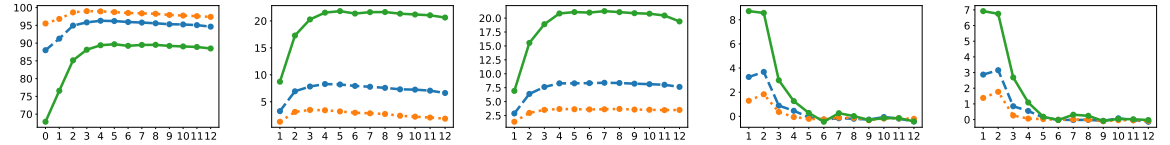
(6) Accuracy German (7) GBP German (8) GCP German (9) LBP German (10) LCP German



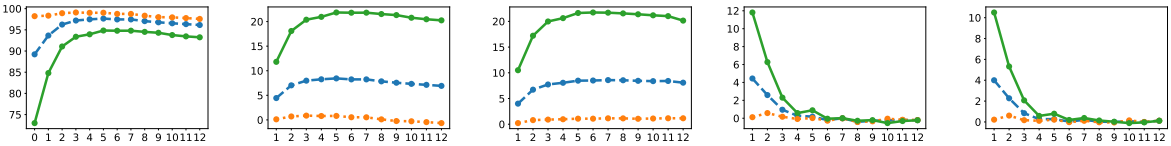
(11) Accuracy English (12) GBP English (13) GCP English (14) LBP English (15) LCP English



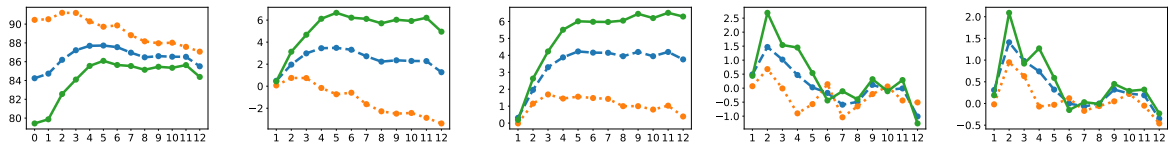
(16) Accuracy Finnish (17) GBP Finnish (18) GCP Finnish (19) LBP Finnish (20) LCP Finnish



(21) Accuracy Hebrew (22) GBP Hebrew (23) GCP Hebrew (24) LBP Hebrew (25) LCP Hebrew



(26) Accuracy Swedish (27) GBP Swedish (28) GCP Swedish (29) LBP Swedish (30) LCP Swedish



(31) Accuracy Turkish (32) GBP Turkish (33) GCP Turkish (34) LBP Turkish (35) LCP Turkish

Figure 5: Plots for all language/metric combinations.

	GBP (& Accuracy)			GCP			LBP			LCP		
	all	MFT	¬MFT	all	MFT	¬MFT	all	MFT	¬MFT	all	MFT	¬MFT
cs	6	4	6	6	4	6	1	3	2	3	3	2
de	8	6	11	10	8	10	1	4	1	1	4	1
en	7	4	7	8	10	8	1	2	1	1	1	1
fi	4	3	4	5	7	5	1	1	1	1	1	1
he	4	3	5	7	8	7	2	2	1	2	2	1
sv	5	3	5	7	11	6	1	2	1	1	2	1
tr	5	2	5	5	3	11	2	2	2	2	2	2
avg	5.6	3.6	6.1	6.9	7.3	7.6	1.3	2.3	1.3	1.6	2.1	1.3

Table 1: Layer of maximum score across metrics and models.