An information theoretic view on selecting linguistic probes

Zining Zhu[1,2], **Frank Rudzicz**[3,4,1,2]
[1] University of Toronto, [2] Vector Institute, [3] Surgical Safety Technologies
[4] Li Ka Shing Knowledge Institute, St Michael's Hospital
{zining,frank}@cs.toronto.edu

Abstract

There is increasing interest in assessing the linguistic knowledge encoded in neural representations. A popular approach is to attach a diagnostic classifier – or "probe" – to perform supervised classification from internal representations. However, how to select a good probe is in debate. Hewitt and Liang (2019) showed that a high performance on diagnostic classification itself is insufficient, because it can be attributed to either "the representation being rich in knowledge", or "the probe learning the task", which Pimentel et al. (2020) challenged. We show this dichotomy is valid information-theoretically. In addition, we find that the methods to construct and select good probes proposed by the two papers, *control task* (Hewitt and Liang, 2019) and *control function* (Pimentel et al., 2020), are equivalent – the errors of their approaches are identical (modulo irrelevant terms). Empirically, these two selection criteria lead to results that highly agree with each other.

1 Introduction

Recently, neural networks have shown substantial progress in NLP tasks (Devlin et al., 2019; Radford et al., 2019). To understand and explain their behavior, a natural question emerge: how much linguistic knowledge is encoded in these neural network systems?

An efficient approach to reveal information encoded in internal representations uses diagnostic classifiers (Alain and Bengio, 2017). Referred to as "probes", diagnostic classifiers are trained on pre-computed intermediate representations of neural NLP systems. The performance on tasks they are trained to predict are used to evaluate the richness of the linguistic representation in encoding the probed tasks. Such tasks include probing syntax (Hewitt and Manning, 2019; Lin et al., 2019; Tenney et al., 2019a), semantics (Yaghoobzadeh et al., 2019), discourse features (Chen et al., 2019; Liu et al., 2019; Tenney et al., 2019b), and commonsense knowledge (Petroni et al., 2019; Poerner et al., 2019).

However, appropriate criteria for selecting a good probe is under debate. The traditional view that high-accuracy probes are better is challenged by Hewitt and Liang (2019), who proposed that the high accuracy could be attributed to either (1) that the representation contains rich linguistic knowledge, or (2) that the probe learns the task. To circumvent this ambiguity, they proposed to use the improvement of probing task performance against a control task (predicting random labels from the same representations), i.e., the "selectivity" criterion. Recently, Pimentel et al. (2020), challenged this dichotomy from an information theoretic viewpoint. They proposed to use an "information gain" criterion, which empirically is the reduction in cross entropy from a "control function task" probing from randomized representation.

In this paper, we show the "non-exclusive-or" dichotomy raised by Hewitt and Liang (2019) is valid information-theoretically. There is a difference between the original NLP model learning the task and the probe learning the task.

In addition, we show that the "selectivity" criterion and the "control function" criterion are comparably accurate. Pimentel et al. (2020) formulated their errors with the difference in a pair of KL divergences. We show that the error of the "selectivity" criterion (Hewitt and Liang, 2019), if measured from cross entropy loss, can be formulated in the difference in a pair of KL divergences as well. When randomizations are perfect, these two criteria differ by only constant terms.

Empirically, on a POS tag probing task on English, French and Spanish translations, we show that the "selectivity" and the "control function" criteria highly agree with each other. We rank experi-

ments with over 10,000 different hyperparameter settings using these criteria. The Spearman correlation of the Hewitt and Liang (2019) vs. Pimentel et al. (2020) criteria are on par with the correlations of "accuracy vs. cross entropy loss" – two very strong baselines.

Overall, we recommend using control mechanisms to select probes, instead of relying on merely the probing task performance. When randomization is done well, controlling the target or representations are equivalent.

2 Related work

Diagnostic probes were originally intended to explain information encoded in intermediate representations (Adi et al., 2017; Alain and Bengio, 2017; Belinkov et al., 2017). Recently, various probing tasks have queried the representations of, e.g., contextualized word embeddings (Tenney et al., 2019a,b) and sentence embeddings (Linzen et al., 2016; Chen et al., 2019; Alt et al., 2020; Kassner and Schütze, 2020; Maudslay et al., 2020; Chi et al., 2020).

The task of probing is usually formulated as a classifier problem, with the representations as input, and the features indicating information as output. A straightforward method to train a classifier is by minimizing cross entropy, which is the approach we follow. Note that Voita and Titov (2020) derived training objectives from minimum description lengths, resulting in cross entropy loss and some variants.

3 Information theoretic probes

3.1 Formulation

We adopt the information theoretic formulation of linguistic probes of (Pimentel et al., 2020), and briefly summarize as follows.

We want to probe true labels T from representations R. An ideal probe should accurately report the code-target mutual information $I(T; R)$, which is unfortunately intractable. We will write $I(T; R)$ in an alternative form.

Let $p(T \mid R)$ be the unknown true conditional distribution, and a diagnostic probe, according to the setting in literature (Alain and Bengio, 2017; Hewitt and Manning, 2019; Maudslay et al., 2020), is an approximation $q_\theta(T \mid R)$ parameterized by θ, then:

$$I(T; R) = H(T) - H(T \mid R)$$
$$H(T \mid R) = -\mathbb{E}_{p(T \mid R)} \log p(T \mid R)$$
$$= -\mathbb{E}_{p(T \mid R)} \log \frac{p(T \mid R) q_\theta(T \mid R)}{q_\theta(T \mid R)}$$
$$= -\mathbb{E}_p \log q_\theta - \mathbb{E}_p \log \frac{p}{q_\theta}$$
$$= H(p, q_\theta) - \mathrm{KL}(p \,\|\, q_\theta),$$

where p and q_θ stand for $p(T \mid R)$ and $q_\theta(T \mid R)$ respectively. We also use $H(p, q_\theta) = -\mathbb{E}_p \log q_\theta$ to represent the cross entropy for simplicity.

3.2 The source of probing error

A valid dichotomy Traditionally, people use the cross entropy loss of the diagnostic probe $H(p, q_\theta)$ to approximate $I(T; R)$. We can derive a source of error by rewinding the above formulations:

$$H(p, q_\theta) = H(T) - I(T; R) + \mathrm{KL}(p \,\|\, q_\theta)$$

The first term on RHS, $H(T)$, is independent of either R or θ. Therefore, a low cross entropy loss $H(p, q_\theta)$ can be caused by either of the two scenarios:

- High code-target mutual information $I(T; R)$, indicating the representation R contains rich information about the target T.
- Low KL-divergence between $p(T \mid R)$ and $q_\theta(T \mid R)$, indicating the probe learns the task.

The two scenarios exactly correspond to the dichotomy of Hewitt and Liang (2019).

A good probe To get a good probe, we want $H_{q_\theta}(T \mid R)$ to approximate $I(T; R)$ as much as possible. This means a good probe should minimize $\mathrm{KL}(p \,\|\, q_\theta)$, as proposed by Pimentel et al. (2020).

However, empirically Pimentel et al. (2020) (as well as many previous articles) used the cross entropy loss $H_{q_\phi}(T \mid R)$ to select good probes, which is insufficient, as described above. Alternatively, Hewitt and Liang (2019) and Pimentel et al. (2020) proposed control tasks and control functions, respectively.

3.3 The control task

The control task (Hewitt and Liang, 2019) sets random targets for the probing task. Let us use $c(T)$ to indicate a "control function" applied on a token

v that originally has label T. The control function could nullify the information of the input, if necessary.

If we measure the difference between cross entropy in the control task and probing task $H(p(c(T)\,|\,R), q_{\theta_c}(c(T)\,|\,R)) - H(p(T\,|\,R), q_\theta(T\,|\,R))$, we can derive a form of error margin in their measurements[1]. Let us use a short-hand notations $H(p_c, q_{\theta_c}) - H(p, q_\theta)$ for clarity. Now, what does the diff between cross entropy on control task and probing task actually contains?

$$
\begin{aligned}
&H(p_c, q_{\theta_c}) - H(p, q_\theta) \\
&= (H(c(T)) - I(c(T); R) + \mathrm{KL}(p_c \,\|\, q_{\theta_c})) \\
&\quad - (H(T) - I(T; R) + \mathrm{KL}(p \,\|\, q_\theta)) \\
&= (H(c(T)) - H(T)) \\
&\quad - I(c(T); R) + I(T; R) \\
&\quad + (\mathrm{KL}(p_c \,\|\, q_{\theta_c}) - \mathrm{KL}(p \,\|\, q_\theta))
\end{aligned}
$$

We already knew that $H(T) = \text{Const}$. According the definition of control function, the output $c(T)$ would be independent of R. Then:

$$
\begin{aligned}
H(c(T)) &= \text{Const} \\
p(c(T), R) &= p(c(T))p(R) \\
I(c(T); R) &= \mathbb{E}\, \frac{p(c(T), R)}{p(c(T))p(R)} = \text{Const}
\end{aligned}
$$

Therefore:

$$
H(p_c, q_{\theta_c}) - H(p, q_\theta) = I(T; R) - \Delta_h
$$

where Δ_h is a short-hand notation for the measurement error in the *control task* criteria:

$$
\Delta_h = \mathrm{KL}(p \,\|\, q_\theta) - \mathrm{KL}(p_c \,\|\, q_{\theta_c}) + \text{Const} \quad (1)
$$

When the probe fits the true distribution to a similar extent on both the control task and probing task, the error Δ_h would be small. Unfortunately, both KL terms are intractable.

3.4 The control function

Control function (Pimentel et al., 2020) introduces a random processor $\mathbf{c}(\cdot)$ on the representation R. To measure the information gain, they used an "information gain" criterion:

$$
\mathcal{G}(T, R, \mathbf{c}) = I(T; R) - I(T; \mathbf{c}(R))
$$

Noticing that mutual information terms are intractable, they approximated the objective with the difference between cross entropy in the control function task (we refer to as "control function" henceforth) and the probing task:

$$
\tilde{\mathcal{G}}(T, R, \mathbf{c}) = H(p_c, q_{\phi_c}) - H(p, q_\phi)
$$

To compute the error of this approximation, they reformulated the terms as following:

$$
\begin{aligned}
&\mathcal{G}(T, R, \mathbf{c}) \\
&= H(T) - H(T\,|\,R) - H(T) + H(T\,|\,\mathbf{c}(R))
\end{aligned}
$$

The two $H(T)$ terms cancel out, then:

$$
\begin{aligned}
H(T\,|\,R) =&\, H(p(T\,|\,R), q_\phi(T\,|\,R)) \\
&- \mathrm{KL}(p(T\,|\,R) \,\|\, q_\phi(T\,|\,R)) \\
=&\, H(p, q_\phi) - \mathrm{KL}(p, q_\phi) \\
H(T|\mathbf{c}(R)) =&\, H(p(T\,|\,\mathbf{c}(R)), q_{\phi_c}(T\,|\,\mathbf{c}(R))) \\
&- \mathrm{KL}(p(T\,|\,\mathbf{c}(R)) \,\|\, q_{\phi_c}(T\,|\,\mathbf{c}(R))) \\
=&\, H(p_c, q_{\phi_c}) - \mathrm{KL}(p_c \,\|\, q_{\phi_c})
\end{aligned}
$$

Where we abbreviate similarly as we did for the control task. Specifically, we write ϕ for the probe parameters of control function to tell apart from θ in the control task.

Pimentel et al. (2020) showed that the error of their approximation, $\Delta_p = \mathcal{G}(T, R, \mathbf{c}) - \tilde{\mathcal{G}}(T, R, \mathbf{c})$, can be expressed as:

$$
\Delta_p = \mathrm{KL}(p \,\|\, q_\phi) - \mathrm{KL}(p_c \,\|\, q_{\phi_c}) \quad (2)
$$

Again, when the probe fits the true distribution to a similar extent on both the target labels distribution and the probing task, the Δ_p will be small. Unfortunately, both KL terms are intractable too.

3.5 Control task vs control function

From Equations 1 and 2, we showed that the *selectivity* criterion of Hewitt and Liang (2019) and the *information gain* criterion of Pimentel et al. (2020), if both measured in cross entropy loss, have very similar errors in approximating information gains.

These errors, Δ_h and Δ_p respectively, appear in very similar forms[2]. The probes selected from these two criteria should be highly correlated to each other, and our experiments will confirm.

---

[1]Note that Hewitt and Liang (2019) used accuracy instead of cross entropy. We discuss cross entropy so as to compare the errors against the control function (Pimentel et al., 2020)

[2]When the two $c(\cdot)$ are ideal, Δ_h and Δ_p differ by only irrelevant terms – we include the derivations in Supplementary Material.

Language	# POS	# Tokens train / dev / test	Correlations (t_acc,f_ent)	(t_acc,t_ent)	(f_acc,f_ent)
English	17	177k / 22k / 22k	0.1615	0.1334	0.1763
French	15	303k / 31k / 8k	0.0906	0.0606	0.1295
Spanish	16	341k / 33k / 11k	0.1360	0.0560	0.1254

Table 1: Spearman correlations between t_acc (the "selectivity" criterion (Hewitt and Liang, 2019)) and f_ent (the "gain" criterion (Pimentel et al., 2020)) are on par with two "accuracy vs. cross entropy" correlations.

4 Experiments

4.1 Setup

We use the same family of probes as Hewitt and Liang (2019) and Pimentel et al. (2020), multiple layers perceptrons with ReLU activations, to show the correlations of their "good probe" criteria (control task and control function, respectively).

Overall, we sweep the probe model hyper-parameters with a unified training scheme on three tasks (probe, control task, control function). The control task (function) setting includes labels (embeddings) drawn from a uniform random sample once before all experiments. In each training, we follow the setting of (Hewitt and Liang, 2019). We save the model with the best dev loss, report the test set loss and accuracy, and average across 4 different random seeds.

Data We use the Universal Dependency (Zeman et al., 2019) dataset loaded with the Flair toolkit (Akbik et al., 2018). We examine three languages: English, French, and Spanish. For the probing task, we use POS with labels provided by SpaCy[3]. We use the embedding of multilingual BERT (mBERT) implemented by huggingface (Wolf et al., 2019). If a word is split into multiple word pieces, we average its representations.

4.2 The "good probes" are good for both

When measuring the qualities of probes using the "selectivity" (Hewitt and Liang, 2019) or "information gain" (Pimentel et al., 2020) criterion, we show that the rules-of-thumb for training good probes largely agree.

- Early stopping before 24,000 gradient steps (approximately 4 epochs) could inhibit probe quality, but longer training procedures do not improve the probe qualities considerably.
- Smaller probes are better in general, but exceptions exist. For example, when weight decay

is set to 0, probes with one hidden layer and 40 hidden neurons are better in both criteria.
- A small weight decay is beneficial.

We include more descriptions, including comprehensive experiment configurations and plots in the Supplementary Material.

4.3 The high correlation between criteria

In addition to the qualitative correlations shown above, we compute the correlations between the two criteria over a grid-search style hyper-parameter sweep of over 10,000 configurations. For each "probe, control task, control function" experiment set, we record the following four criteria:

- **t_acc**: Difference between probing task and control task accuracy. This is the "selectivity" criterion of Hewitt and Liang (2019).
- **f_ent**: Difference between control function and probing task cross entropy. This is the "gain" criterion of Pimentel et al. (2020).
- **t_ent**: Difference between the control task and the probing task cross entropy.
- **f_acc**: Difference between the probing task and control function accuracy.

We collect all experiments of each language according to these criteria, and use Spearman correlation to test three pairs of correlations. As is reported in Table 1, the (t_acc, f_ent) correlations are comparable to two strong baselines, (t_acc, t_ent) and (f_acc, f_ent), the correlations between measurements in accuracy and cross entropy losses.

5 Conclusion

When selecting probes that better approximate $I(T; R)$, we recommend measuring *with a control mechanism* instead of relying on the traditional cross entropy on probing task. We show both information-theoretically and empirically, that controlling the targets and representations are equivalent, as long as the control mechanism is randomized.

---
[3] https://spacy.io

6 Acknowledgement

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *ICLR*, Toulon, France.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING*, pages 1638–1649.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *ICLR*, Toulon, France.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction. In *ACL*. Association of Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? In *ACL*, pages 861–872, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation Benchmarks and Learning Criteria for Discourse-Aware Sentence Representations. In *EMNLP*, pages 649–662, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. In *ACL*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *EMNLP*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *NAACL*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. *ACL*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting Inside BERT's Linguistic Knowledge. *ACL BlackBoxNLP Workshop*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *TACL*, 4:521–535.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A Tale of a Probe and a Parser. In *ACL*. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *EMNLP*, pages 2463–2473, Hong Kong, China. 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 *ACL*. Association of Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. BERT is Not a Knowledge Base (Yet): Factual Knowledge vs. Name-Based Reasoning in Unsupervised QA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):1–24.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *ACL*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. *arXiv preprint arXiv:2003.12298*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.0:1–11.

Yadollah Yaghoobzadeh, Katharina Kann, T J Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings. In *ACL*, pages 5740–5753, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, and Mitchell et al Abrams. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Difference between the two criteria

In Section 3 we show that the error of the two criteria can both be written as difference between a pair of KL divergence (modulo a constant term). Here we further simplify the terms when we assume the control task and functions take perfectly random distributions (i.e., independent of the task and representations, respectively).

$$\Delta_h - \Delta_p = \text{Const}+$$
$$\text{KL}(p(c(T)|R) \,\|\, q_{\theta_c}(c(T)|R)) - \text{KL}(p \,\|\, q_\theta)$$
$$- \text{KL}(p(T|c(R)) \,\|\, q_{\phi_c}(T|c(R))) + \text{KL}(p \,\|\, q_\phi)$$

When we use the same hyperparameter setting, $q_\theta(T \mid R)$ and $q_\phi(T \mid R)$ should be able to approximate $p(T \mid R)$ to the same extent, so $\text{KL}(p \,\|\, q_\theta)$ and $\text{KL}(p \,\|\, q_\phi)$ cancel out. Additionally, following the definitions of the control function and control tasks, we can simplify as follows:

$$p(c(T)|R) = p(c(T)), q_{\theta_c}(c(T)|R) = q_{\theta_c}(c(T))$$
$$p(T|c(R)) = p(T), q_{\phi_c}(T|c(R)) = q_{\phi_c}(T)$$

$$\text{KL}(p(c(T)|R) \,\|\, q_{\theta_c}(c(T)|R)) = \mathbb{E}_{p(c(T))} \frac{p(c(T))}{q_{\theta_c}(c(T))}$$

$$\text{KL}(p(T|c(R)) \,\|\, q_{\phi_c}(T|c(R))) = \mathbb{E}_{p(T)} \frac{p(T)}{q_{\phi_c}(T)}$$

Therefore, the difference between errors of criteria of Hewitt and Liang (2019) and Pimentel et al. (2020) are:

$$\Delta_h - \Delta_p = \text{Const} - \text{KL}(p(T) \,\|\, q_{\phi_c}(T)) + \text{KL}(p(c(T)) \,\|\, q_{\theta_c}(c(T))) \quad (3)$$

In short, these two criteria differ by terms dependent only on the randomization functions and the inherent distributions of task labels, i.e., irrelevant terms.

B Experiments details

Hewitt and Liang (2019) proposed some rules-of-thumb to select good probes, including a "simple probe" suggestion. We sweep the hyper parameters to test whether these rules also apply when measuring probe qualities using the control function (Pimentel et al., 2020).

Hyper-parameter ranges We sweep hyper parameters from the following ranges:

- Learning rate: $\{10^{-4}, 5 \times 10^{-5}, 3 \times 10^{-5}, 10^{-5}, 5 \times 10^{-6}, 3 \times 10^{-6}\}$
- Maximum gradient steps: $\{1500, 3000, 6000, 12000, 24000, 96000, \infty\}$. Their effects are shown in Figure 1.
- Weight decay: $\{0, 0.01, 0.1, 1.0\}$. Their effects are shown in Figures 2, 3 and 4. When sweeping weight decay, max gradient step is set to 24000.

In any configuration mentioned above, we run four experiments with random seeds 73, 421, 9973, 361091, and average the reported results (i.e., accuracy and loss).

Early stopping could inhibit probe quality
Early stopping, if stopped before 24,000 gradient steps (approximately 4 epochs) may inhibit the quality of probes. In addition, we Figure 1 shows high correlation between the "selectivity" (Hewitt and Liang, 2019) and "information gain" (Pimentel et al., 2020) criteria.

Small weight decay is beneficial We find that smaller weight decays (e.g., 0.01) are more beneficial for probes than larger weight decays. While the two criteria rank the capacity of probes similarly, the most simple probes tend to stand out more distinctively with the "selectivity" criterion (Pimentel et al., 2020), as are shown on Figures 2, 3, and 4.

Smaller probes are not necessarily better We find that while smaller probes have higher "selectivity" and "information gain" for mBERT representations, probes with one hidden layer and 40-80 hidden neurons are better than more simplistic probes, as shown in Figures 5, 6 and 7. The plots show consistency between the two criteria. For example, larger models and more layers do not necessary lead to better results. Neither are the smallest probes with 0 hidden layers.

Note that we also swept hyperparameters for FastText, where probes with less parameters do not always outperform more complex probes in either accuracy, loss, selectivity, or information gain. Figures 8 and 9 illustrate these observations.

C Reproducibility

On a T4 GPU card, training one epoch takes around 20 seconds. Without setting maximum gradient steps, 98.6% of experiments finish within 400 epochs. We will open source our codes.
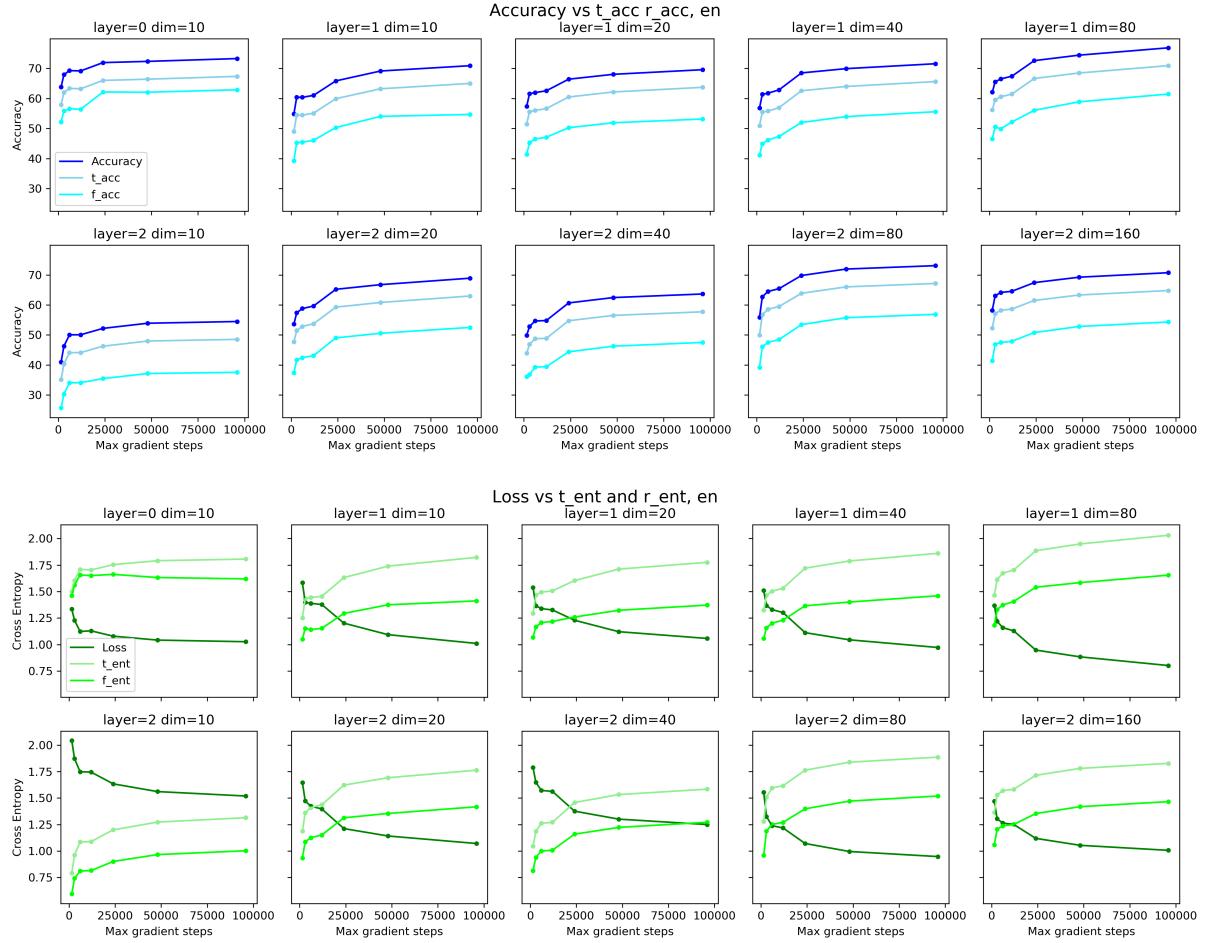
Figure 1: Max gradient step vs accuracy, t_acc, f_acc (blue) and loss, t_ent, f_ent (green) on English. The "t" refers to control task (Hewitt and Liang, 2019), and "h" refers to control function (Pimentel et al., 2020). In these set of experiments, we look for the best learning rate and zero weight decay in each configuration.
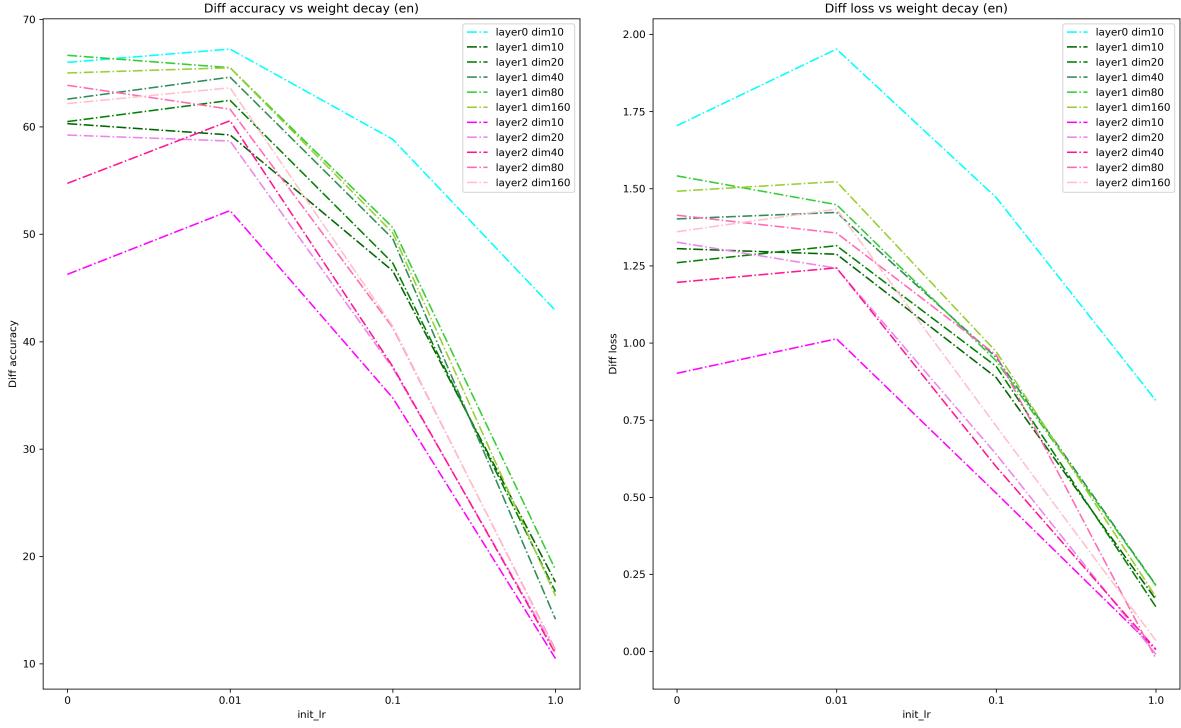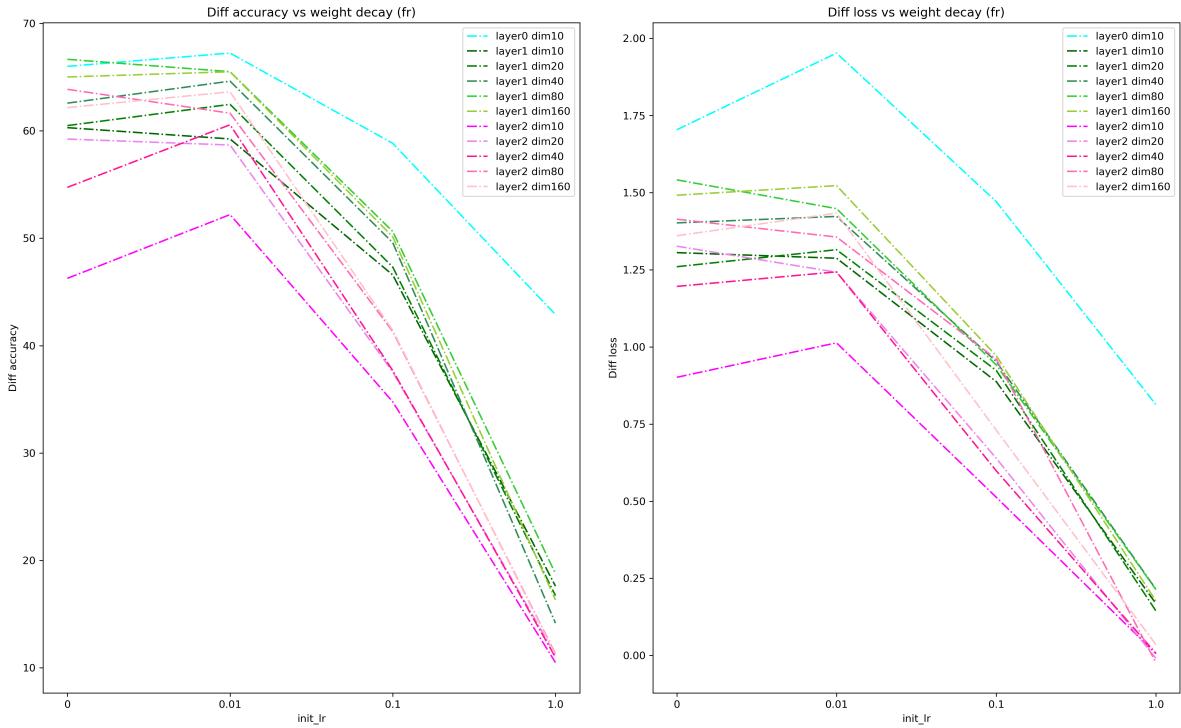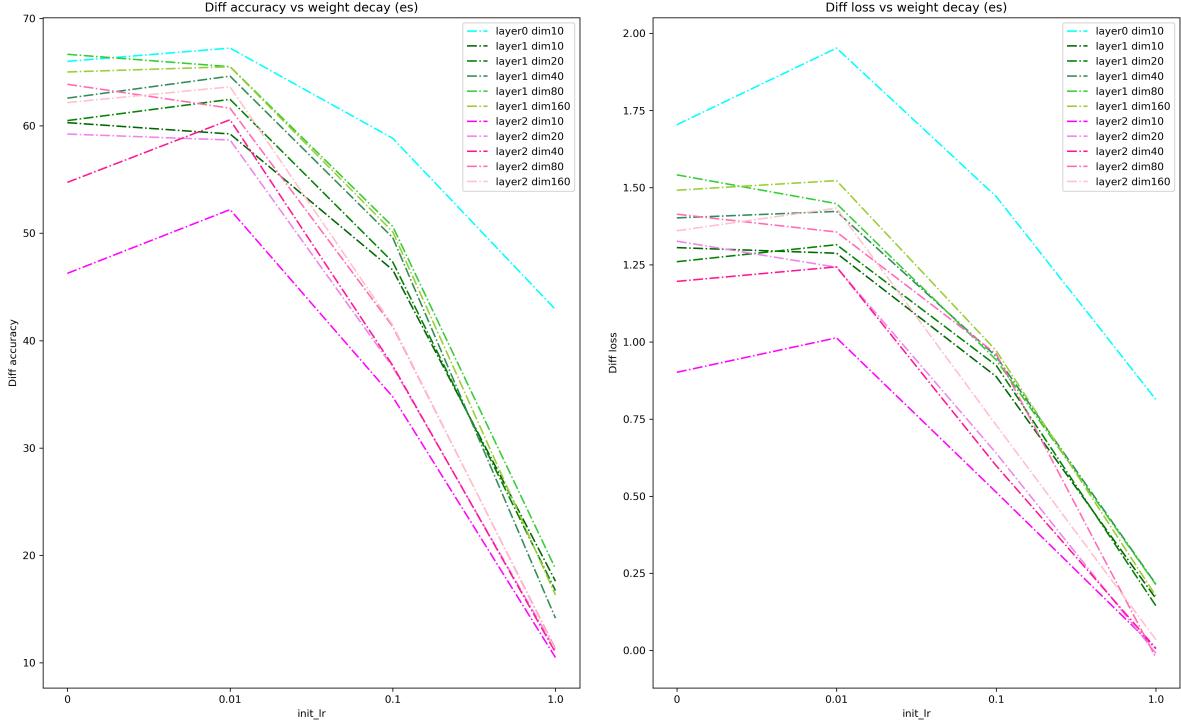
Figure 2: The "difference of accuracy" (Hewitt and Liang, 2019) and the "difference of loss" (Pimentel et al., 2020) criteria against weight decay on model configurations, on UD English. For each configuration, the learning rate leading to the highest accuracy is selected.
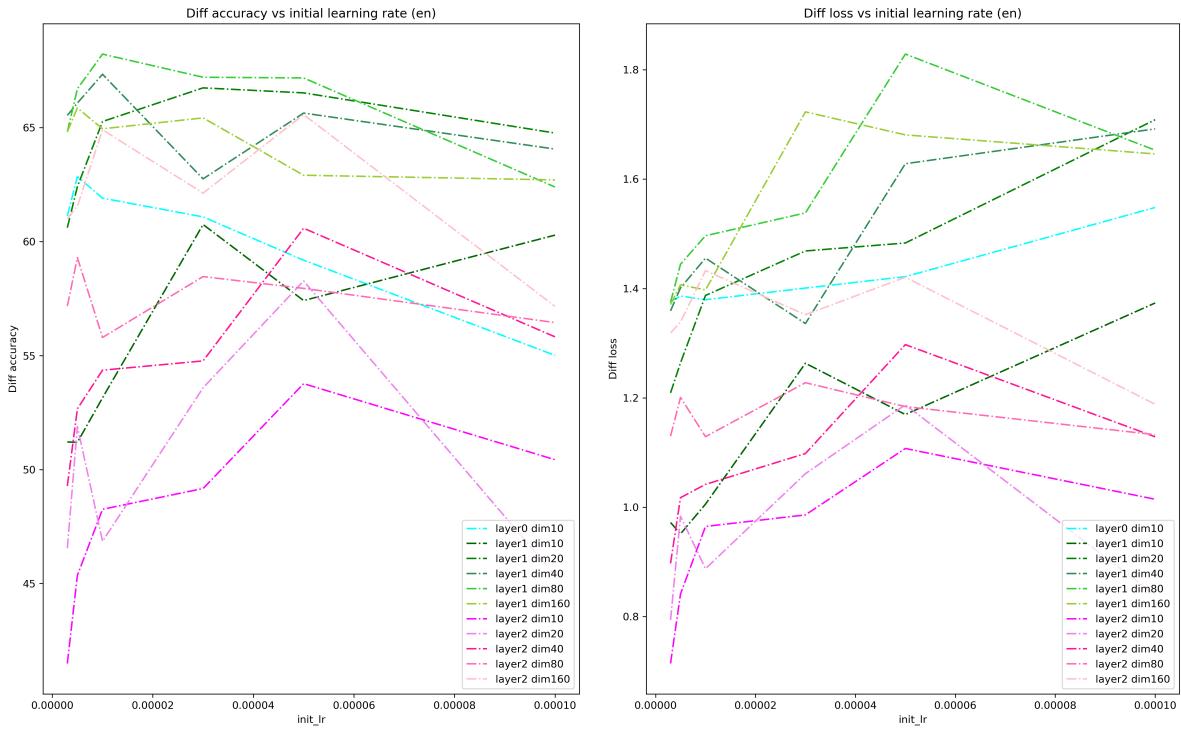

Figure 3: The "difference of accuracy" (Hewitt and Liang, 2019) and the "difference of loss" (Pimentel et al., 2020) criteria against weight decay on model configurations, on UD French. For each configuration, the learning rate leading to the highest accuracy is selected.

Figure 4: The "difference of accuracy" (Hewitt and Liang, 2019) and the "difference of loss" (Pimentel et al., 2020) criteria against weight decay on model configurations, on UD Spanish. For each configuration, the learning rate leading to the highest accuracy is selected.
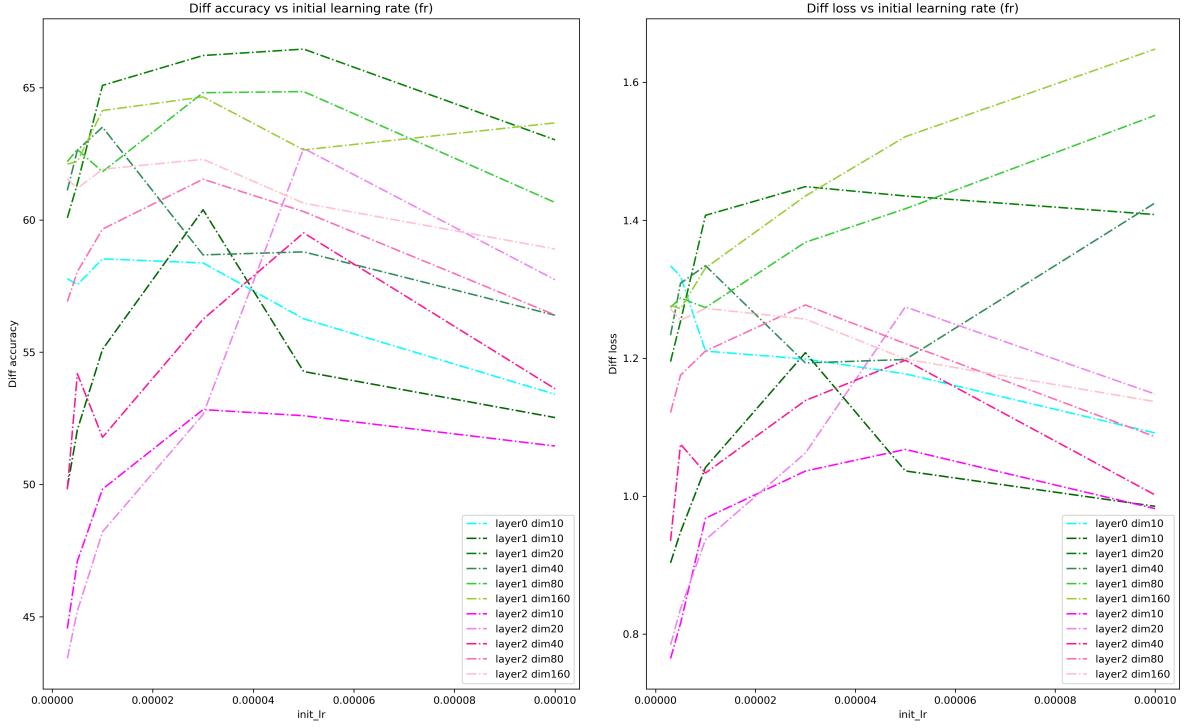

Figure 5: The "difference of accuracy" (Hewitt and Liang, 2019) and the "difference of loss" (Pimentel et al., 2020) criteria with different learning rates on model configurations, on UD English. The weight decay is set to 0.
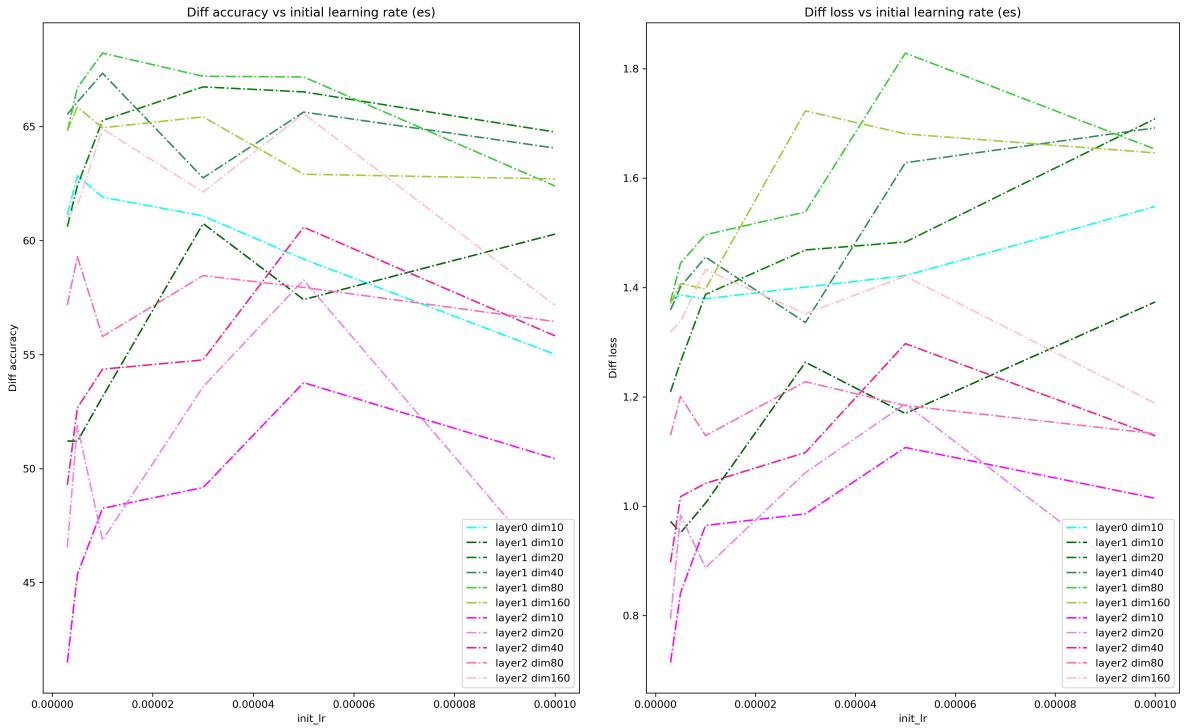
Figure 6: The "difference of accuracy" (Hewitt and Liang, 2019) and the "difference of loss" (Pimentel et al., 2020) criteria with different learning rates on model configurations, on UD French. The weight decay is set to 0.

Figure 7: The "difference of accuracy" (Hewitt and Liang, 2019) and the "difference of loss" (Pimentel et al., 2020) criteria with different learning rates on model configurations, on UD Spanish. The weight decay is set to 0.
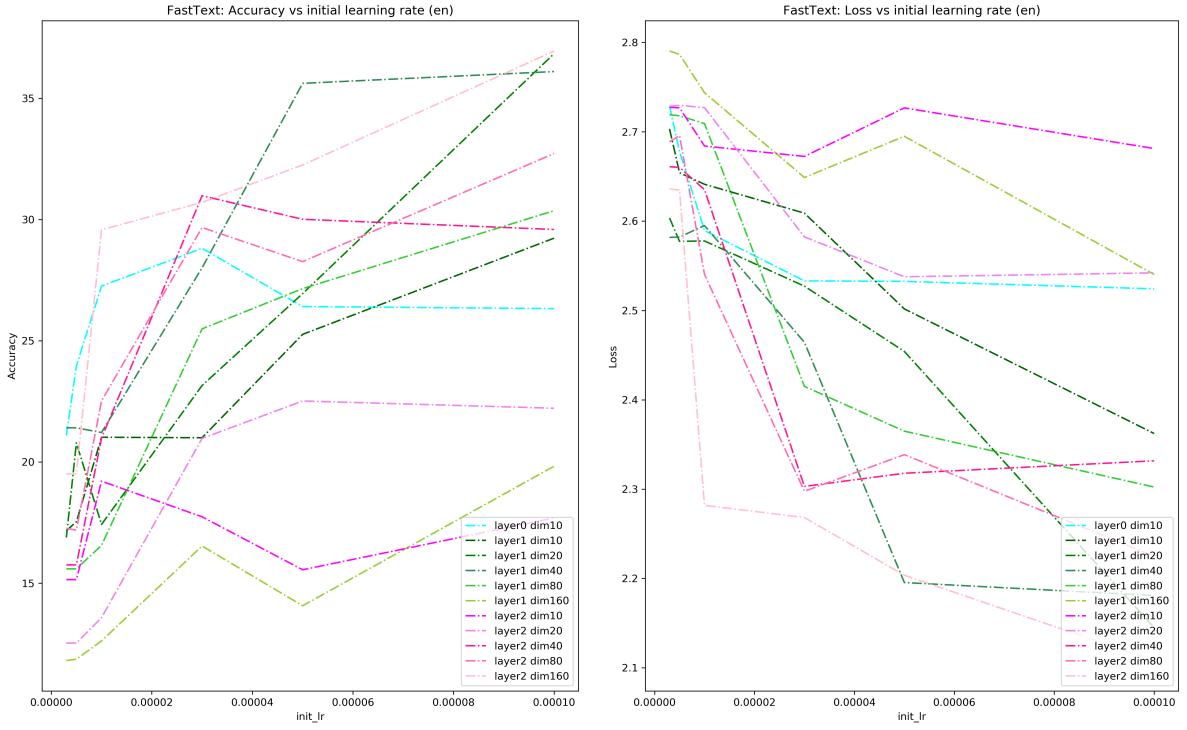
Figure 8: The accuracy and cross entropy loss of probes on FastText. These performances are much worse than those on mBERT, indicating the richness of information encoded in contextuality of mBERT.
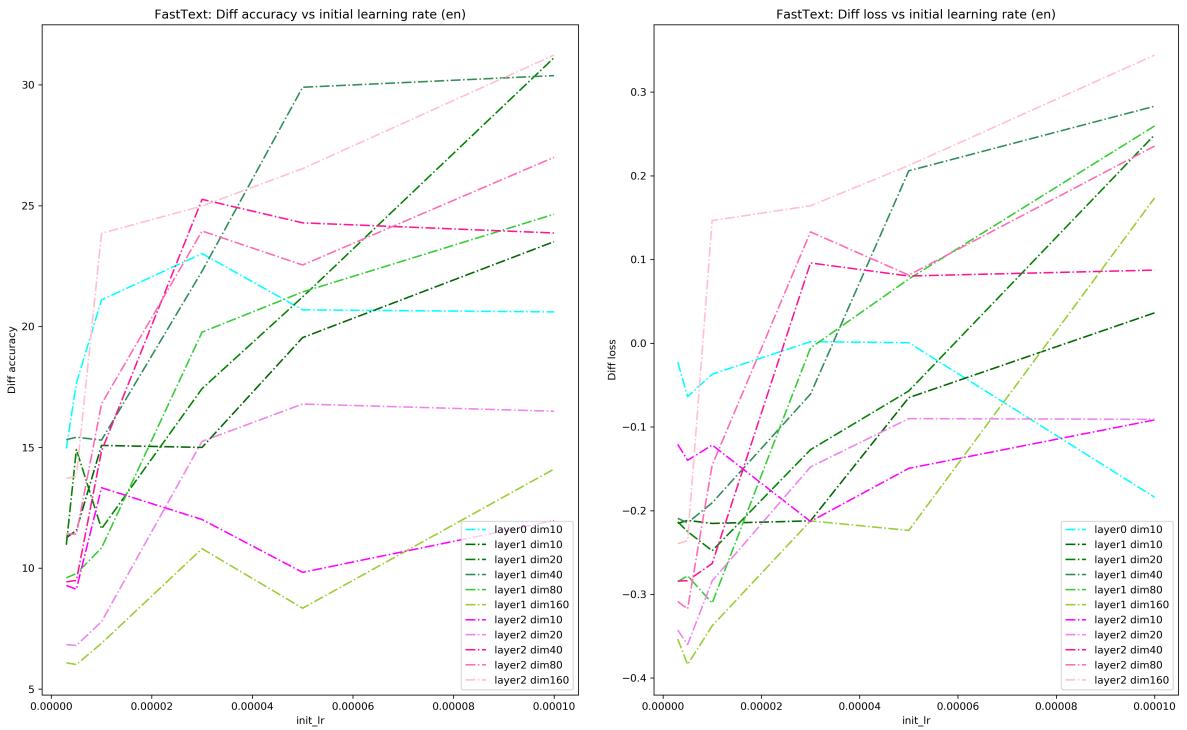

Figure 9: The selectivity (Hewitt and Liang, 2019) and information gain (Pimentel et al., 2020) of probes on FastText. Probes with different capacities are ranked similarly using these two criteria.