While many studies have shown that linguistic information is encoded in hidden word representations, few have studied individual neurons, to show how and in which neuron it is encoded. Among these, the common approach is to use an external probe to rank neurons according to their relevance to some linguistic attribute, and to evaluate the obtained ranking using the same probe that produced it. We show two pitfalls in this methodology:

• It confounds distinct factors: probe quality and ranking quality. We separate them and draw conclusions on each.

• It focuses on encoded information, rather than information that is used by the model. We show that these are not the same.

We compare two recent ranking methods and a simple one we introduce, and evaluate them with regard to both of these aspects.

## Results and Data

We work with three methods to rank neurons according to their importance for a morphological attribute:

- **Linear** [1]: A method that trains a linear classifier on word representations to learn some task $F$. Then, it uses the trained classifier’s weights to rank the neurons according to their importance for $F$.

- **Gaussian** [3]: This method trains a generative classifier on the task $F$, based on the assumption that each dimension in $[z_1, ..., z_d]$ is Gaussian-distributed. Then, it makes use of the decomposability of the multivariate Gaussian distribution to greedily select the most informative neuron, according to the classifier’s performance, at every iteration.

- **Probeless**: A simple method we propose that is based purely on the representations, with no probing involved. For every attribute label $z$, we calculate $\mathbb{E}_p(z)$, the mean vector of all representations of words that possess the attribute and the value $z$. Then, we calculate the element-wise difference between the mean vectors, $r = \sum_{z \in Z} |\mathbb{E}_p(z) - \mathbb{E}_p(z')|$, and obtain a ranking by arg-sorting $r$, i.e., the first neuron in the ranking corresponds to the highest value in $r$.

We use top-to-bottom and bottom-to-top versions of each of the rankings, as well as a random ranking baseline, ending up with 7 rankings overall.

### Data

- 9 different languages from the UD treebanks: Arabic, Bulgarian, English, Finnish, French, Hindi, Russian, Spanish, and Turkish.
- Tasks: predictions of morphological attributes.
- Overall: 156 configs (language, attribute, layer).

### References


## Probing

We probe for morphological attributes in our representations. As probes, we use the two probes proposed by Lin- ear and Gaussian. Each is given a subset of the represen- tation, consisting of neurons selected by ranking of the 7 rankings we work with.

### Interventions

For a representation $h$ with attribute label $z$ and ranking $\Pi(d)$, we modify the $k$-highest-ranked neurons by the ranking, $\Pi(d)_k$, by applying:

$$h_{\Pi(d)_k} = h_{\Pi(d)} + \alpha(z) (z'_{\Pi(d)_k} - z_{\Pi(d)_k}) \quad (2)$$

where $\alpha$ is the same as in Probing ranking, $z'$ is an attribute label such that $z \neq z'$, and $\alpha \in \mathbb{R}$ is a key-scaled coefficients vector in the range $[0, \beta]$, such that the coefficient of the highest-ranked neuron is $\beta$ and that of the lowest-ranked neuron is 0, and $\beta$ is a hyperparameter. Figure 3 is an illustration of the process.

We evaluate our interventions by inspecting the word predicted after applying the intervention. If the predicted word is different than the original one, then the model uses the information we modified. We look for words which have the same lemma as the original one, but a different attribute value, and mark them as CLWV (Correct Lemma, Wrong Value). For example, if the word “makes” becomes “made” when intervening for tense, then it counts as a CLWV, but if it becomes “make” or “prepared” it does not.

### Overlaps

We inspect the 100 top-ranked neurons by Probeless and Gaussian for different configs. We observe significant overlaps between important neurons for the same attribute across languages. We find it reasonable, as we work with a multilingual model.

For Gaussian we do not see the same behaviour, meaning it is less consistent across languages.

### Conclusions

- We compare different methods for ranking neurons according to their importance for a word attribute.
- We show that to evaluate a ranking by probing, one should separate between the ranking itself and the quality of the classifier that is using the ranking.
- While previous work focused on encoded information, we show that it is not the same as used information, by showing that Probeless is superior in the interventions scenario, in contrast to our probing results.