Average Is Not Enough: Caveats of Multilingual Evaluation

Anonymous ACL submission

Abstract

This position paper discusses the problem of multilingual evaluation. Using simple statistics, such as average language performance, might inject linguistic biases in favor of dominant language families into evaluation methodology. We argue that a qualitative analysis informed by comparative linguistics is needed for multilingual results to detect this kind of bias. We show in our case study that results in published works can indeed be linguistically biased and we demonstrate that visualization based on URIEL typological database can detect it.

1 Introduction

The linguistic diversity of NLP research is growing (Joshi et al., 2020; Pikuliak et al., 2021) thanks to improvements of various multilingual technologies, such as machine translation (Arivazhagan et al., 2019), multilingual language models (Devlin et al., 2019; Conneau and Lample, 2019), cross-lingual transfer learning (Pikuliak et al., 2021) or language independent representations (Ruder et al., 2019). It is now possible to create well-performing multilingual methods for many tasks. When dealing with multilingual methods, we need to be able to evaluate how good they really are. Consider the two methods shown in Figure 1 (a). Without looking at the particular languages, *Method A* seems better. It has better results for the majority of languages and its average performance is better as well. However, the trio of languages, where *Method A* is better, are in fact all very similar Iberian languages, while the fourth language is Indo-Iranian. Is the *Method A* actually better, or is it better only for Iberian? Simple average is often used in practice without considering the linguistic diversity of the underlying selection of languages, despite the fact that many corpora and datasets are biased in favor of historically dominant languages and language families.

Additionally, as the number of languages increases, it is harder and harder to notice phenomena such as this. Consider the comparison of two sets of results in Table 1. With 41 languages it is cognitively hard to discover various relations between the languages and their results, even if one has the necessary linguistic knowledge.

In this position paper, we argue that it is not the best practice to compare multilingual methods only with simple statistics, such as average. Commonly used simple evaluation protocols might bias research in favor of dominant languages and in turn hurt historically marginalized languages. Instead, we propose to consider using qualitative results analysis that takes linguistic typology (Ponti et al., 2019) and comparative linguistics into account as an additional sanity check. We believe that this is an often overlooked tool in our research toolkit that should be used more to ensure that we are able to properly interpret results from multilingual evaluation and detect various linguistic biases and problems. In addition to this discussion, which we consider a contribution in itself, we also propose a visualization based on URIEL typological database (Littell et al., 2017) as an example of such qualitative analysis, and we show that it is able to discover linguistic biases in published results.

2 Related Work

Linguistic biases in NLP. Bender (2009) postulated that research driven mainly by evaluation in English will become biased in favor of this language and it might not be particularly language independent. Even in recent years, popular techniques such as *word2vec* or *Byte Pair Encoding* were shown to have worse performance on morphologically rich languages (Bojanowski et al., 2017; Park et al., 2020). Similarly, cross-lingual word embeddings are usually constructed with English as a default hub language, even though this choice might hurt many languages (Anastasopoulos and

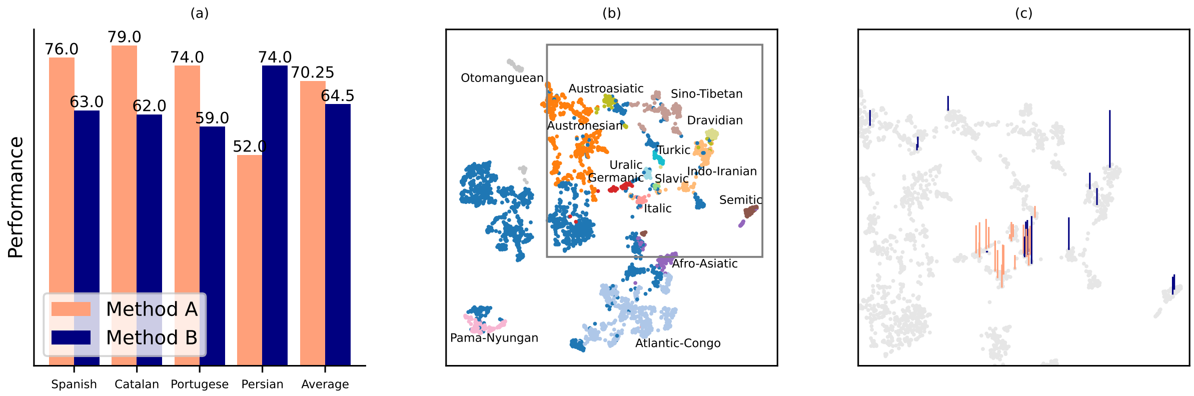

Figure 1: *(a)* Comparison of two methods on unbalanced set of languages. *(b)* Visualization of URIEL languages with certain language families color-coded. *(c)* Comparison of two methods from Rahimi et al. This uses the same map of languages as *b*, but the view is zoomed.

Language	afr	arb	bul	ben	bos	cat	ces	dan	deu	ell	eng	spa	est	pes	fin	fra	heb	hin	hrv	hun	ind
Method A	74	54	54	60	77	79	72	79	64	34	57	76	71	52	69	73	46	58	77	69	61
Method B	59	64	61	70	63	62	62	62	58	61	47	63	64	74	67	57	53	68	61	59	67
Language	ita	lit	lav	mkd	zlm	nld	nor	pol	por	ron	rus	slk	slv	alb	swe	tam	tgl	tur	ukr	vie	**AVG**
Method A	76	75	67	48	63	78	77	77	74	74	36	76	76	76	69	25	57	67	49	48	64.5
Method B	60	62	68	67	66	59	65	61	59	66	53	62	64	69	69	54	66	61	60	55	62.1

Table 1: Comparison of two methods from Rahimi et al. (2019).

Neubig, 2020). Perhaps if the practice of research was less Anglocentric, different methods and techniques would have become popular instead. Our work is deeply related to issues like these. We show that multilingual evaluation with an unbalanced selection of languages might cause similar symptoms.

Benchmarking. Using benchmarks is a practice that came under a lot of scrutiny in the NLP community recently. Benchmark evaluation was said to encourage spurious data overfitting (Kavumba et al., 2019), encourage metric gaming (Thomas and Uminsky, 2020) or lead the research away from general human-like linguistic intelligence (Linzen, 2020). Similarly, benchmarks are criticized for being predominantly focused on performance, while neglecting several other important properties, e.g. prediction cost or model robustness (Ethayarajh and Jurafsky, 2020). Average in particular was shown to have several issues with robustness that can be addressed by using pair-wise instance evaluation (Peyrard et al., 2021). To address these issues, some benchmarks refuse to use aggregating scores and instead report multiple metrics at the same time leaving interpretation of the results to the reader. Gehrmann et al. (2021) is one such benchmark, which proposes to use visualizations to help the intepretation. In this work, we also use visualizations to similar effect.

3 Multilingual Evaluation Strategies

When comparing multilingual methods with non-trivial number of languages, it is cognitively hard to keep track of various linguistic aspects, such as language families, writing systems, typological properties, etc. Researchers often use various simplifying strategies instead:

Aggregating metrics. Aggregating metrics, such as average performance or a number of languages where a certain method achieves the best results provide some information, but as we illustrated in Figure 1 (a), they might not tell the whole story. By aggregating results we lose important information about individual languages and language families. Commonly used statistics usually do not take underlying linguistic diversity into account. This might lead to unwanted phenomena, such as bias in favor of dominant language families. The encoded values of the aggregating metrics might not align with the values we want to express. Average is an example of utilitarianist world view, while using minimal performance might be considered to be a prioritarianist approach (Choudhury and Deshpande, 2021). Even though analyzing the values encoded in metrics is a step towards a fairer evaluation, they still miss a more fine-grained details of the results.

Aggregated metrics for different groups. Another option is to calculate statistics for certain linguistic families or groups. These are steps in the right direction, as they provide a more fine-grained picture, but there are still issues left. It is not clear which families should be selected, e.g. should we average all Indo-European languages or should we average across subfamilies, such as Slavic or Germanic. This selection is ultimately opinionated and different selections might show us different views of the results. In addition, aggregating across families might still hide variance within these families. Grouping languages by the size of available datasets (e.g. low resource vs. high resource) shows us how the models deal with data scarcity, but the groups might still be linguistically unbalanced.

Balanced language sampling. Another option is to construct a multilingual dataset so that it is linguistically balanced. This process is called *language sampling* (Rijkhoff et al., 1993; Miestamo et al., 2016). In practice, this means that a small number of representative languages is selected for each family. The problem with dominant families is solved because we control the number of languages per family. However, selecting which families should be represented and then selecting languages within these families is again an opinionated process. Different families and their subfamilies might have different degrees of diversity. Different selections might favor different linguistic properties and results might vary between them. It is also not clear, how exhaustive given selection is, i.e. how much of the linguistic variety has been covered. Some of the existing works mention their selection criteria: Longpre et al. (2020) count how many speakers the selection covers, Clark et al. (2020) use a set of selected typological properties, Ponti et al. (2020) use the so called *variety language sampling*. Publishing the criteria allows us to do a post-hoc analysis in the future to evaluate, how well did these criteria work.

Qualitative analysis In this paper, we argue that qualitative analysis is an often overlooked, yet irreplaceable evaluation technique. In the following section, we will present our case study of how to perform qualitative analysis.

4 Case Study: Qualitative Analysis through Visualization

In this section we show how to perform a qualitative analysis of multilingual results with a visualization technique based on URIEL typographic database. We show that using this we can (1) uncover linguistic biases in the results, and (2) make sense of results from non-trivial number of languages. As case study, we study results from Rahimi et al. (2019). Our goal is not to evaluate particular methods from this paper, but to demonstrate how linguistically-informed analysis might help researchers gain insights into their results. We analyze the results from this paper not because we want to criticize it, but because it is a well-written paper that actually attempts to do multilingual evaluation for non-trivial number of languages with significantly different methods. The linguistic biases we uncover are already partially discussed in the paper. Here, we only show how to effectively perform qualitative analysis and uncover these biases with appropriate visualization. Appendix A shows similar analysis for another paper (Heinzerling and Strube, 2019) where linguistic biases are visible.

We use URIEL, a typological language database that consists of 289 syntactic and phonological binary features for 3718 languages. We use UMAP feature reduction algorithm (McInnes and Healy, 2018) to create a 2D typological language space. This map is shown in Figure 1 (b). The map is interactive and allows for dynamic filtering of languages and families, as well as inspection of individual languages and their properties[1]. Each point is one language and selected language families are color-coded in the figure. Even though URIEL features used for dimensionality reduction do not contain information about language families, genealogically close languages naturally form clusters in our visualization. Certain geographical relations are captured as well, e.g. Sudanic and Chadic languages are neighboring clusters, despite being from different language families. This evokes the linguistic tradition of grouping languages according to the regions and macroregions. This shows that our visualization is able to capture both intrafamiliar and interfamiliar similarities of languages and is thus appropriate for our use-case.

We visualize results from Rahimi et al. (2019)

[1] Working demo available at Google Colab. Full code will be available at GitHub after acceptance.

on this linguistic map. Rahimi et al. use Wikipedia-based corpus for NER, and they compare various cross-lingual transfer learning algorithms for 41 languages. They use an unbalanced set of languages, where the three most dominant language families – Germanic, Italic and Slavic – make up 55% of all languages. See Appendix A for more details about the paper. We use our URIEL map to visualize a comparison between a pair of methods on all 41 languages from Table 1. In Figure 1 (c) we compare two methods – *Method A* – cross-lingual transfer learning methods using multiple source languages (average performance 64.5), and seemingly worse *Method B* – a low-resource training without any form of cross-lingual supervision (average performance 62.1). We use the same URIEL map, but we superimpose the relative performance of the two methods as colored columns. Orange columns on this map show languages where *Method A* performs better, while blue columns show the same for *Method B*. Height of each column shows how big the relative difference in performance is between the two methods. I.e. taller orange columns mean dominant *A*, taller blue columns mean dominant *B*.

We can now clearly see that there is a pattern in the location of the colored columns. Using average as evaluation measure, *Method A* seems better overall. Here we can see that it is only better in one particular cluster of languages – the cluster of orange columns. All these are related European languages. Most of them are Germanic, Italic or Slavic, with some exceptions being languages that are not Indo-European, but are nevertheless geographical neighbors, such as Hungarian. On the other hand, all the non-European languages actually prefer *Method B*. These are the blue columns scattered in the rest of the space that consists of languages such as Arabic (Semitic), Chinese (Sino-Tibetan) or Tamil (Dravidian).

This shows important fact about the two methods that was hidden by using average. Cross-lingual supervision seemed to have better performance, but it has better performance only in the dominant cluster of similar languages where the cross-lingual supervision is more viable. Other languages, would actually prefer using monolingual low-resource learning, as they are not able to learn from other languages that easily. In this case, average is overestimating the value of cross-lingual learning for non-European languages. This overestimation might cause harm to these languages.

We can also see that there are some exceptions – the blue columns in the orange cluster. These exceptions are Greek, Russian, Macedonian, Bulgarian and Ukrainian – all Indo-European languages that use non-Latin scripts. In this case, different writing systems are probably cause of additional linguistic bias. It might be hard to notice this pattern by simply looking at the table of results, but here we can quickly identify the languages as outliers and then it is easy to realize what they have in common.

Note that we do not expect to see this level of linguistic bias in most papers and we have cherry-picked this particular methods from this particular paper because they demonstrate the case when the linguistic bias in the results is the most obvious. This is caused mainly by unbalanced selection of languages on Wikipedia and in a sense unfair comparison of cross-lingual supervision with low resource learning.

5 Conclusions

Multilinguality in NLP is becoming more common and methodological practice is sometimes lagging behind (Artetxe et al., 2020; Keung et al., 2020; Bender, 2011). Making progress will be inherently hard without proper evaluation methodology. In this work, we argue for necessity for qualitative results analysis and we showed how to use such analysis to improve the evaluation with interactive visualizations. In our case study, we were able to uncover linguistic biases in published results.

Considering the practice in machine learning and NLP, it might be tempting to reduce a multilingual method performance to a single number. However, we believe that intricacies of multilingual evaluation can not be reduced so easily. There are too many different dimensions that need to be taken into consideration and NLP researchers should understand these dimensions. We believe that appropriate level of training in various linguistic fields, such as typology or comparative linguistics, is necessary for proper understanding of multilingual results and for proper qualitative analysis. We argue that qualitative analysis is an oft overlooked approach to results analysis that should be utilized more to prevent various distortions in how we understand linguistic implications of our results.

6 Ethical Considerations

Much of current NLP research is focused on only a small handful of languages. Communities of some language users are left behind, as a result of data scarcity. We believe that our paper might have positive societal impact. It focuses on the issues of these marginalized languages and communities. Following our recommendations might lead to a more diverse and fair multilingual evaluation both in research and in industry. This might in turn led to better models, applications and ultimately quality of life changes for some.

References

Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12710–12718. AAAI Press.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672.

Benjamin Heinzerling and Michael Strube. 2019. Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *CoRR*, abs/2007.15207.

Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426.

Matti Miestamo, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2020. Morphology matters: A multilingual language modeling analysis. *CoRR*, abs/2012.06262.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. 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 2301–2315, Online. Association for Computational Linguistics.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Jan Rijkhoff, Dik Bakker, Kees Hengeveld, and Peter Kahrel. 1993. A method of language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 17(1):169–203.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Rachel Thomas and David Uminsky. 2020. The problem with metrics is a fundamental problem for AI. *CoRR*, abs/2002.08512.

A Details of Analysed Papers

In this appendix, we provide additional information about papers we analysed.

A.1 Rahimi et al.

This is the paper we used for demonstration in the main paper in Section 4. We use results reported in Table 4 in their paper. The languages they use are listed here in Table 2. We can see the apparent dominance of Indo-European languages. There are 14 different methods listed in their paper. We compare the results for these methods in Figure 2. There we can see how the average results for individual methods compare with the average results for non-GIS (Germanic-Italic-Slavic) languages. The numbers correspond to the order of methods listed in the original paper. The two methods compared in Figure 1 (c) are shown as blue and orange, respectively. The orange *Method A* is `BEA`$^{\texttt{tok}}$ in the original paper. The blue *Method B* is called `LSup`. We can see the linguistic bias with this simplistic view as well. All the cross-lingual learning based methods have worse non-GIS results than methods that do not use cross-lingual learning (methods 1 and 2). However, this analysis can not replace the visualization we propose in Section 4. It provides a GIS-centered view, but it can not capture other sources of bias. For example, it does not show various outliers that were seen in the visualization, such as Uralic languages that behave similarly to GIS languages, or Slavic languages with Cyrilic alphabet that behave differently than other Slavic languages.

ISO	Language	Subfamily	Family
bul	Bulgarian	Slavic	Indo-European
bos	Bosnian		
ces	Czech		
hrv	Croatian		
mkd	Macedonian		
pol	Polish		
rus	Russian		
slk	Slovak		
slv	Slovenian		
ukr	Ukrainian		
afr	Afrikkans	Germanic	
dan	Danish		
deu	German		
nld	Dutch		
nor	Norwegian		
swe	Swedish		
cat	Catalan	Italic	
fra	French		
ita	Italian		
por	Portugese		
rom	Romanina		
spa	Spanish		
ben	Bengali	Indo-Iranian	
hin	Hindi		
pes	Iranian Persian		
lit	Lithuanian	Baltic	
lav	Latvian		
ell	Greek		
alb	Albanian		
est	Estonian		Uralic
fin	Finnish		
hun	Hungarian		
ind	Indonesian		Austronesian
tgl	Tagalog		
zlm	Malay		
arb	Standard Arabic		Afro-Asiatic
heb	Hebrew		
vie	Vietnamese		Austroasiatic
tam	Tamil		Davidian
tur	Turkish		Turkic

Table 2: Languages used in Rahimi et al..

A.2 Heinzerling and Strube

Similar linguistic biases can be seen in Heinzerling and Strube as well. They evaluate various representations performance on POS tagging and NER. In Figure 3 we compare POS accuracy of a multilingual model with a shared embedding vocabulary (average performance 96.6, `MultiBPEmb +char +finetune` in the original paper) and a simple BiLSTM baseline with no transfer supervision (average performance 96.4, `BiLSTM` in the original paper). Orange columns are for languages that prefer the multilingual model, blue columns prefer the baseline. In this case, almost all orange columns are in fact GIS languages. Other languages are having significantly worse results with this method and most of them actually prefer the simple baseline with no cross-lingual supervision. This shows the limitations of proposed multilingual

ISO	Language	Subfamily	Family
dan	Danish	Germanic	Indo-European
deu	German		
eng	English		
nld	Dutch		
nor	Norwegian		
swe	Swedish		
bul	Bulgarian	Slavic	
ces	Czech		
hrv	Croatian		
pol	Polish		
slv	Slovenian		
fra	Frech	Italic	
ita	Italian		
por	Portugese		
spa	Spanish		
hin	Hindi	Indo-Iranian	
pes	Iranian Persian		
eus	Basque		Isolate
fin	Finnish		Uralic
heb	Hebrew		Afro-Asiatic
ind	Indonesian		Austronesian

Table 3: Languages used in Heinzerling and Strube.

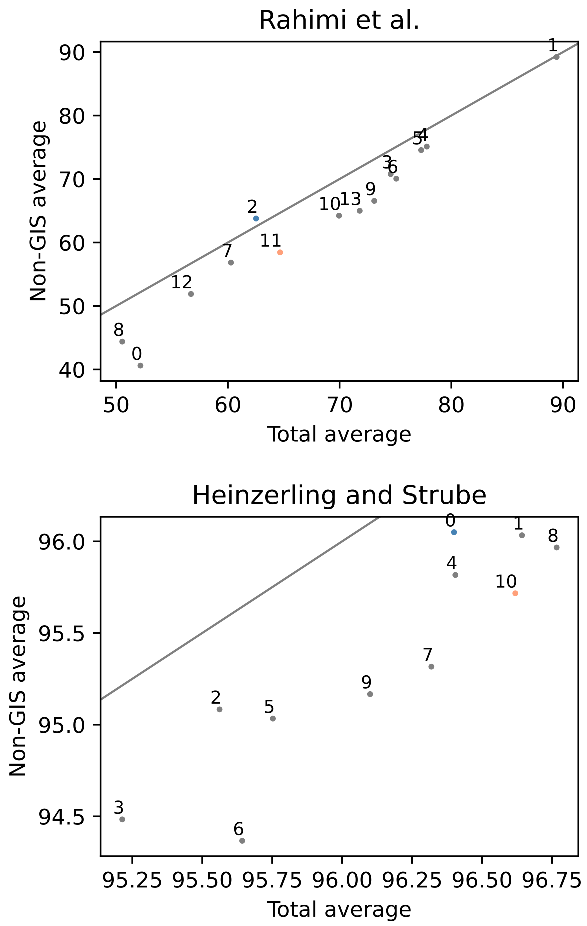

Figure 2: Comparison of method performance. The relation between global average and average on non-GIS languages is shown. Each point represents one method from the papers.

supervision for outlier languages.

We use results reported in Table 5 in their paper. The languages they use are listed here in Table 3. Again, we can see an apparent dominance of GIS languages. There are 11 different methods listed in their paper. We omitted results for additional 6 low resource languages reported in Table 7, because only 4 out of 11 methods were used there. We compare the results for these methods in Figure 2, similarly as in the previous paper. The orange point is the multilingual model, the blue point is the baseline. Now we can see that the BiLSTM baseline is actually the best performing method for non-GIS languages.

B Hyperparameters

We use UMAP python library[2] with the following hyperparameters:

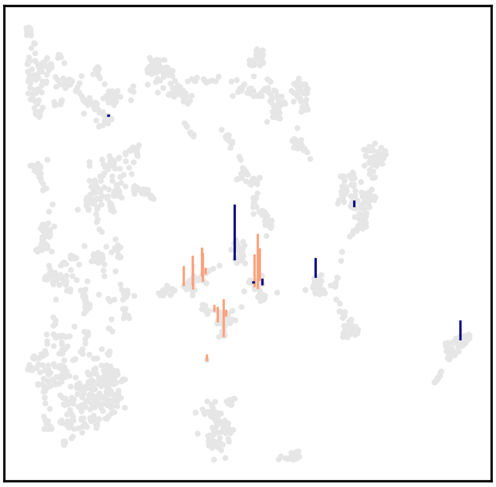

Figure 3: Comparison of two methods from Heinzerling and Strube.

- Number of neighbours (`n_neighbors`): 15
- Distance metric (`metric`): cosine
- Minimal distance (`min_dist`): 0.5
- Random see (`random_state`): 1

C Additional Visualizations

In this Section we show several additional possibilities of using URIEL map of languages to visualize results from multilingual evaluation. Our goal here is to propose additional techniques that can be used for qualitative analysis apart from the comparison of two methods used in Figure 1 in the main body of this paper. This is not an exhaustive list of visualizations. We believe that many other types of visualization can be done using this type of qualitative analysis, based on the needs and requirements of the user.

In Figure 4 we show how to compare more than two methods by visualizing the performance for each method separately. We have created a separate plot for three methods and we can compare their performance visually. We can see that `HSup` method has overall stable high performance. `LSup` has worse performance, but its still quite balanced. Finally, `BWET` has similar performance as `LSup`, but we can see that there are regions where it fails, e.g. the languages in the rightmost part of the figure have visibly worse performance.

In Figure 5 we show yet another type of visualization. In this case, we simply visualize what method is the best performing for each language.

[2] umap-learn.readthedocs.io

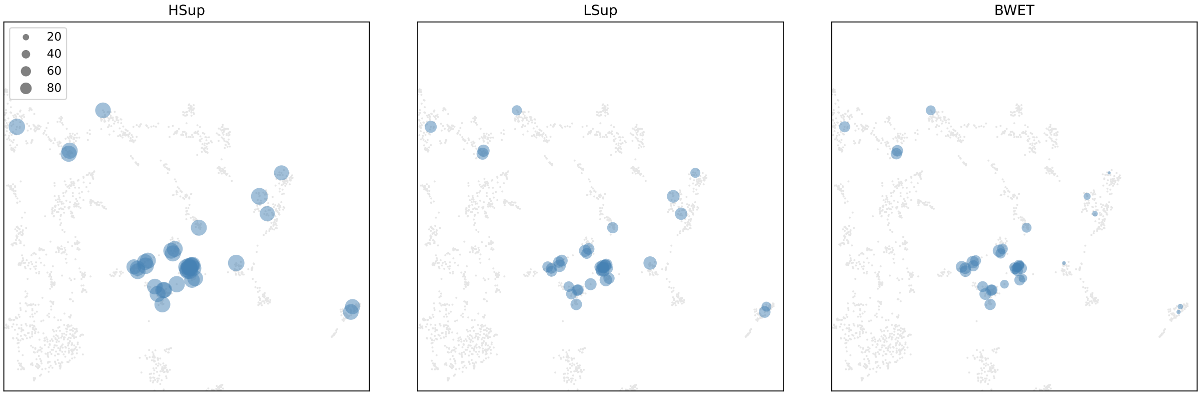

Figure 4: Comparison of multiple methods using size to mark method performance for individual languages. `HSup`, `LSup` and `BWET` are methods reported in (Rahimi et al., 2019).

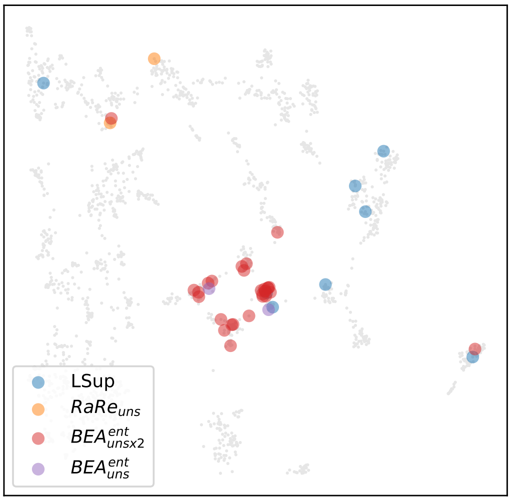

Figure 5: The best performing methods for various languages.

We compare methods using crosslingual supervision and low-resource training (`LSup`). From seven methods, only four achieved the best performance for at least one language and those are shown in the Figure. Again, we can see similar picture as before. One method ($BEA^{ent}_{uns\times 2}$) is the best performing method taking average into account. However, in this visualization we can see that it is actually the best performing method only in the dominant cluster of European languages. Elsewhere, other methods perform better.