Getting The Most Out of Your Training Data: Exploring Unsupervised Tasks for Morphological Inflection

Abhishek Purushothamal Adam Wiemerslage
Katharina von der Wense 1 Georgetown University 2 University of Colorado Boulder 3 Johannes Gutenberg Univers

Abstract

Pretrained transformers such as BERT (Devlin et al., 2019) have been shown to be effective in many natural language tasks. However, they are under-explored for character-level sequence-to-sequence tasks. In this work, we investigate pretraining transformers for the character-level task of morphological inflection in several languages. We compare various training setups and secondary tasks where unsupervised data taken directly from the target task is used. We show that training on secondary unsupervised tasks increases inflection performance even without any external data, suggesting that models learn from additional unsupervised tasks themselves. In addition, we find that standard denoising tasks can hurt in multi-task setups, but using external data for denoising solves this issue.

1 Introduction

Morphological inflection is the task of generating a word form given a lemma and a set of morphological features. For example, given the lemma *walk* and features `V;PST`, the inflected form is *walked*. Morphological inflection is a core task in computational morphology and has been explored in a variety of settings, including as part of the SIGMORPHON-UniMorph shared tasks (Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022; Goldman et al., 2023). In this work, we focus on low-resource morphological inflection, where the amount of training data is limited.

Recent advances in NLP have been driven by pretraining large-scale models on large amounts of text using self-supervised objectives such as masked language modeling (Devlin et al., 2019) and denoising sequence-to-sequence objectives (Lewis et al., 2020; Raffel et al., 2019). While such objectives have been successful for token-level tasks and semantic tasks, they are less explored for character-level sequence-to-sequence tasks such as morphological inflection. At the same time, the computational morphology community is frequently interested in low-resource languages and settings where large external corpora may be unavailable.

We explore whether unsupervised tasks constructed from the available morphological inflection data can improve model performance, and when denoising objectives may hurt in multi-task learning. We investigate:

- Pretraining setups using only the target task data, including two-stage training and multi-task learning.
- Secondary tasks: character-level masked language modeling and character-level autoencoding.
- The effect of using additional unlabeled data from Universal Dependencies (UD) treebanks for denoising tasks.

150-639-2 Language UD teebank used I 150-639-2 Language UD Treebank used afb Arabic, Gulf Arabic-PADT ita Italian Italian-ISDT amh Amharic Amharic-ATT jpn Japanese Japaese-GSD arz Arabic, Egyptian kat Georgian bel Belarusian Belarusian-HSE k1r Khaling dan Danish Danish-DDT mkd Macedonian deu German German-GSD nav Navajo eng English English-Atis rus Russian Russian-GSD fin Finnish Finnish-FTB san Sanskrit Sanskrit-UFAl, fra French French-GSD sme Sami North,Sami-Giella grc Ancient Greek Ancient$_Greek - PerseusspaSpanishSpanish - AnCora hebHebrewHebrew - HTBsqiAlbanian heb(_unvoc)Hebrew, UnvocalizedswaSwahili hunHungarianHungarian-SzegedturTurkishTirrkish - Atis hyeEasternArmenianArmenian - ArmTDP$

Table 1: The 27 typologically diverse languages (Subsection 4.1) from the 2023 shared task, all of which are investigated in this work. We use some UD treebanks for analytical experiments in Subsection 6; the specific treebanks are listed in the final column.

2 Related Work

Multi-task learning (MTL) has long been studied as a way to improve generalization by training on multiple related tasks (Caruana, 1997; Luong et al., 2016). In NLP, intermediate-task and supplementary training can improve downstream performance (Phang et al., 2018; Pruksachatkun et al., 2020). Identifying beneficial task relations has also been explored (Bingel and Søgaard, 2017; Martinez Alonso and Plank, 2017; Fifty et al., 2021).

Denoising objectives such as MLM (Devlin et al., 2019) and sequence-to-sequence denoising (Lewis et al., 2020; Vincent et al., 2010) have been widely used in pretraining. For character-level models, ByT5 (Xue et al., 2022) explores byte-to-byte pretraining. However, the role of such objectives in low-resource character-level tasks is less clear.

Morphological inflection has a rich history, including shared tasks and neural approaches using encoder-decoder models and transformers (Kann and Sch"utze, 2016; Wu et al., 2021). Unlabeled data has been used for morphological generation (Kann and Sch"utze, 2017), and dataset quality and sampling issues have been studied (Kodner et al., 2023; Muradoglu and Hulden, 2022). Noise in morphological inflection has also been investigated (Wiemerslage et al., 2023).

3 Morphological Inflection

We follow the standard formulation: given a lemma and morphological features, generate the inflected form. Inputs and outputs are treated as character sequences. We focus on typologically diverse languages from the SIGMORPHON-UniMorph 2023 shared task (Goldman et al., 2023), and create low-resource training subsets.

4 Data

We use the 27 languages from the SIGMORPHON-UniMorph 2023 shared task (Goldman et al., 2023). For each language, we subsample training data to simulate low-resource scenarios.

5 Baseline Model

Our baseline is a transformer encoder-decoder model operating at the character level (Wu et al., 2021), implemented with `yoyodyne`.

6 Training Methods

We evaluate several training setups.

6.1 Baseline

We train the model on the supervised morphological inflection task only.

6.2 Two-stage pretraining (PT)

We perform two-stage training: first train on an unsupervised objective using unlabeled data derived from the task data, then fine-tune on supervised morphological inflection.

6.3 Multi-task learning (MTL)

We train on the supervised task jointly with an unsupervised auxiliary task. For MTL, the loss is a combination:

$$\mathcal{L} = \mathcal{L} * \mathrm{SUP} + \lambda \mathcal{L} * \mathrm{UNSUP}. \tag{1}$$

We set λ to [ILLEGIBLE].

6.4 Auxiliary tasks

We use two auxiliary tasks:

- Character-level masked language modeling / denoising (CMLM): apply noise to the input sequence and train to reconstruct the original.
- Autoencoding (AE): reconstruct the original sequence from itself.

6.5 Noise

We use a span-mask-based corruption process. Let x be the input sequence. We sample spans and apply replacements [ILLEGIBLE]. The mask sampling rate is a hyperparameter [ILLEGIBLE]. We also explore using external unlabeled data from UD treebanks.

7 Results and Analysis

We report development and test accuracies for each language and model variant.

7.1 When Does Denoising Hurt MTL?

There is a remarkable gap in performance between MTL-AE and MTL-CMLM. The CMLM denoising objective is the worst performing setup, performing below the baseline on average. In further analysis, performing CMLM on external data that is separate from the finetuning data solves this issue, resulting in significantly better performance.

```
Baseline PT-CMLM PT-AE MTL-CMLM MTL-AE Language rso 639-2 Dev Test Dev Test
Dev Test Dev Test Dev Test Arabic, Gulf afb 68.8 69.4 72.2 70.5 72.t 7 t.9
68.8 67.8 72,'7 72.7 Amharic amh 44.6 42.9 48.0 50.8 56.5 66.0 34.9 36.7
56.5 61.4 Arabic, Egyptian atz 82.8 82.5 83.1 83.9 82.3 84.3 80.7 81.4
83.6 83.8 Belarusian bel 61 .2 59.0 62.9 6r.8 61 .5 58.7 s9.8 56.5 64.4
61.7 Danish dan 81 .7 80.  r 81 .7 80.s 81.2 't9.9 80.0 80.7 83.2 82.5
German deu 68.2 7t.2 '70.3 68.7 74.4 '73.t 65.8 65.',7 74,3 73.2 English
eng 91.6 88.2 91.5 88.6 91 .8 90.3 89.5 87.2 92.3 90.9 Finnish fin 74.6
56.7 7s.7 61.6 78.2 61.8 58.9 44.0 81.4 68.6 French fra 15.2 65.2 76.9
68.0 80.6 68.9 69.9 6'7.0 81.1 73.6 Ancient Greek
```
$c54.t33.160.44L.352.834.5 + -1.328.656.640.'7 Hebrewheb74.272.176.676.03'77, 676.t372.272.6180.377.95 Hebrew, Unvocalizedheb_unvoc81.5$

Table 2: The development and test accuracies of the 5 model variants, for all the 27 languages. For each language, the highest development accuracy is underlined and highest test accuracy is bolded.

IMAGE NOT PROVIDED

Figure 1: Figure 1: The distribution of performance (test set accuracy) for each model variant on the various data sizes. Distributions are plotted as violin plots, with box plots visualizing the mean, first and third quartile, and min and max values.

7.2 External Data for Denoising

We prepare additional unlabeled data from UD treebanks and use it for denoising in MTL. Results are shown below.

8 Future Work

The denoising tasks requires hyperparameters for the instrumentation of the noise. Due to this, further work is required in exploring these tasks under different hyperparameter settings with multiple methods to shed light on their sensitivity and ability to improve models for character-level tasks such as morphological inflection and G2P. Future work should also consider exploring more secondary tasks, especially based on particular morphological phenomenon in diverse languages.

Limitations

- Our work is limited to the character-level task of morphological inflection. Thus, findings may not hold for other similar tasks such as G2P and interlinear glossing.
- Considering the sensitivity of training methods to vocabulary and data sizes, it is unclear whether these results can be extrapolated to different scenarios.
- Our work does not explore the disparity of performance of the methods across languages and requires expert analysis over various of linguistic features.

```
Baseline MTL-CMLM MTL-AE MTL-CMLM.UD MTL-AE-UD Language ISO 639-2 Dev Test
Dev Test Dev Test Dev Test Dev Test Arabic, Gulf afb 68.8 69.4 68.8 67.8
72.7 72.7 72.2 72.6 72.8 74.9 Amharic amh 44.6 42.9 34.9 36.7 s6.5 6t.4 56.3
5’,7.7 61.0 66.6 Belarusian bel 6t.2 59.0 59.8 56.5 64.4 6t.7 64.2 61 .5 65.3
62.2 Danish dan 81 .7 80.1 80.0 80.7 83.2 82.5 82.3 80.8 83.7 82.9 German deu
68.2 7t.2 65.8 65.7 74.3 73.2 75.4 74.4 75.4 76.3 English eng 91.6 88.2 89.s
87.2 92.3 90.9 91.3 88.5 9t.9 88.9 Finnish fin 74.6 56.7 58.9 44.0 81.4 68.6
81 .5 70.8 82.7 73.6 French fra 75.2 6s.2 69.9 67,0 81.  I 73.6 82.8 75.2
85.8 74.1. Ancient Greek grc 54.t 33.  I 43.3 28.6 56.6 40.7 64,2 46,5 63.5
47.t Hungarian hun 75.7 65.7 65.4 61.3 80.4 ’7 1.7 81.1 75.2 83.6 78.1 Hebrew
heb 74.2 72.t 72.2 12.61 80.3 77.95 78.6 78.55 79.3 75.73 Eastern Armenian
hye ’79,2 79.4 76.8 76.0 86.9 89.5 90.5 89.0 9t.4 93.0 I J t a a p li a a n n
ese j i a ta p 9 1 0 5 .  .  5 8 8 2 5 0 .  .  7 1 8 4 3 .1 .3 7 5 t .  .  6
4 9 1 4 5 .  .  0 4 2 90 t.  .  9 4 9 3 4 4 .  , 8 3 8 3 8 2 .  .  7 2 9 4 4
4 .  .  3 1 9 4 3 2.  .  8 3 Russian ruS 78.’7 16.6 ’72.7 ’t t.7 80.9 81.8 81
.7 80.1 81.8 82.9 Sanskrit san 55.0 49.0 47.6 s0.5 63.4 56.4 65.4 57.9 65.7
58.3 Sami sme 57.3 43.9 44.2 33.8 70.0 60.4 ’70.2 66.7 74.8 66.3 Spanish spa
88.2 8s.0 19.3 78.9 91.6 90.9 91 .5 90.3 91 .8 91.8 Turkish tur 85.3 85.  r
76.4 73.4 89.7 89.5 87.5 85.9 89.6 89.9 Avg 69.51 64.39 62.45 58.98 ’74.s8
’7t.28 76.3t 72.22 78.09 74.66
```

Table 3: Results for our models by language from the experiments with external data, reporting development and test accuracy. For each language, the highest development accuracy is underlined and highest test accuracy is bolded. Note: results for non “-UD” models are identical to Table 2.

Acknowledgments

We thank the anonymous reviewers for their useful suggestions and feedback and the NALA Lab at the University of Colorado Boulder. This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

References

[1] Sina Ahmadi and Aso Mahmudi. 2023. Revisiting and amending Central Kurdish data on UniMorph 4.0. In Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 38–48, Toronto, Canada. Association for Computational Linguistics.

[2] Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 115–125, Online. Association for Computational Linguistics.

[3] Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kiera6, G6bor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser,

William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaf Zu-maeta Rojas, Didier L6pez Francis, Arturo Oncevay, Juan L6pez Bautista, Gema Celeste Silva Villegas, Lucas Tonoba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritviin Karah6la, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czamowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zafuoh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra [ILLEGIBLE].

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. [ILLEGIBLE].

[5] Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

[6] Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. In Neural Information Processing Systems.

[7] Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON-UniMorph 2023 shared task 0: Typologically diverse morphological inflection. In Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

[8] Katharina Kann and Hinrich Schiitze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 555–560.

[9] Katharina Kann and Hinrich Schtitze. 2017. Unlabeled data for morphological generation with character-based sequence-to-sequence models. In Proceedings of the First Workshop on Subword and Character Level Models in NLP, pages 76–81, Copenhagen, Denmark. Association for Computational Linguistics.

[10] Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3121–3126, PortoroZ, Slovenia. European Language Resources Association (ELRA).

[11] Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, G6bor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kiera5, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah young, and Ekaterina Vylomova. 2022. SIGMORPHON-UniMorph 2022 shared task 0: Generalization

and typologically diverse morphological inflection. In Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, PhonoLogy, and Morpholog), pages 176–203, Seattle, Washington. Association for Computational Linguistics.

[12] Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Litr. 2023. Morphological inflection: A reality check. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.

[13] Kundan Krishna, Saurabh Garg, Jeffrey Bigham, and ZachNy Lipton. 2023. Downstream datasets make surprisingly good pretraining corpora. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12201–12222, Toronto, Canada. Association for Computational Linguistics.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyet. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

[16] Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In International Conference on Learning Representations.

[17] H6ctor Martinez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 44–53, Valencia, Spain. Association for Computational Linguistics.

[18] Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1294–1303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[19] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajid, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 40341043, Marseille, France. European Language Resources Association.

[20] 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.

[21] Jason Phang, Thibault F6vry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088.

[22] Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bemardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, [ILLEGIBLE], and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 229–259, Online. Association for Computational Linguistics.

[23] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. Intermediate-task transfer learning with pre-trained language models: When and why does it work? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5231–5241.

[24] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Lilu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, l-ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

[26] Pascal Vincent, H. Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res., 11:3371–3408.

[27] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, St6fan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Ant6nio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17:261–272.

[28] Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miik:ka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 1–39, Online. Association for Computational Linguistics.

[29] Adam Wiemerslage, Changbing Yang, Garrett Nicolai, Miikka Silfverberg, and Katharina Kann. 2023. An investigation of noise in morphological inflection. In Findings of the Association for Computational Linguistics: ACL 2023, pages 3351–3365, Toronto, Canada. Association for Computational Linguistics.

[30] Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1901–1907, Online. Association for Computational Linguistics.

[31] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics, 10:291–306.

[32] Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noemi Aepli, Hamid Aghaei, Zeljko Agi6, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabriele Aleksandrav-iditte, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranz-abe, Bilge Nas Ancan, Ii6runn Amard6ttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Asgeirsd6ttir, Deniz Baran Aslan, Cengiz Asmazoflu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augusti-nus, Mariana Avelds, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkadur Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shab-nam Behzad, Kepa Bengoetxea, ibrahim Benli, Yifat Ben Moshe, Gdzde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agne Bielinskiene, Kristin Bjarnad6ttir, Rogier Blokland, Victoria Bobicev, Loic Boizou, Emanuel Borges Vdlker, Carl Bdrstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adri-ane Boyd, Anouck Braggaar, Ant6nio Branco, Kristina Brokaite, Aljoscha Burchardt, Marisa Cam-pos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, S6rgio Castro, Tatiana Caval-canti, Gtlqen Cebiroflu Eryilit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomir d6pki, Neslihan Cesur, Savas Cetin, Ozlem Qetinoflu, Fabri-cio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinkov6, Aur6lie Collomb, Qalrr Qdltekin, Miriam Connor, Daniela Corbetta, Fran-cisco Costa, Marine Courtin, Mihaela Cristescu, In-gerid Lpyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz De-rin, Elvis de Souza, AranaaDiaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, TomaZ Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Fa-cundes, RichSrd Farkas, Federica Favero, Jannatul Ferdaousi, Marilia Fernanda, Hector Fernandez A1-calde, Amal Fethi, Jennifer Foster, ClSudia Freitas, Kazunori Fujita, Katar(na Gajdo5ov6, Daniel Gal-braith, Federica Gamba, Marcos Garcia, Moa Giir-denfors, Fabricio Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gdkrrmak, Yoav Goldberg, Xavier G6mez Guinovarl, Berta Gonz6lez Saavedra, Bernadeta Gricifite, Matias Grioni, Loic Grobol, Nor-munds Gruzrtis, Bruno Guillaume, C6line Guillot-Barbance, Tunga Giingdr, Nizar Habash, Hinrik Haf-steinsson, Jan Hajid, Jan Hajid jr., Mika Hiimtil[i-nen, Linh HA My, Na-Rae Han, Muhammad Yud-istira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladk6, Jaroslava HlavSdov6, Florinel Hociung, Petter Hohle, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, An-ton Karl Ingason, Radu lon, Elena Irimia, Olijide Ishola, Artan Islamaj,

Kaoru Ito, Siratun Jannat, Tom65 Jelfnek, Apoorva Jha, Katharine Jiang, An-ders Johannsen, Hildur J6nsd6ttir, Fredrik JOr-gensen, Markus Juutinen, Hiiner Kaqrkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritviin Karah6[a, An-dre KAsen, Tolga Kayadelen, Sarveswaran Kengath-araiyer, Vdclava Kettnerov6, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Kdhn, Abdul-latif Kciksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Kdse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaite, Simon Krek, Parameswari Kr-ishnamurthy, Sandra Ktibler, Adrian Kuqi, O[uzhan Kuyrukgu, Ash Kuzgun, Sookyoung Kwak, Kris Kyle, Veronika Laippala, Lorenzo Lambertino, Ta-tiana Lando, Septina Dian Larasati, Alexei Lavren-tiev, John Lee, Phng L0 Hdng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lind6n, Yang Janet Liu, Nikola Ljube5id, Olga Loginova, Ste-fano Lusito, Andry Luthfi, Mikko Luukko, Olga Lya-shevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Bi.igra Margan, Cltdlina Mdr5n-duc, David Maredek, Katrin Marheinecke, Stella Markantonatou, H6ctor Martinez Alonso, Lorena Martin Rodriguez, Andr6 Martins, Cl6udia Mar-tins, Jan Ma5ek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuin-ness, Gustavo Mendonga, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilli, C5tllin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Mol-niir, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Miiiirisep, Pinkey Nainwani, Mariam Nakhl6, Juan Ignacio Navarro Horfliacek, Anna Nedoluzhko, Gunta Ne5pore-Berzkalne, Manuela Nevaci, Lng Nguy6n Thi, Huybn Nguy6n Thi Minh, Yoshi-hiro Nikaido, Vitaly Nikolaev, Rattima Nitis-aroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda 6lad6ttir, Ad6dayo Ohidkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Osthng, Lrlja @vrelid,
aziyeB.etiilOzateg, MerveOzgelik, ArzucanOzgtir, BalkrzOztiirkBagaran, TeresaPaccosi, AlessioPalme
Passos, Giu−liaPedonese, AngelikaPeljak−tapifska, SiyaoPeng, SiyaoLoganPeng, RitaPereira, SflviaP
AugustoPerez, NataliaPerkova, GuyPerrier, SlavPetrov, DariaPetrova, AndreaPeverelli, JasonPhe−
lan, JussiPiitulainen, YuvalPinter, ClaraPinto, TommiAPirinen, EmilyPitler, MagdalenaPlamada, Bar
maker, MizanurRahoman, TarakaRama, LoganathanRamasamy, JoanaRamos, FamRashel, Moham−
madSadeghRasooli, VinitRavishankar, LivyReal, PetruRebeja, SivaReddy, MathildeRegnault, GeorgRehr
triRizqiyah, LuisaRocha, EirfturRdgnvaldsson, IvanRoksandic, MykhailoRomanenko, RudolfRosa, Vale
son, ManuelaSanguinetti, EzgiSanryar, DageSiirg, MartaSartor, MitsuyaSasaki, BaibaSaulrte, YaninSa
ert, EinarFreyrSigurdsson, JodoSilva, AlineSil−veira, NataliaSilveira, SaraSilveira, MariaSimi, RaduS
nava, TedSither, MariaSkachedubova, AaronSmith, IsabelaSoares −
Bastos, PerErikSolberg, BarbaraSonnenhauser, ShafiSourov, RacheleSprugnoli, Vi −
vianStamou, Steinli6rSteingrimsson, AntonioStella, AbishekStephen, MilanStraka, EmmettStrickland, J
burini, MaryAnnC.Tan, TakaakiTanaka, DiptaTanaya, MirkoTavoni, SamsonTella, IsabelleTellier, Mar
ers, Sveinbjcimli6rdarson, Vilhj6lmurliorsreinsson, SumireUematsu, RomanUntilov, ZdeikaUre5ov6, La
janvanNoord, ViktorVarga, UlianaVedenina, GiuliaVenturi, VeronikaVincze, NataliaVlasova, AyaWaka
gailWalsh, JonathanNorthWashington, MaximilanWendt, PaulWidmer, ShiraWigderson, SriHartatiW
tern, TsegayWoldemariam, Tak−sumWong, AlinaWr6blewska, MaryYako, KayoYamashita, NaokiYama
CZdigitallibraryattheInstituteofFormalandAppliedLinguistics(OFAL), FacultyofMathematicsandPhys

A Data details

A.1 Limitations of UniMorph and SIGMORPHON

The unimorph project is the primary source for the dataset. It draws heavily from Wiktionary[1]](https://www.wiktiona in a semi-automated way based on Kirov et al. (2016). Wiktionary is a collaboratively built resource which, despite processes to promote accuracy, is not a linguistic resource that is considered as gold-standard data. The semi-automated methodology, sources, and broad mandate limits the utility and effectiveness of the dataset. A notable example is Ahmadi and Mahmfii (2023), which discusses this in the context of Sorani (ckb) also known as Central Kurdish (not one of the 27 languages in this work). The limitations of the dataset used in this work, being only very recently released, are not well-studied, and consequently also apply to our work.

A.2 [ILLEGIBLE]

Selection and Sampling

Many features of morphological inflection data, such as overlap and frequency, have been shown to be important factors for model performance (Kodner et a1., 2023). (Muradoglu and Hulden, 2022) demonstrated how data could be sampled using active learning methods to improve model performance. Since we investigate training methods rather than data methods, we perform analysis on data which has been selected specifically for benchmarking purposes. We recommend the readers check Section 4 "Dala preparation" of the shared task paper Goldman et at. (2023) for more information on the data methods used for target-task data selection and splits. We discuss details relevant to our selection and sampling below.

Lemma Overlap The 2023 shared task dataset was specifically designed to prevent lemma overlap between any of dev, train, and test. Since we only sub-sample from train, the lack of lemma overlap is maintained in our datasets, and is thus not a relevant point of analysis as in other work (e.g. Kodner et al. (2023)).

A.3 Preparing Additional Data from UD Treebanks

With a fixed seed, we randomly sample words from the selected UD Treebank to prepare an unlabeled training set of size 2k for each language. We perform sampling only after filtering out NUM and PUNCT tagged and tokenized words (Nivre et al., 2020). We do not otherwise use the token-level annotations from UD, simulating a more realistic data setting than the one UniMorph words represent. Table 1 shows the 19 languages from the shared task for which UD was used for additional training data in our investigation of the denoising task in the MTL setup. We list the specific treebanks used in order to encourage reproducibility. We preserve both the data and corpus information for the selected words. Specifically, we have also collected the token frequency, UPOS frequency, and character frequency for each of the additional data sampled, to be made available with the code for future analysis.

B Models and Experimental Details

B.1 Implementation

All models are implemented with a fork of yoyodyne, which is built over pytorch-lightning (Falcon and The PyTorch Lightning team, 2019). We utilize yoyodyne's existing implementation of the Wu

[1] `[https://www.wiktionary.org/`

et al., 2021 models. We additionally implemented the CMLM objective, two stage training for PT setup, and the MTL setup including data and loss combination using the framework.

B.2 Compute and Infrastructure

For reproducibility, we utilize only Nvidia V100 GPUs for our experiments. The reported models together required ∼180 hours of GPU time.

B.3 Reproducibility

In addition to using a consistent GPU architecture, we use a fixed random seed of 1 for all our model experiments. We also maintain copies of the specific data.

B.4 Morphological Inflection in Japanese

Organizers of the 2023 shared task note the challenges that Japanese presents in morphological inflection, namely due to its extremely large vocabulary size. In our work this persists as most models perform poorly on Japanese and do not meaningfully improve upon the baseline.

C Significance Testing

In order to analyze the significance of our results, we perform a paired permutation test between test accuracies of all the models compared to the baseline. For all these tests, we use the null-hypothesis that the mean difference between the test accuracies for these pairs is 0 and run the tests with 100k sampled permutations of the differences using SciPy (Virtanen et al., 2020).