Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise

Noëmi Aepli[1] and **Rico Sennrich**[1,2]
[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh
{naepli,sennrich}@cl.uzh.ch

Abstract

Cross-lingual transfer between a high-resource language and its dialects or closely related language varieties should be facilitated by their similarity. However, current approaches that operate in the embedding space do not take surface similarity into account. This work presents a simple yet effective strategy to improve cross-lingual transfer between closely related varieties. We propose to augment the data of the high-resource source language with character-level noise to make the model more robust towards spelling variations. Our strategy shows consistent improvements over several languages and tasks: Zero-shot transfer of POS tagging and topic identification between language varieties from the Finnic, West and North Germanic, and Western Romance language branches. Our work provides evidence for the usefulness of simple surface-level noise in improving transfer between language varieties.

1 Introduction

Recent research has achieved impressive results in zero-shot cross-lingual transfer based on multilingual pre-training (Devlin et al., 2019; Conneau and Lample, 2019) or monolingual transfer of embeddings (Artetxe et al., 2020). However, these methods require large amounts of unlabeled data in the target language (Lauscher et al., 2020) and do not take into account surface similarity between languages except for the sharing of subword units in multilingual models. For the transfer between closely related languages and dialects, we deem it desirable to exploit the similarity of surface representations. Specifically, we target orthographic variations that commonly result from pronunciation differences between closely related languages. [1]

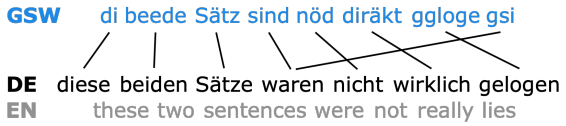

Figure 1: Swiss German (GSW) sentence with corresponding standard German (DE) and English (EN) translations. The sentence shows various spelling differences on the word level, and reordering occurs on the sentence level due to different past-tense formation.

In this paper, we propose to augment the training data of a high-resource language with character-level noise to simulate spelling variations and thus facilitate generalization to closely related[2] low-resource languages.

We test this strategy on two tasks and several language regions. The considered tasks are part-of-speech (POS) tagging on the word level and topic classification on the sentence level; the languages are from the Finnic, West and North Germanic, and Western Romance language branches. We observe that our baseline method for cross-lingual transfer learns undesirable heuristics, e.g., assigning unseen words to open word classes in POS tagging and that injecting noise reduces this bias. Our experiments show absolute accuracy improvements between 1.4 and 22 percentage points over the state of the art, providing evidence that a simple data-augmentation strategy can boost transfer learning for language varieties and dialects with a closely related high-resource language.

2 Related Work

Zero-shot cross-lingual transfer based on multilingual language models (Devlin et al., 2019; Conneau and Lample, 2019) or machine translation models (Siddhant et al., 2020) has turned out

[1]Note that there are also differences on different levels as described in Hollenstein and Aepli (2014) and partly observable in Figure 1 which illustrates a German example sentence with a closely related variant.

[2]Language relatedness is on a continuum, and the difference between dialects and distinct languages is often political. Hence we use a broader term to indicate that the method is not limited to dialects.

Findings of the Association for Computational Linguistics: ACL 2022, pages 4074 - 4083
May 22-27, 2022

to be surprisingly effective. Such representations proved themselves beneficial for a range of diverse tasks (Hu et al., 2020). However, they still require large-scale data sets to train, making them impractical for low-resource languages, to which dialects and language varieties typically belong.

Artetxe et al. (2020) introduce zero-shot cross-lingual transfer by mapping monolingual representations between languages. They also propose adding Gaussian noise to the embeddings during the fine-tuning step. Huang et al. (2021) also operate in the embedding space by constructing robust regions in the embedding space to tolerate noise in the contextual embedding. These are not ideal strategies for closely related languages because words with similar surface forms could still be far from each other in an embedding space.

Surface-level noise such as character substitutions, insertions, and deletions has been proposed as an effective data augmentation strategy for machine translation (Sperber et al., 2017; Heigold et al., 2018; Belinkov and Bisk, 2018; Karpukhin et al., 2019; Vaibhav et al., 2019; Anastasopoulos et al., 2019). Authors report improvements in system accuracy due to more robustness towards speech recognition errors, spelling mistakes, and other naturally occurring noise in text data. Even though cross-lingual transfer between closely related languages has received some attention (Muller et al., 2020; Sakaguchi et al., 2017; Zeman et al., 2017, 2018), it has not been investigated whether this transfer can be improved with character-level noise inserted at training time. We tackle this in our work by adding random character-level noise to the training data of a standard language and applying the model to closely related languages.

Exploiting orthographic similarity to improve cross-lingual transfer between closely related languages is currently an understudied area. Relevant previous work has been done by Sharoff (2018), who used orthographic similarity to refine bilingual dictionary induction.

Transliteration is another line of related work that focuses on improving the transfer between closely related languages with different alphabets (Durrani et al., 2014; Lin et al., 2016; Murikinati et al., 2020; Han and Eisenstein, 2019). On the other hand, our work focuses on languages using the same script. The recent report by Muller et al. (2021) investigating transfer between the same and different alphabets involves a zero-shot task transfer which is, however, preceded by a language model training on (unlabeled) target language data. To the best of our knowledge, we are the first to focus on zero-shot transfer learning techniques for closely related languages.

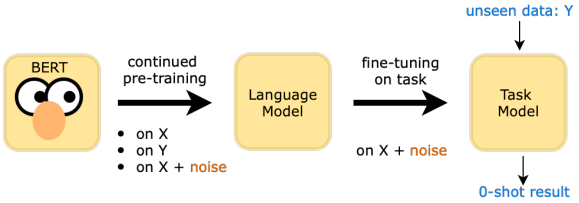

Figure 2: Methodology for zero-shot cross-lingual transfer: We first continue the pre-training of a language model (LM) on text. Then, we fine-tune the adapted LM to task T in high-resource language X. We augment the training data for continued pre-training or fine-tuning with character-level noise and apply the model to task T in a closely related low-resource language Y.

3 Method

Consider a high-resource language X and a closely related low-resource language Y. We perform zero-shot cross-lingual transfer by pre-training a model on unlabeled data from X (and optionally Y if available), then fine-tuning on task T in language X. The resulting model is applied to task T in language Y. This procedure is illustrated in Figure 2. Thus, our question is: How can we best make the model trained on X generalize to Y?

Ideally, such a model would take surface similarity of words into account for generalization. We hypothesize that in closely related languages, unknown words in the low-resource language are likely to correspond to similar known words in the high-resource language in function and meaning.[3] However, state-of-the-art language models represent words through subwords (Sennrich et al., 2016). This representation is sensitive to slight surface variations: minor changes to a string will lead to different segmentations and internal representations.

To account for this problem, we apply noise to the surface representation of words in language X. We implement this through character-level noise, i.e., we randomly[4] select $10\% - 15\%$ of the tokens

[3] Please refer to Table A1 for examples.

[4] We tested more linguistically motivated constraints on the character replacements but did not find it to have a big effect.

Noise	DE-BERT	DE-BERT + GSW	DE-BERT + DE	DE-BERT + DE + Noise
✗	50.66	**72.1**	52.08	53.88
✔	72.77	**82.11**	71.13	70.45

Table 1: POS tag accuracy for Swiss German (GSW) on different language models fine-tuned on German (DE) training data with and without noise.

of a sentence[5] excluding numbers and punctuation. One randomly selected character of the chosen token undergoes one of three possible operations: *delete*, *replace*, *insert*, each with equal probability. The latter two operations work with an additional randomly selected character of the (extended) Latin Alphabet for the source language. The following sentence *This is a short example.* will end up with some sort of a "typo", e.g., as *This i**u**s a short example.* Noise can be applied during pre-training on X, and/or during fine-tuning on a task T for language X.

Another possibility to alleviate the subword representation problem is BPE-dropout (Provilkov et al., 2020), which applies different segmentations to words in a randomized fashion. BPE-dropout was originally motivated to increase robustness for morphological variance. We hypothesize that it is similarly effective for orthographic variance; see Table A1 for examples. Overall, both character-level noise and BPE-dropout encourage the model to learn generalizations across similar surface strings via shared subwords.

A second motivation for character-level noise is that we aim to imbue the model with different inductive biases. For example, a model for POS tagging might learn that only a small set of words can map to closed word classes such as articles, whereas unknown words are likely to belong to an open word class such as named entities. Training with character-level noise will reduce this bias.

4 Experiments

We design our experimental procedure[6] to answer the following question: Does character-level noise improve zero-shot transfer to closely related languages? Within three controlled experiments, we ablate the importance of the noise-augmentation strategy. We select two cross-lingual tasks: 1) POS tagging (15% noise) and 2) topic classification (10% noise). While the former task illustrates the strategy's potential on word level, the latter provides insight into how much it helps on text level.

4.1 POS Tagging for Swiss German Dialects

As base models, we use the German "dbmdz" BERT[7] (DE-BERT) and mBERT (Devlin et al., 2019). We continue pre-training on the *SwissCrawl* corpus (Linder et al., 2020) for the Swiss German (GSW) LM-adaptation and the DE part of *The Credit Suisse News Corpus* (Volk et al., 2018) for the German LM-adaptation. For task fine-tuning, we use a DE UD treebank and for the evaluation a part of *NOAH's Corpus* (Hollenstein and Aepli, 2014).

As shown in Table 1, all settings profit from fine-tuning with noise and bring about improvements of up to 22 percentage points (DE-BERT without pre-training). The best result with an accuracy of 82.11% for zero-shot GSW POS tagging is achieved with a GSW-adapted language model and task fine-tuning on a noised DE corpus. Considering the case where no GSW text data is available for language model adaptation, we still achieve an accuracy of 77.11% for zero-shot GSW POS tagging with mBERT fine-tuned on noised DE data (see Table 2).

4.2 POS Tagging with mBERT

We fine-tune mBERT on a UD corpus of a language already seen during mBERT pre-training: DE, Finnish (FI), Swedish (SV), French (FR), or Icelandic (IS, Arnardóttir et al. (2020)) and test on a closely related language variety absent from mBERT: GSW, Old French (OFR), Livvi (OLO, Pirinen (2019)), Karelian (KRL, Pirinen (2019)), or Faroese (FO, Tyers et al. (2018)). In addition to noise, we added experiments with a BPE-dropout of 0.1 (empirically selected) during the fine-tuning step.

Table 2 illustrates that the method works well for closely related language varieties (upper part) but less for other language pairs, which are more distant (lower part). We do see an occasional improvement for more distant language pairs, but they

[5] We relied on previous work by Vincent et al. (2008) where similar choices were made regarding the amount of noise.

[6] We work with code bases by Wolf et al. (2020) and Wang et al. (2021), multilingual BERT (mBERT), and the data sets' default splits. Most of the corpora we work with were provided by the Universal Dependency project (UD, Nivre et al. (2016)); refer to Appendix A.2 for details.

[7] https://github.com/dbmdz/berts#german-bert

Languages	Baseline	BPE-Dropout	Noise	BPE-Drop-out+Noise
DE→GSW	73.14	76.48	77.11	**78.13**
FI→OLO	69.32	69.66	**73.03**	71.76
FI→KRL	72.44	76.35	**79.18**	78.57
SV→FO	84.76	86.20	**87.63**	87.31
IS→FO	85.94	86.80	87.43	**87.46**
FR→OFR	63.42	66.65	66.73	**67.27**
DE→FO	81.74	81.34	81.38	**82.27**
DE→OLO	**52.63**	52.09	51.10	49.26
DE→KRL	**57.51**	57.47	55.71	53.37
DE→OFR	**44.08**	39.17	38.32	40.03
FR→OLO	56.49	56.72	**58.59**	56.64
FR→KRL	59.46	62.27	**64.52**	64.15
FR→FO	81.13	82.09	81.81	**82.62**

Table 2: Zero-shot POS tagging accuracy of different strategies for several languages (TRAIN→TEST). The training and test languages are closely related in the upper but not in the lower part of the table as indicated by the colors (Finnic, West Germanic, North Germanic, and Western Romance language branches.) Noise consistently adds additional accuracy points beyond BPE-dropout performance increase.

are generally smaller and less consistent than the improvements for the closely related languages we evaluated.

Furthermore, we have to consider the (much) lower baseline where an accuracy gain does not have the same impact. Hence, the two strategies BPE-dropout and noise improve the zero-shot performance for POS tagging over several closely related language pairs. While BPE-dropout shows some performance gain over the baseline, character-level noise adds additional accuracy points.

4.3 Cross-dialect Topic Identification

We work with mBERT and *MOROCO: The Moldavian and Romanian Dialectal Corpus* (Butnaru and Ionescu, 2019). The data set contains 33,564 news domain text samples, each belonging to one of six topics (culture, finance, politics, science, sports, tech). We fine-tune mBERT on topic identification on Moldavian (MD) and evaluate on Romanian (RO) and vice versa. We emphasize the difference between this sentence-level task and the previous word-level task. While POS tagging works on the word form and can benefit from transferring prior information about probable POS sequences, topic classification is mainly meaning-oriented, making a transfer more challenging.

Topic identification results in Table 3 show that

Training	Noise	Test	Accuracy
MD	✗	RO	63.34
	✔		**68.48**
RO	✗	MD	81.65
	✔		**83.01**

Table 3: Results for Moldavian (MD) vs. Romanian (RO) cross-dialect topic identification. Training with noise improves the transfer by 5.1 (MD→RO) respectively 1.3 (RO→MD) percentage points.

fine-tuning with noise consistently improves the accuracy. However, noise-augmented training data appears to have a more substantial effect when transferring from MD to RO (5.1 percentage points) than vice-versa (1.4 percentage points). This is interesting given that RO represents the high-resource standard language in this context (being one of the languages used to train mBERT), while MD is its low-resource variety. We conjecture that this is caused by the fact that the model trained in MD struggles with word meaning and is, therefore, more sensitive to variations than its RO counterpart.

5 Analysis

Figure 3 illustrates the prediction differences of a model trained with and without noise. We observe that the models trained without noise have learned a tendency towards labeling unknown words as open-class words such as names (NE) or adjectives (ADJD), with the label for foreign words (FM) being massively overpredicted, while it tends to undergenerate closed-class tags such as articles (ART) or adverbs (ADV). In contrast, the model trained with noise comes much closer to the gold standard tag distribution. It has learned to rely more on probable POS tag sequences than on the surface form of a token. Consider e.g. the GSW article *d* (DE *die*; the). In the DE training corpus, the token appears only as a foreign word (FM) because it also happens to be a French word, but the model trained with noise is more likely to correctly tag it as an article, relying more on context than just strict mappings.[8]

Figure 4 depicts the per-type F1 change for the most frequent STTS (Schiller et al., 1999) tags. The past participles of the auxiliary verbs (VAPP) form another closed class which profits substantially from a model focusing on tag sequences given the compound structure of the perfect tense. As Swiss German does not have a simple past, the

[8]For more examples, please refer to the Appendix A.1.

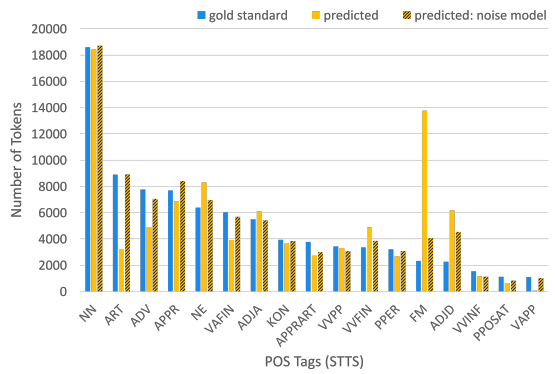

Figure 3: Number of tokens per POS tag in the gold standard vs. predictions of two models, fine-tuned with and without noise. Only the most frequent STTS tags are displayed.

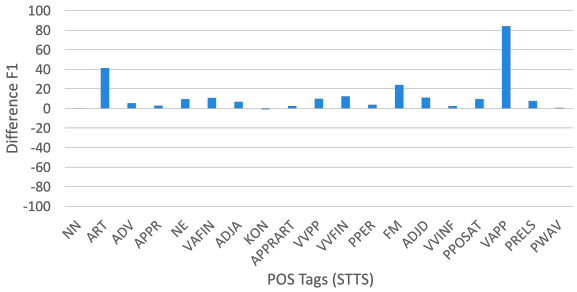

Figure 4: Per-type F1 change of the most frequent STTS tags illustrating which tags profit the most when the model is fine-tuned with noise.

perfect tense is much more frequent than it is in German.

5.1 Conditional Random Field

Given that one potential consequence of adding noise is that the model relies more on surrounding context and probable POS tag sequences (rather than strict word-level mappings), we compare our results to a method that explicitly models tag sequences, Bi-LSTMs+CRF (bidirectional long-short-term memory + conditional random field). This method was used to achieve state-of-the-art performance for POS tagging (Huang et al., 2015). For the implementation, we added a CRF layer on top of BERT.[9]

The added CRF layer did not improve the performance of the fine-tuned mBERT model for zero-shot POS tagging for Swiss German trained on German, as presented in Table 4. In contrast, noise injection has proven effective in both configurations.

[9]Making use of the TorchCRF library: https://github.com/s14t284/TorchCRF.

Noise	mBERT	mBERT+CRF
✗	70.24	69.26
✔	**78.57**	76.90

Table 4: Zero-shot POS tag accuracy for Swiss German on mBERT & mBERT+CRF models trained on German with and without noise.

6 Discussion

Our investigation into cross-lingual zero-shot transfer between closely related languages demonstrates that simple data augmentation with character-level noise can successfully improve transfer, with absolute improvements ranging from 1.4 (RO→MO transfer of topic identification) to 22 (DE→GSW transfer of POS tagging in the case of DE-BERT without pre-training) percentage points.

The examination of prediction errors shows that a baseline BERT model has learned heuristics for unseen words that are undesirable for transfer between closely related languages. In contrast, a model trained with noise can combat this bias without substantial performance losses in the source languages.

We expect that the final effectiveness of using character-level noise for zero-shot cross-lingual transfer depends on the task and language characteristics. We plan to evaluate the effect of character-level noise in a broader range of settings in future work. More broadly, we encourage further research that exploits surface-level word similarity for cross-lingual transfer between related languages and dialects, rather than focusing purely on vector space representations.

Acknowledgments

We thank Yves Scherrer for input regarding data sets and the anonymous reviewers for their valuable comments. This project has received funding from the Swiss National Science Foundation (project nos. 191934 & 176727).

Ethical Considerations

Our work did not involve any new data collection or annotation and did not require in-house workers or introduce any new models and related risks. Instead, we examine how character-level noise can help the transfer between closely related languages, especially in low-resource zero-shot settings.

References

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. 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 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Þórunn Arnardóttir, Hinrik Hafsteinsson, Einar Freyr Sigurðsson, Kristín Bjarnadóttir, Anton Karl Ingason, Hildur Jónsdóttir, and Steinþór Steingrímsson. 2020. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25, Barcelona, Spain (Online). Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR-2018)*, Vancouver, BC, Canada.

Andrei Butnaru and Radu Tudor Ionescu. 2019. Moroco: The moldavian and romanian dialectal corpus. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 688–698.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 148–153, Gothenburg, Sweden. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Ying Lin, Xiaoman Pan, Aliya Deri, Heng Ji, and Kevin Knight. 2016. Leveraging entity linking and related language projection to improve name transliteration. In *Proceedings of the Sixth Named Entity Workshop*, pages 1–10, Berlin, Germany. Association for Computational Linguistics.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2020. Can multilingual language models transfer to an unseen dialect? A case study on north african arabizi. *arXiv preprint arXiv:2005.00318*.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Tommi A Pirinen. 2019. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in Karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136, Paris, France. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. Robsut wrod reocginiton via semi-character recurrent neural network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3281–3287. AAAI Press.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Serge Sharoff. 2018. Language adaptation experiments via cross-lingual embeddings for related languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8854–8861. AAAI Press.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation (IWSLT 2017)*, pages 90–96, Tokyo, Japan.

Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150, Brussels, Belgium. Association for Computational Linguistics.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. 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 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA. Association for Computing Machinery.

Martin Volk, Alena Zwahlen, and Chantal Amrhein. 2018. Credit Suisse News Corpus (Release 05).

XML-Format. A collection of translated news in English, French, German and Italian.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

A.1 Analysis

Table A1 shows Swiss German (GSW) words (and their corresponding standard German (DE) form and English (EN) translation) that had the highest accuracy increase when using a part-of-speech (POS) tagging model trained with character-level noise compared to the model trained without noise. These words were wrongly tagged with open-class tags by the baseline model. However, the model trained with noise was able to reduce this bias and thus correctly tag them with their closed-class tag.

Furthermore, in most cases, one *substitution/insertion/deletion*-operation on the DE word would not suffice to get an exact match with the GSW word. This indicates that it is unnecessary to design a noise function that closely mirrors the linguistic differences between variants.

A.2 Data Sets

A.2.1 Universal Dependencies

Table A2 contains the Universal Dependencies treebanks (UD, Nivre et al. (2016)) we used in this work. The treebanks can be downloaded via `https://universaldependencies.org/#download`.

A.2.2 Other Data Sets

Table A3 shows data sets apart from UD that we used in this work.

GSW	DE	EN	Most frequent POS without noise	Correct POS	Error reduction with noise (relative/absolute)
ond	und	and	FM	KON	99.00% (-104)
worde	geworden	become	VVPP, ADJD	VAPP	98.73% (-156)
dr	der	the	FM, NE, NN	ART	98.46% (-128)
häd	hat	had	VVFIN, FM	VAFIN	98.21% (-55)
gsi	gewesen	been	VVPP, FM	VAPP	98.19% (-434)
vu	von	from	FM, NE	APPR	98.18% (-108)
eme	einem	a	ADJA, FM	ART	97.96% (-48)
grad	gerade	just	ADJD	ADV	97.59% (-81)
vum	vom	from the	FM, APPR	APPRART	96.49% (-55)
de	der	the	FM, NE, ADJA	ART	95.76% (-1558)

Table A1: Swiss German (GSW) words (and their corresponding standard German (DE) form and English (EN) translation) with the highest error reduction using a part-of-speech (POS) tagging model trained with character-level noise compared to the model trained without noise.

Usage	Language (ISO)	Language Branch	Treebank	# Sentences
Training	Finnish (FI)	Finnic	TDT	15K
	French (FR)	Western Romance	GSD	16K
	German (DE)	West Germanic	HDT	190K
	Icelandic (IS)	North Germanic	IcePaHC	39K
	Swedish (SV)	North Germanic	Talbanken	5K
Test	Faroese (FAO)	North Germanic	OFT	1208
	Karelian (KRL)	Finnic	KKPP	228
	Livvi (OLO)	Finnic	KKPP	125
	Old French (OFR)	Western Romance	SRCMF	1927

Table A2: Universal Dependency treebanks we used for our experiments with the number of sentences ("# Sentences") we used for training or testing (specified in "usage").

Corpus name (Language)	Size	Link
Moroco (MD & RO)	33.5K text samples	https://github.com/butnaruandrei/MOROCO
NOAH's Corpus (GSW)	7.3K sentences	https://noe-eva.github.io/NOAH-Corpus/
SwissCrawl (GSW)	500K sentences	https://icosys.ch/swisscrawl
The Credit Suisse News Corpus (DE)	105K sentences	https://pub.cl.uzh.ch/projects/b4c/en/corpora.php

Table A3: Data sets we used for our experiments in addition to the UD treebanks in Table A2.