Backtranslation in Neural Morphological Inflection

Ling Liu and **Mans Hulden**
University of Colorado
first.last@colorado.edu

Abstract

Backtranslation is a common technique for leveraging unlabeled data in low-resource scenarios in machine translation. The method is directly applicable to morphological inflection generation if unlabeled word forms are available. This paper evaluates the potential of backtranslation for morphological inflection using data from six languages with labeled data drawn from the SIGMORPHON shared task resource and unlabeled data from different sources. Our core finding is that backtranslation can offer modest improvements in low-resource scenarios, but only if the unlabeled data is very clean and has been filtered by the same annotation standards as the labeled data.

1 Introduction

Both machine translation (MT) and morphological inflection generation are string transduction tasks: MT is typically treated as word-level (or subword-level) string transduction while morphological inflection generation can be treated as character-level string transduction. MT models and techniques can usually be naturally applied to morphological inflection, as is shown in recent work on morphological inflection (Liu, 2021; Kann and Schütze, 2016; Cotterell et al., 2016, 2017, 2018; Liu et al., 2018; McCarthy et al., 2019; Vylomova et al., 2020; Wu et al., 2020; Moeller et al., 2020, 2021).

Backtranslation (Sennrich et al., 2016) has become a common practice in machine translation in low-resource scenarios (Fadaee et al., 2017; Edunov et al., 2018; Hoang et al., 2018; Xia et al., 2019; Chen et al., 2020; Edunov et al., 2020; Marie et al., 2020; Liu et al., 2021). There has been work on data augmentation for morphological generation in low-resource scenarios (Silfverberg et al., 2017; Bergmanis et al., 2017; Anastasopoulos and Neubig, 2019; Liu and Hulden, 2021), but no previous work has applied the backtranslation technique. In this paper, we propose to apply backtranslation as a data augmentation method in morphological inflection under low-resource circumstances. Our evaluation of the method on six different languages with unlabeled data from different resources indicates that backtranslation can only improve morphological inflection in low-resource scenarios when the unlabeled data set is very clean and has been filtered by the same annotation standards as the labeled data.

(a) Labeled data

LEMMA	INFLECTED FORM	MSD TAG
enchain	enchained	V;V.PTCP;PST
bristle	bristle	V;NFIN
enforce	enforcing	V;V.PTCP;PRS
meagre	meagred	V;V.PTCP;PST
remarry	remarried	V;PST
…	…	…

(b) Morphological inflection

Input

LEMMA	MSD TAG
enchain	V;V.PTCP;PST
bristle	V;NFIN
enforce	V;V.PTCP;PRS
meagre	V;V.PTCP;PST
remarry	V;PST
…	…

Output

INFLECTED FORM
enchained
bristle
enforcing
meagred
remarried
…

(c) Morphological analysis

Input

INFLECTED FORM
enchained
bristle
enforcing
meagred
remarried
…

Output

LEMMA	MSD TAG
enchain	V;V.PTCP;PST
bristle	V;NFIN
enforce	V;V.PTCP;PRS
meagre	V;V.PTCP;PST
remarry	V;PST
…	…

Figure 1: Data example for morphological inflection and morphological analysis.

2 Method

The backtranslation method comes from machine translation. Suppose we need to translate from language A to language B, and we have a parallel text corpus. Suppose further that we have additional monolingual data for B. The idea in backtransla-

Proceedings of the Second Workshop on Insights from Negative Results in NLP, pages 81–88

ISBN 978-1-954085-93-0

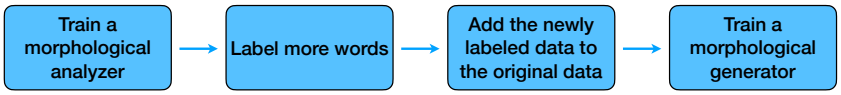

Figure 2: Pipeline for applying backtranslation to morphological inflection.

tion is to train an MT model for B-to-A translation using parallel text, use this MT model to translate our monolingual B data into A, and then add that translated data to the A-to-B parallel text corpus to (re)train an A-to-B translation model. For morphological inflection, the labeled data is usually of the type shown in Figure 1(a), where we are provided with triplets consisting of lemma, inflected form, and a morphosyntactic description (MSD) tag corresponding to the inflected form. In a morphological inflection task, the input is the lemma and the MSD, while the expected output is the inflected form, as shown in Figure 1(b). To apply the backtranslation technique to the morphological inflection task, we can follow a pipeline like the one illustrated in Figure 2: leverage the labeled data to train a morphological analyzer instead of a generator, apply the morphological analyzer to tag more unlabeled words with MSDs, and then add the newly labeled data to the original data to train models for morphological inflection. When training the morphological analyzer, the input is the inflected form and the output is the lemma and the MSD, as is illustrated in Figure 1(c).

3 Experiments and Results

We conduct several experiments to evaluate the performance of morphological inflection with the backtranslation data augmentation technique. The deep learning architecture we use is the Transformer model (Vaswani et al., 2017) as implemented in Fairseq (Ott et al., 2019). For our experiments, we use the same hyperparameter settings as the best-performing system (Liu and Hulden, 2020b,a) in the SIGMORPHON 2020 shared task on inflection (Vylomova et al., 2020). All models have been trained with a single NVIDIA Tesla P100 GPU.

Data Our experiments cover six languages: Czech, Finnish, German, Russian, Spanish and Turkish. These languages are selected to include variety in morphological inflection complexity and difficulty. Finnish and Turkish are agglutinative languages, both of which have vowel harmony and extensive agglutination. Spanish has a rich inflectional system, but is quite regular. Czech is a Slavic language that uses a Latin writing system, and is a fusional language with rich morphology. Russian is also a Slavic language with a rich fusional morphological inflection system, and is written in Cyrillic script. German has a relatively limited inflectional system, but is challenging due to a high rate of syncretism. Table 1 provides more details on the languages.

We follow two settings for our low-resource experiments: 1,000 and 500 training triplets. For the 1,000 training example setting, we use the medium-size setting data from CoNLL-SIGMORPHON 2018 shared task on type-based morphological inflection (Cotterell et al., 2018). For the 500 training triplet setting, we randomly sample 500 examples from the 1,000 setting training examples. The two training data size settings are designed with the consideration that, on the one hand, data augmentation is not necessary when abundant training data is available, and on the other that if training data is too limited, a morphological analyzer of useful quality is not trainable. The development set and test set we use are the 2018 SIGMORPHON shared task development and test sets, unchanged. Each of the development set and the test set for a language contains 1,000 triples respectively.[1]

Our initial experiments used random words from Wikipedia as unlabeled data to be backtranslated with the morphological analyzer, but these pilot experiments showed a significant decrease in the inflection performance after the backtranslated data were added. Table 3 in Appendix B shows the performance of each language with 500 original training examples after adding different amount of backtranslated Wikipedia words. We hypothesized that the reason for the decrease may be that the words available from Wikipedia often represent parts-of-speech (e.g. determiners, adverbs, etc) not found in the labeled data and thus introduce excessive noise. Therefore, we changed the source of our un-

[1]Thanks to one of the reviewers for pointing out that the amount of the development data makes the experiment not really so "low-resource". We agree that 1,000 triples for validation would be very difficult to obtain in an extremely low-resource situation.

Language	Language group	Morphological typology	Writing system
Czech	Indo-European/Balto-Slavic/Slavic/West Slavic	fusional	Latin script
Finnish	Uralic/Finnic	agglutinative	Latin script
German	Indo-European/Germanic/West Germanic	fusional	Latin script
Russian	Indo-European/Balto-Slavic/Slavic/East Slavic	fusional	Cyrillic script
Spanish	Indo-European/Italic/Romance/Western Romance	fusional	Latin script
Turkish	Turkic/Common Turkic/Oghuz	agglutinative	Latin script

Table 1: Language information.

labeled data and conducted further experiments on two sources: inflected words with labels removed in the CoNLL-SIGMORPHON 2018 shared task high-setting training set which are not included in the medium-setting training set, and words from the Universal Dependencies (UD) (version 2.6) corpus (Zeman et al., 2020) for each language, which are of the same parts-of-speech included in the shared task data. Details of the treebank data we use for each language are provided in Table 2 in Appendix A.

Transformer inflection and analyzer performance We first evaluate the base performance of the inflection model trained with only the 500 or the 1,000 triplet set. The accuracy results are presented in Figure 3.

As it has been noted that the quality of the backtranslation model (in our case, the morphological analyzer) is positively correlated to the ability of backtranslation data augmentation to yield improvements (Currey et al., 2017), we present the morphological analyzer accuracy in Figure 4. The development and test data for the morphological analyzer is created by simply reversing the input and output of the development and test set data for morphological inflection.

The reported accuracy for each morphological inflection model and each morphological analysis model are the average of five runs with different random initializations to ensure a good representation of the model performance.

Morphological inflection performance with backtranslation data augmentation We experimented with backtranslating different amounts of delabeled shared task data or UD data. For each data amount, shared task inflected word forms are randomly sampled (uniformly) to match the desired target size for backtranslation (0-9,000 words). Considering that in real low-resource situations the words we can obtain are usually frequently used

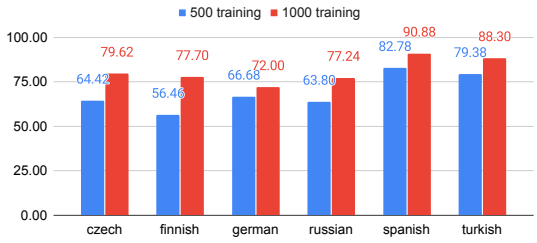

Figure 3: Basic Transformer inflection performance at different training data sizes: 500 or 1,000 training examples.

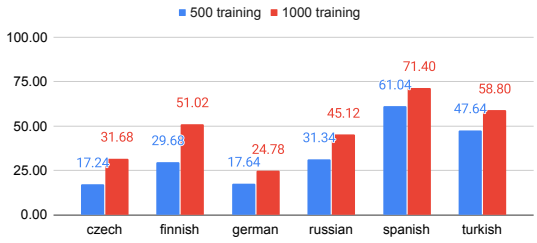

Figure 4: Transformer morphological analyzer performance at different training data sizes: 500 or 1,000 training examples.

ones, when picking from the UD data, we first rank the words from most frequent to least frequent in the corpus and pick the most frequent UD words of the respective parts-of-speech used in the given training data. We use one random morphological analyzer trained in the previous step to label the words, and add the resulting automatically labeled words to the original training data to train the augmented inflection models. Each augmented inflection model is trained with five runs using different random initializations. We use a majority vote by these five models to pick the final prediction.

The inflection performance obtained by adding different amounts of backtranslated data to the original 500 training triplets is presented in Figure 5. The legend in each plot indicates the best accuracy with the corresponding backtranslation data augmentation size for each language. Figure 5(a)

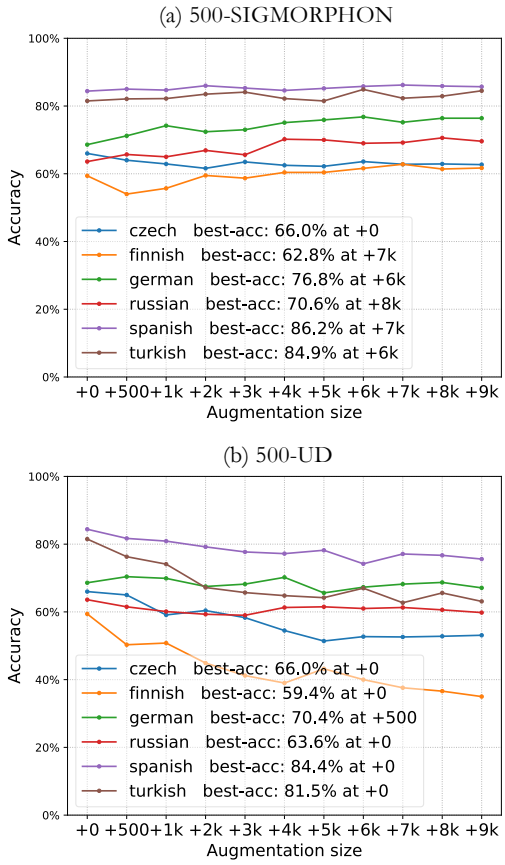

Figure 5: Performance of the Transformer inflection model trained with backtranslated SIGMORPHON shared task data or UD data on top of **500** labeled data points.

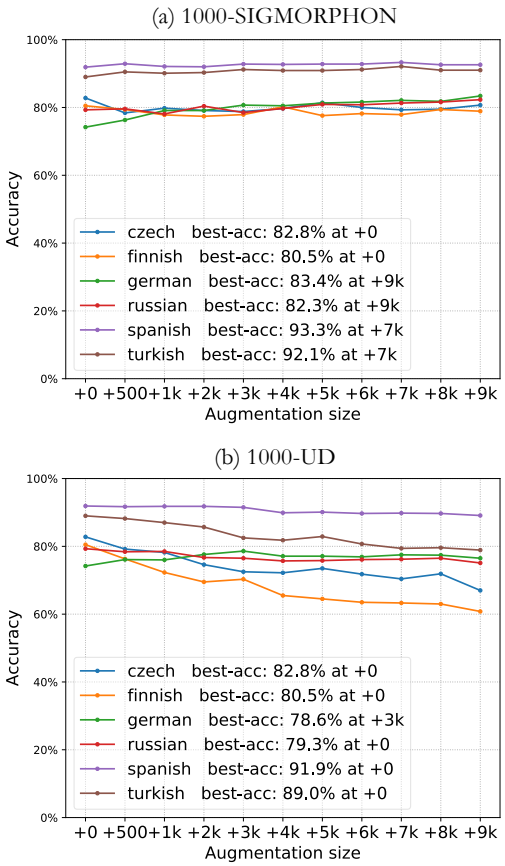

Figure 6: Performance of the Transformer inflection model trained with backtranslated SIGMORPHON shared task data or UD data on top of **1,000** labeled data points.

shows the results for adding backtranslated CoNLL-SIGMORPHON shared task data; here, we see that adding backtranslated data improves the inflection model for all languages except for Czech. However, to our surprise, the results for adding backtranslated UD words, as illustrated in Figure 5(b), show that adding backtranslated data actually hurts the inflection model.

The pattern is similar when the initial training data contains 1,000 examples, shown in Figure 6: though we see that adding backtranslated shared task words improves the inflection model, adding backtranslated UD words causes the model to deteriorate. This opposite tendency goes quite against our expectations, especially considering that the UD words were selected to ensure that they are of the same parts-of-speech covered in the original training data. In order to explain the opposite tendency and answer whether backtranslation could indeed be helpful for morphological generation, we conducted the following experiments on comparing different ways of adding backtranslated data.

Morphological inflection with tagged backtranslation Caswell et al. (2019) show that tagging backtranslated source sentences with an extra distinguishing token can improve the contribution backtranslated data can provide to machine translation. This finding is supported in later work (Marie et al., 2020). Therefore, we hypothesize that adding a special tag to the lemma and MSD tag sequence predicted by the morphological analyzer may improve the performance of the inflection model trained with the combination of the original training data and the backtranslated data.

In order to test the hypothesis, we start with experiments on the 500 training example setting. We train one morphological analyzer for each language, and use the morphological analyzer to label words from CoNLL-SIGMORPHON 2018 shared task not included in the current training set. Then we add the newly labeled data (in differing amounts as in the earlier experiment) in two different ways: (1) add the backtranslated data to the original training data without any special tag; (2) append a

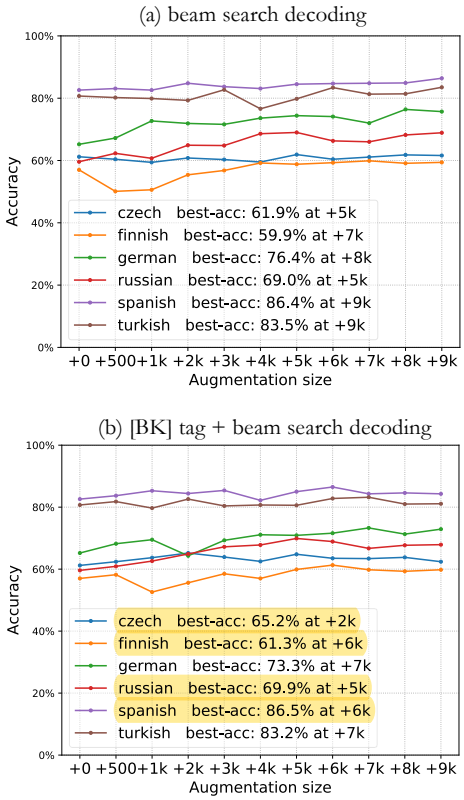

Figure 7: Performance of the Transformer inflection model trained with backtranslated SIGMORPHON shared task data on top of **500** labeled data points. The backtranslated data are added without or without a special tag <BT>.

special tag <BT> at the end of the MSD feature sequence before merging the newly automatically labeled data with the original training data. For this experiment, we report the results of one run in Figure 7. The highest accuracy for each language is presented in the legend of the each plot with the corresponding backtranslation data augmentation size. We highlight the languages which get improved accuracy with tagged backtranslation in Figure 7(b). We see that only two languages (Czech and Finnish) are significantly better with the tagged backtranslation; one language (German) is significantly worse with tagged backtranslation, and there is not a significant difference between tagging or not tagging the backtranslated data for the other three languages (Russian, Spanish and Turkish).[2]

In summary, tagged backtranslation produces similar results to backtranslation without a special tag in our experiments, and thus we would not expect any difference if the words to be analyzed are UD words with tagged backtranslation.

Result analysis In order to understand the performance difference, we examined the delabeled shared task data and UD words. We find that the following two reasons which may contribute the the differences: (1) The delabeled shared task data cover inflected forms where the lemma form is included in the development or test set, while the UD words do not. In other words, some of the delabeled shared task words are for the same lexemes as some words in the development or test set. This reveals a problem in the shared task design, as discussed in Liu and Hulden (2021). (2) There are discrepancies in the UD words and the delabeled shared task data. For example, each of the UD words we used consists of one token, while the delabeled shared task data contains words consisting of multiple tokens. However, multi-token words are common in the shared task development and test sets.

4 Discussion and Conclusion

Though backtranslation has become a common technique in machine translation for data augmentation, our experiments indicate that it is not significantly helpful—at least not by itself—for morphological inflection generation, a character-level string transduction task closely related to MT.

We find small improvements when the backtranslated data is drawn from exactly the same source as the evaluation data, i.e. the SIGMORPHON shared task data. When other sources are used, such as UD or Wikipedia text, backtranslation degrades performance across all data sizes. Though we have controlled the part-of-speech of UD words to match the original training data distribution, adding backtranslated UD words is still unhelpful. Considering that UD data set is labeled with different annotation standards and may also contain some noise, this indicates that unlabeled words used for backtranslation need to be noise-free and have been filtered with the same annotation standards as the labeled data in order to be helpful. Such a strict requirement of data correctness probably renders it unpractical to apply backtranslation to morphological inflection generation in most scenarios.

Further, we did not find any significant difference between the techniques of standard backtranslation and tagged backtranslation in our experiments for morphological inflection.

[2]We used a paired t test to measure whether the difference is statistically significant ($p < 0.05$).

References

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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 984–996, Hong Kong, China. Association for Computational Linguistics.

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39, Vancouver. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Peng-Jen Chen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020. Facebook AI's WMT20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125, Online. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70, Berlin, Germany. Association for Computational Linguistics.

Ling Liu. 2021. Computational morphology with neural network approaches. *arXiv preprint arXiv:2105.09404*.

Ling Liu and Mans Hulden. 2020a. Analogy models for neural word inflection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2861–2878, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ling Liu and Mans Hulden. 2020b. Leveraging principal parts for morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 153–161, Online. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2021. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. *arXiv preprint arXiv:2104.06483*.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 44–50.

Ling Liu, Ilamvazhuthy Subbiah, Adam Wiemerslage, Jonathan Lilley, and Sarah Moeller. 2018. Morphological reinflection in context: CU boulder's submission to CoNLL–SIGMORPHON 2018 shared task. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 86–92, Brussels. Association for Computational Linguistics.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, and Mans Hulden. 2021. To POS tag or not to POS tag: The impact of POS tags on morphological learning in low-resource settings. 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 966–978, Online. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99, Vancouver. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

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, Miikka 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.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv preprint arXiv:2005.10213*.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, et al. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Data details

Language	treebanks
Czech	CAC, CLTT, FicTree, PDT
Finnish	FTB, TDT
German	GSD, HDT
Russian	GSD, SynTagRus, Taiga
Spanish	AnCora, GSD
Turkish	IMST

Table 2: The UD (version 2.6) treebanks sources we use for each language's backtranslation data. We obtain the words from the training set of the treebanks.

B Pilot study results

Augmentation size	Czech	Finnish	German	Russian	Spanish	Turkish
0	66.0	59.4	68.6	63.6	84.4	81.5
500	62.3	50.5	66.9	61.8	75.2	72.4
1k	58.5	45.8	68.9	61.9	71.9	67.5
2k	57.2	41.4	66.2	58.8	68.5	62.5
3k	53.9	41.2	65.9	59.8	66.8	58.9
4k	53.9	35.8	66.2	60.3	66.8	59.3
5k	51.1	36.7	63.0	60.6	63.2	58.3
6k	51.2	36.5	63.0	59.6	61.5	55.3
7k	52.6	37.2	63.9	60.7	60.5	60.4
8k	50.1	33.7	63.8	61.4	60.3	54.0
9k	50.5	33.5	63.0	60.7	57.3	54.7

Table 3: Inflection accuracy (in %) for each language with 500 original training triples after adding different amount of backtranslated Wikipedia data.