Impact of Sequence Length and Copying on Clause-Level Inflection

Badr Jaidi* **Utkarsh Saboo*** **Xihan Wu*** **Garrett Nicolai** **Miikka Silfverberg**
Department of Linguistics
University of British Columbia
badrjd@student.ubc.ca utkarshsaboo45@gmail.com wuxihan@student.ubc.ca
garrett.nicolai@ubc.ca miikka.silfverberg@ubc.ca

Abstract

We present the University of British Columbia's submission to the MRL shared task on multilingual clause-level morphology. Our submission extends word-level inflectional models to the clause-level in two ways: first, by evaluating the role that BPE has on the learning of inflectional morphology, and second, by evaluating the importance of a copy bias obtained through data hallucination. Experiments demonstrate a strong preference for language-tuned BPE and a copy bias over a vanilla transformer. The methods are complementary for inflection and analysis tasks – combined models see error reductions of 38% for inflection and 15.6% for analysis; However, this synergy does not hold for reinflection, which performs best under a BPE-only setting. A deeper analysis of the errors generated by our models illustrates that the copy bias may be too strong - the combined model produces predictions more similar to the copy-influenced system, despite the success of the BPE-model.

1 Introduction

Morphology is often described as the "study of the shape of words", but such a description is not entirely accurate. Without considering the somewhat nebulous definition of a "word", there are clearly inflectional processes that operate on a periphrastic level. For example, in English, the future tense is regularly inflected through the use of an auxiliary: `will` and an infinitive, such as in the case "I will go".

Previous tasks in inflectional morphology (Cotterell et al., 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022) have largely been restricted to generating isolated inflected word forms, which could be viewed as a rather artificial task. While some have included periphrastic constructions (Cotterell et al., 2016), they have largely been constrained to a single part-of-speech. [1] This MRL Shared Task in Multilingual Clause-Level Morphology (Goldman et al., 2022) represents the first attempt to extend inflection generation beyond a single semantic unit to clause-level structures and presents a great opportunity to investigate common inflectional methods in a more realistic morphosyntactic setting.

We augment traditional transformer-based character models with two simple data modifications: we first apply a small BPE-vocabulary to learn common repeated sequences like function words and affixes, hoping to increase performance by reducing the known bias of long character sequences (Neishi and Yoshinaga, 2019). Secondly, we adopt a common data augmentation technique from word-level inflection: adding data that has an identical source and target to bias the model towards the copying of characters (Liu and Hulden, 2022). We find that a combination of these simple techniques improves upon a vanilla transformer for inflection and analysis, while a BPE-only model has the best results for reinflection.

We also contribute a significant error analysis. We investigate the types of errors that inflectional systems are prone to, and how our contributions alleviate them at the clause level; Furthermore, we provide a thorough ablation study that compares errors across inflectional tasks, and how these errors are influenced by sequence length and copy biasing[2].

2 Methods

Studies in neural machine translation have regularly shown character- and subword-level representations outperform word-level ones for morphologically-rich languages (Shapiro and Duh,

*The first three authors contributed equally.

[1] Excepting, of course, those languages where even this distinction is not perfectly clear.

[2] Our data hallucination code is available at https://github.com/mpsilfve/UBCMRL

Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), pages 106 - 114
December 8, 2022

2018), and that optimizing the number of BPE operations can lead to substantial gains in model quality (Araabi and Monz, 2020). Although the sequences in inflectional models are typically shorter, there is evidence that inflection models, like machine translation models, can benefit from grouping common sequences (Peters and Martins, 2022). Similarly, inflectional research has demonstrated that models can be significantly improved by establishing a heavy bias towards copying data directly from input to output (Liu and Hulden, 2022). Many variations of this theme exist, but some of the most successful have included establishing a hard attentional model (Aharoni and Goldberg, 2017), learning an explicit copy bias (Makarov and Clematide, 2018), and augmenting the model with hallucinated data (Anastasopoulos and Neubig, 2019).

For our submission to the shared task, we investigate to what extent these methods are extensible to clausal morphology. Previous work has largely occurred at the word-level, and while it is intuitive that word-level inflection should extend to the clause-level, it is unclear to what extent. As one of the first investigations into clause-level morphology, we investigate the influence of byte pair encoding and copy bias on the production of accurate morphological structures.

2.1 Vanilla system

We build a baseline system using the Fairseq (Ott et al., 2019) implementation of transformers. To distinguish it from the official task baseline, we refer to it as the *vanilla* system. All characters in the input and output are represented as atomic units, and Morphosyntactic descriptors (MSD) are split along semi-colons into inflectional features. Spaces between words in clauses are represented by an underscore (_). An example is provided in Figure 1.

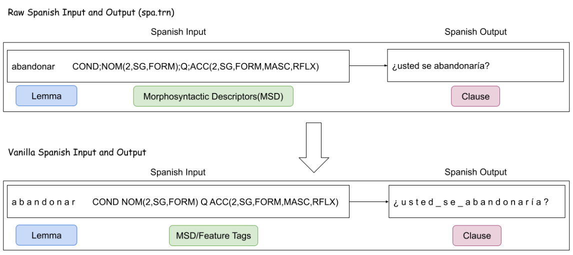


Figure 1: Data representation in the vanilla transformer. The example is from the Spanish data set.

2.2 BPE

Neural models still struggle with long input and output sequences; although great strides have been made in retaining long-distance information, there is still evidence that shorter sequences are easier to represent accurately.

Byte pair encoding (BPE) (Sennrich et al., 2016) reduces the length of both input and output sequences by memorizing frequent symbol sequences and treating them as individual symbols. This typically has a marked positive impact on model performance. In lower-resource settings, however, models can easily overfit if the vocabulary is too large.

We apply BPE to inflection but, in order to avoid over-fitting, we experiment with a very small number of BPE vocabulary merges - 10 to 200. For clause level morphology, we anticipate that these merges will capture only the most common of segments, such as inflectional affixes, pronouns, and function morphemes.

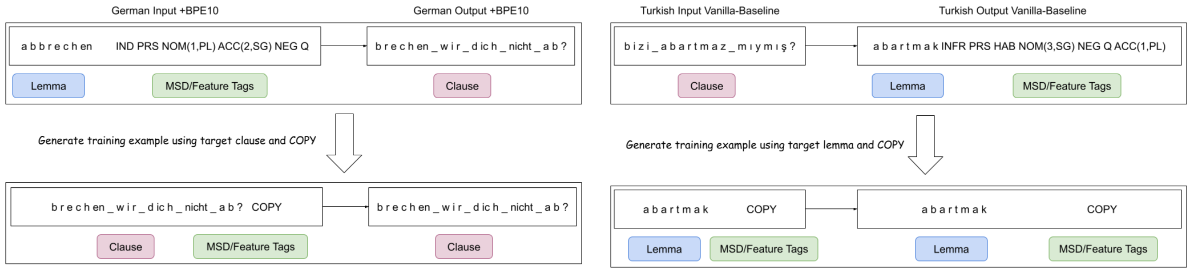
2.3 Copying

When inflecting from a lemma to a surface form, many of the characters in the lemma are often preserved.[3] However, neural models often require a not-insignificant amount of training data to learn this phenomenon. In low-resource inflectional experiments, one process that has repeatedly been shown to improve model stability is the simple expedient of copying the source to the target, without any further modification (Liu and Hulden, 2022). While this copying bias is likely less prevalent at the phrasal level, we believe it still has the opportunity to improve the quality of the inflectional models. Along with strengthening a preference for copying in the model, copying the target data also strengthens the target-side language model. An example of this augmentation for inflection and analysis is shown in Figure 2. For all three tasks, the hallucinated data contains a single `COPY` tag on the source side as MSD.

3 Data / Experiments

The shared task consists of three sub-tasks: (1) inflection, where a lemma and MSD input are converted to an inflected output; (2) reinflection, where an initial clause, input-MSD, and target-MSD are used to generate a target clause, and (3) analysis,

[3]The percentage of characters preserved varies greatly by language.



(a) Copying target-to-target for inflection.

(b) Copying target-to-target for morphological analysis.

Figure 2: Data augmentation via COPY

where an input clause generates a target lemma and MSD. Each task is evaluated across 9 languages.

Each language has a train/dev/test split of 10,000, 1000, and 1000 instances, respectively. [4] Although extra data was allowed for the task, we instead concentrated on optimizing the models without additional data. Each model is evaluated on a single training run; the seed is stabilized to lessen the effect of noise across experiments.

We focus our experiments on an ablation of our proposed data augmentation techniques. The *Vanilla* experiments train models using the vanilla transformer described above. *+BPE* tunes a byte pair encoding vocabulary on each respective language and task; we investigate BPE merges from 10-200, and choose the model that maximizes the results on the development set*+Copy* augments the data with copied target-side data; the size of the BPE vocabulary is tuned individually for each language and amount of additional data.

The transformer was trained with 4 attentional heads over 4 encoder and decoder layers. The Adam$\{0.9, 0.98\}$ optimizer was used, with an initial learning rate of 0.0001, and an inverse square-root learning schedule and a label-smoothed cross entropy criterion. Dropout and attentional dropout of 0.3 were applied to limit over-fitting, and a batch size of 400 was also used. Models were trained for 20,000 updates, with the best model chosen via loss on the development set.

4 Results

We break the discussion of our results down based on the three sub-tasks of the competition: inflection, reinflection, and analysis. All reported results and analysis are on the development set, and are cumulative: "+BPE" applies BPE to the vanilla transformer, and "+Copy" further supplements the model with data hallucination. For official results on the test set, please see the task description paper (Goldman et al., 2022). All systems submitted to the official task were the systems with both BPE and data hallucination. The results report the exact match accuracy of the systems.

4.1 Inflection

Language	Vanilla	+BPE	+Copy
deu	69.0	72.1	75.6
eng	85.4	86.2	89.7
fra	71.6	85.7	89.4
heb	86.9	86.9	86.4
heb_unvoc	63.5	80.6	83.1
rus	80.0	83.4	87.5
spa	87.0	88.6	87.7
swa	82.2	87.0	90.1
tur	81.9	87.0	91.5
Ave.	78.6	84.2	**86.8**

Table 1: Development results for the inflection task (measured in full-form accuracy)

We first report the results for the inflection subtask in Table 1. We observe that both BPE and data hallucination contribute to the quality of the model; on average, adding a small amount of BPE-joined vocabulary reduces the error by more than a quarter. Additionally, providing additional copied data leads to a further 11% error reduction.The BPE vocabulary has the largest impact on French, Swahili, Turkish, and unvocalized Hebrew, while providing smaller gains to the rest of the language set. The only language not to benefit from extra data in training was Spanish. Since Spanish shares a similar morphological makeup to French, which benefits substantially from data hallucination, we

[4] Some languages do not have 10,000 instances exactly, but are of the same magnitude.

do not attribute this finding to the morphological structure of Spanish, but rather to peculiarities of the dataset itself.

4.2 Reinflection

Language	Vanilla	+BPE	+Copy
deu	37.7	49.8	46.6
eng	59.4	73.0	71.4
fra	63.9	68.8	71.1
heb	72.5	80.4	78.6
heb_unvoc	60.6	67.7	63.8
rus	76.8	79.5	78.7
spa	56.1	61.0	72.8
swa	54.9	73.4	65.5
tur	54.6	65.6	63.1
Ave.	59.6	**68.8**	68.0

Table 2: Development results for the reinflection task (measured in full-form accuracy)

In Table 2, it is immediately obvious that reinflection behaves very differently from inflection, despite many conceptual similarities. Although BPE reduces the error of the vanilla transformer to a similar degree as for inflection, adding hallucinated copy data on top of the BPE does not lead to further gains. Again, there seems to be no morphological bias to this trend, with fusional, agglutinative, and templatic languages all behaving similarly.

There is one significant difference between inflection and reinflection that may lead to less success via copy-biasing, however. Although both processes involve the modification of a root, the root is less stable in reinflection. In the inflection task, the input is always the lemma, and identifying the root can largely be generalized over all of the training examples. In reinflection, the input form is inconsistent, and root identification must identify several operations. The problem is exacerbated with larger morphological paradigms, such as clause-level paradigms. While much of the root can be copied, there are also a significant number of substitutions, which may lessen the need for a strong copy bias. For example, in the German data, one example should reinflect *ich würde ihn nicht erschließen* into *es erschlösse sich*. Our copy model instead produces **es erschließe* sich, demonstrating that the copy bias may be too strong.

4.3 Analysis

Table 3 demonstrates the results of our morphological analysis experiments. Conceptually, analysis is the inverse operation of inflection from a lemma (ie, generating a lemma and MSD from an inflected clause), and we observe similar results to those from Section 4.1. Both BPE and data hallucination result in error reductions across most languages, and the effect appears to be cumulative: BPE on its own reduced error by 9.2%, and the extra data leads to a further reduction of 7.1%.

We observe that a significant part of the increase in quality comes from an improved ability to identify the lemma – BPE correctly identifies 2.9% more lemmas than the vanilla system, and the addition of hallucinated data further improves the quality of lemma identification by an absolute 3.1%. This is not surprising, given that once the root has been identified, the generation of the lemma can largely be generalized to a small set of operations, many of which are simple copies.

Somewhat surprisingly, the generation of the MSD also improves from the addition of BPE, despite no modifications to the MSDs in training. We attribute this to the increased quality of lemma generation – in a joint model, the correct identification of part of the output helps with disambiguation of the secondary task. Even with that consideration, it appears that BPE has a larger influence on the production of MSDs than copy biasing.

Language	Vanilla	+BPE	+Copy
deu	83.1	86.1	87.5
eng	89.2	91.0	91.2
fra	93.2	93.2	93.2
heb	92.9	92.9	94.3
heb_unvoc	84.8	87.1	87.7
rus	94.4	94.4	94.1
spa	89.7	89.7	90.6
swa	85.0	87.6	87.9
tur	89.2	89.3	90.8
Average F1	89.1	90.1	**90.8**
Lemma Acc.	67.4	70.3	**73.4**
MSD Acc.	81.7	**82.5**	82.4

Table 3: Development results for the analysis task (measured in F1 Score); the Lemma and MSD Accuracy are averaged over all languages.

5 Analysis / Discussion

In order to better understand the differences in model quality, we perform error analysis along several axes. We first consider the types of inflection errors produced by the BPE, Copy, and BPE+Copy models in Figure 3, while Figure 4 shows error reduction compared to the vanilla transformer when using BPE, Copy, and their combination BPE+Copy.

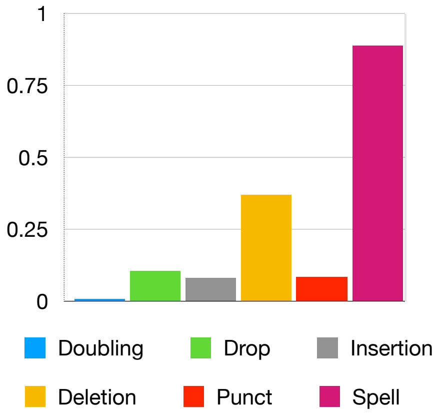


Figure 3: Mean frequencies of various errors across the test languages. Explanation of error types: **Doubling** a character is erroneously doubled ($abc \rightarrow abbc$). **Drop** the second copy of a doubled character is errnouneously dropped ($abbc \rightarrow abc$). **Insertion** a character is erroneously inserted ($abc \rightarrow axbc$). **Deletion** a character is mistakenly deleted ($abc \rightarrow ac$). **Punct** Punctuation is dropped or replaced at the end of a word ($abc. \rightarrow abc$ and $abc. \rightarrow abc?$). **Spell** Total spelling errors affecting the inflected form of the input lemma.

First, we notice that both BPE and Copy individually reduce overall errors (the error type Total in Figure 4). The impact of the methods seems roughly equal, although Copy is slightly more effective on its own. Nevertheless, the combination BPE+Copy clearly outperforms both individual methods. Second, we observe somewhat different influence from the BPE and Copy methods when they are used in isolation - the former significantly improves upon punctuation errors, while the latter removes a number of insertion errors from the vanilla model. Moreover, the most prevalent error type in the vanilla model – deletion – is only moderately reduced by the BPE model, while a far greater error reduction can be seen when Copy is employed. Furthermore, we observe a largely complementary effect - the combined model improves over either individual model for all error categories.

We next run an ablation to investigate the role that each of our contributions has on the quality of the models for each task. The results are plotted in Figure 5. In this graph, we investigate which errors are corrected or introduced by a particular method. BPE and Copy "correct" an error if it was produced by the vanilla model, and "break" an example if it was correctly predicted by the vanilla model, but not the enhanced model. For the model with both BPE and copying, an instance is only considered "corrected" if both the BPE and Copy models produced an incorrect solution. Likewise, it "breaks" a prediction only if both the BPE and Copy models produce the correct solution.

We observe that both BPE and copying lead to large improvements in the model, regardless of the task - far more errors corrected than introduced. For inflection and analysis, both methods appear to contribute roughly equally to the quality of the model. Furthermore, we observe a complementary effect, where the combination of both methods corrects notably more examples than either method on its own. Contrarily, the combined model introduces fewer inflectional errors than either BPE or copying alone.

Interestingly, the trends observed in inflection and analysis do not hold for reinflection. Although BPE and copying alone improve the model, their combination introduces a large number of errors - such that they overwhelm the corrected instances obtained through the combination of methods. A closer inspection reveals that this outlier is largely attributable to a single language - Swahili. When Swahili is excluded, the results trend similar to the other tasks, although BPE still has a stronger influence. There are several areas where Swahili could be contributing to this interesting finding, but lacking experts in the language on our team, we hesitate to make concrete hypotheses.

Figure 5 suggests that the biggest benefit of the combined model is its ability to correctly discern when one of the separate data augmentations correctly produces an inflection, but it isn't quite that clear. Looking at examples where the BPE and Copy models disagree, we observe that the combined inflection model correctly chooses the right solution 72.8% of the time.[5] However, for reinflection and analysis, the correct solution is only chosen slightly more than 50% of the time.

Considering only those instances where the original BPE and Copy models disagree, we investigate the influence of the individual contributions. For inflection, we observe that the combined model produces output identical to the BPE model in 61% of cases, as opposed to only 40% for reinflection and 43% for analysis. It appears that the copied data has an unduly large influence on the combined model for the latter two tasks.

Given that the motivation behind BPE was re-

[5]Note that there is actually no "choosing" occuring, such as might happen in an ensemble. Instead this can be viewed as the influence of a particular addition biasing the model towards a particular prediction.

Figure 4: Mean error reduction across the five test languages for the BPE, Copy and BPE+Copy systems when compared to the baseline system. See caption of Figure 3 for an explanation of the error types.

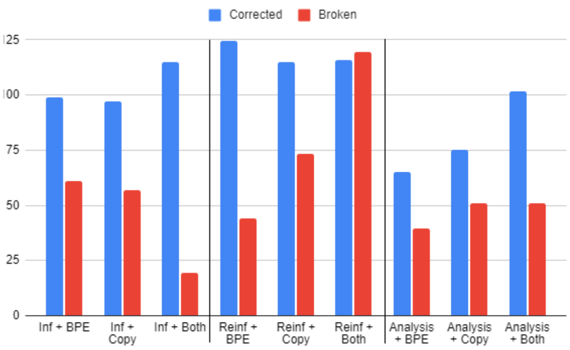


Figure 5: Analysis of errors corrected and introduced by our augmentations over the vanilla model. The y-axis is an absolute scale of the average number of errors corrected and introduced by each model, compared with the vanilla transformer. Inf - Inflection; Reinf - Reinflection.

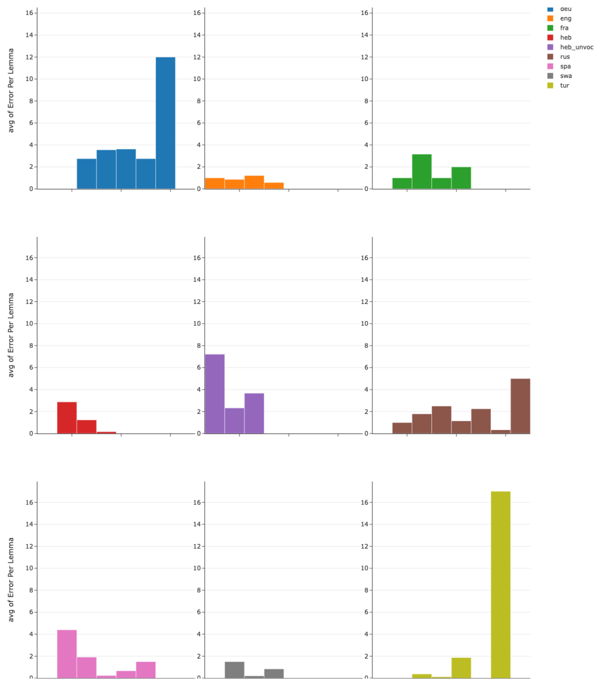


Figure 6: Analysis of errors made by our best model for each language in the inflection task depending on character length of the lemmas. The y-axis is the average number of errors made in the development set and the x-axis is the character length.

ducing the size of input and output segments, we investigate the role that the length of a sequence plays on the quality of a model. Since reinflection and analysis lengths vary based not only on the length of the verb being inflected, but on other factors such as the number of words in the input, etc, we limit this investigation to the inflection task. Figure 6 demonstrates the number of errors produced by our best system, given the length of the input sequence (ie, the lemma). German, Russian, and Turkish show a strong preference for shorter input sequences. Hebrew (both unvocalized and standard) and Spanish instead demonstrate a somewhat surprising preference in the other direction - producing more errors for short sequences.

In an attempt to further explain these conflicting results, we next investigate the relationship between lemmas in the training and development sets. Figure 7 reports the number of errors made by our best system with respect to the distance between the development lemma and the closest analogue in the training data. Now, unsurprisingly, we see that most languages perform better when there is a closely-related lemma in the training data.

Finally, we investigate the efficiency of our copying method by comparing it with two alternatives. Rather than simply taking the training output data and copying it as extra data, `RANDOM` generates random sequences of characters to copy from source to target. Similarly, `LM` creates new copy sequences, but first learns a neural language model from the training data, before generating the sequences. The results for inflection are shown in Figure 8.

We observe that changing the hallucination method from copied training data to randomly-generated sequences greatly improves the quality

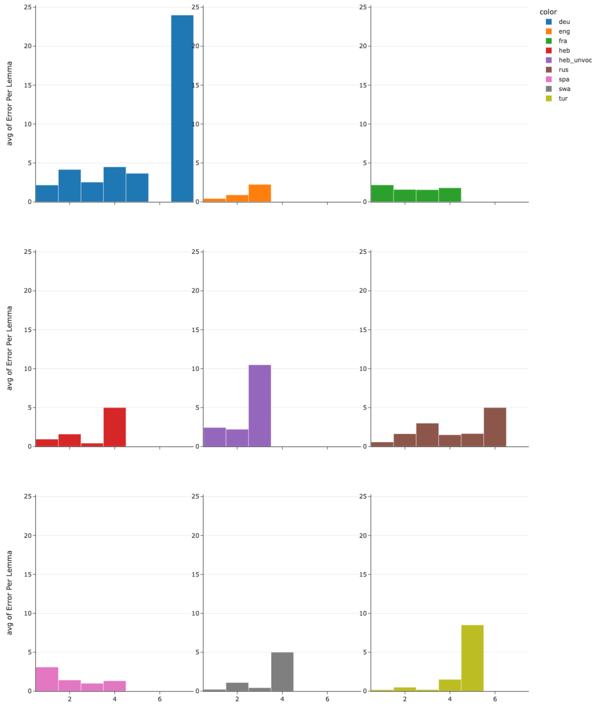


Figure 7: Analysis of errors made by our best model for each language in the inflection task depending on the closest Levenshtein score of any lemma present in the training set. The y-axis is the average number of errors made in the development set and the x-axis is the Levenshtein distance.

of the inflector, with an error reduction of more than 35%, on average. We hypothesize that while the COPY method simply reinforces an existing signal, the RANDOM method introduces new contexts for copying, which allows the model to better generalize the copy operation overall.

6 Conclusion

We have described the submission of the UBC team to the MRL shared task on multilingual clause-level morphology. Experiments on a series of morphologically-diverse languages have demonstrated that BPE and copy-biasing, two methods that have proven successful at the word-level, are largely extensible to clause-level morphology.

We observe that the methods are largely complementary, with one exception - the task of reinflection. Although we observe notable gains over a vanilla transformer when either performing BPE or copy hallucination, combining the two methods leads to a degradation in reinflection quality.

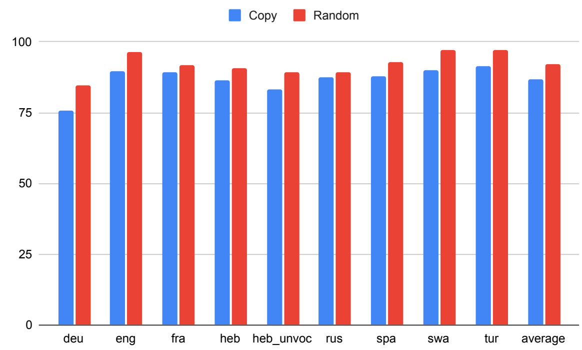


Figure 8: A comparsison of our data hallucination methods using copied training data and randomly generated sequences.

Limitations

The work described in this paper focuses on multilingual representation, but the authors are not familiar with all of the analyzed languages. Hypotheses are based on general linguistic experience, and not necessarily a familiarity with the languages in question.

Deep learning models are stochastic in nature, which may lead to replication difficulties. We have tried to specify relevant hyper-parameters and settings, but random fluctuations in seed values may result in variations in replication studies.

Ethics Statement

We trust that the data used in this paper was ethically-sourced. The models were trained by faculty and students of the Department of Linguistics at the University of British Columbia, and none of the data or models were shared with anyone outside that purview. All contributors to the project are in the author list, or thanked in the acknowledgments. No members of the team received monetary compensation for participating in this task. All participation was voluntary.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

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.

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on 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.

Omer Goldman, Francesco Tinner, Hila Gonen, Benjamin Muller, Victoria Besamova, Shadrack Karimi, Lydia Nishimwe, Benoît Sagot, Djamé Seddah, Reut Tsarfaty, and Duygu Ataman. 2022. The MRL 2022 shared task on multilingual clause-level morphology. In *Proceedings of the 2nd Workshop on Multilingual Representation Learning*, Abu Dhabi. Association for Computational Linguistics.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, 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 Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? Tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93, Santa Fe, New Mexico, USA. 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.

Masato Neishi and Naoki Yoshinaga. 2019. On the relation between position information and sentence length in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 328–338.

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.

Ben Peters and Andre F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–138, Seattle, Washington. Association for Computational Linguistics.

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 Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, 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.

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.

Pamela Shapiro and Kevin Duh. 2018. Bpe and charcnns for translation of morphology: A cross-lingual comparison and analysis. *arXiv preprint arXiv:1809.01301*.

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.