The UniMelb Submission to the SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection

Andrei Shcherbakov
University of Melbourne, AU
ultrasparc@yandex.ru

Abstract

The paper describes the University of Melbourne's submission to the SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection. Our team submitted three systems in total, two neural and one non-neural. Our analysis of systems' performance shows positive effects of newly introduced data hallucination technique that we employed in one of neural systems, especially in low-resource scenarios. A non-neural system based on observed inflection patterns shows optimistic results even in its simple implementation (>75% accuracy for 50% of languages). With possible improvement within the same modeling principle, accuracy might grow to values above 90%.

1 Introduction

According to WALS database 80% of the world's languages morphologically mark verb tense and 65% mark grammatical case (Dryer et al., 2005). Still, until recently most research in natural language processing was focused on a few well-documented languages with modest amount of morphological marking. A great variety of typologically diverse low-resource languages were left outside of NLP investigation and modeling. At the same time, neural systems outperformed non-neural ones on many benchmarks(cite) while being evaluated on a limited (and often not typologically representative) sample of languages. Nevertheless, some of such systems or architectures were stated as "universal". But are they universal? How well models trained a certain sample of language families can generalize outside of it? For instance, a model trained on Indo-European languages might be biased towards suffixing and will be working less well on languages that use infixing or prefixing. The SIGMORPHON 2020 Shared task 0 ("Typologically diverse morphological inflection"; Vylomova et al. (2020)) aims at evaluation of the generalization ability of models. It continues recent trend of increasing linguistic diversity: starting with 10 well-documented languages in Cotterell et al. (2016) up to 103 in Cotterell et al. (2018). These shared tasks demonstrated that neural models outperform non-neural ones but generally struggle in low-resource settings. Therefore, the 2019 Shared Task focused on cross-lingual transfer (McCarthy et al., 2019) and explored transfer of morphological information from a high-resource to a low-resource language. In this paper, we describe three models submitted to the shared task 0. We investigate both generalization ability of models and their performance in low-resource languages. We propose a variation of data hallucination technique that significantly improves the results of neural models in low-resource settings.

2 Task Description

The task was organized in three stages: development, generalization and evaluation. In the development stage participants were provided with initial set of 45 development languages that were used to *develop* their systems. In the next stage, generalization, an extra and more diverse set of 45 languages was released, and participants were asked to fine-tune and optimize their systems on these languages. In both stages, only training and development datasets were released. Test splits for both development and generalization languages were provided in the final, evaluation, stage.

Systems were then evaluated and ranked based on the test set predictions.

3 Data

3.1 Data Format

All shared task data are in UTF-8 and follow UniMorph annotation schema (Sylak-Glassman,

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 177–183
Online, July 10, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

2016). Training and developments samples consist of a lemma, an inflected (target) form, and its morphosyntactic description (tags). Test samples omit the target form.

3.2 Languages

Forty-five languages representing Austronesian, Niger-Congo, Oto-Manguean, Uralic and Indo-European language families were provided in the development stage. Another forty-five (surprise) languages from Afro-Asiatic, Algic, Altaic[1], Dravidian, Indo-European, Niger-Congo, Sino-Tibetan, Siouan, Songhay, Southern Daly, Uralic, and Uto-Aztecan families were provided in the generalization phase one week before the evaluation phase started. Importantly, the dataset sizes are highly imbalanced, ranging from tens of thousand of samples in some Uralic languages to a few hundreds in the Niger-Congo family.

4 Baseline Systems

Two types of baseline systems were provided: neural and non-neural. The **non-neural** baseline was essentially the same as in previous years' tasks (Cotterell et al., 2017, 2018). More specifically, it first extracts possible lemma–form alignments and associates them with corresponding target tags, then majority classifier chooses the most frequent transformation and applies it to a given lemma.

The **neural** baselines include a hard monotonic attention model (Wu and Cotterell, 2019) and a character level transformer (Wu et al., 2020). Both were trained in monolingual and multilingual modes. Organizers also provide a variation of the model that uses data hallucination technique from Anastasopoulos and Neubig (2019) to improve performance in low-resource languages.

5 Evaluation

The systems were evaluated in terms of test accuracy and Levenstein distance between predicted and gold forms. Unlike in earlier shared tasks where systems were ranked based on macro-averaging, here systems were ranked based on statistical significance of differences in their performance.

6 System Description

In terms of the shared task, we experimented with three systems, two neural and one non-neural. Subsections below provide a short description of each.

6.1 A non-neural system based on differently refined alignment patterns

First, we implemented a non-neural system (**`flexica01`**) where possible patterns of lemma-to-inflected form transformation are generated directly by the following simple process:

1) We find all maximal continuous matches between lemma and inflected form; while doing this, we start with the longest possible match and then find matches across the remaining unmatched fragments, recursively. We replace the matches found with groups denoted as `\number`, like in regular expression syntax. Swapped order of groups in inflected forms is allowed. For the simplicity of implementation, we assumed that the number of group is increasing along the lemma word. If multiple matches of the same group lengths are possible for a given lemma - inflection pair, we produce all the respective transformations. However, for the vast majority of samples only a single variant is produced at this stage.
For example, for the past tense of "`to understand`":
`understand → understood`
we extract the following transformation rule: `\0an\1 → \0oo\1`,
where `\0=underst` and `\1=d` are groups. Group substitutions are not stored leaving a transformation as abstract as possible. However, some statistics about group content is used to evaluate the confidence of substitution (see below).

2) Starting with previously generated transformation pattern(s) of maximal abstraction, we generate a set of patterns more specific to a given training word by treating a limited number (0..$ConcreteLetterLimit$, where $ConcreteLetterLimit$ is a hyperparameter) of characters as concrete (i.e. standing outside any group). For our previous example given $ConcreteLetterLimit = 1$ we would finally produce the following set of matching transformations: `\0an\1 → \0oo\1; u\0an\1 → u\0oo\1; \0n\1an\2 → \0n\1oo\2, ... (3 more), \0s\1an\2 → \0s\1oo\2, \0tan\1 → \0too\1, \0and → \0ood`.

All patterns generated for training samples are stored in a trie, which is separate for each combination of grammatical features. The resulting set

[1] Tungusic and Turkic

of tries acts as a model.[2] At prediction phase, a multi-variant search against a given lemma is attempted over the trie for a respective grammatical tag combination. Here, multi-variance means that the search procedure both allows wildcards for possible groups and concrete characters to be matched against. After the search completes, all the candidate transformations found are then sorted by their associated score in order to find the best fit. In the version used to produce prediction submitted to the contest, the score was based on the following three components:

1. A (squashed) frequency f of transformation occurrence in a training set;

2. The diversity d of marginal (the first one and the last one) letters in groups as they occurred in different fits of a given transformation found in the training set. To grasp the underlying idea, take, for example, a `\0` → `\0s` transformation producing plural nouns in English that is considered as highly confident for any possible `\0` value because `\0` was observed to match various strings starting and ending with many different letters in a training set. In contrast, `\0a\1` → `\0oo\1` matches a very limited set of examples such as `stand`→`stood`, `understand`→`understood`, where the last character of `\0` is always `'d'` and the first character of `\1` is `'n'`. Such a poor diversity of characters should signal the predictor that the transformation pattern is not likely to be usable at different group values and it may be better to focus at more specific transformation patterns instead. Technically, we counted d as a product of the number of distinct characters over all start and end positions of groups. Still, if we have an exact match between currently considered substitution letter and one observed at the same position in a training sample, we consider this position exempt from scrutiny by assuming it as having a high "effective" diversity (currently, of 10).

3. Specificity s which here means the number of concrete characters in the pattern (without counting characters falling into groups).

In the submitted version, the score was calculated by the following empirical formula:

$$G = \frac{1}{2}\log_2 f + 6\log_2 d + 12s \qquad (1)$$

Note that in contrast to a conceptually similar approach proposed by Hulden et al. (2014), we didn't encourage the most general paradigms. Instead, we used a trade-off criterion that prefers better confidence but lower amount of abstraction in patterns. Also, we didn't attempt to build whole paradigms. We used an independent alignment process for each form.

Fig. 1 displays accuracy for the model measured across all 90 languages. We additionally show the accuracy that would be achieved in a case of ideal selection criteria (labelled as "+ Ideal Transform Choice" category) for every language. The accuracy equals to the proportion of test samples which succeeded in matching at least one transformation pattern that produces correct prediction. We also note that the proposed scoring formula (mostly inspired by Indo-European languages) does not fit well the Oto-Manguean family. If to speak about the potential ability to cover inflections by directly observable patterns, Finnic languages with their tricky morphology appear to be the most challenging ones.

We also roughly measured potential improvement that may arise from considering correlations between inflection patterns for different grammatical forms of a single lemma (in other words, from paradigm clustering). We trained embeddings for the generated transformations using lemmas as context markers. Then, we used cosine similarity between such embeddings as a candidate transformation selection criterion in cases when a lemma is both present in the train and in the test sets. The proportion of samples where application of such a criterion allowed to turn an incorrect prediction into a correct one, is labelled as "+ Paradigm Search" in Fig. 1.

Generally, the experiments with pattern-based inflection prediction were proposed to verify the following two hypotheses, (1) that it is sufficient to reuse observed substitution patterns for proper modeling of inflection in a wide range of languages, and (2) that candidate inflection pattern selection may be based on a simple statistical

[2] To simplify the implementation, a transformation pattern was stored as a mapping between two plain strings, one for the lemma and another one for the inflected form. Group references were represented by special characters added to the alphabet.

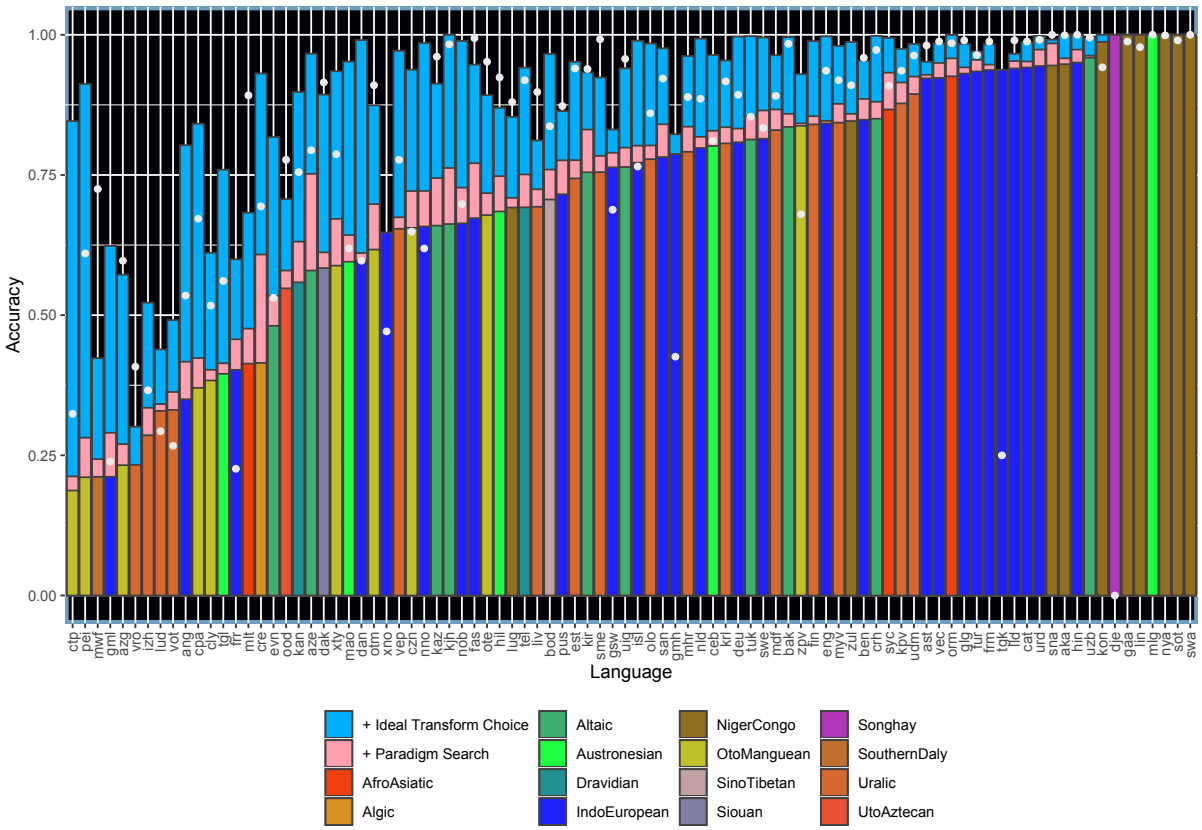

Figure 1: Accuracy for the non-neural `flexica01` solution based on immediately observed transform patterns. Accuracy for the `flexica02` hard attention neural system is also given for comparison (in white points).

criterion (frequency, entropy etc.) While a simple pattern selection rule hasn't yet been discovered, the experimental results largely support the first hypothesis. However, it should be noted that learnt patterns are often too sparse due to the lack of compositionality and abstraction in the initial system design. When an inflection involves complex, phonotactical transformations, it is unlikely to match a quite "similar" sample in a train set. It is especially true if the inflection is irregular which usually implies extreme sparsity of its domain. Another issue that limits pattern search capacity is related to the model size. The experiments have shown that greater values of $ConcreteLetterLimit$ enable greater accuracy figures. However, we had to stick with $ConcreteLetterLimit = 2$ because the choice of greater value led to unacceptably high memory consumption for most of training sets provided. Though, this issue is likely to be addressed by using of ongoing pruning procedures over learnt transformations.

6.2 Neural systems

Multilingual (family-based) learning The neural system (**`flexica02`**; multilingual) is based on hard monotonic attention model proposed in Aharoni and Goldberg (2017), with the same loss function, but with the following differences:

- We combined all the languages belonging to a given family[3] into a single dataset, having added two extra features such as language and genus. The idea was to let the model infer common cross-lingual inflection patterns when a resource for a particular language is low.

- We also made a minor modification of preprocessing. We used maximal continuous sub-string search to organize alignment between lemma and its inflected form in order to advance hard attention state during the learning phase. Compared to the original system,

[3] An exception was Uralic family. Due to excessively high volume of training data, we split this family into 5 subfamilies.

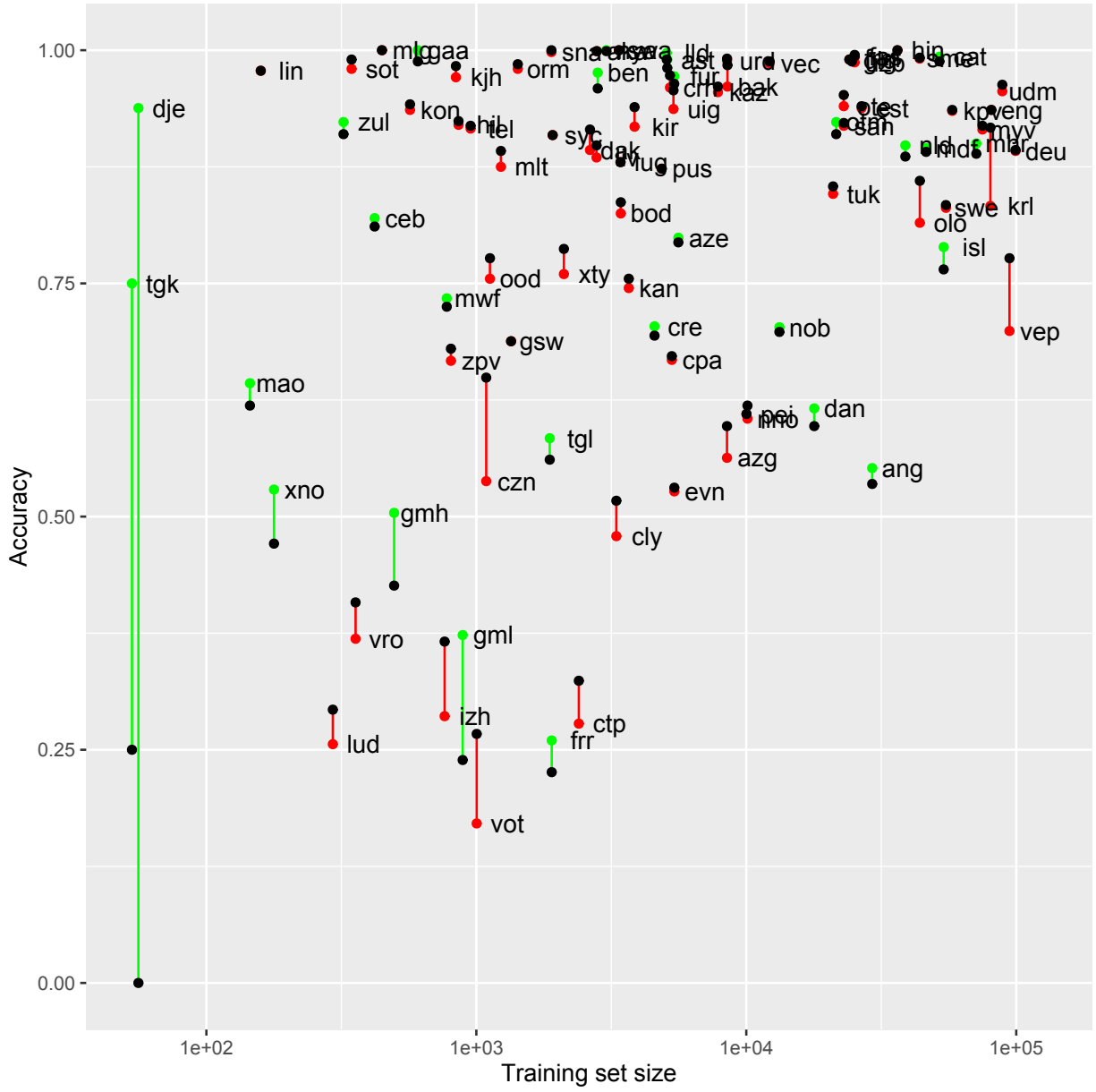

Figure 2: Comparison of accuracy for the proposed neural system with hallucinated data (green or red points for greater or lower accuracy, respectively) and one without hallucinated data (black points)

we abolished one-by-one alignment of mismatching characters, instead letting each mismatching segment to be put into correspondence to a single attention state as a whole.

Hyperparameters are set as follows: hidden and input dimensionality is set to 100, feature dimensionality is 20, the number of layers is 2. The model is trained with AdaDelta (Zeiler, 2012) for 100 and 20 epochs for small-sized and large-size families, respectively.

Adding Hallucinated Data Inspired by Anastasopoulos and Neubig (2019), our last model (**`flexica03`**) is a variation of the above model that uses extra hallucinated samples. We added 200 samples[4] per language per part-of-speech (POS) in order to produce hallucinated inflection samples that look like real. We reused the predictor from **`flexica01`** (presented earlier) with the only difference that now it acts in the reverse direction predicting the best fitting tag–lemma combination for a given inflected form. We also enriched the model with word-generator

[4]We chose this number as an empirical approximation of minimum amount of training data required for the predictor to display stable convergence.

(Shcherbakov et al., 2016) to produce more phonotactically plausible forms. This works in the following way: 1) Word generator trained on inflected forms for a given POS produces samples of hallucinated inflected forms (without distinction of grammatical features); 2) The reverse **`flexica01`** predictor produces tag–lemma for each hallucinated inflected form.

As Fig. 2 shows, supplementing training data with hallucinated samples significantly improved accuracy in low-resource languages (such as Maori, Zarma, Tajik, Anglo-Norman, Middle High/Low German) while for medium to high sized resources we observe less consistency in positive effects.

7 Conclusion

We proposed and tested (1) multilingual training, and (2) pattern-based hallucinated inflections as possible enhancements of sequence-to-sequence morphology modeling for diverse low-resource languages. We also developed a simple non-neural approach based on multi-variant search of common inflection patterns. We explored its suitability for different language families and proposed further improvement options.

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 983–995.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 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.

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, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660, Berlin, Germany. Association for Computational Linguistics.

Matthew Dryer, David Gil, and Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford University Press.

Mans Hulden, Markus Forsberg, and Malin Ahlberg. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578, Gothenburg, Sweden. Association for Computational Linguistics.

Arya D McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian J Mielke, Jeffrey Heinz, et al. 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.

Andrei Shcherbakov, Ekaterina Vylomova, and Nick Thieberger. 2016. Phonotactic modeling of extremely low resource languages. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 84–93.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Ponti, Rowan Hall Maudslay, Ran Zmigrod, Joseph Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrej 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. The SIGMORPHON 2020 Shared Task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1530–1537.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.