The BGU-MeLeL System for the SIGMORPHON 2023 Shared Task on Morphological Inflection

Gal Astrach
Department of Computer Science
Ben Gurion Univeristy
Beer Sheva, Israel
galastra@post.bgu.ac.il

Yuval Pinter
Department of Computer Science
Ben Gurion Univeristy
Beer Sheva, Israel
uvp@cs.bgu.ac.il

Abstract

This paper presents the submission by the MeLeL team to the SIGMORPHON–UniMorph Shared Task on Typologically Diverse and Acquisition-Inspired Morphological Inflection Generation Part 3: Models of Acquisition of Inflectional Noun Morphology in Polish, Estonian, and Finnish. This task requires us to produce the word form given a lemma and a grammatical case, while trying to produce the same error-rate as in children. We approach this task with a reduced-size character-based transformer model, multilingual training and an upsampling method to introduce bias.

1 Background

The SIGMORPHON Shared Task proposed a cross-linguistics modelling of child language acquisition to mediate between the theories of the acquisition of inflectional morphology. Here, unlike previous shared tasks of morphology inflection, the goal is to build a model that shows childlike item-by-item error rates, instead of generating the well-formed inflection.

1.1 Morphological Acquisition

The way that a child or an adult acquires a language is different. Therefore, the way they make mistakes is different. In the past decades there were many studies about the way children acquire a language, but most of the research focus only one language. Granlund et al. (2019) performed a large-scale cross-linguistics study of three languages—Finnish, Estonian and Polish. The research's goal was to find the aspects that indicate what makes children inflect words correctly.

The research found two such aspects: the first is *surface-form frequency*, where the greater the input frequency of the targeted inflectional form (i.e., the exact surface form that the child is attempting to produce in a given context; e.g., Polish książki, 'book-genitive') is, the greater the speed and accuracy of production or recognition. The second is *phonological neighborhood density* (PND), where the greater the number of "neighbours" or "friends"—nouns that are similar in both the base (nominative) form and the relevant target form (e.g., książka → książki; doniczka → doniczki; gruszka → gruszki)—the greater the speed and accuracy of production or recognition.

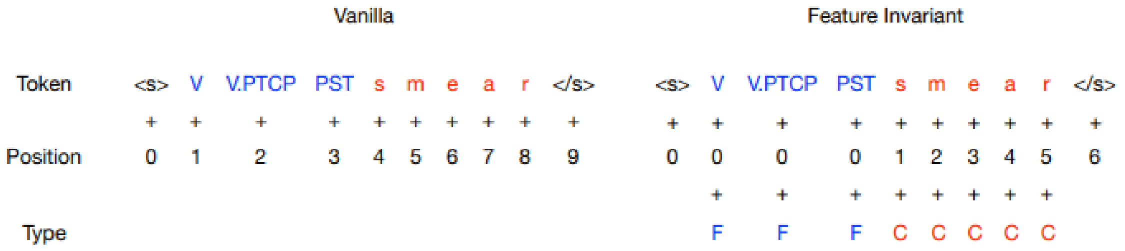
They also describe how these aspects work together: the effect of phonological neighbourhood density is greater for items with low surface-form frequency. Since low-frequency items are less likely to be successfully retrieved from memory, they must be generated by phonological analogy.

1.2 Modeling Acquisition of Inflectional Noun Morphology

The task of morphological inflection (Cotterell et al., 2017; Kodner et al., 2022) is defined as finding an inflected form for a given lemma and list of morphosyntactic attributes. Most state-of-the-art systems for the tasks to date center on character-level transduction and representation, and naturally attempt to predict the correct inflection with maximum performance. The current task, by contrast, requires imperfect generation by design, and thus solicits different approaches than state-of-the-art.

The data format in this task also differs from previous iteration in that it is more faithful to language children are exposed to. Instances are limited to single-feature inflection of lemmas into various grammatical cases (e.g., accusative, nominative, or genitive), and the lemma and the correct inflection are given in both orthographic and phonetic form (using IPA). In addition, the surface-form frequency of the lemma is provided, and the test set also contains children's error-rate of the inflection. The dataset is split such that lemmas in the training set do not appear in the test set (Goldman et al., 2022). The task expects as system output a list of

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 166–170
July 14, 2023 ©2023 Association for Computational Linguistics



Vanilla

Token	<s>	V	V.PTCP	PST	s	m	e	a	r	</s>
	+	+	+	+	+	+	+	+	+	+
Position	0	1	2	3	4	5	6	7	8	9
Type										

Feature Invariant

Token	<s>	V	V.PTCP	PST	s	m	e	a	r	</s>
	+	+	+	+	+	+	+	+	+	+
Position	0	0	0	0	1	2	3	4	5	6
		+	+	+	+	+	+	+	+	
Type		F	F	F	C	C	C	C	C	

Figure 1: Wu et al. feature invariance (taken from the original paper)

top-10 inflections in IPA, alongside their probabilities. As an example, the following is a training data instance for the Polish lemma zdrowie “health”:

```
zdrowie GEN zdrowia zdrɔvjɛ zdrɔvja 6,
```

where the columns represent (in order): the lemma in standard orthography, grammatical case, the inflection in standard orthography, the lemma in IPA, the inflection in IPA and the surface-form frequency.

1.3 Evaluation

In addition to exact-match accuracy and edit distance, correlation-based evaluation was also used for this task. In our development stage, we extracted the top 10 predictions for each instance with their respective probabilities, using beam search. We then calculated the correlation (both Spearman’s and Pearson’s) of the correct inflection’s error rate and the model’s outputs’ probabilities. Due to the data format, this evaluation could only be done on the test set. When the correct form is not in top 10 predictions, we assign it zero probability.

2 Model

The base model that we used is the current state-of-the-art character-based transformer model (Wu et al., 2021). We then modified it to fit the task. The code from the model is forked from the public repository[1] with changes relevant to this task, meaning that the learning rate scheduler, early stopping and various training strategies are the same. Our model accepts the lemma in its IPA form.

The purpose of the original base model was to inflect a lemma form to the correct inflection morphological properties given as input. Our settings differ in that the model should inflect according to the children’s behavior, and not to the correct inflection. We can do that by modifying the model to work with both of the features introduced above, namely **PND** and **surface-form frequency**. We select our model based on the best epoch according to the overall best evaluation (see §1.3) on the test sets.

2.1 Base Model

The transformer (Vaswani et al., 2017) is a sequence-to-sequence model, used for tasks such as machine translation. The transformer-based model we use as a basis for our task (Wu et al., 2021) is tailored for character-level transduction in order to be applied to tasks such as morphological inflection and grapheme-to-phoneme prediction, illustrated in Figure 1 (taken from the original paper). Crucially, the input provided to the model is the concatenation of the characters of the lemma with the morphosyntactic attributes, assigning embeddings to each character and attribute. Their variant, dubbed **feature-invariant transformer**, differs from the original transformer in two aspects: a smaller model and a feature-invariant architecture.

Feature invariance In morphological inflection tasks, the lemma is a sequence of characters mapped to the inflection which is a different sequence of characters, to be predicted according to the list of morphological attributes. The transformer model deals with sequences as they are ordered. However, the portion of the input consisting of a list of morphological attributes is unordered; moreover, the distance between attributes and the characters within the input is irrelevant. These properties may lead to inconsistencies in data representation and generalization when training a sequence model so sensitive to input order. The feature-invariant transformer therefore receives the positional encoding of features as zeroes, and only begins incrementing position count for the lemma’s

[1] https://github.com/shijie-wu/neural-transducer

System	**Accuracy**	**Edit Distance**	**Pearson's**	**Spearman's**
Baseline (Wu et al.)	1.0000	0	−0.029	−0.061
Base + Smaller Model	.8812	0.229	0.078	−0.047
Base + Upsample	.9890	0.015	−0.015	−0.087
Base + Multilingual	.9978	0.002	−0.106	−0.259
Base + Smaller Model + Upsample	.8099	0.359	0.286	0.237
Base + Smaller Model + Multilingual	.6864	0.548	0.379	0.334
Base + Multilingual + Upsample	.9890	0.013	−0.023	−0.318
Base + Small + Upsample + Multiling	.5526	0.814	**0.467**	**0.438**

Table 1: Model variants' results on the test set. Results for models not specified as multilingual are reported are the macro-average for the three languages. Multilinugal models' correlations are calculated on the concatenated test sets of all three languages. The correlations are the metrics of interest. The system in bold was submitted to the shared task.

characters. Additionally, a special token is used to indicate whether a symbol is a word character or a morphosyntactic attribute.

2.2 Surface Form Frequency

According to Granlund et al. (2019), one of the attributes that correlate with accuracy in children is the frequency of the form in the heard corpus they are exposed to. Therefore, we chose to incorporate this information in our model, by a combination of methods, namely **upsampling** and **surface form frequency embeddings**.

Upsampling We manipulate the training dataset synthetically by upsampling each form in direct proportion to the form-frequency as annotated in the dataset. The way we upsample is that when reading the raw dataset, we add the same sample according to the value in the surface-form-frequency column, meaning that if a sample (a lemma, morphological feature and an inflection) has the value n in the surface-form-frequency column, then it will appear n times in the training set.

Surface-form frequency embedding Since in the test set we cannot upsample, we need to also utilize the form-frequency value by itself. We do that by feeding the value of the surface-form frequency into a linear layer, with the layer's output size the same as the other inputs' embedding dimension, and then concatenating it to the embedding's layer's output. The linear layer has no activation function, in order to act like the embedding layer in the transformer. After concatenation, we apply dropout to the new embedding tensor.

2.3 Multilingual

In order to generalize the modeling of language acquisition, we trained the model multilingually. We did that by adding a tag to the morphosyntactic attributes, together with the grammatical case, which indicates the language. The language tag therefore acts like the rest of the morphosyntactic attributes and provided as input to the embedding layer.

2.4 An Even Smaller Model

As mentioned above, the transformer introduced in Wu et al. (2021) is a smaller transformer than the original. Early experiments led us to suspect that further reducing the model size could better approximate children's performance. We use 4 encoder-decoder layers, 2 self-attention heads, a feed-forward layer with hidden size $d_{\text{FF}} = 128$, embedding size $d_{\text{model}} = 256$, dropout rate 0.5, and a batch size of 100.

3 Results and Discussion

We present the results for our models in Table 1. They show that our methods provide substantial improvement over the baseline, which generates perfect inflections, but correlates poorly with the children's error rates. The best improvement in correlation given by a single method was from decreasing model size; the best overall performance was obtained by using all three methods, indicating that their improvement profiles are complementary. We note that multilingual training was mostly beneficial to model performance, suggesting that the language acquisition process is generalizable across languages.

As noted in the background section, there are two aspects relevant to this task of modeling acquisition which are different than normal, well-formed inflection, namely surface-form frequency and phonological neighborhood density (PND). The model we designed captures the former by the upsampling method and frequency embeddings, whereas PND could theoretically be imbued through the transformer's encoder, which embeds the lemma into a hidden state vector given its IPA representation. As such, it is capable of modeling similarity on the phonetic level, so if two words are pronounced similarly, their hidden states can be similar and thus provide means for PND realization.

4 Conclusion

This paper presents the approach taken by the MeLeL team to solving the SIGMORPHON 2023 Shared Task on Typologically Diverse and Acquisition-Inspired Morphological Inflection Generation. To this end, we designed a model for morphological inflection, based on current state of the art. We adapted the model to the task objectives, modifying hyper-parameters to add "forgetfulness", incorporated surface-form frequency information by adding upsampling and embedding the frequency counts, and trained multilingually to generalize cross-lingual features. Our final system, which correlates with child-produced inflection substantially better than the base system, is informed by two aspects previously shown to be relevant to children's inflectional competence, namely surface-form frequency and neighborhood phonetic distance.

References

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.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational Linguistics.

Sonia Granlund, Joanna Kolak, Virve Vihman, Felix Engelmann, Elena V.M. Lieven, Julian M. Pine, Anna L. Theakston, and Ben Ambridge. 2019. Language-general and language-specific phenomena in the acquisition of inflectional noun morphology: A cross-linguistic elicited-production study of polish, finnish and estonian. *Journal of Memory and Language*, 107:169–194.

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.

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

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

Language	In top-10	Acc.	Pear.	Pear-0	Cosine	Cosine-0
Polish	134/150	.73	−0.020	0.231	0.99	0.94
Estonian	121/144	.55	0.547	0.578	0.99	0.94
Finnish	134/162	.44	0.462	0.462	0.98	0.92

Table 2: Submitted model results for each language. "In top-10" means the number of predictions from the test set that were found in the model's top-10 list. "Pear" and "Cosine" are the Pearson's correlation and Cosine Similarity for the predicted probabilities, where the "-0" denotes that when the correct form is not in top-10, the probability assigned is 0.

A Results Per Language

In Table 2 we present the results for each language on the submitted model, as reported in the official task website as of May 18, 2023.