Neural Approaches to Reinflection

David R. Mortensen

February 25, 2025

Reinflection as a Sequence-to-Sequence Task

Seq2seq transduction models take a sequence of symbols as input and return a sequence of symbols as an output. Early seq2seq models included weighted finite state transducers (WFSTs). Neural seq2seq models begin in earnest with RNNs and encoder-decoder architectures.

Many NLP tasks have seq2seq formulations:

- **Machine translation**
- Text normalization
- Automatic grammar correction
- Text summarization
- Automatic speech recognition
- Test to speech synthesis
- Grapheme-to-phoneme transduction

If reinflection is understood in terms of a flat sequence of tokens (letters and morphological properties), it falls naturally into this class. Consider the following example from Totonac ('You (singular) swim').

(1) a. Input: <s> 2 S IPFV p a x </s>
b. Output: <s> p a x a ʔ </s>

where the purple tokens are properties and the violet tokens are letters (graphemes).

In principle, any architecture that can implement seq2seq transduction can be used to perform this task. Architectures that have been tried include

- RNN (Elman Machine)
- LSTM (Long Short-Term Memory)
- GRU (Gated Recurrent Unit)
- Transformer

For some time, the state of the art in this task was LSTM (with GRU remaining competitive). However, Wu et al. showed that Transformers could perform will at this task, if their batch size was large enough.[1] This is shown in Figure 1. When the Transformer models are trained at a relatively

[1] Shijie Wu, Ryan Cotterell, and Mans Hulden. Applying the transformer to character-level transduction. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online, April 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.163. URL https://aclanthology.org/2021.eacl-main.163

small batch size (16–128), their performance lags behind Wu and Cotterell (2019)'s[2] exact hard monotonic attention model. However, when the batch size is over 128, the Transformer models excel. Wu et al. (2021) explore two

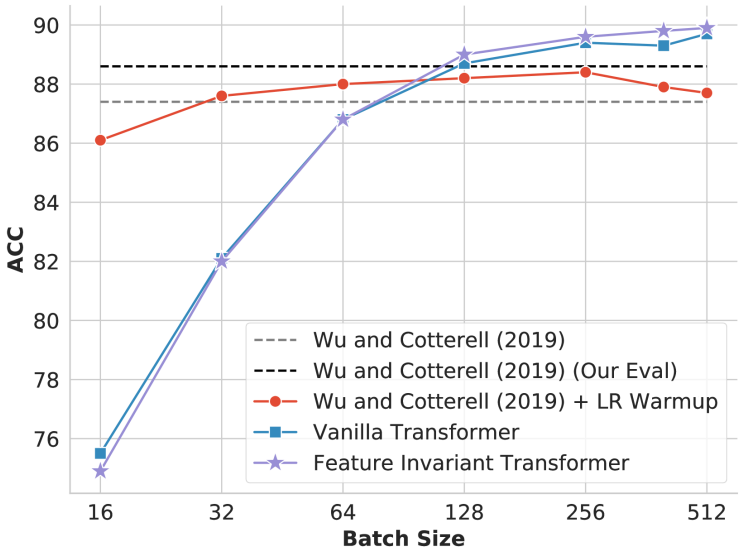

Figure 1: Morphological reinflection accuracy as a function of batch size.

Transformer-based models for character level transduction tasks (including reinflection):

- A Vanilla Transformer
- A Feature Invariant Transformer

To each input embeddings for the Feature Invariant Transformer, a vector encoding whether a token was a feature (property) or a character was concatenated. The feature invariant models performed slightly better.

Wu et al. make a few changes to the Transformer architecture:

(2) Character-level transformer

a. **A Smaller Transformer**. Only 4 encoder-decoder layers with 4 attention heads. Embedding size is 256 and the hidden size of the feed-forward layer is 1024.

b. **Feature Invariance**. The order of the features/properties is rendered irrelevant. They are assigned the same positional encodings, as shown in Figure 2. The features are set to position 0 and counting only begins when the character tokens begin.

Some Discussion

Here are some notes about Wu et al.'s results on the reinflection task:

[2] Shijie Wu and Ryan Cotterell. Exact hard monotonic attention for character-level transduction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy, July 2019. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1148. URL https://aclanthology.org/P19-1148

	Vanilla									Feature Invariant								
Token	<s>	V	1	S	IMPV	p	a	x	</s>	<s>	V	1	S	IMPV	p	a	x	</s>
	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Position	0	1	2	3	4	5	6	7	8	0	0	0	0	0	1	2	3	4
											+	+	+	+	+	+	+	
Type											F	F	F	F	C	C	C	

Figure 2: Comparing the vanilla versus feature invariant inputs from Wu et al, (2021).

(3) a. Slight improvement over earlier (hard attention-based) models

b. Phenomena that were hard for earlier architectures to model are no longer significantly different from more common phenomena, in terms of errors

c. Errors seem to be randomly distributed over languages, words

d. Advantage of transformer diminishes as length of words increases (counterintuitively)

References

Shijie Wu and Ryan Cotterell. Exact hard monotonic attention for character-level transduction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy, July 2019. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1148. URL https://aclanthology.org/P19-1148.

Shijie Wu, Ryan Cotterell, and Mans Hulden. Applying the transformer to character-level transduction. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online, April 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.163. URL https://aclanthology.org/2021.eacl-main.163.