University of Illinois Submission to the SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection

Marc E. Canby, Aidana Karipbayeva, Bryan J. Lunt, Sahand Mozaffari, Charlotte R. Yoder, Julia Hockenmaier
{marcec2,aidana2,bjlunt2,sahandm2,yoder6,juliahmr}@illinois.edu

Our Approach: Bidirectional Decoding

We apply **bidirectional decoding** (Zhou et al., 2019) to Task 0: Morphological Inflection.

- Words are generated simultaneously from left-to-right (L2R) and right-to-left (R2L).
- Each direction is conditioned on the other direction.

Bidirectional decoding works well for machine translation (Zhang et al., 2018; Zhou et al., 2019) because it reduces bias of left-to-right generation.

Bidirectional Decoding for Inflection

In morphological inflection, phonemes or graphemes may depend on either the preceding or the following context, or both.

Regressive assimilation in Kazakh:

kitap + N;ACC;DEF;SG → *kitapty* *kitap* + N;ACC;SG;PSS3S → *kitabı*

In the first case, the initial voiceless *t* of the suffix does not change the voicing of the *p*.
In the second case, *p* voices to *b* to assimilate with the following vowel *ı* of the case ending.

Phonetic conditioning in Latin:

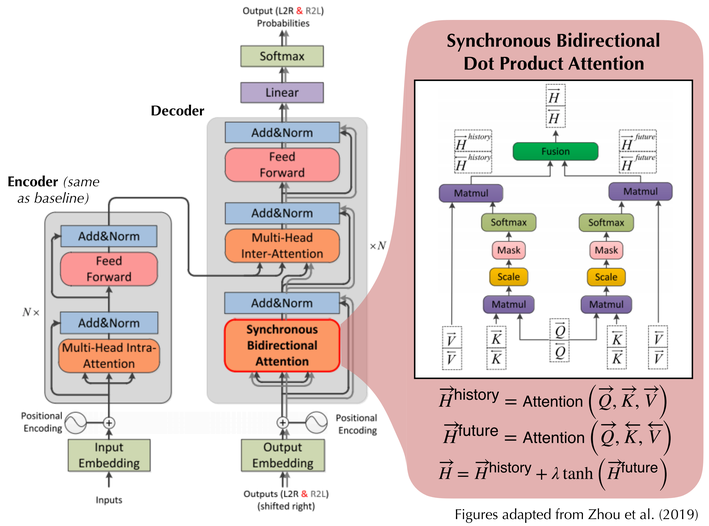
laudō + V;IND;PRS;2;SG → *laudās* *laudō* + V;IND;PRS;3;SG → *laudat*

The underlying morpheme *-ā-* marks the present tense, while *-s* and *-t* mark person.
In the second case, the underlying long vowel *ā* surfaces as short *a* due to the presence of the following stop consonant *t*.

Our Architecture: Transformer with Synchronous Bidirectional Attention

Build upon the baseline transformer of Wu et al. (2020):

1. Lemma & morphosyntactic tags embedded and fed to encoder *(same as baseline model)*.
2. Decoder generates L2R and R2L tokens in parallel at each time step.
3. Both directions share parameters, so the model has the same number of parameters as the unidirectional baseline.
4. Multi-head intra-attention replaced with Synchronous Bidirectional Attention mechanism (see right).



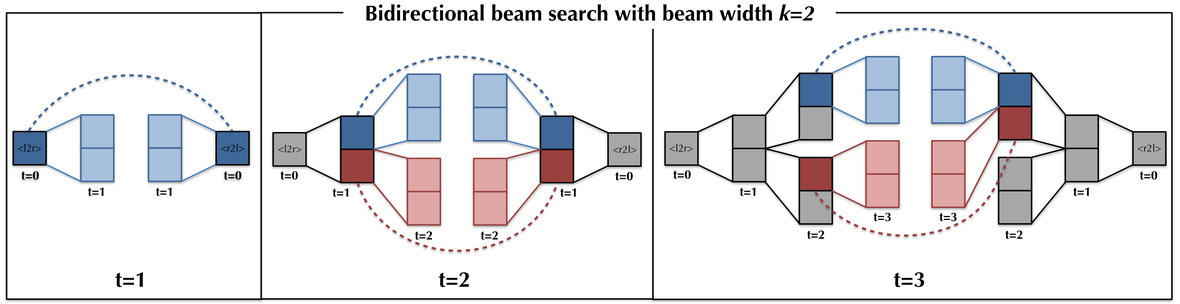
Figures adapted from Zhou et al. (2019)

Training Details:

1. Model trained to optimize likelihood of L2R & R2L outputs.
2. Separate model trained for each language, with hyperparameters selected for each family.

Inference: Bidirectional Beam Search

1. Pursue *k* best L2R and *k* best R2L hypotheses simultaneously.
2. At each step, feed *i*th best L2R hypothesis and *i*th best R2L hypothesis to decoder to generate new L2R and R2L predictions.
3. At the end, select hypothesis with highest probability to length ratio.



Experimental Results

Family	Accuracy			Edit-Distance		
	MONO	TRM	BI-TRM	MONO	TRM	BI-TRM
Afro-Asiatic	92.93	95.67	**96.37**	0.11	**0.05**	**0.05**
Algic	67.20	68.70	**70.30**	1.26	1.20	**1.16**
Australian	61.40	**90.00**	87.80	0.92	0.27	**0.26**
Austronesian	77.66	81.28	**82.30**	0.58	0.44	**0.41**
Dravidian	86.05	**87.10**	85.30	0.48	**0.46**	0.54
Germanic	86.88	**88.00**	87.38	0.30	**0.23**	0.25
Indo-Aryan	97.78	98.02	**98.18**	0.05	0.05	**0.04**
Iranian	63.00	82.50	**82.53**	1.04	**0.42**	0.46
Niger-Congo	97.72	97.72	**97.87**	0.04	0.04	**0.03**
Nilo-Sahan	0.00	87.50	**100.00**	2.88	0.19	**0.00**
Oto-Manguean	82.71	86.59	**87.49**	0.49	0.32	**0.28**
Romance	95.51	**99.25**	98.72	0.12	**0.02**	0.03
Sino-Tibetan	83.20	**84.40**	**84.40**	0.22	**0.20**	0.21
Siouan	92.90	**95.60**	94.90	0.16	**0.08**	0.10
Tungusic	55.30	**58.60**	58.30	1.20	**1.06**	1.09
Turkic	95.33	**95.96**	95.80	0.13	**0.10**	0.11
Uralic	83.21	**88.34**	88.18	0.39	0.29	**0.28**
Uto-Aztecan	76.30	80.80	**82.50**	0.49	0.41	**0.39**

Macro-averages of accuracy and edit distance by language family. Compared against MONO (Wu and Cotterell, 2019) and TRM (Wu et al., 2020).

	Acc. ≥	Acc. >	Avg. Edit Dist. ≤	Avg. Edit Dist. <
Development	27	18	30	14
Surprise	29	13	33	15

Number of languages (out of 45) on which our model equals or outperforms one or both of the neural baselines.

Conclusions & Future Work

Conclusions:

1. Strong performance against baselines make bidirectional decoding a promising direction
2. Some languages appear to strongly favor L2R hypotheses while others favor R2L hypotheses

Questions for Future Work:

1. How does the presence of various types of affixes affect the preferred decoding direction?
2. Initial experiments show the bidirectional transformer converges more quickly than the L2R baseline, despite the same number of parameters. What do further studies show?
3. How can a multilingual model be applied?

References:

1. 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, Florence, Italy. Association for Computational Linguistics.
2. Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction.
3. Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Ron- grong Ji, and Hongji Wang. 2018. Asynchronous bidirectional decoding for neural machine translation. In AAAI Conference on Artificial Intelligence.
4. Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous bidirectional neural machine translation. Transactions of the Association for Computational Linguistics, 7:91–105.