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.

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).

Training Details:

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

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

Regressive assimilation in Kazakh:

kitap + ᵘACC:DEF/SG  ↓  kitap

kaṭy

kitap

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 i of the case ending.

Phonetic conditioning in Latin:

laudās + ᵧIND:PRS:2SG  ↓↓  laudāt

The underlying morpheme -ā- marks the present tense, while -ā 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 r.

Bidirectional Decoding for Inference

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

Figure adapted from Zhou et al. (2019)

Inference: Bidirectional Beam Search

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

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: