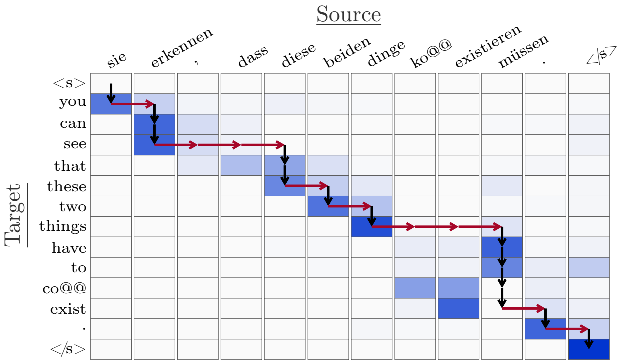
AMMI Reading Group: Beyond Sentences

Laurent Besacier, LIG, Univ. Grenoble Alpes

Beyond Sentences

- For speech transcription / translation / understanding the default granularity is often sentence (or utterance)
- With current encoder-decoder approaches, this results in encoding the full utterance (no more, no less) and decode an output hypothesis
- Not desirable in some use cases !
  - We sometimes have access to less (prefix of the utterance): **online (or simultaneous) speech processing**
  - We sometimes have access to more (full dialog history or full document): **context aware (or dialog level) speech processing**

Online Processing: Example of NMT



- The source-target alignments show that we can decode with an acceptable lagging behind an ideal online translator
- Similar idea for speech transcription / translation with sequence-to-sequence models

Online Sequence-to-Sequence Modeling

Let $(\mathbf{x}, \mathbf{y})$ be a source-target pair and $\mathbf{z} = (z_1, \ldots, z_{|\mathbf{y}|})$ the decoding context sizes.

1. A seq2seq model estimates $p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$:

$$p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \prod_{t=1}^{|\mathbf{y}|} p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x}_{\leq z_t}, \mathbf{z}_{\leq t}).$$

Online Sequence-to-Sequence Modeling

Let $(\mathbf{x}, \mathbf{y})$ be a source-target pair and $\mathbf{z} = (z_1, \ldots, z_{|\mathbf{y}|})$ the decoding context sizes.

1. A seq2seq model estimates $p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$:

$$p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \prod_{t=1}^{|\mathbf{y}|} p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x}_{\leq z_t}, \mathbf{z}_{\leq t}).$$

2. An agent models $p_\varphi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$, the likelihood of a decoding path $\mathbf{z}$ given $(\mathbf{x}, \mathbf{y})$ factorized as a sequence of $\mathrm{READ}/\mathrm{WRITE}$ actions with probabilities ρ_{tj}.

$$\rho_{tj} = p_\varphi(\mathrm{READ} \mid \mathbf{x}_{\leq j}, \mathbf{y}_{\leq t}).$$

Online Sequence-to-Sequence Modeling

Let $(\mathbf{x}, \mathbf{y})$ be a source-target pair and $\mathbf{z} = (z_1, \ldots, z_{|\mathbf{y}|})$ the decoding context sizes.

1. A seq2seq model estimates $p_\theta(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z})$:

$$p_\theta(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}) = \prod_{t=1}^{|\mathbf{y}|} p_\theta(y_t \,|\, \mathbf{y}_{<t}, \mathbf{x}_{\leq z_t}, \mathbf{z}_{\leq t}).$$

2. An agent models $p_\varphi(\mathbf{z} \,|\, \mathbf{x}, \mathbf{y})$, the likelihood of a decoding path $\mathbf{z}$ given $(\mathbf{x}, \mathbf{y})$ factorized as a sequence of $\mathrm{READ}/\mathrm{WRITE}$ actions with probabilities ρ_{tj}.

$$\rho_{tj} = p_\varphi(\mathrm{READ} \,|\, \mathbf{x}_{\leq j}, \mathbf{y}_{\leq t}).$$

3. With the EM algorithm, for an arbitrary distribution q, we optimize a lower-bound of the complete data log-likelihood:

$$\mathrm{Q}(\mathbf{x}, \mathbf{y}, q; \theta, \varphi) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p_{\theta,\varphi}(\mathbf{y}, \mathbf{z} \,|\, \mathbf{x}).$$

Deterministic Decoding Agents (wait-k)

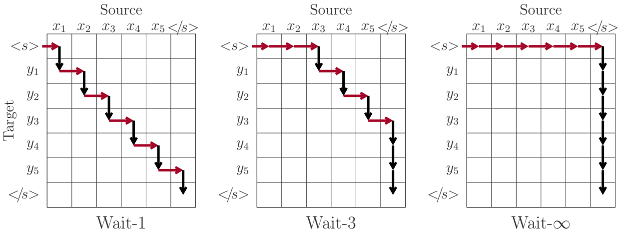
2. The likelihood of $\mathbf{z}$ given $(\mathbf{x}, \mathbf{y})$ is $p_\varphi(\mathbf{z} \,|\, \mathbf{x}, \mathbf{y}) = [\![\mathbf{z} = \mathbf{z}^{\text{wait-}k}]\!]$

$$\forall t \in [1..|\mathbf{y}|], \quad z_t^{\text{wait-}k} = \mathrm{clamp}_{[1..|\mathbf{x}|]}(k + t - 1)\,,$$
$$\rho_{tj} = p_{\text{wait-}k}(\mathrm{READ} \,|\, \mathbf{x}_{\leq j}, \mathbf{y}_{\leq t}) = [\![j < z_t^{\text{wait-}k}]\!],$$

Deterministic Decoding Agents (wait-k)

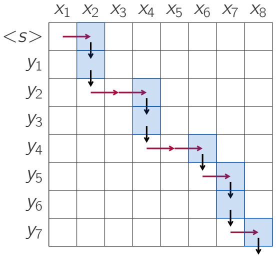
2. The likelihood of $\mathbf{z}$ given $(\mathbf{x}, \mathbf{y})$ is $p_\varphi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) = [\![\mathbf{z} = \mathbf{z}^{\text{wait-}k}]\!]$

$$\forall t \in [1..|\mathbf{y}|], \quad z_t^{\text{wait-}k} = \mathrm{clamp}_{[1..|\mathbf{x}|]}(k + t - 1)\,,$$
$$\rho_{tj} = p_{\text{wait-}k}(\mathrm{READ} \mid \mathbf{x}_{\leq j}, \mathbf{y}_{\leq t}) = [\![j < z_t^{\text{wait-}k}]\!],$$



Dynamic Decoding Agents with Pervasive Attention

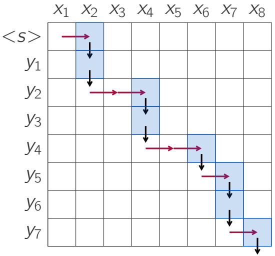
2. The READ/WRITE probabilities are predicted from the hidden states at every grid position (t, j).
3. EM with oracle-estimated q (dynamic programming).



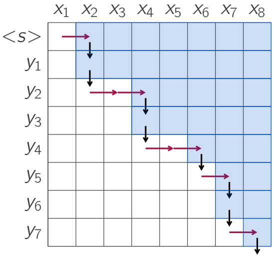
(a) Oracle path $\mathbf{z}^{\star}$

Dynamic Decoding Agents with Pervasive Attention

2. The $\mathrm{READ}/\mathrm{WRITE}$ probabilities are predicted from the hidden states at every grid position (t, j).
3. EM with oracle-estimated q (dynamic programming). The oracle path dictates the optimized writing probabilities (b) and the labels supervising the binary agent (c).



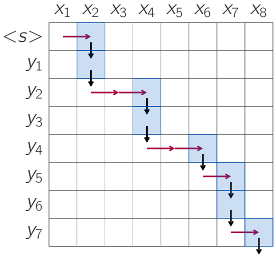
(a) Oracle path $\mathbf{z}^\star$



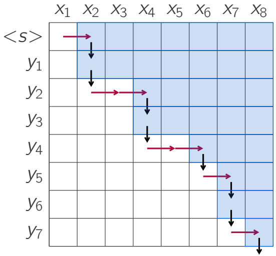
(b) Training the translation model

Dynamic Decoding Agents with Pervasive Attention

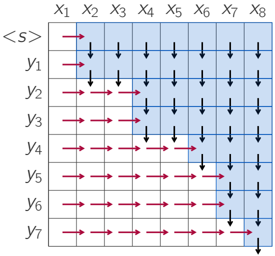
2. The READ/WRITE probabilities are predicted from the hidden states at every grid position (t, j).
3. EM with oracle-estimated q (dynamic programming). The oracle path dictates the optimized writing probabilities (b) and the labels supervising the binary agent (c).



(a) Oracle path $\mathbf{z}^\star$



(b) Training the translation model



(c) READ/WRITE agent supervision

Quality-lagging trade-off: example of NMT

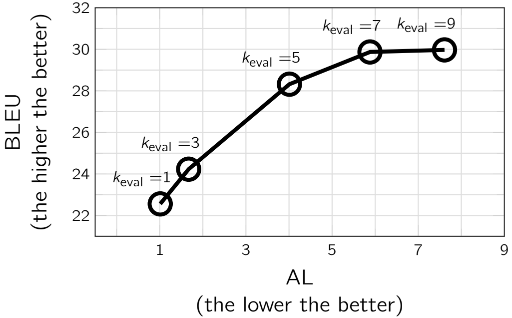
AL measures the lagging behind an ideal wait-0 online translator with $\tau(\mathbf{x}, \mathbf{z})$ the *cut-off* step when we finish reading the source $\mathbf{x}$.

$$\mathrm{AL} = \frac{1}{\tau(\mathbf{x}, \mathbf{z})} \sum_{t=1}^{\tau(\mathbf{x},\mathbf{z})} z_t - \frac{|\mathbf{x}|}{|\mathbf{y}|}(t-1), \quad \tau(\mathbf{x}, \mathbf{z}) = \min\{t \mid z_t = |\mathbf{x}|\}$$

Quality-lagging trade-off: example of NMT

AL measures the lagging behind an ideal wait-0 online translator with $\tau(\mathbf{x}, \mathbf{z})$ the *cut-off* step when we finish reading the source $\mathbf{x}$.

$$\mathrm{AL} = \frac{1}{\tau(\mathbf{x}, \mathbf{z})} \sum_{t=1}^{\tau(\mathbf{x},\mathbf{z})} z_t - \frac{|\mathbf{x}|}{|\mathbf{y}|}(t-1), \quad \tau(\mathbf{x}, \mathbf{z}) = \min\{t \mid z_t = |\mathbf{x}|\}$$



What About Seq2Seq Models for Speech ?

- Streaming automatic speech recognition with the transformer model
  https://arxiv.org/abs/2001.02674

What About Seq2Seq Models for Speech ?

- Streaming automatic speech recognition with the transformer model
  https://arxiv.org/abs/2001.02674
- End-to-End Simultaneous Translation System for the IWSLT2020 using Modality Agnostic Meta-Learning
  https://www.aclweb.org/anthology/2020.iwslt-1.5.pdf

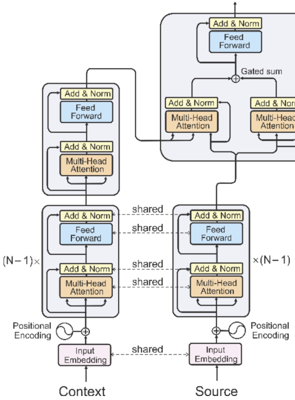
Streaming automatic speech recognition with the transformer model

- Apply time-restricted self-attention to the encoder
- Use triggered attention for the decoder, a CTC objective is used for alignment between encoder state sequence and target label sequence
- Joint CTC-attention training objective and decoding
- Streaming ASR system achieved WERs of 2.8% and 7.2% for the test-clean and test-other data sets of LibriSpeech

End-to-End Simultaneous Translation System for the IWSLT2020 using Modality Agnostic Meta-Learning

- Simultaneous translation for text-to-text(t2t) and speech-to-text(s2t) based on Transformer wait-k model
- Uni-directional encoder Transformer blocks
- Simple adaptation of the deterministic wait-k policy to speech
- Meta-learning approach to address data scarcity of speech-to-text task (sort of supervised pre-training)

Context Aware Processing: Example of NMT



- Take advantage of past context to process an utterance
- Need to model context separately or to model longer sequences (hot topic!)
- For speech, we may want to model dialog context (previous dialog turns)

What About Context Modeling for Speech ?

- Improving Conversation-Context Language Models with Multiple Spoken Language Understanding Models https://www.isca-speech.org/archive/Interspeech_2019/pdfs/1534.pdf

Improving Conversation-Context Language Models with Multiple Spoken Language Understanding Models

- Conversation-context language models (CCLMs) for handling multi-turn conversational ASR tasks
- Problem: conversation-level training datasets are often limited
- Idea: leverage spoken language understanding (SLU) models and integrate them in the CCLM model
- Improvements on contact center dialogue ASR tasks