Self-Attentional Models for Lattice Inputs

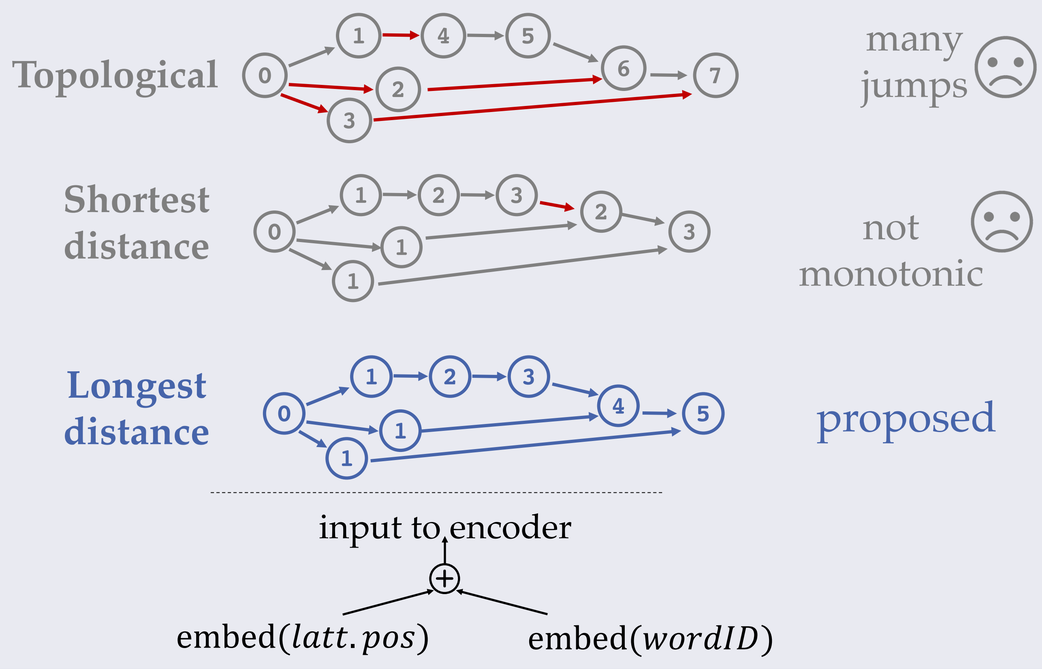
Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, Alex Waibel

Karlsruhe Institute of Technology (Germany) / Carnegie Mellon University (USA)

Motivation

- Lattice self-attention:
  - Fast & easy to implement
  - Can be integrated into existing self-attentional models
  - Leverage lattice confidence scores
- Prior work:
  - Lattice GRUs / Lattice LSTMs
    - Slow, non-trivial implementation
  - Graph Convolutional Nets
    - Local context, need combination with LSTMs
- Here: devise **lattice-to-sequence** model for **speech translation**
- Other usecases: word segmentation latt., alternative video descriptions, morphological analyses, word classes, …

1. Lattice Position Representations



2. Lattice Self-Attention (SA)

Masked self-attention

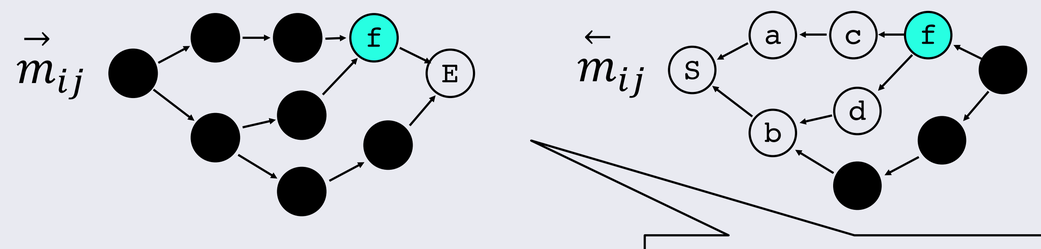
$$e_{ij} = f(\mathbf{x}_i, \mathbf{x}_j) + m_{ij}$$

(query, key, mask)

$$\alpha_i = \underset{\text{normalized score}}{\text{softmax}}(\mathbf{e}_i)$$

$$\mathbf{y}_i = \underset{\text{output}}{\sum_{j=1}^{l}} \alpha_{ij} \underset{\text{value}}{\mathbf{x}_j}$$

We induce lattice structure through reachability masks



Mimics conditioning structure of Lattice-LSTM

Mask type:

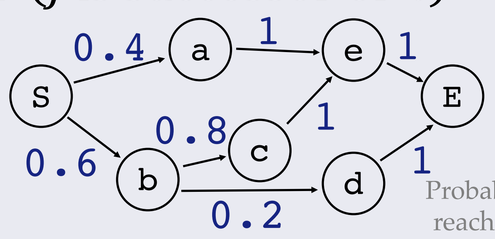
a) Binary masking

$$\vec{m}_{ij} = \begin{cases} 0, & \textbf{if } j \text{ successor of } i \\ -\infty, & \textbf{else} \end{cases}$$

Decrease impact of low-confidence alternatives

b) Probabilistic masking

$$\vec{m}_{ij} = \log P(j \text{ is successor of } i)$$



Probabilistic reachability mask (before log)

$\vec{m}_{ij}$	S	a	b	c	d	e	E
S	1	.4	.6	.48	.12	.88	1
a	0	1	0	1	0	1	1
b	0	0	1	0.8	0.2	0.8	1
c	0	0	0	1	0	1	1
d	0	0	0	0	1	0	1
e	0	0	0	0	0	1	1
E	0	0	0	0	0	0	1

Directionality:

Compute backward masks $\overleftarrow{m}_{ij}$ on reversed lattices

a) **Non-directional model:** merge masks ($\overleftrightarrow{m}_{ij} = \max\{\vec{m}_{ij}, \overleftarrow{m}_{ij}\}$).
b) **Directional model:** dedicate half of attention heads to each direction (multi-head attention) *[Shen+2018]*

3. Main Results

- Fisher/Callhome ES → EN speech translation corpus *[Post+2013]*
  - Fisher: 138k sentences; Callhome: 15k
- XNMT seq2seq toolkit, Transformer encoder (3 layers), MLP cross-attention, LSTM decoder
- Pretrain on sequential data (Fisher), then finetune on provided lattice data (Fisher or Callhome)

Model	*Input*	*Fisher*	*Callh.*
LSTM	1best	35.9	11.8
Seq. SA	1best	35.7	12.4
Seq. SA (bidirect.) *[Shen+2018]*	1best	37.4	13.0
Graph attention *[Veličković+2018]*	Latt.	35.7	11.9
LattLSTM	Latt.	38.0	14.1
Lattice SA (prop.)	Latt.	**38.7**	**14.7**

BLEU scores on Fisher (4 refs.) and Callhome (1 ref.) data.

→ **Proposed SA lattice encoder achieves best BLEU score**

	Training		*Inference*	
	Batch.	*W/sec*	*Batch.*	*W/sec*
Sequential encoders				
LSTM	Manu	4629	-	715
SA	Auto	**5021**	-	**796**
LatticeLSTM & lattice SA				
LSTM	-	178	-	391
LSTM	Auto	710	Auto	538
SA	Manu	**2963**	-	687
SA	Auto	748	Auto	**718**

Speed comparison.
Batched computations: **manual** vs. **disabled** vs. DyNet **autobatching** *[Neubig+2017]*.

→ **Much faster training and inference**

4. Analysis

Feature ablation

reach. mask	*direct. ?*	*mask type*	*pos. repr.*	*Fisher*	*Callh.*
✓	✓	**Prob.**	**Ldist.**	**38.7**	**14.7**
✓	✓	**Prob.**	Topo.	38.3	12.5
✓	✓	Bin.	**Ldist.**	37.5	14.4
✓		**Prob.**	**Ldist.**	35.5	12.8
	✓	Bin.	Topo.	30.6	9.4

BLEU scores, ablation: reachability masks, directional vs. non-directional masking, probabilistic vs. binary masking, longest-distance latt. positions vs. topological positions

→ **All introduced features contribute to gains**

Pretraining & finetuning

Pretrain (sequential data)	*Finetune (lattice data)*	*Fisher test*	*Callh. test*
-	Fisher	1.5	1.8
Callh.	Fisher	34.5	13.0
Fisher	Callh.	35.5	**14.7**
Fisher	Fisher	**38.7**	14.6

BLEU scores when pretraining and finetuning on Fisher vs. the much smaller Callhome training data.

→ **Pretraining essential**
→ **Lattice finetuning data can be small, but should be in-domain**

Conclusion

- **reachability masks** & **lattice position representations** to achieve self-attentional lattice encoder
- Outperform LatticeLSTM (accuracy and speed)
- Easy to implement

Code available: `http://msperber.com/research/acl-lattice-selfatt`

Karlsruhe Institute of Technology

Carnegie Mellon University

www.kit.edu