Recurrent Neural Network Attention Mechanisms for Interpretable System Log Anomaly Detection

Presented By:

Akash Kulkarni

System Log Analysis is complicated

1. Log sources generate TBs of data per day
2. Lack of labelled data (scarce or unbalanced)
3. Actionable information may be obscured (by complex relationships across logging sources)

Need for an aided human monitoring and assessment.

Unsupervised RNN language models

- Models distribution of normal events in system logs (learns complex relationships buried in logs)
- No need of labelled data.
- No feature engineering required (as deep learning learns significant features automatically)

Language Modelling

- Each log-line consists of sequence of T tokens:
  - $x_{(1:T)} = x_{(1)}, x_{(2)}, x_{(3)} \ldots\ldots x_{(T)}$, *each token* $x_{(i)} \in V$
- Language model (like RNNs) assigns probabilities to sequences:
  - $P(x_{(1:T)}) = \prod_{t=1}^{T} P(x_{(t)} | x_{(<t)})$
- Tokenization can be word-based or character-based.

Cyber Anomaly Language Models

1. Event Model (EM)– which applies a standard LSTM to log lines

$$\mathrm{LSTM}(\mathbf{x}_{(1:T)}) = \mathbf{h}_{(1:T)}$$

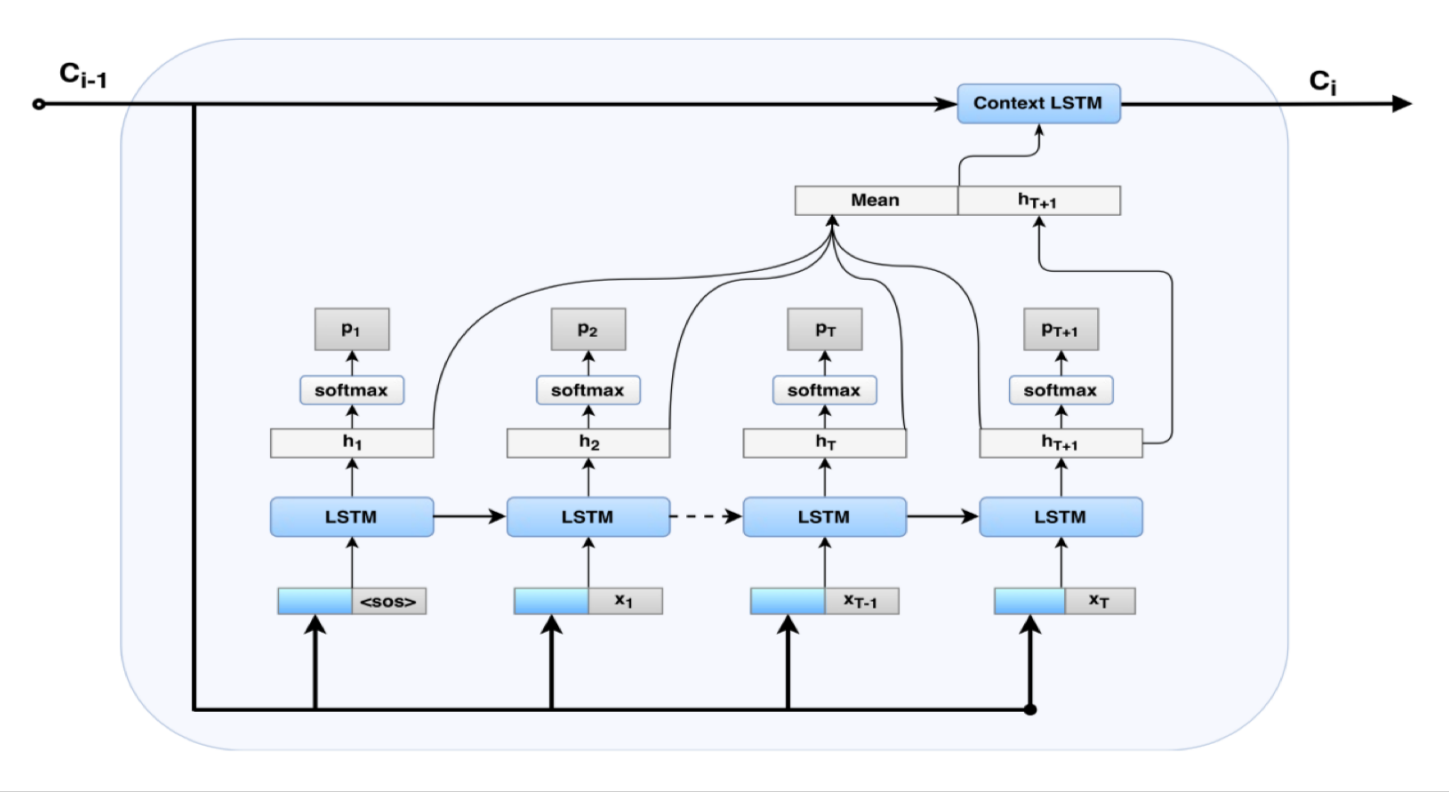
$$\mathbf{p}_{(t)} = \mathrm{softmax}\Big(\underbrace{\mathbf{h}_{(t-1)}}_{L_h} \underbrace{\mathbf{W}}_{L_h \times |\mathbb{V}|} + \underbrace{\mathbf{b}}_{|\mathbb{V}|}\Big) \in \mathbb{R}^{|\mathbb{V}|}$$

2. Bidirectional Event Model (BEM)

$$\mathbf{p}_{(t)} = \mathrm{softmax}\left(\mathbf{h}_{(t-1)}\mathbf{W} + \mathbf{h}^{b}_{(t+1)}\mathbf{W}^{b} + \mathbf{b}\right)$$

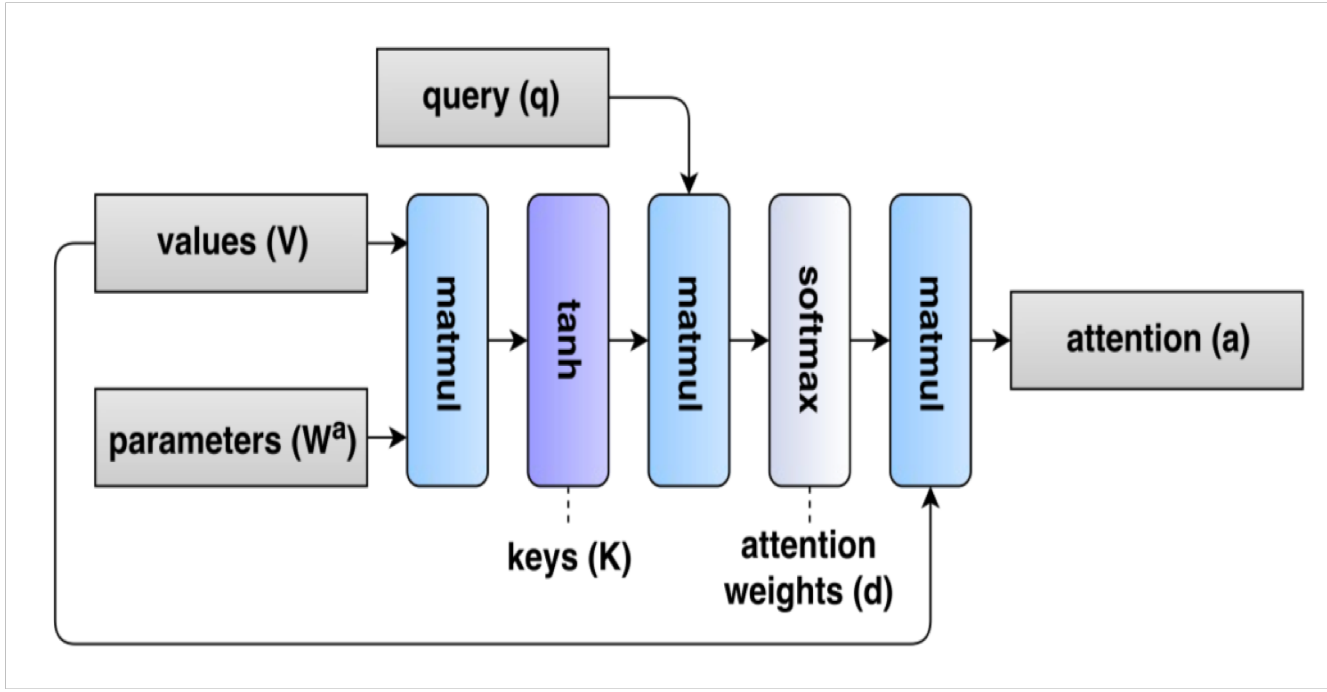
Cyber Anomaly Language Models

3. Tiered Language Model (T-EM or T-BEM)



Attention

- Key matrix, $K = \tanh(VW^a)$
- Weights, $d = softmax(qK^T)$
- Attention, $a = dV$



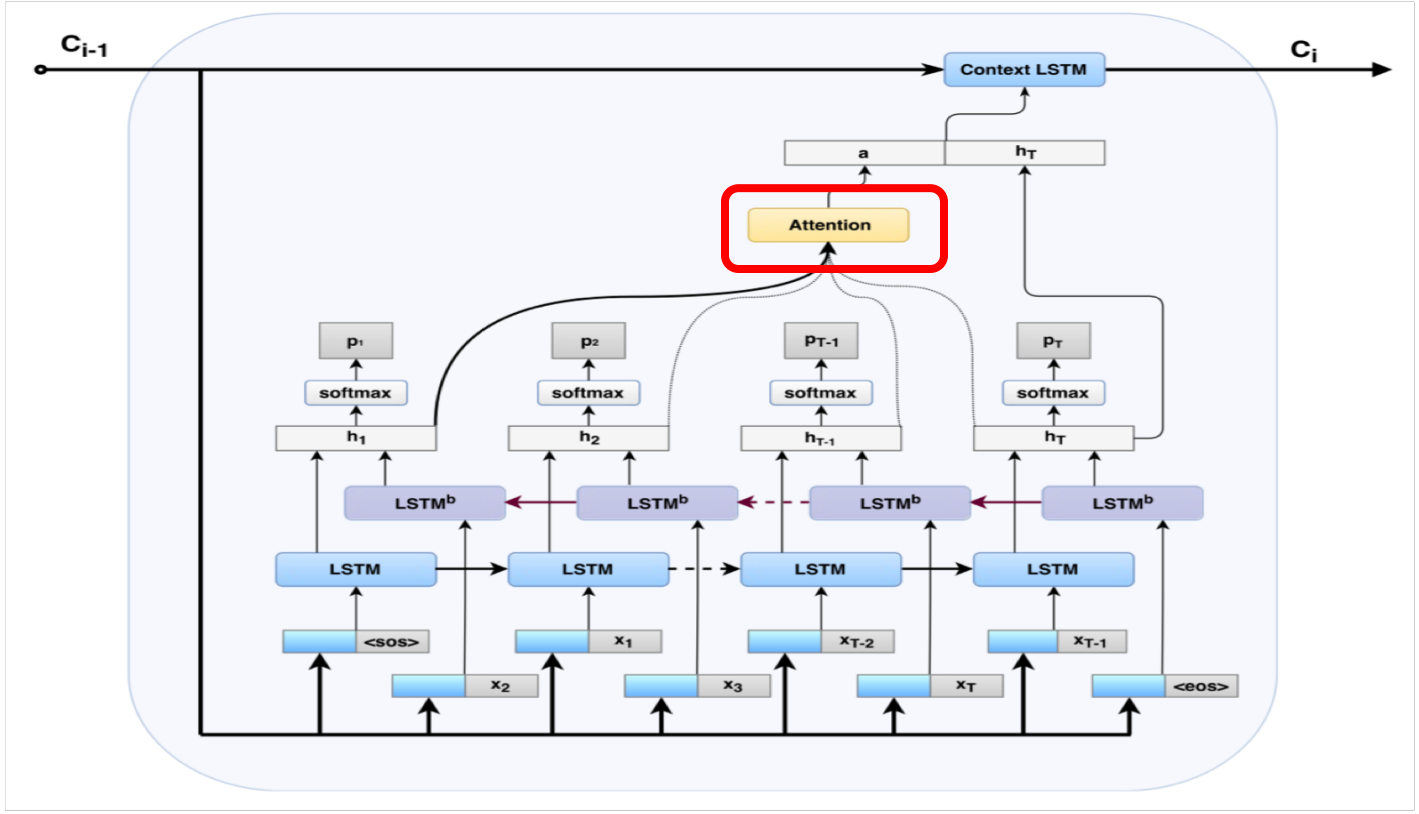
- Prediction, $$\mathbf{p}_{(t)} = \text{softmax}\left(\left[\mathbf{h}_{(t-1)} \quad \mathbf{a}_{(t-1)}\right]\mathbf{W} + \mathbf{b}\right)$$

EM attention variants

1. Fixed Attention:
   - $q(t) = q$
   - Assumes some position in the sequence are more important than others
2. Syntax Attention:
   - $q(t)$ not shared across t
   - Importance depends on the position of the current token in sequence.
3. Semantic Attention 1:
   - $\mathbf{q}_{(t)} = \tanh(\mathbf{h}_{(t)}\mathbf{W}^{sem1})$,
4. Semantic Attention 2:
   - $h'_{(t)}$ = concatentation of $h_{(t)}$ and $q_{(t)}$

EM attention variants

5. Tiered Attention
   - Replaces mean with weighted average via attention



$$
\begin{aligned}
\mathbf{q} &= \tanh(\mathbf{h}_{(T)}\mathbf{W}^{(tier)}) \\
\mathbf{K} &= \tanh(\mathbf{V}\mathbf{W}^{a}) \\
\mathbf{d} &= \mathrm{softmax}(\mathbf{q}\mathbf{K}^{T}) \\
\mathbf{a} &= \mathbf{d}\mathbf{V}
\end{aligned}
$$

Results

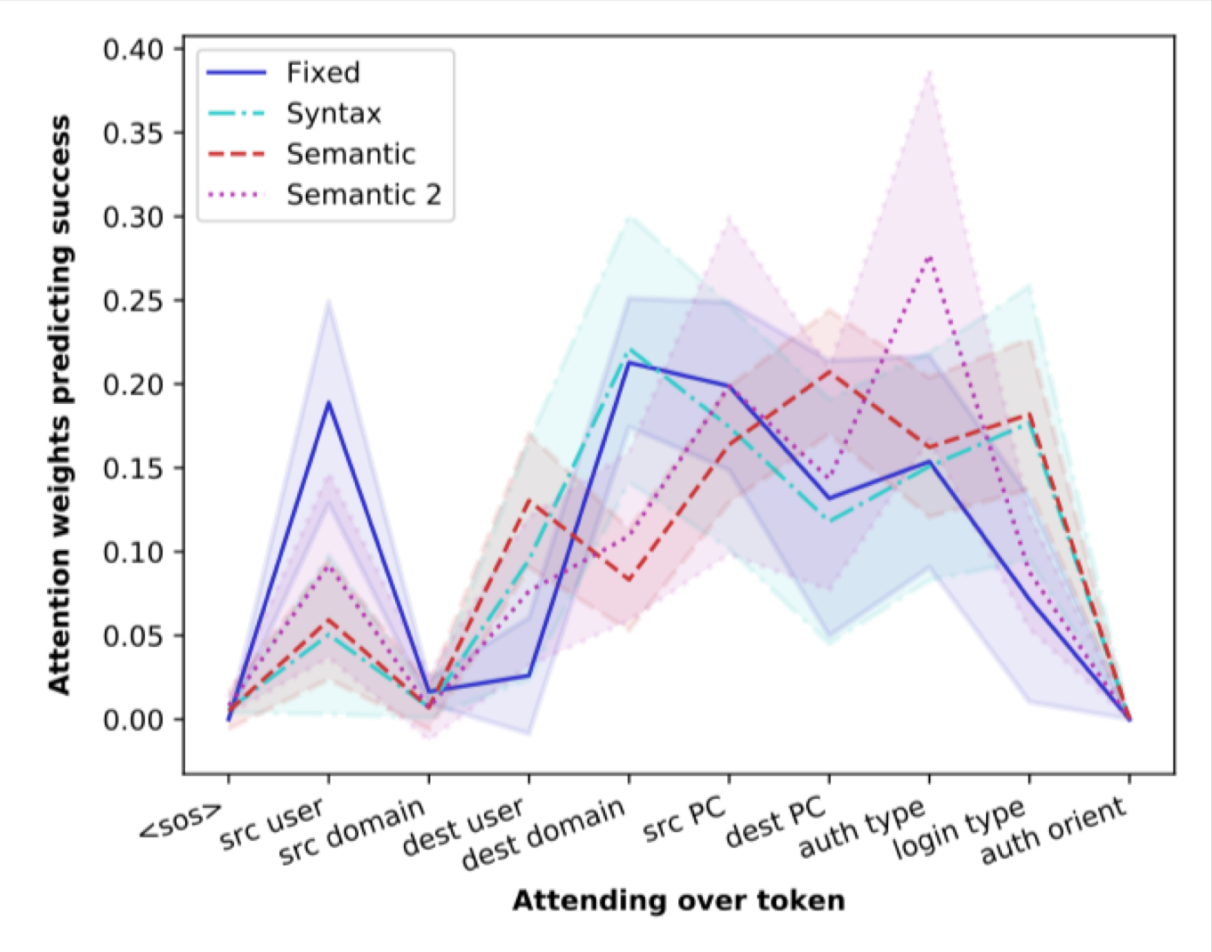
Model	Mean	Max	Min	Std. Dev.
EM	0.968	0.976	0.964	0.005
BEM	0.976	0.981	0.972	0.003
EM with attention				
Fixed	0.974	0.976	0.972	0.001
Syntactic	0.972	0.975	0.967	0.004
Semantic 1	0.975	0.980	0.971	0.004
Semantic 2	0.973	0.976	0.968	0.003
Tiered LSTM variants				
T-EM	0.984	0.989	0.977	0.005
T-BEM	0.987	0.989	0.985	0.002
TA-EM	0.985	0.991	0.979	0.004
TA-BEM	0.988	0.991	0.984	0.003

Table 1. AUC statistics for word tokenization models

Model	Mean	Max	Min	Std. Dev.
EM	0.965	0.969	0.961	0.003
BEM	0.985	0.987	0.982	0.002
EM with attention				
Fixed	0.963	0.971	0.937	0.015
Syntactic	0.967	0.973	0.963	0.004
Semantic	0.975	0.977	0.971	0.003
Semantic 2	0.972	0.977	0.967	0.004
Tiered LSTM variants				
T-EM	0.977	0.988	0.967	0.008
T-BEM	0.992	0.992	0.991	0.000
TA-EM	0.982	0.984	0.979	0.002
TA-BEM	0.991	0.992	0.990	0.001

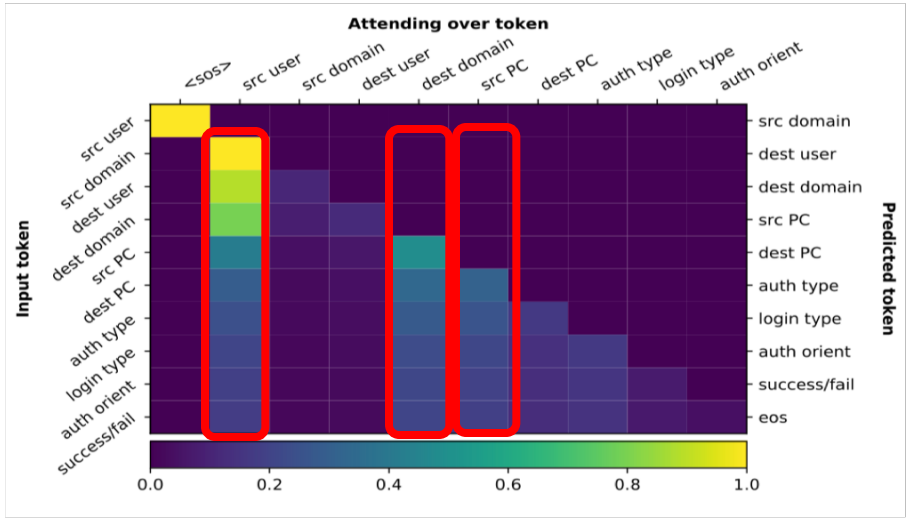
Table 2. AUC statistics for character tokenization models

Analysis

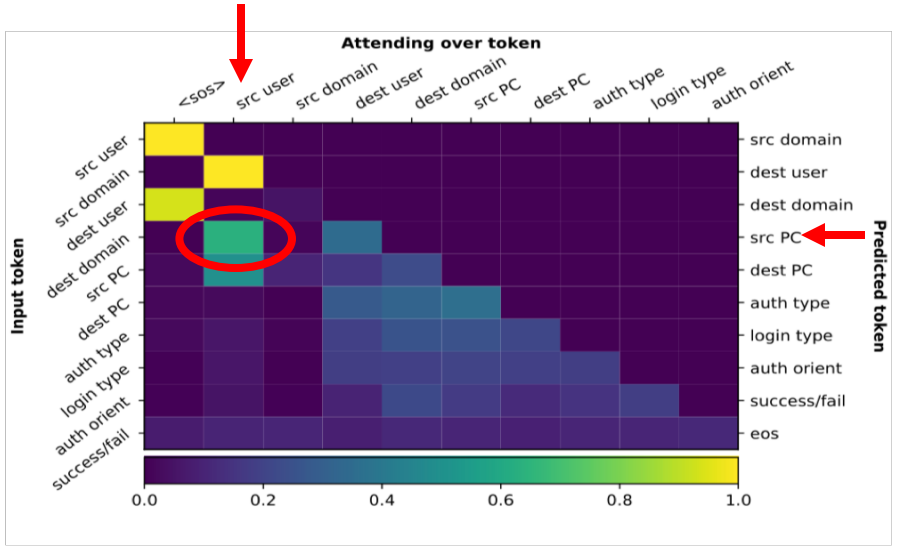


Comparison of attention weights when predicting success/failure token

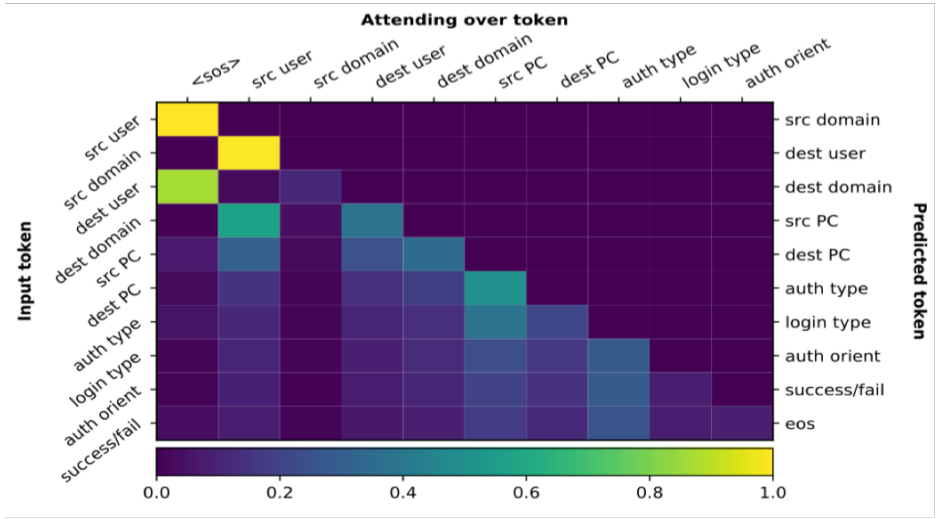
Analysis



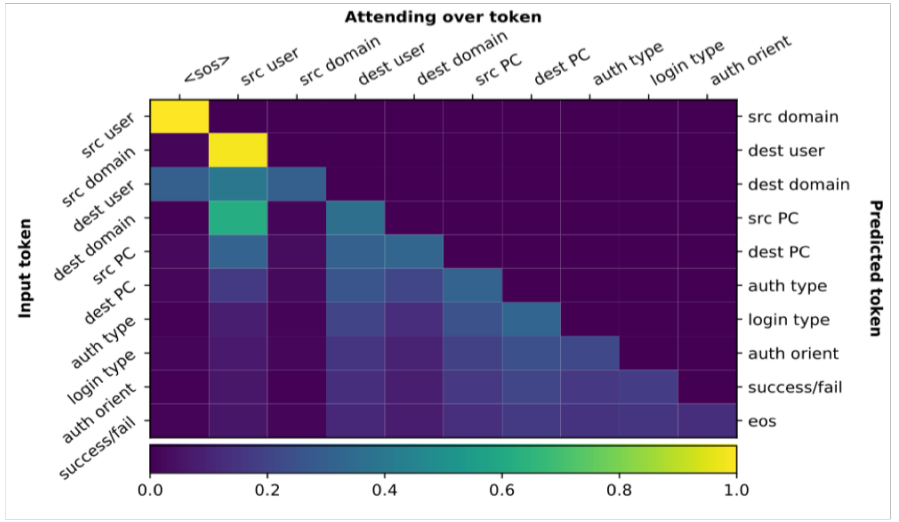
1. Average Fixed attention weights



2. Average Syntax attention weights



3. Average Semantic 1 attention weights

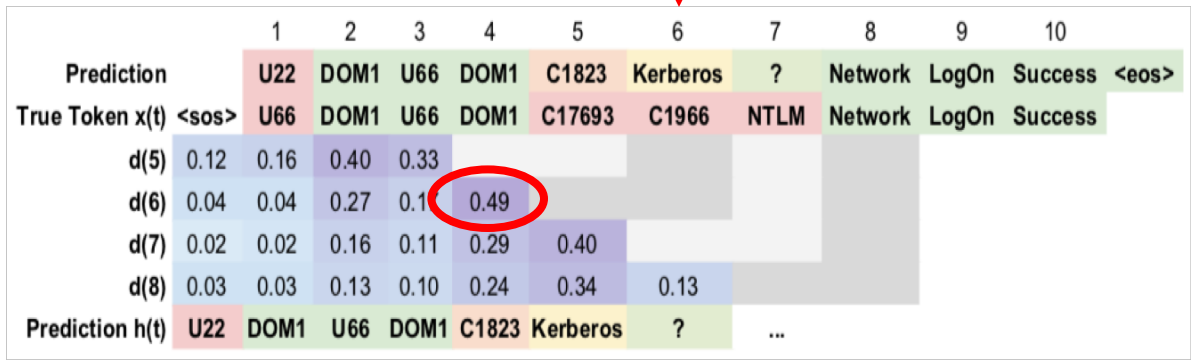


4. Average Semantic 2 attention weights

Analysis

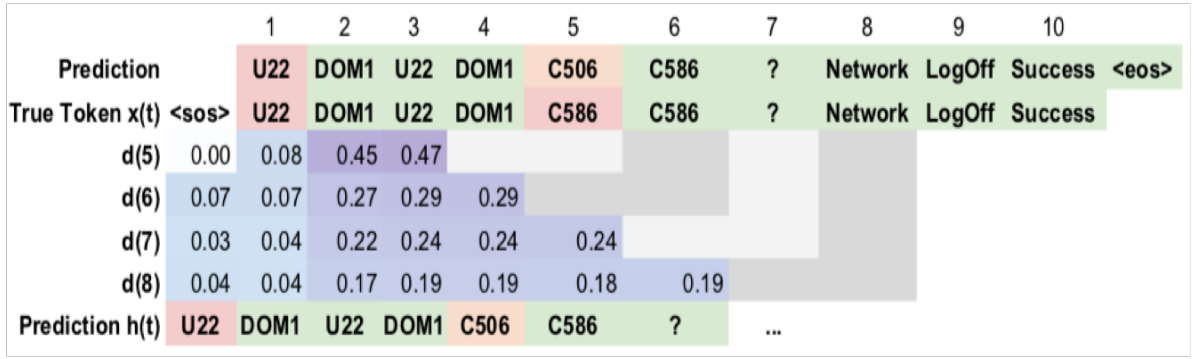
- Tiered attention models
  - For lower forward-directional LSTM, attention weights were nearly 1.0 for 2nd to last hidden state.
  - For lower bidirectional LSTM, attention weights were nearly 1.0 for 1st hidden state and last state
  - Hence, attentions are not needed for this model task.

Case Studies



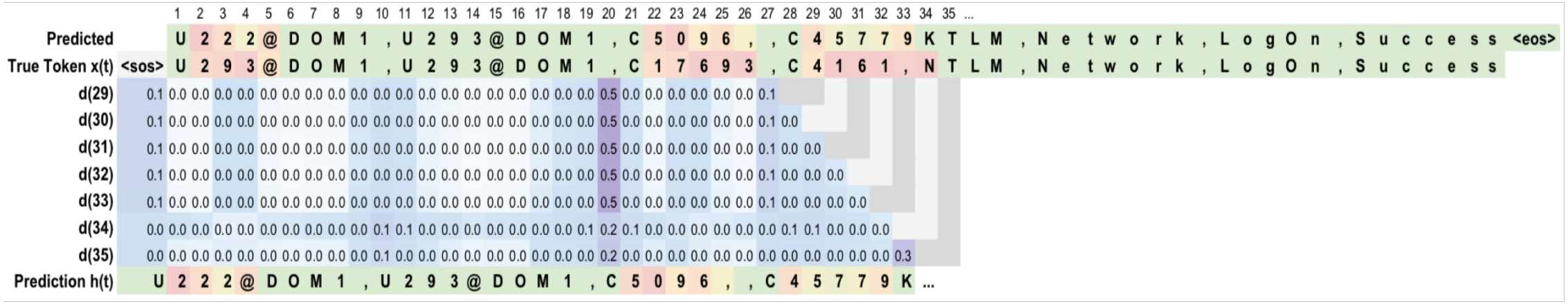
		1	2	3	4	5	6	7	8	9	10	
Prediction		U22	DOM1	U66	DOM1	C1823	Kerberos	?	Network	LogOn	Success	<eos>
True Token x(t)	<sos>	U66	DOM1	U66	DOM1	C17693	C1966	NTLM	Network	LogOn	Success	
d(5)	0.12	0.16	0.40	0.33								
d(6)	0.04	0.04	0.27	0.1	0.49							
d(7)	0.02	0.02	0.16	0.11	0.29	0.40						
d(8)	0.03	0.03	0.13	0.10	0.24	0.34	0.13					
Prediction h(t)	U22	DOM1	U66	DOM1	C1823	Kerberos	?	...				

1. Word case study with Semantic attention



		1	2	3	4	5	6	7	8	9	10	
Prediction		U22	DOM1	U22	DOM1	C506	C586	?	Network	LogOff	Success	<eos>
True Token x(t)	<sos>	U22	DOM1	U22	DOM1	C586	C586	?	Network	LogOff	Success	
d(5)	0.00	0.08	0.45	0.47								
d(6)	0.07	0.07	0.27	0.29	0.29							
d(7)	0.03	0.04	0.22	0.24	0.24	0.24						
d(8)	0.04	0.04	0.17	0.19	0.19	0.18	0.19					
Prediction h(t)	U22	DOM1	U22	DOM1	C506	C586	?	...				

2. Low anomaly word case study with Semantic attention



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	...
Predicted		U	2	2	2	@	D	O	M	1	,	U	2	9	3	@	D	O	M	1	,	C	5	0	9	6	,	,	C	4	5	7	7	9	K	T	L	M	,	N	e	t	w	o	r	k	,	L	o	g	O	n	,	S	u	c	c	e	s	s	<eos>
True Token x(t)	<sos>	U	2	9	3	@	D	O	M	1	,	U	2	9	3	@	D	O	M	1	,	C	1	7	6	9	3	,	C	4	1	6	1	,	N	T	L	M	,	N	e	t	w	o	r	k	,	L	o	g	O	n	,	S	u	c	c	e	s	s	
d(29)	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.1																																			
d(30)	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0																																	
d(31)	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0																																	
d(32)	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0																																	
d(33)	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0																																	
d(34)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0																																	
d(35)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3																															
Prediction h(t)		U	2	2	2	@	D	O	M	1	,	U	2	9	3	@	D	O	M	1	,	C	5	0	9	6	,	,	C	4	5	7	7	9	K	...																																	

3. Case character study with Semantic attention

Conclusions

- Fixed and syntactic attention effective for fixed structure sequences.
- Attention mechanism improve performance and provide feature importance and relational mapping between features.

Future directions

- Explore BEM with attention
- Equipping a lower tier model to attend over upper tier hidden states