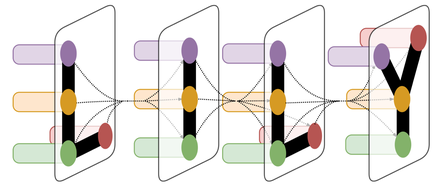
Sequence to Logic with History Gates

Javid Dadashkarimi, Alex Fabbri and Dragomir Radev
{javid.dadashkarimi, alexander.fabbri, dragomir.radev}@yale.edu
Computer Science Department, Yale University



Problem

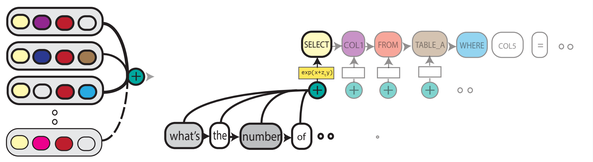


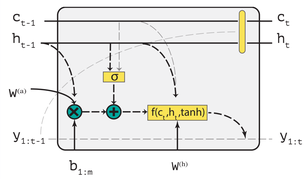
Fig. 1 sequence to SQL.

1. Generating logic form of human language is a fresh way with neural architecture
2. The logic form of the sequence usually preserves information in the human side with similar tokens
3. recently a copying mechanism is proposed where increases the probability of the source tokens through decoding [1]

Introduction

1. Recently [1] proposed an augmented pointer network for sequence to logic
2. This is a very straightforward mechanism to pay attention to the source sequence
   - `x: what rivers flow through colorado ?`
   - `y: answer ( A , ( river ( A ) , traverse ( A , B ) , const ( B , stateid ( colorado ) ) ) )`
3. In this paper we step forward and introduce history gates to take whole the source vocabulary into account for decoding (e.g., '`flow through`' and '`traverse`').

History Gates



$$e_{ji} = [s_j^T W^{(a)} b_i] + [z_t \odot f(s_j, c_j)] \quad (1)$$

$$\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{i'=1}^{m} \exp(e_{ji'})} \quad (2)$$

$$c_j = \sum_{i=1}^{|V_s|} \alpha_{ji} b_i \quad (3)$$

$$P(y_j = w | x, y_{1:j-1}) \propto \exp(U_w [s_j, c_j]) \quad (4)$$

$$P(y_j = \text{copy}[i] | x, y_{1:j-1}) \propto \exp(e_{ji}) \quad (5)$$

1. One simple approximation for $f(s_j, c_j)$ is $f(s_j, c_j) = s_j^T W^{(a)} W^{(h)}$ where $W^{(a)} \in \mathbb{R}^{d \times 4d}$ and $W^{(h)} \in \mathbb{R}^{4d \times |V_s|}$ as the history matrix.
2. $z_t = \sigma([s_j^T U_{z_t}] + [c_j^T W_{z_t}])$ is a reset gate for the context c_j and state s_j where $U_{z_t}, W_{z_t} \in \mathbb{R}^{d \times |V_s|}$.

$f(x, y; \theta)$
$f_1(s_j, c_j) = s_j^T W^a W^{(h)}$
$f_2(s_j, c_j, \sigma) = \sigma(s_j^T W^a W^{(h)})$
$f_3(s_j, c_j, \sigma) = \sigma(s_j^T W^a W^{(h)} + c_j W^{(h)})$
$f_4(s_j, c_j, \sigma) = \sigma(s_j^T U_{z_t})$
$f_5(s_j, c_j, \tanh) = \tanh(s_j^T U_{z_t})$
$f_6(s_j, c_j, \sigma) = \sigma(s_j^T W^a W^{(h)})$ $+(1 - z_t)/z_t \odot \sigma(c_j W^{(h)})$

Table 1: different history functions.

Results

ID	calendar		housing		publications		recipes	
	SEQ	TOK	SEQ	TOK	SEQ	TOK	SEQ	TOK
attn	0.565	0.883	0.476	**0.837**	0.596	0.822	0.741	**0.924**
$f_1(s_j, c_j)$	0.589	0.885	**0.508**	.832	0.602	**0.838**	**0.755**	0.922
$f_2(s_j, c_j, \sigma)$	**0.613**	**0.890**	0.439	0.821	0.621	0.833	0.690	0.904
$f_3(s_j, c_j, \sigma)$	0.548	0.883	0.439	0.821	0.6211	0.833	0.689	0.904
$f_4(s_j, c_j, \sigma)$	0.554	0.873	0.471	0.826	0.627	0.835	0.689	0.904
$f_5(s_j, c_j, \tanh)$	0.554	0.873	0.407	0.806	0.609	0.823	0.690	0.904
$f_6(s_j, c_j, \sigma)$	0.524	0.867	0.407	0.806	**0.640**	0.816	0.690	0.904

x	y	$y_j = \text{copy}[i]$, $y_j \in V_s, y_j \notin x$
`housing units with 2 neighborhoods`	`( call SW.listVal. ( call SW.countComp. ( call SW.getProp. ( call SW.singleton en.housing_unit ) ( string ! type )) ( string neighborhood ) ( string = ) ( number 2 )))`	`number`

Discussion and Conclusion

1. the performance of the method is not independent of the domain.
2. history gates are general form of the traditional copying mechanism.
3. the proposed method generally achieves consistent improvements in token level accuracy.
4. for example in the publication data: `authors of articles published in 2004` both the `data` and `2004` were selected from history not the source sequence. `2004` is already in the source sequence. the reason is that dates and numbers are very common in this domain.

In this paper we introduced the history gates for sequence to sequence encoder/decoder system. The previous copying mechanism could boost performance by giving chance to the source tokens to appear in target. The history gates gives this possibility for all the source vocabulary that might have semantic relationship with current context.

References

[1] Robin Jia and Percy Liang. Data recombination for neural semantic parsing. *arXiv preprint arXiv:1606.03622*, 2016.

[2] Luke S Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, 2012.

Future Research

1. **transfer learning**: for using hidden information from one domain in another.
2. **machine translation**: for out out vocabulary words (e.g., generating Barack Obama in French by an English sentence).
3. **recommender system:** recomending movies for a user based on her/his most recent interests and also a long history.

Contact Information

Web https://yale-lily.github.io, http://dadashkarimi.com
Email javid.dadashkarimi@yale.edu
Phone +1 (475) 731 4330