SEntNet: Source-aware Recurrent Entity Network for Dialogue Response Selection

Jiahuan Pei[1] Arent Stienstra[1] Julia Kiseleva[2] Maarten de Rijke[1]

[1]University of Amsterdam
[2]Microsoft Research AI

Overview

- **Goal**. Select an appropriate response from candidates given a dialogue context for Task-oriented Dialogue Systems (TDSs).
- **Problem**. Obtaining key information from a complex, long dialogue context is challenging, especially when different sources of information are available.
- **Solution**. Employ source-specific memories to exploit differences in the usage of words and syntactic structure from different information sources, i.e., user, system, and knowledge base (KB).

System Response Selection in TDSs

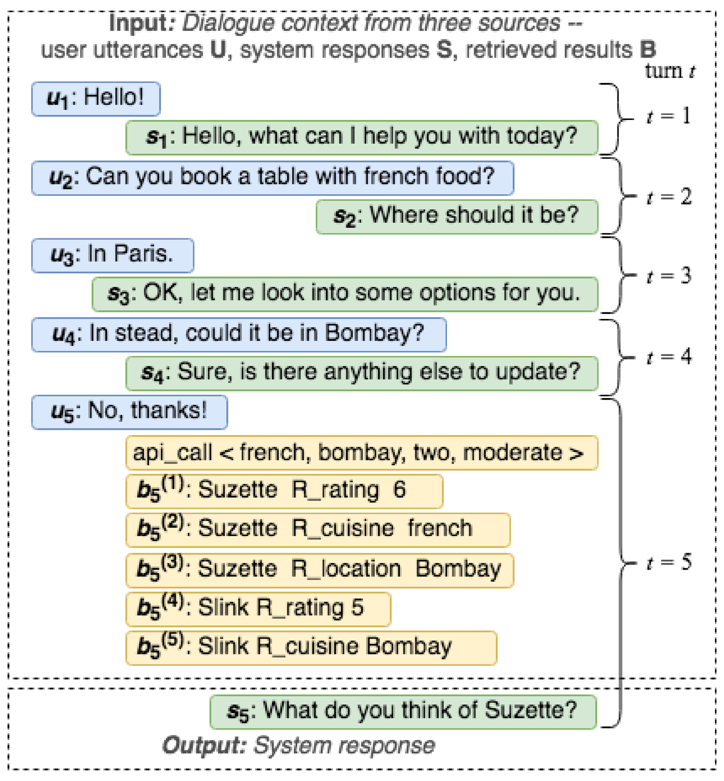


Figure: An example of response selection for booking a restaurant. The top box contains the input for response selection; the bottom box shows the selected response.

- **Given**: a dialogue context $(\mathbb{U}_t, \mathbb{S}_{t-1}, \mathbb{B}_t)$
  - $\mathbb{U}_t = (u_1, u_2, \ldots, u_t)$ are user utterances;
  - $\mathbb{S}_{t-1} = (s_1, s_2, \ldots, s_{t-1})$ are system responses; and
  - $\mathbb{B}_t = (b_t^1, b_t^2, \ldots, b_t^{\lambda})$ is λ-best retrieved results from an external KB.
- **Goal**: select a response s_t from candidates by

$$\psi_\Theta(\mathbb{U}_t, \mathbb{S}_{t-1}, \mathbb{B}_t) \rightarrow s_t. \quad (1)$$

Source-aware Recurrent Entity Network (SEntNet)

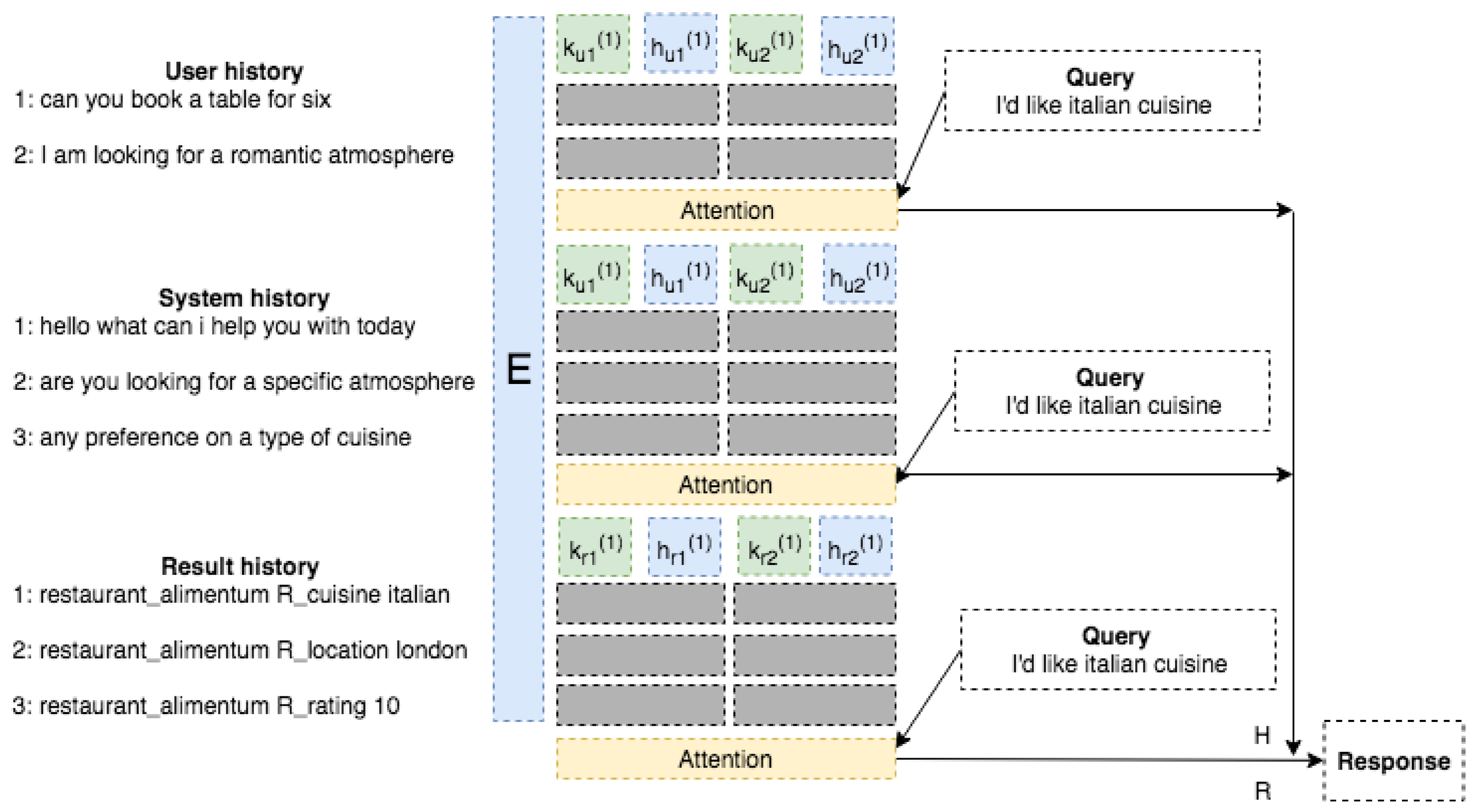


Figure: Schematic representation of SEntNet architecture with separate source-specific memory modules.

SEntNet's functions depend on three modules described below.

- **Input module**. The embedding of the i-th utterance $e_{i(\mathcal{S})}$ for source $\mathcal{S}$ is:

$$e_{i(\mathcal{S})} = \Sigma_x f_x \odot w_x^i + l_x^i \in \mathbb{R}^d. \quad (2)$$

- **Dynamic memory module**. For the i-th utterance from $\mathcal{S}$ in the dialogue, the memory block for the j-th entity is updated as:

$$g_{j(\mathcal{S})}^i = \sigma(e_{i(\mathcal{S})}^T h_{j(\mathcal{S})}^{i-1} + e_{i(\mathcal{S})}^T k_{j(\mathcal{S})}^{i-1}) \in \mathbb{R}^d \quad (3)$$
$$\tilde{h}_{j(\mathcal{S})}^i = \phi(G_{\mathcal{S}} h_{j(\mathcal{S})}^{i-1} + V_{\mathcal{S}} k_{j(\mathcal{S})}^{i-1} + W_{\mathcal{S}} e_{i(\mathcal{S})}) \in \mathbb{R}^d \quad (4)$$
$$h_{j(\mathcal{S})}^i = \frac{h_{j(\mathcal{S})}^{i-1} + g_{j(\mathcal{S})}^i \odot \tilde{h}_{j(\mathcal{S})}^i}{\| h_{j(\mathcal{S})}^{i-1} + g_{j(\mathcal{S})}^i \odot \tilde{h}_{j(\mathcal{S})}^i \|} \in \mathbb{R}^d \quad (5)$$
$$h_{j(\mathcal{S})} = h_{j(\mathcal{S})}^1 \oplus h_{j(\mathcal{S})}^2 \oplus \cdots \oplus h_{j(\mathcal{S})}^n. \quad (6)$$

- **Output module**. Let $q \in \mathbb{R}^d$ be the embedding of the user utterance u_t for the current turn t. The output module is defined as:

$$p_{j(\mathcal{S})} = \text{softmax}(q^T h_{j(\mathcal{S})}) \quad (7)$$
$$z_{\mathcal{S}} = \Sigma_j h_{j(\mathcal{S})} p_{j(\mathcal{S})} \in \mathbb{R}^d \quad (8)$$
$$z = z_{\mathcal{S}_{\mathbb{U}}} \oplus z_{\mathcal{S}_{\mathbb{S}}} \oplus z_{\mathcal{S}_{\mathbb{B}}} \in \mathbb{R}^{3d} \quad (9)$$
$$\tilde{y} = L\phi(q + Hz) \in \mathbb{R}^r \quad (10)$$
$$y = \text{softmax}(\tilde{y}_j). \quad (11)$$

Experimental Setup

Research questions

RQ1: How well does SEntNet predict appropriate responses?
RQ2: How do different embeddings affect SEntNet's performance?
RQ3: How well does SEntNet perform in the case of limited data? And
RQ4: How does lexical diversity affect SEntNet's performance?

- **Datasets**. Dialog bAbI (Bordes&Weston,2017); DSTC2 (Henderson et al.,2014).
- **Evaluation**. Turn-level accuracy – the fraction of correct responses out of all.

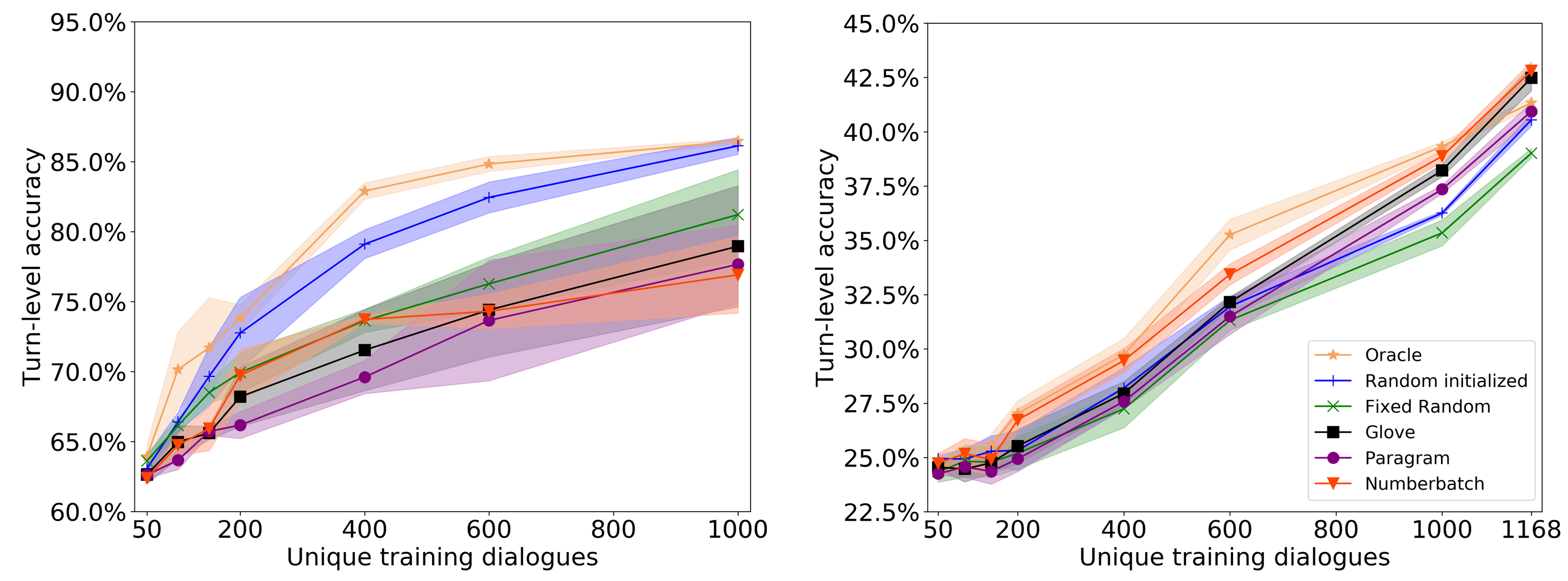
Results

Model	bAbI	DSTC2
TF-IDF	0.040	0.030
Q2A	0.570	0.220
EntNet	0.850	0.388
DQMemNN	0.863	–
HHCN	–	**0.661**
SEntNet	**0.910**	0.412

Table: Comparison with baselines on the bAbI and DSTC2 datasets (**RQ1**).

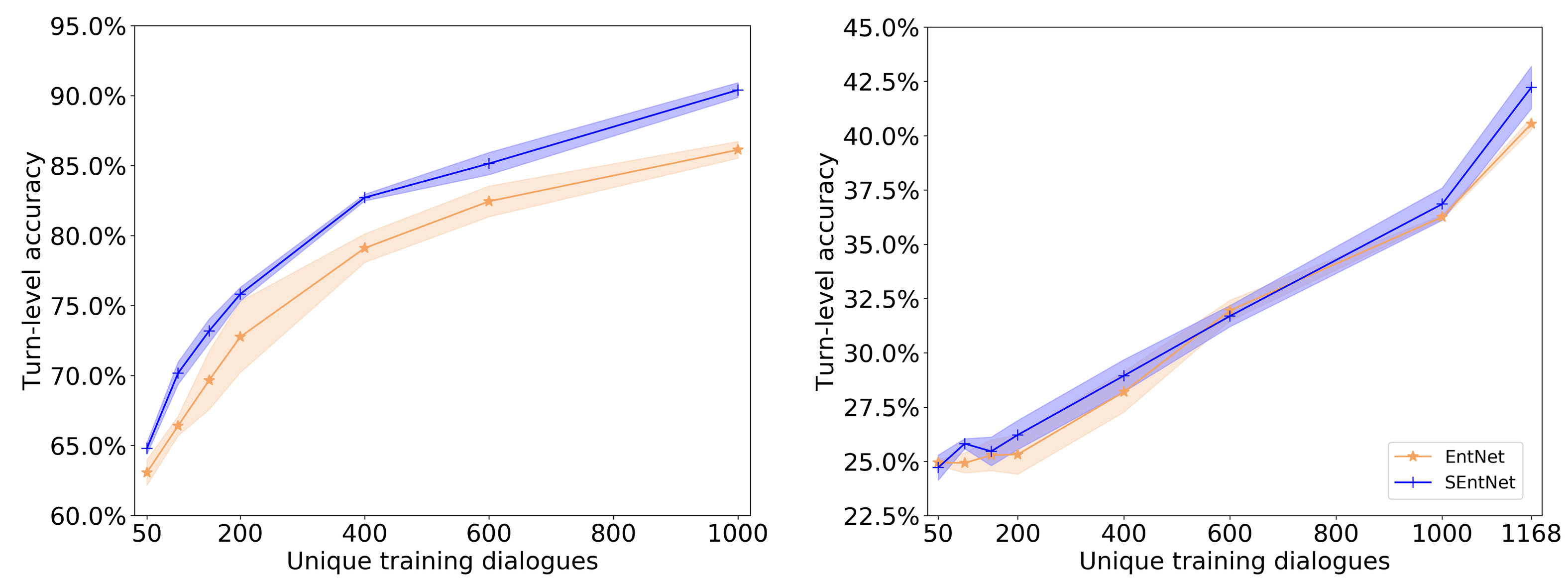
Model	bAbI	DSTC2
EntNet	0.850	0.388
EntNet+POS	0.850	0.398
SEntNet	**0.910**	**0.412**
SEntNet+POS	0.890	0.409

Table: The effect of lexical diversity on EntNet and SEntNet, on the bAbI and DSTC2 datasets (**RQ4**).



(a) bAbI (b) DSTC2

Figure: Turn-level accuracy of SEntNet for different embedding spaces on both datasets. (**RQ2**).



(a) bAbI (b) DSTC2

Figure: Turn-level accuracy of SEntNet on both datasets, when trained with different volumes of training dialogues (**RQ3**).

Conclusion

We propose **SEntNet**, a dialogue response selection model in memory network architecture:

- Select responses aware of source-specific history and consistently outperforms the baselines for end-to-end TDSs.
- Optimizing embeddings while training is useful for the performance.
- Tolerant of sparse data and able to handle different degrees of lexical diversity.
- Increase of learnable parameters by introducing extra memory modules can be addressed with parallel update mechanism design inherited from EntNet.

Acknowledgments. This research was partially supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the China Scholarship Council (CSC), and the Innovation Center for Artificial Intelligence (ICAI). We would like to thank Huawei, Microsoft, Naver and Google for their generous travel support.

University of Amsterdam

 j.pei@uva.nl