FusionNet

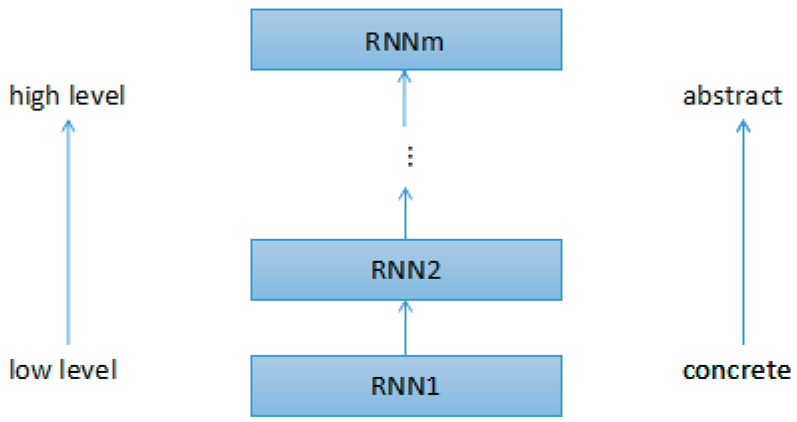
Motivation

1.The encoding for word of the whole history procedure is important.
2.The fully-aware multi-level attention fusion is important.

Contributions

This paper proposes a novel attention mechanism with following three contributions:
1.The concept of "history of words" to build the attention using complete information from lowest word-level embedding up to the highest semantic-level representation.
2.It proposes a novel attention scoring function.
3.It proposes a fully-aware multi-level fusion to exploit information layer by layer.

Multi-Level



Scoring Function

$$S_{ij} = f(U(\mathrm{HoW}_i^A))^T D\, f(U(\mathrm{HoW}_j^B))$$

Fully-Aware Fusion Network

Input Vectors

context: 921-dim {300-dim GloVe embedding, 600-dim contextualized vector, 12-dim POS embedding, 8-dim NER embedding, 1-dim normalized term frequency}
question: 900-dim {300-dim GloVe embedding, 600-dim conextualized vector}
$C : \{w_1^C, ..., w_m^C\} \in R^{900+20+1}$
$Q : \{w_1^Q, ..., w_n^Q\} \in R^{900}$

Fully-aware Multi-level Fusion: Word-level

$$\hat{\boldsymbol{g}}_i^C = \sum_j \alpha_{ij}\, \boldsymbol{g}_j^Q, \quad \alpha_{ij} \propto \exp(S(\boldsymbol{g}_i^C, \boldsymbol{g}_j^Q)), \quad S(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{ReLU}(W\boldsymbol{x})^T\, \mathrm{ReLU}(W\boldsymbol{y})$$

em_i is created for each word in C to indicate wheter the word occurs in the question Q.

$$\tilde{\boldsymbol{w}}_i^C = [\boldsymbol{w}_i^C; \mathrm{em}_i; \hat{\boldsymbol{g}}_i^C]$$

Reading

$$\boldsymbol{h}_1^{Cl}, \ldots, \boldsymbol{h}_m^{Cl} = \mathrm{BiLSTM}(\tilde{\boldsymbol{w}}_1^C, \ldots, \tilde{\boldsymbol{w}}_m^C), \quad \boldsymbol{h}_1^{Ql}, \ldots, \boldsymbol{h}_n^{Ql} = \mathrm{BiLSTM}(\boldsymbol{w}_1^Q, \ldots, \boldsymbol{w}_n^Q),$$

$$\boldsymbol{h}_1^{Ch}, \ldots, \boldsymbol{h}_m^{Ch} = \mathrm{BiLSTM}(\boldsymbol{h}_1^{Cl}, \ldots, \boldsymbol{h}_m^{Cl}), \quad \boldsymbol{h}_1^{Qh}, \ldots, \boldsymbol{h}_n^{Qh} = \mathrm{BiLSTM}(\boldsymbol{h}_1^{Ql}, \ldots, \boldsymbol{h}_n^{Ql}).$$

Hence low-level and high-level concept $\boldsymbol{h}^l, \boldsymbol{h}^h \in \mathbb{R}^{250}$ are created for each word.

Question Understanding

$$U_Q = \{\boldsymbol{u}_1^Q, \ldots, \boldsymbol{u}_n^Q\} = \text{BiLSTM}([\boldsymbol{h}_1^{Ql}; \boldsymbol{h}_1^{Qh}], \ldots, [\boldsymbol{h}_n^{Ql}; \boldsymbol{h}_n^{Qh}]).$$

where $\{\boldsymbol{u}_i^Q \in \mathbb{R}^{250}\}_{i=1}^n$ are the understanding vectors for $\boldsymbol{Q}$.

Fully-aware Multi-level Fusion: Higher-level

$$\text{HoW}_i^C = [\boldsymbol{g}_i^C; \boldsymbol{c}_i^C; \boldsymbol{h}_i^{Cl}; \boldsymbol{h}_i^{Ch}], \ \text{HoW}_i^Q = [\boldsymbol{g}_i^Q; \boldsymbol{c}_i^Q; \boldsymbol{h}_i^{Ql}; \boldsymbol{h}_i^{Qh}] \in \mathbb{R}^{1400},$$

1. Low-level fusion: $\hat{\boldsymbol{h}}_i^{Cl} = \sum_j \alpha_{ij}^l \boldsymbol{h}_j^{Ql}, \quad \alpha_{ij}^l \propto \exp(S^l(\text{HoW}_i^C, \text{HoW}_j^Q)).$
2. High-level fusion: $\hat{\boldsymbol{h}}_i^{Ch} = \sum_j \alpha_{ij}^h \boldsymbol{h}_j^{Qh}, \quad \alpha_{ij}^h \propto \exp(S^h(\text{HoW}_i^C, \text{HoW}_j^Q)).$
3. Understanding fusion: $\hat{\boldsymbol{u}}_i^C = \sum_j \alpha_{ij}^u \boldsymbol{u}_j^Q, \quad \alpha_{ij}^u \propto \exp(S^u(\text{HoW}_i^C, \text{HoW}_j^Q)).$

$$V_C = \{\boldsymbol{v}_1^C, \ldots, \boldsymbol{v}_m^C\} = \text{BiLSTM}([\boldsymbol{h}_1^{Cl}; \boldsymbol{h}_1^{Ch}; \hat{\boldsymbol{h}}_1^{Cl}; \hat{\boldsymbol{h}}_1^{Ch}; \hat{\boldsymbol{u}}_1^C], \ldots, [\boldsymbol{h}_m^{Cl}; \boldsymbol{h}_m^{Ch}; \hat{\boldsymbol{h}}_m^{Cl}; \hat{\boldsymbol{h}}_m^{Ch}; \hat{\boldsymbol{u}}_m^C]).$$

Fully-aware Self-boosted Fusion

$$\mathrm{HoW}_i^C = [\boldsymbol{g}_i^C; \boldsymbol{c}_i^C; \boldsymbol{h}_i^{Cl}; \boldsymbol{h}_i^{Ch}; \hat{\boldsymbol{h}}_i^{Cl}; \hat{\boldsymbol{h}}_i^{Ch}; \hat{\boldsymbol{u}}_i^C; \boldsymbol{v}_i^C] \in \mathbb{R}^{2400}.$$

Then perform fully-aware attention,

$$\hat{\boldsymbol{v}}_i^C = \sum_j \alpha_{ij}^s \boldsymbol{v}_j^C, \;\; \alpha_{ij}^s \propto \exp(S^s(\mathrm{HoW}_i^C, \mathrm{HoW}_j^C)).$$

The final context representation is

$$U_C = \{\boldsymbol{u}_1^C, \ldots, \boldsymbol{u}_m^C\} = \mathrm{BiLSTM}([\boldsymbol{v}_1^C; \hat{\boldsymbol{v}}_1^C], \ldots, [\boldsymbol{v}_m^C; \hat{\boldsymbol{v}}_m^C]).$$

Output

Through the above operations, we can get

$U_C = u_1^C, ..., u_{m'}^C$

$U_Q = u_1^Q, ..., u_n^Q$

Then we get the vector representation of Q

$u^Q = \sum_i \beta_i u_i^Q$

For start,

$$P_i^S \propto \exp((\boldsymbol{u}^Q)^T W_S \boldsymbol{u}_i^C),$$

For end,

$$\boldsymbol{v}^Q = \mathrm{GRU}(\boldsymbol{u}^Q, \textstyle\sum_i P_i^S \boldsymbol{u}_i^C)$$

$$P_i^E \propto \exp((\boldsymbol{v}^Q)^T W_E \boldsymbol{u}_i^C)$$

Experiments

Attention Function	**EM / F1**
Additive (MLP)	71.8 / 80.1
Multiplicative	72.1 / 80.6
Scaled Multiplicative	72.4 / 80.7
Scaled Multiplicative + ReLU	72.6 / 80.8
Symmetric Form	73.1 / 81.5
Symmetric Form + ReLU	**75.3 / 83.6**

Configuration		**Dev EM / F1**
C, Q Fusion	**Self C**	
High-level	None	64.6 / 73.2
FA High-level	None	73.3 / 81.4
FA All-level	None	72.3 / 80.7
FA Multi-level	None	74.6 / 82.7
FA Multi-level	Normal	74.4 / 82.6
FA Multi-level	FA	**75.3 / 83.6**